Speech Recognition for Minority Languages Using HuBERT and Model Adaptation

Tomohiro Hattori and Satoshi Tamura
Department of Computing, Gifu University,1-1 Yanagido, Gifu, Japan

Keywords: Speech Recognition, Hubert, Minority Language, Adaptation.

Abstract: In the field of speech recognition, models and datasets are becoming larger and larger. However, it is difficult to create large datasets for minority languages, which is an obstacle to improve the accuracy of speech recognition. In this study, we attempt to improve the recognition accuracy for minority languages, by utilizing models trained on large datasets of major language, followed by adapting its language model part to the target language. It is believed that deep-learning speech recognition models learn acoustic and language processing parts. Acoustic one may be common among any languages and has fewer differences than language one. Therefore, we investigate whether it is possible to build a recognizer by keeping acoustic processing learned in the other languages and adapting language processing to the minority language.

1 INTRODUCTION

Speech recognition is a technology that captures human speech, analyzes the waveform, and converts the data into spoken text. In recent years, speech recognition technology has become widely used in modern society, improving work efficiency and enhancing our life quality; for example, we utilize the technology to take minutes, transcribe transcriptions, and fill out forms; speech recognition is also used in various home situations, for instance, voice assistants, voice input, and real-time translation applications for smartphones, smart speakers, and any other devices.

For this decade, speech recognition has made remarkable progress with the rapid development of deep learning, and becomes one of the most active research fields. Many researchers and developers have attempted to build deep-learning models with the large number of parameters by using huge data sets, resulting significant improvement of recognition accuracy. In ideal noise-free environments, such models can almost outperform human speech recognition ability. As deep-learning-based speech recognition schemes, many techniques are proposed until now. One of promising methods is to employ an end-to-end architecture, which has been widely studied. It is considered that an end-to-end speech recognition model jointly learn and contain acoustic model and language model, which both were built separately in conventional speech recognition. An acoustic model describes frequency components and temporal variation of phonemes to be recognized, and a language model predicts sentences according to the phoneme or word sequences given from the acoustic model. In the end-to-end model, the first part plays a role in acoustic front-end which may be common among all languages, while the rest part corresponds to the language model which must be quite different from language to language.

Most researchers are focusing on major languages, such as English, Chinese, Spanish, Japanese and German. On the other hand, there are more than 6,000 languages on this planet, most which have not yet been investigated in this field; most minority languages have few mother-tongue people, as well as the small number of spoken data or texts available to develop speech recognizers. Particularly regarding deep-learning-based speech recognition, for those languages, it is quite difficult to collect the large amount of labeled data, and it is thus hard to sufficiently train a large-scale model with the large number of parameters. Because many minority languages are spoken in developing countries, preparing computer resources to adequately use the huge data set for building a model from scratch is also a matter from cost and technical standpoints. Therefore, as long as we know, there are few works focusing on speech recognition for minority languages using state-of-the-art deep learning technology.

This paper proposes an approach to employ an ad-

Hattori, T. and Tamura, S.
Speech Recognition for Minority Languages Using HuBERT and Model Adaptation.
DOI: 10.5220/0011682700003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 350-355
ISBN: 978-989-758-626-2; ISSN: 2184-4313
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

vanced deep-learning speech recognition model built for a particular major language, by adapting the model to a minority language having few resources. As described above, we now have high-performance models such as BERT (Devlin et al., 2018) or its development XLNet(Yang et al., 2019). For major languages, it is relatively easy to obtain end-to-end deep learning models in which the acoustic frontend part can be available for the other languages. Therefore, if we adapt, or conduct fine-tuning to, the language model part in the model for a minority language using its small number of data, we may obtain a speech recognizer for the minority language, employing a state-of-the-art deep learning architecture. We compare the proposed approach with the conventional scheme; a recurrent neural network model is chosen as a competitive model, as the model can be built using the small number of training utterances. We evaluate both methods by recognition accuracy. Note that in this paper, for experimental reasons, we consider English as a major language and Japanese as a minority language. That is, we use only a few Japanese utterances in our experiments.

Our contribution is as follows; we show the possibility to accomplish a deep-learning-based speech recognition model for a minority language, by adapting a pre-trained HuBERT model that is trained using a large-scale corpus in a major language. It finally enables us to improve recognition accuracy, compared with any model that is trained using a small number of dataset in the minority language.

2 RELATED WORK

In recent years, speech recognition using deep learning has made remarkable progress; a decade ago researchers started to employ deep learning mainly to extract acoustic features, followed by composing it to hidden Markov models which were commonly used in order to compute feature observation probabilities instead of Gaussian mixtures. After that, recurrent neural networks such as Long Short-Term Memory (LSTM) were chosen to replace conventional models. As many architectures e.g. attention mechanism and Transformer appeared, speech recognizers have exploited them. For instance, Quartznet (Kriman et al., 2020), Conformer (Gulati et al., 2020), and Contextnet (Han et al., 2020) have improved the accuracy of speech recognition by training large models with the large number of parameters.

To build such the models, it is essential to prepare labeled data for model training. However, only a small portion of the existing speech data is well labeled, while the others have not yet. Therefore, the method of using not only labeled speech data but also unlabeled speech data for training large-scale models has been considered. One example of self-supervised learning that uses unlabeled speech as training data is wav2vec (Schneider et al., 2019), that performs expression learning by contrastive learning. The wav2vec method conducted pre-training with unlabeled speech and then carry out fine-turing with the small number of labeled speech. Another approach, HuBERT (Hsu et al., 2021), was proposed which was an expression learning model that follows wav2vec and performed pre-training by clustering raw speech waveforms.

For minority languages, low-resource speech recognition methods have been studied in the past. One scheme (Bansal et al., 2018) used a model based on LSTM for speech-to-text translation and showed that fine-turning after pre-training with large data improved the accuracy. Multi-lingual speech recognition schemes (Dalmia et al., 2018), (Fathima et al.,) and meta-learning (Hsu et al., 2019) method have been proposed. Furthermore, some attempts have been made to improve the accuracy by performing data augmentation to compensate the lack of training data. For instance, MixSpeech (Hsu et al., 2019) trained a recognition model using a weighted combination of two different speech features as input and improved the accuracy.

3 METHODOLOGY

This section describes details of our proposed scheme.

We employ HuBERT as a recognition model.

As described, this work aims at making a speech recognition model for minority language from a model for major language. Therefore, the HuBERT model is firstly pre-trained on large-scale English speech data. HuBERT is an expression learning model and is considered to have high generalization performance. Hence, HuBERT is expected to well learn the general acoustic feature extraction part in addition to the particular language modeling part for English. After pre-training, we perform fine-tuning to the model on a Japanese speech data set consisting of a smaller amount of data. By carrying out fine-tuning, it is expected to replace the language modeling part for English into a Japanese language model, with keeping the acoustic processing part. It is known that Japanese has fewer phonemes than English, that is, some phonemes used in English like /th/ and /ae/ are missing. The Japanese phoneme set is thus a sub-

set of English one, enabling us to use English acoustic processing for Japanese. Note that, as described in this case, Japanese is regarded as a minority language. Finally, we test the model checking the effectiveness of the acoustic processing part in the model and the performance of the whole system. Details about implementation and experimental condition will appear in the next section.

4 EXPERIMENTS

In this chapter, we explain experimental condition and report experimental results in detail.

4.1 Dataset

In this study, we used the speech recorded in Common Voice (Ardila et al., 2019). The dataset contains speech data of various languages, and was used in this work for voice recording and validation. This dataset was collected on a crowd-sourcing manner, which allows anyone to contribute to the dataset through PCs and smartphones. This framework enhances the scale and sustainability of the dataset. This dataset is also released to the public domain under an easy-to-use license. As described later, by utilizing this dataset we could perform model adaptation from English to Japanese. This may contribute to future development of speech recognition for any minority language.

Each voice read out is a Japanese sentence, for example, “今回のバージョンアップはいい感じ” which literally means “The new version of the software seems good.” In this study, we used the Japanese dataset included in the Common Voice dataset ver7.0. From the dataset, we removed non-Japanese nouns such as person names, and those containing numbers, followed by converting correct answer labels including Chinese characters into Japanese Hiragana characters only. As a result, the training data, evaluation data, and test data had 5,849, 3,438, and 1,927 utterances, respectively. The details of the Japanese dataset of Common Voice ver7.0 are shown in Table 1. Note that the above numbers of sentences for training, validation and test sets differs from those in Table 1, because of the removal of inappropriate sentences as mentioned above.

Table 1: Details of Common Voice Dataset (Japanese).

Data	Dataset name	Common Voice
	Train data	6,390
	Validation data	3,675
	Test data	2,037
Audio	Sampling rate	16,000[Hz]
	File extension	RIFF Waveform Audio (.wav)

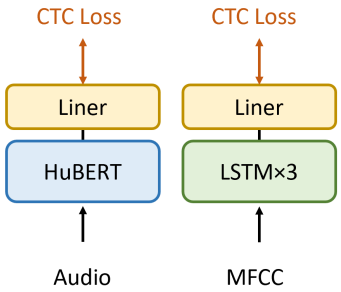

Figure 1: Model architectures.

4.2 Models

4.2.1 HuBERT

In speech recognition using deep learning, Transformer has further improved accuracy from previous methods such as LSTM. However, conventional supervised methods require a large amount of labeled training data. HuBERT is a method to avoid this problem by using self-supervised learning. HuBERT is thus useful in which the large number of labelled training data is not available, such as minority languages. HuBERT uses clustering to create pseudo-labels from speech data, and then performs pre-training with a Masked Language Model (MLM) that masks a portion of the speech data and predicts the pseudo-labels. The pseudo-labels are updated during the learning process, and the feature representation is gradually improved. HuBERT finally learns both acoustic and language model parts by training MLMs on speech data. Because the acoustic feature extraction part is considered to be common irrelated to languages, it is effective to use this part which is trained using different language data.

We used the pre-trained HuBERT model (facebook/hubert-base-ls960) obtained from haggingface (Wolf et al., 2020), which is a BASE-size pre-trained model of HuBERT published by Meta (met,) for English speech recognition. This HuBERT model was pre-trained using Librispeech (Panayotov et al., 2015). Librispeech is a large dataset consisting of 960 hours of English speech data with a sampling frequency of 16,000[Hz], and has been used as a benchmark in many studies in the field of speech recognition.

4.2.2 Comparative Method

LSTM Model. As a comparison method, we built an LSTM model (Hochreiter and Schmidhuber, 1997)

Table 2: Model parameters and condition.

	HuBERT	LSTM
Loss	CTC Loss	
Optimizer	Adam	
Learning rate	0.001	
Batch size	8	32
Early stopping	5	10

with three layers and Connectionist Temporal Classification (CTC). This LSTM model was trained only using Japanese data. Japanese speech data in Common Voice were converted into 40-dimensional MFCCs, and given into the LSTM after SpecAugument (Park et al., 2019). The parameters used for training are shown in Table 2.

HuBERT from Scratch. For comparison, we conducted the same experiment with another HuBERT model without pre-training. Only Japanese data were used for training the model having the same architecture. The parameters used for training, such as data pre-processing and batch size, were the same as those used in the models with pre-training.

4.3 Training

CTC Loss and Adam were used as loss function and optimization method, respectively. The learning rate of Adam was 0.001. In general, as the learning progresses and the loss value approaches a minimum value, the parameters tend to oscillate around the optimal value. In addition, over-learning may occur, in which the parameters are over-fitted to the training data. To avoid these problems, early stopping was used. In this study, early stopping was performed when the loss value for the evaluation data did not fall below the minimum value for five consecutive epochs for the HuBERT model, and for ten consecutive epochs for the LSTM model, respectively. Fine-tuning of HuBERT was performed for the CTC layer and the transformer Encoder4 layer in the model, which is close to the output layer. The parameters used for training are summarized in Table 2.

4.4 Evaluation Metrics

In this study, we used the Character Error Rate (CER) as an evaluation metrics. The CER is calculated as follows:

$$\mathrm{CER} = \frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{\mathrm{N}} \quad (1)$$

$$= \frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{\mathrm{S}+\mathrm{D}+\mathrm{C}} \quad (2)$$

Table 3: Recognition results.

Model	CER[%]
Proposed	5.99
LSTM	7.45
HuBERT from scratch	24.46

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct characters, N is the number of words in the reference (N=S+D+C).

5 RESULTS

Table 3 shows the results of the experiments with pre-trained HuBERT with fine-tuning (proposed approach), LSTM, and the HuBERT from scratch. The proposed scheme resulted in a CER about 1.5% lower than that of LSTM. The HuBERT model only from Japanese data has the highest CER of 24.46%. Examples of recognized sentences are shown in Table 4. It is obvious that the proposed method is the closest to the correct label. For the HuBERT built using only Japanese data, no character was obtained.

In terms of computational time for model training, this study was conducted using one Nvidia RTX3090, and each of those methods required 1-2 hours of training time. There was no significant difference.

As shown in Table 3, pre-trained and fine-tuned HuBERT resulted in a CER about 1.5% lower than LSTM. When the dataset size is small, our approach achieved a better CER, indicating that utilizing a pre-trained model is more effective than a small model from scratch. We believe that this result may be due to the fact that the pre-training was conducted using a large-scale English corpus, resulting well trained acoustic processing in the HuBERT model, in addition, the model was appropriately fine-tuned fitting the new language. On the other hand, a HuBERT model from scratch did not show any progress in learning and had the highest CER. The comparison between the English pre-trained and Japanese fine-tuned HuBERT model and another model only from Japanese data suggests that pre-training is quite effective. We consider that the CER is lower because the acoustic model of the pre-trained model could be used and worked well even for the different language.

Comparing the output results in Table 4, the proposed scheme has fewer errors at the beginning of words than the LSTM model. This may also be due to the higher accuracy of the acoustic model with pre-trained HuBERT, since it is difficult for the language model to predict the next character at the beginning of a sentence or word. The higher CER for HuBERT

Table 4: Comparison for one Japanese sentence.

Label/Model	Recognized characters
Correct label	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
Pre-trained HuBERT with fine-tuning (proposed)	おなじないようのうちしきでも じょうしきとかくとでは ありかてがちがっているる
LSTM	わなじないようのきしきでも きょうしきとくらあくとでは ありかかがけがっているる
HuBERT from scratch	(blank)

from scratch may be due to the fact that the number of parameters of the model was too large for the small dataset, and thus each parameter could not be fully optimized.

6 CONCLUSION

In this study, we proposed a scheme to build a speech recognizer for any minority language, having only a few training data. A pre-trained HuBERT model for a major language, having enough training data, was chosen followed by fine-tuning to its language model part. We conducted experiments using the proposed model, LSTM, and another HuBERT model for the target language on a small number of training data. We compared their CERs and output results. The Fine-Tuned model with pre-trained HuBERT was shown to improve the CER compared to the other models. In addition, it is also found that the number of errors at the beginning of sentences or words was reduced. These indicate that it is effective to use the acoustic processing part in the model that was pre-trained in the other languages.

7 FUTURE WORK

There are three issues to be addressed in the future. The first is to test our method on a variety of languages. In this study, we conducted experiments with a small amount of Japanese data as a minority language. However, Japanese is quite different from English in terms of grammar, vocabulary and letter. We will try to conduct the same experiments based on the above scheme, to clarify the effectiveness of our proposed approach for many minority languages.

Comparison of our proposed scheme with related or similar works is the second work. We conducted preliminary experiments and found a possibility that our scheme is still useful. We would like to further investigate our scheme and the other candidates, e.g. wav2vec and conformer, to clarify advantages and disadvantages of the models in different situations, such as the amount of training data, grammar (SVO, SOV, etc.) and characters, and so on.

The third is to utilize existing language models. In this study, we focused on the reuse of acoustic processing part in the pre-trained models. On the other hand, it is well known that applying a language model to speech recognition can improve the accuracy. Since text data can be collected more easily than speech data, it is possible that better CER can be obtained by integrating language models into our scheme.

REFERENCES

Company info and news — meta. https://about.facebook.com/. Accessed: 2022-02-03.

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*.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., and Goldwater, S. (2018). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Dalmia, S., Sanabria, R., Metze, F., and Black, A. W. (2018). Sequence-based multi-lingual low resource speech recognition. *CoRR*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fathima, N., Patel, T., Mahima, C., and Iyengar, A. Tdnn-based multilingual speech recognition system for low resource indian languages. In *Interspeech*, pages 3197–3201.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu,

J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R., and Wu, Y. (2020). Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*.

Hsu, J., Chen, Y., and Lee, H. (2019). Meta learning for end-to-end low-resource speech recognition. *CoRR*.

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*.

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., and Zhang, Y. (2020). Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS