Audio-Visual Speech Recognition Based on Cross-Lingual Transfer Learning

Fumiya Kondo* and Satoshi Tamura*
* Gifu University, Japan
E-mail: {kondo@asr.,tamura@}info.gifu-u.ac.jp Tel/Fax: +81-58-293-2748

***Abstract*—This paper investigates the effectiveness of cross-lingual transfer learning in audio-only and audio-visual speech recognition. We propose a cross-lingual transfer learning technique and apply it to visual speech recognition, achieving significant improvement. We then extend this technique to audio-only and audio-visual speech recognition in this work. The front-end encoders are inherited from high-performance English models and then fine-tuned using a small amount of Japanese multimodal data, while the other modules are built from scratch. We conducted evaluation experiments and found that our transfer learning scheme can be highly effective, even when only a small amount of audio-visual data is available. Results under noisy conditions also indicate improved performance and robustness in speech recognition.**

I. Introduction

The rapid evolution of Deep Learning (DL) techniques has led to remarkable progress in Automatic Speech Recognition (ASR), Visual Speech Recognition (VSR), and their integration, such as Audio-Visual Speech Recognition (AVSR). Today, ASR has been used on smart devices with promising performance. The accuracy of VSR has drastically improved, compared to lipreading schemes proposed before the DL era. In AVSR, DL can effectively incorporate audio and visual modalities, realizing significant performance and robustness, especially in real-world environments.

The objective of this research paper is to build a high-accuracy Japanese AVSR model. In order to achieve high recognition accuracy, leveraging large-scale DL models and training datasets is essential. Currently, corpora with video data such as LRS2 [1] and LRS3 [2], as well as several high-performance recognition models pre-trained on large-scale datasets in English, are available; these models and related tools can be obtained from development platforms such as GitHub and Hugging Face. Most existing pre-trained models are basically designed for English utterances. On the other hand, it is still challenging to build such a model for other languages, e.g., Japanese. English and Japanese have different acoustic phoneme systems as well as lip and mouth dynamics, along with grammar and vocabulary differences. Only a few Japanese corpora are available for VSR and AVSR, and to our knowledge, there is no large-scale Japanese multimodal dataset for them. That makes it hard to build Japanese VSR and AVSR models from scratch.

Therefore, to address this data scarcity problem, this study proposes a cross-lingual transfer learning approach. We have already succeeded in building a VSR model in our previous work [4]; the inter-language transfer learning technique was applied to obtain a Japanese VSR system from an English lipreading model, using a small Japanese video corpus for fine-tuning. This work thus focuses on AVSR, which includes not only ASR but also VSR architectures. In our method, first, an AVSR model pre-trained on a large-scale English dataset is prepared. Second, the model is fine-tuned using a small Japanese dataset. This scheme enables the model to learn audio patterns appearing in Japanese utterances, as well as visual dynamics made by Japanese speakers, resulting in a high-performance Japanese AVSR model. We conducted experiments to evaluate the effectiveness of our method. We applied the same methodology to ASR and VSR, and compared the results. We also performed recognition experiments in noisy conditions to check the robustness of AVSR.

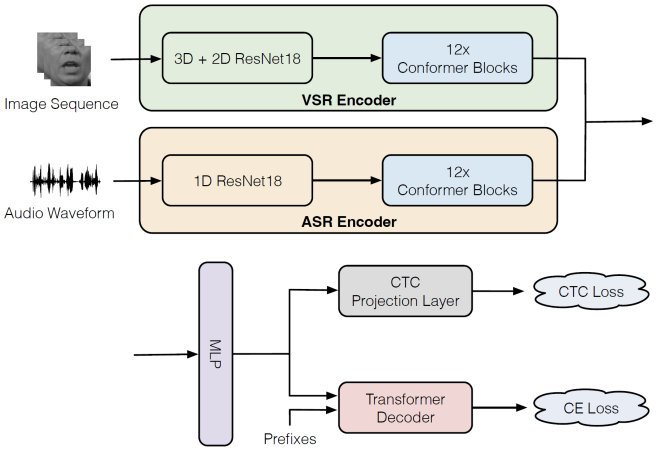


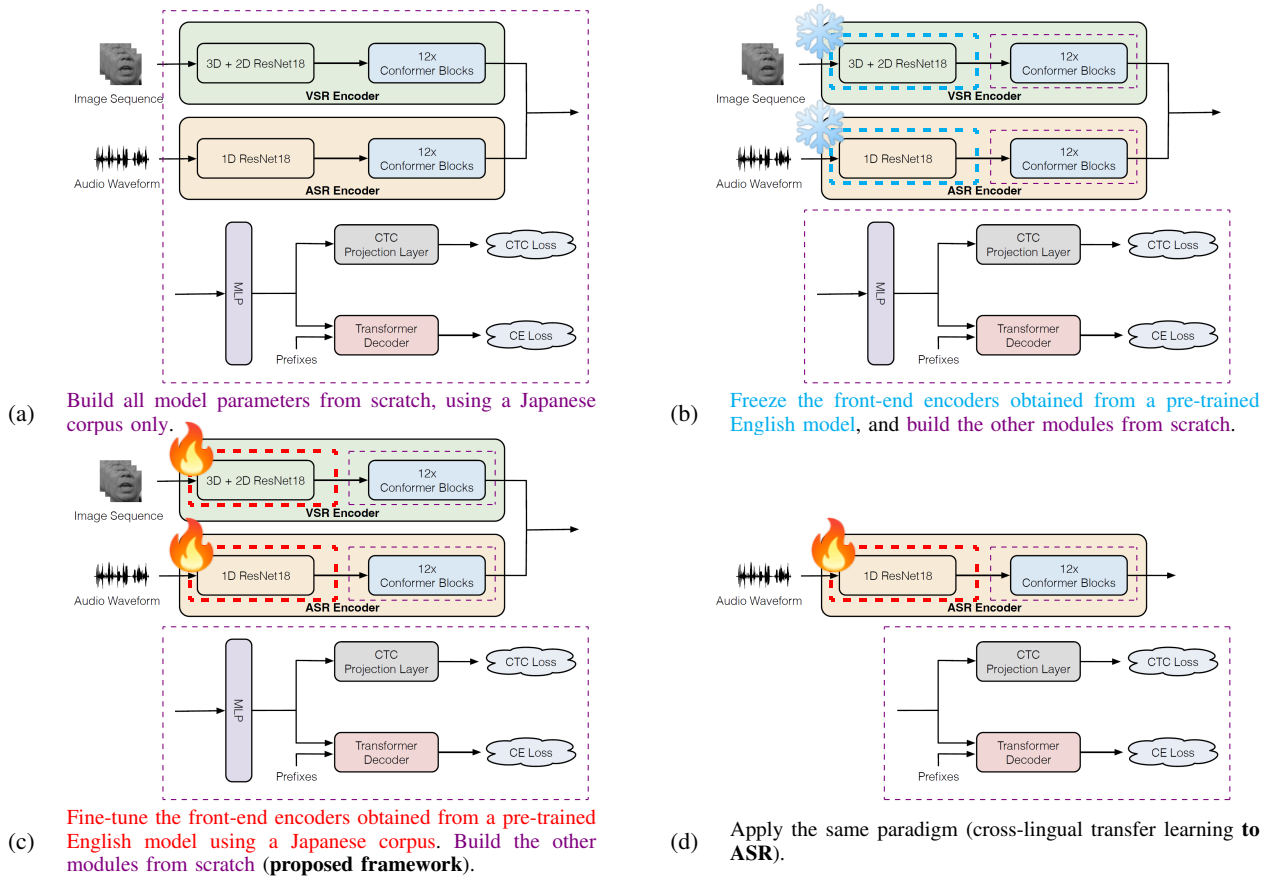
Fig. 1. An overview of the AVSR model. (from the article "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels" [3]).

The rest of this paper is organized as follows. Section II describes related work. The methodology is explained in Section III. The experimental setup is shown in Section IV. Section V presents experimental results and discussion. Finally, Section VI concludes this paper.

II. Related Work

As related work, this section introduces the architecture of the pre-trained model that is employed in this paper, as well as the baseline model used for performance evaluation.

979-8-3315-7206-8/25/$31.00



(a) Build all model parameters from scratch, using a Japanese corpus only.

(b) Freeze the front-end encoders obtained from a pre-trained English model, and build the other modules from scratch.

(c) Fine-tune the front-end encoders obtained from a pre-trained English model using a Japanese corpus. Build the other modules from scratch (**proposed framework**).

(d) Apply the same paradigm (cross-lingual transfer learning **to ASR**).

Fig. 2. Training and evaluation strategies in this paper.

A. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels

The previous study [3] focused on an end-to-end AVSR model [5]. The model architecture is shown in Figure 1. This model has VSR and ASR encoder modules. They chose a 3D convolutional layer and a 2D ResNet-18 [6], [7] as the VSR front-end, as well as a 1D ResNet-18 [8] as the ASR front-end, followed by a Conformer [9] as the back-end, respectively. The features from both modalities are then integrated via a Multi-Layer Perceptron (MLP) fusion module. The AV features are sent to a subsequent Connectionist Temporal Classification (CTC) projection layer and a Transformer decoder [10]. The architecture employs a Hybrid CTC/Attention [11], which combines CTC loss and Cross-Entropy (CE) loss. This model also enables us to train and perform ASR or VSR by exploiting only a single modality.

In the paper [3], they demonstrated that by augmenting training data with automatically generated labels, the performance of speech recognition models could be improved by increasing the amount of data, even if noisy training samples were included. Consequently, the AVSR model achieved extremely low Word Error Rate (WER) of 1.5% on the LRS2 dataset [1] and 0.9% on the LRS3 dataset [2]. Note that a combination of five English datasets totaling 3,448 hours; LRS2 [1], LRS3 [2], LRW [12], VoxCeleb2 [13], and AVSpeech [14] were used.

B. Integrating Pre-Trained Speech and Language Models for End-to-End Speech Recognition

The model "Nue-ASR" proposed in the previous study [15] provides an end-to-end ASR for Japanese. They integrated the pre-trained speech representation model HuBERT (rinna/japanese-hubert-base) [16] with the large language model GPT-NeoX (rinna/japanese-gpt-neox-3.6b) [16]; they also employed an additional bridge network consisting of convolutional layers and linear projections. By leveraging these pre-trained models to reduce training costs and using the approximately 19,000-hour Japanese speech corpus "ReazonSpeech" [17], they achieved high recognition performance compared to other state-of-the-art end-to-end Japanese ASR models. This model and related program codes are publicly available on Hugging Face. We then adopt the model as a baseline for ASR.

III. Methodology (Cross-lingual Transfer Learning)

This section proposes a cross-lingual transfer learning method to develop a high-accuracy Japanese AVSR model. As already mentioned, it is difficult to build a high-performance AVSR system by training a model from scratch because of the lack of large-scale Japanese multimodal datasets. Our scheme involves fine-tuning an AVSR model pre-trained on a large-

scale English dataset using a small amount of Japanese AV data.

A. Transfer Learning

The cross-lingual transfer is done in two ways, as follows. For the front-end encoders, parameters are inherited from the original English model followed by fine-tuning, while the other modules are built using Japanese data. Figure 2 (c) depicts our scheme.

- **Front-end encoders**
  In the VSR front-end (3D convolutional layer and 2D ResNet-18) and ASR front-end (1D ResNet-18), encoders are obtained using the original English model and Japanese data for fine-tuning. These components are expected to extract general-purpose features that are less dependent on any specific language, such as human lip movements and fundamental acoustic characteristics. Nevertheless, these front-end encoders are still based on English, as the training was done using English data. Fine-tuning is thus needed to adjust the model parameters to the target language. Therefore, by inheriting the feature extraction capabilities acquired from large-scale English data and fine-tuning with a small amount of Japanese data, we believe that we can efficiently adapt the model to Japanese pronunciation and lip dynamics, resulting in better encoders.
- **Back-end encoders, MLP and decoders**
  The VSR/ASR back-end (Conformer) encoders and subsequent modules including MLP, CTC projection, and Transformer decoder are built from scratch. They are responsible for capturing temporal contexts of the extracted features and learning language-specific grammatical structures in addition to phonological rules. These linguistic features are significantly dependent on the target language, in this case, Japanese. Therefore, the parameters in these modules are not inherited but overwritten using Japanese data.

In terms of the decoder part, the output dimension must be adjusted based on the vocabulary size of a given training corpus or the target language. The original decoder is designed for English terms, while a character-based Japanese recognizer can be built by reducing the vocabulary size; the dimension is thus changed from 5,000 (for English words) to 87 (for Japanese characters, i.e., Katakana). In terms of loss functions, let us denote a CTC loss and a CE loss by L_{CTC} and L_{CE}, respectively. The CTC loss L_{CTC} measures the discrepancy between the sequence predicted by the model and the correct sequence. Another loss, L_{CE}, is used in classification tasks to maximize the probability of the correct token at each time point. We then combine them to define the loss function of the whole architecture as:

$$L = \alpha\, L_{CTC} + (1 - \alpha)\, L_{CE} \tag{1}$$

where α is a scaling factor, for which this paper employs 0.1.

TABLE I
SUBSETS IN THE ROHAN DATASET.

Subset	# of sentences
Training	3,400
Validation	400
Test	400

B. Competitive Training Strategies

To validate the effectiveness of the proposed method, several training strategies are tested in this paper. Figure 2 illustrates these schemes as well as our method.

(a) **Training from scratch using Japanese data only**
All model parameters are trained from scratch using Japanese data only.

(b) **Freezing front-end parameters**
The front-end encoders employ pre-trained model parameters, which are frozen (not updated) during the following training. The parameters from the back-end encoders to the end are trained from scratch on Japanese data.

(c) **Fine-tuning front-end parameters**
This corresponds to our proposed approach. The front-end encoders are first initialized using pre-trained parameters and subsequently fine-tuned using Japanese data. The parameters from the back-end encoders to the end are trained from scratch on Japanese data, as described in the previous scheme.

C. ASR and VSR Models

In order to evaluate our AVSR model, ASR and VSR models are also built and used as in [4]. This allows us to discuss the comparative analysis and effectiveness of cross-lingual transfer learning for each modality. We employ the same model architecture as AVSR, then remove the visual encoder for ASR and the acoustic encoder for VSR, respectively. These models are trained under the same conditions as the AVSR model described above, e.g., Figure 2 (d). Note that the MLP layer in the AVSR model is skipped for ASR and VSR; the MLP layer is used to integrate audio and visual features, which is not needed for unimodal recognition schemes.

IV. EXPERIMENT

A. Datasets

In this research work, we used the ROHAN [18] Japanese multimodal dataset. This open-source dataset consists of paired audio-video data corresponding to read-aloud sentences, enabling us to train ASR, VSR, and AVSR models. The entire 4,200 pairs were divided into training, validation, and test sets, respectively. Table I shows the subsets. The training data from this set amounts to approximately 7.7 hours. This is roughly 1/450th of the total training time (3,448 hours) of the major English datasets [1], [2], [12]–[14]. This indicates a significant scarcity of Japanese multimodal datasets compared to their English counterparts.

TABLE II
ASR RESULTS OF THE BASELINE MODEL AND OUR MODELS

Method	Input Type	Training Data (Total Hours)	CER (%) (↓)
Baseline (Rinna/Nue-ASR) [15]	Audio-only (16,000Hz)	ReazonSpeech (19,000h) [17]	4.04
Ours (a), *built from scratch*		ROHAN (7.7h) [18]	6.51
Ours (b), *English front-end*			4.08
Ours (c), *cross-lingual transfer*			**3.88**

TABLE III
ASR, VSR, AND AVSR RESULTS WITH THE JAPANESE DATASET

Input Type	Method & CER (%) (↓)		
	(a) *Built from scratch*	(b) *English front-end*	(c) *Cross-lingual transfer*
Audio-only (16,000Hz)	6.51	4.08	3.88
Audio-only (**19,200Hz**)	6.15	3.75	3.34
Visual-only (25fps)	27.73	23.39	19.70
Audio-Visual (**19,200Hz**, 25fps)	12.48	2.96	**2.78**

B. Pre-processing

- **Video data**
  For the video data, the frame rate was standardized to 25 fps. The Regions of Interest (ROIs) were cropped using a 96×96 pixel bounding box, so that the cropped image included only the mouth region.
- **Audio data**
  The sampling rate for the audio data was standardized. Note that this paper chose 19,200Hz for AVSR, though 16,000Hz is commonly used for training speech recognition models; applying the sampling rate of 16,000Hz to the AVSR model with the dataset [18] caused extensive zero-padding during audio-video synchronization, which significantly reduced training efficiency. For comparison, we also tested the conventional 16,000Hz sampling for ASR.

C. Reference Label

In order to prepare reference labels for model training, we chose transcribed sentences consisting only of Japanese Katakana characters. A SentencePiece [19] model was subsequently applied using the Katakana sentences to uniquely assign an ID to each character. Finally, we obtained 87 unique IDs in total.

D. Experimental Setup

For model training, we adopted AdamW [20] as an optimizer and a cosine learning rate scheduler. A warm-up was performed for 5 epochs, with the peak learning rate set to 1×10^{-4}. Some hyper-parameters were empirically modified for each modality. For the ASR and VSR models, the number of epochs was set to 60, and the maximum number of frames per batch was 1,600. In contrast, for the AVSR model, the number of epochs was set to 40, and the maximum number of frames was 1,000. We utilized a single NVIDIA GeForce RTX 3090 throughout our experiments.

E. Evaluation Metric

Character Error Rate (CER) was chosen as an evaluation metric. CER indicates the percentage of incorrectly predicted characters. CER is defined by the following Equation (2).

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2}$$

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 ($N = S + D + C$).

V. RESULTS

A. ASR Results

We conducted speech recognition experiments using the baseline [15] and our ASR models. Table II shows a performance comparison. It is found that our proposed model could achieve almost the same performance as the baseline, despite an overwhelming difference in the amount of training data; while the baseline model was trained using 19,000-hour data, our scheme used only 7.7 hours. As mentioned, the same paradigm has already been applied to VSR, achieving significant lipreading performance [4]. We then found that our proposed approach is also effective in ASR, although many large-scale Japanese speech corpora are available for ASR. This result also suggests the potential to apply this framework to low-resource languages for ASR and VSR.

B. Results of Our Proposed Scheme

Table III shows recognition results of our proposed models in ASR, VSR, and AVSR. The first row, ASR with 16,000Hz sampling, corresponds to ours in Table II. The VSR results in the third row also correspond to our previous results in [4].

- **Discussion of ASR and VSR results**
  In the ASR results, the performance difference between the 16,000Hz and 19,200Hz sampling rates is not significant. This indicates that we can adopt the 19,200Hz sampling rate to incorporate audio and visual modalities efficiently in AVSR. The VSR performance

TABLE IV
ASR AND AVSR RESULTS IN NOISY ENVIRONMENTS

Input Type	Noise Level & CER (%) (↓)		
	Clean	SNR=10dB	SNR=5dB
Audio-only (16,000Hz)	3.88	26.02	40.13
Audio-only (19,200Hz)	3.34	34.74	55.48
Audio-Visual (19,200Hz, 25fps)	2.78	**11.95**	**16.11**

was lower than that of ASR. This usually happens because of fewer cues in the visual modality. According to previous lipreading research, we believe that 20% CER is sufficiently good for VSR tasks.

- **Comparison of ASR and AVSR results**
  Focusing on the cross-lingual transfer results, we found that the ASR model (19,200Hz) achieved a lower CER of 3.34%, demonstrating high recognition performance as a unimodal system. Furthermore, by incorporating the visual modality, our AVSR model improved the CER to 2.78%. Although the absolute performance gain was modest, it is concluded that our AVSR model successfully reduced roughly 17% of errors compared to ASR. Consequently, this result suggests that combining multimodal information (audio and video) with the fine-tuning of front-end parameters via cross-lingual transfer learning is extremely effective for speech recognition tasks.
- **Effectiveness of cross-lingual transfer learning and front-end fine-tuning**
  Across all input types, the proposed method with cross-lingual transfer learning clearly reduced the CER compared to the scratch models. On the other hand, focusing on the scratch models, we found that the CER of the AVSR model (12.48%) was notably higher than that of the ASR model (6.15%). Since the AVSR model has more parameters than the ASR model, we consider this occurred due to the lack of training data. This also indicates the usefulness of cross-lingual transfer learning, as the AVSR model shows higher performance than ASR after applying the transfer learning. By comparing the results of (b) and (c), it turns out that fine-tuning contributed to performance improvement, particularly for VSR (CER 23.39% $\rightarrow$ 19.70%). This indicates that optimizing the front-end for the target language is crucial for capturing complex features like lip movements.

C. Noise Experiments

AVSR has the advantage of high performance and robustness. On the other hand, it also has the disadvantage of additional computational costs, such as video data processing and modality integration. Here, we focus on whether the AVSR model justifies these costs. We conducted ASR and AVSR recognition experiments in noisy conditions. We added white noise to the Japanese audio data (400 test items from ROHAN [18]) for evaluation. Table IV compares the performance of the ASR and AVSR models at various SNR levels under the condition of (c) cross-lingual transfer learning with front-end fine-tuning.

- **Comparison of ASR and AVSR results**
  In ideal, noiseless environments, there is no significant performance difference between ASR and AVSR. However, in noisy environments such as SNR=10dB and SNR=5dB, the AVSR model demonstrated lower CER — and thus higher recognition accuracy — compared to the ASR models. The result at SNR=5dB is particularly remarkable; the ASR models had more than 40% CER, while our AVSR scheme achieved a better CER of 16.11%. This clearly indicates that visual information can contribute immensely to improving recognition accuracy when the audio signal is degraded by noise.
- **Discussion about SNR Levels**
  By comparing the results at 10dB and 5dB, it is found that significant performance degradation was observed in the ASR methods. In contrast, only a 4-point CER increase (from 11.95% to 16.11%) was observed in AVSR. This result indicates the robustness of AVSR frameworks against acoustic noise.

VI. CONCLUSION

This paper investigated the effectiveness of cross-lingual transfer learning in audio-visual speech recognition, in addition to speech recognition and lipreading. In our method, front-end modules in ASR and VSR encoders are inherited from an English model and then fine-tuned using a target-language corpus. Evaluation experiments were carried out to compare the ASR and AVSR models with and without transfer learning, and to evaluate the robustness of AVSR models against noise. The results show that our cross-lingual transfer learning method can work effectively, improving both performance and noise robustness in AVSR.

Our future work is as follows. The optimal cross-lingual transfer scheme will be examined, including which inherited parameters in the model, such as Conformers and the MLP fusion module, should be fine-tuned. Evaluation of AVSR performance under different kinds of acoustic noise, as well as in visually degraded environments, is expected. We will further investigate the effectiveness of our scheme using larger-scale corpora and various models. Finally, we would like to apply the proposed method to other languages, particularly low-resource ones.

References

[1] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2017, pp. 6447–6456.

[2] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[3] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-visual speech recognition with automatic labels," in *ICASSP*, 2023, pp. 1–5.

[4] F. Kondo and S. Tamura, "Inter-language transfer learning for visual speech recognition toward under-resourced environments," in *SIGUL (Special Interest Group on Under-resourced Languages)*, 2024, pp. 149–154.

[5] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021, pp. 7613–7617.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[7] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *INTERSPEECH*, 2017, pp. 3652–3656.

[8] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *ICASSP*, 2018, pp. 6548–6552.

[9] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020, pp. 5036–5040.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[11] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[12] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Computer Vision–ACCV 2016*, Springer, 2017, pp. 87–103.

[13] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018, pp. 1086–1090.

[14] A. Ephrat, I. Mosseri, O. Lang, *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[15] Y. Hono, K. Mitsuda, T. Zhao, K. Mitsui, T. Wakatsuki, and K. Sawada, "Integrating pre-trained speech and language models for end-to-end speech recognition," in *ACL*, 2024, pp. 13 289–13 305.

[16] K. Sawada, T. Zhao, M. Shing, *et al.*, "Release of pre-trained models for the Japanese language," in *LREC-COLING*, 2024, pp. 13 898–13 905.

[17] Y. Yin, D. Mori, and S. Fujimoto, "ReazonSpeech: A free and massive corpus for Japanese ASR," in *Annual Conference of the Association for Natural Language Processing in Japan*, 2023, pp. 1134–1139.

[18] M. Morise, "ROHAN: Mora-balanced Japanese corpus for text-to-speech synthesis," *Journal of Acoustical Society of Japan (in Japanese)*, vol. 79, no. 1, pp. 9–17, 2022.

[19] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *ACL*, 2018, pp. 66–75.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019, pp. 1–8.