What kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced ASR task?

Péter Mihajlik[1,2], *Máté Soma Kádár*[1,2], *Gergely Dobsinszki*[1,2], *Yan Meng*[1], *Meng Kedalai*[1], *Julian Linke*[3], *Tibor Fegyó*[1,4], *Katalin Mády*[2]

[1] Budapest University of Technology and Economics, Hungary,
[2] Research Centre for Linguistics, Hungary,
[3] Graz University of Technology, Austria, [4] SpeechTex, Hungary

mihajlik@tmit.bme.hu

Abstract

Most languages are under-resourced for Automatic Speech Recognition (ASR), and most relevant tasks are related to the transcription of spontaneous speech. The application of cross- or multi-lingual pre-training is inevitable, however, the selection of the best pre-trained model or data/method is not straightforward. In this paper, we introduce a case study for Hungarian, targeting good quality spontaneous speech while monitoring the ASR performance of read speech. Transformer/conformer-based end-to-end neural models with supervised cross-lingual, self-supervised cross- and (massively) multi-lingual and weakly supervised multi-lingual pre-training are fine-tuned and evaluated. Surprisingly, a relatively small-scale tri-lingual (SSL pre-trained) model won the competition by a large margin over very large-scale models trained on more Hungarian data. The results revealed that the composition of pre-training data in terms of language and speech style was essential, bigger size or higher number of languages did not necessarily come with improvement, and no transcription was required in the pre-training for the best performance.

Index Terms: automatic speech recognition, less-resourced languages, pre-training, spontaneous speech, SSL, weak-supervision, conformer, wav2vec2.0, Hungarian.

1. Introduction

Recently ASR (Automatic Speech Recognition) of smaller, under-resourced languages has gained more support by the introduction of (massively) multi-lingual speech recognition models, such as Whisper [1], USM [2] and MMS [3]. These large-scale developing models are evaluated on multi-lingual benchmarks, e.g., on FLEURS [4]. Improvements in overall performance, however, tell us little about ASR accuracy for a given task in a given (less-resourced) language and speech style. For a specific task – in our case, spontaneous Hungarian speech recognition –, the best practice is still to use large-scale pre-trained models and fine-tune them with in-domain data. To the question in the title – what kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced ASR task – we could not find an up-to-date answer, so we conducted several experiments including (but not limited to) the latest publicly available pre-trained models.

We investigated recent neural architectures and training schemes, such as Conformer [5] using only supervised training from scratch and also with cross-language supervised pre-training + fine-tuning [6] on the BEA-Base [7] Hungarian training set. In large-scale weak supervision based experiments (Whisper) [1] we report zero-shot and fine-tuned results. Finally, classic and most recent self-supervised wav2vec2.0 based pre-training setups [8, 9, 10, 3] were also fine-tuned on the BEA-Base training data and evaluated on spontaneous – and contrastively on read/repeated – speech. Additionally, supplementary evaluations are reported on the Hungarian CommonVoice (CV) v12.0 test set when available. We show that even if training (fine-tuning) does involve a large proportion of spontaneous speech, ASR of this speech style is still challenging if compared to read speech. Based on the Hungarian BEA-Base [7] where spontaneous (including conversational) and read/repeated speech is collected from each speaker under the same conditions, and evaluation subsets are defined correspondingly, a clear contrast between speech registers (spontaneous vs. non-spontaneous) in ASR was measured.

One of our key findings is that pre-training data with the highest proportion of spontaneous-like speech (such us parliamentary debates) in the target language led to the optimum performance – even though other approaches used additional target language data.

This work is a significant extension of our earlier study [7] introducing the BEA-Base benchmark and various ASR baselines. In this paper, we apply more recent approaches that clearly outperform all previous results, and we provide new insights into the improvements and pre-training model selection.

2. Data sets and ASR task

2.1. Database statistics

For supervised end-to-end acoustic model training/fine-tuning, we always used the "train-114" subset of BEA-Base v0.1 and applied the "dev_spont" as validation set (see details in Table 1). For evaluations, we primarily used the spontaneous (mostly conversational) "eval_spont" speech subset. For more general conclusions, the non-spontaneous (read+repeated) "eval_repet" and "dev_repet" and CV Hungarian (v12.0) [11] test sets are reported regarding WER (Word Error Rate) and CER (Character Error Rate) as well, where both reference and hypothesis transcriptions were normalized. (For further details on the exact composition of the subsets, see [7].)

To train a language model (LM), the spoken language (SPOK) sub-corpus of the Hungarian Gigaword Corpus (HGC) [7] was used. Refer to Table 1 for the statistics of the corpora applied and for SPOK-trained word 3-gram based perplexity results obtained by using the KenLM [12] tool.

2.2. Speech data visualisation

To check the composition of training/evaluation speech data, similarity analysis is carried out based on the quantized latent representations of a pre-trained wav2vec2.0 model. We follow the recipe of [8], but we calculate the codebook frequency vectors on a per speaker basis instead of per language (Figure 1,

Table 1: *Main characteristics of data sets used in the experiments.*

	HGC SPOK	BEA-Base train-114	dev-repet	dev-spont	eval-repet	eval-spont	CV test
Length [hours]	-	71.2	0.65	4.02	0.95	4.91	6.8
Num of speakers	-	114	10	10	16	16	220
Num of segments	-	76 881	568	4 893	858	5 693	4 871
Num of characters	516.84M	3.1M	28 467	154 994	43 448	197 738	250 709
Num of words	56.13M	0.56M	4 110	27 939	6 229	35 178	35 485
3-gram PPL	-	-	924	771	846	857	2 387
OOV rate [%]	-	-	1.6	1.9	1.4	1.7	3.1

left). A simple mean is calculated within each speech style cluster to get the centroids as can be seen in Figure 1 (right). For the visualization experiments, we used the Uralic wav2vec2.0-large model[1]. More details on the speaker-wise similarity-based visualization can be found in [13]. As Figure 1 shows, there is a significant overlap between various speech styles, but spontaneous (Discourse, Interview, Summarization, and Opinion) module centroids are clearly separated from non-spontaneous ones (Readsent, Readtext and Repeat) confirming the appropriateness of BEA-Base subset definitions. For more information on BEA modules and subsets, see [7].

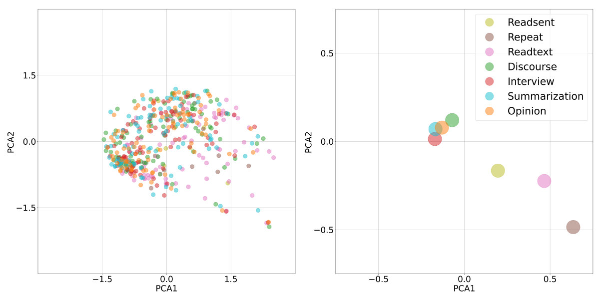

Figure 1: *Quantized latent representation based data visualisation on a per speaker (left) and per speech type (right) basis.*

3. Supervised learning based results

In the following, we introduce the experimental results obtained with various Conformer [5] approaches. In all configurations the NVIDIA NeMo toolkit v1.6.2 was applied with the default hyper-parameter settings – unless mentioned otherwise. We used CTC loss [14] and a simple convolutional decoder. For data augmentation, SpecAugment [15] and speed perturbation was applied. In all experiments we used a batch size of 32, a total epoch number of 200, and a beam size of 150 for LM rescoring. For character based setups, the same word 3-gram language model trained on HGC-SPOK was used as in Table 1. (Note that our LM is different from the one used in the baseline [7], trained on more data but on a text independent of BEA-Base.) A SentencePiece tokenizer [16] was trained on the train-114 transcriptions with 128 unigram units and was applied on the HGC-SPOK before training subword 6-gram LM. Two RTX A6000 GPUs served as hardware accelerators.

[1]https://github.com/facebookresearch/voxpopuli

3.1. Training from scratch

First, we wanted to set up Conformer baselines and check if they can outperform the previous convolutional QuartzNet [17] baselines published in [7]. For this, we trained Conformer models with subword output labels for both the small and medium sizes with a learning rate of 1. As can be seen in Table 2, greedy (no LM) results are clearly better than the convolutional baseline, and adding a 6-gram subword language model improved the accuracies further (unlike in the case of the QuartzNet model).

3.2. Pre-training & cross-language transfer learning

Second, we applied models pre-trained with English data by NVIDIA and fine-tuned them on the Hungarian train-114 set. All the pre-trainin details and models can be accessed through the NVIDIA Catalog[2] using model name formats such as *stt_en_conformer_ctc_large*. The results are shown in Table 3. In spite of the significant difference between the acoustics of English and Hungarian, the positive effects of cross-language transfer learning can be clearly observed.

4. Large-scale weak supervision based results

Once the multi-lingual Whisper [1] ASR models became available, it was an obvious task to test and fine-tune them on the BEA-Base data set. Since Whisper training covers Hungarian, beyond fine-tuning, zero-shot experiments were carried out. For the ASR experiments we used the SpeechBrain toolkit [18] and the same fine-tuning setup (e.g., same augmentation) as described in the previous section. According to SpeechBrain's CV recipe for Whisper fine-tuning, the encoder part of the models was frozen in our experiments. We applied 20 epochs on the BEA-Base training set with a batch size of 12, an initial learning rate of 0.00003 and a max decode ratio of 0.1 – all other hyper-parameters were unchanged. No LM was used since the end-to-end model itself applied a heavy decoder. Only medium and large models were used because we wanted to achieve the highest accuracies possible.

As shown in Table 4, the results are somewhat disappointing. On BEA-Base, even the largest (v2) model with fine-tuning could not outperform the previous Conformer model trained by simple cross-lingual transfer learning and having less than one tenth of parameters. In terms of CV test results, Whisper showed better accuracies which can be attributed to its noise robust (pre-)training.

[2]https://catalog.ngc.nvidia.com/

Table 2: *CER(%) / WER(%) results with supervised training from scratch.*

Model / Num of parameters	**LM**	**BEA-Base** dev-repet	dev-spont	eval-repet	eval-spont	**CV** test
QuartzNet15x3 [7] / 12.7M	-	2.20 / 9.73	8.33 / 25.20	2.91 / 11.56	8.84 / 26.70	-
	3-gram	1.86 / 6.50	10.0 / 25.50	2.36 / 6.86	10.76 / 26.83	-
Conformer-Small / 13M	-	2.13 / 10.71	7.77 / 23.90	2.87 / 12.73	8.21 / 25.31	14.47 / 49.83
	6-gram	1.53 / 7.27	7.14 / 21.44	2.01 / 7.98	7.62 / 22.78	**13.03 / 42.70**
Conformer-Medium / 30.5M	-	2.10 / 9.93	7.77 / 23.25	2.61 / 10.98	8.14 / 24.93	14.72 / 49.80
	6-gram	**1.26 / 5.67**	**7.00 / 19.74**	**1.53 / 5.65**	**7.30 / 21.01**	13.32 / 42.98

Table 3: *CER(%) / WER(%) results based on cross-lingual (English to Hungarian) pre-training + fine-tuning.*

Model / Num of parameters	**LM**	**BEA-Base** dev-repet	dev-spont	eval-repet	eval-spont	**CV** test
QuartzNet15x5 [7] / 18.9M	-	1.96 / 8.93	7.55 / 23.55	2.58 / 10.63	7.96 / 24.87	-
	3-gram	1.66 / 5.99	9.52 / 24.29	1.92 / 5.83	9.62 / 25.23	-
Conformer-Small / 13M	-	1.92 / 9.64	6.14 / 20.02	2.51 / 11.22	6.48 / 21.39	10.34 / 40.78
	6-gram	1.21 / 5.43	5.53 / 17.00	1.31 / 4.96	5.82 / 17.77	9.23 / 34.77
Conformer-Medium / 30.5M	-	1.73 / 8.56	5.58 / 18.45	1.88 / 8.17	5.83 / 19.60	8.36 / 35.55
	6-gram	1.09 / 4.53	**5.06** / 15.94	1.15 / 4.40	5.27 / 16.52	**7.46 / 30.42**
Conformer-Large / 121M	-	1.14 / 5.48	5.09 / 16.44	1.26 / 5.20	5.29 / 17.24	8.77 / 34.79
	6-gram	**0.97 / 4.45**	5.08 / **15.64**	**0.98 / 3.66**	**5.24 / 16.25**	8.02 / 30.82

Table 4: *CER(%) / WER(%) results based on large-scale weak supervision.*

Model / Num of parameters	**BEA-Base** dev-repet	dev-spont	eval-repet	eval-spont	**CV** test
Whisper-medium zero-shot / 769M	4.82 / 21.92	17.97 / 37.18	5.18 / 22.33	19.46 / 38.67	6.91 / 27.61
Whisper-large-v2 zero-shot / 1550M	3.74 / 17.54	17.06 / 33.17	3.99 / 18.04	17.06 / 32.76	**5.27 / 20.41**
Whisper-medium fine-tuned / 769M	1.31 / 5.38	7.96 / 18.83	1.50 / 4.90	9.33 / 20.60	7.83 / 27.93
Whisper-large-v2 fine-tuned / 1550M	**1.01 / 4.45**	**7.10 / 16.96**	**1.23 / 4.37**	**8.46 / 18.69**	6.19 / 23.69

5. Self-supervised pre-training based results with wav2vec2.0

As could be observed, supervised pre-training even on distant languages (English vs. Hungarian) improved the results significantly. Supervised (or weakly supervised) pre-training, however, will always have its limits due to the price, amount, quality and language of the available transcription data. Therefore, the introduction of Self-Supervised Learning based pre-training (SSL) at a large scale [19, 8, 9] made a real breakthrough in ASR. Currently, one of the most popular approaches is the Transformer based [20] wav2vec2.0 [19] framework. Several thousands of pre-trained/fine-tuned models are available in public model repositories (e.g., HuggingFace[3]). A major question is, which one to apply and fine-tune for the given down-stream task (ASR of spontaneous Hungarian). As [7] pointed out, using an already fine-tuned model is not effective in our case, therefore in this study we restrict the question to the selection among purely SSL-trained wav2vec2.0 large models with 300 million parameters (other structures were not considered due to lack of performance or computational resources).

[3]https://huggingface.co/

5.1. SSL pre-trained models

At first, we selected the model trained purely on English [19] so that the results may be comparable to the previous cross-lingual setup. Then we applied various multilingual models [8, 9, 3]. Each of these models were also trained on Hungarian speech data, including VoxPopuli (European parliamentary) speech [10] in the latter two cases. Finally, the only wav2vec2-large model left trained partially on Hungarian was the Uralic model [10] where (untranscribed) training data encompassed 10.6k hours of Estonian, 14.2k hours of Finnish and 17.7k hours of Hungarian speech.

5.2. Fine-tuning

We adopted an architecture described in SpeechBrain's CV recipe: a wav2vec2.0 encoder paired with an attentional GRU decoder. Again, we used the SentencePiece tokenizer [16] on the train-114 set with a unigram vocabulary size of 600. Data augmentation, i.e., speed perturbation with 0.95, 1.0 and 1.05 factors and SpecAugment [15] was employed during the fine-tuning phase. Joint CTC+Attention loss [21] with a CTC weight of 0.4 was calculated in the first 20 epochs and only attentional loss with label smoothing [22] in the remaining 80 epochs. The effective batch size was 12. Separate optimizers were utilized for the wav2vec2.0 encoder part, i.e., Adam [23] (alpha=1e-4,

Table 5: *CER(%) / WER(%) results based on self-supervised pre-training + fine-tuning with wav2vec2.0 encoder + attentional decoder.*

Model / Pre-train data [hours]	**Langs.**	**LM**	**BEA-Base**				**CV**
			dev-repet	**dev-spont**	**eval-repet**	**eval-spont**	**test**
wav2vec2-large-lv60 [19] / 60k	1	-	3.19 / 8.61	5.45 / 18.01	2.59 / 8.46	5.94 / 19.17	11.21 / 36.48
wav2vec2-large-xlsr-53 [8] / 56k	53	-	1.12 / 5.09	5.17 / 16.24	2.09 / 5.81	5.53 / 16.62	10.49 / 34.18
wav2vec2-xls-r-300m [9] / 440k	128	-	1.15 / 5.28	4.70 / 14.95	2.39 / 6.16	5.11 / 15.61	8.57 / 30.53
wav2vec2-mms-300 [3] / 491k	1406	-	1.15 / 5.40	5.29 / 17.07	2.22 / 6.65	5.83 / 18.82	9.13 / 34.89
wav2vec2-uralic [10] / 42.5k	3	-	*0.74 / 3.50*	*3.56 / 11.63*	*1.67 / 4.24*	*3.68 / 11.55*	*5.77 / 21.26*
		Neural	**0.67 / 3.09**	**3.22 / 10.47**	**0.67 / 2.42**	**3.32 / 10.50**	**4.47 / 17.21**

beta1=0.9 and beta2=0.999), and Adadelta [24] for all the additional layers (alpha=1.0, rho=0.95). We found a value of 0.15 for the Transformer dropout as optimal. In all other aspects we kept the recipe unchanged.

5.3. Decoding with Transformer LM

This time, a GPT-based architecture [25] was applied. LM pre-training was performed on the tokenized HGC-SPOK corpus, followed by domain-specific fine-tuning on the BEA-Base training set. The Transformer LM encompassed 14 stacked encoder blocks with 16 attention-heads per block. The dimension of the embedding layer was set to 1024 and the inner fully-connected layer size was fixed at 3072 totaling in 149.3M trainable parameters. The LM training lasted for 20 epochs with an effective batch size of 512. For the fine-tuning phase, the first 4 layers were frozen.

To integrate the LM to the wav2vec2.0 acoustic model shallow fusion was performed [26]. The evaluation of the ensemble model relied on beam search with a beam size of 8 taking the CTC loss into account with a factor of 0.014. The output probabilities of the Transformer LM were added with a weight of 0.285. The most probable beams were normalized by their lengths [27] to deter the system from preferring shorter sequences. Both the acoustic and the language model's output were sampled with a temperature of 1.05. Setting the end-of-sentence threshold [28] to 2.5 and coverage penalty [29] to 3.0 seemed to produce the best validation results in our experiments. In the LM-free experiments the beam width of beam search was set to 80. NVIDIA A6000 and RTX 3090 GPU's were used both for acoustic and LM fine-tuning and tests. For the results, see Table 5.

5.4. SSL-based Results & Discussion

Regarding monolingual (English) pre-training + fine-tuning to Hungarian, it can be confirmed that self-supervised pre-training is not necessarily superior to supervised or weakly supervised pre-training – as spontaneous WER results are worse than in the best of previous cases (Conformer-large/Whisper-large-v2 fine-tuned).

The effect of multi-lingual pre-training looked convincing: both the xlsr-53 (53 languages) and xls-r (128 languages) setup provided comparable or better results (withouth LM) then the Conformer-large supervised cross-lingual approach. The massively multi-lingual model (MMS), however, performed slightly worse than the other SSL multilingual models. This was unexpected because it used all the training data of the xls-r model (including Hungarian). Surprisingly, the relatively small-scale Uralic model drastically outperformed all other approaches even without a language model. Adding a Transformer LM reduced the error rates further – almost halving the error rates of the lv60 (cross-lingual SSL) approach. Interestingly, the Uralic pre-training (+ fine-tuning on BEA-Base) provided almost the best results on the CV test set although no CV data of any language was involved in either pre-training or fine-tuning. The most straightforward explanation of the results may be that not the absolute quantity of all/target language speech matters but the relative quantity (the higher the better) and quality (the more spontaneous the better) of the target language. Possibly, adding related or structurally similar languages to the SSL pre-training enhances the results substantially. The best, wav2vec2-uralic-based fine-tuned model along with the Transformer LM was made available for the Research Community (after registration)[4].

6. Conclusions

Despite the recent advances in large-scale multi-lingual ASR, the recognition of spontaneous speech in a less-resourced language still remains challenging. BEA-Base provided a unique opportunity to compare ASR results on spontaneous and non-spontaneous subsets directly since they had been recorded from the same speakers in identical conditions. We investigated powerful state-of-the-art cross- and multi-lingual pre-training approaches to decrease primarily the spontaneous Hungarian error rates. Large-scale weakly supervised and massively multilingual self-supervised pre-trained models were outperformed significantly by a relatively small-scale tri-lingual model. We think that the superior results are due to the highest density of target language and speech style (Hungarian spontaneous speech) in the pre-training data set. The results suggest that increasing data sizes and number of languages in multilingual pre-trained models may not necessarily result in lower error rates for specific under-resourced tasks, and so the development of mono- or oligo-lingual pre-trained models seems unavoidable. Adding more spontaneous speech to SSL data sets in general (without the need for transcription), however, has the potential to improve ASR results in real-life applications in a cost-effective way.

7. Acknowledgement

This work was supported partially by NKFIH-828-2/2021(MILab), by the NVIDIA Academic Hardware Grant and by the NKFIH K143075 and K135038 projects of the NRDI Fund.

[4]https://phon.nytud.hu/bea/bea-base.html

8. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[2] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[3] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[4] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[6] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," *arXiv preprint arXiv:2005.04290*, 2020.

[7] P. Mihajlik, A. Balog, T. E. Graczi, A. Kohari, B. Tarján, and K. Mady, "BEA-base: A benchmark for ASR of spontaneous Hungarian," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1970–1977. [Online]. Available: https://aclanthology.org/2022.lrec-1.211

[8] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[9] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[10] C. Wang, M. Riviere, 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," *arXiv preprint arXiv:2101.00390*, 2021.

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[12] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[13] J. Linke, M. S. Kádár, G. Dobsinszki, P. Mihajlik, G. Kubin, and S. Barbara, "What do self-supervised speech representations encode? an analysis of languages, varieties, speaking styles and speakers," in *InterSpeech2023*. ISCA, 2023.

[14] 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, pp. 369–376.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[16] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[17] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[22] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[26] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 369–375.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[28] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," *arXiv preprint arXiv:1904.02619*, 2019.

[29] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.