INJECTING TEXT AND CROSS-LINGUAL SUPERVISION IN FEW-SHOT LEARNING FROM SELF-SUPERVISED MODELS

Matthew Wiesner¹, Desh Raj², Sanjeev Khudanpur¹,²

¹Human Language Technology Center of Excellence, Johns Hopkins University, USA
²Center for Language and Speech Processing, Johns Hopkins University, USA
{wiesner,khudanpur}@jhu.edu, draj@cs.jhu.edu

ABSTRACT

Self-supervised model pre-training has recently garnered significant interest, but relatively few efforts have explored using additional resources in fine-tuning these models. We demonstrate how universal phonetic acoustic models can leverage cross-lingual supervision to improve transfer of pretrained self-supervised representations to new languages. We also show how target-language text can be enabled and improve fine-tuning with the lattice-free maximum mutual information (LF-MMI) objective. In three low-resource languages these techniques greatly improved few-shot learning performance.

Index Terms— Self-supervised, few-shot learning, lattice-free MMI, cross-lingual ASR

1. INTRODUCTION

In the last decade, automatic speech recognition (ASR) systems have benefited from supervised training of deep neural networks on transcribed speech at scale. However, since annotated resources for low-resource languages are often scarce or expensive, considerable efforts have been made towards unsupervised and semi-supervised learning for ASR, resulting in methods such as representation learning, pseudo-labeling, local prior matching, and adversarial training [1, 2, 3, 4]. Recently, self-supervised approaches [5, 6, 7, 8, 9, 10] have become popular for learning general representations of the input distribution without supervision. In particular, wav2vec2.0 (W2V2) [11] has been used for downstream tasks such as cross-lingual ASR [12], speech translation [13], and speaker and language identification [14].

Prior work has investigated improved self-supervised pre-training of models using side information including (i) using a text-to-speech (TTS) module to inject text [15], (ii) using teacher models to improve representation learning [16], and (iii) moving average based learning from a fixed supervised teacher [17]. However, research on few-shot fine-tuning with these models has received less attention. In this work, we develop ASR models for new languages using very small amounts (say, 15 mins.) of transcribed speech. However, we have access to additional resources, including cross-lingual transcribed and untranscribed speech, as well as unpaired target language data. We explore the best strategies to incorporate these in fine-tuning large self-supervised pre-trained models.

1.1. Related Work

Most recent work on fine-tuning of self-supervised models is in the “end-to-end” framework by learning a linear output layer using the connectionist temporal classification (CTC) objective [18]. In contrast, “hybrid” ASR systems which combine acoustic, pronunciation, and language models for decoding with weighted finite state transducers (WFSTs) [19] are popular in industry due to their compositionality and efficiency. We focus on such hybrid ASR models.

Acoustic model transfer in hybrid ASR has used the lattice-free MMI (LF-MMI) objective [20] — in conjunction with methods such as weight transfer, multi-task training, and teacher-student learning [21, 22] for fine-tuning models. One outstanding problem in LF-MMI based model transfer is how best to construct the denominator graph when using only small amounts of domain-matched transcribed speech. This issue was raised in [23], but has not been explicitly addressed in a multilingual setting.

Recently, Vyas et al. [24] studied the application of supervised fine-tuning using LF-MMI on self-supervised acoustic models, and showed that representations learned on English transferred well to low-resource languages such as Swahili. However, there have been no investigations about the use of additional resources, such as cross-lingual supervision or unpaired target language text, in fine-tuning from large self-supervised models.

We leverage ideas from prior work [25, 26, 27] which explored language-independent approaches for cross-lingual transfer by relying on a universal phoneset, such as the IPA [28], to train phone models by mixing training data from all languages. More recently, the sequence training of universal phonetix models using LF-MMI was explored with various neural architectures [29, 30], while [31] explored cross-lingual transfer of acoustic models trained using CTC. Our main contributions in this paper are three-fold:

1. We propose using a cross-lingual universal phoneset model for
few-shot learning. This significantly outperforms the conventional method of fine-tuning by constructing a new output layer.

2. We leverage untranscribed speech by exploring the choice of pre-trained, self-supervised models used in fine-tuning. Experiments show that using matched-domain and matched-language data in pre-training significantly improves downstream performance.

3. We simultaneously leverage unpaired, target language text while addressing the question of how to construct the denominator graph in LF-MMI based fine-tuning, using unpaired text to improve the LF-MMI denominator graph phone language model.

2. TECHNICAL PRELIMINARIES

The hybrid HMM-DNN framework for speech recognition follows the Bayes decomposition of the ASR objective. Given a speech input \( x^{(u)} \), the most likely sequence \( \hat{w}^{(u)} \) is given as

\[
\hat{w}^{(u)} = \arg \max_w p(w|x^{(u)}) = \arg \max_w p(x^{(u)}|w)p(w) \\
= \arg \max_w \prod_{t=1}^{T} p(x_t^{(u)}|f_t^{(w)})p(f_t^{(w)})p(w),
\]

where each term, \( H, L, G \), are estimated independently. \( H \) is a collection of hidden Markov models (HMM), one per (context-dependent) phone, where a neural network models HMM state emission probabilities. Together, the HMMs and neural network are termed the acoustic model. The acoustic model scores can be combined with a pronunciation lexicon, \( L \), and language model (grammar), \( G \), to produce scores for hypothesis sequences, \( w \). One advantage of these models is the ability to optimize each component independently. In this work, this property enables the re-use of acoustic models trained to produce units from a universal phone set.

The acoustic model is often trained using a sequence discriminative objective such as maximum mutual information (MMI), which aims to maximize the the point-wise mutual information between the speech signal and reference transcripts:

\[
F_{\text{MMI}} = \sum_{u \in U} \log \frac{p_L(x^{(u)}|M_u^{(w)})}{p_L(x^{(u)})p(x^{(u)}|M_{\text{dnn}})},
\]

Here, \( M_u^{(w)} \), called the numerator graph, is a graph representing the set of all alignments and pronunciations corresponding to the ground-truth sequence \( w \). \( M_{\text{dnn}} \), called the denominator graph, is ideally, an HMM modeling all possible word sequences. The LF-MMI objective [20] maximizes a lower-bound on the mutual information [21] by using un-normalized distributions parameterized by a neural network. The denominator graph is approximated with a 4-gram language model. In other words, the LF-MMI objective is globally normalized, as opposed to CTC, which assumes a uniform prior between output units and is locally normalized. \( \lambda \) is the set of all HMM parameters, including the DNN that estimates the globally normalized, as opposed to CTC, which assumes a uniform prior between output units and is locally normalized. \( \lambda \) is the set of all HMM parameters, including the DNN that estimates the probability of HMM states at each time step.

Contrastive self-supervised models aim to learn representations of speech that maximize a lower bound on the mutual information between a chunk of audio \( x_t \) at time \( t \) and its surrounding context \( c_t \) using the InfoNCE objective. In wav2vec 2.0 [11], a convolutional feature extractor generates latent speech representations \( z \) from the waveform, which are then processed by a Transformer to produce contextualized representations \( e \). A quantization module converts the continuous latent vectors \( z \) to discrete representations \( q \). The pre-training objective predicts the correct quantized vector \( q_t \) by contrasting against distractors. The emission probabilities in DNN-HMMs can be replaced by a linear transformation of the outputs \( c \), of self-supervised models. General practice, which we follow, is to freeze the parameters of the feature extractor during fine-tuning.

3. METHOD

In this section, we will describe our method for injecting each of the 3 different “additional” resources (shown in Fig. [1]) — cross-lingual transcribed speech, untranscribed speech, and unpaired target language text — into few-shot learning.

3.1. Cross-lingual transcribed speech

We use transcribed cross-lingual speech to train a universal phoneme-sequence model, which we refer to as “multi-phoneset” (as opposed to a “mono-phoneset” obtained from just the target language). We hypothesize that multi-phoneset models can be used, to an extent, in new languages without transcribed speech.

To verify this, we conducted some preliminary experiments on \( \text{hat} \), where we analyzed the performance of wide ResNet (WRN) based mono- and multi-phoneset models when fine-tuned on various amounts of data (Fig. [2]). We found that in the few-shot scenario (<1k utterances), the multi-phoneset model outperformed the conventional approach of retraining a monolingual output layer. Our goal is to replicate this behavior using larger self-supervised models capable of producing better pre-trained representations.

3.2. Untranscribed speech

We explore the value of using matched-domain and matched-language untranscribed speech in self-supervised pre-training. [33] demonstrated the clear value of pre-training on matched acoustic channel untranscribed speech. Furthermore, some of the improvements of self-supervised contrastive learning appear to transfer across language [34, 35], and [12] demonstrated that representations trained on large and diverse cross-lingual speech could outperform those trained on small amounts of matched-language speech. Using pre-trained models on speech in new languages results in both mismatch in language and/or acoustic channel (different auditory environments). It is not immediately clear which, if any, of these are more harmful for downstream performance. To investigate these questions, we performed fine-tuning experiments with publicly available W2V2 models pre-trained on 4 different kinds of data:

- **LV60**: Trained on 53.2k hours of English audiobook recordings
- **https://hf.co/facebook/wav2vec2-large-lv60**
- **Mono-phoneset**
- **Multi-phoneset**
from Libri-Light \[1\]. Few-shot transfer using this model performs remarkably well on LibriSpeech \[36\].

2. \textbf{Large-Robust} \[3\] Trained on English audiobooks (LibriLight \[1\]), conferences (TED-Lium \[37\]), and telephone conversations (Switchboard) \[43\].

3. \textbf{VoxPopuli-100k} \[4\] Trained on 100k hours of speech in 23 languages from the VoxPopuli corpus \[38\]. The data is obtained from European parliamentary recordings (and only contains European languages).

4. \textbf{XLSR-53} \[4\] Trained on 56k hours of speech in 53 languages, including CommonVoice and BABEL corpora. It has previously shown strong cross-lingual transfer performance for ASR and phoneme recognition tasks \[12\] \[39\].

3.3. \textbf{Unpaired target language text}

Conventionally, unpaired target language text is used to train language models which are then fused with acoustic model outputs during decoding or rescorign. In contrast, we explore how to use these resources during training of the acoustic model. Recall from \[2\] that the LF-MMI objective, contrary to the CTC objective (often used in fine-tuning self-supervised models), enables us to incorporate a language model in training through the denominator graph \(M_{\text{den}}\), which can be supplemented with additional unpaired, or cross-lingual text. We conjecture that such denominator graph expansion may lead to better loss computation by estimating a prior which is closer to the true target distribution (as opposed to CTC’s uniform prior), particularly in the few-shot scenario.

4. \textbf{EXPERIMENTAL SETUP}

4.1. \textbf{Data}

Prior work in \textit{few-shot} transfer has mainly used clean, read speech. A motivation of this work was to study how these representations transfer in moderately challenging, conversational, and multilingual speech. The BABEL corpus \[41\] was developed to support research in ASR and keyword search in such conditions. The speech data are 8kHz telephone conversations in 25 languages. Each language has two training sets: the full language pack (FLP) (between 40h and 80h); and the limited language pack (LLP) (10h subset of FLP). A 10h development split (dev10h) is also defined for each language. We report performance on dev10h as separate evaluation sets are not publicly available. To study cross-lingual transfer in the \textit{few-shot} scenario, we created 200-utterance subsets from the LLP set for each language (or 15 min. of speech total from 2-4 speakers).

Pronunciation lexicons containing X-SAMPA \[42\] pronunciations of all words seen in the FLP and dev10h sets are provided. X-SAMPA phonemes are an ASCII representation of the IPA \[23\], which was constructed to be a cross-lingual finite symbol set representing all attested phonemes in the world. This label-set enables training universal phonset models.

We use 21 languages (XL-21) as our cross-lingual supervision, and report performance of all our systems on held-out languages, Pashto (\textit{pus}), Haitian (\textit{hat}), and Georgian (\textit{kat}). Importantly, the FLP audio from these languages are \textit{included} in the XLSR-53 training data. In Table \[1\] we report the state-of-the-art WERs on these languages using various methods from literature. We used a 1h subset of the Haitian dev10h for early stopping and hyper-parameter tuning.

4.2. \textbf{Training details}

We used 2 types of acoustic models: 28-10 wide ResNets (WRN) \[43\], and the W2V2-based models and architectures. All monolingual and cross-lingual supervised baselines are reported using the WRN model; for example, XL-21 WRN refers to the WRN model pre-trained on the 21 BABEL languages with supervision.

We conducted fine-tuning experiments on the few-shot, LLP, and FLP subsets, where the training alignments for the few-shot and LLP were obtained by training GMM-HMMs on the LLP subset, and those for FLP by training on the FLP data, respectively. We use the cross-lingual GMM-HMMs, occasionally remapping missing phonemes in the target language to a similar phoneme seen in training to create the targets for fine-tuning with the "Multi" models. We trained all acoustic models with PyTorch \[44\] and used Kaldi \[45\] for decoding. We used Pychain for the LF-MMI implementation \[46\]. All decoding was performed with a 3-gram language model trained on the FLP transcripts in srilm \[47\]. This conforms with a common scenario where larger amounts of unpaired text are more readily available than transcribed speech and is different from denominator graph expansion, which we used during \textit{training}. Our implementation and recipes are publicly available \[4\]. The denominator graph phone-LMs for the "multi" models combined the cross-lingual and target language training data, weighting the target-language data by a factor of 10 as in \[23\].

5. \textbf{RESULTS AND DISCUSSION}

5.1. \textbf{Effect of phone set expansion}

Table \[2\] shows the \textit{few-shot} transfer of multilingual baselines in the three target languages. We find that transferring the supervised cross-lingual representation from WRN models (row-2) works better than reconstructing the output layer (row-1). Training the cross-lingual supervised model with the W2V2 neural architecture (row-3) performed slightly worse than the WRN baseline. Self-supervised

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|}
\hline
Language & Model & Data (h) & WER \\
\hline
Pashto (\textit{pus}) & CTC-BLSTM + lang. adversarial & 0.2k & 43.4 \\
Haitian (\textit{hat}) & Multi-lingual CTC-BLSTM & 0.5k & 44.3 \\
Georgian (\textit{kat}) & XLSR-53 w/ CTC fine-tuning & 56k & 31.1 \\
\hline
\end{tabular}
\caption{Previous SOTA WERs (using FLP training) for the languages used for fine-tuning experiments in this paper.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|l|}
\hline
Model & Pre-training (h) & \textit{pus} & \textit{hat} & \textit{kat} \\
\hline
XL-21 WRN Mono & 0 / 1.2k & 86.2 & 81.5 & 89.2 \\
XL-21 WRN Multi & 0 / 1.2k & 75.6 & 72.9 & 73.1 \\
XL-21 Wav2Vec 2.0 (random) & 0 / 1.2k & 77.2 & 77.9 & 76.0 \\
XLSR-53 & 56k / 0 & 75.9 & 74.7 & 81.5 \\
XLSR-53 + XL-21 & 56k / 1.2k & 71.0 & 67.8 & 64.5 \\
XLSR-53 (frozen) + XL-21 Multi & 56k / 1.2k & \textbf{63.6} & \textbf{58.9} & \textbf{58.2} \\
\hline
\end{tabular}
\caption{Effect of cross-lingual supervision on downstream \textit{few-shot} WER performance, using the 15-min train subset. Column 2 shows the amount of (labeled / unlabeled) data in pre-training.}
\end{table}
### Table 3. Effect of pre-training with untranscribed cross-lingual data on the downstream fine-tuning WER. †The WERs can be reduced to 38.8, 36.9, and 32.2, respectively, if we decode with an expanded vocabulary with no OOVs.

| Model                  | Pre-training (hours) | Few-shot (~15m) | LLP (~10h) | FLP (~80h) |
|------------------------|----------------------|-----------------|------------|------------|
|                        |                      | Labeled | Unlabeled |             | Labeled | Unlabeled |             | Labeled | Unlabeled |             | Labeled | Unlabeled |             |
| Random WRN             | 0                    | 0       | 93.2      | 95.4       | 61.1     | 57.7       | 57.6       | 50.1     | 46.6       | 49.5       |
| XL-21 WRN Mono         | 1.2k                 | 0       | 86.2      | 81.5       | 56.6     | 52.3       | 55.1       | 44.7     | 43.3       | 46.2       |
| LV60                   | 0                    | 60k     | 81.4      | 84.2       | 86.0     | 50.3       | 50.3       | 50.5     | 42.8       | 40.8       | 43.1       |
| Large-Robust           | 0                    | 63k     | 81.2      | 80.1       | 93.0     | 49.5       | 49.1       | 48.7     | 41.8       | 40.4       | 42.2       |
| VoxPopuli-100k         | 0                    | 100k    | 80.3      | 77.2       | 85.6     | 48.9       | 48.0       | 47.9     | 41.1       | 39.3       | 41.5       |
| XLSR-53                | 0                    | 56k     | 75.9      | 74.7       | 81.5     | 45.5       | 45.3       | 45.1     | 39.4†       | 37.7†       | 40.0†       |

Fig. 3. Effect of denominator graph expansion (regular vs. expanded) for LF-MMI based few-shot cross-lingual transfer of the XLSR-53 model.

### 5.2. Analysis of pre-trained W2V2 models

Self-supervised pre-training outperformed XL-21 based supervised pre-training in all scenarios, as shown in Table 3 but we found a clear hierarchy among pretrained models. (i) **Matched unlabeled data is king** – XLSR-53 was the best performing base-model and improved over a cross-lingual supervised WRN baseline by 10-15% relative across the board. It was trained on the smallest amount of, but matched acoustic and language, data. The LV60 model is trained on a similar amount of data, but only clean English audio from LibriLight. It was by far the worst performing model. (ii) **Linguistic diversity, quantity of data, and matched acoustic conditions are important.** For instance, the VoxPopuli-100k model, like the LV60 model, is trained on clean wide-band speech. However, presumably due to the quantity, and multilingual (but not matched language) nature, of training data, it was the second best performing model. Finally, the Robust model, despite training on much less, and purely monolingual English speech, performed comparably to the VoxPopuli model, likely thanks to the telephone speech (similar to BABEL) that it was exposed to in training. Increasing the amount of fine-tuning data, reduced the difference in performance among models.

### 5.3. Effect of denominator graph expansion

Using additional, unpaired text to train the denominator graph phone-LM reduces overfitting. Fig. 3 shows that using the resulting "expanded" LF-MMI denominator graphs improves the WER across all languages in the few shot setting. However, using lower-order n, in the phone-LM n-gram model would result in less over-fitting. We analyzed the trade-off between modeling-power and unpaired text weight in the Haitian few-shot scenario with XLSR-53 (Table 4).

The best performing systems use n = 2, demonstrating the benefit of using LF-MMI in fine-tuning as opposed to CTC, which corresponds to the n = 0 case with a different HMM topology. We also found that including the unpaired text with a small weight improved performance across all, but especially higher-order n. Using additional text enables the use of higher-order n-gram models even in low-resource scenarios. We also see that the improvements from the universal phoneset model cannot be attributed solely to the inclusion of cross-lingual phone sequences when training the denominator graph phone-LM.

### 6. CONCLUSION

In this work, we have shown how cross-lingual supervision can be used during fine-tuning in few-shot learning to significantly improve transfer of self-supervised models. We confirmed the importance of using matched-domain speech in pre-training, and we demonstrated how to incorporate external text in the phone-LM of the LF-MMI denominator graph in order to help guide fine-tuning. These techniques resulted in significantly improved few-shot performance in three low-resource test conditions.
