ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446286

UPDATED CORPORA AND BENCHMARKS FOR LONG-FORM SPEECH RECOGNITION

Jennifer Drexler Fox[1], *Desh Raj*[2], *Natalie Delworth*[1], *Quinn McNamara*[1], *Corey Miller*[1], *Migüel Jetté*[1]

[1]Rev.com; [2]Center for Language and Speech Processing, Johns Hopkins University, USA.

ABSTRACT

The vast majority of ASR research uses corpora in which both the training and test data have been pre-segmented into utterances. In most real-word ASR use-cases, however, test audio is not segmented, leading to a mismatch between inference-time conditions and models trained on segmented utterances. In this paper, we re-release three standard ASR corpora—TED-LIUM 3, Gigapeech, and VoxPopuli-en—with updated transcription and alignments to enable their use for long-form ASR research. We use these reconstituted corpora to study the train-test mismatch problem for transducers and attention-based encoder-decoders (AEDs), confirming that AEDs are more susceptible to this issue. Finally, we benchmark a simple long-form training for these models, showing its efficacy for model robustness under this domain shift.

Index Terms— Long-form ASR, datasets, segmentation, transducers.

1. INTRODUCTION

Most ASR research uses corpora in which both the training and test data have been pre-segmented into utterances. Real-world audio, on the other hand, often occurs as *long-form* unsegmented recordings, leading to a mismatch between inference-time conditions and models trained on segmented utterances. This mismatch problem for long-form ASR has been well established in the literature [1, 2], and researchers have sought to tackle it through better segmentation [3, 4], large context acoustic modeling [5, 6], or rescoring with appropriate language models [7, 8].

A significant fraction of these long-form modeling techniques have only been evaluated on in-house or simulated data. In Fig. 1, we present corpus statistics from 36 published papers[1] on long-form ASR, showing that 26.9% and 5.7% used in-house and simulated data, respectively. Even when publicly available "true" long-form corpora were used, they were often multi-speaker (21.2%; e.g. AMI [10] and SwitchBoard [11]), non-English (32.6%), or contained missing segments (11.5%; GigaSpeech and TED-LIUM). This obscures the real long-context modeling problem with orthogonal issues such as overlapped speech, tokenization, or incorrect evaluation.

[1]These papers were manually selected based on an approximate depth-first search on the citation graph of a few seed papers, such as [9].

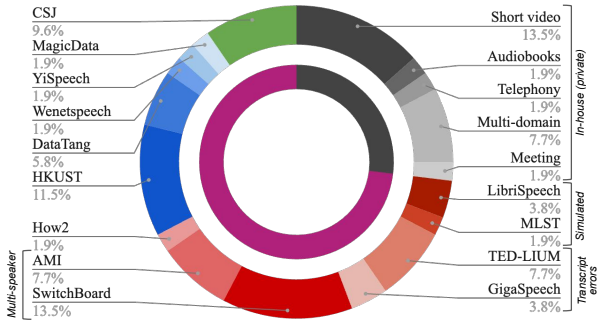


Fig. 1: Statistics of in-house (gray) and public (colored) datasets used in long-form ASR research. Color shades represent languages: English, Mandarin, and Japanese.

To enable fundamental research on this problem, we release long-form versions of three English ASR corpora: TED-LIUM 3 [12], Gigapeech [13], and VoxPopuli-en [14]. Although the original releases for these datasets provide full recordings, the completeness of their transcriptions varies significantly, creating several challenges towards their use for long-form ASR. For instance, portions of the recording may be untranscribed, or some segments may have been removed due to alignment problems or non-verbatim transcription. We reconstitute these long-form corpora through *linking* and *expansion* techniques (Section 3).

Finally, we use these reconstituted corpora to demonstrate the train/inference mismatch problem using ASR models trained on the original short-form segments, for both transducers [15] and attention-based encoder-decoders (AEDs) [16]. We show that incorporating long-form training can significantly improve performance when using chunk-wise overlapped inference. Our reconstituted versions of the corpora, along with word-level alignments, are publicly available as Lhotse manifests [17].[2]

2. RELATED WORK

There is a large body of work addressing long-form ASR from several perspectives. On the modeling front, researchers have extended end-to-end ASR for large context handling through strategies such as: using historical (or contextual) utterances to modify the encoder representation [5, 18, 19]), context expansion through preceding audio [1, 6, 20], and summarizing

[2]https://github.com/revdotcom/speech-datasets/tree/main/longform_reconstitution/

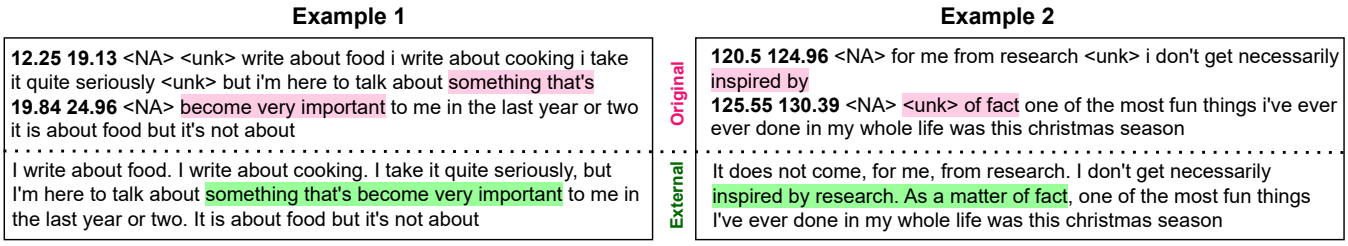


Fig. 2: Linkability determined by reference to external transcription

context through embeddings [21, 22]. Chiu et al. [9] compared popular end-to-end models for long-form ASR, finding that transducers are more robust than AEDs to the train-test mismatch. Often, this mismatch can be partially alleviated through techniques such as random utterance concatenation [23], minimum word error rate (MWER) training [2], and strong regularization [24]. OpenAI's Whisper model [25] takes a simpler approach to match training and inference conditions: in both cases, all audio is segmented into 30s chunks without any external VAD or diarizer.

Chunk-wise overlapped inference [9] is commonly used for offline decoding of long recordings. The related problem of segmentation has been addressed by using CTC-predicted blanks [26], a jointly trained continuous-integrate-and-fire (CIF) module [4], or using special tokens to predict segment boundaries [3]. Language models (LMs) trained with expanded context have been used in first-pass decoding [27], or more commonly for second-pass rescoring [7, 8, 28–30].

Despite such interest, there is little consensus about best practices for training/decoding in long-form ASR, partially because of a lack of common benchmarks. Although Earnings21 [31] and Earnings22 [32] were proposed to bridge this gap, they do not have any training data included, which makes it difficult to perform controlled investigations.

3. RECONSTITUTING LONG-FORM DATA

Our premise is that a "true" long-form corpus has long audio files and accompanying transcriptions. We used GigaSpeech, TED-LIUM, and VoxPopuli (`en` subset) as the base datasets for long-form reconstitution, since they provide such long recordings. These corpora have `train`, `dev`, and `test` partitions, but they are based on short segments and transcriptions (cut from the original recordings). In this section, we describe our reconstitution process for converting an eligible long-form corpus into a true long-form corpus. This process has two possible realizations, *linking* and *expansion*. We view linking as a long-form repackaging of an existing corpus, whose results are directly comparable with results on the original corpora. In contrast, we view expansion as a new version of an existing corpus, since we have added new audio segments or transcriptions to the existing data.

3.1. Linking

We define linking as concatenating original segments to make longer ones if no speech or transcriptions lie in between.

GigaSpeech comes with sequentially numbered segments that can be joined with previous and following segments when available. In some cases, segments were missing in the sequence, and thus were presumed to be untranscribed and would not be able to be linked across.

In contrast, internal resources were insufficient to allow for linking TED-LIUM. We observed several words in the audio that were not included in the transcriptions.[3] Most of these missing transcriptions were, however, present in the transcripts from a scrape of `ted.org`.[4] Mapping TED-LIUM talks to the scrape was largely automatic, but a remainder of files needed to be associated by a semi-automatic method. By referring to these externally-sourced complete transcriptions, we were able to link adjacent segments in the original partitions when there was no missing text between the segments. Fig. 2 shows two representative segment pairs. In Example 1, the external transcriptions (at bottom) indicate that there were no missing transcriptions in between, so linking is possible. In Example 2, the external transcriptions indicate that there were in fact missing transcriptions between the segments and thus they cannot be linked, unless the corpus is expanded. We will describe this expansion process in the following section.

3.2. Expansion

Expansion is an optional process involving the addition of speech and/or transcriptions to an existing corpus.

In VoxPopuli, purportedly exhaustive transcriptions are present in the original release, but 57% of transcribed segments are not in the partitions. Segments were marked invalid when an ASR system got >20% CER. We listened to several of these "invalid" segments and decided that their audio quality was not markedly different from other segments. For any paragraph used in a particular partition, we resuscitated formerly invalid segments, allowing longer sequences to be reconstituted.

For TED-LIUM, expansion involved using the scraped transcriptions described in 3.1 to replace the original transcriptions which had gaps.

[3] Previous work on long-form ASR using TED-LIUM seems to have missed or ignored this issue [26].

[4] `https://www.kaggle.com/datasets/thegupta/ted-talk`

Table 1: Statistics of reconstituted data: total size of the set (HH:MM) and average segment length (seconds). †For long-form TED-LIUM, the numbers outside and within parentheses represent linked and expanded versions, respectively.

	Dataset	Original		Long-form	
		Size	Length	Size	Length
GigaSpeech	Train (M)	999:56	4.0	1077:25	11.8
	Train (200h)	195:26	4.0	223:10	295.0
	Dev	11:50	6.6	10:29	1510.9
	Test	39:39	5.7	39:18	1088.2
TED-LIUM	Train	453:48	6.1	441:59 (514:23)	64.0 (827.4)
	Dev	1:35	11.3	1:31 (1:35)	459.3 (815.3)
	Test	2.37	8.2	2:24 (2:42)	576.1 (972.7)
VoxPopuli	Train	536:08	10.6	1111:46	143.7
	Dev	5:06	10.5	7:31	129.5
	Test	5:04	9.9	18:01	108.5

3.3. Statistics of reconstituted data

Table 1 provides summary statistics contrasting the original and reconstituted long-form versions of the corpora. Since GigaSpeech reconstitution is simply linking, the extra corpus size is entirely between-segment silence. We created two long-form versions: M and 200h. The former is obtained by linking segments of the original GigaSpeech-M (GS-M), which is approximately 1000 hours. Since GS-M is a *random* subset of GS-XL, it does not have many consecutive segments; as a result, the reconstituted long-form version only has an average length of 11.8s. To solve this issue, we created GS-200h specifically out of the longest consecutive segments in GS-XL, resulting in segments that are at least 240s. The `dev` and `test` have no missing references; therefore, their long-form versions are the full recordings.

For TED-LIUM, we show statistics for both the *linked* and *expanded* versions of the reconstituted data (we used the former for ASR experiments in Section 4). The increase in partition size for the expanded corpora results primarily from inclusion of inter-segment silence, not new references.

For VoxPopuli, expansion resulted in the addition of a substantial amount of new data. Compared to GS and TL, long-form segments are relatively short because we used the original paragraph segmentation.

3.4. Alignment

In addition to extended transcriptions, we also provide word-level timestamps obtained through forced alignments. For the linked corpora, we used a HMM-GMM model trained using Kaldi [33] to align the original short segments. For the expanded corpora, we modified the Fairseq aligner [34] to provide start and end times and accompanying scores for each word. This aligner was run on the complete TED-LIUM talks and the VoxPopuli paragraphs. We used these word-level timestamps to create fixed-length chunks for training (e.g., 15 or 30 seconds) by concatenating subsequent words until the segment exceeded the specific length. We have supplied these alignments in our distribution to enable users to experiment with other segment lengths or dynamic segmentation.

4. ASR BENCHMARKS

4.1. Models

For this paper, we benchmarked two common end-to-end ASR model architectures. The first, **neural transducers** [15] use a training loss which marginalizes over all possible alignments between input audio frames and output label sequences using a blank label to account for differences in sequence length. This *frame-synchronous* behavior may result in more robustness to train-test length mismatch. Furthermore, it also allows the estimation of token-level time-stamps at inference.

Conversely, we also trained a *label-synchronous* **attention-based encoder-decoder** (AED) in the joint CTC-attention framework [35]. The attention head is trained with a label-wise cross-entropy loss, whereas the CTC head is trained with a sequence-level alignment-free criterion [36].

4.2. Overlapped chunk-wise decoding

We follow the overlapped chunk-wise decoding strategy [9]. Given a long recording, we chunk it into fixed-length segments of size ℓ_{ch}, and extend them by an additional ℓ_{ex} on each side to avoid edge effects. These segments are decoded using the transducer/AED model to obtain time-stamped tokens. We discard the edge tokens which belong to the extra regions in each segment. Finally, we concatenate the segment-level tokens to obtain the transcript for the recording. Unlike [37], we do not need to align the overlapped regions of consecutive segments, but the models are required to estimate token-level time-stamps.

4.3. Long-form inference with attention decoding

We make two changes to the standard attention decoding paradigm to enable long-form inference. First, to combat this problem of high deletion rates on longer segments, we remove short hypotheses after beam search - any hypothesis with 10 or more tokens fewer than the longest hypothesis is removed from consideration. We use this setting for all AED decoding results. Second, because token-level time-stamps are required for the overlapped inference method described above, we obtain these from a constrained forward pass with the CTC head after decoding with the attention head.

4.4. Experimental Setup

For our experiments with neural transducers, we modified the standard Zipformer-transducer recipe in icefall[5], trained using the pruned RNN-T loss [38]. The encoder consists of 6 Zipformer blocks [39], which are subsampled by up to 8x, and contain multiple self-attention layers (with shared attention weights). The prediction network is a 1D-convolutional layer

[5]`https://github.com/k2-fsa/icefall`

Table 2: WER results on original test sets.

Model	Size (M)	TL		GS		VP	
		Dev	Test	Dev	Test	Dev	Test
Transducer	65.5	6.38	5.86	14.49	13.98	8.03	8.29
AED	109.8	9.11	8.48	15.34	15.31	13.63	14.07

Table 3: WER results on long-form evaluation. †Long-form training for VoxPopuli contains non-verbatim transcripts.

Model	Training data	TL		GS		VP†	
		Dev	Test	Dev	Test	Dev	Test
Transducer	Original	7.02	6.08	17.09	17.06	16.37	19.33
	+ Long-form	6.25	5.71	16.21	16.36	26.18	29.84
AED	Original	60.58	62.77	45.05	45.50	34.21	39.06
	+ Long-form	18.88	23.89	20.17	20.78	27.83	33.02

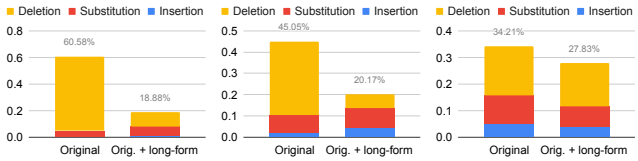
with bigram context. We used greedy search for the chunk-wise decoding strategy. AEDs are implemented as part of a joint CTC/attention model using the Wenet toolkit[6]. They consist of a 12-layer conformer [40] encoder and 6-layer bidirectional transformer decoder, although only the 3 forward decoder layers are used for inference. We use beam search for all decoding conditions. For both models, 80-dim log Mel filter-banks are used as acoustic input, and the output units are BPEs. We trained all models using SpecAugment [41].

We experimented with two training datasets - original and original+long-form - per corpus. For TED-LIUM, the long-form segments came from the linked version of the corpus. For Gigaspeech, the original condition used original segments from both the M and 200h subsets of the corpus; the original+long-form condition used the original segments from the M subset and long-form segments from the 200h subset. Due to GPU memory constraints, the long-form partitions were split into 30s segments for both training and evaluation. We evaluated the models on the original segments to compare word error rate (WER) performance with standard segmentation, and then on the reconstituted long-form sets to measure robustness to train-test mismatch.

4.5. Results

Table 2 shows the performance of the different model types when trained and evaluated on the original segments. Across all three corpora, the transducer model performed best, but both model types gave reasonable results. However, when these models were used to decode the reconstituted long-form data, their performance varied significantly, as seen in the "Original" rows of Table 3. As expected, the transducer model degraded only slightly (e.g., 6.38%→7.02% on TED-LIUM dev), whereas **AED degraded significantly** on all test suites, driven predominantly by high deletion rates. The breakdown of AED WERs into insertion, substitution, and deletion errors can be seen in Figure 3.

[6]https://github.com/wenet-e2e/wenet



(a) TED-LIUM **(b)** GigaSpeech **(c)** VoxPopuli

Fig. 3: Error rates on the `dev` sets for AED long-form inference.

Table 3 also shows the results for long-form training using the updated `train` sets. For both TED-LIUM and GigaSpeech, this training led to small improvements in transducer performance and large improvements in AED performance, although the transducer was still significantly better than the AED for long-form inference.

From Fig. 3, we see that long-form training resulted in **large reductions in the deletion rate**, leading to better performance.

The inclusion of the long-form VoxPopuli data degraded performance for the transducer model, and only improved the AED model slightly. This is likely due to our being overly permissive with the included data. Some of the VoxPopuli references in the original transcripts were heavily edited (i.e., non-verbatim transcription), including reordering of words. Since the transducer model assumes monotonic alignments, training with such transcripts could potentially deteriorate the model. Future work is needed to find the appropriate balance between including additional data needed for long-form training and rejecting low-quality references. Alternatively, recently proposed techniques such as bypass temporal classification [42], which allow training with imperfect transcripts, could be explored for making the best use of this data.

5. CONCLUSION

In this work, we released updated long-form versions of three popular English datasets — TED-LIUM, GigaSpeech, and VoxPopuli. This was achieved using a general "reconstitution" recipe comprising linking and expansion stages. To accompany this release, we presented baseline results using two commonly used models, transducers and AEDs. Across all three datasets, we demonstrated that transducers are more robust than AEDs to the train/test mismatch, when trained on segmented utterances. Finally, we showed that a simple strategy of combining original and long-form segments for training is effective at reducing the performance gap. Nevertheless, more research into training and modeling strategies is required to make long-form ASR robust in real scenarios, and we believe our public benchmarks would be important to measure progress.

Acknowledgments. This work was started during JSALT 2023, hosted at Le Mans University, France, and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, and Microsoft. D.R. acknowledges funding by NSF CCRI Grant No. 2120435 and a JHU-Amazon AI2AI fellowship.

6. REFERENCES

[1] A. Narayanan, R. Prabhavalkar, C.-C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," in *IEEE ASRU*, 2019.

[2] Z. Lu, Y. Pan, T. Doutre, P. Haghani, L. Cao, et al., "Input length matters: Improving RNN-T and MWER training for long-form telephony speech recognition," *ArXiv*, 2021.

[3] W. R. Huang, S.-y. Chang, D. Rybach, R. Prabhavalkar, T. N. Sainath, et al., "E2E segmenter: Joint segmenting and decoding for long-form ASR," in *InterSpeech*, 2022.

[4] Y. Shu, H. Luo, S. Zhang, L. Wang, and J. Dang, "A CIF-based speech segmentation method for streaming E2E ASR," *IEEE Signal Processing Letters*, vol. 30, 2023.

[5] X. Gong, Y. Wu, J. Li, S. Liu, R. Zhao, X. Chen, and Y. Qian, "LongFNT: Long-form speech recognition with factorized neural transducer," in *IEEE ICASSP*, 2022.

[6] T. Hori, N. Moritz, et al., "Transformer-based long-context end-to-end speech recognition," in *InterSpeech*, 2020.

[7] W. Xiong, L. Wu, J. Zhang, and A. Stolcke, "Session-level language modeling for conversational speech," in *EMNLP*, 2018.

[8] T. Chen, C. Allauzen, et al., "Large-scale language model rescoring on long-form data," in *IEEE ICASSP*, 2023.

[9] C.-C. Chiu, W. Han, Y. Zhang, R. Pang, S. Kishchenko, et al., "A comparison of end-to-end models for long-form speech recognition," in *IEEE ASRU*, 2019.

[10] J. Carletta, S. Ashby, et al., "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, 2005.

[11] J. J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *IEEE ICASSP*, 1992.

[12] F. Hernandez et al., "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *SPECOM*, 2018.

[13] G. Chen, S. Chai, G.-B. Wang, et al., "GigaSpeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio," in *InterSpeech*, 2021.

[14] C. Wang, M. Rivière, A. Lee, et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.

[15] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Representation Learning Workshop*, 2012.

[16] J. Chorowski, D. Bahdanau, D. Serdyuk, et al., "Attention-based models for speech recognition," in *NIPS*, 2015.

[17] P. Żelasko, D. Povey, J. Y. Trmal, and S. Khudanpur, "Lhotse: A speech data representation library for the modern deep learning ecosystem," in *NeurIPS Data-centric AI Workshop*, 2021.

[18] R. Masumura, T. Tanaka, T. Moriya, Y. Shinohara, T. Oba, and Y. Aono, "Large context end-to-end automatic speech recognition via extension of hierarchical recurrent encoder-decoder models," in *IEEE ICASSP*, 2019.

[19] S. Kim and F. Metze, "Dialog-context aware end-to-end speech recognition," in *IEEE SLT*, 2018.

[20] A. Schwarz, I. Sklyar, and S. Wiesler, "Improving RNN-T ASR accuracy using context audio," in *InterSpeech*, 2020.

[21] M. Cui, J. Kang, J. Deng, X. Yin, Y. Xie, et al., "Towards effective and compact contextual representation for conformer transducer speech recognition systems," in *InterSpeech*, 2023.

[22] K. Wei, Y. Zhang, S. Sun, L. Xie, and L. Ma, "Leveraging acoustic contextual representation by audio-textual cross-modal learning for conversational ASR," in *InterSpeech*, 2022.

[23] Y. Y. Lin, T. Han, H. Xu, V. T. Pham, et al., "Random utterance concatenation based data augmentation for improving short-video speech recognition," in *InterSpeech*, 2022.

[24] C.-C. Chiu, A. Narayanan, W. Han, R. Prabhavalkar, Y. Zhang, N. Jaitly, R. Pang, T. N. Sainath, P. Nguyen, L. Cao, and Y. Wu, "RNN-T models fail to generalize to out-of-domain audio: Causes and solutions," *IEEE SLT*, 2020.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[26] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with CTC-based voice activity detection," *IEEE ICASSP*, 2020.

[27] X. Chen, S. Parthasarathy, W. A. Gale, S. Chang, and M. Zeng, "LSTM-LM with long-term history for first-pass decoding in conversational speech recognition," *ArXiv*, 2020.

[28] S. R. Chetupalli and S. Ganapathy, "Context dependent RNNLM for automatic transcription of conversations," in *InterSpeech*, 2020.

[29] K. Irie, A. Zeyer, R. Schlüter, and H. Ney, "Training language models for long-span cross-sentence evaluation," in *IEEE ASRU*, 2019.

[30] S.-H. Chiu, T.-H. Lo, and B. Chen, "Cross-sentence neural language models for conversational speech recognition," in *IJCNN*, 2021.

[31] M. D. Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, et al., "Earnings-21: A practical benchmark for ASR in the wild," in *InterSpeech*, 2021.

[32] M. D. Rio, P. Ha, Q. McNamara, C. Miller, and S. Chandra, "Earnings-22: A practical benchmark for accents in the wild," *ArXiv*, 2022.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[34] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, et al., "Scaling speech technology to 1,000+ languages," *Arxiv*, 2023.

[35] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE ICASSP*, 2016.

[36] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[37] T. G. Kang, H.-G. Kim, M.-J. Lee, J. Lee, and H. Lee, "Partially overlapped inference for long-form speech recognition," *IEEE ICASSP*, 2021.

[38] F. Kuang, L. Guo, W. Kang, L. Lin, M. Luo, Z. Yao, and D. Povey, "Pruned RNN-T for fast, memory-efficient ASR training," in *InterSpeech*, 2022.

[39] D. Povey, ," `https://github.com/k2-fsa/icefall/blob/master/egs/librispeech/ASR/zipformer/zipformer.py`.

[40] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech 2020*, 2020.

[41] D. S. Park, W. Chan, et al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *InterSpeech*, 2019.

[42] D. Gao, M. Wiesner, H. Xu, L. P. Garcia, D. Povey, and S. Khudanpur, "Bypass temporal classification: Weakly supervised automatic speech recognition with imperfect transcripts," 2023.