SCALING ASR IMPROVES ZERO AND FEW SHOT LEARNING

Alex Xiao, Weiyi Zheng*, Gil Keren, Duc Le, Frank Zhang, Christian Fuegen*
Ozlem Kalinli, Yatharth Saraf, Abdelrahman Mohamed

Facebook AI

ABSTRACT

With 4.5 million hours of English speech from 10 different sources across 120 countries and models of up to 10 billion parameters, we explore the frontiers of scale for automatic speech recognition. We propose data selection techniques to efficiently scale training data to find the most valuable samples in massive datasets. To efficiently scale model sizes, we leverage various optimizations such as sparse transducer loss and model sharding. By training 1-10B parameter universal English ASR models, we push the limits of speech recognition performance across many domains. Furthermore, our models learn powerful speech representations with zero and few-shot capabilities on novel domains and styles of speech, exceeding previous results across multiple in-house and public benchmarks. For speakers with disorders due to brain damage, our best zero-shot and few-shot models achieve 22% and 60% relative improvement on the AphasiaBank test set, respectively, while realizing the best performance on public social media videos. Furthermore, the same universal model reaches equivalent performance with 500x less in-domain data on the SPGISpeech financial-domain dataset.

Index Terms— large-scale, semi-supervised learning, transfer learning

1. INTRODUCTION

Using massive datasets to train neural models with ever-increasing sizes has spurred rapid progress in many fields of machine learning, such as natural language processing [1, 2], computer vision [3], and automatic speech recognition (ASR) [4, 5, 6]. The size of the training dataset and the number of model parameters are mutual bottlenecks and must be scaled in tandem [7]. In this paper, we explore and overcome the limitations of these two dimensions in ASR.

The abundance of publicly available text on the internet enabled the large-scale training of language representation models of up to 175B parameters on hundreds of billions of tokens [1]. On the other hand, supervised ASR datasets and models have been orders of magnitude smaller, and only recently, billion parameter ASR models are used with semi-/self-supervised methods [6, 8, 9] or through pooling together data from many sources [4].

This paper pushes these ideas to the extreme by pooling data from 10 different sources and employing semi-supervised training through pseudo-labeling. Our data contains 4.5 million hours of speech, most notably 4 million hours of unlabelled public social media videos on Facebook, uploaded from 120 countries and containing a wide variety of content and acoustic conditions. We propose data selection strategies to emphasize data diversity while reducing the computation cost of working with the whole dataset.

Data Source	**Transcriber**	**Hours**	**Hours After Augmentation**
LibriSpeech [18]	Human	960	5760
Common Voice [19]	Human	500	3000
Libri-Light [20]	Model	60000	360000
Fisher [21]	Human	1960	11760
Assistant*	Human	12600	41400
Conversational*	Human	780	6600
Calling Names	TTS	640	3840
Dictation*	Model	880	7920
Portal †	Human	1350	8100
Video †	Human	18000	108000
Portal †	Model	4800	28800
Video †	Model	4009400	4009400

Table 1. Our 4.5M hour dataset consists of 10 sources. Data sources marked with * are collected through third-party vendors. Those marked with † are collected from Facebook products.

Following prior work on scaling Transformer models [1, 10, 11], we scale the encoder of an E2E VGG-transformer transducer model [12, 13] up to 10B parameters. We leverage several techniques to train our transducer models efficiently on GPUs: FairScale model sharding [14], sparse alignment restricted transducer loss [15], mixed-precision training [16], and large batch sizes [17].

Prior work [4, 6, 9, 22, 23] explored mixing many datasets to train large multi-domain speech models but was limited to under 100K total hours and 1B model parameters. [24] analyzed scaling trends for acoustic models but did not go beyond 10K hours and 100M parameters. With a focus on multi-lingual models, [5] scaled ASR models up to 10B parameters but only demonstrated less than 0.5% relative improvement compared to 1B parameter models. This paper expands these efforts to show that English speech has sufficient difficulty to merit scaling to 10B parameters and shares a recipe to train models at this scale efficiently.

While videos on social media are abundant, other scenarios severely lack audio resources. For example, AphasiaBank [25], the largest source for aphasic speech recognition, contains under 100 hours of audio data. By pushing the limits of scale for ASR, we can improve ASR not just for domains with large datasets but also low resource domains like aphasic speech. Pre-training large models on a universal dataset shows impressive zero-shot 22% WER improvement on AphasiaBank. Transfer to other novel domains with zero, limited, and large-scale fine-tuning conditions exceed previously reported results, e.g., SPGISpeech [26] and an in-house dataset of long-form videos. We find scaling model size to 1B parameters to significantly improve zero and few-shot performance, even in low resource conditions.

*Equal contribution

2. DATA SCALING

2.1. Multi-domain Data Sources

Our first method of constructing a large speech recognition dataset is to pool data from various sources. Table 1 lists out the data sources used. The data can be grouped into four categories:

- Publicly released datasets: LibriSpeech [18], Common Voice [19], Libri-Light [20], and Fisher [21].
- In-house datasets collected from third-party vendors via crowd-sourced volunteers responding to artificial prompts with mobile devices. The content varies from voice assistant commands to a simulation of conversations between people.
- In-house datasets from Facebook products: public Facebook videos and voice commands to Portal. Videos used are from 120 different countries.
- Data generated from an in-house TTS model to increase the diversity of sentence patterns in our training data.

All in-house datasets are de-identified with no personally identifiable information (PII). Depending on the source, the data was further augmented with various distortion methods: speed perturbation [27], simulated reverberation, and randomly sampled additive background noise extracted from public Facebook videos.

We retain punctuation and casing from in-house datasets, which introduces inconsistency with some public datasets but allows the final model to output richer information. For evaluation, we use hand-transcribed data from the LibriSpeech, Portal, Video, and Conversational data sources, ranging from 3K-15K utterances with no overlap with training. We split up Video into "Standard" and "Challenging" subsets, where the "Challenging" subset contains videos with more noise and music.

2.2. Semi-supervised Labeling

The key to our data scaling strategy is leveraging 4 million hours of unlabelled audio with pseudo-labeling. The majority of our data in Table 1 comes from public Facebook videos labeled by a 1B parameter model trained on a smaller supervised and semi-supervised dataset from similar sources. We use a language identification model to select videos predicted to be English. These videos came from the same source as the supervised videos but may contain more challenging data such as singing and foreign speech.

2.3. Data Selection

4M hours of pseudo-labels present many challenges, including noisy labels and audio, viral videos dominating most of the content, and infrastructure requirements to work with such a massive dataset. Data selection is a common technique to address these issues [28]. We propose the following strategies for data selection to bring the dataset down to 1.3M hours only:

- *Words per Second*: Remove pseudo-labels with fewer than 0.5 words per seconds, which correlates with noisy music videos or foreign language.
- *Confidence Score*: Remove data with a bottom 20% confidence score to remove low confidence pseudo-labels.
- *Model Disagreement*: Re-decode the unlabeled data with an 80M parameter streaming model. We compute the edit distance between the two hypotheses to filter out data within the bottom and top 20% of disagreement to avoid too easy and too noisy utterances.
- *Segmentation + Alignment*: A hybrid model [29] is used for the alignment restricted loss [15] to segment data into 10s segments and filter out empty segments or ones that fail to align.
- *Rare Data*: Compute the cumulative word frequency distribution based on the supervised data and a consider a word to be rare if it is not in the top 90% most frequent words. Video segments with W words and R rare words are preserved if $R >= min(2, 0.25 * W)$. We keep all videos not uploaded from United States, Great Britain, Canada, or Australia to maintain data diversity.

Parameters	Hidden Size	Layers	Attention Heads
100M	512	36	8
1B	1152	60	16
10B	3072	90	48

Table 2. Hyper-parameters for our Transformer encoders.

3. MODEL SCALING

3.1. Model Architecture

Our model architecture is a non-streamable full-context Transformer-Transducer [12] with a VGG-Transformer encoder [13], a 19M parameter 2-layer LSTM predictor, and a 4M parameter feed-forward joiner. We focus on increasing the size of the Transformer encoder, which showed the most promise in initial experiments. Three encoders of 100M, 1B, and 10B parameters are constructed by varying the number of transformer layers and hidden dimensions. FFN dimension is always set to 4 times the hidden dimension. 3 VGG blocks are applied at the encoder input [13] for an inter-frame length of 80ms. We use 0.1 dropout in all Transformer blocks. Details for each model size are listed in Table 2. We also experimented with Mixture of Experts encoders [30] of up to 40B parameters. We did not see improvements, hence, we leave its exploration for future work.

3.2. Model Convergence

Due to convergence stability challenges in large mixed-precision models [31] with gradients or activations overflowing we recommend the following strategies:

- Pre-layernorm [32] avoids gradient explosion and enables better gradient flow.
- Scale the weight of the second linear layer in the FFN block by $\frac{1}{\sqrt{2n}}$, where n is the number of Transformer blocks [11].
- Set β_2 in the Adam optimizer to 0.98 to avoid network activations overflowing beyond the FP16's range [2].

3.3. Model Training Efficiency

Training transducer models with a billion or more parameters with distributed data parallel training is prohibitively slow. We leverage multiple optimizations to make it feasible to train such models in a reasonable amount of time. Large batch sizes can speed up training by improving the efficiency of GPU kernels and reducing the number of inter-GPU communication rounds required [17]. We use a global batch size of 23 hours. To fit such a large batch size into GPU memory, we leverage the following optimizations:

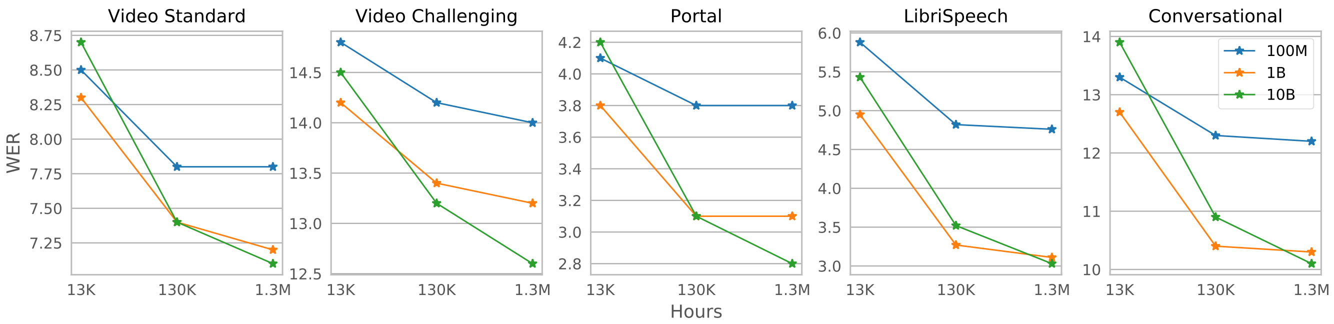

Fig. 1. WER results of a single universal model as we vary model size (100M, 1B, and 10B parameters) and dataset size (13K, 130K, and 1.3M hours). LibriSpeech results are reported on the test-other set. Increasing model size from 1B to 10B parameters only helps in the largest data setting: we see an average relative WER change of 7.32%, 2.19%, and -4.03% on 13K, 130K, and 1.3M hours respectively.

- *Alignment Restricted Transducer Loss* [15] utilizes word level alignments to reduce the memory required for transducer loss from $O(B \times T \times U \times D)$ to $O(B \times (T + U \times (b_l + b_r)) \times D)$, where B is the batch size, T the number of timesteps, U the number of target symbols, D the vocabulary size, and b_l and b_r are the left and right buffers. We set $b_l = 15$ and $b_r = 15$.
- *Fully Sharded Data-Parallel* [14] shards model weights, gradients, and optimizer states to reduce the memory consumption of large models. We only shard optimizer state and gradients to reduce the communication overhead.
- *Activation Checkpointing* [33] reduces activation memory by recomputing them in the backward pass.
- *Mixed Precision Training* [16] utilizes GPU Tensor Cores for more efficient compute and reduces the GPU memory and communication bandwidth required.

4. EXPERIMENTS

4.1. Experiment Details

We use 80-D log Mel features computed every 10ms with a window of 25ms. SpecAugment [34] with the LibriSpeech Double policy is applied to the input features. We train our models for 200,000 updates, linearly increasing the learning rate to $4e^{-4}$ in the first 20,000 updates and exponentially decaying by $1e^{-2}$ over the remaining updates. We use Adam with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e^{-6}$ and normalize the global gradient norm to 2. The vocabulary is set to 4095 BPE units. All training is done in Fairseq [35]. Our largest 10B parameter model is trained with 128 A100 GPUs for 25 days and needs $8.41 * 10^6$ PFLOPs for the encoder.

4.2. Impact of Scaling

To analyze the interaction between data and model size, we train a universal model with three different sizes on three datasets of 13K hours, 130K hours, and 1.3M hours. Results are plotted in Figure 1. We find that there is a benefit when scaling dataset size and model size together. At 1.3M hours, WER reduction is correlated with the model size where the 10B model obtains on average a 4.03% relative WER reduction compared to the 1B model and 20.00% relative reduction compared to the 100M model. Similarly, WER reduction is correlated with dataset size at 10B parameters. Increasing 130K hours to 1.3M hours improves the average relative WER of the 100M model by 0.01%, the 1B model by 1.67%, and the 10B model by 8.46%. These results suggest that scaling the model and dataset together is the key to further improvement.

Model Size	Data Size (h)	WER	Rare WER
10B	3.2M	7.21	11.00
10B	1.3M	**7.16**	**10.55**
1B	3.2M	7.56	11.53
1B	1.3M	**7.46**	**10.88**

Table 3. Effect of data selection with *Rare Data* and *Model Disagreement* after 100k updates on average WER and Rare WER.

4.3. Data Selection

Applying all data selection methods described in Section 2 reduced the original data from 4.5M hours to 1.3M hours. Without applying *Rare Data* and *Model Disagreement* filtering on the pseudo-labels, the dataset is about 3.2M hours. The goal of these two techniques is to reduce cost by removing unnecessary data while improving performance on the long tail. To measure the impact of these two methods, we introduce the rare WER metric, which measures WER only on words outside the top 90% cumulative word frequency distribution – computed on the supervised data. These words are often proper nouns and more important to the meaning of the utterance than common words like articles. Table 3 shows that reducing the data by 1.9M hours not only maintains the overall WER but also improves rare WER by 4-6% relative. Although we previously found increasing dataset size beneficial, these findings suggest that quality is more important than quantity: it is crucial to pick diverse samples when scaling up dataset size.

4.4. Zero-shot and Few-shot ASR

To understand how our models generalize to novel domains, we perform zero-shot and few-shot experiments on three datasets: AphasiaBank [25], SPGISpeech [26], and an in-house long-form videos dataset. We conduct few-shot learning by fine-tuning the universal models from Figure 1 further on each respective dataset. Our models achieve strong zero-shot performance and demonstrate impressive few-shot performance by exceeding baseline results by 16% to 60% relative (Table 4). In all cases, few-shot learning on top of our universal model is significantly superior to training on the relevant domain from scratch, enabling low-resource domains to enjoy the benefits of large models.

Our experiments also show that zero-shot and few-shot learning benefit from scaling from 100M to 1B parameters. The 10B results, however, are less consistent and points to overfitting during fine-tuning. We highlight our results below.

Dataset		AphasiaBank		Long-Form Video		SPGISpeech				
		Overall	**Fold 1**	**Short**	**Long**	**5000h**	**100h**	**10h**	**1h**	**10m**
Prior Work		37.37 [25]	-	-	-	2.3* [26]	-	-	-	-
From Scratch	100M	**53.39**	**51.72**	**13.52**	**8.94**	2.6(2.5)	18.5(18.5)	-	-	-
	1B	54.32	52.51	13.56	9.01	2.6(2.5)	**17.1**(17.0)	-	-	-
	10B	56.69	54.81	15.30	10.14	**2.4**(2.4)	27.9(27.8)	-	-	-
Universal	100M	30.29	29.63	12.99	9.17	7.1(4.9)				
	1B	**29.06**	**28.55**	11.88	**8.72**	6.5(4.4)				
	10B	30.05	29.33	**11.43**	9.33	**6.4**(4.3)				
+ Fine-tuning	100M	16.44	15.59	12.68	8.45	2.0(2.0)	2.7(2.6)	3.0(2.9)	3.5(3.4)	4.2(3.9)
	1B	**14.83**	**13.98**	11.18	**7.45**	**1.8**(1.8)	**2.2**(2.2)	**2.4**(2.3)	**2.7**(2.6)	**3.9**(3.4)
	10B	15.76	15.11	**11.09**	8.21	**1.8**(1.7)	**2.2**(2.1)	**2.4**(2.4)	2.9(2.8)	4.0(3.4)

Table 4. Results on novel domains. We benchmark 3 types of models: universal (trained on the general 4.5M hour dataset), from scratch (trained on the in-domain dataset), and fine-tuned (fine-tune the universal model on the in-domain dataset). Fine-tuning is significantly better than training from scratch and enables 1B+ models on lower resource domains. We also report WER computed with an in-house reference normalization in parentheses. *Private test we don't have access to.

4.4.1. AphasiaBank

Aphasia is an acquired speech-language disorder due to damages to portions of the brain, most commonly resulting from a stroke. It impairs verbal communication and makes it difficult for ASR systems to understand aphasic speech [36, 37]. Transcribed aphasic speech is also scarce: a large-scale aphasic speech dataset like AphasiaBank [25] only contains about 100 hours of recorded interactions between clinicians and persons with aphasia (PWAs). These challenges motivate leveraging transfer learning from a large, diverse dataset like ours. We hope that high-quality ASR for aphasic speech will allow PWAs to enjoy the benefits of ASR technologies while enabling medical analyses that rely on ASR [37].

We follow the same normalization, data folds, and data splits from [36]. Results aggregated across all four folds are shown in Table 4. When trained from scratch, large E2E models cannot achieve WER better than 50%. On the other hand, universal and fine-tuned models perform quite well; the fine-tuned 1B parameter model achieves a 60% relative WER improvement compared to the baseline hybrid model in [37] and a 72% relative WER improved compared to our own baseline, both of which were trained from scratch on AphasiaBank. These results indicate that few-shot learning benefits low resource domains like aphasic speech. More work needs to be done, however, to avoid overfitting for the 10B model.

4.4.2. SPGISpeech

SPGISpeech [26] contains 5,000 hours of financial audio from corporate earnings calls. We use the `norm` setting to analyze generalization to a more formal setting with financial jargon. The test set is private, so we split half of the 100h validation to create our own test set. While our results are not strictly comparable, the test sets are drawn from the same distribution.

Table 4 demonstrates that universal models perform somewhat reasonably but still struggle relative to [26]. Many errors are from jargon like “GAAP” or the mismatch in transcription conventions: 1/4 of the errors are from inserting “uh” and “um.” When using our in-house reference normalization, which avoids counting fillers as errors, the WER drops by about 30% relative. After fine-tuning, our 10B model enjoys a 23% relative improvement compared to [26] without any extra normalization.

We create smaller training sets with as little as 10 minutes of data to stress test low-resource adaptation. Our models display powerful adaptation capabilities: only 10 hours of fine-tuning data is needed to match the training performance from scratch on the original 500x larger dataset. Furthermore, few-shot learning improves WER by 20% relative using only 10 minutes of data. The 1B model performs the best while the 10B model overfits in ultra low-resource conditions. In contrast, the 1B model's extra capacity improves generalization from 5K hours to 1 hour; the improvement relative to 100M parameters steadily increases from 13% to 23%, suggesting a sweet spot for fine-tuning towards low resource domains.

4.4.3. Long-Form Video

We use 18000 hours of human-labeled long-form videos from social media to test the ability of our models to generalize to different lengths. These videos were in the original 4.5M hour dataset but with different lengths. We segment the training data to 45s instead of 10s and do not segment evaluation data. The evaluation videos are at most 5 minutes in length. Table 4 breaks down the results into short videos (less than 45s) and long videos (more than 45s).

Within the universal models, the 10B model does the best on short videos but the worst on long videos, which suggests that although our huge dataset may be diverse in some areas, length diversity is still a blind spot for the large models. Fine-tuning alleviates this problem, but the 1B model still has a 9% lower WER on long videos. This observation highlights the need for including length diversity when building large-scale datasets or regularization techniques to avoid overfitting to specific lengths [38].

5. CONCLUSION

In this work, we pushed the boundaries of large-scale speech recognition. We proposed an efficient recipe to train models of up to 10B parameters on 4.5M hours of audio. These large models demonstrated powerful zero-shot and few-shot learning capabilities across several domains, even with limited in-domain data. We also identified issues related to generalization and over-fitting in our current paradigm for scaling to 10B parameters. For future work, we plan to explore better low-resource transfer learning techniques for huge models. We will also investigate ways to improve data diversity and training objectives when working with massive datasets.

6. REFERENCES

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al., "Language models are few-shot learners," in *NeurIPS*, 2020.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, et al., "Self-supervised pretraining of visual features in the wild," *arXiv preprint arXiv:2103.01988*, 2021.

[4] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[5] Bo Li, Ruoming Pang, Tara N Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W Ronny Huang, and Min Ma, "Scaling end-to-end models for large-scale multilingual asr," *arXiv preprint arXiv:2104.14830*, 2021.

[6] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," in *NeurIPS*, 2020.

[7] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, et al., "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *arXiv preprint arXiv:2106.07447*, 2021.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[11] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[12] Ching-Feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, et al., "Transformer-transducer: End-to-end speech recognition with self-attention," *arXiv preprint arXiv:1910.12977*, 2019.

[13] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al., "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP*, 2020.

[14] Mandeep Baines, Shruti Bhosale, Vittorio Caggiano, Naman Goyal, Siddharth Goyal, Myle Ott, et al., "Fairscale: A general purpose modular pytorch library for high performance and large scale training," , 2021.

[15] Jay Mahadeokar, Yuan Shangguan, Duc Le, Gil Keren, Hang Su, Thong Le, et al., "Alignment restricted streaming recurrent neural network transducer," in *SLT*, 2021.

[16] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, et al., "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[17] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, et al., "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[19] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[20] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, et al., "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020.

[21] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: A resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[22] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Interspeech 2021*.

[23] Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve, "Rethinking Evaluation in ASR: Are Our Models Robust Enough?," in *Interspeech 2021*.

[24] Jasha Droppo and Oguz Elibol, "Scaling laws for acoustic models," *arXiv preprint arXiv:2106.09488*, 2021.

[25] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[26] Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, et al., "Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv preprint arXiv:2104.02014*, 2021.

[27] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015.

[28] Hank Liao, Erik McDermott, and Andrew Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *ASRU*, 2013.

[29] Duc Le, Xiaohui Zhang, Weiyi Zheng, Christian Fügen, Geoffrey Zweig, and Michael L. Seltzer, "From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition," in *ASRU*, 2019.

[30] William Fedus, Barret Zoph, and Noam Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*, 2021.

[31] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W Cottrell, and Julian McAuley, "Rezero is all you need: Fast convergence at large depth," in *UAI*, 2021.

[32] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, et al., "On layer normalization in the transformer architecture," in *ICML*, 2020.

[33] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin, "Training deep nets with sublinear memory cost," *arXiv preprint arXiv:1604.06174*, 2016.

[34] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.

[35] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, et al., "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL-HLT*, 2019.

[36] Duc Le and Emily Mower Provost, "Improving automatic recognition of aphasic speech with aphasia bank," in *Interspeech*, 2015.

[37] Duc Le, Keli Licata, and Emily Mower Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, pp. 1–12, 2018.

[38] Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, et al., "Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions," in *SLT*, 2021.