Interspeech 2024
1-5 September 2024, Kos, Greece

Exploring the limits of decoder-only models trained on public speech recognition corpora

Ankit Gupta, George Saon, Brian Kingsbury

IBM Research

ankitgupta.iitkanpur@gmail.com, gsaon@us.ibm.com, bedk@us.ibm.com

Abstract

The emergence of industrial-scale automatic speech recognition (ASR) models such as Whisper and USM, trained on 1M hours of weakly labelled and 12M hours of audio only proprietary data respectively, has led to a stronger need for large scale public ASR corpora and competitive open source pipelines. Unlike the said models, large language models are typically based on Transformer decoders, and it remains unclear if decoder-only models trained on public data alone can deliver competitive performance. In this work, we investigate factors such as choice of training datasets and modeling components necessary for obtaining the best performance using only public English ASR corpora. Our **D**ecoder-**O**nly **T**ransformer for **A**SR (DOTA) model comprehensively outperforms the encoder-decoder open source replication of Whisper (OWSM) on nearly all English ASR benchmarks and outperforms Whisper large-v3 on 6 out of 15 test sets. We release our codebase and model checkpoints under permissive license.

Index Terms: speech recognition, public corpora, Transformer

1. Introduction

Attention-based models [1] have been successful across many areas of machine learning [2, 3]. In particular, large language models (LLMs) comprising decoder-only Transformers pre-trained on large amounts of unlabelled text have become the standard in natural language processing, exhibiting impressive amounts of linguistic and world knowledge [4].

In contrast to LLMs, the best performing ASR models are typically based on Conformer acoustic encoders [5] and are trained with a connectionist temporal classification [6] or an RNN transducer [7] objective. These methods have been highly successful and are employed in nearly all of the best performing ASR models [8, 9, 10, 11, 12] exemplified by USM [13] which also serves as the speech encoder of Gemini v1 [14]. However, as these methods use a monotonic inductive bias specific to ASR, it is natural to investigate if Transformers trained to autoregressively generate text can deliver competitive performance. This was answered with the introduction of Whisper, a Transformer encoder-decoder model that is highly competitive with the best performing CTC-Conformer and RNN-T-Conformer pipelines on several speech recognition and translation benchmarks across multiple languages [15].

Since Whisper (large-v3) is trained on 1M hours of proprietary speech-text data and 4M hours of pseudo-labelled audio, it is natural to investigate if similar performance could be achieved via public ASR data alone. Secondly, unlike Whisper which is an encoder-decoder model, most LLMs are decoder-only and it is important to compare the performance of decoder-only models with encoder-decoder models in a data and compute-matched setting. The OWSM model [16] is a step towards this direction and comprises a Whisper-style encoder-decoder model trained on a compilation of public multilingual ASR corpora. However, as OWSM models are encoder-decoder Transformers trained using an additional CTC-based loss, it remains to be answered if conventional Transformer decoder training, similar to that for LLMs [17], suffices for competitive performance.

In this work we investigate the performance of Transformer decoders and prefix LMs [18] as well as the individual utility of public English ASR datasets by combining them into a large 93K hour paired speech-text corpus. Unlike OWSM, we train decoder-only models solely using cross-entropy loss. We train models varying over a wide range of hyperparameters such as 1) model size, 2) bidirectionality over audio frames, 3) downsampling rate of audio frames, 4) audio augmentation, and, 5) the datasets included in the training set. In addition, we also evaluate the performance of the trained models at low audio bitrates using recent neural codecs such as DAC [19].

We find that our best **D**ecoder-**O**nly **T**ransformer for **A**SR (DOTA) model outperforms Whisper large-v3 on 6 out of 15 test sets (Figure 1), and OWSM medium-v3.1 on nearly all test sets, while having fewer than half as many parameters as Whisper. Additionally, our best DOTA model uses twice the audio frame downsampling rate versus OWSM, making it faster.

We have open-sourced our codebase and trained models at `https://github.com/ag1988/mel-asr`.

2. Method

2.1. Data Preparation

To create a large-scale supervised ASR corpus we downloaded all major public English ASR datasets and organized them into a common format. Audio was resampled to 16kHz and stored as single-channel signed 16-bit integers in HDF5 format for memory-mapped access. Our training data consists of MultilingualLibriSpeech (English) [20], PeoplesSpeech [21], GigaSpeech [22], SPGISpeech [23], CommonVoice 11.0 [24], LibriSpeech [25], Fisher [26], TedLium 3 [27], AMI [28], FLEURS (English) [29], VoxPopuli (English) [30], LJ Speech [31], VoiceMail [32], and VCTK [33]. This resulted in 93K hours of speech-text pairs (top row of Table 3). In addition, we report results on an internal test set called Payload which contains 7.3h of speech (72k words) from the callcenter domain.

Text Processing We normalized the transcripts using Whisper's `EnglishTextNormalizer` module which converts text to lower case, removes punctuation and applies several other case-by-case transformations (see [15] for further details). We further remove the newline character `'\n'` and insert space between consecutive digits (e.g. $21 \mapsto 2\ 1$). We then tokenize the text using the

10.21437/Interspeech.2024-565

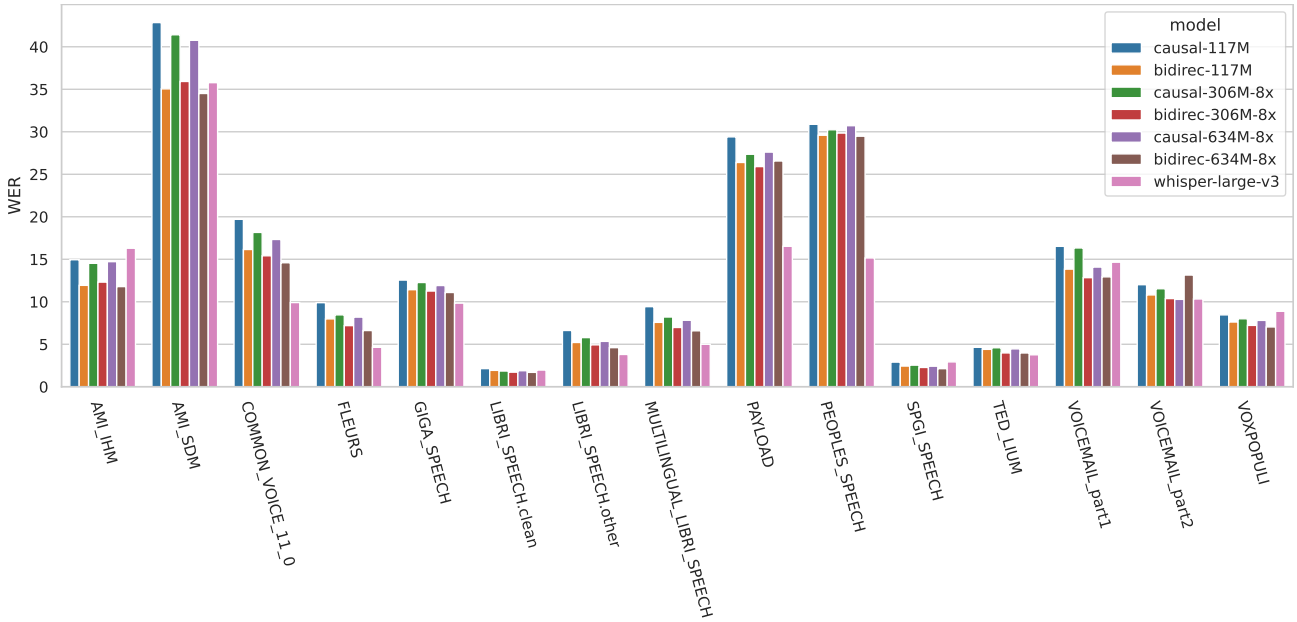

Figure 1: *Word error rates of* DOTA *models compared to Whisper on English test sets.*

Table 1: *Configurations of our* DOTA *models in Table 2. Feed-forward dimension is* $4\times$ *the model dimension and audio inputs are 30sec long. Audio frame rate for the model is the reciprocal of time shift.*

model	params (M)	layers	dimension	attention heads	embedding dim	time shift (ms)	vocab size
DOTA 117M	117	16	768	12	128	40	30522
DOTA 306M	306	24	1024	16	128	80	30522
DOTA 634M-8x	634	32	1280	20	128	80	30522
DOTA 634M-12x	634	32	1280	20	128	120	30522
OWSM medium v3.1	1020	18	1024	16	1024	40	50K
Whisper medium	769	24	1024	16	1024	20	51865
Whisper large v3	1550	32	1280	20	1280	20	51866

`bert-base-uncased` tokenizer [34]. Text inputs to the model are truncated to 146 tokens.

2.2. Audio Processing

The audio input to the model is a 30sec waveform consisting of 480K floats. Shorter instances are 0-padded to this length and longer waveforms are truncated. Log-mel spectrograms are computed using a window size of 25ms, a hop size of 10ms, and 80 mel bins. This results in 3000 audio frames per instance. For efficiency, we stacked every 4 frames into a single frame in the 117M model resulting in 750 audio frames. In the wider 306M and 634M models, we stacked every 8 (or 12) frames into a single frame further reducing the number of frames to 375 (or 250) (Table 1).

Audio Augmentation To make the model more robust to out of distribution instances we applied each of the following augmentations to the audio waveforms during training

- with probability (w.p.) 10^{-3} we applied speed augmentation with a factor sampled uniformly from [0.9,1.1].
- w.p. 0.2 tempo augmentation with a factor sampled uniformly from [0.9,1.1] [35].
- single-pole low pass filter w.p. 10^{-3}
- reverberation w.p. 10^{-3} using the `sox` "reverb" effect with default parameters
- given a partially formed training instance, we randomly concatenated the next sample from the same dataset to it w.p. $p = 0.25$ or else 0-padded it to 30secs w.p. $1 - p$, applying this process repeatedly until the length of the partially formed instance was at least 30sec. This allowed the model to explore the full 30sec input limit, preventing it from overfitting to instances mostly containing trailing silence.

2.3. Model

In this work, we are interested in exploring the limits of Transformer decoder-only architectures inspired by their success as language models [4]. As shown in Figure 2, our models comprise a single decoder stack with a causal attention mask. The input to the model is a sequence of audio frames followed by text token embeddings. Each of these vectors are mapped to the model dimension using linear projections. Each frame attends to itself and the frames on its left. However, for the prefix LMs (denoted by "bidirec") we allow audio frames to attend to the audio frames on their right as well. For simplicity, we used sinusoidal positional embeddings.

Table 2: *ASR performance (WER) on English test sets.* Notation*: "bidirec": prefix LM where causal attention mask is not used across audio frames allowing past audio frames to view future audio frames, "no-people": PeoplesSpeech was excluded from training set, "no-mls": MultilingualLibriSpeech was excluded from training set, "eval DAC 6kbps": test set audio was compressed via 16kHz 6kbps DAC. Whisper v3 uses 1M hours of weakly-labelled data and 4M hours of pseudo-labelled audio, OWSM uses 180K hours of public multilingual data and* DOTA *models use 93K hours of public English data. OWSM results are reported directly from [36] whereas the Whisper results are computed using the official repository.* DOTA *uses greedy decoding. For Whisper and* DOTA*, waveforms longer than 30sec are decoded after splitting them into 30sec segments and concatenating the corresponding outputs.*

Model	MultilingualLibriSpeech.en	PeoplesSpeech	GigaSpeech	SPGISpeech	CommonVoice 11.0.en	LibriSpeech.test-clean	LibriSpeech.test-other	TED-LIUM3	AMI-IHM	AMI-SDM	VoxPopuli.en	Fleurs.en	Voicemail.part1	Voicemail.part2	Payload (internal)
DOTA causal-117M	9.4	30.9	12.6	2.9	19.7	2.1	6.6	4.6	14.9	42.8	8.5	9.9	16.5	12.0	29.4
DOTA causal-117M-no-augmentation	9.2	31.2	12.4	3.0	19.2	2.7	6.7	4.8	12.9	39.9	8.5	10.0	20.5	27.0	43.5
DOTA causal-117M-no-people	9.2	31.3	12.6	2.9	19.3	2.2	6.1	4.9	19.0	47.0	9.1	9.2	16.4	12.4	29.3
DOTA causal-117M-no-people-no-mls	12.5	30.8	13.0	3.0	21.2	3.0	9.0	5.4	18.6	45.0	10.1	10.5	17.8	14.0	32.6
DOTA bidirec-117M	7.6	29.6	11.4	2.4	16.1	1.9	5.2	4.4	11.9	35.0	7.6	8.0	13.8	10.8	26.4
DOTA bidirec-117M (eval DAC 6kbps)	8.4	32.3	12.2	3.0	18.2	2.1	6.1	4.5	16.0	50.7	7.9	8.2	15.1	11.5	28.3
DOTA causal-306M-8x	8.2	30.2	12.3	2.5	18.2	1.9	5.8	4.6	14.5	41.4	8.0	8.5	16.3	11.5	27.4
DOTA bidirec-306M-8x	7.0	29.8	11.3	2.3	15.4	1.7	4.9	4.0	12.3	35.9	7.2	7.2	12.8	10.4	25.9
DOTA causal-634M-8x	7.8	30.7	11.9	2.4	17.3	1.9	5.3	4.5	14.7	40.8	7.8	8.2	14.1	**10.3**	27.6
DOTA bidirec-634M-8x	6.6	29.5	11.1	**2.1**	14.6	**1.7**	4.6	4.0	**11.8**	**34.5**	**7.0**	6.6	12.9	13.1	26.6
DOTA bidirec-634M-8x (eval DAC 6kbps)	7.2	31.2	11.5	2.5	16.2	1.9	5.1	4.1	14.7	49.0	7.2	7.1	13.2	13.0	27.9
DOTA bidirec-634M-12x	6.8	29.7	11.4	2.2	15.4	1.8	4.6	4.0	12.6	35.9	7.3	7.1	**12.7**	10.6	28.4
OWSM v3.1 medium (1.02B)	7.1				12.6	2.4	5.0	5.1			8.4	9.0			
Whisper large v2 (1.55B)	6.6	18.2	10.4	3.8	10.4	2.6	5.1	3.9	17.8	38.2	7.0	5.4	16.0	10.9	20.2
Whisper large v3 (1.55B)	**5.0**	**15.2**	**9.8**	2.9	**9.9**	2.0	**3.8**	3.8	16.3	35.8	8.9	4.6	14.7	**10.3**	**16.5**
Whisper large v3 (eval DAC 6kbps)	5.2	16.0	10.0	2.9	11.1	2.1	4.2	**3.6**	16.3	38.2	**7.0**	**4.1**	16.0	10.8	17.5

2.4. Training

Our models are trained for 1M steps with a batch size of 128 using ithe AdamW optimizer with moment scales $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The peak learning rate was 0.0002 linearly warmed up over 1K steps and decayed to 0 using a cosine schedule. Weight decay of 0.1 was used but was not applied to biases and layer norm parameters. Training was performed in bfloat16 precision and the gradients were clipped to norm 1.0. We did not use early stopping and used the final checkpoint in all evaluations. The largest DOTA models, bidirec-634M-12x and bidirec-634M-8x, were trained on 8 40GiB A100's for 85 and 120 hours respectively. During training, all datasets except MultilingualLibriSpeech, PeoplesSpeech, GigaSpeech, SPGISpeech and LibriSpeech, were oversampled by $2\times$ to compensate for the large differences in the dataset sizes. We experimented with various model and training configurations as shown in Table 2 with the corresponding model sizes in Table 1.

3. Speech Recognition Performance

To evaluate the performance of the trained models, we performed inference on the test and training sets using greedy decoding and computed the word error rates (WER) as shown in Tables 2 and 3. Due to the large sizes of training sets, we randomly sampled 24K instances for the training sets of size larger than this amount. During evaluation we did not perform any data augmentation and we removed any instances where the normalized transcript was empty. While evaluating DOTA and Whisper on the test sets, the waveforms longer than 30sec were split into 30sec long chunks and their corresponding predictions were concatenated. During training, waveforms longer than 30sec were removed.

As shown in Table 2, our DOTA models are competitive with the best performing ASR models such as Whisper, while having significantly fewer parameters. Concretely, our best model DOTA bidirec-634M-8x outperforms Whisper large-v3 on 6 out of 15 test sets despite the latter being trained on $10\times$ more data. Moreover, our bidirec-634M-8x model outperforms OWSM on all the test sets on which its authors reported results with the exception of CommonVoice (En).

Bidirectionality Comparing the performance of the "causal" DOTA models vs the corresponding prefix LMs ("bidirec") we find that bidirectionality over audio frames is critical to high performance across model scales. Surprisingly, even the smallest prefix LM DOTA bidirec-117M outperforms a significantly larger causal model DOTA causal-634M-8x.

Dataset ablation Our baseline causal model causal-117M is trained on 93K hours of paired data. To determine the utility of the largest components of our training data viz. MutlilingualLibriSpeech (MLS) and PeoplesSpeech of sizes 41K hours and 29K hours respectively we trained two analogous variants after ablating these datasets. As shown in Table 2, excluding PeoplesSpeech (causal-117M-no-people) did not result in significant degradation on most test sets with the exception of AMI test sets containing meeting recordings. However, further excluding MLS (causal-117M-no-people-no-mls) resulted in a significant degradation w.r.t. the baseline model.

Augmentation We trained a variant causal-117M-no-

Table 3: *WER computed over at most 24K samples chosen randomly from each training set.*

	MultilingualLibriSpeech.en	PeoplesSpeech.clean	PeoplesSpeech.clean.sa	PeoplesSpeech.dirty	PeoplesSpeech.dirty.sa	GigaSpeech	SPGISpeech	CommonVoice 11.0.en	Fisher	LibriSpeech.clean.100	LibriSpeech.clean.360	LibriSpeech.other.500	TED-LIUM3	AMI-IHM	AMI-SDM	VoxPopuli.en	Fleurs.en_us	Voicemail.part1	Voicemail.part2	LJSpeech	VCTK
dataset size (hours)	41K	6K	1K	22K	2K	10K	5K	1.5K	2K	100	360	500	453	78	77	522	7	15	15	23	82
number of samples $\times 1000$	9978	1501	257	5477	548	8283	1927	949	2228	28.5	104	149	268	108	107	182	2.6	1.8	2.0	13.1	88.2
DOTA causal-117M	2.6	5.9	5.6	17.9	17.7	3.5	2.4	5.3	8.6	1.5	1.6	2.2	3.8	8.1	32.3	5.5	5.0	7.0	8.1	0.9	0.5
DOTA causal-117M-no-augment	2.7	5.8	5.5	17.7	17.4	4.0	2.6	7.6	9.6	1.6	1.7	2.3	4.5	9.3	33.0	5.9	6.0	8.9	10.0	1.1	0.7
DOTA causal-117M-no-people	2.3	10.1	9.3	24.7	24.0	3.0	2.1	3.4	7.3	1.2	1.3	1.8	3.3	9.6	34.7	5.2	3.9	5.4	6.6	1.7	0.3
DOTA causal-117M-no-people/mls	4.9	10.4	9.6	25.1	24.4	2.1	1.6	1.3	3.9	1.4	1.5	2.1	1.9	6.3	24.6	3.5	2.1	2.6	3.7	1.7	0.1
DOTA bidirec-117M	2.4	5.6	5.2	17.2	16.7	3.0	2.1	4.7	7.9	1.3	1.4	1.8	3.3	5.8	24.0	5.2	4.7	6.2	7.3	0.7	0.4
DOTA causal-306M-8x	2.2	5.6	5.2	16.9	16.5	2.7	1.9	2.8	6.6	1.1	1.1	1.6	2.7	5.0	24.9	4.5	3.4	4.6	5.6	0.6	0.2
DOTA bidirec-306M-8x	2.1	5.4	5.0	16.4	16.0	2.5	1.8	3.1	6.9	1.1	1.1	1.5	2.7	4.1	19.9	4.4	3.4	4.4	5.6	0.5	0.2
DOTA causal-634M-8x	2.0	5.2	4.9	16.3	15.9	2.1	1.6	1.6	5.0	0.9	0.9	1.3	2.0	3.8	20.5	3.6	2.5	3.0	4.5	0.4	0.1
DOTA bidirec-634M-8x	2.0	5.1	4.6	15.8	15.3	2.1	1.6	1.7	5.1	0.9	1.0	1.2	1.9	2.1	13.9	3.5	2.5	2.9	4.2	0.4	0.1
DOTA bidirec-634M-12x	1.9	5.2	4.7	16.1	15.6	2.2	1.6	1.9	5.4	0.9	1.0	1.2	2.2	2.6	15.8	3.7	2.5	3.1	4.6	0.3	0.1
Whisper large v3	2.3	9.6	9.5	23.8	23.1	4.3	2.9	6.3	15.1	1.6	1.6	2.0	5.1	20.0	41.4	7.4	4.6	9.9	10.8	1.7	1.2

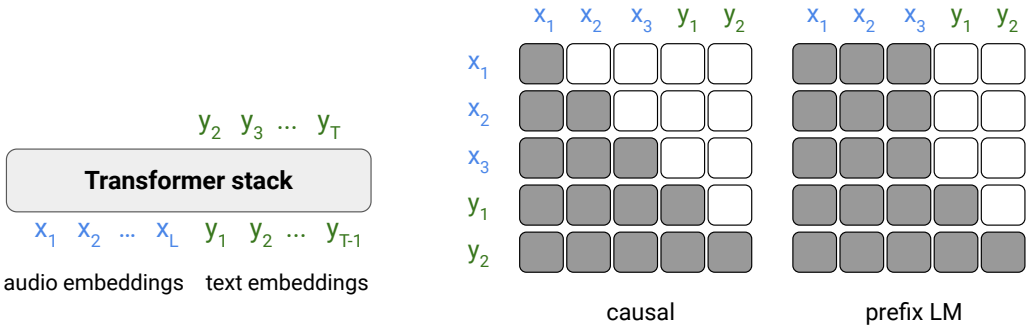

Figure 2: DOTA *models comprise a single Transformer stack that receives a sequence of concatenated audio and text embeddings as input. They are trained using the next (text) token prediction objective. The self attention layers can be purely unidirectional (causal) or allow audio positions to attend to future audio positions (prefix LM).*

augmentation of our baseline causal model without using the augmentations described in §2.2 and found their performance to be similar on the test sets.

Performance on training sets The performance on the training sets is summarized in Table 3. As our models are trained on these sets, the errors rates are low on most training sets with the exception of the "dirty" splits of PeoplesSpeech. Compared to our DOTA models, Whisper reports significantly higher error rates with performance gaps as large as 10 on Fisher (telephony) and 25 on AMI-SDM (distant microphone). This outcome suggests that Whisper was not trained on these public datasets.

Audio at low bitrate We also evaluated our trained models after encoding the audio with the DAC neural audio codec [19], which has reported significantly better performance at lower bitrates than commonly used codecs such as Opus. We used the 16kHz 6kbps version of DAC and include the results in Table 2. For our DOTA models, we observe significant degradation on test sets such as AMI (meetings) highlighting the importance of exposing the codec to diverse modalities during training. However, we observe no such degradations in case of Whisper pointing to the possibility of Whisper being exposed to low bitrate audio during training.. This is an interesting development as it is significantly easier to scale up dataset sizes at lower bitrates due to the smaller download size and disk footprint. We plan on mixing in lower bitrate audio while training future DOTA models.

4. Limitations and Future Work

In this work, we introduced DOTA, a single-stack Transformer trained using conventional cross-entropy loss on public ASR corpora and demonstrated it to match the performance of state of the art ASR models such as Whisper on several benchmarks.

However, unlike Whisper and OWSM, our DOTA models currently only support English and only perform ASR and we plan on extending our models to other languages and tasks.

Secondly, as shown for USM [13], audio pretraining can significantly reduce the amount of supervised data required for obtaining high performance. Given the abundance of audio-only data we plan on utilizing audio pretraining for further improving the performance of our models.

Acknowledgments We thank Samuel Thomas for helping us detect an error in our data loading script.

5. References

[1] A. Vaswani, N. Shazeer *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg *et al.*, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] J. M. Jumper, R. Evans *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583 – 589, 2021.

[3] D. Kondratyuk, L. Yu *et al.*, "VideoPoet: A large language model for zero-shot video generation," *ArXiv*, vol. abs/2312.14125, 2023.

[4] A. Q. Jiang, A. Sablayrolles *et al.*, "Mixtral of experts," *ArXiv*, vol. abs/2401.04088, 2024.

[5] A. Gulati, J. Qin *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020.

[6] A. Graves, S. Fernández *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the 23rd International Conference on Machine learning*, 2006.

[7] A. Graves, "Sequence transduction with recurrent neural networks," *ArXiv*, vol. abs/1211.3711, 2012.

[8] O. Kuchaiev, J. Li *et al.*, "NeMo: a toolkit for building ai applications using neural modules," *ArXiv*, vol. abs/1909.09577, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202712805

[9] E. G. Ng, C.-C. Chiu *et al.*, "Pushing the limits of non-autoregressive speech recognition," *arXiv preprint arXiv:2104.03416*, 2021.

[10] Y. Huang, G. Ye *et al.*, "Rapid speaker adaptation for conformer transducer: Attention and bias are all you need." in *INTERSPEECH*, 2021.

[11] M. Zeineldeen, J. Xu *et al.*, "Conformer-based hybrid ASR system for Switchboard dataset," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7437–7441.

[12] T. N. Sainath, Y. He *et al.*, "Improving the latency and quality of cascaded encoders," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8112–8116.

[13] Y. Zhang, W. Han *et al.*, "Google USM: Scaling automatic speech recognition beyond 100 languages," *ArXiv*, vol. abs/2303.01037, 2023.

[14] Google-Gemini-Team, "Gemini: A family of highly capable multimodal models," *ArXiv*, vol. abs/2312.11805, 2023.

[15] A. Radford, J. W. Kim *et al.*, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[16] Y. Peng, J. Tian *et al.*, "Reproducing Whisper-style training using an open-source toolkit and publicly available data," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2023.

[17] T. B. Brown, B. Mann *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato *et al.*, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[18] C. Raffel, N. Shazeer *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[19] R. Kumar, P. Seetharaman *et al.*, "High-fidelity audio compression with improved RVQGAN," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[20] V. Pratap, Q. Xu *et al.*, "MLS: A large-scale multilingual dataset for speech research," *ArXiv*, vol. abs/2012.03411, 2020.

[21] D. Galvez, G. Diamos *et al.*, "The People's Speech: A large-scale diverse English speech recognition dataset for commercial usage," *CoRR*, vol. abs/2111.09344, 2021.

[22] G. Chen, S. Chai *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Proc. Interspeech 2021*, 2021.

[23] P. K. O'Neill, V. Lavrukhin *et al.*, "SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv e-prints*, 2021.

[24] R. Ardila, M. Branson *et al.*, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[25] V. Panayotov, G. Chen *et al.*, "LibriSpeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[26] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

[27] F. Hernandez, V. Nguyen *et al.*, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer*. Springer International Publishing, 2018, pp. 198–208.

[28] J. Carletta, "Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 11 2007.

[29] A. Conneau, M. Ma *et al.*, "FLEURS: Few-shot learning evaluation of universal representations of speech," *arXiv preprint arXiv:2205.12446*, 2022.

[30] C. Wang, M. Riviere *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[31] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[32] M. Padmanabhan, G. Saon *et al.*, "Automatic speech recognition performance on a voicemail transcription task," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 433–442, 2002.

[33] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[34] J. Devlin, M.-W. Chang *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[35] T. Ko, V. Peddinti *et al.*, "Audio augmentation for speech recognition." in *INTERSPEECH*, 2015.

[36] Y. Peng, J. Tian *et al.*, "OWSM v3.1: Better and faster open Whisper-style speech models based on E-Branchformer," *arXiv preprint arXiv:2401.16658*, 2024.