Speech Recognition for Indigenous Language Using Self-Supervised Learning and Natural Language Processing

Satoshi Tamura, Tomohiro Hattori, Yusuke Kato and Naoki Noguchi
Gifu University, 1-1 Yanagido, Gifu, Japan

Keywords: Speech Recognition, Self-Supervised Learning, Neural Machine Translation, Transformer, HuBERT, Indigenous Language, Under-Resourced Language.

Abstract: This paper proposes a new concept to build a speech recognition system for an indigenous under-resourced language, by using another speech recognizer for a major language as well as neural machine translation and text autoencoder. Developing the recognizer for minor languages suffers from the lack of training speech data. Our method uses natural language processing techniques and text data, to compensate the lack of speech data. We focus on the model based on self-supervised learning, and utilize its sub-module as a feature extractor. We develop the recognizer sub-module for indigenous languages by making translation and autoencoder models. We conduct evaluation experiments for every systems and our paradigm. It is consequently found that our scheme can build the recognizer successfully, and improve the performance compared to the past works.

1 INTRODUCTION

Automatic Speech Recognition (ASR) is a technique to transcribe human speech. In recent years, speech recognition technology has been widely used in a lot of environments, improving work efficiency and enhancing quality of life. Voice assistance and voice input are often employed on smartphones and smart speakers. ASR is also used to take minutes in online meetings and on-site conferences. Speech translation based on ASR is expected to make our international communication richer and easier. ASR has made remarkable progress for this decade, with the rapid development of Deep Learning (DL). Many researchers have attempted to build DL models, resulting in significant improvements in recognition accuracy. Several languages with large populations, such as English, Mandarin, and Spanish, now have high-performance ASR technology, as DL requires huge data sets, and it is relatively easier to do so. Developers are also interested in having such the techniques for languages spoken in emerging markets, such as Hindi, Bahasa Indonesia, and so on.

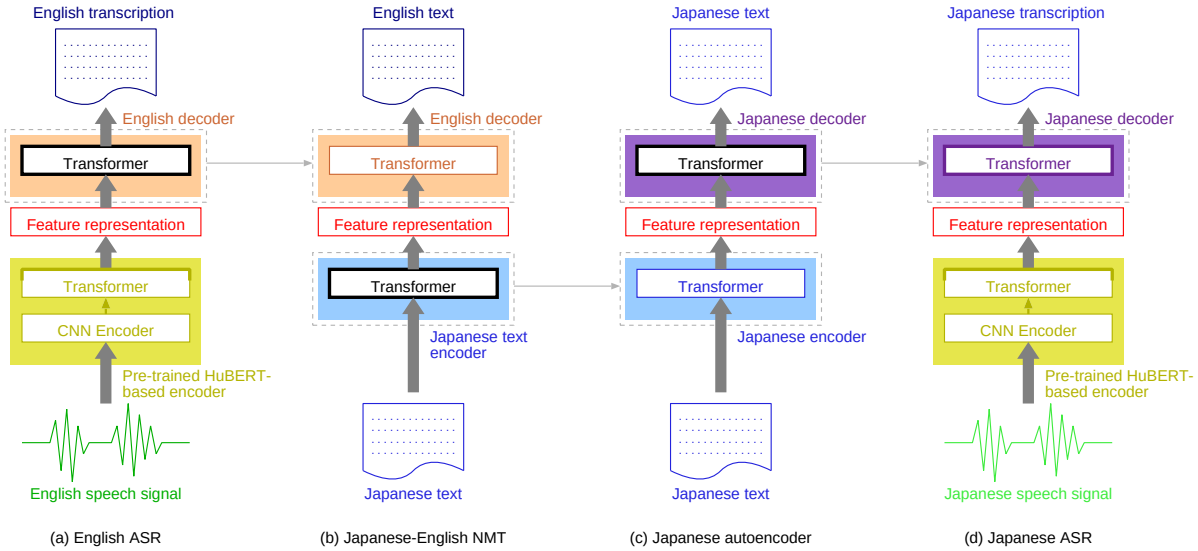
It is known that there are more than 7,000 languages on this planet. However, only a few languages have been well studied for DL-based ASR, while most indigenous languages having small populations used in limited areas have not yet, due to the lack of spoken data. To discuss this issue, UNESCO, ELRA (European Language Resource Association) and several societies jointly organized a conference on Language Technologies for All (LT4All) in 2019 (UNESCO, 2019). We need to strongly encourage researchers and developers to build ASR systems for these minor languages. And in order to do so, we should develop and improve DL techniques in under-resourced conditions.

Recently, Self-Supervised Learning (SSL) has been focused on in the DL field. SSL allows us to utilize unlabeled data or low-quality data, and to obtain effective feature representation for subsequent tasks. In ASR research works, several SSL techniques such as HuBERT (Hsu et al., 2021) and wav2vec (Baevski et al., 2020) are often employed. Focusing on the HuBERT model, we have developed a speech recognition scheme for under-resourced languages (Hattori and Tamura, 2023). In our previous work, we firstly prepared an English ASR system consisting of a pre-trained HuBERT and a shallow DL model for recognition. We secondly applied fine-tuning to the system using Japanese speech data, to obtain the recognizer for Japanese.

This paper proposes a new paradigm of ASR for indigenous languages, enhancing the recognition performance. In our work, we employ several Natural Language Processing (NLP) techniques, such as Neural Machine Translation (NMT) and text autoencoder. We assume that the HuBERT-based English

Tamura, S., Hattori, T., Kato, Y. and Noguchi, N.
Speech Recognition for Indigenous Language Using Self-Supervised Learning and Natural Language Processing.
DOI: 10.5220/0012396300003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 779-784
ISBN: 978-989-758-684-2; ISSN: 2184-4313



(Japanese as an indigenous language and English as a major language. Bold indicates model training or fine-tuning)

Figure 1: Our proposed concept to build an ASR method for an indigenous language.

ASR can be divided into two sub-modules: a feature extraction module based on HuBERT and a recognition module. By using the recognition module as an English decoder and adding a new Japanese encoder, we build a Japanese-English NMT system. Subsequently, we choose the Japanese encoder and prepare a new Japanese decoder to make a text autoencoder. Finally, we can develop a Japanese ASR system by employing the HuBERT feature extractor and the Japanese decoder. The novelty of this work is that, we can build an ASR system using less or ideally no speech data of a target indigenous language, thanks to state-of-the-art SSL and NLP technology. In addition, using NLP techniques is expected to enhance the semantic feature representation, improving the ASR performance. As far as the authors know, there is no other work regarding indigenous ASR incorporating HuBERT, NMT and text autoencoder.

The rest of this paper is organized as follows. Section 2 explains our concept in detail. Experiments are reported in Section 3. Section 4 concludes this work.

2 METHODOLOGY

Figure 1 illustrates our proposed scheme to build an ASR sysem for an indigenous language. In this paper, we use Japanese as an indigenous minor language and English as a major language; though Japanese has a large population, the reason why Japanese is chosen in this work is that, we can compare our scheme with conventional DL-based ASR methods as large-size Japanese corpora are available, and the authors can easily conduct subjective evaluation to the results.

Table 1: Training and fine-tuning settings.

	(a) E. ASR	(b) J-E NMT	(c) J. AE	(d) J. ASR
# epochs	40	50	50	40
Batch size	8	256	256	8
Optimizer	RAdam	Adam	Adam	RAdam
Learning rate	1e-4			
Loss function	Cross entropy loss			

(a) First of all, a DL-based SSL model is prepared, that is trained using a number of speech data of the major language. The model is used as a feature extractor from speech data. We employ a pre-trained English HuBERT model in this work. We then build an English ASR system using the model followed by a transformer-based decoder. We believe that a feature vector extracted by the first part includes contextual or semantic information, and the second model can be performed as a speech recognizer to generate English sentences.

(b) Suppose a text encoder for an indigenous language, in this case Japanese. Using the encoder and the decoder introduced above, we build an NMT system from the indigenous language (Japanese) to the major language (English). When training the system, all the parameters in the English decoder are fixed, while the model parameters in the Japanese encoder are optimized using Japanese-English parallel text data.

Table 2: Examples of English ASR results in Figure 1 (a). Underlined words mean recognition errors.

Reference	Recognized characters
THE WHOLE CAMP WAS COLLECTED BEFORE A ROT CABIN ON THE OUTER EDGE OF THE CLEARING	THE WHOLE CAMP WAS COLLECTED BEFORE A RUDE CABIN ON THE OUTER EDGE OF THE CLEARING
INTO THE LAND BEYOND THE SYRIAN DESERT BUT EITHER OF THEM DREAMED THAT THE SCATTERED AND DISUNITED TRIBES OF ARABIA WOULD EVER COMBINE OR BECOME A SERIOUS DANGER	INTO THE LAND BEYOND THE SYRIAN DESERT BUT NEITHER OF THEM DREAMED THAT THE SCATTERED AND DISUNITED TRIBES OF ARABIA WOULD EVER COMBINE OR BECOME A SERIOUS DANGER

(c) We further introduce a new text decoder for the indigenous language. Next, we make an text autoencoder for the minor language, consisting of the above encoder as well as the decoder. We apply model training using the text data written in the indigenous language, only to the decoder.

(d) Finally, we use the HuBERT encoder as a feature extractor and the text decoder for the indigenous language, to build an ASR system for the minor language. It is said that English phonemes fully cover Japanese vowels and consonants, therefore, the English feature extractor is expected to also work for Japanese speech data as well. Note that in this paper, to improve the performance we apply fine-tuning not only to the decoder but also to a part of the encoder.

The advantage of this scheme is that, ideally we do not need any speech data for the indigenous language, or only a few data may be significant for fine-tuning to finalize the ASR model. As mentioned above, it is hard to collect speech data for such a minor language with a small population. In our scheme we utilize a pre-trained SSL-based feature extractor, that is originally built for different languages, because a human speech production system is language independent . Furthermore, for indigenous languages, it is relatively easier to collect text data than speech data; we can obtain text data from official government documents, textbooks, news sites and internet articles such as Wikipedia.

3 EXPERIMENT

We conducted experiments to evaluate the effectiveness of our proposed approach. First, we report preliminary experimental results about training data size for an indigenous language. Second, we examine our NMT and autoencoder performance to check Japanese encoder and decoder. Finally, we evaluate our Japanese ASR. Table 1 shows model training and fine-tuning settings in the following experiments.

3.1 Preliminary Experiments

3.1.1 Machine Translation

In our previous work, we investigated the influence of training data size and model complexity in NMT. We used the MultiUN and Wikipedia data sets provided in OPUS (Tiedemann, 2012), in order to obtain parallel sentences. We then chose German-English sentence pairs as a training data set. Though German has a large population, in this experiment German is treated as an indigenous language, while English was a major language. We employed a pre-trained NMT model provided by OpenNMT (Klein et al., 2017), that was based on a tiny transformer; the encoder and decoder had six layers respectively. A transformer model was then explored with different settings, such as the number of layers in the encoder and decoder parts, and the number of training sentences.

It turns out that, with the small data set, we can build an NMT model, which achieves roughly the same performance as the pre-trained model, by adjusting the hyperparameters; we should make an encoder for indigenous language smaller to maintain the translation performance, while the decoder should still be large because it directly affects the performance. It is well known that the larger the training data set becomes, the better NMT performance is. On the other hand, it is sometimes hard to obtain larger data sets. According to our preliminary results, in this work we decided to use 10,000 sentences in the following experiments, which is quite small compared to the data set used in existing works.

3.1.2 English Speech Recognition

Next, we tested an English ASR shown in Figure 1 (a). We adopted an English HuBERT model provided by Facebook, which was trained using 960-hour spoken data from Librispeech (Panayotov et al., 2015). The transformer consisted of a CNN encoder and a 12-layer transformer. As an English text decoder, we employed a two-layer transformer, each having 12 attention heads. When building the ASR system, in the encoder transformer, we fixed the eight layers on the

Table 3: Examples of input, reference and translated sentences in Figure 1 (b).

Input Japanese text	Reference text	Translated English text
かれにあいたいの	i — w a n t — t o — s e e — h i m	i — — — — — — m e e t — h i m
わたしのむすこがちょうど ぴあのをはじめたんです	m y — s o n — s t a r t e d — p l a y i n g — p i a n o	m y — s o n — — — — — — — — — — — p i a n o
よかったらつかってください	p l e a s e — u s e — i t	u s e — — — — — — — — — — — —

† '—' indicates space.

Table 4: Examples of Japanese autoencoder results in Figure 1 (c).

Reference	Reconstructed characters
かれにあいたいの	かれはあううすす
わたしのむすこがちょうどぴあのをはじめたんです	わたしののののぴあのはちかかがはじめたです
よかったらつかってください	よかかかつかううういい

input side, while fine-tuning the remaining four layers on the output side. The decoder was trained from scratch. We used the Librispeech test-clean-100 data set to re-train the model, and a one-hour subset of LibriLightLimited (Liu et al., 2019) for evaluation.

Table 2 shows recognition result examples. As a result, the English ASR achieved a word error rate of 11.04%. We finally confirmed that the feature encoder and the English text decoder were properly prepared for the following experiments.

3.2 Experiments in NLP

3.2.1 Japanese-English Machine Translation

First, we checked our machine translation model, depicted in Figure 1 (b). We utilized a JESC corpus (Pryzant et al., 2018), consisting of Japanese-English subtitle pairs. From the corpus, we randomly selected 10,000 pairs for model training, 1,000 for validation, and 1,000 for evaluation. As explained in Section 2, only the Japanese encoder was trained, while the decoder was derived from the English ASR system. The architecture of the Japanese encoder was the same as that of the English decoder. Note that our Japanese encoder only accepted Japanese hiragana characters. Japanese characters were firstly converted to IDs, followed by the embedding process to obtain a 768-dimensional vector for each character, the size of which was the same as the input/output size of the English decoder.

Table 3 shows examples of translation results, where the BLEU score is 0.20. Although it was not sufficient, it can be seen that our model can correctly translate several words that probably appeared in the training data set. It is generally acceptable that words which did not appear in the training data set can hardly be translated correctly, because the model did not know the terms. It is finally concluded that the translation system was trained, and can translate Japanese sentences into English sentences to some extent.

3.2.2 Japanese Text Autoencoder

Second, we investigated our Japanese text autoencoder, illustrated in Figure 1 (c). We used the same data set as in the NMT experiment above; only Japanese sentences were used this time. The encoder was the same as the Japanese encoder in NMT, which was fixed throughout this experiment. The decoder had the same model architecture, and was optimized using the Japanese sentences.

Table 4 indicates the results, and the BLEU score is 0.24. Similar to the NMT results, some words could be reconstructed correctly, while the other parts could not. In spite that the output sentences seem to be inappropriate perhaps due to the lack of vocabulary in the data set, it can be said that semantic information may still be retained in our model.

In NMT and autoencoder experiments, we observed still lower BLEU performance, mainly due to the lack of vocabulary. It is of course needed to improve the scores, however, it was unknown that such the systems could contribute to our final goal, to build a better ASR system. We then moved to the next ASR experiment using the above NLP models.

3.3 Experiments in ASR

3.3.1 Our Proposed Method

Finally, we evaluated our Japanese ASR system, shown in Figure 1 (d). For fine-tuning, we prepared 5,880 Japanese spoken sentences from the Common Voice 7.0 Japanese data set (Ardila et al., 2019). Regarding a recognition model, the English HuBERT-based feature extractor was chosen, in addition to the Japanese text decoder. The whole decoder and the four layers on the output side in the encoder were

Table 5: Examples of recognition results of our proposed method in Figure 1 (d).

Reference (English translation)	Recognized characters
がめんがこうこくだらけでみにくてしょうがない (The screen is full of ads, making it difficult to see.)	がめんがこうぽくだらけでみにくてしょうがない
こまっているひとはほっておけないせいかく (I have a personality that cannot leave people in need.)	こまっているひとはほうっておうけないせいかくう
かれらはゆうびんはいたついんにわいろをわたしなんとかそのてがみをてにはいれました (They bribed the postman and managed to get the letter.)	かれらはゆうびんはいったついんにわいろをわたしなんとかそんてがめをてにいでました
おなじないようのちしきでもじょうしきとかがくとではありかたがちがっている (Even for the same content knowledge, common sense and science have different ways of being.)	おなじないようのちしきでもじょうしきとかがくとではありかたがちがっている

Table 6: Examples of recognition results of the competitive method.

Reference (English translation)	Recognized characters
りかはにがてだがどっぷらーこうかだけはおぼえてる (I'm not good at science, but I only remember the Doppler effect)	でぃかーはにがてだがどっぷらーこうかだけはおぼえてる
なぜならさいこうのゆうじんはかれしかいないから (Because he is the best friend I have.)	なぜならさいこうのゆうじんはかでしかいないから
ふたりはたけーなといった (Two said it was expensive.)	ふたりはたけいなとをいった
どうめいのすーぱーでもおみせごとにしなぞろえにとくちょうがある (Even within the same chain of supermarkets, each store has its own unique selection of products.)	どうめいのすーぱーでもおみせごとにしなぞくちょうがある

Table 7: Character error rates of Japanese ASR systems.

Model	CER [%]
Proposed	20.18
Baseline	27.97

then fine-tuned. We also obtained test data, i.e. 1,928 Japanese spoken sentences, from the same data set.

Table 5 shows examples of Japanese recognition results. The whole character error rate is 20.18%. It is found that the recognition results seem to be acceptable; they are not perfect, but in many cases we can easily guess the meaning.

3.3.2 Competitive Baseline Method

For comparison, we also built another Japanese ASR model as a baseline, which was similar to (Hattori and Tamura, 2023). We prepared a model consisting of the same architecture as our proposed scheme: the English HuBERT-based feature extractor and the Japanese recognizer. Similar to our previous work, we carried out fine-tuning to the four layers in the encoder, and trained the decoder from scratch. Table 6 indicates recognition examples. We then obtained a character error rate of 27.97%, which means that our scheme achieved a 7.79% absolute improvement or a 27.85% relative error reduction over the baseline.

Table 7 summarizes both accuracy. Still, the performance needs to be improved for practical use. However, as mentioned it is hard to collect speech data for any minor language, and the lack of the data set causes low ASR accuracy. It is found that our scheme could compensate the performance by employing NLP technology and text data. Consequently, we believe that the effectiveness of our proposed concept to build a better indigenous ASR method using another SSL-based ASR for a major language as well as NLP techniques is clarified.

4 CONCLUSION

This paper proposed a novel framework to build an ASR system for under-resourced languages by utilizing an SSL-based ASR system for a major language, as well as NLP technology such as NMT and text autoencoder. First, we made English ASR, Japanese-English NMT and Japanese text autoencoder. We checked their performance, and confirmed that we had developed them well. Second, we built a Japanese ASR using sub-modules of the above systems. We conducted evaluation experiments, and it is found that we could successfully develop the system with acceptable accuracy.

As our future works, we need to improve models to achieve higher BLEU scores using larger data sets. Investigating the relationship between the data size and the performance is also useful, since collecting a larger database for indigenous languages requires higher costs. We will also compare our scheme with the state-of-the-art NLP and ASR system so that we could know the performance upper limit, which is useful for future improvement. It is also obvious that applying the proposed technique to real indigenous languages is included in our future tasks.

REFERENCES

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Hattori, T. and Tamura, S. (2023). Speech recognition for minority languages using hubert and model adaptation. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 350–355.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Pryzant, R., Chung, Y., Jurafsky, D., and Britz, D. (2018). JESC: Japanese-english subtitle corpus. *arXiv preprint arXiv:1710.10639*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2214–2218.

UNESCO (2019). LT4All. https://lt4all.org/en/index.html.