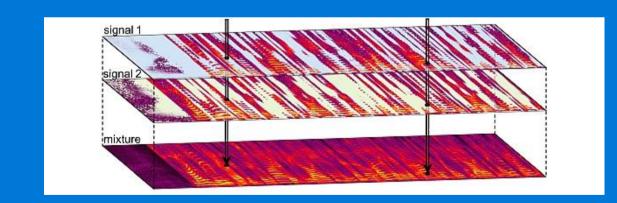
Sequence Modeling

in Unsupervised Single-channel Overlapped Speech Recognition

Zhehuai Chen and Jasha Droppo

chenzhehuai@sjtu.edu.cn jdroppo@microsoft.com







Outline

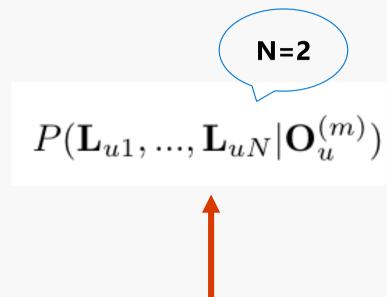
- Introduction
 - Cocktail party problem
 - PIT-TS framework and discriminative training
- Proposed methods
 - Temporal Correlation Modeling
 - Integrating Language Model
- Experiments
- Conclusion

Introduction

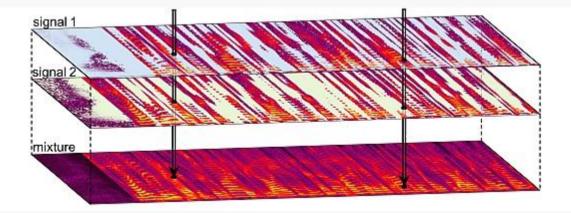
Cocktail-party problem







$$\mathbf{O}_u^{(m)} = \sum_{n=1}^N \mathbf{O}_{un}^{(r)}$$



Assignment error:

e.g. ch-a: how oh you

ch-b: are no

Cross talk error:

e.g. ch-a: how are you

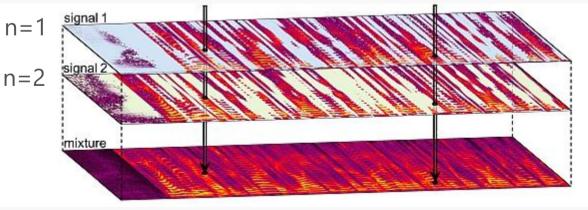
ch-b: oh are no

Label assignment problem

$$P(\mathbf{L}_{u1}, ..., \mathbf{L}_{uN} | \mathbf{O}_{u}^{(m)}) \approx \prod_{n=1}^{N} P(\mathbf{L}_{un}^{(r)} | \mathbf{O}_{u}^{(m)})$$

$$\underline{\mathbf{Label}}$$
Independence

$$\mathbf{O}_{u}^{(m)} = \sum_{n=1}^{N} \mathbf{O}_{un}^{(r)}$$



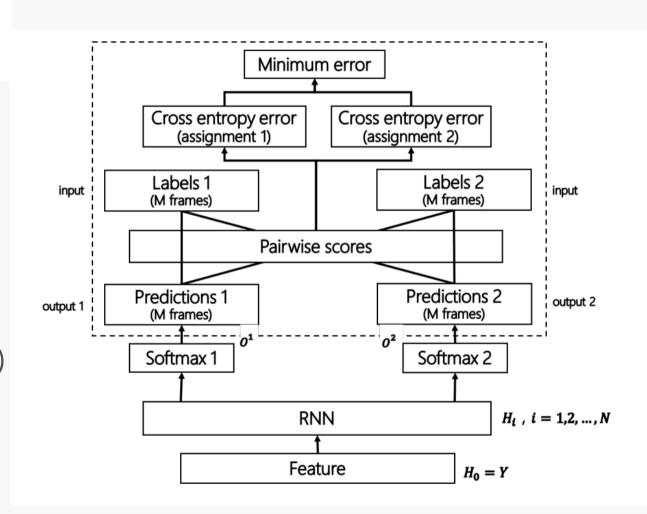
Permutation Invariant Training for ASR

$$P(\mathbf{L}_{u1}, ..., \mathbf{L}_{uN} | \mathbf{O}_u^{(m)}) \approx \prod_{n=1}^N P(\mathbf{L}_{un}^{(r)} | \mathbf{O}_u^{(m)})$$
 (2)

$$\mathcal{J}_{\text{CE-PIT}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
(4)

Disadvantage

- Model Complexity (3 hardest problems)
- Frame CE → Utt. Problem
- No Linguistics



PIT + Transfer Learning (TS)

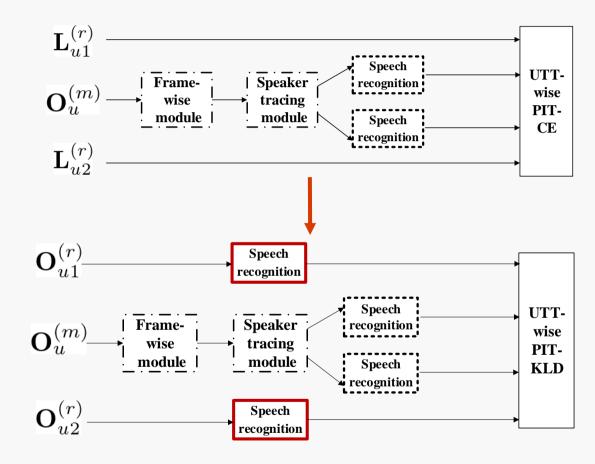
$$\mathcal{J}_{\text{CE-PIT}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \qquad (4)$$

$$\mathcal{J}_{\text{KLD-PIT}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} \qquad (8)$$

$$KLD(P(l_{utn}^{(c)}|\mathbf{O}_{un}^{(r)}), P(l_{utn}^{(s')}|\mathbf{O}_{u}^{(m)}))$$

$$Clean infer. \qquad \text{PIT model infer.}$$

- Better model convergence
 - Domain adaptation v.s. from scratch



[1] Zhehuai Chen, Jasha Droppo, Jinyu Li, Wayne Xiong, Progressive Joint Modeling in Unsupervised Single-channel Overlapped Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 184-196, Jan. 2018. doi: 10.1109/TASLP.2017.2765834.

Linguistics - Multi-outputs Seq. Disc. Training

- Motivation:
 - Both ASR & speaker tracing → sequential
 - Implicit integrating language model
- Formulation:

$$\mathcal{J}_{\text{CE-PIT}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
(4)

- Key challenges:
 - Design the multi-output search space
 - Integrate with label assignment

$$\mathcal{J}_{\text{SEQ-PIT}} = \sum_{u} \min_{s' \in \mathbf{S}} \frac{1}{N} \sum_{n \in [1, N]} - \mathcal{J}_{\text{SEQ}} \mathbf{L}_{un}^{(s')}, \mathbf{L}_{un}^{(r)})$$
(12)

[1] Zhehuai Chen, Jasha Droppo, Jinyu Li, Wayne Xiong, Progressive Joint Modeling in Unsupervised Single-channel Overlapped Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 184-196, Jan. 2018. doi: 10.1109/TASLP.2017.2765834.

Proposed methods

Follow PIT-TS diagram

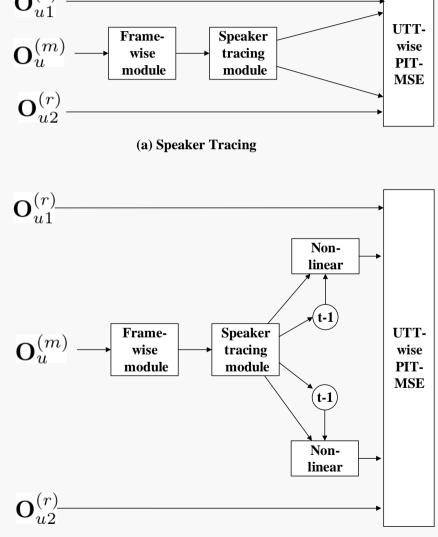
- Motivation
 - improve sequence modeling & language model
- Method
 - Implicit correlation modeling → explicit
 - Integrate linguistic information

Acoustics – Temporal Correlation Modeling

- Motivation
 - Sequential correlation v.s. stream decorrelation
 - the frequency bins between adjacent frames of the same speaker are correlated

Assignment error: e.g. ch-a: how oh you ch-b: are no

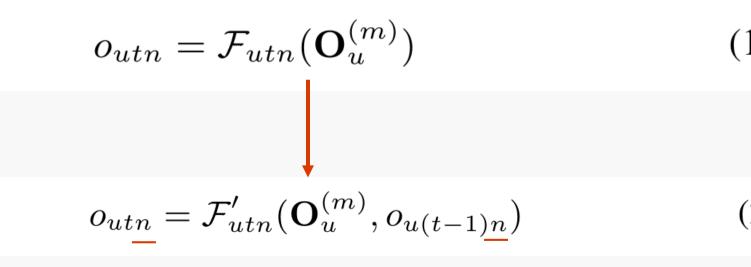
Last inference can improve current inference

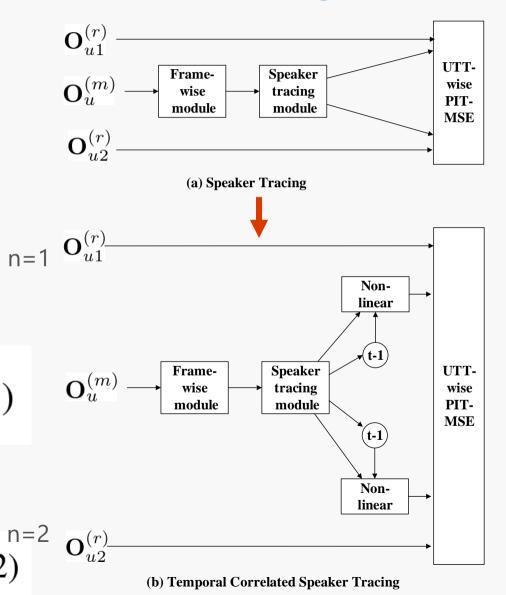


(b) Temporal Correlated Speaker Tracing

Acoustics – Temporal Correlation Modeling

- Motivation
 - Sequential correlation v.s. stream decorrelation
 - Last inference can improve current inference
- Sequential labels correlation

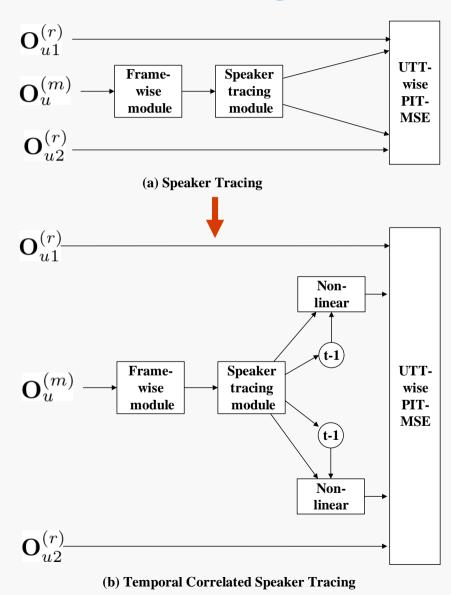




Acoustics – Temporal Correlation Modeling

- Motivation
 - Sequential correlation v.s. stream decorrelation
 - last inference can improve current inference
- Sequential labels correlation
- alleviates the assignment & cross talk errors

Assignment error: e.g. ch-a: how oh you ch-b: are no



- Motivation:
 - Improve assignment decision by combining LM in training stage
 - Still train a pure acoustic model and integrate it with more powerful word level language model in evaluation stage
- Original PIT-CE

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
(3)

- Motivation:
 - Improve assignment decision by combining LM in training stage
 - Still train a pure acoustic model and integrate it with more powerful word level language model in evaluation stage
- Original PIT-CE

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
• PIT-MAP:

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | \mathbf{O}_u^{(m)}) / P(l) \cdot P(l_{utn}^{(r)} | \mathbf{L}_{u(t-1)n}^{(s')})}{P(\mathbf{O}_u^{(m)})}$$

Discriminative training

$$\approx \frac{P(l_{utn}^{(r)}|\mathbf{O}_{u}^{(m)})}{P(l)} \cdot \left(P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})\right)^{\lambda}$$

Proposed method

- Motivation:
 - Improve assignment decision by combining LM in training stage
 - Still train a pure acoustic model and integrate it with more powerful word level language model in evaluation stage
- Original PIT-CE

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
(3)

• Proposed:

$$CE(\cdot)$$
 PIT-trained AM

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \underbrace{\frac{P(l_{utn}^{(r)} | \mathbf{O}_u^{(m)})}{P(l)}}_{P(l)} (P(l_{utn}^{(r)} | \mathbf{L}_{u(t-1)n}^{(s')}))^{\lambda}$$
(4)

- Motivation:
 - Improve assignment decision by combining LM in training stage
 - Still train a pure acoustic model and integrate it with more powerful word level language model in evaluation stage
- Original PIT-CE

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
(3)

• Proposed:

$$E(\cdot) \longrightarrow MAP(\cdot)$$
 Senone level NNLM

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | \mathbf{O}_u^{(m)})}{P(l)} \cdot \left(P(l_{ut\underline{n}}^{(r)} | \mathbf{L}_{u(t-1)\underline{n}}^{(s')}) \right)^{\lambda}$$
(4)

- Motivation:
 - Improve assignment decision by combining LM in training stage
 - Still train a pure acoustic model and integrate it with more powerful word level language model in evaluation stage
- Original PIT-CE

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum \min_{s' \in \mathbf{S}} \sum_{i} \frac{1}{N} \sum_{lutn} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$
• Proposed:
$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | \mathbf{O}_{u}^{(m)}) / P(l) \cdot P(l_{utn}^{(r)} | \mathbf{L}_{u(t-1)n}^{(s')})}{P(\mathbf{O}_{u}^{(m)})}$$

$$\approx \frac{P(l_{utn}^{(r)}|\mathbf{O}_{u}^{(m)})}{P(l)} \cdot \left(P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})\right)^{\lambda}$$

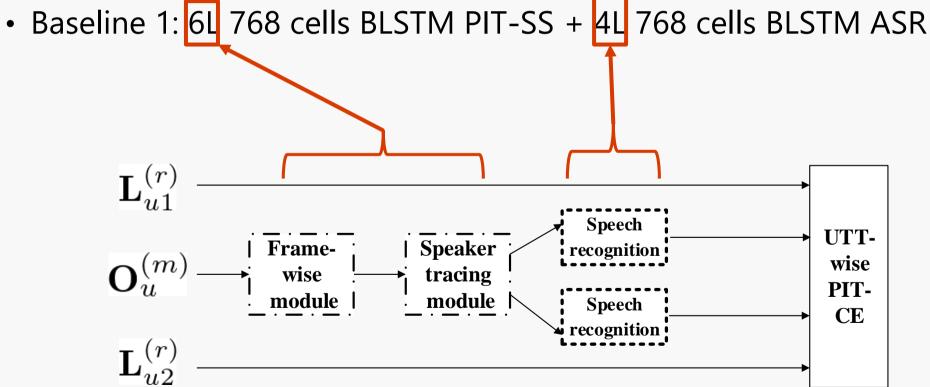
system Opt. Ass. CE **Proposed MAP** CE Disc. Train MAP MAP

Discriminative training

Proposed method

Experiments

- Setup and baselines:
 - Artificial overlapped SWBD 300 \rightarrow 150 (\rightarrow 50); hub5e-swb 1831 \rightarrow 915 utts
 - 9000 senones; clean speech alignment;
 Baseline 1: 61 768 cells BLSTM PIT-SS + 41 768 cells BLSTM ASE



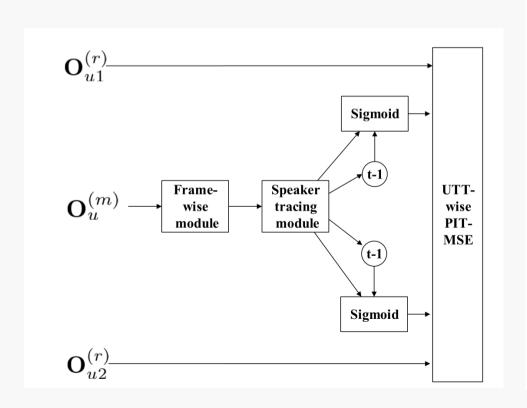
Experiments

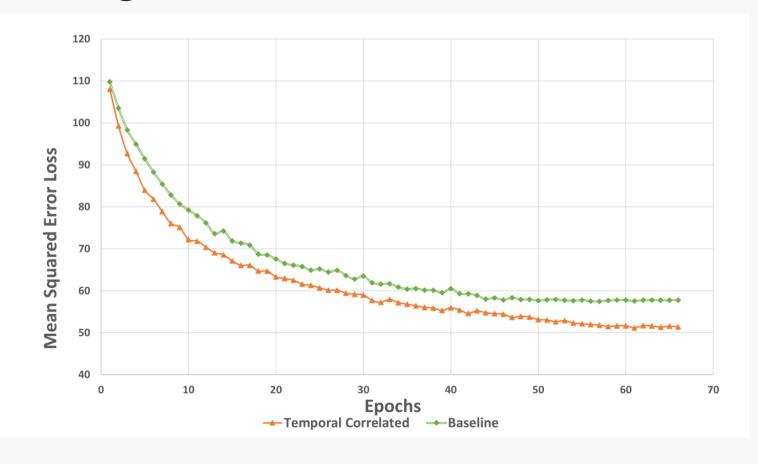
- Setup and baselines:
 - Artificial overlapped SWBD 300 \rightarrow 150 (\rightarrow 50); hub5e-swb 1831 \rightarrow 915 utts
 - 9000 senones; clean speech alignment;
 - Baseline 1: 6L 768 cells BLSTM PIT-SS + 4L 768 cells BLSTM ASR
 - Baseline 2: + transfer learning (TS, taught by clean teacher)

Neural network	Model	WER
	PIT-ASR	57.5
6 BLSTM + 4 BLSTM	progressive joint training + clean teacher	38.9

Experiments – Temporal Correlated

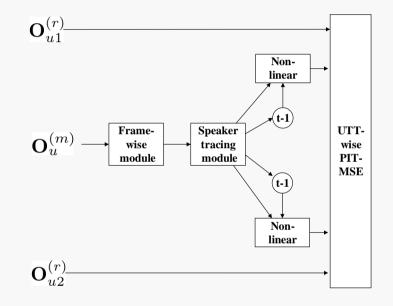
- Baseline: modularization + clean teacher WER=38.9
- Improve in Speaker Tracing:





Experiments – Temporal Correlated

- Baseline: modularization + clean teacher WER=38.9
- Improve in Speaker Tracing
- WER improve after joint training



Temporal Correlated	# of Sigmoid	WER	Rel. (%)
×	0	38.9	0
	0	37.5	-3.6
$\sqrt{}$	1	35.8	-8.0
	2	36.7	-5.7

Experiments – LM Integration

• Baseline: modularization + clean teacher WER=38.9

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$

$$CE(\cdot) \longrightarrow MAP(\cdot)$$
(3)

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | \mathbf{O}_u^{(m)})}{P(l)} \cdot (P(l_{utn}^{(r)} | \mathbf{L}_{u(t-1)n}^{(s')}))^{\lambda}$$
(4)

		50 hours		150	150 hours	
Assign.	Opt.	WER	Rel. (%)) WER	Rel. (%)	
CE	CE	38.9	0	32.8	0	
MAP	CE	37.3	-4.1	30.9	-5.8	

Experiments – LM Integration

• Baseline: modularization + clean teacher WER=32.8

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | \mathbf{O}_u^{(m)})}{P(l)} \cdot (P(l_{utn}^{(r)} | \mathbf{L}_{u(t-1)n}^{(s')}))^{\lambda}$$
(4)

			50 hours) hours
Assign.	Opt.	WER	Rel. (%)	WER	Rel. (%)
CE	CE	38.9	0	32.8	0
MAP	CE	37.3	-4.1	30.9	-5.8

- with more data, the improvement becomes larger
 - AM becomes stronger
 - Assignment decision is not over-fit to the LM

Experiments – Compare with disc. training

			50 hours	
system	Assign.	Opt.	WER	Rel. (%)
baseline	CE	CE	38.9	0
LM integration	MAP	CE	37.3	-4.1
LF-DC-bMMI	MAP	MAP	35.6	-8.5

Discriminative training

- Differences:
 - optimization stage
- $MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(\mathbf{O}^{(m)}|\mathbf{L}^{(r)}) \cdot P(\mathbf{L}^{(r)})}{P(\mathbf{O}^{(m)})}$ $= \frac{P(l_{utn}^{(r)}|\mathbf{O}_{u}^{(m)})/P(l) \cdot P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})}{P(\mathbf{O}_{u}^{(m)})}$ $\approx \frac{P(l_{utn}^{(r)}|\mathbf{O}_{u}^{(m)})}{P(l)} \cdot (P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')}))^{\lambda}$
- NNLM v.s. N-gram in discriminative training
- hardness in modeling $P(\mathbf{O}_u^{(m)})$

Proposed method

Experiments – Combination

Method	WER	Rel. (%)
baseline	38.9	0
+ Temporal Correlated	35.8	-8.0
+ LM Integration	34.4	-11.5
+ LF-DC-bMMI	31.6	-18.8

• Operate in different levels -> can be combined

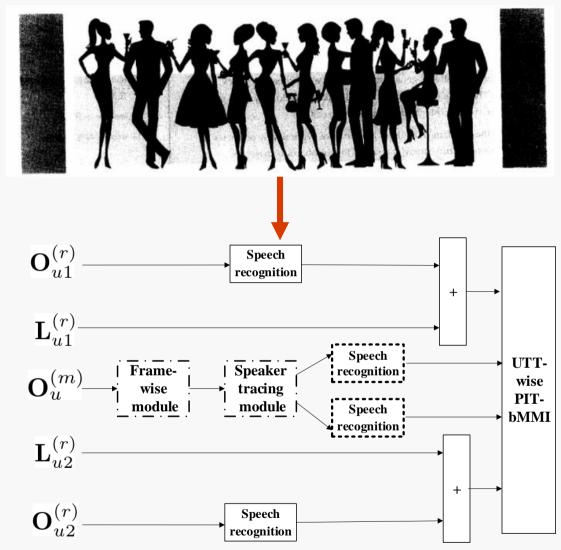
Experiments – Combination

Method	WER	Rel. (%)
baseline	38.9	0
+ Temporal Correlated	35.8	-8.0
+ LM Integration	34.4	-11.5
+ LF-DC-bMMI	31.6	-18.8
+ MMI clean teacher	35.8	-8.0
+ LF-DC-bMMI	35.2	-9.5

- Operate in different levels > can be combined
- Better than only utilize TS + discriminative training

Our final system

- Acoustics
 - Modular Initialization 4%
 - CNN 10%
 - Transfer Learning Based Joint Training 20%
 - Temporal Correlation Modeling 8%
- Linguistics
 - Multi-outputs Sequence Discriminative Training 8%
 - Integrating Language Model in Assignment Decision 4%

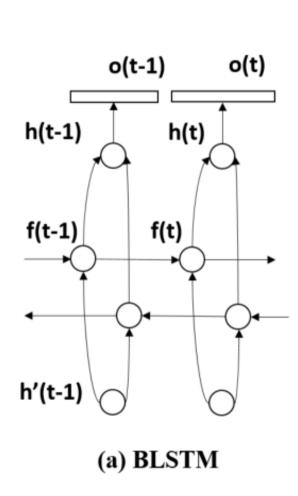


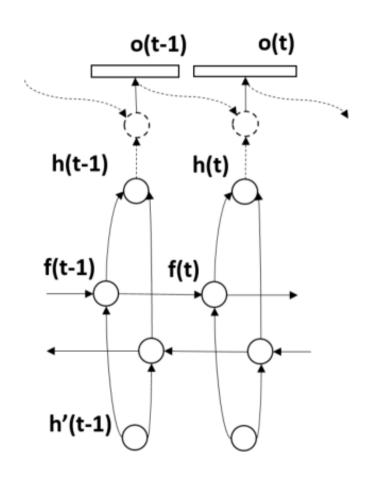
[1] Zhehuai Chen, Jasha Droppo, Sequence Modeling in Unsupervised Single-channel Overlapped Speech Recognition, accepted by IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Calgary, Canada, 2018.

[2] Zhehuai Chen, Jasha Droppo, Jinyu Li, Wayne Xiong, Progressive Joint Modeling in Unsupervised Single-channel Overlapped Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 184-196, Jan. 2018. doi: 10.1109/TASLP.2017.2765834.

Backup materials

Temporal correlation modeling in BLSTM





(b) Temporal Correlated BLSTM

Experiments – Example 50hrs (F-F)

- Clean ASR (90+WER)
- 1 PIT-CE
- 2 Transf.
- 3 +MMI teacher
- 4 +seq. disc. tr.



```
1 id: (sw 4776 a-006)
Labels: <>
  File: sw_4776
  Channel: a
  REF: i just ** **** ** **** GO WALKING AND I USUALLY TRY TO go for about an hour (e-) and i HAVE TRIED DOING IT
                                                                                                               EVERY DAY BUT i mean some times I do NOT EVEN DO
                                                                                                                                                              IT ONCE DURING THE WEEK but you know MOST OF THE TIME I T
                                                      1 PIT-CE
 Scores: (#C #S #D #I) 18 22 11 4
                                                      AS WE go for about an hour
  HYP: i just SO LONG AS THEY HAVE COME
                                                                                                                   ANY WAY i mean some times * do *** IT LONGER THAN A
             II III S S
                                        S S S
                                                                                                                                            D D S S
                                                                                                                                                               5 5 5
                                                       2 Transf.
12 Scores: (#C #S #D #I) 34 7 10 1
                           and i usually *** HAVE go for about an hour
                                                                       *** i **** tried ***** IN every day but i mean some times i do not even do it once during the week but you know **** ** *** *** * *** ONCE WE TURN DEFINITE
13 HYP: i just go OUT YOU
                                                                                                                                                                            D D D D D D S S S S
                                                                  +MMI teacher
16 Scores: (#C #S #D #I) 36 9 6 2
                                        LIKE to HAVE go for about an hour
17 HYP: i just go OUT HERE and i HAD
                                                                           *** i **** ***** AM TRYING every day but i mean some times i do not even do it once during the week but you know **** ** *** WHAT REALLY TRYING to get
                T S
                                                                           D D D S S
                                                                                                                                                                                   D D D S S
20 Scores: (#C #S #D #I) 37 9 5 1
1 HYP: i just go OUT HERE
                          and i usually HAVE to go for about an hour
                                                                      and i **** tried **** IN ANOTHER day but i mean some times i do not even do it once during the week but you know *** ** WE ARE REALLY TRYING to get *** IT
22 Eval:
                                        S
                                                                                     D 5 5
                                                                                                                                                                            D DSSS
24 id: (sw 4854 b-009)
25 Labels: <>
26 File: sw_4854
27 Channel: b
29 REF: well I (don-) i DO NOT CAMP NEAR AS MUCH AS WE USED TO WE
                                                              USED TO
                                                                         go *** TO THE LAKE all the time (%hesitation) WE (USE-) WE ALL SKI and everything *** WE USED TO ALL GO TO THE LAKE ALL the TIME AND MY PARENTS H.
31 Scores: (#C #S #D #I) 20 22 20 2
                    i ** *** **** ** ** ** ** ** ** ** ** THINK AND HERE'S go FOR ABOUT AN HOUR all the time
                                                                                                                    I WOULD TRADE OFF NOW and everything BUT I REALLY DO NOT I MEAN I AM
33 Eval:
  Scores: (#C #S #D #I) 42 9 11 0
                                                                                                                    we all *** SCAN everything ** **** ** THESE DOGS OVER LIKE all the time and my parents had a cabin IN the
                    i ** *** CAMPED near as much as we used to we **** ** ** FEEL LIKE all the time
                                                                                                                                            D D D D S S S S
37 Eval:
                      D D S
                                                             D D D D S S
                                                                                                                          D S
9 Scores: (#C #S #D #I) 42 10 10 0
                    i ** THINK CAMPED near as much as we used to we **** ** ** FEEL LIKE all the time
                                                                                                                      we all SCREAM and everything ** **** ** YOU STILL HAVE LIKE all the time and my parents had a cabin IN
                     D S S
                                                               D D D D S S
                                                                                                                                                D D D D S S S S
41 Eval:
         D
43 Scores: (#C #S #D #I) 47 8 7 0
  HYP: well *
                    i ** THINK CAMPED near as much as we used to we used to go to *** lake all the time
                                                                                                                     we all SCREAM and everything ** **** ** THESE DOGS BUT LIKE all the time and my parents had a cabin IN
45 Eval:
                      D S S
                                                                                                                                               D D D D S S S S
```

Experiments – Example 150hrs (F-F)

- Clean ASR (90+WER)
- 1 PIT-CE
- 2 Transf.
- 3 +CNN
- 4 +seq. disc. tr.



```
id: (sw 4776 a-006)
 2 Labels: <>
  File: sw 4776
 Channel: a
  REF: i just go WALKING and i usually TRY TO
                                                 go for about an hour (e-) AND i have tried DOING IT EVERY DAY but i mean some times i do not even do it once during the week but you know MOST OF THE TIME i TRY to get out there
                                                                   1 PIT-CE
8 Scores: (#C #S #D #I) 38 5 8 0
   HYP: i just go ON
                         and i usually *** THOUSAND go for about an hour
                                                                         *** i have tried ***** ** TO NAME but i mean some times i do not even do it once during the week but you know **** ** *** i LIKE to get out there
                                     D S
                                                                                                                                                                                D D D D S
                                                                                  2 Transf.
12 Scores: (#C #S #D #I) 39 6 6 2
                                                                            *** i WOULD TRY doing it every day but i mean some times i do not even do it once during the week *** you know **** ** *** WE WE try to get out there
HYP: i just go OUT TO EAT
                               and i usually *** HAVE go for about an hour
                                           D S
                                                                                                                                                                                   D D D S S
16 Scores: (#C #S #D #I) 42 5 4 1
                                                                         and i **** tried doing it every day but i mean some times i do not even do it once during the week but you know **** ONCE the **** SAME TRYING to get out then
17 HYP: i just go OUT AND
                            and i usually *** HAVE go for about an hour
                                                                                                                                                                                            D 5 5
                 T S
                                         D S
                                                                                                            +seq. disc. tr.
20 Scores: (#C #S #D #I) 41 8 2 0
21 HYP: i just go WATCHING and i usually DO NOT go for about an hour
                                                                     and i **** tried doing it every day but i mean some times i do not even do it once during the week but you know **** ONCE IN SOME WAY TRYING to get out there
22 Eval:
                                      5 5
                                                                           D
                                                                                                                                                                           D S S S S S
24 id: (sw 4854 b-009)
25 Labels: <>
26 File: sw 4854
27 Channel: b
29 REF: well I (don-) i DO NOT camp NEAR as much as we used to we used TO GO to THE LAKE all the time (%hesitation) we (use-) we all SKI AND everything WE USED TO ALL GO TO THE LAKE ALL THE time AND my parents HAD a CABIN AT THE
31 Scores: (#C #S #D #I) 36 10 16 0
                                                                                                                       we all STAY IN everything ** **** ** *** ** ** ** YOU STILL ILLEGAL time *** mv parents GOT a ***** **
                     i ** THINK camp HERE as much as we used to we used ** ** to *** LEAVE all the time
                                                                                                                             5 5
                                                                                                                                               D D D D D D D S S S
33 Eval:
                                                                                                                                                                                            D
35 Scores: (#C #S #D #I) 42 10 10 0
                     i AM A camp near as much as we used to we **** ** ** FEEL LIKE all the time
                                                                                                                     we all *** SCAN everything ** **** ** THESE DOGS ARE LIKE all the time and my parents had a cabin IN the l
                                                                                                                           D S
                                                                                                                                             37 Eval:
                      5 5
                                                             D D D D S S
                                                                                                                                                                                                                  S
39 Scores: (#C #S #D #I) 40 9 13 0
                     i ** THINK camp near as much as we used to we **** ** ** FEEL LIKE all the time
                                                                                                                       we *** WILL SCAN everything ** **** ** *** ** THESE DOGS LIKE all the time and my parents had a cabin ON the
                                                               D D D D S S
                                                                                                                         D S S
                                                                                                                                                D D D D D S S S
                                                                                                                                                                                                                    5
41 Eval:
43 Scores: (#C #S #D #I) 46 9 7 0
44 HYP: well *
                    i ** THINK CAMPED near as much as we used to we used to go to *** lake all the time
                                                                                                                        we all *** SCAN everything ** **** ** WITH THESE DOGS OVER LIKE all the time and my parents had a cabin IN
                                                                                                                                                D D D S S S S S
45 Eval:
                      D S S
```