Unsupervised Music Understanding based on Nonparametric Bayesian Models

Kazuyoshi Yoshii

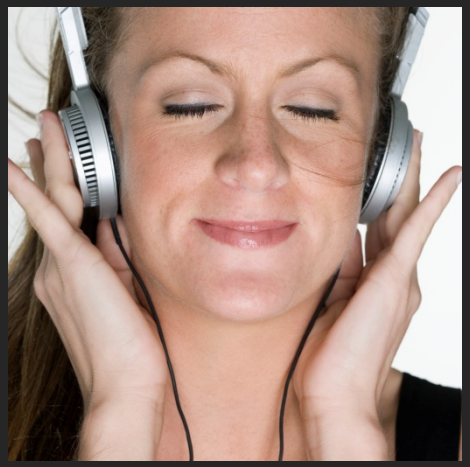
Masataka Goto

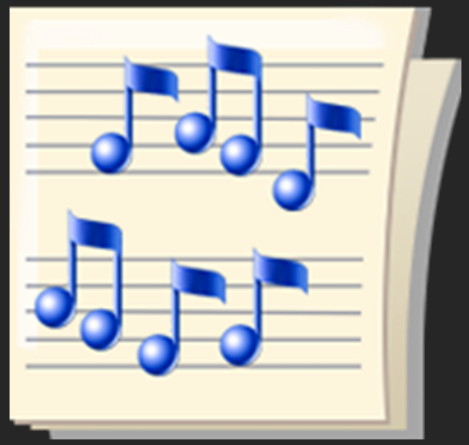
AIST, Japan

Why can we recognize music as music?

Is this because we are taught music theory?

Definitely No!



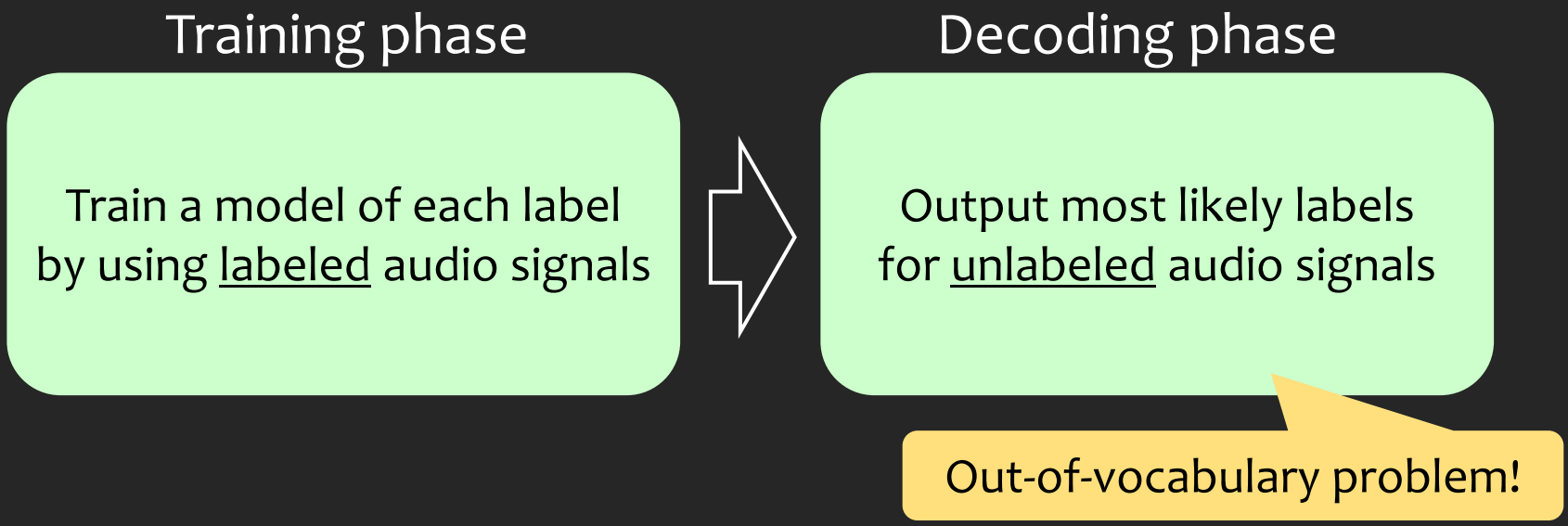


Why "Unsupervised Understanding"?

- Even musically-untrained people can intuitively understand and enjoy music
  - Examples:
    - People can notice that multiple sounds (musical notes) having different F0s are contained in music
      - Even if they do not know the number of notes in advance
    - People can distinguish whether given sound mixtures (chords) are harmonic or inharmonic
      - Even if they are not taught labels (chord names)
  - Structural patterns over musical notes can be discovered by listening to a large amount of music
    - Simultaneous and temporal patterns (chords and progressions)

Supervised Approach

- The most popular approach in previous studies
  - Examples:
    - Music transcription
      - The number of musical notes is given in advance
    - Chord recognition
      - A vocabulary of chord labels (e.g., maj, min, dim, aug) is defined



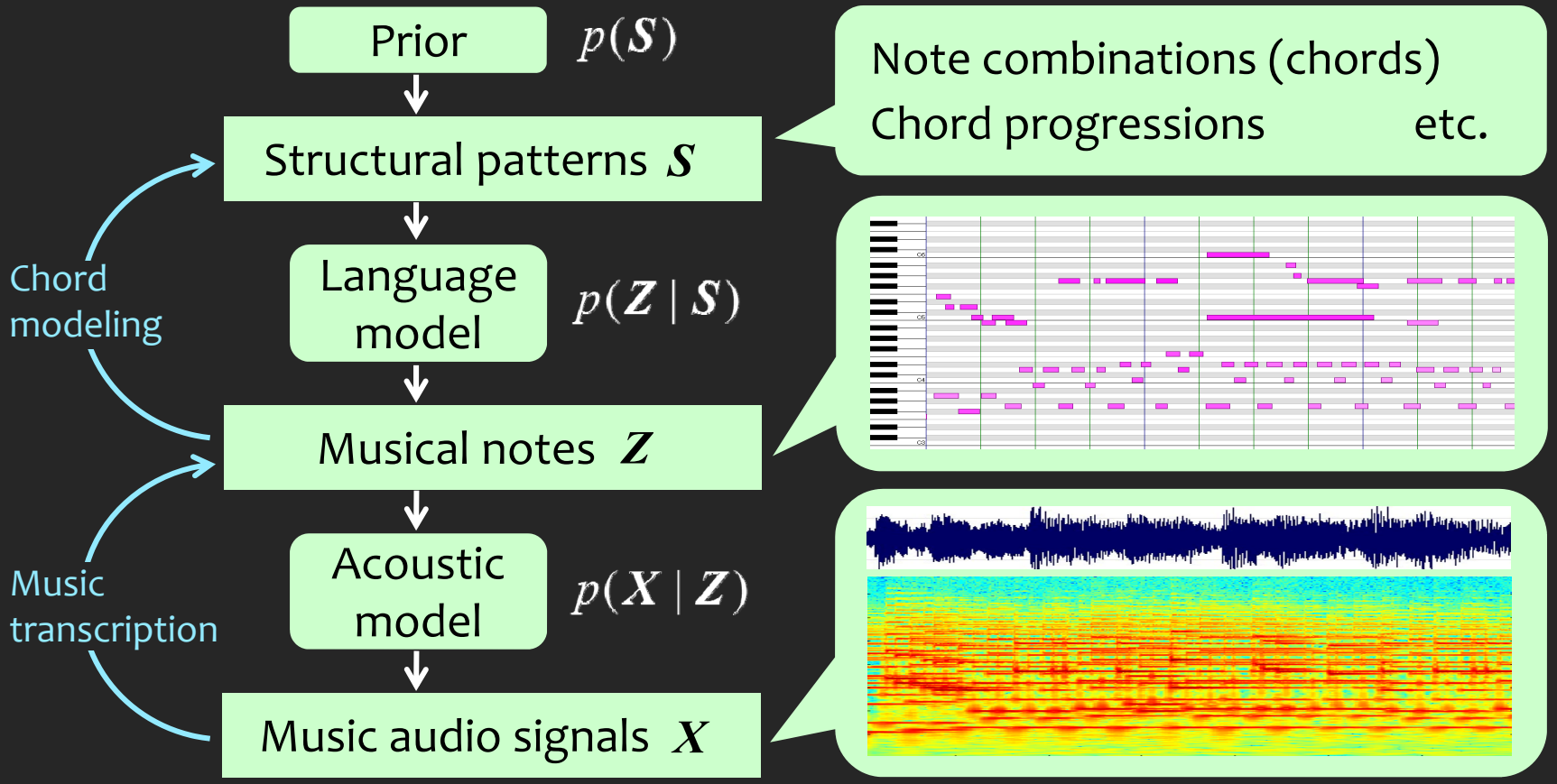
Unsupervised Approach

- Our goal is to find musical notes and discover their latent structural patterns at the same time
  - No finite vocabularies are defined
    - The number of musical notes is not given
    - Any chord labels (e.g., maj, min, dim, aug) are not given
  - Only polyphonic audio signals are available
    - We aim to find an appropriate number of musical notes
    - We aim to form chords directly from musical notes
      - No out-of-vocabulary problem!
  - "Understanding" is to infer the most likely hierarchical organization of music audio signals
    - A probabilistic framework is promising

chords and chord progressions

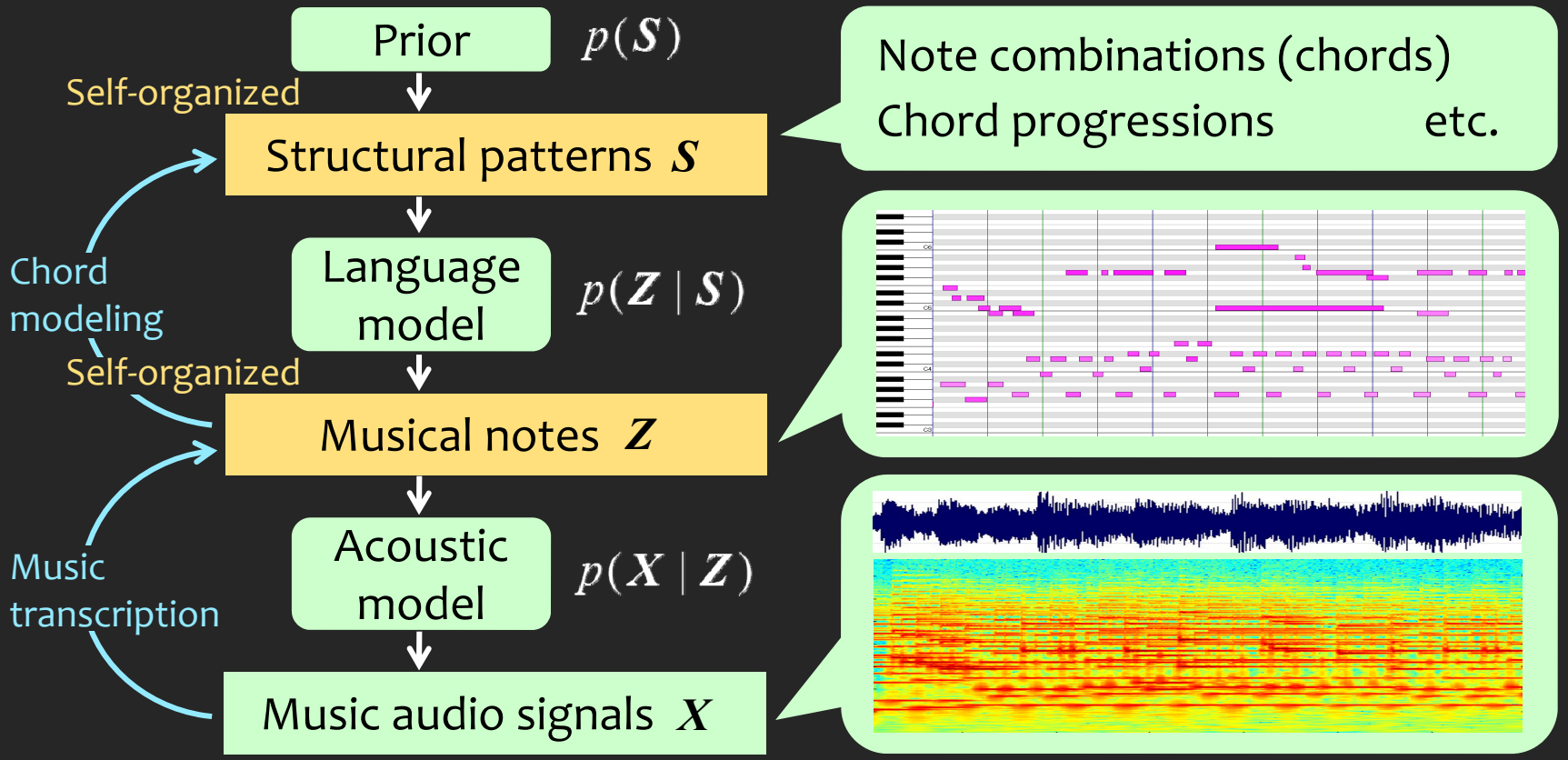
The Big Picture

- Integration of language and acoustic models
  - Hierarchical Bayesian formulation



A Key Feature

- Joint unsupervised learning of both models
  - Find the most likely latent structures in a data-driven way
    - What we usually call "chords" could be statistically defined as typical note combinations by maximizing the evidence p(X)



Model Selection

- How to determine appropriate numbers of notes and chords so that p(X) is maximized?
  - Naïve combinatorial complexity control (grid search) is computationally prohibitive!

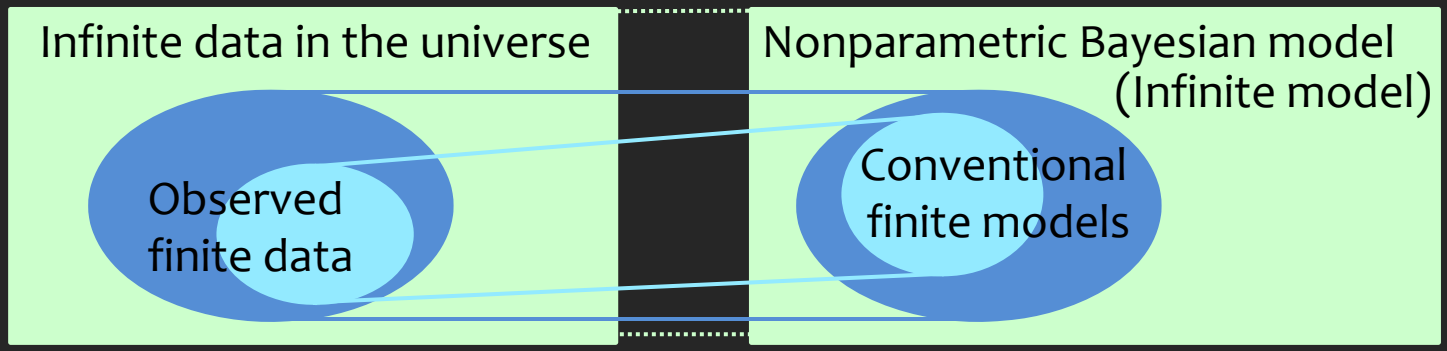
The number of musical notes

The number of types of chords	10	20	30	40	…
10	-20,000	-9,000	-8,000	-8,000	
20	-10,000	-8,500	**-7,000**	-7,500	
30	-9,000	-8,300	-7,500	-7,900	
40	-8,800	-8,400	-8,000	-8,500	
…					

Log-evidence is maximized

Why Nonparametric Bayes?

- Principled approach to structure learning
  - "L0-regularized" sparse learning in an infinite space
    - Infinite types of musical units (e.g., notes & chords) would be required to represent the variety of the whole music data in the universe
    - Only limited types of musical units are actually instantiated for explaining available finite data
  - No model selection
    - Effective model complexities (the numbers of musical units required) can be inferred at the same time



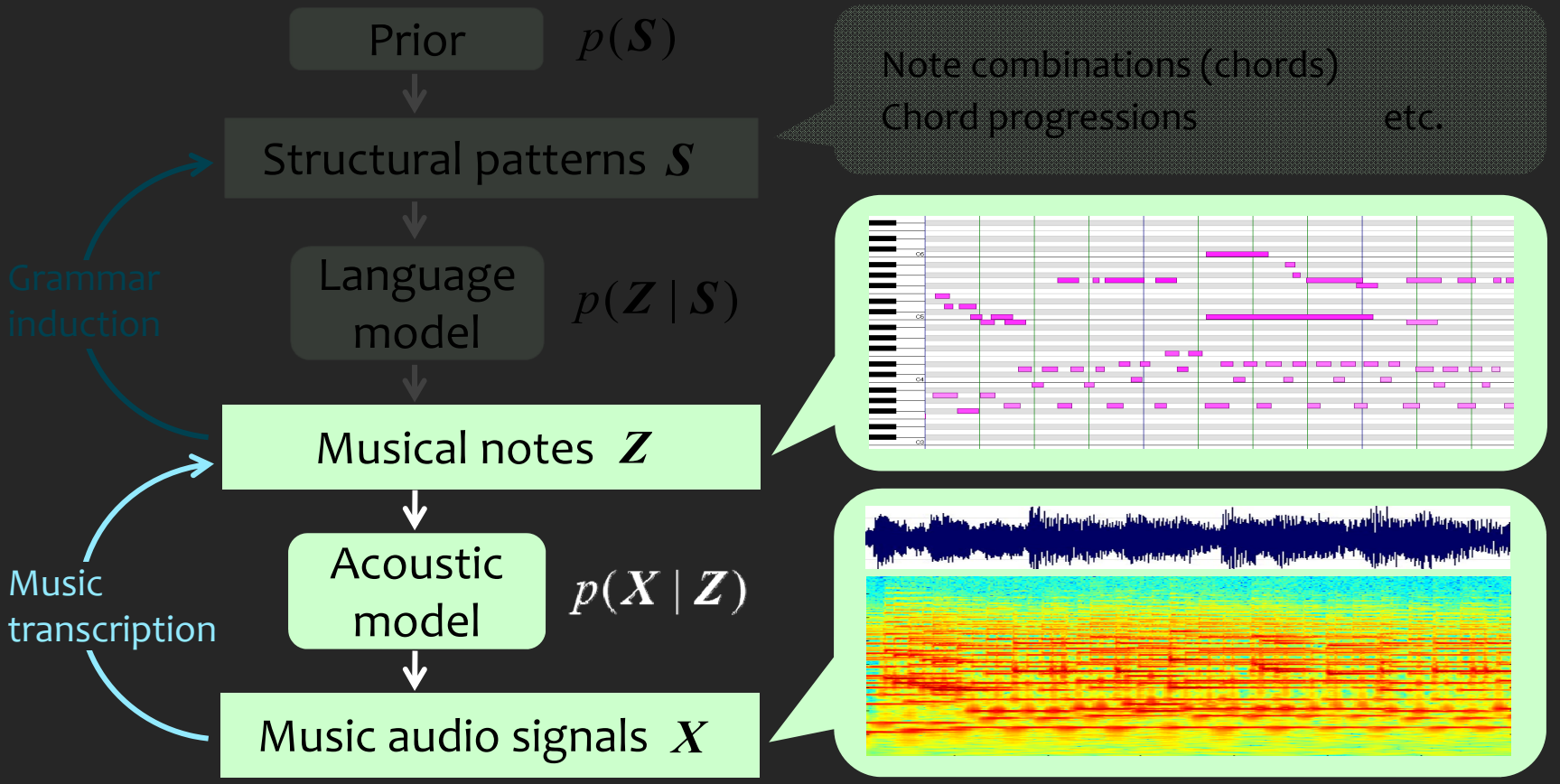
Latest Achievements

- We have developed nonparametric Bayesian acoustic and language models
  - Multipitch analysis for music audio signals
    - Infinite latent harmonic allocation [Yoshii ISMIR2010]
      - Infinite number of musical notes allowed
  - Chord progression modeling for musical notes
    - Vocabulary-free infinity-gram model [Yoshii ISMIR2011]
      - Infinite kinds of note combinations allowed

Language \ Acoustic	Mixture model (e.g., PLCA)	Factorial model (e.g., NMF)
Chain-structured model (e.g., n-gram model)	Yoshii ISMIR2010 Yoshii ISMIR2011	?
Tree-structured model (e.g., PCFG)	?	Nakano WASPAA2011 Nakano ICASSP2012

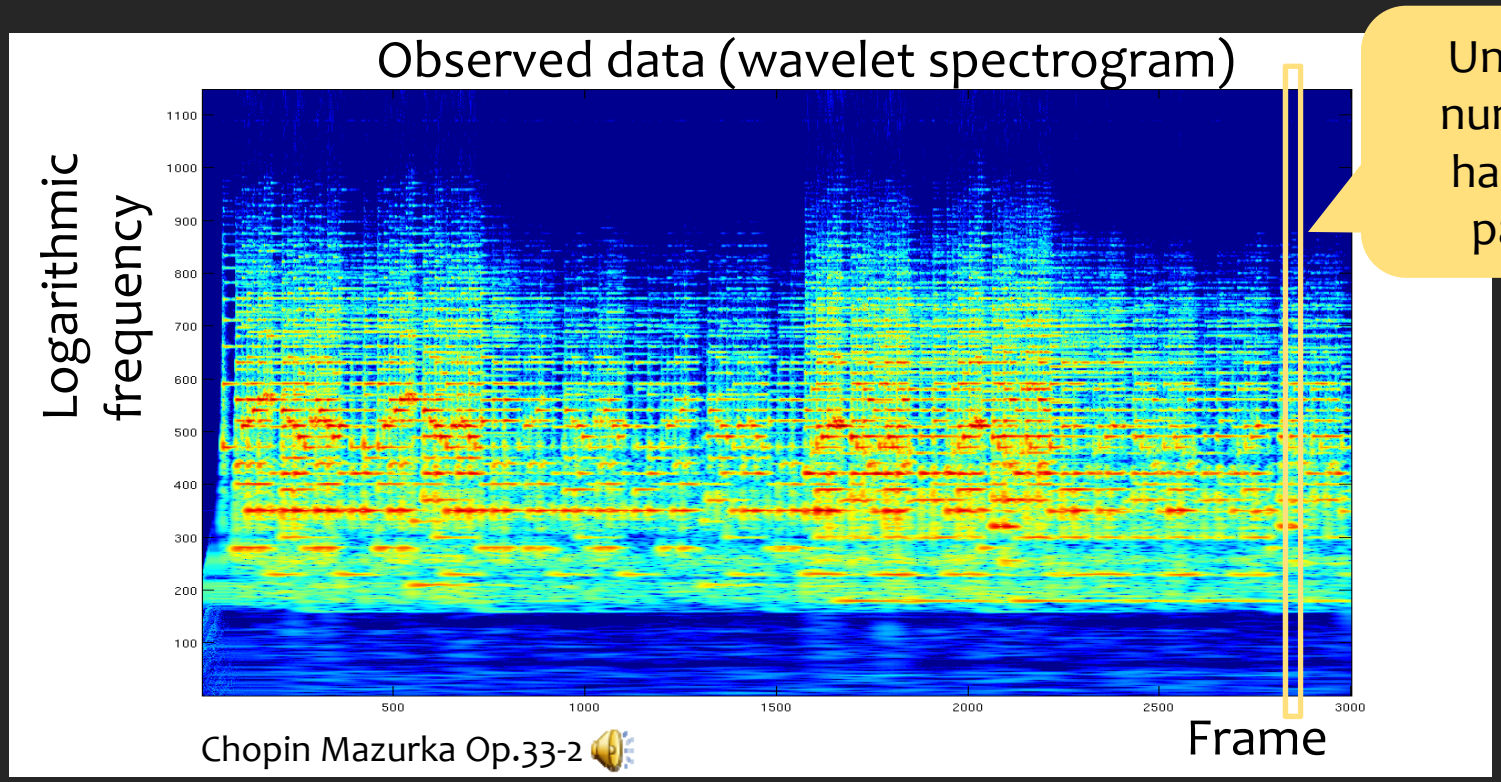
The Big Picture

- Integration of language and acoustic models
  - Hierarchical Bayesian formulation



Nonparametric Bayesian Acoustic Modeling

- Objective: Multipitch analysis
  - Detect multiple fundamental frequencies (F0s) at each frame from polyphonic audio signals
    - Unknown number of musical notes



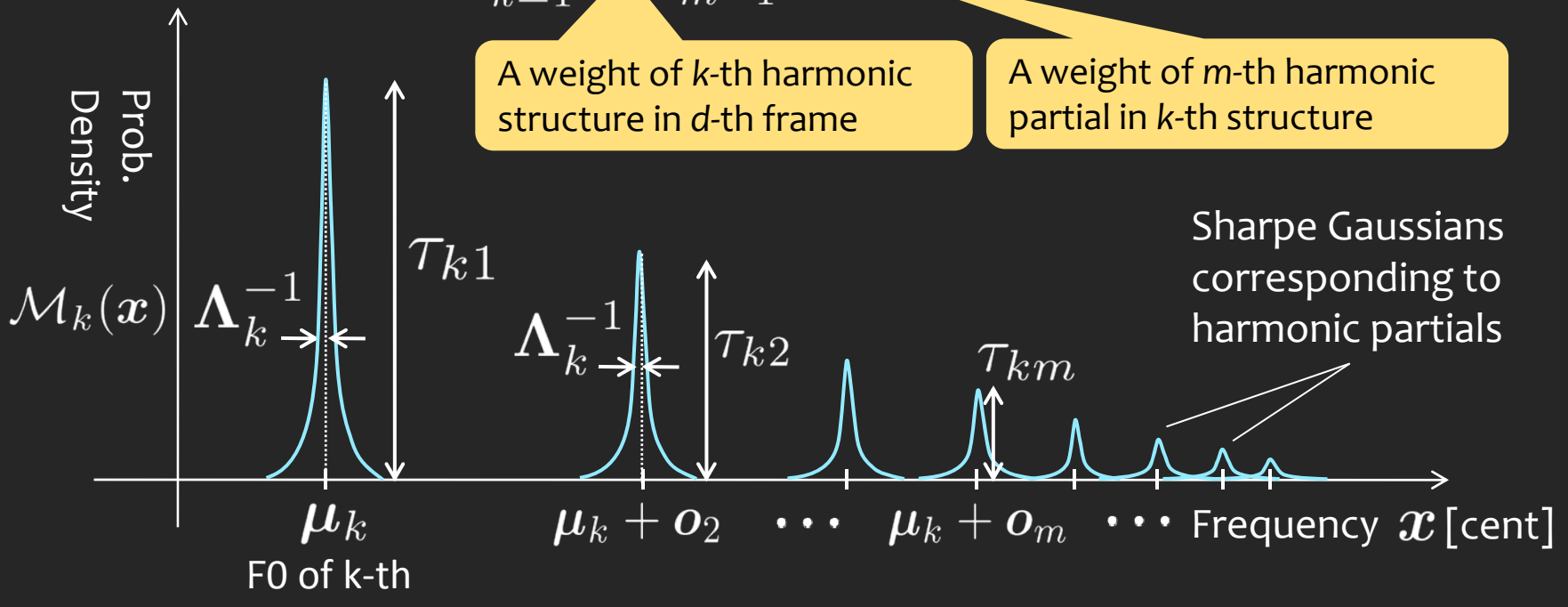
Conventional Finite Modeling

- A nested Gaussian mixture model [Goto 1999]
  - A spectrum at each frame is assumed to consist of K harmonic structures (GMMs)
  - Each harmonic structure is assumed to contain M harmonic partials (Gaussians)

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{dk} \sum_{m=1}^{M} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)$$

A weight of *k*-th harmonic structure in *d*-th frame

A weight of *m*-th harmonic partial in *k*-th structure



Conventional Model Selection

- We need to model polyphonic mixtures
  - How many numbers of musical notes (*K*)?
- We need to model harmonic structures
  - How many numbers of harmonic partials (*M*)?

The number of musical notes

The number of harmonic partials	10	20	30	40	…
5	-20,000	-9,000	-8,000	-8,000	
10	-10,000	-8,500	**-7,000**	-7,500	
15	-9,000	-8,300	-7,500	-7,900	
20	-8,800	-8,400	-8,000	-8,500	
…					

Computationally prohibitive!

Taking the Infinite Limit

- Infinite latent harmonic allocation (iLHA) [Yoshii ISMIR2010]
  - We do not need to specify model complexities
    - Unknown number of musical notes (K)
    - Unknown number of harmonic partials (M)

Finite nested GMM

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{dk} \sum_{m=1}^{M} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)$$

⇩ Let K & M diverge to infinity

Infinite nested GMM

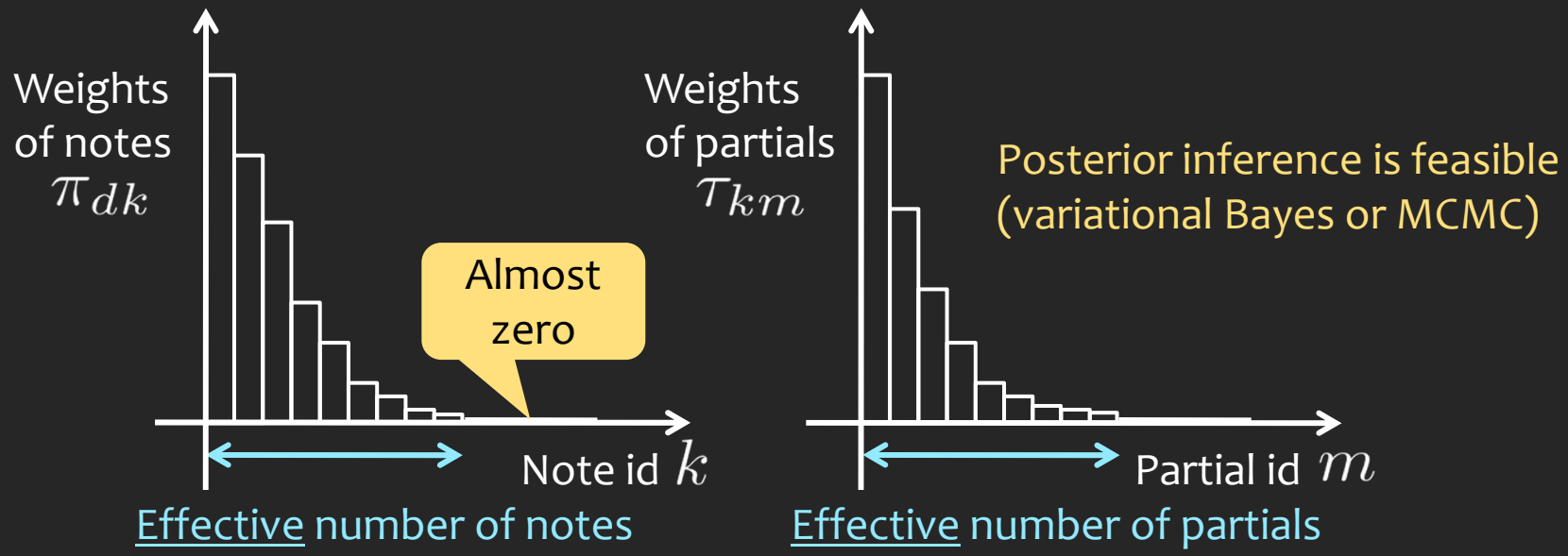
$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_{dk} \sum_{m=1}^{\infty} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)$$

Only limited numbers of musical notes and harmonic partials are actually contained in a finite amount of observed data

Sparse Learning

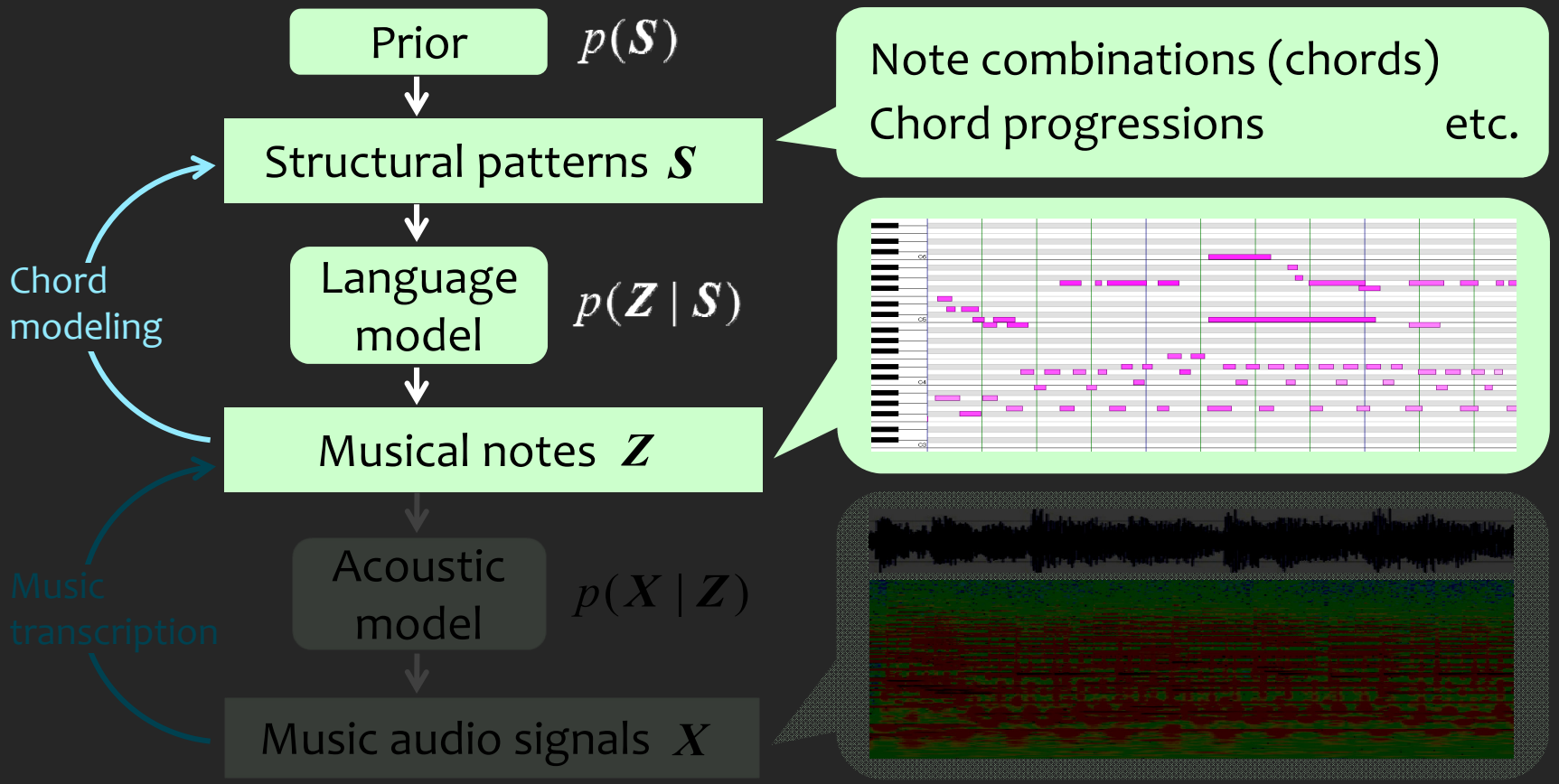
- Incorporate Dirichlet process (DP) prior
  - An infinite number of exponentially-decayed mixture weights can be stochastically generated
    - All weights sum to unity
    - Almost all weights are very close to zero

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_{dk} \sum_{m=1}^{\infty} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)$$



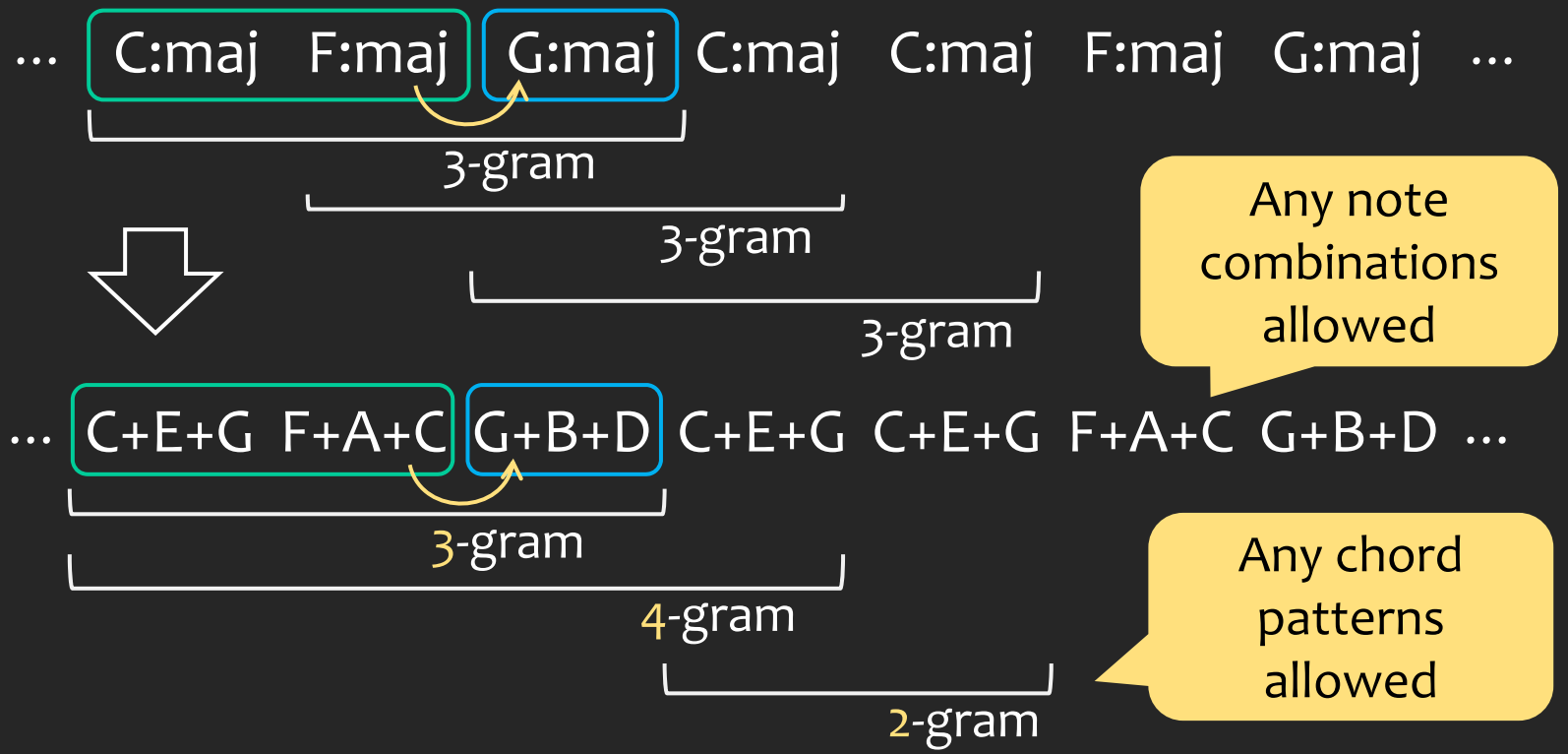
The Big Picture

- Integration of language and acoustic models
  - Hierarchical Bayesian formulation



Nonparametric Bayesian Language Modeling

- Objective: Chord progression modeling
  - Learn an *n*-gram model directly from musical notes
    - Without using a vocabulary of conventional chord labels
    - Without specifying the value of *n* (a vocab. of chord patterns)



Conventional Model Selection

- We need to formulate a variable-order model
  - How to determine a context length (*n*) for each chord?

… C:maj F:maj G:maj C:maj C:maj F:maj G:maj …

The value of *n*

…					
C:maj	1	2	3	4	
F:maj	1	2	3	4	
G:maj	1	2	3	4	
C:maj	1	2	3	4	…
F:maj	1	2	3	4	…
G:maj	1	2	3	4	…

Testing all combinations is infeasible!

Is 3-gram always the best?

Taking the Infinite Limit

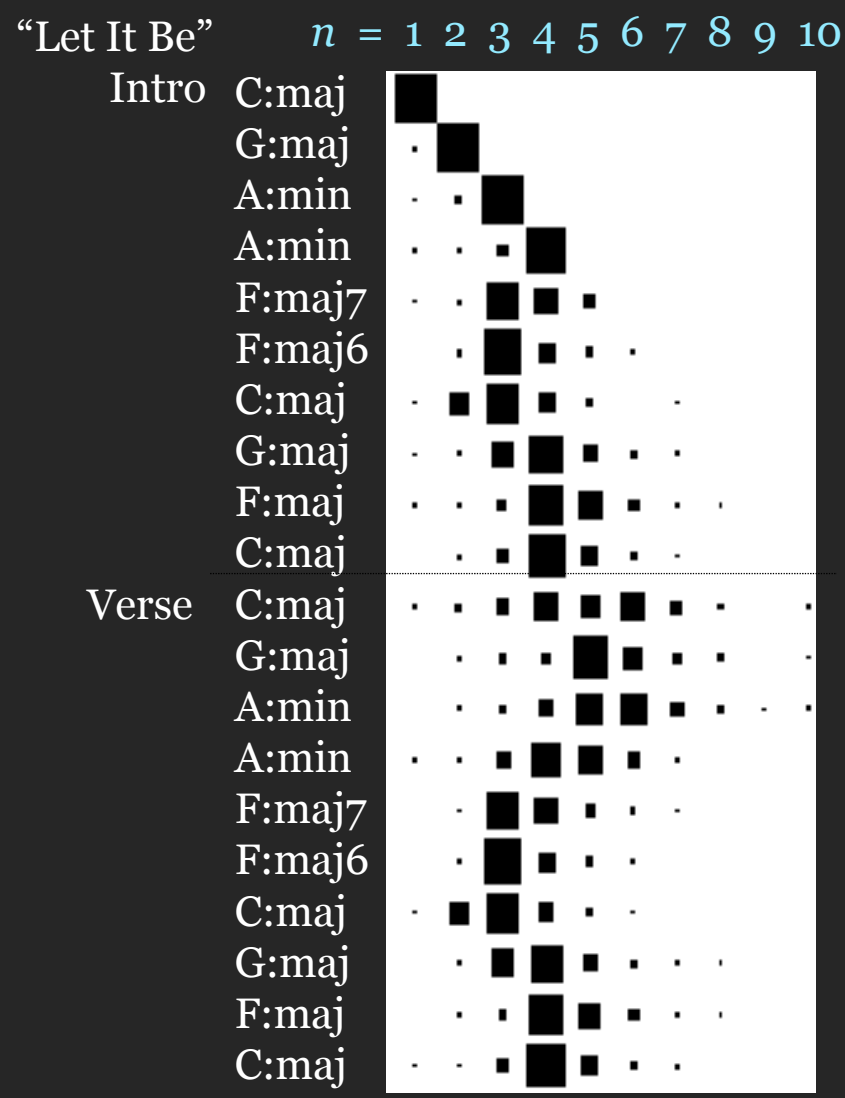
- Vocabulary-free infinity-gram model [Yoshii ISMIR2011]
  - Hierarchical Bayesian smoothing
    - Based on generative model of *n*-grams [Teh 2006]
    - Pitman-Yor process (PY)
  - Diverge the value of *n* to Infinity
    - All possibilities of *n* are considered [mochihashi 2007]
    - Dirichlet process (DP)
  - Allow any note combinations to form "chords"
    - No out-of-vocabulary problem!
      - C + E + G is a chord
      - C + D + E is another chord (no corresponding conventional label)

… C+E+G F+A+C G+B+D C+E+G C+E+G F+A+C G+B+D …

We consider a generative model of note combinations and integrate it to the n-gram model

Inference Results

- Discover chord patterns from Beatles songs



(represented by conventional chord labels for readability)

Probability	n	Stochastically-coherent chord patterns (in C major scale)
0.701	3	C:7 F:7 C:7
0.682	3	B:maj F:maj G:maj
0.656	3	A:min C:7 F:maj
0.647	3	F:min G:maj C:maj
0.645	4	F:maj F:maj G:maj C:maj
0.632	3	E:min C:7 F:maj
0.630	3	C:maj7 D:min7 E:min7
0.623	4	B:maj F:maj G:maj C:maj
0.622	3	D:min7 G:sus4 G:maj
0.620	5	D:min G:maj C:maj F:maj C:maj

Conclusion

- Unsupervised music understanding
  - The new research framework proposed
    - Integration of language and acoustic models
    - Hierarchical nonparametric Bayesian modeling
  - Current progress:
    - Nonparametric Bayesian acoustic modeling
      - Infinite latent harmonic allocation [Yoshii ISMIR2010]
    - Nonparametric Bayesian language modeling
      - Vocabulary-free infinity-gram model [Yoshii ISMIR2011]
  - Remaining issues:
    - Joint learning of both models
    - Bridging the gap between multipitch analysis and music transcription