ARE E2E ASR MODELS READY FOR AN INDUSTRIAL USAGE?

Valentin Vielzeuf, Grigory Antipov

ARE E2E ASR MODELS READY FOR AN INDUSTRIAL USAGE?

Valentin Vielzeuf, Grigory Antipov

Orange, 4 rue du Clos Courtel, Cesson-Sévigné, France
{valentin.vielzeuf,grigory.antipov}@orange.com

ABSTRACT

The Automated Speech Recognition (ASR) community experiences a major turning point with the rise of the fully-neural (End-to-End, E2E) approaches. At the same time, the conventional hybrid model remains the standard choice for the practical usage of ASR. According to previous studies, the adoption of E2E ASR in real-world applications was hindered by two main limitations: their ability to generalize on unseen domains and their high operational cost. In this paper, we investigate both above-mentioned drawbacks by performing a comprehensive multi-domain benchmark of several contemporary E2E models and a hybrid baseline. Our experiments demonstrate that E2E models are viable alternatives for the hybrid approach, and even outperform the baseline both in accuracy and in operational efficiency. As a result, our study shows that the generalization and complexity issues are no longer the major obstacle for industrial integration, and draws the community's attention to other potential limitations of the E2E approaches in some specific use-cases.

Index Terms— Benchmark, Industry, ASR, E2E

1. INTRODUCTION

In recent years, Automatic Speech Recognition (ASR) has experienced an impressive breakthrough of performances measured as Word Error Rate (WER), especially on LibriSpeech, the most popular academic benchmark of English read speech [1]. As testified by the results aggregated by the website PapersWithCode[1], the End-To-End (E2E) fully-neural models have significantly outperformed conventional hybrid approaches [2] (the neural part of which is limited to an acoustic model). The dazzling progress of the E2E models has been mainly due to the proposal of new neural architectures (such as ContextNet [3] or Conformer [4]), to exploitation of large amounts of non-annotated speech data via semi- or self-supervised learning [5, 6], and to the new data augmentation techniques [7].

However, despite the fact that the progress of the E2E models is undeniable, hybrid models still remain a default option when building ASR systems for practical usage. Indeed, recent work highlights at least two major concerns hindering the adoption of such models in an industrial context, namely: (a) their generalization ability, and (b) their computational complexity (and therefore operational costs).

More precisely, several studies [8, 9] demonstrate that the scores on academic datasets such as LibriSpeech can be deceptive and poorly generalize on other speech domains. In particular, Szymanski et al. [8] urge the community to create new benchmarks, and illustrate that there is a huge gap between the WERs measured on popular academic datasets, and the WERs measured on private ones for various real-life use-cases. In the same spirit, Likhomanenko et al. [9] show that there is little generalization between performances of the contemporary E2E ASR models across public benchmark datasets, and that the models trained on LibriSpeech particularly struggle to transfer to other domains. Thus, a major milestone to better quantify this lack of generalization consists in building comprehensive evaluation datasets composed of speech of various nature (*i.e.* multi-domain evaluation). This is in line with very recent work proposing to aggregate different existing datasets [10]. The mentioned studies allow to clearly identify the generalization problem and Aksenova et al. [11] and some earlier work [12, 13, 14] logically propose to address it by augmenting the diversity of the training datasets (*i.e.* multi-domain ASR training).

On the other hand, E2E models are often associated with a larger computational burden. For instance, recent models [6] reach 300M parameters which represents around 30 times as much as the size of the acoustic part of the traditional hybrid models usually used in industry [2]. This strongly motivates the community to focus on the reduction of the computational cost of the E2E ASR approaches. For example, there is a strong interest in methods allowing online E2E ASR decoding without latency [15, 16]. Another branch of the literature targets lighter architectures which may help to reduce both training and inference time in several use-cases. Indeed, several efficient convolutional models have been proposed using the depthwise convolution and the simple CTC loss [17, 18]. Transformer-based models have also been studied in depth. For instance, at least three "efficient" Conformers [19, 20, 21] have been proposed recently.

Summarizing, the above-mentioned industrial constraints face the community with a trade-off between a high and reliable ASR accuracy and a low resource consumption (in the spirit of the Occam's Razor principle). In this paper, we demonstrate that there are contemporary E2E models which perfectly match the presented compromise and, therefore, show that the generalization and efficiency are no longer the major barrier to the industrial adoption of the E2E models. To this end, as illustrated in Figure 1, we benchmark promising E2E architectures comparing with a standard hybrid ASR model used for business applications by (a) performing both training and evaluation in a multi-domain context; and (b) measuring both the accuracy and the efficiency in a real-world aware manner.

2. MULTI-DOMAIN E2E VS. HYBRID BENCHMARK

2.1. Benchmark dataset

Multi-domain training & evaluation We follow the recommendations of the recent studies mentioned in Section 1 by designing a multi-domain dataset for our benchmark. We choose to work in English because it is the language with the largest choice of public ASR datasets. More precisely, we construct a collection of datasets issued from various application domains, namely: read speech (LibriSpeech [1]), phone conversations (SwitchBoard [22]), dictation (WSJ) [23], prepared talks (TED-LIUM [24]) and spontaneous non-

[1]Link to PapersWithCode.

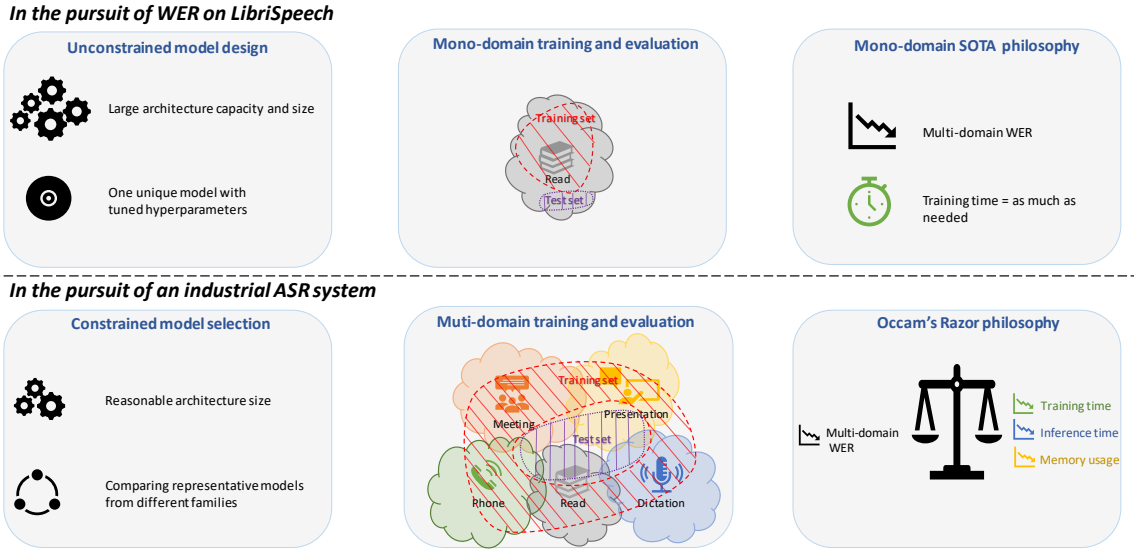

Fig. 1. Comparison between designing a mono-domain SOTA ASR model (top), and a general ASR system dedicated for an industrial usage (bottom). We adopt the latter strategy in order to benchmark SOTA-level representative E2E ASR models vs. a standard hybrid approach by performing a multi-domain training / evaluation while measuring both the transcription accuracy and the potential deployment costs.

native speech (we use a private dataset called Franglish, composed of the meeting recordings of French natives speaking English in a spontaneous manner). Thus, our collection is composed of 4 public well-studied datasets of various nature and of a private dataset. The latter (Franglish) is added to the benchmark as, to the best of our knowledge, there are no public datasets with non-native spontaneous English speech. At the same time, such speech includes grammatical imperfections as well as hesitations, repeated words and interrupted phrases, augments the acoustic variability of the data, and therefore, represents a real challenge for ASR systems (which perfectly fits our objective of evaluating the generalization of the compared models).

Data preparation Before training the E2E models on our multi-domain dataset, we normalize their annotations with the help of the Nemo Text Normalization tool [25]. We follow the previous studies and, for simplicity, leave Voice Activity Detection (VAD) out of the scope of the present benchmark by employing the ground-truth (oracle) speech segments both for training and evaluation. Finally, the training / valid / test splits of the public datasets of our collection follow the respective protocols established in the community. Franglish dataset is randomly split in proportion 80% / 10% / 10%.

2.2. Compared ASR models

Hybrid system In this benchmark, we use a standard hybrid approach from [2] as a baseline for the evaluated ASR models. Roughly speaking, it consists of 2 parts: an Acoustic Model (AM) and a Language Model (LM). AM is a TDNN [26] which is used to predict a posterior distribution over the tied Hidden Markov Model (HMM) states corresponding to context-dependent phonemes (biphones). These posterior distributions are then combined with a pronunciation dictionary (*i.e.* the lexicon) and a n-gram LM in order to construct a search graph in a form of WFST [27]. During the inference, the decoding is done via the beam search which looks for the best paths in the constructed graph.

E2E models Obviously, it is infeasible to evaluate all SOTA ASR models in the frame of one benchmark. Therefore, we select several SOTA-level representatives of the 3 large families of E2E models, namely: the recurrent ones, the fully-convolutional ones and the Transformer-based ones. More precisely, hereafter, we present the selected *encoder* architectures, while the same CTC decoder [28] is used for all E2E ASR models in the present benchmark.

Recurrent ASR encoder. Being natural candidates for modelling sequential data (such as speech recordings), RNN-based models are notorious for their computational complexity. For this benchmark, we choose a popular **CRDNN** architecture, which is a combination of a CNN, a RNN and a MLP composed of 120M trainable weights. In particular, we use the public implementation of this architecture [2].

Fully-convolutional ASR encoder is another promising family of models composed of convolutional blocks, which, in contrast to recurrent models, allow fast training and inference while also obtaining decent ASR performances. A SOTA-level model in this family is Citrinet [18] which is a CTC version of the Contextnet approach [3]. We evaluate 2 versions of Citrinet: **Citrinet-small** and **Citrinet-medium** (10M and 30M parameters, respectively).

Transformer-based ASR encoder. Transformer-based models are the SOTA in ASR today. But their computationally cost is high due to the quadratic complexity of the self-attention mechanism w.r.t. the input size. In this benchmark, we employ the Conformer [4] encoder combining self-attention and convolutional layers. In particular, we evaluate two versions of this architecture of varying complexity, namely: **Conformer-small** and **Conformer-medium** (13M and 30M parameters, respectively).

2.3. Evaluated metrics

Accuracy WER (defined as the ratio between the sum of the substitution S, deletion D and insertion I errors and the total number of words N in the ground-truth transcription: $WER = \frac{S+D+I}{N}$) is by far the most widely adopted metric for evaluation of the ASR systems. Therefore, we use it in our benchmark for evaluation of the multi-domain accuracy.

[2] Link to the CRDNN model description.

Models		*Mutli-domain Accuracy (WER in %)*							*Computational Cost*			
									Training	iRTF		#
		FR	LS_c	LS_o	SB	TED	WSJ	Overall	time (days)	CPU	GPU	params
Hybrid		37.2	11.0	25.4	25.3	12.1	9.3	20.0	7	2	N/A	8M
Conformer (small)	**Greedy**	30.1	5.8	13.7	20.1	9.4	6.7	14.3	7	**33**	50	13M
	+ LM	27.0	4.9	11.8	18.7	8.1	5.5	12.6				
Conformer (medium)	**Greedy**	32.3	6.2	13.9	21.2	9.9	6.8	15.1	7	17	50	30M
	+ LM	28.0	4.9	11.7	19.1	8.1	5.5	12.9				
Citrinet (small)	**Greedy**	33.8	5.0	12.9	22.1	9.2	6.5	14.9	7	25	50	10M
	+ LM	32.0	4.4	11.4	21.2	8.1	5.1	13.7				
Citrinet (medium)	**Greedy**	28.6	**4.0**	**10.2**	19.3	**7.6**	**5.0**	**12.5**	7	10	50	21M
	+ LM	26.3	6.5	10.6	19.7	7.7	8.6	13.2				
CRDNN	**Greedy**	**25.5**	5.5	15.1	**18.1**	8.3	**5.1**	12.9	14	2.5	50	120M
	+ LM	27.2	7.0	17.2	21.8	11.0	6.3	15.1				

Table 1. Summary of the principal benchmark results: Multi-domain Accuracy and Computational Cost. ASR WERs for all compared models are provided on the datasets described in Subsection 2.1, namely: Franglish (FR), LibriSpeech-clean (LS_c), LibriSpeech-other (LS_o), SwitchBoard (SB), TED (TED-LIUM), and WSJ. The column "Overall" is the average of the 6 scores. For each compared model, the WER scores are provided for the greedy and language model (LM)-based decoding. The latter is performed with a 3-gram LM and a beam size of 4. The training times are measured on contemporary work stations equipped with 4 Nvidia 2080 Ti GPUs. The reported inverted RTFs (iRTF) correspond to the average iRTFs over the test dataset which are measured in virtual machines where the compared models are run either on a single CPU or GPU (Nvidia 2080 Ti). The number of trainable parameters is reported in the last column.

Computational cost In addition to the number of parameters, the compared ASR models are evaluated according to 3 criteria which are particularly important for the integration of ASR systems, namely: the training time, the inference time and the required RAM. We allocate the same training time budget of 7 days of calculation on a modern workstation equipped with 4 Nvidia 2080 Ti GPUs for all E2E models. This value corresponds to the time required by the baseline hybrid model for convergence on the selected collection of training datasets. We make a single exception to this training budget for the CRDNN model which (due to its computational complexity) requires at least twice as much time to converge to competitive ASR performances. For the inference time, we employ the popular inverted Real Time Factor (iRTF) metric measuring the ratio between the real time of the input recording and the time spent by an ASR system for its transcription. Moreover, the inference time and the memory requirements obviously depend on the duration of the input audio recordings. Therefore, in our benchmark, we also evaluate both criteria by varying the size of inputs in order to evaluate the scalability of the compared ASR systems.

3. BENCHMARK RESULTS

3.1. Multi-Domain accuracy

The principal results of our benchmark are summarized in Table 1. The WER scores significantly vary depending on the evaluation dataset (and hence, on the target domain) which corroborates with the previous studies discussed in Section 1. As one might expect, the best transcription results are witnessed on the read speech (from 4% to 11% of WER on LibriSpeech_clean) which is widely recognized as the easiest ASR use-case. The results on the 16kHz-sampled prepared speech are somewhat close to those of the read speech (from 5% to 12% on TED-LIUM and WSJ). On the contrary, the accuracy drastically drops on the 8kHz-sampled phone speech (from 18% to 25% on SwitchBoard) and, above all, on the spontaneous accented speech (from 25% to 37% on Franglish) which clearly represents the biggest challenge among the domains included in the benchmark.

The results in Table 1 are unequivocal in terms of the ASR accuracy comparison between the E2E and hybrid models. Indeed, as one may observe, *all* compared E2E ASR models outperform the hybrid one by a large margin on *all* evaluation datasets. The relative WER improvements brought by the E2E models w.r.t. the hybrid one vary from about 30% to 65% depending on the dataset. In other words, the superiority of the E2E models does not limit to LibriSpeech, but is a rather general tendency on all ASR use-cases.

When comparing the E2E models between each other, the "Overall" column of Table 1 demonstrates that the Citrinets slightly outperform the Conformers of the similar sizes. However, it should be noted that (as explained in Subsection 3.2) we set the same training time budget of 7 days for all E2E models except for CRDNN. And the Conformers appear to converge a little slower than the Citrinets. Therefore, we suspect that the Conformers might be slightly underfit, which partially explains the worse accuracies than those obtained by the Citrinets. This is particularly true for the Conformer-medium which is outperformed by its simpler (but better converged) Conformer-small counterpart. At last, CRDNN obtains excellent WERs which are on-par with the ones of Citrinet-medium, but, as discussed in Subsection 3.2, at much greater cost than the latter.

Finally, it is worth noting that all compared E2E models perform reasonably well even with the simplest greedy decoding. This confirms that E2E models do not exclusively focus on the acoustics of the speech (as it is the case for the acoustic neural network of the hybrid model), but rather jointly learn acoustic and linguistics aspects of the language. Moreover, the WER scores of the most accurate E2E models, namely Citrinet-medium and CRDNN, are even deteriorated by the added LM. This can be explained by the fact that a simple 3-gram LM is used in our experiments, and probably Citrinet-medium and CRDNN implicitly learn a better representation of the language than the one provided by such a trivial LM.

3.2. Computational Cost

Currently the average cost of one hour rental of a workstation equipped with 4 contemporary GPUs is around $ 2. For one training run, it means an average cost of $ 336 for the Conformers, Citrinets,

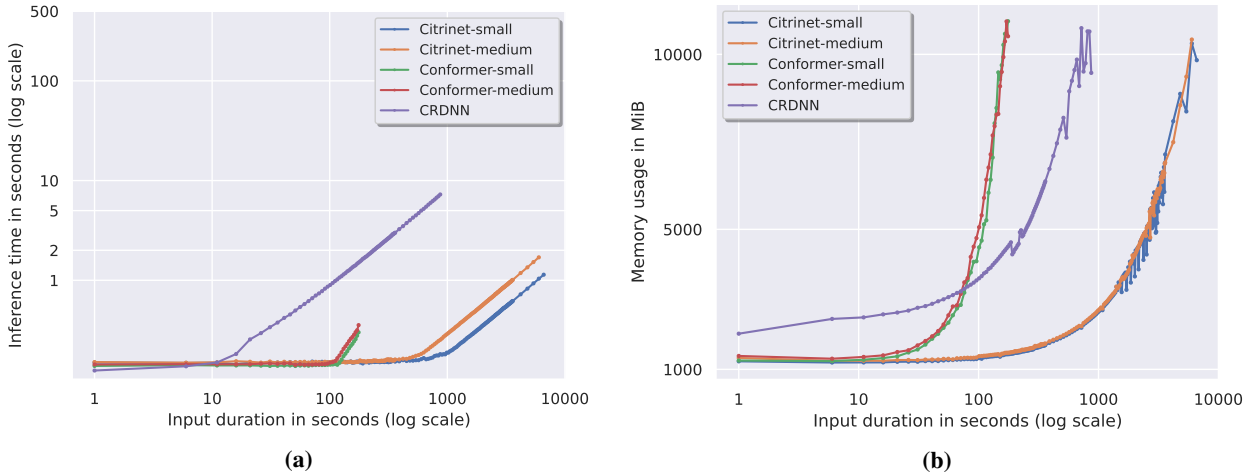

Fig. 2. Greedy inference time **(a)** and memory usage **(b)** vs. the input recording's length measured for the compared ASR E2E models on a contemporary workstation equipped with a single Nvidia 2080 GPU Ti.

and for the hybrid system and twice as much (*i.e.* $ 672) for the CRDNN. The WER improvements brought by CRDNN which are reported in Table 1 seem marginal comparing to its training cost. Moreover, one must keep in mind that several training runs are often needed to maintain the model with the most recent data or to adapt it to specific use-cases, multiplying the original cost.

When the model is finally trained and delivered for industrial usage, the main concern is its inference cost. In other words, how long does it take to process one standard input with a given hardware? From Table 1, we can see that all chosen models are faster than real-time, even using only one modern CPU. Yet, CRDNN and the hybrid system are only 2 times faster than real-time, while Conformer-small is 33 times faster, meaning that Conformer-small would need about 16 times less resources to guarantee the same rapidity of the transcription as the one provided by the hybrid model. Given an everyday intensive usage, such enormous difference may represent a very large cost gain arguing in favor of the deployment of light E2E models such as Conformer-small or Citrinet-small.

One may notice that according to the results in Table 1, GPUs do not seem to accelerate the inference of the E2E ASR models. However, this is only due to our evaluation protocol, which processes input sequences one by one and not in batches. In order to quantify the potential benefits brought by the GPU usage at inference, batches or longer sequences should be fed to the model. Hence, in Figure 2-(a) we extend the experiment to longer sequences. One may observe that for very short sequences (less than 10 seconds of audio), it's difficult to compare the E2E models, as the GPU is not optimally used. For longer sequences, the difference between CRDNN and the other E2E models becomes obvious, the former being much slower (even though keeping a very decent iRTF). The longest sequences perfectly illustrate that the Conformers are slower than the Citrinets, which (as discussed in Subsection 2.2) is due to the squared complexity of the self-attention w.r.t. the input duration. At last, one may observe an outstanding result of Citrinet-small which manages to process an hour of speech in less than a second.

The number of parameters is another important concern, as it is directly connected to the minimal required disk storage. For instance CRDNN is around 10 times larger than the smallest E2E models and therefore, does not imply the same industrial constraints. Concerning the specific case of the hybrid model, the reported parameters only include the acoustic neural network part and therefore, do not represent the real storage requirements. Moreover, for the E2E models, Figure 2-(b) shows that the problem is not only about the number of parameters, but also about the dependency between the memory usage and the length of the processed inputs. Indeed, the Conformers introduce an enormous memory burden when the sequences are too long, quickly saturating the GPU storage, while even the largest convolutional or recurrent models like the CRDNN are able to process such sequences while fitting on a modern GPU.

4. DISCUSSION AND CONCLUSION

In this work, we have proposed a multi-domain training and evaluation benchmark and studied 2 reported practical limitations of E2E ASR models: their generalization ability and computational cost. The evaluated E2E models have consistently outperformed the strong hybrid baseline system in terms of the multi-domain WER. The estimated computational costs also testify in favor of the E2E models. Indeed, all evaluated E2E models (except for CRDNN) significantly reduce both training time and inference costs w.r.t. the hybrid approach. We have also shown that E2E models scale well w.r.t. the input recordings duration (processing up to one hour at once for the Citrinets). As a result, our experiments demonstrate that generalization and efficiency can no longer be considered as the central issue preventing the industrial usage of E2E ASR, which allows us to positively answer the question put in the paper's title.

As a side result, the benchmark has pointed the Citrinets as a better trade-off (than the Conformers and CRDNN) between the resulting ASR accuracy and the training / inference complexity, and that a LM-free greedy decoding is sufficient to obtain decent performances on all tested use-cases.

Finally, we have left the problem of VAD out of the scope of the present benchmark, and all E2E experiments have been done with a trivial 3-gram LM. Therefore, the study on the impact of VAD and / or complex LM integration on the E2E model's accuracy and efficiency constitutes an important direction for future work. Another promising path of research would consist in further extending the evaluation protocol in order to assess the models' adaptability. Indeed, the hybrid ASR systems are known to be easily adaptable to a new lexicon, but can we say likewise regarding the E2E ones?

5. REFERENCES

[1] Vassil Panayotov, Guoguo Chen, Daniel Povey, et al., "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[2] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, et al., "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Interspeech*, 2016.

[3] Wei Han, Zhengdong Zhang, Yu Zhang, et al., "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," in *Interspeech*, 2020.

[4] Anmol Gulati, James Qin, Chung-Cheng Chiu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[5] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, et al., "Iterative pseudo-labeling for speech recognition," in *Interspeech*, 2020.

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

[7] Daniel S Park, William Chan, Yu Zhang, et al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.

[8] Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, et al., "Wer we are and wer we think we are," in *EMNLP*, 2020.

[9] Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, et al., "Rethinking evaluation in asr: Are our models robust enough?," *arXiv preprint arXiv:2010.11745*, 2020.

[10] Solene Evain, Ha Nguyen, Hang Le, et al., "Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech," *arXiv preprint arXiv:2104.11462*, 2021.

[11] Alëna Aksënova, Daan van Esch, James Flynn, et al., "How might we create better benchmarks for speech recognition?," in *Workshop on Benchmarking*, 2021.

[12] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ICML*, 2016.

[13] Arun Narayanan, Ananya Misra, Khe Chai Sim, et al., "Toward domain-invariant speech recognition via large scale training," in *SLT*, 2018.

[14] Naoyuki Kanda, Guoli Ye, Yu Wu, et al., "Large-scale pre-training of end-to-end multi-talker asr for meeting transcription with single distant microphone," *arXiv preprint arXiv:2103.16776*, 2021.

[15] Jiahui Yu, Wei Han, Anmol Gulati, et al., "Dual-mode asr: Unify and improve streaming asr with full-context modeling," in *ICLR*, 2020.

[16] Jiahui Yu, Chung-Cheng Chiu, Bo Li, et al., "Fastemit: Low-latency streaming asr with sequence-level emission regularization," in *ICASSP*, 2021.

[17] Vineel Pratap, Awni Hannun, Qiantong Xu, et al., "Wav2letter++: A fast open-source speech recognition system," in *ICASSP*, 2019.

[18] Somshubra Majumdar, Jagadeesh Balam, Oleksii Hrinchuk, et al., "Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition," *arXiv preprint arXiv:2104.01721*, 2021.

[19] Xiong Wang, Sining Sun, Lei Xie, et al., "Efficient conformer with prob-sparse attention mechanism for end-to-endspeech recognition," *arXiv preprint arXiv:2106.09236*, 2021.

[20] Shengqiang Li, Menglong Xu, and Xiao-Lei Zhang, "Efficient conformer-based speech recognition with linear attention," *arXiv preprint arXiv:2104.06865*, 2021.

[21] Maxime Burchi and Valentin Vielzeuf, "Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition," in *ASRU*, 2021.

[22] John J. Godfrey and Edward Holliman, "Switchboard-1 release 2," Linguistic Data Consortium, 1993.

[23] John S. Garofolo, David Graff, Doug Paul, et al., "Csr-i (wsj0) complete," Linguistic Data Consortium, 1993.

[24] François Hernandez, Vincent Nguyen, Sahar Ghannay, et al., "Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *SPECOM*, 2018.

[25] Yang Zhang, Evelina Bakhturina, Kyle Gorman, et al., "Nemo inverse text normalization: From development to production," *arXiv preprint arXiv:2104.05055*, 2021.

[26] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech*, 2015.

[27] Mehryar Mohri, "Weighted automata algorithms," in *Handbook of weighted automata*, 2009.

[28] Alex Graves, Santiago Fernández, Faustino Gomez, et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.