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

Matthew Wiesner[1], *Desh Raj*[2], *Sanjeev Khudanpur*[1,2]

[1]Human Language Technology Center of Excellence, Johns Hopkins University, USA
[2]Center for Language and Speech Processing, Johns Hopkins University, USA

{wiesner,khudanpur}@jhu.edu, draj@cs.jhu.edu

ABSTRACT

Self-supervised model *pretraining* has recently garnered significant interest. However, using additional resources in *fine-tuning* these models has received less attention. We demonstrate how universal phoneset 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 used to enable 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 primarily from *supervised* training of deep neural networks on transcribed speech at scale. However, the dearth of annotated data in low-resource languages spurred interest in unsupervised, semi-supervised and self-supervised learning [1]. Recently, self-supervised approaches [2, 3, 4, 5, 6, 7] have been used to learn general representations of raw data *without* supervision. In particular, wav2vec2.0 (W2V2) [8] has been used for downstream tasks such as cross-lingual ASR [9], speech translation [10], and speaker and language identification [11].

Some prior work has focused on improving self-supervised pretraining of models using side information including (i) using a text-to-speech (TTS) module to inject text [12], (ii) using teacher models to improve representation learning [13], and (iii) moving average based learning from a fixed supervised teacher [14]. However, using side information in 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. We also assume access to additional resources, including cross-lingual transcribed and untranscribed speech, as well as unpaired target language data. We explore strategies to incorporate these data when *fine-tuning* self-supervised, pretrained models.

1.1. Related Work

Most recent research on fine-tuning of self-supervised models is in the "end-to-end" framework and consists of training an additional linear output layer using the connectionist temporal classification (CTC) objective [15], or multilingual transfer learning using encoder-decoder models as in [16, 17, 18]. In contrast, "hybrid" ASR systems which combine acoustic, pronunciation, and language
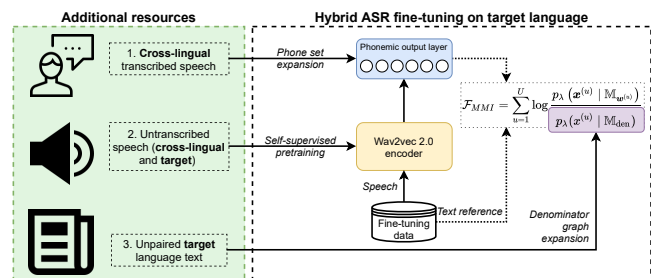

Fig. 1. Our proposed integration of cross-lingual transcribed speech and unpaired target language text in LF-MMI-based fine-tuning of large self-supervised models. Cross-lingual supervision is injected using a sequence-trained universal phone recognizer at the output layer. We use unpaired target language text to estimate the denominator graph used for LF-MMI computation.

models for decoding with weighted finite state transducers (WF-STs) [19] are popular in industry due to their compositionality and efficiency. We focus on fine-tuning self-supervised representations for use in 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, no work, to our knowledge, has explored using additional resources, such as cross-lingual supervision or unpaired target language text, in fine-tuning from self-supervised models.

As in prior work [25, 26, 27], we enable language-*independent* cross-lingual acoustic model transfer via a universal phoneset, such as the IPA [28]. We train phone models by mixing data from all languages. More recently, the sequence training of universal phoneset models using LF-MMI was explored with various neural architectures [29, 30], while [31] explored *cross*-lingual transfer of acoustic models trained using CTC.

Finally, prior work [9] demonstrated that representations trained on large and diverse cross-lingual speech could outperform those trained on smaller amounts of matched-language speech and [32, 33]

showed that some of the improvements of self-supervised contrastive learning appear to transfer across language. However, [34] also demonstrated the clear value of including at least some matched acoustic channel untranscribed speech in pretraining. It is unclear however, the extent to which both acoustic channel and language mismatch of pretrained models affects downstream ASR performance. 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 explore the choice of pretrained, self-supervised models used in fine-tuning as a means of leveraging untranscribed speech. Experiments show that using matched-domain and matched-language data in pretraining significantly improves downstream performance.

3. We demonstrate how to use unpaired text in the LF-MMI denominator graph phone language model for improved fine-tuning.

2. TECHNICAL PRELIMINARIES

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

$$\hat{\mathbf{w}}^{(u)} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{x}^{(u)}) = \arg\max_{\mathbf{w}} p(\mathbf{x}^{(u)}|\mathbf{w})p(\mathbf{w}) \quad (1)$$

$$= \arg\max_{\mathbf{w},\ell} \underbrace{p(\mathbf{x}^{(u)}|\ell)}_{H} \underbrace{p(\ell|\mathbf{w})}_{L} \underbrace{p(\mathbf{w})}_{G}, \quad (2)$$

where in Equation 2, the Viterbi approximation is used, and ℓ represents a latent subword-unit sequence into which words are decomposed. Each term, H, L, G, is 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, $\mathbf{w}$. 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 phoneset.

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:

$$\mathcal{F}_{\text{MMI}} = \sum_{u \in U} \log \frac{p_\lambda(\mathbf{x}^{(u)}|\mathbb{M}_w^{(u)})}{p_\lambda(\mathbf{x}^{(u)}|\mathbb{M}_{\text{den}})}. \quad (3)$$

Here, $\mathbb{M}_w$, called the numerator graph, is a graph representing the set of all alignments and pronunciations corresponding to the ground-truth sequence w. $\mathbb{M}_{\text{den}}$, 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 [35] by using un-normalized distributions parameterized by a neural network. The denominator graph is approximated with a 4-gram phone 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. λ is the set of all HMM parameters, including the DNN that estimates the probability of HMM states at each time step.
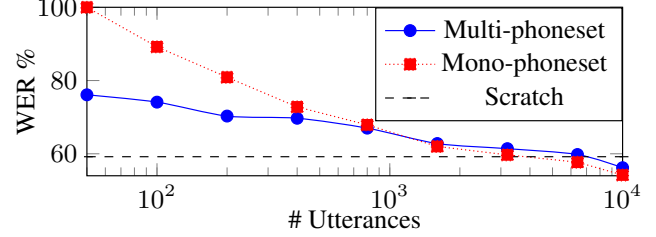

Fig. 2. Comparison between multi and mono phoneset outputs with increasing training data on WRN models on `hat`. In the few-shot scenario (< 1k utterances), using a shared, universal phoneset provides significant WER improvements over a monolingual baseline.

Contrastive self-supervised models aim to learn representations of speech that maximize a lower bound on the mutual information between a chunk of audio $\mathbf{x}_t$ at time t and its surrounding context $\mathbf{c}_t$ using the InfoNCE objective. In wav2vec 2.0 [8], a convolutional feature extractor generates latent speech representations $\mathbf{z}$ from the waveform, which are then processed by a Transformer to produce contextualized representations $\mathbf{c}$. A quantization module converts the continuous latent vectors $\mathbf{z}$ to discrete representations $\mathbf{q}$. The pretraining objective predicts the correct quantized vector $\mathbf{q}_t$ by contrasting against distractors. The emission probabilities in DNN-HMMs can be replaced by a linear transformation of the outputs $\mathbf{c}$, of self-supervised models. General practice, which we follow, is to freeze the parameters of the feature extractor during fine-tuning.

3. METHOD

We 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 the BABEL Haitian (`hat`) corpus, described in more detail in Section 4.1.. 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 pretrained representations.

3.2. Untranscribed speech

Using pretrained models on speech in new languages results in both mismatch in language and/or acoustic channel. 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 pretrained on 4 different kinds of data:

Table 1. Previous SOTA WERs (using FLP training) for the languages used for fine-tuning experiments in this paper.

Language	Model	Data (h)	WER
Pashto (pus)	CTC-BLSTM + lang. adversarial [40]	0.2k	43.4
Haitian (hat)	Multi-lingual CTC-BLSTM [31]	0.5k	44.3
Georgian (kat)	XLSR-53 w/ CTC fine-tuning [9]	56k	31.1

1. **LV60**[1]: Trained on 53.2k hours of English audiobook recordings from Libri-Light [1]. Few-shot transfer using this model performs remarkably well on LibriSpeech [36].

2. **Large-Robust**[2]: Trained on English audiobooks (LibriLight [1]), conferences (TED-Lium [37]), and telephone conversations (Switchboard) [34].

3. **VoxPopuli-100k**[3]: 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. **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 [9, 39].

3.3. 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 rescoring. 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 $\mathbb{M}_{den}$, 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. EXPERIMENTAL SETUP

4.1. Data

Prior work in *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) (40-80h); 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 **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 [28],

---

[1] https://hf.co/facebook/wav2vec2-large-lv60
[2] https://hf.co/facebook/wav2vec2-large-robust
[3] https://hf.co/facebook/wav2vec2-large-100k-voxpopuli
[4] https://hf.co/facebook/wav2vec2-large-xlsr-53

Table 2. Effect of cross-lingual supervision on downstream **few-shot** WER performance, using the 15-min train subset. Column 2 shows the amount of (unlabeled / labeled) data in pretraining.

Model	pretraining (h)	pus	hat	kat
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	**63.6**	**58.9**	**58.2**

which was constructed to be a cross-lingual finite symbol set representing all attested phonemes in the world. This label-set enables training universal phoneset models.

We use 21 languages (XL-21) as our cross-lingual supervision, and report performance of all our systems on held-out languages, Pashto (pus), Haitian (hat), and Georgian (kat). Importantly, the FLP audio from these languages are *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. 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 pretrained 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 *training*. Our implementation and recipes are publicly available[5]. 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. RESULTS AND DISCUSSION

5.1. Effect of phone set expansion

Table 2 shows the 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

---

[5] https://github.com/m-wiesner/nnet_pytorch

Table 3. Effect of pretraining 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	Pretraining (hours)		Few-shot ($\sim$15m)			LLP ($\sim$10h)			FLP ($\sim$80h)		
	Labeled	Unlabeled	pus	hat	kat	pus	hat	kat	pus	hat	kat
Random WRN	0	0	93.2	95.4	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	89.2	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**†

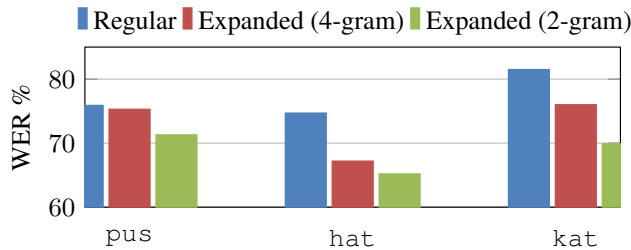**Fig. 3**. Effect of denominator graph expansion (regular vs. expanded) for LF-MMI based few-shot cross-lingual transfer of the XLSR-53 model.

Table 4. Effect of n-gram order and unpaired text weight, α, on WER in LF-MMI few-shot learning of Haitian ASR models.

Weight (α)	n=0	n=1	n = 2	n = 3	n = 4
$\alpha = 0.0$	81.9	66.2	65.2	68.2	74.7
$\alpha = 0.1$	81.9	66.1	**65.0**	66.3	67.6
$\alpha = 0.2$	81.9	66.3	65.1	66.1	66.8
$\alpha = 0.5$	81.9	65.9	65.2	66.3	67.2

vs. supervised pretraining (rows 3 v. 4) made little difference in the few-shot scenario. There were likely not enough data to retrain the output mapping from self-supervised representations to phonemes.

We compare 2 approaches that combine the supervised, sequence-level universal phoneset model, with the pretrained self-supervised models. In the XLSR-53 + XL-21 (row-5), we performed two sequential fine-tuning steps: first on the cross-lingual representations, and then on the target language few-shot supervisions. We also try freezing the XLSR-53 network when fine-tuning on the XL-21 data (row-6). This was the best-performing approach, almost matching the *supervised* models trained on 10h of audio! We suspect that freezing the W2V2 model prevents catastrophic forgetting during cross-lingual fine-tuning, while simultaneously enabling us to learn a universal phoneset model. Few-shot learning by transferring a universal output layer worked significantly better than reconstructing the output layer using small amounts of speech as in [8, 9].

5.2. Analysis of pretrained W2V2 models

Self-supervised pretraining outperformed XL-21 based supervised pretraining 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 [1]. 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. Fi-

nally, 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 preforming 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 pretraining, 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.

7. REFERENCES

[1] J. Kahn et al., "Libri-light: A benchmark for asr with limited or no supervision," in *IEEE ICASSP*, 2020.

[2] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *IEEE ICASSP*, 2020.

[3] A. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," 2017.

[4] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[5] Y-A. Chung, W-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Proc. INTERSPEECH*, 2019.

[6] D. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[7] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[9] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. INTERSPEECH*, 2021.

[10] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève, and L. Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *Proc. INTERSPEECH*, 2020.

[11] Z. Fan et al., "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. INTERSPEECH*, 2021.

[12] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, G. Wang, and P. Moreno, "Injecting text in self-supervised speech pretraining," *arXiv preprint arXiv:2108.12226*, 2021.

[13] W-N. Hsu, Y-H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?," in *IEEE ICASSP*, 2021.

[14] Z. Zhang, Y. Song, M-H. Wu, X. Fang, and L. Dai, "XLST: Cross-lingual self-training to learn multilingual representation for low resource speech recognition," *ArXiv*, vol. abs/2103.08207, 2021.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[16] Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky, "Massively multilingual adversarial speech recognition," *arXiv preprint arXiv:1904.02210*, 2019.

[17] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.

[18] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[19] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer, Speech, and Language*, 2002.

[20] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Proc. INTERSPEECH*, 2016.

[21] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for asr using lf-mmi trained neural networks," in *Proc. IEEE ASRU*, 2017.

[22] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," in *IEEE SLT*, 2018.

[23] V. Manohar, D. Povey, and S. Khudanpur, "JHU Kaldi system for arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning," in *Proc. IEEE ASRU*, 2017.

[24] A. Vyas, S. Madikeri, and H. Bourlard, "Lattice-free MMI adaptation of self-supervised pretrained acoustic models," in *IEEE ICASSP*, 2021.

[25] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *European Conf. on Speech Communication and Technology*, 1997.

[26] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998.

[27] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, 2001.

[28] Peter Ladefoged, "The revised international phonetic alphabet," *Language*, vol. 66, no. 3, pp. 550–552, 1990.

[29] Sibo Tong, Philip N Garner, and Hervé Bourlard, "An investigation of multilingual asr using end-to-end lf-mmi," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6061–6065.

[30] S. Madikeri et al., "Lattice-free maximum mutual information training of multilingual speech recognition systems.," in *INTERSPEECH*, 2020, pp. 4746–4750.

[31] S. Dalmia, R. Sanabria, F. Metze, and A. Black, "Sequence-based multi-lingual low resource speech recognition," in *IEEE ICASSP*, 2018.

[32] K. Kawakami et al., "Learning robust and multilingual speech representations," *Findings of EMNLP*, 2020.

[33] M. Riviere, A. Joulin, P-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *IEEE ICASSP*, 2020.

[34] W-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. INTERSPEECH*, 2021.

[35] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5171–5180.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015.

[37] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation," *ArXiv*, vol. abs/1805.04699, 2018.

[38] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.

[39] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *ArXiv*, vol. abs/2109.11680, 2021.

[40] J. Yi, Jianhua Tao, and Y. Bai, "Language-invariant bottleneck features from adversarial end-to-end acoustic models for low resource speech recognition," in *IEEE ICASSP*, 2019.

[41] Mary P Harper, "Data resources to support the babel program intelligence advanced research projects activity (IARPA)," .

[42] John C Wells, "Computer-coding the IPA: a proposed extension of SAMPA," *Revised draft*, vol. 4, no. 28, pp. 1995, 1995.

[43] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[44] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," *ArXiv*, vol. abs/1912.01703, 2019.

[45] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[46] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, "PyChain: A fully parallelized PyTorch implementation of LF-MMI for end-to-end ASR," in *INTERSPEECH*, 2020.

[47] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proc. Intl. Conf. Spoken Language Processing*, 2002.