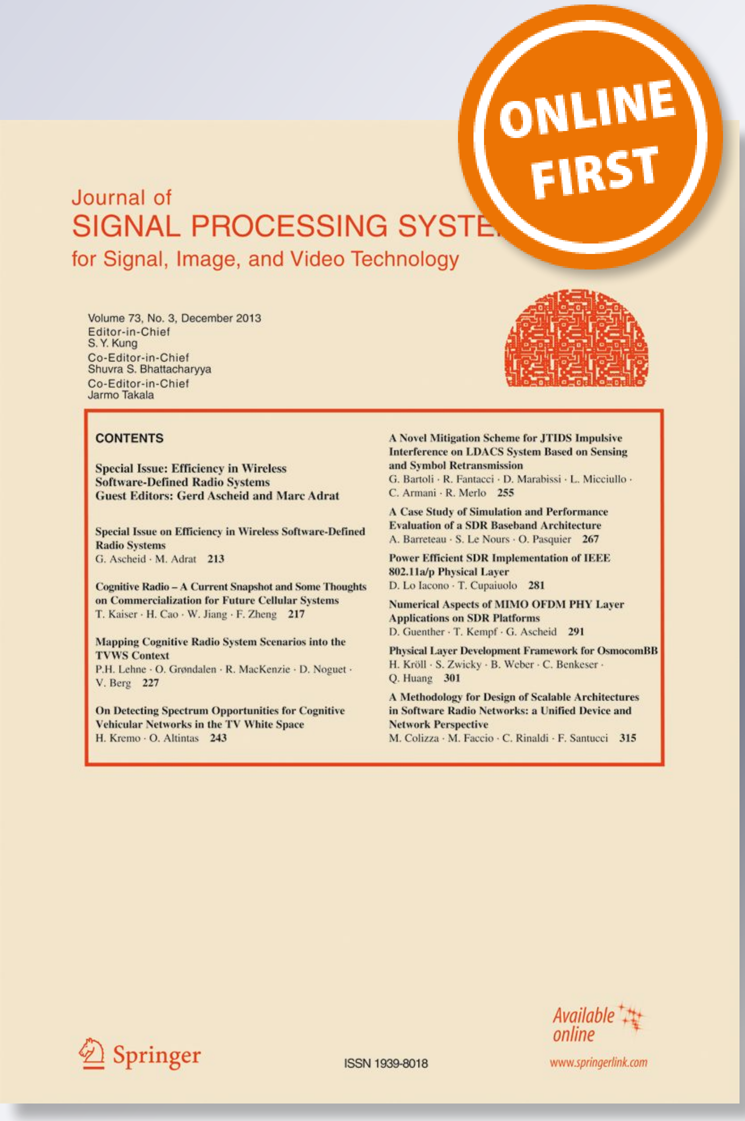
CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition

Jiangyan Yi, Zhengqi Wen, Jianhua Tao, Hao Ni & Bin Liu

Journal of Signal Processing Systems
for Signal, Image, and Video Technology
(formerly the Journal of VLSI Signal
Processing Systems for Signal, Image,
and Video Technology)

ISSN 1939-8018

J Sign Process Syst
DOI 10.1007/s11265-017-1291-1



ONLINE FIRST

Journal of
SIGNAL PROCESSING SYSTE
for Signal, Image, and Video Technology

Volume 73, No. 3, December 2013
Editor-in-Chief
S. Y. Kung
Co-Editor-in-Chief
Shuvra S. Bhattacharyya
Co-Editor-in-Chief
Jarmo Takala

Springer

ISSN 1939-8018

Available online
www.springerlink.com

Springer

J Sign Process Syst
DOI 10.1007/s11265-017-1291-1

CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition

Jiangyan Yi[1,2] · Zhengqi Wen[1] · Jianhua Tao[1,2,3] · Hao Ni[1,2] · Bin Liu[1]

Received: 15 January 2017 / Revised: 17 August 2017 / Accepted: 18 September 2017

Abstract This paper proposes a novel regularized adaptation method to improve the performance of multi-accent Mandarin speech recognition task. The acoustic model is based on long short term memory recurrent neural network trained with a connectionist temporal classification loss function (LSTM-RNN-CTC). In general, directly adjusting the network parameters with a small adaptation set may lead to over-fitting. In order to avoid this problem, a regularization term is added to the original training criterion. It forces the conditional probability distribution estimated from the adapted model to be close to the accent independent model. Meanwhile, only the accent-specific output layer needs to be fine-tuned using this adaptation method. Experiments are conducted on RASC863 and CASIA regional accented speech corpus. The results show that the proposed method obtains obvious improvement when compared with LSTM-RNN-CTC baseline model. It also outperforms other adaptation methods.

Keywords multi-accent · Mandarin speech recognition · LSTM-RNN-CTC · model adaptation · CTC regularization

1 Introduction

Accent is one of the key factors to result in poor performance of automatic speech recognition (ASR) system [1–3]. It is particularly severe for Mandarin ASR system with regional accented speech. The accented speech is determined by the speaker's native language or dialect [3, 4].

There are about seven major dialects in China: Guanhua dialect, Wu dialect, Xiang dialect, Gan dialect, Kejia dialect, Yue dialect and Min dialect [5]. There are many Chinese speakers who learn Mandarin as a second language. Their pronunciations are influenced by their native dialects. Statistics [6] show that over 79.6% of Mandarin speakers have regional accents, and 44.0% of them have heavy accents. Therefore, it is very common that Mandarin speech has accent. The performance of the Mandarin ASR systems is not well on accented speech, especially on multiple-accented speech [4].

To solve this problem, many adaptation methods have been proposed. These methods can be roughly classified into two categorizations: lexicon adaptation method [1, 3, 10–13] and model adaptation method [4, 5, 7–9, 14]. The former focuses on phonetic variations. The common methods are to extend phone set or augment pronunciation dictionary [10]. However, these methods increase lexical confusions [11]. Thus they do not gain obvious performance improvement. The pronunciation modeling technique proposed to reduce phonetic confusions. This modeling method is at either phone level or HMM

✉ Zhengqi Wen
zqwen@nlpr.ia.ac.cn

Jiangyan Yi
jiangyan.yi@nlpr.ia.ac.cn

Jianhua Tao
jhtao@nlpr.ia.ac.cn

Hao Ni
hao.ni@nlpr.ia.ac.cn

Bin Liu
liubin@nlpr.ia.ac.cn

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China

[2] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China

[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Published online: 23 September 2017

state level [12, 13]. The latter focuses on acoustic variations, the most direct way is to build acoustic models for each accent using a large amount of accent speech [14]. Previous studies have shown that the model adaptation method is more effective than the lexicon adaptation method. In this paper, we focus on the model adaptation method for multi-accent Mandarin speech in ASR task.

There are many literatures on the model adaptation for accented speech recognition. Previously, the simplest proposed adaptation method is maximum likelihood linear regression (MLLR). This method is based on GMM-HMM model. Huang et al. [15] proposed standard speaker MLLR adaptation to a Microsoft Whisper system that has been trained on speech from speakers living in the Beijing area. In [15, 16], MLLR is adapted not just to the single accented test speaker, but to a larger number of accented speakers. Research in [15–17] shows the effectiveness of MLLR or Maximum A Posteriori (MAP) adaptation on accented speech. Y. L. Zheng et al. [8] combined MLLR and MAP to perform accent adaptation for Shanghai-Accented Mandarin ASR task. Their experimental results show that this approach can improve the performance of accented speech recognition.

Recently, deep neural networks (DNNs) have become dominant methods for acoustic modeling in ASR system [18–20]. DNNs have layer-by-layer invariant and can extract high level representation features [21, 22]. Therefore, they can help improve the performance of ASR system on the accented speech. However, there are still large performance gap between the accented speech and the native speech for the deep neural network hidden Markov model (DNN-HMM) based acoustic model [23].

More recently, the multi-accent deep neural network with accent-specific top layer is proposed to gain improvement for recognizing foreign accented speech [23]. The method is inspired by the multilingual speech recognition using multitask learning technology [24]. Furthermore, they have been proved to be effective and efficient. Moreover, i-vectors have also been used to perform accent adaptation for Mandarin speech [25]. The results show that this method can achieve promising results. These methods have been conducted for DNN-HMM based acoustic model.

Most recently, long short term memory (LSTM) recurrent neural networks (RNNs) outperform the state-of-the-art DNN-HMM systems [26]. There are many literatures on LSTM RNN based speaker adaptation [27–29]. A few studies on accent adaptation are conducted based on LSTM RNN. It has been reported that accent-specific bottleneck features can improve the performance of multi-accent Mandarin ASR task [30].

However, the above models are hybrid models. These DNN-HMM or RNN-HMM based acoustic models are trained with a cross-entropy (CE) loss function. DNNs or RNNs are used to classify speech frames into clustered context-dependent (CD) states (i.e., senones). The frame-level senones are generated from GMM-HMM based model. Thus the training procedures of ASR are very complex. Graves et al. [31] introduce the connectionist temporal classification (CTC) loss function to infer speech-label alignments automatically. This CTC technique is further investigated in [32–35] on large-scale acoustic modeling tasks. These end-to-end acoustic models show promising results.

Naturally, we adopt LSTM RNN based acoustic model trained with a CTC loss function (LSTM-RNN-CTC) to estimate posteriors of initial and final (I/F) sequences. This method does not need alignments from GMM-HMM models. The results show that this model can achieve promising results and speed up decoding [36]. Nevertheless, adjusting the network parameters directly with a small adaptation data may lead to over-fitting [23–25].

In order to avoid over-fitting, this paper proposes a novel regularized adaptation method to improve the performance of multi-accent Mandarin speech recognition task. This method is based on LSTM-RNN-CTC model. This regularized method is inspired by KLD regularized adaptation with a CE loss function [37]. In [37], the method is used for speaker adaptation. A regularization term is added to the original training criterion. It forces the probability distribution over senones estimated from the adapted model to be close to speaker independent (SI) model. We apply this idea to accent adaptation with a CTC loss function. Thus this paper proposes a CTC regularized adaptation method for LSTM-RNN-CTC base acoustic model.

Furthermore, inspired by the method in [23], we also propose a multi-accent LSTM-RNN-CTC model with accent-specific output layer. This method only needs to adjust the parameters of the output layer of the accent independent (AI) model with the accent-specific adaptation set. The parameters of all the hidden layers of the AI model are shared for all accented speech.

Experiments are conducted on RASC863 [38] and CASIA [39] corpus. The results demonstrate that the proposed method gains 37.7%, 10.5%, 5.8%, 6.5% and 13.2% for Beijing (BJ), Shanghai (SH), Guangzhou (GZ), Chongqing (CQ) and Xiamen (XM) accent speech relative word error rate (WER) reduction against 260 hours LSTM-RNN-CTC based acoustic model. The results also show that the proposed method outperforms other adaptation methods.

Our main contributions are summarized as follows:

- A novel regularized adaptation method is proposed to improve the performance of multi-accent Mandarin speech recognition task. This method is called CTC regularized adaptation.
- A multi-accent LSTM-RNN-CTC model with accent-specific output layer is proposed for Mandarin speech recognition. The output labels of the LSTM-RNN-CTC model are initials and finals (I/F) sequences.

- The results show that the proposed CTC regularized adaptation method is effective. When the adaptation data is small, this method can help avoid over-fitting. The regularization weight is a hyper-parameter which can be adjusted according to the amount of the adaptation data.

The rest of this paper is organized as follows. In section 2, the proposed regularized adaptation method is introduced in detail. Section 3 describes the framework of our proposed model adaptation method. In section 4, experiments and results are presented at some length. Section 5 discusses the results. Section 6 concludes the paper.

2 Proposed Regularized Method

In this section, connectionist temporal classification (CTC) loss function is reviewed briefly at first. Then, the proposed CTC regularized model adaptation is introduced in detail.

2.1 Connectionist temporal classification

LSTM RNN with CTC can be used as a classifier by selecting the most probable label sequences for a given input sequence [31, 32].

Let S denotes a set of training samples. The input space X is the set of all sequences of m dimensional real value vectors. The output space Z is the set of all sequences over the alphabet L of labels. In general, each training sample in S is defined as a pair of sequences (x, z).

Formally, for an input sequence x of length T, define a LSTM RNN with m inputs and n outputs as a continuous map $f : y = f(x)$. y is the output sequence of the network. The length of y is T. x^t in x is the m dimensional inputs at time t and y^t in y is the n dimensional outputs at time t. Then $y^t_{\pi_t}$ is interpreted as the probability of observing label π_t at time t, which defines a distribution over the set L'^T of length T sequences:

$$p(\pi|x) = \prod_{t=1}^{T} y^t_{\pi_t}, \forall \pi \in L'^T \tag{1}$$

where the alphabet L' is defined as $L \cup \{blank\}$, the element of L'^T is referred as path and is denoted as π. π_t denotes the element π at time t.

B is defined as a many-to-one map: $L'^T \to L^{\le T}$, where L'^T is the set of all the paths π and $L^{\le T}$ is the set of possible output labels z. Then all blanks and repeated labels can be removed from the path π. For example, $B(c-cd-) = B(cc--cdd) = ccd$, $c-cd-$ and $cc--cdd$ are the paths π. $-$ denotes blank. ccd is the possible output labels z. So B^{-1} is defined as a one-to-many map: $L^{\le T} \to L'^T$, the inverse of B, such as $B^{-1}(ccd) = c-cd-$ or $B^{-1}(ccd) = cc--cdd$. Finally, the conditional probability of a given label sequences $z \in L^{\le T}$ is defined as the sum of the probabilities of all the paths π corresponding to it [31]:

$$p(z|x) = \sum_{\pi \in B^{-1}(z)} p(\pi|x) \tag{2}$$

The aim of maximum likelihood training is to simultaneously maximise the log probabilities of all the correct classifications in the training set. This means minimising the following objective function:

$$L(S) = -ln \prod_{(x,z)\in S} p(z|x) = -\sum_{(x,z)\in S} lnp(z|x) \tag{3}$$

where $(x, z) \in S$ denotes training samples. The training is carried out by back propagation through time (BPTT) algorithm.

2.2 Proposed CTC regularized adaptation

A direct method to adapt neural network is to fine-tune all the parameters of accent independent (AI) LSTM-RNN-CTC based acoustic model with adaptation sets. Nevertheless, it may distort the probability distribution of the AI model and cause over-fitting problem, especially if the adaptation set is small.

To avoid over-fitting, the network should be adapted conservatively. Thus we propose a CTC regularization method. The output labels z is initials and finals (I/F) sequences. So what estimated from the model are not the senone posteriors distribution but the I/Fs distribution. I/Fs form the fundamental elements in Mandarin pinyin. For example "间jian", 间 is a Chinese character. Jian is the pinyin of the Chinese character间. j is initial (I) and ian is final (F) of the pinyin.

The intuition behind the proposed method is that the conditional probability distribution over I/F sequences estimated from the adapted model should be close to the AI model.

Therefore, this constraint is realized by adding a penalty ρ

$\left(-\sum_{(x,z)\in S} lnp^{AI}(z|x)\right)$ as a regularization term to eq. (3). Thus we get the adaptation criterion as follow:

$$\hat{L}(S) = (1-\rho)L(S) + \rho\left(-\sum_{(x,z)\in S} lnp^{AI}(z|x)\right) \tag{4}$$

where $\ln p^{AI}(z|x)$ is the log probability of I/F label sequences estimated using the AI model and ρ is the regularization weight. This adaptation method is called as CTC regularized adaptation. Then, eq. (4) can be reorganized to eq. (5) and (6):

$$\hat{L}(S) = -\sum_{(x,z)\in S} \left((1-\rho)lnp(z|x) + \rho lnp^{AI}(z|x)\right) \tag{5}$$

$$\hat{L}(S) = -\sum_{(x,z)\in S} ln\hat{p}(z|x) \qquad (6)$$

By eq. (5) and (6), and removing the unrelated terms, we can define as follow:

$$ln\hat{p}(z|x) \triangleq (1-\rho)lnp(z|x) + \rho lnp^{AI}(z|x) \qquad (7)$$

where $\ln p(z|x)$ is the log probability of I/F sequences estimated from the model using the original training criterion with adaptation data. $ln\hat{p}(z|x)$ is a linear interpolation of the log probability $\ln p(z|x)$ and $\ln p^{AI}(z|x)$.

By comparing equation (3) with equation (5) and (6), we can see that applying the CTC regularization to the original training criterion $L(S)$ is equivalent to changing the log probability distribution from $\ln p(z|x)$ to $ln\hat{p}(z|x)$.

The CTC regularization constrains the log probability of the I/F sequences rather than model parameters themselves. Therefore, the normal BPTT algorithm can be directly used to adapt the AI model. Then, the only thing needs to be changed is the error signal at the output layer. The error signal is computed According to the log probability $ln\hat{p}(z|x)$ of the I/F sequences.

When the AI model is adapted with the adaptation set, the regularization weight ρ in equation (7) can be adjusted using a development set. When $\rho = 1$, it indicates that the AI model is trusted completely and all new information is ignored from the adaptation data. When =0, it indicates that the AI model is only used to initialize the adapted model and the information from the adaptation set is trusted completely. Intuitively, a small regularization weight ρ can be used for a large adaptation set. A large regularization weight ρ should be used for a small adaptation set.

3 Framework of the Proposed Method

This section describes the framework of the proposed CTC regularized adaptation method. The AI model should be trained using all kinds of accented speech at first. Then, the accent adaptation is performed based on the AI model with different accent adaptation data.

The AI model is based on LSTM-RNN-CTC. The input data of the AI model is the training set of all kinds of accented speech data. The output labels of the AI model are I/Fs, such as "d ai k ou y in d e y u y in sh i b ie zh un q ue l v z en m e y ang". The AI model is used to initialize the adapted model. Then the adapted model is fine-tuned with different adaptation sets for different accented speech. The network architecture of the AI model is depicted at the left of Fig. 1.

The accent adaptation is performed by starting from the AI model. This method only needs to adjust the parameters of the output layer of the AI model with the accent-specific adaptation set. The parameters of all the hidden layers of the AI model are shared for all accented speech. We do not need to store different models for all kinds of accented speech. We only need to store different parameters of the accent-specific output layers and the shared hidden layers. Thus the storage cost can be reduced for accent-specific models. The framework of the proposed accent adaptation method for multi-accent Mandarin speech recognition is depicted at the right of Fig. 1.

In Fig. 1, the AI model is trained to be as the initialized model at first. The AI model is trained with all kinds of accented speech data. Then accent adaptation is performed with different kinds of accented speech adaptation set. The output labels of the accented speech (*Accent 1*) are the same as other three accented speech (*Accent 2*, *Accent 3* and *Accent 4*). Different accented speech has different output layer. All accent-specific layers share the parameters of the hidden layers of the AI model. For *Accent 1*, the adaptation is performed to fine-tune the output layer of the AI model with the adaptation set of *Accent 1*. For *Accent 2*, the adaptation is performed to adjust the output layer of the AI model with the adaptation set of *Accent 2*. The adaptation set of each accent is a part of the training data of the AI model.

The proposed framework has two advantages. First, the number of I/Fs is very small compared with the number of senones. Thus training cost and decoding cost can be significantly reduced. Second, the number of senones may change with augmenting the adaptation data in the real ASR system, especially when the adaptation data is large. In general, the set of I/Fs is stationary for Mandarin speech.

4 Experiments

In this section, experimental data is described at first. Then experimental setup and baseline models are introduced. Finally, a series of experiments and results are described in detail.

4.1 Data description

Our experiments are conducted on RASC863 [38] and CASIA [39] regional accented speech corpus. RASC863 corpus is 863 annotated 4 regional accented speech, namely Shanghai (SH), Guangzhou (GZ), Chongqing (CQ) and Xiamen (XM). The corpus consists of spontaneous speech, read speech and selected dialectical words. CASIA corpus is developed by Institute of Automation, Chinese Academy of Sciences. The corpus consists of southern and northern accented speech.

There are five kinds of regional accented speech selected to form the experimental data. Four kinds of regional accented speech are selected from RASC863. Beijing (BJ) accented speech is selected from CASIA. The accented speech in our experiments is spoken by native residents from BJ, SH, GZ, CQ and XM respectively. BJ accented speech is very close to

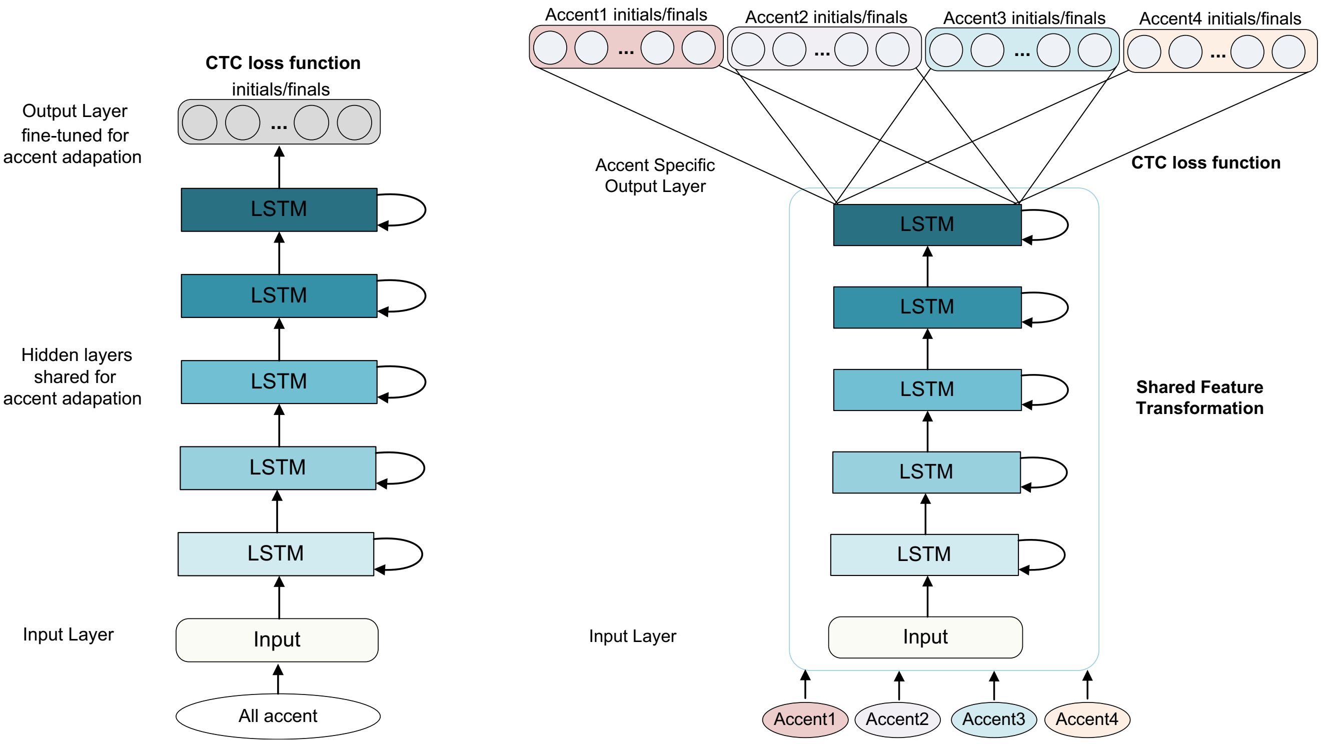


Figure 1 Framework of the proposed adaptation method. Left: Network architecture of accent independent (AI) model for Mandarin speech recognition. Right: Framework of multi-accent LSTM-RNN-CTC based model adaptation for speech recognition. Note: The labels of output layer are initials and finals which are same for different kinds of accented speech.

Mandarin speech while the others are seriously affected by Wu dialect, Yue dialect, CQ dialect and Min dialect. These four kinds of accented speech are very different from Mandarin speech.

In all our experiments, the experimental data contains three parts: training data, development data and test data. Moreover, training data, development data and test data are mutually exclusive. The parameters of the model are updated on the training data. The learning rate is adjusted on the development data. The training terminates when the performance fails to decrease by 0.1% between two successive epochs on the development data. The word error rate (WER) of all models is calculated on the test data.

For all AI models, the training data is about 260 hours. The development data is about 25 hours. The test data is about 14.7 hours. The statistics of the training data, the development data and the test data used to train AI models in our experiments are listed in Table 1.

For all adapted models, the adaptation data is selected from the training data and development data that are used to train the AI models. The adaptation data used to update the parameters of the adapted models is called *adaptation_tr* data. The adaptation data used to select adapted models, adjust the learning rate and the regularized weight is called *adaptation_cv* data. The *adaptation_tr* data is selected from the training data randomly for each accent in Table 1. The *adaptation_cv* data is selected from the development data randomly for each accent in Table 1.

There are three kinds of adaptation data used for accent adaptation: *Apt.1 k, Apt.10 k* and *Apt.tot*. These adaptation data are selected from the training data of Table 1. *Apt.1 k*

Table 1 Statistics of the training data, the development data and the test data used to train AI models in our experiments.

Accent	Training data		Development data		Test data	
	#utterances	#hours	#utterances	#hours	#utterances	#hours
BJ	70,996	55	7099	5.1	4004	3.1
SH	33,576	63	3357	6.2	1894	3.6
GZ	13,448	25	1344	2.4	758	1.4
CQ	30,835	67	3083	6.5	1738	3.8
XM	30,059	50	3005	4.8	1693	2.8
TOTAL	178,914	260	17,888	25	10,087	14.7

denotes that the *adaptation_tr* data contains 1000 utterances for each accented speech. *Apt.10 k* denotes that the *adaptation_tr* data contains 10,000 utterances for each accented speech. *Apt.tot* denotes that the *adaptation_tr* data contains all utterances for each accented speech. The statistics of these three kinds of adaptation data for conducting accent adaptation in our experiments are listed in Table 2.

The test data for the adapted models is the same as the test data for the AI models. This means that the test data in Table 1 is used in all our experiments.

4.2 Experimental setup

In our experiments, the CTC regularization method is implemented based on the eesen toolkit [33]. The open source Kaldi speech recognition toolkit [40] is used to compare the performance among different baseline models.

The sampling frequency of speech data is 16 KHz. The feature vector is 40-dimensional filter bank (FBANK) features plus their first and second order derivatives. The frame length is 25 ms and the frame shift is 10 ms. The features are normalized via mean subtraction and variance normalization on the utterance basis.

At first, we train the AI model using the training data of all five kinds of accented speech in Table 1. Then, we adjust the parameters of the output layer of the AI model for each accent using the *adaptation_tr* data of the corresponding kind of accented speech. The learning rate and the regularized weight are adjusted on the *adaptation_cv* data of the corresponding kind of accented speech. The adaptation procedure usually takes 1 or 2 iterations before the parameters of the accent-specific output layer converge. The training of all neural network based models is speeded up by multiple GPUs across different machines. This parallel implementation is based on Message Passing Interface (MPI).

The language model (LM) used in all experiments is a 3-g LM trained with the transcriptions of the training data about 13 M. However, it excludes the transcriptions of the test data. The vocabulary used has 80 K words. Our decoding is weighted finite-state transducers (WFST) based approach.

4.3 Baseline model

This section describes different modeling technique among Gaussian mixture model hidden Markov model (GMM-HMM), DNN-HMM, LSTM-RNN-HMM and LSTM-RNN-CTC.

These models are trained using the training data and the development data of all five kinds of accented speech data in Table 1. The training data is about 260 hours. The development data is about 25 hours. The GMM-HMM model has 90,000 Gaussian components and 18,251 senones optimized with the maximum likelihood estimation (MLE) procedure. The speaker-independent crossword triphones use the common 3-state topology. The DNN-HMM model uses a sliding context window of 11 frames. It is trained with 6 hidden layers and each layer has 1024 nodes. The output labels are 18,251 nodes which are identical to the senones generated from the GMM-HMM model.

The LSTM-RNN-HMM model uses a single frame as input. It has 4 stacked LSTM layers with projection, and each layer has 640 memory cells and 320 output units. The output labels are 18,251 nodes which are identical to the senones generated from the GMM-HMM model.

The LSTM-RNN-CTC model uses a single frame as input. It has 4 stacked LSTM layers with projection, and each layer has 640 memory cells and 320 output units. Particularly, it adopts a CTC loss function to infer the alignments between speech and target sequences. The output labels are I/Fs, such as such as "*y u y in sh i b ie blank*". There are 61 output nodes in the softmax layer. The number of I/Fs is 60. One of the output labels is *blank*. The output labels are independent.

For the hybrid models (DNN-HMM and LSTM-RNN-HMM) are trained with a CE loss function. They are trained using stochastic gradient decent (SGD). DNN-HMM, LSTM-RNN-HMM and LSTM-RNN-CTC are trained using 4 GPUs across 2 machines. The initial learning rate of DNN-HMM and LSTM-RNN-HMM is set to 2×10^{-5}. For LSTM-RNN-

Table 2 Statistics of the three kinds of the adaptation data for conducting accent adaptation in our experiments.

Accent	Apt.1 k		Apt.10 k		Apt.tot	
	#adaptation_tr utterances	#adaptation_cv utterances	#adaptation_tr utterances	# adaptation_cv utterances	#adaptation_tr utterances	# adaptation_cv utterances
BJ	1000	100	10,000	1000	70,996	7099
SH	1000	100	10,000	1000	33,576	3357
GZ	1000	100	10,000	1000	13,448	1344
CQ	1000	100	10,000	1000	30,835	3083
XM	1000	100	10,000	1000	30,059	3005

CTC, the initial learning rate and momentum are set to 5×10^{-7} and 0.9 respectively.

Table 3 and Fig. 2 list word error rate (WER) for different acoustic models on five kinds of accented speech test sets. The results demonstrate that the LSTM-RNN-CTC model outperforms DNN-HMM and GMM-HMM on all test data. The LSTM-RNN-HMM model obtains the best performance. The reason is that RNN can model long-term dependency via its recurrent structure.

Although the LSTM-RNN-HMM model obtains the best performance, the loss function of this model is CE. In this paper, the regularized adaptation method is based on the CTC loss function. Therefore, we select the LSTM-RNN-CTC model as our baseline model. The baseline model is trained using the training data of all five kinds of accented speech. The baseline model is used as the AI model for accent adaptation. The CTC regularization adaptation is evaluated based on the AI model in the rest of experiments. The adaptation is conducted on the *adaptation_tr* data for each accent. In addition, all the hyper parameters are adjusted on the *adaptation_cv* data, such as the regularized weight ρ and the learning rate etc.

4.4 Regularization weight of adaptation

Starting from the AI model (LSTM-RNN-CTC), two groups of experiments are conducted to find out how the improvement is influenced by different CTC regularization weights with different accented speech. The adapted models are evaluated using the test data of five kinds of accented speech in Table 1.

One group of experiments are designed to adapt model with different adaptation data through setting different regularization weights, such as [0 0.25]. There are three kinds of adaptation data used to conduct experiments: *Apt.1 k*, *Apt.10 k* and *Apt.tot*. Figure 3 depicts the WER of three kinds of adaptation data for BJ accent with different CTC regularization weights. Figure 4 depicts the WER of three kinds of adaptation data for SH accent with different CTC regularization weights.

The other group of experiments are conducted to adjust the model with *Apt.10 k* adaptation data for GZ, CQ and XM accent with different regularization weights. The regularized weights are set to be within [0 0.25]. The results of the experiments are shown in Fig. 5.

Table 3 Word error rate (WER %) of four acoustic models on five accented speech test sets.

Accent	GMM-HMM	DNN-HMM	LSTM-RNN-HMM	LSTM-RNN-CTC
BJ	4.18	2.78	0.51	0.61
SH	7.36	5.84	3.82	4.24
GZ	10.72	7.56	3.05	3.27
CQ	7.50	4.72	2.61	2.76
XM	9.54	7.04	2.95	3.25
TOTAL	7.15	5.10	2.01	2.59

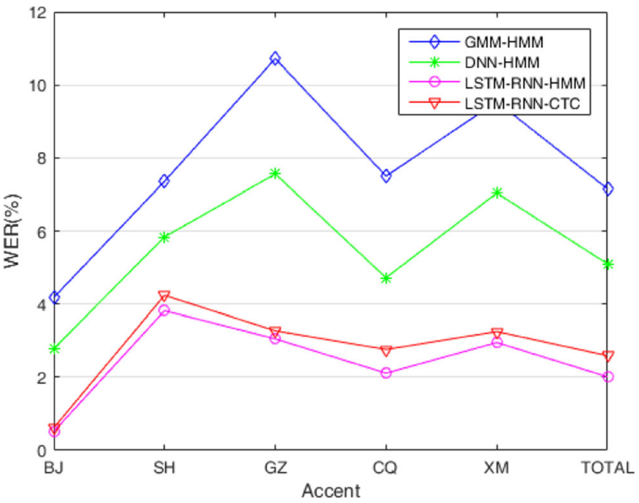


Figure 2 Word error rate (WER %) of four acoustic models on five accented speech test sets.

We conduct these experiments for 10 trials. The adaptation data (*Apt.1 k, Apt.10 k and Apt.tot*) is selected randomly from the training data in Table 1 for each trial. So the adaptation data may be different between two trails for one experiment. The result of the experiment is the average of the results from 10 trails.

The orange dashed line is the WER of the AI model for each accent in Figs. 3, 4, and 5. We get the following observations by studying in Fig. 3 and Fig. 4 together:

- When the regularization weight is $\rho = 0$, the WER curves of the adapted model are above the orange dashed line both for BJ and SH accent on the adaptation data *Apt.1 k*. It means that directly adjusting the parameters of the AI model is prone to cause over-fitting with a small adaptation set.
- When the regularization weight is set to (0, 0.25), the WER curves of the adapted model are all below the orange dashed line both for BJ and SH accent on three adaptation data.

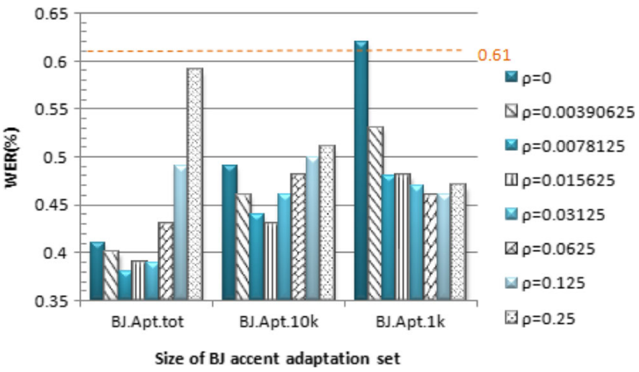


Figure 3 WER (%) of Apt.1 k, Apt.10 k and Apt.tot adaptation data for BJ accent with different CTC regularization weights.

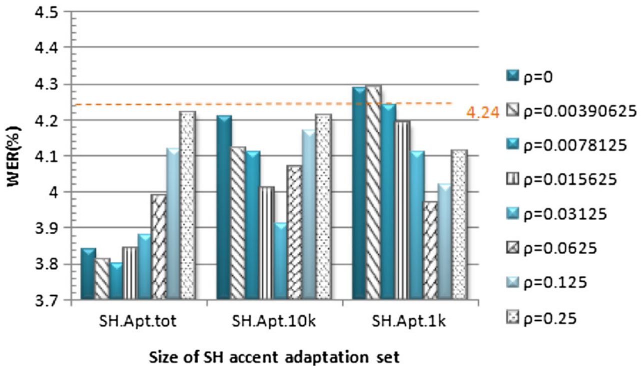


Figure 4 WER (%) of Apt.1 k, Apt.10 k and Apt.tot adaptation data for SH accent with different CTC regularization weights.

- When the adaptation data is *Apt.tot*, the best WER reduction can be obtained with a smaller ρ. For BJ accent, it can gain 37.7% relative WER reduction with $\rho = 0.0078125$. For SH accent, it can achieve best relative WER reduction of 10.4% with $\rho = 0.0078125$.
- When the adaptation data is *Apt.1 k*, the best WER reduction can be obtained with a larger ρ. For BJ accent, it can gain 24.6% relative WER reduction with $\rho = 0.125$. For SH accent, it can achieve the relative WER reduction of 6.4% with $\rho = 0.0625$.
- When the adaptation data is *Apt.tot* or *Apt.10 k*, the WER curves of the adapted model are all below the orange dashed line both for BJ and SH accent on three kinds of adaptation data with different regularization weight ρ.

From Fig. 5, when we perform the experiments of the adaptation on the adaptation data *Apt.10 k*, we can see that:

- The WER curves of the adapted model are all below the orange dashed line for GZ, CQ and XM accent with different regularization weight ρ.
- The best performance for GZ, CQ and XM accent is achieved when the weight ρ is set to 0.03125, 0.03125 and 0.0625 respectively.
- The regularized weight is robust for all kinds of accented speech.

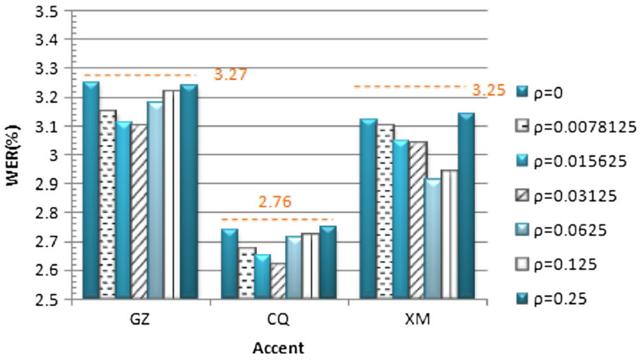


Figure 5 WER (%) of Apt.10 k adaptation data for GZ, CQ and XM accent with different CTC regularization weights.

4.5 Amount of accent adaptation data

This series of experiments investigate how the performance is affected by different amount of adaptation data with CTC regularized adaptation method. The proposed regularized adaptation is not only compared with the baseline model, but also compared with the other adaptation methods. The other adaptation methods are as follows.

L2-Reg: L2 regularization adaptation is proposed by Li et al. [41]. They produce an adapted model by using a regularizer that penalizes distance from unadapted model. In this paper, we generate an accent dependent model by using a L2 regularizer to penalize distance from the accent dependent baseline model.

LHN: It is a linear transform adaptation [27, 42]. This method is performed by inserting an accent-specific linear layer on top of each LSTM layer to transform the hidden activations. The linear layer is inserted at different hidden layers which denotes LHN-N ($N = 1, 2, 3, 4$). For example, LHN-1 denotes the linear layer inserted on top of the first LSTM layer in the baseline model. We only adjust the linear layer using adaptation utterances.

The experiments are conducted using the adaption data with different utterances (*Apt.1 k, Apt.10 k* and *Apt.tot*) for BJ, SH, GZ, CQ and XM accent. The adapted models are evaluated using the test data of five accented speech.

These experiments are conducted for 10 trials. The adaptation data may be different between two trails for one experiment. To our proposed adaptation, we select the best result from the adapted model with different regularized weight for each trial. To L2-Reg adaptation, the value of L2 is set to 0.002, 0.0001, 0.000002 relative to the best result with *Apt.1 k, Apt.10 k, Apt.tot* adaptation utterances respectively. To LHN adaptation, the training is early stopped by 2 iterations. The results of the experiments are the average of the best results from 10 trails. The results of the experiments are summarized in Table 4. From Table 4, we can get some observations:

- Compared with the baseline model (the AI model), all the adapted models outperform the baseline model for all five kinds of accent.
- The best relative WER reduction is obtained for BJ accent even with a small adaptation data.
- When the adaptation data is *Apt.1 k*, the relative WER reduction of CQ accent is not obvious.
- When the adaptation data is *Apt.tot*, the relative WER reduction of GZ accent is the smallest. It is because that the total utterances of GZ accented speech data is the smallest.
- The relative WER reduction of XM accent increases from 1.5% to 13.2% with the adaptation data increased from 1 k to the total utterances.

Table 4 WER (%) and relative WER reduction (% in parentheses) of the adaption data with different utterances(Apt.1 k, Apt.10 k and Apt.tot) for five kinds of accent test data.

Accent	Baseline	Amount	L2-Reg	LHN-1	LHN-2	LHN-3	LHN-4	Proposed
BJ	0.61	Apt.1 k	0.58(4.9)	0.69(−13.1)	0.65(−6.6)	0.63(−3.3)	0.62(−1.6)	**0.46(24.6)**
		Apt.10 k	0.51(16.4)	0.58(4.9)	0.54(11.5)	0.52(14.8)	0.49(19.7)	**0.41(32.8)**
		Apt.tot	0.46(24.6)	0.49(19.7)	0.46(24.6)	0.45(26.2)	0.41(32.8)	**0.38(37.7)**
SH	4.24	Apt.1 k	4.23(0.2)	4.47(−5.4)	4.39(−3.5)	4.32(−1.9)	4.29(−1.2)	**3.97(6.4)**
		Apt.10 k	4.12(2.8)	4.19(1.2)	4.16(1.9)	4.15(2.1)	4.12(2.8)	**3.91(7.8)**
		Apt.tot	3.91(7.8)	3.97(6.4)	3.86(9.0)	3.81(10.1)	3.81(10.1)	**3.80(10.4)**
GZ	3.27	Apt.1 k	3.25(0.6)	3.59(−9.8)	3.48(−6.4)	3.43(−4.9)	3.42(−4.6)	**3.17(3.1)**
		Apt.10 k	3.18(2.8)	3.28(−0.3)	3.25(0.6)	3.21(1.8)	3.16(3.4)	**3.10(5.2)**
		Apt.tot	3.14(4.0)	3.19(2.4)	3.12(4.6)	3.09(5.5)	3.08(5.8)	**3.08(5.8)**
CQ	2.76	Apt.1 k	2.75(0.4)	2.91(−5.4)	2.89(−4.7)	2.85(−3.3)	2.79(−1.1)	**2.74(0.7)**
		Apt.10 k	2.65(4.0)	2.71(1.8)	2.69(2.5)	2.65(4.0)	2.64(4.3)	**2.62(5.1)**
		Apt.tot	2.62(5.1)	2.75(0.4)	2.75(0.4)	2.66(3.6)	2.60(5.8)	**2.58(6.5)**
XM	3.25	Apt.1 k	3.24(0.3)	3.33(−2.5)	3.31(−1.8)	3.28(−0.9)	3.27(−0.6)	**3.20(1.5)**
		Apt.10 k	3.05(6.2)	3.11(4.3)	3.11(4.3)	3.06(5.8)	3.04(6.5)	**2.91(10.5)**
		Apt.tot	2.94(9.5)	3.02(7.1)	3.01(7.4)	2.94(9.5)	2.92(10.2)	**2.82(13.2)**

- The adaptation model can gain better performance with augmenting the adaptation data.
- The LHN method can obtain better performance than the L2-Reg method with the large adaptation set.
- For the LHN adaptation, we can achieve the best performance when adapting the linear layer at the top hidden layer. But there is very little gain by adapting the linear transformation at the bottom hidden layer. The conclusion is consistent to the result in [27].
- Compared with the L2-Reg and LHN method, all the proposed adapted models outperform the other adapted model. It is easy to cause over-fitting when performing adaptation using LHN method using the small adaptation data. However, the proposed adaptation can avoid this over-fitting problem.

4.6 Adaptation based on LSTM-RNN-HMM model

In this section, the proposed regularized adaptation is compared with the KLD regularized adaptation [37] base on LSTM-RNN-HMM model. The baseline model is LSTM-RNN-HMM model.

The KLD adaptation is started from the AI baseline model. We conduct these experiments for 10 trials. The adaptation data (*Apt.1 k, Apt.10 k and Apt.tot*) is selected randomly from the training data in Table 1 for each trial. So the adaptation data may be different between two trails for one experiment. We also select the best result from the adapted model with different regularized weight for each trial. The result of the experiment is the average of the results from 10 trails. The regularized weights are set to be within [0 0.25]. The results of the experiments are shown in Table 5.

The relative WER reduction of the proposed adaptation and KLD adaptation are compared in Figs. 6, 7, and 8. The Fig. 6 denotes that the model is adapted with *Apt.1 k* utterances. The Fig. 7 denotes that the model is adapted with *Apt.10 k* utterances. The Fig. 8 denotes that the model is adapted with *Apt.tot* utterances.

From Table 5 and Table 4, we can find that both the proposed and the KLD adaptation can obtain improvement when compared with the AI baseline model.

From Fig. 6 to Fig. 8, we can see that the proposed adaptation can obtain more performance gain over the KLD adaptation based on the LSTM-RNN-HMM model.

5 Discussions

In summary, the above experiments explore how the improvement is affected by the regularization weight, the amount of adaptation data and the accent of adaptation data. The results

Table 5 WER (%) and relative WER reduction (% in parentheses) of the adaption data with different utterances(Apt.1 k, Apt.10 k and Apt.tot) for the KLD adaptation.

Acc- ent	Base- line	Apt.1 k	Apt.10 k	Apt.tot
BJ	0.51	0.40(21.6)	0.38(25.5)	0.36(29.4)
SH	3.82	3.71(2.9)	3.60(5.8)	3.55(7.1)
GZ	3.05	2.97(2.6)	2.90(4.9)	2.85(6.6)
CQ	2.61	2.61(0.0)	2.58(1.1)	2.54(2.7)
XM	2.95	2.91(1.4)	2.82(4.4)	2.74(7.1)

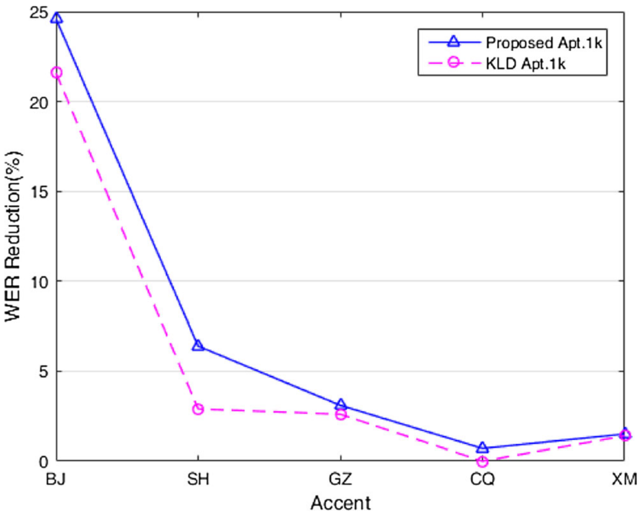


Figure 6 Curves of relative WER reduction (%) of the adaption data with Apt.1 k utterances for five accent test data.

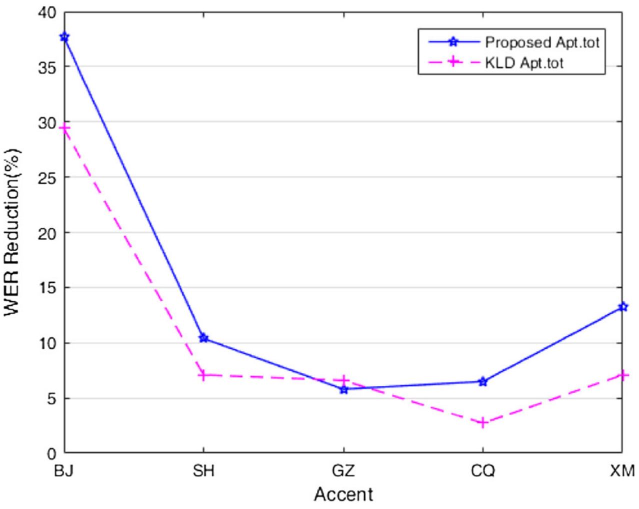


Figure 8 Curves of relative WER reduction (%) of the adaption data with Apt.tot utterances for five accent test data.

show that the proposed method is effective. We make some interesting observations as follow.

The CTC regularized adaptation method can help avoid over-fitting with a small adaptation data. The regularization weight is set to zero, which means that directly adjusting the parameters of the AI model. It causes over-fitting with a small adaptation data. When the regularization weight is set to (0.0.25), the WER of the adapted model reduces. It is because that the proposed regularized method makes the conditional probability distribution over I/F sequences estimated from the adapted model to be not far from the AI model.

The regularization weight should be adjusted with the amount of the adaptation data. A large regularization weight ρ should be set for a small adaptation data to avoid over-fitting. A small regularization weight ρ can be set for a large adaptation data. If the adaptation data is smaller, the AI model will get more trust than the new knowledge from the adaptation data. If the adaptation data is larger, the AI model will get less trust than the new information from the adaptation data.

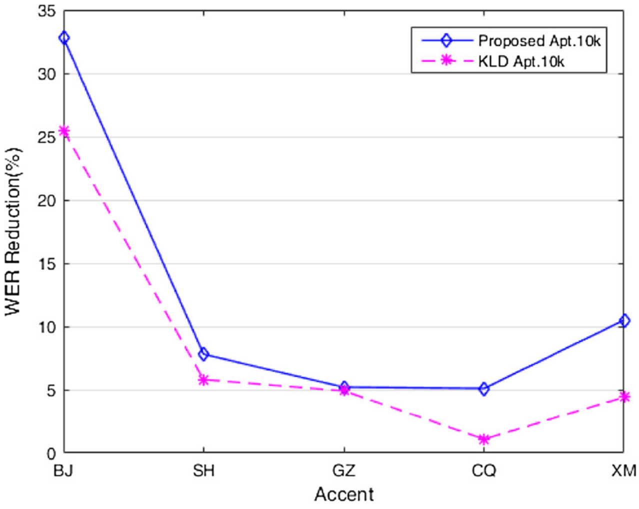


Figure 7 Curves of relative WER reduction (%) of the adaption data with Apt.10 k utterances for five accent test data.

Better performance can be achieved with increasing the adaptation data. When augmenting the adaptation data of BJ, SH, CQ and XM accent, the relative WER reduction is obvious. However, when the adaptation data is *Apt.tot*, the relative WER reduction of GZ accent is the smallest. It is because that the total utterances of GZ accented speech data is the smallest.

The proposed method also outperforms the L2-Reg and LHN adaptation. It is easy to cause over-fitting when performing adaptation using LHN method using the small adaptation data. However, the proposed adaptation can avoid this over-fitting problem.

Pronunciation dictionary is a key factor to affect the performance of the accent adaptation. The pronunciation dictionary used in our experiments is of Mandarin. The pronunciation dictionaries of Wu dialect, Yue dialect and Min dialect are a pretty different from Mandarin. The pronunciation dictionary of CQ dialect is similar to Mandarin. So when the adaptation data is small, the relative WER reduction of SH, GZ and XM accent are improved obviously while the improvement of CQ accent are moderate. In fact, articulation of BJ dialect is very similar to Mandarin. Therefore, the performance of the AI model for BJ accent is very high. So the best relative WER reduction is obtained for BJ accent even with a small adaptation data.

6 Conclusions

This paper proposes a novel regularized adaptation method for LSTM-RNN-CTC based acoustic model to improve the performance of multi-accent Mandarin ASR task. The basic idea of this method is that the distribution over I/F sequences estimated from the adapted model should be close to the distribution estimated from the AI model. This constraint is realized by adding a regularization to the original training criterion. Meanwhile, hidden layers should not be adjusted, but only the accent-specific output layer needs to be fine-tuned using the proposed CTC regularized method. Experiments are conducted on RASC863 and CASIA corpus. The results show that the proposed method gains 37.7%, 10.5%, 5.8%, 6.5% and 13.2% for BJ, SH, GZ, CQ and XM accent speech WERR against 260 hours LSTM-RNN-CTC based acoustic model. The results show that the proposed the CTC regularized adaptation method is effective. When the adaptation data is small, this method can help avoid over-fitting. The regularization weight should be adjusted with the amount of the adaptation data. Better performance can be achieved with increasing the adaptation data. Pronunciation dictionary is a key factor to affect the performance of the accent adaptation. The results also show that the proposed method outperforms other adaptation methods. In future studies, we are going to apply the proposed method to other different tasks, such as speaker adaptation, environment adaptation.

Acknowledgements This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386, No. 61305003), the Strategic Priority Research Program of the CAS (GrantXDB02080006) and the Major Program for the National Social Science Fund of China (13&ZD189).

References

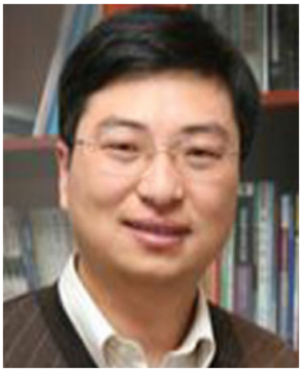
1. Huang, C., Chen, T., & Chang, E. (2004). Accent Issues in Large Vocabulary Continuous Speech Recognition. *Int J Speech Technol, 7*(2), 141–153.
2. Wang, Z., Schultz, T., & Waibel, A. (2013). Comparison of Acoustic Model Adaptation Techniques on Non-native Speech. In the Proceedings of the 2013 I.E. International Conference on Acoustics, Speech, and Signal Processing (ICASSP).
3. Arslan, L. M., & Hansen, J. L. (1997). A study of the temporal features and frequency characteristics in American english foreign accent. *Journal of the Acoustical Society of America, 102*(1), 28–40.
4. Liu, Y., & P. Fung (2006). Multi-accent Chinese Speech Recognition. In the Proceedings of Interspeech.
5. Fung, P., & Liu, Y. (Nov. 2005). Effects and Modeling of Phonetic and Acoustic Confusions in Accented Speech. *J Acoust Soc Amer, 118*(4), 3279–3293.
6. Leading Group Office of Survey of Language Use in China (2006). *In survey of language use in China*. Beijing: Yu Wen Press (in Chinese).
7. Davis S. B., & Mermelstein, P. (2013) Reliable Accent-Specific Unit Generation With Discriminative Dynamic Gaussian Mixture Selection for Multi-Accent Chinese Speech Recognition. *IEEE Trans Acoustics Speech Signal Process, 21* (10), 2073–2084.
8. Zheng, Y. L., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., & Yoon, S. (2005). Accent Detection and Speech Recognition for Shanghai-accented Mandarin. In the Proceedings of Interspeech.
9. Vergyri, D., Lamel, L., & Gauvain, L. (2010). Automatic Speech Recognition of Multiple Accented English Data. In the Proceedings of Interspeech.
10. Ding, G. H. (2008). Phonetic Confusion Analysis and Robust Phone Set Generation for Shanghai-Accented Mandarin Speech Recognition. In the Proceedings of Interspeech.
11. Fosler-Lussier, E., Amdal, I., & Kuo, H.-K. J. (2005). A Framework for Predicting Speech Recognition Errors. *Speech Communication, 46*(2), 153–170.
12. Fosler-Lussier, E. (1999). Dynamic Pronunciation Models for Automatic Speech Recognition. Ph.D. dissertation, Int. Comput. Sci. Inst., Berkeley, CA, USA.
13. Hain, T., & Woodland, P. C. (1999). Dynamic HMM Selection for Continuous Speech Recognition. In Proc. Eurospeech, pp. 1327–1330.
14. V. Fisher et al. (1998). Speaker-Independent Upfront Dialect Adaptation in A Large Vocabulary Continuous Speech Recognition. In Proc. Int. Conf. Spoken Lang. Process.
15. Wang, Z., Schultz, T., & Waibel, A. (2003). Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech. In ICASSP 2003. IEEE, pp. 540–543.
16. Mayfield Tomokiyo, L., & Waibel, A. (2001). Adaptation Methods for Non-Native Speech," in Proceedings of Multilinguality in Spoken Language Processing, Aalborg.
17. Huang, C., Chang, E., Zhou, J., & Lee, K.-F. (2000). Accent Modeling Based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition. In ICSLP 2000, Beijing, pp. 818–821.
18. Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. *IEEE Trans Audio Speech Lang Process, 1*(1), 33–42.
19. Seide, F., Li, G., & Yu, D. (2012). Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In the Proceedings of Interspeech.
20. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., NSainath, T., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process Mag, 29*(6), 82–97.
21. Yu, D., Seltzer, M., Li, J., Huang, J., & Seide, F. (2013). Feature learning in Deep Neural Networks - Studies on Speech Recognition Tasks. In the Proceedings of 2013 International Confernece on Learning Representation.
22. Goodfellow, I. J., Le, Q. V., Saxe, A. M., Lee, H., & Ng, A. Y. (2009). Measuring Invariances in Deep Networks. Advances in Neural Information Processing Systems (NIPS) 22.
23. Huang, Y., Yu, D., Liu, C. J., & Gong, Y. F. (2014). Multi-Accent Deep Neural Network Acoustic Model with Accent-Specific Top Layer Using the KLD-Regularized Model Adaptation. In the Proceedings of Interspeech.
24. Huang, J., Li, J., Yu, D., Deng, L., & Gong, Y. F. (2013). Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers. In the Proceedings of the 2013 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
25. Chen, M. M., Yang, Z. Y., Liang, J. Z., Li, Y. P., Liu, W. J. (2015). Improving Deep Neural Networks Based Multi-Accent Mandarin Speech Recognition Using I-Vectors and Accent-Specific Top layer. In the Proceedings of Interspeech.
26. Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. In the Proceedings of Interspeech.
27. Liu, C., Wang, Y., Kumar, K., & Gong, Y. F. (2016). Investigations on Speaker Adaptation of LSTM RNN Models for Speech

Recognition. In the Proceedings of the 2016 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

28. Huang, Z., Tang, J., Xue, S., & Dai, L. (2016). Speaker Adaptation of RNN-BLSTM for Speech Recognition Based on Speaker Code. In the Proceedings of the 2016 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
29. Tan, T., Qian, Y., Yu, D., Kundu, S., & Lu, L. (2016). Speaker-Aware Training of LSTM-RNNs for Acoustic Modelling. In the Proceedings of the 2016 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
30. Yi, J., Ni, H., Wen, Z. H., & Tao, J. (2016). Improving BLSTM RNN Based Mandarin Speech Recognition Using Accent Dependent Bottleneck Features. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.
31. Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In ICML, Pittsburgh, USA.
32. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech Recognition With Deep Recurrent Neural Networks. In the Proceedings of the 2013 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6645–6649.
33. Graves, A., & Jaitly, N. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1764–1772.
34. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deepspeech: Scaling up End-To-End Speech Recognition. arXiv preprint arXiv:1412.5567.
35. Hannun, A. Y., Maas, A. L., Jurafsky, D., & Ng, A. Y. (2014). First-Pass Large Vocabulary Continuous Speech Recognition Using Bi-Directional Recurrent DNNs. arXiv preprint arXiv: 1408.2873.
36. Miao, Y. J., Gowayyed, M. & Metze, F. (2015). EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding. In the Proceedings of ASRU.
37. Yu, D., Yao, K., Su, H., Li, G., & Seide, F. (2013). KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition. In the Proceedings of the 2013 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
38. (2003). RASC863: 863 annotated 4 regional accent speech corpus. Chinese Academy of Social Sciences. Available: http://www.chineseldc.org/doc/CLDC-SPC-2004-005/intro.htm.
39. (2003). CASIA: CASIA northern accent speech corpus. Chinese Academy of Sciences. Available: http://www.chineseldc.org/doc/CLDC-SPC-2004-015/intro.htm.
40. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y. M., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi SpeechRecognition Toolkit. In the Proceedings of ASRU.
41. Li, X., & Bilmes, J. (2006). Regularized adaptation of discriminative classifiers. In the Proceedings of the 2013 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
42. Miao, Y., Metze, F. (2015). On Speaker Adaptation of Long Short-Term Memory Recurrent Neural Networks. In the Proceedings of Interspeech.

Jiangyan Yi is currently a PhD Candidate at University of Chinese Academy of Sciences (UCAS), Beijing. She received her the M.A. degree from Graduate School of Chinese Academy of Social Sciences (CASS), Beijing, in 2010. She was a senior R&D engineer at Alibaba Group, Beijing, during 2011 to 2014. She current research interests include speech recognition, distributed computing and deep learning.

Zhengqi Wen received his the B.S. degree from University Of Science and Technology of China (USTC), Heifei, in 2008 and received his the Doctor degree from Chinese Academy of Sciences (CAS), Beijing, in 2013. He is currently Associate Researcher in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He current research interests include speech recognition and speech synthesis.



Jianhua Tao received his PhD from Tsinghua University in 2001, and got his M.S. from Nanjing University in 1996. He is currently a Professor in NLPR, Institute of Automation, Chinese Academy of Sciences. His current research interests include speech synthesis and coding methods, human computer interaction, multimedia information processing and pattern recognition. He has published more than eighty papers on major journals and proceedings including IEEE Trans. on ASLP, and got several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He serves as the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC etc. He also serves as the steering committee member for IEEE Transactions on Affective Computing, associate editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions, Deputy Editor-in-chief for Chinese Journal of Phonetics.

Hao Ni received his the B.S. degree from Wuhan University of Technology, Wuhan, in 2014 and received his the M.S. degree from University of Chinese Academy of Sciences (UCAS), Beijing, in 2017. He current research interests include language modeling and natural language processing.

Bin Liu received his the B.S. degree and the M.S. degree from Beijing institute of technology (BIT), Beijing, in 2007 and 2009 respectively. He received his the Doctor degree from Chinese Academy of Sciences (CAS), Beijing, in 2015. He is currently Research Assistant in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He current research interests include speech coding and speech enhancement.