A Proposal for Creating Syllabic Datasets for Japanese Language Lipreading by Using Machine Learnings

Rui Kitahara¹,ᵃ, Lifeng Zhang¹,ᵇ

¹ Faculty of Engineering, Kyusyu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu, Fukuoka, 804-8550, Japan

ᵃ<kitahara.rui166@mail.kyutech.jp>,ᵇ<zhang@elcs.kyutech.ac.jp>

Abstract

Although lip-reading using image processing and machine learning have been performed at the word level, LipNet,¹ a network that enables recognition at the sentence level, improves the recognition accuracy over the former method. However, this was the case for English speakers, and no extra experimental result has been reported for Japanese speakers. This study aims to create an experimental database for Japanese speech scenes containing all 50 Japanese sounds and evaluate the recognition accuracy using LipNet.¹

Keywords: Machine Learning, Lipreading.

1. Introduction

Machine lip reading using image processing and machine learning has been researched for the long term. But still, it is quite difficult to recognize untrained words than the trained ones. If the recognition system for machine lip reader can be made for speech recognition, the following may be possible. First, lip-reading itself is one of the tools for communicating with the hearing impaired, allowing for smoother communication. Secondly, it can be used in situations where it is difficult to hear voices in remote or noisy environments, such as through security cameras or in a stadium, where the ability to solve crimes is expected. Although it is slightly different from security cameras, recently, a video assistant referee (VAR) system has been introduced in soccer matches. Verbal abuse is often a problem in soccer matches. In soccer stadiums, there is a lot of cheering, and it is difficult to judge what was said exactly by video judgments. Therefore, the use of machine-reader lips in this system may prevent misjudgments.

In recent years, voice input software, such as Siri and Google Assistant, has become popular due to the proliferation of smartphones.

Fig. 1. search by typing character

Fig. 2. Voice input in public

© 2021 The Institute of Industrial Applications Engineers, Japan.

DOI: 10.12792/iciae2021.004
according to Fig. 2, 67.6% of male respondents and 74.6% of female respondents felt embarrassed when they were in front of others. Therefore, more than half of the respondents felt embarrassed about speech input in front of others. Therefore, if the Japanese language machine reading is as good as speech recognition, it may be possible to enter characters without uttering a word.

2. Principle

2.1 Data Preprocessing

The GRID corpus is processed with Dlib’s face detector and the iBug face landmark predictor with 68 landmarks in combination with an online Kalman filter to apply an affine transform. Then, the image around the lips is extracted at a size of 100x50 pixels per frame.

2.1.1 GRID

The GRID corpus used in LipNet is a free public lip-reading corpus provided by the University of Sheffield, UK. The language is English and consists of six elements:

- command + color + preposition + letter + digit + adverb

where

- command is {"bin", "lay", "place", "set"},
- color is {"blue", "green", "red", "white"},
- the preposition is {"at", "by", "in"},
- letter is 25 different types of letters,
- digit is 10 types, from “zero” to “nine”,
- adverb is {"again", "now", "please", and "soon"}.

There were 34 subjects. Each subject was recorded with 10,000 different utterances. The length of all the videos was 3 seconds, and the frame rate was 25 fps. The actual number of usable speech scenes was 32,746 due to the existence of corrupted data. In addition, an alignment is provided for each movie. An example of an alignments show in Fig. 3. Where “sil” and “sp” mean silent time.

2.1.2 Dlib

Dlib is a general-purpose, cross-platform open source software written in C++. It includes components that handle many tasks such as networking, threading, GUIs, data structures, linear algebra, statistical machine learning, image processing, data mining, XML and text analysis, numerical optimization, and Bayesian networks. Much of the recent development has focused on the creation of statistical machine learning tools.

2.1.3 iBug Facial Landmark Predictor

The trained model is by using the iBug 300-W dataset. Using this model and the previously mentioned Dlib face detector, we obtain the 68 landmarks in Fig. 4.

2.2 LipNet

Traditionally, lip-reading has been done by word classification only. LipNet is the first lip-reading model to allow lip-reading at the sentence level. It also achieved an accuracy of 95.2%, surpassing the 86.4% accuracy of experienced human lip readers and the previous word level accuracy of 86.4%. The Fig. 4 shows the LipNet network.
The sequence of T frames is inputted and processed in three layers of Spatiotemporal Convolutional Neural Networks STCNN, which are convolutional neural networks in space-time. The GRU output is processed in both directions by the Gated Recurrent Unit (GRU), a type of Neural Network (RNN). LipNet is trained with Connectionist Temporal Classification (CTC).

2.2.1 STCNN

STCNN (Spatio-temporal convolutional neural networks) can process video data by folding time into spatial dimensions.

\[
\text{stconv}(x, w)[c', t'ij] = \sum_{c=1}^{C} k_c \sum_{t'=1}^{t} \sum_{j'=1}^{j} x_{c,t'+j'} + \sum_{t'=1}^{t} \sum_{j'=1}^{j} w_{c', t'j'}
\]

2.3 WER and CER

The recognition accuracy data in LipNet is WER (Word Error Rate) and CER (Character Error Rate) are calculated and used. The definitions of WER and CER are as follows.

\[
\text{WER} = \frac{i_w + s_w + d_w}{n_w}
\]

\[
i_w : \text{Number of words inserted in the reference text}
\]

\[
s_w : \text{Number of replaced words in the reference text}
\]

\[
d_w : \text{Number of deleted words in the reference text}
\]

\[
n_w : \text{Number of words in the reference text}
\]

\[
\text{CER} = \frac{i + s + d}{n}
\]

\[
i : \text{Number of characters inserted}
\]

\[
s : \text{Number of replaced characters}
\]

\[
d : \text{Deleted characters}
\]

\[
n : \text{Total number of characters}
\]

3. Experiment Preparation

3.1 Data Set Creation

The dataset was created in the same way with the GRID corpus structure used in LipNet, in the order “command + color + color + preposition + letter + digit + adverb”, the GRID corpus contains all 26 letters of the alphabet. Therefore, the words were chosen so that they contained all 50 Japanese words. Fig. 3 shows the structure of the GRID corpus. The data was created on an iPad (7th generation). The frame rate of the video was 30 fps. Then, to apply the LipNet structure, the image around the lips was extracted at a size of 100×50 pixels per frame using a combination of the Dlib face detector and the iBug face landmark predictor with 68 landmarks. Furthermore, since the frames were different for each speech scene, we decided to fit all speech scenes to 400 frames by padding the final frame with closed lips.

4. Experiments and Discussions

4.1 Experiment 1

First, we filmed the speech scenes in Table for 15 subjects (225 speech scenes). Then, we applied it to the LipNet.
structure for training. In addition, we used 12 of the 15 participants (180 speech scenes) for learning and three of the 15 participants (45 speech scenes) for verification. We tested the trained models created by the training. Three new subjects (45 speech scenes) were used for the tests. The input and output words, as well as WER and CER are shown in Table 2.

As shown in Table 2, the recognition accuracy is very low. One possible reason for this is that LipNet itself is not suitable for all Japanese sounds. The reason for this is that LipNet itself is inappropriate for all Japanese sounds. Here, we will focus on the small number of data and padding of the data, which is a method of increasing the amount of data set when performing machine learning. Methods for padding data include “contrast adjustment”, “scaling”, and “flipping”.

### 4.2 Experiment 2

In the padding of the data described above, we thought that a left-right inversion was effective because human faces are generally asymmetrical. Therefore, the data for 30 people (450 speech scenes) were applied to the LipNet structure, including the original data from Experiment 1, by performing a left-right inversion. In this study, we applied the LipNet structure to the data of 30 people (450 speech scenes). Afterwards, we tested the trained model using the test data used in Experiment 1. The results are shown in Table 3.

We can see that the accuracy has increased slightly compared to Experiment 1. Therefore, under the conditions of this experiment, we can consider that the data’s padding by left-right flipping is effective, and the accuracy increases as the number of data increases.

### 5. Conclusion and Future Work

In this study, we proposed creating a dataset covering all the syllables for Japanese machine lip reading and trained it with LipNet. In terms of accuracy, it was not as successful as English recognition. As for future work, since the dataset we created in this study ignored overlapping words and sounds, it was necessary to create a well-balanced dataset. As shown in the experiments, there is a possibility that the accuracy may increase as the data increases. Therefore, a large number of data sets like the GRID corpus will be needed to achieve the same level of accuracy as in English recognition. One way to increase accuracy with a small amount of data is to use transference learning; if a model trained using the GRID corpus can be transfer-learned, it may be possible to increase the accuracy with a small amount of data.
References

(1) Yannis M. Assael and Brendan Shillingford and Shimon Whiteson and Nando de Freitas “LipNet: End-to-End Sentence-level Lipreading” https://arxiv.org/pdf/1611.01599.pdf, 2016

(2) Dlib C++ Library, http://dlib.net/intro.html, Access date: 2020.11.20

(3) Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou and Maja Pantic, “300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge”, https://ibug.doc.ic.ac.uk/media/uploads/documents/sagonas_iccv_2013_300_w.pdf

(4) GRID, https://spandh.dcs.shef.ac.uk/gridcorpus, Access date: 2020.11.20

(5) Dietrich Klakow and Jochne Peters, “Testing the correlation of word error rate and perplexity”, Speech Communication, Volume 38, Issues 1-2, pp.19-28, September 2002, DOI: 10.1016/S0167-6393(01)00041-3

(6) Awareness Survey on Japanese Voice Operations 2017(in Japanese), https://news.kddi.com/kddi/corporate/newsrelease/2017/10/05/2726.html, KDDI. Access date: 2020.11.20

(7) 300 Faces In-the-Wild Challenge (300-W), (https://ibug.doc.ic.ac.uk/resources/300-W), ICCV 2013, Access date: 2020.11.20