ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095644

DOMAIN AND LANGUAGE ADAPTATION USING HETEROGENEOUS DATASETS FOR WAV2VEC2.0-BASED SPEECH RECOGNITION OF LOW-RESOURCE LANGUAGE

Kak Soky[†], *Sheng Li*[‡], *Chenhui Chu*[†], *Tatsuya Kawahara*[†]

[†]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
[‡]National Institute of Information and Communications Technology (NICT), Kyoto, Japan

ABSTRACT

We address the effective finetuning of a large-scale pretrained model for automatic speech recognition (ASR) of low-resource languages with only a one-hour matched dataset. The finetuning is composed of domain adaptation and language adaptation, and they are conducted by using heterogeneous datasets, which are matched with either domain or language. For effective adaptation, we incorporate auxiliary tasks of domain identification and language identification with multi-task learning. Moreover, the embedding result of the auxiliary tasks is fused to the encoder output of the pretrained model for ASR. Experimental evaluations on the Khmer ASR using the corpus of ECCC (the Extraordinary Chambers in the Courts of Cambodia) demonstrate that first conducting domain adaption and then language adaption is effective. In addition, multi-tasking with domain identification and fusing the domain ID embedding gives the best performance, which is a CER improvement of 6.47% absolute from the baseline finetuning method.

Index Terms— Speech recognition, low-resource language, domain adaptation, language adaptation, Khmer language, self-supervised pretraining.

1. INTRODUCTION

Large-scale pretrained models based on self-supervised learning (SSL) [1, 2, 3, 4, 5, 6] have been intensively studied in speech and language processing communities. In speech processing, pretrained models such as wav2vec 2.0 [1], XLSR-53 [3], and XLS-R [4] have been successfully applied to many downstream tasks, including ASR [7, 8, 9, 10], speaker recognition (SRE) [11, 12], language identification (LID) [13], and speech emotion recognition (SER) [14]. Among them, XLS-R, which was trained with speech data from many languages, has shown impressive performance in ASR of low-resource languages. Finetuning the SSL model requires a much smaller amount of labeled data than training conventional end-to-end (E2E) networks [15, 16], which need massive amounts of data [17]. However, it is shown that it still requires a considerable amount of labeled data, like 10 hours, for finetuning the pretrained model to achieve satisfactory performance. [7, 8, 9, 10]. This is the case for languages that are not well covered by the pretrained model, such as XLS-R.[1] Since it is still difficult to collect such an amount of labeled data for many low-resource languages, in this paper, we address effective finetuning with the target dataset of only one hour.

The finetuning process for ASR involves domain adaptation and language adaptation. Here domain adaptation is concerned with application systems, speaking style, and input environments. Although the dataset matched with both domain and language of the target task is very limited (i.e., one hour), we often have access to other datasets, which are matched only with the domain (but in different languages) or only with the language (but in different domains). In this study, we explore effective domain and language adaptation using these kinds of heterogeneous datasets.

A straightforward method is to conduct domain adaptation using matched-domain datasets and then language adaptation with the target-language datasets. We also investigate the incorporation of auxiliary tasks, such as language identification and domain identification, with a framework of multi-task learning (MTL). This will allow the adaptation process to use the different kinds of datasets selectively. Moreover, we also investigate fusing the result of the auxiliary tasks by means of embedding domain ID or language ID to the encoder output of the SSL model before the final ASR step.

A variety of adaptation methods are evaluated in the ASR task of the Khmer language using the corpus of Extraordinary Chambers in the Courts of Cambodia (ECCC), where the domain is defined as transcription of court speech. It is shown that the two-step adaptation to the domain and the language using the heterogeneous datasets with multi-task learning and fusion results in improved performance.

The rest of this paper is structured as follows. We briefly overview the related work in Section 2. We then present our proposed method in Section 3. Section 4 describes the setup of the experiments and presents the result of all experimental evaluations. We conclude the paper in Section 5.

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m

2. RELATED WORK

Finetuning a large-scale SSL pretrained model has been intensively studied for many tasks. In low-resource language ASR, Yi et al. [7] showed the effectiveness of applying a wav2vec2.0 model pretrained using English speech to ASR of various spoken languages, which were recorded in different scenarios from the speech used in pretraining. Similarly, Krishna et al. [8] investigated the effectiveness of many kinds of self-supervised pretrained models for low-resource ASR tasks, showing that the multi-lingual model finetuned with about 20-hour speech data gave a competitive performance for both seen and unseen languages. Fatehi et al. [10] demonstrated the improvement of low-resource ASR by two-step finetuning: first pretraining a model in a high-resource language datasets and then finetuning with the low-resource language datasets to obtain language-dependent lexical units. Meanwhile, Yi et al. [9] improved the ASR system by fusing the encoders of wav2vec 2.0 and BERT [6] together.

Speaker recognition was also well adapted by finetuning a large-scale SSL pretrained model as investigated by Baskar et al. [11] and Vaessen et al. [12]. On the other hand, Tjandra et al. [13] showed an improvement in the language identification task by finetuning a large-scale SSL pretrained model. The SER task could also be improved by applying the MTL with ASR. The finetuning was investigated by Cai et al. [14]

There are several studies on improving ASR with auxiliary tasks such as speaker recognition. For example, Soky et al. [18] investigated the use of speaker ID embedding for ASR.

3. PROPOSED METHOD

In this section, we describe our proposed methods of two-step adaption and multi-task learning and fusion, which are illustrated in Fig. 1. Each adaptation step uses different heterogeneous datasets. Domain adaptation uses datasets that are matched to the domain but can be in different languages. Alternatively, language adaptation uses datasets that are of the same target language but can be in different domains.

In Fig. 1, domain adaptation is conducted in the first step, and then language adaptation is done in the second step, but they can be performed in a different order. The pretrained model used in the first step is the original XLS-R, a wav2vec 2.0-based multilingual SSL speech representation model. Then we use the finetuned model from the first step as the pretrained model for the second step.

In languages with limited resources, it is reasonable to use these kinds of heterogeneous datasets. It is often the case, even in major languages, that the matched dataset of the target task is limited, but datasets of different domains can be exploited. In this study, they are combined with the auxiliary task of domain identification, which is expected to guide the network to use the datasets selectively.

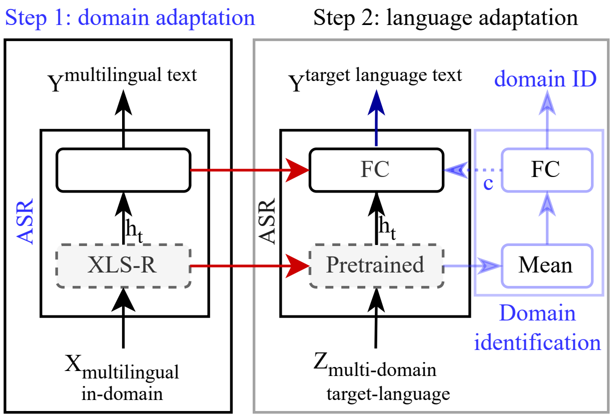
Step 1: domain adaptation
Step 2: language adaptation
$Y^{\text{multilingual text}}$
ASR
h_t
XLS-R
$X_{\text{multilingual}}$
in-domain
$Y^{\text{target language text}}$
domain ID
FC
c
FC
ASR
h_t
Pretrained
Mean
Domain identification
$Z_{\text{multi-domain}}$
target-language

Fig. 1. The proposed method of two-step adaptation using heterogeneous datasets. Step 1: domain adaption with multilingual in-domain datasets, Step 2: language adaptation with multi-domain datasets of the same target language. Domain adaptation and language adaptation can be made in different orders. There are two options for MTL: simple MTL and MTL with ID embedding, where ID can be the domain ID or language ID.

3.1. First-step adaptation

Step 1 in Fig. 1 shows domain adaptation using the domain-matched multi-lingual datasets. An input speech X is fed to a pre-trained XLS-R model, which extracts the features and encodes them with a Transformer to generate a contextual representation over the input continuous speech. Finally, we add a dense layer for the ASR task. For the output text, the Connectionist Temporal Classification (CTC) loss [19] is applied to finetune the entire network except for the feature extraction module. Note that simple language adaptation can be made in the same manner by using the relevant datasets of the same language.

3.2. Second-step adaptation with multi-task learning

Step 2 in Fig. 1 shows language adaptation using the multi-domain target-language datasets. In this step, the pretrained model is the result of the finetuning in Step 1. ASR is conducted in the same manner as Step 1. Here, we can also incorporate multi-task learning (MTL), in which domain identification is performed as an auxiliary task, which is depicted in the right-most part of the figure. It aggregates the features of the encoder output and applies a dense layer for identification. MTL is expected to guide the network to use the datasets selectively, namely, use the in-domain dataset and the out-of-domain dataset in a different way.

MTL is effective by sharing the encoder and employing dual decoders. The encoder is based on the Transformer architecture, whereas the ASR decoder is based on CTC, and the domain identification comprises pooling, linear, and normalization layers followed by the softmax layer. For MTL, we jointly optimize ASR and identification losses, defined as:

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{CTC} + \alpha * \mathcal{L}_{CE}, \tag{1}$$

where α is the weight of the identification task, $\mathcal{L}_{CTC}$ is the CTC loss of the ASR task, and $\mathcal{L}_{CE}$ is the cross-entropy loss of the identification task. The loss is used to finetune the entire network except for the feature extraction module of the pretrained model.

For more explicit guidance, the domain ID embedding as the result of domain identification is fused to the encoder output of the pretrained model. In this case, we add two more layers, a linear layer, and a normalization layer [20], for the output of identification. Then, the summation of the vectors is used in fusing with the encoder output. Here, we introduce a weighted sum of the domain ID embedding c and the encoder output of the pretrained model h_t to compute the final output h'_t used for ASR.

$$h'_t = h_t + \gamma * c, \tag{2}$$

where γ is the weight of the identification task.

Note that domain adaptation can be conducted in combination with language identification in the same manner by using the relevant datasets. Moreover, language adaptation and domain adaptation with MTL can be made in a single step. They are compared in the experiments. On the other hand, we apply MTL only in the second step because the first step is regarded as pretraining.

4. EXPERIMENTAL EVALUATIONS

4.1. Datasets

In this study, we conduct experiments of finetuning to ASR of Khmer, one of the low-resource languages. In particular, the target task is automatic transcription of Khmer speech in ECCC (Extraordinary Chambers in the Courts of Cambodia) [18]. From ECCC, a trilingual speech translation corpus (TriECCC) was also compiled for Khmer, English, and French [21, 22], thus we can use English and French datasets for domain adaptation.

In addition, we have access to another Khmer speech dataset of Google text-to-speech corpus.[2] It is matched with the language, but much different in terms of the domain, such as vocabulary, speaking styles, and recording environments. Thus, it is used for language adaptation.

For training datasets, we randomly select one hour from ECCC [18] as the target-task dataset (ECCC_KM), one hour per language of English (ECCC_EN) and French (ECCC_FR) from TriECCC [22], and one hour of Khmer speech from Google text-to-speech (Google_KM), whereas the testing and validation sets are the original data from ECCC [21] about 10 hours per each.

Table 1. The datasets used in this work

dataset	#hour	description
ECCC_KM	1	In-domain target Khmer
ECCC_EN	1	In-domain English for domain adaptation
ECCC_FR	1	In-domain French for domain adaptation
Google_KM	1	Out-of-domain Khmer for language adaptation

In summary, domain adaptation is conducted with the three-hour speech of ECCC_KM, ECCC_EN, and ECCC_FR, which can be used for three-language identification of Khmer, English, and French. On the other hand, language adaptation is conducted with the two-hour speech of ECCC_KM and Google_KM, which can be used for two-domain identification of ECCC (court) and Google (read).

4.2. System settings

We conducted experiments using XLS-R (wav2vec2-xls-r-300m), which is a large-scale wav2vec 2.0-based multilingual pretrained model for speech. It is a Transformer-based model comprised of 7 convolutional neural network (CNN) layers (each layer has 512 channels) and 24 encoder layers (each hidden layer size is $1,024$). The implementation is based on the Transformers [23]. During finetuning, we froze the CNN layers, which are primarily for feature extraction and had already been sufficiently trained during pretraining. A linear layer is added on top of the Transformer encoder layers. This linear layer takes the contextualized output of the encoder and converts them to tokens for ASR with the softmax operation. The CTC loss, which does not require the alignment information between the output sequences and the input speech, was used as the objective loss function of ASR. In this study, there are 76 and 112 output characters in Khmer and multi-lingual settings, respectively.

In MTL of ASR and domain/language identification, the weight α in Eqn. (1) was set to 0.01, and the weight to the domain/language ID embedding γ in Eqn. (2) was set to 0.01.

To speed up the training time, we group samples of similar input lengths into one batch to reduce the overall number of useless padding tokens passed through the model. The seed of learning rate was set to $3e$-4 to warm up until the finetuning has become stable. During training, SpecAugment [24] was also applied by masking some time frames and channels, and the last 2 checkpoints were saved asynchronously for every 500 training step. Each checkpoint was used to decode the validation set and evaluated with the character error rate (CER). Due to the large memory consumption, we used 16 batch sizes in each GPU with 2-step gradient accumulation on 2 GPUs. The total training batch size was 64, with $5,000$ in training steps for all models.

[2] https://openslr.org/42/

Table 2. CER performance of finetuning with single-step adaptation of domain or language, with or without MTL

method	#hour	CER(%)
Single step finetuning		
in-domain target: ECCC_KM (Baseline)	1	21.74
out-of-domain: Google_KM	1	35.01
language adaptation: {ECCC, Google}_KM	2	16.12
domain adaptation: ECCC_{KM, EN, FR}	3	17.75
One step finetuning with MTL		
language adaptation with domain identification	2	16.47
+ w/ domain ID embedding	2	16.23
domain adaptation with language identification	3	17.67
+ w/ language ID embedding	3	**16.11**

Table 3. CER Performance of the proposed method of two-step adaptation with or without MTL. (n, m) represents the number of hours in adaptation data of the first step (n) and the second step (m).

method	#hour	CER(%)
two-step finetuning		
domain → language adaptation	$(3, 2)$	15.59
language → domain adaptation	$(2, 3)$	16.00
two-step finetuning with MTL		
domain → language adaptation	$(3, 2)$	16.02
+ w/ domain ID embedding	$(3, 2)$	**15.27**
language → domain adaptation	$(2, 3)$	16.04
+ w/ language ID embedding	$(2, 3)$	15.95

At inference time, an input speech sample is decoded with a single step by the finetuned ASR system. In experimental evaluations, we tested various combinations of domain and language adaptation in different orders and also in a single step only.

4.3. Results

We evaluate the performance of all ASR models based on the character error rate (CER) of the 10-hour test set of Khmer ECCC [18]. Table 2 presents the results of individual adaptations to a domain or language in a single step.

The baseline is finetuning with ECCC_KM, the target dataset, which is matched with both domain and language but has only one-hour labeled speech data. Its CER is 21.74%. For reference, when we finetune with the Google_KM dataset, the CER is much worse (35.01%), which confirms a serious mismatch in terms of the domain.

When we conduct domain adaptation by using ECCC_EN and ECCC_FR, a large improvement (3.99% absolute) is gained from the baseline despite the use of speech data from different languages. The result confirms the significance of domain adaptation. When we conduct language adaptation by using Google_KM, an even larger improvement (5.62% absolute) is achieved. The result shows the effect of language adaptation is larger than that of domain adaptation. This is partly because the target language is not covered well in the pretrained model of XLS-R.

The lower part of Table 2 presents the effect of multi-task learning (MTL) with or without ID embedding. The performance of the domain adaptation is significantly improved by MTL with language ID embedding, which allows the model training to use the Khmer speech selectively. The result is comparable to the case of the language adaptation. On the other hand, domain identification does not help language adaptation. It is noted that both domain and language identifications were done almost 100% correctly, as they are very easy tasks.

Table 3 presents the results of the proposed two-step adaptation methods. When we compare the results of the upper part with those of Table 2, the two-step adaptation always gives an additional improvement. Among them, domain adaptation followed by language adaptation obtained the largest improvement. As the language adaptation dataset is more effective than the domain adaptation dataset, as shown in Table 2, the better-matched dataset must be used in the final finetuning.

The lower part of Table 3 presents the effect of MTL with or without ID embedding. Here, domain identification is conducted with language adaptation, and language identification is conducted with domain adaptation in the second step. The fusion of domain ID embedding in language adaptation results in a significant improvement, achieving the best performance in all settings. The CER improvement from the baseline is 6.47% absolute. On the other hand, language ID embedding does not help this time when the language adaptation was already conducted in the first step.

In all cases, we do not observe the effect of MTL alone, but the use of domain/language ID embedding in ASR is necessary for improved performance. This is the most critical part of the proposed method and finding in the experimental results.

5. CONCLUSION

We have presented effective finetuning strategies using matched data of only one hour in low-resource languages. In the proposed two-step adaptation scheme, domain adaptation, and language adaptation are conducted by using heterogeneous datasets, which are matched either in domain or language. Multi-task learning with an auxiliary task of domain/language identification is incorporated. Moreover, the result of the identification is fused into the ASR module. It is demonstrated that the fusion is effective for adaptation and achieves a significant improvement. In the future, we will investigate the effectiveness of the proposed method in other low-resource languages and other settings.

6. REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proc. NeurlPS*, 2020, vol. 33, pp. 12449–12460.

[2] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *CoRR*, vol. abs/1910.05453, 2019.

[3] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[4] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[5] Tao Xu Greg Brockman Christine McLeavey Ilya Sutskever Alec Radford, Jong Wook Kim, "Robust speech recognition via large-scale weak supervision.," 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[7] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu, "Applying wav2vec2.0 to speech recognition in various low-resource languages," *ArXiv*, vol. abs/2012.12121, 2020.

[8] Krishna D. N, Pinyi Wang, and Bruno Bozza, "Using Large Self-Supervised Models for Low-Resource Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2436–2440.

[9] Cheng Yi, Shiyu Zhou, and Bo Xu, "Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 788–792, 2021.

[10] Kavan Fatehi, Mercedes Torres Torres, and Ayse Kucukyilmaz, "ScoutWav: Two-Step Fine-Tuning on Self-Supervised Automatic Speech Recognition for Low-Resource Environments," in *Proc. Interspeech*, 2022, pp. 3523–3527.

[11] Murali Karthick Baskar, Tim Herzig, Diana Nguyen, Mireia Diez, Tim Polzehl, Lukas Burget, and Jan Černocký, "Speaker adaptation for Wav2vec2 based dysarthric ASR," in *Proc. Interspeech*, 2022, pp. 3403–3407.

[12] Nik Vaessen and David A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *Proc. ICASSP*, 2022, pp. 7967–7971.

[13] Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli, "Improved language identification through cross-lingual self-supervised learning," in *Proc. ICASSP*, 2022, pp. 6877–6881.

[14] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, "Speech Emotion Recognition with Multi-Task Learning," in *Proc. Interspeech*, 2021, pp. 4508–4512.

[15] Ashish Vaswani, Noam Shazeer abd Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is All You Need," in *Proc. NeurlPS*, 2017.

[16] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020.

[17] Jinyu Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, April 2022.

[18] Kak Soky, Sheng Li, Masato Mimura, Chenhui Chu, and Tatsuya Kawahara, "On the use of speaker information for automatic speech recognition in speaker-imbalanced corpora," in *Proc. APSIPA ASC*, 2021.

[19] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Shmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006.

[20] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer Normalization," *preprint arXiv:1607.06450*, 2016.

[21] Kak Soky, Masato Mimura, Tatsuya Kawahara, Sheng Li, Chenchen Ding, Chenhui Chu, and Sethserey Sam, "Khmer Speech Translation Corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC)," in *Proc. O-COCOSDA*, 2021.

[22] Kak Soky, Masato Mimura, Tatsuya Kawahara, Chenhui Chu, Sheng Li, Chenchen Ding, and Sethserey Sam, "TriECCC: Trilingual Corpus of the Extraordinary Chambers in the Courts of Cambodia for Speech Recognition and Translation Studies," *International Journal of Asian Language Processing*, vol. 31, no. 03n04, pp. 2250007, 2022.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush, "Transformers: State-of-the-art natural language processing," in *Proc. ACL-EMNLP*, Oct. 2020, pp. 38–45.

[24] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019.