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

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

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

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

(2)

\[
\mathbf{O}_{u}^{(m)} = \sum_{n=1}^{N} \mathbf{O}_{un}^{(r)}
\]
Permutation Invariant Training for ASR

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

\[ J_{CE-PIT} = \sum_u \min_{s^i \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} CE(l^{(s^i)}_{utn}, l^{(r)}_{utn}) \] (4)

**Disadvantage**
- Model Complexity (3 hardest problems)
- Frame CE \( \rightarrow \) Utt. Problem
- No Linguistics
\[ J_{\text{CE-PIT}} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} CE(l_{u_{tn}}^{(s')}, l_{u_{tn}}^{(r)}) \] (4)

\[ J_{\text{KLD-PIT}} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} KLD(P(l_{u_{tn}}^{(c)} | O_u^{(r)}), P(l_{u_{tn}}^{(s')} | O_u^{(m)})) \] (8)

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

Linguistics - Multi-outputs Seq. Disc. Training

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

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

• 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 $\rightarrow$ 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

![Diagram of Speaker Tracing and Temporal Correlated Speaker Tracing](image-url)
Acoustics – Temporal Correlation Modeling

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

\[ o_{utn} = \mathcal{F}_{utn}(O_u^{(m)}) \]  \hspace{2cm} (1)

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

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

• **PIT-MAP:**

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

- **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

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

• Proposed:

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

\[ J_{U-PIT-CE} = \sum_{u} \min_{s' \in S} \sum_{t} \frac{1}{N} \sum_{n \in [1,N]} CE(l^{(s')}_{utn}, l^{(r)}_{utn}) \] (3)

• Proposed:

\[ MAP(l^{(s')}_{utn}, l^{(r)}_{utn}) = \frac{P(l^{(r)}_{utn} | O^{(m)}_u)}{P(l)} \cdot \left( P(l^{(r)}_{utn} | L^{(s')}_{u(t-1)n}) \right)^{\lambda} \] (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
\[ J_{U-PIT-CE} = \sum_u \min_{s' \in S} \sum_t \frac{1}{N} \sum_{n \in [1, N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \]

Proposed:
\[ MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)}|O_u^{(m)})/P(l) \cdot P(l_{utn}^{(r)}|L_u^{(s')}}{P(O_u^{(m)})} \]
\[ \approx \frac{P(l_{utn}^{(r)}|O_u^{(m)})}{P(l)} \cdot (P(l_{utn}^{(r)}|L_u^{(s')}})^\lambda \]

<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>
<tr>
<td>Disc. Train</td>
<td>MAP</td>
<td>MAP</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>0</td>
<td>37.5</td>
<td>-3.6</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

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

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

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

<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

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

<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>MAP</td>
<td>35.6</td>
</tr>
</tbody>
</table>

MAP($l_{u | t_n}^{s'}$, $l_{u | t_n}^{r}$) = \frac{P(O^{(m)}_{l | t_n} | L^{(r)}_{l | t_n}) \cdot P(L^{(r)}_{l | t_n})}{P(O^{(m)}_{l | t_n})}

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

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

**Discriminative training**

**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><strong>31.6</strong></td>
<td><strong>-18.8</strong></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.

+ MMI teacher

+ 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.