INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Comparing CTC and LFMMI for out-of-domain adaptation of wav2vec 2.0 acoustic model

Apoorv Vyas[1 2], *Srikanth Madikeri*[1], *Hervé Bourlard*[1 2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
{avyas, msrikanth, bourlard}@idiap.ch

Abstract

In this work, we investigate if the wav2vec 2.0 self-supervised pretraining helps mitigate the overfitting issues with connectionist temporal classification (CTC) training to reduce its performance gap with flat-start lattice-free MMI (E2E-LFMMI) for automatic speech recognition with limited training data. Towards that objective, we use the pretrained wav2vec 2.0 BASE model and fine-tune it on three different datasets including out-of-domain (Switchboard) and cross-lingual (Babel) scenarios. Our results show that for supervised adaptation of the wav2vec 2.0 model, both E2E-LFMMI and CTC achieve similar results; significantly outperforming the baselines trained only with supervised data. Fine-tuning the wav2vec 2.0 model with E2E-LFMMI and CTC we obtain the following relative WER improvements over the supervised baseline trained with E2E-LFMMI. We get relative improvements of 40% and 44% on the clean-set and 64% and 58% on the test set of Librispeech (100h) respectively. On Switchboard (300h) we obtain relative improvements of 33% and 35% respectively. Finally, for Babel languages, we obtain relative improvements of 26% and 23% on Swahili (38h) and 18% and 17% on Tagalog (84h) respectively.

Index Terms: speech recognition, wav2vec 2.0, e2e-lfmmi, ctc, cross-lingual adaptation

1. Introduction

Self-supervised training methods to learn powerful acoustic representations from untranscribed audio data have received a lot of attention recently [1, 2, 3, 4, 5, 6]. These methods aim to learn models that can extract good representations from the audio signal. The learnt representations or the model can later be adapted using supervised data to achieve state-of-the-art performance for automatic speech recognition (ASR) while greatly reducing the amount of transcribed training data which is both expensive and time-consuming to obtain.

Self-supervised training methods can broadly be grouped into two categories: (1) auto-regressive models that predict the future representations given only the the past inputs [2, 4] and (2) bidirectional models that predict masked parts of the input by looking at the full input [1, 3].

Currently, bidirectional models outperform autoregressive self-supervised models for ASR [1, 3]. In [1], the authors train a transformer model, wav2vec 2.0, that learns representations from raw audio data using contrastive learning. The model is trained on 1000 hours of unsupervised Librispeech [7] data and is later adapted on a 100 hour supervised subset of Librispeech data to achieve state-of-the-art performance. While wav2vec 2.0 model achieves state-of-the-art performance on Librispeech 100h subset, the authors only consider Connectionist Temporal Classification (CTC) [8] for acoustic model training. Moreover, they only investigate the performance of supervised adaptation on subsets of Librispeech dataset which was used for pretraining.

In [9], it was shown that self-supervised pretraining with reconstruction based masked acoustic modeling [3] remains useful on the out-of-domain datasets when the pretrained model is adapted with flat-start lattice free MMI (E2E-LFMMI) [10]. In this work, we not only investigate the performance of wav2vec 2.0 pretraining in such scenarios but we also look into the role of the training criterion for supervised adaptation of wav2vec 2.0 model.

The contribution of this work is twofold. First, we compare the effect of sequence discriminative training criterion for supervised adaptation. We show that fine-tuning the wav2vec 2.0 model with E2E-LFMMI and CTC criterion yield similar performances with neither consistently better than the other. Second, we adapt the wav2vec 2.0 model on out-of-domain conversational speech and on cross-lingual data to show that the wav2vec 2.0 pretraining provides significant gains over the models trained only with supervised data.

Specifically, we fine-tune the wav2vec 2.0 BASE model [1] pretrained on 1000 hours of Librispeech untranscribed data on three different datasets using E2E-LFMMI and CTC criteria. We first consider a 100h subset of Librispeech which was also seen during the pretraining stage. We next train on the three hundred hours of Switchboard data, which is conversational telephonic speech sampled at 8 KHz as opposed to the pretraining data which is read speech sampled at 16 KHz. Finally, we also evaluate on two low resource languages, Tagalog and Swahili, from the Babel dataset. To the best of our knowledge, we are the first ones to show that both E2E-LFMMI and CTC training achieve similar results on low resource languages when fine-tuned with a pretrained model.

The rest of the paper is organized as follows: In Section 2, we describe the details such as model architecture, fine-tuning learning rate, and decoding hyper-parameters. In Section 3, we present the details of the data preparation and baselines we considered. Finally, in Section 3.3, we present the results on comparison between E2E-LFMMI and CTC on Librispeech, Switchboard, and Babel datasets.

2. Our Method

To investigate the effectiveness of wav2vec 2.0 [1] pretraining for cross-lingual transfer and out of domain adaptation, we choose the BASE model as our pretrained model. This model is pretrained on 1000 hours of Librispeech data. The model contains 12 transformer layers each with 8 attention heads. The embedding dimension is set to 96 for each head and feed-forward dimension is set to 3072. In the following subsections we pro-

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1683

vide details of acoustic model training and decoding.

2.1. Acoustic Model Training

The input to the wav2vec 2.0 BASE model is the raw speech signal sampled at 16 KHz. During supervised adaptation, we pass the output of the wav2vec 2.0 transformer encoder to a seven layered factorized time-delay neural network (TDNNF). We fine-tune the wav2vec 2.0 BASE model together with TDNNF layers using E2E-LFMMI and CTC criteria.

Our E2E-LFMMI model is trained with biphone units while our CTC model is trained using character units. We train for a maximum of 30000 and 75000 updates for E2E-LFMMI and CTC criterion respectively where each update is over 1500 seconds of speech input. All our models are trained with 3 GTX-1080 Ti GPUs with gradient synchronized training. We use gradient accumulation to obtain a batch size of 1500 seconds.

For both E2E-LFMMI and CTC, we update the BASE model parameters with a learning rate that is linearly increased to 3e-5 over 10% of the updates, then held constant for 40% of the updates, and linearly decreased for 50% of the updates. For TDNNF model parameters, we use a learning rate that is 20 times the current learning rate for BASE model updates. We also use the natural gradient update [11] for training with the E2E-LFMMI objective.

All our models are trained with PyTorch [12] [1]. For fine-tuning with CTC, we use the Fairseq toolkit [13] and for E2E-LFMMI, we use the Espresso toolkit [14] which uses PyChain [15] for the implementation of LFMMI loss. We use the PyTorch implementation for natural gradient update from [16].

2.2. Decoding

For all models trained with E2E-LFMMI, we use the WFST decoder from [17] with a beam width of 15. For the models trained with CTC, we use the decoder from [18] with a beam width of 500. We always use the language model from Kaldi recipes [17] which are trained with SRILM [19]. We found this to give better results than KenLM [20].

3. Experiments

In the following, we discuss in detail the datasets and augmentation that we used, baselines we compare against, and finally the results for the experiments.

3.1. Datasets

We evaluate the performance of the wav2vec 2.0 BASE model on the same datasets as [9]. The three datasets are selected in increasing order of difficulties. We first consider the 100 hour clean subset of Librispeech [7]. This is the easiest setting because the training set is a part of the 1000 hours of Librispeech pretraining data. We next consider the Switchboard [21] dataset with 300 hours of transcribed data. In contrast to Librispeech pretraining data, Switchboard has conversational speech sampled at 8 KHz making it a more difficult out-of-domain setting. We finally fine-tune on two of the Babel [22] languages: Tagalog (84h) and Swahili (38.5h). We consider this to be the hardest setting as there is both language and acoustic conditions mismatch. For all our experiments we apply speed and volume perturbation to increase the dataset by three times.

[1] We will open-source our training scripts and models

3.2. Baselines

We reproduce the supervised baselines from [9]. We train from scratch a twelve layered TDNNF model using 80 dimensional filter bank features. Our baseline is trained with the E2E-LFMMI objective. For all TDNNF models, we set hidden layer dimension to 1024 and bottleneck dimension to 128.

We also compare against pretrained model trained with Masked Acoustic Modeling (MAM) objective [3]. For the MAM pretraining, the input to the network is a sequence of masked or noise corrupted acoustic features. The model attempts to reconstruct the original input given the corrupted input and is trained with L_1 loss. This is different from wav2vec 2.0 pretraining which masks input segments and uses cross-entropy loss to contrast between learned representation at the masked time-steps with representations at other time steps.

For comparison against MAM pretraining, we refer to results from [9] where the pretrained model was fine-tuned with E2E-LFMMI objective on the same datasets. Note that this model had 6 attention heads per layer in comparison to wav2vec 2.0 base model which has 8 attention heads per layer. This results in a model of smaller capacity. We believe that this doesn't affect our conclusions as there is no significant performance difference between models of different capacity that are pretrained with MAM [3].

Finally, when available, we provide results for models trained with CTC criterion using only the supervised data. Note that the dataset augmentation and model architectures can differ significantly from our TDNNF baseline trained with E2E-LFMMI criterion and they might not be strictly comparable.

3.3. Results

In the following we compare the Word Error Rate (WER) achieved by supervised adaptation of pretrained wav2vec 2.0 BASE model using E2E-LFMMI and CTC objectives. We refer to the 12 layered TDNNF baseline trained only on supervised data as *TDNNF*. Our wav2vec 2.0 BASE model with seven TDNNF layers is referred to as *wav2vec2-base*. The masked acoustic pretrained model from [9] is referred to as *MAM*.

3.3.1. Librispeech (100 hours)

In this experiment, we discuss the case when the supervised training data was seen during the pretraining. We present our main results in Table 1. For evaluation, we select the model that achieves lowest WER on the *dev-other* set. For E2E-LFMMI models, we first decode using a 3-gram language model and then rescore using a 4-gram language model. For the CTC model, we directly decode with the 4-gram language model as done in [1]. The beam search decoder from [18] can efficiently decode with 4-gram language model on a single GPU.

In rows (a) and (b), we present the results for training with only supervised data using E2E-LFMMI and CTC. From the comparison on the *dev* set it can be seen that CTC training requires additional regularization and modifications to the deep neural network training to reach a similar level of performance.

In rows (d) and (e), we compare the performance of fine-tuning the wav2vec 2.0 BASE model with E2E-LFMMI and CTC loss. It can be seen that both models reach similar level of performance providing $\sim$ 12.7% and $\sim$ 11.5% absolute WER improvements over the supervised *TDNNF* baseline on the noisy portion of the test set. Note that we did not apply any additional regularization or changes to train with the CTC loss.

Furthermore it can be noticed from (c) that wav2vec 2.0

Table 1: *Comparison of word error rates (WER) (in %) on the clean and other parts of the Librispeech test set with and without 4-gram language model rescoring. Fine-tuning the pretrained wav2vec 2.0 BASE model with E2E-LFMMI or CTC gives similar level of performance without any additional regularization techniques. Both models outperform the masked-acoustic modeling (MAM) pretraining and fully supervised baselines.*

			dev	**test**			
			3-gram	**3-gram**		**4-gram**	
	Architecture	Criterion	clean	clean	other	clean	other
			Supervised Only				
(a)	TDNNF-large	E2E-LFMMI	8.3	8.6	26.3	6.0	19.9
(b)	Bi-LSTM [23]	CTC	11.1	-	-	-	-
	+ max perturbation	CTC	9.8	-	-	-	-
	+ cascade dropout	CTC	7.9	8.7	26.1	-	-
			Pretraining + Supervised (others)				
(c)	MAM [9]	E2E-LFMMI	-	7.8	20.2	5.3	14.7
			Pretraining + Supervised (ours)				
(d)	wav2vec2-base	E2E-LFMMI	-	**4.4**	**8.9**	3.5	**7.3**
(e)	wav2vec2-base	CTC	-	-	-	**3.3**	8.5

BASE model fine-tuned with either loss significantly outperforms the model pretrained with masked acoustic modeling.

*3.3.2. Switchboard (*300 *hours)*

In this experiment, we explore the out-of-domain scenario in which pretraining data and supervision data share the same language however they are dissimilar with respect to content, and acoustic conditions. Switchboard dataset comprises of telephonic conversations which are recorded at 8 KHz. This is different from Librispeech pretraining data which comprises of read speech sampled at 16 KHz.

For fine-tuning with wav2vec 2.0 BASE model, we resample the Switchboard recordings at 16 KHz. To train the baseline *TDNNF* acoustic model using only the transcribed data, we use the 8 KHz recordings.

For evaluation we select the model that gives smallest WER on the held out development set. Once again, for E2E-LFMMI models, we first decode using a 3-gram language model followed by rescoring with 4-gram language model trained on Switchboard and Fisher transcripts. For model fine-tuned with CTC, we directly decode with the 4-gram language model.

Table 2 compares the WER for the E2E-LFMMI and CTC models trained from scratch as well as those fine-tuned from models pretrained on Librispeech data. Consistent with Librispeech experiment, we see that for models trained using only the supervised data, CTC requires additional regularization techniques to reach the same level of performance as E2E-LFMMI. Furthermore, the CTC baseline presented in [24] applies fMLLR transformation which typically provides additional gains for end-to-end ASR [25].

Once again, it can be seen that fine-tuning wav2vec 2.0 BASE model with E2E-LFMMI or CTC leads to comparable performances. Both models significantly outperform the baselines trained only with supervised data as well as the model pretrained with masked acoustic modeling. Similar to the previous experiment, we do not use any additional regularization techniques for CTC training. Note that we get absolute WER improvements of $\sim 3.7\%$ and $\sim 7\%$ over the *TDNNF* baseline on the switchboard and callhome portion of evaluation sets.

Table 2: *Comparison of word error rates (WER) (in %) on eval2000 test set for the 300 hours Switchboard task. The 3-gram language model is based on Switchboard, whereas the 4-gram employs Switchboard+Fisher training set transcripts. Fine-tuning wav2vec 2.0 BASE model with E2E-LFMMI or CTC improves the performance over the baselines with no additional regularization needed for the CTC model.*

		Hub5'00 (eval2000)			
		3-gram		**4-gram**	
Model	**Criterion**	**SW**	**CH**	**SW**	**CH**
	Supervised Only				
TDNNF-large	E2E-LFMMI	11.8	22.5	10.3	20.3
Bi-LSTM [24]	CTC	-	-	12.2	21.8
+ seq. noise	CTC	-	-	10.9	21.3
+ regularization	CTC	-	-	10.6	19.5
	Pretraining + Supervised (Others)				
MAM [9]	E2E-LFMMI	10.9	20.4	9.4	18.2
	Pretraining + Supervised (Ours)				
wav2vec2-base	E2E-LFMMI	**7.3**	**14.5**	6.7	13.7
wav2vec2-base	CTC	-	-	**6.6**	**13.2**

3.3.3. Babel: Swahili and Tagalog

In this experiment, we evaluate the effectiveness of wav2vec 2.0 pretraining when the BASE model is fine-tuned on cross lingual data. For our evaluation, we consider two low resource languages, Swahili and Tagalog, from the Babel dataset. Once again, we resample the audio recordings at 16 KHz for fine-tuning the pretrained model. For our *TDNNF* baseline trained only supervised data, we use the original recordings sampled at 8 KHz.

For both languages, we report the results on the *dev10h* development part due to the lack of a separate evaluation set. We use the 2 hour development set for model selection. For both E2E-LFMMI and CTC models, we use 3-gram language model for decoding using the previously described hyperparameters

for beam search. We do not consider the non-language symbols for scoring on these datasets.

Table 3 compares the WER for the models trained from scratch to the models pretrained on the Librispeech dataset. Note that we could not find any CTC baseline that is trained only on the supervised data and provides competitive performance to E2E-LFMMI training.

Once again, we see that for both Swahili and Tagalog, wav2vec 2.0 BASE model fine-tuned with E2E-LFMMI and CTC obtain similar performance. Both models outperform the baselines trained only with supervised data as well as the model pretrained with masked acoustic modeling.

We additionally report the WER for the XLSR-10 model from [26]. This is a large wav2vec 2.0 multilingual model pretrained on 10 languages. As can be seen, we get a much better word error rate on Swahili and a comparable performance on Tagalog. We think that the results might not be directly comparable because we use speed and volume perturbation for data augmentation and do not score on non-language symbols. Additionally, XLSR-10 uses KenLM for decoding while we use SRILM. In our experiments, we noticed a significant degradation in WER using KenLM. Despite these differences, it is clear from our results that wav2vec 2.0 BASE model pretrained on Librispeech still offers a very competitive baseline to the large multilingual model.

Table 3: *Comparison of word error rates (WER) (in %) on dev10h set for the Swahili and Tagalog languages of the Babel dataset. Fine-tuning the pretrained wav2vec 2.0 BASE model significantly outperforms the monolingual and MAM baselines. Note that while we use SRILM, XLSR-10 model uses KenLM for decoding and does not use speed or volume perturbation.*

Model	Criterion	Swahili	Tagalog
Supervised Only			
TDNNF	E2E-LFMMI	39.5	44.9
Pretraining + Supervised (Others)			
MAM [9]	E2E-LFMMI	36.7	43.4
XLSR-10 (Large) [26]	CTC	35.5	37.3
Pretraining + Supervised (Ours)			
wav2vec2-base	E2E-LFMMI	**29.4**	**36.9**
wav2vec2-base	CTC	30.4	37.3

4. Conclusions

In this work, we investigate the effects of the sequence discriminative training criteria for the supervised adaptation of pretrained wav2vec 2.0 BASE model. We show that fine-tuning wav2vec 2.0 BASE model with either E2E-LFMMI or CTC gives similar performance with no additional regularization needed for CTC training. We further show that wav2vec 2.0 pretraining provides significant gains and outperforms models pretrained with masked acoustic modeling even for out-of-domain and cross-lingual adaptation.

In future, we will compare the performance of the monolingual wav2vec 2.0 model to the multilingual model of similar capacity to understand the advantages of multilingual pretraining. We will additionally explore supervised fine-tuning with multilingual data to further improve the performance in low resource settings.

5. Acknowledgments

The research leading to these results has received funding from Swiss National Science Foundation project SHISSM (Sparse and hierarchical Structures for Speech Modeling), grant agreement 200021-175589. The research is also partially based upon the work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract #FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

6. References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the international conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

[3] A. T. Liu, S.-W. Li, and H. yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *arXiv*, 2020.

[4] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, 2019.

[5] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019.

[6] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *Proceedings of ICASSP*, 2020.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proceedings of ICASSP*, 2015.

[8] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

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

[10] H. Hadian *et al.*, "Flat-start single-stage discriminatively trained hmm-based models for asr," *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[11] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with Natural Gradient and Parameter Averaging," *arXiv*, 2014.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019.

[13] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[14] Y. Wang, T. Chen, H. Xu, S. Ding, H. Lv, Y. Shao, N. Peng, L. Xie, S. Watanabe, and S. Khudanpur, "Espresso: A fast end-to-end neural speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[15] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, " PyChain: A Fully Parallelized PyTorch Implementation of LF-MMI for End-to-End ASR," in *Proc. Interspeech*, 2020.

[16] S. Madikeri, S. Tong, J. Zuluaga-Gomez, A. Vyas, P. Motlicek, and H. Bourlard, "Pkwrap: a pytorch package for lf-mmi training of acoustic models," *arXiv*, 2020.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[18] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2letter++: A fast open-source speech recognition system," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[19] A. Stolcke, "Srilm - an extensible language modeling toolkit." in *Proceedings of Interspeech*, 2002.

[20] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011.

[21] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech, and Signal Processing,*, 1992.

[22] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued." in *The Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2014.

[23] J. Billa, "Improving lstm-ctc based asr performance in domains with limited training data," *arXiv*, 2017.

[24] K. Audhkhasi, G. Saon, Z. Tüske, B. Kingsbury, and M. Picheny, "Forget a Bit to Learn Better: Soft Forgetting for CTC-Based Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019.

[25] N. Tomashenko and Y. Estève, "Evaluation of Feature-Space Speaker Adaptation for End-to-End Acoustic Models," in *LREC 2018*, 2018.

[26] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv*, 2020.