Cross-Lingual Self-training to Learn Multilingual Representation for Low-Resource Speech Recognition

Zi-Qiang Zhang\textsuperscript{1} · Yan Song\textsuperscript{1} · Ming-Hui Wu\textsuperscript{1} · Xin Fang\textsuperscript{1} · Ian McLoughlin\textsuperscript{2} · Li-Rong Dai\textsuperscript{1}

Received: 10 August 2021 / Revised: 29 May 2022 / Accepted: 30 May 2022 / Published online: 23 July 2022 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Representation learning or pre-training has shown promising performance for low-resource speech recognition which suffers from the data shortage. Recently, self-supervised methods have achieved surprising performance for speech pre-training by effectively utilizing large amount of un-annotated data. In this paper, we propose a new pre-training framework, Cross-Lingual Self-Training (XLST), to further improve the effectiveness for multilingual representation learning. Specifically, XLST first trains a phoneme classification model with a small amount of annotated data of a non-target language and then uses it to produce initial targets for training another model on multilingual un-annotated data, i.e., maximizing frame-level similarity between the output embeddings of two models. Furthermore, we employ the moving average and multi-view data augmentation mechanisms to better generalize the learned representations. Experimental results on downstream speech recognition tasks for 5 low-resource languages demonstrate the effectiveness of XLST. Specifically, leveraging additional...
100 h of annotated English data for pre-training, the proposed XLST achieves a relative 24.8% PER reduction over the state-of-the-art self-supervised methods.

**Keywords** Multilingual representation learning · Cross-lingual self-training · Low-resource speech recognition · Speech pre-training

### 1 Introduction

Modern successful automatic speech recognition (ASR) systems are usually trained with hundreds of hours of annotated data [24, 53], while most languages in the world lack such amount of resource [4]. Languages with limited annotated data are termed the low-resource languages. Recently, pre-training techniques have shown to be promising [7, 40] to tackle the data shortage problem of low-resource ASR. Generally, speech pre-training is also regarded as speech representation learning [34], with the goal to learn a speech encoder from other materials (e.g., non-target annotated data or un-annotated speech data) that is transferable to downstream tasks like ASR.

Existing pre-training methods can be conducted in either supervised or unsupervised manner. Supervised methods try to solve the same task (i.e., ASR) using annotated data from one or more non-target languages [7, 11, 14, 17, 19, 22, 42, 44–48, 50, 51] for pre-training. The pre-trained models could be used as feature extractors [22, 44–47, 50] or the initialization of final ASR systems [7, 11, 14, 17, 19, 51], both usages have shown good performance for low-resource languages. Unsupervised pre-training has the advantage of employing large amount of un-annotated speech data, which is much cheaper to obtain than annotated data in practice. In the scenario of speech recognition, unsupervised pre-trained models are expected to learn acoustic representations carrying more phonetic information. In recent years, self-supervised methods [2, 3, 8, 9, 20, 26–30, 34, 39, 41] stand out as being surprising effective performing unsupervised speech pre-training. Most self-supervised tasks are defined as predicting speech signals auto-regressively [9, 34, 41], or retrieving masked frames from their contexts [3, 26, 27]. Here, speech signals to be predicted or retrieved are served as training targets, which are derived from the model itself. Among them, Wav2vec 2.0 [3] is one of the state-of-the-art methods, and is demonstrated good performance for multi-lingual pre-training [10].

This paper considers a multilingual scenario where pre-training or representation learning is conducted on multilingual un-annotated data, and then, the pre-trained model is fine-tuned for multiply low-resource languages. We argue that it is a better way to both incorporate some phonetic prior knowledge from annotated data and leverage large amount of un-annotated data like [3, 16]. So in this paper we propose a new multilingual pre-training framework, Cross-Lingual Self-Training (XLST), that uses an external model with phonetic prior knowledge to improve the representation learning on multilingual data. We show that such external model could be simply a phoneme classifier trained on a small amount of annotated data from a non-target language, as phonetic information is known to be transferrable across languages [17, 44]. To leverage the external model as well as multilingual un-annotated data, XLST uses BYOL-like [16] parallel networks: one for producing frame-level embeddings.
Fig. 1 XLST model architecture and training procedure. In stage 1, the Target Network is first trained with a small amount of non-target speech data ($x^o$) with phoneme labels. In stage 2, the Main Network is trained with multilingual un-annotated data ($x^u$) by maximizing the similarity between the two networks. Then, the Main Network is used to fine-tune for low-resource ASR of any target language and is initialized by the external model, and another for training to predict these embeddings, as illustrated in Fig. 1. They are denoted as the Target Network and the Main Network, respectively, in the rest of paper. These two networks take different corrupted versions of a same sentence as input through multi-view augmentation, and the Target Network is online refined in a moving average (MA) way [16].

The concept of using a prior model to produce targets for unsupervised data is also similar to self-training [18, 23, 25, 31, 38]. In the original self-training, a trained teacher model decodes pseudo-labels for unlabeled speech to train a student model. Different from self-training, XLST is a pre-training framework using high-level extracted embeddings instead of pseudo-labels. Such embeddings are free of phoneme inventory, which masks the pseudo-labels unavailable as there are unseen phonemes for different target languages.

We summary our contributions as threefold: First, we introduce XLST, a multilingual representation learning framework which leverages a non-target phoneme classifier to improve the multilingual pre-training. Second, we investigate the key components of XLST, including the data needed by the Target Network, the online moving average (MA) mechanism and multi-view augmentation. Third, we evaluate XLST performance on downstream low-resource speech recognition tasks—demonstrating significant improvement over previous self-supervised methods like Wav2vec 2.0 [10].

2 Related Work
2.0 [3]. XLST differs primarily in the way we derive training targets for un-annotated data, leading to a very different network design. More recently, UniSpeech [52] has incorporated CTC [15] loss into Wav2vec 2.0 framework using multi-task learning with additional annotated data. Unlike UniSpeech, XLST uses the annotated data only for obtaining a target producer in the early pre-training stage. Table 1 highlights the main difference between XLST and other previous multilingual representation learning methods.

**Self-training**, also known as teacher-student learning, has been proposed for semi-supervised ASR [18, 38] and domain adaptation [21, 23, 31]. Self-training involves generating pseudo-labels with a teacher model, and we introduce this idea to multilingual pre-training, i.e., leveraging a non-target phoneme classifier to guide the learning of the main model. Cross-lingual phoneme labels or posteriorgrams are also used in the field of unsupervised sub-word modeling [13, 43]. As there exists unseen phonemes in target languages which restricts the decoding of pseudo-labels, XLST simply employs the high-level embeddings as the training targets.

**Self-supervised learning** is a more general task which aims to learn informative representations from un-annotated data [5, 12, 16, 34]. Our work is most related to BYOL [16], as our model architecture is built upon it. BYOL was first designed for visual tasks [16] and then extended to audio representation learning [33]. The main difference with our work is that we incorporate an additional prior model in XLST. What is more, XLST learns from continuous speech at sequence level instead of instance level in [16, 33].

### 3 Methods

In this section, we will describe the details of our proposed pre-training method. For better understanding, we begin with a brief formulation of the general representation learning problem and then describe the proposed XLST.

#### 3.1 General Formulation of Representation Learning

Given an unlabeled dataset \( \{x^1, x^2, \ldots, x^N\} \) with \( x \) being a input sample, representation learning aims to optimize a model \( f_\theta \) able to encode \( x \) into a structured embedding \( e = f_\theta(x) \). The effectiveness of \( e \) can be evaluated by its performance when applied to downstream tasks, e.g., linear classification of phonemes. In self-supervised methods, the model is trained by maximizing the output similarity between two models taking different augmented views of a same sample as input, which is formulated as:

\[
\mathcal{L} = \text{sim}(f_\theta(A(x)), g_\phi(A'(x)))
\]  

(1)

where \( A \) and \( A' \) indicate random data augmentations, and \( g_\phi, f_\theta \) denote two models with parameters \( \phi \) and \( \theta \), respectively. \( g_\phi \) could be considered as a target producer for the main model \( f_\theta \). For example, in visual representation learning, \( g_\phi \) could simply be \( f_\theta \), i.e., \( g = f \) and \( \phi = \theta \) [5, 6], which means the model derives training targets
### Table 1 A comparison of our proposed XLST with other two previous multilingual pre-training methods

| Method       | Usage of annotated data                          | Usage of un-annotated data                                      |
|--------------|-------------------------------------------------|---------------------------------------------------------------|
| XLSR [10]    | Predicting masked features by contrastive loss  |                                                               |
| UniSpeech [52] | Multi-task learning with CTC [15]               | Predicting masked features by contrastive loss                 |
| XLST (ours)  | Learning a prior phoneme classifier              | Maximizing the similarity between the outputs of two networks  |
itself, while in speech representation learning, existing methods like [3, 41] define \( g \) as a front part of \( f \), specifically a stack of convolutional layers. However, starting with randomly initialized \( g_\phi \) may result in “model collapse,” where the outputs of \( f_\theta \) and \( g_\phi \) are degenerated to the same and be independent of the input, resulting in meaningless representations [6, 16].

To prevent collapse, existing speech representation learning methods [3, 41] introduced negative samples to formulate the contrastive loss [34]. Another solution in visual tasks is using a moving averaging target producer \( g_\phi \) instead of training it [16]. In XLST, we propose to initialize the target producer \( g_\phi \) with some prior knowledge. Without breaking the unsupervised restriction of target languages, the prior knowledge could be obtained by a phoneme classifier trained with few annotated data from a non-target language. We find it can bring significant improvement to the final learned representations, as well as making the training much simpler, e.g., with no need for contrastive loss.

3.2 XLST for Multilingual Speech Representation Learning

The goal of XLST is to learn multilingual representations for low-resource speech recognition, which can be regarded as obtaining a well pre-trained model that can be transferred to ASR tasks for multiple target languages. Here the multilingual setting is inherited from [10], where there are \( M \) languages and only un-annotated data of these languages is provided for pre-training. To perform XLST, we additionally use a small annotated dataset from an non-target language, and details will be found in 4.

As illustrated in Fig. 1, XLST inherits the basic idea of Eq. 1, with the essential difference of adopting a prior model (Target Network) to produce training targets for un-annotated data. Therefore, XLST contains two stages of pre-training. The first stage is to obtain the Target Network, and the second stage is to obtain the final pre-trained model (Main Network). At the first stage, the Target Network could be simply trained by phoneme classification of an non-target language (other options are discussed in Sect. 3.3). Then, at the second stage, the Target Network is assumed to be non-trainable (\( g_\phi \)) and only the Main Network needs to be optimized (\( f_\theta \)).

Two networks have the same architectures – a stack of convolutional layers, Transformer blocks and a nonlinear projector. As described in 3.1, generating two different views of an same input is essential for leveraging model consistency [5, 6, 16], so we also use augmenters \( A(\cdot) \) and \( A'(\cdot) \) to create randomly corrupted views for speech inputs. Let an input sequence of speech features be \( x^u \in R^{T \times F} \) (\( T \) and \( F \) indicate the time and frequency dimensions, respectively), the augmenter can be implemented by randomly masking consecutive time and frequency bins, or mixing with another sentence [49, 54]. It encourages the model to learn from context information in both the time and frequency dimensions.

Similar to Eq. 1, the training of the Main Network could be simple and straightforward by mimicking its Target Network. Specifically, let \( e \) and \( z \) denote the output embeddings from the Target Network and the Main Network, respectively, as in [16], we minimize the similarity loss that is computed as the normalized frame-level squared
Euclidean distance

$$
e = f_\theta \left( A \left( x^u \right) \right); \quad z = g_\phi \left( A' \left( x^u \right) \right)$$

\( L_{\text{sim}} \left( e, z \right) = \sum_{i=1}^{T} \left\| \bar{z}_i - \bar{e}_i \right\|_2^2 = \sum_{i=1}^{T} 2 - 2 \frac{\langle z_i, e_i \rangle}{\|z_i\|_2 \|e_i\|_2}$$

where \( z = z_1, z_2, \ldots, z_T \) and \( e = e_1, e_2, \ldots, e_T \) are frame-level embeddings with dimension of \( d \).

When two stages of pre-training are done, the Main Network could be further fine-tuned for specific ASR tasks. As illustrated in Fig. 1, we drop the nonlinear projector of the Main Network and add a few language specific layers on the top of it, e.g. a linear classification layer.

### 3.3 Obtaining and Refining the Target Network

In Sect. 3.2, we obtain the Target network by training a non-target phoneme classifier, while it is worth noting that the choice of the Target network is not unique. In visual tasks like [16], even a random Target Network can result in good performance. However, our primary experiments indicate that this is not a good choice for speech tasks (see Table 3). So we use a phoneme classifier of a non-target language. The basic principle in this paper is to use as few annotated data as possible, since it is far more expensive to obtain than un-annotated data in practical.

Due to language mismatch, the non-target phoneme model may not be able to effectively extract the information of multilingual speech. Therefore, the Target Network is refined during training, producing better targets for the Main Network. This can be achieved either by iteratively assigning the parameters of the Main Network (\( \theta \)) to the Target Network (\( \phi \)), or done in an online way. In this work, we use the online moving average (MA) [16] mechanism with a moving coefficient \( \lambda \). Specifically, the following operation is executed after every \( \Lambda \) training steps of the Main Network:

$$
\phi \leftarrow \lambda \phi + (1 - \lambda) \theta
$$

### 4 Experimental Details

For un-annotated pre-training data, we follow the data preparation in [10] which uses 1350 h of un-annotated speech in the CommonVoice corpus [1], including 557 h of English data and 793 h of multilingual data from 10 languages detailed in [10]. Here English is not treated as a target language and we also expand the English data to 1350 h for building the English baseline. For annotated pre-training data, we chose English as

---

1 As the older version is no longer available, we use the December 2019 release, maintaining the same number of hours of data as [10] for 11 languages. CommonVoice corpus is publicly available in https://commonvoice.mozilla.org.
Table 2: The abbreviations of different pre-trained models

| Pre-trained Model | Description |
|-------------------|-------------|
| T-En#*            | Target Network pre-trained with #* hours of annotated English data |
| ST-En# (T-En#*)   | Main Network pre-trained with # hours of un-annotated English data, using T-En#* as the Target Network |
| ST-En#            | Main Network pre-trained with # hours of un-annotated English data, using T-En100 as the default Target Network |
| XLST-M            | Main Network pre-trained with 793 h of un-annotated multilingual data, using ST-En557 as the Target Network |
the non-target language and used the LibriSpeech corpus [36]. Both LibriSpeech and CommonVoice contain reading style speech. To keep the total data amount constant, when we used annotated data, the un-annotated English data are correspondingly reduced by the same amount.

For model architecture, our backbone is a VGG-Transformer [32] model with 2 VGG convolutional blocks and 12 Transformer blocks. The attention dimension is 512 and the feed-forward dimension is 2048. The projector is an MLP with a single hidden layer of 2048 units, ReLU activation and output dimension of 256. The total number of parameters is around 45 millions. Frame-level batch normalization is applied both at the end of the VGG blocks and in the hidden layer of the projector. Input acoustic features are composed of 80-dimensional filter bank and 3-dimensional pitch coefficients. The VGG layers have a downsampling factor of 2 so that the output embeddings are 20 ms per frame.

As noted in Sect. 3.2, XLST has two stages of pre-training: supervised pre-training with frame-level cross-entropy (CE) loss, and unsupervised pre-training to minimize the similarity loss (Eq. 3). The ADAM optimizer is used for all pre-training stages. For convenience, we denote the learning rate schedule as \((l_r, w_1, w_2, w_3)\), indicating the maximum learning rate and its warming up, holding and exponential decay periods. It is set to \((10^{-3}, 0.2, 0, 0.8)\) for supervised pre-training at the first stage and \((5 \times 10^{-4}, 0, 0.5, 0.5)\) for unsupervised pre-training at the second stage. The batch size is about 200 seconds per GPU and is varied by changing the number of GPUs (12 as default). The total training epoch is 200/100/100/50 for training the Target Network on 10/50/100/460 h of annotated data, and is 60/50/40 for training the Main Network on 557/793/1350 h of un-annotated data. Fairseq [35] is used to implement all the experiments.

The effectiveness of the pre-training is evaluated by fine-tuning the Main Network for a downstream phoneme recognition task [10, 40]. We also fine-tune the Target Network for comparison. When fine-tuning, the projector is removed, VGG layers are fixed and a linear classifier concatenating 2 successive frames is added on the top. CTC [15] loss is computed with non-aligned phoneme transcriptions. In this paper, we examine 5 target languages, Spanish (es), French (fr), Italian (it), Russian (ru) and Tatar (tt). Each language has 1-h training data. We use the respective validation sets to select fine-tuned models and report the phone error rate (PER) on the test sets of these languages.

5 Results and Analysis

In this section, we first investigate 3 key components of XLST, which are the Target Network, the online moving average (MA) mechanism and the multi-view augmentation. Then, we discuss the final performance of XLST in comparison with previous
Table 3 A comparison of pre-training models using different English data; annotated data are indicated by ∗

| Pre-trained model       | Pre-training data (h) | Fine-tuning PER (%) |
|-------------------------|-----------------------|---------------------|
|                         | es       | fr       | it       | ru       | tt       | avg      |
| –                       | –        | 34.2     | 41.4     | 39.9     | 42.1     | 34.6     | 38.4     |
| ST-En557 (T-random)     | 557      | 18.4     | 22.3     | 24.3     | 24.8     | 16.1     | 21.2     |
| T-En10                  | 10∗      | 21.7     | 26.2     | 27.3     | 27.5     | 19.1     | 24.4     |
| T-En50                  | 50∗      | 15.6     | 19.1     | 20.4     | 20.6     | 13.5     | 17.8     |
| T-En100                 | 100∗     | 14.1     | 17.9     | 18.7     | 19.1     | 12.6     | 16.5     |
| T-En460                 | 460∗     | 12.6     | 16.0     | 16.8     | 17.5     | 11.0     | 14.8     |
| ST-En557 (T-En10)       | 10∗ + 547 | 14.3    | 18.6     | 19.2     | 19.6     | 12.6     | 16.9     |
| ST-En557 (T-En50)       | 50∗ + 507 | 12.4    | 16.2     | 17.0     | 17.4     | 10.9     | 14.8     |
| ST-En557 (T-En100)      | 100∗ + 447 | 12.1    | 15.8     | 16.5     | 17.0     | 10.6     | 14.4     |
| ST-En557 (T-En460)      | 460∗ + 97  | 11.9    | 15.3     | 16.3     | 16.6     | 10.2     | 14.1     |
| ST-En1350(T-En100)      | 100∗ + 1250 | 11.6   | 14.9     | 15.9     | 16.6     | 10.0     | 13.8     |

PER is evaluated after fine-tuning over 5 downstream target languages

5.1 Effectiveness of the Target Network

To validate how a prior phoneme classifier benefits the pre-training and how much annotated data we need, we pre-train different models with respect to the data amount. Total 557 h of data is randomly selected from the LibriSpeech and CommonVoice English corpus, including different amounts of annotated data. The Target Network is pre-trained at the first stage with annotated data and then the Main Network is pre-trained at the second stage with all 557 h of data. Table 3 reports the fine-tuning PERs of various pre-trained models.

The first row in Table 3 is a random baseline without pre-training, the PER is high due to the shortage of fine-tuning data. The second row indices using a random Target Network, which is in the scope of unsupervised methods. It is noticed that a random Target Network is useful for pre-training, reducing the fine-tuning PER from 38.4% to 21.2%. However, supervised pre-training (T-EnXX series) is much more effective than the unsupervised baseline. That confirms our motivation that a prior model with phoneme knowledge of a non-target language is of great benefit to pre-training. Then, 2-stage pre-training continues to improve the performance as large amount of un-annotated data are used (ST-En557 series).

Meanwhile, there is an interesting phenomenon where the ST-En557 series get improved much slower than T-EnXX with increasing annotated data. Especially, the

---

4 As suggested in [16], for random initialization experiment we add an extra MLP predictor (same architecture as the projector) on the top of the Main Network, and moving average mechanism ($\lambda = 0.99, \Lambda = 32$) is applied.
Table 4  Ablation study on moving average (MA) mechanism

| Pre-trained model             | λ of MA | Fine-tuning PER (%) |
|-------------------------------|---------|---------------------|
|                               |         | es  | fr  | it  | ru  | tt  | avg |
| ST-En1350 (T-En100)          | –        | 11.6| 14.9| 15.9| 16.6| 10.0| 13.8|
| ST-En1350 (T-En100)          | 0.99     | 11.2| 15.2| 15.6| 17.0| 10.2| 13.8|
| ST-En1350 (ST-En557)         | –        | 10.8| 14.4| 15.2| 16.1| 9.6 | 13.2|
| XLST-M                       | –        | 9.2 | 10.8| 12.6| 14.2| 9.3 | 11.2|
| XLST-M                       | 0.98     | 6.6 | 8.1 | 9.9 | 11.6| 7.3 | 8.7 |

For experiments using MA, λ is chosen from (0.96, 0.98, 0.99) according to the fine-tuning PER on validation sets

PER only gets reduced from 14.4 to 14.1% with annotated data increased from 100 to 460h. We attribute the reason to a balance between phoneme modeling ability and cross-lingual generalization. Supervised pre-training with annotated data helps the former but quickly reaches its saturation point, while unsupervised pre-training with un-annotated data helps the latter and needs a much larger data amount. It could be validated by further increasing the amount of un-annotated data, as the last line in Table 3, the pre-trained model could still get improved (14.4% to 13.8%).

5.2 Importance of the Target Network Refining

Online refining the Target Network is important since it is randomly initialized in [16]. We wonder if it is still necessary to refine the Target Network when it is not randomly initialized. In the following experiments, we use either T-En100 or ST-En557 from Table 2 as the Target Network to pre-train the Main Network with different refining mechanisms. For the experiment using MA, λ is set to 32 and λ is searched from (0.96, 0.98, 0.99) according to validation PER. The results are shown in Table 4.

It is observed that for models pre-trained with only English data, MA does not lead to better performance. While, directly offline assigning the Main Network to the Target Network (i.e. using ST-En557 as the Target Network for an additional stage of pre-training on 1350h of data) tends to perform better. However, in the multi-lingual setting, MA brings significant improvement for the fine-tuning performance (11.2% to 8.7%). The inconsistency between English and multilingual settings confirms that online refining the Target Network is crucial when pre-training on multilingual data.

5.3 Effectiveness of Multi-view Augmentation

In self-supervised methods like [5, 6], model’s consistency on different augmented views of a same input is the key to learn meaningful representations. For speech pre-training, XLST can leverage model consistency by performing context-level augmentation, including time masking, frequency masking [37] and sentence mixup [49, 54]. In the following experiments, we apply frequency masking by selecting two mask windows with length ranging from 0 to 27. For time masking, we follow the strategy
Table 5  Ablation study on multi-view augmentation

| Pre-trained model                      | Fine-tuning PER (%) |
|----------------------------------------|---------------------|
|                                        | es | fr | it | ru | tt | avg |
| T-En100                                | 14.1|17.9|18.7|19.1|12.6|16.5 |
| T-En100, w/o mixup                     | 15.1|18.7|19.3|20.1|13.2|17.3 |
| T-En100, w/o mixup, F-mask             | 16.3|20.2|20.5|21.3|14.8|18.6 |
| T-En100, w/o mixup, F-mask, T-mask     | 21.5|26.6|26.9|28.2|22.1|25.1 |
| ST-En557                               | 12.1|15.8|16.5|17.0|10.6|14.4 |
| XLST-M, w/o mixup                      | 9.2 |10.8|12.6|14.2| 9.3 |11.2 |
| XLST-M, w/o mixup, F-mask              | 9.7 |12.0|14.0|15.3| 9.5 |12.1 |
| XLST-M, w/o mixup, F-mask, T-mask      | 12.0|15.7|16.6|17.4|10.9|14.5 |

MA is not performed and ST-En557 is used as the Target Network to train XLST-M. F/T-mask denotes the time or frequency masking, respectively.

in [3] but with a mask length of 10 (100ms) and a masking proportion of 40%. For mixup [54], two input sequences are mixed and padded to the same length for the Main Network, while their targets are individually drawn by the Target Network (mixing phoneme labels in supervised pre-training). The mixing weight $\beta$ is sampled from a uniform distribution $\beta \sim U(0, 1)$.

The results are shown in Table 5. XLST-M uses ST-En557 as the Target Network and is pre-trained without MA mechanism. Considering ST-En557 already achieves a fine-tuning PER of 14.4%, XLST-M would achieve no improvement without any augmentation. In contrast, XLST-M achieves a PER of 11.2% if time and frequency masking are applied. This confirms multi-view augmentation is still necessary for the unsupervised pre-training stage in XLST. Meanwhile, Table 5 also shows that all kinds of augmentations can benefit the supervised pre-training, while mixup does not work well for unsupervised pre-training which might be due to the distance-based similarity loss.

5.4 General Performance of XLST

The general results of our pre-trained models are shown in Table 6. We use the best strategy discussed above to obtain our final model. Specifically, we use T-En100 as the Target Network to train ST-En557 on 557h of English data, and then use ST-En557 as the Target Network to train XLST-M on 793h of multilingual data using MA ($\lambda = 0.98$). Multi-view augmentation is applied for all pre-training stages except mixup for the unsupervised stage. We compare our results with several recent works under similar data settings [10, 52]. Both UniSpeech [52] and XLSR [10] use Wav2vec 2.0 [3] framework, whose model has similar components but is twice as large as ours.

Table 6 shows that:

(1) English pre-trained models perform well on other languages. Compared with XLSR-English, our ST-En1350 brings a 20.5% relative reduction of PER at a cost of using 100h annotated data (7.4% of the total). It also surpasses the fully
### Table 6: Overall results of XLST

| Model                        | Size (M) | Total pre-training data          | Fine-tuning PER (%) |
|------------------------------|----------|----------------------------------|---------------------|
|                              |          |                                  | es      | fr      | it      | ru      | tt      | avg     |
| Results reported in [52]     |          |                                  |         |         |         |         |         |         |
| CTC-Transfer                 | 90       | 1350* (En)                       | 12.6    | 16.7    | 16.4    | 17.5    | 11.2    | 14.9    |
| UniSpeech                    | 90       | 1350* (En)                       | 10.9    | 14.8    | 15.2    | 16.1    | 9.6     | 13.3    |
| Results reported in [10]     |          |                                  |         |         |         |         |         |         |
| XLSR-English                 | 90       | 1350 (En)                        | 13.7    | 20.0    | 19.1    | 18.6    | 11.5    | 16.6    |
| XLSR-10                      | 90       | 557 (En) + 793 (Multi)           | 9.4     | 14.2    | 14.1    | 11.0    | 7.6     | 11.3    |
| English pre-trained models (Ours) |      |                                  |         |         |         |         |         |         |
| T-En100                      | 45       | 100* (En)                        | 14.1    | 17.9    | 18.7    | 19.1    | 12.6    | 16.5    |
| ST-En457                     | 45       | (100* + 457) (En)                | 12.1    | 15.8    | 16.5    | 17.0    | 10.6    | 14.4    |
| ST-En1350                    | 45       | (100* + 1250) (En)               | 10.8    | 14.4    | 15.2    | 16.1    | 9.6     | 13.2    |
| Multilingual pre-trained models (Ours) |      |                                  |         |         |         |         |         |         |
| XLST-M (MA)                  | 45       | (100*+457)(En) + 793 (Multi)     | 6.7     | 8.1     | 9.9     | 11.6    | 7.3     | 8.7     |
| XLST-M (MA, balanced)        | 45       | (100*+457)(En) + 793 (Multi)     | 6.6     | 9.0     | 9.4     | 10.8    | 6.5     | 8.5     |

PERs are reported on 5 target languages after fine-tuning the pre-trained models. *Indicates annotated data.
supervised CTC baseline, confirming that purely supervised pre-training is insufficient for cross-lingual transferring. Meanwhile, ST-En1350 is also on par with UniSpeech which uses fully annotated data.

(2) Multilingual pre-trained models further improve the performance of downstream ASR. A 35.6% relative reduction of PER (XLST-M) over the best English pre-trained model (ST-En1350) is observed in Table 6. This indicates that even starting from an English model, XLST-M can still learn important knowledge of other languages only using their un-annotated speech. Again, XLST-M achieves a 24.8% relative reduction of PER compared with the self-supervised model XLSR-10.

6 Conclusion

This work has proposed and investigated a multilingual pre-training framework, XLST. XLST begins with a prior model and then generalizes its representations using multilingual un-annotated data by maximizing the similarity between two networks. We demonstrated that a suitable prior model could be obtained with minimum cost, by training a phoneme classifier with a small amount of annotated data of a non-target language. Multi-view augmentations and moving average mechanism were confirmed crucial to successfully pre-train XLST models. Experiments also showed that XLST significantly outperformed the previous self-supervised methods, even with a smaller model size. In the future, other kinds of prior models like unsupervised models, as well as more elaborate augmentations, could be further investigated.

Acknowledgements This work was supported by the Leading Plan of CAS (Grant No. XDC08030200).

Data Availability The training datasets generated during the current study are available in the CommonVoice (https://commonvoice.mozilla.org) and the LibriSpeech corpus (http://www.openslr.org/12). The evaluation datasets generated during this study are included in this published article [40].

References

1. R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, Common voice: a massively-multilingual speech corpus, in Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association (2020), pp. 4218–4222
2. A. Baevski, S. Schneider, M. Auli, vq-wav2vec: Self-supervised learning of discrete speech representations, in International Conference on Learning Representations (2020)
3. A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12449–12460 (2020)
4. L. Besacier, E. Barnard, A. Karpov, T. Schultz, Automatic speech recognition for under-resourced languages: a survey. Speech Commun. 56, 85–100 (2014). https://doi.org/10.1016/j.specom.2013.07.008
5. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in Proceedings of the 37th International Conference on Machine Learning, ed. by H.D.A. Singh III, vol. 119 (PMLR, 2020), pp. 1597–1607
6. X. Chen, K. He, Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566 (2020)
7. J. Cho, M.K. Baskar, R. Li, M. Wiesner, S.H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, T. Hori, Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling, in SLT (2018), pp. 521–527. https://doi.org/10.1109/SLT.2018.8639655
8. J. Chorowski, R.J. Weiss, S. Bengio, A. van den Oord, Unsupervised speech representation learning using wavenet autoencoders. IEEE/ACM Trans. Audio Speech Lang. Process. 27(12), 2041–2053 (2019). https://doi.org/10.1109/TASLP.2019.2938863
9. Y.A. Chung, W.N. Hsu, H. Tang, J. Glass, An unsupervised autoregressive model for speech representation learning, in Proceedings of Interspeech 2019 (2019), pp. 146–150
10. A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979 (2020)
11. S. Dalmia, R. Sanabria, F. Metze, A.W. Black, Sequence-based multi-lingual low resource speech recognition, in ICASSP (2018), pp. 4909–4913. https://doi.org/10.1109/ICASSP.2018.8461802
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers) (2019), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
13. S. Feng, T. Lee, Exploiting cross-lingual speaker and phonetic diversity for unsupervised subword modeling. IEEE/ACM Trans. Audio Speech Lang. Process. 27(12), 2000–2011 (2019). https://doi.org/10.1109/TASLP.2019.2937953
14. A. Ghoshal, P. Swietojanski, S. Renals, Multilingual training of deep neural networks, in ICASSP (2013), pp. 7319–7323. https://doi.org/10.1109/ICASSP.2013.6639084
15. A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in Proceedings of the 23rd International Conference on Machine Learning (2006), pp. 369–376
16. J.B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent—a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284 (2020)
17. J. Huang, J. Li, D. Yu, L. Deng, Y. Gong, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, in ICASSP (2013), pp. 7304–7308. https://doi.org/10.1109/ICASSP.2013.6639081
18. J. Kahn, A. Lee, A. Hannun, Self-training for end-to-end speech recognition, in ICASSP 2020 (2020), pp. 7084–7088. https://doi.org/10.1109/ICASSP40776.2020.9054295
19. M. Karafiát, M.K. Baskar, S. Watanabe, T. Hori, M. Wiesner, J. Černocký, Analysis of multilingual sequence-to-sequence speech recognition systems, in Proceedings of Interspeech 2019 (2019), pp. 2220–2224. https://doi.org/10.21437/Interspeech.2019-2355
20. S. Khurana, A. Laurent, W.N. Hsu, J. Chorowski, A. Lancucki, R. Marxer, J. Glass, A convolutional deep Markov model for unsupervised speech representation learning, in Proceedings of Interspeech 2020 (2020), pp. 3790–3794. https://doi.org/10.21437/Interspeech.2020-3084
21. S. Khurana, N. Moritz, T. Hori, J.L. Roux, Unsupervised domain adaptation for speech recognition via uncertainty driven self-training, in ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021), pp. 6553–6557. https://doi.org/10.1109/ICASSP39728.2021.9414299
22. P. Lal, S. King, Cross-lingual automatic speech recognition using tandem features. IEEE Trans. Audio Speech Lang. Process. 21(12), 2506–2515 (2013). https://doi.org/10.1109/TASL.2013.2277932
23. J. Li, M.L. Seltzer, X. Wang, R. Zhao, Y. Gong, Large-scale domain adaptation via teacher-student learning, in Proceedings of Interspeech 2017 (2017), pp. 2386–2390. https://doi.org/10.21437/Interspeech.2017-519
24. J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, S. Liu, On the comparison of popular end-to-end models for large scale speech recognition, in Proceedings of Interspeech 2020 (2020), pp. 1–5. https://doi.org/10.21437/Interspeech.2020-2846
25. J. Li, R. Zhao, J.T. Huang, Y. Gong, Learning small-size DNN with output-distribution-based criteria, in Interspeech (2014), pp. 1910–1914
26. S. Ling, Y. Liu, Decor 2.0: deep contextualized acoustic representations with vector quantization. arXiv preprint arXiv:2012.06659 (2020)
27. S. Ling, Y. Liu, J. Salazar, K. Kirchhoff, Deep contextualized acoustic representations for semi-supervised speech recognition, in ICASSP 2020 (2020), pp. 6429–6433. https://doi.org/10.1109/ICASSP40776.2020.9053176

28. A.H. Liu, Y.A. Chung, J. Glass, Non-autoregressive predictive coding for learning speech representations from local dependencies. arXiv preprint arXiv:2011.00406 (2020)

29. A.T. Liu, S.W. Li, H.Y. Lee, Tera: self-supervised learning of transformer encoder representation for speech. arXiv preprint arXiv:2007.06028 (2020)

30. A.T. Liu, S.W. Yang, P.H. Chi, P.C. Hsu, H.Y. Lee, Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders, in ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020), pp. 6419–6423. https://doi.org/10.1109/ICASSP40776.2020.9054458

31. Z. Meng, J. Li, Y. Gaur, Y. Gong, Domain adaptation via teacher-student learning for end-to-end speech recognition, in ASRU (2019), pp. 268–275. https://doi.org/10.1109/ASRU46091.2019.9003776

32. A. Mohamed, D. Okhonko, L. Zettlemoyer, Transformers with convolutional context for ASR. arXiv preprint arXiv:1904.11660 (2019)

33. D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, K. Kashino, Byol for audio: self-supervised learning for general-purpose audio representation. arXiv preprint arXiv:2103.06695 (2021)

34. A.V. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

35. M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, fairseq: a fast, extensible toolkit for sequence modeling, in Proceedings of NAACL-HLT 2019: Demonstrations (2019)

36. D.S. Park, W. Chan, Y. Zhang, C.C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, SpecAugment: a simple data augmentation method for automatic speech recognition, in Proceedings of Interspeech 2019 (2019), pp. 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

37. D.S. Park, Y. Zhang, Y. Jia, W. Han, C.C. Chiu, B. Li, Y. Wu, Q.V. Le, Improved noisy student training for automatic speech recognition, in Proceedings of Interspeech 2020 (2020), pp. 2817–2821. https://doi.org/10.21437/Interspeech.2020-1470

38. S. Pascual, M. Ravanelli, J.Serrá, A. Bonafonte, Y. Bengio, Learning problem-agnostic speech representations from multiple self-supervised tasks, in Proceedings of Interspeech 2019 (2019), pp. 161–165. https://doi.org/10.21437/Interspeech.2019-2605

39. S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: unsupervised pre-training for speech recognition, in Proceedings of Interspeech 2019 (2019), pp. 3465–3469. https://doi.org/10.21437/Interspeech.2019-1873

40. T. Sercu, C. Puhrsch, B. Kingsbury, Y. LeCun, Very deep multilingual convolutional neural networks for LVCSR, in ICASSP (2016), pp. 4955–4959. https://doi.org/10.1109/ICASSP.2016.7472620

41. H. Shibata, T. Kato, T. Shinozaki, S. Watanabet, Composite embedding systems for zerospeech2017 track1, in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2017), pp. 747–753. https://doi.org/10.1109/ASRU.2017.8269012

42. A. Stolcke, F. Grezl, Hwang, Mei-Yuh. Lei, Xin. N. Morgan, D. Vergyri, Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons, in ICASSP, vol. 1 (2006). https://doi.org/10.1109/ICASSP.2006.1660022

43. S. Thomas, S. Ganapathy, H. Hermansky, Cross-lingual and multi-stream posterior features for low resource LVCSR systems, in Interspeech (2010)

44. S. Thomas, S. Ganapathy, H. Hermansky, Multilingual MLP features for low-resource LVCSR systems, in ICASSP (2012), pp. 4269–4272. https://doi.org/10.1109/ICASSP.2012.6288862

45. S. Thomas, M.L. Seltzer, K. Church, H. Hermansky, Deep neural network features and semi-supervised training for low resource speech recognition, in ICASSP (2013), pp. 6704–6708. https://doi.org/10.1109/ICASSP.2013.6638959

46. S. Tong, P.N. Garner, H. Bourlard, Cross-lingual adaptation of a CTC-based multilingual acoustic model. Speech Commun. 104, 39–46 (2018). https://doi.org/10.1016/j.specom.2018.09.001

47. V. Verma, A. Lamb, J. Kannala, Y. Bengio, D. Lopez-Paz, Interpolation consistency training for semi-supervised learning. arXiv preprint arXiv:1903.03825 (2019)
50. K. Veselý, M. Karafiát, F. Grézl, M. Janda, E. Egorova, The language-independent bottleneck features, in SLT (2012), pp. 336–341. https://doi.org/10.1109/SLT.2012.6424246
51. N.T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, H. Bourlard, Multilingual deep neural network based acoustic modeling for rapid language adaptation, in ICASSP (2014), pp. 7639–7643. https://doi.org/10.1109/ICASSP.2014.6855086
52. C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, X. Huang, Unispeech: unified speech representation learning with labeled and unlabeled data, in Proceedings of the 38th International Conference on Machine Learning, vol. 139. PMLR (2021), pp. 10937–10947
53. Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig, M.L. Seltzer, Transformer-based acoustic modeling for hybrid speech recognition, in ICASSP 2020 (2020), pp. 6874–6878. https://doi.org/10.1109/ICASSP40776.2020.9054345
54. H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.