CCF ADL

计算机学会 学科前沿讲习班

数据高效的多语言与跨语言语音识别

欧智坚

清华大学·语音处理与机器智能(SPMI)实验室

http://oa.ee.tsinghua.edu.cn/ouzhijian/

2022/11/20

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. Conditional random fields and sequence discriminative training

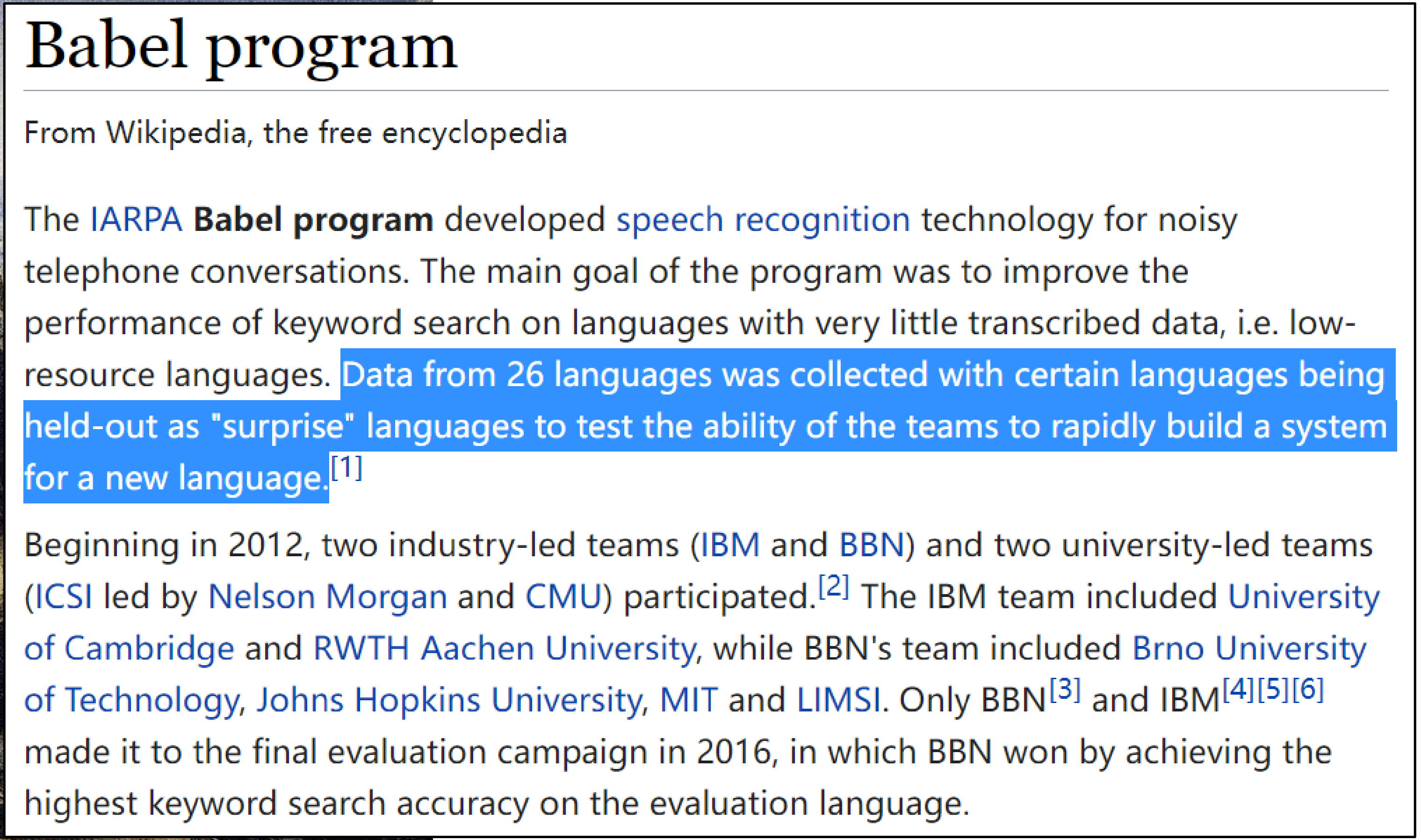
QA + break 15-min

II. Improving end-to-end speech recognition (15*3=45 min)

1. Data-efficiency
2. Multilingual and crosslingual ASR
3. Language Model Integration for Speech Recognition

III. Open questions and future directions + QA (15 min)

Tower of Babel 巴比塔

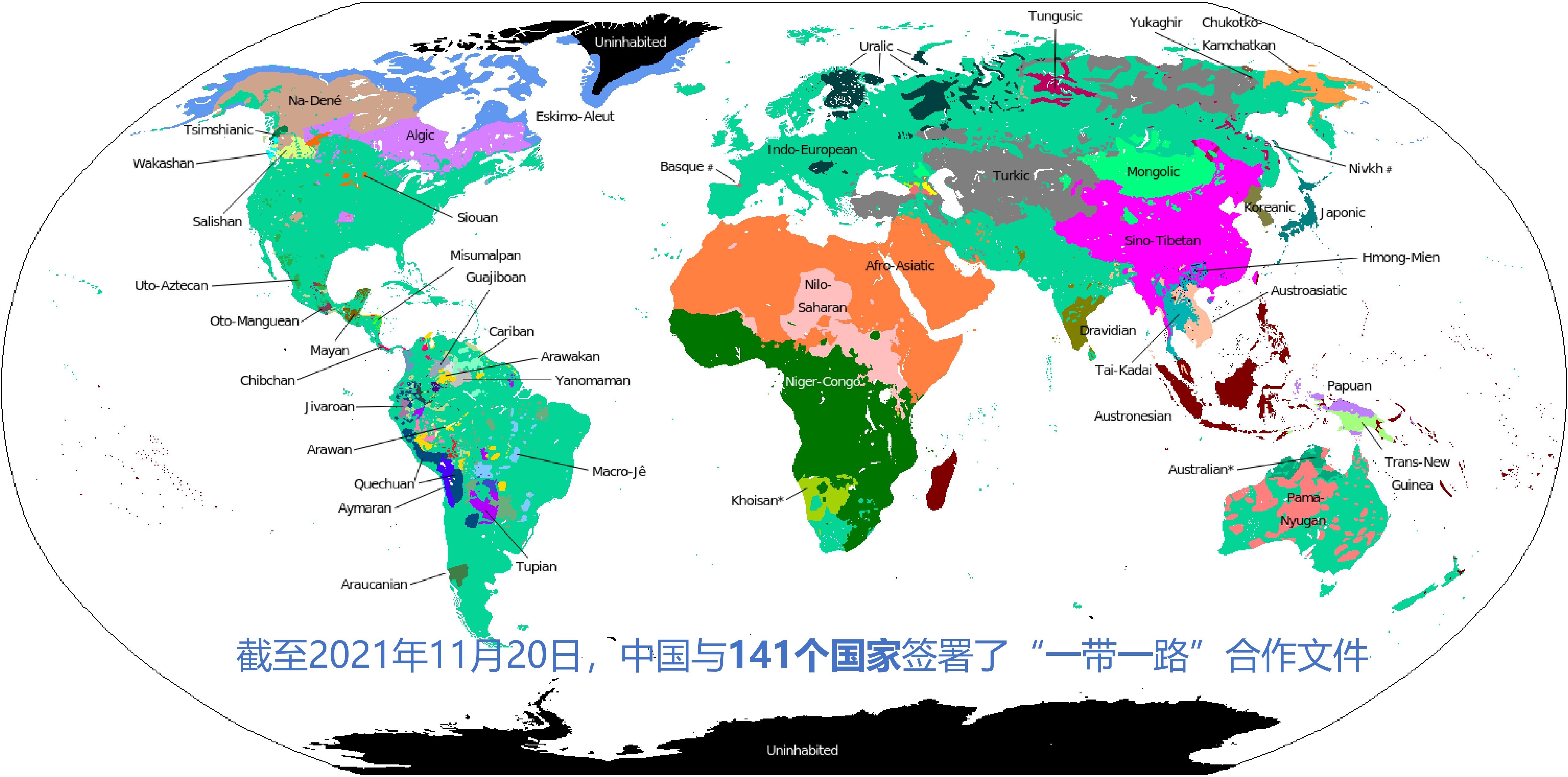


Babel program

From Wikipedia, the free encyclopedia

The IARPA **Babel program** developed speech recognition technology for noisy telephone conversations. The main goal of the program was to improve the performance of keyword search on languages with very little transcribed data, i.e. low-resource languages. Data from 26 languages was collected with certain languages being held-out as "surprise" languages to test the ability of the teams to rapidly build a system for a new language.[1]

Beginning in 2012, two industry-led teams (IBM and BBN) and two university-led teams (ICSI led by Nelson Morgan and CMU) participated.[2] The IBM team included University of Cambridge and RWTH Aachen University, while BBN's team included Brno University of Technology, Johns Hopkins University, MIT and LIMSI. Only BBN[3] and IBM[4][5][6] made it to the final evaluation campaign in 2016, in which BBN won by achieving the highest keyword search accuracy on the evaluation language.



截至2021年11月20日，中国与**141个国家**签署了“一带一路”合作文件

There are 7,139 living human languages distributed in 142 different language families.

ASR brief history

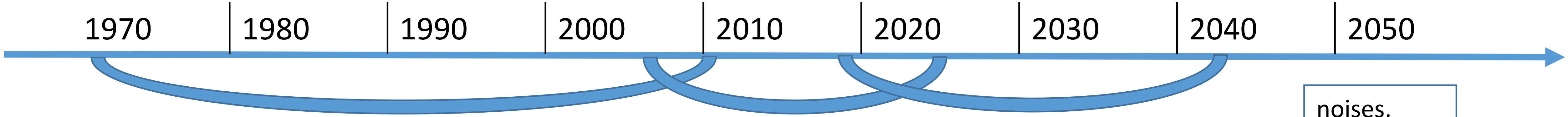
1970 – 2010: 1st Generation	
HMM	• F. Jelinek, “Continuous speech recognition by statistical methods”, Proc. of the IEEE, 1976. • J. Baker, “The DRAGON system--An overview”, T-ASSP, 1975.
GMM	• B.H. Juang, “Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains”, AT&T Technical Journal, 1985.
N-gram, Smoothing	• F. Jelinek & R.L. Mercer, “Interpolated estimation of Markov source parameters from sparse data”, Proc. Workshop on Pattern Recognition in Practice, 1980. • F. Jelinek, “The development of an Experimental Discrete Dictation Recognizer”, Proc. of the IEEE, 1985.
Tree based state tying	• S. Young, J.J. Odell, P.C. Woodland, “Tree-based state tying for high accuracy acoustic modeling”, HLT workshop, 1994.
MAP, MLLR	• C.H. Lee, C.H. Lin, B.H. Juang, “A study on speaker adaptation of the parameters of continuous density hidden Markov models”, T-IP, 1991. • C.J. Leggetter & P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, 1995.
fMLLR, Speaker adaptive training	• M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition”, Computer Speech and Language, 1998.
WFST	• M. Mohri. Finite-State Transducers in Language and Speech Processing. Computational Linguistics, 1997. • M. Mohri, F. Pereira, and M. Riley, “Speech Recognition with Weighted Finite-State Transducers”, 2008.
Discriminative Training, MMI, MPE	• D. Povey, “Discriminative training for large vocabulary speech recognition”, Ph.D. dissertation, 2003.

ASR brief history

2011 – now: 2nd Generation	
DNN-HMM	• A. Mohamed, G. Dahl, and G. Hinton, “Deep belief networks for phone recognition”, NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009. • G. Dahl, et al, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition”, T-ASLP, 2012. • F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks”, Interspeech, 2011. • D. Povey, et al, "Purely sequence-trained neural networks for ASR based on lattice-free MMI", Interspeech 2016.
NN-LM	• Bengio, et al, “A Neural Probabilistic Language Model”, NIPS, 2001. • Mikolov, et al, "Recurrent neural network based language model", Interspeech, 2010.
CTC	• A. Graves, et al, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks”, ICML, 2006. • H. Sak, et al, “Learning acoustic frame labeling for speech recognition with recurrent networks”, ICASSP, 2015. • Y. Miao, et al, “EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding”, ASRU, 2015.
Attention seq2seq	• D. Bahdanau, et al, “Neural machine translation by jointly learning to align and translate”, ICLR 2015. • J. K. Chorowski, et al, “Attention-based models for speech recognition,” NIPS, 2015. • W. Chan, et al @ google, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition”, ICASSP, 2016.
RNN Transducer	• A. Graves, “Sequence transduction with recurrent neural networks,” ICML 2012 Workshop on Representation Learning. • E. Battenberg, et al @ Baidu, “Exploring neural transducers for end-to-end speech recognition”, ASRU 2017. • K. Rao, et al @ Google, “Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer”, ASRU 2017
Transformer	• A. Vaswani, et al @ google, "Attention Is All You Need", NIPS, 2017.
CRF	• H. Xiang, Z. Ou. "CRF-based Single-stage Acoustic Modeling with CTC Topology", ICASSP, 2019.

欧智坚，“第三代语音识别技术初探”，全国声学大会, 2021/3/29, 上海

New-generation ASR



- HMM
- GMM
- N-gram, Smoothing
- Tree based state tying
- MAP, MLLR
- fMLLR, Speaker adaptive training
- WFST
- Discriminative Training, MMI, MPE

- DNN-HMM
- NN-LM
- CTC
- Attention seq2seq
- RNN Transducer
- Transformer
- CRF

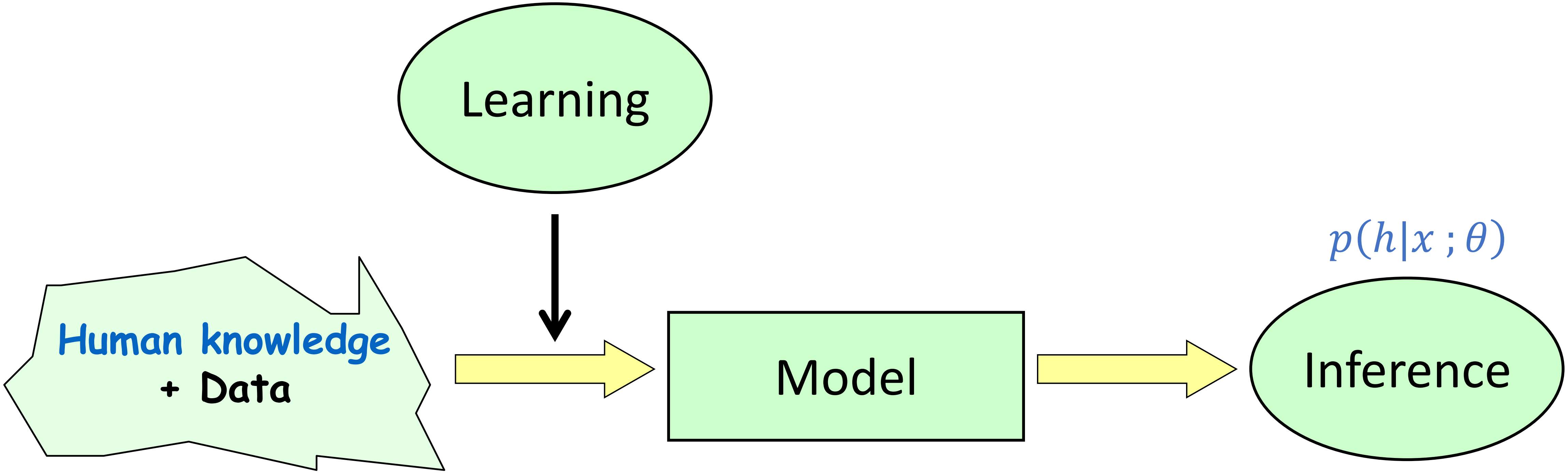
- Data-efficient
- AutoML
- Trustworthy AI

noises,
accents,
languages,
scenarios,
domains,
...

- Greater representational capability of DNNs 算法
- Larger amounts of labeled speech data for supervised training 数据
- Powerful hardware such GPUs 算力

欧智坚，“第三代语音识别技术初探”，全国声学大会, 2021/3/29, 上海

Probabilistic Framework



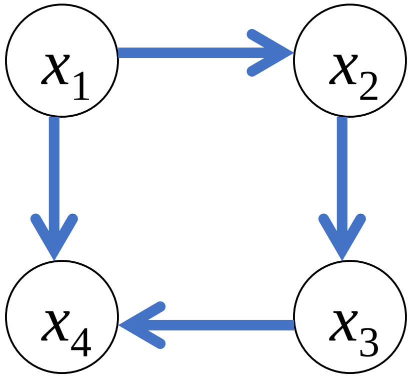
$p(x,h;\ \theta)$: Generative model, e.g., Hidden Markov Model (HMM)

$p(h|x;\ \theta)$: Discriminative model, e.g., Conditonal Random Field (CRF)

We need probabilistic models, besides neural nets.

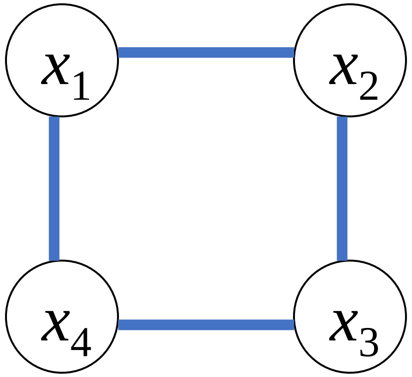
Probabilistic Graphical Modeling (PGM) Framework

- **Directed Graphical Models / Bayesian Networks (BNs)**
  - Self-normalized/Local-normalized
  - e.g. Hidden Markov Models (HMMs), Neural network (NN) based classifiers, Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs), auto-regressive models (e.g. RNNs/LSTMs)



$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_1, x_3)$$

- **Undirected Graphical Models / Random Fields (RFs) / Energy-based models**
  - Involves the normalizing constant Z / Globally-normalized
  - e.g. Ising model, Conditional Random Fields (CRFs)



$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\Phi(x_1, x_2)\Phi(x_2, x_3)\Phi(x_3, x_4)\Phi(x_1, x_4)$$

DGM example - Neural Net (NN) based classifier

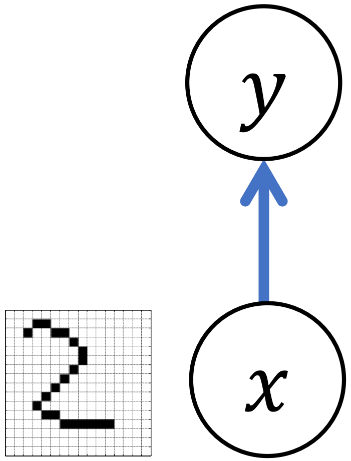
- Multi-class logistic regression

Consider observation/features $x \in \mathbb{R}^d$, class label $y \in \{1, \cdots, K\}$

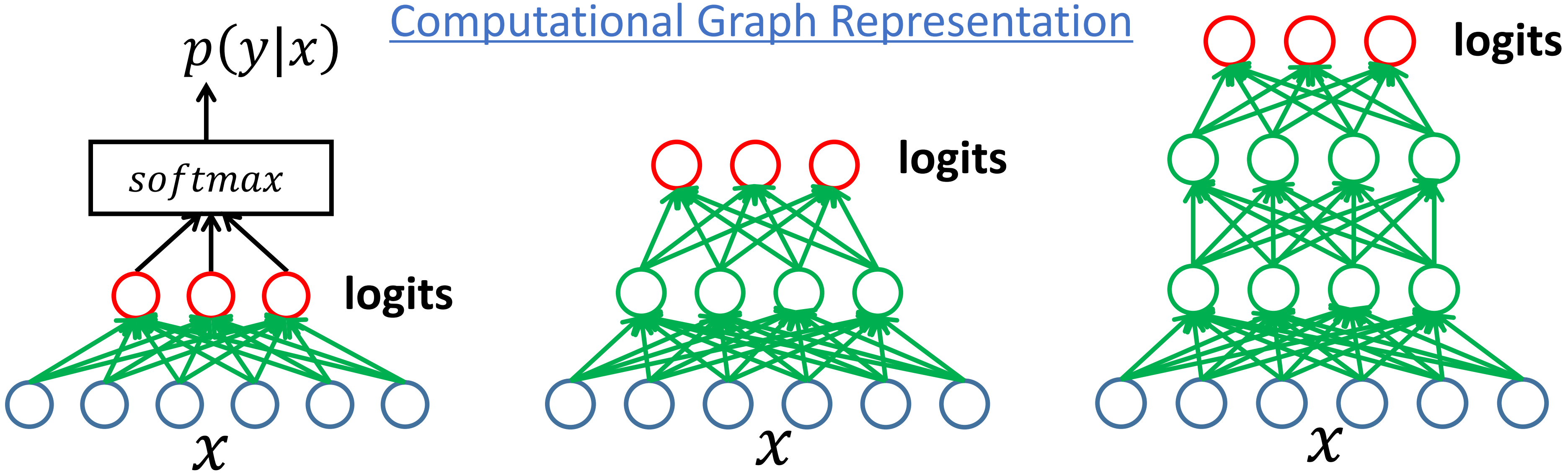
$$p(y = k|x) = \frac{exp(z_k)}{\sum_{j=1}^{K} exp(z_j)} \triangleq softmax(z_k)$$

where $z_k = w_k^T x + b_k, k = 1, \cdots, K$, often called **logits**

GM Representation



Computational Graph Representation



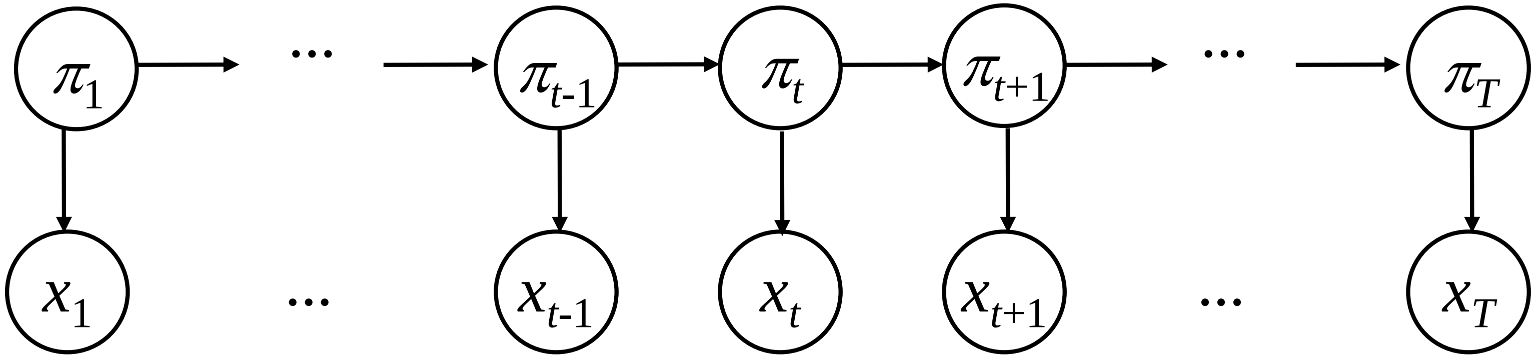
1-layer NN

2-layer NN

3-layer NN

(NNs as feature extractors)

HMM Viewed as Directed Graphical Model



The joint probability distribution of a hidden Markov model (HMM) :

$$p(\pi_{1:T}, x_{1:T}) = \underbrace{p(\pi_1)}_{\text{State Initial Distr.}} \prod_{t=1}^{T-1} \underbrace{p(\pi_{t+1}|\pi_t)}_{\text{State Transition Distr.}} \prod_{t=1}^{T} \underbrace{p(x_t|\pi_t)}_{\text{State Observation Distr.}}$$

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. → Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. Conditional random fields and sequence discriminative training

QA + break 15-min

II. Improving end-to-end speech recognition (15*3=45 min)

1. Data-efficiency
2. Multilingual and crosslingual ASR
3. Language Model Integration for Transducer based Speech Recognition

III. Open questions and future directions + QA (15 min)

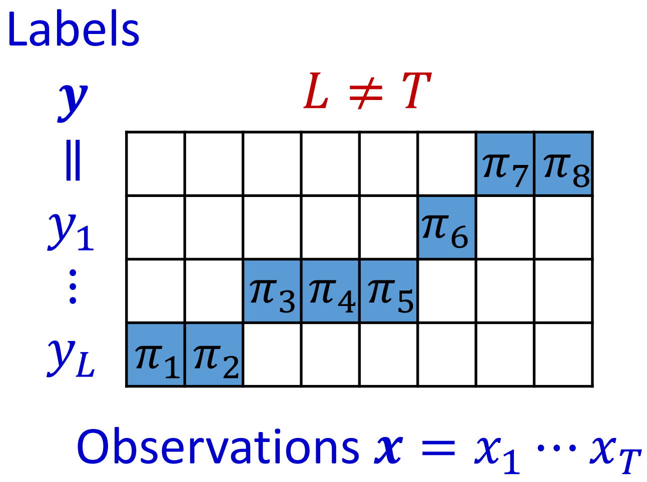
ASR: Basics

ASR (Automatic Speech Recognition) is a seq. discriminative problem

- For acoustic observations $\boldsymbol{x} \triangleq x_1, \cdots, x_T$, find the most likely labels $\boldsymbol{y} \triangleq y_1, \cdots, y_L$

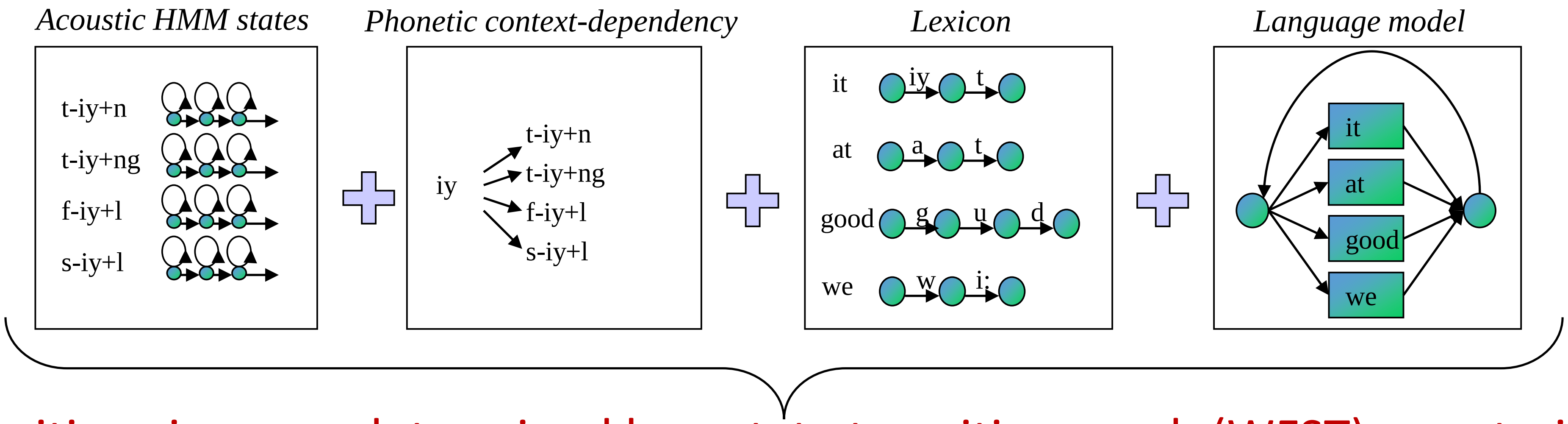
1. How to obtain $p(\boldsymbol{y} \mid \boldsymbol{x})$
2. How to handle alignment, since $L \neq T$

Separate
neural network architectures
and probabilistic model definitions !

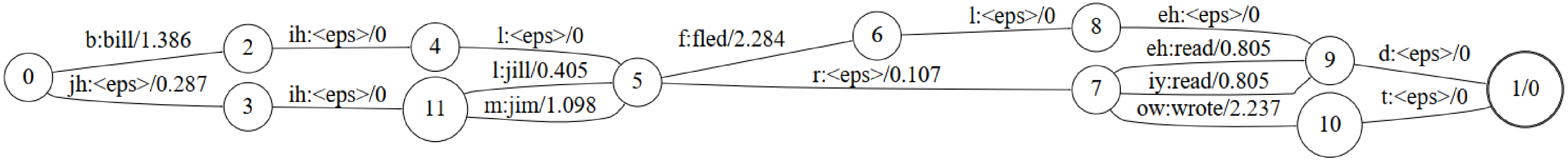

Labels
$\boldsymbol{y}$
$\parallel$
y_1
$\vdots$
y_L
$L \neq T$
π_1 π_2 π_3 π_4 π_5 π_6 π_7 π_8
Observations $\boldsymbol{x} = x_1 \cdots x_T$

Example of alignment

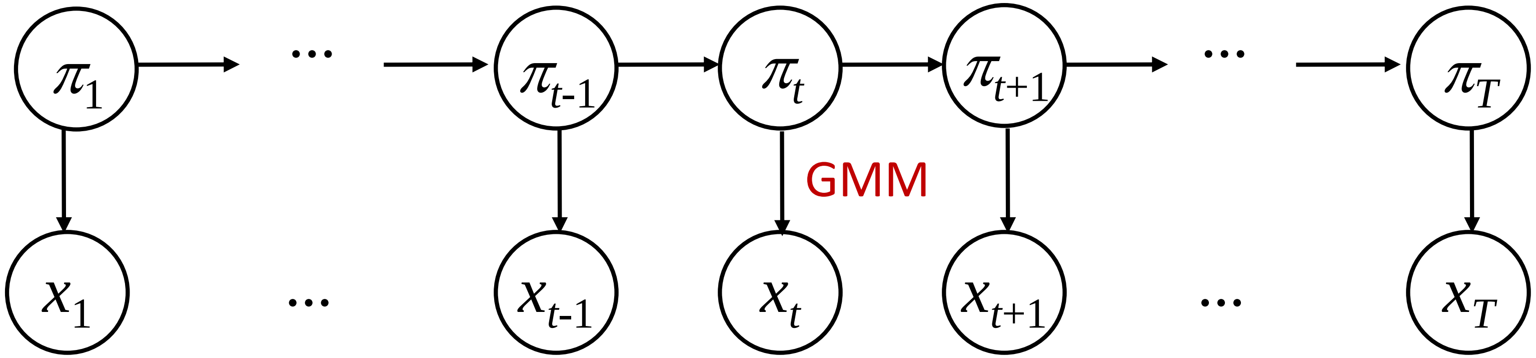
GMM-HMM: state transitions



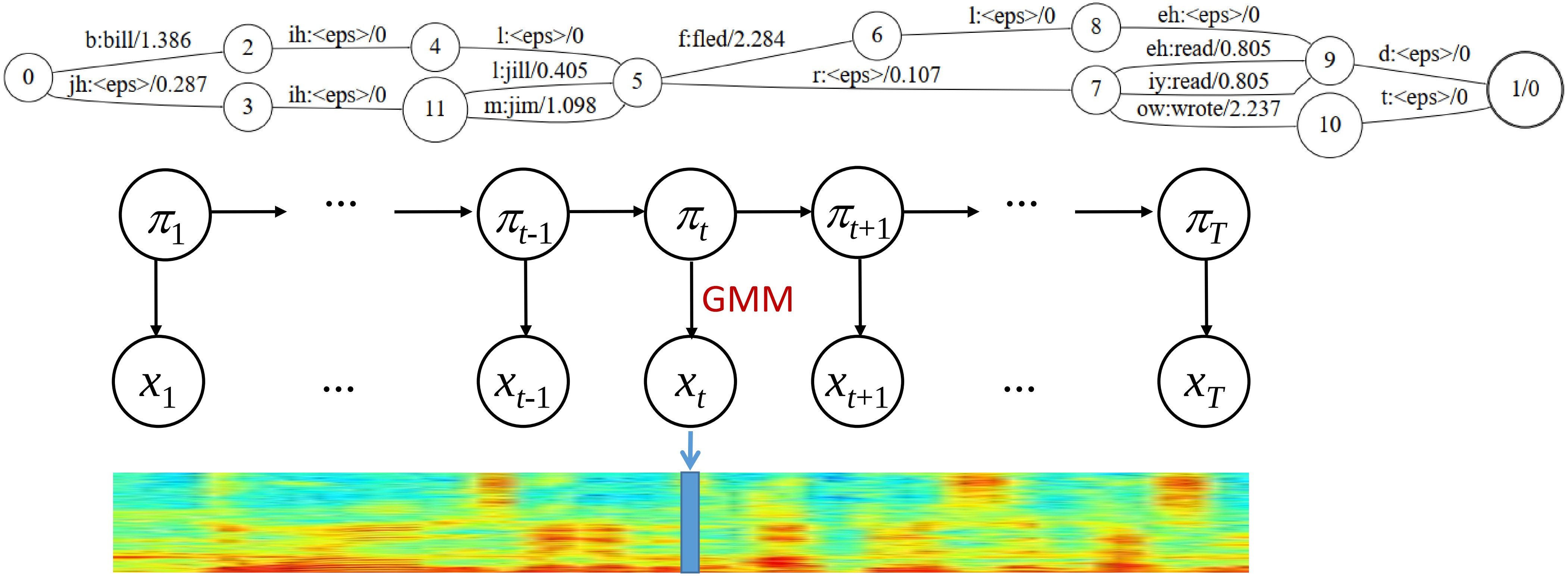
State transitions in $\boldsymbol{\pi}$ are determined by a state transition graph (WFST), constrained by ↑



A path $\boldsymbol{\pi} \triangleq \pi_1, \cdots, \pi_T$ uniquely determines a label sequence $\boldsymbol{y}$, but not vice versa.



GMM-HMM



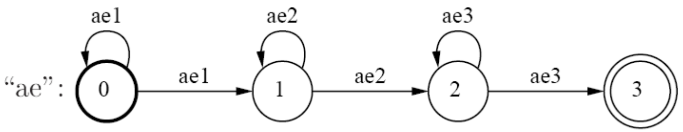
A path $\boldsymbol{\pi}$ uniquely determines $\boldsymbol{y}$ via mapping $\mathcal{B}_{HMM}$

- ▶ Training: Maximum likelihood $p(\boldsymbol{y}, \boldsymbol{x}) = \sum_{\boldsymbol{\pi}:\, \mathcal{B}_{HMM}(\boldsymbol{\pi})=\boldsymbol{y}} p(\boldsymbol{\pi}, \boldsymbol{x})$ via the forward-backward algo.
- ▶ Inference: Viterbi Decoding via $\max_{\boldsymbol{\pi}} p(\boldsymbol{\pi}, \boldsymbol{x})$

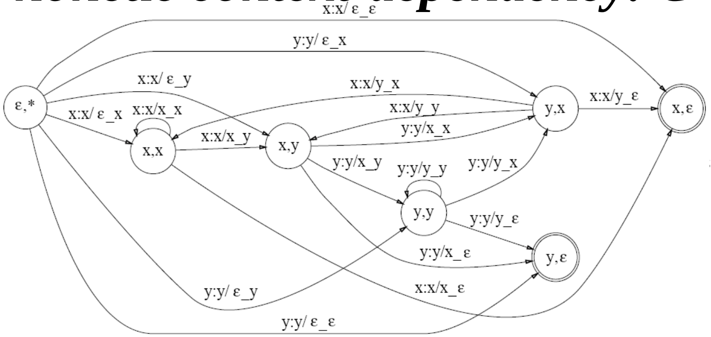
WFST

- WFSTs (weighted finite-state transducers) for Viterbi decoding
  - Pioneered by AT&T in late 1990's [Mohri et al., 2008]

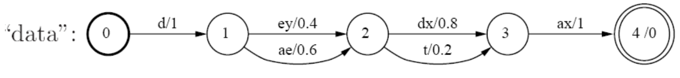
Acoustic HMMs: H



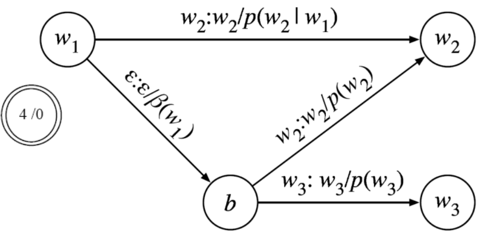
Phonetic context-dependency: C



Lexicon: L



Language model: G


w_1
$w_2{:}w_2/p(w_2 \mid w_1)$
w_2
$\varepsilon{:}\varepsilon/\beta(w_1)$
$w_2{:}w_2/p(w_2)$
b
$w_3{:}\ w_3/p(w_3)$
w_3

Composed and optimized into a single WFST

$$N = min\left(det\left(H \circ det\left(C \circ det\left(L \circ G\right)\right)\right)\right)$$

which represents $p(\pi_{t+1}|\pi_t)$ and is used in Viterbi decoder.

Well implemented in Kaldi toolkit https://github.com/kaldi-asr/kaldi

M. Mohri, et al., "Speech Recognition with Weighted Finite-State Transducers", 2008.

DNN-HMM

- ASR state-of-the-art: DNNs of various network architectures (MLP, LSTM, CNN, Transformer, etc.), initially DNN-HMM

State posterior prob.
estimated from the DNN, which needs
frame-level alignments

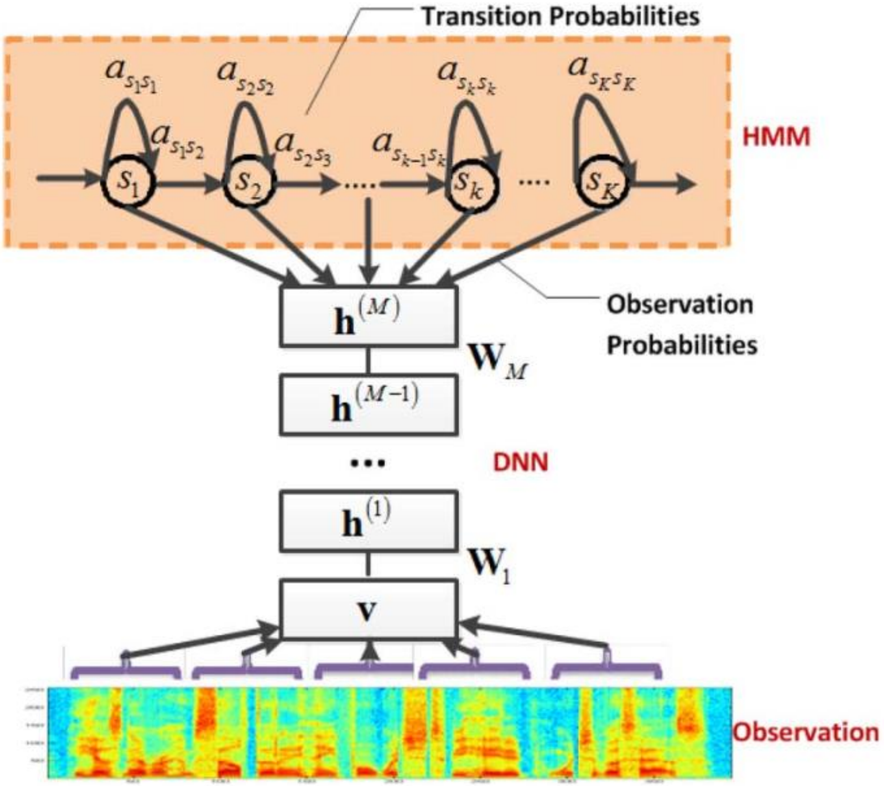
Can be ignored.

$$p(x_t|\pi_t) = \frac{p(\pi_t|x_t)p(x_t)}{p(\pi_t)}$$

State prior prob.
estimated from the training data

- Conventionally, multi-stage

  monophone GMM-HMM
  → alignment & triphone tree building
  → triphone GMM-HMM
  → alignment
  → triphone DNN-HMM



[Dahl, et al., TASLP 2012]

G. Dahl, et al., "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition", TASLP, 2012.

Advancing to end-to-end ASR: motivation

- End-to-end in the sense that:
  - Eliminate the construction of GMM-HMMs and phonetic decision-trees, and can be trained from scratch (flat-start or single-stage)
- In a more strict/ambitious sense:
  - Remove the need for a pronunciation lexicon and, even further, train the acoustic and language models jointly rather than separately
  - Trained to optimize criteria that are related to the final evaluation metric that we are interested in (typically, word error rate)
- Motivation
  - Simplify system pipeline, reduce expert knowledge and labor (such as compiling the ProLex, building phonetic decision trees)

Advancing to end-to-end ASR: techniques

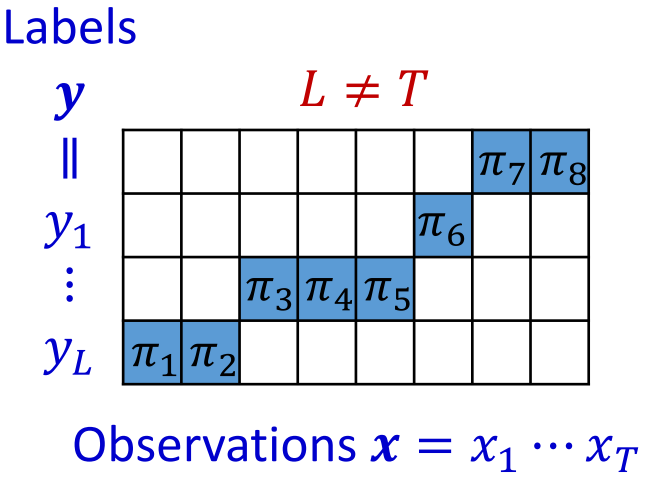
ASR is a *sequence discriminative* problem

- For acoustic observations $\boldsymbol{x} \triangleq x_1, \cdots, x_T$, find the most likely labels $\boldsymbol{y} \triangleq y_1, \cdots, y_L$

1. How to obtain $p(\boldsymbol{y} \mid \boldsymbol{x})$
2. How to handle alignment, since $L \neq T$

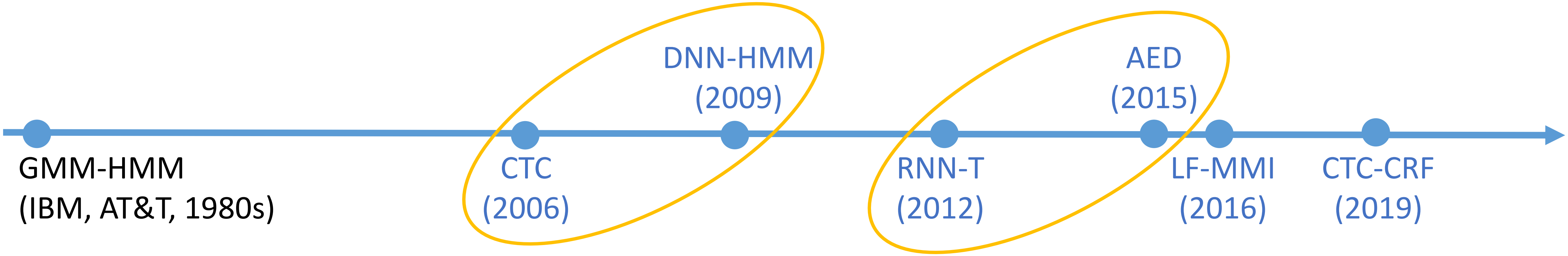
- Need a differentiable sequence-level loss of mapping acoustic sequence $\boldsymbol{y}$ to label sequence $\boldsymbol{x}$

- **Explicitly**: introduce hidden state sequence $\boldsymbol{\pi}$, as in Connectionist Temporal Classification (CTC), RNN Transducer (RNNT), CRF
- **Implicitly**: as in Attention based Encoder-Decoder (AED)



Example of explicit alignment

History



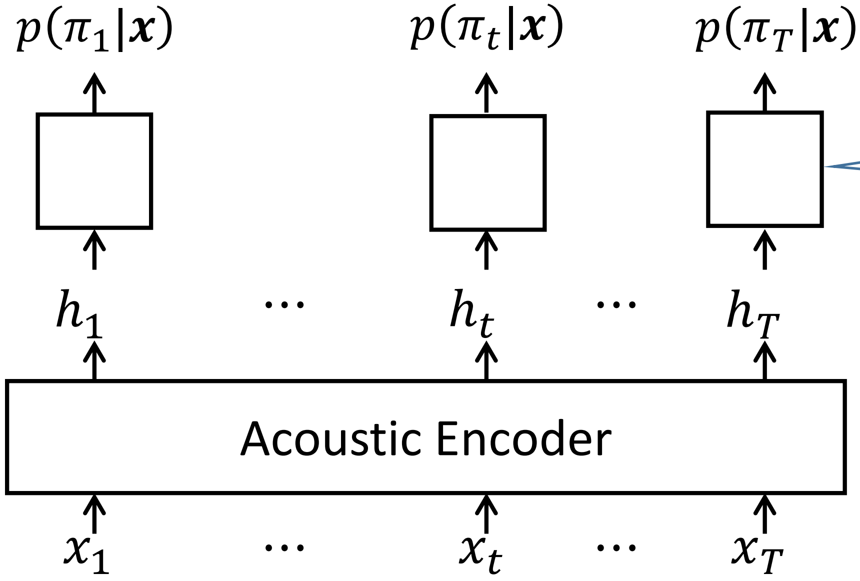
- [CTC] Graves, et al., “Connectionist Temporal Classification: Labelling unsegmented sequence data with RNNs”, ICML 2006.
- [DNN-HMM] A. Mohamed, et al., “Deep belief networks for phone recognition”, NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.
- [RNNT] A. Graves, “Sequence transduction with recurrent neural networks”, ICML 2012 Workshop on Representation Learning.
- [AED] D. Bahdanau, et al., “Neural machine translation by jointly learning to align and translate”, ICLR 2015.
- [LF-MMI] D. Povey, et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI", INTERSPEECH 2016.
- [CTC-CRF] Xiang&Ou. "CRF-based Single-stage Acoustic Modeling with CTC Topology", ICASSP, 2019.

CTC: introducing blank symbol

- Motivation: training $p(\boldsymbol{y} \mid \boldsymbol{x})$ without the need for frame-level alignments between the acoustics $\boldsymbol{x}$ and the transcripts $\boldsymbol{y}$
  - Introduce a state sequence $\boldsymbol{\pi} \triangleq \pi_1, \cdots, \pi_T$, where $\pi_t \in$ the-alphabet-of-labels $\cup$ <b>

Path posterior State posterior

$$p(\boldsymbol{\pi} \mid \boldsymbol{x}) = \prod_{t=1}^{T} p(\pi_t \mid \boldsymbol{x})$$



Linear&Softmax Layer

$z_t = W h_t \in \mathbb{R}^{K+1}$

$p(\pi_t = k|\boldsymbol{x}) = \frac{exp(z_t^k)}{\sum_i exp(z_t^i)} \triangleq p_t^k$: the prob. of observing label k at time t

The un-normalized outputs z_t are often called **logits**.

Graves, et al., “Connectionist Temporal Classification: Labelling unsegmented sequence data with RNNs”, ICML 2006.

CTC topology

- **State topology** refers to the state transition structure in $\boldsymbol{\pi}$, which basically determines the mapping $\mathcal{B}_{CTC}$ from $\boldsymbol{\pi}$ to $\boldsymbol{y}$

CTC topology : a mapping $\mathcal{B}_{CTC}$ maps $\boldsymbol{\pi}$ to $\boldsymbol{y}$ by

1. reducing repetitive symbols to a single symbol;
2. removing all blank symbols.

$$\mathcal{B}(-CC--AA-T-) = CAT$$

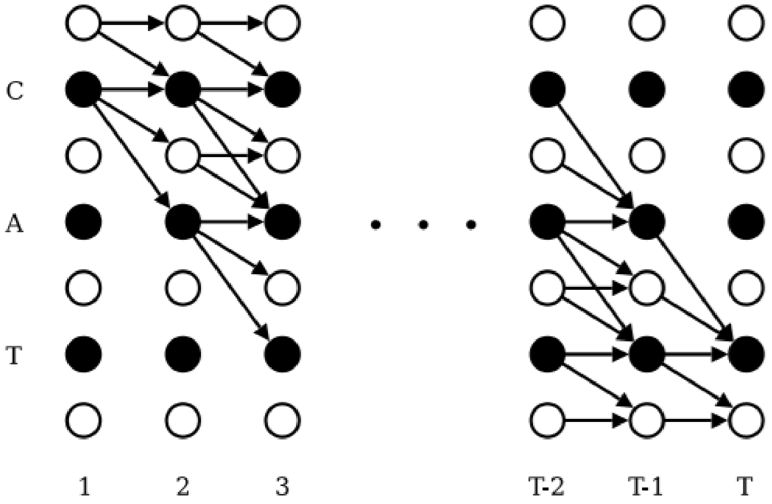
Path posterior

$$p(\boldsymbol{\pi}|\boldsymbol{x}) = \prod_{t=1}^{T} p(\pi_t|\boldsymbol{x})$$

Label-seq posterior

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}:\,\mathcal{B}_{CTC}(\boldsymbol{\pi})=\boldsymbol{y}} p(\boldsymbol{\pi}|\boldsymbol{x})$$

Summing over all possible paths, which map to $\boldsymbol{y}$



```
B B c B B a a B B t
B c c B a B B B B t
          . . .
B c B B a B B t t B
```

CTC: the gradient & the forward-backward algorithm

For logit $z_t^k, 1 \leq t \leq T$

$$\frac{\partial logp(\boldsymbol{y}|\boldsymbol{x})}{\partial z_t^k} = E_{p(\boldsymbol{\pi}|\boldsymbol{x},\boldsymbol{y})}\left[\frac{\partial logp(\boldsymbol{\pi}|\boldsymbol{x})}{\partial z_t^k}\right]$$ ∵ Fisher Equality [Ou, arxiv 2018]

$$= E_{p(\boldsymbol{\pi}|\boldsymbol{x},\boldsymbol{y})}\left[\frac{\partial logp_t^{\pi_t}}{\partial z_t^k}\right] \quad \because p(\boldsymbol{\pi}|\,\boldsymbol{x}) = \prod_{t=1}^{T} p_t^{\pi_t}$$

$$= E_{p(\boldsymbol{\pi}|\boldsymbol{x},\boldsymbol{y})}[\delta(\pi_t = k) - p_t^k]$$

$$= p(\pi_t = k|\boldsymbol{x},\boldsymbol{y}) - p_t^k$$

i.e., the **error signal** received by the acoustic encoder NN during training

i.e., γ_t^k, the posterior **state occupation probability**, calculated using the alpha-beta variables from the forward-backward algorithm [Rabiner, 1989]

Providing easy derivation and giving insight, not appeared in [Graves, et al., 2006] and elsewhere

Z. Ou. "A Review of Learning with Deep Generative Models from Perspective of Graphical Modeling", arXiv, 2018.

CTC: LM integration with WFSTs

- Best-path-decoding or Prefix-search-decoding

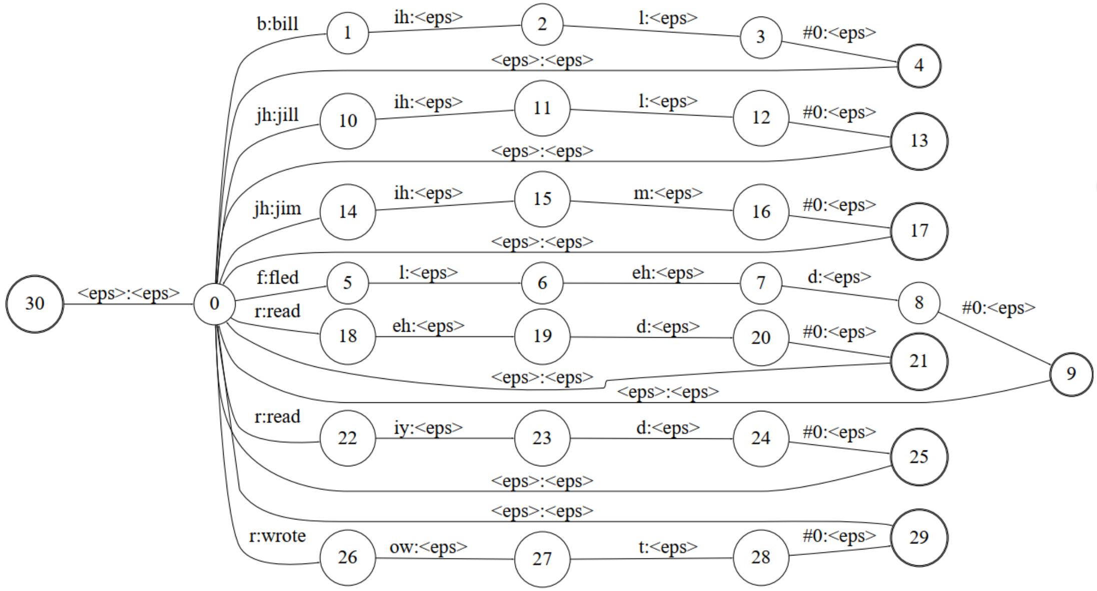
$$\max_{\boldsymbol{\pi}} p(\boldsymbol{\pi}|\boldsymbol{x}) \qquad\qquad \max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x})$$

- Incorporate lexicon and LM to improve best-path-decoding

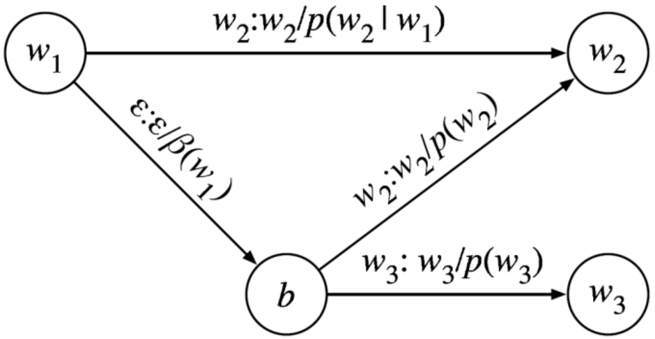
$$\max_{\boldsymbol{\pi}} p(\boldsymbol{\pi}|\boldsymbol{x}) LM(\mathcal{B}_{CTC}(\boldsymbol{\pi}))$$

WFST representing CTC topology: T

Lexicon: L



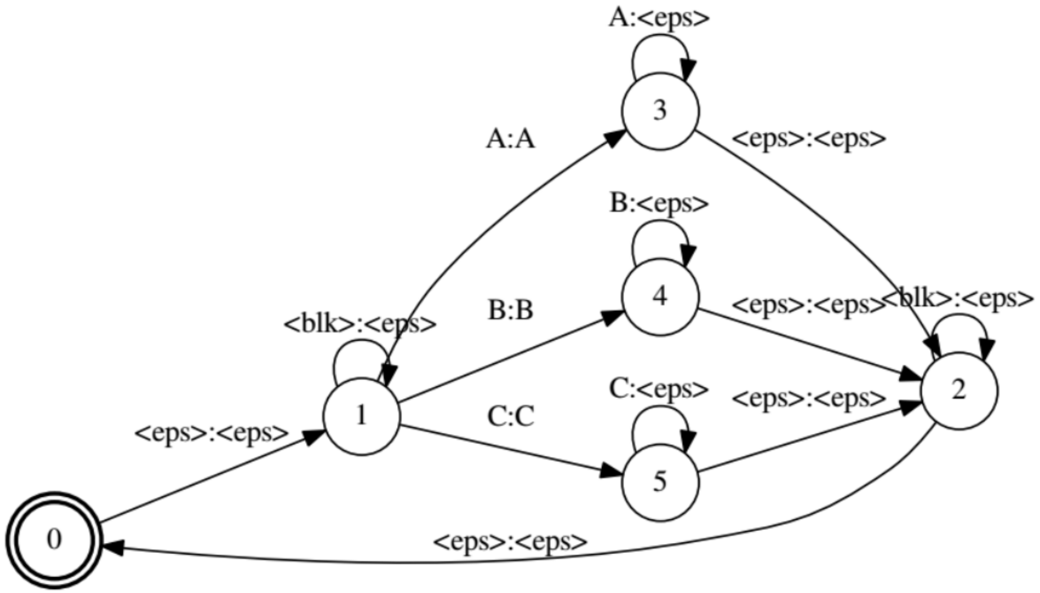
Language model: G



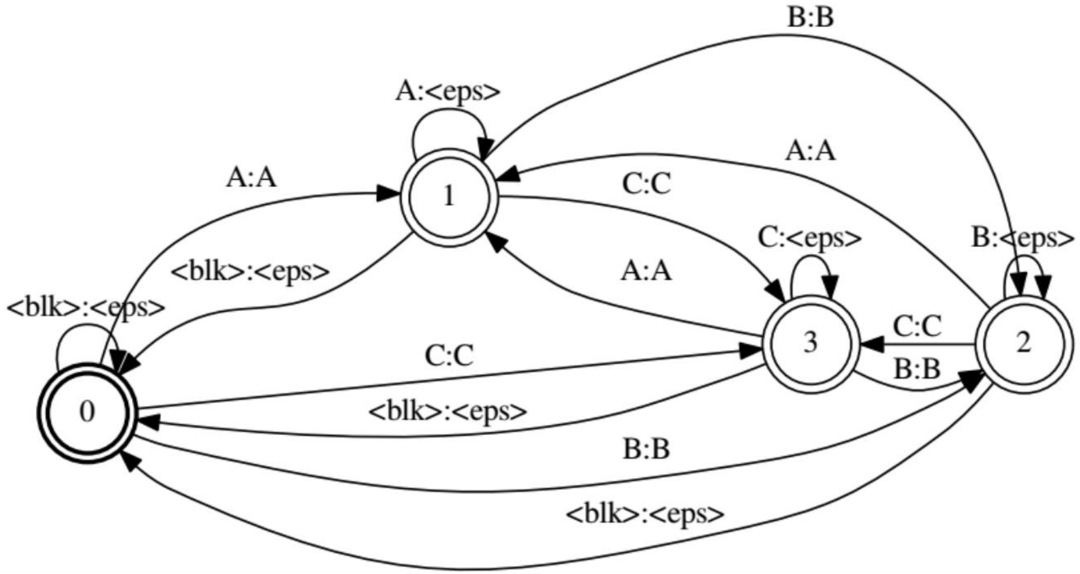
w_1
$w_2{:}w_2/p(w_2 \mid w_1)$
w_2
$\varepsilon{:}\varepsilon/\beta(w_1)$
$w_2{:}w_2/p(w_2)$
b
$w_3{:}\ w_3/p(w_3)$
w_3

Composed and optimized into a single WFST

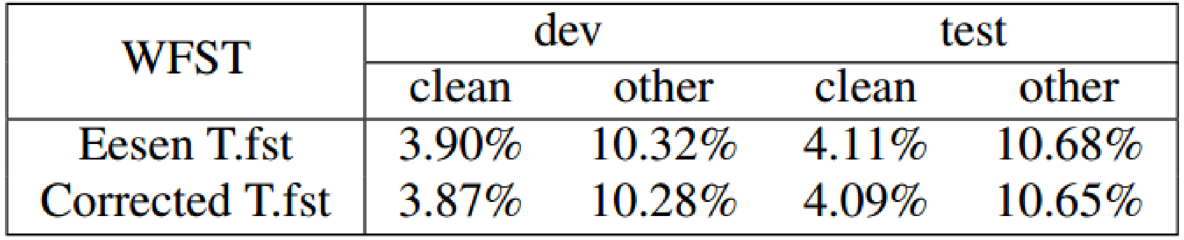
WFST representation of CTC topology [Xiang&Ou, 2019]



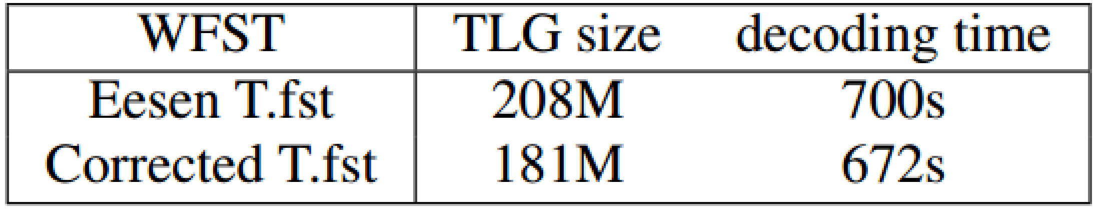
EESEN T.fst



Corrected T.fst



WFST	dev		test	
	clean	other	clean	other
Eesen T.fst	3.90%	10.32%	4.11%	10.68%
Corrected T.fst	3.87%	10.28%	4.09%	10.65%



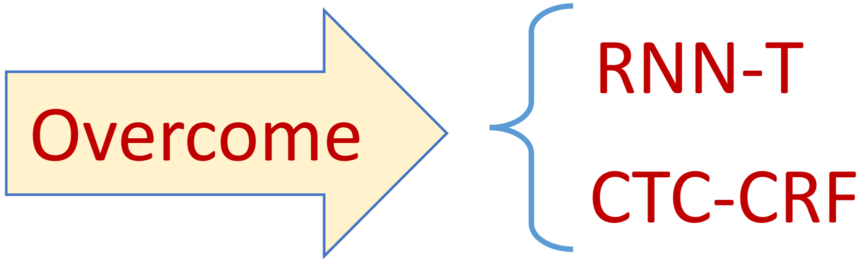
WFST	TLG size	decoding time
Eesen T.fst	208M	700s
Corrected T.fst	181M	672s

Using corrected T.fst performs slightly better; The decoding graph size smaller, and the decoding speed faster.

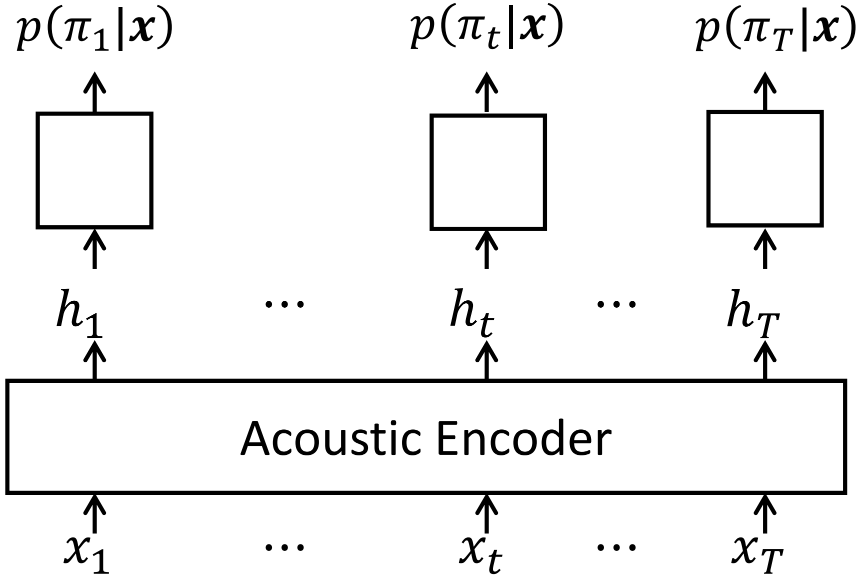
- Miao, et al., “EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding ”, ASRU 2015.
- Xiang&Ou. "CRF-based Single-stage Acoustic Modeling with CTC Topology", ICASSP, 2019.

CTC: shortcoming

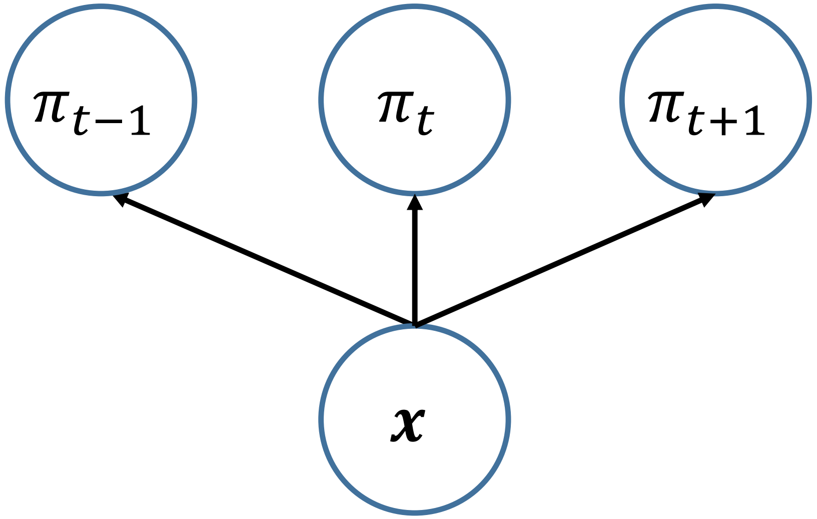
- Conditional independence assumption



$$p(\boldsymbol{\pi}|\ \boldsymbol{x}) = \prod_{t=1}^{T} p(\pi_t|\boldsymbol{x})$$


$p(\pi_1|\boldsymbol{x})$
$p(\pi_t|\boldsymbol{x})$
$p(\pi_T|\boldsymbol{x})$
h_1
…
h_t
…
h_T
Acoustic Encoder
x_1
…
x_t
…
x_T

Computational flow



Graphical Model Representation

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. → Conditional random fields and sequence discriminative training

QA + break 15-min

II. Improving end-to-end speech recognition (15*3=45 min)

1. Data-efficiency
2. Multilingual and crosslingual ASR
3. Language Model Integration for Transducer based Speech Recognition

III. Open questions and future directions + QA (15 min)

Sequence discriminative training

- Historically
  - GMM-HMMs are generative models
  - DNN-HMMs are interpreted as generative models (interpreting $p(x_t|\pi_t) = \frac{p(\pi_t|x_t)p(x_t)}{p(\pi_t)}$ as pseudo-likelihood), though strictly not
- A large body of works to improve GMM-HMMs and DNN-HMMs, by using sequence-discriminative criteria, like
  - Maximum Mutual Information (MMI), boosted MMI (BMMI), Minimum Phone Error (MPE), Minimum Bayes Risk (MBR) [Karel, et al., 2013]
  - Minimum Word Error Rate (MWER) [Stolcke, et al., 1997]

- V. Karel, et al., "Sequence-discriminative training of deep neural networks", INTERSPEECH 2013.
- A. Stolcke, et al., "Explicit word error minimization in N-best list rescoring", Eurospeech, 1997.

MMI and CML

- MMI training of a GMM-HMM, for acoustic input $\boldsymbol{x}$ and transcript $\boldsymbol{y}$, is equivalent to CML (conditional maximum likelihood) training of a CRF (using 0/1/2-order features in potential definition) [Heigold, et al., 2011].

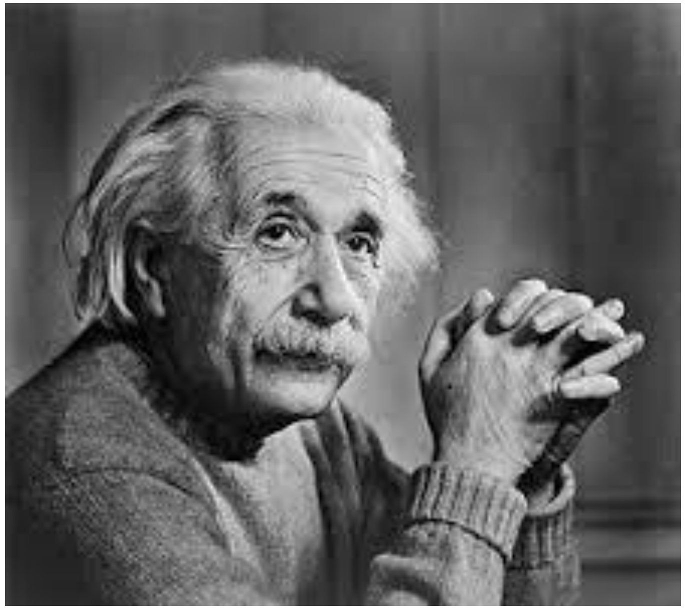
$$J_{MMI} = log\frac{p(\boldsymbol{x} \mid \boldsymbol{y})}{p(\boldsymbol{x})} = log\frac{p(\boldsymbol{y} \mid \boldsymbol{x})}{\cancel{p(\boldsymbol{y})}}$$

$$J_{CML} = logp(\boldsymbol{y} \mid \boldsymbol{x})$$

- LF-MMI: no division by the prior, uniform transition probabilities, using log-softmax prob. of states as the log of a pseudo-likelihood [Povey, et al., 2016]

- For the two manners - indirectly formulated as MMI training of a pseudo HMM [Povey, et al., 2016] or directly formulated as CML training of a CRF, it would be conceptually simpler to adopt the later manner.

- G. Heigold, et al., "Equivalence of generative and log-linear models", TASLP, 2011.
- D. Povey, et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI", INTERSPEECH 2016.



- Everything should be made as simple as possible, but not simpler.
- When the solution is simple, God is answering.

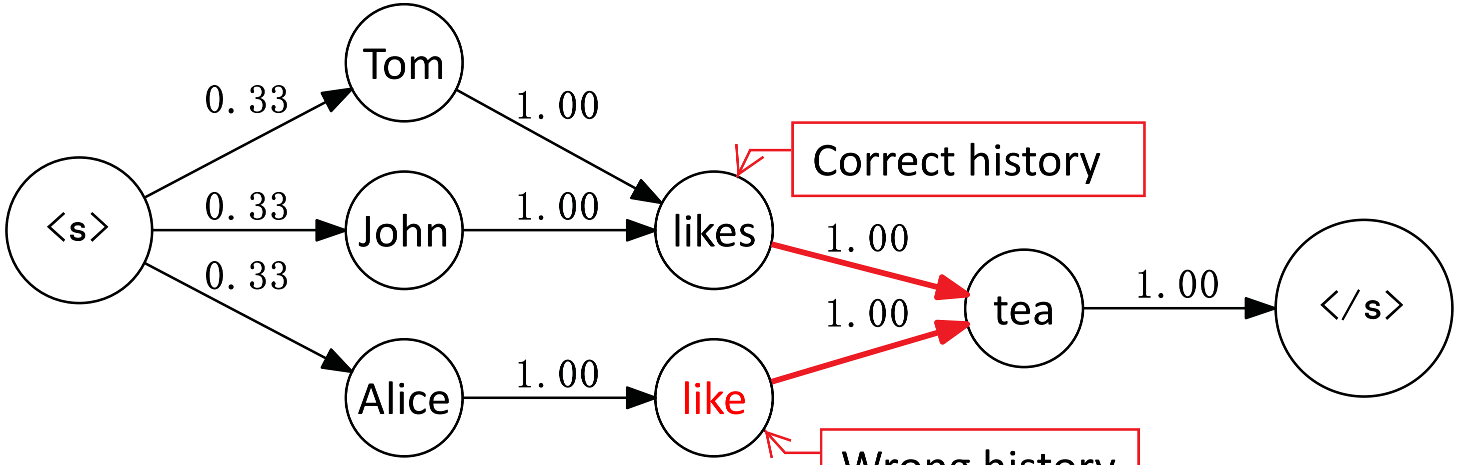
—— Albert Einstein

Label bias

▶ Word probabilities at each time-step are locally normalized, so successors of incorrect histories receive the same mass as do the successors of the true history. [Wiseman, et al., 2016]

Training data

Tom likes tea
John likes tea
Alice like tea



Correct history

Wrong history

▶ [Andor, et al., 2016]

- "Intuitively, we would like the model to be able to revise an earlier decision made during search, when later evidence becomes available that rules out the earlier decision as incorrect."
- "the label bias problem means that locally normalized models often have a very weak ability to revise earlier decisions."
- A proof that globally normalized models are strictly more expressive than locally normalized models.

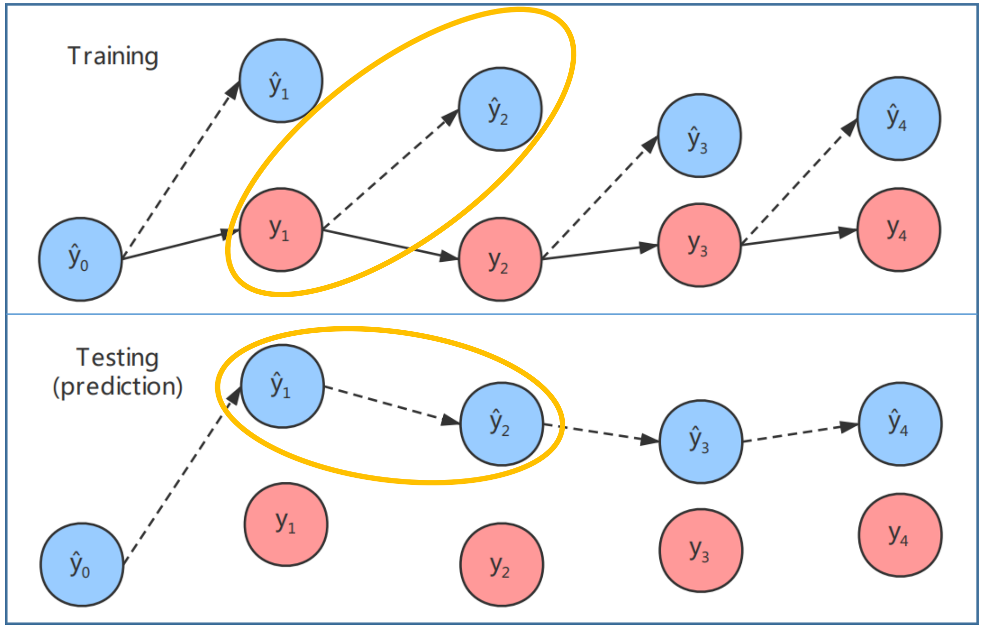
- Wiseman, et al., "Sequence-to-sequence Learning as Beam-Search Optimization", EMNLP, 2016.
- Andor, et al., "Globally Normalized Transition-Based Neural Networks", ACL, 2016.

Exposure bias

▶ Mismatch between **training** (teacher forcing) and **testing** (prediction) of locally-normalized sequence models [Wiseman, et al., 2016]:

- **Training**: maximize the likelihood of each successive target word, conditioned on the gold history of the target word.
- **Testing**: the model predict the next step, using its own predicted samples in testing.

▶ **Exposure bias** results from training in a certain way (maybe alleviated by scheduled sampling), **Label bias** results from properties of the model itself.



训练时 $y_t \rightarrow \hat{y}_{t+1}$，预测时 $\hat{y}_t \rightarrow \hat{y}_{t+1}$，不匹配

- Guo et al, A new GAN-based end-to-end TTS training algorithm, Interspeech 2019.

Conditional random field (CRF)

A CRF define a conditional distribution over output sequence y^l given input sequence x^l of length l :

$$p_\theta(y^l|x^l) = \frac{1}{Z_\theta(x^l)} \exp(u_\theta(x^l, y^l)) \qquad Z_\theta(x^l) = \sum_{y^l} \exp(u_\theta(x^l, y^l))$$

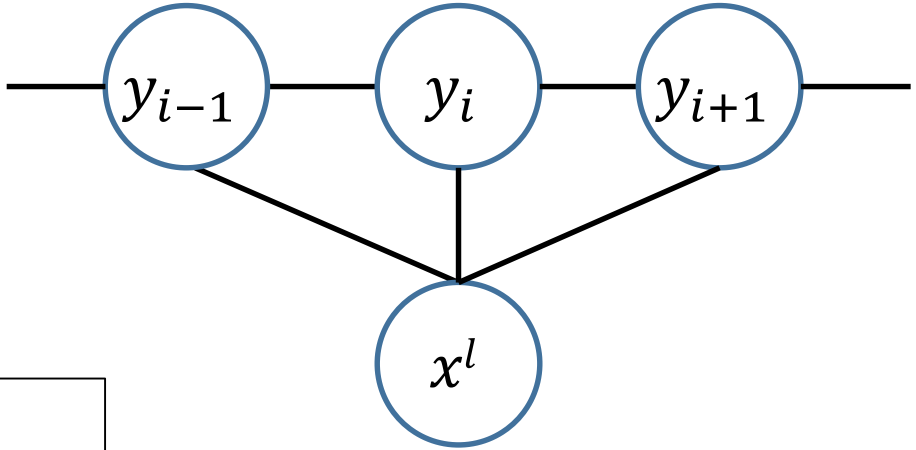
Potential for linear-chain:

Node potential

Edge potential

$$u_\theta(x^l, y^l) = \sum_{i=1}^{l} \phi_i(y_i, x^l) + \sum_{i=1}^{l} \psi_i(y_{i-1}, y_i, x^l)$$

☺ CRFs can overcome “label bias” and “exposure bias”.



Example of a linear-chain CRF

Successfully applied for sequence labeling in NLP, less so for ASR

- CRFs was explored for phone classification, using zero, first and second order features [Gunawardana, et al., 2005].
- CTC-CRF: the first CRF successfully developed for end-to-end ASR

A. Gunawardana, et al.,"Hidden conditional random fields for phone classification", Europspeech, 2005.

Training of Neural (linear-chain) CRFs

Model $p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp[u_\theta(x,y)]$, where $u_\theta(x,y) = \sum_t \phi_t^{y_t} + \sum_t A_{y_{t-1},y_t}$

For potential value $\phi_t^k, 1 \le t \le T, 1 \le k \le K$

$$\frac{\partial logp(y|x)}{\partial \phi_t^k} = \delta(y_t = k) - E_{p(y|x)}[\delta(y_t = k)]$$

$$= \delta(y_t = k) - p(y_t = k|x)$$

i.e., the **error signal** received by the NN feature extractor during training

i.e., γ_t^k, the posterior **state occupation probability**, calculated using the alpha-beta variables from the forward-backward algorithm [Rabiner, 1989]

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. Conditional random fields and sequence discriminative training

---

II. Improving end-to-end speech recognition (15*3=45 min)

QA + break 15-min

→ 1. Data-efficiency
2. Multilingual and crosslingual ASR
3. Language Model Integration for Transducer based Speech Recognition

III. Open questions and future directions + QA (15 min)

Section Content

1. Motivation
2. Related work
3. Method: **CTC-CRF**
4. Experiments
5. Conclusion

- H. Xiang, Z. Ou. "CRF-based Single-stage Acoustic Modeling with CTC Topology", ICASSP, 2019.
- K. An, H. Xiang, Z. Ou. "CAT: A CTC-CRF based ASR Toolkit Bridging the Hybrid and the End-to-end Approaches towards Data Efficiency and Low Latency", INTERSPEECH, 2020.
- Fan, et al., "The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines", SLT, 2021.
- H. Zheng, W. Peng, Z. Ou, J. Zhang. "Advancing CTC-CRF Based End-to-End Speech Recognition with Wordpieces and Conformers", arXiv:2107.03007, 2021.

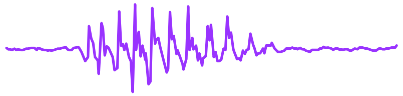
数据高效 Data-efficient

$$效率 = \frac{收益}{数据人工标注成本}$$

- 目前语音识别技术，过度依赖有监督学习和大量人工标注数据
- 这里的效率，不是指机器计算的效率（MIPS，million instructions per second），也不是指机器的能耗效率（MIPS/Watt），而是指

  —— 机器学习的效率
- 谱系化的数据高效的建模与学习方法
  - ✓ 模型架构
  - ✓ 无监督、半监督、自监督学习
  - ✓ 预训练
  - ✓ 迁移学习
  - ✓ 主动学习
  - ✓ 元学习

研究背景



语音信号

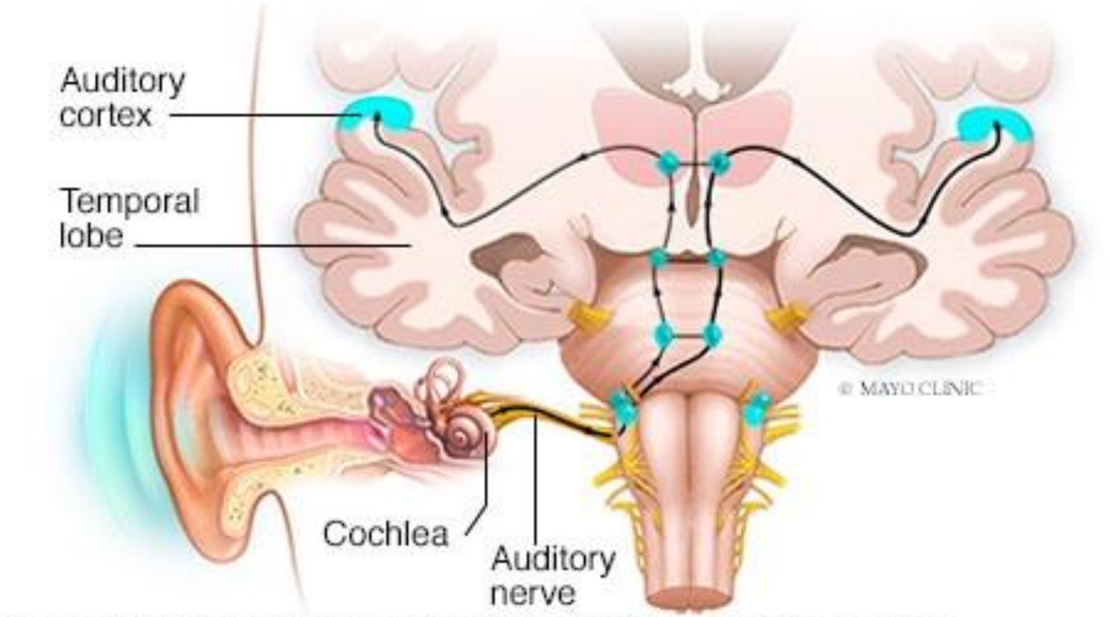
$X = x_1x_2\cdots$

$$P(Y|X) = \frac{\overbrace{P(X|Y)}^{\text{声学模型}}\,\overbrace{P(Y)}^{\text{语言模型}}}{\cancel{P(X)}}$$

欢迎参观清华电子系

词序列

$Y = y_1y_2\cdots$



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

题海

- 当前技术依赖

N种声学场景 * M种语言领域

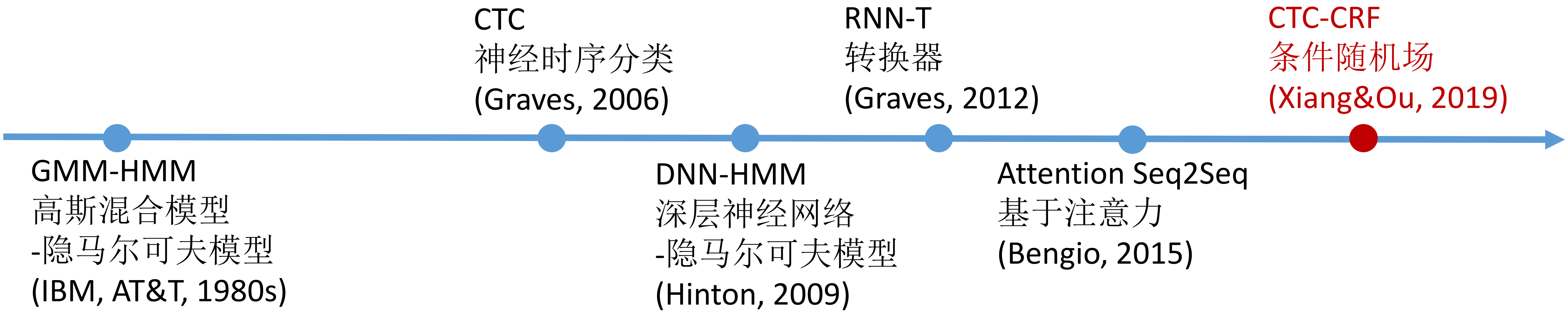
大量标注下有监督训练

- 适度模块化实现高效学习，

保留声学模型、语言模型的必要分解

技术挑战

语音识别模型 $P(X|Y)$ 发展历史：具有不同的图结构，不断进步



CTC
神经时序分类
(Graves, 2006)

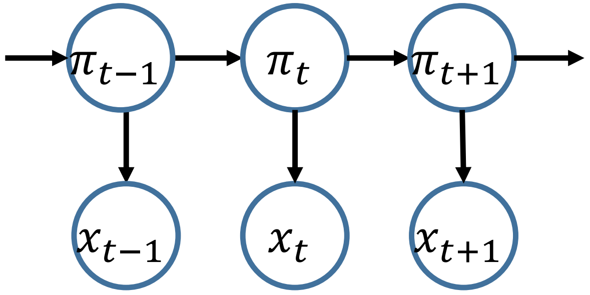
RNN-T
转换器
(Graves, 2012)

CTC-CRF
条件随机场
(Xiang&Ou, 2019)

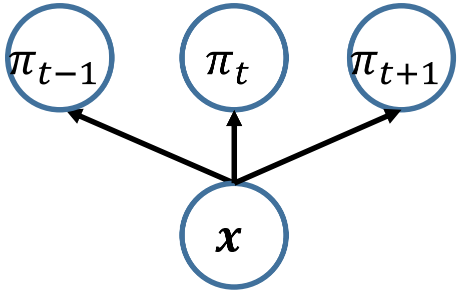
GMM-HMM
高斯混合模型
-隐马尔可夫模型
(IBM, AT&T, 1980s)

DNN-HMM
深层神经网络
-隐马尔可夫模型
(Hinton, 2009)

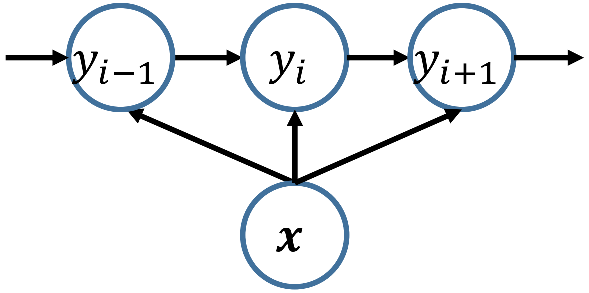
Attention Seq2Seq
基于注意力
(Bengio, 2015)



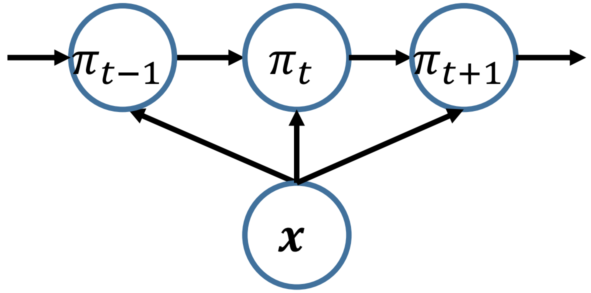
DNN-HMM
缺陷：多阶段



CTC
缺陷：$\{\pi_t\}$条件独立性

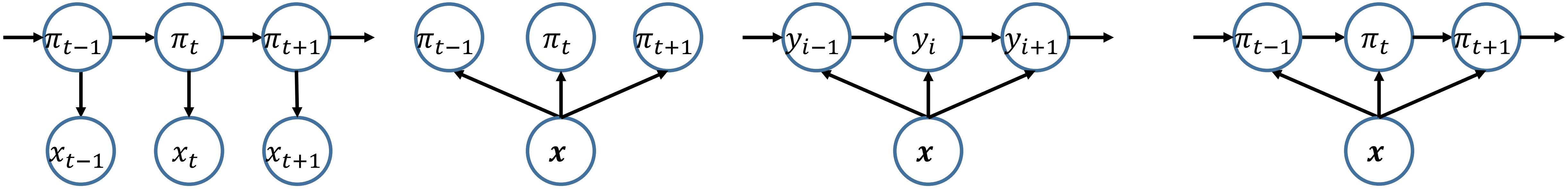


Attention
缺陷：$\{y_i\}$有向图序列模型
曝光偏置缺陷



RNN-T
缺陷：$\{\pi_t\}$有向图序列模型
曝光偏置缺陷

基于条件随机场的声学模型

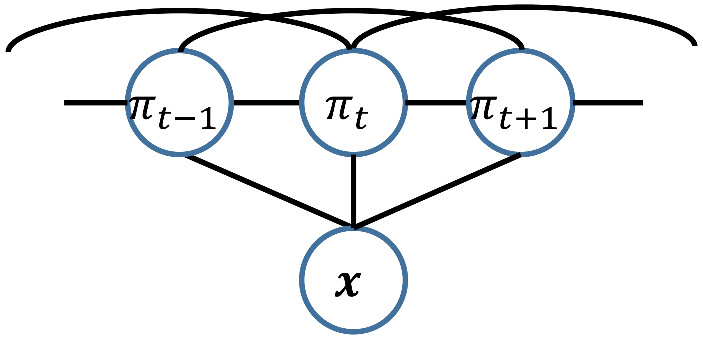


DNN-HMM
缺陷：多阶段

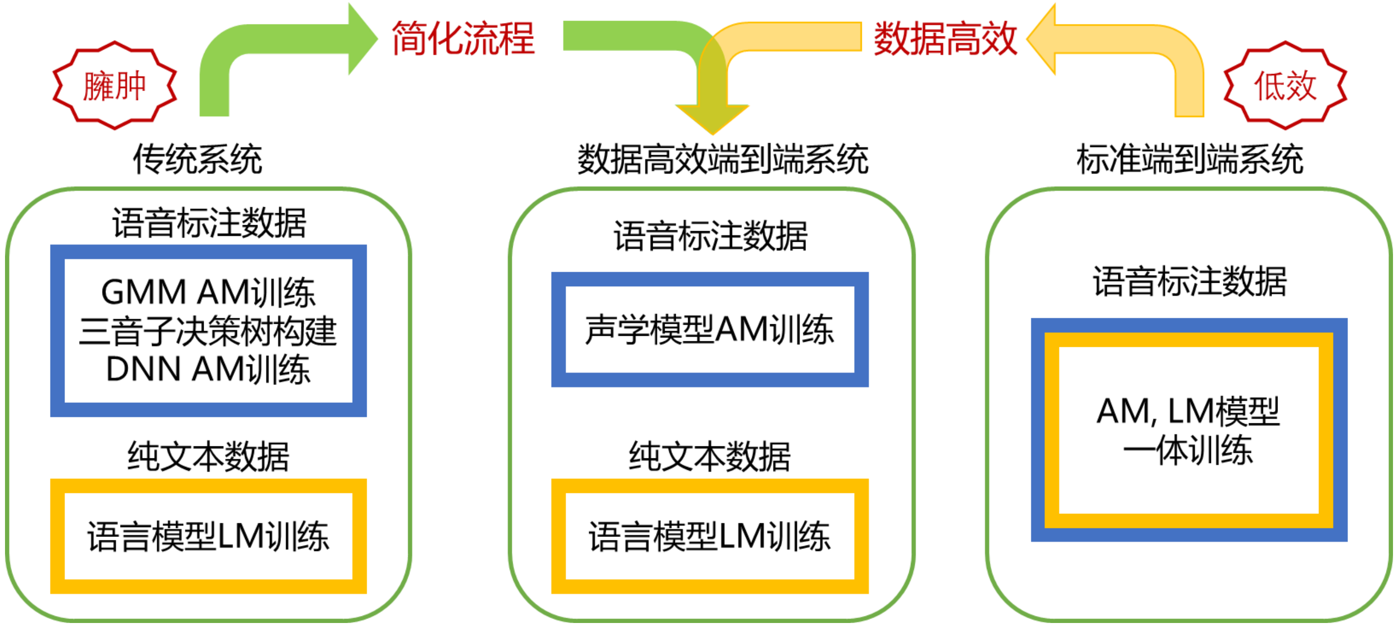
CTC
缺陷：$\{\pi_t\}$条件独立性

Attention
缺陷：$\{y_i\}$有向图序列模型

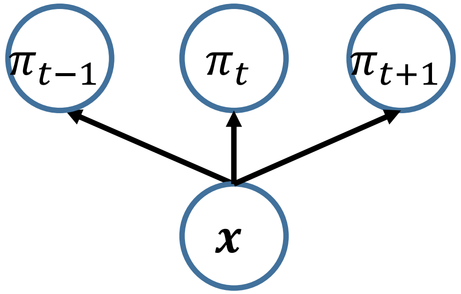
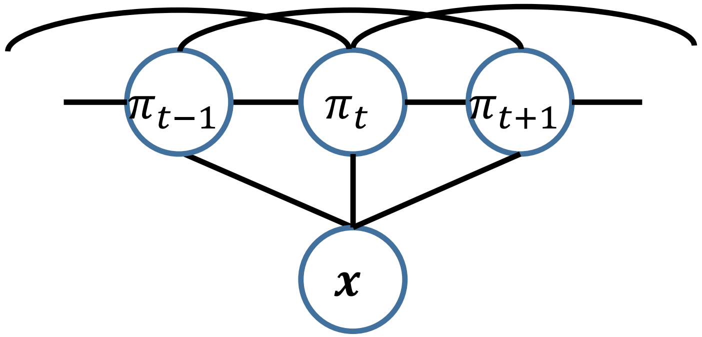
RNN-T
缺陷：$\{\pi_t\}$有向图序列模型



提出 CTC-CRF，占有独特位置，
克服了历史上各类模型的缺陷，助力数据高效，



CTC vs CTC-CRF

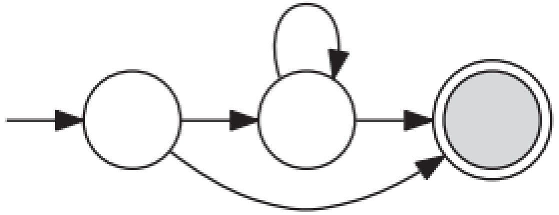
CTC	CTC-CRF
$p(\boldsymbol{y}\|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}:\mathcal{B}(\boldsymbol{\pi})=\boldsymbol{y}} p(\boldsymbol{\pi}\|\boldsymbol{x})$, using CTC topology $\mathcal{B}$	
State Independence $p(\boldsymbol{\pi}\|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\pi_t\|\boldsymbol{x})$	$p(\boldsymbol{\pi}\|\boldsymbol{x};\boldsymbol{\theta}) = \frac{e^{\phi(\boldsymbol{\pi},\boldsymbol{x};\boldsymbol{\theta})}}{\sum_{\boldsymbol{\pi}'} e^{\phi(\boldsymbol{\pi}',\boldsymbol{x};\boldsymbol{\theta})}}$ $\phi(\boldsymbol{\pi},\boldsymbol{x};\boldsymbol{\theta}) = \sum_{t=1}^{T} \begin{pmatrix} \log p(\pi_t\|\boldsymbol{x}) \\ +\log p_{LM}\ (\mathcal{B}(\boldsymbol{\pi})) \end{pmatrix}$ Node potential, by NN Edge potential, by n-gram denominator LM of labels, like in LF-MMI
$\frac{\partial \log p(\boldsymbol{y}\|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{\pi}\|\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial \log p(\boldsymbol{\pi}\|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$	$\frac{\partial \log p(\boldsymbol{y}\|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{\pi}\|\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}\left[\frac{\partial \phi(\boldsymbol{\pi},\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] - \mathbb{E}_{p(\boldsymbol{\pi}'\|\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial \phi(\boldsymbol{\pi}',\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$
	

Related work (SS-LF-MMI/EE-LF-MMI)

- Single-Stage (SS) Lattice-Free Maximum-Mutual-Information (LF-MMI)
  - 10 - 25% relative WER reduction on 80-h WSJ, 300-h Switchboard and 2000-h Fisher+Switchboard datasets, compared to CTC, Seq2Seq, RNN-T.
  - Cast as MMI-based discriminative training of an HMM (generative model) with

    Pseudo state-likelihoods calculated by the bottom DNN,

    Fixed state-transition probabilities.
  - 2-state HMM topology
  - Including a silence label



CTC-CRF

- Cast as a CRF;
- CTC topology;
- No silence label.

Hadian, et al., “Flat-start single-stage discriminatively trained HMM-based models for ASR”, T-ASLP 2018.

SS-LF-MMI vs CTC-CRF

	SS-LF-MMI	CTC-CRF
State topology	HMM topology with two states	CTC topology
Silence label	Using silence labels. Silence labels are randomly inserted when estimating denominator LM.	No silence labels. Use <blk> to absorb silence. ☺ No need to insert silence labels to transcripts.
Decoding	No spikes.	The posterior is dominated by <blk> and non-blank symbols occur in spikes. ☺ Speedup decoding by skipping blanks.
Implementation	Modify the utterance length to one of 30 lengths; use leaky HMM.	☺ No length modification; no leaky HMM.

Experiments

- We conduct our experiments on three benchmark datasets:
  - WSJ 80 hours
  - Switchboard 300 hours
  - Librispeech 1000 hours
- Acoustic model: 6 layer BLSTM with 320 hidden dim, 13M parameters
- Adam optimizer with an initial learning rate of 0.001, decreased to 0.0001 when cv loss does not decrease
- Implemented with Pytorch.
- Objective function (use the CTC objective function to help convergences):

$$\mathcal{J}_{CTC-CRF} + \alpha \mathcal{J}_{CTC}$$

- Decoding score function (use word-based language models, WFST based decoding):

$$\log p(\boldsymbol{l}|\boldsymbol{x}) + \beta \log p_{LM}(\boldsymbol{l})$$

H. Xiang, Z. Ou. "CRF-based Single-stage Acoustic Modeling with CTC Topology", ICASSP, 2019.

Experiments (Comparison with CTC, phone based)

WSJ 80h

Model	Unit	LM	SP	dev93	eval92	
CTC	Mono-phone	4-gram	N	10.81%	7.02%	44.4%
CTC-CRF	Mono-phone	4-gram	N	6.24%	3.90%	

Switchboard 300h

Model	Unit	LM	SP	SW		CH	
CTC	Mono-phone	4-gram	N	12.9%	14.7%	23.6%	11%
CTC-CRF	Mono-phone	4-gram	N	11.0%		21.0%	

Librispeech 1000h

Model	Unit	LM	SP	Dev Clean	Dev Other	Test Clean		Test Other	
CTC	Mono-phone	4-gram	N	4.64%	13.23%	5.06%	19.1%	13.68%	22.1%
CTC-CRF	Mono-phone	4-gram	N	3.87%	10.28%	4.09%		10.65%	

SP: speed perturbation for 3-fold data augmentation.

Experiments (Comparison with STOA)

Switchboard 300h

Model	SW	CH	Average	Source
Kaldi chain triphone	9.6	19.3	14.5	IS 2016
Kaldi e2e chain monophone	11.0	20.7	15.9	ASLP 2018, 26M
Kaldi e2e chain biphone	9.8	19.3	14.6	ASLP 2018, 26M
CTC-CRF monophone	10.3	19.7	15.0	ICASSP 2019, BLSTM, 13M
CTC-CRF monophone	**9.8**	**18.8**	**14.3**	**IS 2020, VGG BLSTM, 16M**

10%

RWTH IS 2018, “Improved training of end-to-end attention models for speech recognition”.
RWTH IS 2019, “RWTH ASR Systems for LibriSpeech Hybrid vs Attention -- Data Augmentation”.
IBM IS19, “Forget a Bit to Learn Better Soft Forgetting for CTC-based Automatic Speech Recognition”.
Espnet ASRU19, “Espresso: A Fast End-to-end Neural Speech Recognition Toolkit”.
Google IS19, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”.

Experiments (Comparison with STOA)

Librispeech 1000h

Model	Test Clean	Test Other	Source
Kaldi chain triphone	4.28	-	IS 2016
CTC-CRF monophone	**4.0**	**10.6**	**ICASSP 2019, BLSTM (6,320), 13M**

RWTH IS 2018, “Improved training of end-to-end attention models for speech recognition”.
RWTH IS 2019, “RWTH ASR Systems for LibriSpeech Hybrid vs Attention -- Data Augmentation”.
IBM IS19, “Forget a Bit to Learn Better Soft Forgetting for CTC-based Automatic Speech Recognition”.
Espnet ASRU19, “Espresso: A Fast End-to-end Neural Speech Recognition Toolkit”.
Google IS19, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”.

Mandarin Aishell-1 results

- 170 hours mandarin speech corpus
- 400 speakers from different accent areas
- 15% CER reduction compared with LF-MMI
- 5% CER reduction compared with end-to-end transformer

Model	%CER
LF-MMI with i-vector [1]	7.43
Transformer [2]	6.7
CTC-CRF [3]	6.34

https://github.com/thu-spmi/ASR-Benchmarks

Measure the progress in a more scientifically way!

[1] D. Povey, A. Ghoshal, and et al, “The Kaldi speech recognition toolkit,” ASRU 2011.

[2] S. Karita, N. Chen, and et al, “A comparative study on transformer vs RNN in speech applications,” ASRU 2019.

[3] Keyu An, Hongyu Xiang, and **Zhijian Ou**, “CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency,” INTERSPEECH 2020.

2021 SLT CHILDREN SPEECH RECOGNITION CHALLENGE (CSRC)

- 400 hours of data, targeting to boost children speech recognition research.
- Evaluated on 10 hours of children’s reading and conversational speech.
- 3 baselines (Chain model, Transformer and CTC-CRF) are provided.

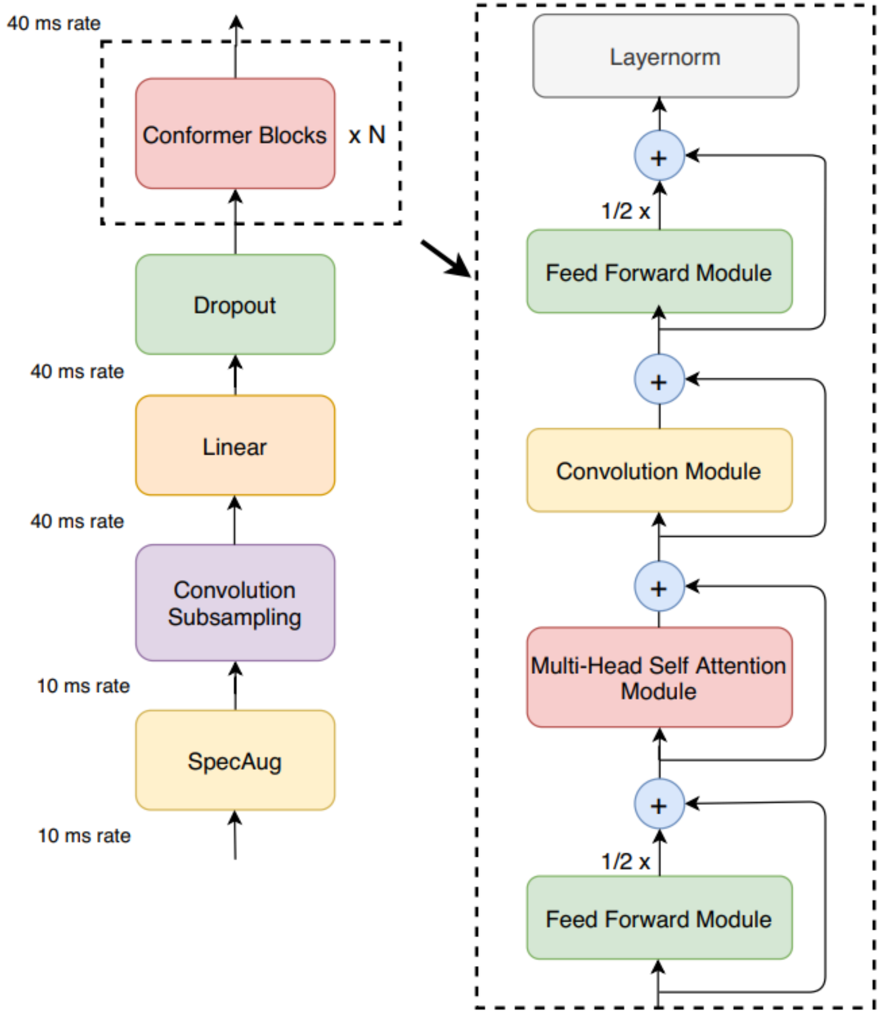
model	Chain model	Transformer	CTC-CRF
CER%	28.75	27.28	**25.34**

Fan Yu, Zhuoyuan Yao, Xiong Wang, Keyu An, Lei Xie, **Zhijian Ou**, Bo Liu, Xiulin Li, Guanqiong Miao. The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines. SLT 2021.

Advancing CTC-CRF Based End-to-End Speech Recognition with Wordpieces and Conformers

Huahuan Zheng, Wenjie Peng, **Zhijian Ou** and Jinsong Zhang, arXiv:2107.03007

Basic Units of Labels	**Label Sequence**
phoneme	DH AE1 T N IY1 DH ER0 AH1 V DH EH1 M HH AE1 D K R AO1 S T DH AH0 TH R EH1 SH OW2 L D S IH1 N S DH AH0 D AA1 R K D EY1
character /grapheme	t h a t _ n e i t h e r _ o f _ t h e m _ h a d _ c r o s s e d _ t h e _ t h r e s h o l d _ s i n c e _ t h e _ d a r k _ d a y _
subword /wordpiece	that_ ne i ther_ of_ them_ had_ cro s sed_ the_ th re sh old_ sin ce_ the_ d ar k_ day_
word	that neither of them had crossed the threshold since the dark day



Experiments (Comparison between different units, WER%)

Switchboard 300h

Model	Unit	LM	Augmentation	Eval2000	SW	CH
Conformer (this work)	monophone	4-gram	SP, SA	12.1	7.9	16.1
	monophone	Trans.*	SP, SA	10.7	6.9	14.5
	wordpiece	4-gram	SP, SA	12.7	8.7	16.5
	wordpiece	Trans.*	SP, SA	11.1	7.2	14.8

Librispeech 1000h

Model	Unit	LM	Augmentation	Test Clean	Test Other
Conformer (this work)	monophone	4-gram	SA	3.61	8.10
	monophone	Trans.**	SA	2.51	5.95
	wordpiece	4-gram	SA	3.59	8.37
	wordpiece	Trans.**	SA	2.54	6.33

SP: speed perturbation for 3-fold data augmentation.

SA: our implementation of SpecAug with ratio

* Latest Kaldi Transformer LM rescoring

** RWTH 42-layer Transformer

English: a low degree of grapheme-phoneme correspondence

Experiments (Comparison between different units, WER%)

CommonVoice German 700h

Model	#params	unit	LM	Augmentation	Test
Conformer (This work)	25.03	char	4-gram	SP, SA	12.7
	25.03	char	Trans.	SP, SA	11.6
	25.03	monophone	4-gram	SP, SA	10.7
	25.03	monophone	Trans.	SP, SA	10.0
	25.06	wordpiece	4-gram	SP, SA	10.5
	25.06	wordpiece	Trans.	SP, SA	9.8

German: a high degree of grapheme-phoneme correspondence

Experiments (Comparison with STOA)

Switchboard 300h

Model	#params	LM	unit	SW	CH	Eval2000
RNN-T, 2021 [10]	57	RNN LM	char	6.4	13.4	9.9
Conformer [9]	44.6	Trans.	bpe	6.8	14.0	10.4
TDNN-F [11]	-	Trans.*	triphone	7.2	14.4	10.8
TDNN-F [11]	-	Trans.**	triphone	6.5	13.9	10.2
VGGBLSTM [2]	39.15	RNN LM	monophone	8.8	17.4	[13.0]
Conformer (This work)	51.82	Trans.	monophone	6.9	14.5	10.7
	51.85	Trans.	wordpiece	7.2	14.8	11.1

* N-best rescoring, ** Iterative lattice rescoring

[2] “CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency,” INTERSPEECH 2020.
[9] “Conformer: Convolution-augmented Transformer for Speech Recognition”, Interspeech 2020.
[10] “Advancing RNN transducer technology for speech recognition,” ICASSP 2021.
[11] “A paralleliz- able lattice rescoring strategy with neural language models,” ICASSP, 2021

Section Conclusion

- The CTC-CRF framework inherits the data-efficiency of the hybrid approach and the simplicity of the end-to-end approach.
- CTC-CRF significantly outperforms regular CTC on a wide range of benchmarks, and is on par with other state-of-the-art end-to-end models.
  - English WSJ-80h, Switchboard-300h, Librispeech-1000h; Mandarin Aishell-170h; …
- Flexibility
  - Streaming ASR <- INTRESPEECH 2020
  - Neural Architecture Search <- SLT 2021
  - Children Speech Recognition <- SLT 2021
  - Wordpieces, Conformer architectures
  - Multilingual and Crosslingual <- ASRU2021
  - CUSIDE: streaming ASR <- INTERSPEECH 2022
  - LODR: LM integration <- INTERSPEECH 2022

https://github.com/thu-spmi/cat

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. Conditional random fields and sequence discriminative training

II. Improving end-to-end speech recognition (15*3=45 min)

QA + break 15-min

1. Data-efficiency
2. → Multilingual and crosslingual ASR
3. Language Model Integration for Transducer based Speech Recognition

III. Open questions and future directions + QA (15 min)

Section Content

1. Motivation
2. Related work
3. Method: **JoinAP**
4. Experiments
5. Conclusion

结合声学（Acoustic）和音韵学（Phonology），促进多语言信息共享与迁移

- Chengrui Zhu, Keyu An, Huahuan Zheng, Zhijian Ou. “Multilingual and Crosslingual Speech Recognition using Phonological-Vector based Phone Embeddings”, ASRU 2021.

Motivation

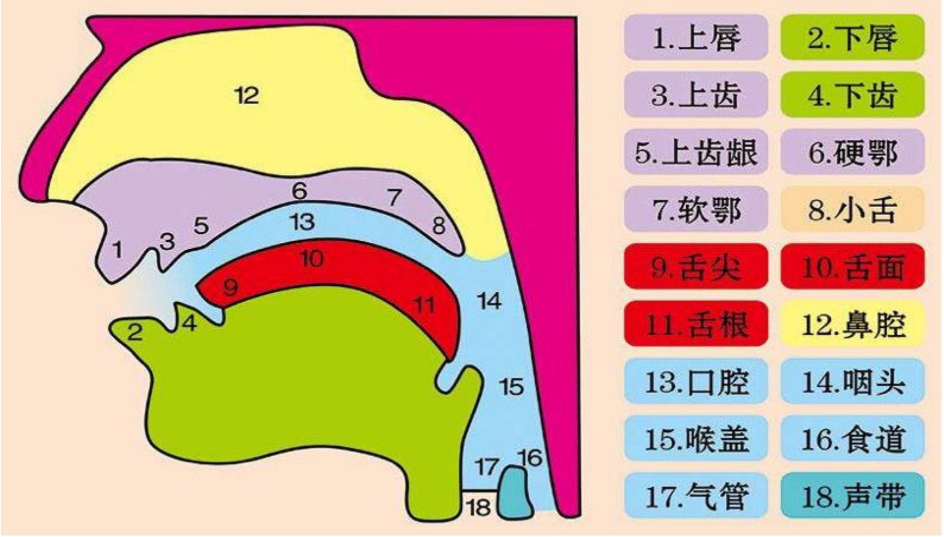
- There are more than 7100 languages in the world, and most of them are low-resourced languages.
- Multilingual speech recognition
  - Training data from a number of languages **(seen languages)** are merged to train a multilingual AM.
- Crosslingual speech recognition
  - The target language is **unseen** in training the multilingual AM.
  - In few-shot setting , the AM can be finetuned on limited target language data.
  - In zero-shot setting , the AM is directly used without finetuning*.

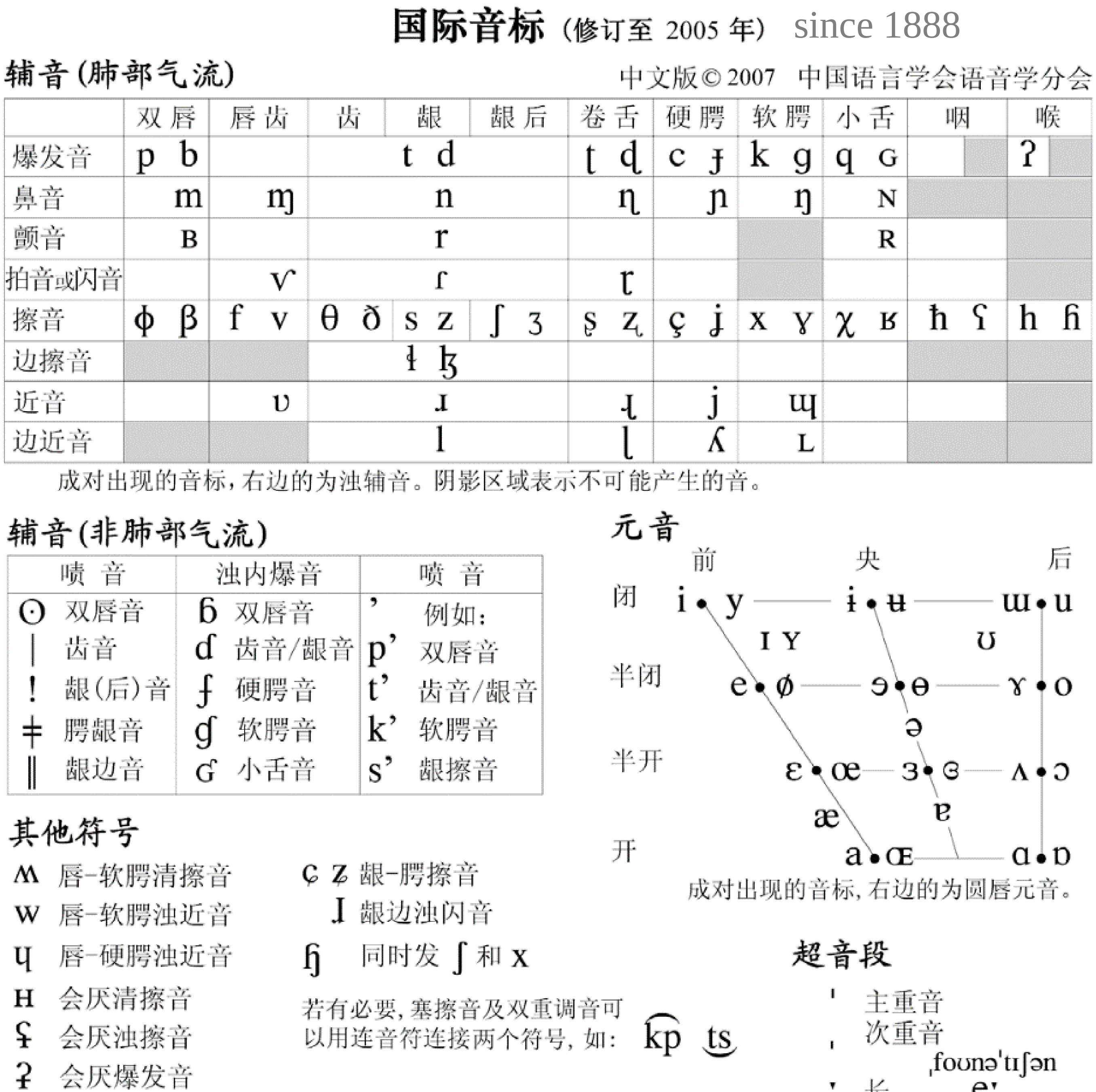
* Suppose that text corpus from the target language are available.

Intuitively, the key to successful multilingual and crosslingual recognition is
to promote the information sharing in multilingual training
and maximize the knowledge transferring from the well trained multilingual model to the model
for recognizing the utterances in the new language.

Universal Phone Set

无论哪种人类语言，都是人类的一套发音器官发出来的音





国际音标（修订至 2005 年） since 1888

辅音(肺部气流)

中文版© 2007 中国语言学会语音学分会

	双唇	唇齿	齿	龈	龈后	卷舌	硬腭	软腭	小舌	咽	喉
爆发音	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
鼻音	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
颤音	ʙ			r				▨	ʀ		▨
拍音或闪音		ⱱ		ɾ		ɽ		▨			▨
擦音	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
边擦音	▨	▨		ɬ ɮ						▨	▨
近音		ʋ		ɹ		ɻ	j	ɰ			▨
边近音	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

成对出现的音标，右边的为浊辅音。阴影区域表示不可能产生的音。

辅音(非肺部气流)

啧 音		浊内爆音		喷 音	
ʘ	双唇音	ɓ	双唇音	ʼ	例如:
ǀ	齿音	ɗ	齿音/龈音	pʼ	双唇音
ǃ	龈(后)音	ʄ	硬腭音	tʼ	齿音/龈音
ǂ	腭龈音	ɠ	软腭音	kʼ	软腭音
ǁ	龈边音	ʛ	小舌音	sʼ	龈擦音

元音

	前		央		后
闭	i • y		ɨ • ʉ		ɯ • u
		ɪ ʏ		ʊ	
半闭	e • ø		ɘ • ɵ		ɤ • o
			ə		
半开	ɛ • œ		ɜ • ɞ		ʌ • ɔ
		æ	ɐ		
开	a • ɶ				ɑ • ɒ

成对出现的音标，右边的为圆唇元音。

其他符号

ʍ	唇-软腭清擦音
w	唇-软腭浊近音
ɥ	唇-硬腭浊近音
ʜ	会厌清擦音
ʢ	会厌浊擦音
ʡ	会厌爆发音
ɕ ʑ	龈-腭擦音
ɺ	龈边浊闪音
ɧ	同时发 ʃ 和 x

若有必要，塞擦音及双重调音可以用连音符连接两个符号，如: k͡p t͜s

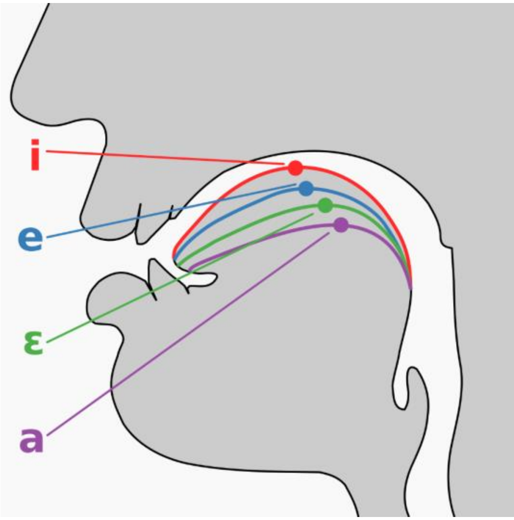
超音段

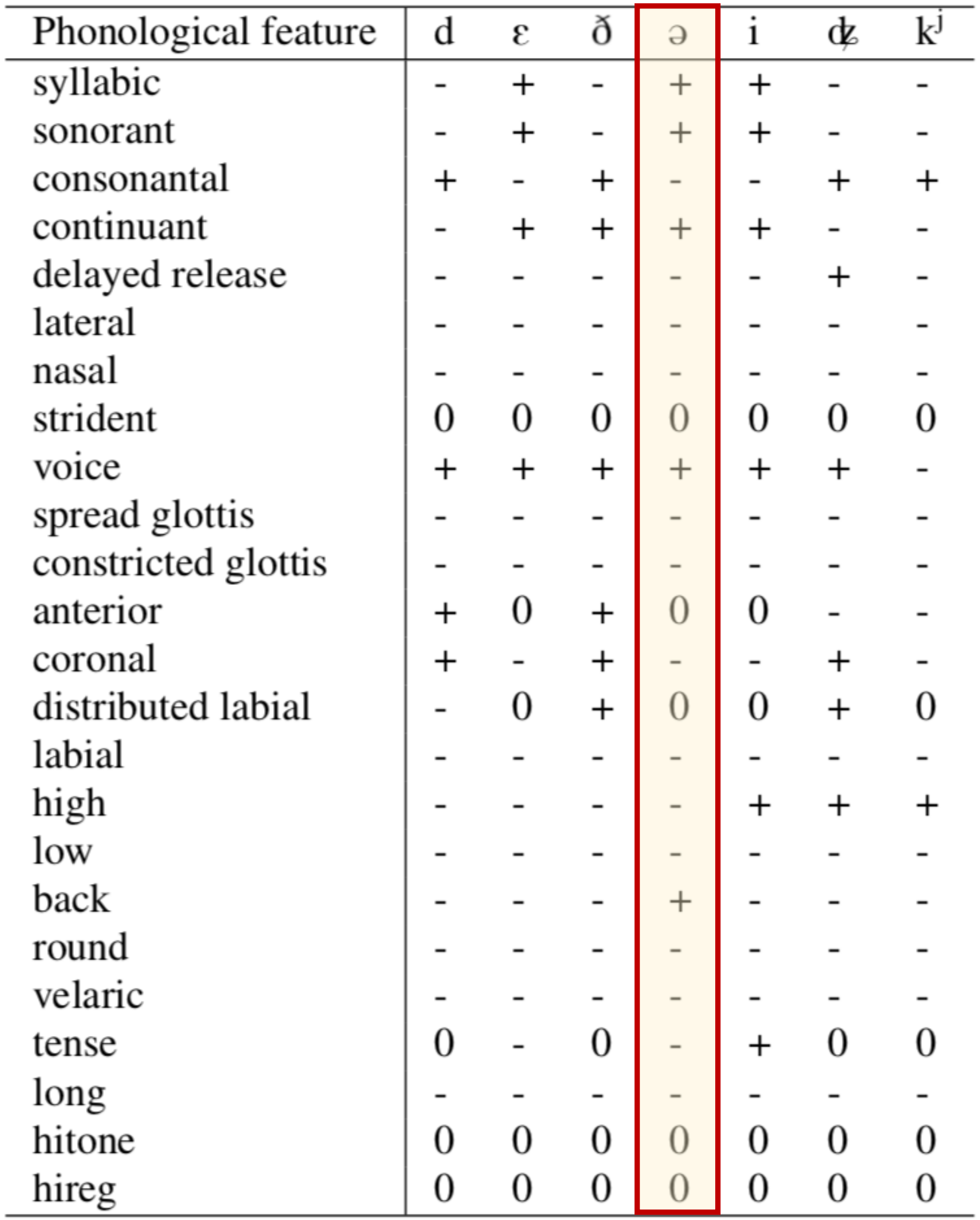
ˈ	主重音
ˌ	次重音
	ˌfoʊnəˈtɪʃən
ː	长 eː

Phonological features

- Often phones are seen as being the “atoms” of speech.
- But it is now widely accepted in phonology that phones are decomposable into smaller, more fundamental units, sharable across all languages, called phonological (distinctive) features.

- Describe phones by phonological features
  - Vowels
    - vowel height
    - vowel backness
  - Consonants
    - Place of articulation
    - Manner of articulation

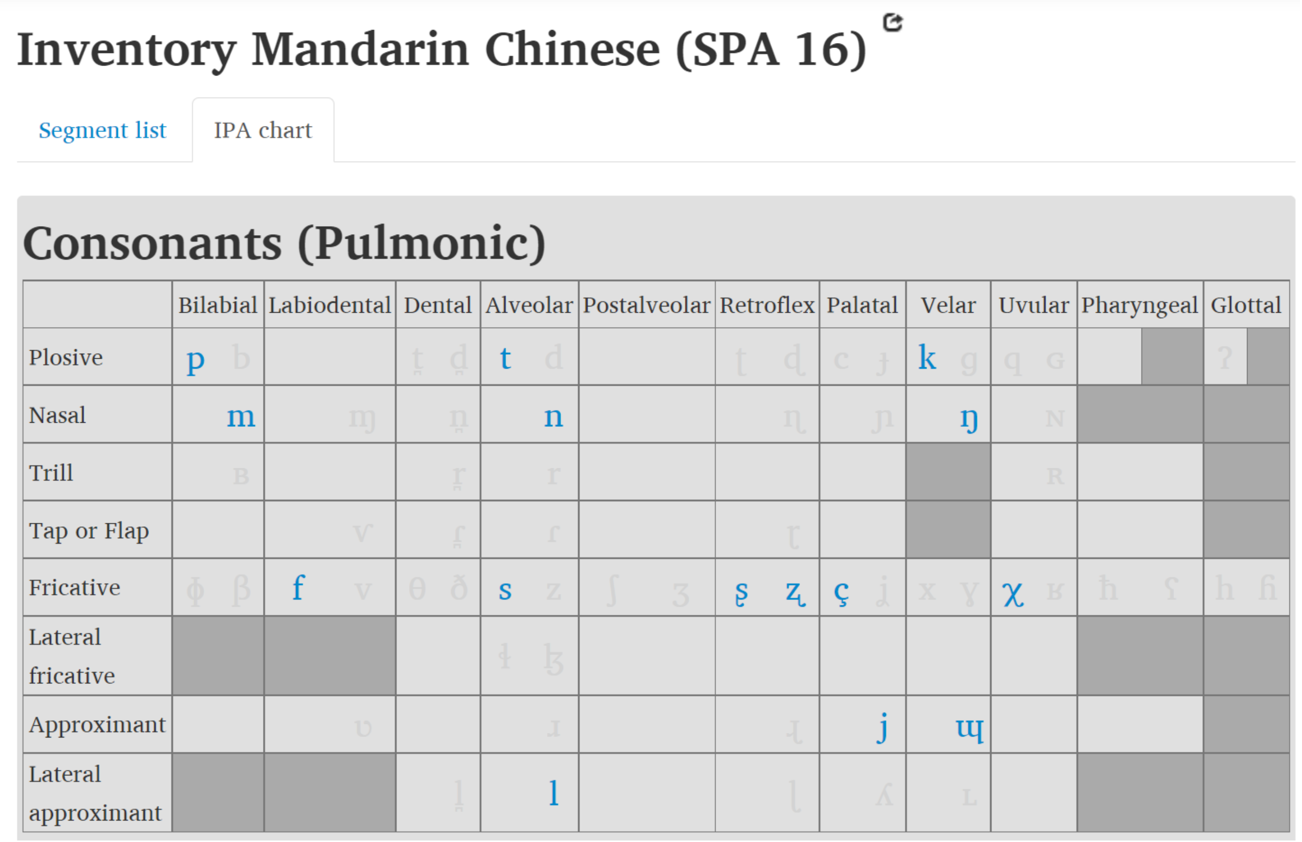




Phonological feature	d	ɛ	ð	ə	i	ʥ	kʲ
syllabic	-	+	-	+	+	-	-
sonorant	-	+	-	+	+	-	-
consonantal	+	-	+	-	-	+	+
continuant	-	+	+	+	+	-	-
delayed release	-	-	-	-	-	+	-
lateral	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-
strident	0	0	0	0	0	0	0
voice	+	+	+	+	+	+	-
spread glottis	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-
anterior	+	0	+	0	0	-	-
coronal	+	-	+	-	-	+	-
distributed labial	-	0	+	0	0	+	0
labial	-	-	-	-	-	-	-
high	-	-	-	-	+	+	+
low	-	-	-	-	-	-	-
back	-	-	-	+	-	-	-
round	-	-	-	-	-	-	-
velaric	-	-	-	-	-	-	-
tense	0	-	0	-	+	0	0
long	-	-	-	-	-	-	-
hitone	0	0	0	0	0	0	0
hireg	0	0	0	0	0	0	0

https://phoible.org/

- Steven Paul Moran, “Phonetics Information Base and Lexicon”, PhD Thesis, UofW, 2012.
- Release 2.0 from 2019 includes 3020 inventories that contain 3183 segment types found in **2186** distinct languages.
- In addition to **phoneme inventories**, PHOIBLE includes **distinctive feature data** for every phoneme in every language.

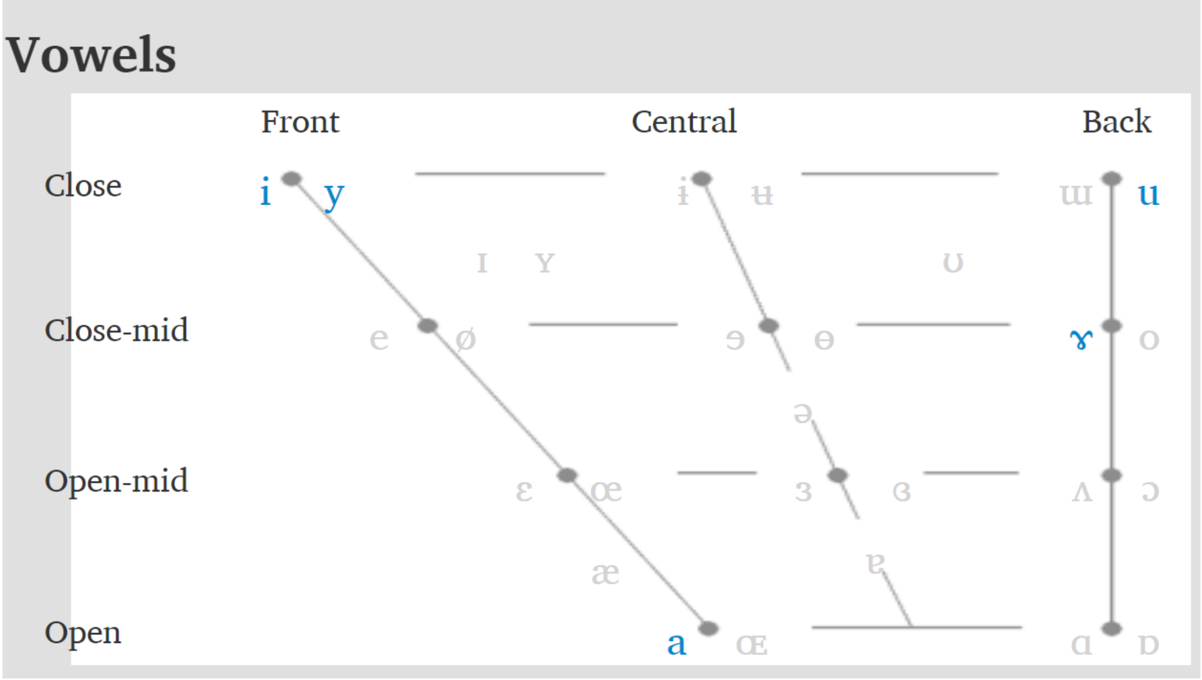


Inventory Mandarin Chinese (SPA 16)

Segment list | IPA chart

Consonants (Pulmonic)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t̪ d̪	t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ	n̪	n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ		r̪	r					ʀ		
Tap or Flap		ⱱ	ɾ̪	ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant			l̪	l		ɭ	ʎ	ʟ			



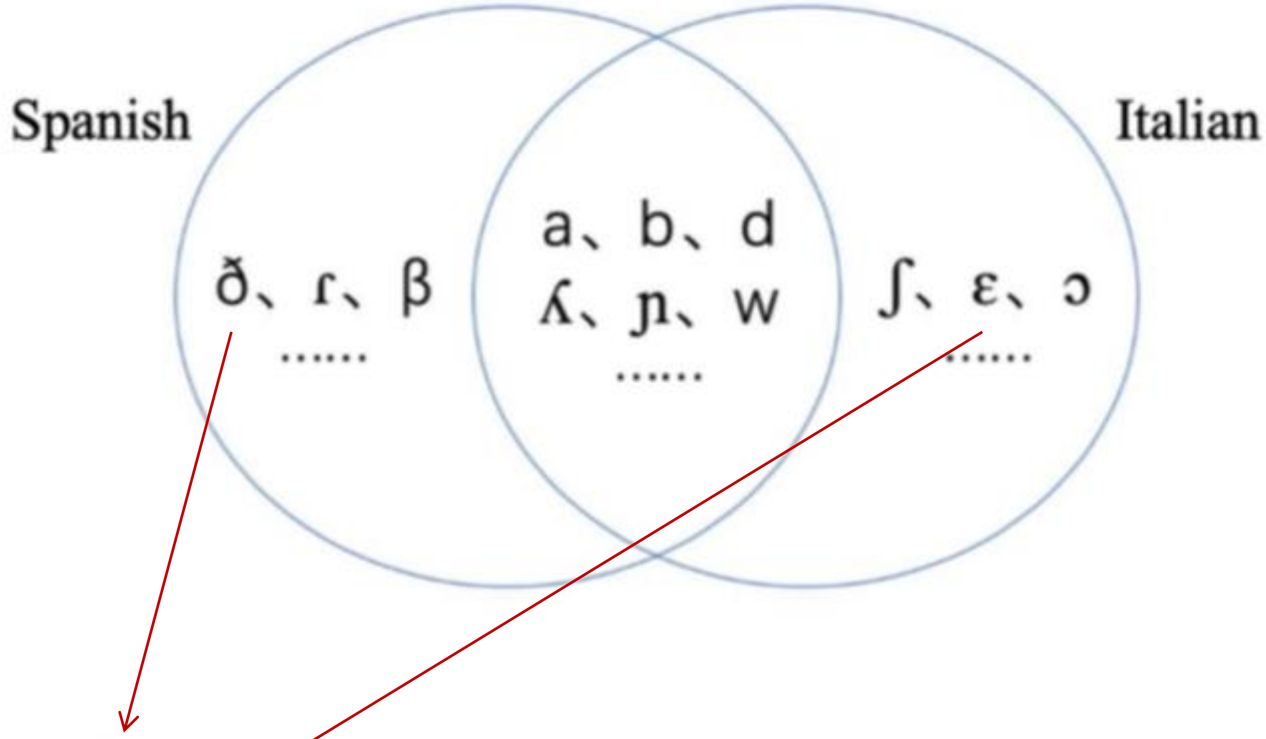
Phonological features: micro-decomposition of phones

- Like atoms could be split into nucleus and electrons, phones can be expressed by phonological features.

物质 Matter	语音 Speech
元素 Atoms	音素 Phones
元素周期表 Periodic table of elements	国际音标表 IPA table
原子核、电子 Nucleus, electrons	音韵特征 Phonological features

Phonological features: promote information sharing

- Even language-specific phones are connected by using phonological features.

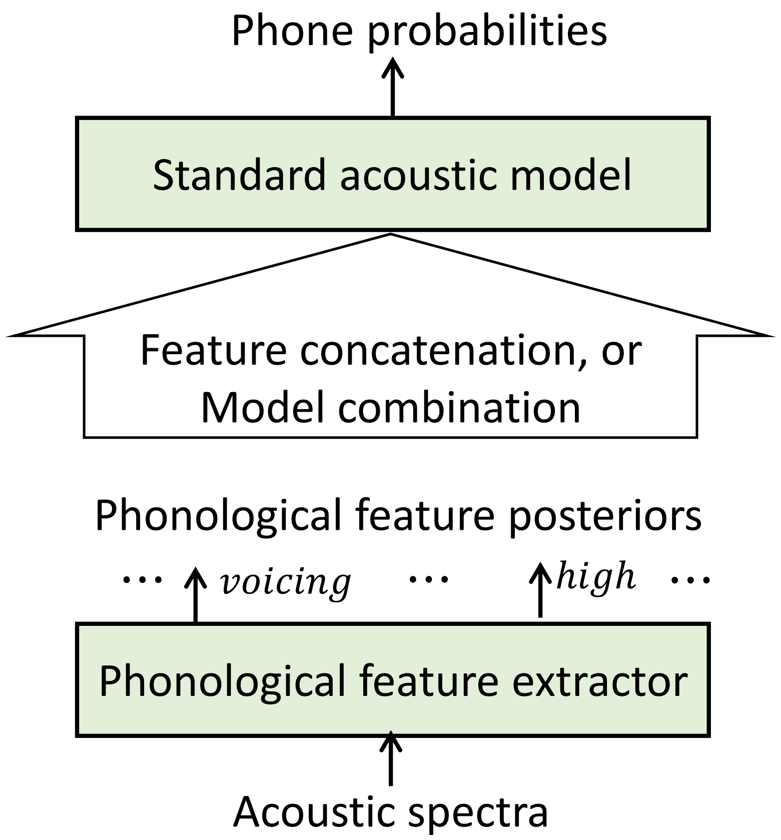


ð：-,-,+,+,-,-,-,0,+,-,-,+,+,+,-,-,-,-,-,-,0,-,0,0

ɛ：+,+,-,+,-,-,-,0,+,-,-,0,-,0,-,-,-,+,-,-,+,-,0,0

Related work

- Phonological features(PFs) have been applied in multilingual and crosslingual ASR
- Previous studies generally take a bottom-up approach, and suffer from:
  - The acoustic-to-PF extraction in a bottom-up way is itself difficult.
  - Do not provide a principled model to calculate the phone probabilities for unseen phones from the new language towards zero-shot crosslingual recognition.



From phonological features to phonological-vector

- Phonological-vector
  - Encode each phonological feature by a 2-bit binary vector. (24PFs -> 48bits)

+	-	0
10	01	00

  - Plus 3 bits to indicate <blk>, <spn>, <nsn>
  - Phonological-vector: Total 51 bits

Joining of Acoustics and Phonology (JoinAP)

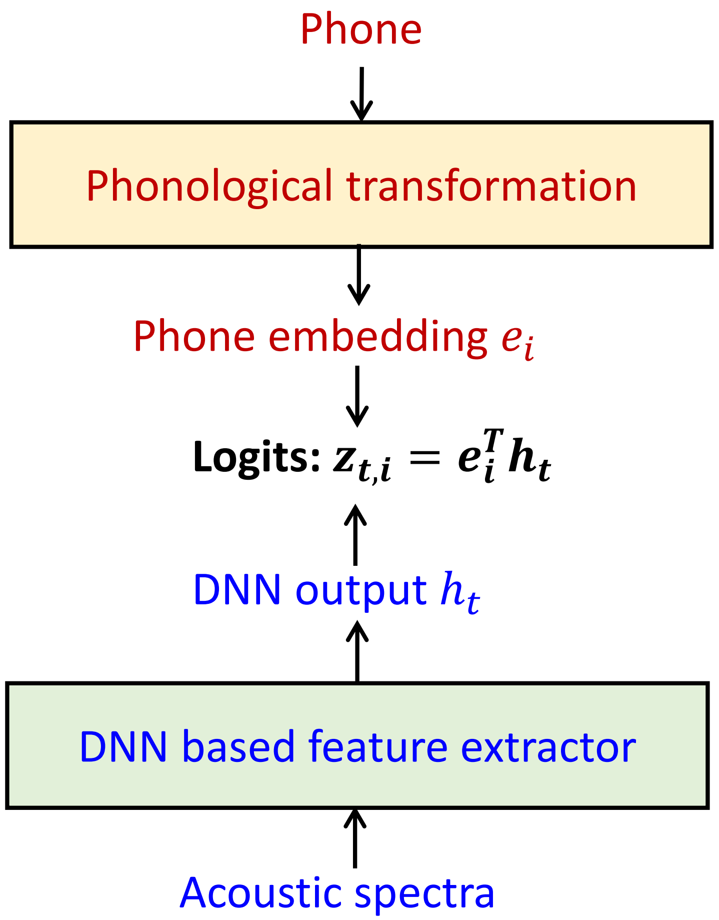
- The JoinAP method
  - DNN based acoustic feature extraction (bottom-up) and phonology driven phone embedding (top-down) are joined to calculate the **logits**.

- JoinAP-Linear
  - Linear transformation of phonological-vector p_i to define the embedding vector for phone i:

$$e_i = Ap_i \in \mathbb{R}^H$$

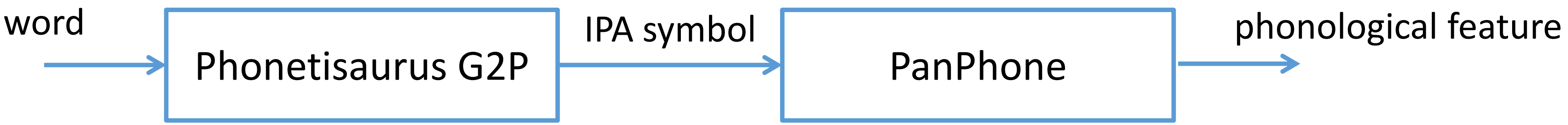
- JoinAP-Nonlinear
  - Apply nonlinear transformation, multilayered neural networks:

$$e_i = A_2\sigma(A_1 p_i) \in \mathbb{R}^H$$

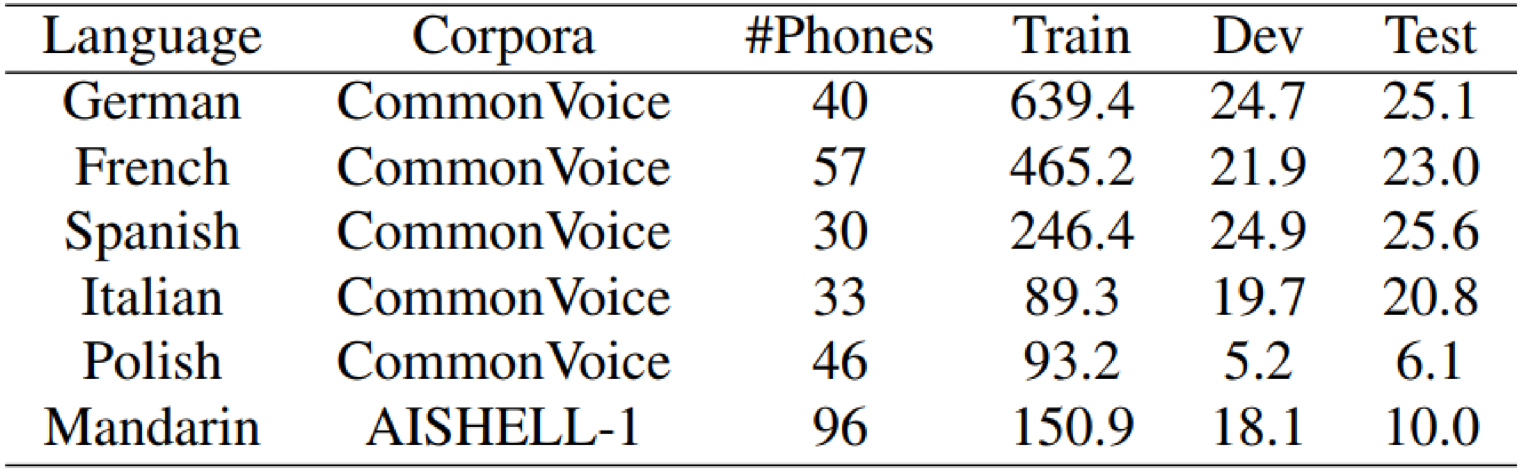

Phone
Phonological transformation
Phone embedding e_i
Logits: $\boldsymbol{z_{t,i} = e_i^T h_t}$
DNN output h_t
DNN based feature extractor
Acoustic spectra

Experiments

- Train multilingual AM on German, French, Spanish and Polish.
- Zero-shot and few-shot crosslingual ASR on Polish and Mandarin.



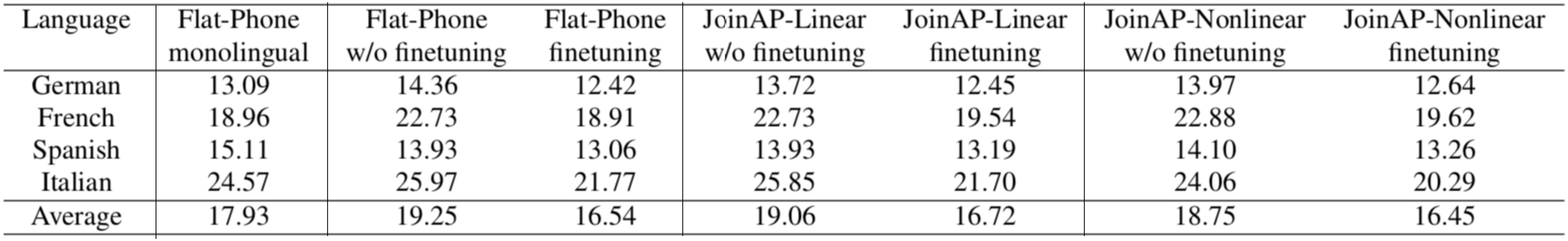
- Use CTC-CRF based ASR toolkit, CAT
  - Acoustic model: 3 layer VGGBLSTM with 1024 hidden dim
  - Adam optimizer: with an initial learning rate of 0.001, decreased to 1/10 until less than 0.00001
  - Dropout 0.5



Language	Corpora	#Phones	Train	Dev	Test
German	CommonVoice	40	639.4	24.7	25.1
French	CommonVoice	57	465.2	21.9	23.0
Spanish	CommonVoice	30	246.4	24.9	25.6
Italian	CommonVoice	33	89.3	19.7	20.8
Polish	CommonVoice	46	93.2	5.2	6.1
Mandarin	AISHELL-1	96	150.9	18.1	10.0

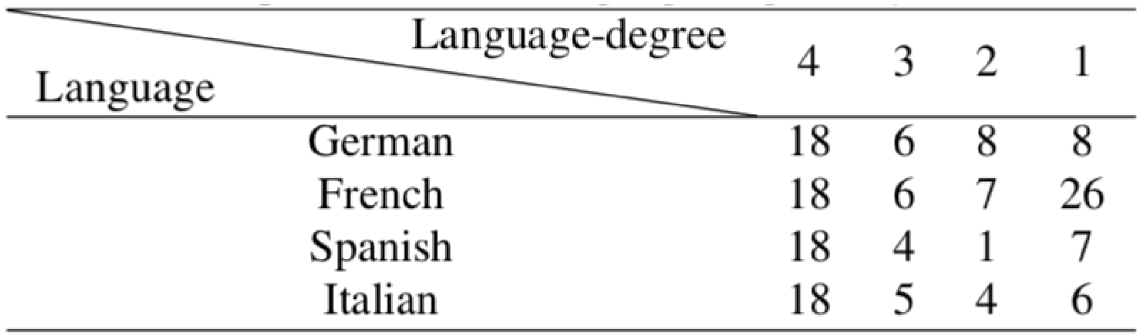
Experiments

- Multilingual experiments



Language	Flat-Phone monolingual	Flat-Phone w/o finetuning	Flat-Phone finetuning	JoinAP-Linear w/o finetuning	JoinAP-Linear finetuning	JoinAP-Nonlinear w/o finetuning	JoinAP-Nonlinear finetuning
German	13.09	14.36	12.42	13.72	12.45	13.97	12.64
French	18.96	22.73	18.91	22.73	19.54	22.88	19.62
Spanish	15.11	13.93	13.06	13.93	13.19	14.10	13.26
Italian	24.57	25.97	21.77	25.85	21.70	24.06	20.29
Average	17.93	19.25	16.54	19.06	16.72	18.75	16.45

- Language-degree of a phone: how many languages a phone appears

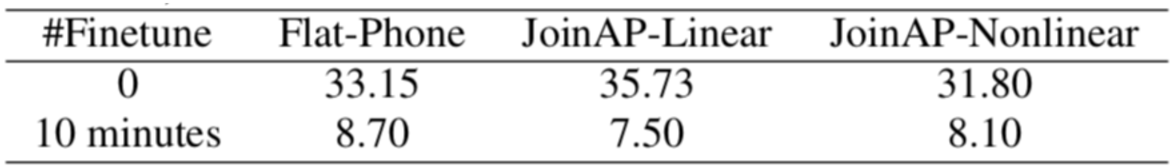


Language-degree Language	4	3	2	1
German	18	6	8	8
French	18	6	7	26
Spanish	18	4	1	7
Italian	18	5	4	6

On average, both JoinAP-Nonlinear and JoinAP-Linear perform better than Flat-Phone, and JoinAP-Nonlinear is the strongest.

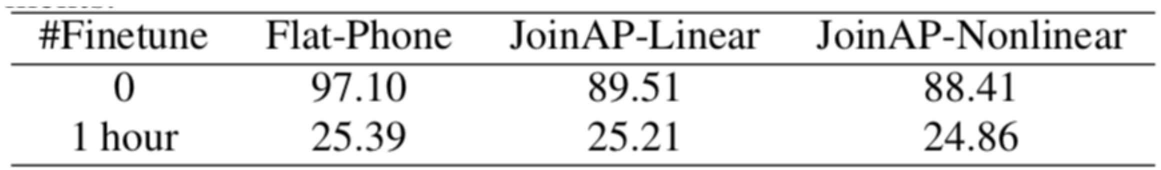
Experiments

- Crosslingual experiments
  - Polish:



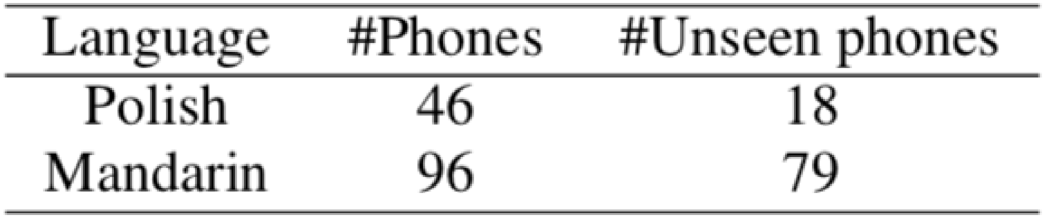
#Finetune	Flat-Phone	JoinAP-Linear	JoinAP-Nonlinear
0	33.15	35.73	31.80
10 minutes	8.70	7.50	8.10

  - Mandarin:



#Finetune	Flat-Phone	JoinAP-Linear	JoinAP-Nonlinear
0	97.10	89.51	88.41
1 hour	25.39	25.21	24.86

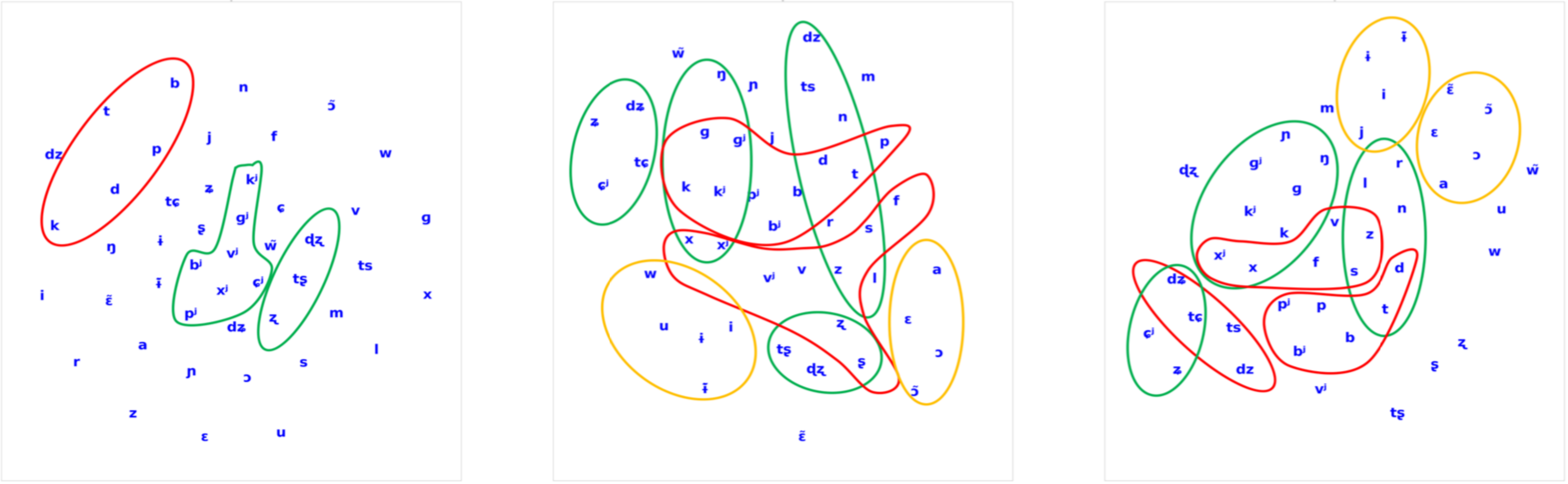
  - Statistics about Polish and Mandarin:



Language	#Phones	#Unseen phones
Polish	46	18
Mandarin	96	79

On average, both JoinAP-Nonlinear and JoinAP-Linear perform better than Flat-Phone, and JoinAP-Nonlinear is the strongest.

t-SNE map of Polish phone embeddings

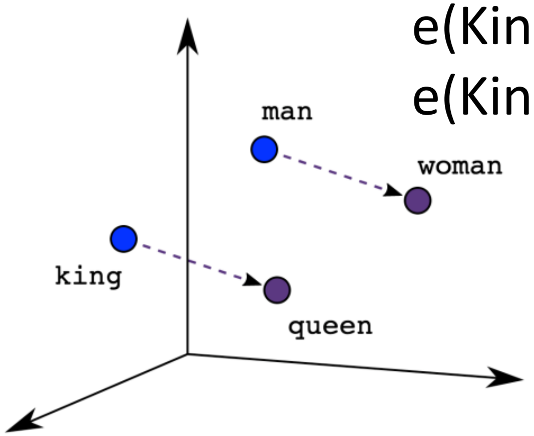


(a) Flat

(b) JoinAP-Linear

(c) JoinAP- Nonlinear

Consonants with the same manner of articulation; Consonants with the same place of articulation; Vowel with similar height



e(King) - e(queen) ≈ e(man) – e(woman)

e(King) ≈ e(man) - e(woman) + e(queen)

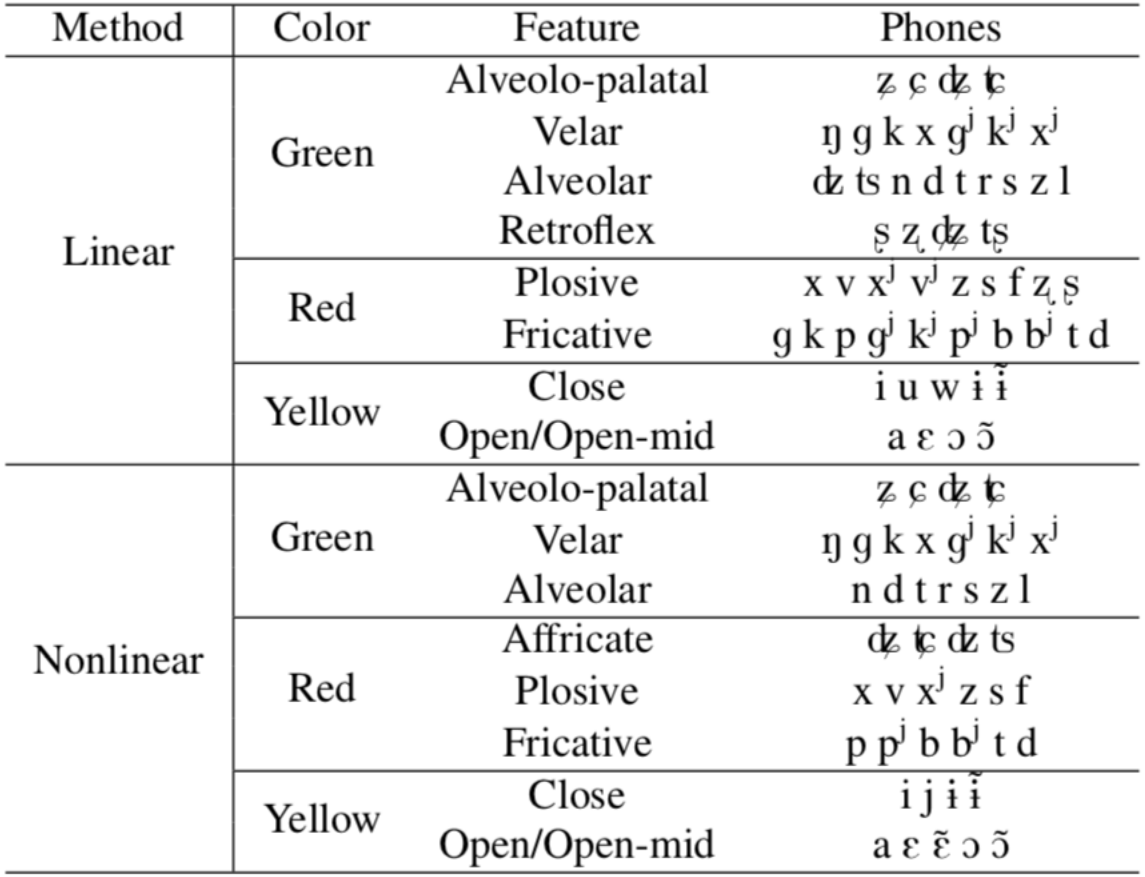
Unvoiced - Voiced: $e([k]) - e([g]) = e([p]) - e([b])$

Aspirated - Unaspirated: $e([p^h]) - e([p]) = e([k^h]) - e([k])$

Li, et al., "Hierarchical Phone Recognition with Compositional Phonetics", INTERSPEECH, 2021.

Experiments

- t-SNE map of Polish phone embeddings
  - Detailed explanation



Method	Color	Feature	Phones
Linear	Green	Alveolo-palatal	ʑ ɕ ʥ ʨ
		Velar	ŋ g k x gʲ kʲ xʲ
		Alveolar	ʣ ʦ n d t r s z l
		Retroflex	ʂ ʐ ʥ tʂ
	Red	Plosive	x v xʲ vʲ z s f ʐ ʂ
		Fricative	g k p gʲ kʲ pʲ b bʲ t d
	Yellow	Close	i u w ɨ ɨ̃
		Open/Open-mid	a ɛ ɔ ɔ̃
Nonlinear	Green	Alveolo-palatal	ʑ ɕ ʥ ʨ
		Velar	ŋ g k x gʲ kʲ xʲ
		Alveolar	n d t r s z l
	Red	Affricate	ʥ ʨ ʣ ʦ
		Plosive	x v xʲ z s f
		Fricative	p pʲ b bʲ t d
	Yellow	Close	i j ɨ ɨ̃
		Open/Open-mid	a ɛ ɛ̃ ɔ ɔ̃

Section Conclusion

- In the multilingual and crosslingual experiments, JoinAP-Nonlinear generally performs better than JoinAP-Linear and the traditional flat-phone method on average. The improvements for target language depend on its data amount and language-degree.

- Our JoinAP method provides a principled, data-efficent approach to multilingual and crosslingual speech recognition.

- Promising directions: exploring DNN based phonological transformation, and pretraining over increasing number of languages.

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. Conditional random fields and sequence discriminative training

---

II. Improving end-to-end speech recognition (15*3=45 min)

QA + break 15-min

1. Data-efficiency
2. Multilingual and crosslingual ASR
3. Language Model Integration for Transducer based Speech Recognition

III. Open questions and future directions + QA (15 min)

An Empirical Study of Language Model Integration for Transducer based Speech Recognition

Huahuan Zheng[1], Keyu An[1], Zhijian Ou[1],
Chen Huang[2], Ke Ding[2], Guanglu Wan[2]

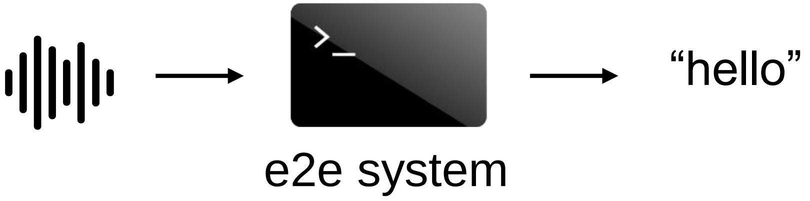
[1]Speech Processing and Machine Intelligence (SPMI) Lab,
Tsinghua University, Beijing, China.
[2]Meituan, China.

Presented at INTERSPEECH 2022

Content

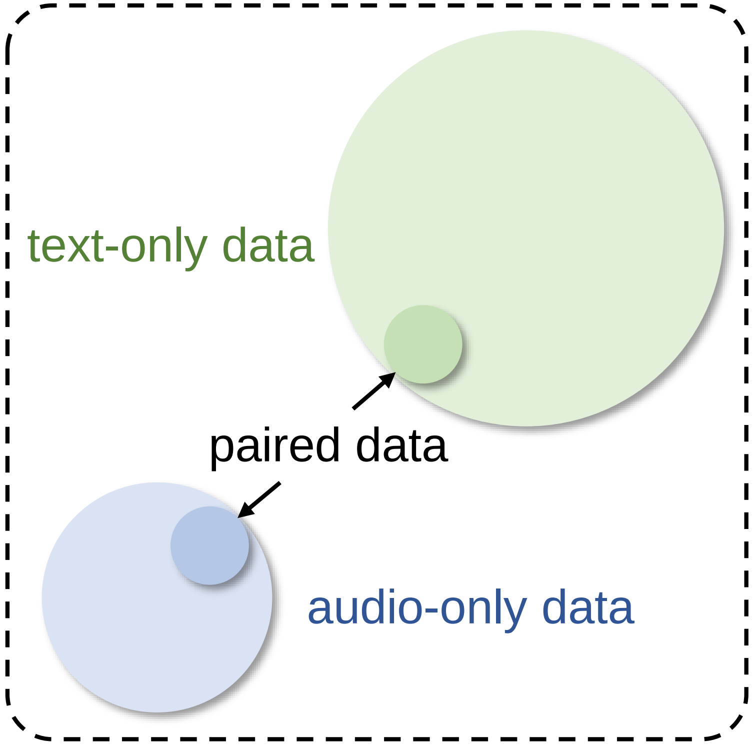
1. **Introduction**
2. Low Order Density Ratio (LODR)
3. Experiments
4. Conclusions

“Data efficiency” in speech recognition:
towards utilizing the text-only data



- End-to-end (e2e) speech recognition is “data hungry”, whose performance relies on the amount of paired speech-text data.
- Text-only & audio-only data are more easily available, compared to paired ones (a.k.a. the labeled data).

How to utilize the text?
Language Model (LM) integration!



Amount of available data.

[1] Li, Jinyu. "Recent Advances in End-to-End Automatic Speech Recognition." arXiv preprint arXiv:2111.01690 (2021).

LM integration in Transducer:
some intuition and heuristic experience

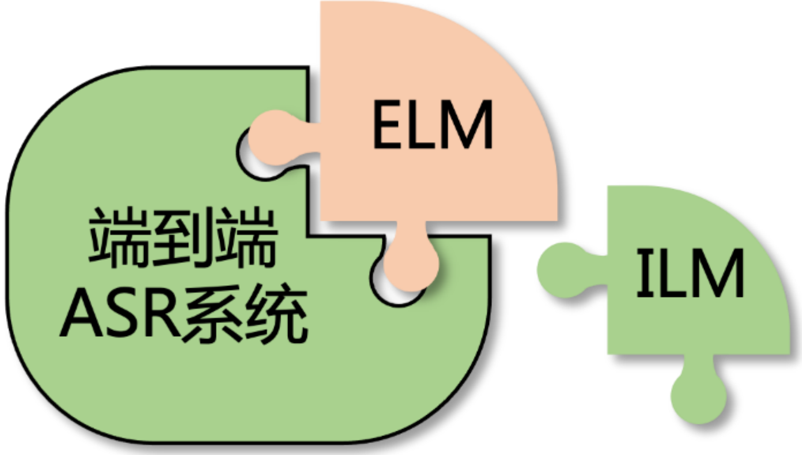
X: speech data, Y: corresponding label sequence.

Hybrid model (e.g., DNN-HMM):

$$\hat{Y} = \arg\max_{Y} \left[P_{\mathrm{AM}}(X|Y) P_{\mathrm{ELM}}(Y) \right]$$

E2E model (e.g., RNNT, AED):

$$\hat{Y} = \arg\max_{Y} \left[\frac{P_{\mathrm{RNN\text{-}T}}(Y|X)}{P_{\mathrm{ILM}}(Y)} P_{\mathrm{ELM}}(Y) \right]$$



[1] A. Graves, “Sequence transduction with recurrent neural networks,” arXiv preprint arXiv:1211.3711, 2012.
[2] Z. Meng, and et al., “Internal language model estimation for domain-adaptive end-to-end speech recognition,” SLT 2021.

Related work:
shallow fusion, density ratio and ILME

1. shallow fusion (SF):

$$Y^* = \arg\max_Y \left(\log P_{\mathrm{RNNT}}(Y|X) + \lambda_1 \log P_{\mathrm{ELM}}(Y) + \beta|Y|\right)$$

2.1 density ratio (DR):

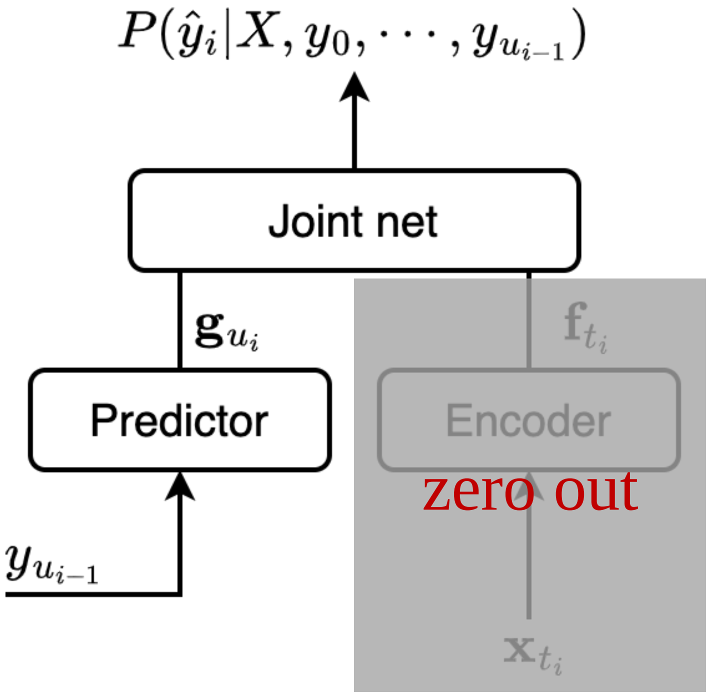
$$Y^* = \arg\max_Y \left(\log P_{\mathrm{RNNT}}(Y|X) + \boxed{\lambda_0 \log P_{\mathrm{ILM}}(Y)} + \lambda_1 \log P_{\mathrm{ELM}}(Y) + \beta|Y|\right)$$

ILM is approximated via a separate NN LM trained with the same linguistic information as RNN-T (transcript of the audio data).

2.2 ILME (Internal Language Model Estimation):

linear approximation $J(\mathbf{g}_u, \mathbf{f}_t) \approx J(\mathbf{g}_u, \mathbf{0}) + J(\mathbf{0}, \mathbf{f}_t)$

$$\longrightarrow \quad P_{\mathrm{ILM}}(y_{u+1}|y_{0:u}) \propto \exp\left(J(\mathbf{g}_u, \mathbf{0})\right)$$



[1] E.McDermott, and et al., "A density ratio approach to language model fusion in end-to-end automatic speech recognition," ASRU 2019.
[2] Z. Meng, and et al., "Internal language model estimation for domain-adaptive end-to-end speech recognition," SLT 2021.

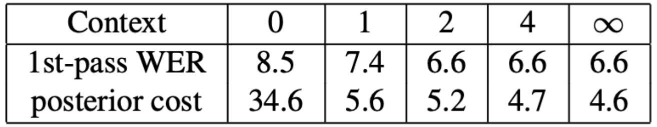
Content

1. Introduction
2. **Low Order Density Ratio (LODR)**
3. Experiments
4. Conclusions

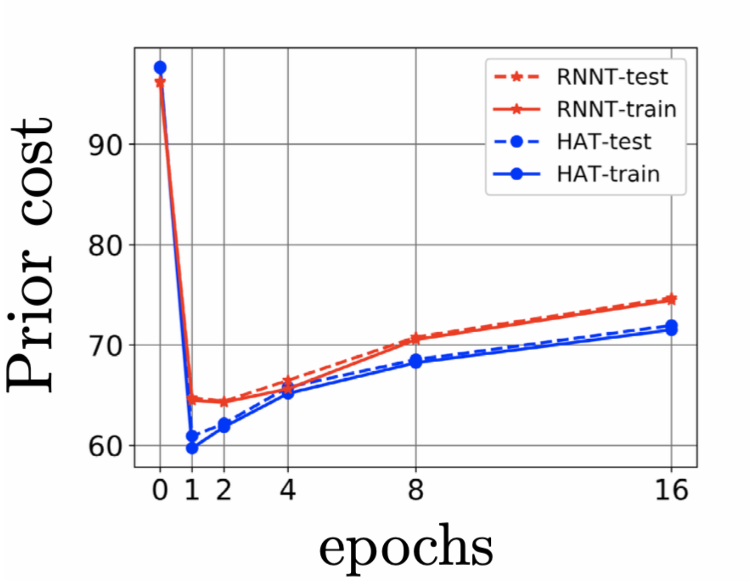
A brief summary of observation about the Predictor.

1. The Predictor is commonly very shallow neural network. (e.g. 1x LSTM);
2. The Predictor only makes use of limited context (Table 1);
3. The ILM estimated from Predictor performs poorly when evaluated as normal LM.

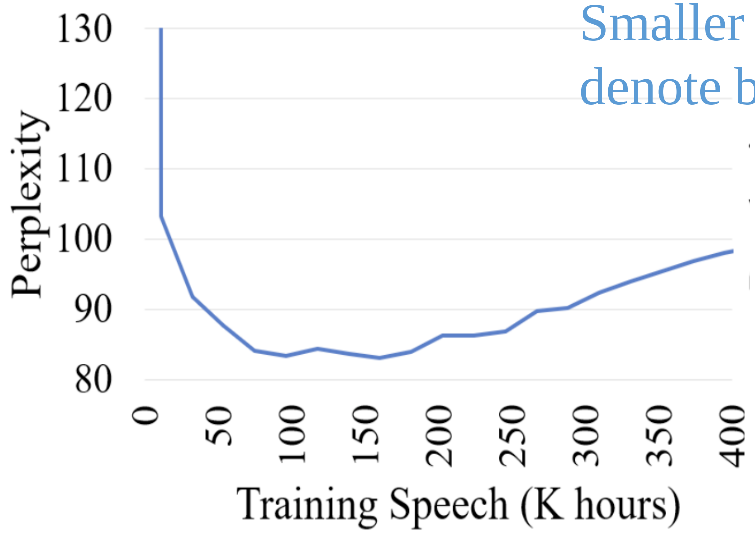
Table 1. Effect of limited context history [1].



Context	0	1	2	4	∞
1st-pass WER	8.5	7.4	6.6	6.6	6.6
posterior cost	34.6	5.6	5.2	4.7	4.6



(a) Prior cost of estimated ILM from HAT [1];
The “prior cost” measures the $-\log P(Y)$.

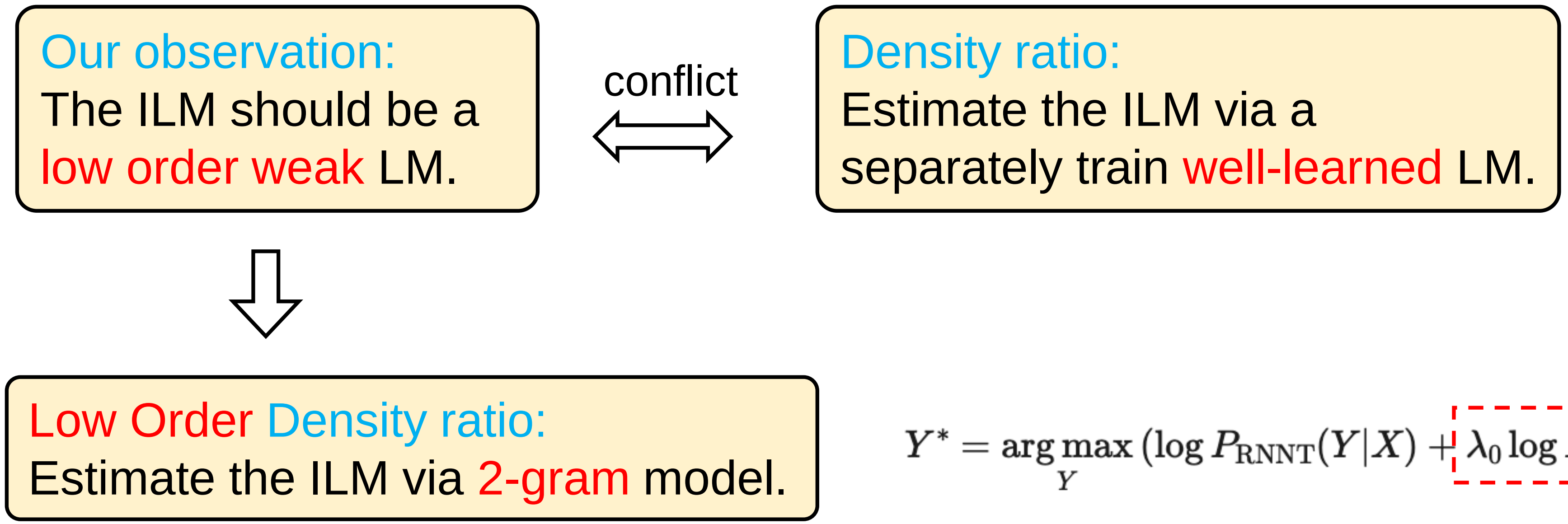


Smaller prior cost and perplexity denote better LM performance.

(b) Perplexity of estimated ILM from ILME [2].
A “normal” LM trained on the transcript has a perplexity of 30.1

[1] E. Variani, and et al, “Hybrid autoregressive transducer (HAT),” in ICASSP 2020.
[2] Z. Meng, and et al., “Internal language model estimation for domain-adaptive end-to-end speech recognition,” in SLT 2021.

Low Order Density Ratio (LODR)



Our observation:
The ILM should be a low order weak LM.

conflict

Density ratio:
Estimate the ILM via a separately train well-learned LM.

Low Order Density ratio:
Estimate the ILM via 2-gram model.

$$Y^* = \arg\max_{Y} \left(\log P_{\mathrm{RNNT}}(Y|X) + \boxed{\lambda_0 \log P_{\mathrm{ILM}}(Y)} + \lambda_1 \log P_{\mathrm{ELM}}(Y) + \beta|Y|\right)$$

Low order LM

In practice, we obtain the ILM as follows:

1. Prepare the training corpus: we use the transcript only;
2. Train a 2-gram LM on the corpus using KenLM with some prunes if required*.

* The size of context could be different according to the granularity of the modeling units.

[1] https://github.com/kpu/kenlm

Content

1. Introduction
2. Low Order Density Ratio (LODR)
3. **Experiments**
4. Conclusions

Experiments: **in-domain** evaluation with large amount of text corpus

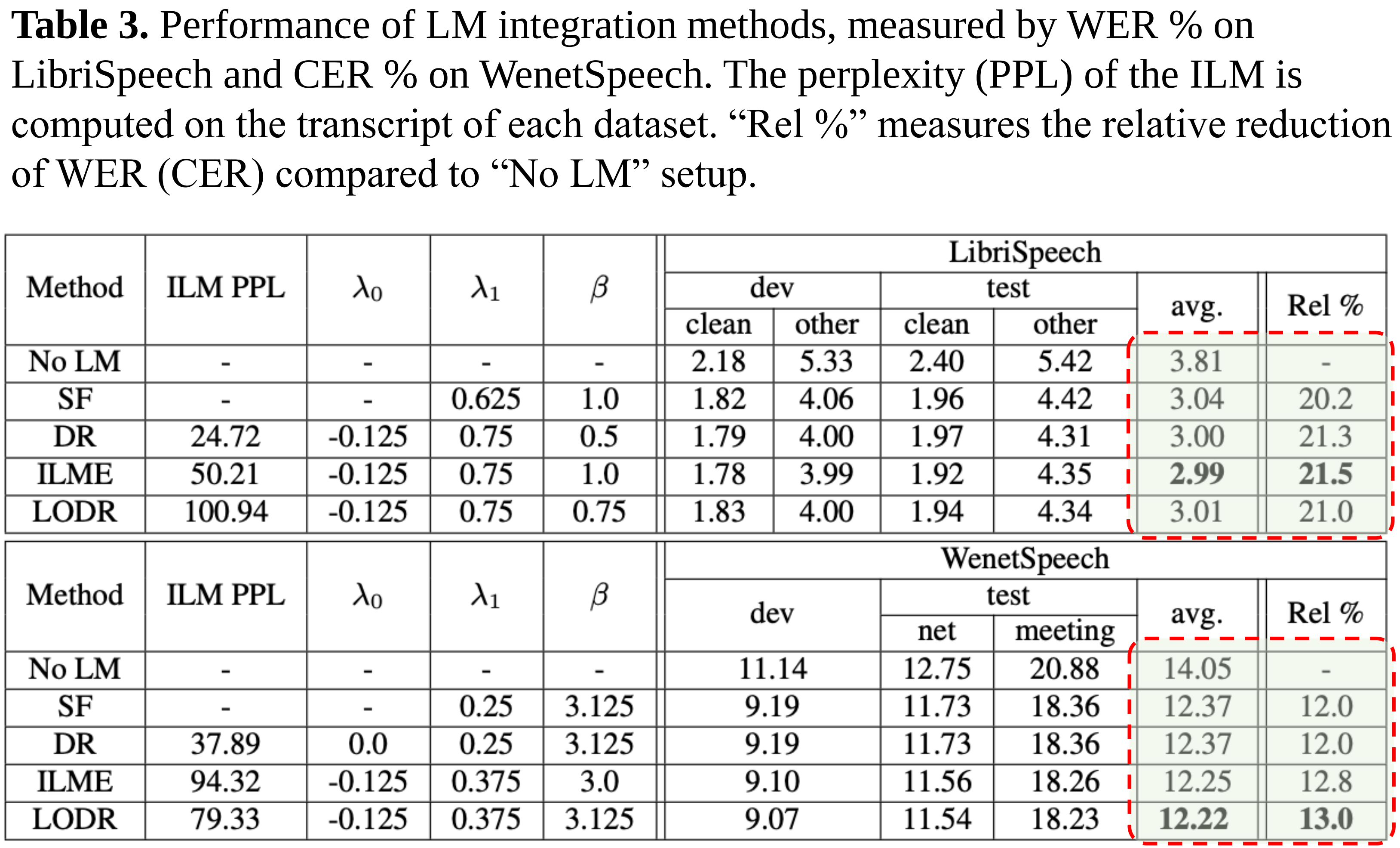


Table 3. Performance of LM integration methods, measured by WER % on LibriSpeech and CER % on WenetSpeech. The perplexity (PPL) of the ILM is computed on the transcript of each dataset. "Rel %" measures the relative reduction of WER (CER) compared to "No LM" setup.

Method	ILM PPL	λ_0	λ_1	β	LibriSpeech					
					dev		test		avg.	Rel %
					clean	other	clean	other		
No LM	-	-	-	-	2.18	5.33	2.40	5.42	3.81	-
SF	-	-	0.625	1.0	1.82	4.06	1.96	4.42	3.04	20.2
DR	24.72	-0.125	0.75	0.5	1.79	4.00	1.97	4.31	3.00	21.3
ILME	50.21	-0.125	0.75	1.0	1.78	3.99	1.92	4.35	**2.99**	**21.5**
LODR	100.94	-0.125	0.75	0.75	1.83	4.00	1.94	4.34	3.01	21.0

Method	ILM PPL	λ_0	λ_1	β	WenetSpeech				
					dev	test		avg.	Rel %
						net	meeting		
No LM	-	-	-	-	11.14	12.75	20.88	14.05	-
SF	-	-	0.25	3.125	9.19	11.73	18.36	12.37	12.0
DR	37.89	0.0	0.25	3.125	9.19	11.73	18.36	12.37	12.0
ILME	94.32	-0.125	0.375	3.0	9.10	11.56	18.26	12.25	12.8
LODR	79.33	-0.125	0.375	3.125	9.07	11.54	18.23	**12.22**	**13.0**

Size of extra corpus:

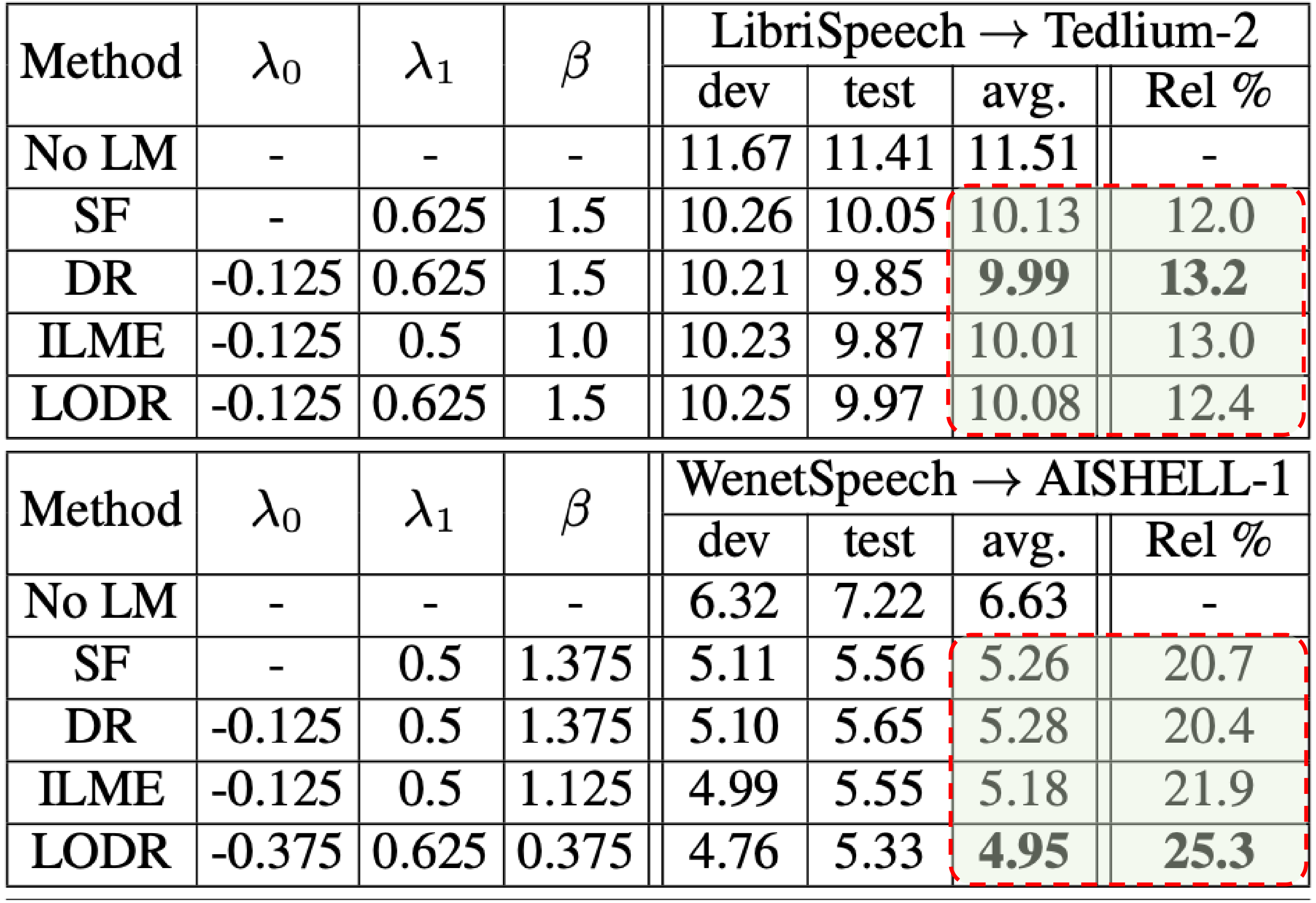
English: 800 million words (9.4M words in transcript)

Chinese: 200 million chars (17M chars in transcript)

All methods subtracting ILM perform better than the shallow fusion consistently.

Experiments: **cross-domain** evaluation and discussion

Table 4. Performance of LM integration methods evaluated on cross-domain scenarios.



Method	λ_0	λ_1	β	LibriSpeech → Tedlium-2			
				dev	test	avg.	Rel %
No LM	-	-	-	11.67	11.41	11.51	-
SF	-	0.625	1.5	10.26	10.05	10.13	12.0
DR	-0.125	0.625	1.5	10.21	9.85	**9.99**	**13.2**
ILME	-0.125	0.5	1.0	10.23	9.87	10.01	13.0
LODR	-0.125	0.625	1.5	10.25	9.97	10.08	12.4

Method	λ_0	λ_1	β	WenetSpeech → AISHELL-1			
				dev	test	avg.	Rel %
No LM	-	-	-	6.32	7.22	6.63	-
SF	-	0.5	1.375	5.11	5.56	5.26	20.7
DR	-0.125	0.5	1.375	5.10	5.65	5.28	20.4
ILME	-0.125	0.5	1.125	4.99	5.55	5.18	21.9
LODR	-0.375	0.625	0.375	4.76	5.33	**4.95**	**25.3**

Size of extra corpus:
English (Tedlium-2):
2.2M words (9.4M words in transcript)

Chinese (AISHELL-1):
1.7M chars (17M chars in transcript)

Librispeech (960 hours). Streaming encoder + stateless Transducer.

Decoding method	λ_1	λ_2	test-clean	WERR	test-other	WERR
Modified beam search	-	-	2.73	-	7.15	-
+ SF	0.3	-	2.42	11.4%	6.46	9.7%
+ ILME	0.3	-0.05	2.36	13.6%	6.23	12.9%
+ LODR (bi-gram)	0.3	-0.16	**2.28**	**16.5%**	**5.94**	**16.9%**

Librispeech + Gigaspeech (10k hours). Non-streaming encoder + pruned & stateless Transducer.

Decoding method	λ_1	λ_2	test-clean	WERR	test-other	WERR
Modified beam search	-	-	2.00	-	4.63	-
+ SF	0.3	-	1.96	2.0%	4.18	9.7%
+ ILME	0.3	-0.05	**1.82**	**9.0%**	4.10	11.4%
+ LODR (bi-gram)	0.4	-0.14	1.83	8.5%	**4.03**	**13.0%**

在K2实验中，LODR表现优秀！

*Results are reported on icefall, a repo maintained by the K2 team.

[1] https://github.com/k2-fsa/k2

[2] https://github.com/k2-fsa/icefall

Conclusions

1. We propose a LODR method, which uses low order and weak LM as the estimated ILM for the original DR method, with the observation that the ILM of Transducer indeed only captures limited linguistic information.

2. The LODR method is evaluated on both in-domain and cross-domain scenarios, and compared with existing methods.

- Our proposed LODR consistently outperforms the SF, and performs better than the original DR in most tests with less extra parameters introduced.

- As compared to ILME, our LODR method has close performance but avoids feeding the labels to the text encoder twice.

Content

I. Basics for end-to-end speech recognition (15*3=45 min)

1. Probabilistic graphical modeling (PGM) framework
2. Hybrid DNN-HMM and Connectionist Temporal Classification (CTC)
3. Conditional random fields and sequence discriminative training

QA + break 15-min

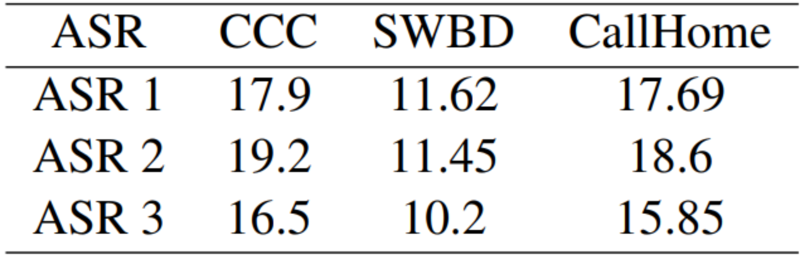
II. Improving end-to-end speech recognition (15*3=45 min)

1. Data-efficiency
2. Multilingual and crosslingual ASR
3. Language Model Integration for Transducer based Speech Recognition

→ III. Open questions and future directions + QA (15 min)

“WER we are and WER we think we are”

> “The conclusions are clear: we are definitely not where we think we are in terms of WERs (Word Error Rates).”



ASR	CCC	SWBD	CallHome
ASR 1	17.9	11.62	17.69
ASR 2	19.2	11.45	18.6
ASR 3	16.5	10.2	15.85

Table 1: WER [%] comparison on benchmarks

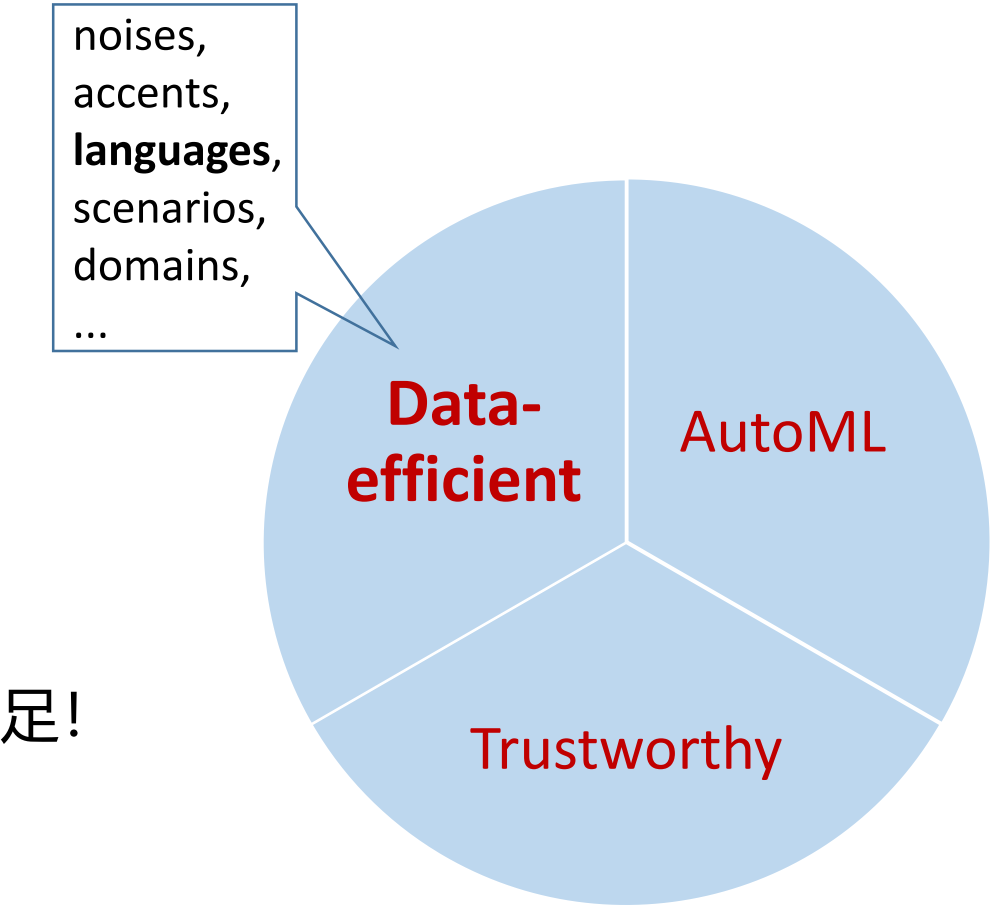
- Test: three different state-of-the-art commercial ASR solutions
- Call Center Conversations (CCC)
- The commercial ASR systems in our evaluation achieve nearly double the error rates (reported in the literatures) on both HUB’05 evaluation subsets.

Szymański, et al., "WER we are and WER we think we are", EMNLP 2020.

Summary

新一代语音识别技术的若干特点

✓ **Data-efficient**, AutoML, Trustworthy



noises,
accents,
languages,
scenarios,
domains,
…

数据高效的多语言与跨语言语音识别

▶ CTC-CRF：支持分立的AM与LM

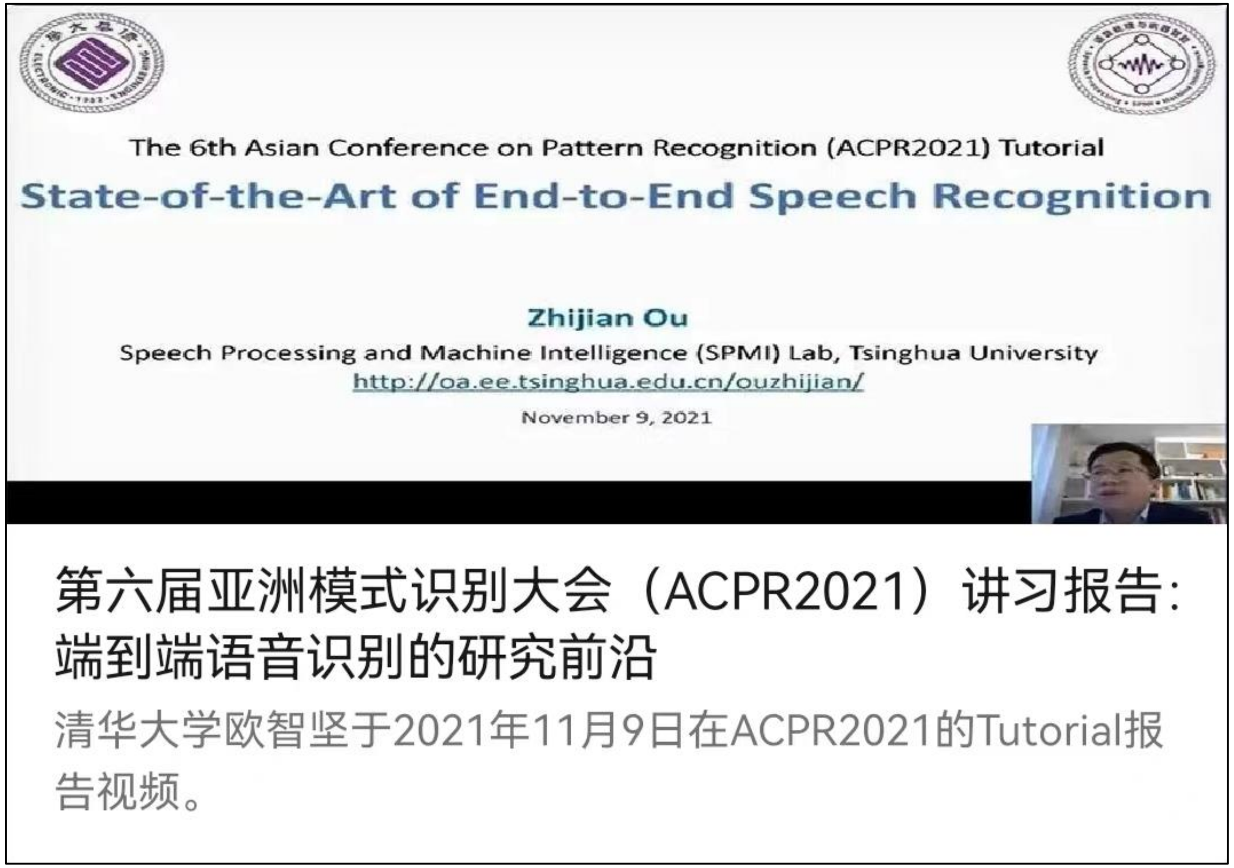
- 在原理上克服了历史上各类序列鉴别模型的不足！
- 减少对大量人工标注语音数据的依赖

▶ JointAP：联合声学与音韵学

- 促进多语言训练时信息共享以及跨语言语音识别时信息迁移

▶ LODR：一种更好、更轻量的语言模型融合新方式

- 如何更好利用纯文本数据，是数据高效ASR的重要特征



The 6th Asian Conference on Pattern Recognition (ACPR2021) Tutorial

State-of-the-Art of End-to-End Speech Recognition

Zhijian Ou

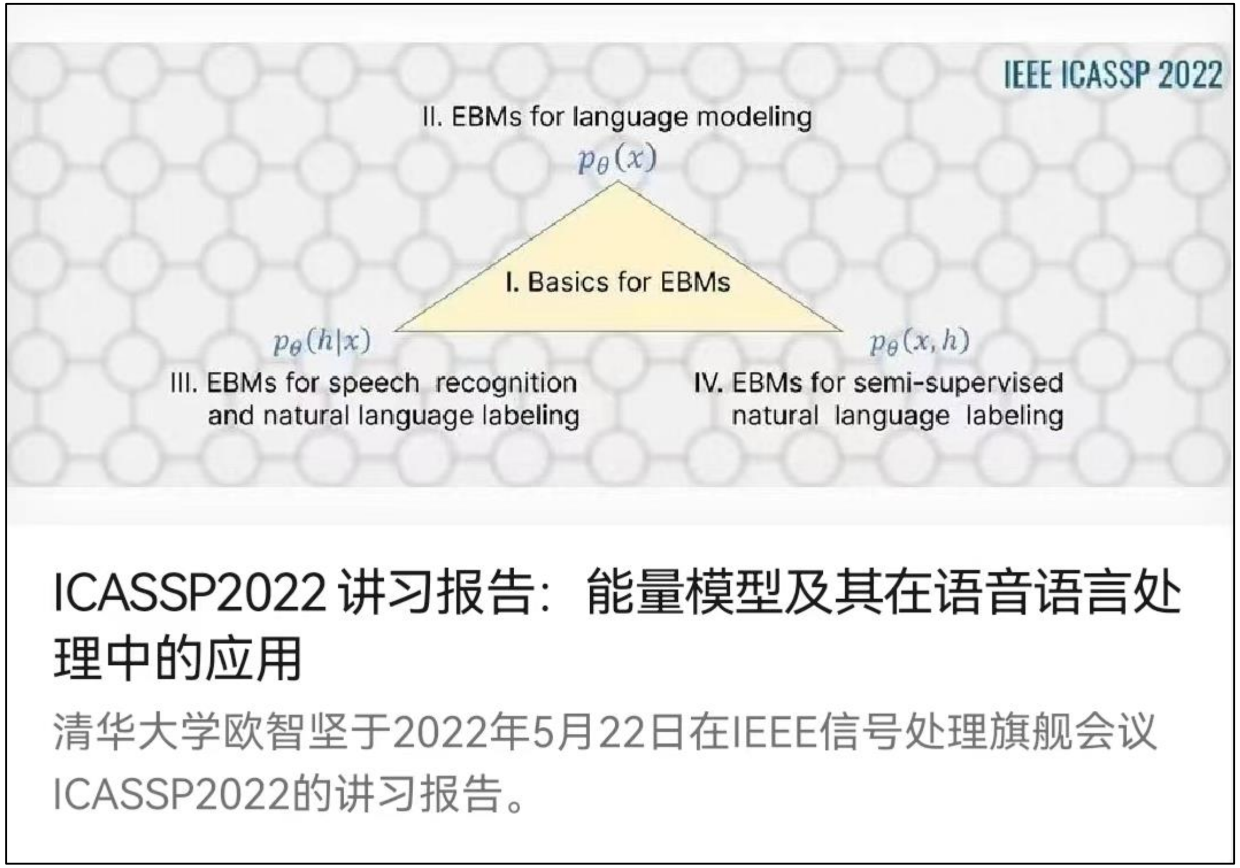
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University

http://oa.ee.tsinghua.edu.cn/ouzhijian/

November 9, 2021

第六届亚洲模式识别大会（ACPR2021）讲习报告：端到端语音识别的研究前沿

清华大学欧智坚于2021年11月9日在ACPR2021的Tutorial报告视频。



IEEE ICASSP 2022

II. EBMs for language modeling

$p_\theta(x)$

I. Basics for EBMs

$p_\theta(h|x)$

III. EBMs for speech recognition and natural language labeling

$p_\theta(x,h)$

IV. EBMs for semi-supervised natural language labeling

ICASSP2022 讲习报告：能量模型及其在语音语言处理中的应用

清华大学欧智坚于2022年5月22日在IEEE信号处理旗舰会议ICASSP2022的讲习报告。

Slides and video: http://oa.ee.tsinghua.edu.cn/~ouzhijian/news.htm

感谢聆听！