Data-efficient Automatic Speech Recognition

Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab

Tsinghua University

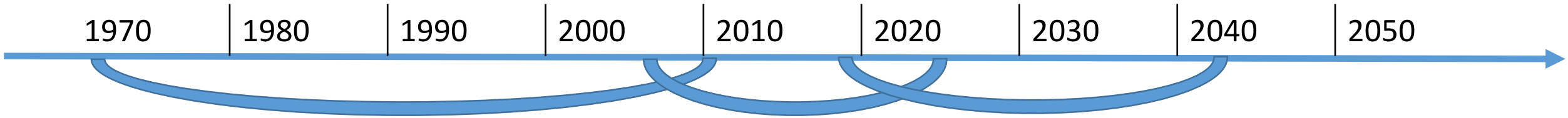
http://oa.ee.tsinghua.edu.cn/ouzhijian/

2021/7/9

Content

1. Motivation
2. Related work
3. Method : CTC-CRF
   - ✓ ICASSP2019, Interspeech2020
   - ✓ Children Speech Recognition (SLT2021)
   - ✓ Wordpieces and Conformers (submitted ASRU2021)
   - ✓ Multilingual and Crosslingual ASR (submitted ASRU2021)
4. Experiments
5. Conclusion

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
CTC
Attention seq2seq
RNN Transducer
Transformer

Data-efficient
AutoML
Trustworthy AI

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

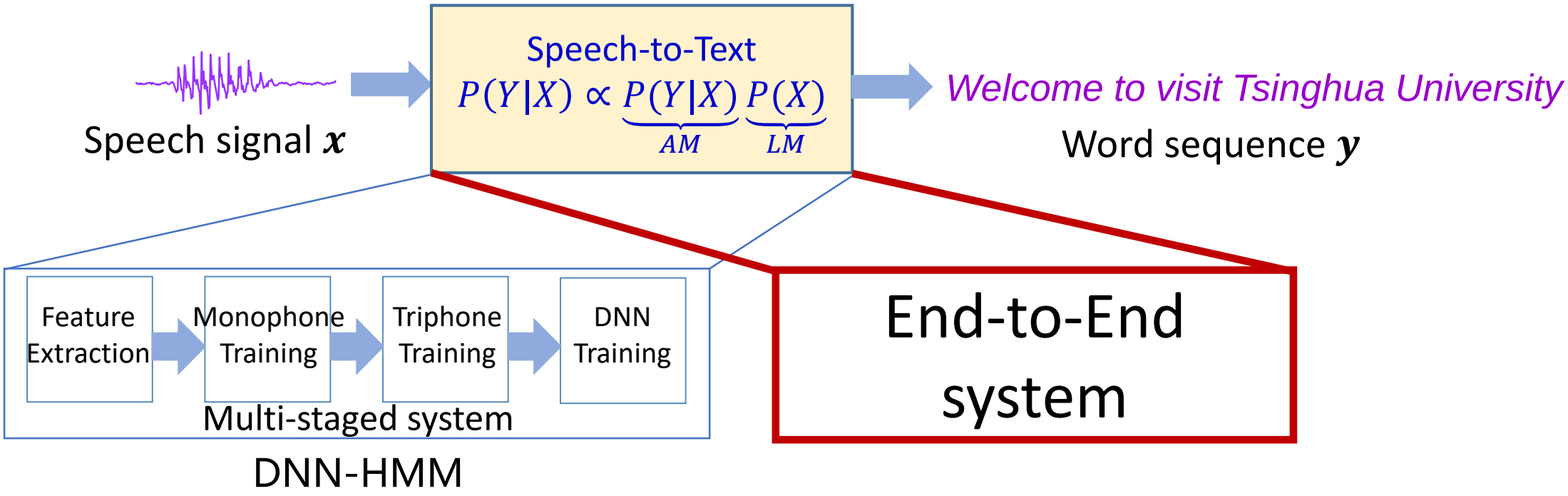
Data-efficient $Efficiency = \frac{Performance}{Cost}$

- Current ASR: heavy reliance on supervised learning and large amounts of manually-labeled data
- Different from: computation-efficient (MIPS, million instructions per second), power-efficient (MIPS/Watt)

  —— Efficiency of learning by machines
- A spectrum of data-efficient modeling and learning methods
  - ✓ Model architecture
  - ✓ unsupervised, semi-supervised, self-supervised learning
  - ✓ Pre-training
  - ✓ Transfer learning
  - ✓ Active learning
  - ✓ Meta-leaning

Related work

- ASR is a discriminative problem
  - For acoustic observations $\boldsymbol{x} \triangleq x_1, \cdots, x_T$, find the most likely labels $\boldsymbol{y} \triangleq y_1, \cdots, y_L$
- ASR state-of-the-art: DNNs of various network architectures (Hinton NIPSw2009, Microsoft IS2011)



Related work

- End-to-end system:
  - Eliminate GMM-HMM pre-training and triphone tree building, and can be trained from scratch (flat-start or single-stage).
- In a more strict sense:
  - Remove the need for a pronunciation lexicon and, even further, train the acoustic and language models jointly rather than separately
  - Data-hungry

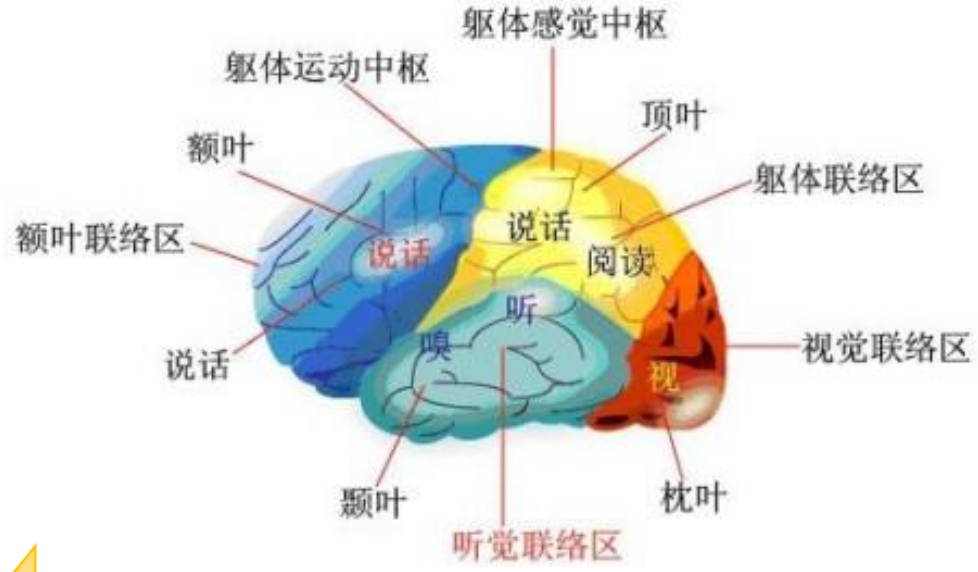
We advocate data-efficient end2end speech recognition, which uses a separate language model (LM) with or without a pronunciation lexicon.

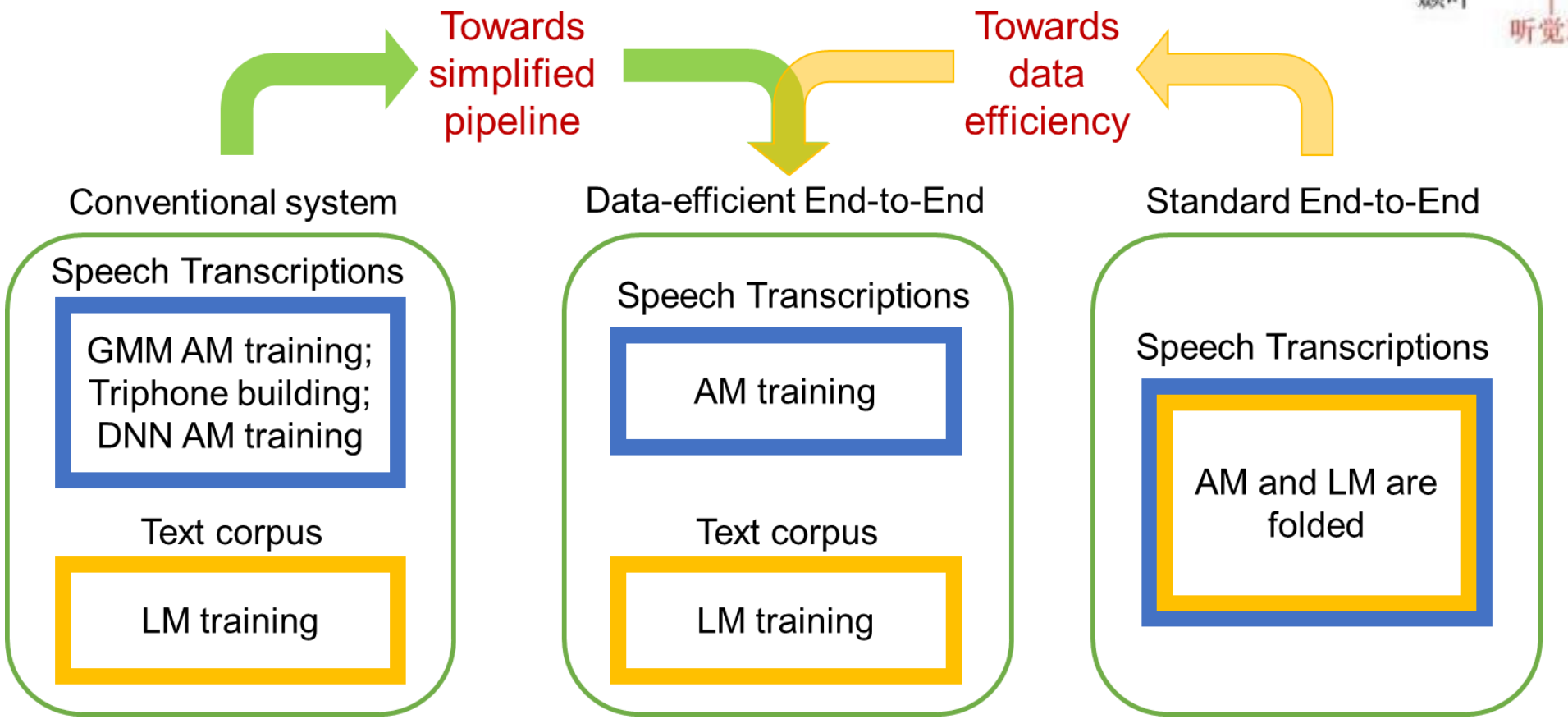
- Text corpus for language modeling are cheaply available.
- Data-efficient

Motivation

Modularization promote Data-efficiency

- ✓ Not pursue unduly end-to-end
- ✓ Keep necessary factorization of AM and LM





Related work

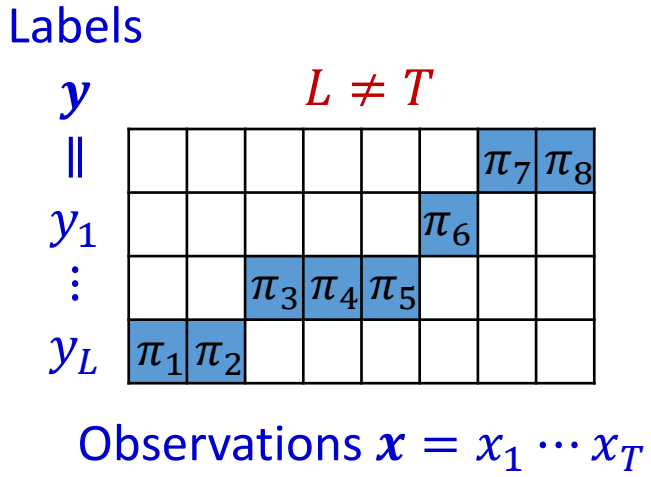
ASR is a discriminative problem

- 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 $L \neq T$

- Explicitly by state sequence $\boldsymbol{\pi} \triangleq \pi_1, \cdots, \pi_T$ in HMM, CTC, RNN-T, or implicitly in Seq2Seq



Related work How to handle alignment, since $L \neq T$

- Explicitly by state sequence $\boldsymbol{\pi} \triangleq \pi_1, \cdots, \pi_T$ in HMM, CTC, RNN-T, or implicitly in Seq2Seq
- State topology : determines a mapping $\mathcal{B}$, which map $\boldsymbol{\pi}$ to a unique $\boldsymbol{l}$

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\boldsymbol{l})} p(\boldsymbol{\pi}|\boldsymbol{x})$$

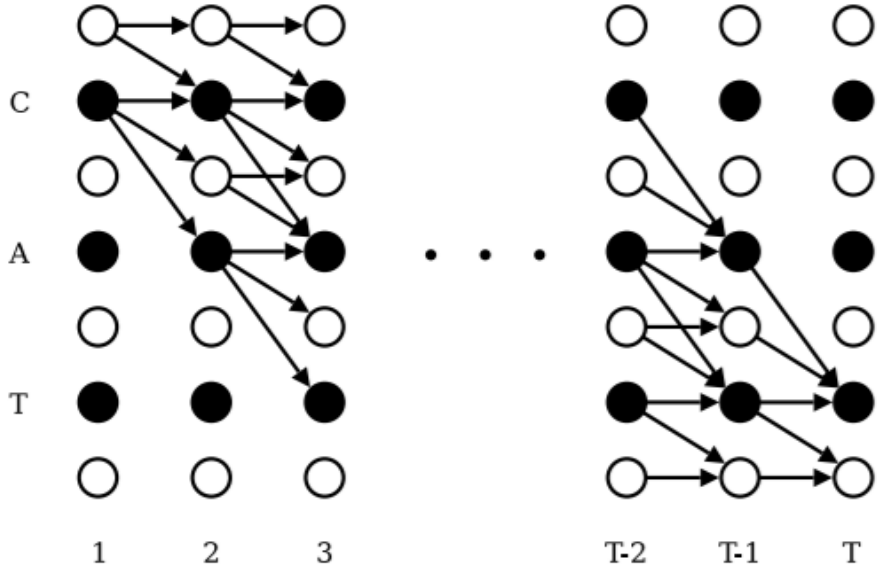
CTC topology : a mapping $\mathcal{B}$ maps $\boldsymbol{\pi}$ to $\boldsymbol{l}$ by

1. removing all repetitive symbols between the blank symbols.
2. removing all blank symbols.

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

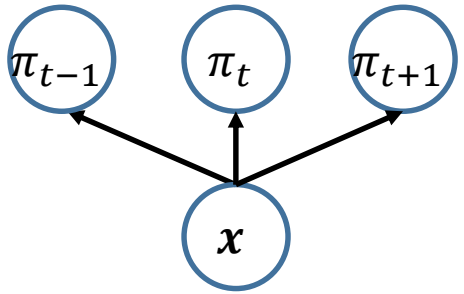
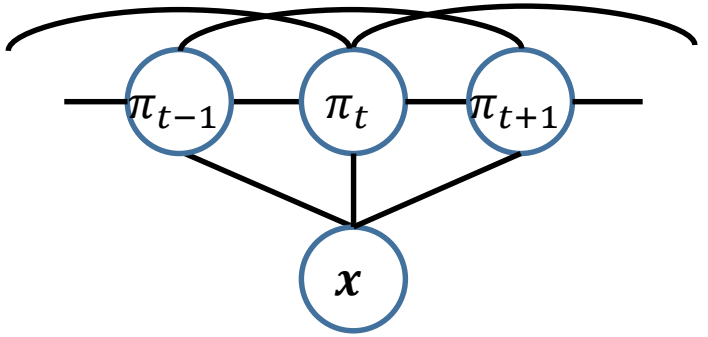
☺ Admit the smallest number of units in state inventory, by adding only one <blk> to label inventory.

☺ Avoid ad-hoc silence insertions in estimating denominator LM of labels.



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

CTC vs CTC-CRF

CTC	CTC-CRF
$p(\boldsymbol{l}\|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}\in\mathcal{B}^{-1}(\boldsymbol{l})} p(\boldsymbol{\pi}\|\boldsymbol{x})$, 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
$\frac{\partial \log p(\boldsymbol{l}\|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{\pi}\|\boldsymbol{l},\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial \log p(\boldsymbol{\pi}\|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$	$\frac{\partial \log p(\boldsymbol{l}\|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{\pi}\|\boldsymbol{l},\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial \phi(\boldsymbol{\pi},\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] - \mathbb{E}_{p(\boldsymbol{\pi}'\|\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial \phi(\boldsymbol{\pi}',\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$
	

Related work How to obtain $p(\boldsymbol{y} \mid \boldsymbol{x})$

GMM-HMM
(IBM, AT&T, 1980s)

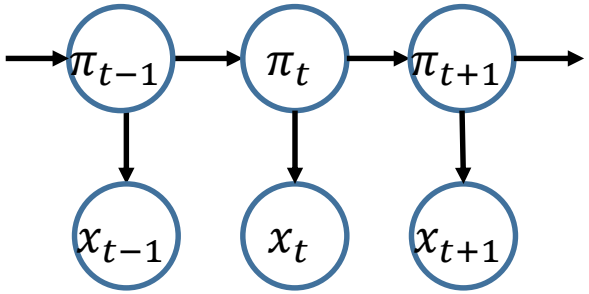
DNN-HMM
(Hinton NIPSw09; Microsoft IS11)

CTC
(Graves ICML06; Google IC15)

Attention Seq2Seq
(Bengio NIPS15;
Google IC16)

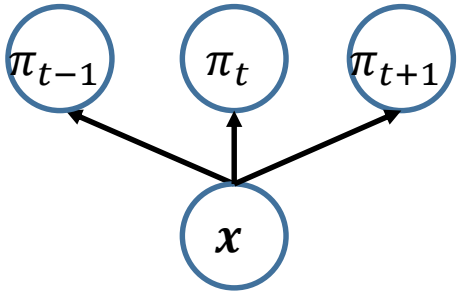
RNN-T
(Graves ICMLw12;
Baidu, Google ASRU17)

CTC-CRF
(Xiang&Ou IC19)



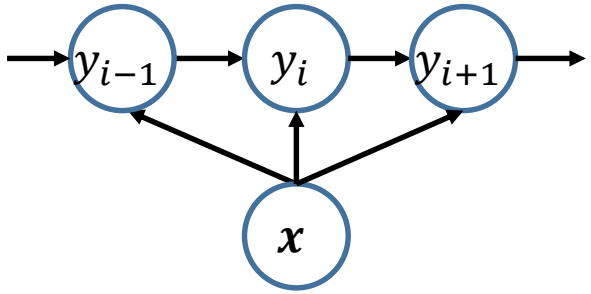
DNN-HMM

☹ Multi-staged



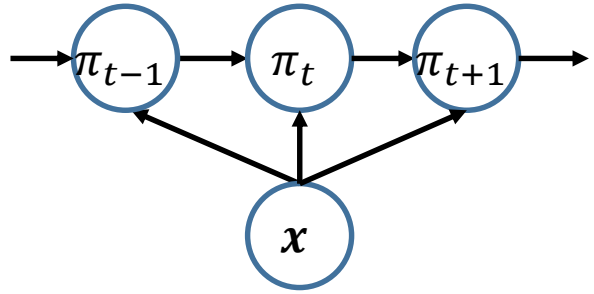
CTC

☹ $\{\pi_t\}$ Conditional Independence



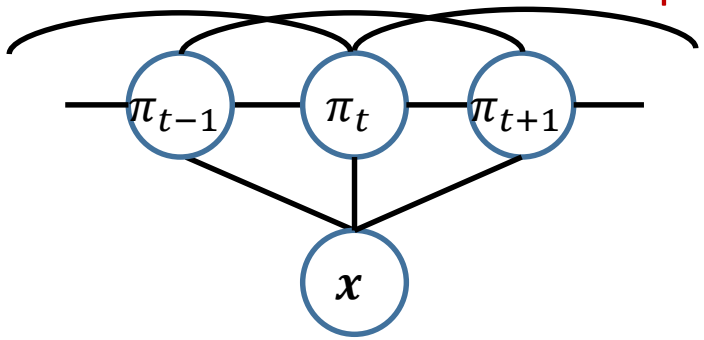
Attention

☹ $\{y_i\}$ Directed sequence modeling



RNN-T

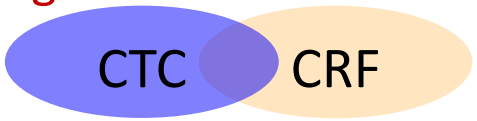
☹ $\{\pi_t\}$ Directed sequence modeling



CTC-CRF

CTC-CRF occupies a unique position!

For the first time, successfully combine neural network and undirected graph model for speech recognition, and principally overcome the shortcomings of various models in history!



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):

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

Experiments (Comparison with CTC, phone based)

WSJ 80h

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

44.4%

Switchboard 300h

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

14.7% 11%

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%	13.68%
CTC-CRF	Mono-phone	4-gram	N	3.87%	10.28%	4.09%	10.65%

19.1% 22.1%

SP: speed perturbation for 3-fold data augmentation.

- 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.

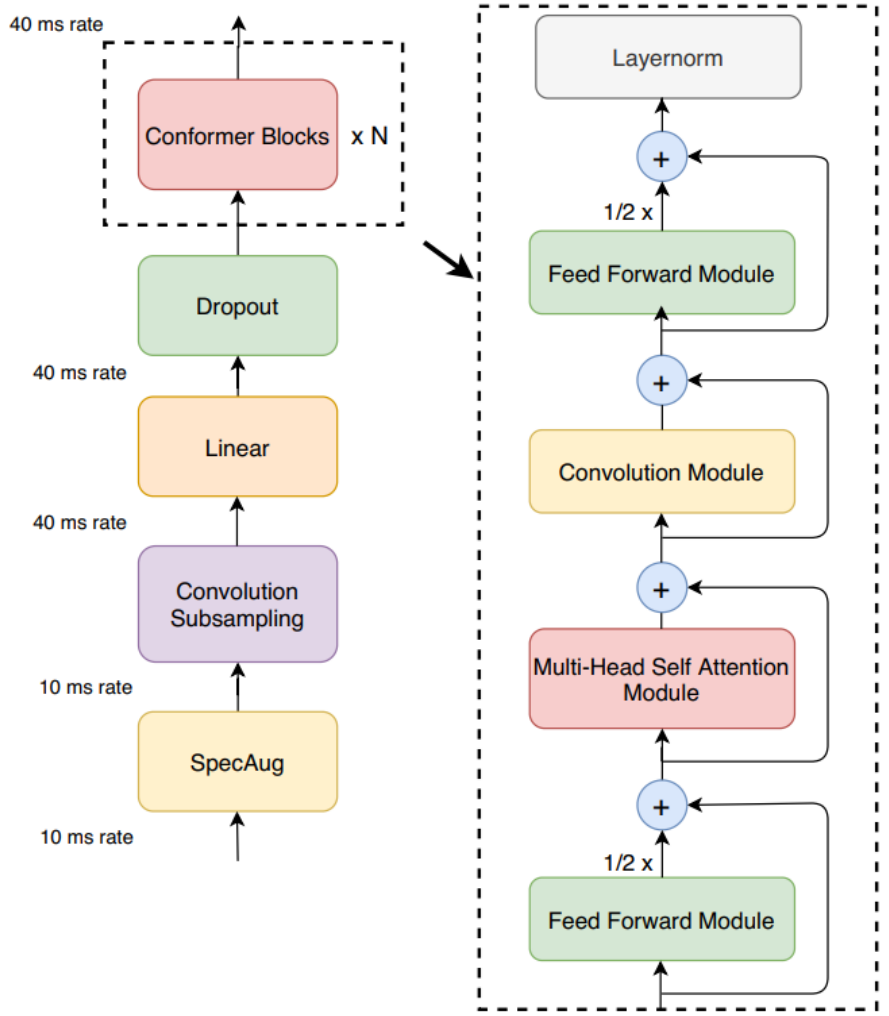
model	Chain model	Transformer	CTC-CRF
CER%	28.75	27.28	**25.34**

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

Basic Unit	Segmented Sequence
word	that neither of them had crossed the threshold since the dark day
character	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	that_ ne i ther_ of_ them_ had_ cro s sed_ the_ th re sh old_ sin ce_ the_ d ar k_ 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

Multilingual and Crosslingual Speech Recognition using Phonological-Vector based Phone Embeddings

Chengrui Zhu, Keyu An, Huahuan Zheng, **Zhijian Ou**

Content

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

Motivation

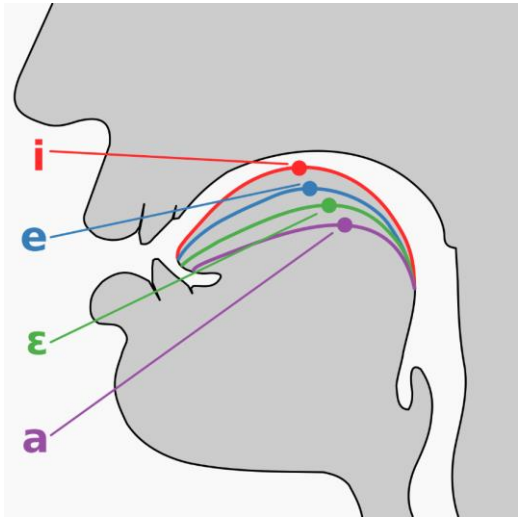
- There are more than 7100 languages in the world, and most of them are low-resourced languages.
- Multilingual speech recognition
  - Training data from a number of languages (seen languages) are merged to train a multilingual AM.
- Crosslingual speech recognition
  - The target language is unseen in 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 directely used without finetuning*.

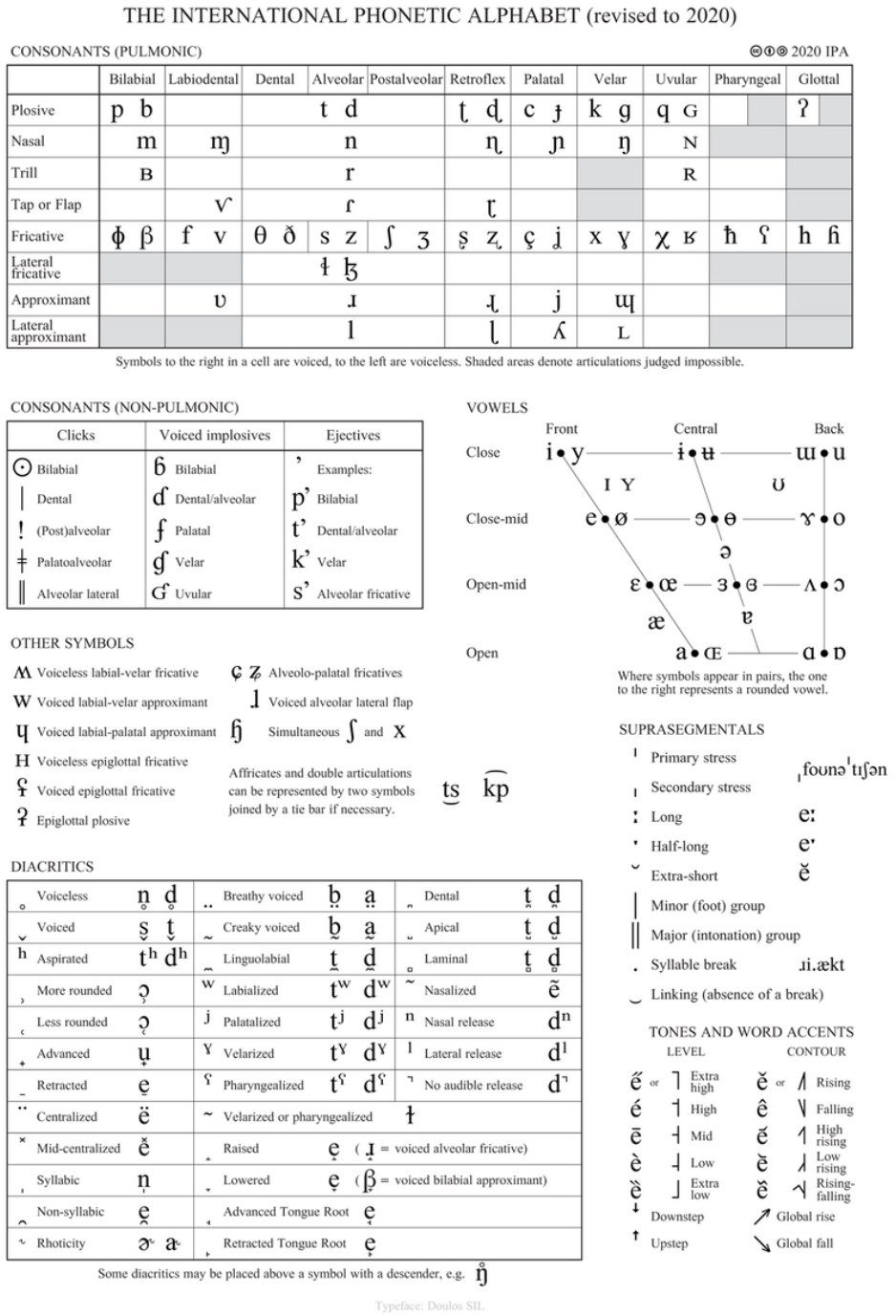
* Suppose that text corpus from the target language are available.

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

- International Phonetic Alphabet (IPA)
  - Classify phones by phonological features
  - Vowels
    - vowel height
    - vowel backness
  - Consonants
    - Place of articulation
    - Manner of articulation





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.

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

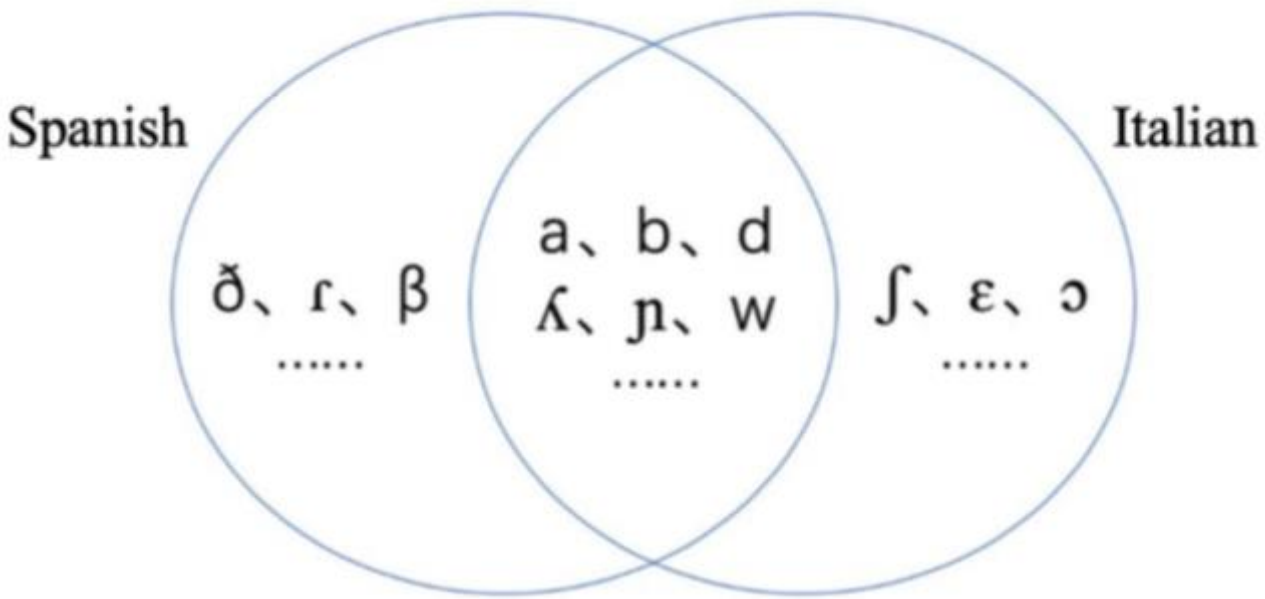
+	-	0
10	01	00

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

Phonological feature	d	ɛ	ð	ə	i	ʤ	k^j
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

Phonological features

- Unseen phones are connected by using phonological-vector.

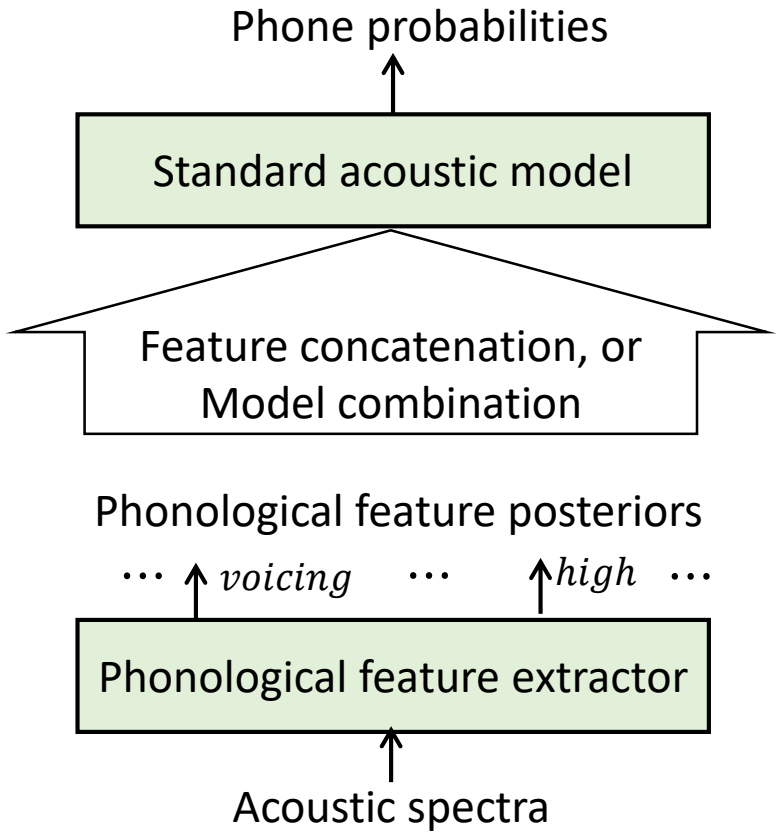


ð：-,-,+,+,-,-,-,0,+,-,-,+,+,+,-,-,-,-,-,-,0,-,0,0

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

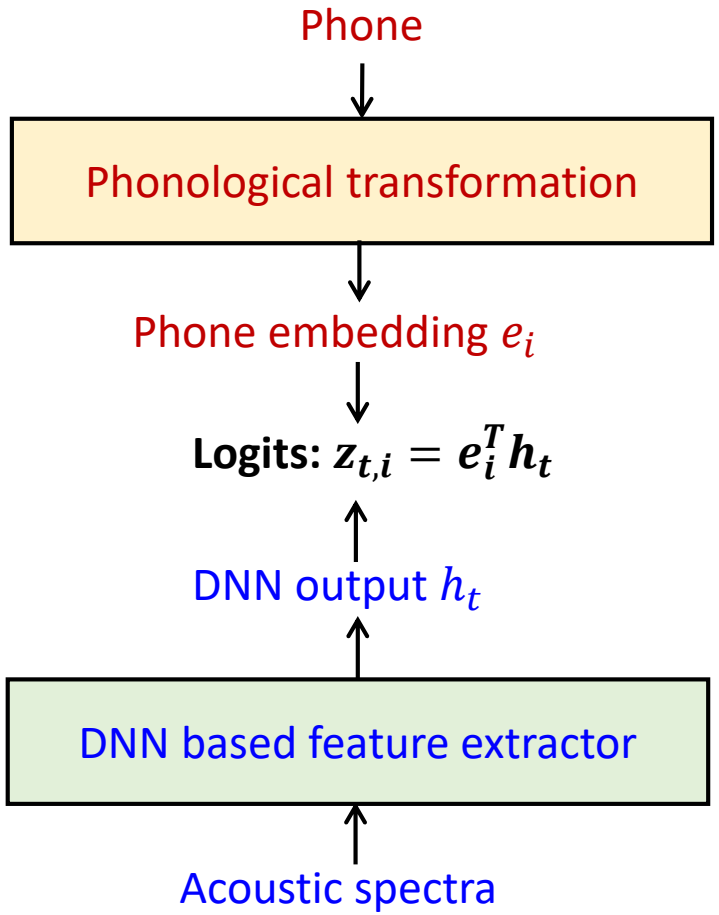
Related work

- 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.



Joining of Acoustics and Phonology (JoinAP)

- The JoinAP method
  - Phonology driven phone embedding (top-down) and DNN based acoustic feature extraction (bottom-up) 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:

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

Experiments

- Train multilingual AM on German, French, Spanish and Polish.
- Zero-shot and few-shot crosslingual ASR on Polish and Mandarin.
- Employ Phonetisaurus G2P to generate IPA lexicons
- 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 \ Language-degree	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:

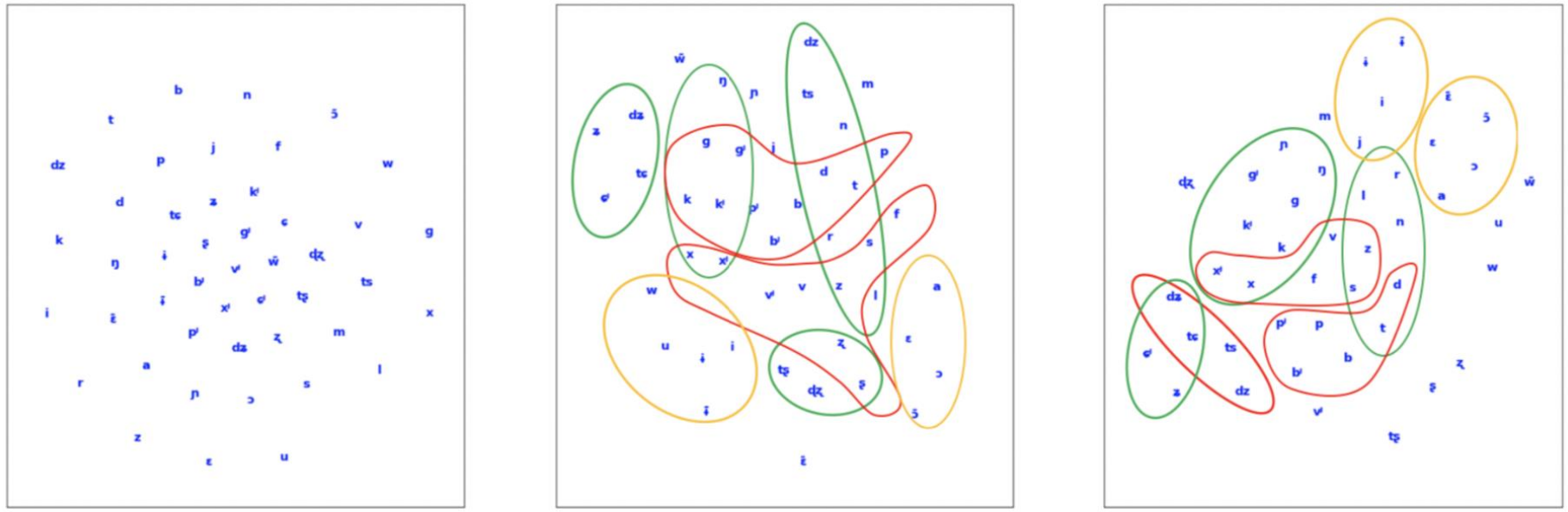
Language	#Phones	#Unseen phones
Polish	46	18
Mandarin	96	79

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

Experiments

- t-SNE map of Polish phone embeddings

  (obtained from un-finetuned multilingual models)



(a) (b) (c)

(a) Flat phone embeddings, (b) JoinAP-Linear phone embeddings, (c) JoinAP- Nonlinear phone embeddings.

Consonants with the same manner of articulation

Consonants with the same place of articulation

Vowel with similar height

Experiments

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

Method	Color	Feature	Phones
Linear	Green	Alveolo-palatal	ʑ ɕ dʑ tɕ
		Velar	ŋ ɡ k x ɡʲ kʲ xʲ
		Alveolar	dz ts n d t r s z l
		Retroflex	ʂ ʐ dʐ tʂ
	Red	Plosive	x v xʲ vʲ z s f ʐ ʂ
		Fricative	ɡ k p ɡʲ 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 dʑ tɕ
		Velar	ŋ ɡ k x ɡʲ kʲ xʲ
		Alveolar	n d t r s z l
	Red	Affricate	dʑ tɕ dz ts
		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

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.

Conclusions

- 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.
- Flexibility and future work
  - Streaming ASR <- INTRESPEECH 2020
  - Neural Architecture Search <- SLT 2021
  - Children Speech Recognition <- SLT 2021
  - Wordpieces, Conformer architectures <- submitted to ASRU2021
  - Multilingual and Crosslingual <- submitted to ASRU2021
  - …

Thanks for your attention !