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

A comparison of supervised and unsupervised pre-training of end-to-end models

Ananya Misra, Dongseong Hwang, Zhouyuan Huo, Shefali Garg, Nikhil Siddhartha, Arun Narayanan, Khe Chai Sim

Google LLC, USA

{amisra,dongseong,zhhuo,shefgarg,nikhilsid,arunnt,khechai}@google.com

Abstract

In the absence of large-scale in-domain supervised training data, ASR models can achieve reasonable performance through pre-training on additional data that is unlabeled, mismatched or both. Given such data constraints, we compare pre-training end-to-end models on matched but unlabeled data (unsupervised) and on labeled but mismatched data (supervised), where the labeled data is mismatched in either domain or language. Across encoder architectures, pre-training methods and languages, our experiments indicate that both types of pre-training improve performance, with relative WER reductions of 15-30% in the domain mismatch case and up to 15% in the language mismatch condition. We further find that the advantage from unsupervised pre-training is most prominent when there is no matched and labeled fine-tuning data, provided that a sufficient amount of mismatched data is still available for supervised fine-tuning.

Index Terms: speech recognition, cross-domain, cross-lingual, low-resource, pre-training, self-supervised learning, supervised training, unsupervised training

1. Introduction

The absence of large-scale matched and labeled training data can diminish speech recognition accuracy. However, other available data that is either unlabeled, mismatched (in acoustics, language, or application domain), or both, can be used to mitigate this gap. Pre-training is one way to use additional data of this sort. Here, pre-training refers to any training of a network before explicitly training it for its final task. We describe comprehensive experiments comparing the relative performance of unsupervised versus supervised pre-training approaches, given particular data constraints.

Recent results have shown that self-supervised training with a contrastive loss can learn meaningful representations [1]. Unsupervised pre-training of this form has improved speech recognition accuracy without much supervised data; examples include autoregressive predictive coding [2], wav2vec [3], and wav2vec 2.0 [4]. We evaluate some related approaches with supervised pre-training on mismatched data.

Domain, loosely referring to a logical group of utterances that share some common characteristics, serves as one dimension for data mismatch [5]. Examples of domain include application domains (such as voice-search and telephony), noise conditions, or other categorical groupings. Transfer learning [6, 7], adaptation of a background model through fine-tuning [8], factorized hidden layer [8], and generative adversarial networks [9, 10, 11] have previously been applied to handle domain mismatch.

Language mismatch can be considered a special case of domain mismatch. Cross-lingual and multilingual training have been explored to improve the performance of low-resource languages by learning common bottleneck features [12, 13, 14]. Shared hidden layers have also served as feature extractors, feeding separate classifiers [15] or softmax layers [16] in multilingual setups, potentially aided by adversarial training [17]. While this paper also builds on shared hidden layers, the underlying model is sequence-based. Existing work has also demonstrated the ability to train multilingual sequence-based models [18, 19, 20]. Unlike these examples, we use pre-training to bootstrap a monolingual model. Previous work comparing cross-lingual pre-training approaches reports phone accuracy [21]. In contrast, this work looks at the word error rate of RNN-T [22] models as well as considering other forms of domain mismatch.

This work contributes comprehensive, large-scale experiments that study both domain and language mismatch. It examines various pre-training strategies including recent advances in self-supervised training, when applied to state-of-the-art RNN-T LSTM and transformer architectures. The results are especially relevant because mismatched data is easy to obtain, particularly when expanding to different languages or new domains, even when matched data is limited.

The rest of the paper specifies the data sets (Sec. 2) and models (Sec. 3) used, and goes on to describe the pre-training methods under comparison (Sec. 4). Experiments and results are presented in Sec. 5. We conclude in Sec. 6.

2. Data configurations

We use large multi-domain (MD) data sets [5] in English ($\mathrm{MD_E}$) and German ($\mathrm{MD_G}$), both described in Tab. 1. The $\mathrm{MD_E}$ training set has 400k hours of data, including 200k semi-supervised hours from YouTube [23]. The $\mathrm{MD_G}$ training set has close to 50k hours, including 8k semi-supervised. We focus on a few sets as target domains. For English experiments, the target domain consists of medium-form ($\mathrm{MF_E}$) utterances, which have an average duration of 10.4 seconds. The $\mathrm{MD_E}$ training set includes 26k hours of $\mathrm{MF_E}$. For German experiments, the target domain is a mix of medium-form ($\mathrm{MF_G}$) and short-form ($\mathrm{SF_G}$, average duration 4.6 seconds) utterances. Together these are referred to as $\mathrm{Mix_G}$ and represent 35k hours of $\mathrm{MD_G}$. Target-domain utterances are anonymized and hand-transcribed. We assume constraints on the quantity of $\mathrm{MF_E}$ and $\mathrm{Mix_G}$ utterances, and simulate the lack of labeled data by arbitrarily sampling subsets of these sets.

For cross-domain experiments, we define an out-of-domain (OOD) set by removing the target-domain utterances from the full MD sets for supervised pre-training, and an additional limited in-domain (LinD) subset of the target-domain data for fine-tuning. For unsupervised pre-training, we use the full MD training set. For cross-language experiments, we use $\mathrm{MD_E}$ for pre-

http://dx.doi.org/10.21437/Interspeech.2021-654

Table 1: *Overview of training data sets.*

Language	Data set	Hours
English	Multi-domain ($\mathrm{MD_E}$)	400k
English	Medium-form ($\mathrm{MF_E}$), subset of $\mathrm{MD_E}$	26k
German	Multi-domain ($\mathrm{MD_G}$)	50k
German	Short-form ($\mathrm{SF_G}$), subset of $\mathrm{MD_G}$	21k
German	Medium-form ($\mathrm{MF_G}$), subset of $\mathrm{MD_G}$	13k
German	Mixed ($\mathrm{Mix_G}$) = $\mathrm{SF_G}$ + $\mathrm{MF_G}$	35k

training, and varying-size subsets of $\mathrm{Mix_G}$ for fine-tuning. Word error rates are reported on disjoint test sets sampled from the target domains for both English and German.

3. Models

We use an RNN-Transducer (RNN-T) [22, 24] as the ASR model, trained on all or a subset of the MD training data. The acoustic input consists of 128-dimensional log-Mel features [25]. The log-Mel filterbank energies are computed on 32-msec windows with a 10-msec hop. Features from 4 contiguous frames are stacked and subsampled by a factor of 3. For robustness, we employ multi-style training [26], SpecAugment [27], and arbitrary sample rate conversions to 8 kHz or 16 kHz. A ground-truth domain ID is appended to the acoustic input, with the option of being ignored depending on the model [28].

The encoder operates on the acoustic input and contains either 8 LSTM layers [28] or 15 transformer layers [29]. Each LSTM layer is unidirectional with 2048 units and a 640-unit projection layer [28]. Each transformer layer is also streaming and contains masked multi-head attention with relative positional encoding, with 16 attention heads of dimension 64 each, and 2 feed-forward dense layers of 2048 and 1024 units respectively [29]. The prediction network contains 2 LSTM layers and takes an embedding of the labels as input. A joint network with a single feed-forward layer of 640 units combines the outputs of the encoder and prediction network before a softmax loss. The label embedding, prediction network and joint network together comprise the RNN-T decoder. In total, the LSTM-encoder model has 120M parameters while the transformer-encoder model has 116M parameters.

4. Pre-training approaches

For supervised pre-training, we train the entire RNN-T on the OOD data, and transfer either the encoder or the entire pre-trained model for fine-tuning. For unsupervised pre-training, we implement: (a) a version of Wav2vec 2.0 [4], which we refer to as `Mel2vec` (Sec. 4.1), with a transformer model, or (b) autoregressive predictive coding [2] with either a transformer or LSTM encoder. With unsupervised pre-training, we initialize only the encoder from the pre-trained model.

4.1. Mel2vec

We pre-train the RNN-T encoder via Wav2vec 2.0 [4] pre-training. As the original work takes raw waveform as input, its architecture has a feature encoder transforming the raw waveform to a latent representation z. Our model does not have a feature encoder, because we use stacked log-Mel spectrograms as both the input [30] and the latent representation z. We refer to this as `Mel2vec`.

4.2. Autoregressive predictive coding (APC)

Autoregressive predictive coding (APC) was proposed in [2, 31] as a generative pre-training approach for speech representation learning. Different from a contrastive loss in Wav2vec2.0, APC is trained to encode the given speech utterances and predict future frames, where l_2 or l_1 loss is applied. In the paper, we use l_2 loss and add a total variation regularization between consecutive frames. The regularization weight is 0.1 in all experiments.

5. Experiments and results

A range of experiments across mismatch types, languages, encoder architectures and pre-training techniques are described. The experiment configurations are summarized in Tab. 2.

Table 2: *Experiment configurations. The data sets are described in Sec. 2. "Trans" refers to a transformer architecture.*

Mismatch type	Encoder type	OOD set	LinD set source	Table ref
Domain	LSTM	$\mathrm{MD_E}$ - $\mathrm{MF_E}$	$\mathrm{MF_E}$	3
Domain	Trans	$\mathrm{MD_E}$ - $\mathrm{MF_E}$	$\mathrm{MF_E}$	4
Domain	LSTM	$\mathrm{MD_G}$ - $\mathrm{Mix_G}$	$\mathrm{Mix_G}$	5
Domain	Trans	$\mathrm{MD_G}$ - $\mathrm{Mix_G}$	$\mathrm{Mix_G}$	6
Language	LSTM	$\mathrm{MD_E}$	$\mathrm{Mix_G}$	7
Language	Trans	$\mathrm{MD_E}$	$\mathrm{Mix_G}$	8

5.1. Domain mismatch

Assuming a limited quantity of LinD data, we examine the effects of pre-training with: (a) labeled but mismatched data (i.e. OOD only), and (b) unlabeled data that includes matched in-domain examples. Given previous results that multi-domain training performs at least as well as domain-specific training [5, 25], we focus on using LinD as a fine-tuning set rather than as a standalone training set. We include supervised baselines trained on OOD sets alone, which also serve as the supervised pre-trained models. Unsupervised pre-training is performed on the full MD set as matched examples may be more easily available without labels. Fine-tuning typically uses the LinD set combined with the labeled OOD set.

5.1.1. Cross-domain results in English

We first examine the results from supervised pre-training on OOD data. Tab. 3 shows the gain over an English LSTM OOD baseline as we fine-tune on increasing amounts of LinD data combined with OOD data. As the quantity of LinD data increases, the benefit from additional LinD data diminishes. Adding 260 hours of LinD data reduces the word error rate (WER) by almost 30% relative. Table 3 also shows the $\mathrm{MD_E}$ baseline WER, which continues to have a gap as it is trained with all $\mathrm{MD_E}$ utterances, including an accented $\mathrm{MF_E}$ subset that is not in the LinD set.

We also study side-by-side results for supervised and unsupervised pre-training approaches. Tab. 4 compares supervised and unsupervised pre-training methods for an English transformer model. All pre-trained models are fine-tuned on OOD + LinD, with varying quantities of LinD data. In the supervised case, we again observe the largest gain from a small amount of LinD fine-tuning data. The benefit of unsupervised pre-training approaches is highest when there is no matched fine-tuning data

Table 3: *Effect of fine-tuning quantities on English LSTM model with supervised OOD pre-training and fine-tuning on OOD + LinD, where LinD is a subset of MF_E.*

LinD fine-tune hours	**MF_E word error rate %**
0 (MD_E - MF_E)	5.8
260	4.1
1.3k	3.8
2.6k	3.7
13k	3.7
Multidomain (MD_E baseline)	3.2

Table 4: *Pre-training comparisons: English transformer, fine-tuning on OOD + LinD, where LinD is a subset of MF_E.*

LinD fine-tune	**MF_E word error rate %**		
hours	**Supervised**	**Mel2vec**	**APC**
0 (MD - MF_E)	5.3	4.6	4.6
512	4.1	4.1	4.1
2.5k	3.9	3.7	3.8
26k	3.3	3.3	3.3

(i.e. LinD quantity is 0). When we add LinD data, the gap between supervised and unsupervised pre-training is reduced or closed. Experimentally, we found it optimal to weight the LinD data relative to other sources by scaling the original MF_E weight by $(k/100)^{0.4}$ when using $k\%$ of MF_E as the LinD set.

5.1.2. Cross-domain results in German

We present similar comparisons across pre-training methods for the domain mismatch scenario in German. Tab. 5 shows results with an LSTM model, confirming that supervised pre-training on OOD followed by fine-tuning on LinD produces a lower WER than training on LinD alone. When fine-tuning on only the 3.5k-hour LinD set, supervised pre-training on OOD outperforms unsupervised pre-training on MD_G due to seeing more labeled examples. However, if fine-tuning also includes the OOD set, unsupervised pre-training yields better results.

Tab. 6 presents the results for supervised and unsupervised pre-training with the German transformer model. This displays a similar pattern to Tab. 5, where unsupervised pre-training followed by fine-tuning on OOD + LinD yields the best results. The table also confirms that given only labeled OOD data (no LinD), unsupervised pre-training still improves the performance on the target-domain test sets, as with English (Tab. 4).

Overall, the German results show a larger gain from unsupervised pre-training than in English, if we can fine-tune on OOD + LinD. This may reflect the smaller amount of total German data and the fact that both SF_G and MF_G are excluded from the supervised OOD set.

5.2. Language mismatch

In the language mismatch scenario, a limited quantity (3.5k hours) of supervised utterances sampled from Mix_G is available in the target language (German). In addition, we have access to: (a) a large labeled set in a different language (MD_E), or (b) a large unlabeled set in the target language (MD_G). We therefore consider fine-tuning on the target set after (i) supervised pre-training on the mismatched language, (ii) unsupervised pre-training on the mismatched language, or (iii) unsupervised pre-training on the matched language. We assume no more than the LinD set of labeled in-language data, hence all fine-tuning uses only the LinD data set. However, as the LinD set is identical to the one in Tab. 5 and Tab. 6, we can use those tables as a reference for what can be accomplished with more domain-mismatched but language-matched supervised data. Additionally, we report the numbers for supervised training on all available in-language and in-domain utterances (35k hours of Mix_G) as a measure of the gains achievable with more domain-matched and language-matched supervised data.

Table 5: *Pre-training comparisons: German LSTM. LinD is a 3.5k-hour subset of Mix_G.*

Fine-tuning	**Pre-training**	**Word error rate %**	
set		**MF_G**	**SF_G**
None	Supervised MD_G	9.2	12.8
None	Supervised OOD	21.7	20.8
LinD	None	14.3	18.5
LinD	Supervised OOD	12.1	15.4
LinD	APC MD_G	13.3	17.4
OOD + LinD	Supervised OOD	11.5	15.9
OOD + LinD	APC MD_G	**11.3**	**15.3**

Table 6: *Pre-training comparisons: German transformer. LinD is a 3.5k-hour subset of Mix_G.*

Fine-tuning	**Pre-training**	**Word error rate %**	
set		**MF_G**	**SF_G**
None	Supervised MD_G	9.9	13.1
None	Supervised OOD	18.8	19.5
LinD	None	15.0	18.8
LinD	Supervised OOD	12.5	16.0
LinD	Mel2vec MD_G	13.7	17.5
LinD	APC MD_G	13.3	17.4
OOD	Mel2vec MD_G	17.7	17.6
OOD	APC MD_G	16.8	17.6
OOD + LinD	Supervised OOD	12.5	15.4
OOD + LinD	Mel2vec MD_G	**10.4**	14.3
OOD + LinD	APC MD_G	10.6	**14.1**

5.2.1. Layers transferred

In most cross-language experiments, we initialize only the encoder from the pre-trained model before fine-tuning. For unsupervised pre-training, this is an artifact of the method itself. For supervised pre-training, as the pre-training is limited to language-mismatched data, intuitively the pre-trained encoder will capture more low-level language-independent features while the later layers, trained with label embeddings and targets from a language-specific symbol table, will be more language-dependent. Hence, we report numbers initializing only the encoder from supervised pre-training as well. However, we found that for the LSTM model, initializing the encoder and the prediction network from the supervised MD_E model resulted in the best final word error rate (Tab. 7). This suggests that the weights for the LSTM encoder and RNN-T decoder are tightly coupled, to the extent that initializing the

Table 7: *Cross-lingual pre-training comparisons with an LSTM model. The target language data is artificially limited to 3.5k hours (10%) of Mix*$_\text{G}$. *We also report numbers using the 100% Mix*$_\text{G}$.

		Fine-tuning hours	**3.5k (limited)**		**35k (full)**	
Pre-training set	**Pre-training type**	**Layers transferred**	**Word error rate (WERR)**		**Word error rate (WERR)**	
			MF$_\text{G}$	**SF$_\text{G}$**	**MF$_\text{G}$**	**SF$_\text{G}$**
None	None	None	14.3	18.5	9.4	12.9
Mismatched (MD$_\text{E}$)	Supervised	Enc	13.5 (6%)	17.7 (4%)	9.4 (0%)	12.9 (0%)
Mismatched (MD$_\text{E}$)	Supervised	Enc + Dec RNN	**12.1** (15%)	**16.3** (12%)	9.4 (0%)	12.9 (0%)
Matched (MD$_\text{G}$)	APC	Enc	13.3 (7%)	17.4 (6%)	9.4 (0%)	12.9 (0%)

Table 8: *Cross-lingual pre-training comparisons with a transformer model. The data configuration matches Tab. 7.*

	Fine-tuning hours	**3.5k (limited)**		**35k (full)**	
Pre-training set	**Pre-training type**	**Word error rate (WERR)**		**Word error rate (WERR)**	
		MF$_\text{G}$	**SF$_\text{G}$**	**MF$_\text{G}$**	**SF$_\text{G}$**
None	None	15.0	18.8	10.0	13.4
Mismatched (MD$_\text{E}$)	Supervised	**12.8** (15%)	**16.2** (14%)	9.5 (5%)	12.6 (6%)
Mismatched (MD$_\text{E}$)	APC	14.7 (2%)	19.0 (−1%)	9.4 (6%)	12.8 (4%)
Matched (MD$_\text{G}$)	APC	13.3 (11%)	17.4 (7%)	**9.2** (8%)	**12.4** (7%)

LSTM layers of the decoder from the mismatched language yields a better starting point than random initialization. In contrast, in the transformer model, initializing only the encoder yielded the best cross-language results.

5.2.2. LSTM results

Tab. 7 compares supervised pre-training on the mismatched language versus unsupervised pre-training on the matched language for LSTM models. With 3.5k hours of LinD data, both forms of pre-training yield improvements over no pre-training. When only the pre-trained encoder is transferred, unsupervised pre-training on the matched data (MD$_\text{G}$) slightly outperforms supervised pre-training on mismatched data (MD$_\text{E}$). However, initializing the encoder as well as the predictor network from the supervised mismatched pre-training yields the best performance. When more in-language fine-tuning data is available, pre-training does not make a difference.

5.2.3. Transformer results

Tab. 8 reports similar experiments on the transformer model. As with earlier results, with limited target-language data, any pre-training tends to help. If only language-mismatched data is available, supervised pre-training shows an advantage over unsupervised pre-training, even though only the pre-trained encoder is transferred. Unsupervised pre-training on language-matched data approaches the performance of supervised pre-training on language-mismatched data, but does not appear to close the gap with 3.5k hours of LinD data. This suggests that the ASR loss trains the encoder in cross-lingually applicable ways not captured by unsupervised pre-training, which can be recovered only when enough supervised fine-tuning data is available. The last row of Tab. 8 can be compared to the last row of Tab. 6, which shows a gain from adding more in-language fine-tuning examples even if they are out-of-domain. With the 35k-hour in-domain fine-tuning set, unsupervised pre-training is competitive with supervised pre-training, and unsupervised pre-training on language-matched data outperforms supervised pre-training on language-mismatched data.

6. Conclusion

Extensive experiments on pre-training end-to-end models with different data constraints, model architectures and languages suggest that both supervised and unsupervised pre-training are beneficial. Which method works better depends on the data constraints. Overall, unsupervised pre-training on matched data is competitive with or better than supervised pre-training on mismatched data, provided enough (even mismatched) fine-tuning data is available. In the language mismatch case, when the supervised fine-tuning data in a target language is very limited, supervised pre-training on a mismatched language yields better results. However, this changes if we obtain more in-language fine-tuning data, even from a mismatched domain.

These experiments can be further validated by studying a wider range of mismatched languages and considering smaller-scale data. A natural extension is to check whether the gains from these two forms of pre-training are complementary.

7. Acknowledgements

We thank Parisa Haghani, Hasim Sak, Manasa Prasad and Lillian Zhou for helpful discussion.

8. References

[1] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2019. [Online]. Available: https://arxiv.org/pdf/1807.03748.pdf

[2] Y. Chung and J. Glass, “Generative pre-training for speech with autoregressive predictive coding,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3497–3501.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec:

Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[5] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 441–447.

[6] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[7] 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. ASRU*, 2017.

[8] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition," in *Proc. Interspeech 2018*, 2018, pp. 892–896.

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 4.

[11] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator CycleGAN for unsupervised non-parallel speech domain adaptation," 2018. [Online]. Available: https://arxiv.org/pdf/1804.00522.pdf

[12] N. T. Vu, W. Breiter, F. Metze, and T. Schultz, "Initialization schemes for multilayer perceptron training and their impact on ASR performance using multilingual data," in *Proc. Interspeech 2012*, 2012, pp. 2586–2589.

[13] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 246–251.

[14] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6704–6708.

[15] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. II. IEEE, 2013, pp. 8619–8623.

[16] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7304–7308.

[17] K. Hu, H. Sak, and H. Liao, "Adversarial training for multilingual acoustic modeling," 2019. [Online]. Available: https://arxiv.org/pdf/1906.07093.pdf

[18] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4909–4913.

[19] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4904–4908.

[20] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," in *Proc. Interspeech 2019*, 2019, pp. 2130–2134.

[21] M. Rivière, A. Joulin, P. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7414–7418.

[22] A. Graves, "Sequence transduction with recurrent neural networks," in *International Conference on Machine Learning: Representation Learning Workshop*, 2012.

[23] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 368–373.

[24] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 193–199.

[25] A. Narayanan, R. Prabhavalkar, C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 920–927.

[26] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. Interspeech 2017*, 2017, pp. 379–383.

[27] 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," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[28] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S. Chang, W. Li, R. Alvarez, Z. Chen, C. Chiu, D. Garcia, A. Gruenstein, K. Hu, A. Kannan, Q. Liang, I. McGraw, C. Peyser, R. Prabhavalkar, G. Pundak, D. Rybach, Y. Shangguan, Y. Sheth, T. Strohman, M. Visontai, Y. Wu, Y. Zhang, and D. Zhao, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6059–6063.

[29] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7829–7833.

[30] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020. [Online]. Available: https://arxiv.org/pdf/2010.10504.pdf

[31] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," 2019. [Online]. Available: https://arxiv.org/pdf/1904.03240.pdf