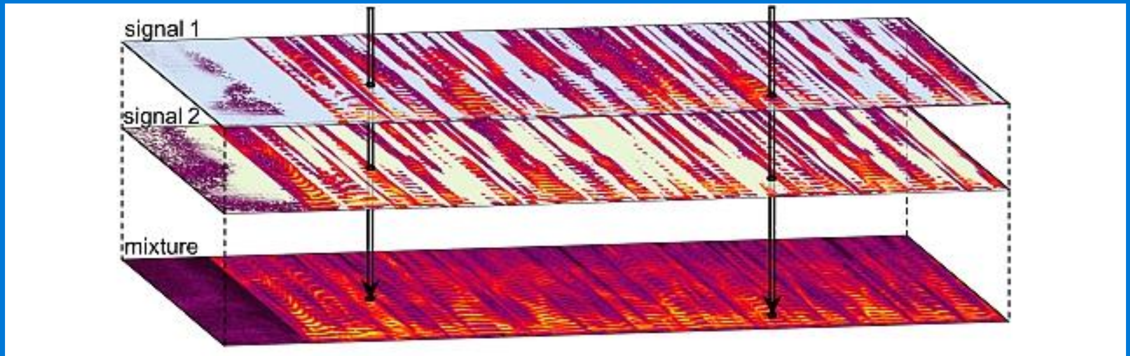
Advances in Cocktail-party Problem

Unsupervised Single-channel Overlapped Speech Recognition

Zhehuai (Tom) Chen
Research Intern
chenzhehuai@sjtu.edu.cn

Mentor: Jasha Droppo



Microsoft

SJTU SPEECH LAB
上海交通大學智能語音實驗室

Outline

- Introduction
  - Cocktail-party problem
  - Permutation Invariant Training (baseline)
- Acoustics
  - Modular Initialization
  - Transfer Learning Based Joint Training
  - Temporal Correlation Modeling
- Linguistics
  - Multi-outputs Sequence Discriminative Training
  - Integrating Language Model in Assignment Decision
- Experiments
- Conclusion & Future Directions

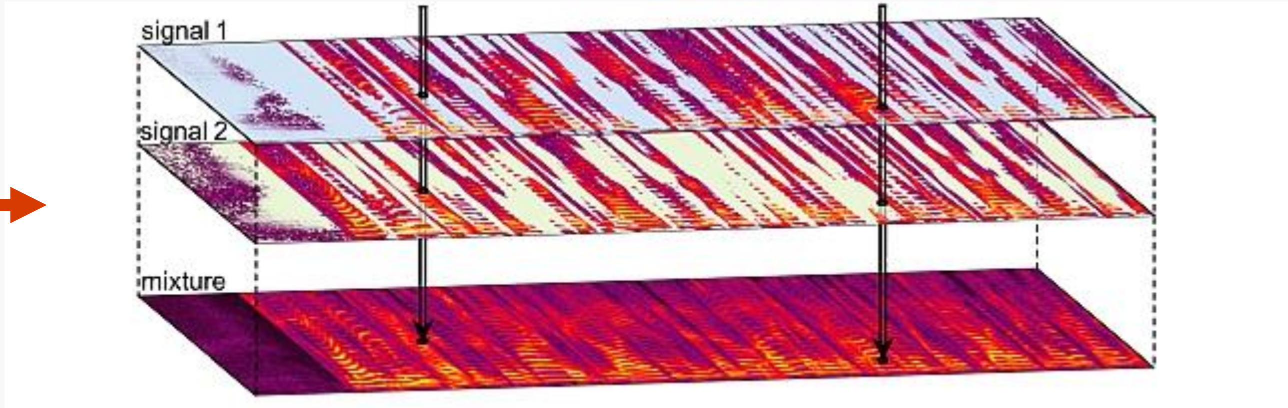
Introduction

- Cocktail-party problem

N=2

$$P(\mathbf{L}_{u1}, \ldots, \mathbf{L}_{uN} | \mathbf{O}_u^{(m)})$$

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



Assignment error:

e.g. ch-a: how oh you

ch-b: are no

Cross talk error:

e.g. ch-a: how are you

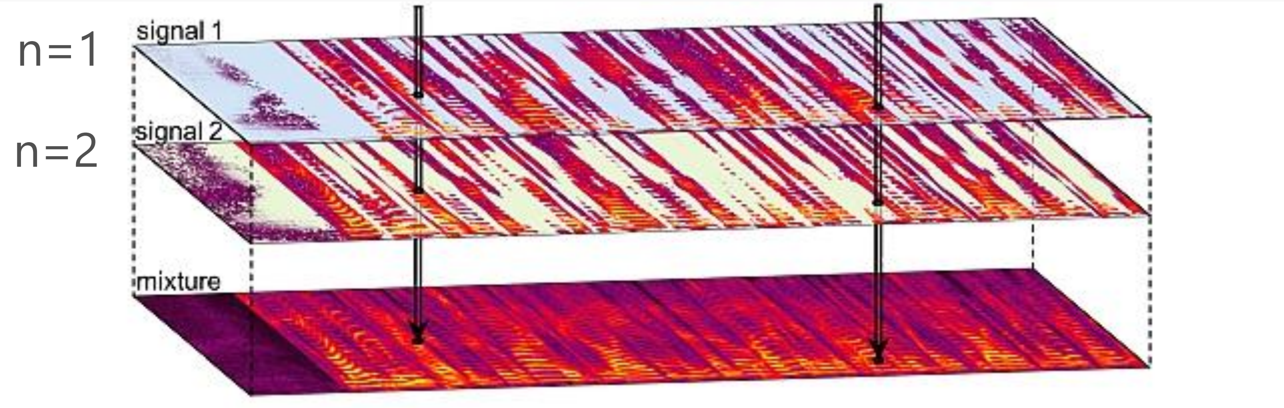
ch-b: oh are no

Label assignment problem

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

Label Independence

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



Introduction

$$P(\mathbf{L}_{u1}, \ldots, \mathbf{L}_{uN} | \mathbf{O}_u^{(m)}) \approx \prod_{n=1}^{N} P(\mathbf{L}_{un}^{(r)} | \boxed{\mathbf{O}_{un}^{(r)}}) \qquad (3)$$

Feature Independence

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

Label Independence

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

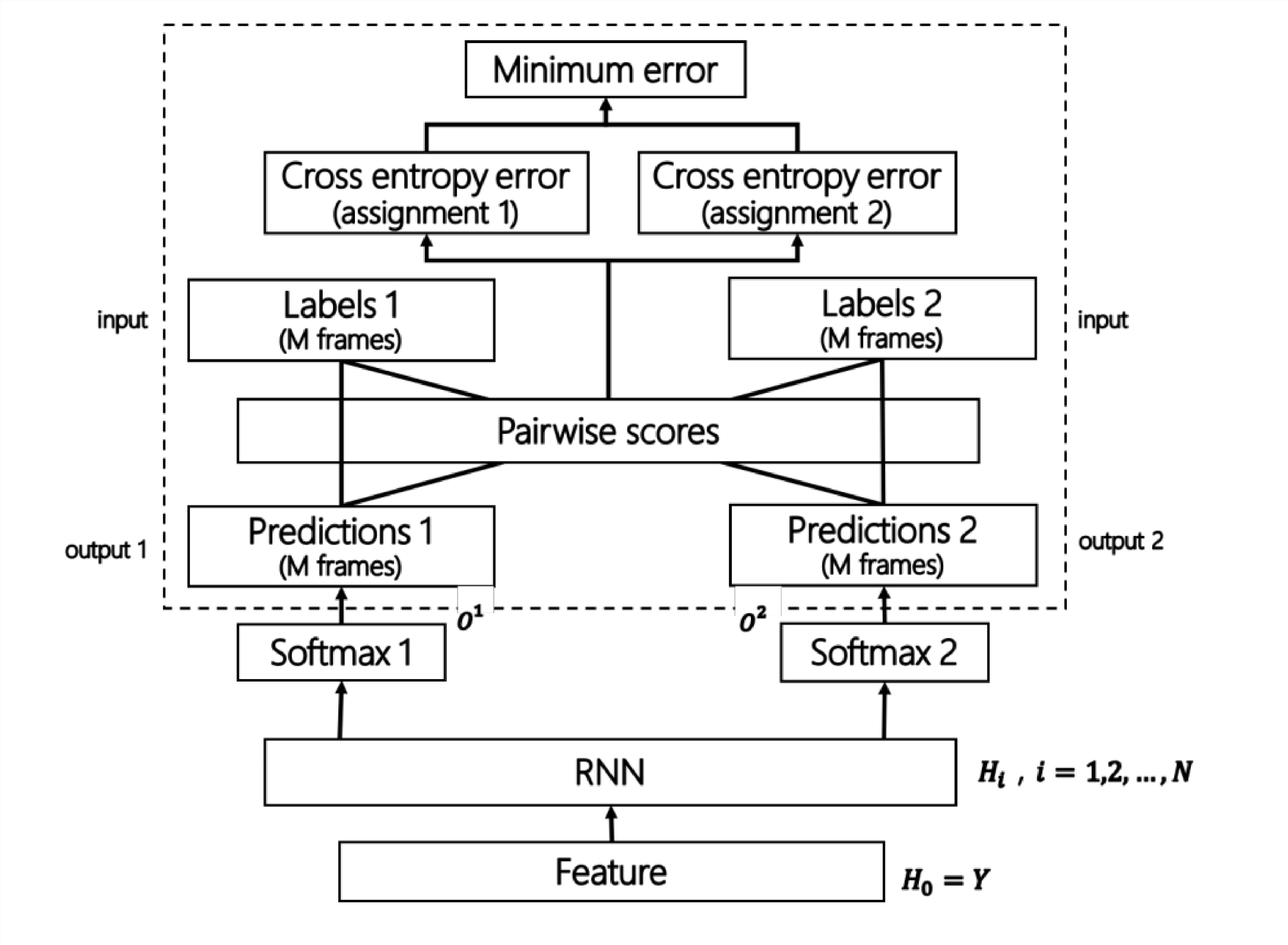
Introduction

- Cocktail-party problem
  - Label Independence
    - Label assignment problem (hard)
  - Feature Independence
    - Speech separation & recognition are independent (bad)

Introduction

- Cocktail-party problem
  - Speech Separation + Speech-to-text **(feature independence)**
    - Before deep learning: Computational Auditory Scene Analysis (CASA)
    - Deep learning based: Deep Clustering (DPCL)
      - NN to produce spectrogram embedding of separated speech
    - Permutation Invariant Training for Speech Separation
  - Joint Modeling **(label independence)**
    - Permutation Invariant Training for ASR

Permutation Invariant Training for ASR

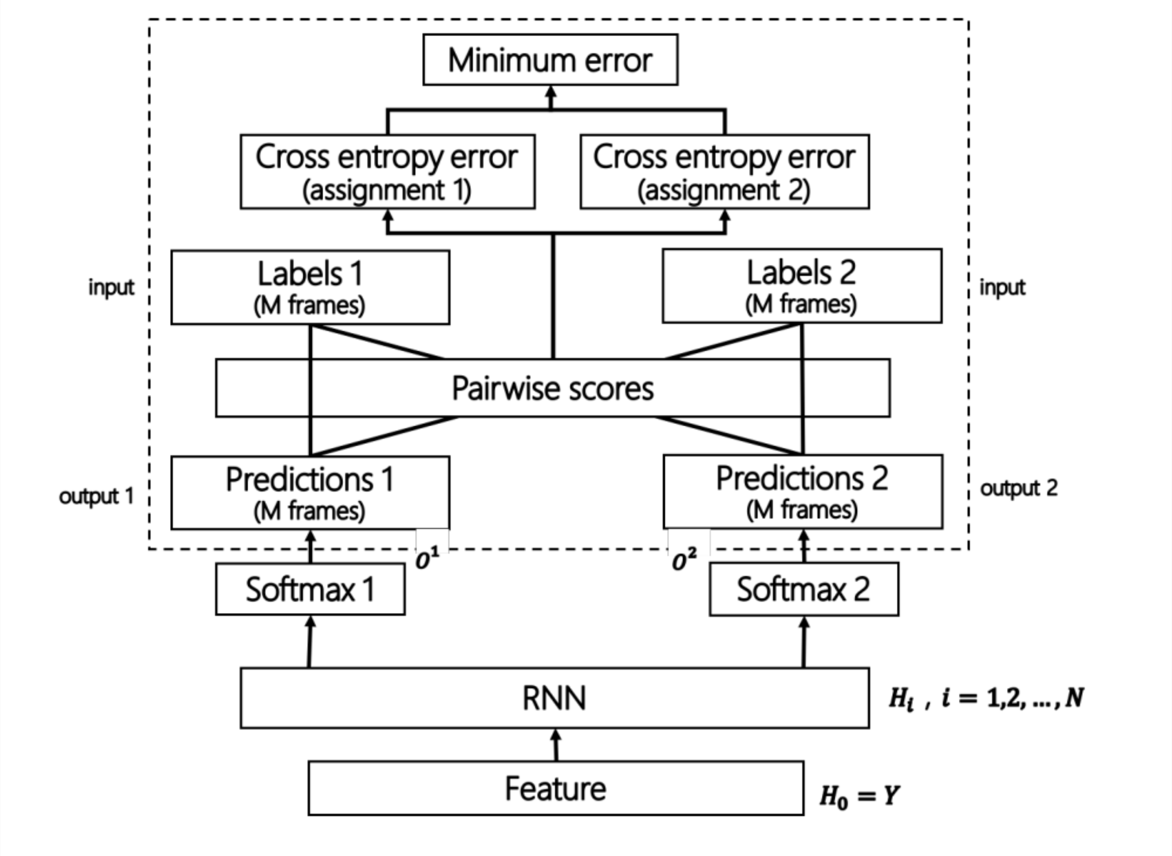


Permutation Invariant Training for ASR

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

$$\mathcal{J}_{\text{CE-PIT}} = \sum_u \boxed{\min_{s' \in \mathbf{S}}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \qquad (4)$$

- Disadvantage
  - Model Complexity (3 hardest problems)
  - Frame CE → Utt. Problem
  - No Linguistics
- Result
  - WER 50+% → still far road



We propose:

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

Acoustics - Modular Initialization

- Frame-wise **interpreting** (swapped segments)
  - Local feature extraction → CNN
- Speaker **Tracing** (no swap)
  - Temporal modeling → RNN
- Speech-to-text

$$\mathcal{J}_{\text{F-PIT}} = \sum_u \sum_t \frac{1}{N} \min_{s' \in \mathbf{S}} \sum_{n \in [1,N]} MSE(o_{utn}^{(s')}, o_{utn}^{(r)}) \quad (5)$$

$$\mathcal{J}_{\text{U-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} MSE(o_{utn}^{(s')}, o_{utn}^{(r)}) \quad (6)$$

Acoustics - Modular Initialization

- Speech-to-text (details)
  - Clean speech Force-alignment by seed GMM-HMM → triphone state for each fr.
  - Train NN with the state alignment → 9000 senone (clustered triphone state)
  - ASR Decoding

NN Acoustic model

$$P(\mathbf{W}|\mathbf{O}) = \max_{\mathbf{W}} P(\mathbf{W})P(\mathbf{s}|\mathbf{W}) \cdot P(s_0)\prod_{t=1}^{T} \frac{P(s_t|\mathbf{O})}{P(s)} \cdot P(s_t|s_{t-1})$$

State prior

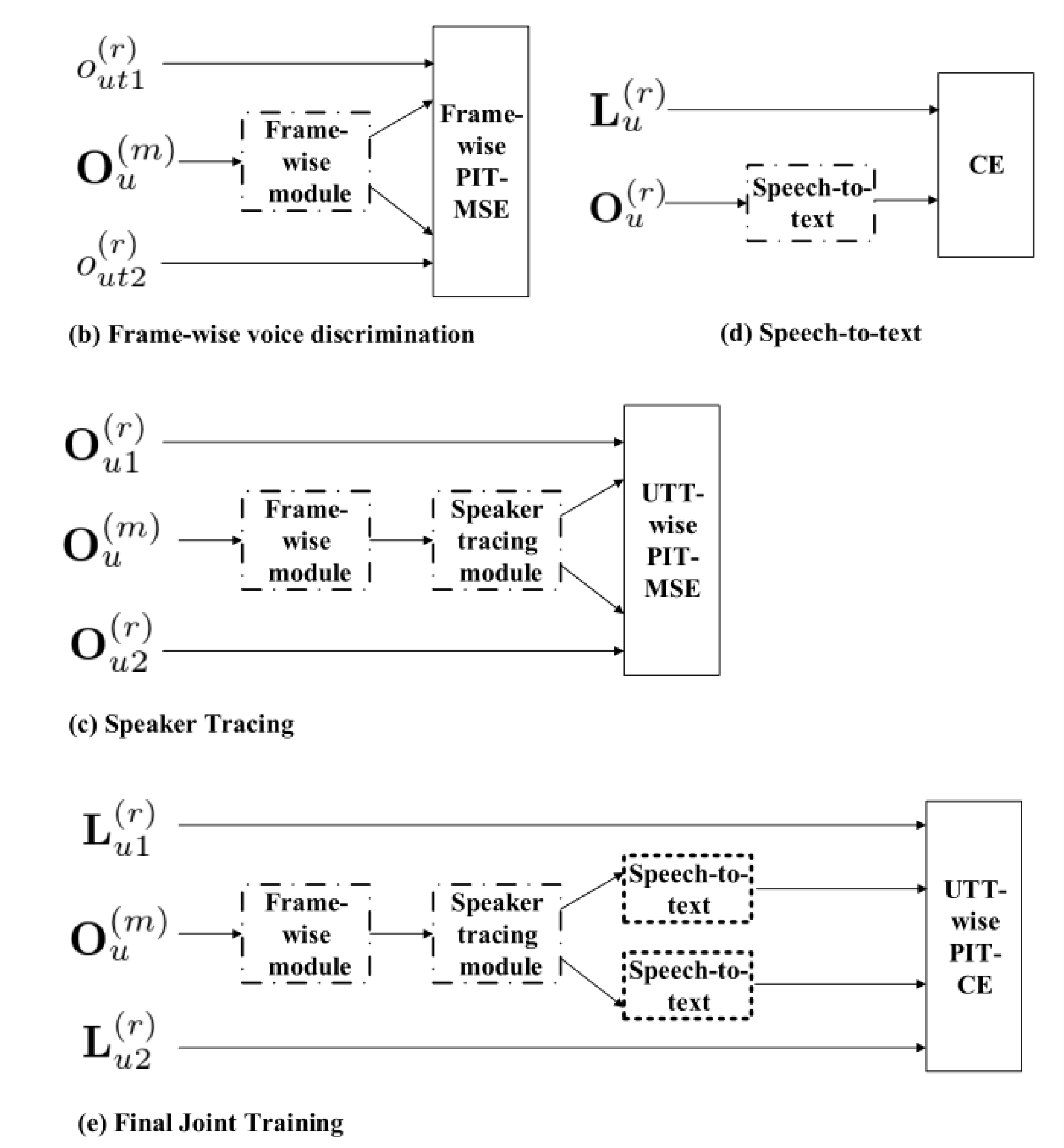
Lexicon

HMM

Language model

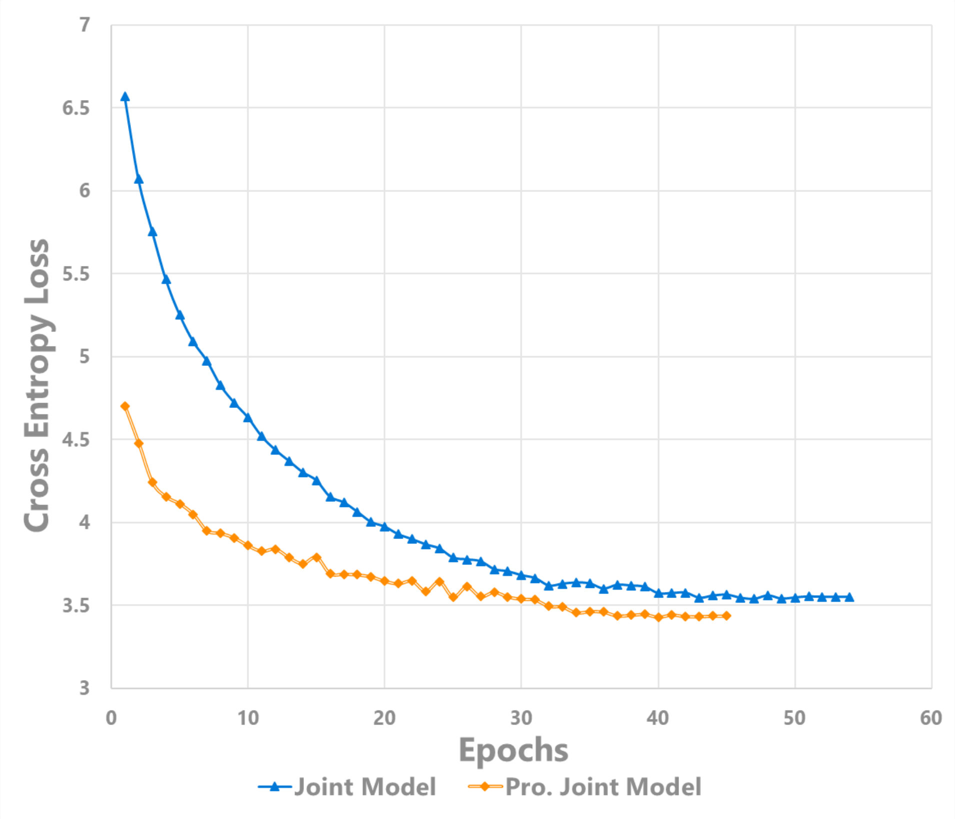
Acoustics - Modular Initialization

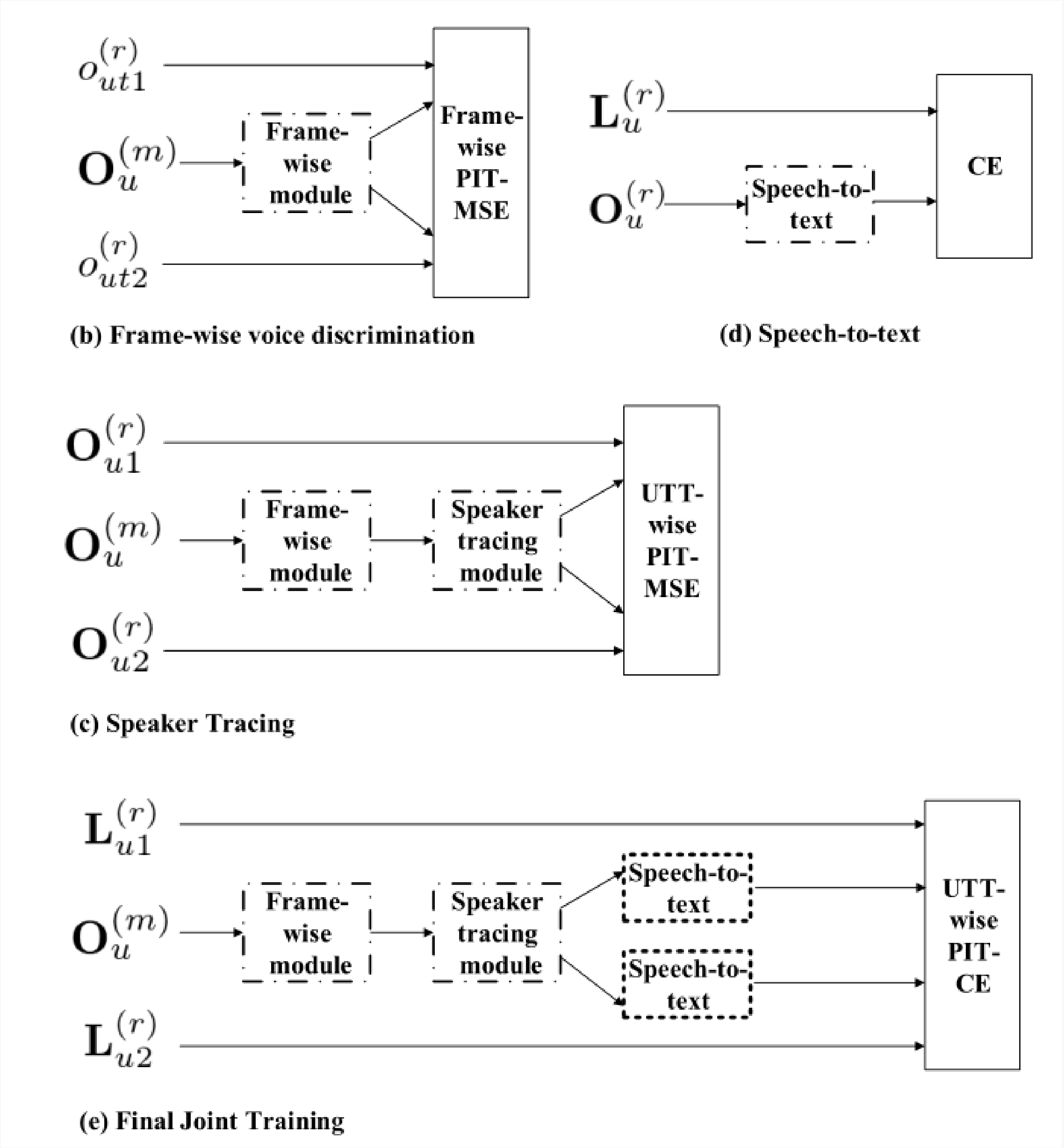
- Progressive joint training
  - Curriculum learning theory
  - The harder task, the larger NN (stacking)



Acoustics - Modular Initialization

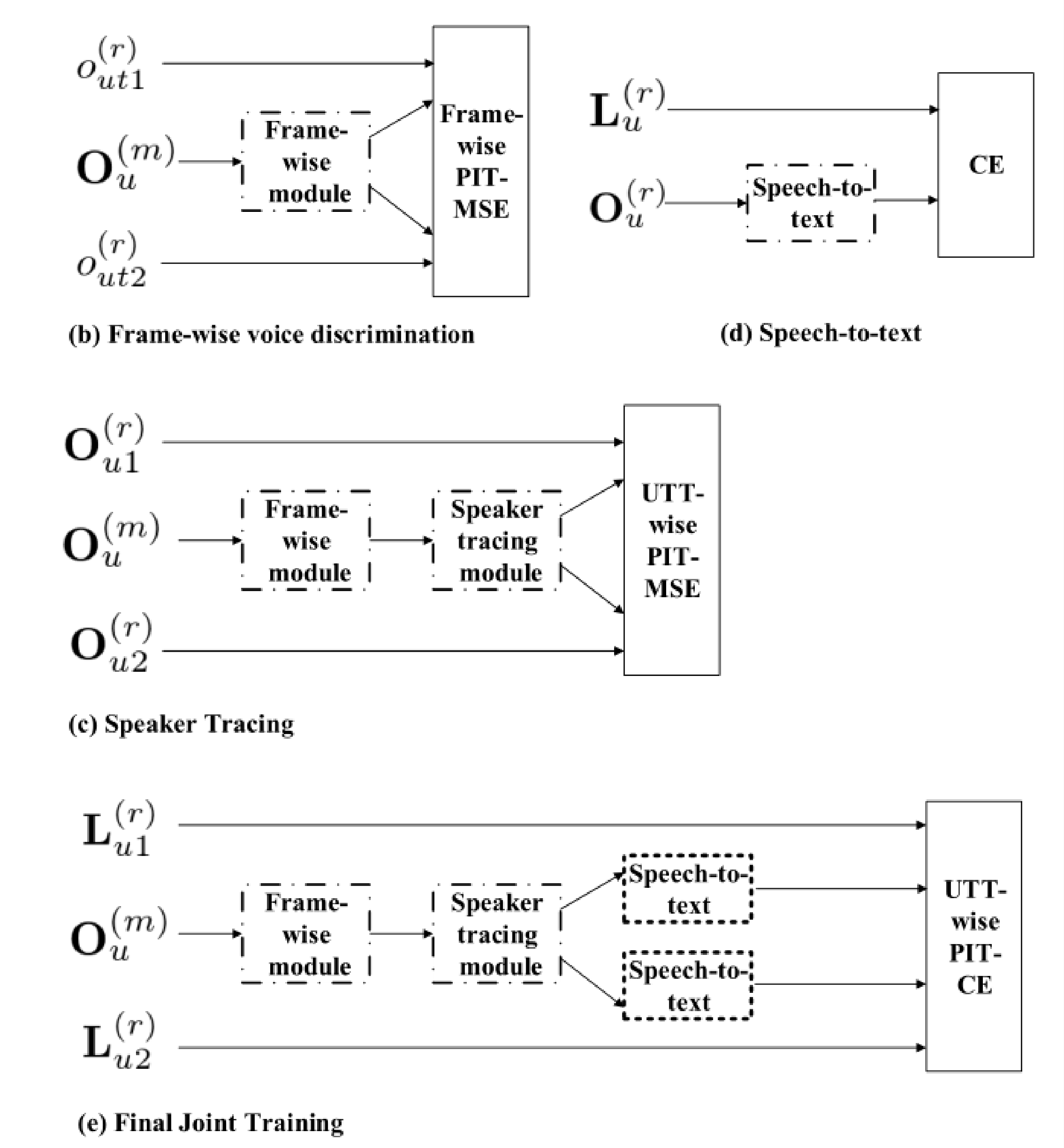
- Less Model Complexity
  - Speed of convergence
  - Better local minima





Acoustics - Modular Initialization

- Less Model Complexity
  - Speed of convergence
  - Better local minima
- Data Efficiency
- Combine with other tech.
  - Sequence disc. training on speech-to-text
  - Integrate LM

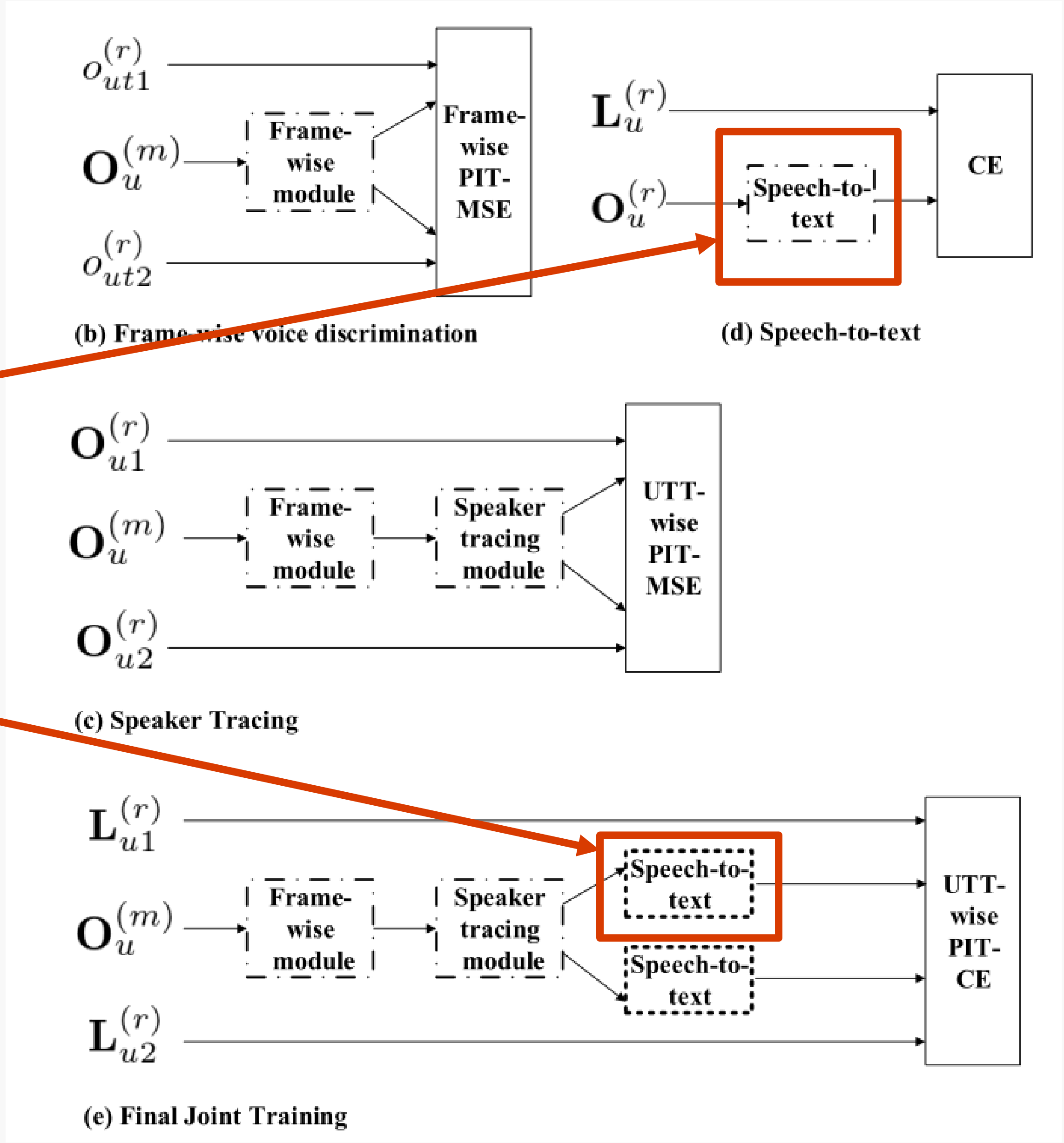


We propose:

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

Acoustics - Transfer Learning x Joint Training

- Transfer Learning
  - To solve the **distribution mismatch** problem in feature space
  - Model → speech-to-text module
  - Source domain → clean speech
  - Target domain → output of speaker tracing module (enhanced feat.)



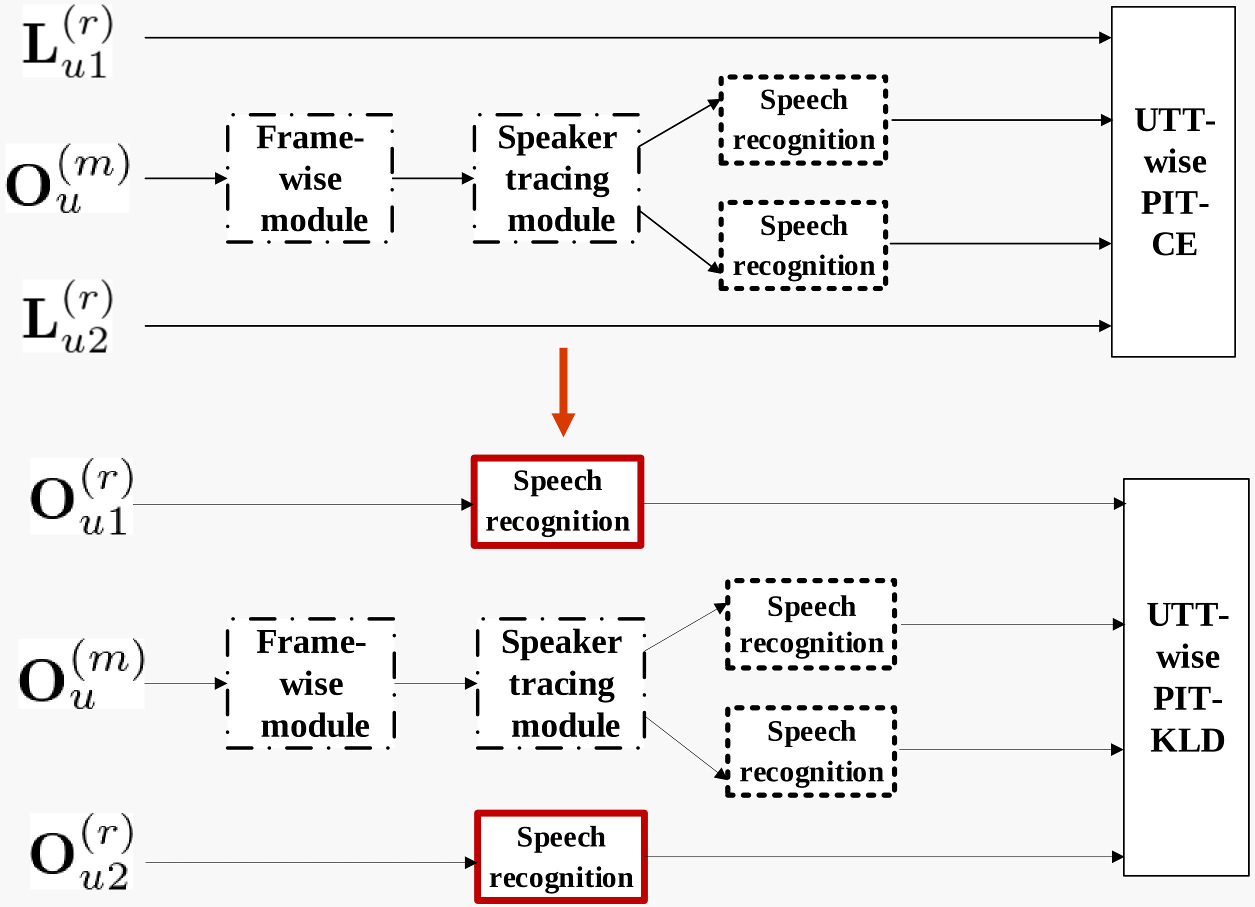
Acoustics - Transfer Learning x Joint Training

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

$$\mathcal{J}_{\text{KLD-PIT}} = \sum_{u} \min_{s' \in \mathbf{S}} \sum_{t} \frac{1}{N} \sum_{n \in [1,N]} \boxed{KLD}(P(l_{utn}^{(c)}|\mathbf{O}_{un}^{(r)}), P(l_{utn}^{(s')}|\mathbf{O}_{u}^{(m)})) \quad (8)$$

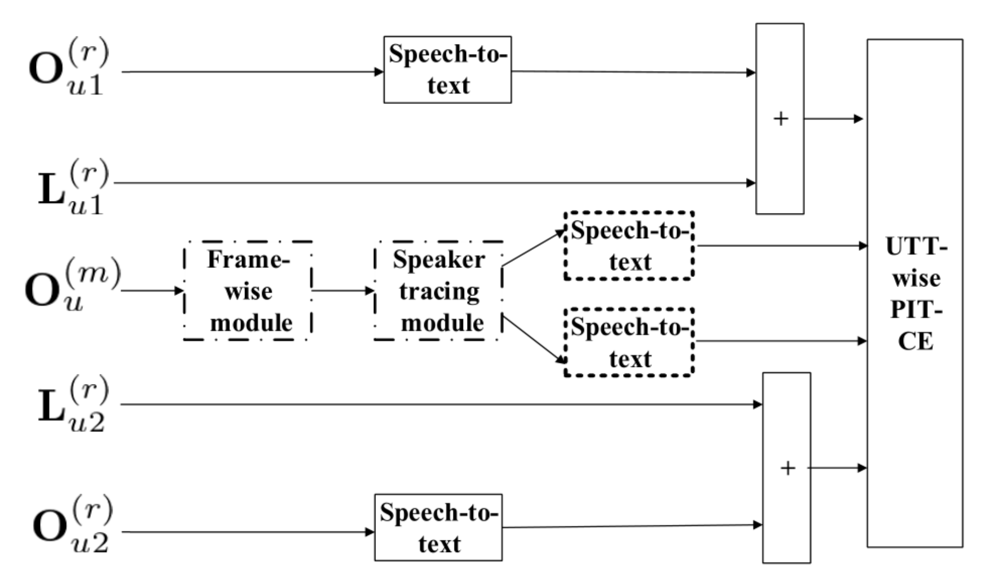
Clean infer.

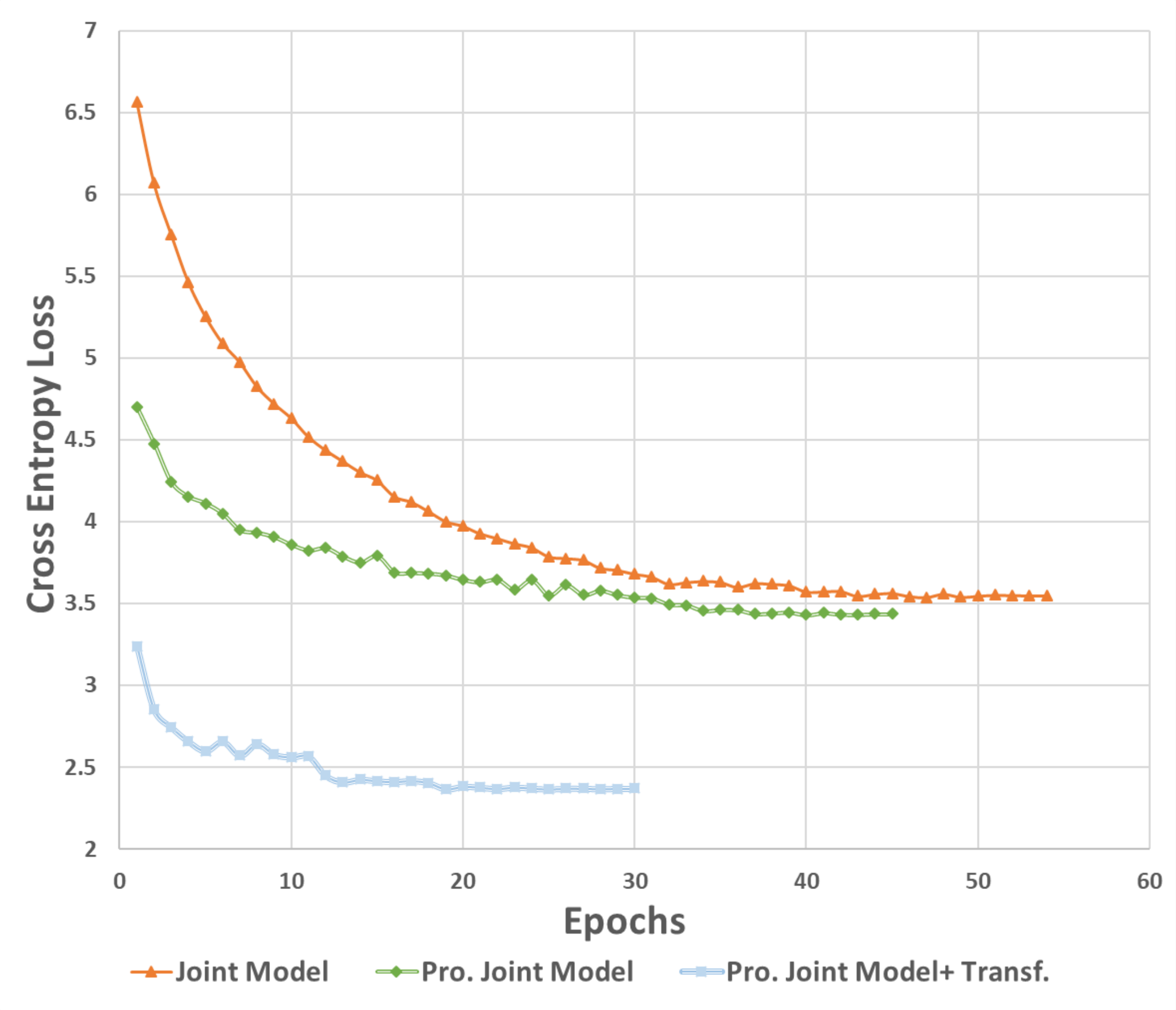
PIT model infer.



Acoustics - Transfer Learning x Joint Training

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

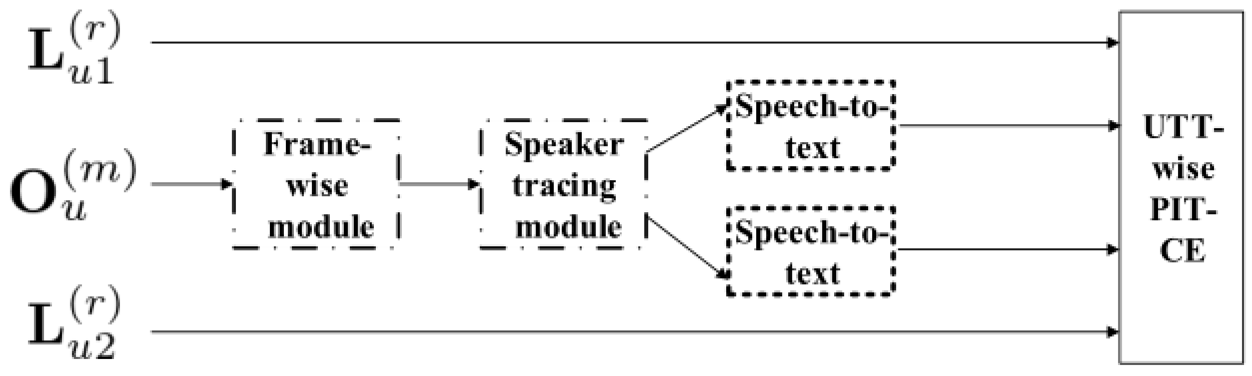




Acoustics - Transfer Learning x Joint Training

- Further: learn from ensemble
  - Different structure has different abilities in this task
    - e.g. 6 layer in the bottom + 4 layer in the top v.s. 10 layer in the bottom

- Motivation
  - Learn different abilities
  - Model compression



(e) Final Joint Training

We propose:

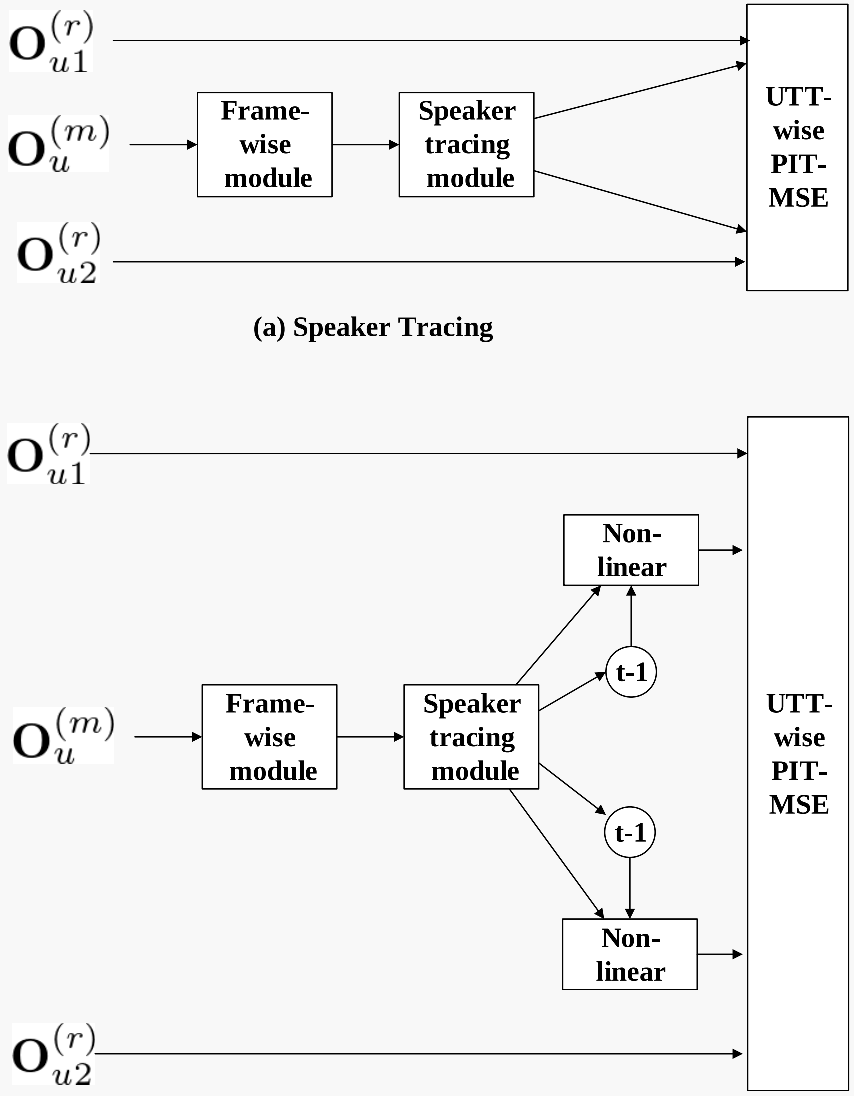
- Acoustics
  - Modular Initialization 4-10%
  - Transfer Learning Based Joint Training 20%
  - **Temporal Correlation Modeling** 8%
- Linguistics
  - Multi-outputs Sequence Discriminative Training 8%
  - Integrating Language Model in Assignment Decision 4%

Acoustics – Temporal Correlation Modeling

- Motivation
  - Sequential correlation v.s. stream de-correlation
    - 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



(a) Speaker Tracing

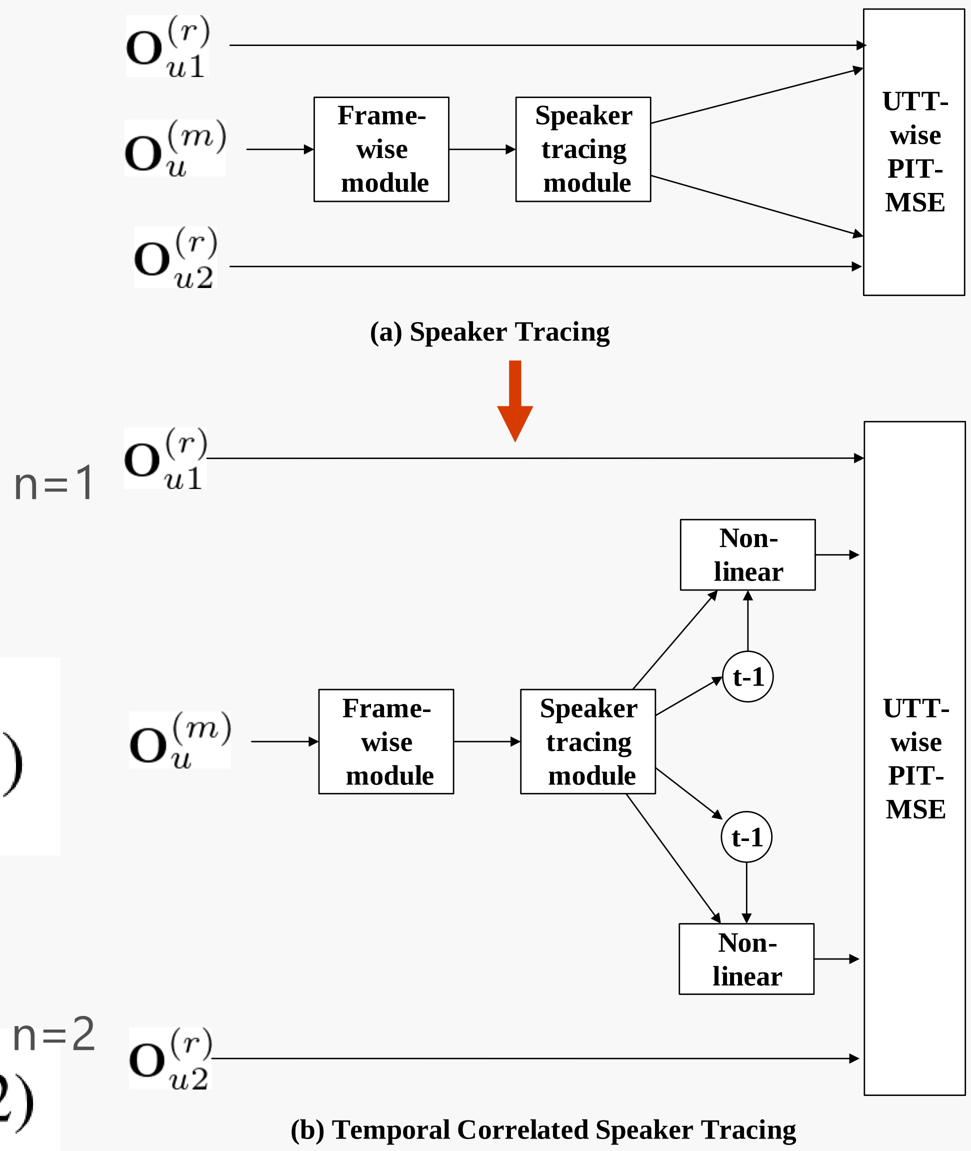
(b) Temporal Correlated Speaker Tracing

Acoustics – Temporal Correlation Modeling

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

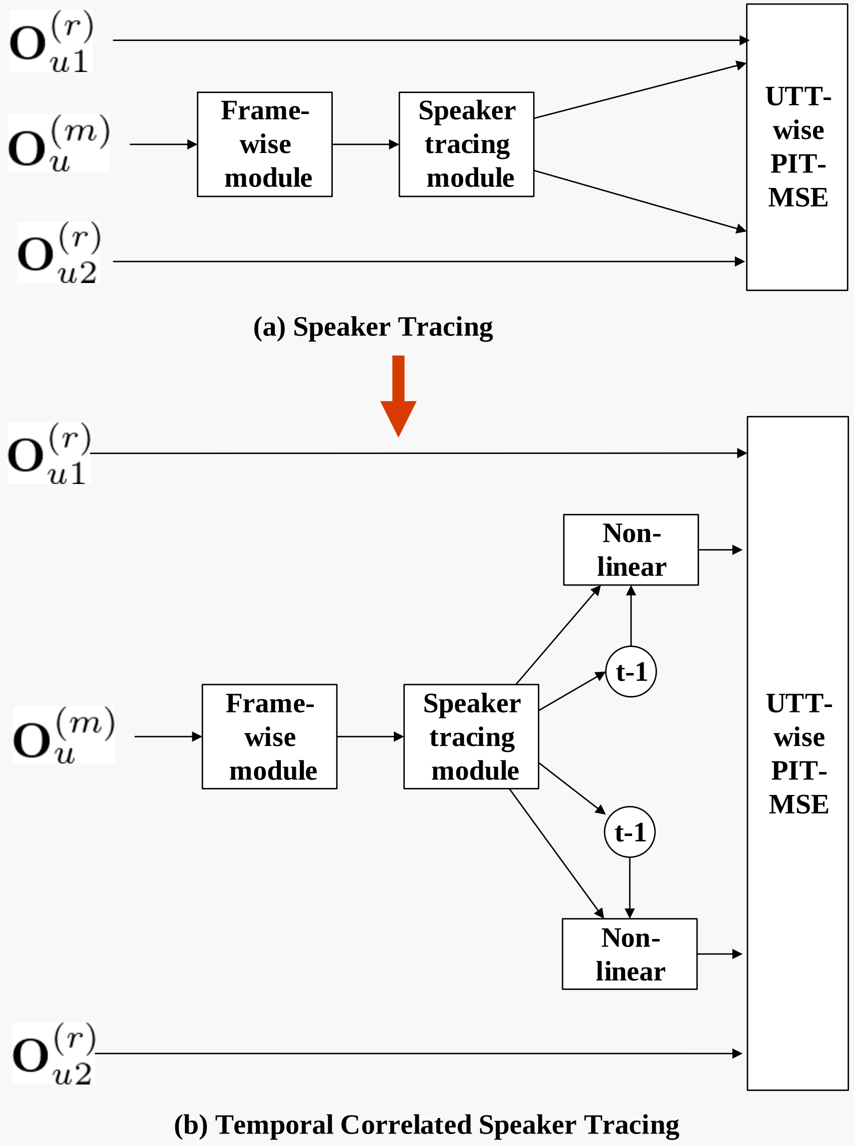
$$o_{utn} = \mathcal{F}_{utn}(\mathbf{O}_u^{(m)}) \quad (1)$$

$$o_{\underline{utn}} = \mathcal{F}'_{utn}(\mathbf{O}_u^{(m)}, o_{u(t-1)\underline{n}}) \quad (2)$$



Acoustics – Temporal Correlation Modeling

- Motivation
  - Sequential correlation v.s. **stream de-correlation**
  - 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

We propose:

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

Linguistics - Multi-outputs Seq. Disc. Training

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

Linguistics - Multi-outputs Seq. Disc. Training

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

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

$$\mathcal{J}_{\text{CE-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \qquad (4)$$

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

Linguistics - Multi-outputs Seq. Disc. Training

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

$$\mathcal{J}_{\text{CE-PIT}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \quad (4)$$

For MMI

$$\mathcal{J}_{\text{SEQ}}(\mathbf{L}_u, \mathbf{L}_u^{(r)}) = \log P(\mathbf{L}_u^{(r)} | \mathbf{O}_u) \quad (11)$$

i.e.

$$\begin{aligned}\mathcal{J}_{\text{LF-MMI}} &= \sum_u \mathcal{J}_{\text{SEQ}}(\mathbf{L}_{un}^{(s')}, \mathbf{L}_{un}^{(r)}) \\ &= \sum_u \log \frac{\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^\kappa P(\mathbf{L}_u)}{\sum_{\mathbf{L}} p(\mathbf{O}_u|\mathbf{L})^\kappa P(\mathbf{L})}\end{aligned} \quad (13)$$

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

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling

$$
\begin{aligned}
\mathcal{J}_{\text{LF-MMI}} &= \sum_u \mathcal{J}_{\text{SEQ}}(\mathbf{L}_{un}^{(s')}, \mathbf{L}_{un}^{(r)}) \\
&= \sum_u \log \frac{\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^{\kappa} P(\mathbf{L}_u)}{\boxed{\sum_{\mathbf{L}}} p(\mathbf{O}_u|\mathbf{L})^{\kappa} P(\mathbf{L})}
\end{aligned} \tag{13}
$$

ASR Insertion
ASR Deletion
ASR substitution
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

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling

$$\mathcal{J}_{\text{LF-MMI}} = \sum_u \mathcal{J}_{\text{SEQ}}(\mathbf{L}_{un}^{(s')}, \mathbf{L}_{un}^{(r)})$$
$$= \sum_u \log \frac{\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^\kappa P(\mathbf{L}_u)}{\boxed{\sum_{\mathbf{L}}} p(\mathbf{O}_u|\mathbf{L})^\kappa P(\mathbf{L})} \tag{13}$$

- Search space modeling
  - With lattice
    - Permutation keep changing → update lattice each epoch
  - √ Lattice-free
    - Pre-pruned senone LM as the search space
    - Both outputs of all utt. share the same denominator graph

ASR Insertion
ASR Deletion
ASR substitution
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

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling
  - Search space modeling
  - Swapped word modeling

ASR Insertion
ASR Deletion
ASR substitution
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

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling
  - Search space modeling
  - Swapped word modeling
    - Motivation: $\mathbf{L}$ doesn't contain swapped words

$$\mathcal{J}_{\text{LF-MMI}} = \sum_u \mathcal{J}_{\text{SEQ}}(\mathbf{L}_{un}^{(s')}, \mathbf{L}_{un}^{(r)})$$
$$= \sum_u \log \frac{\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^\kappa P(\mathbf{L}_u)}{\sum_{\mathbf{L}} p(\mathbf{O}_u|\mathbf{L})^\kappa P(\mathbf{L})} \quad (13)$$

ASR Insertion
ASR Deletion
ASR substitution
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

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling
  - Search space modeling
  - Swapped word modeling
    - Motivation: $\mathbf{L}$ doesn't contain swapped words
    - Method 1: artificial swapped words in LM
      - In LM generation, make texts:
        - Swapped senone segments

ASR Insertion

ASR Deletion

ASR substitution

Assignment error

e.g. ch-a: how oh you

ch-b: are no

Cross talk error

e.g. ch-a: how are you

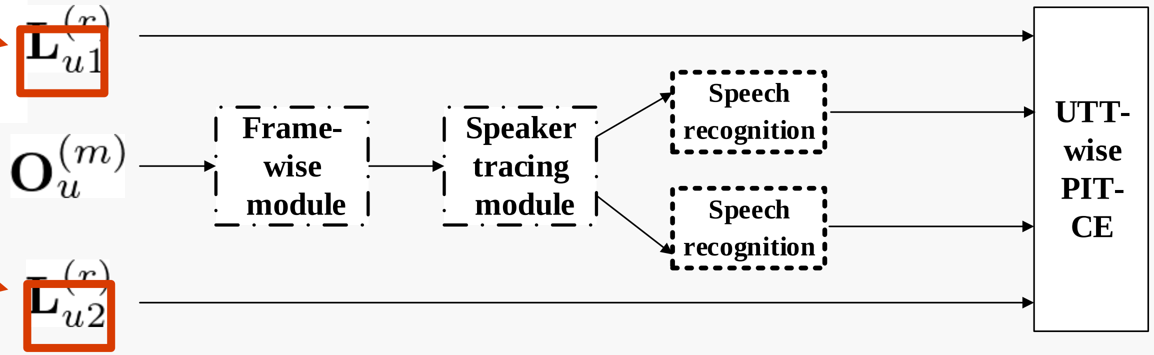
ch-b: oh are no

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling
  - Search space modeling
  - Swapped word modeling
    - Motivation: $\mathbf{L}$ doesn't contain swapped words
    - Method 1: artificial swapped words in LM
    - Method 2: add $\mathbf{L}_{\hat{u}}$ into $\mathbf{L}$

ASR Insertion
ASR Deletion
ASR substitution
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

$$\mathcal{J}_{\text{LF-DC-MMI}} = \sum_u \log\Big[\frac{\sum_{\boxed{\mathbf{L}_u}} p(\mathbf{O}_u|\mathbf{L}_u)^\kappa P(\mathbf{L}_u)}{\left(\sum_{\mathbf{L}} p(\mathbf{O}_u|\mathbf{L})^\kappa P(\mathbf{L})\right)^{1-\lambda}} \cdot \frac{1}{\left(\sum_{\boxed{\mathbf{L}_{\hat{u}}}} p(\mathbf{O}_u|\mathbf{L}_{\hat{u}})^\kappa P(\mathbf{L}_{\hat{u}})\right)^{\lambda}}\Big] \qquad (14)$$



Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling
  - Search space modeling
  - Swapped word modeling
    - Motivation: $\mathbf{L}$ doesn't contain swapped words
    - Method 1: artificial swapped words in LM
    - Method 2: add $\mathbf{L}_{\hat{u}}$ into $\mathbf{L}$

$$\mathcal{J}_{\text{LF-DC-MMI}} = \sum_{u} \log\Big[\frac{\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^{\kappa} P(\mathbf{L}_u)}{\big(\sum_{\mathbf{L}} p(\mathbf{O}_u|\mathbf{L})^{\kappa} P(\mathbf{L})\big)^{1-\lambda}} \cdot \boxed{\frac{1}{\big(\sum_{\mathbf{L}_{\hat{u}}} p(\mathbf{O}_u|\mathbf{L}_{\hat{u}})^{\kappa} P(\mathbf{L}_{\hat{u}})\big)^{\lambda}}}\Big] \quad (14)$$

ASR Insertion

ASR Deletion

ASR substitution

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

Linguistics - Multi-outputs Seq. Disc. Training

- Hypothesis Modeling
  - Search space modeling
  - Swapped word modeling
    - Motivation: $\mathbf{L}$ doesn't contain swapped words
    - Method 1: artificial swapped words in LM
    - Method 2: add $\mathbf{L}_{\hat{u}}$ into $\mathbf{L}$
    - Method 3: boost errors of $\mathbf{L}_{\hat{u}}$ (bMMI)

$$\mathcal{J}_{\texttt{LF-DC-MMI}} = \sum_u \log\Big[\frac{\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^\kappa P(\mathbf{L}_u)}{\left(\sum_{\mathbf{L}} p(\mathbf{O}_u|\mathbf{L})^\kappa P(\mathbf{L})\right)^{1-\lambda}} \cdot \frac{1}{\left(\sum_{\mathbf{L}_{\hat{u}}} p(\mathbf{O}_u|\mathbf{L}_{\hat{u}})^\kappa P(\mathbf{L}_{\hat{u}})\right)^{\lambda}}\Big] \tag{14}$$

$$\mathcal{J}_{\texttt{LF-DC-bMMI}} = \sum_u \log\Big[\sum_{\mathbf{L}_u} p(\mathbf{O}_u|\mathbf{L}_u)^\kappa P(\mathbf{L}_u) \cdot \frac{1}{\sum_{\mathbf{L}} p(\mathbf{O}_u|\mathbf{L})^\kappa P(\mathbf{L}) e^{-b \, \max_{\mathbf{L}_u} A(\mathbf{L},\mathbf{L}_u) - \hat{b} \, \max_{\mathbf{L}_{\hat{u}}} (1 - A(\mathbf{L},\mathbf{L}_{\hat{u}}))}}\Big] \tag{16}$$

ASR Insertion
ASR Deletion
ASR substitution
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

We propose:

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

Linguistics – Language Model Integration

- 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

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \tag{3}$$

Linguistics – Language Model Integration

- 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

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

- PIT-MAP:

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

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

Discriminative training

Proposed method

Linguistics – Language Model Integration

- 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

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \quad (3)$$

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

- Proposed:

PIT-trained AM

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

Linguistics – Language Model Integration

- 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

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \quad (3)$$

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

- Proposed:

Senone level NNLM

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

Linguistics – Language Model Integration

- 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

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

- Proposed:

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

Discriminative training

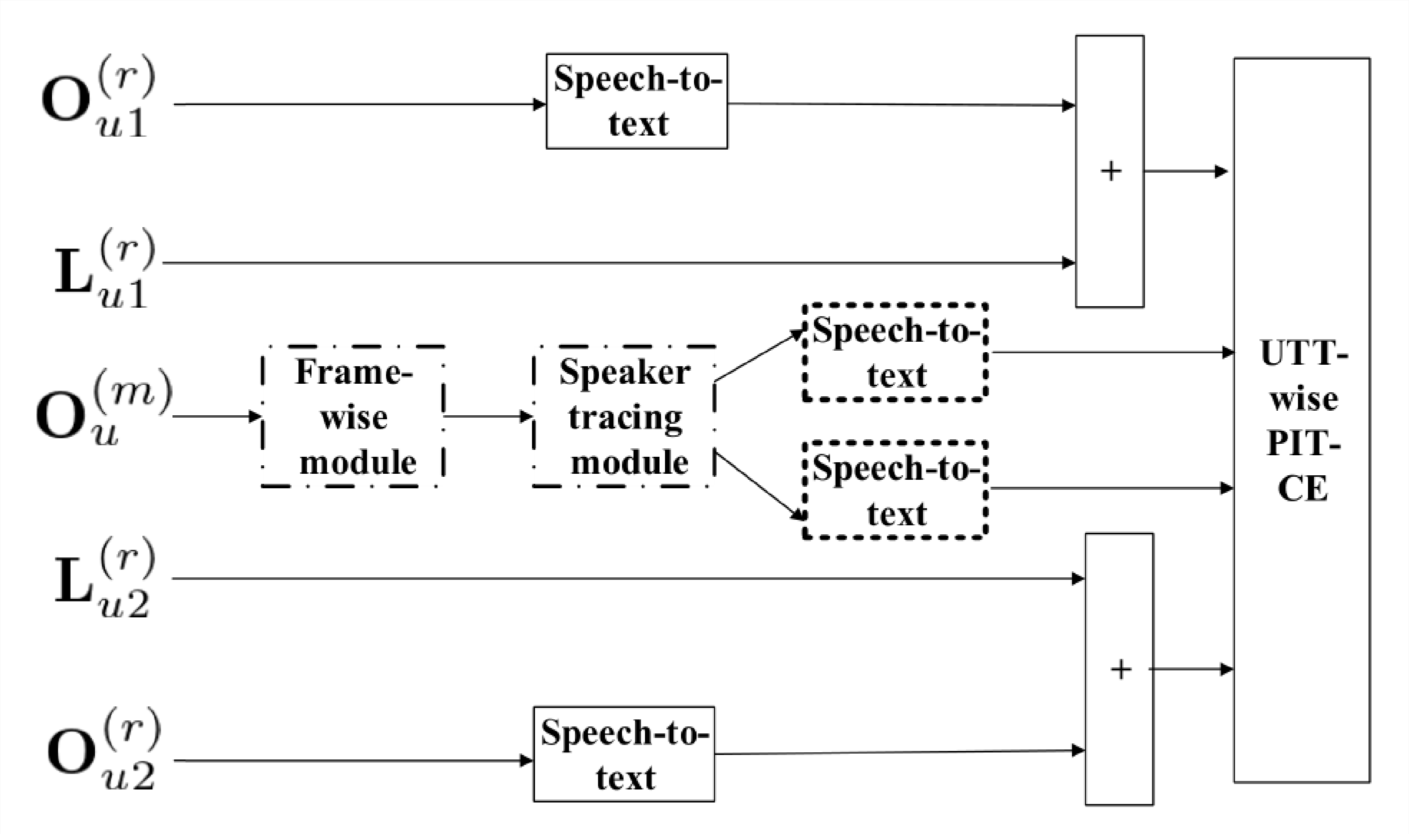
$$\approx \frac{P(l_{utn}^{(r)} | \mathbf{O}_u^{(m)})}{P(l)} \cdot \left(P(l_{utn}^{(r)} | \mathbf{L}_{u(t-1)n}^{(s')}) \right)^{\lambda}$$

Proposed method

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

Brief Summary

- Acoustics
  - Modular Initialization
  - Transfer Learning Based Joint Training
  - Temporal Correlation Modeling
- Linguistics
  - Multi-outputs Sequence Discriminative Training
  - Integrating Language Model in Assignment Decision



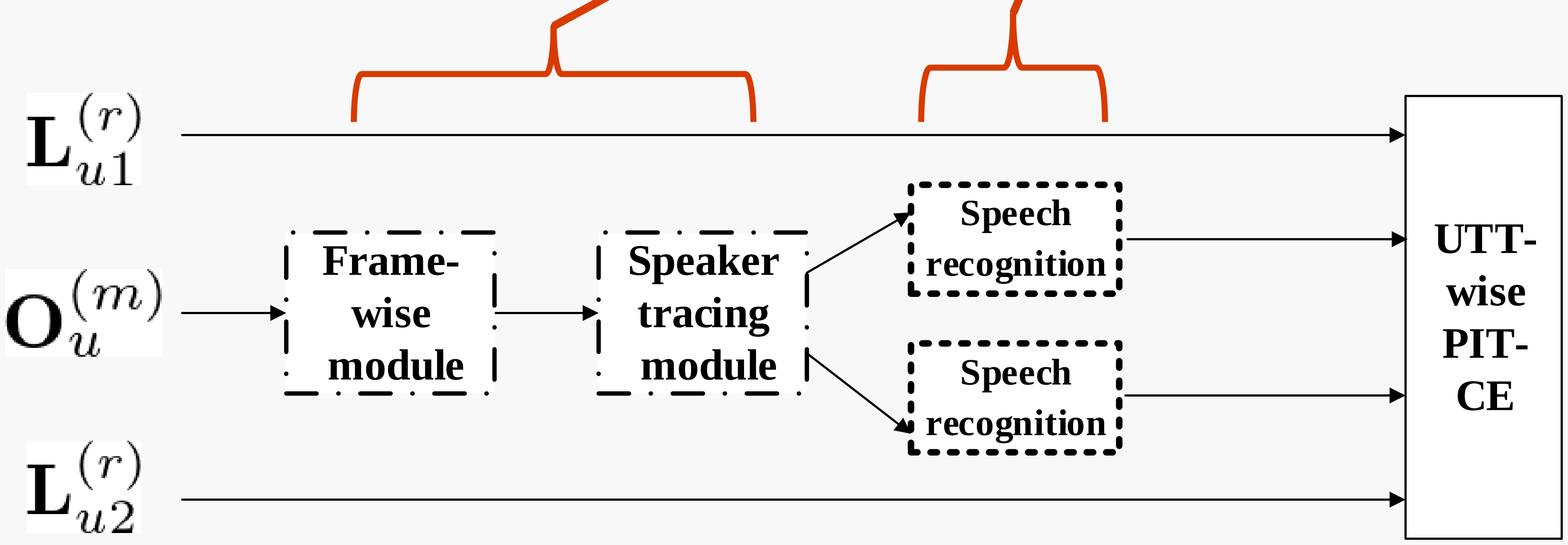
Experiments

- Setup:
  - Artificial overlapped SWBD 300→150 (→50); hub5e-swb 1831 → 915 utts
  - 9000 senones; clean speech alignment;
  - Baseline 1: 10L 768 cells BLSTM PIT-ASR model
  - Baseline 2: 6L 768 cells BLSTM PIT-SS + 4L 768 cells BLSTM ASR
    - All with large # of params. in the original paper

Experiments - Modularization

- Better model generalization

Layers	Modular	Fine-tune ST	Fine-tune ASR	WER	Rel. (%)
10 · 0	×	×	×	57.5	0
6 · 4	×	×	×	52.8	-8.2
	√	×	×	93.4	+62.4
	√	√	×	51.3	-10.7
	√	√	√	50.2	-12.7



Experiments - Modularization

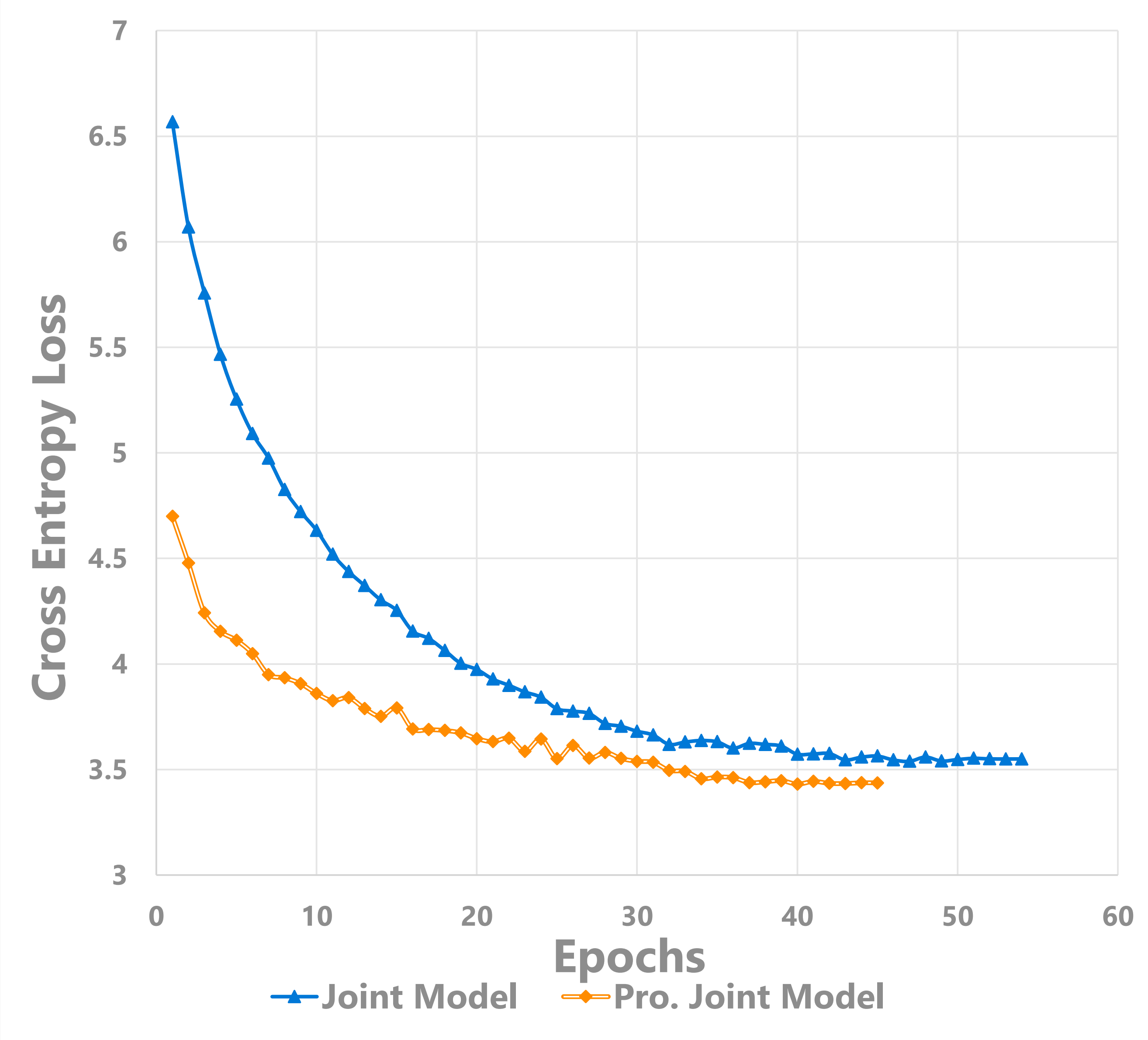
- Better model generalization

Layers	Modular	Fine-tune ST	Fine-tune ASR	WER	Rel. (%)
10 · 0	×	×	×	57.5	0
6 · 4	×	×	×	52.8	-8.2
	√	×	×	93.4	+62.4
	√	√	×	51.3	-10.7
	√	√	√	50.2	-12.7

Better structure for ASR

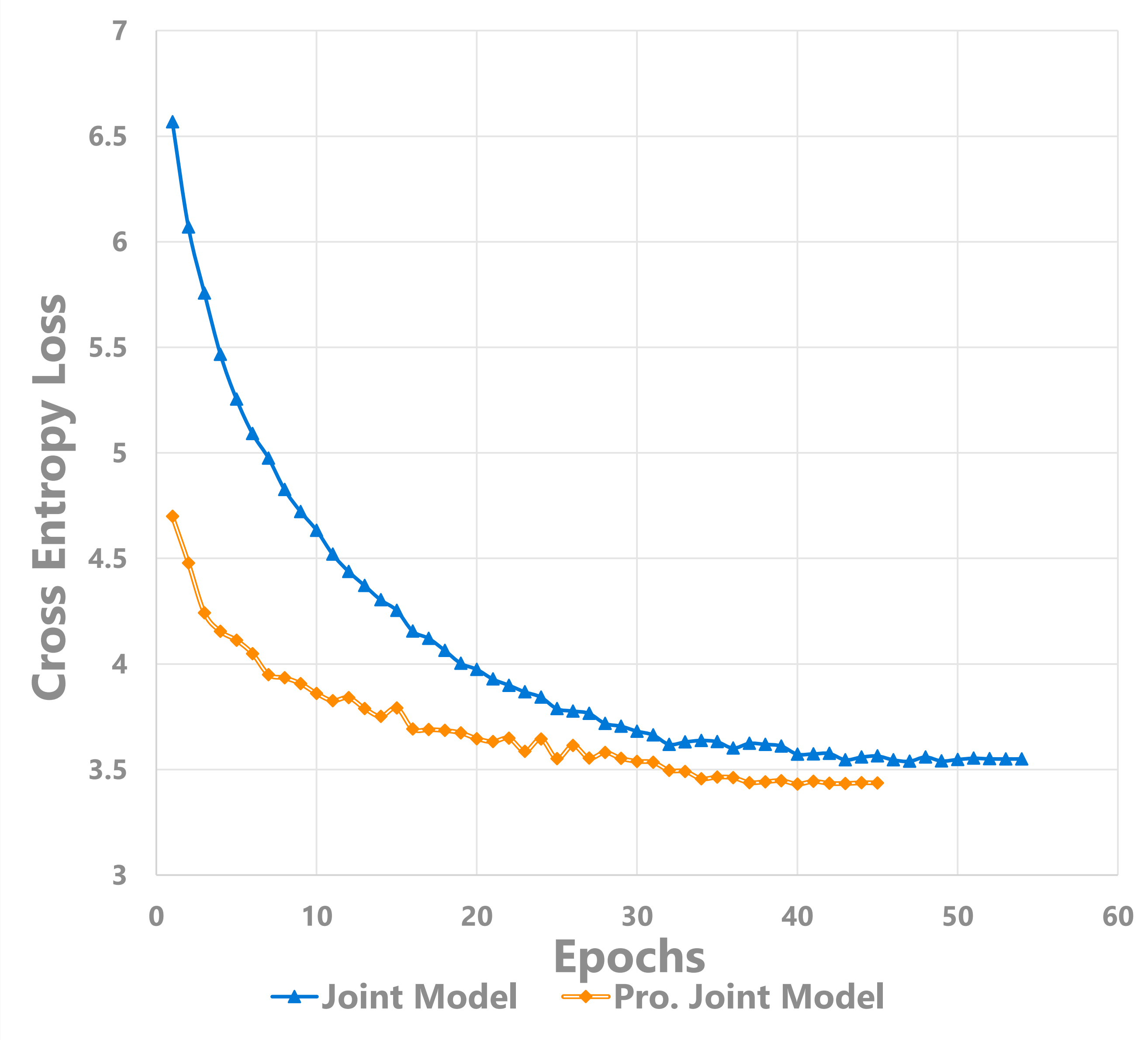
Progressive joint training

Experiments - Modularization



- why
  - Better starting point
  - Better convergence

Experiments - Modularization



- why
  - Better starting point
  - Better convergence
- Better structure
  - Frame-wise interpreting→CNN
  - Speaker Tracing→BLSTM
  - ASR→BLSTM

10 BLSTM	50.3	-4.9
1 LACE + 9 BLSTM	47.4	-10.2

Experiments – Transfer Learning based

Layers	Modular	teacher	WER	Rel. (%)
10·0	$\times$	$\times$	57.5	0
	$\times$	$9{\cdot}1 \oplus 6{\cdot}4 \oplus 3{\cdot}7$	55.0	-4.4
	$\times$	clean	52.5	-8.7
6·4	$\times$	$\times$	52.8	-8.2
	$\times$	clean	47.1	-18.0
	$\surd$	clean	38.9	-32.4
	$\surd$	MMI clean	35.8	-37.7

Experiments – Transfer Learning based

Learn from clean teacher

Layers	Modular	teacher	WER	Rel. (%)
10·0	×	×	57.5	0
	×	9·1 ⊕ 6·4 ⊕ 3·7	55.0	-4.4
	×	clean	52.5	-8.7
6·4	×	×	52.8	-8.2
	×	clean	47.1	-18.0
	√	clean	38.9	-32.4
	√	MMI clean	35.8	-37.7

Experiments – Transfer Learning based

Learn from clean teacher + modularization

Layers	Modular	teacher	WER	Rel. (%)
10·0	×	×	57.5	0
	×	$9{\cdot}1 \oplus 6{\cdot}4 \oplus 3{\cdot}7$	55.0	-4.4
	×	clean	52.5	-8.7
6·4	×	×	52.8	-8.2
	×	clean	47.1	-18.0
	√	clean	38.9	-32.4
	√	MMI clean	35.8	-37.7

Experiments – Transfer Learning based

Learn from clean teacher + modularization

ASR From scratch v.s.
Domain adaptation

Layers	Modular	teacher	WER	Rel. (%)
10·0	×	×	57.5	0
	×	9·1 ⊕ 6·4 ⊕ 3·7	55.0	-4.4
	×	clean	52.5	-8.7
6·4	×	×	52.8	-8.2
	×	clean	47.1	-18.0
	√	clean	38.9	-32.4
	√	MMI clean	35.8	-37.7

Experiments – Transfer Learning based

learn from MMI teacher

Layers	Modular	teacher	WER	Rel. (%)
10·0	$\times$	$\times$	57.5	0
	$\times$	$9{\cdot}1 \oplus 6{\cdot}4 \oplus 3{\cdot}7$	55.0	-4.4
	$\times$	clean	52.5	-8.7
6·4	$\times$	$\times$	52.8	-8.2
	$\times$	clean	47.1	-18.0
	$\surd$	clean	38.9	-32.4
	$\surd$	MMI clean	35.8	-37.7

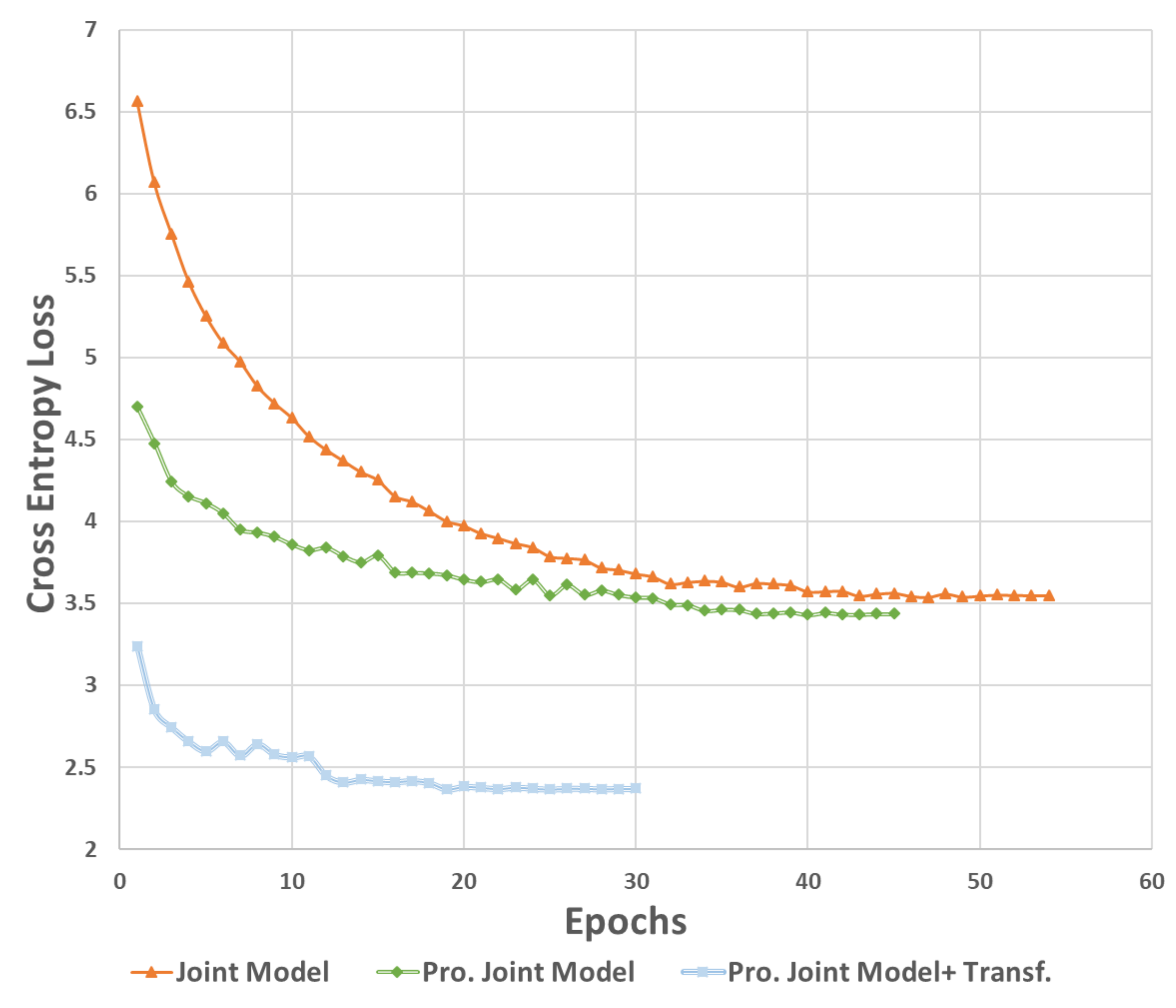
Experiments – Transfer Learning based

Learn from ensemble

Layers	Modular	teacher	WER	Rel. (%)
10·0	×	×	57.5	0
	×	$9{\cdot}1 \oplus 6{\cdot}4 \oplus 3{\cdot}7$	55.0	-4.4
	×	clean	52.5	-8.7
6·4	×	×	52.8	-8.2
	×	clean	47.1	-18.0
	√	clean	38.9	-32.4
	√	MMI clean	35.8	-37.7

Experiments – Transfer Learning based

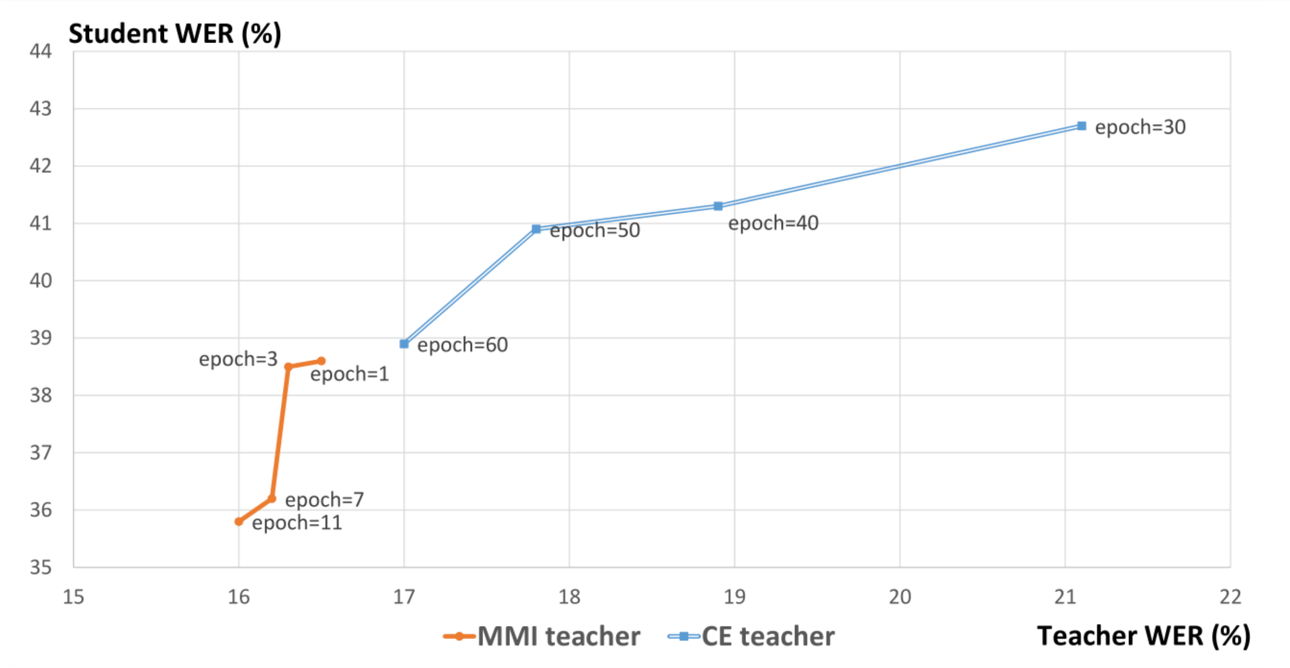
- Why
  - Even better starting point & model convergence

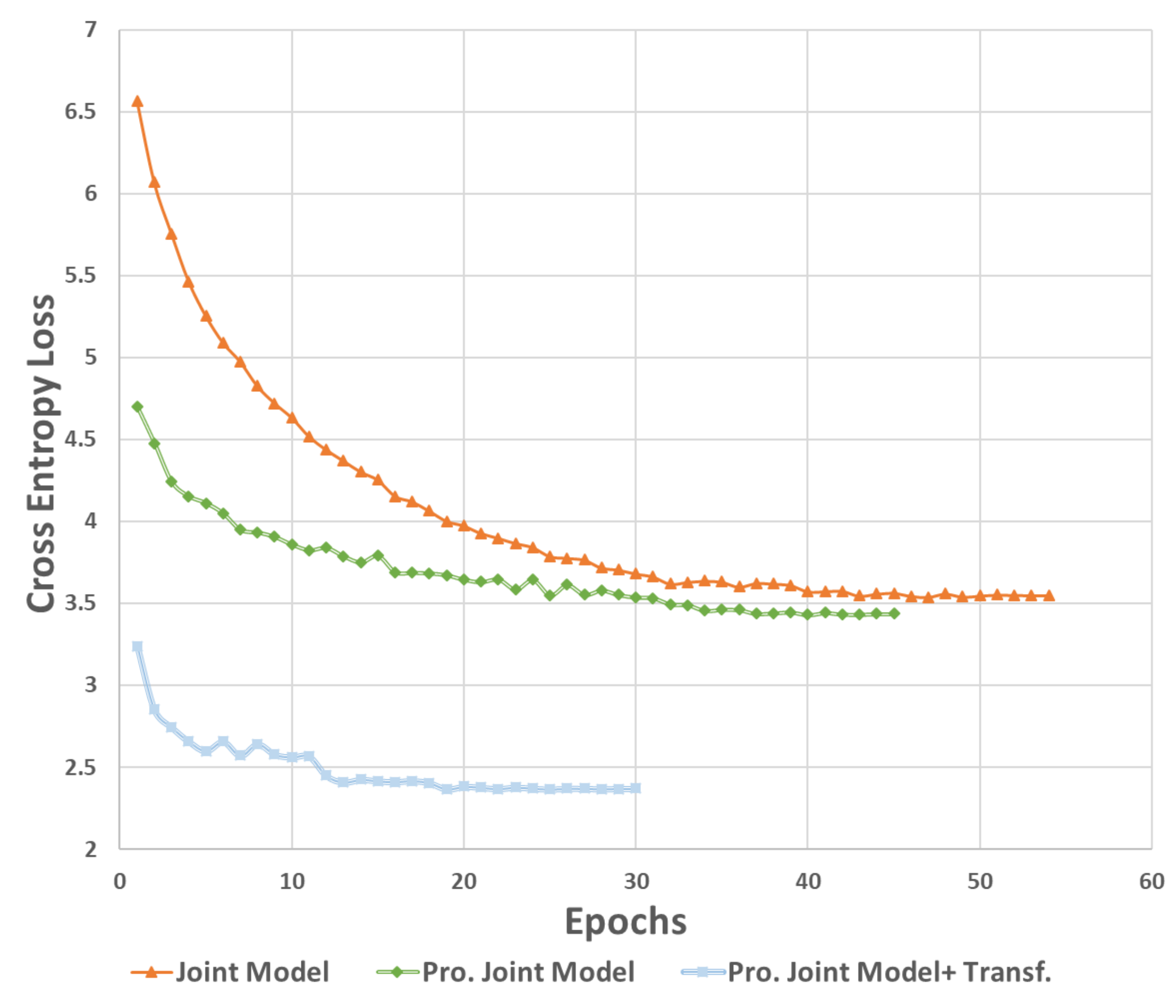


Experiments – Transfer Learning based

- Why
  - Even better starting point & model convergence

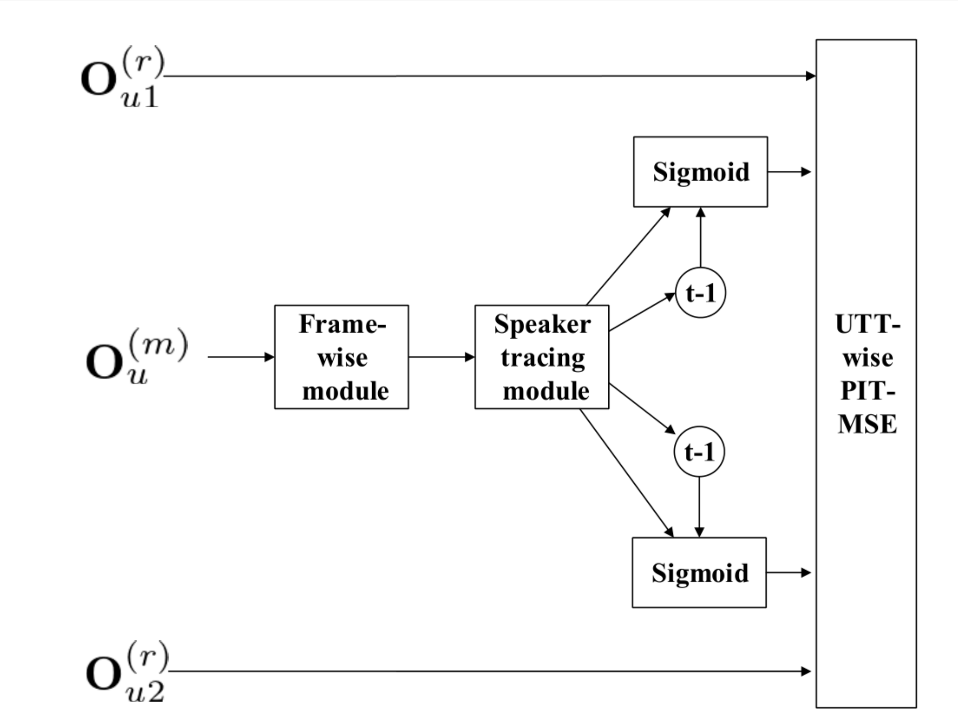
- Relation between Tea. & Stu.

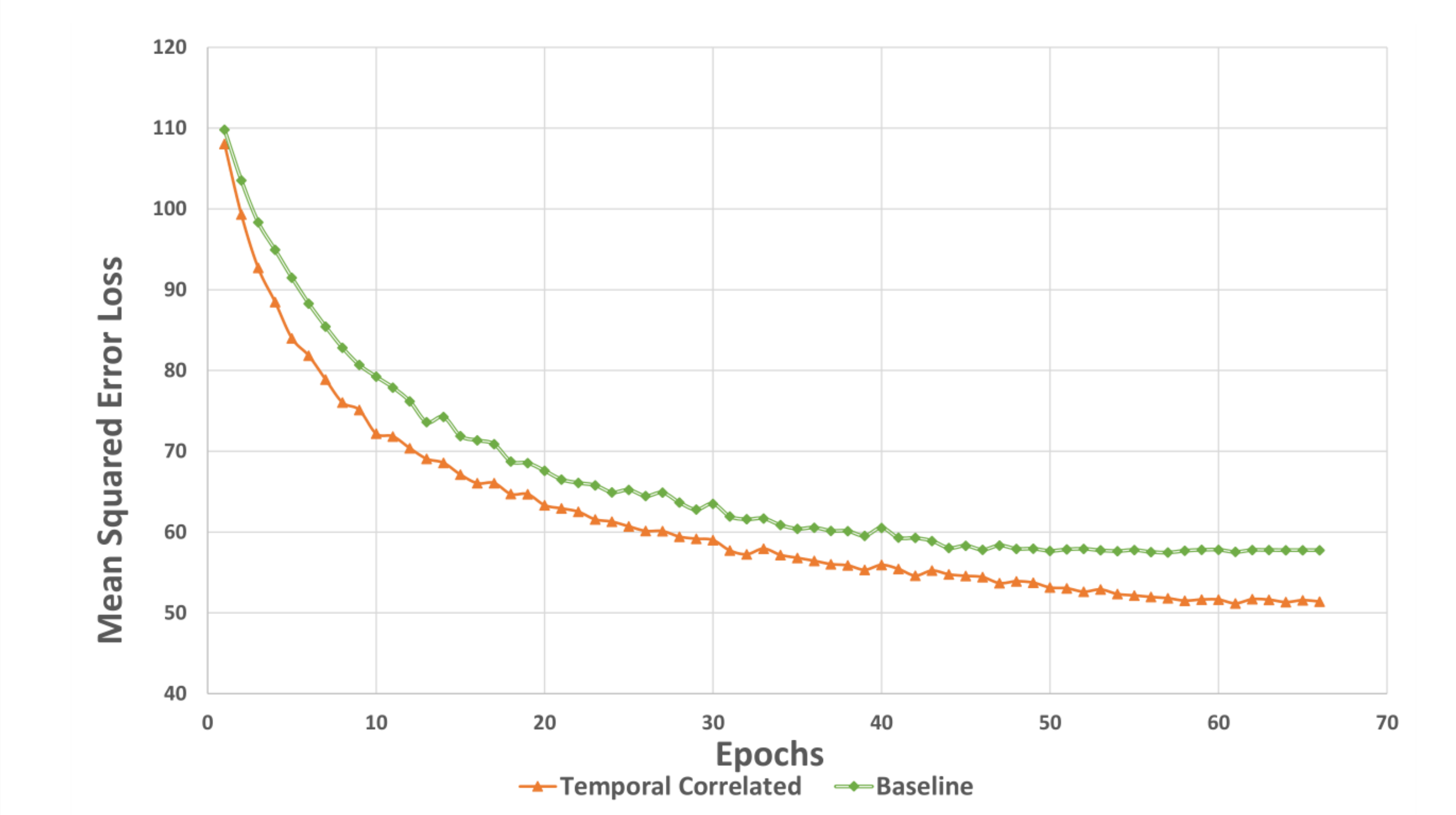




Experiments – Temporal Correlated

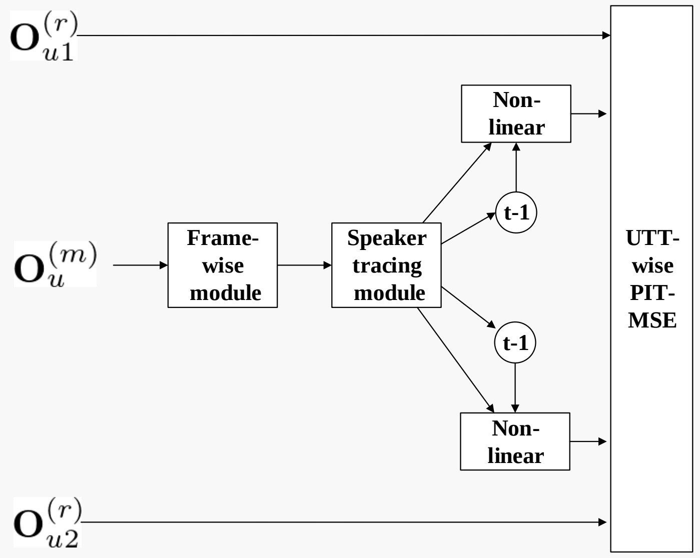
- Baseline: modularization + clean teacher WER=38.9
- Improve in Speaker Tracing:





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
	1	**35.8**	**-8.0**
	2	36.7	-5.7

Experiments – Seq. Disc. Training

6.1% & 7.9% improvement on clean teacher

Performance Summary in SWBD 150 Hours Dataset

Neural network	Model	WER	Rel. (%)
10·0 BLSTM	PIT-CE	42.2	0
6·4 BLSTM	progressive joint training	41.0	-2.9
	+ clean teacher	32.8	-22.3
	+ LF-DC-bMMI	30.8	-27.0
1 LACE + 5·4 BLSTM	progressive joint training	39.4	-6.6
	+ clean teacher	30.4	-27.9
	+ LF-DC-bMMI	28.0	-33.6

Experiments – Seq. Disc. Training

Also improve MMI teacher

Performance Summary in SWBD 50 Hours Dataset

Neural network	Model	WER	Rel. (%)
10·0 BLSTM	PIT-CE	57.5	0
6·4 BLSTM	progressive joint training	50.2	-13
	+ clean teacher	38.9	-32.4
	+ MMI clean teacher	35.8	-37.7
	+ LF-DC-bMMI	35.2	-38.8
1 LACE + 5·4 BLSTM	progressive joint training	47.4	-17.5
	+ clean teacher	36.0	-37.4
	+ MMI clean teacher	34.6	-39.8
	+ LF-DC-bMMI	34.0	-40.9

Experiments – LM Integration

- Baseline: modularization + clean teacher WER=38.9

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \tag{3}$$

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

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

		50 hours		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

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

		50 hours		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**

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

Experiments – Compare with disc. training

system	Assign.	Opt.	50 hours WER	Rel. (%)
baseline	CE	CE	38.9	0
LM integration	MAP	CE	37.3	**-4.1**
LF-DC-bMMI	MAP	$\boxed{\text{MAP}}$	35.6	**-8.5**

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(\mathbf{O}^{(m)}|\mathbf{L}^{(r)}) \cdot P(\mathbf{L}^{(r)})}{P(\mathbf{O}^{(m)})}$$

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

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

Discriminative training

Proposed method

- Differences:
  - optimization stage
  - NNLM v.s. N-gram in discriminative training
  - hardness in modeling $P(\mathbf{O}_u^{(m)})$

Experiments – Combination

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

- 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

Experiments – Summary

- 50 hours
  - WER: 57.5 → 34.0 (-40.9%)
- 150 hours
  - WER: 42.2 → 28.0 (-33.6%)
- *In WSJ, best DPCL sys.
  - WER: 30.8 (joint tr.)
  - WER: 29.7 (Spatial feat. No joint tr.)

  (Although there are lots of diff: corpus, SNR, clean ASR perf…)

Performance Summary in SWBD 50 Hours Dataset

Neural network	Model	WER	Rel. (%)
10·0 BLSTM	PIT-CE	57.5	0
6·4 BLSTM	progressive joint training	50.2	-13
	+ clean teacher	38.9	-32.4
	+ MMI clean teacher	35.8	-37.7
	+ LF-DC-bMMI	35.2	-38.8
1 LACE + 5·4 BLSTM	progressive joint training	47.4	-17.5
	+ clean teacher	36.0	-37.4
	+ MMI clean teacher	34.6	-39.8
	+ LF-DC-bMMI	34.0	-40.9

Performance Summary in SWBD 150 Hours Dataset

Neural network	Model	WER	Rel. (%)
10·0 BLSTM	PIT-CE	42.2	0
6·4 BLSTM	progressive joint training	41.0	-2.9
	+ clean teacher	32.8	-22.3
	+ LF-DC-bMMI	30.8	-27.0
1 LACE + 5·4 BLSTM	progressive joint training	39.4	-6.6
	+ clean teacher	30.4	-27.9
	+ LF-DC-bMMI	28.0	-33.6

Experiments – Example 50hrs (F-F)

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

```
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
HYP:  i just SO LONG AS THEY HAVE COME    HERE AS MUCH    AS  WE go for about an hour      and i **** ***** WAS   READING IN    ANY WAY i mean some times * do *** IT   LONGER THAN A    WOMAN  AND STUFF but you know **** ** *** **** * *
Eval:        I  I    I  I    S    S       S    S  S       S   S                                  D    D     S     S       S     S   S                      D    D   S    S      S    S    S      S   S                  D    D  D   D    D D
```

2 Transf.

```
Scores: (#C #S #D #I) 34 7 10 1
HYP:  i just go OUT YOU     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
Eval:           I   S                     D   S                               D     D          D     S                                                                                        D    D  D   D    D D   S    S   S    S
```

3 +MMI teacher

```
Scores: (#C #S #D #I) 36 9 6 2
HYP:  i just go OUT HERE    and i HAD     LIKE to HAVE go for about an hour      *** 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
Eval:           I   S             S       S       I                                D     D    D     S     S                                                                                            D    D  D   S    S      S
```

4 +seq. disc. tr.

```
Scores: (#C #S #D #I) 37 9 5 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
Eval:           I   S                     S                                       D          D     S  S                                                                                           D    D  S   S    S      S             D   S

id: (sw_4854_b-009)
Labels: <>
File: sw_4854
Channel: b

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

Scores: (#C #S #D #I) 20 22 20 2
HYP:  well *        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   SURE the **** *** ** ******* *
Eval:       D          D  D   D    D    D  D    D  D  D    D  S     S     S         I   S     S   S                               S  S      S     S   S                  I   S  S      S  S   S  S    S   S    S        D    D   D  D       D

Scores: (#C #S #D #I) 42 9 11 0
HYP:  well i        i ** *** CAMPED near as much as we used to we **** ** ** ** FEEL LIKE all the time              we        we all *** SCAN everything ** **** ** *** THESE DOGS OVER LIKE all the time and my parents had a cabin IN th
Eval:                  D  D   S                                    D    D  D  D  S    S                                            D   S               D  D    D  D   S     S    S    S                                              S

Scores: (#C #S #D #I) 42 10 10 0
HYP:  well *        i ** THINK CAMPED near as much as we used to we **** ** ** ** FEEL LIKE all the time              we         we all SCREAM and everything ** **** ** *** YOU STILL HAVE LIKE all the time and my parents had a cabin IN
Eval:       D          D  S     S                                    D    D  D  D  S    S                                                S                     D  D    D  D   S   S     S    S                                              S

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        we all SCREAM and everything ** **** ** *** THESE DOGS BUT LIKE all the time and my parents had a cabin IN
Eval:       D          D  S     S                                                  D                                                    S                     D  D    D  D   S     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)
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

```
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
Eval:           S                     D   S                                  D                D     D  S     S                                                                                D    D  D   D      S
```

2 Transf.

```
Scores: (#C #S #D #I) 39 6 6 2
HYP:  i just go OUT TO EAT     and i usually *** HAVE go for about an hour      *** 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
Eval:           I   I  S                     D   S                              D     S     S                                                                                       D            D    D  D   S    S
```

3 +CNN

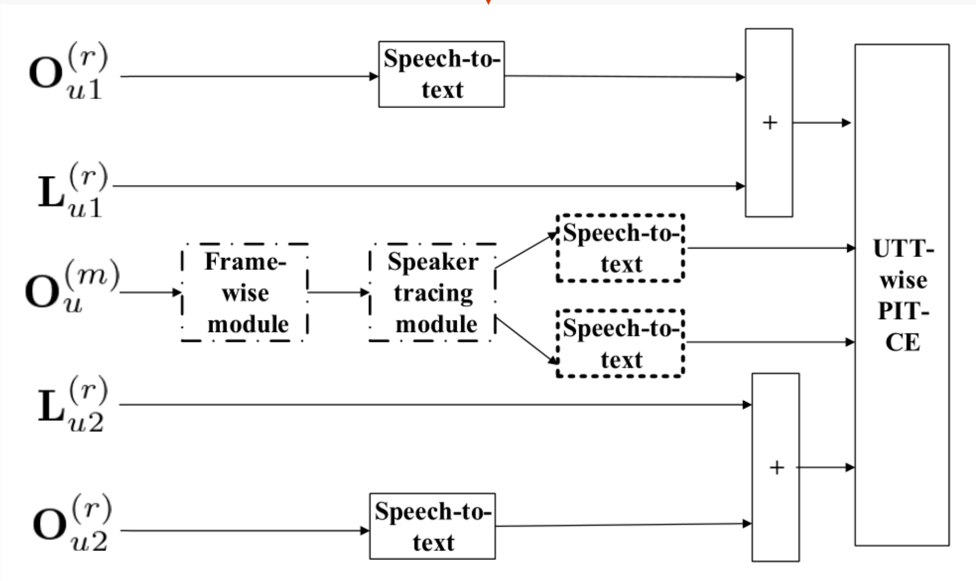
```
Scores: (#C #S #D #I) 42 5 4 1
HYP:  i just go OUT AND     and i usually *** HAVE go for about an hour      and i **** tried doing it every day but i mean some times i do not even do it once during the week but you know **** ONCE the **** SAME TRYING to get out ther
Eval:           I   S                     D   S                                    D                                                                                                         D    S        D    S    S
```

4 +seq. disc. tr.

```
Scores: (#C #S #D #I) 41 8 2 0
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
Eval:           S                      S   S                                   D                                                                                                         D    S    S   S    S   S

id: (sw_4854_b-009)
Labels: <>
File: sw_4854
Channel: b

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 TH

Scores: (#C #S #D #I) 36 10 16 0
HYP:  well *        i ** THINK camp HERE as much as we used to we used ** ** to *** LEAVE all the time               we        we all STAY IN  everything ** **** ** *** ** ** *** YOU  STILL ILLEGAL time *** my parents GOT a ***** ** **
Eval:      D          D  S          S                                  D  D     D   S                                                 S    S              D  D    D  D   D  D  D   S    S     S            D              S     D     D  D

Scores: (#C #S #D #I) 42 10 10 0
HYP:  well *        i AM A   camp near as much as we used to we **** ** ** ** FEEL LIKE all the time              we        we all *** SCAN everything ** **** ** *** THESE DOGS ARE LIKE all the time and my parents had a cabin IN the l
Eval:      D          S  S                                      D    D  D  D  S    S                                                D   S              D  D    D  D   S     S    S   S                                            S

Scores: (#C #S #D #I) 40 9 13 0
HYP:  well *        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
Eval:      D          D  S                                        D    D  D  D  S    S                               D            D   S    S               D  D    D  D   D  S     S    S                                            S

Scores: (#C #S #D #I) 46 9 7 0
HYP:  well *        i ** THINK CAMPED near as much as we used to we used to go to *** lake all the time               we        we all *** SCAN everything ** **** ** WITH THESE DOGS OVER LIKE all the time and my parents had a cabin IN
Eval:      D          D  S     S                                                  D                                                    D   S               D  D    D  S    S     S    S    S                                            S
```

Conclusion

- 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%



Future works

- Better Modular Initialization
  - e.g. DPCL, beamforming with spatial features
- Sequence Modeling
  - e.g. CTC
- Linguistic Information
  - Joint decoding
  - Joint AM & LM modeling (end-to-end)

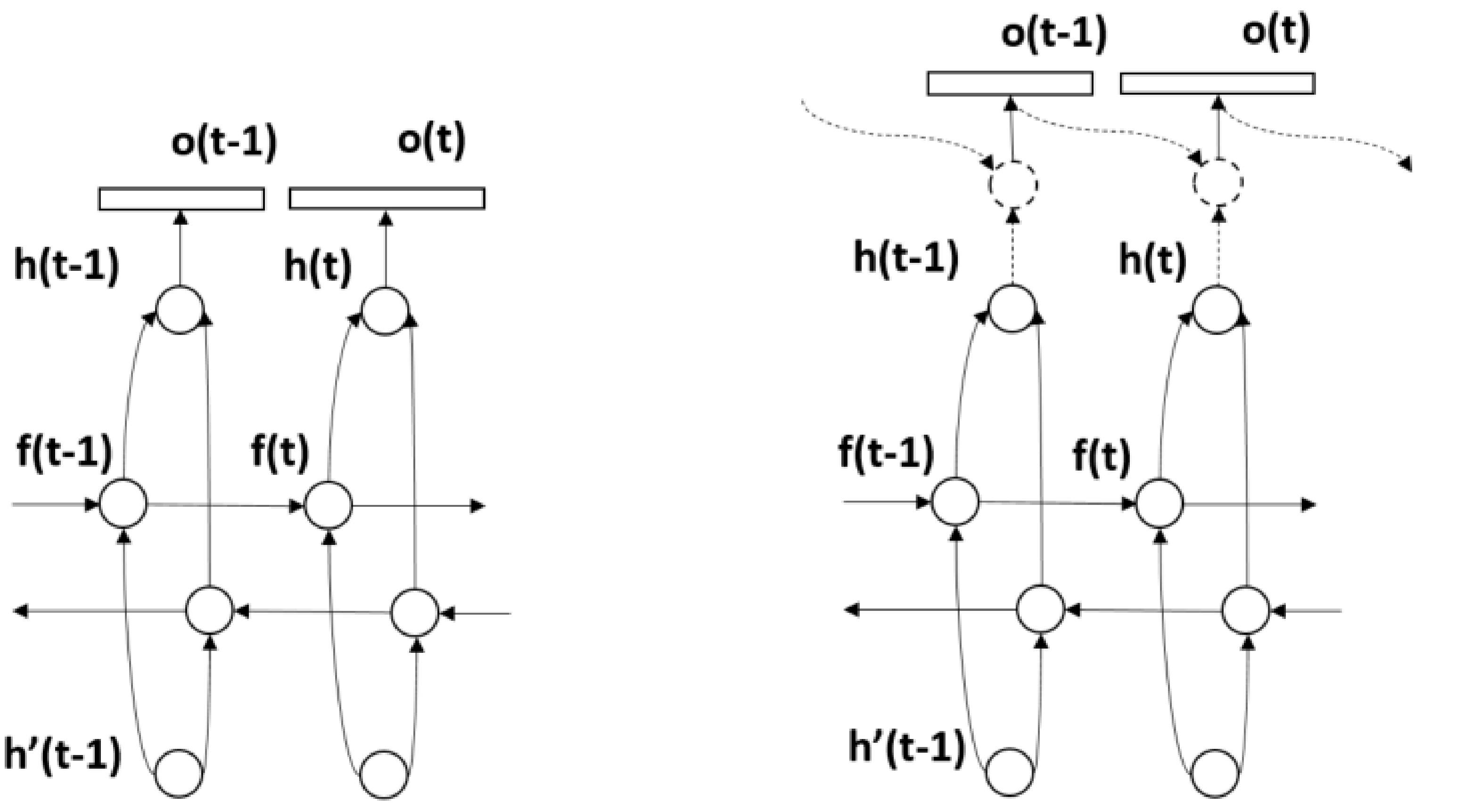
Future works

- Better Modular Initialization
  - e.g. DPCL, beamforming with spatial features
- Sequence Modeling
  - e.g. CTC
- Linguistic Information
  - Joint decoding
  - Joint AM & LM modeling (end-to-end)
- Multi-channel and front-end

- 2000 hrs swb+fsh → 2 weeks → approaching 20%?

Backup materials

Temporal correlation modeling in BLSTM



(a) BLSTM

(b) Temporal Correlated BLSTM