



CTC-CRF

CRF-based single-stage acoustic modeling with CTC topology

Hongyu Xiang, Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab

Tsinghua University

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

Content



1. Introduction

- Related work

2. CTC-CRF

3. Experiments

4. Conclusions



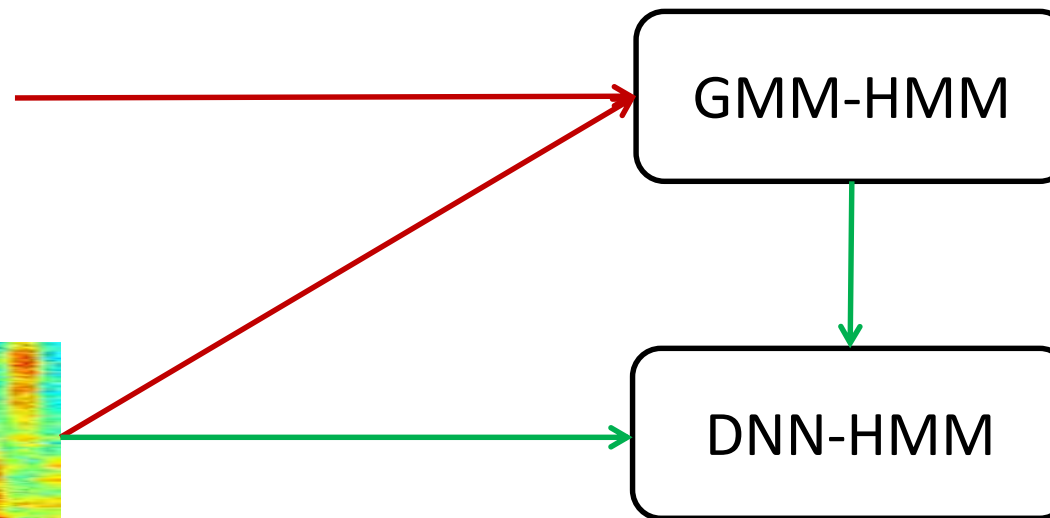
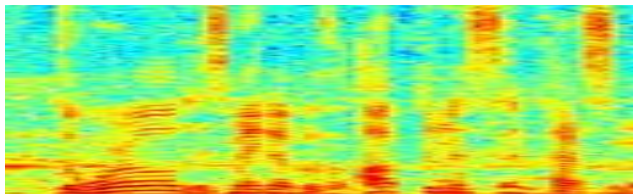
Introduction

- ASR is a discriminative problem
 - For acoustic observations $\mathbf{x} \triangleq x_1, \dots, x_T$, find the most likely labels $\mathbf{l} \triangleq l_1, \dots, l_L$
- ASR state-of-the-art: DNNs of various network architectures
- Conventionally, multi-stage
 - Monophone \rightarrow alignment & triphone tree building \rightarrow triphone \rightarrow alignment \rightarrow DNN-HMM

Labels \mathbf{l} :

Nice to meet you.

Acoustic features \mathbf{x} :





Motivation

- End-to-end system:
 - Eliminate GMM-HMM pre-training and 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 are interested in advancing single-stage acoustic models, which use a separate language model (LM) with or without a pronunciation lexicon.

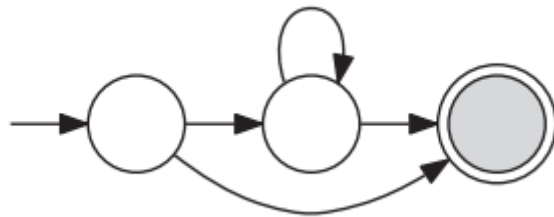
- Text corpus for language modeling are cheaply available.
- Data-efficient



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.

Related work



ASR is a discriminative problem

- For acoustic observations $\mathbf{x} \triangleq x_1, \dots, x_T$, find the most likely labels $\mathbf{l} \triangleq l_1, \dots, l_L$

1. How to obtain $p(\mathbf{l} | \mathbf{x})$
2. How to handle alignment, since $L \neq T$



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

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

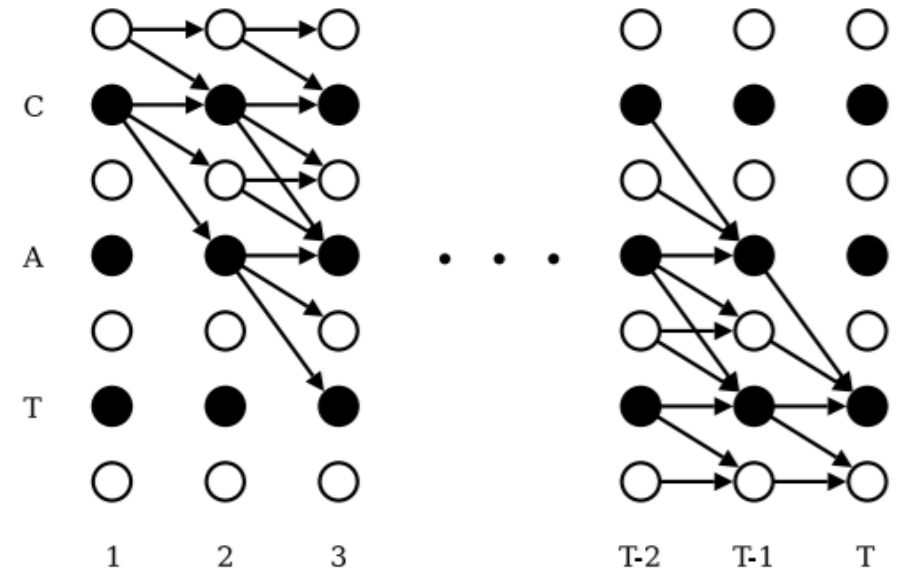
$$p(l|x) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(l)} p(\boldsymbol{\pi}|x)$$

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

- removing all repetitive symbols between the blank symbols.
- 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.



Related work

How to obtain $p(\mathbf{l} | \mathbf{x})$

■ Directed Graphical Model/Locally normalized

- DNN-HMM : Model $p(\boldsymbol{\pi}, \mathbf{x})$ as an HMM, could be discriminatively trained, e.g. by $\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{l} | \mathbf{x})$
- CTC : Directly model $p(\boldsymbol{\pi} | \mathbf{x}) = \prod_{t=1}^T p(\pi_t | \mathbf{x})$
- Seq2Seq : Directly model $p(\mathbf{l} | \mathbf{x}) = \prod_{i=1}^L p(l_i | l_1, \dots, l_{i-1}, \mathbf{x})$

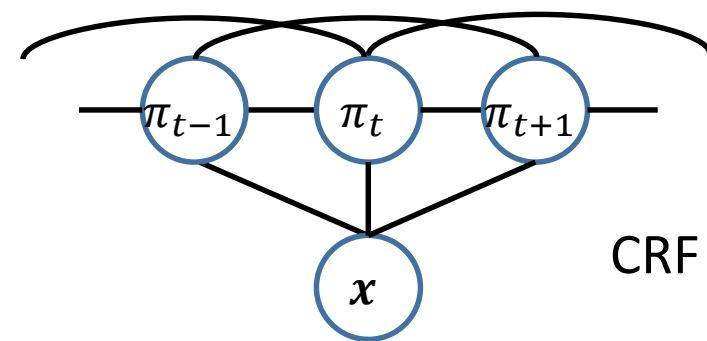
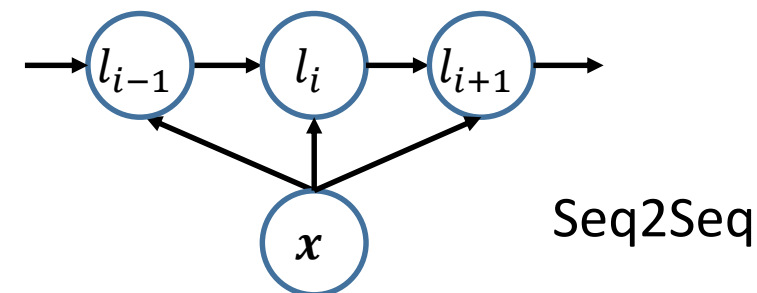
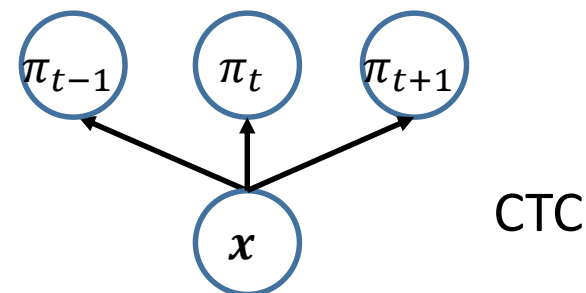
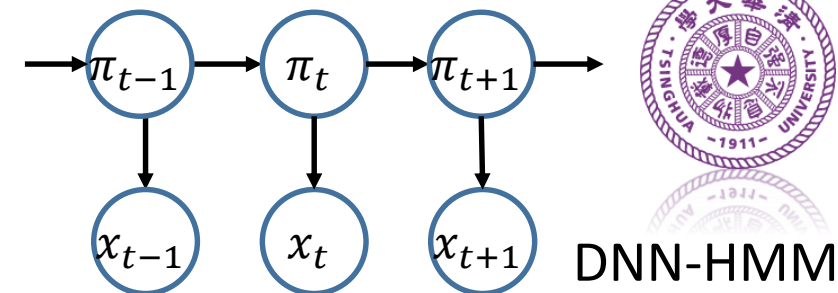
■ Undirected Graphical Model/Globally normalized

- CRF : $p(\boldsymbol{\pi} | \mathbf{x}) \propto \exp[\phi(\boldsymbol{\pi}, \mathbf{x})]$

MMI training of GMM-HMMs is equiv. to

CML training of CRFs (using 0/1/2-order features in potential definition).

Heigold, et al., "Equivalence of generative and log-linear models", T-ASLP 2011.



Related work (summary)



Model	State topology	Training objective	Locally/globally normalized
Regular HMM	HMM	$p(\mathbf{x} \mathbf{l})$	Local
Regular CTC	CTC	$p(\mathbf{l} \mathbf{x})$	Local
SS-LF-MMI	HMM	$p(\mathbf{l} \mathbf{x})$	Local
CTC-CRF	CTC	$p(\mathbf{l} \mathbf{x})$	Global
Seq2Seq	-	$p(\mathbf{l} \mathbf{x})$	Local

- To the best of our knowledge, this paper represents the first exploration of CRFs with CTC topology.

Content

1. Introduction

- Related work

2. CTC-CRF

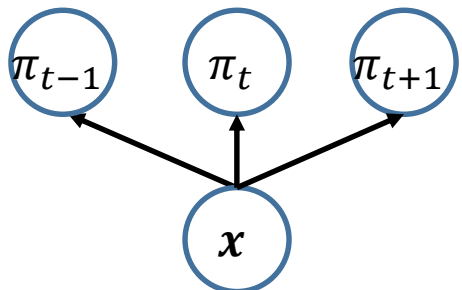
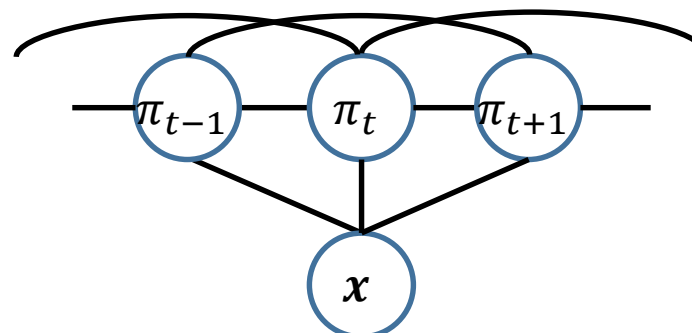
3. Experiments

4. Conclusions



CTC vs CTC-CRF



CTC	CTC-CRF
$p(\mathbf{l} \mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi \mathbf{x}), \text{ using CTC topology } \mathcal{B}$	
<p>State Independence</p> $p(\pi \mathbf{x}; \theta) = \prod_{t=1}^T p(\pi_t \mathbf{x})$	$p(\pi \mathbf{x}; \theta) = \frac{e^{\phi(\pi, \mathbf{x}; \theta)}}{\sum_{\pi'} e^{\phi(\pi', \mathbf{x}; \theta)}}$ <p style="text-align: right; color: red;">Node potential, by NN</p> $\phi(\pi, \mathbf{x}; \theta) = \sum_{t=1}^T \left(\log p(\pi_t \mathbf{x}) + \log p_{LM}(\mathcal{B}(\pi)) \right)$ <p style="text-align: right; color: red;">Edge potential, by n-gram denominator LM of labels, like in LF-MMI</p>
$\frac{\partial \log p(\mathbf{l} \mathbf{x}; \theta)}{\partial \theta} = \mathbb{E}_{p(\pi \mathbf{l}, \mathbf{x}; \theta)} \left[\frac{\partial \log p(\pi \mathbf{x}; \theta)}{\partial \theta} \right]$	$\frac{\partial \log p(\mathbf{l} \mathbf{x}; \theta)}{\partial \theta} = \mathbb{E}_{p(\pi \mathbf{l}, \mathbf{x}; \theta)} \left[\frac{\partial \phi(\pi, \mathbf{x}; \theta)}{\partial \theta} \right] - \mathbb{E}_{p(\pi' \mathbf{x}; \theta)} \left[\frac{\partial \phi(\pi', \mathbf{x}; \theta)}{\partial \theta} \right]$
	

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.

Content

1. Introduction

- Related work

2. CTC-CRF

3. Experiments

4. Conclusions



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):

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



Experiments (Comparison with CTC, phone based)

WSJ

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

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

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.

Experiments (Comparison with SS-LF-MMI, phone based)



WSJ

Model	Unit	LM	SP	dev93	eval92
SS-LF-MMI	Mono-phone	4-gram	Y	6.3%	3.1%
SS-LF-MMI	Bi-phone	4-gram	Y	6.0%	3.0%
CTC-CRF	Mono-phone	4-gram	Y	6.23%	3.79%

Switchboard

Model	Unit	LM	SP	SW	CH
SS-LF-MMI	Mono-phone	4-gram	Y	11.0%	20.7%
SS-LF-MMI	Bi-phone	4-gram	Y	9.8%	19.3%
CTC-CRF	Mono-phone	4-gram	Y	10.3%	19.7%
Seq2Seq	Subword	LSTM	N	11.8%	25.7%

Red arrows indicate improvements: 6.4% for SW and 4.8% for CH.

Librispeech

Model	Unit	LM	SP	Dev Clean	Dev Other	Test Clean	Test Other
LF-MMI	Tri-phone	4-gram	Y	-	-	4.28%	-
CTC-CRF	Mono-phone	4-gram	N	3.87%	10.28%	4.09%	10.65%
Seq2Seq	Subword	4-gram	N	4.79%	13.1%	4.82%	15.30%

Red arrow indicates improvement: 4.4% for Test Clean.

Experiments (Comparison with SS-LF-MMI, phone based)



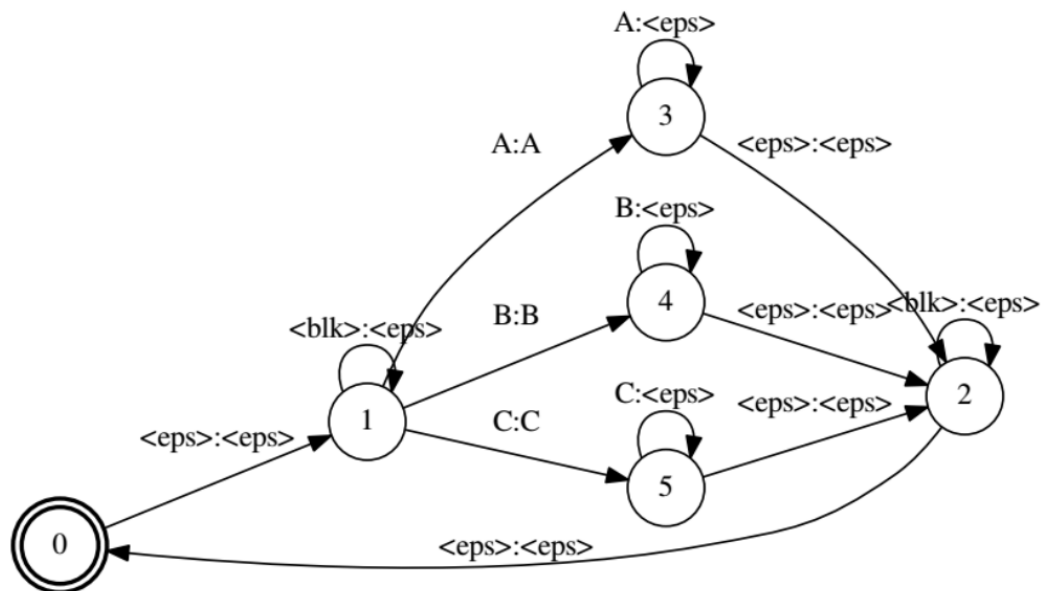
Switchboard (After camera-ready version)

Model	Unit	LM	SP	SW	CH
SS-LF-MMI	Mono-phone	4-gram	Y	11.0%	20.7%
SS-LF-MMI	Bi-phone	4-gram	Y	9.8%	19.3%
CTC-CRF	Mono-phone	4-gram	Y	10.3%	19.7%
Seq2Seq	Subword	LSTM	N	11.8%	25.7%
CTC-CRF	Clustered Bi-phone	4-gram	Y	9.8%	19.0%

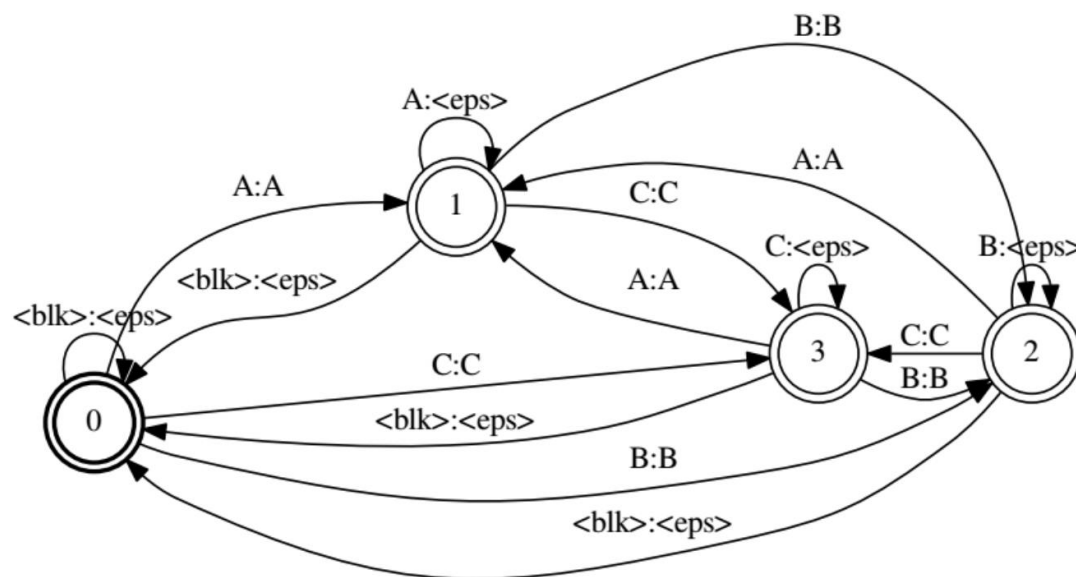
Red arrows indicate performance improvements: 5% improvement in SW (from 10.3% to 9.8%) and 4% improvement in CH (from 19.7% to 19.0%) for the Clustered Bi-phone model compared to the Mono-phone CTC-CRF model.

Bi-phones clustering from 1213 to 311 according to frequencies

WFST representation of CTC topology



EESSEN 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.

Content

1. Introduction

- Related work

2. CTC-CRF

3. Experiments

4. Conclusions





Conclusions

- We propose a framework for single-stage acoustic modeling based on CRFs with CTC topology (CTC-CRF).
- CTC-CRFs achieve strong results on WSJ, Switchboard and Librispeech datasets.
 - CTC can be significantly improved by CTC-CRF;
 - CTC-CRF significantly outperforms attention-based Seq2Seq;
 - CTC-CRF outperforms the SS-LF-MMI in both cases of mono-phones and mono-chars (except in WSJ);
 - Conceptually simple, and avoids some ad-hoc operations in SS-LF-MMI (randomly inserting silence labels in estimating denominator LM, length modification, leaky HMM).
- Going to release Crf-based Asr Toolkit (CAT) for reproducing this work.



Thanks for your attention !

Hongyu Xiang, Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab

Tsinghua University

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