17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Japanese Sentence Dataset for Lip-reading

Tatsuya Shirakata and Takeshi Saitoh
Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka, Japan
saitoh@ai.kyutech.ac.jp

Abstract

This research is about lip-reading for Japanese sentences. Research on English sentences is actively pursued due to the extensive datasets. However, a sufficient dataset for Japanese sentences has not been released. Therefore, this paper builds a Japanese sentence dataset. A Transformer model is used for the recognition task. Three recognition target levels: phoneme, mora, and vowel, are set, and recognition experiments show that they can be recognized.

1 Introduction

In the research field on lip-reading, recognition targets are roughly divided into three categories: single sounds, words, and sentences. Single sounds such as vowels have been conducted mainly in the 1980s. Many word-level lip-reading methods have been proposed [1, 2]. However, recent subjects are shifting from words to sentences [3, 4, 5, 6, 7].

Assael et al. proposed a well-known model LipNet, which enables sentence-level lip-reading with an end-to-end model [3]. LipNet consists of ST-CNN and pooling layers in the first stage and two bidirectional GRU layers in the second stage. The word error rate of 11.4% was obtained in the unseen speaker recognition task with GRID [8], which is an English sentence speech scene dataset. Chung et al. constructed a database LRS2 that collects English speech scenes from BBC, proposed a Watch, Listen, Attend and Spell network, and proposed a method of recognizing each character. A character error rate of 39.5% is obtained using only video data [4]. Regarding lip-reading for English sentences, large-scale datasets such as GRID, LRS2, and LRS3 have been released, so they have been actively promoted in recent years [3, 4, 5, 7].

On the other hand, as for the Japanese sentence, Komai et al. [9] collected speech scenes of ATR phoneme-balanced words from one speaker and reported the phoneme recognition results. The parameters of the active appearance model are extracted as features and recognized by HMM. 216 words × 10 sets were used for training, and unknown 100 words × 1 set was used for testing. They showed a phoneme-level accuracy rate of 40.7%. Noda et al. [10] proposed a combined method with CNN and HMM. They collect speech scenes of ATR phoneme balance words from six speakers, conducts recognition experiments for 40 phonemes, and report an average recognition rate of 58%. Unfortunately, these datasets are private and are not available to us.

Compared to English sentences such as GRID and LRS2, a large-scale Japanese speech scene dataset is not released, and the sentence-level lip-reading for Japanese is not sufficiently accurate. This paper originally collects Japanese speech scenes from 32 speakers and experiments with the sentence-level recognition results using the Transformer model [5]. Our challenge is whether it is possible to estimate the sentence even in the Japanese speech scene.

2 Publicly Available Datasets

There are several publicly available sentence-level English databases.

GRID [8] is a syntax of six kinds of words such as "put red at G 9 now", four commands, four colors, four prepositions, 25 alphabets, ten digits, and four adverbs. Each word is randomly assigned, and each speaker recorded 1,000 speech scenes. The number of the speaker is 33. The duration time of all scenes is three seconds. The total duration time of the entire database is 27.5 hours.

LRS2 [4] contains English speech scenes from BBC. LRS2 provides both training and test data. Training data consists of 70,783 scenes, and its total duration time is about 102 hours. Test data consists of 48,165 scenes, and its total duration time is about 29 hours. Not only video data but also speech text is provided. Only the face region is extracted from the video data as preprocessing, the image size is 160 × 160 [pixels], and the frame rate is 25 fps. Furthermore, label information of each word's utterance start/end time is also provided regarding the training data.

LRS3 [6] consists of thousands of spoken sentences from TED and TEDx videos. In this dataset, pretrain data (118516 scenes), trainval data (31982 scenes), and test data (1321 scenes) are provided. The total duration time is about 438 hours. The image size is 224 × 224 [pixels], the frame rate is 25 fps, and the speech scene of the face area and the content text are provided similar to LRS2.

As mentioned above, a large-scale speech scene dataset of Japanese sentences has not been released.

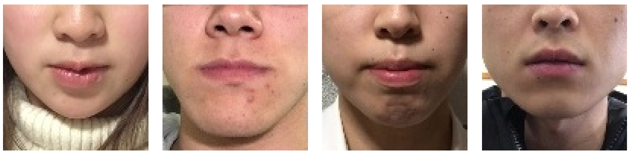


Figure 1. Extracted lipROI.

Table 1. Collected Japanese speech scenes.

corpus	ATR	ITA
# sentences	503	424
# speakers	26 (13M+13F)	6 (1M+5F)
# collected scenes	13,078	3,444
total length [h:m:s]	24:39:02.96	5:50:00.32
average length [s]	6.79	6.11

3 Transformer-based Lip-reading

Afouras et al. [5] proposed three deep neural network models for sentence-level lip-reading. Each model consists of two modules: a spatio-temporal visual front-end and outputs one feature vector per frame; and a sequence processing module that inputs the sequence of per-frame feature vectors and outputs a sentence character by character. In [5], the visual front-end is common across the three models. However, the sequence processing module is different, and the Transformer model obtained the highest accuracy. Thus, this paper adopts this model. We briefly describe each module in the following.

Our task is to predict the sentences from a silent video of a talking face. The input data of the model is a sequence of lipROI. This a square region cropped around the mouth of a face image. All lipROIs are resized to 96 × 96 [pixels] as shown in Fig. 1.

The network applies a spatio-temporal convolution on the input image sequence, with a filter width of five frames, followed by a 2D ResNet that gradually decreases the spatial dimensions with depth. The output is a 512-dimensional feature vector for every input frame.

The Transformer model proposed by Vaswani et al. [11] is used. It is an encoder-decoder model that enables parallel calculation using the Attention mechanism without using RNN to improve the calculation speed. The Encoder and Decoder convert input word into word embeddings through the Embedding layer. At this time, the Transformer does not use RNN, so there is no word order concept. Therefore, the position information is considered by creating an embedded expression that considers the word's position by Positional Encoding. In the Encoder model, one Encoder component consists of two sub-layers, a Multi-Head Attention layer, a fully-connected layer, and behind each sub-layer is a residual and a normalization layer. Six of these components are stacked. The Attention used in Transformer uses Scaled Dot-Product Attention, which calculates Attention by using a query that is a search target word, a key of a word that is an answer, and a pair of values and key-value. Besides, the accuracy is improved by arranging multiple Attentions in Multi-Head Attention. One Decoder component of the Decoder model has a Masked Multi-Head Attention layer at the beginning, a residual coupling, regularization layer, and then consists of the same Multi-Head Attention layer and fully coupling layer as the Encoder component. Masked Multi-Head Attention is Attention that is masked so as not to use the previous prediction. In the Decoder model, similar to the Encoder model, six of these components are stacked.

4 Evaluation Experiment

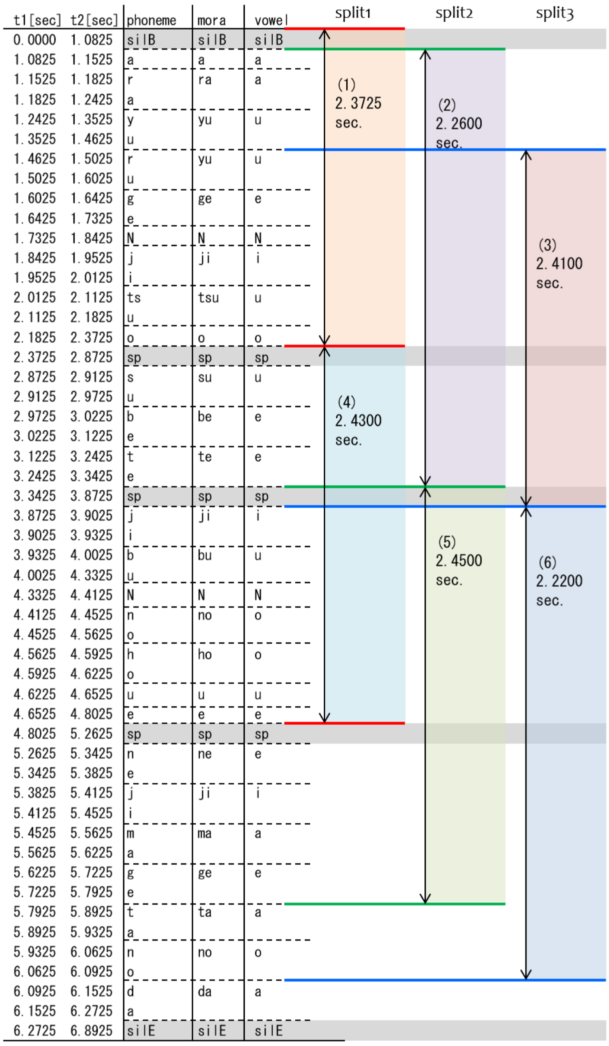
4.1 Speech Scene Collection

Regarding a reference [12], we collected Japanese sentence speech scenes of ATR phoneme balance sentences (ATR) and ITA corpus (ITA) using the tablet. ATR [13] consists of 503 sentences created by balancing the phoneme environment based on about 10,000 sentences randomly selected from newspapers, magazines, novels, letters, textbooks. ITA is a new corpus created to extract a chain of two consecutive phonemes in Japanese and cover all of them. ATR was collected from 26 speakers, and ITA was collected from six speakers. All speakers were healthy people.

Our objective is to propose sentence-level lip-reading for Japanese. In this case, since it is necessary to prepare label information for each character, the speaker was made to speak aloud instead of silently. Besides, it was manually confirmed whether or not the recorded speech scene's speech content was correct. If the speech was incorrect, the speech scene was re-taken so that the speech content was correct. The collected speech scene information is shown in Table 1. Although it is not large enough compared to LRS2, the data scale is larger than other researches on Japanese sentence lip-reading [9, 10].

4.2 Annotation

As mentioned above, the speech contents of ATR and ITA collected in this research are known. On the other hand, the length of the model input is constant F frames. In this paper, $F = 75$ frame is used. Since many recorded speech scenes are longer than F, it is necessary to divide the speech scene. Even if the speech content is known, the speech speed differs depending on the speaker and the content. Thus, we apply phoneme alignment that automatically recognizes phonemes' start and end times to divide the



t1[sec]	t2[sec]	phoneme	mora	vowel
0.0000	1.0825	silB	silB	silB
1.0825	1.1525	a	a	a
1.1525	1.1825	r	ra	a
1.1825	1.2425	a		
1.2425	1.3525	y	yu	u
1.3525	1.4625	u		
1.4625	1.5025	r	yu	u
1.5025	1.6025	u		
1.6025	1.6425	g	ge	e
1.6425	1.7325	e		
1.7325	1.8425	N	N	N
1.8425	1.9525	j	ji	i
1.9525	2.0125	i		
2.0125	2.1125	ts	tsu	u
2.1125	2.1825	u		
2.1825	2.3725	o	o	o
2.3725	2.8725	sp	sp	sp
2.8725	2.9125	s	su	u
2.9125	2.9725	u		
2.9725	3.0225	b	be	e
3.0225	3.1225	e		
3.1225	3.2425	t	te	e
3.2425	3.3425	e		
3.3425	3.8725	sp	sp	sp
3.8725	3.9025	j	ji	i
3.9025	3.9325	i		
3.9325	4.0025	b	bu	u
4.0025	4.3325	u		
4.3325	4.4125	N	N	N
4.4125	4.4525	n	no	o
4.4525	4.5625	o		
4.5625	4.5925	h	ho	o
4.5925	4.6225	o		
4.6225	4.6525	u	u	u
4.6525	4.8025	e	e	e
4.8025	5.2625	sp	sp	sp
5.2625	5.3425	n	ne	e
5.3425	5.3825	e		
5.3825	5.4125	j	ji	i
5.4125	5.4525	i		
5.4525	5.5625	m	ma	a
5.5625	5.6225	a		
5.6225	5.7225	g	ge	e
5.7225	5.7925	e		
5.7925	5.8925	t	ta	a
5.8925	5.9325	a		
5.9325	6.0625	n	no	o
6.0625	6.0925	o		
6.0925	6.1525	d	da	a
6.1525	6.2725	a		
6.2725	6.8925	silE	silE	silE

Figure 2. Phoneme alignment and scene division.

speech scene. In this process, audio data was extracted from the speech scene, and phoneme alignment was performed using the Julius phoneme segmentation toolkit[1]. Figure 2 shows an example of phoneme alignment results. t1 and t2 are the phoneme start-time and end-time in second, respectively.

Chung et al. [4] recognize English sentences in alphabetical units rather than word units. Although Japanese is targeted in this paper, the recognition level can be phoneme, mora, and vowel, as shown in Table 2. A mora is a phonological unit which is the unit of rhythm in Japanese. In the table, `silB` means a silent section before the start of utterance, `silE` means a silent section after the utterance, and `sp` means a short pause. The mora has the longest time among these units, and the phoneme has the shortest time.

Table 2. Recognition target level.

Japanese	/あ/ら/ゆ/る/げ/ん/じ/つ/を/, /
phoneme	/silB/a/r/a/y/u/r/u/g/e/N/j/i/ts/u/o/sp/
mora	/silB/a/ra/yu/ru/ge/N/ji/tsu/o/sp/
vowel	/silB/a/a/u/u/e/N/i/u/o/sp/

The Julius phoneme segmentation toolkit has obtained phoneme level labels. Based on this, the labels for the mora unit and vowel unit are computed. In addition, since these labels are based on audio data, the phoneme, mora, and vowel labels are given for each frame concerning the frame rate of the utterance scene.

4.3 Scene Division

Based on the previous annotation process, the scene is divided in advance. Since the longest unit is the mora, in this experiment, the speech scene is divided so that one scene is 50 $\sim$ 75 frame based on the mora information.

Here, there are roughly two types of scene division approaches, no overlap in which the divided scenes do not overlap on the time axis, and overlap in which the divided scenes overlap on the time axis. Furthermore, when considering overlap, the overlap ratio is also a parameter. This paper defines three approaches, as shown in Fig. 2.

- split1: The scene is divided including `silB` at the start of the speech.
- split2: A method of dividing the scene from the speech start time of the first sound excluding `silB`.
- split3: Divide the scene by shifting about 50% of split 1.

split1 is used for division without overlap.

On the other hand, all scenes divided by the three types of split1, split2, and split3 are used when considering the overlap. In Fig. 2, two split scenes (1) and (4) are used for no overlap, and six split scenes (1) to (6) are used for overlap. Post-padding is applied to adjust the number of frames to F for input to the model. The divided speech scene information obtained by applying the division process is shown in Table 3.

4.4 Experimental Condition

The character error rate (CER) at each recognition level was used as the evaluation. CER is defined by $CER = (I + S + D)/N$, where I is the number of inserted characters I, S is the number of replaced characters, D is the number of deleted characters, and N is the number of correct characters. Since CER is an error rate, the smaller the value, the higher the accuracy.

[1] https://github.com/julius-speech/segmentation-kit

Table 3. Divided speech scenes.

without overlap	ATR	ITA
# scenes	27,587	6,212
length [h:m:s]	21:23:52.68	4:48:03.48
average length [s]	2.79	2.78
with overlap	**ATR**	**ITA**
# scenes	70,415	15,046
length [h:m:s]	54:23:18.76	11:34:22.92
average length [s]	2.78	2.77

A speaker-dependent recognition (SD) task and a speaker-independent recognition (SI) task are recognition tasks. In this experiment, both SD task and SI task experiments were carried out.

Regarding the training data of the Transformer model, we consider two types; one is when only ATR dataset is used, and the other is when both ATR and ITA datasets are used. Furthermore, for the divided scenes, we experimented under two conditions, with and without overlap, for a total of four conditions. For the test data, we used speech scenes with only no overlap ATR dataset.

As mentioned above, there are three types of recognition target levels: phoneme, mora, and vowel, and the number of labels for each level were 44, 150, and 13, respectively.

4.5 Experimental Result

The recognition experiment results of the SD task and SI task are shown in Table 4(a),(b), respectively. The recognition accuracy of the SD task is higher than that of the SI task. Although the training data of both the SD task and SI task are the same, it is easy to imagine that the recognition accuracy will be higher because the test data of the SD task is included in the training data.

The SD task and SI task tended to differ in terms of the difference in the training data and with/without overlap. Since condition (4) has the largest number of training data, it was assumed that the recognition accuracy would be high, but in the SD task, the recognition accuracy of condition (3) is high. It is presumed that because the test data was included in the training data, and the test data had no overlap ATR, so there was too much training data under condition (4). On the other hand, in the SI task of condition (4), the recognition accuracy was high, and the tendency was confirmed as expected.

Regarding the recognition level, the SD task has higher recognition accuracy in the order of phoneme, mora, and vowel, and the SI task has higher recognition accuracy in the order of mora, phoneme, and vowel. In general, as the number of labels to be recognized increases, the recognition accuracy decreases, so the SI task obtained the expected tendency. On the other hand, regarding the SD task, the factors behind the tendency of recognition accuracy have not been clarified.

Table 4. Recognition result（CER[%]）
(a) Speaker depdendent recognition task

dataset		ATR		ATR+ITA	
overlap		×	○	×	○
# training scenes		23,392	29,604	70,415	85,461
# test scenes		4,240			
condition		(1)	(2)	(3)	(4)
target level	phoneme	41.1	**35.5**	37.9	38.7
	mora	40.3	40.3	**11.4**	28.4
	vowel	35.0	25.4	**7.4**	8.2

(b) Speaker independent recognition task

dataset		ATR		ATR+ITA	
overlap		×	○	×	○
# training scenes		23,392	29,604	70,415	85,461
# test scenes		4,195			
condition		(1)	(2)	(3)	(4)
target level	phoneme	80.2	78.4	81.3	**75.3**
	mora	95.4	90.9	**82.6**	82.8
	vowel	55.0	49.4	46.2	**40.0**

5 Conclusion

In research on lip-reading, a large-scale dataset of speech scenes of English sentences has been published. On the other hand, research on Japanese sentences' speech scenes has not progressed because a certain scale dataset has not been prepared. Therefore, we collected ATR and ITA Japanese sentence speech scenes from 29 speakers and prepared them in this paper. Furthermore, we conducted a recognition experiment using the collected Japanese sentence speech scenes for the lip-reading model using the Transformer model. As a result, the CER of the phoneme, mora, and vowel in the SI task were 75.3%, 82.6%, and 40.0%, respectively. Although the recognition accuracy was not sufficient, it showed the possibility of reading Japanese sentences.

As a future task, it is considered that the database of Japanese sentence speech scenes will increase. Build a database of the same scale as LRS2 and investigate the accuracy of Japanese sentences.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 19KT0029.

References

[1] J. S. Chung and A. Zisserman. Lip reading in the wild. In *ACCV*, 2016.

[2] M. Iwasaki, M. Kubokawa, and T. Saitoh. Two features combination with gated recurrent unit for visual speech recognition. In *IAPR MVA*, pp. 300–303, 2017.

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. LipNet: End-to-end sentence-level lipreading. In *arXiv:1611.01599*, 2016.

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

[5] T. Afouras, J. S. Chung, and A. Zisserman. Deep lip reading: A comparison of models and an online application. In *Interspeech 2018*, 2018.

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

[7] X. Zhang, F. Cheng, and S. Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *ICCV*, pp. 713–722, 2019.

[8] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, Vol. 120, No. 5, pp. 2421–2424, 2006.

[9] Y. Komai, C. Miyamoto, T. Takiguchi, and Y. Araki. Phoneme analysis of image feature in utterance recognition using AAM in lip area. MIRU, pp. 1771–1778, 2010. (in Japanese)

[10] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Lipreading using convolutional neural network. In *INTERSPEECH*, pp. 1149–1153, 2014.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv:1706.03762*, 2017.

[12] Takeshi Saitoh and Michiko Kubokawa. SSSD: Speech scene database by smart device for visual speech recognition. In *24th International Conference on Pattern Recognition (ICPR)*, pp. 3228–3232, 2018.

[13] Y. Sagisaka and N. Uratani. ATR spoken language database. Acoustical Society of Japan, Vol. 48, No. 12, pp. 878–882, 1992. (in Japanese)