Replay to Remember: Continual Layer-Specific Fine-Tuning for German Speech Recognition

Theresa Pekarek Rosin[(✉)] and Stefan Wermter

Knowledge Technology, Department of Informatics, University of Hamburg,
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
{theresa.pekarek-rosin,stefan.wermter}@uni-hamburg.de
http://www.knowledge-technology.info

Abstract. While Automatic Speech Recognition (ASR) models have shown significant advances with the introduction of unsupervised or self-supervised training techniques, these improvements are still only limited to a subsection of languages and speakers. Transfer learning enables the adaptation of large-scale multilingual models to not only low-resource languages but also to more specific speaker groups. However, fine-tuning on data from new domains is usually accompanied by a decrease in performance on the original domain. Therefore, in our experiments, we examine how well the performance of large-scale ASR models can be approximated for smaller domains, with our own dataset of German Senior Voice Commands (SVC-de), and how much of the general speech recognition performance can be preserved by selectively freezing parts of the model during training. To further increase the robustness of the ASR model to vocabulary and speakers outside of the fine-tuned domain, we apply Experience Replay [20] for continual learning. By adding only a fraction of data from the original domain, we are able to reach Word-Error-Rates (WERs) below 5% on the new domain, while stabilizing performance for general speech recognition at acceptable WERs.

Keywords: Automatic Speech Recognition · Transfer Learning · Continual Learning · Domain Adaptation

1 Introduction

Automatic Speech Recognition (ASR) models have previously reached unseen state-of-the-art performance after the introduction of unsupervised and self-supervised pre-training methods from raw audio data, which allowed models to utilize a larger amount of speech data for training [3,5]. However, this has been accompanied by state-of-the-art models increasing in size and requiring thousands of hours of speech data to be trained properly. A recent example is Whisper [19] which, in its largest release, contains 1550 M parameters and is trained on 680,000 h of multilingual speech.

Fortunately, it is not necessary to train such a model from scratch for different languages and domains. Multilingual models like Whisper, or XLSR-53 [5] and
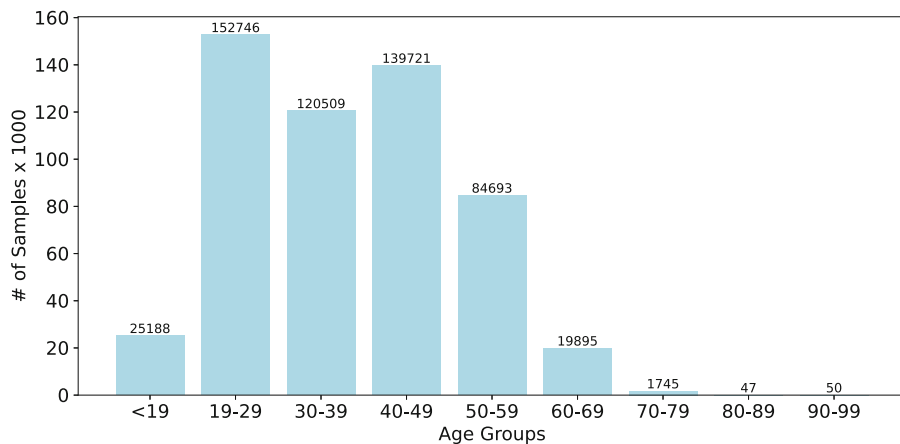
Fig. 1. The age distribution for Common Voice DE 10.0 [1]. As can be seen, of the labeled samples (ca. 70%), the majority are between 19 and 59 years old. Older adults only constitute a fraction of the available samples.

its successor XLS-R [2], generally perform better on low-resource languages than monolingual models trained from scratch, since similarities between languages can be leveraged. However, there is still improvement to be gained by fine-tuning for a specific language. For example, we observe a Word-Error-Rate (WER) of 15.2% for Whisper-small [19] on Common Voice German 10.0 (CV-de) [1] without any adaptation, still, through fine-tuning on additional hours of German speech this can be improved to 11.2% [10].

However, often a more specific adaptation for sub-groups or speakers is necessary due to the fact that performance usually is much lower for speech that differs from the norm, e.g. due to accents, age, or speech disorders [16,17]. This is due to the demographic distribution in most available datasets, where the majority of speakers are male, white, and middle-aged [17]. As can be seen in Fig. 1, this issue transcends languages, as older age groups are similarly underrepresented in CV-de [1], the most commonly used resource to train German speech recognition models. The same problem exists for the distribution of gender: of the subset labeled with additional demographic information in CV-de (ca. 70%), female and diverse speakers only constitute 14%.

To address this problem and thereby create more reliable ASR models, we can facilitate the knowledge contained in large-scale models, similar to how multilingual models can be utilized to improve ASR for low-resource languages. However, End-to-End ASR models also suffer from catastrophic forgetting [18], even for within-language adaptation, which usually destroys the performance of general speech recognition [23]. Therefore, a careful combination of transfer learning, i.e. leveraging the information contained in pre-trained models to facilitate learning on new domains, and continual learning, i.e. preventing the deterioration of performance on previously learned domains, is required.

We collect a dataset of German Senior Voice Commands (SVC-de) and compare the performance of Whisper [19], XLSR-53 [5], and XLS-R [2], three state-of-the-art multilingual speech recognition models. We follow research for layer-specific fine-tuning [12,21] and examine how unfreezing different layer configurations influences the performance of the ASR model. Since domain adaptation usually leads to a decrease in performance on the original domain, we utilize Experience Replay (ER) [20] for continual learning to lessen the drop in performance for general speech recognition, and thereby increase the ASR model's robustness to out-of-domain vocabulary and speakers.

2 Related Work

2.1 Multilingual Speech Recognition

The availability of pre-trained multilingual models in ASR has enabled transfer learning approaches for domains with limited data. This has been especially beneficial for improving speech recognition for non-standard speech and low-resource languages.

XLSR-53 [5] and its successor XLS-R [2] are based on the wav2vec 2.0 [3] architecture and offer large-scale cross-lingual speech recognition. Pre-training in multiple languages, 53 for XLSR-53 and 128 for XLS-R, improves speech recognition across different languages since similarities between them are exploited during training. Whisper [19] is a recent large-scale multilingual model, trained in an unsupervised manner for zero-shot cross-lingual speech recognition, speech translation, and language identification across 97 different languages with 680 000 h of speech data. The underlying architecture is a simple encoder-decoder transformer [24].

The results presented alongside these models show that multilingual ASR models usually perform better than monolingual models on low-resource languages. However, for languages where a large number of transcribed speech data is available, these models are outperformed by models utilizing supervised training [2,19]. This shows that it is beneficial to combine unsupervised pre-training with language- or domain-specific supervised fine-tuning.

2.2 Layer-Specific Fine-Tuning

While the transfer learning capabilities of large-scale speech recognition models have been demonstrated for multilingual [2,5,19] as well as monolingual adaptations [14,16], the question remains if it is necessary to adapt the entire model during the fine-tuning process, especially for very specific or smaller domains.

Shor et al. [21] fine-tune different layer combinations in Listen, Attend, and Spell (LAS) models [4] and RNN-T models [6] to find the subset of layers encoding the most information. For the LAS model, the best results are achieved

through fine-tuning the entire model, but for the RNN-T model, 91% of relative WER improvement is achieved by only fine-tuning the joint layer and the first layer of the encoder.

Similarly, Huang et al. [12] look at the influence of different layer configurations on the performance of a Conformer-Transducer [11] model in the context of efficient speaker adaptation. They observe that adaptation of the mid and bottom layers of the Conformer [9] encoder offers a slight decrease in WER over adaptation of the top layers.

Shrivasta et al. [22] examine how much model performance depends on trained weights in the encoder and decoder of RNN-T [6] and Conformer models [9]. They randomly initialize different parts of the model and confirm that, while randomly initializing the encoder immediately hurts model performance, there is no significant difference in results for randomly initializing the decoder.

While research on layer-specific fine-tuning has been mainly focused on performance approximations for new domains, the loss of performance for general speech recognition in monolingual layer-specific adaptations has not been examined in detail. The occurrence of catastrophic forgetting might be greatly dependent on the number of updated parameters, while the performance of attention-based models on the fine-tuned domain might not be. Therefore, we examine layer-specific fine-tuning for both domains, to see how much knowledge is gained for the new domain and lost for out-of-domain speech recognition in each configuration.

2.3   Experience Replay

Experience Replay (ER) [20] is a rehearsal-based continual learning (CL) method that aims to counteract catastrophic forgetting [18] by including a small fraction of data from the original domain in the training data for the new domain. While CL for speech recognition is still relatively unexplored, ER has been utilized successfully for monolingual Dutch accent adaptation before [23]. One advantage of rehearsal-based CL methods is that as long as data from the original domain is available or can be generated, the approaches can be used in a model-agnostic fashion.

3   Experiments

3.1   Data

We fine-tune the models on the German Senior Voice Commands (SVC-de) dataset, a dataset we collected for the development of an ASR system for German senior citizens in the context of a home assistant system. The data has been collected with the approval of the Ethics Commission at the University of Hamburg.

SVC-de consists of short speech commands recorded by German speakers between the ages of 50 and 99. Overall 30 people (21 female, 9 male) recorded 52

Table 1. An overview of the WER (%) of our baseline models (-*de* indicates fine-tuning on CV-de), evaluated on the test-split of Common Voice DE 10.0 [1] and our own German Senior Voice Commands (SVC-de) dataset.

Model	CV-de 10.0 test	SVC-de	# of Params
XLSR-53-large-de [7]	12.8	18.4	315 M
XLS-R-1B-de [8]	11.6	20.0	962 M
XLS-R-300m-de [15]	22.8	31.3	315 M
Whisper-base-de	20.4	25.7	74 M
Whisper-small-de [10]	11.2	18.4	244 M

sentences each with two microphones, for a total of 3 h 9 m, with approximately 6–7 min of audio data per speaker. The recorded sentences were manually cut and transcribed afterward to give a realistic estimation of the examined ASR models' performance. We use 70% of the dataset for training, 10% for validation, and the remaining 20% for testing.

Common Voice DE (CV-de) [1] is one of the largest and most utilized German speech datasets, and features a large variety of recording conditions and speakers due to the crowd-sourced nature of the collection. We utilize CV-de 10.0, which has 1136 validated hours of audio from 16,944 different speakers and contains additional demographic data (e.g. age group, gender, accent) for about 70% of the samples. We use the predefined dataset splits for training and testing.

3.2 Base Models

As can be seen in Table 1, the performance for German speech varies even for large-scale ASR models. While fine-tuning on data like CV-de improves the average performance for German speech, the improvement does not immediately translate to elderly speech. Additionally, a higher number of parameters does not seem to automatically lead to a better performance. The performance of XLS-R-1B-de [2], a model with approximately 1 B parameters is comparable to its predecessor XLSR-53-large [5] and to Whisper-small [19], with only 244 M parameters, after fine-tuning on CV-de.

In our experiments, we utilize a selection of pre-trained models from the publicly available checkpoints in Huggingface's[1] model repository. All models are approximately the same size and have been adapted to German speech with CV-de. We include a pre-trained version of XLS-R, with 300 M parameters [15], a pre-trained XLSR-53-large model [7], and a pre-trained Whisper-small model [10]. XLSR-53-large and XLS-R-300M both consist of 24 encoder layers and use character-based tokenization. Whisper-small consists of 12 encoder- and 12 decoder-layers and utilizes a byte-level BPE text tokenizer for an output vocabulary size of 51,865. All models include punctuation to some degree, but

---

[1] https://huggingface.co/.

to enable a fair comparison, we normalize the generated transcripts before the evaluation.

3.3   Experiments

In all our experiments, unless specifically stated otherwise, we train our models for five epochs with a batch size of 128 and AdamW [13] optimizer. The learning rate is set to 3e-4 for XLS-R and XLSR-53, and to 3e-5 for Whisper. It decays linearly after a warm-up of 50 steps. We set the dropout for XLSR-53 and XLS-R to 0.1, and use mean CTC loss reduction. All hyperparameters were determined empirically by comparing the behavior of the models during the layer-specific fine-tuning experiments. We train our models on an NVIDIA A100 80G graphics card.

Transfer Learning. The transfer learning capabilities of large pre-trained speech recognition models have been proven for adaptation to non-standard speech before [14,21]. Therefore, we establish a baseline by fine-tuning the entirety of our selected models on the SVC-de dataset. For XLSR-53 and XLS-R, we keep the feature extractor frozen.

Then, following similar approaches [12,21,22], we fine-tune different layer combinations to determine the most efficient subset for the adaptation of the model. Table 2 shows the layer configurations for our baseline models. Since XLSR-53-large [5] and XLS-R-300m [2] share a network structure of 24 encoder layers, we can apply the same configurations to both models. While Whisper contains the same number of layers, the network structure is split into 12 encoder and 12 decoder layers. Therefore, we examine the layer configurations for the encoder and decoder in separate experiments and then apply them to both parts of the model simultaneously. For example, in the encoder-decoder fine-tuning scenario, we would apply the 'first 6' configuration to both the encoder and the decoder, leading to a total of 12 adaptable layers.

We examine how much the performance differs between layer configurations for SVC-de and how much the performance for CV-de degrades due to domain adaptation. This should serve as an indicator as to which parts of the model are essential for the creation of general speech representations and therefore more sensitive to change, and which parts can be adapted for another domain without affecting the performance of the original dataset too drastically.

Continual Learning. To reduce the loss of knowledge regarding general speech recognition, we implement Experience Replay (ER) [20] for continual learning. However, instead of including a fixed number of samples from the original domain in each batch, we include either 10% or 20% of the original domain in the SVC-de training data spread out over all batches. We examine these data splits for the models with the best layer configurations, regarding their WER reduction on CV-de and their WER and convergence on SVC-de. We compare the performance between our best models with and without ER for both datasets.

Table 2. The fine-tuning layer configurations for our baseline models. XLSR-53-large [5] and XLS-R-300m [2] share the same number of encoder layers and therefore we can apply the layer configurations to both models. Due to the encoder-decoder architecture of Whisper [19], we apply these configurations first to the 12 layers of the encoder and the 12 layers of the decoder separately, and then to both simultaneously.

XLS-R & XLSR-53		Whisper	
Name	Layer Configuration	Name	Layer Configuration
first 12	[0, 1, ..., 10, 11]	first 6	[0, 1, 2, 3, 4, 5]
last 12	[12, 13, ..., 22, 23]	last 6	[6, 7, 8, 9, 10, 11]
f4-i4-l4	[0, ..., 3, 10, ..., 13, 20, ..., 23]	f1-i2-l1	[0, 5, 6, 11]
f2-i2-l2	[0, 1, 11, 12, 22, 23]	f2-i2-l2	[0, 1, 5, 6, 10, 11]
last 6	[18, 19, 20, 21, 22, 23]	last 3	[9, 10, 11]

4   Results and Discussion

4.1   Layer-Specific Fine-Tuning

As can be seen in Fig. 2, fine-tuning the entire model generally leads to the best performance for all examined models. This aligns with the observations by Shor et al. [21] in their experiments with LAS. However, Whisper shows a clear difference in performance between layer configurations that adapt only the layers of the encoder or the decoder. Fine-tuning only the encoder layers leads to a final average WER of 15.5%, which is an average increase of 13.3% compared to the final WER obtained by fine-tuning the entire model (2.2%). The encoder-decoder layer configurations reach an average WER of 4.8% and the decoder configurations follow close behind with a final average WER of 6.2% after five epochs.

The closest approximation of fine-tuning the entire Whisper model is obtained by fine-tuning the last six layers of the encoder and the decoder at the same time (WER: 3.1%), followed closely by fine-tuning only the decoder (WER: 3.5%). For XLSR-53, adapting only the first 12 layers (WER: 6.6%) or configuration 'f4-i4-l4' (WER: 7.1%) offers a close approximation of the best model performance (WER: 5.5%). Meanwhile, XLS-R shows the largest gap in WER between fine-tuning the entire model (WER: 7.4%) and the next best 'f4-i4-l4' configuration (WER: 12.2%), but also the largest improvement on SVC-de compared to its performance before the adaptation (Table 1). However, Whisper outperforms both XLS-R and XLSR-53 on average after five epochs of training, despite an initial spike in WER on SVC-de.

As expected, the performance of CV-de deteriorates as a result of the fine-tuning process. Figure 3 shows a drastic increase in WER for all layer configurations and all examined models. However, the most forgetting occurs when the entire model is trained, and fine-tuning only a reduced number of layers generally leads to a lower WER for CV-de. This is especially interesting for cases, where adapting a smaller selection of layers is a close approximation of the original
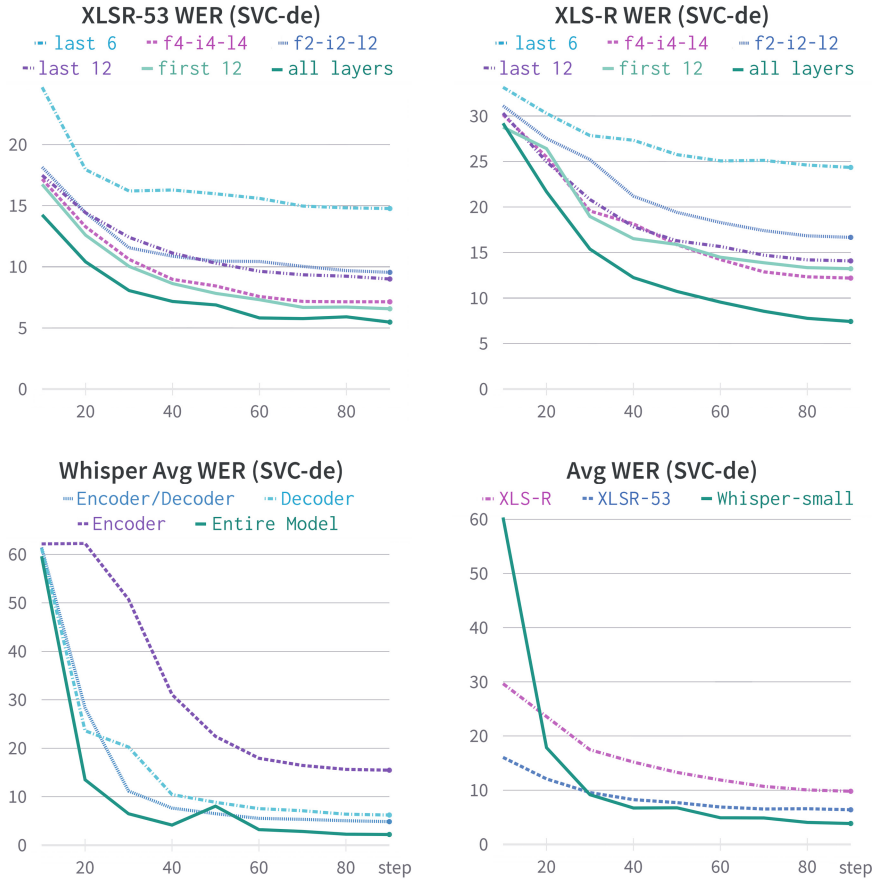
Fig. 2. The results of the layer-specific fine-tuning on SVC-de. For all models, the largest increase in performance can be observed after fine-tuning the entire model. However, for Whisper-small this performance can be approximated by layer configurations that only adapt the decoder or both model parts in unison. While XLSR-53 also offers a close approximation for some layer configurations, this is not the case with XLS-R. On average, Whisper's best layer configurations outperform both XLSR-53 and XLS-R after five epochs of training.

model performance. For example, fine-tuning only the last 6 layers of Whisper's encoder and decoder achieves a similar WER on SVC-de as adapting the entire model, with a difference of only 0.9%. The WER on CV-de, however, is approximately 5% lower for the smaller selection (24.5%) compared to the entire model (29.1%), which indicates that adapting only a smaller layer configuration is beneficial for preserving the performance of the original domain.

For XLS-R and XLSR-53 the behavior is similar, as most forgetting occurs when the entire model is fine-tuned. But, compared to Whisper, the WER on CV-de does not show any major changes after the first 10 optimization steps and is
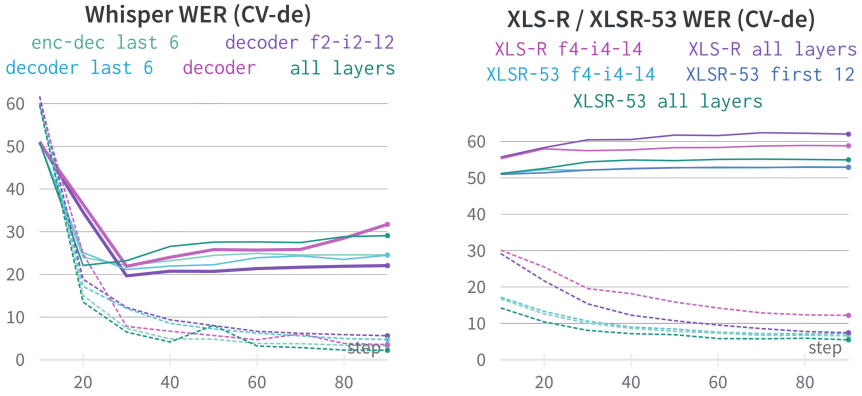
Fig. 3. The performance decay of CV-de during fine-tuning on SVC-de, measured in WER (lower is better, dashed lines indicate corresponding results on SVC-de). The most forgetting occurs for Whisper if the entire model is fine-tuned or all decoder layers are fine-tuned. For XLS-R and XLSR-53 the largest decay of performance happens (1) within the first 10 optimization steps and (2) when the entire model is fine-tuned on SVC-de.

generally much higher. This is due to the selected learning rate. While a learning rate of 3e-3 leads to a better performance on SVC-de, the decay on CV-de is even more drastic and the learning process is overall more unstable. In comparison, a learning rate of 3e-4 offered the best trade-off between performance on the new and the old domain.

4.2   Experience Replay

After applying ER during the fine-tuning process, we observe that ER with as little as 10% of original data not only helps to stabilize Whisper's training on SVC-de but also diminishes the performance decrease on CV-de. As can be seen in Table 3, fine-tuning only the last 6 layers of the encoder and the decoder leads to our best performance, with a final WER of 18.1% on CV-de and 3.0% on SVC-de, after five epochs of training. This is closely followed by adapting only the last 6 layers of the decoder with ER on 20% of CV-de. XLS-R and XLSR-53 also experience a WER reduction from ER, even though they do not reach the same level of performance as Whisper.

Table 3. A comparison of our best models with and without Experience Replay (ER). While all models benefit from ER, a trade-off can be observed if we increase the percentage of samples from CV-de. Of all examined models, Whisper is the only one that can be stabilized at an acceptable WER for CV-de, while showing vast improvements for SVC-de.

Model	Layer Config	ER (%)	SVC-de WER (%)	CV-de 10.0 WER (%)
XLSR-53	all layers	–	**5.5**	55.0
XLSR-53	first 12	–	6.6	52.9
XLSR-53	all layers	10	**5.6**	42.7
XLSR-53	all layers	20	5.9	**31.6**
XLS-R	all layers	–	**7.4**	62.0
XLS-R	all layers	10	8.1	53.5
XLS-R	all layers	20	8.1	**47.3**
Whisper-small	all layers	–	**2.2**	29.1
Whisper-small	enc/dec last 6	–	3.2	24.5
Whisper-small	enc/dec last 6	10	**3.0**	18.1
Whisper-small	dec last 6	10	5.0	17.6
Whisper-small	dec last 6	20	4.9	**16.8**

5   Conclusion and Future Work

In this work, we demonstrate the effectiveness of combining layer-specific fine-tuning and continual learning to improve performance for under-represented speaker groups, while keeping the performance for general speech recognition from deteriorating in the process. Adapting smaller layer sub-groups for specific domains can, depending on the choice of model and configuration, approximate the performance of a model that has been fine-tuned in its entirety. Additionally, since fewer parameters are adapted during training, the performance decay on the original domain is decreased. We show that utilizing Experience Replay (ER) [20] with only a small fraction of data from the original domain can lead to vast improvements in WER for the original, as well as minor improvements for the new domain.

Our best model is a pre-trained German Whisper-small architecture [10, 19], fine-tuned on SVC-de with 10% ER, which reduces the WER for SVC-de from 18.4% to 3.0%. By adapting only the last six layers of the encoder and the decoder, we are able to stabilize the performance of CV-de at 18.1% WER. By adding more data from the original domain, the WER on the original domain can be lowered further. However, we observe that at 20% ER a trade-off starts to happen, where the performance on CV-de can only be improved with detriment to the performance of the new domain.

While we utilize our own novel dataset of elderly German speech (SVC-de) in our experiments, the methods we use are model- and dataset-independent, which

indicates that our approach could be applied to other domains (e.g. dialects) as well. Additionally, since the vocabulary in SVC-de is limited, our approach promises more robustness for out-of-domain words and a larger variety of speakers than traditional fine-tuning approaches.

References

1. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (2020)
2. Babu, A., et al.: XLS-R: self-supervised Cross-lingual Speech Representation Learning at Scale. In: Proceedings of INTERSPEECH 2022, pp. 2278–2282. ISCA, Incheon, Korea (2022)
3. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Curran Associates Inc., Vancouver, BC, Canada (2020)
4. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: LIsten, attend and spell: a neural network for large vocabulary conversational speech recognition. In: Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964. IEEE Press, Shanghai, China (2016)
5. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. In: Proceedings of INTERSPEECH 2021, pp. 2426–2430. ISCA, Brno, Czechia (2021)
6. Graves, A.: Sequence transduction with recurrent neural networks. In: ICML 2012 Workshop on Representation Learning (2012)
7. Grosman, J.: Fine-tuned XLSR-53 Large model for speech recognition in German. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german (2021)
8. Grosman, J.: Fine-tuned XLS-R 1B model for speech recognition in German. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-german (2022)
9. Gulati, A., et al.: Conformer: convolution-augmented transformer for speech recognition. In: Proceedings of INTERSPEECH 2020, pp. 5036–5040. ISCA, Shanghai, China (2020)
10. Huang, B.: Fine-tuned whisper model for speech recognition in German. https://huggingface.co/bofenghuang/whisper-small-cv11-german (2022)
11. Huang, W., Hu, W., Yeung, Y.T., Chen, X.: Conv-transformer transducer: low latency, low frame rate, streamable end-to-end speech recognition. In: Proceedings of INTERSPEECH 2020, pp. 5001–5005. ISCA, Shanghai, China (2020)
12. Huang, Y., Ye, G., Li, J., Gong, Y.: Rapid speaker adaptation for conformer transducer: attention and Bias are all you need. In: Proceedings of INTERSPEECH 2021, pp. 1309–1313. ISCA, Brno, Czechia (2021)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of 7th International Conference on Learning Representations (ICLR). New Orleans, LA, USA (2019)

14. MacDonald, R.L., et al.: Disordered speech data collection: lessons learned at 1 million utterances from project Euphonia. In: Proceedings of INTERSPEECH 2021, pp. 3066–3070. ISCA, Brno, Czech Republic (2021)
15. McDowell, A.: Fine-tuned XLS-R 300M model for speech recognition in German. https://huggingface.co/AndrewMcDowell/wav2vec2-xls-r-300m-german-de (2022)
16. Moro-Velazquez, L., et al.: Study of the performance of automatic speech recognition systems in speakers with Parkinson's Disease. In: Proceedings of INTER-SPEECH 2019, pp. 3875–3879. ISCA, Graz, Austria (2019)
17. Ngueajio, M.K., Washington, G.: Hey ASR system! Why aren't you more inclusive? In: Chen, J.Y.C., Fragomeni, G., Degen, H., Ntoa, S. (eds.) HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence. HCII 2022. LNCS, vol. 13518. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21707-4_30
18. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: a review. Neural Netw. **113**, 54–71 (2019)
19. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv:2212.04356 (2022)
20. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), pp. 348–358. Curran Associates Inc, Vancouver, BC, Canada (2019)
21. Shor, J., et al.: Personalizing ASR for Dysarthric and accented speech with limited data. In: Proceedings of INTERSPEECH 2019, pp. 784–788. ISCA, Graz, Austria (2019)
22. Shrivastava, H., Garg, A., Cao, Y., Zhang, Y., Sainath, T.N.: Echo state speech recognition. In: Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5669–5673. IEEE Press, Toronto, ON, Canada (2021)
23. Vander Eeckt, S., Van Hamme, H.: Continual learning for monolingual end-to-end automatic speech recognition. In: Proceedings of 30th European Signal Processing Conference (EUSIPCO), pp. 459–463. IEEE Press, Belgrade, Serbia (2022)
24. Vaswani, A., et al.: attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), pp. 5998–6008. Curran Associates Inc, Long Beach, CA, USA (2017)