ReazonSpeech: A Free and Massive Corpus for Japanese ASR

Yue Yin[1] Daijiro Mori[1] Seiji Fujimoto[2]
[1]Reazon Holdings, Inc. [2]Clear Code, Inc.
{yue_yin, daijiro_mori}@reazon.jp fujimoto@clear-code.com

Abstract

ReazonSpeech is a 15,000-hour and continuously growing corpus collected from Japanese TV shows free for commercial usage. The automatic speech recognition (ASR) model trained on ReazonSpeech achieves state-of-the-art results with 8.23% character error rate (CER) on JSUT basic5000[1] and 9.93% on Common Voice[2] v8.0 test set, on par with the recently released Whisper[3] large-v2 model. We released the dataset creation toolkit under Apache License 2.0 and made both the corpus and the pretrained ASR model freely available[1)].

1 Introduction

End-to-end (E2E) automatic speech recognition has recently drawn attention for its state-of-the-art results on benchmark datasets in terms of accuracy and its promise for not needing in-depth expertise compared to hybrid models[4]. With the release of large-scale corpora and models for English, the development of English E2E ASR is flourishing more than ever. Compared to English, Japanese is still resource-scarce and the development of E2E ASR is largely hindered by this scarcity. This paper introduces our attempt to construct a freely available large-scale high-quality Japanese ASR corpus from TV recordings via bootstrapping labeling seeding from a small-scale corpus. The release of the corpus is bundled with the creation pipeline free-and-open-source for continual evolution. The initial release of the corpus consists of 15,735 hours of audio data and is by far the largest freely available Japanese ASR corpus of our knowledge and the construction is intended to be continuous. With the Amendment to the Copyright Act[2)] encouraging the use of digital contents in AI going on in Japan since 2019, we hope to ride the crest of the wave and bring Japanese E2E ASR to a new stage by open-sourcing the project.

2 Related Work

There has been a line of research on creating Japanese ASR corpora recently. High-quality corpora that require human efforts such as CSJ[5], and Common Voice[2] have been explored extensively for Japanese ASR for the last decades. While highly supervised corpora are fundamental for hybrid models and remain essential to E2E ASR as benchmark datasets for evaluation, the expensive human labor cost casts a limit on the scalability of such highly supervised corpora in terms of E2E training. These days, the popularity of bootstrapping audio-transcription pairs for ASR training from media like audiobooks and subtitled videos using existing ASR resources has taken a rise. Corpora like The People's Speech [6] and GigaSpeech[7] have shown the competency of automatically generated large-scale corpora for ASR training. For Japanese, LaboroTVSpeech[8] and JTubeSpeech[9] have been introduced in recent years and built up a solid foundation for Japanese E2E ASR. More recently, with the release of Whisper[3], sacrificing the supervision for scale is shown to be promising for E2E ASR. In this work, we applied the automatic construction techniques to TV recordings and further seeks the potential of constructing large-scale datasets without depending on existing resources.

3 Corpus Construction

A lot of Japanese TV shows are aired with subtitles that can be leveraged for ASR corpus creation. However, the subtitles cannot be directly used to construct an ASR corpus mainly due to the following reasons:

1. timestamps are inaccurate, especially for news and live TV programs
2. transcriptions are missing for commercials and speech

1) Sample data and relevant resources can be found at the project page https://research.reazon.jp/projects/ReazonSpeech
2) Detailed information can be found at https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

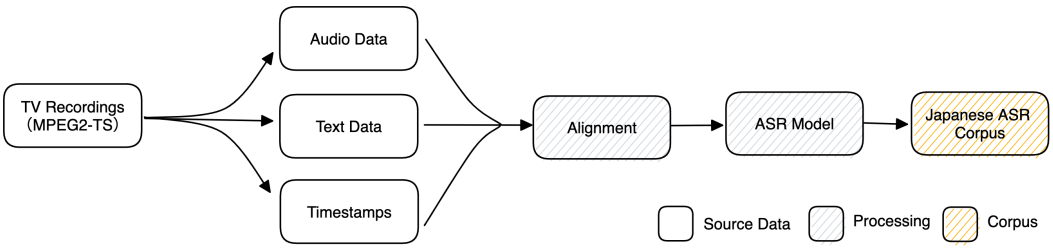

Figure 1 Corpus construction workflow

that has its transcription embedded in the video

3. transcriptions are inaccurate, especially when the speech is informal (eg. deletion of filler phrases)

Of the above limitations, we aim to address 1 and 2 by refining the alignment between text transcription and audio. We leave limitation 3 out of account because of its complexity and the amount of human effort required. The workflow of the construction is illustrated in Figure 1.

3.1 Pretrained ASR Models for Alignment

ASR based As suggested in [7, 8, 10], one can automatically generate a corpus using ASR-based alignment and cleansing by leveraging the divide-and-conquer. LaboroTVSpeech[8] and GigaSpeech[7] use the strategy proposed in [11] which consists of a pretrained acoustic model and a biased language model through Kaldi[12] interface. Inspired by this approach, we tried ASR-based alignments for the corpus. We broke down the audio into utterances using voice activity detection through pyannotate[13] and resegmented subtitles into complete sentences using GiNZA[14] through SpaCy framework[15]. We then transcribed the audio segments using an ASR model pretrained on LaboroTVSpeech[16] and matched the audio and transcription chunks based on the ASR results using a dynamic programming-based algorithm as illustrated in Figure 5 in the appendix. By listening to the extracted utterances, we noticed a lot of insertion/deletion at the beginning/end, suggesting that this approach is inadequate to resolve the missing transcription issue 2. Besides, certain words occur repeatedly throughout the entire TV show, making the matching error-prone. In our experience with this method, the quality of the constructed corpus and extraction efficiency largely depends on the accuracy and agreement of the intermediate processes, and this dependency unnecessarily aggregates errors and introduces overheads in computation.

CTC segmentation based Finding out that using ASR results overcomplicate the task and suffers from missing transcriptions drove us to rethink the possibility to leverage the inaccurate timestamps that have been ignored. JTubeSpeech[9] has suggested the viability of anchoring text in a longer audio segment for corpus creation on Japanese subtitled YouTube videos. The success of this work hinted at the possibility to reframe the task to anchor the text in a longer audio segment which is a superset of the actual utterance by utilizing the fuzzy timestamps in the subtitles. As suggested in this work of German corpus construction[17], there are several established methods for alignment:

- Acoustics based: Montreal Forced Aligner[18], Kaldi [12]based scripts, etc.
- Text-to-speech based: Aeneas[19]
- Connectionist Temporal Classification outputs based: CTC segmentation[17]

Amongst the above methods, we chose to experiment with CTC through ESPnet2[20] interface as we have access to a noise-tolerant pretrained model under the same-domain corpus LaboroTVSpeech[16]. Moreover, this method is independent of in-depth language-specific knowledge and compatible with our eventual E2E training goal. To test the compatibility of the CTC segmentation approach with our task, we used a model from [16] and added 25 seconds before the starting point of the subtitled timestamp as the inaccuracy usually happens at the beginning. The issues we noticed in ASR based approach are considerably relieved. The positive results confirmed the viability of this method on our task.

3.2 Bootstrapping Labeling for Alignment

Approach The success in using CTC segmentation [17] shows the promise of the approach leveraging robust pretrained ASR models for corpus creation. However, de-

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

pending on resources that have non-permissive licenses casts restrictions on the constructed corpus, not to mention that such language resource is a luxury that some languages might not have. Therefore we further seek the possibility of loosening the assumption of having an ASR model by leveraging public-domain small-scale data. Bootstrapping labeling methods are shown to be effective in various tasks where there is little data available. The process of bootstrapping labeling usually follows this procedure: 1. obtain a small amount of labeled data and train a model based on it 2. use the model to label more data 3. select high-quality labeled data to add to the training data 4. retrain the model 5. repeat 2-4 until satisfactory results are obtained.

Although this technique has been applied to different tasks including ASR [21, 22, 23], its efficacy on alignment has not been well-studied. We applied this approach to our task by starting with training an initial model based on Common Voice[2] dataset and grew the aligner iteratively following the procedure described in Figure 2:

1. train 0th-gen ASR model on Mozilla Common Voice[2] v8.0 (47.7 hrs)
2. use ASR model CTC alignment[17]
3. select high-quality aligned utterances
4. retrain ASR model
5. repeat 2-4 until extraction efficiency reaches a satisfactory level

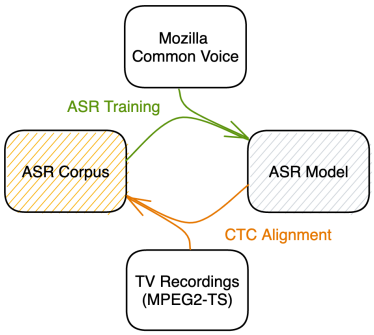

Figure 2 bootstrapping labeling flow

To speed up the development cycle, we orchestrated a pipeline where ASR training and aligning can be carried out in parallel. The paragraphs below describe how high-quality alignments are selected and how the extraction efficiency is measured to monitor the improvements. Figure 3 and 4 show the improvements in extraction efficiency and ASR performance for the models throughout the process.

Alignment quality measurement We listened to 200 utterances of variable lengths created by CTC segmentation using the ASR model pretrained on LaboroTVSpeech [16], and 79 utterances are deemed appropriate. Instead of thresholding CTC scores which is a byproduct of the alignment as suggested in JTubeSpeech [9], we use CER(ASR(audio), transcription) ≤ 0.33 as the threshold to select quality data as it has a higher level of agreement with human judgment: F1 score of 0.87 while using CTC score > -1 only has an F1 score of 0.42.

Extraction efficiency evaluation To further quantify the extraction efficiency, we define the metric as

$$\text{extraction rate} = \frac{\text{length of text in constructed corpus}}{\text{length of text in subtitles}}$$

Note that the extraction rate is an approximate metric as the potential of TV recordings as ASR corpus largely depends on the TV program genre. However, by calculating this metric using the same TV recordings, we are able to compare the extraction efficiency of different models.

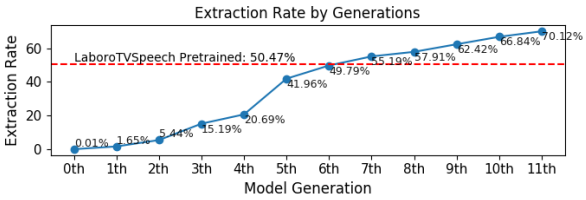

Figure 3 Extraction rate for every generation on one-day amount of unseen recordings

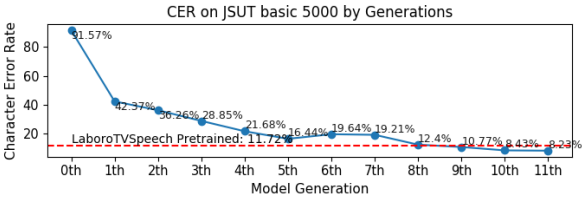

Figure 4 CER on JSUT basic5000 for every generation

4 Experimental Results

Corpus creation results We used the recordings in 2022.12.27 to compare the extraction rate for every generation of models. As shown in Figure 3, the extraction rate surpassed the LaboroTVSpeech pretrained model[16] starting from the 7th generation and reached 70.12% for the released model. The consistently improving extraction rate showed the effectiveness of bootstrapping labeling in our task, eventually allowing us to release the model and corpus freely available. It also revealed the possibility of

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

constructing large-scale corpora for more languages with few labeled resources.

ASR results We trained an E2E ASR model using a recipe provided by ESPnet2 [24] on ReazonSpeech and evaluated the system on the following benchmark datasets:

- JSUT[1] basic 5000: a high-quality dataset covers common words originally recorded for text-to-speech.
- Common Voice[2] v8.0 test set：4483 utterances with slightly noisier environments.

We used CER after removing punctuation marks and converting numbers to words using num2words[25] to measure the ASR performance. The model trained on ReazonSpeech achieves state-of-the-art performance in terms of CER, on par with the Whisper[3] model trained on 680,000 hours of multilingual data and 7043 hours of Japanese. The comparison on model sizes perspective is attached in the appendix. The competitive results we have on the benchmark datasets showed the usefulness of our corpus. Also by observing the improvements in ASR performance throughout the bootstrapping process, we confirmed our pipeline's capability of fostering continuous evolution.

Table 1 Comparison of Model CER (%)

	JSUT[1]	Common Voice[2]
Whisper small [3]	14.35	15.18
Whisper medium [3]	9.89	11.42
Whisper large-v2[3]	8.17	9.70
ESPnet LaboroTVSpeech[16]	11.72	12.56
ESPnet **ReazonSpeech**(15k hrs)	8.23	9.93

5 Released Corpus information

We used 1seg TV recordings from 2021.05.27 to 2022.12.25 to construct the initial release of ReazonSpeech which consists of 10,390,151 utterances (15,735 hours of speech and 227.6M characters in the transcriptions). The alignment of the corpus is done by the 10th-gen ASR created as the procedure described in 3.2(trained on 10,000 hrs, extraction rate 66.84%, with a threshold at CER(transcription, ASR(audio)) ≤ 0.33). The duration of utterance was capped at 14 seconds and the dictionary size at 2600 for training efficiency. The full corpus without duration or dictionary constraints contains 19k hours of audio. Details about the corpus and TV program genre can be found on our project page. To comply with the Copyright Act, we shuffled the data at the utterance level to prevent the reconstruction of the TV shows for purposes except for ASR researches. We will delete the relevant utterances upon request from the authors of the content.

Table 2 Comparison of Japanese ASR Corpora

	Duration(hrs)	Commercial	Creation
JSUT[1]	10	Free	Recorded
Common Voice v8.0[2]	48	Free	Recorded
CSJ[5]	600	Charged	Recorded
JTubeSpeech[9]	1300	Free	Automatic
LaboroTVSpeech[8]	2050	Negotiable	Automatic
ReazonSpeech	15735	Free	Automatic

Audio processing To enrich the variability in leading blanks and noises and avoid the abrupt starting of speech, we allowed more span in the beginning during audio segmentation to reduce the possibility of introducing systematic errors brought by CTC segmentation. To ensure no confounding speech gets prepended to the audio, the midpoint between the end of the previous segment and the start of the current segment is set to be the adjusted starting point, and the prepending duration is capped at 3 seconds.

Text processing We unified half-width and full-width alphanumerics and removed special characters from the transcriptions. Further processing is left to the judgment of researchers who use the corpus. The normalized transcriptions cover 2599 unique characters and 359,608 unique surface forms under tokenization by MeCab[26] with mecab-ipadic-NEologd[27] as the dictionary.

6 Conclusion

This paper introduces our efforts on the automatic construction of a freely available large-scale Japanese ASR corpus from TV recording fully from scratch without dependency on pretrained models. By confirming the strong ASR performance, we showed the effectiveness of our proposed method and corpus. By open-sourcing the project, we hope to lower the barrier of entry for E2E ASR training, allowing more individuals and organizations to participate.

7 Future Work

- We plan to maintain the project and release corpus updates periodically.
- We currently use CER based on ASR results for filtering. There might be more appropriate and computationally efficient metrics.
- Bootstrapping from pretrained multilingual models could make the learning curve less steep as they have a robust acoustic model.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Acknowledgement

We would like to express our gratitude to Taichi Kakinuma for providing legal guidance on the project. We are also grateful to the datasets and softwares we extensively used including but not limited to ESPnet[20] toolkit, recipes and models, Common Voice[2], LaboroTVSpeech[8].

References

[1] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. **arXiv preprint arXiv:1711.00354**, 2017.

[2] 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 **Proceedings of the Twelfth Language Resources and Evaluation Conference**, 2020.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. **arXiv preprint arXiv:2212.04356**, 2022.

[4] Jinyu Li. Recent advances in end-to-end automatic speech recognition. **arXiv preprint arXiv:2111.01690**, 2021.

[5] Kikuo Maekawa. Corpus of spontaneous japanese: Its design and evaluation. In **Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)**, 2003.

[6] Daniel Galvez, Greg Diamos, and Juan et al. Ciro. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. **arXiv preprint arXiv:2111.09344**, 2021.

[7] Guoguo Chen, Shuzhou Chai, and Guan-Bo Wang et al. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In **Proceedings of Interspeech 2021**, pp. 3670–3674, 2021.

[8] Shintaro Ando and Hiromasa Fujihara. Construction of a large-scale japanese asr corpus on tv recordings. In **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2021)**, 2021.

[9] Shinnosuke Takamichi, Ludwig Kurzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe. JTubeSpeech: 音声認識と話者照合のために YouTube から構築される日本語音声コーパス. 言語処理学会第 28 回年次大会, 2022.

[10] Jeong-Uk Bang, Mu-Yeol Choi, Sang-Hun Kim, and Oh-Wook Kwon. Automatic construction of a large-scale speech recognition database using multi-genre broadcast data with inaccurate subtitle timestamps. **IEICE Transactions on Information and Systems**, 2020.

[11] Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur. Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning. In **IEEE 2017 Workshop on Automatic Speech Recognition and Understanding (ASRU2017)**, 2017.

[12] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, and Burget et al. The kaldi speech recognition toolkit. In **IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU2011)**, December 2011.

[13] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, and Gelly et al. pyannote.audio: neural building blocks for speaker diarization. In **IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP2020)**, 2020.

[14] Mai Hiroshi and Masayuki. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会第 25 回年次大会, 2019.

[15] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[16] Shinji Watanabe. ESPnet2 pretrained model, Shinji Watanabe/laborotv_asr_train_asr_conformer2_latest33_raw_char_sp_valid.acc.ave, fs=16k, lang=jp, 2020. https://zenodo.org/record/4304245.

[17] Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. Ctc-segmentation of large corpora for german end-to-end speech recognition. In **Speech and Computer**, 2020.

[18] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In **Proceedings of Interspeech 2017**, 2017.

[19] Alberto Pettarin et al. Aeneas, 2017. https://www.readbeyond.it/aeneas/.

[20] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In **Proceedings of Interspeech 2018**, 2018.

[21] Daniel Waegel. A survey of bootstrapping techniques in natural language processing. 2013.

[22] Paul Albert, Diego Ortego, Eric Arazo, Noel E. O'Connor, and Kevin McGuinness. Relab: Reliable label bootstrapping for semi-supervised learning. **arXiv preprint arXiv:2007.11866**, 2020.

[23] Manuel Giollo, Deniz Gunceler, Yulan Liu, and Daniel Willett. Bootstrap an End-to-End ASR System by Multilingual Training, Transfer Learning, Text-to-Text Mapping and Synthetic Audio. In **Proceedings of Interspeech 2021**, 2021.

[24] Espnet/egs2/laborotv/asr1, 2021. https://github.com/espnet/espnet/tree/master/egs2/laborotv/asr1.

[25] Virgil Dupras et al. num2words library: Convert numbers to words in multiple languages, 2021. https://github.com/savoirfairelinux/num2words.

[26] Taku Kudo. Mecab : Yet another part-of-speech and morphological analyzer, 2006.

[27] Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会, 2017.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A Appendix

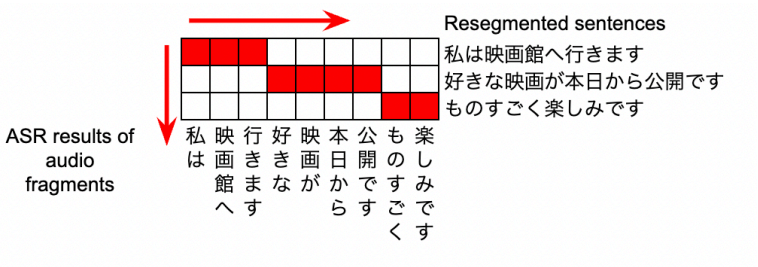

Figure 5 Example of Aligning using Dynamic Programming

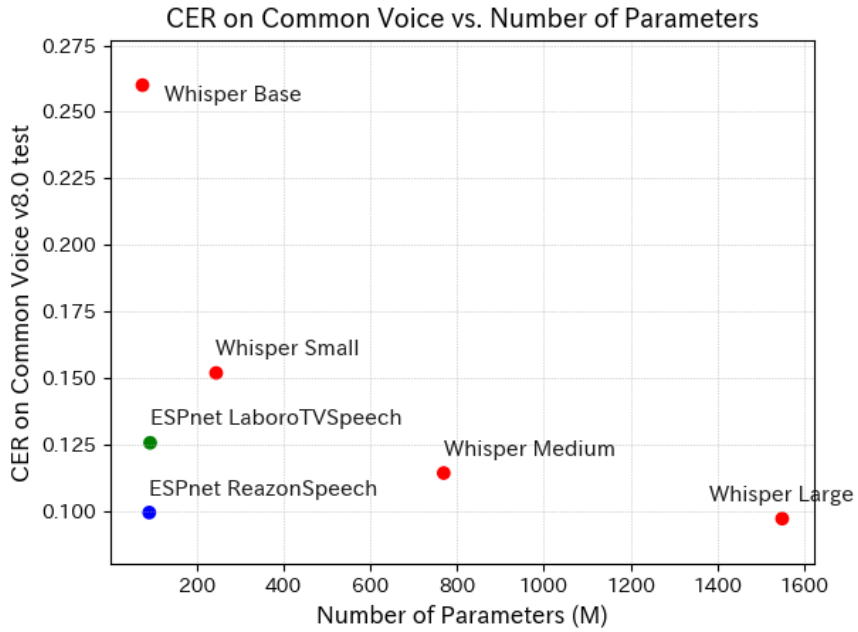

Figure 6 CER vs. Number of Model Parameters

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).