Rapid Speaker Adaptation Based on D-code Extracted from BLSTM-RNN in LVCSR

Shaofei Xue¹, Zhijie Yan¹, Zhiying Huang², Lirong Dai²

¹Alibaba Inc
²University of Science and Technology of China

Sept. 20, 2016 @ Tianjin
Outline

- Introduction
- Method
- Experiments
- Conclusions
Outline

Introduction

Method

Experiments

Conclusions
Introduction

Background
Background

- **Speaker code** based adaptation have been applied to unsupervised speaker adaptation for NN models.
- About **8%-15%** relative reduction in WER/CER on different tasks.
- Two-pass decoding is needed.
Motivation

- Obtain final results with one-pass decoding.
- Improve accuracy when adaptation data is especially limited.
D-code extraction based BLSTM

Classification of speakers

Input

T_1
T_2
T_3
T_4
\ldots
T_{N-1}
T_N

Sum &\& Normalization

D-code

Output

O^{(3)}
O^{(2)}
O^{(1)}
Method

**LC-BLSTM training** (Yu Zhang, et al. 2015)

time

![Diagram of LC-BLSTM training](image-url)
Method

Speaker clustering BLSTM

Problem

- Target speakers often 100,000+ in huge task.
- Sometimes even no speaker information.
- Implementing speaker classification with NNs often meets problem.
Speaker clustering BLSTM

- Hierarchical clustering based i-vector
- Training cluster speaker-BLSTM is fast.
- Improves ASR performance.
Method

D-code interpolation

➢ Problem

➢ WER/CER increases visibly when data is extreme limited (e.g. one sentence).

➢ Solution

➢ Use D-codes of training set.

➢ Interpolate new speaker’s D-code with N most likely D-codes from training set through

\[
\bar{S}_{test} = \frac{\alpha S_{test} + (1-\alpha) \sum_{i=1}^{N} \beta_i S_{train,i}}{\alpha + (1-\alpha) \sum_{i=1}^{N} \beta_i}
\]
Outline

- Introduction
- Method
- Experiments
- Conclusions
Experiments

Experimental setup

- 309 hour Switchboard-I and 20 hour Call Home English training set.
- Hub5e evaluation set.
- Features: 36 dimensional FBANK features, plus their first and second derivatives.
- A standard 8882 tri-phones GMM/HMM model for force alignment.
- 4-gram LM using training and Fisher English Part 1 transcripts.

Baselines

- ReLU-DNN(3x1024)
- ReLU-DNN(6x2048)
- hybrid BLSTM-RNN(3BLSTM+2ReLU-DNN)
Experiments

Table 1: Adaptation performance using different d-code extractions on a 3-layer ReLU-DNN.

<table>
<thead>
<tr>
<th>Speaker-BLSTMs</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>15.6</td>
</tr>
<tr>
<td>BLSTM-1hid*200cell</td>
<td>14.2(9%)</td>
</tr>
<tr>
<td>BLSTM-1hid*400cell</td>
<td>14.2</td>
</tr>
<tr>
<td>BLSTM-1hid*1000cell</td>
<td>14.1</td>
</tr>
<tr>
<td>BLSTM-2hid*200cell</td>
<td>14.2</td>
</tr>
<tr>
<td>LSTM-1hid*200cell</td>
<td>14.8</td>
</tr>
<tr>
<td>LSTM-2hid*200cell</td>
<td>14.7</td>
</tr>
</tbody>
</table>

D-code from speaker-BLSTMs outperforms speaker-LSTMs and the size of speaker-BLSTMs has no conspicuous influence.
Table2 Training time for speaker-BLSTMs and WER(%) of adaptation with different speaker cluster number.

<table>
<thead>
<tr>
<th>Number of cluster</th>
<th>Training time (1 epoch)</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4803 (no clustering)</td>
<td>1.76h</td>
<td>14.2</td>
</tr>
<tr>
<td>400</td>
<td>0.83h</td>
<td>14.0</td>
</tr>
<tr>
<td>800</td>
<td>0.9h</td>
<td>13.9</td>
</tr>
<tr>
<td>1200</td>
<td>0.96h</td>
<td>13.8 (2.8%)</td>
</tr>
<tr>
<td>1600</td>
<td>1h</td>
<td>14.0</td>
</tr>
</tbody>
</table>

Speaker clustering not only speeds up the training of speaker-BLSTMs (about two times) but also benefits the ASR performance.
**Table 3** WER(%) of D-code interpolation method on a 3-layer ReLU-DNN.

<table>
<thead>
<tr>
<th>D-code</th>
<th>α</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>speaker-level</td>
<td>13.9</td>
<td>13.8</td>
<td><strong>13.7</strong></td>
<td>13.8</td>
<td>13.8</td>
</tr>
<tr>
<td>utterance-level</td>
<td><strong>14.0</strong></td>
<td>14.0</td>
<td>14.2</td>
<td>14.3</td>
<td>14.3</td>
</tr>
</tbody>
</table>

D-code interpolation improve the performance when data is especially limited.
Table 4 Comparison of different adaptation strategies on better baselines.

<table>
<thead>
<tr>
<th>Models</th>
<th>Adaptation strategies</th>
<th>WER(%)</th>
<th>Decoding pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU-DNN(6x2048)</td>
<td>baseline</td>
<td>13.9</td>
<td>one</td>
</tr>
<tr>
<td></td>
<td>d-code</td>
<td>12.7</td>
<td>one</td>
</tr>
<tr>
<td></td>
<td>standard SAT-SC</td>
<td>12.7</td>
<td>two</td>
</tr>
<tr>
<td></td>
<td>i-vector</td>
<td>13.0</td>
<td>one</td>
</tr>
<tr>
<td>hybrid BLSTM-RNN</td>
<td>baseline</td>
<td>13.0</td>
<td>one</td>
</tr>
<tr>
<td></td>
<td>d-code</td>
<td>11.9</td>
<td>one</td>
</tr>
<tr>
<td></td>
<td>standard SAT-SC</td>
<td>11.8</td>
<td>two</td>
</tr>
<tr>
<td></td>
<td>i-vector</td>
<td>12.2</td>
<td>one</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Method
- Experiments
- Conclusions
Conclusions

- An effective speaker adaptation method named D-code adaptation is proposed.
- Speaker clustering is introduced to accelerate training speed and improves ASR performance.
- Interpolation method that make use of D-codes from training set is provided to improve the recognition accuracy.
Thank you!

Q&A

Shaofei Xue, 薛少飞
Shaofei.xsf@alibaba-inc.com