ADVERSARIAL MULTILINGUAL TRAINING FOR LOW-RESOURCE SPEECH RECOGNITION

Jiangyan Yi¹,², Jianhua Tao¹,²,³, Zhengqi Wen¹, Ye Bai¹,²

¹National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
²School of Artificial Intelligence, University of Chinese Academy of Sciences, China
³CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, China

ABSTRACT

This paper proposes an adversarial multilingual training to train bottleneck (BN) networks for the target language. A parallel shared-exclusive model is also proposed to train the BN network. Adversarial training is used to ensure that the shared layers can learn language-invariant features. Experiments are conducted on IARPA Babel datasets. The results show that the proposed adversarial multilingual BN model outperforms the baseline BN model by up to 8.9% relative word error rate (WER) reduction. The results also show that the proposed parallel shared-exclusive model achieves up to 1.7% relative WER reduction when compared with the stacked share-exclusive model.

Index Terms— Speech recognition, low-resource, deep neural networks, bottleneck features, adversarial multilingual training

1. INTRODUCTION

Multilingual training is an effective approach to improve the performance of automatic speech recognition (ASR) systems for low-resource languages [1, 2, 3].

Previously, deep neural network (DNN) based acoustic models trained jointly on several languages are used to train bottleneck (BN) networks [4, 5]. Sercu et al. [6] propose to utilize deep convolutional neural networks (CNN) to train multilingual BN networks. More recently, Hartmann et al. [7] use very deep CNNs and bi-directional long-short term memory networks (BLSTM) to train BN feature extractors. The BN extractors have shared and exclusive layers. The shared layers are used to learn language-invariant features. While the exclusive layers are used to capture language-dependent features. The BN features are extracted from the shared layers. Previous studies [8, 9, 10] have shown that acoustic models trained using BN features outperform models trained only on the target language data, especially when the amount of labelled data from the target language is limited. However, the BN shared features may contain some unnecessary language-specific information.

Inspired by the success of adversarial training on domain adaptation [11], this paper proposes an adversarial multilingual training to alleviate this problem. A parallel shared-exclusive model is also proposed to train the BN network using multitask learning [12]. Adversarial training [13] is used to ensure that the shared layers can learn language-invariant features.

Adversarial learning of DNNs is one of the hottest topics in many tasks recently. Ganin et al. [11] proposed to use adversarial strategy for domain adaptation in image tasks. More recently, Chen et al. [14] use adversarial multi-criteria learning for Chinese word segmentation in text tasks. Shinohara [15] and Saon et al. [16] utilizes adversarial multi-task learning for noise robustness and speaker adaptation respectively. These methods use adversarial multi-task learning to improve the performance of the primary task. The results show that they achieve state-of-the-art performance. However, this paper uses the adversarial learning to train multilingual BN networks. The BN networks are used to extract features for the target languages. There has been no work, to the best of our knowledge, that uses adversarial multilingual learning for lower-resource speech recognition.

The main contributions of this paper are as follows. 1) A parallel shared-exclusive BN model is proposed to extract features for the target language. 2) An adversarial training is used to force the shared layers to learn language-invariant features. Experiments are conducted on IARPA Babel datasets. The results show that the proposed adversarial multilingual BN model outperforms the baseline BN model by up to 8.9% relative word error rate (WER) reduction. The results also show that the proposed parallel shared-exclusive model achieves up to 1.7% relative WER reduction when compared with the stacked share-exclusive model.

The rest of this paper is organized as follows. Section 2 introduces multilingual bottleneck models. Section 3 describes adversarial training for shared layers. Section 4
presents the experiments. The results are discussed in Section 5. This paper is concluded in Section 6.

2. MULTILINGUAL BOTTLENECK MODELS

Two conventional DNN based BN models are introduced at first. Then the proposed BN model is presented. The three BN models are shown in Fig.1. The BN models are used to extract BN features for the target languages.

SSE-Model [6]. The shared and exclusive layers of this model are parallel. The outputs of the shared and exclusive layers are concatenated as the inputs of the output layers.

Given a dataset with \( N_m \) training samples \( \{ x_i^{(m)}, y_i^{(m)} \}^{N_m}_{i=1} \) for the \( m \)-th language, where \( \{ x_i^{(m)}, y_i^{(m)} \} \) is the training samples (frame-level), \( x_i^{(m)} \in \mathbb{R}^d \) is a feature vector, e.g. filterbank coefficients, \( d \) is the dimension of the feature vector, \( y_i^{(m)} \in \{ 1, \ldots, C_y^{(m)} \} \) is the senone label, \( C_y^{(m)} \) is the number of senone labels. The multilingual BN model is trained to minimize the cross-entropy on all the languages. The loss function of multilingual training can be defined as:

\[
L_{Mul}(\theta^s, \theta^m) = -\sum_{m=1}^{M} \sum_{i=1}^{N_m} \log P(y_i^{(m)} \mid x_i^{(m)}; \theta^s, \theta^m) \quad (1)
\]

where \( \theta^s \) denotes the parameters of the shared layers, \( \theta^m \) denotes the parameters of the exclusive layers for the \( m \)-th language, \( M \) is the number of all the languages.

3. ADVERSARIAL TRAINING FOR SHARED LAYERS

In order to learn language-invariant features, the adversarial training is used to optimize the shared layers of SSE-Model (Adv-SSE-Model) and PSE-Model (Adv-PSE-Model) as shown in Fig.2. Thus the shared layers are prevented from learning the language-specific features.

In adversarial training procedure, a language discriminator is used to recognize the language label using the shared features. An additional language label is given for each training sample \( \{ x_i^{(m)}, y_i^{(m)} \} \), \( m \in \{ 1, \ldots, M \} \) denotes the language label for each frame, and \( M \) is the number of language labels. The language discriminator loss function \( L_{Adv}(\theta^s, \theta^a) \) is defined as:

\[
L_{Adv}(\theta^s, \theta^a) = -\sum_{m=1}^{M} \sum_{i=1}^{N_m} \log P(m \mid x_i^{(m)}; \theta^s, \theta^a) \quad (2)
\]

where \( \theta^a \) denotes the parameters of the top sub-network of the language discriminator.

The gradient reversal layer (GRL) [11, 19] is introduced to ensure the feature distributions over all the languages are indistinguishable as possible for the language discriminator. Thus the shared layers can learn language-invariant features. At the feed-forward stage, the GRL acts as an identity transformation. During the back-propagation, however, the GRL takes the gradient from the subsequent level and changes its sign, i.e., multiplying by -1. The GRL has no parameters associated with it.

Thus, the adversarial multilingual training is to optimize the above mentioned two loss functions: \( L_{Mul}(\theta^s, \theta^m) \) and \( L_{Adv}(\theta^s, \theta^a) \).

The gradient w.r.t. the parameters are computed via back-
Adv-SSE-Model and Adv-PSE-Model denote SSE-Model and PSE-Model with adversarial training respectively.

\[ \theta^m \leftarrow \theta^m - \alpha \frac{\partial L_{\text{Mul}}}{\partial \theta^m} \]  
\[ \theta^a \leftarrow \theta^a - \alpha \lambda \frac{\partial L_{\text{Adv}}}{\partial \theta^a} \]  
\[ \theta^s \leftarrow \theta^s - \alpha \left( \frac{\partial L_{\text{Mul}}}{\partial \theta^s} - \lambda \frac{\partial L_{\text{Adv}}}{\partial \theta^s} \right) \]

where \( \alpha \in \mathbb{R} \) is the learning rate, \( \lambda \in \mathbb{R} \) is the loss weight, \( \lambda \) is gradually increased from 0 to 1 as epoch increases so that the model is stably trained [11].

4. EXPERIMENTS

4.1. Datasets

Our experiments are conducted on IARPA Babel datasets. The Babel datasets consist of conversational telephone speech for 25 languages collected across a variety of environments. The total amount of transcribed audio data varies depending on the language and condition. We select 4 languages from the datasets as the source languages: Assamese, Bengali, Kurmanji, and Lithuanian. The source languages are the full language pack (FLP), which are only used to train the multilingual BN networks. We also select 3 languages from the datasets as the target languages: Pashto, Turkish, and Vietnamese. The target languages have the FLP and the limited language pack (LLP).

4.2. Experimental setup

Our experiments are conducted using Kaldi speech recognition toolkit [20] and TensorFlow [21]. We follow the officially released Kaldi recipe to build a Gaussian mixture model hidden Markov model (GMM-HMM) at first. The features are extracted with a 25-ms sliding window with a 10-ms shift. Input features for the GMM-HMM model consist of 3-dimensional pitch features and 13-dimensional MFCC and their delta and delta-delta. We use the GMM-HMM models to generate frame-level state alignments for DNN models. All the DNN models use a sliding context window of 11 consecutive speech frames as inputs. Each frame is represented by 3-dimensional pitch features and 40-dimensional log mel-filter bank (Fbank) features plus their delta and delta-delta.

The four source languages are only used to train BN models. The size of the BN layer is 40, which is set inspired by [7]. The BN features are extracted from the BN models for three target languages respectively. The BN features are concatenated with Fbank and pitch features to train the DNN models for the target languages.

The three target languages are utilized to train DNN based monolingual models. For LLP systems, the DNN models have 5 hidden layers with 2048 nodes in each layers. For FLP systems, the DNN models have 6 hidden layers with 2048 nodes in each layer. The 3-gram language model (LM) is trained using the transcriptions of the training data for each language. We use the officially released vocabulary from IARPA Babel datasets. At the decoding stage, decoding is performed using fully composed 3-gram weighted finite state transducers.

4.3. Baseline model

At first, we train two models only using Fbank with or without pitch features. Then we use four source languages to train two BN models: SHL-Model-5L and SHL-Model-7L. They denote SHL-Model has 5 hidden layers and 7 hidden layers respectively. Each layer has 2048 nodes. The results on the LLP and FLP datasets are listed in Table 2 and Table 3 respectively.

The results show that the models with pitch features outperform the models without pitch features, especially for

Table 1. Overall experimental data distributions (hours).

<table>
<thead>
<tr>
<th>Language (Id)</th>
<th>Dataset</th>
<th>Training</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assamese (102)</td>
<td>FLP</td>
<td>61</td>
<td>10</td>
</tr>
<tr>
<td>Bengali (103)</td>
<td>FLP</td>
<td>62</td>
<td>10</td>
</tr>
<tr>
<td>Kurmanji (205)</td>
<td>FLP</td>
<td>41</td>
<td>10</td>
</tr>
<tr>
<td>Lithuanian (304)</td>
<td>FLP</td>
<td>42</td>
<td>10</td>
</tr>
<tr>
<td>Pashto (104)</td>
<td>FLP</td>
<td>78</td>
<td>10</td>
</tr>
<tr>
<td>Turkish (105)</td>
<td>LLP</td>
<td>77</td>
<td>10</td>
</tr>
<tr>
<td>Vietnamese (107)</td>
<td>FLP</td>
<td>88</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>LLP</td>
<td>11</td>
<td>10</td>
</tr>
</tbody>
</table>
the Vietnamese language. This is because the Vietnamese is a tonal language. The results also show that SHL-Model-7L achieves the best performance. Therefore, SHL-Model-7L is selected as our baseline BN model.

### 4.4. Adversarial multilingual BN models

In this group of experiments, we use four source languages to train two shared-exclusive multilingual BN models and their adversarial models. The SSE-Model and PSE-Model both have 5 shared hidden layers and 2 exclusive hidden layers. Each hidden layer has 2048 nodes. The network configurations of the Adv-SSE-Model and Adv-PSE-Model are similar to SSE-Model and PSE-Model respectively. The only difference is that the adversarial models add the language discriminator. The results of the models using BN features concatenated with Fbank and pitch features on the LLP and FLP datasets are shown in Table 4 and Table 5 respectively.

The results show that all the models with the adversarial BN features perform better than the models with BN features. Adv-PSE-Model achieves up to 5.5\% relative WER reduction when compared with PSE-Model on the LLP. PSE-Model outperforms SSE-Model by up to 1.7\% relative WER reduction. Adv-PSE-Model perform better than Adv-SSE-Model.

### 5. DISCUSSIONS

The above experimental results show that the proposed adversarial multilingual training is effective. Some interesting observations are made as follows.

The stacked and parallel shared-exclusive models both outperform the shared hidden layers models. The main reason may be that the shared BN features contained mixed language-specific information when the model only has shared hidden layers.

The proposed parallel shared-exclusive models outperform the stacked share-exclusive models for all the target languages. The possible reason is that the shared features contain less language-dependent information when the shared and exclusive layers are parallel.

The proposed adversarial multilingual BN models perform better than multilingual BN models. This is because the adversarial training makes the shared layers to prevent from learning the language-specific features. Thus the shared layers can learn more language-invariant features.

### 6. CONCLUSIONS

This paper proposes an adversarial multilingual training to train BN feature extractors for the target languages. A parallel shared-exclusive model is also proposed to train the BN network. Adversarial training is used to ensure that the shared layers can extract language-invariant features. Experiments are conducted on IARPA Babel datasets. The results show that the proposed adversarial multilingual BN model outperforms the baseline BN model by up to 8.9\% relative WER reduction. The results also show that the proposed parallel shared-exclusive model achieves up to 1.7\% relative WER reduction when compared with the stacked share-exclusive model. In future work, we plan to train CNN or BLSTM based adversarial multilingual BN models.

### 7. 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).
8. REFERENCES


