Sequence Modeling
in Unsupervised Single-channel Overlapped Speech Recognition

Zhehuai Chen and Jasha Droppo

chenzhehuai@sjtu.edu.cn
jdropp@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

\[ O_u^{(m)} = \sum_{n=1}^{N} O_u^{(r)} \]

\[ P(L_{u1}, \ldots, L_{uN} | O_u^{(m)}) \]
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

\[ P(L_{u1}, \ldots, L_{uN} | O_u^{(m)}) \approx \prod_{n=1}^{N} P(L_{un}^{(r)} | O_u^{(m)}) \] (2)

\[ O_u^{(m)} = \sum_{n=1}^{N} O_{un}^{(r)} \]
Permutation Invariant Training for ASR

\[
P(L_{u1}, \ldots, L_{uN} | O_u^{(m)}) \approx \prod_{n=1}^{N} P(L_{un}^{(r)} | O_u^{(m)})
\]  \hspace{1cm} (2)

\[
J_{CE-PIT} = \min_{u \in S} \sum_{t} \frac{1}{N} \sum_{n \in [1, N]} CE(t_{utn}^{(s')}, t_{utn}^{(r')})
\]  \hspace{1cm} (4)

- **Disadvantage**
  - Model Complexity (3 hardest problems)
  - Frame CE \Rightarrow Utt. Problem
  - No Linguistics
PIT + Transfer Learning (TS)

\[
J_{CE-PIT} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} CE(l_{unt}^{(s')}, l_{unt}^{(r)})
\]

\[
J_{KLD-PIT} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} KLD(P(l_{un}^{(c)} | O_{un}^{(r)}), P(l_{un}^{(s')} | O_u^{(m)}))
\]

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

Linguistics - Multi-outputs Seq. Disc. Training

• Motivation:
  • Both ASR & speaker tracing $\rightarrow$ sequential
  • Implicit integrating language model

• Formulation:

$$J_{CE-PIT} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})$$  \hspace{1cm} (4)$$

$$J_{SEQ-PIT} = \sum_u \min_{s' \in S} \frac{1}{N} \sum_{n \in [1,N]} \mathcal{J}_{SEQ}(L_{un}^{(s')}, L_{un}^{(r)})$$  \hspace{1cm} (12)$$

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

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 de-correlation
    - the frequency bins between adjacent frames of the same speaker are correlated
  - Last inference can improve current inference

Assignment error:
- e.g. ch-a: how oh you
- ch-b: are no
Acoustics – Temporal Correlation Modeling

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

\[ o_{utn} = F_{utn}(O_u^{(m)}) \]  \hspace{1cm} (1)

\[ o_{utn} = F'_{utn}(O_u^{(m)}, o_u(t-1)n) \]  \hspace{1cm} (2)

---

(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
• alleviates the assignment & cross talk errors

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

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

(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

\[
J_{U-PIT-CE} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)})
\]  

(3)

• PIT-MAP:

\[
MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)}|O_u^{(m)})/P(l) \cdot P(l_{utn}^{(r)}|L_{u(t-1)n}^{(s')})}{P(O_u^{(m)})}
\]

\[
\approx \frac{P(l_{utn}^{(r)}|O_u^{(m)})}{P(l)} \cdot (P(l_{utn}^{(r)}|L_{u(t-1)n}^{(s')}))^\lambda
\]

Discriminative training

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

\[ J_{U-PIT-CE} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{s'}, l_{utn}^{(r)}) \] (3)

Proposed:

\[ MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | O_u^{(m)})}{P(l)} \left( P(l_{utn}^{(r)} | L_{u(t-1)n}^{(s')}) \right)^\lambda \] (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

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


\[ CE(\cdot) \rightarrow MAP(\cdot) \]

• Proposed:

\[ MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)}|O_{u}^{(m)})}{P(l)} \cdot \left( \frac{P(l_{utn}^{(r)}|L_{u(t-1)n}^{(s')})}{P(l_{utn}^{(r)})} \right)^{\lambda} \]
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

\[ J_{\text{U-PIT-CE}} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \]

• Proposed:

\[ MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | O_u^{(m)})/P(l) \cdot P(l_{utn}^{(r)} | L_{u(t-1)n}^{(s')})}{P(O_u^{(m)})} \]

\[ \approx \frac{P(l_{utn}^{(r)} | O_u^{(m)})}{P(l)} \cdot \left( P(l_{utn}^{(r)} | L_{u(t-1)n}^{(s')}) \right)^\lambda \]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PIT</td>
<td>CE</td>
<td>CE</td>
</tr>
<tr>
<td>Proposed</td>
<td>MAP</td>
<td>CE</td>
</tr>
</tbody>
</table>

Discriminative training

Proposed method
Experiments

• Setup and baselines:
  • Artificial overlapped SWBD 300→150 (→50); hub5e-swb 1831 → 915 utts
  • 9000 senones; clean speech alignment;
  • Baseline 1: 6L 768 cells BLSTM PIT-SS + 4L 768 cells BLSTM ASR
Experiments

- Setup and baselines:
  - Artificial overlapped SWBD 300 → 150 (→ 50); hub5e-swb 1831 → 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)

<table>
<thead>
<tr>
<th>Neural network</th>
<th>Model</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 BLSTM + 4 BLSTM</td>
<td>PIT-ASR</td>
<td>57.5</td>
</tr>
<tr>
<td></td>
<td>progressive joint training + clean teacher</td>
<td>38.9</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Temporal Correlated</th>
<th># of Sigmoid</th>
<th>WER</th>
<th>Rel. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>0</td>
<td>38.9</td>
<td>0</td>
</tr>
<tr>
<td>√</td>
<td>1</td>
<td>35.8</td>
<td>-8.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>36.7</td>
<td>-5.7</td>
</tr>
</tbody>
</table>
Experiments – LM Integration

- Baseline: modularization + clean teacher WER=38.9

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

\[
CE(\cdot) \rightarrow MAP(\cdot)
\]

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

<table>
<thead>
<tr>
<th>Assign.</th>
<th>Opt.</th>
<th>50 hours</th>
<th>150 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WER</td>
<td>Rel. (%)</td>
</tr>
<tr>
<td>CE</td>
<td>CE</td>
<td>38.9</td>
<td>0</td>
</tr>
<tr>
<td>MAP</td>
<td>CE</td>
<td>37.3</td>
<td>-4.1</td>
</tr>
</tbody>
</table>
Experiments – LM Integration

- Baseline: modularization + clean teacher WER=32.8

\[
MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)} | O_u^{(m)})}{P(l)} \cdot (P(l_{utn}^{(r)} | L_u^{(s')}_{u(t-1)n}))^\lambda
\] (4)

<table>
<thead>
<tr>
<th>Assign.</th>
<th>Opt.</th>
<th>50 hours</th>
<th>150 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WER</td>
<td>Rel. (%)</td>
</tr>
<tr>
<td>CE</td>
<td>CE</td>
<td>38.9</td>
<td>0</td>
</tr>
<tr>
<td>MAP</td>
<td>CE</td>
<td>37.3</td>
<td>-4.1</td>
</tr>
</tbody>
</table>

- with more data, the improvement becomes larger
  - AM becomes stronger
  - Assignment decision is not over-fit to the LM
Experiments – Compare with disc. training

<table>
<thead>
<tr>
<th>system</th>
<th>Assign.</th>
<th>Opt.</th>
<th>50 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>CE</td>
<td>CE</td>
<td>38.9</td>
</tr>
<tr>
<td>LM integration</td>
<td>MAP</td>
<td>CE</td>
<td>37.3</td>
</tr>
<tr>
<td>LF-DC-bMMI</td>
<td>MAP</td>
<td></td>
<td>35.6</td>
</tr>
</tbody>
</table>

\[
MAP(l_{un}^{(s')}, l_{un}^{(r)}) = \frac{P(O_{u}^{(m)}|L^{(r)}) \cdot P(L^{(r)})}{P(O_{u}^{(m)})} \\
= \frac{P(l_{un}^{(r)}|O_{u}^{(m)})/P(l) \cdot P(l_{un}^{(r)}|L_{u(t-1)n}^{(s')})}{P(l)} \\
\approx \frac{P(l_{un}^{(r)}|O_{u}^{(m)})}{P(l)} \cdot (P(l_{un}^{(r)}|L_{u(t-1)n}^{(s')}) )^{\lambda} \\
\]

- **Differences:**
  - optimization stage
  - NNLM v.s. N-gram in discriminative training
  - hardness in modeling

**Proposed method**
### Experiments – Combination

<table>
<thead>
<tr>
<th>Method</th>
<th>WER</th>
<th>Rel. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>38.9</td>
<td>0</td>
</tr>
<tr>
<td>+ Temporal Correlated</td>
<td>35.8</td>
<td>-8.0</td>
</tr>
<tr>
<td>+ LM Integration</td>
<td>34.4</td>
<td>-11.5</td>
</tr>
<tr>
<td>+ LF-DC-bMMI</td>
<td>31.6</td>
<td>-18.8</td>
</tr>
</tbody>
</table>

- Operate in different levels ➔ can be combined
Experiments – Combination

<table>
<thead>
<tr>
<th>Method</th>
<th>WER</th>
<th>Rel. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>38.9</td>
<td>0</td>
</tr>
<tr>
<td>+ Temporal Correlated</td>
<td>35.8</td>
<td>-8.0</td>
</tr>
<tr>
<td>+ LM Integration</td>
<td>34.4</td>
<td>-11.5</td>
</tr>
<tr>
<td>+ LF-DC-bMMI</td>
<td>31.6</td>
<td>-18.8</td>
</tr>
<tr>
<td>+ MMI clean teacher</td>
<td>35.8</td>
<td>-8.0</td>
</tr>
<tr>
<td>+ LF-DC-bMMI</td>
<td>35.2</td>
<td>-9.5</td>
</tr>
</tbody>
</table>

- 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%**

---


Backup materials
Temporal correlation modeling in BLSTM

(a) 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 PIT-CE

2 Transf.

3 + MMI teacher

4 + seq. disc. tr.
Experiments – Example 150hrs (F-F)

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

2 Transf.

3 +CNN

4 +seq. disc. tr.

REF: I just go WALKING and I usually TRY TO go for about an hour (or) 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.

Scores: (MC #4 D #1) 36 16 0
Hyp: Well *, I ** THINK camp HERE as much as we used to use we used ** ** to *** LEAVE all the time
Eval: D S S S D D D D S S D S D D D D

Scores: (MC #4 D #1) 42 10 0
Hyp: Well * I am away camp near as much as we used to use we ***** ** FEEL LIKE all the time
Eval: D D D D S S D S D D D D

Scores: (MC #4 D #1) 40 9 0
Hyp: Well * I ** THINK camp near as much as we used to use ***** ** FEEL LIKE all the time
Eval: D D D D S S D S D D D D

Scores: (MC #4 D #1) 46 9 0
Hyp: Well * I ** THINK CAMPED near as much as we used to use *** lake all the time
Eval: D D S S