On Modular Training of Neural Acoustics-to-Word Model for LVCSR

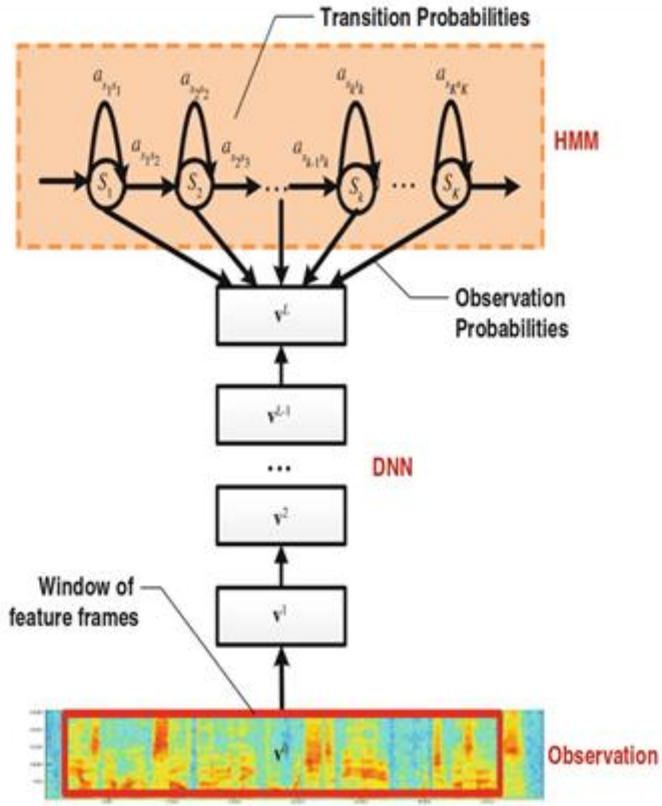
Zhehuai Chen, Qi Liu, Hao Li and Kai Yu

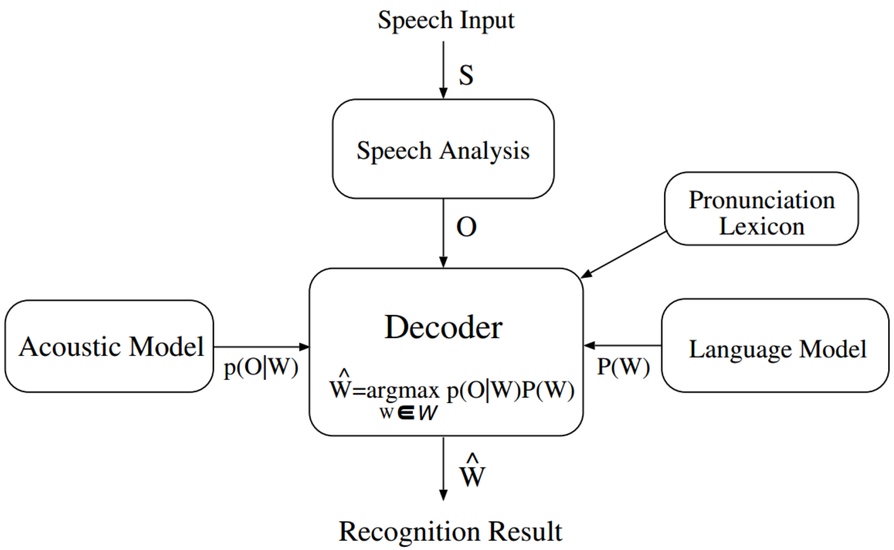
Outline

- Review of End-to-End (E2E) ASR
- Motivation and our Target
- Modular training strategy
  - Framework
  - Analysis
  - Implementation
- Experiment

Review
ASR and DNN-HMM hybrid system

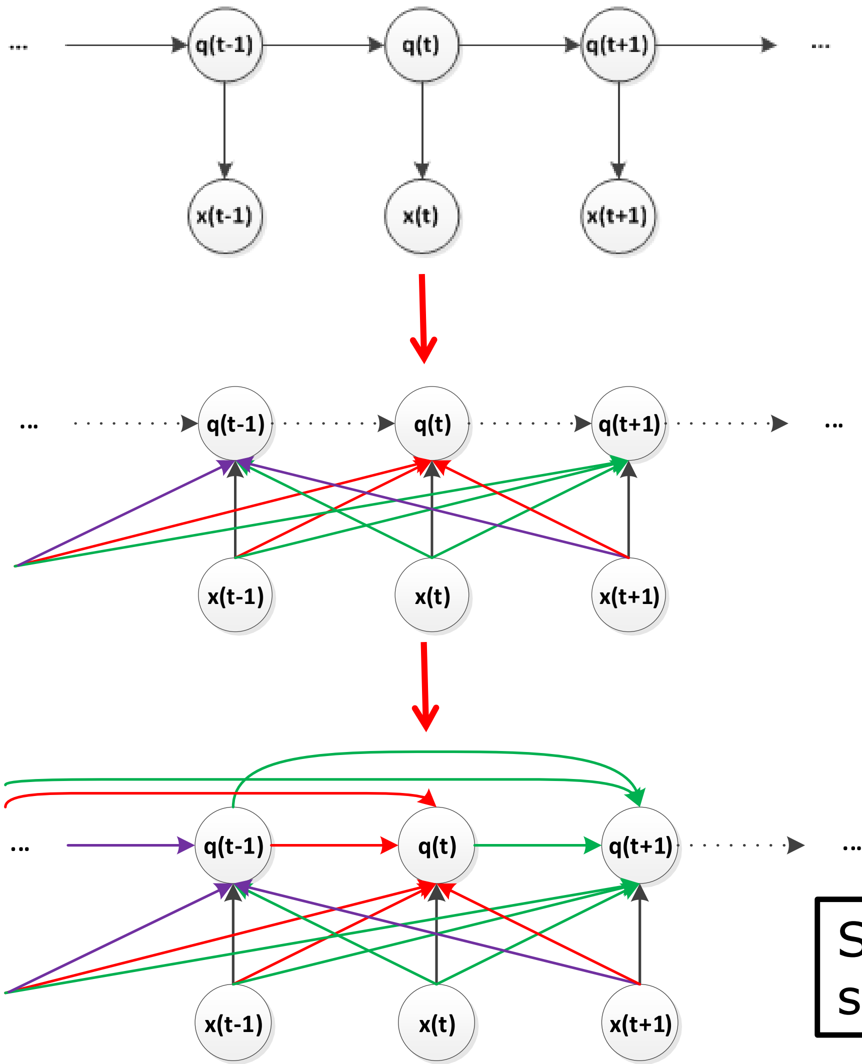
- Acoustic, pronunciation, and language model
- Separate optimization
- Alignment from an existing model
- Decoder to combine them and find the best hypothesis

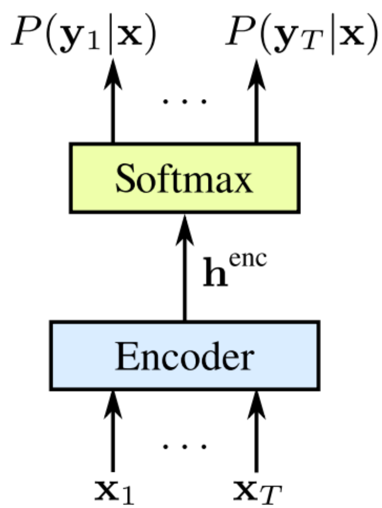




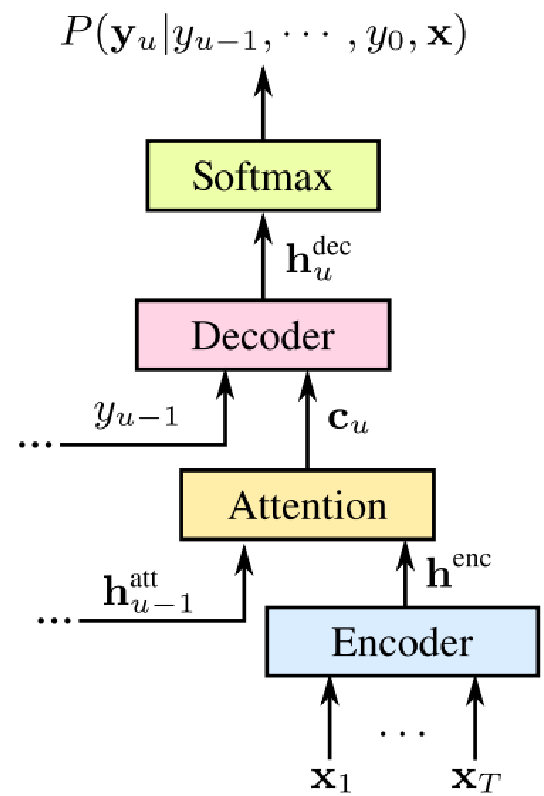
Review
End-to-End (E2E) ASR

- HMM→CTC→S2S





Connectionist Temporal Classification (CTC)



Sequence-to-sequence (S2S)

Motivation and our Target

- Characteristics:

  End-to-End optimization + End-to-End inference (decoding)

- Advantages:
  - Better sequential modeling: better WER (Soltau et al.2017)
  - Simpler and faster decoding: 3-5X speedup (Chen et al.2017)

Motivation and our Target

- Characteristics:
  End-to-End optimization + End-to-End inference (decoding)

- Advantages:
  - Better sequential modeling: better WER (Soltau et al.2017)
  - Simpler and faster decoding: 3-5X speedup (Chen et al.2017)

- Disadvantages:

Big data? But why?

Motivation and our Target

- Characteristics:

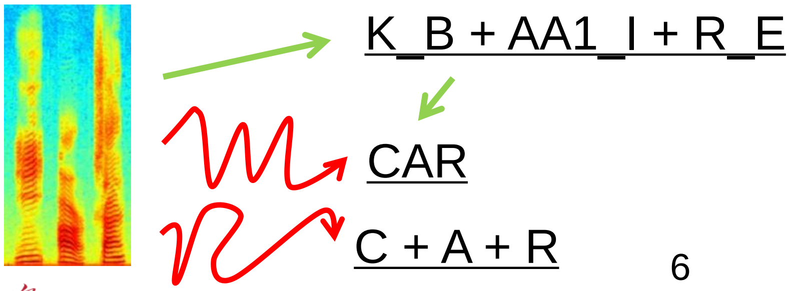
  End-to-End optimization + End-to-End inference (decoding)

- Advantages:
  - Better sequential modeling: better WER (Soltau et al.2017)
  - Simpler and faster decoding: 3-5X speedup (Chen et al.2017)

- Disadvantages:

Big data? But why?

  - Acoustic data and text data usage
  - AM and LM both infer grapheme/word
  - Hard to apply prior arts



Motivation and our Target

- Characteristics:
  End-to-End optimization + End-to-End inference (decoding)

- Disadvantages:
  - Acoustic data and text data usage
  - AM and LM both infer grapheme/word
  - Hard to apply prior arts

Our Solution

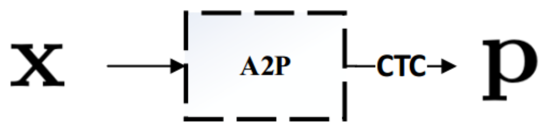
- Step 1: utilize different sources to train each building block (for performance)
- Step 2: retaining end-to-end decoding by final joint optimization (for speed)

Modular training strategy
Framework

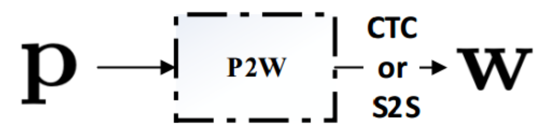
$$P(\mathbf{w}|\mathbf{x}) \approx \max_{\mathbf{p}} [\, P(\mathbf{w}|\mathbf{p}) \cdot PSD(\, P(\mathbf{p}|\mathbf{x})\,)\,]$$

- utilizing acoustic and text data in E2E ASR modeling by modular training strategy

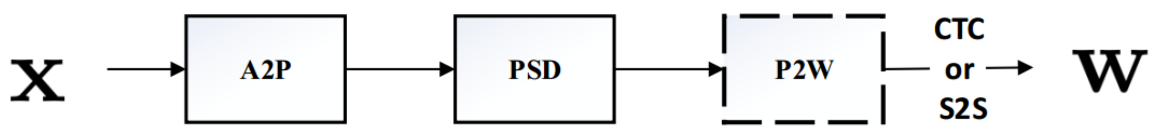
- combining modules into an acoustics-to-word model (A2W) by phone synchronous decoding (PSD, Chen et al.2017) and joint optimization



(a) Acoustic-to-phoneme Module



(b) Phoneme-to-word Module



(c) PSD-based Joint Training

Modular training strategy
Analysis

$$P(\mathbf{w}|\mathbf{x}) \approx \max_{\mathbf{p}} [\, P(\mathbf{w}|\mathbf{p}) \cdot PSD(\, P(\mathbf{p}|\mathbf{x})\,)\,]$$

- Compared with Multi-modal Training ♤ :
  - modularizing the end-to-end speech recognition by Bayesian theorem
  - utilizing respective inference units for acoustic and language modeling
  - the LM generalizes word sequences and lexicons jointly.

♤ Multi-model Training refers to methods utilizing multi-source data to augment the ASR training corpus

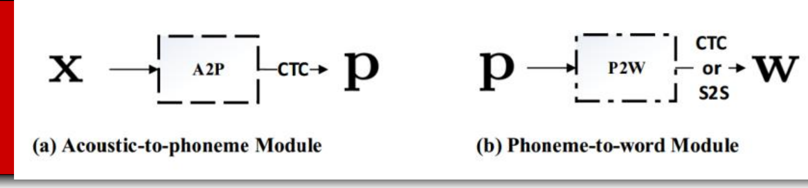
Modular training strategy
Analysis

$$P(\mathbf{w}|\mathbf{x}) \approx \max_{\mathbf{p}} [\, P(\mathbf{w}|\mathbf{p}) \cdot PSD(\, P(\mathbf{p}|\mathbf{x}) \,) \,]$$

- Compared with Multi-modal Training:
  - modularizing the end-to-end speech recognition by Bayesian theorem
  - utilizing respective inference units for acoustic and language modeling
  - the LM generalizes word sequences and lexicons jointly.
- What we expect:
  - easier and faster model convergence due to modularization and initialization
  - easy to utilize traditional AM and LM techs using text and acoustic data respectively.

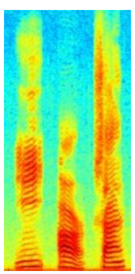
Modular training strategy

Modularization



(a) Acoustic-to-phoneme Module (b) Phoneme-to-word Module

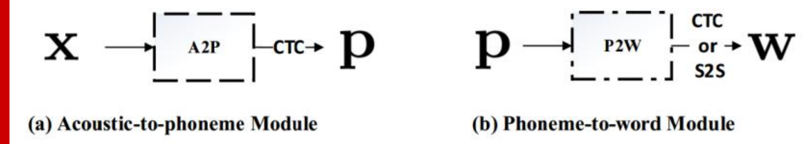
- Still take phoneme as the mediator between acoustics and words
- Using **acoustic** data, train a phoneme recognition model, $P(\mathbf{p}|\mathbf{x})$, e.g. the standard mono-phone CTC or LFMMI.



K_B + AA1_I + R_E

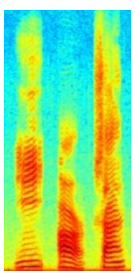
Modular training strategy

Modularization



(a) Acoustic-to-phoneme Module
(b) Phoneme-to-word Module

- Still take phoneme as the mediator between acoustics and words
- Using **acoustic** data, train a phoneme recognition model, $P(\mathbf{p}|\mathbf{x})$, e.g. the standard mono-phone CTC or LFMMI.
- Using **text** data, train a phoneme-to-word system, $P(\mathbf{w}|\mathbf{p})$, e.g. CTC or S2S.

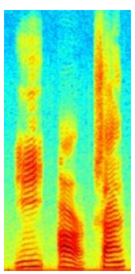


K_B + AA1_I + R_E → CAR

Modular training strategy
Modularization

(a) Acoustic-to-phoneme Module (b) Phoneme-to-word Module

- Still take phoneme as the mediator between acoustics and words
- Using **acoustic** data, train a phoneme recognition model, $P(\mathbf{p}|\mathbf{x})$, e.g. the standard mono-phone CTC or LFMMI.
- Using **text** data, train a phoneme-to-word system, $P(\mathbf{w}|\mathbf{p})$, e.g. CTC or S2S.
  - P2W model v.s. LM:
    - implicitly doing the phoneme tokenization
    - always easier than LM, as P2W gets more phoneme hints from the next word
    - trained by sequence criteria → learn phoneme-word alignment
  - Adding word boundary unit <wb> to help tokenization



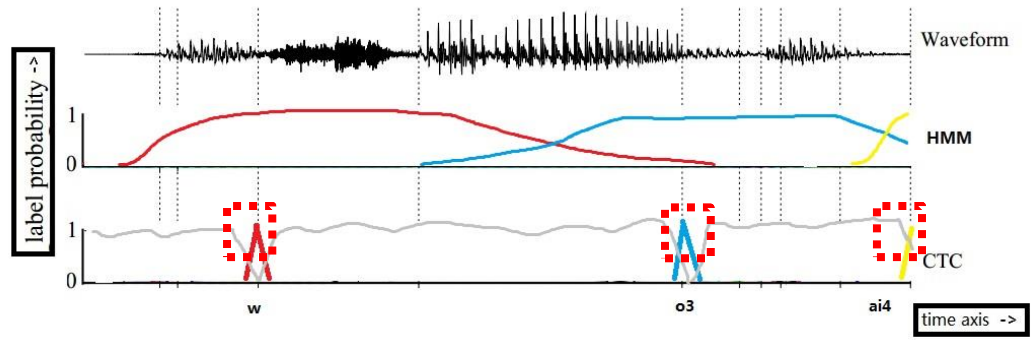
Oh, god: OW1_S <wb> G_B AA1_I D_E <wb>

K_B + AA1_I + R_E → CAR

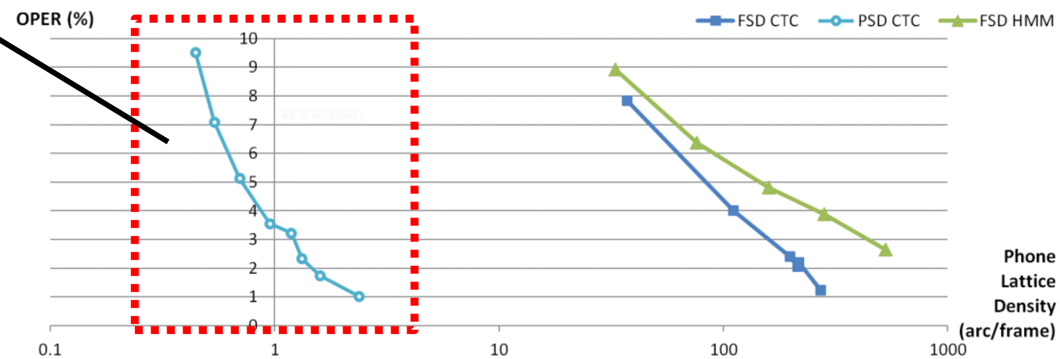
Modular training strategy

Phone Synchronous Decoding and Joint Optimization

- Motivation:
  - Different information rate in acoustics and phoneme
  - long sequence is hard for S2S (for speech, avg. 500 tokens)
  - Speedup training and decoding



Reduce information rate without precision loss



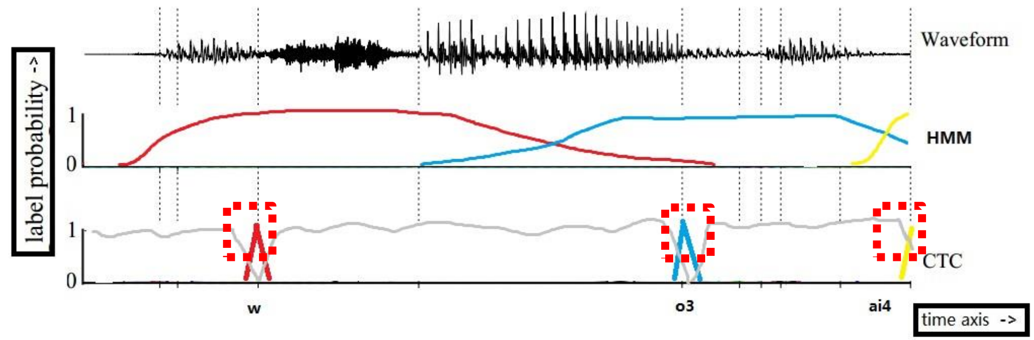
[1] Chen, Zhehuai, et al. "Phone synchronous speech recognition with ctc lattices." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (2017): 90-101.

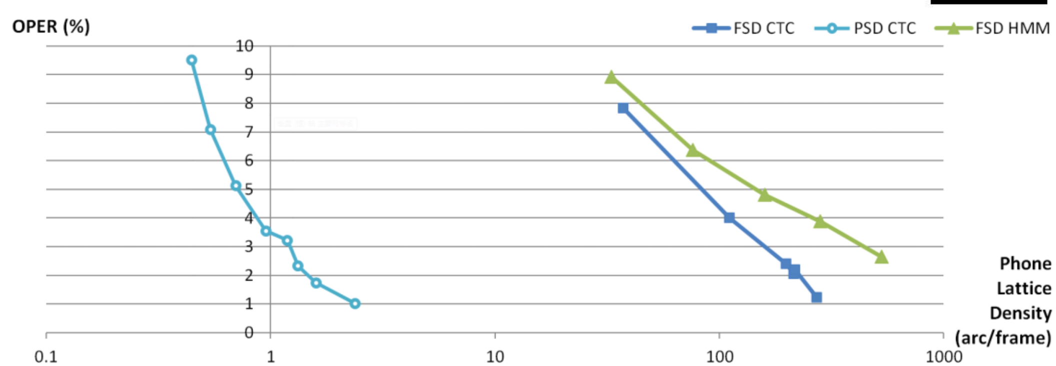
Modular training strategy

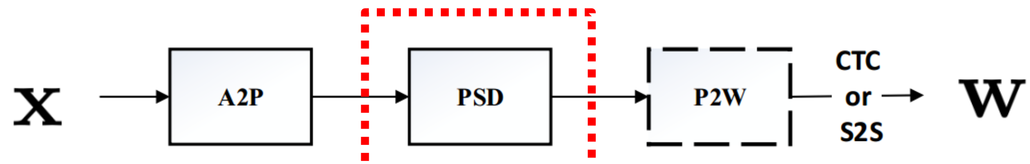
Phone Synchronous Decoding and Joint Optimization

- Motivation:
  - Different information rate in acoustics and phoneme
  - long sequence is hard for S2S (for speech, avg. 500 tokens)
  - Speedup training and decoding
- Procedure:

- A2P inference
- PSD sub-sampling
- P2W inference
- Back propagation
- fine-tune P2W only







(c) PSD-based Joint Training

Experiment
Setup

- Switchboard 300 corpus
- A2P model
  - CTC
  - 36-d fbank
  - 45 mono-phones and a blank and <wb>
  - 5X1024(P=256) LSTMs
- P2W model
  - CTC / S2S
  - 30K vocabulary size
- 3-gram SWBD LM without Fisher interpolation
  - Hybrid CE **baseline**
  - Mono-phone CTC **baseline**
- Direct A2W **baseline**

- More details in our paper

Experiment
Modularization

- Performance of each module in the validation set

Module	Model	Inf. Label	Word bound.	PER/WER CV (%)
A2P	CTC	phoneme	×	13.0
			√	**12.0**
P2W	CTC	word	×	16.0
			√	**4.3**
	S2S	word	×	13.9
			√	**2.8**

  - <wb> doesn't hurt the A2P performance (prediction error=4%)

Experiment
Modularization

- Performance of each module in the validation set

Module	Model	Inf. Label	Word bound.	PER/WER CV (%)
A2P	CTC	phoneme	×	13.0
			√	**12.0**
P2W	CTC	word	×	16.0
			√	**4.3**
	S2S	word	×	13.9
			√	**2.8**

  - <wb> doesn't hurt the A2P performance (prediction error=4%)
  - <wb> significantly helps P2W

Experiment
Modularization

- Performance of each module in the validation set

Module	Model	Inf. Label	Word bound.	PER/WER CV (%)
A2P	CTC	phoneme	×	13.0
			√	**12.0**
P2W	CTC	word	×	16.0
			√	**4.3**
	S2S	word	×	13.9
			√	**2.8**

  - <wb> doesn't hurt the A2P performance (prediction error=4%)
  - <wb> significantly helps P2W
  - S2S is consistently better thanks to removal of conditional independent assumption in CTC

Experiment
Baseline

- CI-phone CTC v.s. CD-phone CE is similar to other research in this corpus

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST ♤	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5

♤ "WFST" in P2W is compiled from a 3-gram LM trained by SWBD corpus.

Experiment
Baseline

- CI-phone CTC v.s. CD-phone CE is similar to other research in this corpus
- Direct A2W CTC with phoneme initialization but without GloVe in [1]

Name	E2E Opt.	Modularization A2P	P2W	WER (%) swbd	callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5

[1] Audhkhasi K, Ramabhadran B, Saon G, et al. Direct Acoustics-to-Word Models for English Conversational Speech Recognition[J]. Proc. Interspeech 2017, 2017: 959-963.

Experiment

Effects of Modular Training Strategy

- Proposed modular training significantly improves the baseline
  - Easier and faster model convergence
  - Better to capture the LM knowledge source

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5

Experiment
Effects of Phone Synchronous Decoding

- Training speedup
  - PSD reduces the sequence length to be processed by P2W in each sequence
  - As the sequence length is reduced, more sequences can be loaded into GPU memory for parallel training

Name	PSD	Training Speed Seq./GPU ♤	Training Speed fr./s. ◊	WER (%) swbd	WER (%) callhm
Mod. CTC	×	5	1027	32.0	42.5
	√	**30**	**5851**	**24.9**	**36.5**

♤ "seq./GPU" denotes the number of streams used in parallel LSTM training.
◊ "fr./s." denotes the number of acoustics frames processed per second.

Experiment
Effects of Phone Synchronous Decoding

- Training speedup
  - PSD reduces the sequence length to be processed by P2W in each sequence
  - As the sequence length is reduced, more sequences can be loaded into GPU memory for parallel training
- Performance improvement
  - Reduced sequence length (some researches cope it by pyramid model structure)

Name	PSD	Training Speed Seq./GPU ♤	Training Speed fr./s. ◊	WER (%) swbd	WER (%) callhm
Mod. CTC	×	5	1027	32.0	42.5
	√	**30**	**5851**	**24.9**	**36.5**

♤ "seq./GPU" denotes the number of streams used in parallel LSTM training.
◊ "fr./s." denotes the number of acoustics frames processed per second.

Experiment
More Comparisons

- Decoding with external LM still helps
  - Current P2W modeling is still not perfect (conditional independent assumption in CTC)

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5
	√	CTC	+WFST♤	23.0	35.1
Mod. S2S	√	CTC	S2S	31.2	40.5

♤ "WFST" in P2W is compiled from a 3-gram LM trained by SWBD corpus.

Experiment
More Comparisons

- Decoding with external LM still helps
  - Current P2W modeling is still not perfect
  - The overall improvement is similar to the optimization in [1]

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5
	√	CTC	+WFST	23.0	35.1
Mod. S2S	√	CTC	S2S	31.2	40.5

[1] Audhkhasi K, Ramabhadran B, Saon G, et al. Direct Acoustics-to-Word Models for English Conversational Speech Recognition[J]. Proc. Interspeech 2017, 2017: 959-963.

Experiment
More Comparisons

- Unlike in P2W task, S2S shows no improvement:
  - S2S is prone to the phoneme recognition errors from the A2P module

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5
	√	CTC	+WFST	23.0	35.1
Mod. S2S	√	CTC	S2S	31.2	40.5

Experiment
More Comparisons

- Overall, the gap between E2E ASR and traditional CTC is reduced to relative 15% (in [1], 21.7 → 14.5, 30% gap)
  - Modular strategy could be better to catch up the gap

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5
	√	CTC	+WFST	23.0	35.1
Mod. S2S	√	CTC	S2S	31.2	40.5

[1] Audhkhasi K, Ramabhadran B, Saon G, et al. Direct Acoustics-to-Word Models for English Conversational Speech Recognition[J]. Proc. Interspeech 2017, 2017: 959-963.

Experiment
More Comparisons

- Our new results
  - The gap can finally disappeared (still retaining E2E decoding)
  - Modular training is easy to combine with prior arts

Name	E2E Opt.	Modularization A2P	Modularization P2W	WER (%) swbd	WER (%) callhm
CD-phone CE	×	HMM	WFST	14.9	27.6
CI-phone CTC	×	CTC	WFST	19.4	33.5
Word CTC	√	n/a	n/a	29.6	41.7
Mod. CTC	√	CTC	CTC	24.9	36.5
	√	CTC	+WFST	23.0	35.1
Mod. S2S	√	CTC	S2S	31.2	40.5
new results					
better CTC	√	CTC	CTC	19.8	34.0
+ I-vector, etc.	√	CTC	CTC	16.5	30.5
better S2S	√	CTC	S2S	24.4	37.2

Experiment
Examples Analysis

- Stronger language and context modeling
- Less robustness

```
id: (sw_4390_a-001)
Labels: <o,sw,m,sw-m>
File: sw_4390
Channel: a
REF:  well (%hesitation) CURRENTLY (%hesitation) i (li-) ACTUAL   live in virginia and VIRGINIA    does have the DEATH penalty but (%hesitation)
Scores: (#C #S #D #I) 18 1 0 0
HYP:  well               currently               i       ACTUALLY live in virginia and virginia does have the death penalty but
Eval:                                                    S
Scores: (#C #S #D #I) 14 5 0 2
HYP:  well %hesitation   CRIME      BELIE-       i- I     ACTUALLY live in virginia and THEN IT       does NOT have the death penalty but %hesitation
Eval:                    S          S               S     S                             I    S              I

id: (sw_4390_a-020)
Labels: <o,sw,m,sw-m>
File: sw_4390
Channel: a
REF:  i **** *** *** WANNA CONTINUE along the LINES OF   fair AND speedy trial
Scores: (#C #S #D #I) 6 5 0 3
HYP:  i WILL PUT CAN SEE   YOU      along the LINE  THEM fair IN  speedy trial
Eval:   I    I   I   S     S                  S     S         S
Scores: (#C #S #D #I) 9 2 0 0
HYP:  i WOULD continue along the lines of fair IN  speedy trial
Eval:   S                                      S
```

Mod. E2E CTC

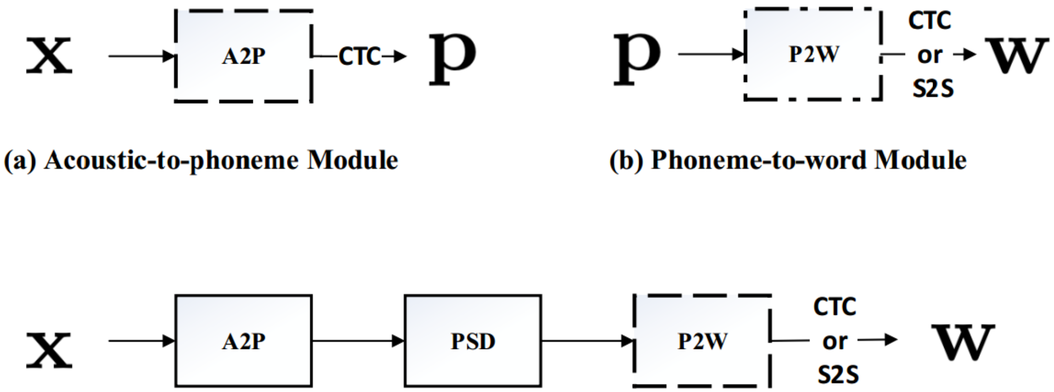
CI-phone CTC+WFST

Mod. E2E CTC

CI-phone CTC+WFST

Conclusion

- Utilizing different sources to train each building block for easier and faster model convergence
- retaining end-to-end decoding by final joint optimization
- Phone Synchronous Decoding helps both performance and speed



(a) Acoustic-to-phoneme Module

(b) Phoneme-to-word Module

(c) PSD-based Joint Training

- Promising to:
  - solve “big data” problem
  - utilize traditional AM and LM techs using text and acoustic data respectively

Backup materials

Experiment
Effects of Phone Synchronous Decoding

- Training speedup
- Performance improvement
- Compared to A2W baseline
  - Benefit: better convergence and knowledge integration
  - Harm: information loss from modularization

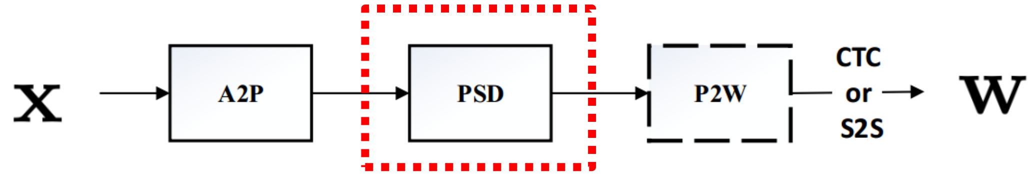
Name	PSD	Training Speed Seq./GPU	Training Speed fr./s.	WER (%) swbd	WER (%) callhm
Word CTC (baseline)	-	-	-	29.6	41.7
Mod. CTC	×	5	1027	32.0	42.5
	√	**30**	**5851**	**24.9**	**36.5**

"fr./s." denotes the number of acoustics frames processed per second.
"seq./GPU" denotes the number of streams used in parallel LSTM training.

Modular training strategy
Phone Synchronous Decoding and Joint Optimization

- **Why we only fine-tune P2W:**
  - the A2P module, mono-phone level CTC model, can always achieve good modeling effects for phoneme recognition.
  - take distribution but not one-hot
  - fixing A2P and combining PSD module can greatly speed up the joint optimization, which we will show in experiments
- Procedure:

- A2P inference
- PSD sub-sampling
- P2W inference
- Back propagation
- fine-tune P2W only



(c) PSD-based Joint Training

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY