Interspeech 2022
18-22 September 2022, Incheon, Korea

Adaptive multilingual speech recognition with pretrained models

Ngoc-Quan Pham[1] *Alex Waibel*[1,2] *Jan Niehues*[1]

[1]Interactive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Carnegie Mellon University, Pittsburgh PA, USA

ngoc.pham@kit.edu

Abstract

Multilingual speech recognition with supervised learning has achieved great results as reflected in recent research. With the development of pretraining methods on audio and text data, it is imperative to transfer the knowledge from unsupervised multilingual models to facilitate recognition, especially in many languages with limited data. Our work investigated the effectiveness of using two pretrained models for two modalities: wav2vec 2.0 for audio and MBART50 for text, together with the adaptive weight techniques to massively improve the recognition quality on the public datasets containing CommonVoice and Europarl. Overall, we noticed an 44% improvement over purely supervised learning, and more importantly, each technique provides a different reinforcement in different languages. We also explore other possibilities to potentially obtain the best model by slightly adding either depth or relative attention to the architecture.

Index Terms: speech recognition, multilingual, transformer, lstm, weight factorization, weight decomposition, pre-training, wav2vec, bart

1. Introduction

The sequence-to-sequence approach is widely used in speech recognition (SR) nowadays [1, 2, 3], and many research works are dedicated to show that their capabilities relying on a single architecture often match or are even better than traditional hybrid or CTC systems with separately optimized components [4, 3]. Moreover, this approach is also successfully realized in multilingual speech recognition, allowing one model to contain the information of many languages shared in one single body of neural networks [5].

Despite such promising potential, this approach inevitably requires abundant training data because it combines acoustic models, language models and alignment into the same model [6, 7]. In fact, the labeled data necessary for the task is limited, and even more limited outside of English and other mainstream languages, making building a reliable speech recognizer for many languages even more challenging. One can even question if sequence-to-sequence is a sustainable approach without fully utilizing data as well as the hybrid systems.

On the other hand, the availability of unlabeled data is virtually unlimited, and more importantly they can also be available under the form of pretrained models. Since the release of Bidirectional Transformers or BERT [8] with masked language model pretraining and its success in applying for natural language processing, many Transformers [9] based architectures have been the silver bullet to tackle low-resource problems. Recently, in the speech domain, Transformers are also well adopted in supervised learning [2, 10] and unsupervised pretraining with contrastive predicting methods [11, 12]. Most importantly, these models can be effortlessly trained on a multilingual dataset and acquire the acoustic or syntactic information of many languages. Replacing the components in an e2e model with the pretrained ones is trivial and can potentially gain largely for multilingual recognition.

With the pretrained models available in our arsenal, in this paper we set out to explore the possibilities of combining them for multilingual ASR. While the application wav2vec 2.0 and MBART50 individually is ubiqutuous and the combination of them has been observed in Speech Translation [13], to the best of our knowledge this is the first attempt in multilingual ASR, especially in a large scale with many languages with different data size.

More importantly, there are many improvements for the Transformers over the years that improve either modeling long range dependencies in self-attention [14, 10] or adaptive components for multilingual models [15, 16]. It is promising to also combine these techniques with the pretrained models for the best results.

With such motivation, we carried out experiments with 32 languages ranging from high to very low resource and explore the possibilities of using pretrained models. Our contribution is as follows:

- First, using multilingual pretrained acoustic and language models for the encoder and decoder respective brings a large improvement as a whole. However each pretrained module has a different influence on different languages, namely encoder pretraining is more impactful for languages with higher resources while the decoder counterpart is more effective for languages with medium-low resources. Surprisingly, many languages with extremely low resource (less than 5 hours) do not benefit much from this combination.

- Second, the language specific modulation techniques such as language adapters [15] and factorized adaptive weights [16] complement the two pretrained modules very well and have a strong impact on especially the low-resource languages mentioned above.

Moreover, there are different possibilities to integrate knowledge from pretrained models, and it is not necessary by simply replacing components. We also provided further analysis to the architecture, by showing that there are benefits to either improving the self-attention mechanism by adding relative positions during fine-tuning, or stacking the MBART encoders to the wav2vec counterpart.

This is the continuation of the line of work building multilingual ASR systems based on sequence-to-sequence neural networks [17, 5, 16]. Our work is available for public at *https://github.com/quanpn90/NMTGMinor* providing a highly CUDA-optimized implementation for both wav2vec and MBART which is potentially useful for the community.

10.21437/Interspeech.2022-872

2. Modeling

Transformers [9] are a class of sequence-to-sequence models with an encoder and a decoder with attention. Both components are equipped with self-attention layers being able to well handle long-short range dependencies in the input, output or alignment between sequences. These properties make this class of model appeal for Speech Recognition [2, 3].

2.1. Pretrained Acoustic representation

Transformer encoders can be used to learn useful representation from input masking and construction which is demonstrated in masked language model [8]. Following this trend, they are inevitably applied for audio signals, starting from learning to reconstruct the log-mel frequency features [18] to using quantization to learn latent variables [11].

In our work, the wav2vec 2.0 model [12] is selected to replace the typically randomly initialized encoder. It consists of three main components: a convolutional feature extractor that convolves and downsamples the raw audio input, a deep Transformer encoder that learn high level representation from the downsample sequence, and a quantization module to generate latent variables. During pre-training, the network optimizes for a contrastive loss function while masking the speech input. wav2vec 2.0 showed that it can outperform other semi-supervised learning approaches using finetuning with a CTC model.

2.2. Pretrained Multilingual BART

BART [19] was designed to learn syntactic features in languages using Transformers. The network is tasked to reconstruct a particular sentence at the decoder given a noisy version at the encoder side. MBART [20] and its extension MBART50 [21] took the BART training scheme and applied to 50 languages. This proved to be very useful for multilingual translation by finetuning the network on parallel data.

In speech recognition, data scarcity is even more problematic when the number of sentences fall short compared to other natural language tasks. The presence of MBART in the decoder is promising, especially for the self-attention parts which play the role of a language model.

2.3. Language Adaptive Components

The development of multilingual models for either machine translation, speech translation or speech recognition often concern between versatility versus specialization [22]. The motivation comes from the assumption that there are features being shared between languages and at the same time each language requires to selectively represented, and networks are encouraged to change "modes" depending on the language being processed [23]. Since then, multingual model designers opt to use specific network components being presented for each language, ranging from weight generator [24] to adapters [15, 22] or recently adaptive weights adding scales and biases to each weight matrix in the whole architecture [16]. In this paper, the last two options are selected for investigation thanks to being computationally manageable.

On the one hand, Adapters plugged into pretrained models were introduced in computer vision [25] and later natural language processing [26] and recently in Transformers for text/speech translation [15, 22]. They are materialized with a small multilayer perceptrons (MLP) with one hidden layer that acts as a *downsampler* (for parameter efficiency). This MLP is serialized at the end of each layer in the Transformer to help the network changes the feature distribution based on languages.

On the other hand, adaptive weights [16] was proposed based on the observation that neural networks evolve rapidly yet the core remains to be matrix multiplication. Therefore, it is possible to separate the weight matrix into a shared component W_S and language dependent *adaptive* scale W_{ML} and bias W_{BL}. The simple matrix multiplication $Y = WX$ becomes:

$$Y = (W_S \cdot W_{ML} + W_{BL})^\top X \tag{1}$$

In order to encourage the model to share parameters as well as keeping the parameters efficient, the adaptive weights are *factorized* by using the form of 1-rank matrices [27] which can be compactly represented as a dot-product between two vectors. This factorization can be established with k vectors per language so that there are k independent weight factors followed by a summation, which increases the rank of the additional weight matrices.

$$\bar{W} = \sum_{i}^{k} r_i s_i^\top \tag{2}$$

Equation 2 applies for all scale and bias matrices in the network.

Comparing two approaches, the advantage of the adapters is that they can increase the depth of representation in the network thanks for having an activation function in the downsampled layer. In contrast, the adaptive weights have the advantage to directly affect each layer function, such as the QKV-projection layer in self-attention, instead of applying a new function on the output of the layer. In the Transformers specifically this particular combination between a pretrained encoder and a decoder, the cross-attention layers in the decoder are left with weights untrained to connect two modalities audio and text. By being able to directly alter this function, the advantage of the adaptive weights are even more considerable.

2.4. Related Work

Using pretrained models is very effective for low-resource machine translation [20, 21] and recently speech translation [13] which can even handle zero translation from speech. In ASR, the presence of pretrained acoustic models [12, 28] allows recognition to be possible with very little data. The combination of pretrained acoustic and language models are recently investigated via learning to relax the modality mismatch [29, 30] yet still requires CTC and limited in a monolingual setup. In terms of multilingual ASR, there are various results in training a single model [5] that can overcome the monolingual performance. In this paper, we combined most prominent techniques to improve the results for many languages.

3. Experiments

There are 32 languages We report the error rates on the test sets of CommonVoice and Europarl. It is notable that both of the pretrained model (wav2vec) and the available supervised training data are mostly read speech, leading to the curiosity about the performance on a more spontaneous setting.

Our speech recognition experiments are conducted using the public dataset including CommonVoice [31] and Europarl [32] as training data.

For the progressive comparison, we trained a competitive supervised model using the Transformer large configuration [9]

with 24 encoder layers, 8 decoder layers and relative attention [10]. For transfer learning, we used the wav2vec 2.0 model pretrained with 53 languages [33] with the large configuration that has the same hidden size with our initial model. It is notable that the data used in pretraining is heavily biased to read speech including CommonVoice and Multilingual Librispeech [34]. For pretrained language models, MBART50 [21] with the same hidden size is used. For language specific modules, we use adapters with hidden layer size 512 and adaptive weights with $k = 8$ for bias and $k = 1$ for scale matrices, so that they have the same number of additional parameters per language. For training, we used an effective batch size of around 2.84 hours of data per update, together with a linear decay learning rate that peaks at 0.001[1]. The supervised model takes $150K$ updates to converge, while the model with transfer learning takes $50K$ updates.[2]

The performance impact of the pretrained modules are fully presented in Table 1. The test data here is the combination of CommonVoice and Europarl wherein the latter is available. Since the languages widely vary in terms of data size, we divide them into three groups that have less than 10 hours (very low), between 10-100 hours (low) and more than 100 hours (medium-large) of training data. Notably, our supervised model outperformed the previously reported error rates [16, 35].

3.1. Impact of acoustic pretraining

Compared to the Transformer model without any pretraining (**TF**), having the encoder pretrained with XLS-R (**W**) brings a substantial improvement to the average error rates by 18%, and this enables many languages to reach 10% errors or lower. Across the three groups however, there is a clear difference in impact. While the high-medium group enjoys a 20% decrease in error rate, this figure drops to 16% in the first group, and the third group is only improved by 4%.
While this is rather surprising, compared to the previously reported of wav2vec 2.0 pretrained models on very low resource settings [12], it is explainable by the difference of the approaches used in their and our works. The pretrained acoustic model has often been used directly with the CTC loss function to generate characters which heavily requires an external language model. In our setting, we are limited in both acoustic and text resources for the languages in the third group, and data scarcity makes learning to align from attention [7] even harder.

3.2. Effects from pretrained language model

With that observation, initializing the decoder with the MBART pretrained model is expected to alleviate the data scarcity problem. In fact, comparing the model with two pretrained modules (**WM**) with the previous one showed a benefit of 15% error reduction for the former.

Some languages have worse performance with the MBART decoder, such as Arabic, Turkish or Thai. This can be explained as a negative effect of the large byte-pair encoding [36] model shared between many languages originally used with MBART training. A large vocabulary size of $250K$ allow for large granularity which is far from characters or phonemes, the units that speech recognition models are often trained upon. Nevertheless, this is apparently not a problem for most languages.

Analyzing the individual performance of each group, the medium-large group is not impressively improved with just a 5% reduction. This result shows that having an additional six layers of decoders and even with pretrained weights only has a minimal effect on the result and network depth has a clearer impact on the source side than the target side in end-to-end speech recognition [2]. Probably the model does not struggle with the test data in terms of syntax, despite the fact that the text data here is not comparable to typical language models. Nevertheless, the second group receives a clearer merit from the pretrained language model, by a 27% improvement, and totally 39% improvement compared to the original multilingual Transformer. Even with a mismatched pretrained weights for the cross-attention module, it is still a noticeable improvement coming from the pretrained self-attention, feed-forward and layer normalization weights. The languages benefiting the most are Romanian, Estonian, Czech, Turkish, Indonesian, Swedish and Ukrainian.

The effect on the third group is modest at 11% reduction and the error rates remain very high for Urdu (ur), Kazakh (kk), Finnish (fi), Latvian (lv) or Vietnamese (vi). Many of the languages are also syntactically with many morphological word forms, such as Finnish, making recognition even more challenging. The most surprising improvement, however, comes from Slovenian that massively reduces from 26% to 14.6%. The overall struggling mainly comes from data scarcity which is not adequate for the decoder cross-attention layers.

3.3. Effect of the language-specific modules

As can be seen from Table 1, both techniques are able to help the model generalize better in all language groups. Most importantly, the very low resource group witnesses 27% and 26% improvement using adaptive weights and adapters respectively, compared to the baseline model with two pretrained modules.

In order to quantify their impact, we proceeded to freeze all of the pretrained parameters and only fine-tuned the language-specific parameters. There is a difference in how each technique handles this situation. With the presence of only adapters, the errors in all languages deteriorate rapidly in all language groups. Many languages in the low-group experience very bad results including Romanian, Arabic, Chinese, Lithuanian and especially many members within the very-low group exceed 90% error rates. While this is unexpected, we can see that the main problem here is the cross-attention layer which is not familiar with the inputs coming from two modalities. The adapters are not able to drive the bad context vectors (weighted-sum of the encoder inputs) into meaningful representation in the low resource condition.

The adaptive factorized weights do not have this problem because they directly alter cross-attention. As a result, the performance is much better than the adapters even though they still fall behind the baseline without language-specific modules.

3.4. Further analysis

In the previous sections, we presented the most important enhancement for our multilingual setup coming from the pretrained modules and the language adaptive components. The success of using the language adaptive components in various places, either breaking the *layer dynamics* with adapters or the *function dynamics* with the adaptive weights suggests that further improvement can be found by adding more information to the system instead of treating the pretrained model as an immovable black box.

Here we follow the implementation in [10] to add rela-

[1]The decaying equation follows the same in Attention is all you need [9]

[2]Training was possible using 4 NVIDIA A100 GPUs.

Table 1: *Performance(WER↓) on the CommonVoice-Europarl dataset. Models include Transformers (**TF**), with wav2vec pre-training (**W**), with wav2vec and MBART50 (**WM**), with adapters (**WMA**) and factorized weights (**WMF**) and the version w/ frozen pt. weights.*

Language	Hours	**TF**	**W**	**WM**	**WMA**	**WMF**	**FWMA**	**FWMF**
(de)	1050	10.4	8.1	7.8	7.7	**7.2**	11.8	10.0
(nl)	150	13.2	8.4	7.7	7.4	**6.8**	10.2	9.0
(fr)	800	15.2	12.7	12.12	11.62	**11.2**	16.7	15.1
(it)	325	11.5	8.7	8	7.8	**6.5**	12.8	10.5
(fa)	293	5.5	4.8	3.9	4.2	**4.0**	6.4	7.2
(pl)	145	11.5	10.5	9.2	9.1	**7.6**	14.8	12.2
(pt)	120	14.0	10.9	10.0	9.5	**6.0**	18.5	12.8
(es)	400	10.9	8.0	7.6	7.4	**6.2**	11.2	9.8
(ru)	148	10.0	8.7	5.5	5.7	**5.4**	20.1	10.1
(ta)	198	28.6	24	20.2	**20.7**	21.0	31.4	31.5
(th)	133	2.8	2.6	3.3	3.4	**3.2**	5.1	4.5
Average		12.1	10.3	9.5	9.3	**8.6**	15.8	13
(ro)	45	18.1	21.2	15.8	13.7	**10.12**	42.0	15.7
(ar)	85	21.1	**15.8**	18.7	17.6	18.2	31.2	23.2
(et)	32	30.4	22.1	14.8	15.1	**13.2**	32.5	21.5
(ja)	26	13.0	10.3	8.3	7.91	**7.9**	21.8	11.5
(zh)	63	25.9	16.7	15.2	**14.6**	14.7	37.6	18.2
(cs)	49	19.8	15.8	10.0	9.2	**8.4**	16.4	12.7
(lt)	16	43.3	37.9	31.9	26.7	**25.5**	71.3	30.0
(tr)	30	10.4	13.6	7.5	8.4	**7.5**	9.8	10.1
(id)	23	14.0	13.6	7.5	7.5	**6.6**	9	8.7
(mn)	12	49.8	35.1	26.2	26.0	**24.3**	90.4	32.0
(sv)	35	24.7	20.6	14.3	13.0	**12.3**	17.5	16.3
(uk)	56	14	13.4	7.6	8.3	**7.4**	11.4	14.8
Average		23.7	20	14.5	13.7	**12.5**	32.7	17.4
(lv)	6	41.9	57.0	41.01	**22.3**	**22.3**	79.1	25.0
(vi)	3	49.5	53.5	46.1	35.6	**34.0**	104.1	38.3
(ka)	6	58.6	48.0	48.3	33.0	**32.09**	130.3	39.0
(sl)	9	20.5	26.5	14.6	10.4	**9.1**	10.5	12.8
(fi)	6	54.2	48.5	41.0	31.4	**30.0**	109.2	35.2
(hi)	8	46.1	46.4	36.4	28.6	**27.6**	99.5	31.1
(gl)	7	26.5	15.3	15.3	11.2	**9.8**	48.8	16.0
(ur)	0.6	78.0	68.2	62.3	**56.6**	57.0	104.1	70.0
(kk)	0.73	86.5	76.8	86.4	**60.6**	61.0	104.3	70.03
Average		51.3	48.9	43.5	32.2	**31.4**	87.8	37.5
Overall		30	24.8	20.9	17.4	**16.6**	41.6	21.6

tive positions into the wav2vec model, by adding the content-position interaction together with the content bias and position bias to self-attention to all self-attention layers of wav2vec with factorized weights. The additive information allowed for the reduction of the average error rate to 16.04% (3% improvement) and especially an 12.7% on the large-medium group. This evidently helps the model learns better with adequate data.

Surprisingly, an alternative attempt to enhance the encoder simply by stacking the MBART50 encoder on top of the wav2vec encoder (without any length conversion) yields similar results, compared to the baseline **WM**, it improves the large-medium group by an impressive 18.3%, yet only 3.2% overall. Not only does stacking the encoder increase the encoder depth, it also helps the cross-attention layers because they are familiar with the output of the text encoder during training. Training a stacked model with the adaptive techniques would probably result in the best model in our experiments, however it was unfortunately beyond our computational tolerance.

4. Conclusion

In this work, we massively improved multilingual ASR using a combination of three techniques: encoder pretraining, decoder pretraining and adaptive weights. Our empirical results indicate that each technique has a different impact on different languages varying in linguistic characteristic and data size. Nevertheless, at the very low condition the model still struggles to reach an acceptable error rate. Our analysis in using adaptive weights shed light on the future work, in which multimodal pretraining [37] is potentially beneficial to address the cross-attention layers.

5. Acknowledgements

The projects on which this paper is based were funded by the Federal Ministry of Education and Research (BMBF) of Germany under the numbers 01IS18040A and 01EF1803B.

6. References

[1] 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*, 2019.

[2] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, and A. Waibel, "Very Deep Self-Attention Networks for End-to-End Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 66–70.

[3] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[4] T.-S. Nguyen, S. Stueker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," *arXiv preprint arXiv:1910.13296*, 2019.

[5] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.

[6] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[10] N.-Q. Pham, T.-L. Ha, T.-N. Nguyen, T.-S. Nguyen, E. Salesky, S. Stüker, J. Niehues, and A. Waibel, "Relative Positional Encoding for Speech Recognition and Direct Translation," in *Proc. Interspeech 2020*.

[11] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[12] 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.

[13] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[14] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[15] A. Bapna, N. Arivazhagan, and O. Firat, "Simple, scalable adaptation for neural machine translation," *arXiv preprint arXiv:1909.08478*, 2019.

[16] N.-Q. Pham, T.-N. Nguyen, S. Stueker, and A. Waibel, "Efficient weight factorization for multilingual speech recognition," *arXiv preprint arXiv:2105.03010*, 2021.

[17] 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," *arXiv preprint arXiv:1909.05330*, 2019.

[18] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving transformer-based speech recognition using unsupervised pre-training," *arXiv preprint arXiv:1910.09932*, 2019.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[21] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv preprint arXiv:2008.00401*, 2020.

[22] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Lightweight adapter tuning for multilingual speech translation," *arXiv preprint arXiv:2106.01463*, 2021.

[23] J. B. Hampshire II and A. Waibel, "The meta-pi network: Building distributed knowledge representations for robust multisource pattern recognition," *IEEE Computer Architecture Letters*, 1992.

[24] E. A. Platanios, M. Sachan, G. Neubig, and T. Mitchell, "Contextual parameter generation for universal neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[25] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[26] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[27] Y. Wen, D. Tran, and J. Ba, "Batchensemble: an alternative approach to efficient ensemble and lifelong learning," *arXiv preprint*, 2020.

[28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[29] G. Zheng, Y. Xiao, K. Gong, P. Zhou, X. Liang, and L. Lin, "Wav-bert: Cooperative acoustic and linguistic representation learning for low-resource speech recognition," *arXiv preprint arXiv:2109.09161*, 2021.

[30] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9161–9168.

[31] 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.

[32] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP*, 2020.

[33] 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.

[34] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[35] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki, "Exploiting adapters for cross-lingual low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2021.

[36] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[37] J. Ao, R. Wang, L. Zhou, S. Liu, S. Ren, Y. Wu, T. Ko, Q. Li, Y. Zhang, Z. Wei *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," *arXiv preprint arXiv:2110.07205*, 2021.