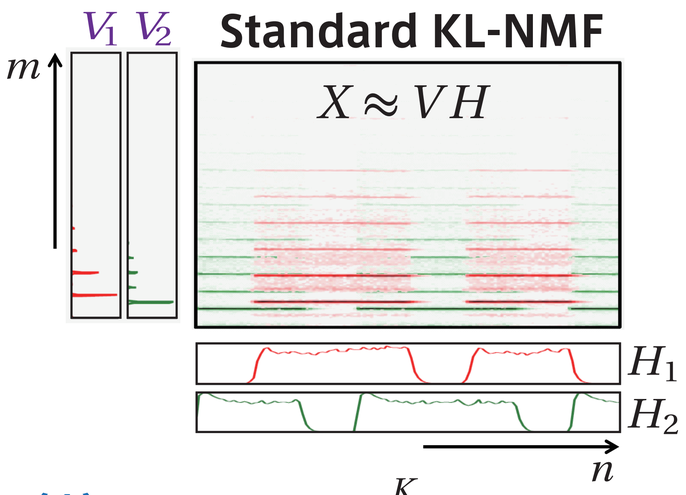
Infinite Composite Autoregressive Models for Music Signal Analysis

Don't be scared of Bayesian!

Kazuyoshi Yoshii and Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST)

Objectives: Multiple F0 estimation and timbre-based source separation

We aim to overcome three fundamental limitations of the standard NMF



(1) A large number of unconstrained spectral bases are needed to fully represent the timbral variations of instrument sounds

Idea: Factorizing spectral bases as the products of sources and filters

(2) An independent post-processing step is needed to determine the existence of a F0 and estimate its value from each basis

Idea: Parametrizing F0s and harmonic structures of sources

(3) The number of spectral bases should be specified in advance

Idea: Using Bayesian nonparametrics for sparse learning in an infinite space

Yasuraoka MUS 2012 · Kameoka ISCAS 2009 · Hennequin DAFX 2010 · Hoffman ICML 2010

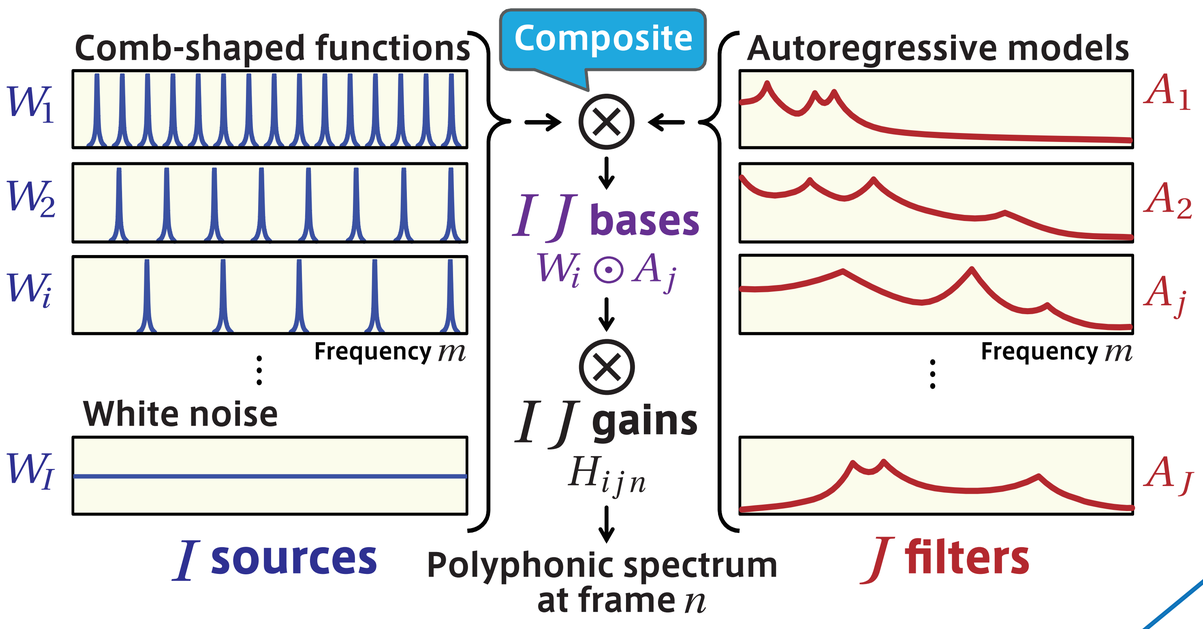
We integrate all these techniques into a unified Bayesian model

We propose an ultimate probabilistic framework for joint estimation of F0s, timbres, the number of F0s, and the number of timbres (Point!)

(1)
$$|X_{mn}| \approx \sum_{k=1}^{K} \underbrace{V_{km}}_{\text{Basis}} \underbrace{H_{kn}}_{\text{Gain}}$$

Approach: Nonparametric Bayesian formulation of source-filter NMF

We design prior distributions and parametric functions for individual factors



(3) **We take the limit as I and J diverge to infinity**

$$|X_{mn}| \text{ or } |X_{mn}|^2 \approx \sum_{i,j=1}^{I,J\to\infty} \theta_i \phi_j \underbrace{W_{im} A_{jm}}_{\text{Source Filter}} H_{ijn} = Y_{mn}$$

Amplitude / Power · Basis · Gain · Reconst. (model)

Glogal weight of source i · Glogal weight of filter j

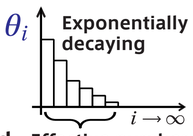
Two variants of likelihood functions for X

$$\begin{cases} |X_{mn}| \sim \text{Poisson}(Y_{mn}) \\ |X_{mn}|^2 \sim \text{Exponential}(Y_{mn}) \end{cases} \text{ equiv. to } \begin{cases} \min \text{KL}\left(|X_{mn}| \middle| Y_{mn}\right) \\ \min \text{IS}\left(|X_{mn}|^2 \middle| Y_{mn}\right) \end{cases}$$

Theoretically justified to some degree

Gamma process priors on θ, ϕ

$$\theta_i \sim \text{Gamma}(\alpha/I, \alpha)$$
$$\phi_j \sim \text{Gamma}(\gamma/J, \gamma)$$



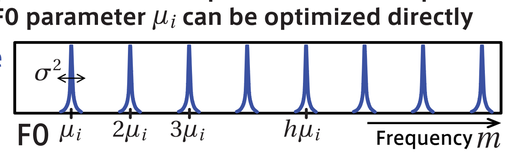
Larger α means heavier-tailed
α is insensitive to results

As the truncation level I diverges to infinity, the vector θ approximates an infinite sequence drawn from a GaP with shape parameter α

Comb-shaped functions for W

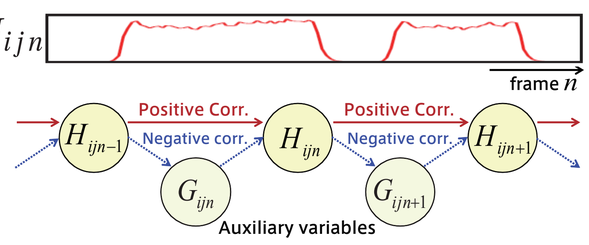
$$W_{im} = \sum_{h=1}^{H} \exp\left(-\frac{(m - h\mu_i)^2}{2\sigma^2}\right) \quad (2)$$

Sharp Gaussians correspond to harmonic partials
The F0 parameter μ_i can be optimized directly



Gamma chain priors on H

$$G_{ijn} \sim \text{Gamma}\left(\beta, \beta H_{ijn-1}\right)$$
$$H_{ijn} \sim \text{Gamma}\left(\beta, \beta G_{ijn}\right)$$

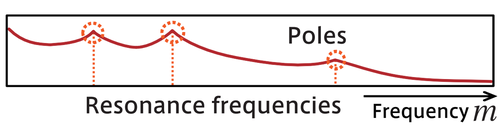


All-pole transfer functions for A

$$\begin{cases} A^{\text{KL}}_{jm} = \sqrt{A^{\text{IS}}_{jm}} \\ A^{\text{IS}}_{jm} = \dfrac{1}{\left|\sum_{p=0}^{P} a^j_p e^{-2\pi \frac{m}{2M} pi}\right|^2} \end{cases}$$

$\{a^j_0, a^j_1, \cdots, a^j_P\}$: Linear predictive coefficients

There are $P/2$ m's corresponding to poles

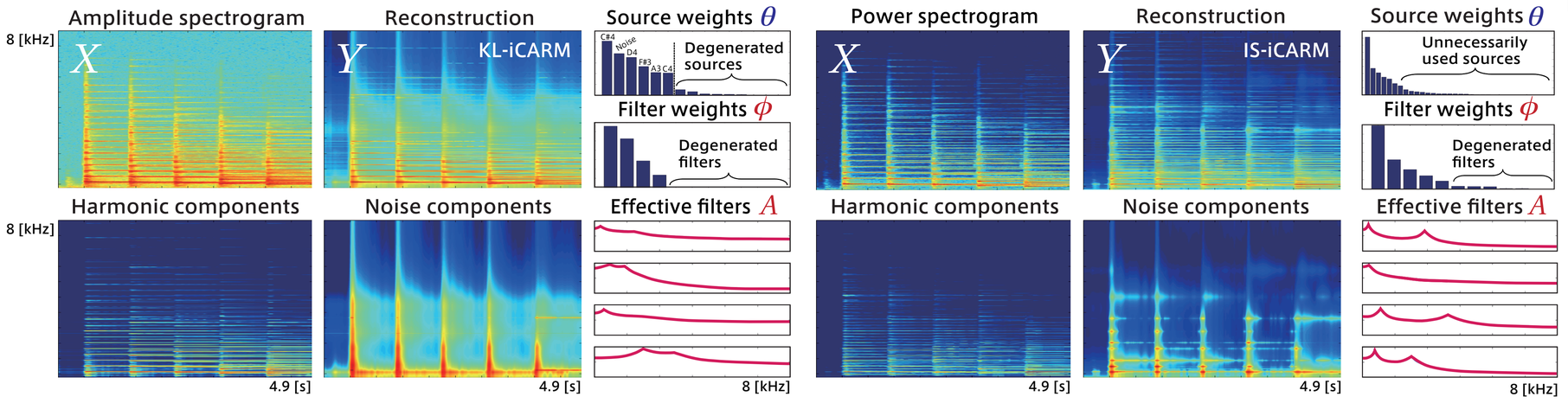


Our models are optimized by combination of variational Bayes (VB) and multiplicative update (MU)

Experiments: Source-filter factorization of piano and popular music

Multiple F0 estimation for MAPS Piano Database (analysis of sources)

Frame-level F-measure for 30 pieces:
48.4% (KL-iCARM) 35.1% (IS-iCARM)



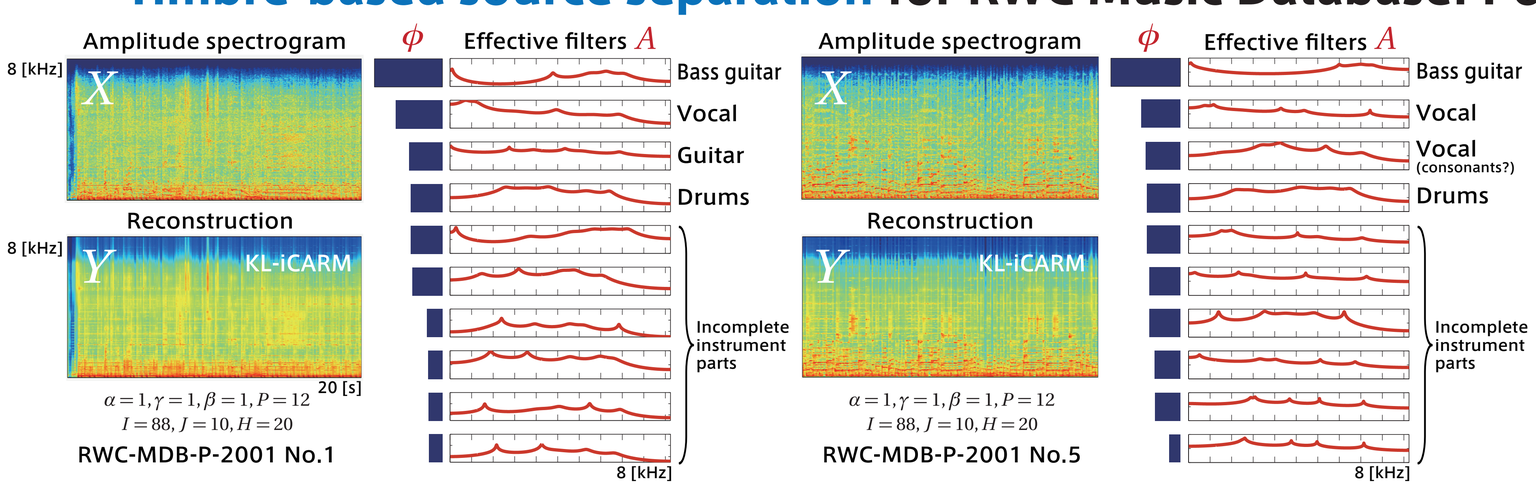
The KL-iCARM is capable of discovering the reasonable numbers of sources and filters in a data-driven manner

- Harmonic and noise components are associated with different sources

The IS-iCARM tends to overestimate the numbers of sources and filters

- The reconstructed power is allowed to exceed the observed power with smaller penelty

Timbre-based source separation for RWC Music Database: Popular Music (analysis of filters)



In many songs, the most significant filter corresponds to the bass guitar and the second one to the vocal

Percussive sounds (bass drums, snare drums, and hi-hats) are modeled by several filters

Future work:

- Formulate iCARMs on the logarithmic frequency domain to deal with wavelet spectrograms
- Integrate language models with acoustic models to directly estimate musical notes for transcription