Inter-language Transfer Learning for Visual Speech Recognition toward Under-resourced Environments

Fumiya Kondo, Satoshi Tamura
Gifu University
1-1 Yanagido, Gifu, 501-1193 Japan
kondo@asr.info.gifu-u.ac.jp, tamura@info.gifu-u.ac.jp

Abstract

In this study, we introduce a method of inter-language transfer learning for under-resourced visual speech recognition. Deploying speech-related technology to all languages is a quite important activity. However, applying state-of-the-art deep-learning techniques requires huge-size labeled corpora, which makes it hard for under-resourced languages. Our approach leverages a small amount of labeled video data of the target language, and employs inter-language transfer learning using a pre-trained English lip-reading model. By applying the proposed scheme, we build a Japanese lip-reading model, using the ROHAN corpus, the size of which is about one 450th of the size of English datasets. The front-end encoder part of the pre-trained model is fine-tuned to improve the acquisition of pronunciation and lip movement patterns unique to Japanese. On the other hand, the back-end encoder and the decoder are built using the Japanese dataset. Although English and Japanese have different language structures, evaluation experiments show that it is possible to build the Japanese lip-reading model efficiently. Comparison with competitive schemes demonstrates the effectiveness of our method.

Keywords: visual speech recognition, lip-reading, transfer learning

1. Introduction

In recent years, extensive research works have been conducted in the fields of Automatic Speech Recognition (ASR), Visual Speech Recognition (VSR), and Audio-Visual Speech Recognition (AVSR). The advancement of deep learning techniques has led to significant improvements in recognition accuracy for these studies. One key factor behind this success is the utilization of large-scale models and datasets. Several languages having high demands and populations, such as English and Mandarin, are well investigated using huge datasets. On the other hand, we should still investigate techniques in under-resourced conditions, in order to enhance the recognition performance.

This study focuses on VSR or lip-reading, which transcribes visual speech activities, e.g. changes in lip movements, shapes, and facial expressions. This technique can serve as an effective mode of communication, even in environments at high levels of noise. VSR also contributes to our society, particularly in providing communication support for individuals with hearing or speech impairments.

Our final goal is to build a VSR system for under-resourced languages. Similar to ASR, numerous English lip-reading models, trained on extensive datasets, are now available for public use. In contrast, VSR research works for the other languages are still insufficient. For example, there is a notable absence of a Japanese large-scale lip-reading dataset, making it significant challenges to create an accurate Japanese lip-reading model.

The objective of this study is to develop a Japanese lip-reading model through inter-language transfer learning, using a limited resource. English VSR models are primarily designed to analyze English pronunciation and lip movements, which may be partially or fully common for all languages. We introduce a method of inter-language transfer learning that leverages a small amount of Japanese data applied to a pre-trained English lip-reading model. This approach enables the model to acquire pronunciation and lip movement patterns unique to Japanese, facilitating the more efficient development of a Japanese lip-reading model.

2. Proposed Method

2.1. Lip-reading Model

We use an end-to-end lip-reading model composed of a front-end encoder part, a back-end encoder, a decoder, and predictors, as shown in Figure 1. The model is based on a pre-trained English version from the paper (Ma et al., 2023). The pre-trained model was trained on five English language datasets; LRW (Chung and Zisserman, 2017), LRS2 (Chung et al., 2017), LRS3 (Afouras et al., 2018), Voxceleb2 (Chung et al., 2018), and AVSpeech (Ephrat et al., 2018). These datasets comprise a total of 3,448 hours of video data, providing a substantial volume of training data. It is reported that the model achieved Word Error Rate (WER) of 14.6% on the LRS2 test dataset and 19.1% on the LRS3 test dataset, demonstrating high recognition performance across both datasets.

The model is designed as follows;

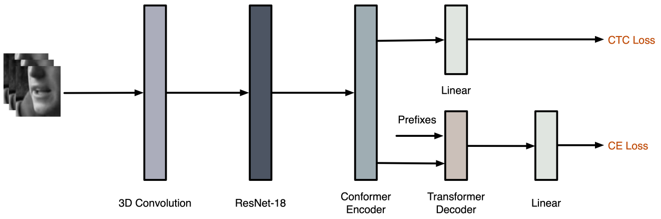

Figure 1: A schematic diagram of lip-reading model (Quoted from paper (Ma et al., 2023)).

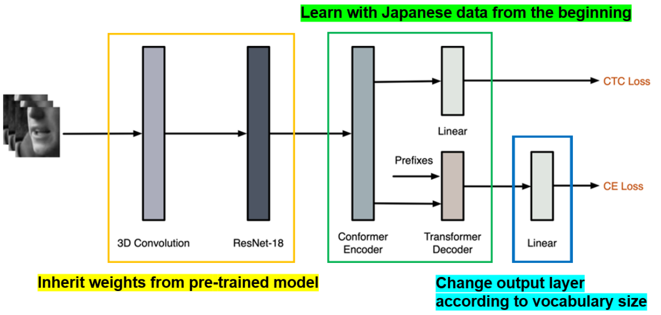
Learn with Japanese data from the beginning
CTC Loss
Linear
Prefixes
CE Loss
3D Convolution
ResNet-18
Conformer Encoder
Transformer Decoder
Linear
Inherit weights from pre-trained model
Change output layer according to vocabulary size

Figure 2: A schematic diagram of inter-language transfer learning from English to Japanese.

- Front-end encoder
  This part consists of 3D convolution layers and the ResNet-18 (He et al., 2016) model. The front-end encoder part aggregates and outputs visual features as a 512-dimensional feature vector.

- Back-end encoder
  We employ the conformer (Gulati et al., 2020). The conformer encoder incorporates transformer and CNN models to successfully capture both long-range dependencies between frame sequences as well as local features in each frame.

- Decoder
  The transformer decoder (Vaswani et al., 2017) is chosen in this work. The attention mechanism in the decoder enables us to predict appropriate tokens by considering both visual features and contextual information.

2.2. Inter-language Transfer Learning

In this study, inter-language transfer learning in addition to model training is applied to develop a Japanese lip-reading model from the English pre-trained model. Figure 2 illustrates a schematic diagram of the inter-language transfer learning.

First, the front-end encoder part is initialized with the weights from the pre-trained model. This part enables us to efficiently extract language-independent visual features, such as lip shape, and the speed and extent of mouth opening and closing. It is further expected that the recognition accuracy can be improved by adjusting these encoders to Japanese data with fine-tuning, since the model can fit the pronunciation and lip movements unique to Japanese, while those unique to English may be discarded.

Second, the back-end encoder and the decoder are built from scratch, keeping the structure of the pre-trained model. According to the similar work for ASR (Hattori and Tamura, 2023), such a recognizer implicitly consists of two modules; a feature extraction module and a recognition module. The latter module relies on vocabulary and grammar of the target language while the former one is language-independent. It is obvious that sentence structures of English and Japanese are markedly different, and the linguistic features derived from visual cues show low similarity. Therefore, we train these submodules only using Japanese datasets.

Regarding the linear layer following the transformer decoder, we change the model setting to the target language; the layer was originally designed for English words, on the other hand, in this paper, the output layer is modified to Japanese character-based labels. The dimension of the output layer is thus changed from 5,000 to 87.

2.3. Loss Function

The loss function is Hybrid CTC/Attention (Watanabe et al., 2017) loss, as in the pre-trained model. Let us denote an input sequence by $\mathbf{x} = [x_1, ..., x_T]$ where x_i indicates a video frame, and an output sequence by $\mathbf{y} = [y_1, ..., y_L]$, where y_j corresponds to a word, character or phoneme, respectively. The loss function, combining Connectionist Temporal Classification (CTC) (Graves et al., 2006) and attention mechanism approaches, is defined as Equation (1), using CTC loss and Cross Entropy (CE) loss.

$$\mathcal{L}_{VSR} = \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{CE} \tag{1}$$

Table 1: Subsets in ROHAN corpus.

Subset	# sentences
Training	3,400
Validation	400
Test	400

In this study, the hyper-parameter α in Equation (1) is set to 0.1.

The CTC loss measures the discrepancy between the sequence predicted by the model and the correct sequence. By using this loss function, we can build the model even when the temporal correspondence between the input and output data is unknown. In the pre-trained model, the linear layer following the conformer encoder is trained using this CTC loss, which is defined by the following Equation (2).

$$\mathcal{L}_{CTC} = -\log P_{CTC}(\mathbf{y}|\mathbf{x}) \tag{2}$$

The CE loss, on the other hand, is a loss function primarily used in classification tasks to maximize the probabilitiy of the correct token at each time point. In the pre-trained model, the linear layer following the transformer decoder is trained using this loss function, which is defined by the following Equation (3).

$$\mathcal{L}_{CE} = -\log P_{CE}(\mathbf{y}|\mathbf{x}) \tag{3}$$

3. Dataset and Pre-processing

3.1. ROHAN Dataset

In this study, we use a Japanese dataset ROHAN (Morise, 2022) for lip-reading. ROHAN consists of 4,600 sentences, which are collected to cover almost all the Japanese moras (the minimum set of combination of acoustic units). The dataset contains video data corresponding to each sentence, which can be used to train lip-reading models. Speech signals are not included, while cropped mouth sequences are composed in the video data. Note that as of February 2024, there are 4,200 video data available to the public. The dataset is divided into three subsets, as shown in Table 1. The total duration of the training data is 7.7 hours, which is explicitly a small amount of data, equivalent to one 450th of the datasets used in the pre-trained model. Additionally, the test dataset includes only one speaker. We point this out in particular because the number of speakers may affect the recognition results of the lip-reading model.

3.2. Reference Label

In order to prepare reference labels for model training, we choose transcribed sentences from the dataset, which consist only of Japanese katakana characters. After splitting the sentences into katakana characters, we assign a unique ID to each katakana character. A SentencePiece (Kudo, 2018) model is developed using the katakana sentences, to uniquely assign an ID to each character. Finally, we get 87 unique IDs in total, which corresponds to the number of Japanese vocabulary in this study.

Table 2: Model training condition.

Optimizer	AdamW
Learning rate	0.0001
Warm-up epoch	5
Weight decay	0.03
Epochs	60
Maximum number of frames	1,600
Loss function	Hybrid CTC/Attention

Table 3: Character error rates with/without inter-language transfer learning.

Method	CER
Proposed (w/ inter-lang. transfer)	0.197
Competitive (w/o inter-lang. transfer)	0.277

3.3. Video Data

Pre-processing of video data is performed in the following order. First, the size of all video data is changed from 300x300 to 96x96, and the frame rate is unified at 25 frames per second. Next, we normalize pixel values from the range of (0, 255) to (0, 1).

We apply random cropping and adaptive time masking to the training data to facilitate spatial and temporal data augmentation. Random cropping involves cutting a random portion from given images to create new images of size 88x88. Adaptive time masking randomly obscures several parts of each frame within a certain time frame. For the validation and test data, center cropping yields image frames of the same size, cropped from the center to the size of 88x88.

Additionally, all the video data are converted to gray-scale to reduce computational costs. In order to enhance the robustness against environmental changes, the pixel value distribution is adjusted so that the new distribution has the mean of 0.421 and the standard deviation of 0.165.

4. Experiment

In order to evaluate the effectiveness of our proposed approach to build a lip-reading scheme for an under-resourced language, we conducted the following experiments.

Table 4: Comparison of our and competitive Japanese lip-reading schemes.

Item	Proposed	Baseline
Input image size	96x96	96x96
Front-end encoder part	3D-CNN+ResNet-18	3D-CNN+ResNet-34
Back-end encoder	Conformer	Conformer
Decoder	Transformer	Transformer
Number of classes	87	166
Dataset (Japanese corpus)	ROHAN	ROHAN+ITA
CER	0.197 (katakana)	0.373 (mora-level)

4.1. Evaluation Metric

Character Error Rate (CER) was chosen as an evaluation metric. CER is a measure of the percentage of incorrectly predicted characters. CER is calculated by the following Equation (4).

$$CER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \tag{4}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correctly recognized characters, and N is the number of characters in the reference ($N{=}S{+}D{+}C$), respectively.

4.2. Experimental Setup

Experimental setup for model training is shown in Table 2. We employed AdamW (Loshchilov and Hutter, 2017) as an optimizer. This method is an extension version of the widely-used Adam (Kingma and Ba, 2014) algorithm in the field of deep learning, accomplishing a weight decay more effectively. During the warm-up, the learning rate was set lower than the value in Table 2 for the first 5 epochs, and then gradually increased to the normal learning rate. The number of epochs was set to 60 and the maximum number of frames to 1,600. The batch size is defined by the number of frames. This means that up to 1,600 frames of the data can be processed per batch. The Hybrid CTC/Attention loss introduced in Equation (1) was used as the loss function. A single NVIDIA GeForce RTX 3090 machine was used in this experiment.

4.3. Result and Discussion

4.3.1. Recognition performance

Effectiveness of inter-language transfer learning

We compared our proposed method to a scheme without the inter-language transfer learning, in which the entire lip-reading model network was trained from scratch using Japanese data only. Note that the other conditions, such as the model architecture, dataset, pre-processing and hyper-parameters for model training were the same as those of the proposed method. The experimental results are shown in Table 3.

Table 3 shows that the proposed method achieved 8% lower CER than the competitive scheme without transfer learning. It is thus found that inter-language transfer learning with a small amount of training data is effective for building a lip-reading model in under-resourced environments, using the pre-trained English high-performance lip-reading model. As already mentioned, the English pre-trained scheme recorded WER of 19.1% in the LRS3 test dataset. Though we cannot directly compare these results, it turns out that our proposed method can achieve enough performance.

Regarding computational time, it took approximately five hours to build the proposed model. Training the competitive model needed almost the same time. The fact that the proposed method can be effectively built within practical time and no significant difference between the proposed and competitive schemes suggests its practicality and efficiency.

Comparison of Japanese lip-reading methods

We also evaluated our scheme in Japanese lip-reading; we focus on another baseline (Arakane et al., 2022), in which the Japanese corpus ROHAN and ITA (Koguchi et al., 2021) were used to develop a conformer-based Japanese lip-reading model. A comparison of the architecture and recognition accuracy between our proposed method and the baseline lip-reading model is presented in Table 4. The front-end encoder part of the proposed method was pre-trained using five English datasets, while the baseline front-end encoder part was pre-trained solely with the LRW dataset.

We tried to compare both results in CER. In the former work they employed a mora-based recognizer; in spite the number of moras varies in several papers, they said the total number is about 170. On the other hand, the number of Japanese katakana characters used in our scheme is approximately 90. A mora is a basic phonological unit, and often identical to a Japanese katakana character; however, there are differences in some units. Table 5 shows the difference between katakana notation and mora-level notation in one sentence. Though it

Table 5: The difference between katakana notation and mora-level notation.

Character	Sentence
Katakana (Proposed)	ナガシギリガカンゼンニハイレバ、 デバフノコウカガフヨサレル。(Japanese)
Mora-level (Baseline)	/silB/na/ga/shi/gi/ri/ga/ka/N/ze/N/ni/ha/i/re/ba/sp /de/ba/fu/no/ko/o/ka/ga/fu/yo/sa/re/ru/silE/
Latin	na ga shi gi ri ga ka n ze n ni ha i re ba , de ba fu no ko u ka ga fu yo sa re ru .
English	If the swift slash is executed perfectly, the debuff effect will be applied.

is hard to directly compare the results, our method achieved approximately 17% lower than the former baseline. Even taking the different numbers of classes into account, the results suggest the significant performance improvement achieved by our proposed method.

4.3.2. Analysis of recognized sentences

Examples of recognition results obtained the proposed and competitive methods as well as the correct transcription and corresponding English sentence are shown in Table 6. Characters in red indicate errors in substitution, deletion, and insertion.

Comparing the results of the proposed method with the sentences from another scheme without pre-training, it is found that the proposed method can generate more accurate results, especially in recognizing characters at the beginning of sentences. It is also observed that our scheme can more correctly recognize consecutive characters having the same vowel sounds. On the other hand, we sometimes found the same substitution, deletion, and insertion errors in both results, indicating that fine-tuning was not sufficient to avoid such errors. Looking at the results in Table 6, we can guess the meaning from the output of our proposed scheme. This suggests our system may be useful in practical use.

In conclusion, as also shown in the recognition performance, it is clarified that our proposed method can generally output more correct sentences, that are closer to the correct labels. This means our approach is useful to compensate the lack of training data in VSR, reaching better recognition performance.

Table 6: An example of recognition results (Red characters indicate recognition errors).

Correct sentence	ヒメジジョウノグニャグニャトマガリクネッタコミチワ、セメコマレニククスルクフウデアル。(Japanese) hi me ji j yo u no gu n ya gu n ya to ma ga ri ku ne t ta ko mi chi wa , se me ko ma re ni ku ku su ru ku fu u de a ru . 'Zig-zag winding pathways to Himeji castle are designed to make it more difficult to be attacked.'

Method	Predicted sentence
Proposed	ヒメチソ_オノグニャグニャトマガリクダッタコミチワ、セメコマレニグクスルキュフクデアル。(Japanese) hi me chi so _ o no gu n ya gu n ya to ma ga ri ku da t ta ko mi chi wa , se me ko ma re ni gu ku su ru k yu fu ku de a ru .
Competitive	シメキゾ_ウトズザ_ルニャトマガリクダッタコミチワ、セメコマレニウ_スグチュフクデアル。(Japanese) shi me ki zo _ u to zu za _ ru n ya to ma ga ri ku da t ta ko mi chi wa , se me ko ma re ni u _ su gu ch yu fu ku de a ru .

5. Conclusion

This paper proposed how to build a high-performance lip-reading recognizer for under-resourced languages based on inter-language transfer learning. This scheme was inspired by the success of the similar strategy in ASR. We applied the technique to make a Japanese VSR system using a pre-trained English VSR model. Experimental results show the effectiveness of our method in constructing a lip-reading model using a small amount of video data. Finally, we achieved roughly 20% CER performance, which may be acceptable in practical use.

Our future work includes the application of our scheme to the other languages. Through experiments in different language and data settings, we will clarify the effectiveness of our scheme in detail. Employing Large Language Models (LLM) to further improve the results is also interesting. Building an AVSR system by combining our approach and ASR will be explored.

6. Bibliographical References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Taiki Arakane, Takeshi Saitoh, Ryuuichi Chiba, Masanori Morise, and Yasuo Oda. 2022. Conformer-based lip-reading for Japanese sentence. In *International Conference on Image and Vision Computing*, pages 474–485. Springer.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *International Conference on Computer Vision and Pattern Recognition*, pages 6447–6456.

Joon Son Chung and Andrew Zisserman. 2017. Lip reading in the wild. In *International Conference on Computer Vision*, pages 87–103. Springer.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference 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, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Tomohiro Hattori and Satoshi Tamura. 2023. Speech recognition for minority languages using HuBERT and model adaptation. In *International Conference on Pattern Recognition Applications and Methods*, pages 350–355.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Junya Koguchi, Ikuya Kanai, Yasuo Oda, Takeshi Saitoh, Masanori Morise, et al. 2021. ITA corpus: Construction and basic evaluation of a Japanese text corpus composed of phoneme-balanced sentences from the public domain. *In Proceedings of IPSJ SIGMUS (in Japanese)*, 2021(31):1–4.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Masanori Morise. 2022. ROHAN: Morae-balanced Japanese corpus for text-to-speech synthesis. *Journal of Acoustical Society of Japan (in Japanese)*, 79(1):9–17.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.