Probabilistic Modeling of Speech and Language

Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab,

Department of Electronic Engineering, Tsinghua University, Beijing, China.

Now Visiting Scholar at Beckman Institute, UIUC.

6/16/2015, IBM

What is this talk about?

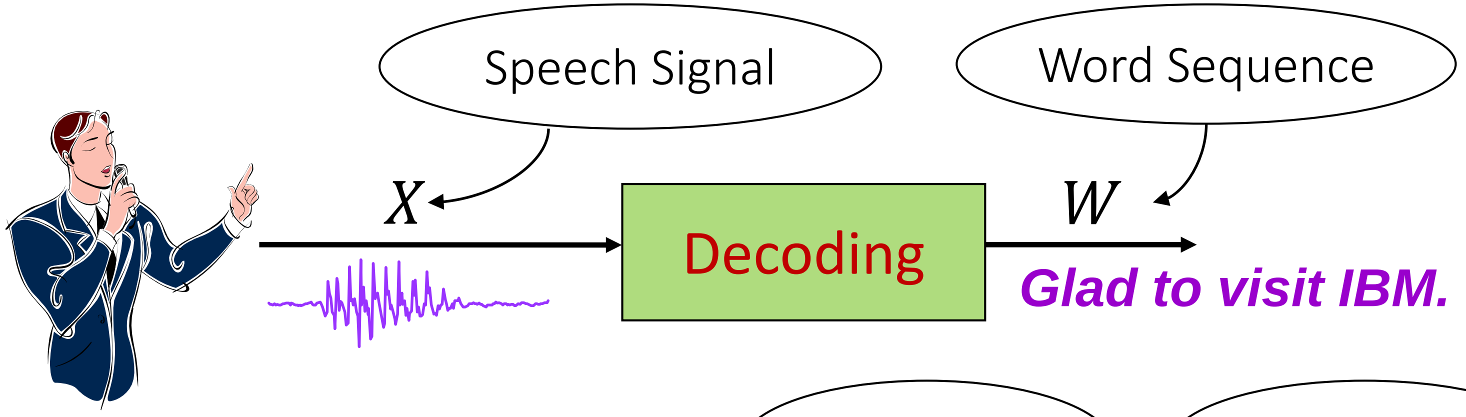
- Brief introduction to SPMI lab

- Motivation
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.
- Random field approach to language modeling, ACL 2015.

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 - Probabilistic Modeling of Speech and Language



Speech Model Language Model

$$\underset{W}{\operatorname{argmax}} P(W|X) = \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(W)}{\cancel{P(X)}}$$

- Speech Models: Speech recognition, pitch estimation, source separation, …
- Language Models: Speech recognition, machine translation, handwriting recognition, …
- The more scientific the models are, the better we can do for speech and language processing.

What is this talk about?

- Brief introduction to SPMI lab

- Motivation
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.
- Random field approach to language modeling, ACL 2015.

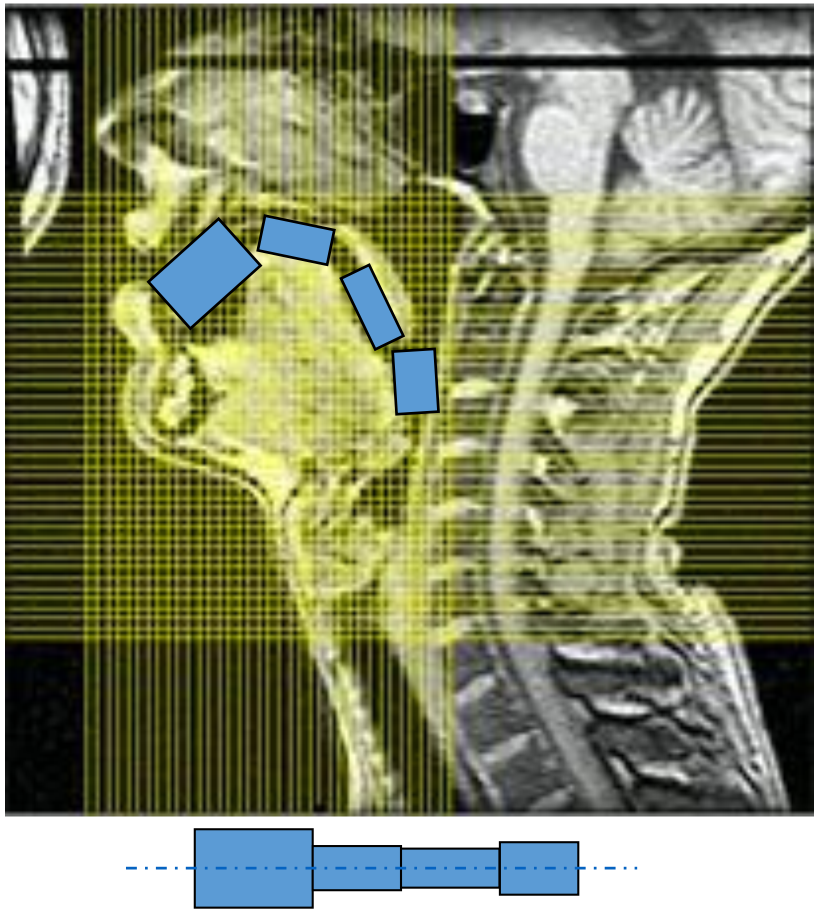
Our trial-and-error efforts

- Relax the state independent assumption in HMMs
  - ICASSP 2002, ICSLP 2002, INTERSPEECH 2004.
- 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.
- Eigenvoice Speaker Modeling + VTS-based Environment Compensation for Robust Speech Recognition
  - ICASSP 2012
- PAT Models
  - AISTATS 2012, ICASSP 2014

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, Cepstrum, Correlogram
- Only a few directly model the spectrogram
  - Reyes-Gomez, Jojic, Ellis, 2005; Bach and Jordan, 2005; Kameoka et al. 2010; Hershey et al. 2010; Deng et al. 2006.
- None of them fully respect the physical acoustic tube model

Important speech elements

- Pitch
- Glottal source
- Vocal tract response
- Aspiration noise
- Phase

Motivation

- Drawback: Speech analysis is inaccurate, making great troubles for back-end inference
  - Chicken and egg effect [1]
  - Entangled variation/randomness
  - e.g. Vocal tract estimate (e.g. LPC and MFCC) corrupted by ‘spectral tilt’ due to glottal pulse
- A complete model of speech
  - Disentangle the underlying elements of variation, knowledgeably vs blindly.
  - Provide strong constraints/priori knowledge [2]

[1] Kameoka, Ono, Sagayama, “Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency”, 2010.

[2] Simsekli, Le Roux, Hershey, “Non-negative source-filter dynamical system for speech enhancement”, 2014.

Motivation

- Previous efforts
  - Additive deterministic-stochastic model, (Serra & Smith 1990)
  - STRAIGHT model, (Kawahara, et al. 2008)
  - Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis, (Degottex, et. al 2013)
  - Non-negative source-filter dynamical system for speech enhancement, (Simsekli, Le Roux, Hershey, 2014)
- Probabilistic Acoustic Tube (PAT)
  - Jointly consider breathiness, glottal excitation and vocal tract in a probabilistic modeling framework, and notably with phase information.

PAT1: Probabilistic acoustic tube: A Probabilistic Generative Model of Speech for Speech Analysis/Synthesis.
(Ou, Zhang. AISTATS 2012)

PAT2: Improvement of PAT Model for Speech Decomposition.
(Zhang, Ou, Hasegawa-Johnson. ICASSP 2014)

PAT3: Incorporating AM-FM effect in voiced speech for PAT model.
(Zhang, Ou, Hasegawa-Johnson. Submitted)

PAT2 Model

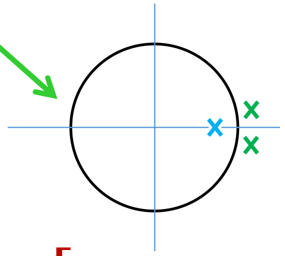
speech

Serra & Smith 1990, Degottex et al 2013

Impulse response of vocal tract

$$s[t] = v[t] + u[t]$$
$$= (a \cdot e_v\,[t] + b \cdot e_u[t]) * h[t]$$

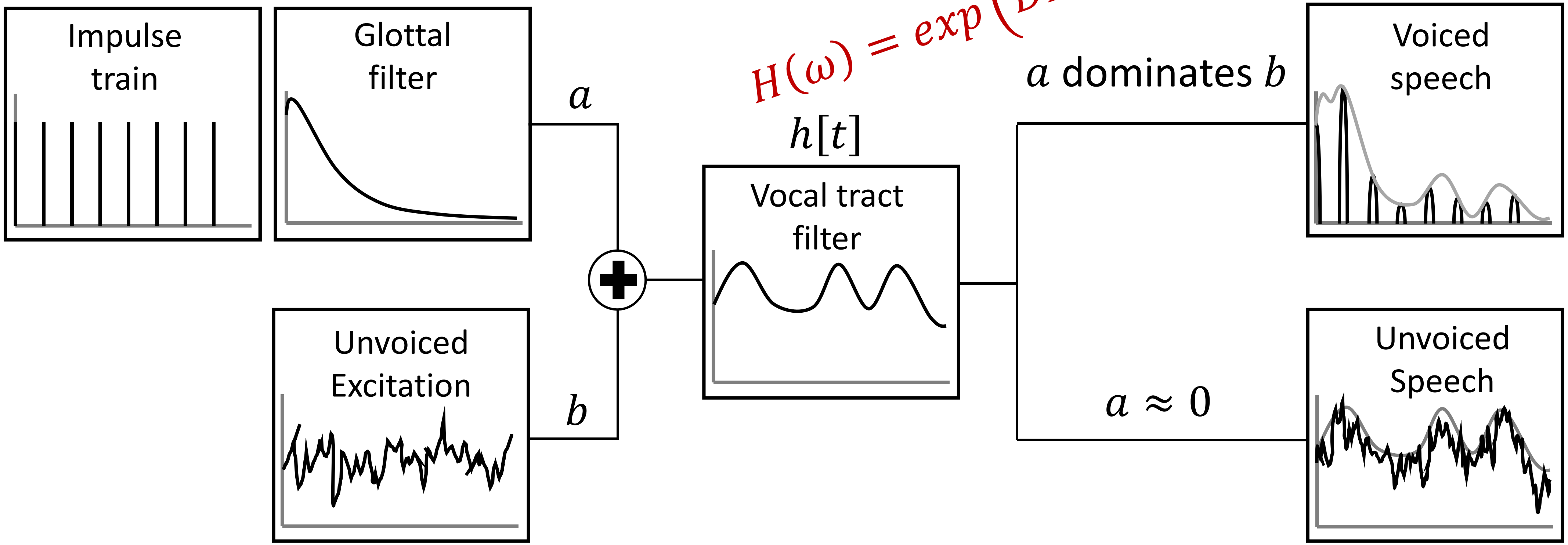
Doval et al 2013



$$e_v[t] = \sum_d real\left[G(d\omega_0) \cdot e^{jd\omega_0(t-\tau)}\right]$$

26-dim Complex Cepstrum $\hat{h}$ with quefrency $\hat{t}$

$$H(\omega) = exp\left(DFT\left(\hat{h}[\hat{t}]\right)\right)$$



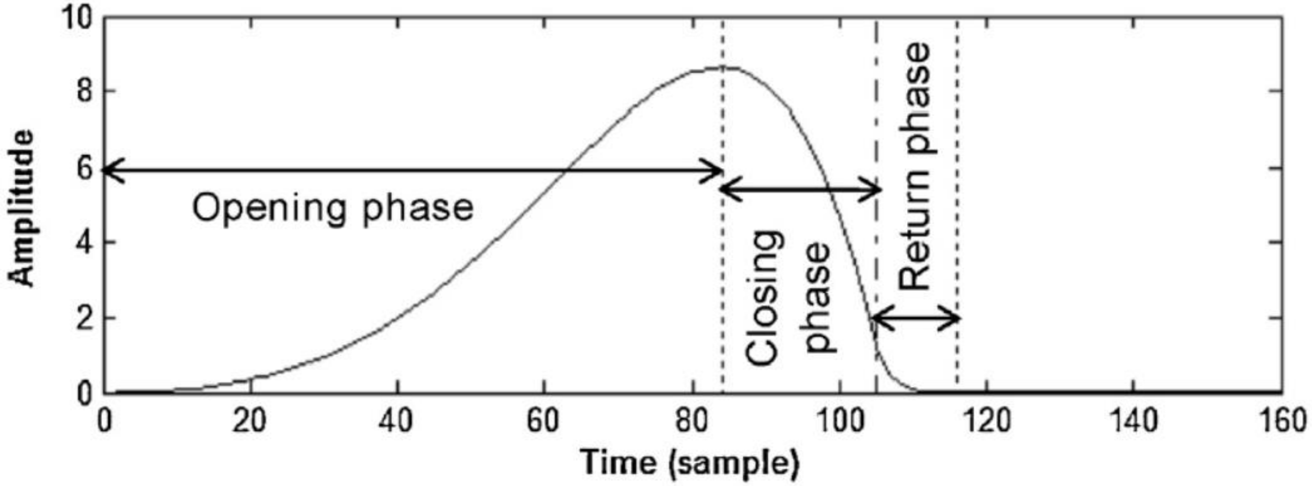
$$e_u[t] \sim \mathcal{N}(0,1), i.e.\ WGN$$

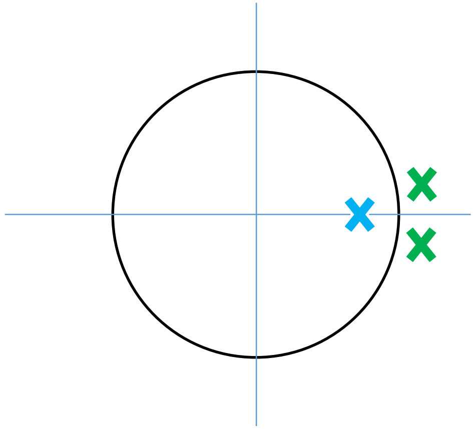
Three-pole Model for Glottal Pulse (Doval et al 2013)

Glottal Flow Waveform

$$e_v[t] = \sum_d real\left[G(d\omega_0) \cdot e^{jd\omega_0(t-\tau)}\right]$$

Glottal Pulse

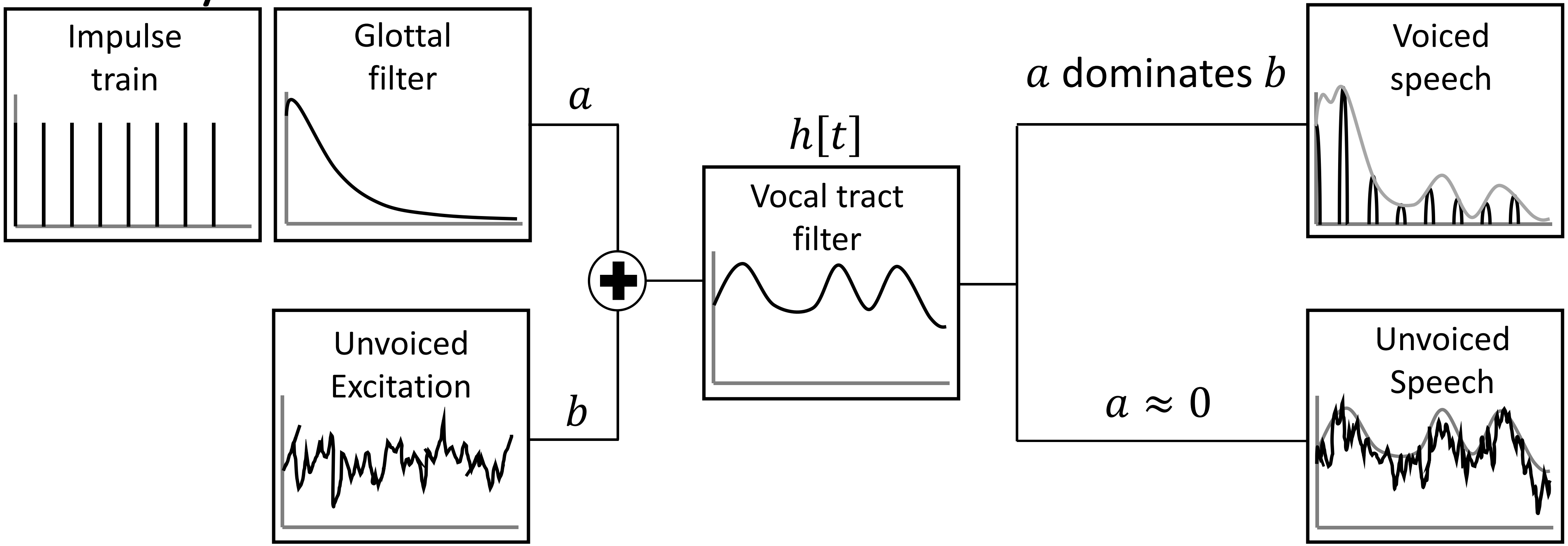




$$G(\omega) = \frac{1}{[1 + 2g_1\cos(\beta)e^{-j\omega} + g_1^2 e^{-2j\omega}][1 + g_2 e^{-j\omega}]}$$

parameterized by $\vec{g} = \{g_1, \beta, g_2\}$

PAT2 Summary



Time domain: $s[t] = v[t] + u[t] = (a \cdot e_v\,[t] + b \cdot e_u[t]) * h[t]$

$$\Big\downarrow\; vec\left[DFT[s[t]]\right]$$

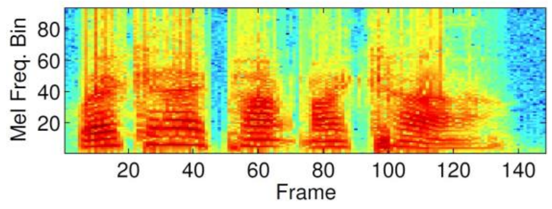
Frequency domain: $\vec{s} = a \cdot vec(\omega_0, \tau, \vec{g}, \hat{h}) + b \cdot vec\left[DFT[h[t]]\right] \boxdot vec[DFT[WGN]]$

Hidden variables: $z = \{a, b, \omega_0, \tau, \vec{g}, \hat{h}\} \in R^{31}$

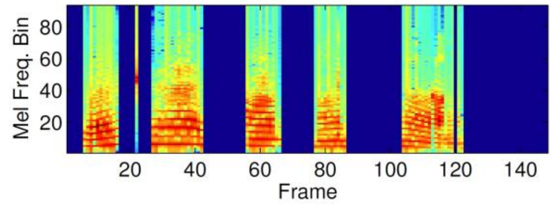
MAP inference $p(z|\vec{s}) \propto p(\vec{s}|z)p(z)$ by Monte Carlo sampling and L-BFGS search.

Experimental Results

Voiced Reconstruction

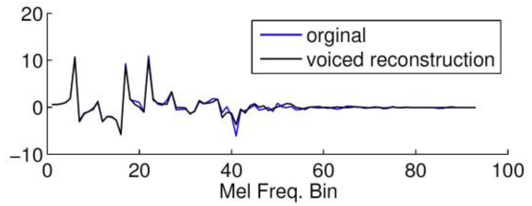


Original Spectrogram

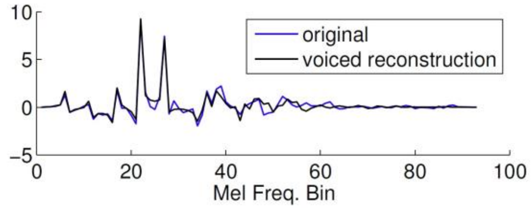


Voiced Reconstruct

Voiced Reconstruction – Single Frame

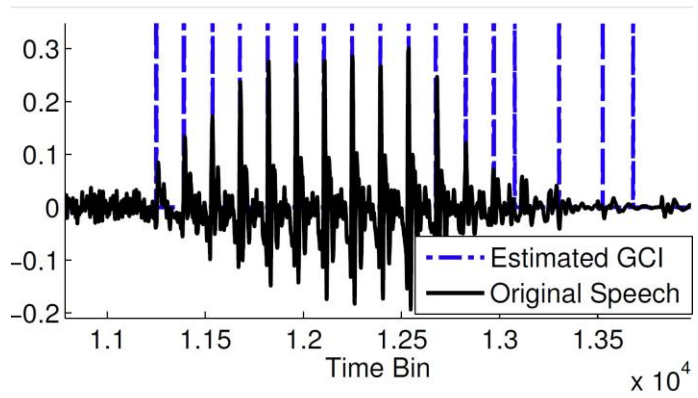


Real Spectrum

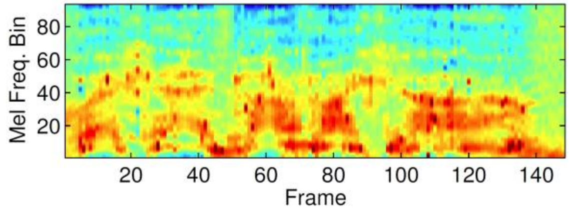


Imaginary Spectrum

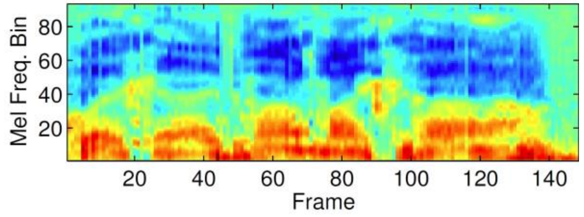
GCI Location Estimation



Vocal Tract Filter Estimation

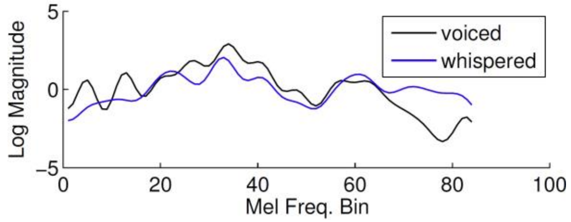


PAT2

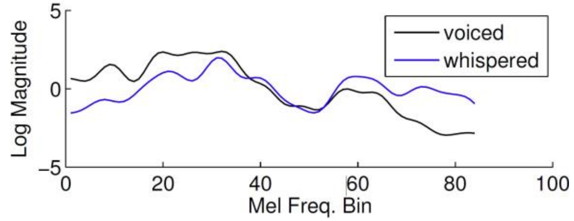


MFCC

Voiced vs Whispered



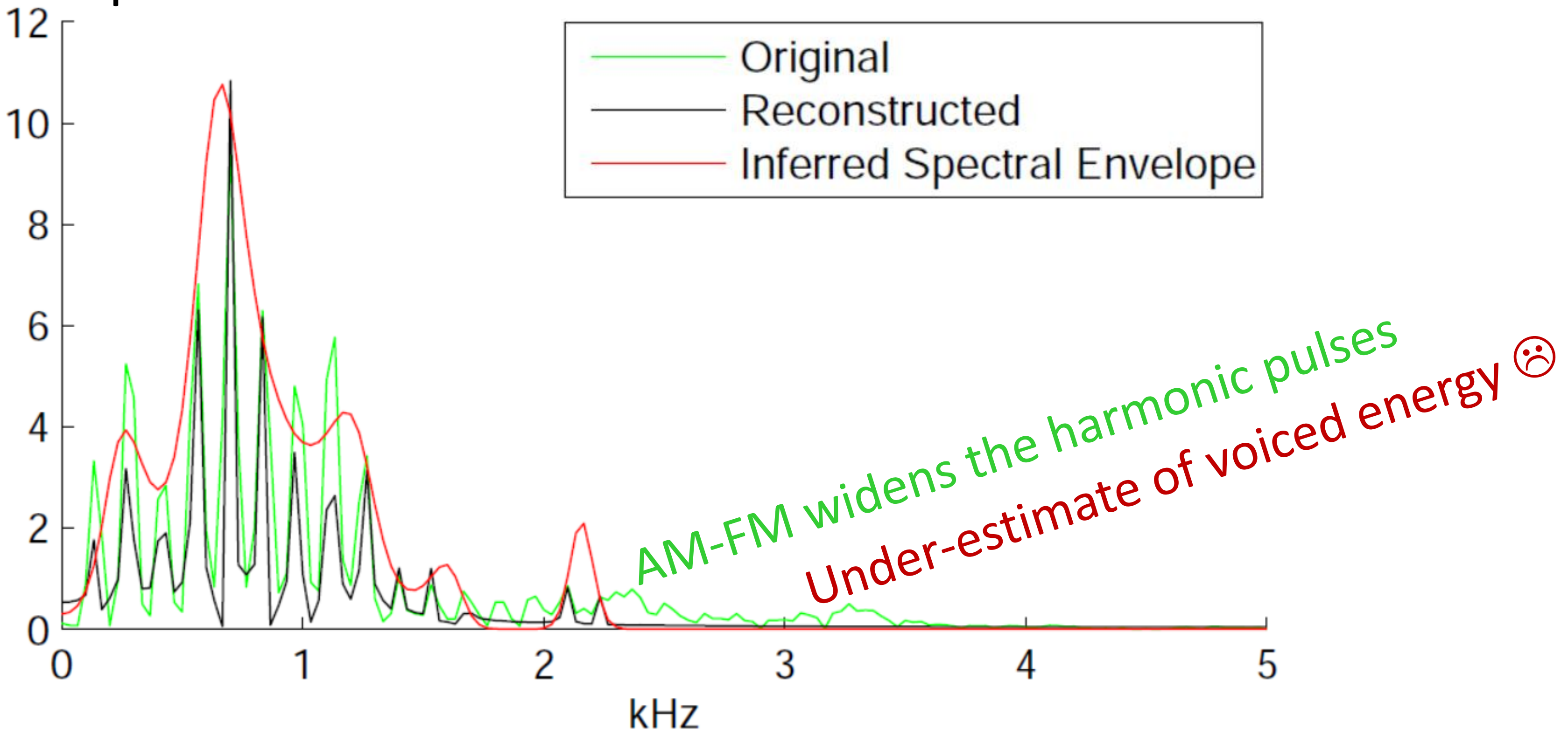
PAT2



MFCC

PAT3 Motivation

- To incorporate AM-FM effect in voiced speech
  - Harmonic part is assumed to be strictly periodic.
  - Variations within a single voiced frame are common and non-negligible.
- Two main variations are pitch jitter and amplitude shimmer
  - Give voiced speech its naturalness



PAT2 Model

$$v[t] = \sum_d real\left[\alpha_d e^{jd\omega_0 t}\right]$$

where $\alpha_d = aH(d\omega_0)G(d\omega_0)e^{-jd\omega_0\tau}$

PAT3 Model

$$v[t] = \sum_d real\left[\alpha_d \eta_d[t] e^{jd\omega_0 t + jd\phi[t]}\right]$$

Amplitude perturbation

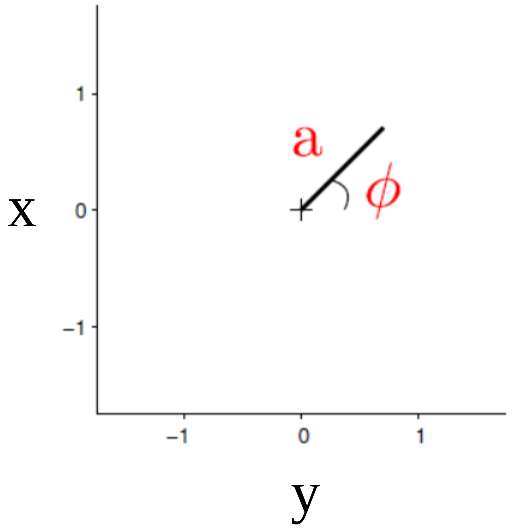
Phase perturbation

$$v[t] = \sum_d x_d[t]^T \xi_d[t]$$

$$\begin{cases} x_d[t] = \begin{pmatrix} |\alpha_d| cos(d\omega_0 t + \angle\alpha_d) \\ |\alpha_d| sin(d\omega_0 t + \angle\alpha_d) \end{pmatrix}, \text{ the strictly periodic signal} \\ \xi_d[t] = \begin{pmatrix} \eta_d[t] cos(d\phi[t]) \\ \eta_d[t] sin(d\phi[t]) \end{pmatrix}, \text{ the amplitude and phase perturbation, phasor} \end{cases}$$

Phasor representation

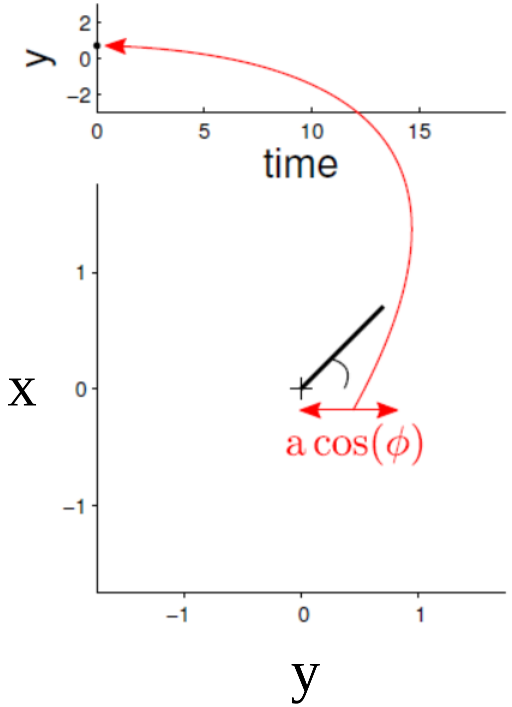
An AM-FM sinusoid $\mathrm{y}(t) = \Re\left(\mathrm{a}(t)\exp(i\phi(t))\right)$



Qi, Minka, Picara, “Bayesian spectrum estimation of unevenly sampled nonstationary data”, ICASSP 2002.
Turner and Sahani, “Probabilistic amplitude and frequency demodulation”, NIPS 2011.

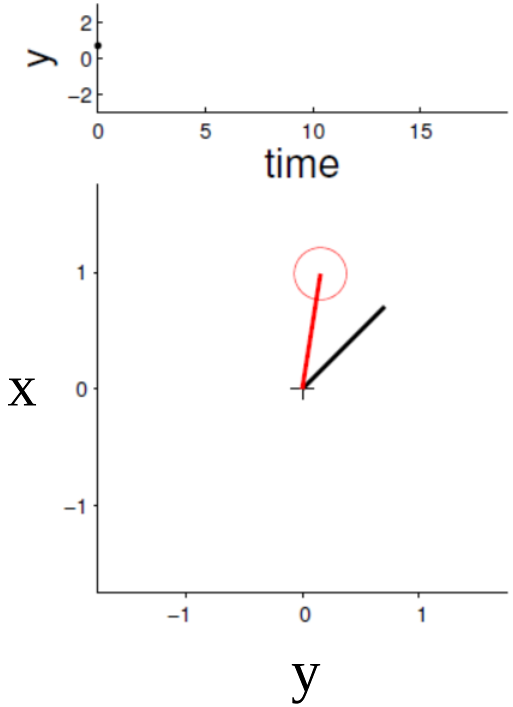
Phasor representation

$$y(t) = \Re\left(a(t)\exp(i\phi(t))\right)$$



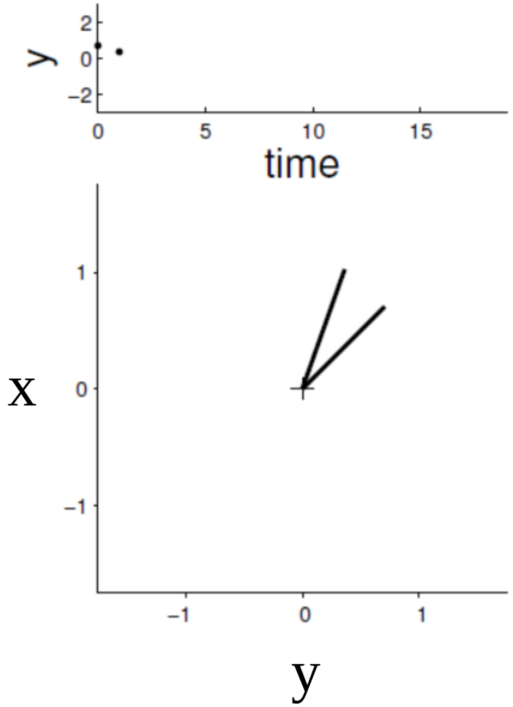
Phasor representation

$$y(t) = \Re\left(a(t)\exp(i\phi(t))\right)$$



Phasor representation

$$\mathrm{y}(t) = \Re\left(\mathrm{a}(t)\exp(i\phi(t))\right)$$



Phasor representation

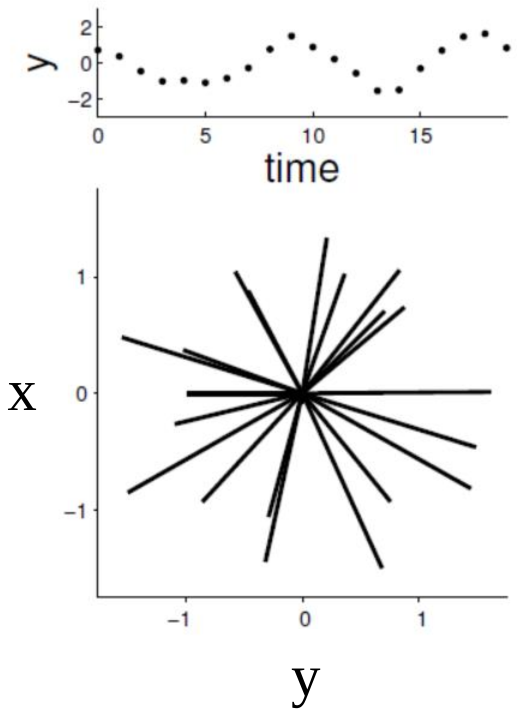
If $\phi(t) = \omega t + \theta(t)$, then

$$y(t) = real\left[a(t)e^{j\omega t + j\theta(t)}\right]$$
$$= f(t)^T \xi(t)$$

$f(t) = \begin{pmatrix} cos(\omega t) \\ sin(\omega t) \end{pmatrix}$, a fixed freq signal

$\xi(t) = \begin{pmatrix} a(t)cos(\theta(t)) \\ a(t)sin(\theta(t)) \end{pmatrix}$, a phasor

$y(t) = \Re(a(t)\exp(i\phi(t)))$



Theorem: If $\theta(t)$ is uniform distributed, $a(t)$ is Rayleigh distributed,

Then $\xi(t) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$

PAT3 Model

$$v[t] = \sum_d x_d[t]^T \xi_d[t] \quad \begin{cases} x_d[t] = \begin{pmatrix} |\alpha_d| cos(d\omega_0 t + \angle\alpha_d) \\ |\alpha_d| sin(d\omega_0 t + \angle\alpha_d) \end{pmatrix}, \text{ the strictly periodic signal} \\ \xi_d[t] = \begin{pmatrix} \eta_d[t] cos(d\phi[t]) \\ \eta_d[t] sin(d\phi[t]) \end{pmatrix}, \text{ the amp. \& phase perturbation, phasor} \end{cases}$$

$$\xi_d(t) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_d^2 \begin{pmatrix} 1 & 0 \\ 0 & \rho_d \end{pmatrix}\right)$$

It can be shown that $\begin{cases} \sigma_d = \frac{c}{\sqrt{1 - e^{-2 \cdot d \cdot \delta}}} \\ \rho_d = tanh(2 \cdot d \cdot \gamma) \end{cases}$

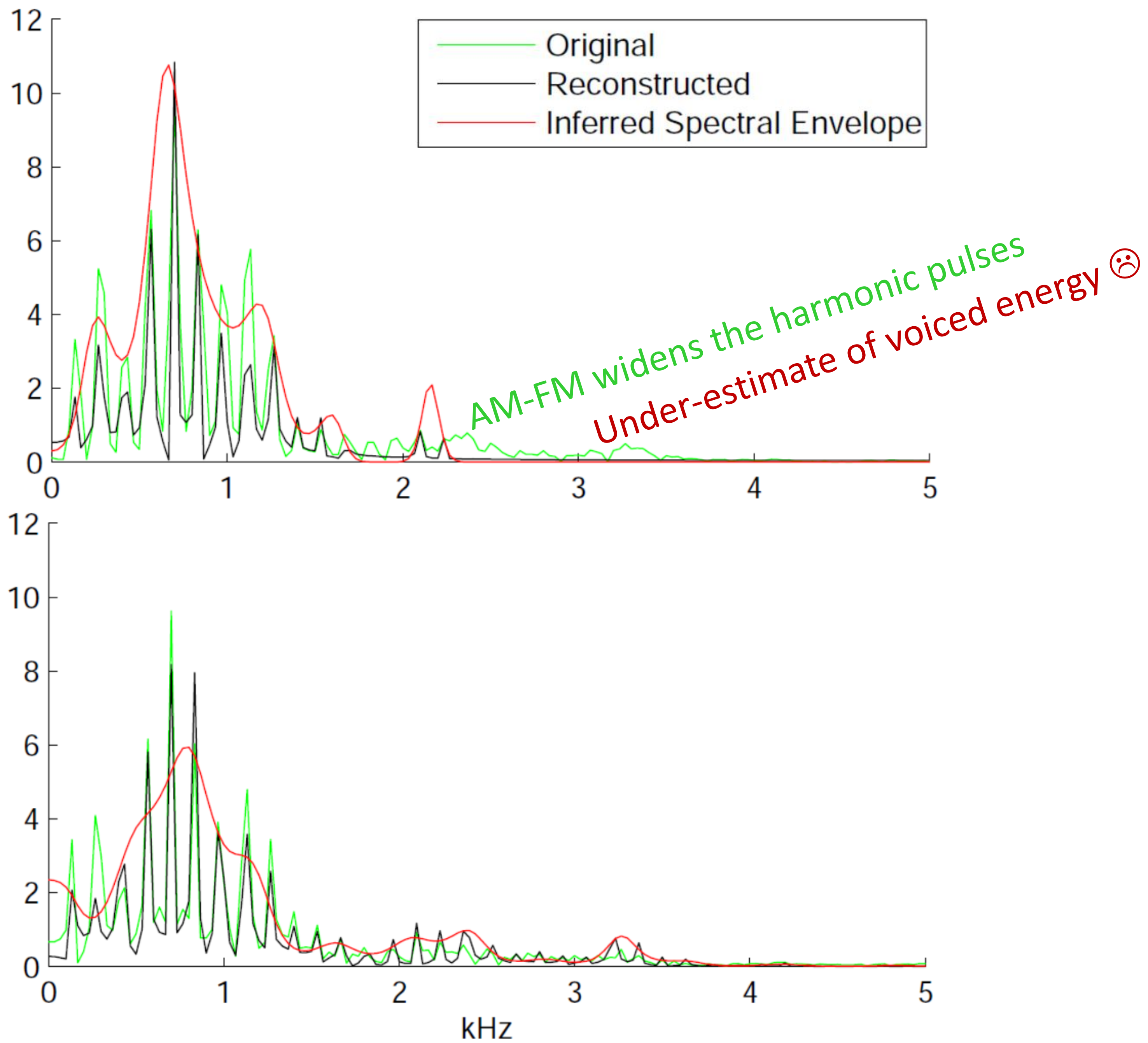
Time domain: $s[t] = v[t] + u[t] = v[t] + (b \cdot e_u[t]) * h[t]$

Frequency domain: $\vec{s} = vec(\omega_0, \tau, \vec{g}, \hat{h}; \delta, \gamma) + b \cdot vec\left[DFT[h[t]]\right] \boxdot vec[DFT[WGN]]$

Hidden variables: $z = \{a, b, \omega_0, \tau, \vec{g}, \hat{h}; \delta, \gamma\} \in R^{31+2}$

MAP inference $p(z|\vec{s}) \propto p(\vec{s}|z)p(z)$ by Monte Carlo sampling and L-BFGS search.

Experiment - Reconstruction of Voiced Speech with Heavy AM/FM Effect



PAT – Summary

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

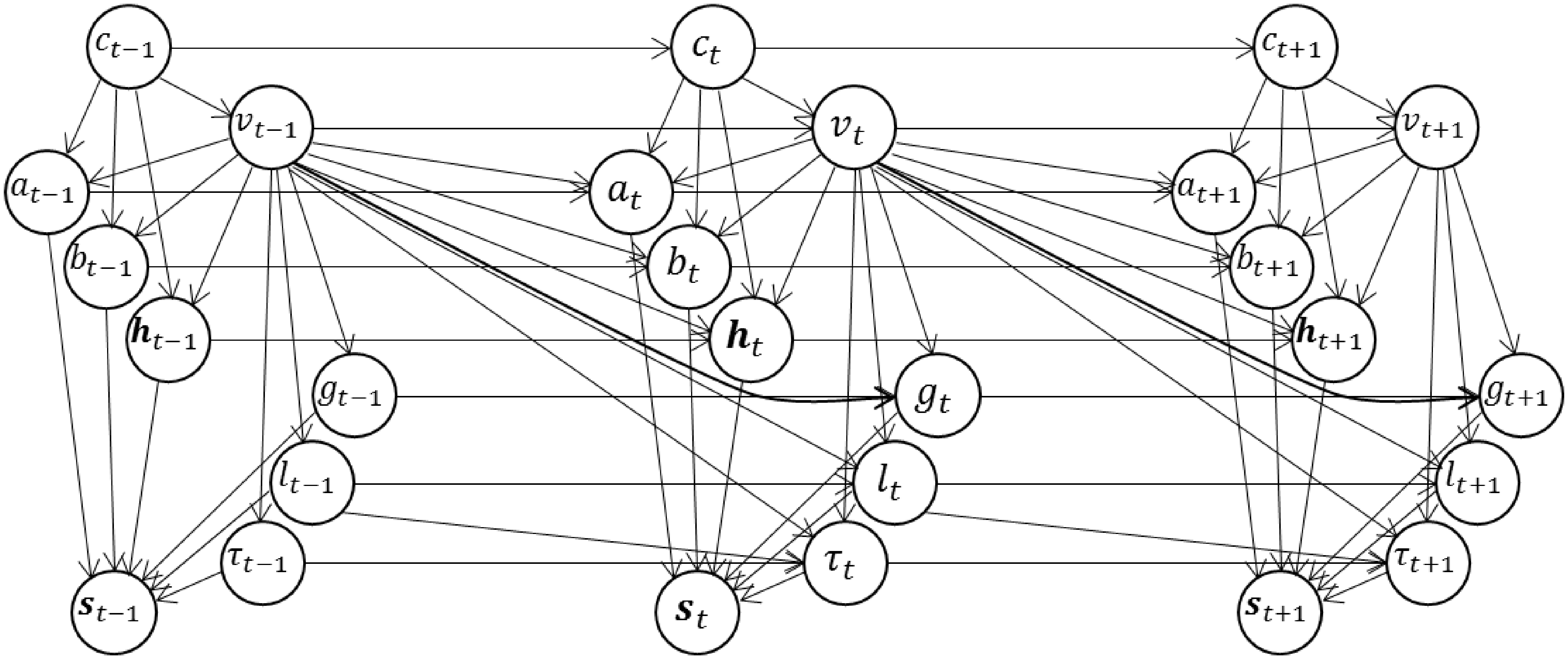
UTML TR 2006–004

To Recognize Shapes, First Learn to Generate Images

Geoffrey Hinton
Department of Computer Science, University of Toronto

PAT – Future work

- PAT: On the way …
  - A sequential inference algorithm for nonlinear state-space model
  - Large scale experiments



What is this talk about?

- Brief introduction to SPMI lab

- Motivation
- Probabilistic Acoustic Tube (PAT) Model, AISTATS 2012, ICASSP 2014.
- Random field approach to language modeling, ACL 2015.

Content

Random Field Language Models (RFLMs) – brand new

- State-of-the-art LMs - review
    - N-gram LMs
    - Neural network LMs
- Motivation - why
- Model formulation - what
- Model Training - breakthrough
- Experiment results - evaluation
- Summary

N-gram LMs

- Language modeling (LM) is to determine the joint probability of a sentence, i.e. a word sequence.
- Dominant: Conditional approach

Current word

All previous words/history

$$p(x_1, x_2, \cdots, x_l) = \prod_{i=1}^{l} p(x_i | x_1, \cdots, x_{i-1})$$

Previous $n-1$ words

$$\approx \prod_{i=1}^{l} \underline{p(x_i | x_{i-n+1}, \cdots, x_{i-1})}$$

- Using Markov assumption leads to the N-gram LMs
  - One of the state-of-the-art LMs

Neural network LMs

- Another state-of-the-art LMs

history

$x_1, \cdots, x_{i-1}$ → Neural Network → $\phi[x_1, \cdots, x_{i-1}] \triangleq \phi \in R^h$

$$p(x_i | x_1, \cdots, x_{i-1}) \approx p(x_i | \phi[x_1, \cdots, x_{i-1}])$$

$$p(x_i = k | x_1, \cdots, x_{i-1}) \approx \frac{\phi^T w_k}{\sum_{k=1}^{V} \phi^T w_k}$$ where V is lexicon size, $w_k \in R^h$

☹ Computational very expensive in both training and testing [1]

e.g. $V = 10k{\sim}100k, h = 250$

[1] Partly alleviated by using un-normalized models, e.g. through noise contrastive estimation training.

RFLMs – Motivation (1)

$$p(x_1, x_2, \cdots, x_l) = ?$$

Dominant:

Conditional approach / Directed

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_l$$

Alternative:

Random field approach / Undirected

$$x_1 - x_2 - x_3 - \cdots - x_l$$

☹ Difficulty in model training

☺ A rule in language cognition: employ context for reading and writing

The cat is on the table.

The cat is in the house.

☺ Breakthrough in training with a number of innovations

Fixed-dimensional (e.g. image) -> Trans-dimensional (sequential modeling)

RFLMs – Motivation (2)

- Drawback of N-gram LMs
  - N-gram is only one type of linguistic feature/property/constraint
  - meeting on Monday

$$P(w_i = Monday | w_{i-2} = meeing, w_{i-1} = on)$$

  - What if the training data only contain ‘meeting on Monday’ ?
  - New feature ‘meeting on DAY-OF-WEEK’, using class
  - New feature ‘party on *** birthday’, using skip
  - New features ….
- Jelinek 1995: put language back into language modeling

RFLMs – Formulation

- Intuitive idea
  - Features ($f_i, i = 1,2, \dots, F$) can be defined arbitrarily, beyond the n-gram features.
  - Each feature brings a contribution to the sentence probability $p(x)$
- Formulation

$$p(x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{F} \lambda_i f_i(x) \right), x \triangleq (x_1, x_2, \cdots, x_l)$$

$$f_i(x) = \begin{cases} 1, & \text{'meeting on DAY−OF−WEEK' appears in } x \quad \Rightarrow \lambda_i \text{ is activated} \\ 0, & \text{Otherwise} \quad \Rightarrow \lambda_i \text{ is removed} \end{cases}$$

☺ More flexible features, beyond the n-gram features, can be well supported in RFLMs.
☺ Computational very efficient in computing sentence probability.

WSME - Introduction

- Whole-sentence maximum entropy (WSME)
  - Rosenfeld, Chen, Zhu. “Whole-sentence exponential language models: a vehicle for linguistic-statistical integration”. Computer Speech & Language, 2001.

$$p(x;\lambda) = \frac{1}{Z(\lambda)} \exp[\lambda^T f(x)]$$

- The empirical results of previous WSME models are not satisfactory
  - After incorporating lexical and syntactic information, 1% and 0.4% respectively in perplexity and in WER is reported for the resulting WSEM (Rosenfeld et al., 2001).
  - Amaya and Benedi. “Improvement of a whole sentence maximum entropy language model using grammatical features”, ACL 2001.
  - Ruokolainen, Alumae, Dobrinkat. “Using dependency grammar features in whole sentence maximum entropy language model for speech recognition”. HLT 2010.

WSME – Difficulty in model training

$$p(x;\lambda) = \frac{1}{Z(\lambda)} \exp[\lambda^T f(x)]$$

Normalization constant:

$$Z(\lambda) = \sum_x \exp\left(\sum_{i=1}^{F} \lambda_i f_i(x)\right)$$

- Maximum-likelihood training

$$\frac{\partial LogLikelihood}{\partial \lambda} = E_{\tilde{p}(x)}[f_i(x)] - E_{p(x;\lambda)}[f_i(x)] = 0$$

Expectation under empirical distribution $\tilde{p}(x)$

Expectation under model distribution $p(x;\lambda)$

RFLMs vs WSME

- Whole-sentence maximum entropy (WSME)

$$p(l, x^l; \lambda) = \frac{1}{Z(\lambda)} \exp[\lambda^T f(x^l)], \qquad x = (l, x^l), \qquad x^l \triangleq (x_1, x_2, \cdots, x_l)$$

Essentially a mixture distribution with unknown weights (differ from each other greatly, 10^{40}) !
Poor sampling → poor estimate of gradient → poor fitting

$$p(l, x^l; \lambda) = \frac{Z_l(\lambda)}{Z(\lambda)} \cdot \frac{1}{Z_l(\lambda)} \cdot \exp[\lambda^T f(x^l)], \; Z_l(\lambda) = \sum_{x^l} \exp[\lambda^T f(x^l)]$$

RFLMs vs WSME

- Whole-sentence maximum entropy (WSME)

$$p(l, x^l; \lambda) = \frac{1}{Z(\lambda)} \exp[\lambda^T f(x^l)], \qquad x \triangleq (l, x^l), \qquad x^l \triangleq (x_1, x_2, \cdots, x_l)$$

Essentially a mixture distribution with unknown weights (differ from each other greatly, 10^{40}) !
Poor sampling → poor estimate of gradient → poor fitting

$$p(l, x^l; \lambda) = \frac{Z_l(\lambda)}{Z(\lambda)} \cdot \frac{1}{Z_l(\lambda)} \cdot \exp[\lambda^T f(x^l)], \; Z_l(\lambda) = \sum_{x^l} \exp[\lambda^T f(x^l)]$$

- We propose a trans-dimensional RF model

$$p(l, x^l; \lambda) = \pi_l \cdot \frac{1}{Z_l(\lambda)} \cdot \exp[\lambda^T f(x^l)], \qquad l = 1, \cdots, m$$

Empirical length probabilities in the training data
Serve as a control device to improve sampling from multiple distributions !

Introduction to Stochastic Approximation (SA)

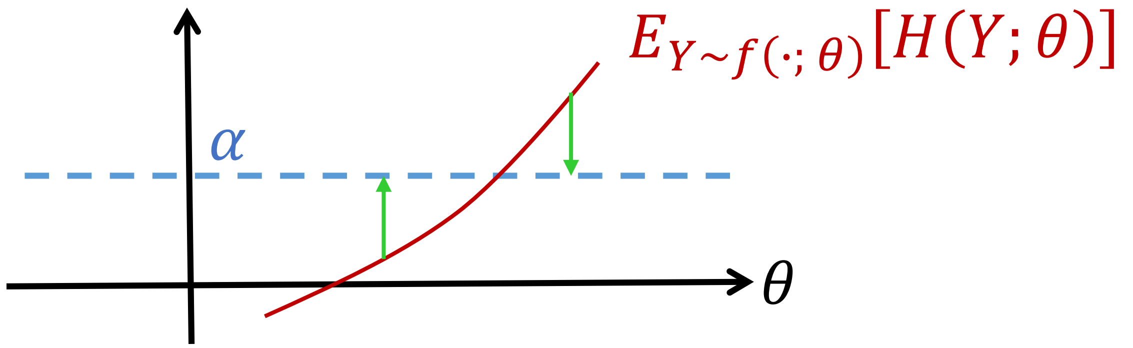
Problem: The objective is to find a solution θ to $E_{Y\sim f(\cdot;\,\theta)}[H(Y;\theta)] = \alpha$, where $\theta \in R^d$, noisy observation $H(Y;\theta) \in R^d$

Method:

(1) Generate $Y_t \sim K(Y_{t-1}, \cdot\,; \theta_{t-1})$, a Markov transition kernel that admits $f(\cdot;\ \theta_{t-1})$ as the invariant distribution.

(2) Set $\theta_t = \theta_{t-1} + \gamma_t\{\alpha - H(Y_t;\theta_{t-1})\}$

$$e.g.\ \gamma_t = \frac{1}{t_0 + t}$$



Robbins and Monro (1951). A stochastic approximation method. Ann. Math. Stat.
Chen (2002), Stochastic Approximation and Its Applications, Kluwer Academic Publishers.

Apply SA to RFLM training

- The trans-dimensional RF model

$$p(l, x^l; \lambda) = \pi_l \cdot \frac{1}{Z_l(\lambda)} \cdot \exp[\lambda^T f(x^l)] \quad (1)$$

$$E_{\tilde{p}(x)}[f_i(x)] - E_{p(x;\lambda)}[f_i(x)] = 0, \qquad x \triangleq (l, x^l)$$

- Consider the joint distribution of the pair (l, x^l)

$$p(l, x^l; \lambda, \zeta) \propto \pi_l \cdot \frac{1}{e^{\zeta_l}} \cdot \exp[\lambda^T f(x^l)] \quad (2)$$

where ζ_l is hypothesized values of the true $\zeta_l^*(\lambda) = log Z_l(\lambda)$.

The marginal probability of length l is: $p(l; \lambda, \zeta) = \frac{\pi_l e^{-\zeta_l + \zeta_l^*(\lambda)}}{\sum_j \pi_l e^{-\zeta_j + \zeta_j^*(\lambda)}}$.

- SA is used to find $\zeta_l^* = \zeta_l^*(\lambda^*)$ and λ^* that solves

$$\begin{cases} \pi_l = p(l; \lambda, \zeta), \qquad l = 1, \cdots, m \\ 0 = E_{\tilde{p}(x)}[f_i(x)] - E_{p(l, x^l; \lambda, \zeta)}[f_i(x)] \end{cases}$$

Zhiqiang Tan. 2015. Optimally adjusted mixture sampling and locally weighted histogram. In Technical Report, Dept. of Statistics, Rutgers Univ.

RFLMs – Breakthrough in training (1)

- Propose Joint Stochastic Approximation (SA) Training Algorithm
  - Simultaneously updates the model parameters and normalization constants

Algorithm 1 Joint stochastic approximation

Input: training set

set initial values $\lambda^{(0)} = (0, \dots, 0)^T$ and $\zeta^{(0)} = \zeta^*(\lambda^{(0)}) - \zeta_1^*(\lambda^{(0)})$

for $t = 1, 2, \dots, t_{max}$ **do**

  set $B^{(t)} = \emptyset$

  set $(L^{(t,0)}, X^{(t,0)}) = (L^{(t-1,K)}, X^{(t-1,K)})$

  Step I: MCMC sampling

  for $k = 1 \to K$ **do**

    sampling (See Algorithm 3) $(L^{(t,k)}, X^{(t,k)}) = SAMPLE(L^{(t,k-1)}, X^{(t,k-1)})$

    set $B^{(t)} = B^{(t)} \cup \{(L^{(t,k)}, X^{(t,k)})\}$

  end for

  Step II: SA updating

  Compute $\lambda^{(t)}$ based on (13)

  Compute $\zeta^{(t)}$ based on (14) and (15)

end for

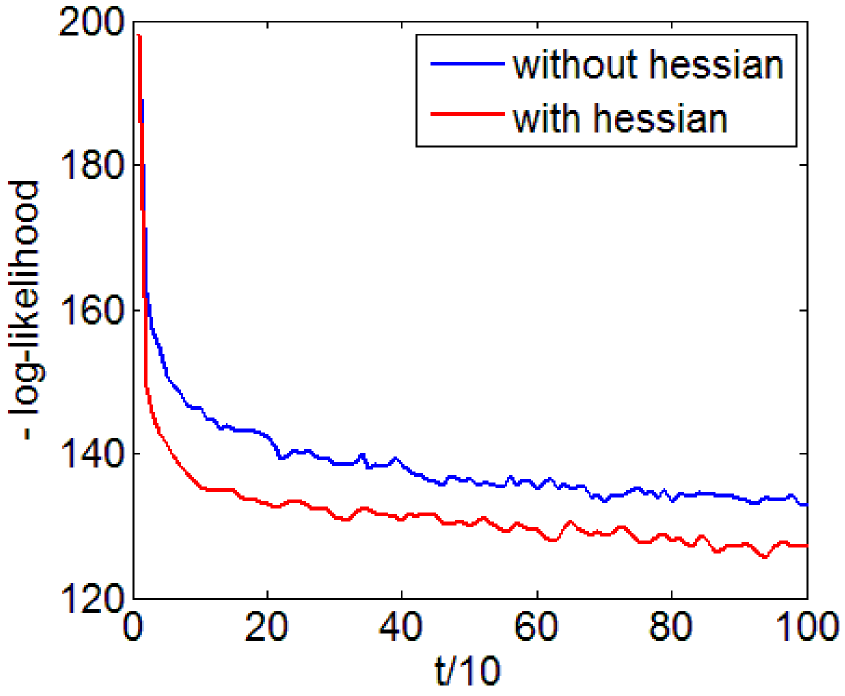
RFLMs – Breakthrough in training (2)

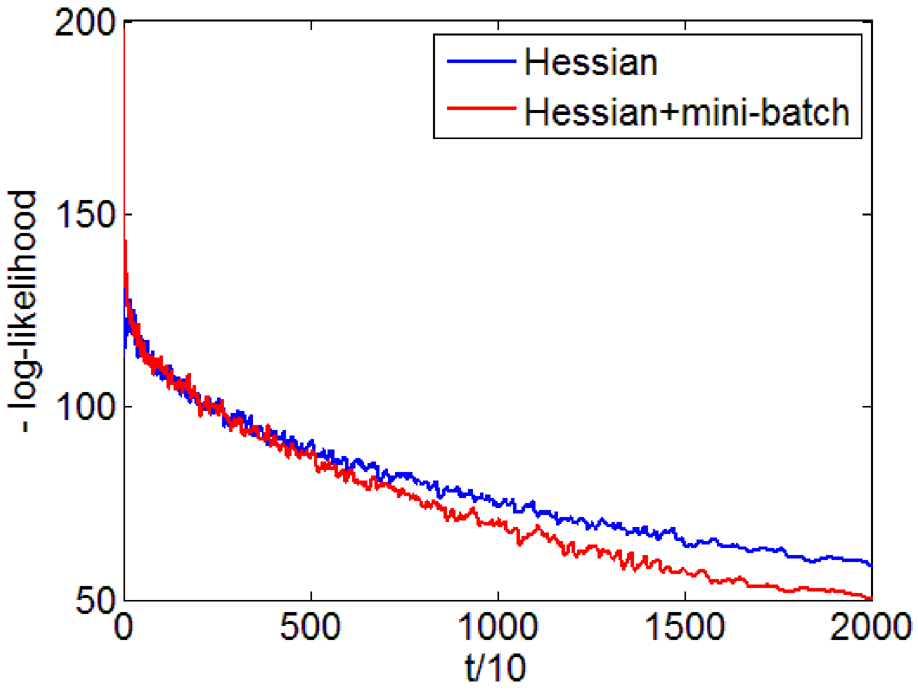
- Propose Trans-dimensional mixture sampling
  - Sampling from $p(l, x^l; \lambda, \zeta)$, a mixture of RFs on subspaces of different dimensions.
  - Formally like RJ-MCMC.

function SAMPLING($(L^{(t-1)}, X^{(t-1)})$)
  set $k = L^{(t-1)}$
  set $L^{(t)} = k$
  set $X^{(t)} = X^{(t-1)}$
  Stage I: Local jump
  generate $j \sim \Gamma(k, \cdot)$
  if $j = k + 1$ **then**
    generate $Y \sim g_{k+1}(y|X^{(t-1)})$ (equ.24)
    set $L^{(t)} = j$ and $X^{(t)} = \{X^{(t-1)}, Y\}$ with probability equ.22
  end if
  if $j = k - 1$ **then**
    set $L^{(t)} = j$ and $X^{(t)} = X^{(t-1)}_{1:k-1}$ with probability equ.23
  end if
  Stage II: Markov move
  for $i = 1 \to L^{(t)}$ **do**
    $a \sim p(L^{(t)}, \{X^{(t)}_{1:i-1}, \cdot, X^{(t)}_{i+1:L^{(t)}}\}; \Lambda, \zeta)$
    $X^{(t)}_i \leftarrow a$
  end for
  return $(L^{(t)}, X^{(t)})$
end function

RFLMs – Breakthrough in training (3)

- Exploit Hessian diagonal in SA
- Introduce training set mini-batching





Improve the convergence !

Content

Random Field Language Models (RFLMs) – brand new

- State-of-the-art LMs - review
  - N-gram LMs
  - Neural network LMs
- Motivation - why
- Model formulation - what
- Model Training - breakthrough
- Experiment results - evaluation
- Summary

Experiment setup

- LM Training — Penn Treebank portion of WSJ corpus
  - Vocabulary : 10K words
  - Training data : 887K words, 42K sentences
  - Development data : 70K words
  - Testing data : 82K words
- Test speech — WSJ'92 set (330 sentences)
  - By rescoring of 1000-best lists
- Various LMs
  - KN4 (Kneser-Ney)
    - 4gram LMs with modified Kneser-Ney smoothing
  - RNNLMs (Recurrent Neural Network LMs)
    - Trained by the RNNLM toolkit of Mikolov
    - The dimension of hidden layer = 250. Mini-batch size=10, learning rate=0.1, BPTT steps=5.
    - 17 sweeps are performed before stopping (takes about 25 hours). No word classing is used.
  - RFLMs
    - A variety of features based on word and class information

Feature Definition

Type	Features
w	$(w_{-3}w_{-2}w_{-1}w_0)(w_{-2}w_{-1}w_0)(w_{-1}w_0)(w_0)$
c	$(c_{-3}c_{-2}c_{-1}c_0)(c_{-2}c_{-1}c_0)(c_{-1}c_0)(c_0)$
ws	$(w_{-3}w_0)(w_{-3}w_{-2}w_0)(w_{-3}w_{-1}w_0)(w_{-2}w_0)$
cs	$(c_{-3}c_0)(c_{-3}c_{-2}c_0)(c_{-3}c_{-1}c_0)(c_{-2}c_0)$
wsh	$(w_{-4}w_0)\ (w_{-5}w_0)$
csh	$(c_{-4}c_0)\ (c_{-5}c_0)$
cpw	$(c_{-3}c_{-2}c_{-1}w_0)\ (c_{-2}c_{-1}w_0)(c_{-1}w_0)$

w / c : the word/class ngram features up to order 4
ws / cs : the word/class skipping ngram features up to order 4
wsh / csh : the higher-order word/class features
cpw : the crossing class-predict-word features up to order 4

Word Error Rate (WER) results for speech recognition

model	WER	PPL (± std. dev.)	#feat
KN4	8.71	295.41	1.6M
RNN	7.96	256.15	5.1M
RFLMs (100c)			
w+c	8.56	268.25±3.52	2.2M
w+c+ws+cs	8.16	265.81±4.30	4.5M
w+c+ws+cs+cpw	8.05	265.63±7.93	5.6M
w+c+ws+cs+wsh+csh	8.03	276.90±5.00	5.2M
RFLMs (200c)			
w+c	8.46	257.78±3.13	2.5M
w+c+ws+cs	8.05	257.80±4.29	5.2M
w+c+ws+cs+cpw	**7.92**	264.86±8.55	6.4M
w+c+ws+cs+wsh+csh	**7.94**	266.42±7.48	5.9M
RFLMs (500c)			
w+c	8.72	261.02±2.94	2.8M
w+c+ws+cs	8.29	266.34±6.13	5.9M

Table 3: The WERs and PPLs on the WSJ'92 test data. "#feat" denotes the feature number. Different RFLMs with class number 100/200/500 are reported (denoted by "100c"/"200c"/"500c")

- Encouraging performance
  - The RFLM using the "w+c+ws+cs+cpw" features with class number 200 performs comparable to the RNNLM, but is computationally more efficient in computing sentence probability.

    Re-ranking of the 1000-best list for a sentence takes 0.16 sec. vs 40 sec.
  - The WER relative reduction is 9.1% compared with the KN4, and 0.5% compared with the RNNLM.
- Efficient in training
  - Training the RFLM with up to **6 million** features, takes 15 hours.

Summary

Contribution

- Breakthrough in training with a number of innovations.
- Successfully train RFLMs and make performance improvements.

	Computation efficient in training	Computation efficient in test	Bidirectional context	Flexible features	Performance
N-gram LMs	✓	✓	✗	✗	✗
Neural network LMs	✗	✗	✗	✓	✓
RFLMs	✗	✓	✓	✓	✓

Future work

- Train RFLMs with richer features on larger-scale corpus.
- Features selection strategy such as L1 regularization.

Thanks:

Yang Zhang, Bin Wang, Mark Hasegawa-Johnson, Zhiqiang Tan.

Thanks for your attention !