Injecting Text and Cross-Lingual Supervision in Few-shot Learning from Self-Supervised Models

Matthew Wiesner, Desh Raj, Sanjeev Khudanpur

Human Language Technology Center of Excellence, Center for Language and Speech Processing,
The Johns Hopkins University, Baltimore, MD 21218, USA

Summary

We train Hybrid ASR Models by fine-tuning Wav2Vec 2.0 with LF-MMI, leveraging ***ALL*** available data, including cross-lingual supervisions and monolingual text

Few-shot learning with a cross-lingual universal phoneset model significantly outperforms fine-tuning by training a monolingual output layer from scratch.

Model	labeled / unlabeled pretrain (h)	pus	hat	kat
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**

Table 1. WER of few-shot Wav2Vec 2.0 systems on the BABEL pus, hat, and kat dev10h sets, using different cross-lingual fine-tuning methods. XLSR-53 (frozen) + XL-21 Multi trains the phonemic output layer on cross-lingual data while freezing the XLSR-53 model before jointly fine-tuning on target language speech.

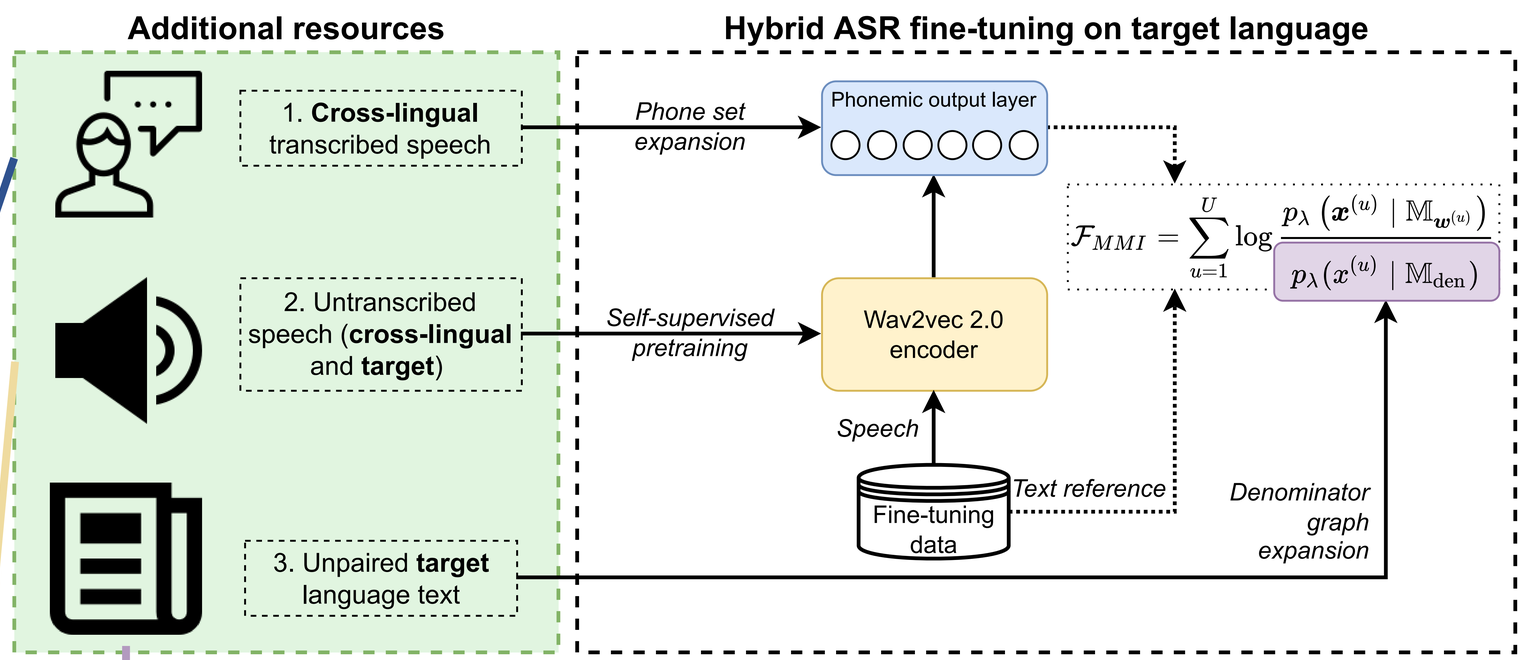
Pretrained models (XLSR-53) that are matched in both language AND acoustic channel work best.

Model	Pretraining (hours)		Few-shot (∼15m)			LLP (∼10h)			FLP (∼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**†

Table 1. WER of different fine-tuned Wav2Vec 2.0 systems on the BABEL pus, hat, and kat dev10h sets compared to WideResnet (WRN) monolingual and cross-lingual baselines. Fine-tuned Wav2Vec2.0 systems outperform the XL-21 cross-lingual WRN baseline.

Model	Telephone Speech	Matched Language	Multilingual	Largest
LV60	❌	❌	❌	❌
Large Robust	✅	❌	❌	❌
VoxPopuli-100k	❌	❌	✅	✅
XLSR-53	✅	✅	✅	❌

- Pre-training on telephone speech (matched acoustics) is helpful (**Large Robust**)
- Pre-training on a lot of multilingual speech is helpful (**VoxPopuli-100k**)
- Pre-training on matched acoustic and language speech is best (**XLSR-53**)



Few-shot training with LF-MMI using unpaired target language monolingual text to train the denominator phone-LM improves performance.

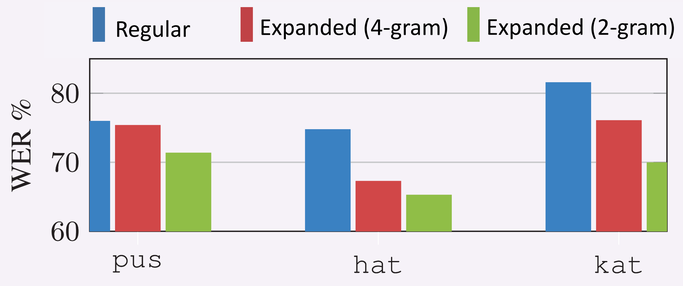


Figure 1. WER when fine-tuning using LF-MMI and expanding the denominator graph with extra, unpaired, monolingual target language text.

- Denominator expansion is helpful
- 2-gram phone-LM for denominator graph seems better
- Why?

Analysis

- Does the denominator graph LM help (LF-MMI vs. CTC)?
  - Yes. 0-gram LM performs much worse than 1,2,3,4-gram
- Extra-text reduces overfitting of denominator language model
  - Possibly why lower order n-gram models are helpful
- Using unpaired text in the denominator may bias the model to never predict sequences present in the unpaired text, but absent from the transcripts
  - Use the unpaired text with a smaller weight

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

Table 3. WER on Haitian dev10h set when fine-tuning with LF-MMI using a language model trained using expanded denominator graph with varying n-gram phone-LMs and external text weight

JOHNS HOPKINS UNIVERSITY

human language technology center of excellence

CENTER FOR LANGUAGE AND SPEECH PROCESSING