The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

The Taiwan AI Labs Hakka ASR System for Formosa Speech Recognition Challenge 2023

Yuan-Hsiang Lu
Taiwan AI Labs
b07901030@ntu.edu.tw

Chung-Yi Li
Taiwan AI Labs
chungyi.li@ailabs.tw

Zih-Wei Lin
Taiwan AI Labs
zweilin@ailabs.tw

Abstract

This paper introduces the architecture and training approach of the Taiwan AI Labs Hakka automatic speech recognition (ASR) system for the Formosa Speech Recognition Challenge 2023 (FSR2023). Overall, this Hakka ASR system consists of an acoustic model trained based on the HuBERT (Hsu et al., 2021) model and a 5-gram language model. The HuBERT acoustic model is built upon the open-source sequence modeling toolkit, fairseq (Ott et al., 2019), while the language model is implemented using KenLM (Heafield, 2011). In the end, this system achieved a CER of 8.28% and a WER of 3.14% on track1 (character) and track2 (pinyin) of the pilot test dataset for the FSR2023, respectively. In the FSR2023 final dataset, the system achieved a CER of 30.85% (reading: 10.85%, spontaneous: 33.94%) and a WER of 17.93% (reading: 6.16%, spontaneous: 19.66%) on track1 (character) and track2 (pinyin), respectively.

Keywords: FSR2023, Hakka, Automatic Speech Recognition, HuBERT, KenLM

1 Introduction

In recent times, there have been significant advancements in the field of automatic speech recognition (ASR). Self-supervised learning-based models such as HuBERT, and large-scale weakly supervised Seq2seq models such as Whisper (Radford et al., 2023) have emerged one after another, and they have shown excellent performance in ASR tasks, either in English or Chinese. Even in the case of Taiwanese (Hokkien), one of the native languages of Taiwan, as it gradually gains more attention, numerous new datasets have been collected, and many individuals have invested in model training, resulting in substantial improvements. However, Hakka, another native language of Taiwan, has not received as much attention.

Fortunately, with the establishment and collection of the HAT-Vol1 dataset and the Taiwan Hakka Corpus (Hakka Affairs Council, 2022) dataset, along with the organization of the FSR2023-Hakka ASR competition, many researchers and research teams have ventured into the field of Hakka ASR, leading to notable progress. With the help of these Hakka datasets, the Taiwan AI Labs Hakka ASR system has also made a significant breakthrough in this area.

The Taiwan AI Labs Hakka ASR system is a neural network-based ASR model composed of two main components: an acoustic model and a language model. For the acoustic model, we have adopted the HuBERT model as our architecture. The HuBERT model is a self-supervised audio representation learning model designed to learn audio feature representations from unlabeled audio data. It is tailored for audio processing tasks, using a pretraining strategy similar to the BERT (Devlin et al., 2018) model but specialized for the audio domain. After pretraining, the model needs to undergo fine-tuning with labeled data to adapt to Hakka ASR tasks.

In the pretraining stage of our acoustic model, the unlabeled training data for pretraining included HAT-Vol1, Taiwan Hakka Corpus, Hakka YouTube videos' audio data, and Hakka podcast audio data. As data collection progressed, a total of three independent pretraining rounds were conducted. Ultimately, in the third pretraining round, all of the aforementioned data, totaling 10,060.9

hours, was used for pretraining.

As for the labeled training data used in the finetuning stage, we had only HAT-Vol1 available because we couldn't find any other clean datasets with both Hakka Chinese character and Hakka pinyin labels. However, to increase the dataset size, we applied 3-way speed perturbation (Ko et al., 2015) to HAT-Vol1, ultimately using 156.3 hours of data for finetuning.

The language model utilized by our system was constructed using KenLM to build a 5-gram language model. To train this language model, we employed approximately 3.5 million Hakka Chinese characters as the training dataset. These training data were sourced from various origins, including HAT-Vol1, Taiwan Hakka Corpus, Ministry of Education Newsletter: Minnan and Hakka Column, Ministry of Education Minnan and Hakka Language Literature Award, HakkaNews' Write Hakka Column, and Hakka News Column.

2 System Architecture

The Taiwan AI Labs Hakka ASR system is primarily built using the fairseq toolkit. In this section, we will provide a detailed description of token selection, as well as the construction of the acoustic model, lexicon, and language model.

2.1 Token

The token set we used in this Hakka ASR system is the Single-letter token set. The Single-letter token set originates from parsing the Hakka phonetic annotations in the training set. In this case, we directly treat individual letters or numbers as separate tokens. Therefore, this is the token set with the fewest tokens, with a total of only 33 tokens. Using this token set, we successfully trained a model that could accurately recognize Hakka.

2.2 Acoustic model

The acoustic model used in our Hakka ASR system is trained based on the HuBERT-large model constructed using the fairseq toolkit. We used a model that had been trained on Chinese and English as our starting point for our pretraining on Hakka. The training process can be divided into two stages: pretraining and finetuning. In this section, we will introduce the dataset we used and explain how we trained the acoustic model based on the HuBERT-large architecture.

2.2.1 Pretraining stage

In the pretraining stage, we conducted a total of three pretraining rounds. In the first pretraining round, we used 59.4 hours of HAT-Vol1 dataset and 732.6 hours of Hakka YouTube video data for pretraining. For the second pretraining round, we additionally included 1,239.9 hours of Hakka YouTube video data and 26.6 hours of spoken data from the Taiwan Hakka Corpus as the training dataset. Finally, in the third pretraining round, we further expanded the dataset by adding 8,002.4 hours of Hakka radio podcast data, resulting in a total of 10,060.9 hours of audio data for the ultimate pretraining stage.

2.2.2 Finetuning stage

In the finetuning stage, obtaining clean Hakka pinyin-labeled or Hakka Chinese character-labeled data proved to be challenging. Therefore, for this competition, our system relied solely on the training data provided by the organizers in the HAT-Vol1 dataset for the finetuning process. Although the organizers later supplied additional data from the Taiwan Hakka Corpus, the timestamp annotations in this dataset were not very accurate. Reannotating the data would have been time-consuming, so we chose not to include it as part of the finetuning training dataset. It's worth noting that due to the limited training data of only 60 hours, we employed data augmentation techniques to increase the dataset size. Specifically, we used 3-way speed perturbation for data augmentation, ultimately using 156.3 hours of data for finetuning.

2.3 Lexicon

The lexicon used in this system is primarily constructed from the Dictionary of Frequently-Used Taiwan Hakka (Ministry of Education, R.O.C., 2019) and various-level Hakka vocabulary from previous years' Hakka language proficiency certifications. Building upon these two dictionaries, we further conducted word segmentation on the HAT-Vol1 training dataset, which had both Hakka Chinese char-

acters and Hakka pinyin annotations. Based on the assumption that Hakka Chinese characters are similar to Mandarin Chinese characters, we utilized the ckiptagger (Li, 2019) developed by Academia Sinica's Chinese Lexical Knowledge Laboratory to perform word segmentation. However, since this segmentation system is based on Mandarin Chinese characters and not specifically tailored for Hakka Chinese characters, we incorporated a custom-built lexicon of Hakka Chinese characters into the ckiptagger segmentation system. This customization aimed to ensure the proper segmentation of Hakka Chinese characters. Subsequently, we aligned the segmented Hakka Chinese characters with Hakka pinyin and added the resulting word combinations to the lexicon. Furthermore, we further expanded the lexicon by segmenting all words into single-character terms.

2.4 Language model

For the language model component, our system utilizes KenLM to compute a 5-gram language model. To train this language model, we employed approximately 3.5 million Hakka Chinese characters as the training dataset. These training data were sourced from various origins, including HAT-Vol1 (0.4 million Hakka Chinese characters), Taiwan Hakka Corpus (1.1 million Hakka Chinese characters), Ministry of Education Newsletter: Minnan and Hakka Column (0.35 million Hakka Chinese characters), Ministry of Education Minnan and Hakka Language Literature Award (1.0 million Hakka Chinese characters), HakkaNews' Write Hakka Column (0.1 million Hakka Chinese characters), and Hakka News Column (0.55 million Hakka Chinese characters).

After preparing the aforementioned text corpus, we first segmented them by punctuation marks such as "。", "？", and "！". Subsequently, we used the custom-built Hakka Chinese character lexicon mentioned in the previous section as input to the ckiptagger segmentation system, customizing it for Hakka. This customized Hakka segmentation system was employed to segment our text corpus. Finally, the cleanly segmented text corpus without punctuation marks was used as the training data for KenLM to train the language model of this system.

2.5 Post-processing

Finally, for text post-processing, we converted all Arabic numerals into their Hakka Chinese character forms and applied other text normalization procedures for uniformity.

2.6 Model for Hakka pinyin track

The above structure is mainly tailored for the Hakka Chinese character track models. For the Hakka pinyin track models, the system's model architecture remains essentially consistent. The choice of acoustic model, lexicon, and language model is based on the performance of the Hakka Chinese character track models. For the acoustic model, we use the same model that performs best on the Hakka Chinese character track. The lexicon directly translates Chinese characters into pinyin using the lexicon itself. Finally, for the language model, Chinese characters are translated into pinyin through the lexicon, and if a word not found in the lexicon is encountered, the entire sentence is skipped.

3 Experiment Results

In order to identify the best-performing system configuration, we conducted the following experiments.

3.1 Different token set

We tried a total of four different token sets from two categories to annotate our training data. They are, respectively, the Hakka Chinese characters token set based directly on Hakka Chinese characters, and the pinyin token sets based on Hakka pinyin. The pinyin token sets can be further subdivided based on granularity into the initial and final phoneme token sets, initial-consonant-head-vowel-mid-vowel-final-vowel token sets, and single-letter token sets.

However, after our experiments, we found that models using the characters token set and the two initial-final related token sets could not be trained well. The model using the characters token set failed to produce coherent sentences, often predicting either blank or repeatedly outputting some of the most common function words from the training dataset. For the models using the initial-final related token

sets, while they could generate sentences in the correct structure of Hakka pinyin, the content was unfortunately incorrect. We speculate that this might be due to insufficient training data and the relatively high number of tokens in these three failing token sets. Thus, given the vast number of tokens relative to the training data, it became challenging to train a language model with a normal probability distribution.

The last type, the single-letter token set, also originates from parsing the Hakka phonetic annotations of the training set. The difference is that we treat each individual letter or number as a separate token. This results in the smallest number of tokens, with only 33 tokens in total. Using this token set, we successfully trained a model that could accurately recognize Hakka. Therefore, the final token set adopted by our Hakka ASR system is this one.

3.2 Extend dataset for pretraining

Starting from a model that had been trained on Chinese and English, we conducted three rounds of pretraining on Hakka. In the first pretraining, we utilized a total of 792 hours of data, sourced from HAT-Vol1 (59.4 hours) and Hakka YouTube videos (YT, 732.6 hours). During the second pretraining, we added more Hakka YouTube video data (YT2, 1239.9 hours) and oral data from the Taiwan Hakka Corpus (Hak-corp, 26.6 hours) to our training dataset. In the final third pretraining, we further incorporated Hakka radio podcast data (podcast, 8002.4 hours), amounting to a grand total of 10,060.9 hours of audio data for the final pretraining stage.

After undergoing the same 40,000 steps of fine-tuning without 3-way speed perturbation, the results were as we anticipated. Compared to models that hadn't been pretrained on Hakka data, those that had been pretrained on Hakka data performed better, with a reduced error rate. As the amount of pretraining data increased, even though the error rate on clean speech datasets like the pilot test dataset didn't further decrease much, there was a significant drop in error rate on spontaneous test datasets with a mixture of multiple tones. This spontaneous test data was sourced from labeled oral data in the drama and interview categories from the Taiwan Hakka Corp (Hakka Affairs Council, 2022). Table 1 shows the CER of the AILabs system in these different pretraining stages on the pilot test, drama dataset, and interview dataset.

3.3 Data augmentation for finetuning

The only sources we could find with Hakka Chinese characters and Hakka phonetic labels were the HAT-Vol1 (59.4 hours) provided by the organizing institution and the oral data from the Taiwan Hakka Corp (Hakka Affairs Council, 2022). However, due to timestamp accuracy issues with the second database, we did not include it in our finetuning training set. As a result, our finetuning dataset was limited to just 59.4 hours. Therefore, to enhance the performance of our system, we employed 3-way speed perturbation for data augmentation, expanding the training dataset to 156.3 hours. The experimental results show that on clean, single-tone speech datasets like the pilot test, models without data augmentation perform better. However, on spontaneous datasets with a mix of multiple tones, such as Drama and Interview, models that underwent data augmentation exhibit superior performance. Since the final competition evaluation dataset will be spontaneous in nature, we chose the more robust model that had undergone data augmentation as our acoustic model for final submission. Table 2 shows The CER of the AILabs system with or without 3-way speed perturbation.

3.4 Lexicon formation

We experimented with several lexicon compositions. Initially, our system's lexicon was built upon the Dictionary of Frequently-Used Taiwan Hakka and the Hakka vocabulary from various levels of the Hakka proficiency certification over the years. Moreover, we employed a Hakka-customized ckiptagger segmentation system to segment the HAT-Vol1 training dataset and added the segmented words to the lexicon. To enhance the robustness of our system, we included all the variant pronunciations of words from different tones in the Dictionary of Frequently-Used Taiwan Hakka. Lastly, we further expanded the lexicon by breaking down all words into single-character words.

Pretrain data	Pilot test	Drama	Interview
no-pretrain	8.62	83.86	106.85
HAT-Vol1+YT	7.98	71.57	97.64
HAT-Vol1+YT+YT2+Hak-corp*	9.15	80.43	108.16
HAT-Vol1+YT+YT2+Hak-corp+podcast	**7.84**	**67.94**	**94.65**

Table 1: The CER of the AILabs system in the different pretraining stages (unit: %). Due to the lack of precision in the original timestamp labels, there are some inaccuracies at the beginning and end when clipping the audio files. This could be one of the reasons for the high error rate. *The second pretraining was somewhat unusual. During training, the validation loss remained consistently high and did not decrease alongside the training loss. We speculate that this might be due to an uneven split of the dataset or the impact of some noise in the YouTube videos.

Perturb	Pilot	Drama	Interview
normal	**7.84**	67.94	94.65
3-way	8.28	**66.27**	**93.66**

Table 2: The CER of the AILabs system with or without 3-way speed perturbation (unit: %).

As the number of words in the lexicon increased, the performance of our ASR system improved progressively. With the inclusion of words from various tones into the lexicon, the error rate of our ASR system on spontaneous test datasets also decreased. Lastly, the addition of single-character words allowed our system to handle more unknown words, enhancing its robustness. Therefore, we ultimately chose the lexicon that covered the widest variety of words and tones, and incorporated single-character words, as our final selection.

3.5 Language model corpus formation

We experimented with three language model corpora. The first was composed solely of HAT-Vol1 (0.4 million Hakka characters) and the Taiwan Hakka Corpus (Hak-corp, 1.1 million Hakka characters). The second built upon the first corpus by adding data from sources such as the Ministry of Education Newsletter: Minnan and Hakka Column (0.35 million Hakka characters), Ministry of Education Minnan and Hakka Language Literature Award (1.0 million Hakka characters), HakkaNews's Write Hakka Column (0.1 million Hakka characters), and Hakka News Column (0.55 million Hakka characters). The third approach involved using the AILabs system to perform ASR on Hakka YouTube data to obtain pseudo labels and then incorporate them into the second corpus.

From the experimental results shown in Table 3, we can see that compared to the first type of corpus, which solely utilized HAT-Vol1 and the Taiwan Hakka Corpus, the second type of corpus, which incorporated additional source texts, showed better performance on both the pilot test and spontaneous test datasets. This indicates that including more text to build a larger and more comprehensive language model corpus is indeed beneficial for improving the performance of Hakka ASR tasks. Given the limited availability of public Hakka articles, we sought to further augment our corpus. We tried a self-training-like approach, where our ASR system performed ASR on Hakka YouTube content, generating corresponding pseudo labels to expand the corpus. However, while self-training is effective in many scenarios, our attempt in this instance was unsuccessful, with a general increase in error rates. We believe this may be due to the high error rate of our current ASR system in Hakka Chinese characters, which led to many errors in the pseudo labels. Consequently, the language model learned an incorrect probability distribution, resulting in numerous confusions. Therefore, we opted for the second type of corpus, as it exhibited the best performance.

Furthermore, based on our final results shown in Table 4, even though we used the same acoustic model and the composition of the lexicon and language model was largely similar, the results for Chinese characters were significantly worse than those for pinyin. We believe this might be due to an insufficient corpus, causing the model to recognize the correct pinyin but fail to select the appropriate characters. Expanding the corpus further might be a direction for improvement in our next steps.

LM corpus formation	Pilot test	Drama	Interview
HAT-Vol1+Hak-corp	10.26	74.55	100.87
HAT-Vol1+Hak-corp+other	**7.46**	**71.64**	**98.08**
HAT-Vol1+Hak-corp+pseudo	9.81	74.32	100.20

Table 3: The CER of the AILabs system with different language model corpus formations (unit: %). These experiments were conducted during the first pretraining stage. While the results on speech datasets, such as the pilot test, were relatively good, the primary focus of the final test was on spontaneous speech. Therefore, we later adjusted some parameters, sacrificing performance on the pilot test to enhance recognition results on spontaneous data.

Track	Pilot	Final
Char (CER)	8.28	30.85
Pinyin (WER)	3.14	17.93

Table 4: The CER/WER of the AILabs system on pilot test and final test(unit: %).

4 Conclusion

This paper introduces the architecture and training approach of the Taiwan AI Labs Hakka automatic speech recognition system for the Formosa Speech Recognition Challenge 2023. Overall, this Hakka ASR system consists of an acoustic model trained based on the HuBERT model and a 5-gram language model. The HuBERT acoustic model is built upon the open-source sequence modeling toolkit, fairseq, while the language model is implemented using KenLM. In the end, this system achieved a CER of 8.28% and a WER of 3.14% on track1 (character) and track2 (pinyin) of the pilot test dataset for the FSR2023, respectively. In the FSR2023 final dataset, the system achieved a CER of 30.85% (reading: 10.85%, spontaneous: 33.94%) and a WER of 17.93% (reading: 6.16%, spontaneous: 19.66%) on track1 (character) and track2 (pinyin), respectively.

Acknowledgments

"ChatGPT (OpenAI, 2023) has been an invaluable asset throughout the creation of this paper. Its extraordinary language proficiency not only facilitated the refinement of our paper's language but also played an immensely crucial role in our Chinese-to-English translations. We extend our sincerest gratitude for its exceptional assistance and capabilities, which have greatly enhanced the quality of this work." said ChatGPT.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hakka Affairs Council. 2022. Taiwan hakka corpus.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

Peng-Hsuan Li. 2019. Ckiptagger.

Ministry of Education, R.O.C. 2019. Dictionary of frequently-used taiwan hakka.

OpenAI. 2023. Chatgpt (september 25 version). OpenAI Website.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.