Probabilistic Modeling of Speech

Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab

Department of Electronic Engineering, Tsinghua University, Beijing, China

Now Visiting Scholar at Beckman Institute, UIUC

2015-3-11, Beckman, UIUC

What is this talk about?

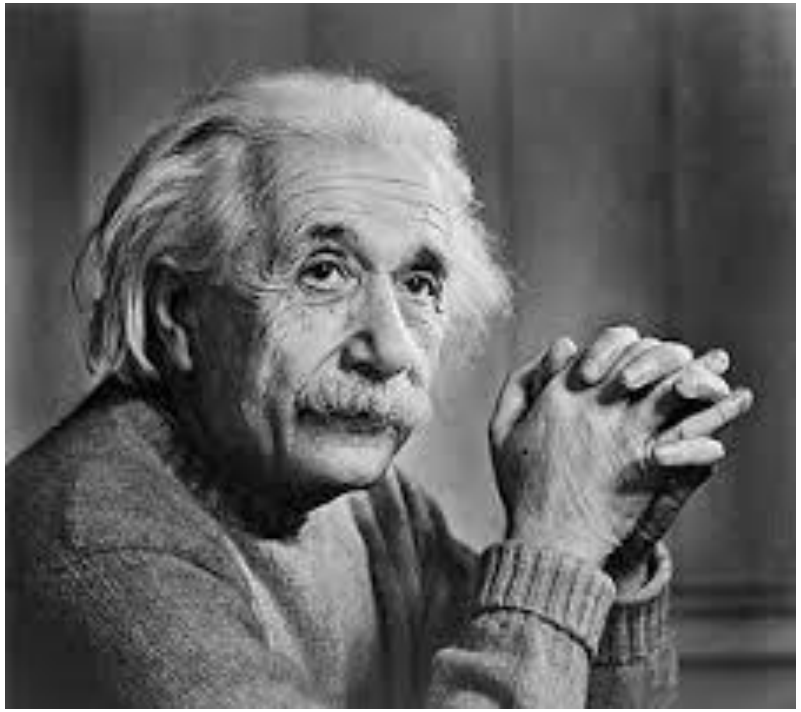
- Brief introduction to SPMI lab

- Motivation
- Bayesian HMM modeling of speech, ICASSP 2007.
- Variational nonparametric Bayesian HMM, ICASSP 2010.
- NMF modeling of voice in song, and a monaural voice and accompaniment separation system, ICASSP 2011.
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.

Overview of SPMI Lab

- Setup the lab, since 2003.
- 2 master and 2 ph.d. students (Current), 7 master students (Graduated).
- Research interests
  - Speech Signal and Information Processing
    - Speech recognition and understanding (LVCSR - Mandarin, English)
    - Source separation
    - Speaker recognition
    - Natural language processing
    - Microphone array
  - Statistical Machine Intelligence
    - Construct probabilistic models of the studied phenomenon using human knowledge and machine learning algorithms;
    - Find efficient ways of implementing probabilistic inference with those models.

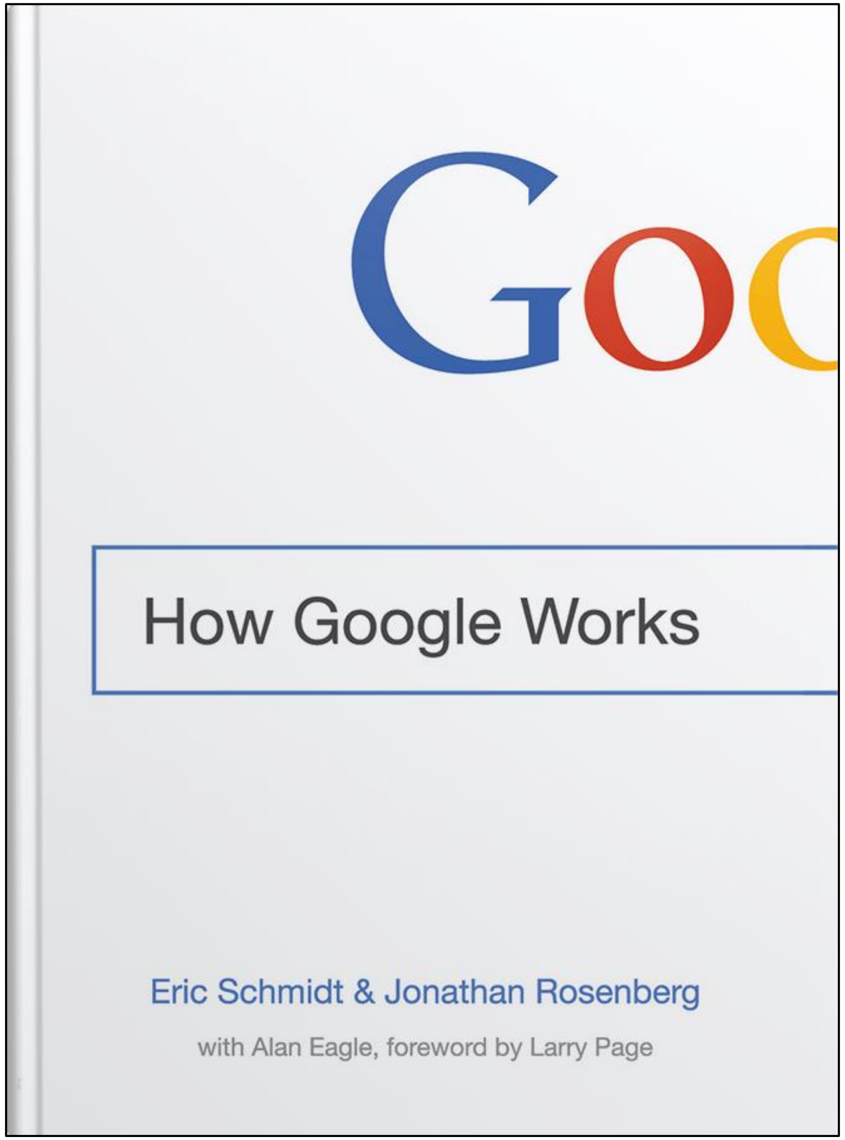
Motivation



As far as the laws of mathematics refer to reality, they are not certain;
and as far as they are certain, they do not refer to reality.

—— Albert Einstein

Motivation



… to think from first principles and real-world physics rather than having to accept the prevailing "wisdom."

—— Larry Page

Motivation - Probabilistic Modeling of Speech

- Dealing with uncertainty + Thinking from physics

- Most speech processing tasks (e.g. pitch estimation, speech recognition, source separation and so on) require a probabilistic model of speech.
- The more scientific the model is, the better we can do for speech processing.

What is this talk about?

- Brief introduction to SPMI lab

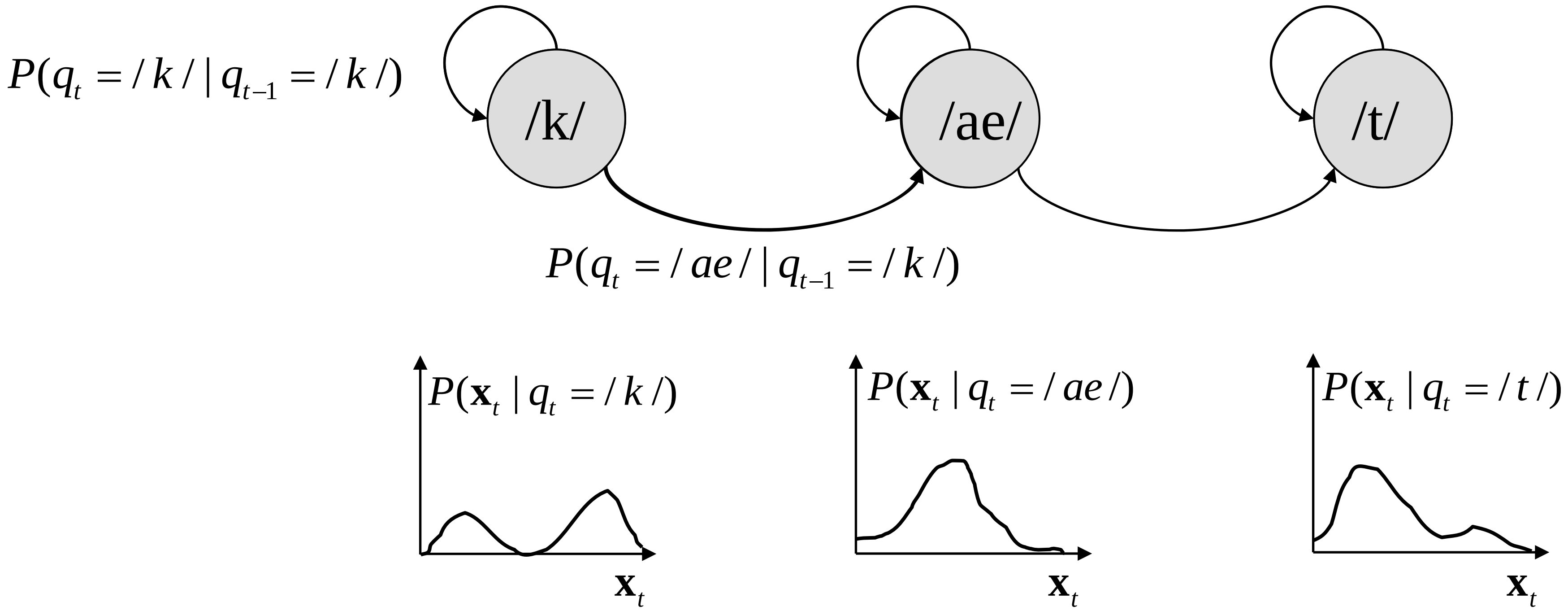
- Motivation
- Bayesian HMM modeling of speech, ICASSP 2007.
- Variational nonparametric Bayesian HMM, ICASSP 2010.
- NMF modeling of voice in song, and a monaural voice and accompaniment separation system, ICASSP 2011.
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.

HMM based Acoustic Model

$$P(\vec{\mathbf{x}},\vec{q}) = P(q_0)P(\mathbf{x_0} \mid q_0)\prod_{t=1}^{T} P(q_t \mid q_{t-1})P(\mathbf{x_t} \mid q_t)$$

Feature Vector Sequence

Phone Sequence



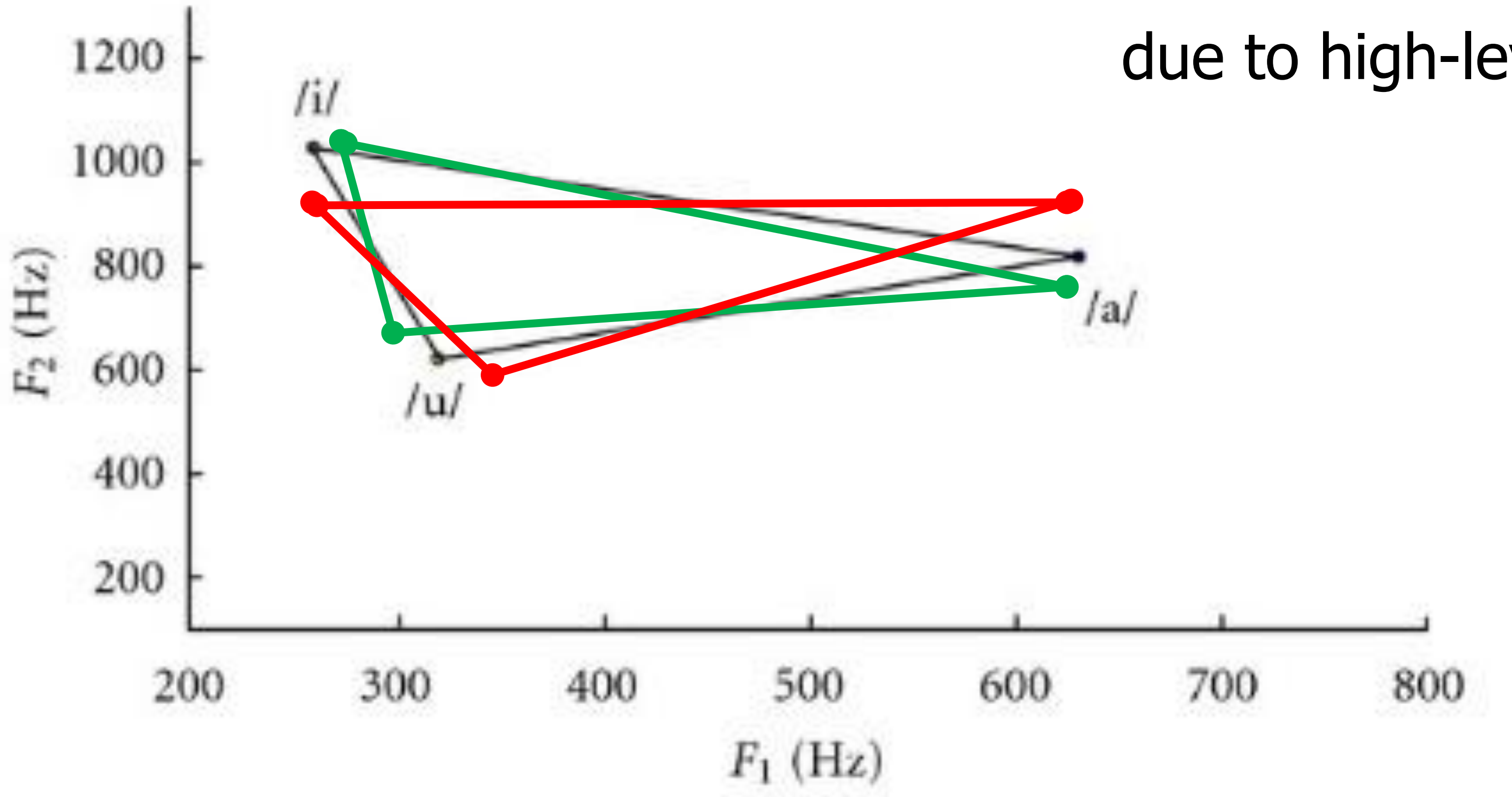
$\mathbf{x}_t$ is the Front-End Feature at time t

Correlation between different sounds

3000 states * 32 gaussians in R^{39}

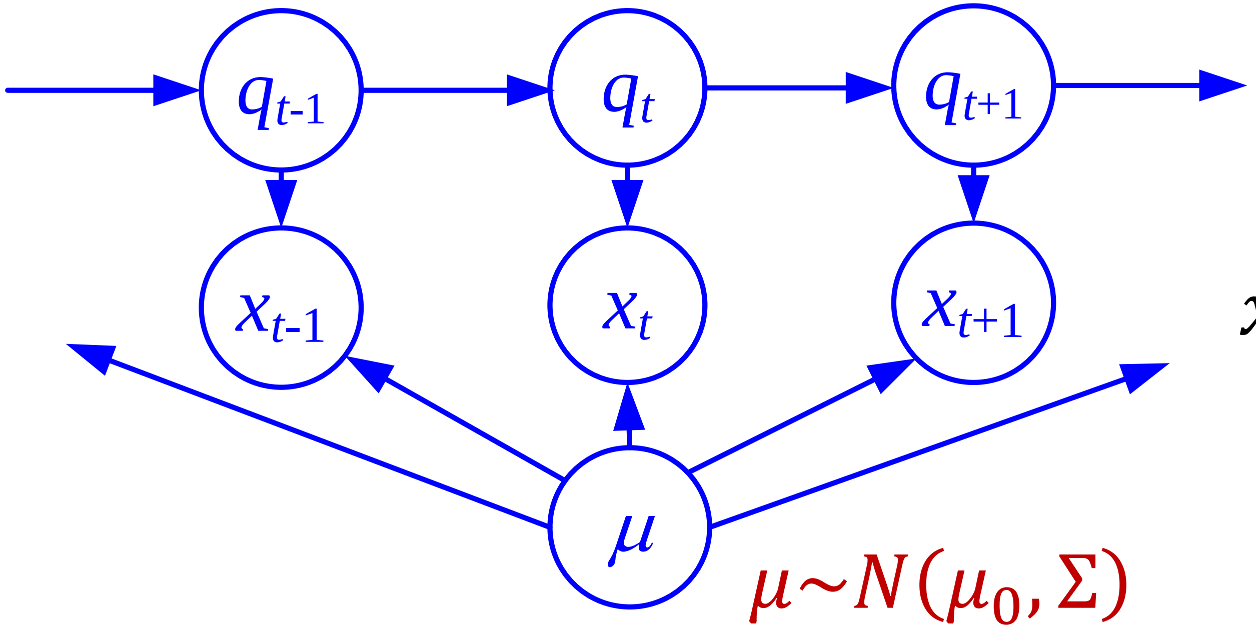
3 states * 1 gaussian in R^{2}

Correlation between the Gaussian means of different sounds due to high-level factors (e.g. speaker).



Bayesian HMM modeling of speech

Bayesian Network Representation of the Generative Model of Speech, incorporating the Supervector Variable μ.



$$x_t | q_t = i \sim N(\mu_i, C_i)$$

$$\mu \sim N(\mu_0, \Sigma)$$

- Use Variational EM algorithm to learn $\Theta = \{\mu_0, \Sigma, \{C_i\}\}$.
- Use ICM to adapt and recognize

$$\max_{q_1 \cdots q_T} p(q_1 \cdots q_T | x_1 \cdots x_T, \mu, \Theta), \max_{\mu} p(\mu | x_1 \cdots x_T, q_1 \cdots q_T, \Theta).$$

Experimental Results – ICASSP 2007

- OGI Numbers: 30-word vocabulary
- 39-dim feature : (12 MFCCs, Energy)+Δ+ΔΔ
- 26 monophone + sil + pause, each modeled by 3 states.

Word Error Rates

Mixture num per state		1	2	4
Baseline		20.86	16.85	13.34
Speaker adaptation	MLLR	20.71	16.79	13.25
	MAP	20.75	16.83	13.32
	MLLR+EV	20.79	16.27	12.59
	EM+EV	**18.42**	**15.76**	**12.44**
Utterance adaptation	MLLR	20.71	16.80	13.29
	MAP	20.75	16.86	13.24
	MLLR+EV	20.81	16.62	13.20
	EM+EV	**18.31**	**15.20**	**11.97**

i-vector in speaker recognition (2010)

Motivation to the next work

When applying HMMs, how many states should we use, and how the states are connected ?

Can we infer the state-transition structure from data ?

What is this talk about?

- Brief introduction to SPMI lab

- Motivation
- Bayesian HMM modeling of speech, ICASSP 2007.
- Variational nonparametric Bayesian HMM, ICASSP 2010.
- NMF modeling of voice in song, and a monaural voice and accompaniment separation system, ICASSP 2011.
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.

Variational Nonparametric Bayesian HMM

Differences from other existing nonparametric Bayesian HMM

iHMM: Beal, Ghahramani, Rasmussen, “The infinite hidden Markov model,” NIPS 2002.

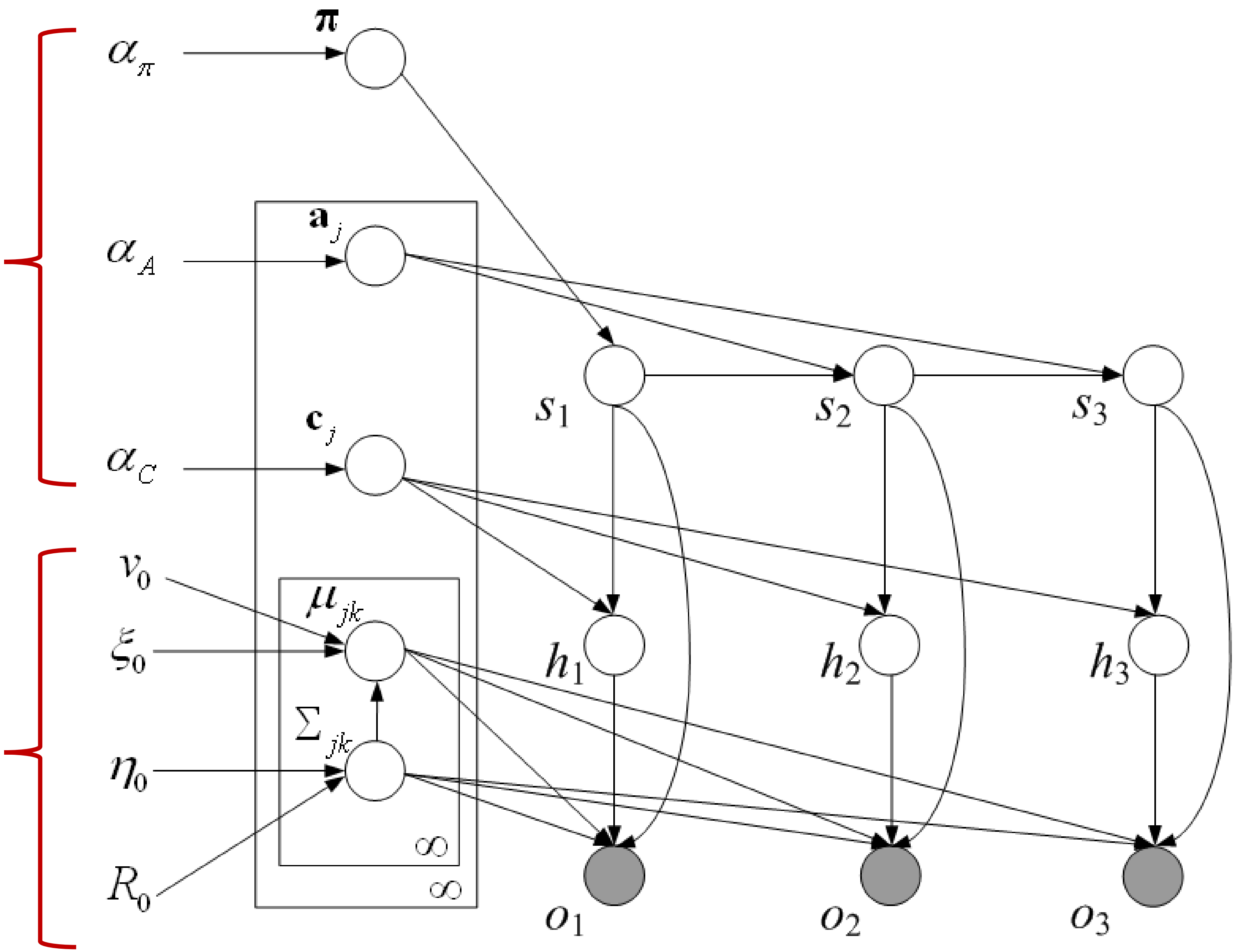
HDP-HMM: Teh, Jordan, Beal, Blei, “Hierarchical Dirichlet processes,” JASA 2006.

1	iHMM and HDP-HMM employ sampling based inference.	We apply the efficient variational inference for the NBHMM.
2	iHMM deals only with discrete observations.	NBHMM supports continuous observations via (infinite) Gaussian mixtures.
3	The transition distribution in iHMM and HDP-HMM is generated from HDP	In the NBHMM, directly created from a stickbreaking construction, simpler

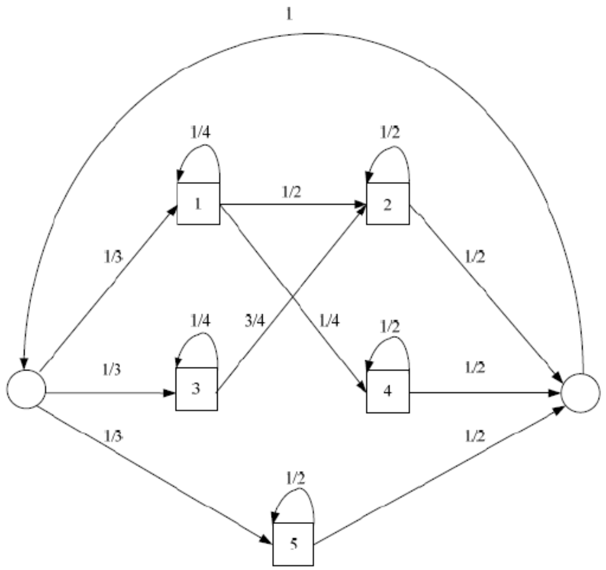
Graphical Model of the Nonparametric Bayesian HMM

A stickbreaking construction of Dirichlet Process prior for the infinite-length multinomial distributions

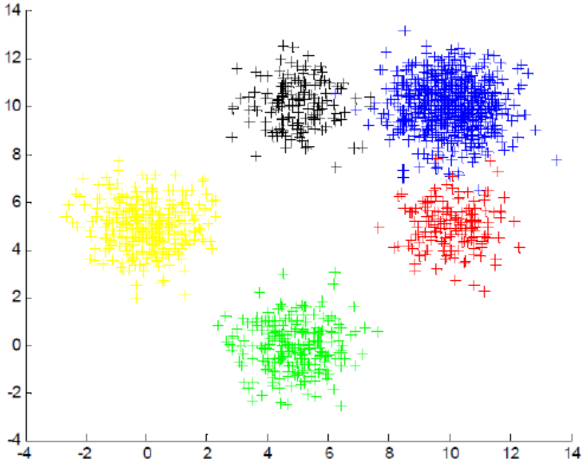
Gaussian-Gamma prior for the Gaussian means and variances



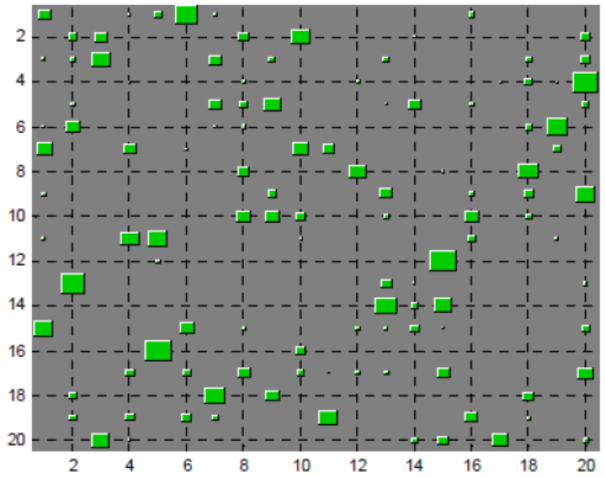
Experimental Results



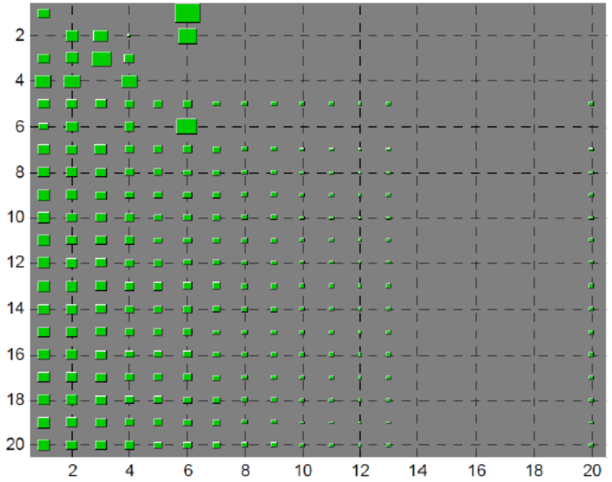
(a) Synthetic Markov machine.



(b) Synthetic observations



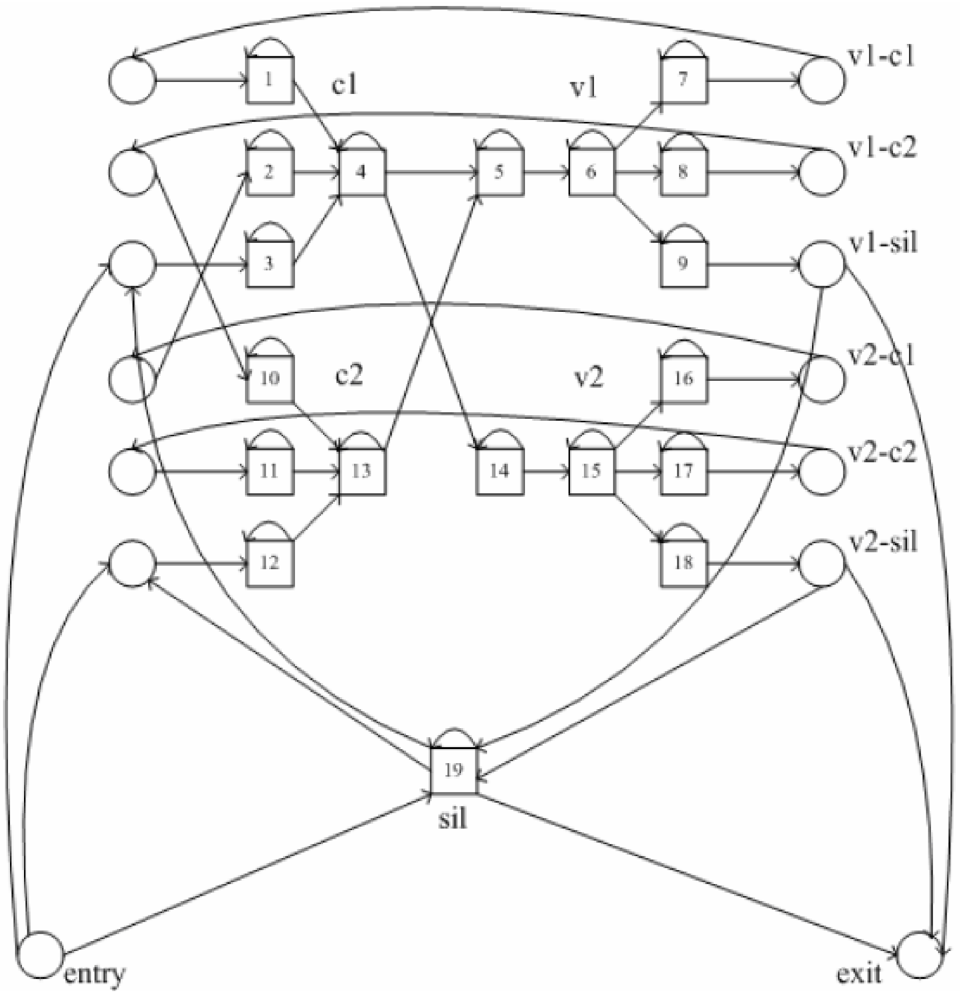
(c) Hinton graph for classical HMM



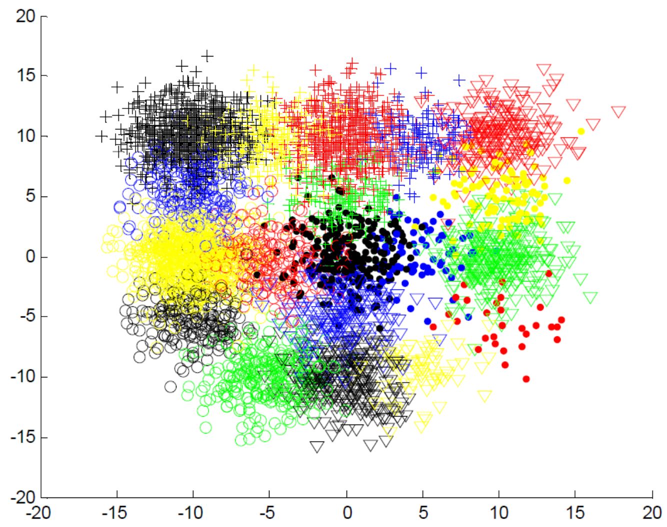
(d) Hinton graph for NBHMM

- A toy example of continuous speech recognition which uses four phonetic states (no.1-4) plus a silence state (no. 5).
- The data contains 50 chains, and the length of each chain is 20.
- The classical HMM with the size of state-space N = 20.
- The NBHMM with the truncation level L = 20.

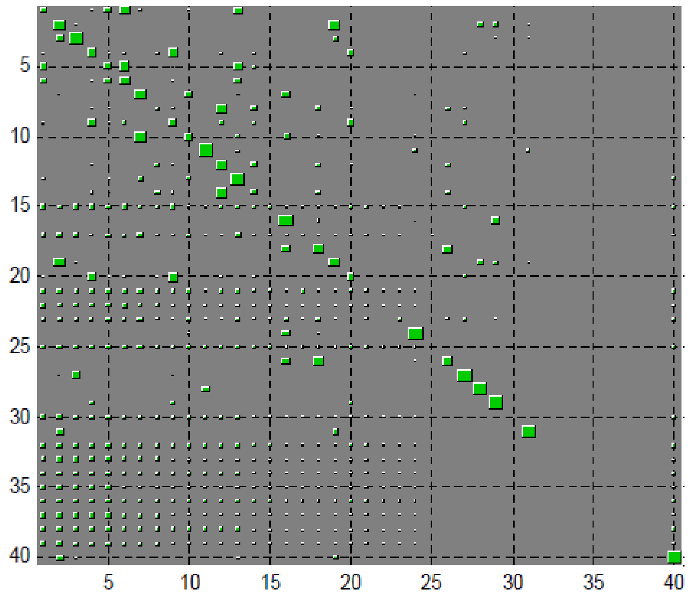
Experimental Results



A “triphone” structure



Synthetic observations



Hinton graph for NBHMM

What is this talk about?

- Brief introduction to SPMI lab

- Motivation
- Bayesian HMM modeling of speech, ICASSP 2007.
- Variational nonparametric Bayesian HMM, ICASSP 2010.
- NMF modeling of voice in song, and a monaural voice and accompaniment separation system, ICASSP 2011.
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.

Introduction

- **Block diagram of voice/accompaniment separation systems**

Mixed Signal → Melody extraction → Time-frequency masking → Voice & Accompaniment

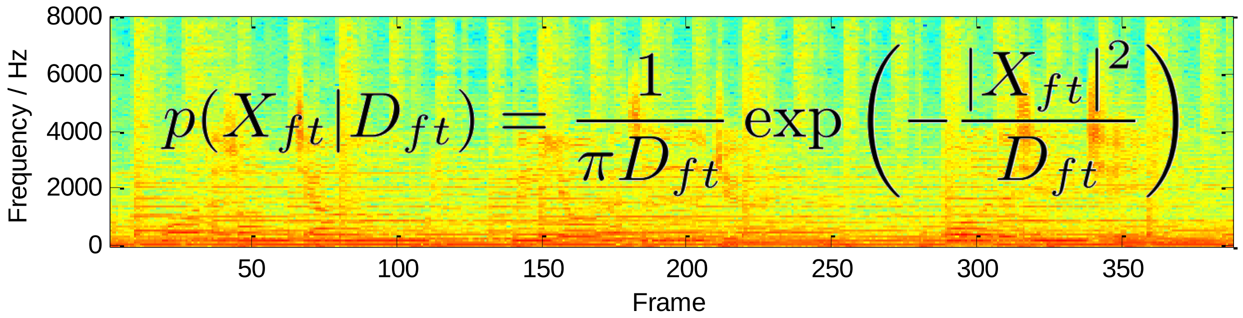
- **Various implementations**

	Melody extraction	**T-F masking**
D.L.Wang [IEEE ASLP 2007]	HMM	Hard
Hsu(1) [IEEE ASLP 2010]	Dressler (Neither HMM nor NMF)	Hard
Hsu(2) [ISMIR 2009]	**HMM**	
Virtanen [ISCA 2008]	HMM	NMF Soft
Durrieu [ICASSP 2009]	NMF	**NMF Soft**
Ours	**HMM**	**NMF Soft**

Flaw

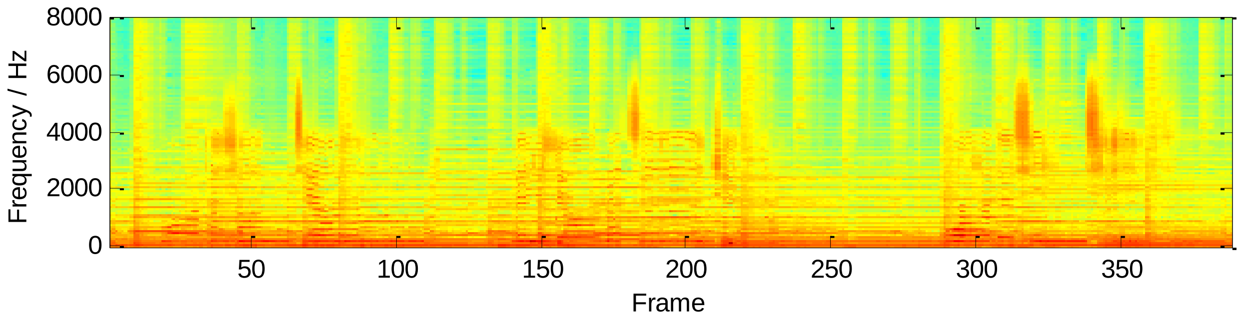
NMF based Acoustic Model

- **Observed spectogram *X*** as a stochastic process



$$p(X_{ft}|D_{ft}) = \frac{1}{\pi D_{ft}} \exp\left(-\frac{|X_{ft}|^2}{D_{ft}}\right)$$

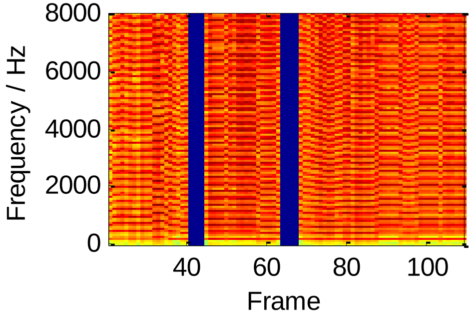
- **Power spectrogram *D*** as its variance parameters



- Main task: Estimate ***D*** with NMF constraints to maximize $p(\boldsymbol{X}|\boldsymbol{D})$

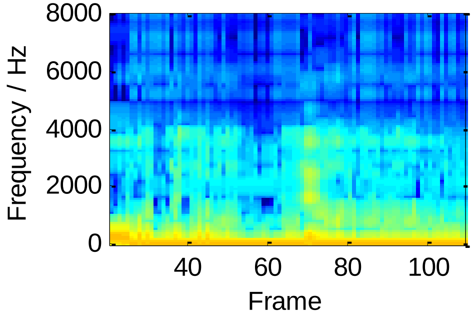
NMF based Acoustic Model

Glottal excitation

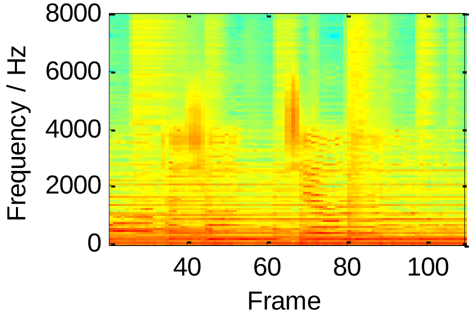


.*

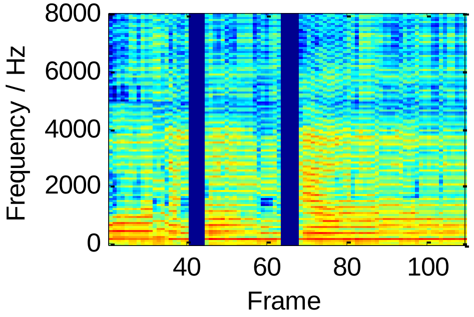
Vocal tract



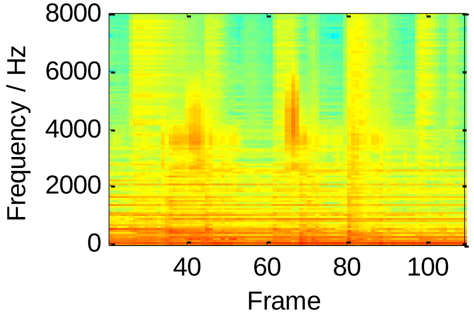
$$\boldsymbol{D} = \underbrace{(\boldsymbol{B}_F \boldsymbol{A}_F) .* (\boldsymbol{B}_K \boldsymbol{A}_K)}_{\boldsymbol{D}_V} + \underbrace{(\boldsymbol{B}_M \boldsymbol{A}_M)}_{\boldsymbol{D}_M}$$



=



+

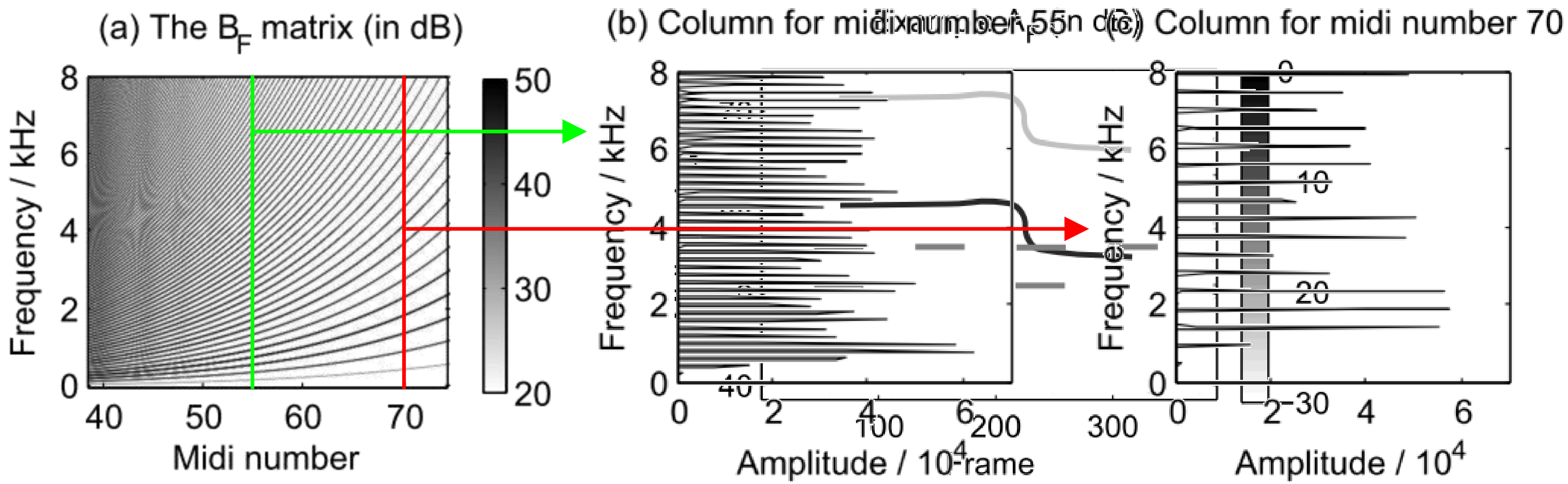


NMF based Acoustic Model

Glottal excitation　　Vocal tract　　Music

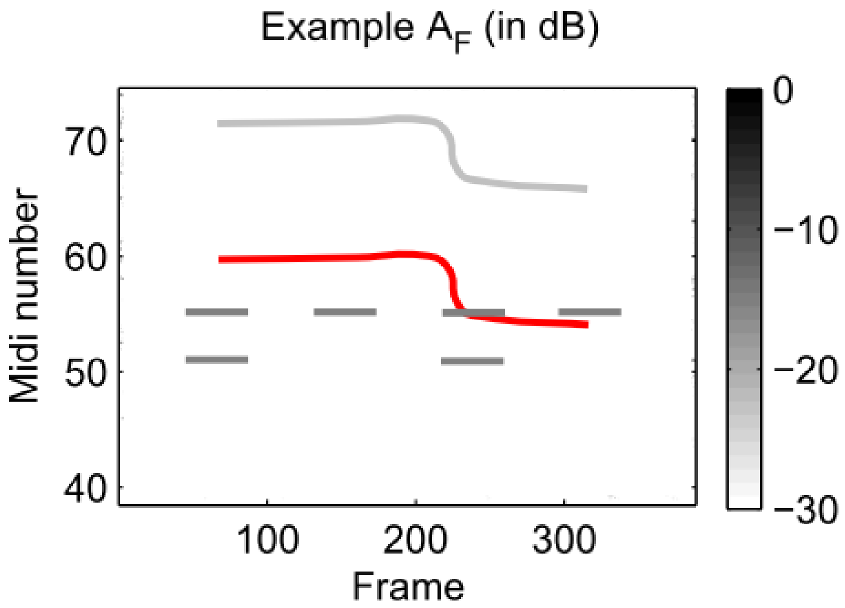
$$\boldsymbol{D} = \underbrace{(\boldsymbol{B}_F \boldsymbol{A}_F) .* (\boldsymbol{B}_K \boldsymbol{A}_K)}_{\boldsymbol{D}_V} + \underbrace{(\boldsymbol{B}_M \boldsymbol{A}_M)}_{\boldsymbol{D}_M}$$

- ***B*** matrices are "codebooks"
- ***A*** matrices are linear combination coefficients



NMF-based melody extraction and separation

- Fix B_F, estimate $\Theta = \{A_F, B_K, A_K, B_M, A_M\}$ under max likelihood
- Find the strong continuous pitch trajectory on A_F , using DP

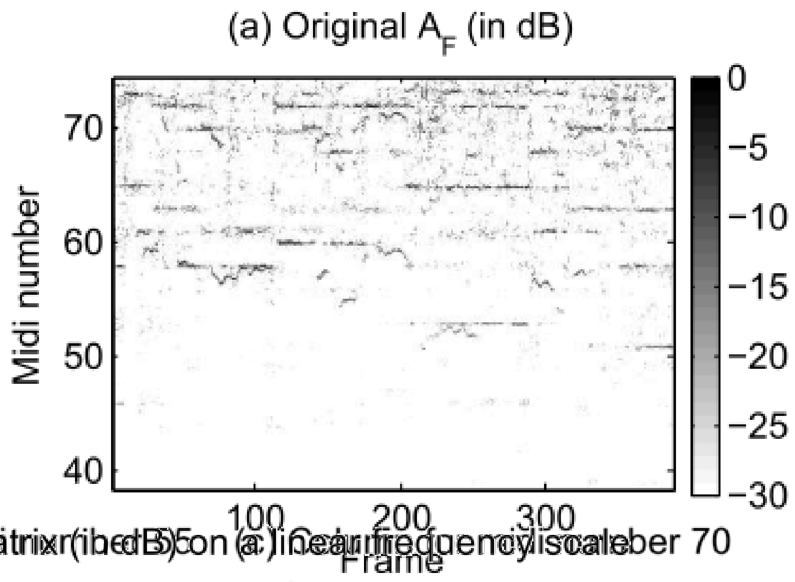


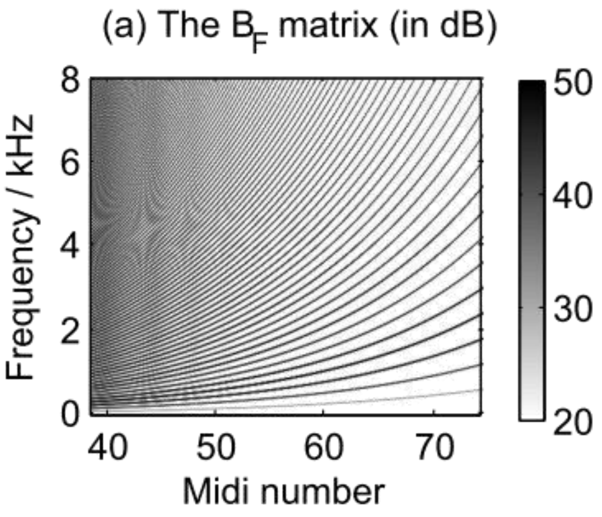
- Fix B_F and A_F , Re-estimate and Soft masking

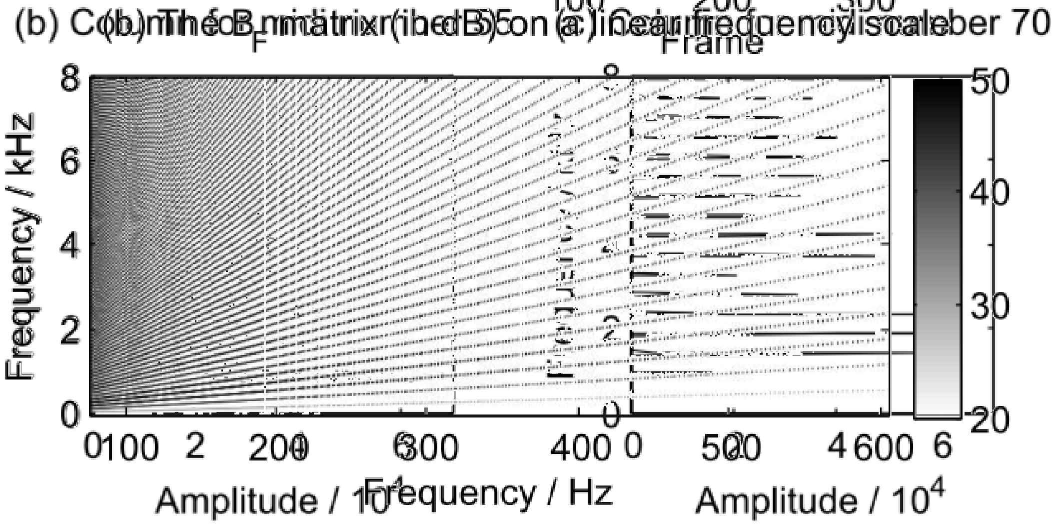
$$\hat{\boldsymbol{X}}_V = \frac{\boldsymbol{D}_V}{\boldsymbol{D}_V + \boldsymbol{D}_M}\boldsymbol{X} \qquad \hat{\boldsymbol{X}}_M = \frac{\boldsymbol{D}_M}{\boldsymbol{D}_V + \boldsymbol{D}_M}\boldsymbol{X}$$

Flaw of NMF-based melody extraction

- **Imbalance in A_F**
- **Two causes:**
  - Non-linearity of midi number scale
  - Columns of B_F unnormalized

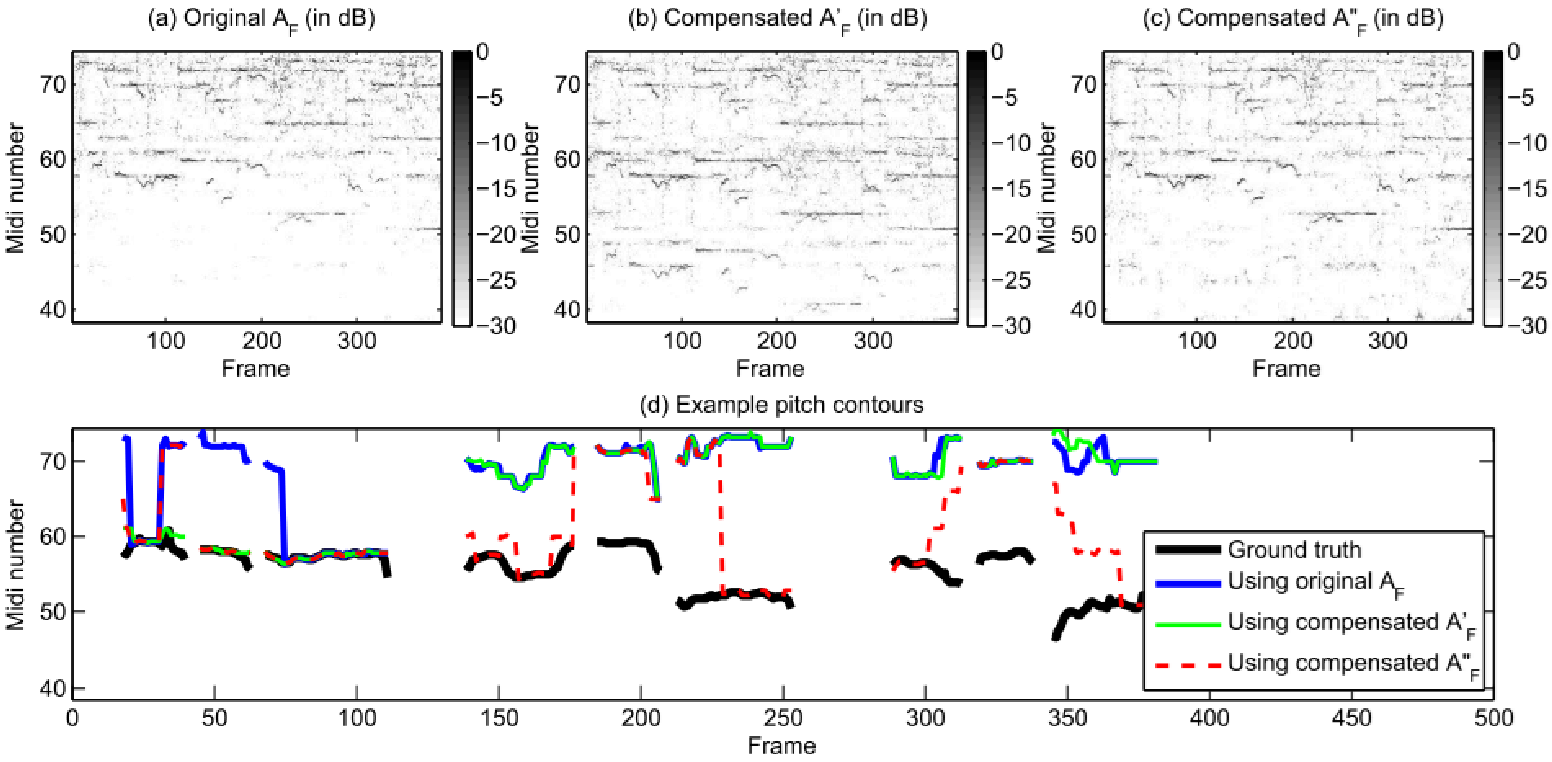






Flaw of NMF melody extraction

- Durrieu's compensation: $(\boldsymbol{A}'_F)_{n,t} = (\boldsymbol{A}_F)_{n,t} + 0.5(\boldsymbol{A}_F)_{n+12,t}$
- Our compensation: $(\boldsymbol{A}''_F)_{n,t} = (\boldsymbol{A}_F)_{n,t} \cdot \frac{1}{f'(n)} \cdot \sum_i (\boldsymbol{B}_F)_{i,n}$
- Compensation cannot eliminate imbalance!



Experimental Results

Mixing ratio	H1 system			Our system	
	Ideal masks	Annot. pitch	Extr. pitch	Annot. pitch	Extr. pitch
-5 dB	10.62	7.5	-0.5	10.34	4.03
0 dB	8.36	6.0	0.9	8.70	5.31
5 dB	5.82	3.0	0.2	6.53	4.09

Table 1. Comparison of Hsu's SDR gains (in dB) on the MIR-1K database for the H1 system (cited from [2]) and our system

Clip	Original		Durrieu		Our system	
	Voice	Acc.	Voice	Acc.	Voice	Acc.
Bearlin	-5.37	5.37	6.2	11.6	3.44	8.76
Tamy	0.51	-0.51	11.5	11.0	4.17	3.66
Bent	0.01	-0.01	5.5	5.6	8.46	8.45
Chevalier	-6.79	6.79	1.5	8.3	2.72	9.50
Love	0.28	-0.28	8.6	8.4	5.17	4.89
Matter	-4.72	4.72	8.0	12.7	4.52	9.24

Table 2. Comparison of Durrieu's SDRs (in dB) for voice and accompaniment on Durrieu's database for Durrieu's system using compensated A'_F (cited from Durrieu's website) and our system

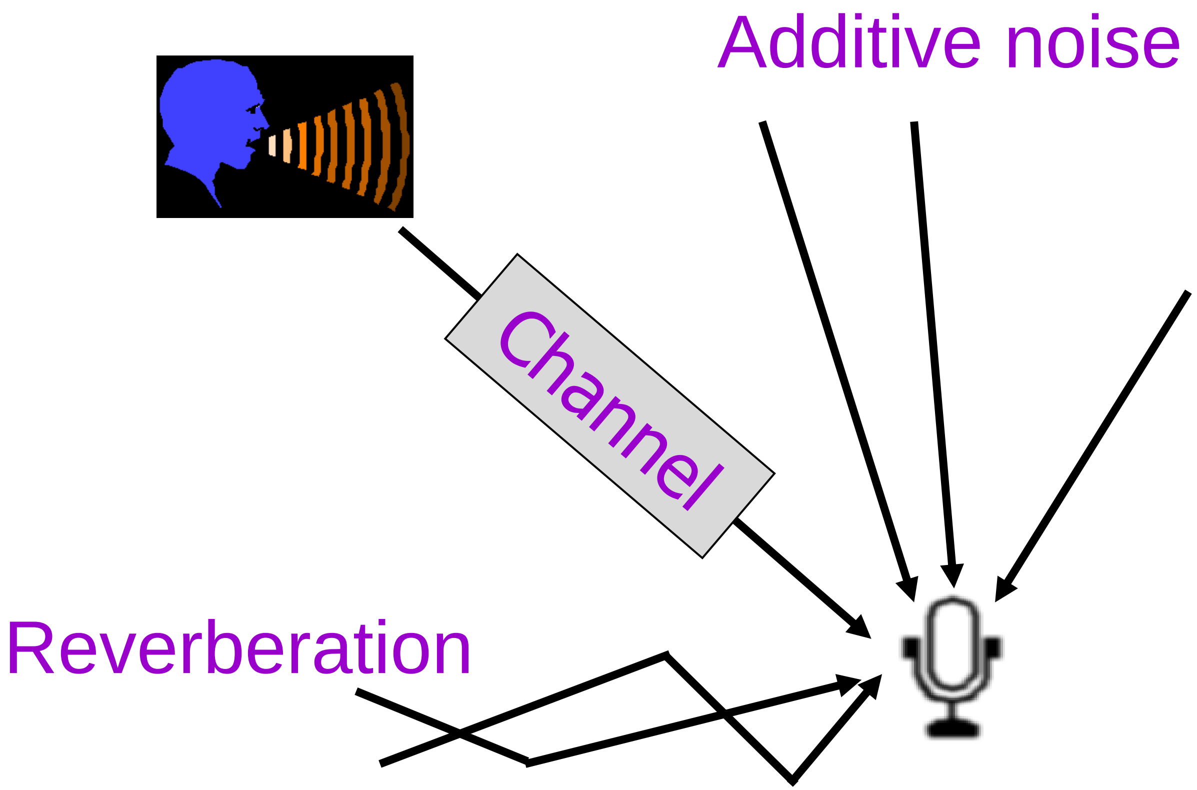
What is this talk about?

- Brief introduction to SPMI lab

- Motivation
- Bayesian HMM modeling of speech, ICASSP 2007.
- Variational nonparametric Bayesian HMM, ICASSP 2010.
- NMF modeling of voice in song, and a monaural voice and accompaniment separation system, ICASSP 2011.
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.

Grand Challenge

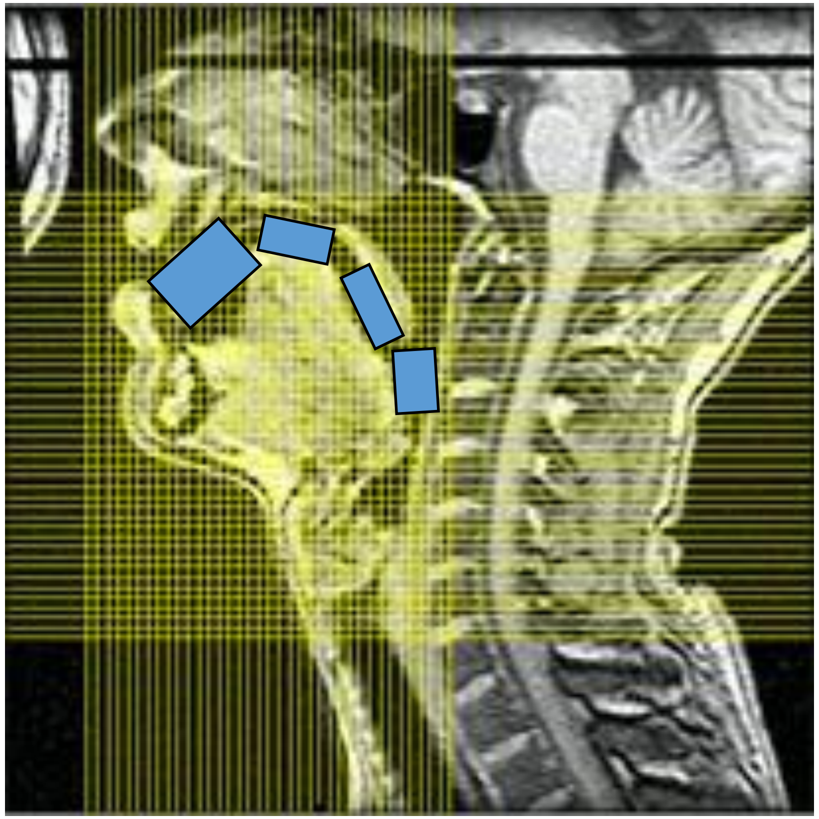
Make Intelligent Machines That Can Hear,
Especially In Complex Acoustic Environment Like Cocktail Party.

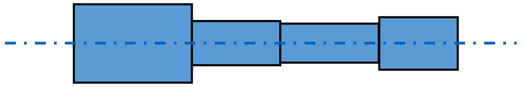


Motivation

What is the basic physical model of speech production ?

—— The Acoustic Tube Model, a.k.a Source-Filter Model.





Motivation

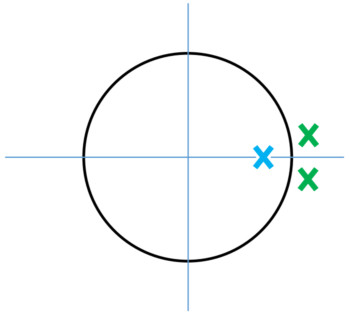
Are there any generative models of speech?

Motivation

- Most of them are actually generative models of the speech features (e.g. Magnitude, Correlogram, Cepstrum).
- Only a few directly model the spectrogram (Reyes-Gomez et al. 2005, Bach and Jordan 2005, Kameoka et al. 2006, Hershey et al. 2010).
- None of them fully respect the physical acoustic tube model
  - Pitch, Glottal source, Vocal tract response, Aspiration noise, Phase
- Drawback: Speech analysis is incomplete, inaccurate or even incorrect.
  - Chicken and egg effect;
  - Vocal tract estimate (e.g. LPC and MFCC) corrupted by spectral tilt.

PAT Model

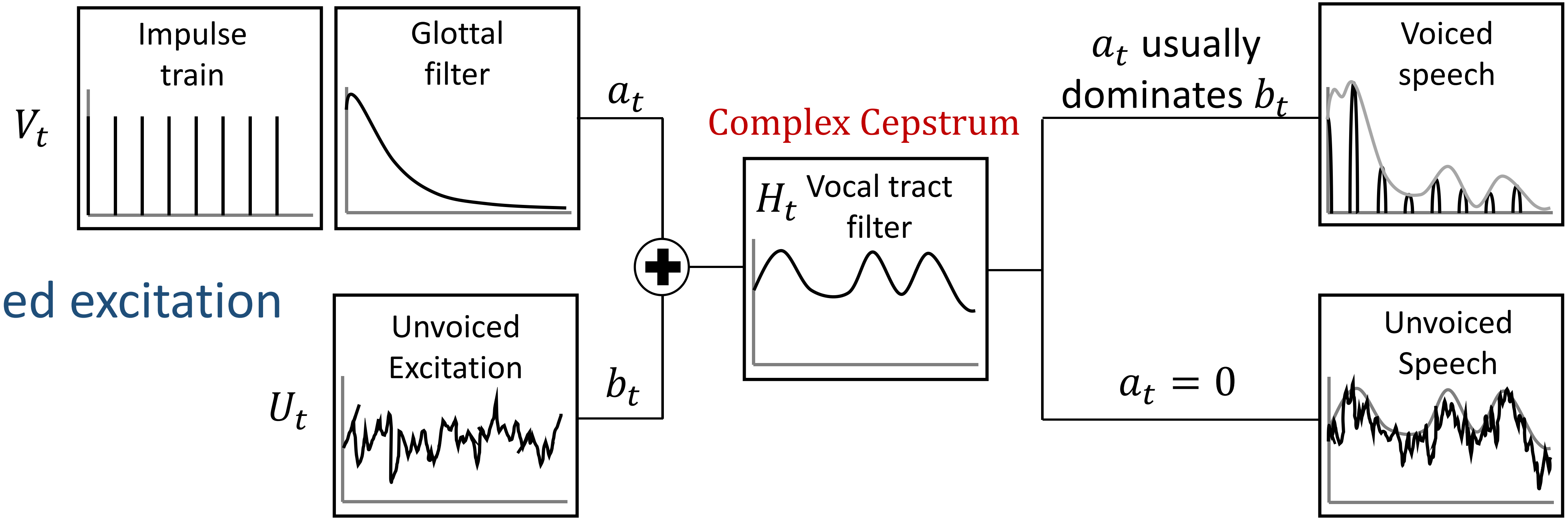
$$S_t(\omega) = [a_t V_t(\omega) + b_t U(\omega)] H_t(\omega) \circledast W_t(\omega) + N_t(\omega)$$



Voiced excitation

$$V_t(\omega) = G_t(\omega) e^{-j\omega\tau_t} \sum_k \delta(\omega - k\omega_{0t})$$

Unvoiced excitation



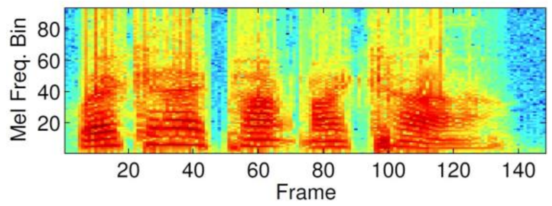
Highlight of PAT

- PAT is based on the fundamental physics of speech production.

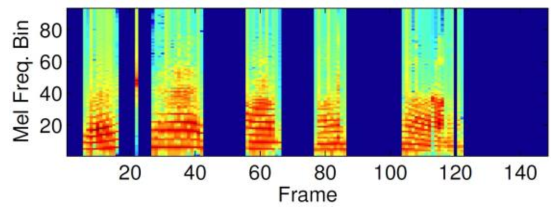
- A **probabilistic generative model** that **jointly** considers all important speech parameters;
- Incorporates **breathiness** and **glottal source**;
- Incorporates **phase modeling** and so completely defines a probabilistic model for the complex spectrum of speech;
- **Makes U/V states a continuum** by introducing voiced amplitude and unvoiced amplitude, which is closer to the nature of speech.

Experimental Results

Voiced Reconstruction

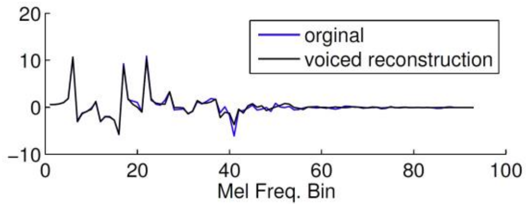


Original Spectrogram

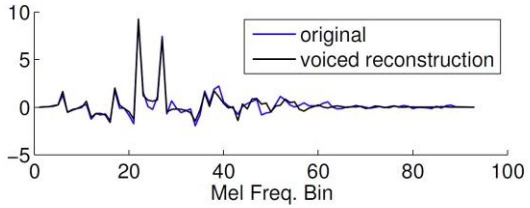


Voiced Reconstruct

Voiced Reconstruction – Single Frame

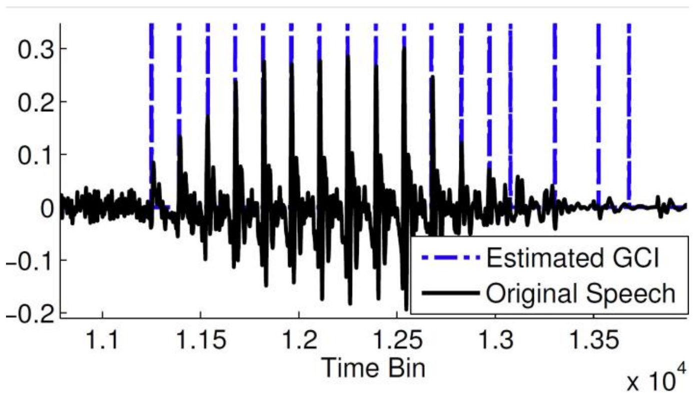


Real Spectrum

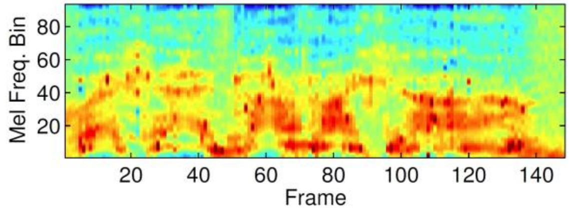


Imaginary Spectrum

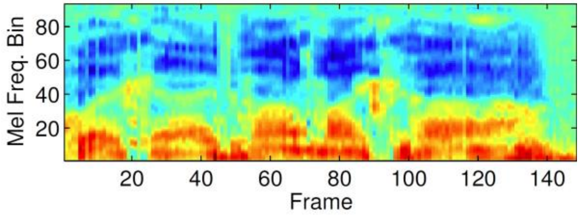
GCI Location Estimation



Vocal Tract Filter Estimation

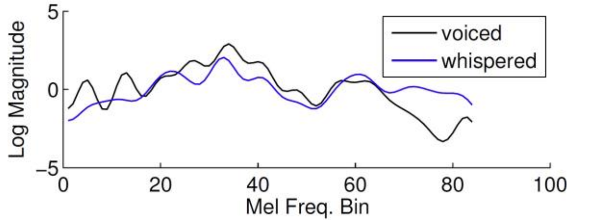


PAT2

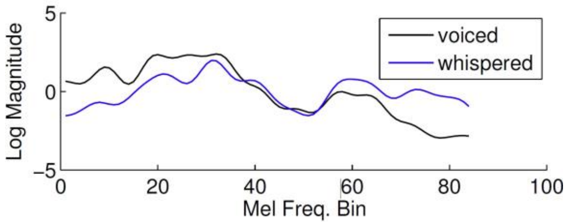


MFCC

Voiced vs Whispered



PAT2



MFCC

Summary - Probabilistic Modeling of Speech

- PAT: On the way …
- One of the reviewers comments "to my knowledge the most complete attempt on developing a true generative model for speech".

- Bayesian HMM modeling of speech, ICASSP 2007

  -> Put a prior over model parameters to account for high-level factors (e.g. the speaker, utterance style).

- Variational nonparametric Bayesian HMM, ICASSP 2010

  -> Discover the state-transition structure according to data.

- NMF modeling of voice, ICASSP 2011

  -> feasible

Thanks:

Jun Luo, Nan Ding, Yun Wang, Yang Zhang, Mark Hasegawa-Johnson.

Thanks for your attention !