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

WhisperHakka: A Hybrid Architecture Speech Recognition System for Low-Resource Taiwanese Hakka

Ming-Hsiu Chiang **Chien-Hung Lai** **Hsuan-Sheng Chiu**

Advanced Technology Laboratory,
Telecommunication Laboratories,
Chunghwa Telecom Co., Ltd.,
Taiwan
{kurt, jhlai, samhschiu}@cht.com.tw

Abstract

Deep learning-based automatic speech recognition (ASR) design has been growing in popularity. Besides the ASR model depends on the reliable speech representation offered by the self-supervised learning (SSL) model, Whisper is also a powerful model that makes use of the knowledge from large-scale labeled datasets and the self-attention mechanism. However, its inherent training method decreases the potential to expand on low-resource language because of the difficulty of acquiring labeled data. As a result, we integrated both the Wav2vec2 and Whisper into an ASR system not only to provide extra abundant information on features but also to have the ability to train on unlabeled data through the SSL model while retaining the capacity of Whisper. Experimental Results show that the proposed hybrid architecture system outperforms the vanilla Whisper in the reading speech scenario, achieving a roughly 21% improvement in recognition rate.

Keywords: ASR, Whisper, Wav2vec2, Taiwanese Hakka, Self-Supervised Learning

1 Introduction

With the development of technology in full swing, traditional interaction interfaces have gradually progressed from keyboards, mice, and touchscreens to a new generation of interaction interfaces based on natural conversation. Both businesses and individuals are attempting to use such technology to enhance the convenience and comfort of people's lives. Speech recognition is unquestionably one of the most important technologies used in the current process of natural conversation interaction. Typically, the system needs to transcribe the user's speech into text before applying natural language processing to the text. Due to the significance of ASR, the technology's advancement over the past ten years has been extremely rapid. In particular, the introduction of deep learning (DL) has resulted in a significant decrease of more than 50% in the word error rate in common languages (Prabhavalkar et al., 2023).

Sadly, despite the advances in technology, the effectiveness of speech recognition systems is still constrained by the sheer amount and diversity of training data. As a result, only roughly 100 languages are currently covered by ASR models, compared to the more than 7,000 known languages in the world. Specifically, even Hakka, which is spoken by the third-largest population in Taiwan after Chinese and Hokkien, is nevertheless listed by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as a severely endangered level. Language is not just a means of communication, it also plays a significant role in preserving cultural legacy. Therefore, it can be claimed that language is a valuable resource for people, which is why it is crucial to protect endangered languages. From a digital standpoint, it is believed that AI technology can benefit these endangered languages through speech synthesis, which can turn text into conversational speech, and speech recognition, which can extract text from speech to aid understanding. Excitingly, Meta's Massively Multilingual Speech (MMS) (Pratap et al., 2023) project has successfully recognized 1107 languages with varied degrees of recognition rate.

Traditional speech recognition systems are mostly based on a classic architecture composed of four primary components, includ-

ing an acoustic feature extractor, acoustic model, language model, and search based on the Bayes decision rule (Jelinek, 1998). Mel-frequency cepstral coefficients (MFCC) and filter bank (FBank) are common acoustic feature extractors that are in charge of extracting acoustic features from speech signals. Moreover, the hidden Markov model (HMM) (Rabiner, 1989) and its variations are the most widely applied acoustic models, which are used to infer the corresponding phonemes from the audio features of each frame. The N-gram model (Goodman, 2001) is the most frequently employed language model, used to assess the probability of the input text sequences and select the text sequence with the highest overall probability as the final output result.

Later, as DL technology flourished, DL models increasingly took the place of both the HMM model as well as N-gram, which are used in acoustic and language modeling separately (Bourlard and Morgan, 1994) (Seide et al., 2011) (Nakamura and Shikano, 1988) (Bengio et al., 2000) (Schwenk and Gauvain, 2002), respectively. Furthermore, the traditional statistical architectures are then replaced by a variety of end-to-end designs, including the Connectionist Temporal Classification (CTC) (Graves et al., 2006), Recurrent Neural Network Transducer (RNN-T) (Graves, 2012), Listen Attend and Spell (LAS) (Chan et al., 2016), and Conformer (Gulati et al., 2020), etc. To deal with the issue of varying input and output lengths, CTC is proposed as an alternative to the lattice free maximal mutual information (LFMMI) (Hannun et al., 2014) used in traditional acoustic models. CTC categorizes each input audio sequence and also introduces the concept of blank as a space unit. These characteristics enable CTC to combine many units of the same classification result and remove blanks to obtain results that are the same length as the output. Although CTC resolves the issue of alignment among audio and text sequences, the dependence between the outputs is unable to be modeled because each frame is an independent output. By constructing a prediction network to create the dependencies between the sequences, RNN-T improves on CTC by jointly training it with the output of the original encoder to increase the performance even more. Afterward, a prototype of the attention mechanism is put forth as the NLP field developed (Bahdanau et al., 2014). To enhance the performance of ASR, LAS adopted the attention mechanism to perform effective alignment, which is unlike CTC. Moreover, the mechanism is also able to take into consideration all of the contextual information to produce superior outcomes. As for the Conformer, it combines the Transformer (Vaswani et al., 2017) and convolutional layer, strengthening the capacity to extract local features while preserving the ability to capture lengthy sequential dependencies.

In addition, pre-training models based on SSL have also been designed and applied to speech tasks. With its self-supervised training method, a robust pre-trained model can be trained on large-scale datasets consisting of relatively simple-to-obtain unlabeled data, then use the pre-trained model as an acoustic feature extractor, fine-tuning with a small amount of labeled data to get favorable outcomes. Thus, these kinds of models are also called speech foundation models (Latif et al., 2023). Some of the well-liked pre-training models applied for speech recognition include Wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and BEST-RQ (Chiu et al., 2022). Wav2vec2 uses diversity loss and contrastive loss during training. Contrastive loss measures how similar the output is to the quantized vector, whereas diversity loss reflects the diversity of the quantized vector itself. With regard to the HuBERT and BEST-RQ, they will use different methods to obtain a codebook before attempting to reduce the distance between the output category and the corresponding category in the codebook. Currently, both the MMS and the Universal Speech Model (UMS) (Zhang et al., 2023b) that utilize pre-trained models as acoustic feature extractors reach SOTA performance in ASR, while the former applies Wav2vec2 and the latter uses BEST-RQ. In contrast to MMS and UMS, Whisper (Radford et al., 2023) directly employs around 680k hours of enormous amounts of labeled data for training on Transformer is also achieves SOTA outcomes. Moreover, many recent studies have sought to adopt multi-modal data as the input of the model,

which may contain many types of data including audio, text, and images, in order to further explore the possibilities of the model. For now, there are different methods of integration existing for multimodal models. In the ASR field, multimodal models UMS, AudioPaLM (Rubenstein et al., 2023), and SeamlessM4T (Barrault et al., 2023) have all attained or are very near the SOTA.

The main purpose of this paper is to build a reliable ASR system for the low-resource language Taiwanese Hakka with two kinds of transformer-based models. Both Whisper and Wav2vec2 are adopted in the system as a backbone and an extra feature extractor. With this design, a more comprehensive information of input speech is considered, while retaining the capacity of Whisper, leading to significant improvements compared to the baseline. To the best of our knowledge, our research is the first to integrate the Whisper and the SSL model, which is the paper's main contribution.

2 System Implementation

In this section, the methods of data augmentation will be introduced first and then the overall ASR system will be described, including the extra feature extractor and the system backbone.

2.1 Data Augmentation

Since Taiwanese Hakka is a low-resource language, there isn't much-labeled data to train on. In addition to the roughly 82.5 hours of labeled training data supplied by relevant organizers, an additional 64 hours of labeled training data are used. Although extra datasets are employed, bringing the total number of training datasets hours to 146.5 hours, the amount of expanded training datasets remains relatively small in contrast to the datasets of common languages that perform well in ASR systems. Therefore, a data augmentation on our datasets was performed. We picked the time stretch approach to accomplish data augmentation because of the aim to have more speech data with diverse speech characteristics of different speakers. Furthermore, the noise data from the MUSAN (Snyder et al., 2015) corpus were also added in the final training round to improve the robustness of the ASR system. The stretching coefficients α are set to $0.9(\pm 0.05)$ and $1.1(\pm 0.05)$, respectively. After completing the process, the total hours of data grew by an extra two times, from 146.5 hours to 439.5 hours. With regard to the approach of adding noise data within the final training round, the probability of the data being chosen to mix with noise data is set at 0.5 with the signal-to-noise ratio (SNR) ranging from 3 to 8.

2.2 Feature Extractor

SSL Models, as the foundation model of speech, not only aid in improving different speech tasks to variable degrees but also exhibit the ability to reach SOTA performance in ASR tasks with only a small amount of labeled data. It is clear that SSL Models can learn from a vast amount of unlabeled data, extracting speech representations that are crucial for downstream tasks effectively. Among the available open-source SSL pre-training models, the Wav2vec2 model used in MMS is pre-trained using about 500k hours of the corpus in approximately 1,400 languages, allowing it to accommodate low-resource languages. Not only that, but Hakka is one of these 1,400 languages, as a result, we employed the model to offer a speech representation containing significant information within speech.

2.3 ASR Backbone

Whisper is employed as the backbone of the ASR system due to its tremendous performance. The model demonstrates that even without utilizing a pre-trained model, as long as there is enough labeled data, its performance can reach the SOTA. Although Whisper performs admirably, its training strategy of using labeled data may make it challenging to expand to low-resource languages. Therefore, we merge these two types of models in our work. On the one hand, the speech representation from the SSL model can be leveraged to get more extensive speech features, at the same time, the flexibility to use unlabeled data also makes it easier to apply to low-source languages. On the other hand, it is able to sustain the powerful performance of Whisper and the comprehension of 680k labeled data. The overall architecture of the ASR system is shown in Fig. 1. The original Whisper performs feature

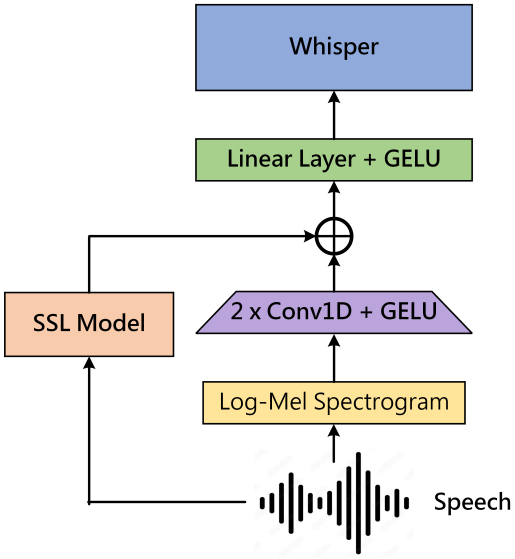

Figure 1: The proposed hybrid architecture. Both the output of the SSL model and the downsampling via convolution layers are considered. The linear layer then down the embedding dimension back to the original dimension. Lastly, the new features are employed by the model to train with cross-entropy loss.

extraction with MFCC and then downsamples through the convolution layer to obtain the final features. According to Fig. 1, before feeding the original features into the encoder, we concatenate the features with the speech representation output by Wav2vec2, passing them through a linear layer to extract key information. Finally, the new features are delivered to the model for training.

3 Experimental Results

3.1 Datasets

In our work, we used roughly 59.5 hours of the training datasets in HAT-Vol1 and approximately 23 hours of training data provided by the Hakka Affairs Council. In addition, TLHAK datasets comprising around 64 hours of recordings speech data and internet speech data collected by ourselves are utilized. Moreover, the evaluation datasets in HAT-Vol1 that are composed of reading speech data are also employed in final training, which contains around 10 hours. As for the FSR-2023-Finals evaluation datasets which consist of both reading and spontaneous speech data, we take it as the final evaluation to assess the performance of the ASR system. Each dataset contains two formats of text, Taiwanese Hakka characters and Taiwanese Hakka pinyin, which are used to train and evaluate the corresponding model. Table 1 shows the specific contents of each dataset.

Datasets	Hours	#Sents
HAT-Vol1(Train)	59.5	20611
HAT-Vol1(Eval)	10	3598
FSR-2023-Finals	17	5913
Hakka Affairs Council	23	9688
TLHAK	64	48409

Table 1: Details of Datasets.

Models	BS	LR	Epoch
Whisper Medium	7	6e-06	9
Whisper Large v2	3	4e-06	9

Table 2: Base Training Setting, in which BS indicates batch size, LR denotes learning rate.

3.2 Training Settings

In our trial, two distinct versions of Whisper, medium and large v2 are used for fine-tuning training with the SSL model wav2vec2. The base settings of the two sizes of models during fine-tuning can be found in Table 2. As for the optimizer, both models employ the same AdamW (Loshchilov and Hutter, 2019). Furthermore, as a baseline for our comparison, we adopt a vanilla Whisper medium model which is fine-tuned by the AdaLoRA (Zhang et al., 2023a) approach rather than directly updating the parameters of the whole model to achieve a superior result.

3.3 Evaluation

In order to properly evaluate whether the addition of the SSL Model is capable of boosting the entire system performance, we compared the proposed models with the benchmark. For different forms of ground truth, the character error rate (CER) is used to evaluate Taiwanese Hakka characters, while the syllable error rate (SER) is used to evaluate Taiwanese Hakka pinyin. Moreover, both the out-domain evaluation datasets HAT-Vol1(Eval) and FSR-2023-Finals datasets are used to evaluate during the experiment. Table 3 and Table 4 shows the comparison results respectively. The results in Table 3 illustrate the usage of SSL models enhances the overall ASR system significantly

Models	TS	AN	Char	Pinyin
Baseline	x	x	9.73	8.61
Medium	x	x	7.64	-
Large v2	x	x	6.68	4.95

Table 3: The evaluation results of Taiwanese Hakka characters with CER and Taiwanese Hakka pinyin with SER on HAT-Vol1 (Eval) datasets, where TS represents time stretch and AN denotes adding noise. Note that the results of the models used HAT-Vol1 (Eval) as training datasets aren't listed in the table.

Models	TS	AN	Char	Pinyin
Baseline	x	x	27.53	**30.35**
Medium	x	x	23.97	-
Large v2	x	x	23.60	49.11
Large v2	v	x	23.58	39.77
Large v2	v	v	**21.19**	39.13

Table 4: The evaluation results of Taiwanese Hakka characters with CER and Taiwanese Hakka pinyin with SER on FSR-2023-Finals datasets in which TS indicates time stretch and AN denotes adding noise.

in reading speech scenarios. There is about a 21% improvement between the baseline and Whisper medium which combines SSL models. Furthermore, employing the Whisper large v2 instead of the Whisper medium yields a 12% improvement.

Unlike the former Table 3, Table 4 doesn't demonstrate a similar tendency in the results of Taiwanese Hakka pinyin when the evaluation data not only includes reading speech but also comprises more spontaneous speech data. During our experiment, it appears that the Whisper large v2 with SSL model for Taiwanese Hakka pinyin is unable to work successfully in a more complex scenario. In contrast, the model for Taiwanese Hakka characters is still capable of operating in difficult settings, yielding a 14% improvement over the baseline.

The effects of data augmentation that we perform were also evaluated on the Whisper large v2 with SSL model, the results can be found in Table 4. Based on the results, a 1.6% to 19% enhancement can be achieved after applying the data augmentation depending on different methods of data augmentation and scenarios.

4 Conclusion

In this work, we propose a hybrid architecture system for ASR, which includes the Wav2vec2 as an extra feature extractor and the Whisper as a backbone. The proposed ASR system is able to get more comprehensive features by effectively capturing both the information provided by the feature extractor and the original path as much as possible. In accordance with the experimental results, the proposed system performs better in the low-resource language Taiwanese Hakka no matter on characters or pinyin than the baseline in reading speech scenarios. For spontaneous speaking scenarios, our vanilla Whisper medium model and Whisper large v2 with SSL model are able to achieve error rates of 21.19% and 30.35% for characters and pinyin respectively. In summary, our work shows the combination of Whisper and Wav2vec2 has the potential to promote the capability of recognition in most cases. In future works, we will continue to explore and enhance our proposed model, particularly for the model of Taiwanese Hakka pinyin in complex scenarios, by additional data collecting, various methods of data augmentation, and further pre-training of Wav2vec2.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, pages 1–15.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, pages 1–102.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13, pages 932–938. MIT Press.

Herve A Bourlard and Nelson Morgan. 1994. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE Intl. Conf. on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. pages 3915–3924. PMLR.

Joshua T Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, pages 1–9.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of the 23rd Intl. Conf. on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Awni Y Hannun, Andrew L Maas, Daniel Jurafsky, and Andrew Y Ng. 2014. First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns. *arXiv preprint arXiv:1408.2873*, pages 1–7.

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 Trans. on Audio, Speech, and Language Processing*, 29:3451–3460.

Frederick Jelinek. 1998. *Statistical methods for speech recognition*. MIT press.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*, pages 1–34.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Intl. Conf. on Learning Representations*, pages 1–18.

Masami Nakamura and Kiyohiro Shikano. 1988. A study of english word category prediction based on neural networks. *The Journal of the Acoustical Society of America*, 84(S1):S60–S61.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schluter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *arXiv*, abs/2303.03329:1–27.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, pages 1–41.

L.R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286.

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

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, pages 1–27.

Holger Schwenk and Jean-Luc Gauvain. 2002. Connectionist language modeling for large vocabulary continuous speech recognition. In *2002 IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages I–765–I–768.

Frank Seide, Gang Li, and Dong Yu. 2011. Conversational speech transcription using context-dependent deep neural networks. In *Proc. Interspeech 2011*, pages 437–440.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, pages 1–4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–11. Curran Associates, Inc.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh Intl. Conf. on Learning Representations*, pages 1–17.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023b. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, pages 1–20.