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

Shaofei Xue[1], Zhijie Yan[1], Zhiying huang[2], Lirong Dai[2]

[1]Alibaba Inc
[2]University of Science and Technology of China

Sept. 20, 2016 @ Tianjin

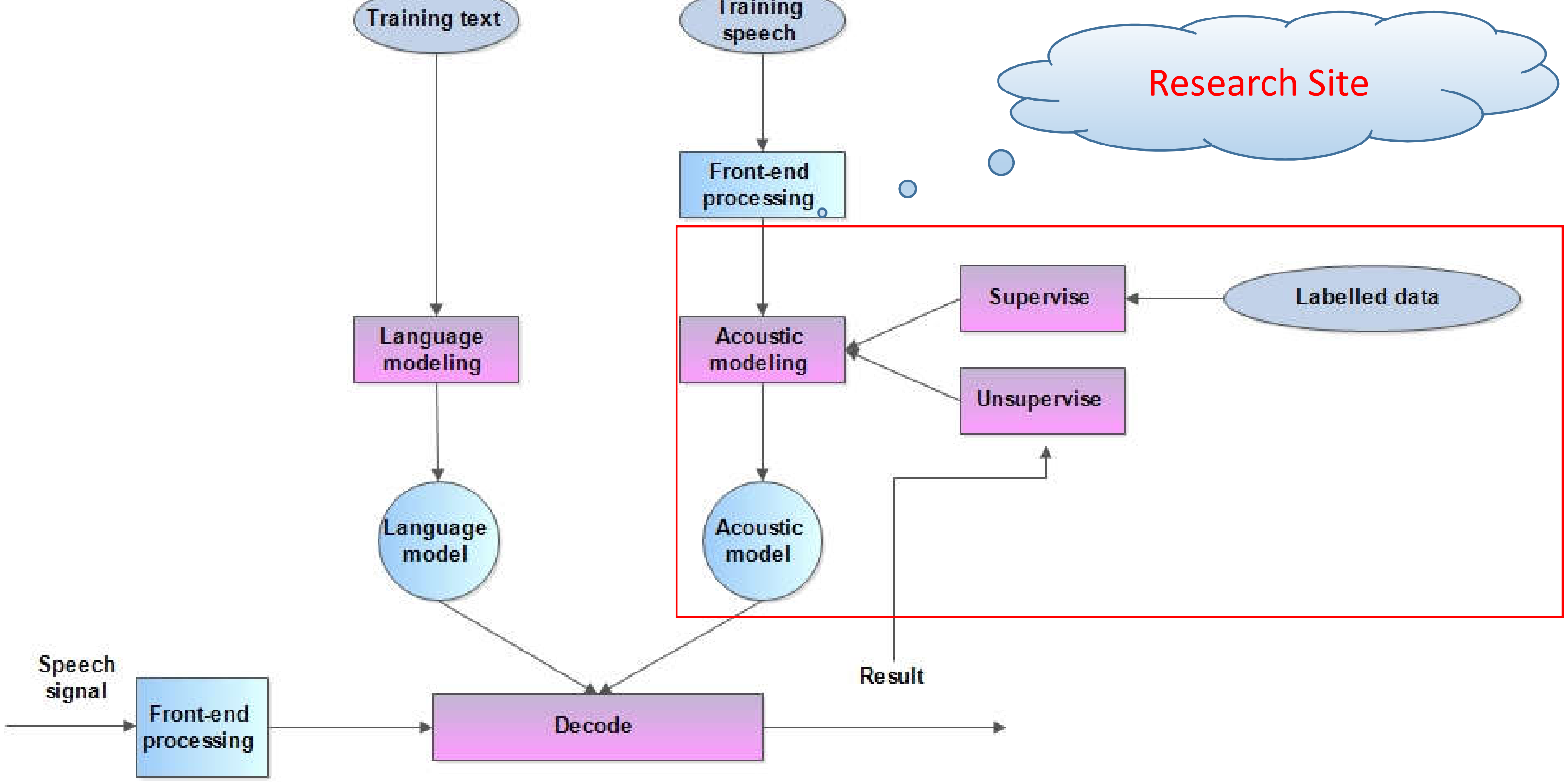
Outline

- **Introduction**
- **Method**
- **Experiments**
- **Conclusions**

Outline

- **Introduction**
- Method
- Experiments
- Conclusions

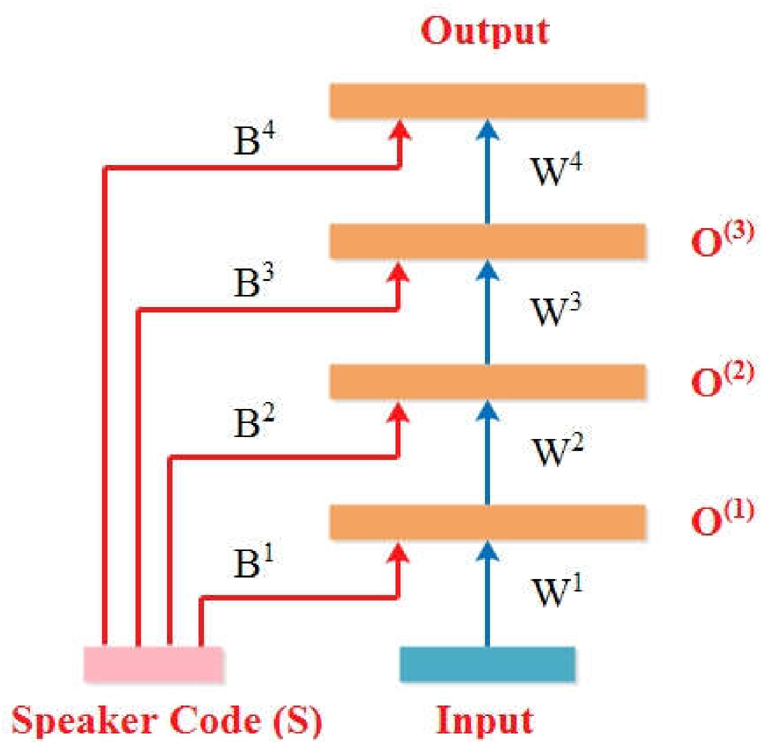
Background


Training text
Training speech
Research Site
Front-end processing
Supervise
Labelled data
Language modeling
Acoustic modeling
Unsupervise
Language model
Acoustic model
Speech signal
Front-end processing
Decode
Result

Introduction

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

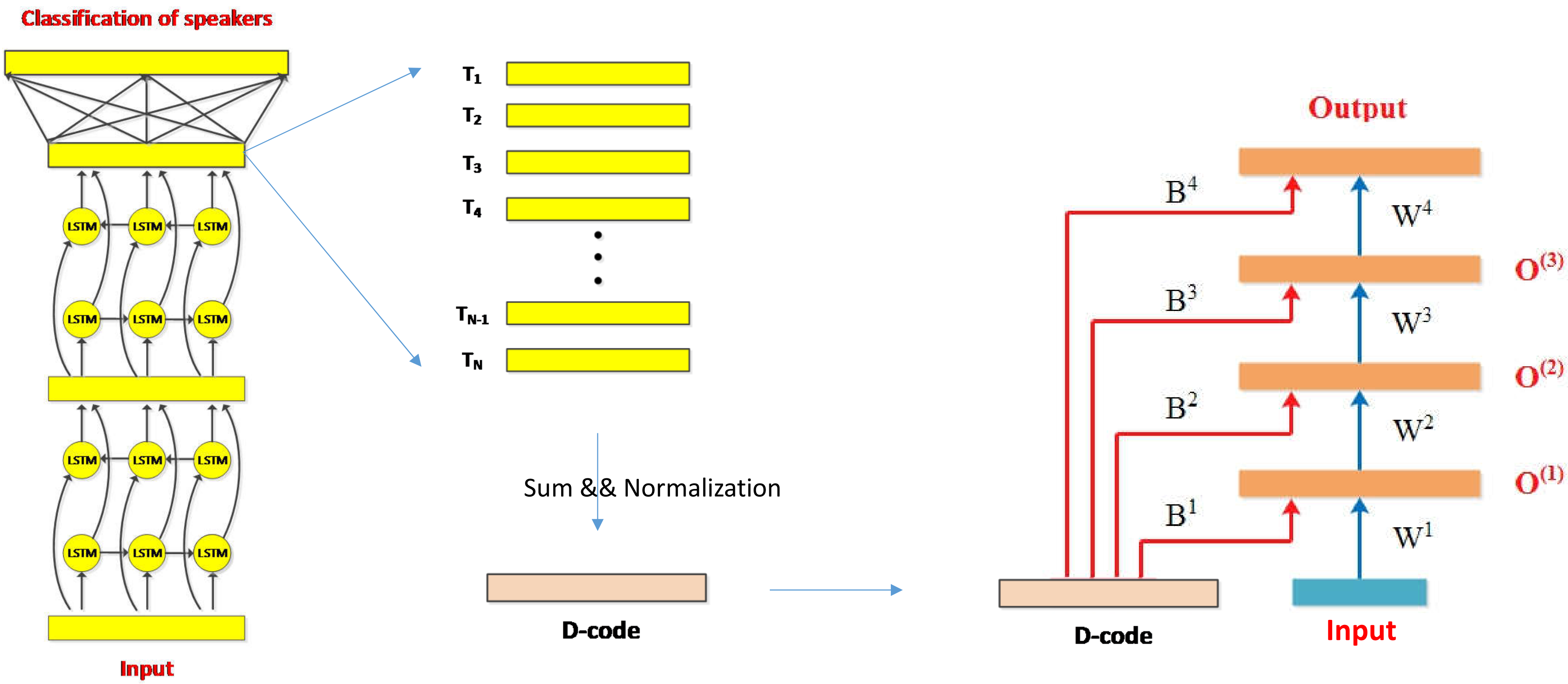
- Obtain final results with one-pass decoding.
- Improve accuracy when adaptation data is especially limited.

Outline

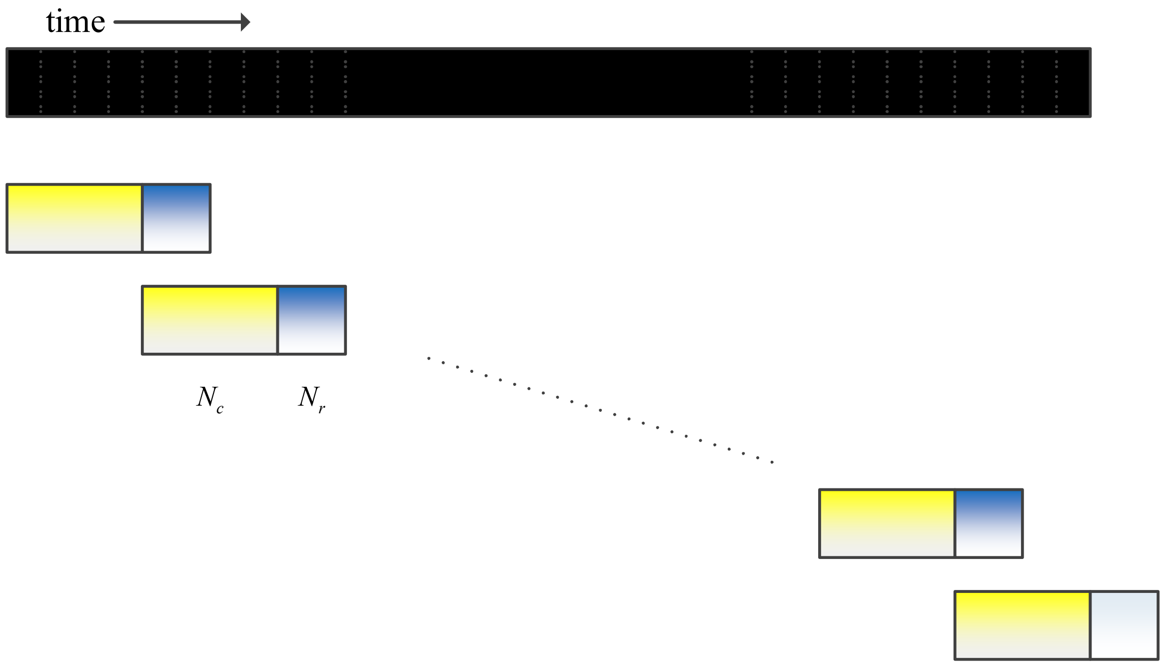
- Introduction
- **Method**
- Experiments
- Conclusions

Method

D-code extraction based BLSTM



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



Method

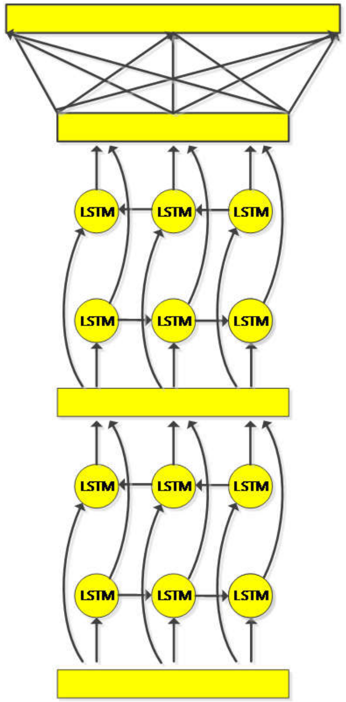
Speaker clustering BLSTM

Problem

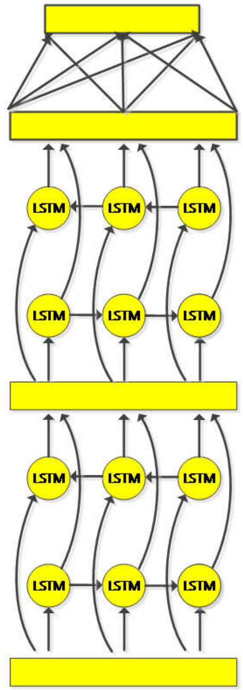
- Target speakers often 100,000+ in huge task.
- Sometimes even no speaker information.
- Implementing speaker classification with NNs often meets problem.

Method

Speaker clustering BLSTM



Hierarchical clustering based i-vector



- Training cluster speaker-BLSTM is fast.
- Improves ASR performance.

D-code interpolation

Problem

- WER/CER increases visibly when data is extreme limited (eg. 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

$$\bar{\boldsymbol{S}}_{\text{test}}=\frac{\alpha \boldsymbol{S}_{test}+(1-\alpha)\sum_{i=1}^{N}\beta_i \boldsymbol{S}_{train\,i}}{\alpha+(1-\alpha)\sum_{i=1}^{N}\beta_i},$$

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

Table1 Adaptation performance using different d-code extractions on a 3-layer ReLU-DNN.

Speaker-BLSTMs	WER(%)
baseline	15.6
BLSTM-1hid*200cell	14.2(9%)
BLSTM-1hid*400cell	14.2
BLSTM-1hid*1000cell	14.1
BLSTM-2hid*200cell	14.2
LSTM-1hid*200cell	14.8
LSTM-2hid*200cell	14.7

D-code from speaker-BLSTMs outperforms speaker-LSTMs and the size of speaker-BLSTMs has no conspicuous influence.

Experiments

Table2 Training time for speaker-BLSTMs and WER(%) of adaptation with different speaker cluster number.

Number of cluster	Training time(1 epoch)	WER(%)
4803(no clustering)	1.76h	14.2
400	0.83h	14.0
800	0.9h	13.9
1200	0.96h	13.8(2.8%)
1600	1h	14.0

speaker clustering not only speeds up the training of speaker-BLSTMs (about two times) but also benefits the ASR performance.

Experiments

Table3 WER(%) of D-code interpolation method on a 3-layer ReLU-DNN.

D-code	α				
	0	0.25	0.5	0.75	1
speaker-level	13.9	13.8	13.7	13.8	13.8
utterance-level	14.0	14.0	14.2	14.3	14.3

D-code interpolation improve the performance when data is especially limited.

Experiments

Table4 Comparison of different adaptation strategies on better baselines.

Models	Adaptation strategies	WER(%)	Decoding pass
ReLU-DNN(6x2048)	baseline	13.9	one
	d-code	12.7	one
	standard SAT-SC	12.7	two
	i-vector	13.0	one
hybrid BLSTM-RNN	baseline	13.0	one
	d-code	11.9	one
	standard SAT-SC	11.8	two
	i-vector	12.2	one

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.**

Reference

[1] Yu Zhang, et al. Highway Long Short-Term Memory RNNs for Distant Speech Recognition[J]. 2016.

Thank you!
Q&A

Shaofei Xue, 薛少飞

Shaofei.xsf@alibaba-inc.com