Machine Intelligence with Speech and Language
语音和语言的机器智能

欧智坚

清华大学电子工程系

语音处理与机器智能实验室

Speech Processing and Machine Intelligence (SPMI) Lab

2016-1-19

内容安排

1. State-of-the-art – Where we are
2. Basic thoughts – What we believe
3. Highlight – What we do
    - Probabilistic Acoustic Tube (PAT) Model
    - Random field approach to language modeling
4. Summary

1955，McCarthy, **人工智能**：制造智能机器的科学与工程

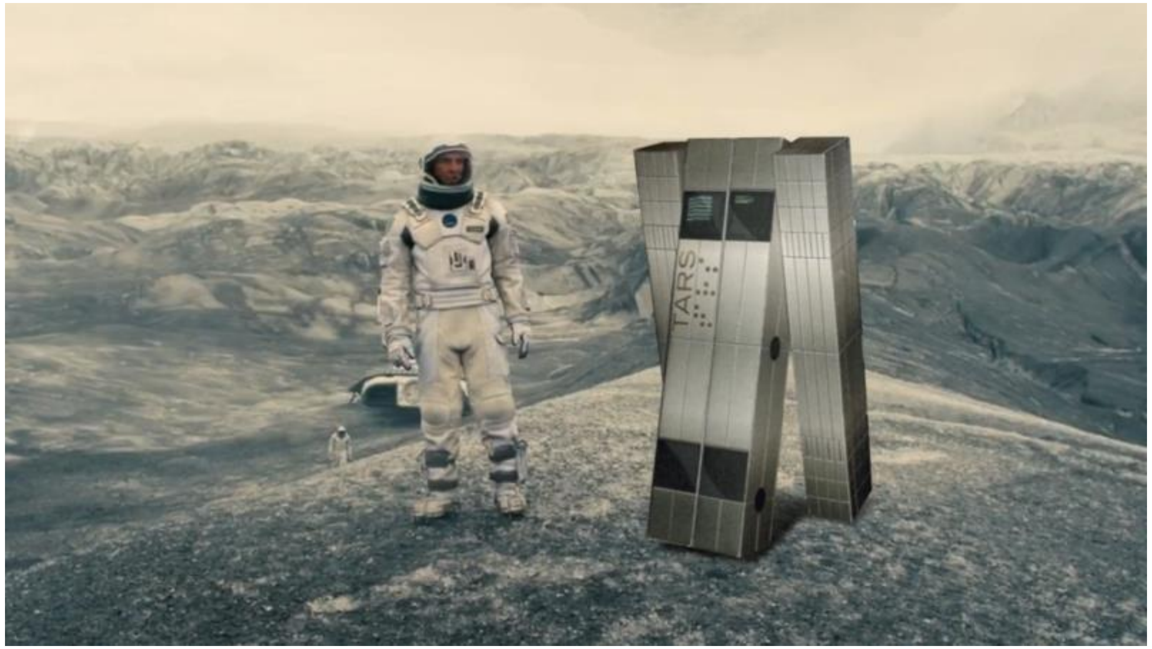
- 语音和语言—是人类智能最突出的表现

2006，Hinton,《Science》, 深度神经网络（DNN）

- 语音识别、图像识别、自然语言理解等

2013

欧盟Human Brain Project（十年），
美国BRAIN（十年），
Google、Microsoft、Facebook，
BAT - 百度、阿里巴巴、腾讯，
……



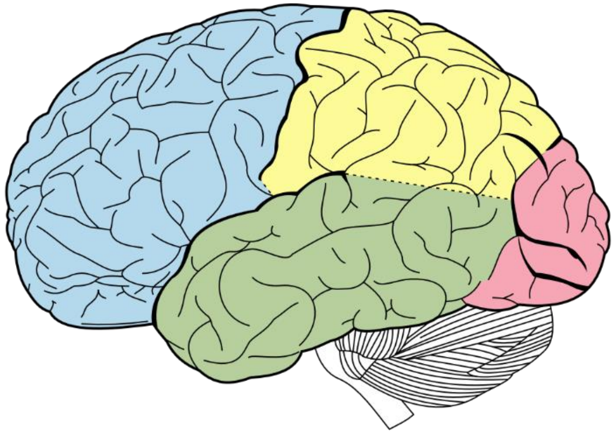
2015，现代任何高级的机器在**语音和语言的智能**表现仍远不如人类。

技术现状

机器识别准确率、功耗效率均远逊于人类 ☹

语音识别任务	机器错误率	人错误率
安静环境, 朗读新闻	3%	0.9%
常见现实情况：嘈杂环境(10db), 自然谈话	20%	4%

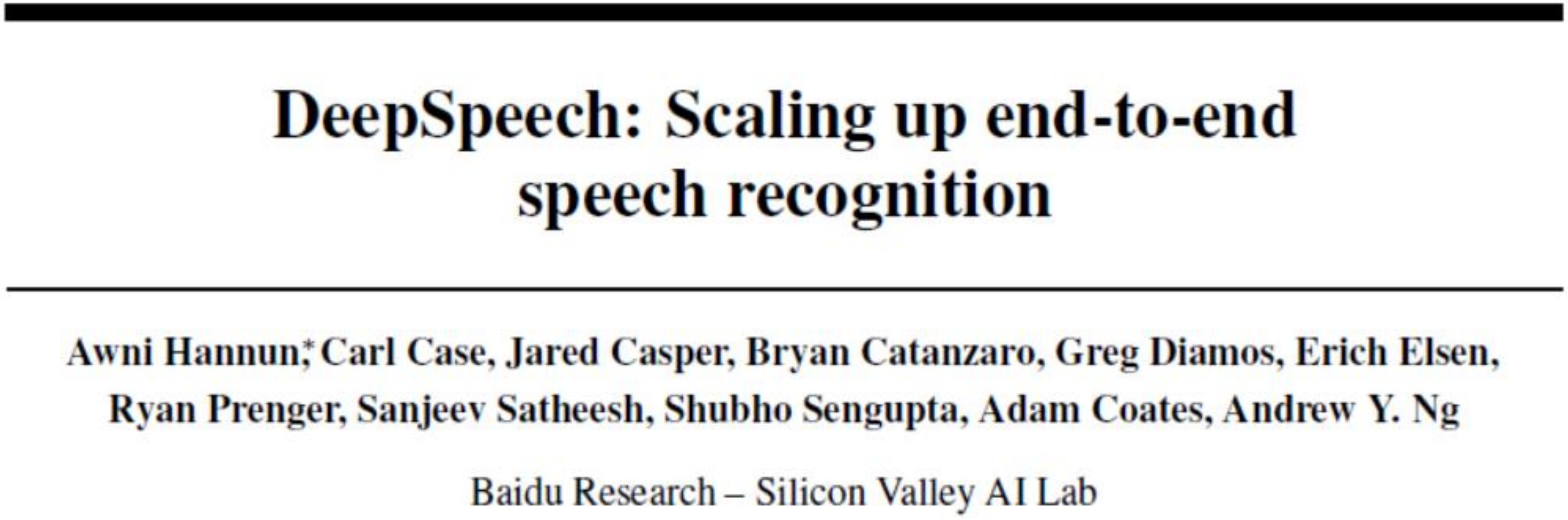
目前DNN技术的最大效用—30%的错误率相对降低 ↓ 14%



Human brain, 20 watts

10^6倍

Google Brain, 1000 servers (16000 CPU cores)

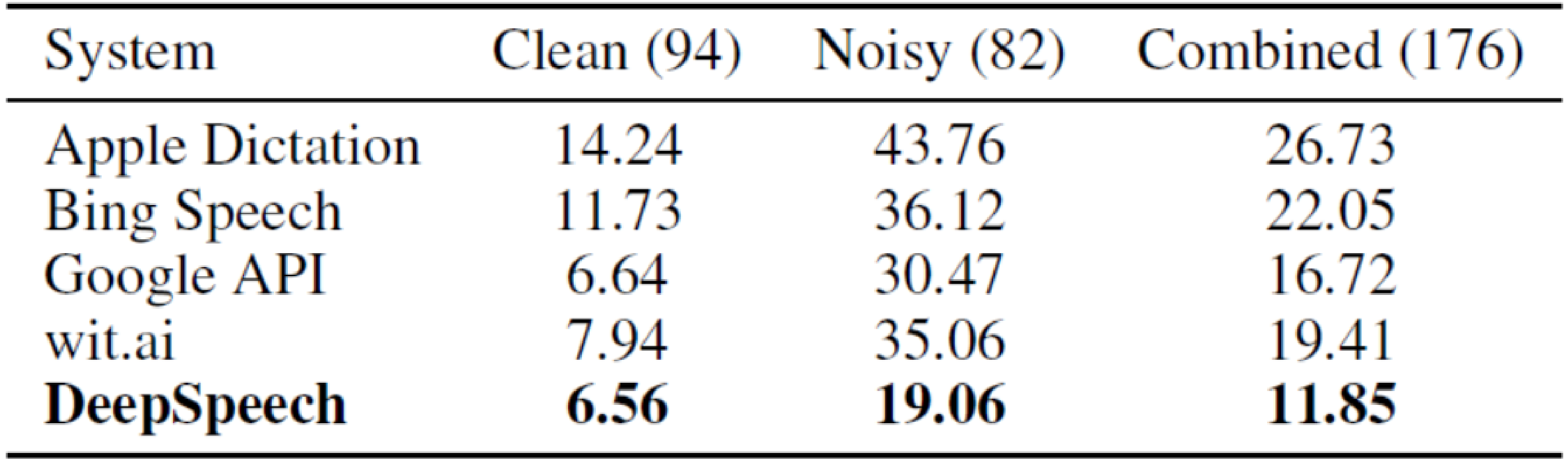


DeepSpeech: Scaling up end-to-end speech recognition

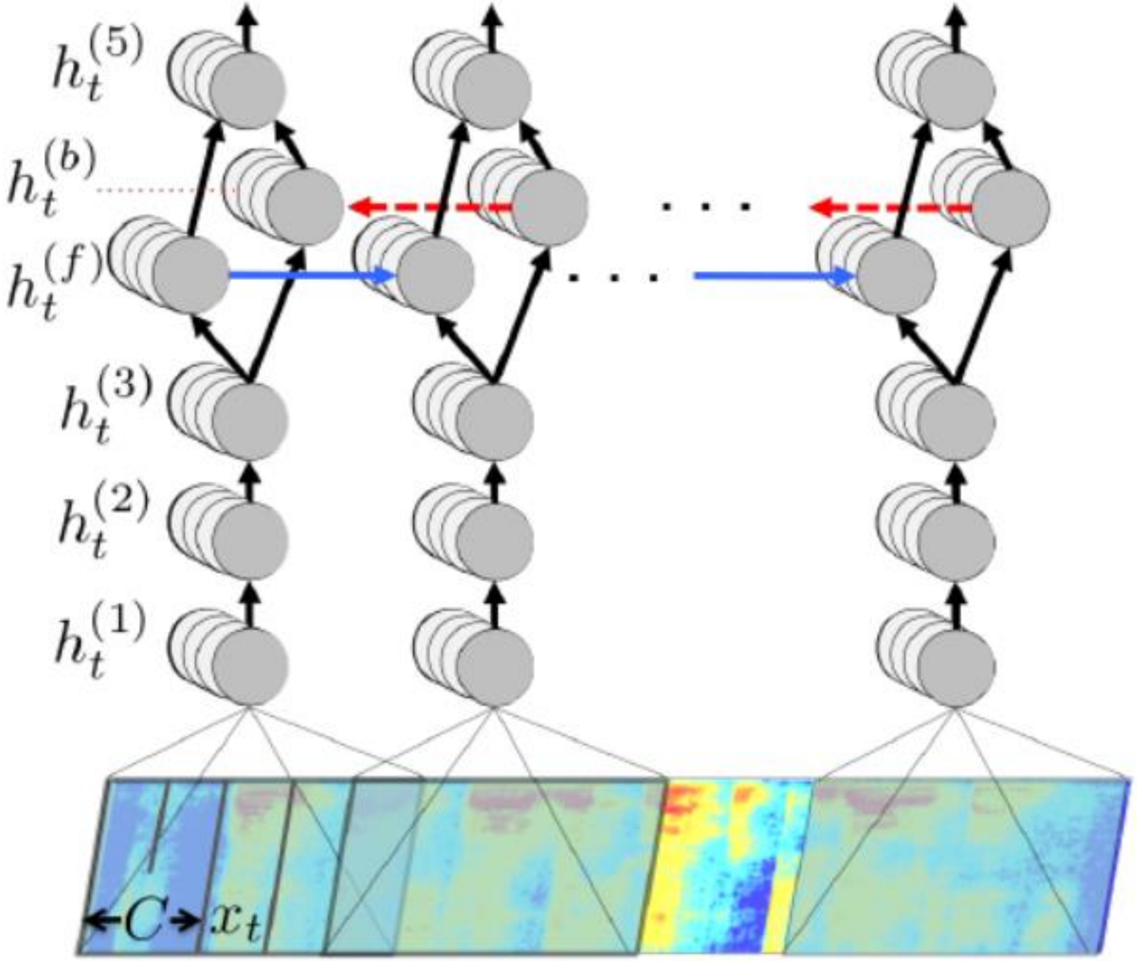
Awni Hannun[*], Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen,
Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng

Baidu Research – Silicon Valley AI Lab

March, 2015



System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
DeepSpeech	**6.56**	**19.06**	**11.85**



An ensemble of 6 RNNs each with 5 hidden layers of 2560 neurons

100,000 hours speech ÷ (365*12 hour) = 23 year

研究历史

- 在1998年国家“863”汉语连续语音识别评测中，我们研制的听写机系统第三次蝉联冠军
    - 对于1993、1994年《人民日报》语料的朗读语音数据，字正确率达到98.7%。

- Pierce objects to
    - attempts to sell science as something other than it is (e.g., applications),
    - as well as attempts to misrepresent progress with misleading demos and/or mindless metrics (such as the kinds of evaluations that are routinely performed today).

研究现状

- Kenneth Church,《钟摆摆得太远》(A Pendulum Swung Too Far), 2007.

- On the positive side, pattern recognition makes it possible to make progress on applications by finessing many hard scientific questions.
- On the other hand, pattern recognition makes it hard to make progress on the key scientific questions because short-term finesses distracts long-term science.

  短期的取巧分散了领域的精力，无法顾及真正有意义的长远科学目标。

Paths to the future

内容安排

1. State-of-the-art – Where we are
2. Basic thoughts – What we believe
3. Highlight – What we do
   - Probabilistic Acoustic Tube (PAT) Model
   - Random field approach to language modeling
4. Summary

Basic thought 1 – Empiricism and Rationalism

Kenneth Church

Education: MIT (1974-1983)

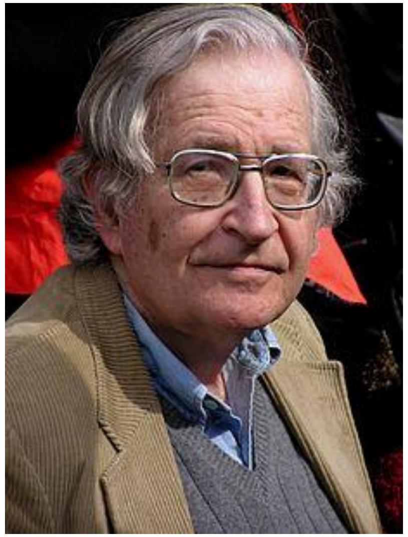
AT&T (1983-2003), Microsoft (2003-2009), JHU (2009-2011), IBM (2011-

《钟摆摆得太远》(A Pendulum Swung Too Far)

发表于Linguistic Issues in Language Technology – LiLT, May 2007.

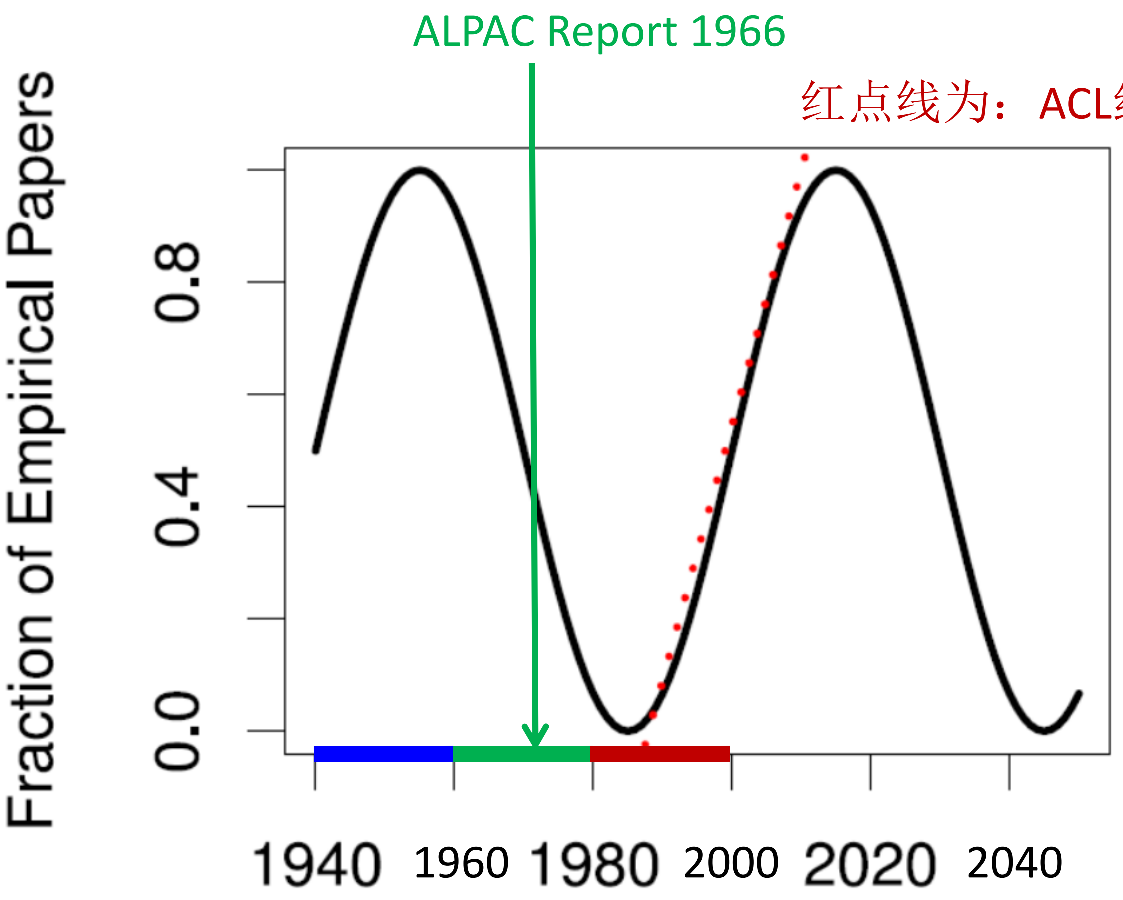
John R. Pierce 1910-2002
Bell, JPL, Stanford
ALPAC Report 1966

Marvin Minsky 1928 –
"A founding father of AI"
Turning Prize 1969



Noam Chomsky 1928 –
"the father of modern linguistics,"

Basic thought 1 – Empiricism and Rationalism



ALPAC Report 1966

红点线为：ACL经验主义文章数的实际拟合

1950s: Empiricism
(Shannon, Skinner, Firth, Harris)

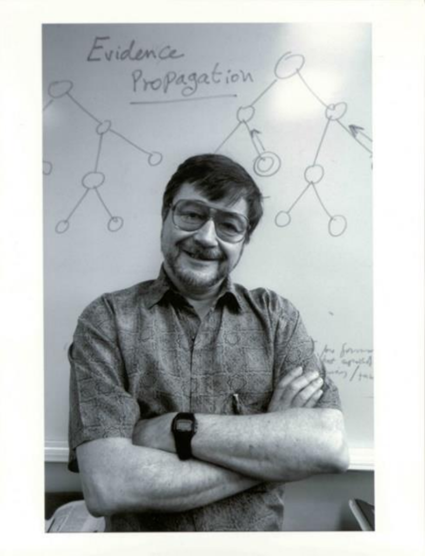
1970s: Rationalism
(Chomsky, Minsky)

1990s: Empiricism
(IBM, Bell)

Cited from : Kenneth Church, 《钟摆摆得太远》(A Pendulum Swung Too Far), 2007

Basic thought 2 – Deal with uncertainty

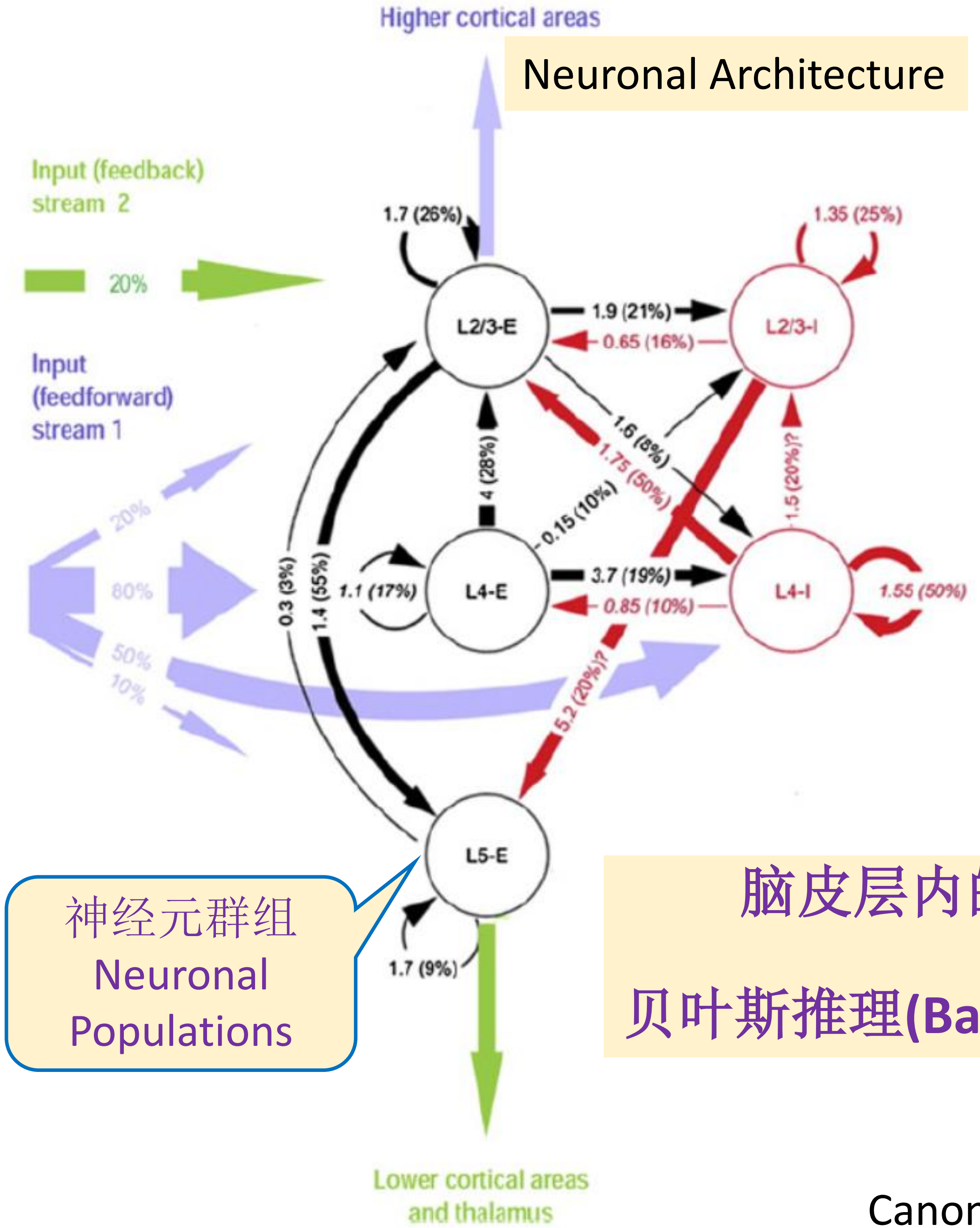
- 智能的本质是什么？这是至今为止仍很难回答的基础科学问题。
- 处理不确定性是智能的一种重要表现，得到了众多的研究者的充分肯定和践行。



- **2012年图灵奖: UCLA的Judea Pearl教授**
  - 开创性的工作——贝叶斯网络、消息传递概率推理；
  - Revolutionized AI；
  - 让人们认识到**处理不确定性**对**建造人工智能系统**的作用。

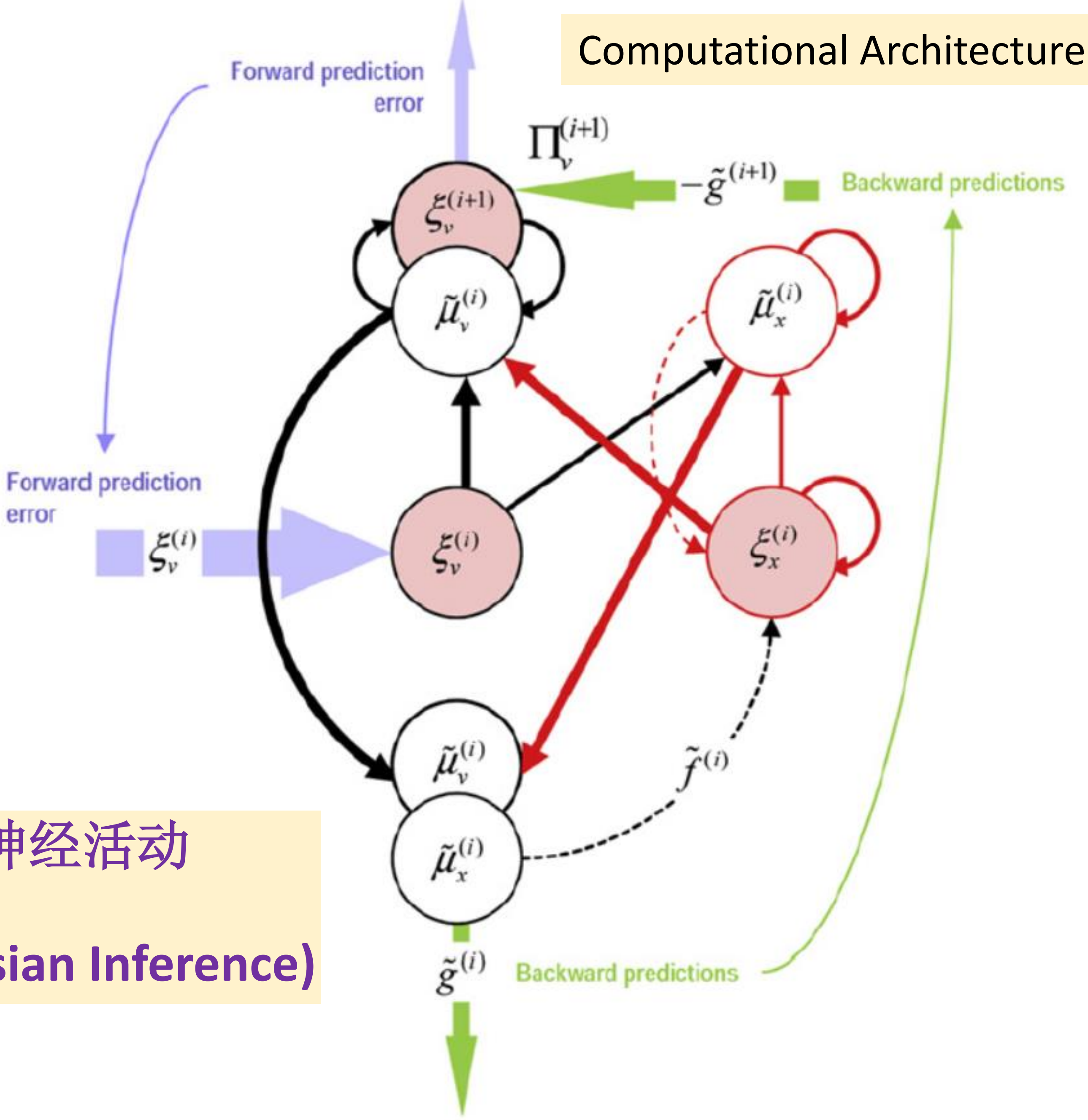
- **计算神经科学：贝叶斯心智（Bayesian Mind）**
  - 人类思维行为与贝叶斯分析非常相近 - Griffiths, Tenenbaum, 2006.
  - 语言和认知是随机现象 - Chater, Manning, 2006.

Neuronal Architecture



神经元群组
Neuronal
Populations

Computational Architecture

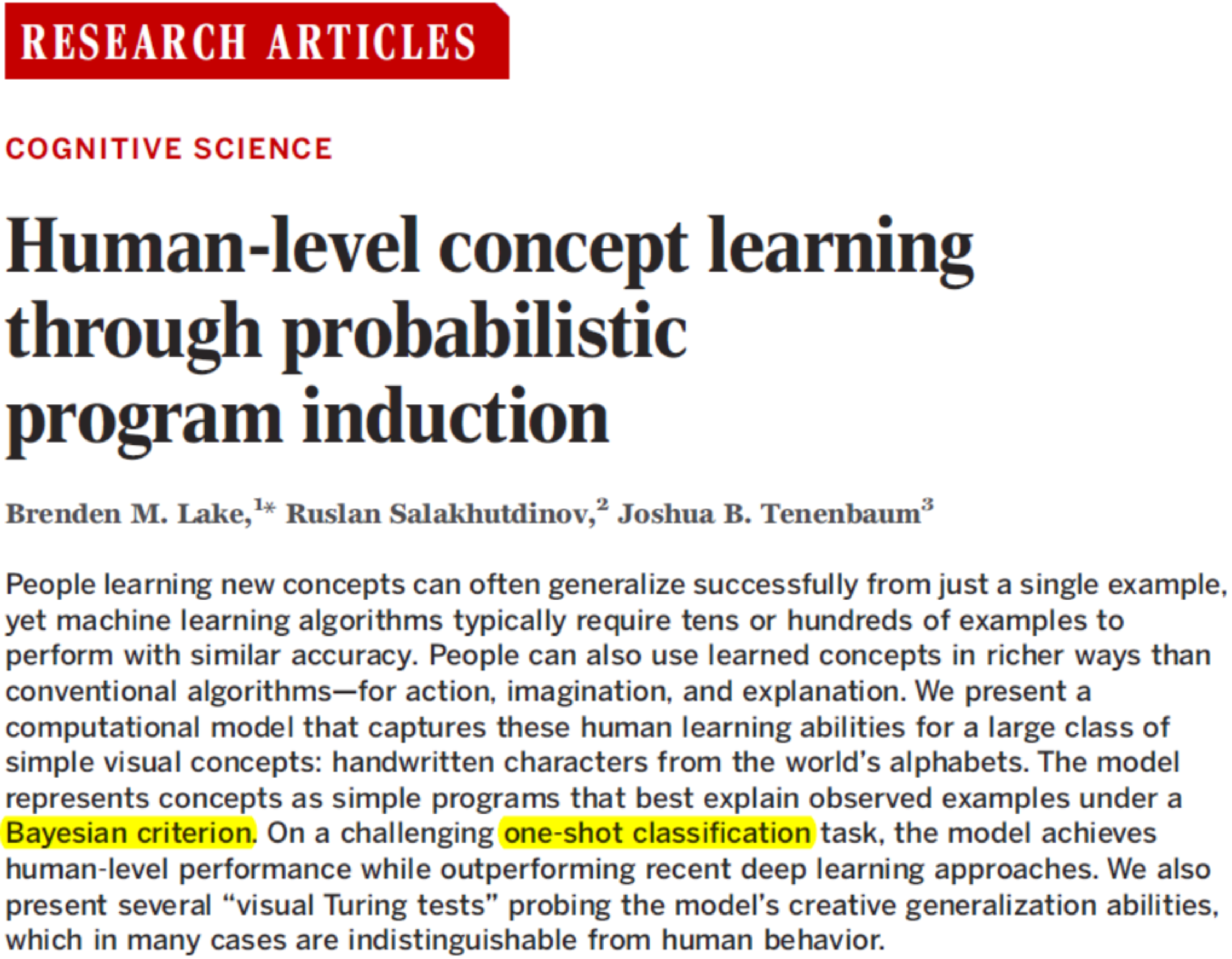


脑皮层内的神经活动

⇕

贝叶斯推理(Bayesian Inference)

Canonical Microcircuits for Predictive Coding, Neuron, 2012.



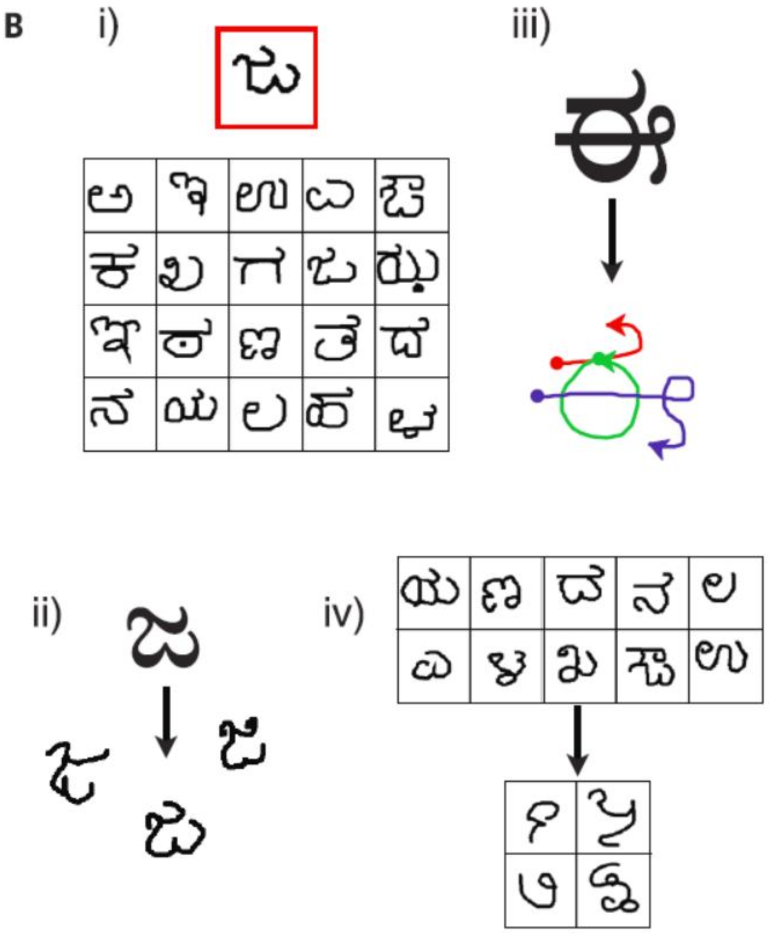
RESEARCH ARTICLES

COGNITIVE SCIENCE

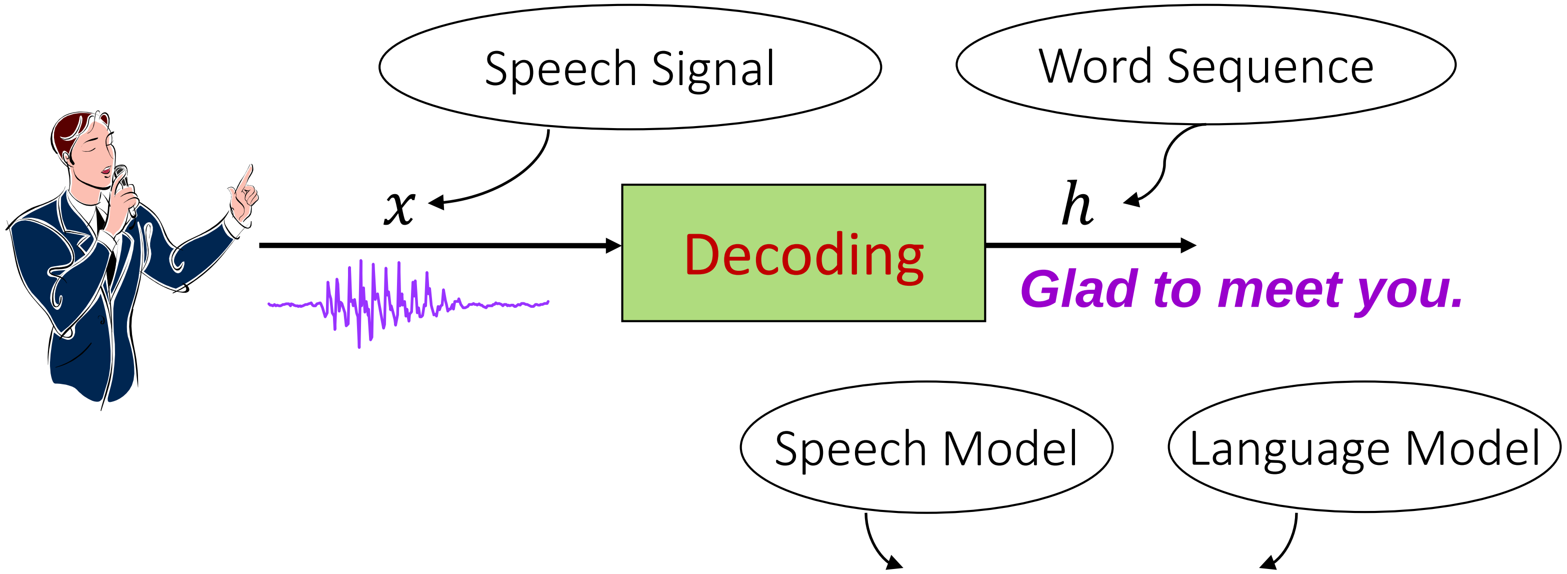
Human-level concept learning through probabilistic program induction

Brenden M. Lake,[1*] Ruslan Salakhutdinov,[2] Joshua B. Tenenbaum[3]

People learning new concepts can often generalize successfully from just a single example, yet machine learning algorithms typically require tens or hundreds of examples to perform with similar accuracy. People can also use learned concepts in richer ways than conventional algorithms—for action, imagination, and explanation. We present a computational model that captures these human learning abilities for a large class of simple visual concepts: handwritten characters from the world's alphabets. The model represents concepts as simple programs that best explain observed examples under a Bayesian criterion. On a challenging one-shot classification task, the model achieves human-level performance while outperforming recent deep learning approaches. We also present several "visual Turing tests" probing the model's creative generalization abilities, which in many cases are indistinguishable from human behavior.



Probabilistic Modeling of Speech and Language



$$\underset{h}{\operatorname{argmax}}\, P(h|x) = \underset{x}{\operatorname{argmax}} \frac{P(x|h)P(h)}{\cancel{P(x)}}$$

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

内容安排

1. State-of-the-art – Where we are
2. Basic thoughts – What we believe
3. Highlight – What we do
   - Probabilistic Acoustic Tube (PAT) Model
   - Random field approach to language modeling
4. Summary

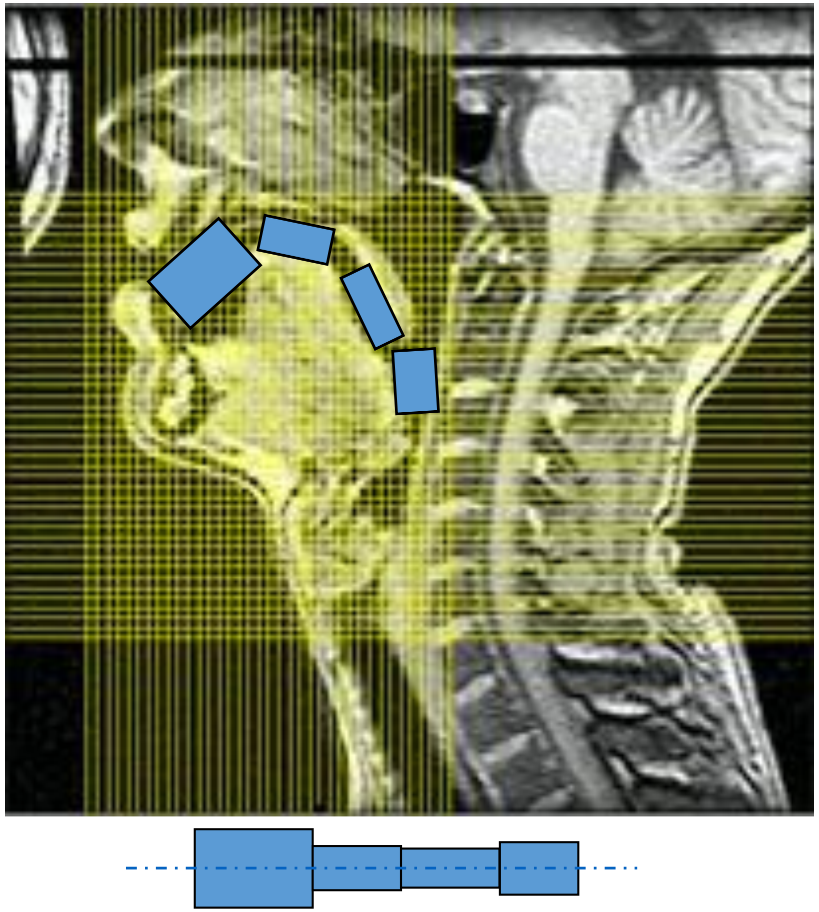
Our trial-and-error efforts

- 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, WASPAA 2015

Motivation

Q1: What is the basic physical model of speech production ?

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



Motivation

Q2: 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

Important speech elements

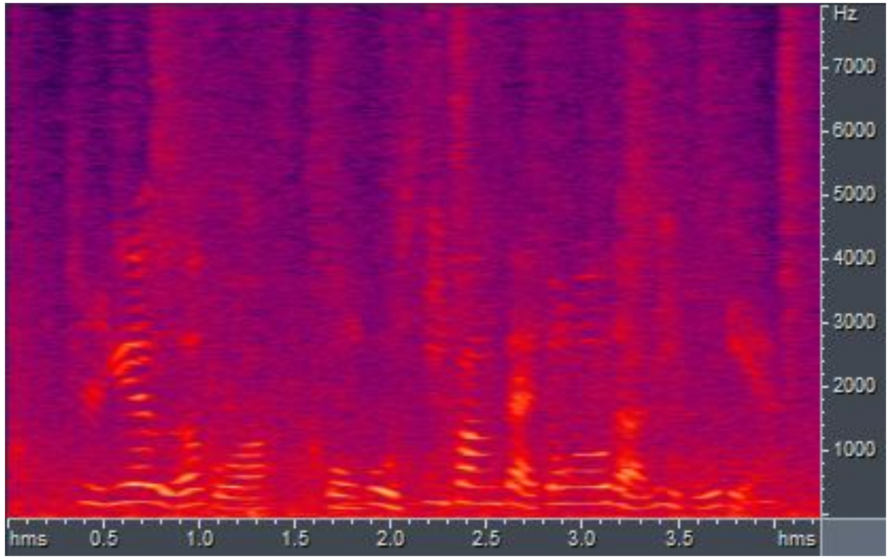
Pitch

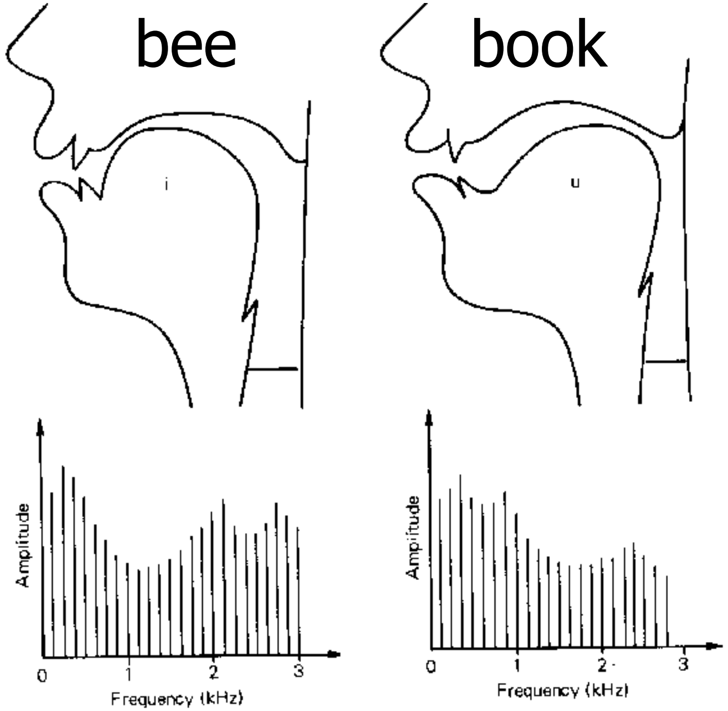
Glottal source

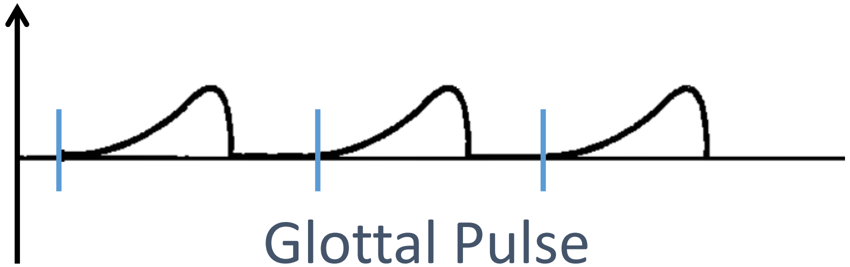
Vocal tract response

Aspiration noise

Phase







Glottal Pulse

Motivation

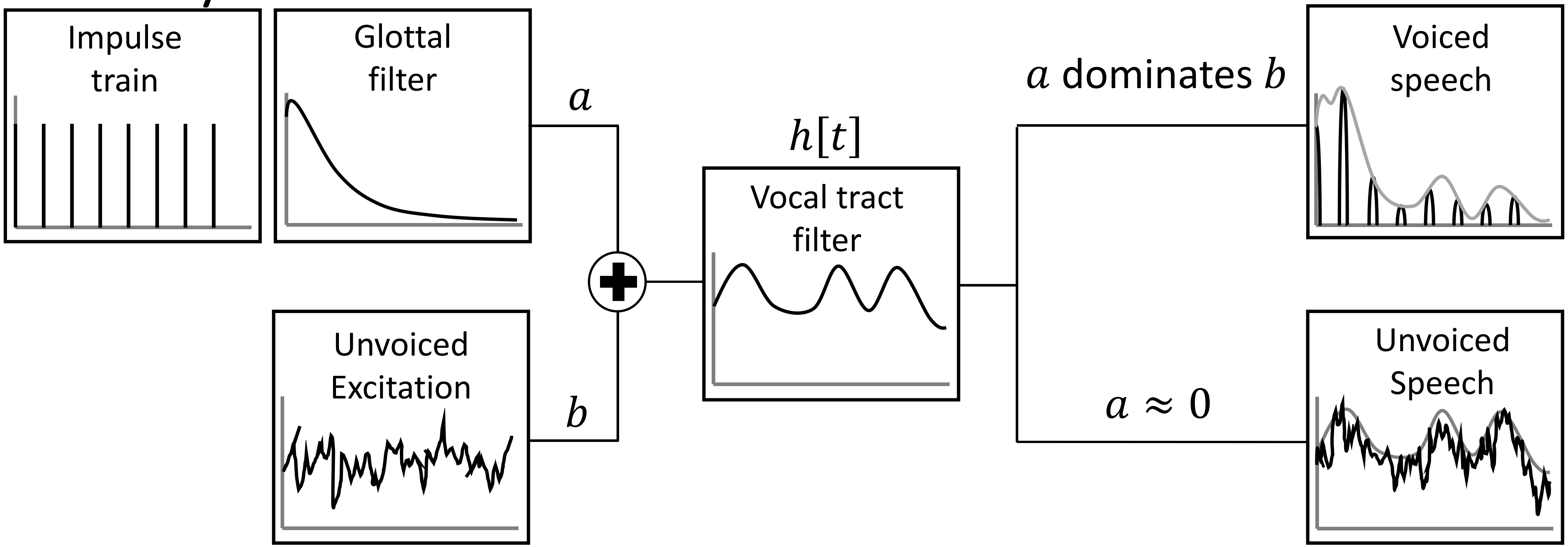
- Previous efforts
  - Additive deterministic-stochastic model, (Serra & Smith 1990)
  - STRAIGHT model, (Kawahara, et al. 2008)
  - Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis, (Degottex, et. al 2013)
  - Non-negative source-filter dynamical 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. WASPAA 2015)

PAT2 Summary



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

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

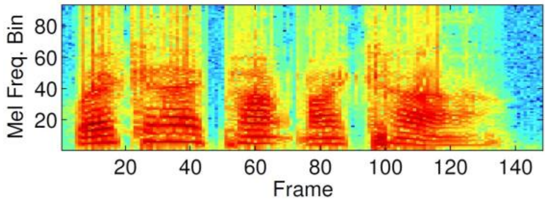
Frequency domain: $\vec{s} = a \cdot vec\left(\omega_0, \tau, \vec{g}, \hat{h}\right) + b \cdot vec\left[DFT\left[h[t]\right]\right] \boxdot vec\left[DFT[WGN]\right]$

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

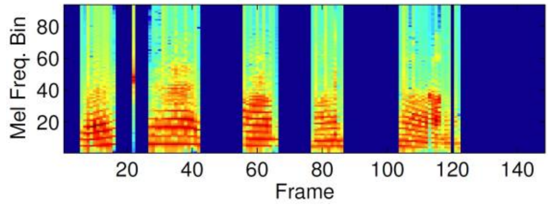
MAP inference $p(z|\vec{s}) \propto p(\vec{s}|z)p(z)$ by Monte Carlo sampling and L-BFGS search.

Experimental Results

Voiced Reconstruction

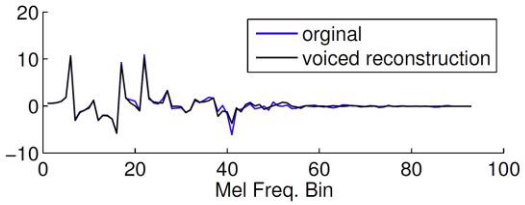


Original Spectrogram

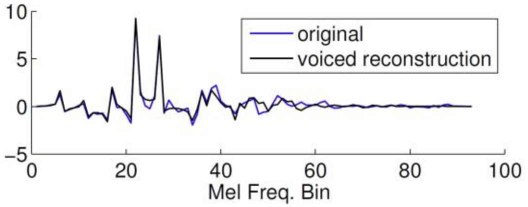


Voiced Reconstruct

Voiced Reconstruction – Single Frame

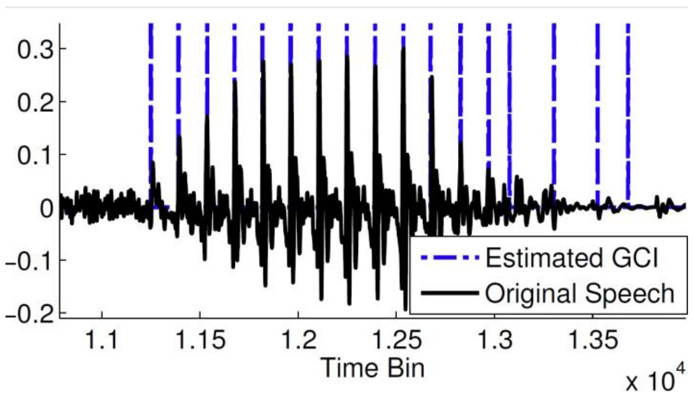


Real Spectrum

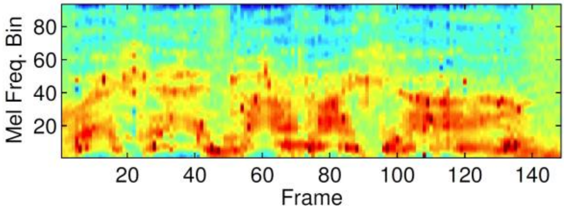


Imaginary Spectrum

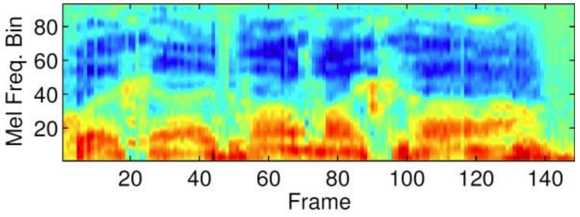
GCI Location Estimation



Vocal Tract Filter Estimation

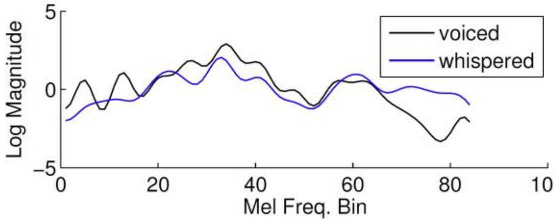


PAT2

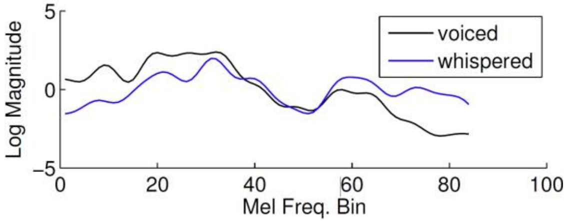


MFCC

Voiced vs Whispered



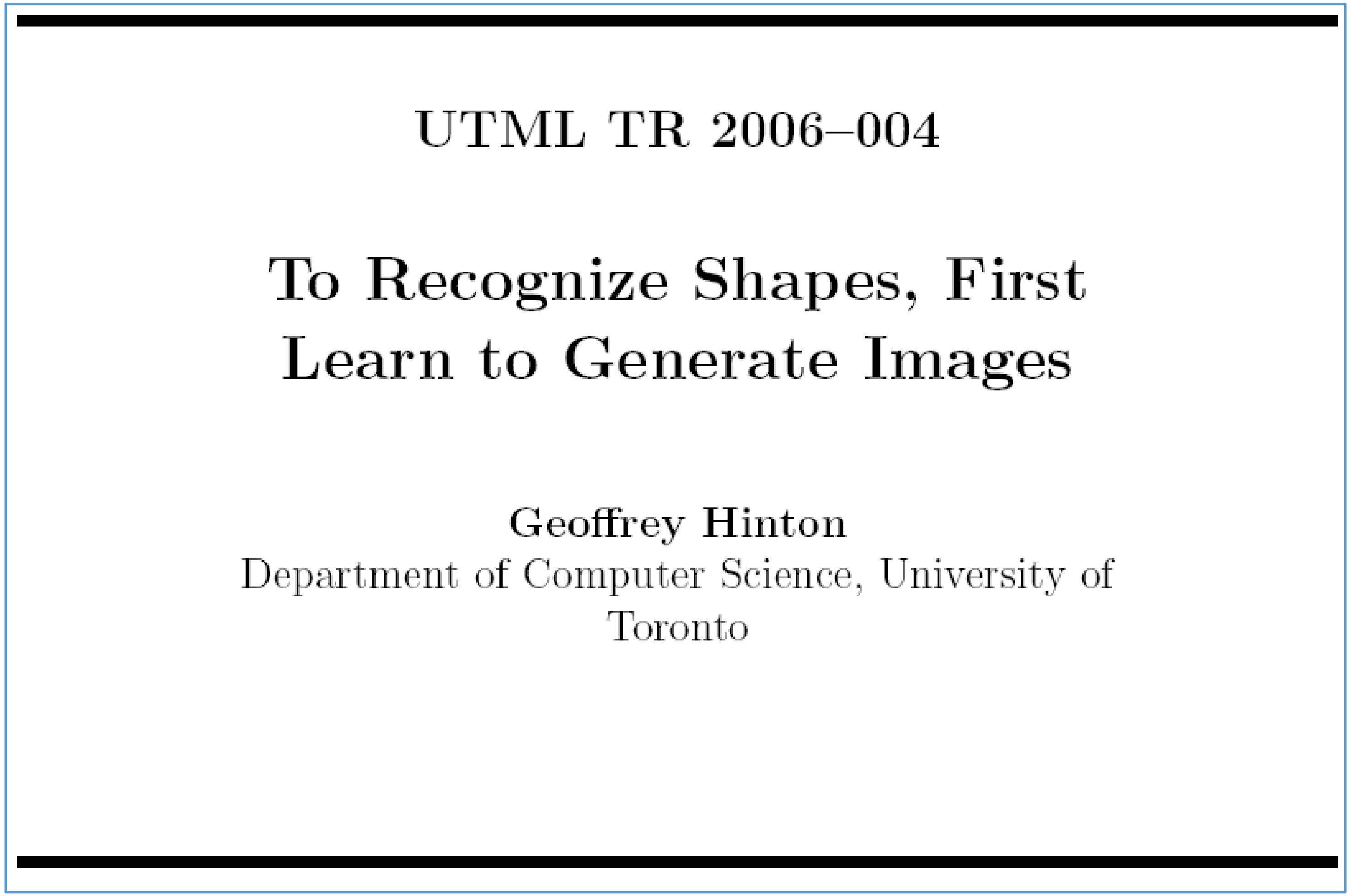
PAT2



MFCC

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

内容安排

1. State-of-the-art – Where we are
2. Basic thoughts – What we believe
3. Highlight – What we do
   - Probabilistic Acoustic Tube (PAT) Model
   - Random field approach to language modeling
4. Summary

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

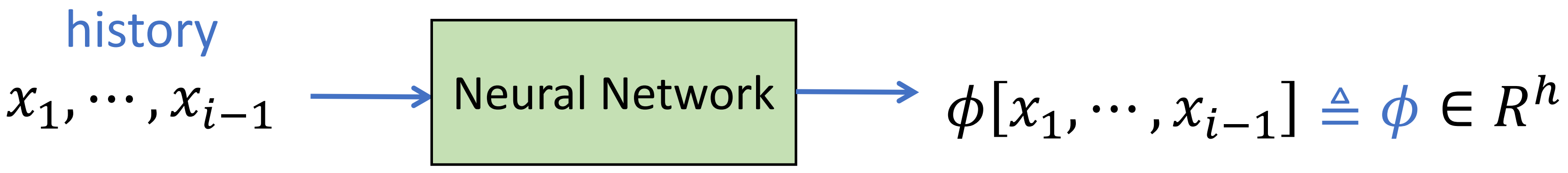
$$p(x_1,x_2,\cdots,x_l) = \prod_{i=1}^{l} p(\overset{\text{Current word}}{x_i}|\overset{\text{All previous words/history}}{x_1,\cdots,x_{i-1}})$$

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

- 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. $V = 10k \sim 100k, h = 250$

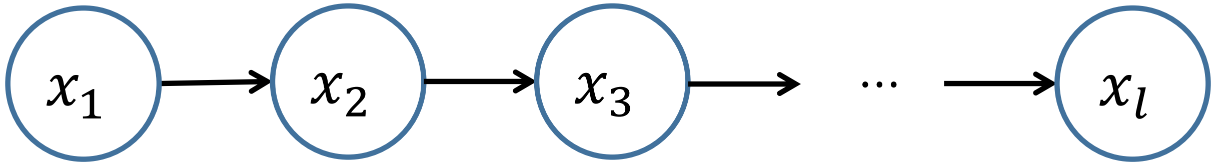
[1] Partly alleviated by using un-normalized models, e.g. through noise contrastive estimation training.

RFLMs – Motivation (1)

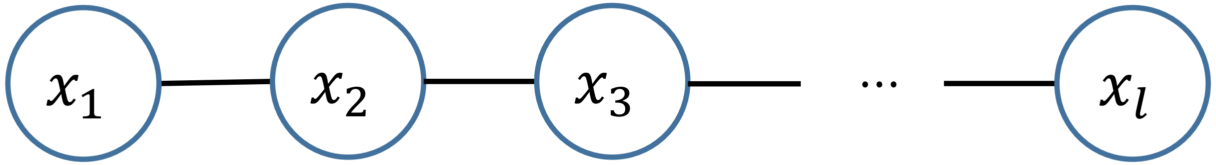
一句话

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

现状： 有向概率图模型



新思路： 无向概率图模型



☹ 模型学习有难度

☺ 上下文的彼此影响，长跨度

☺ **突破模型学习：定维情形（如固定尺寸）-> 跨维情形（如序列建模）**

Bin Wang, Zhijian Ou, Zhiqiang Tan. Trans-dimensional Random Fields for Language Modeling. **Association of Computational Linguistics (ACL Long Paper)** 2015.

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 ….

F. Jelinek, 1932 – 2010

- 1985: Every time I fire a linguist, the performance of the speech recognizer goes up.
- 1995: put language back into language modeling.

RFLMs – Formulation

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

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

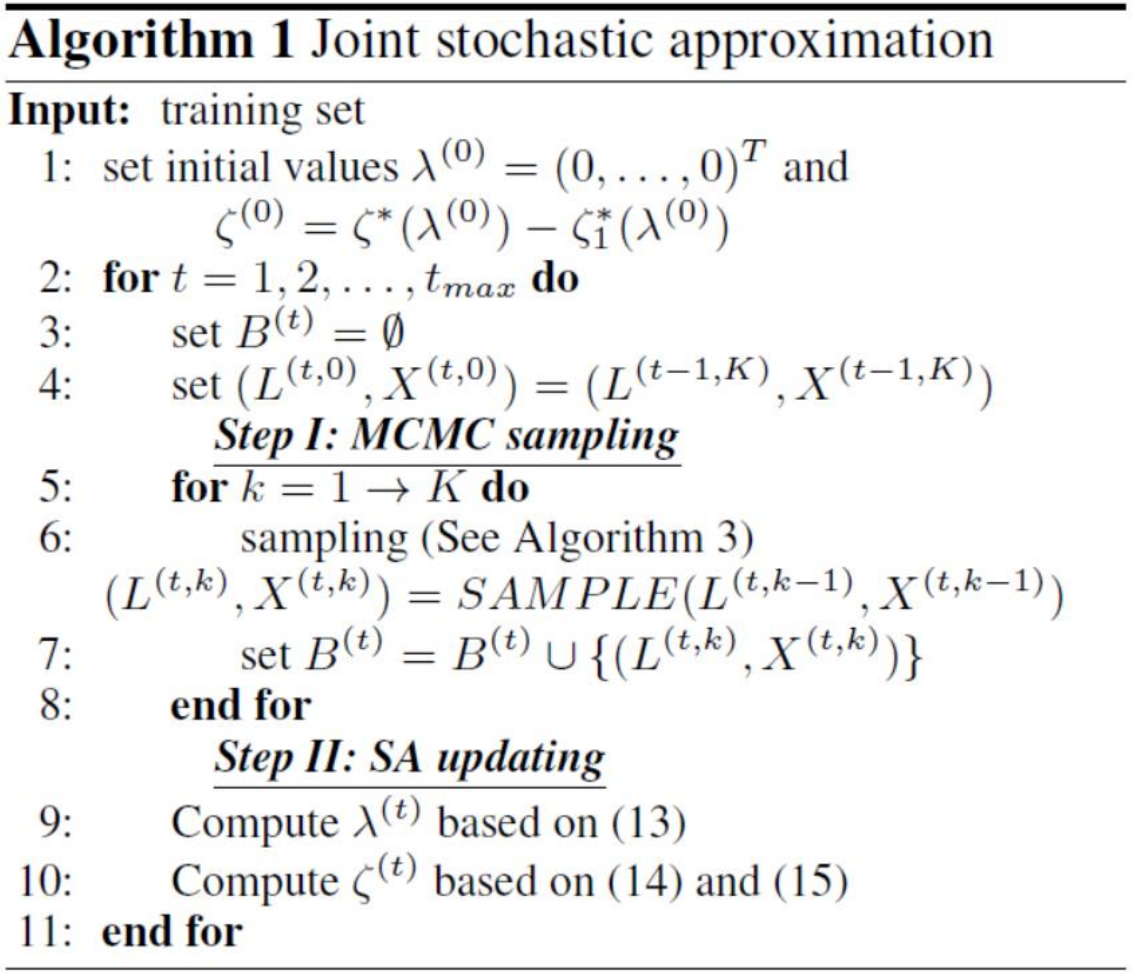
$$f_i(x) = \begin{cases} 1, & \text{‘meeting on DAY−OF−WEEK’ appears in } x \quad \Rightarrow \lambda_i \text{ is activated} \\ 0, & \text{Otherwise} \quad \Rightarrow \lambda_i \text{ is removed} \end{cases}$$

☺ More flexible features, beyond the n-gram features, can be well supported in RFLMs.

☺ Computational very efficient in computing sentence probability.

RFLMs – Breakthrough in training (1)

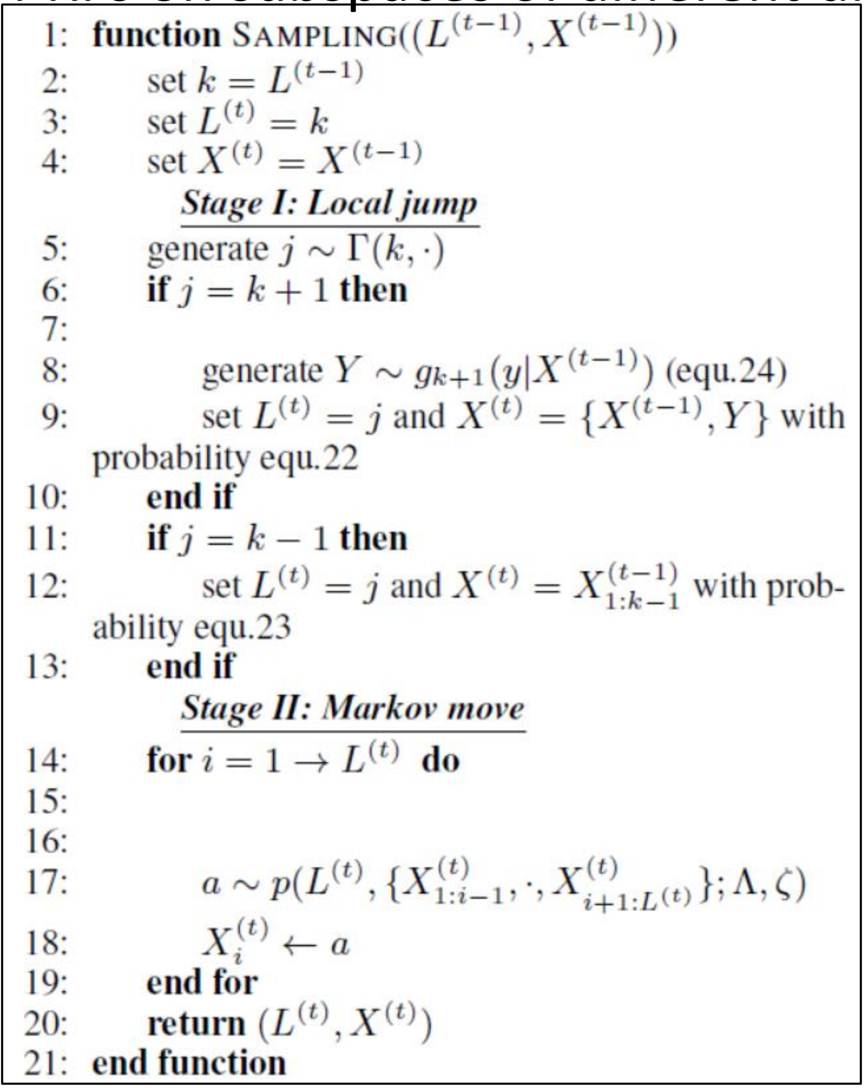
- Propose Joint Stochastic Approximation (SA) Training Algorithm
  - Simultaneously updates the model parameters and normalization constants



```
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 (Green, 1995).



```
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
```

Content

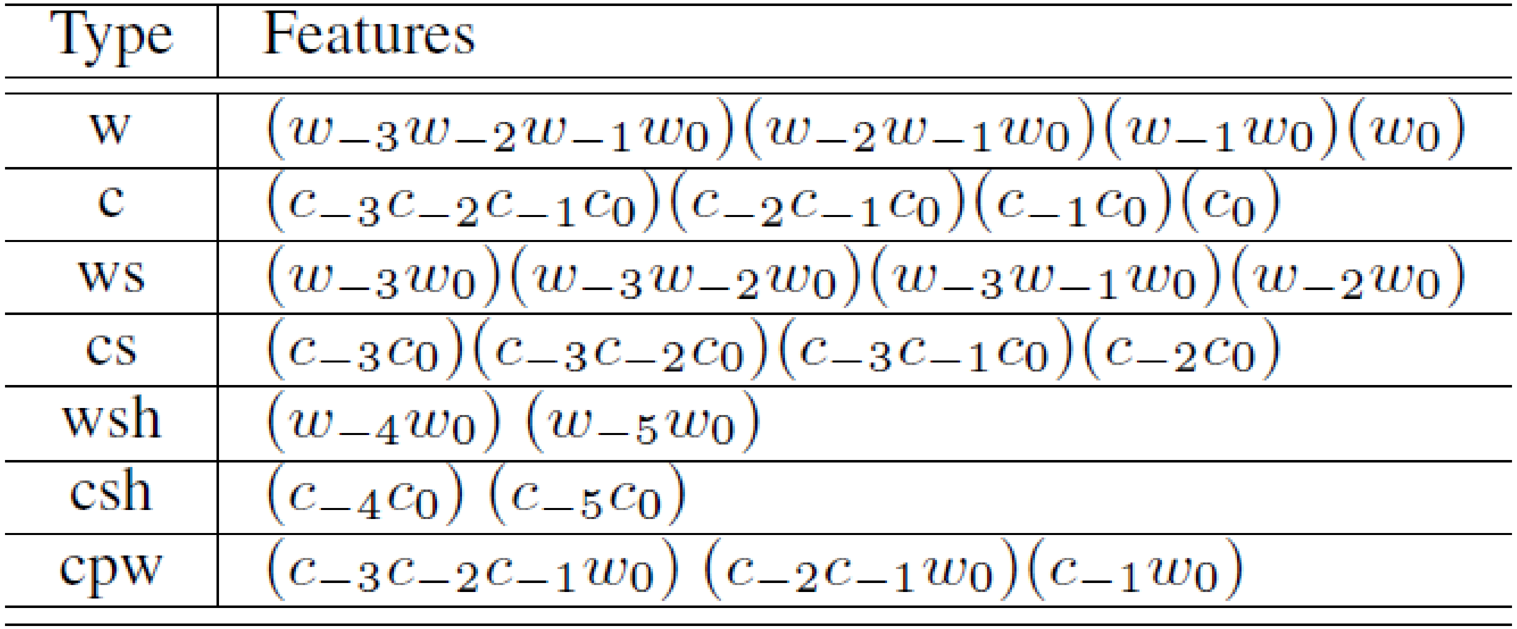
Random Field Language Models (RFLMs) – brand new

- 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

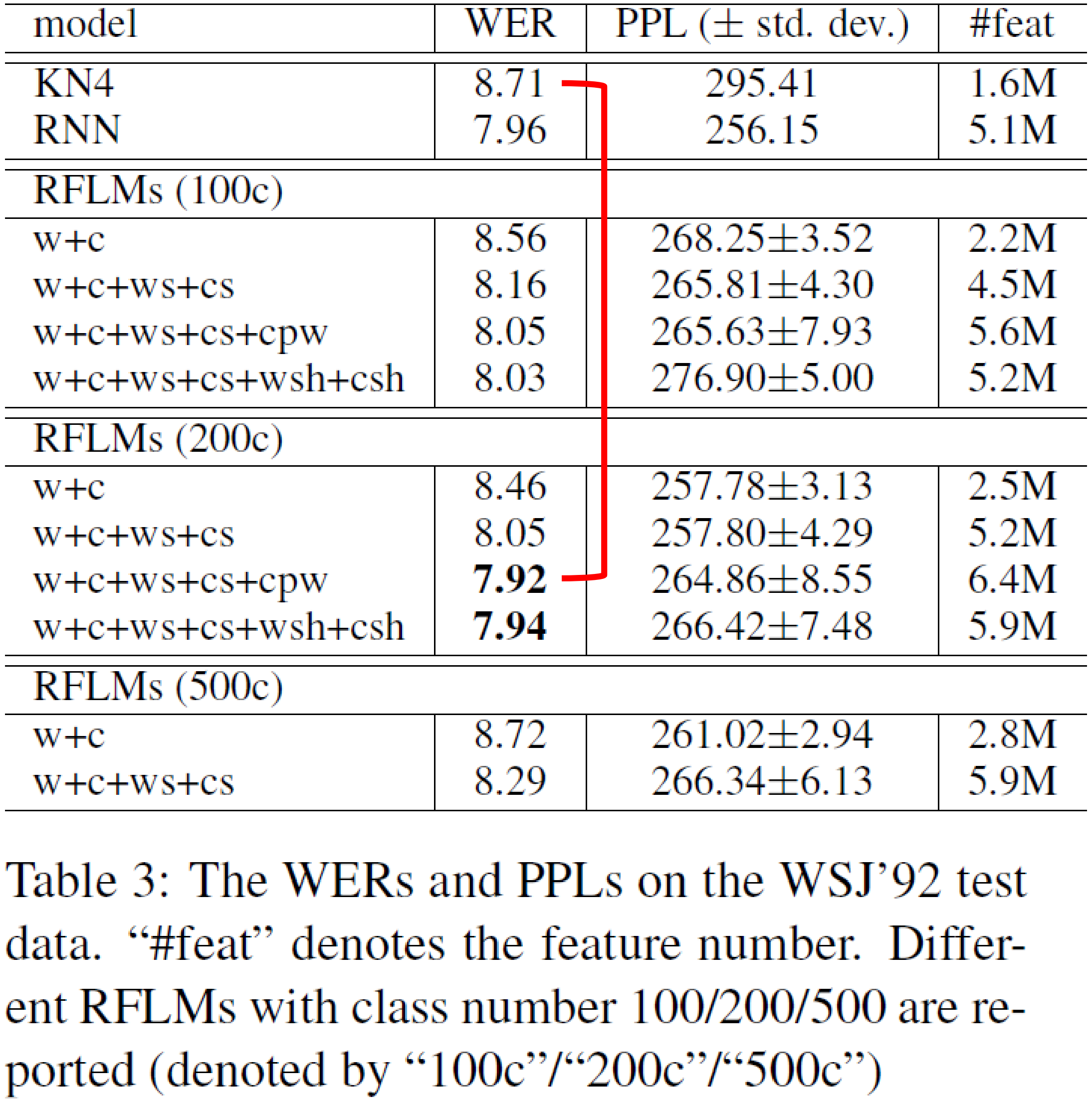
Feature Definition



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. **(200x faster !)**
  - 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	✗	✓	✓	✓	✓

内容安排

1. State-of-the-art – Where we are
2. Basic thoughts – What we believe
3. Highlight – What we do
    - Probabilistic Acoustic Tube (PAT) Model - 体现语音产生客观规律
    - Random field approach to language modeling - 实现数据驱动和语言学知识相结合
4. Summary

Thanks:

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

Thanks for your attention !