LiP25w: Word-level Lip Reading Web Application for Smart Device

Takeshi Saitoh and Michiko Kubokawa

Kyushu Institute of Technology
680–4 Kawazu, Iizuka, Fukuoka, 820–8502, JAPAN.
saitoh@ces.kyutech.ac.jp

Abstract

Lip reading technology is expected for the next-generation interface. However, there is no practical system or interface by lip reading that can be easily used by anyone. This paper develops a highly practical web application named LiP25w that can be used anytime and anywhere by using smart devices. The fundamental method of LiP25w is based on our previous researches. Our application is open to the public. We introduced our application at some events and performed about 1200 trials in 200 days. We analyzed all trial results and obtained an average recognition accuracy of 73.4%. We confirmed that our application is highly practical.

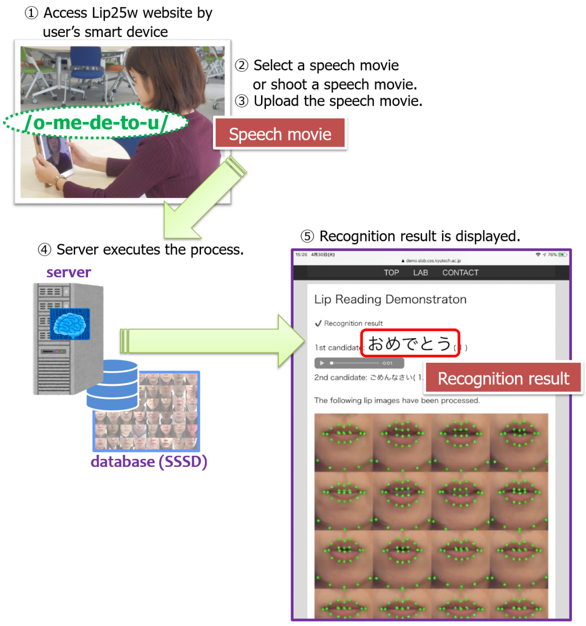
Index Terms: lip reading, web application, smart device, SSSD.

1. Introduction

The lip reading technology or visual speech recognition (VSR) technology that estimates speech content only using visual information without using audio information is expected as an interface for disable people who cannot speak well. Recently, the audio-based speech recognition (ASR) technology changed our lives comfortably. Some typical practical examples are text input on PC or smart device, command input for car navigation or game, and conversation with robot or agent. However, speech disorders or speech impediments cannot produce the sound and they cannot use these ASR systems. In addition, ASR is difficult to use in a high noisy environment or in a public place such as the waiting room in the hospital. Lip reading technology has not been put to practical use yet. In this research, our aim is to develop an interface using lip reading, and in this paper, we develop a novel web application named LiP25w which is implemented word-level lip reading method.

We built a publicly available speech scene database by smart device named SSSD [1] for development a lip reading-based interface. This database contains 36k speech scenes by 48 speakers uttering 25 Japanese words. On the other hand, in our previous research [2], a region of interest (ROI) around the lip is extracted and two types of image-based feature and motion-based feature are proposed, and after recognizing each feature independently, the late fusion approach is applied, which integrates output values of a Recurrent Neural Network (RNN). Although the stacked convolutional autoencoder (SCAE) is applied to obtain image-based feature. These prior studies [2, 1] were either database construction or proposal of method. In this paper, we develop a practical lip reading web application that integrates previous studies. Here, the application developed in this paper recognizes words or phrases registered in the database. We demonstrate the development application using smart device at some events and confirm the usefulness of the application.

The rest of this paper is organized as follows: in Sect. **2** introduces some related researches. Section **3** provides details of our application. In Sect. **4**, experimental results are described. This paper concludes in Sect. **5**.


① Access Lip25w website by user's smart device
/o-me-de-to-u/
② Select a speech movie or shoot a speech movie.
③ Upload the speech movie.
Speech movie
④ Server executes the process.
server
database (SSSD)
⑤ Recognition result is displayed.
Recognition result

15:26 4月30日(火)

demo.slab.ces.kyutech.ac.jp

TOP LAB CONTACT

Lip Reading Demonstraton

✓ Recognition result

1st candidate: おめでとう(1)

2nd candidate: ごめんなさい(1.

The following lip images have been processed.

Figure 1: *Overview of LiP25w.*

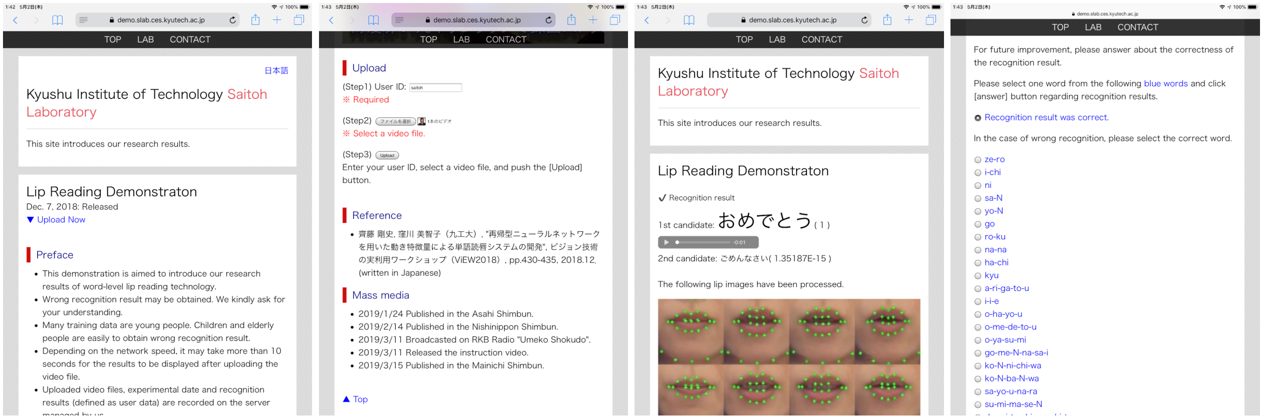
2. Related Research

Regarding research on lip reading, there are many researches propose lip reading methods that mainly aim for higher recognition accuracy than system development such as interface.

In the conventional approaches, various hand-craft features, such as Active Appearance Model (AAM) features [3] or Discreate Cosine Transform (DCT) coefficients, are proposed and these features are fed to Hidden Markov Model (HMM) [4, 5].

Recently, deep learning techniques have been successfully applied to learn features from either audio-visual data or visual data for the tasks of lip reading. Noda et al. proposed to apply a convolutional neural network (CNN) as the visual feature extraction mechanism for lip reading [6]. HMM with Gaussian mixtures were used for the task of recognizing isolated words. Saitoh et al. proposed a sequence image representation method called concatenated frame image (CFI), and evaluated with a public database OuluVS2 [7]. Chung and Zisserman created several datasets of The Oxford-BBC Lip Reading in the Wild (LRW) [8], and The Oxford-BBC Lip Reading Sentences

10.21437/AVSP.2019-17



日本語

Kyushu Institute of Technology Saitoh Laboratory

This site introduces our research results.

Lip Reading Demonstraton

Dec. 7, 2018: Released

▼ Upload Now

Preface

- This demonstration is aimed to introduce our research results of word-level lip reading technology.
- Wrong recognition result may be obtained. We kindly ask for your understanding.
- Many training data are young people. Children and elderly people are easily to obtain wrong recognition result.
- Depending on the network speed, it may take more than 10 seconds for the results to be displayed after uploading the video file.
- Uploaded video files, experimental date and recognition results (defined as user data) are recorded on the server managed by us.

(a)

Upload

(Step1) User ID: saitoh

※ Required

(Step2) ファイルを選択 1本のビデオ

※ Select a video file.

(Step3) Upload

Enter your user ID, select a video file, and push the [Upload] button.

Reference

- 齊藤 剛史, 窪川 美智子（九工大）, "再帰型ニューラルネットワークを用いた動き特徴量による単語読唇システムの開発", ビジョン技術の実利用ワークショップ（ViEW2018）, pp.430-435, 2018.12. (written in Japanese)

Mass media

- 2019/1/24 Published in the Asahi Shimbun.
- 2019/2/14 Published in the Nishinippon Shimbun.
- 2019/3/11 Broadcasted on RKB Radio "Umeko Shokudo".
- 2019/3/11 Released the instruction video.
- 2019/3/15 Published in the Mainichi Shimbun.

▲ Top

(b)

Kyushu Institute of Technology Saitoh Laboratory

This site introduces our research results.

Lip Reading Demonstraton

✓ Recognition result

1st candidate: おめでとう (1)

2nd candidate: ごめんなさい(1.35187E-15)

The following lip images have been processed.

(c)

For future improvement, please answer about the correctness of the recognition result.

Please select one word from the following blue words and click [answer] button regarding recognition results.

● Recognition result was correct.

In the case of wrong recognition, please select the correct word.

○ ze-ro
○ i-chi
○ ni
○ sa-N
○ yo-N
○ go
○ ro-ku
○ na-na
○ ha-chi
○ kyu
○ a-ri-ga-to-u
○ i-i-e
○ o-ha-yo-u
○ o-me-de-to-u
○ o-ya-su-mi
○ go-me-N-na-sa-i
○ ko-N-ni-chi-wa
○ ko-N-ba-N-wa
○ sa-yo-u-na-ra
○ su-mi-ma-se-N

(d)

Figure 2: *Operation screen of LiP25w.*

2 (LRS2) [9]. They use these datasets to propose lip reading methods for words and sentences.

As for the system using lip reading technology, Saitoh developed a complete communication support system for speech and hearing disorders using lip reading [10]. His system recognizes 50 words registered in advance. He implements the system in a commercial laptop. However, his method is for specific speaker, and there are problems in practical use. Shin et al. developed a real time lip reading system for isolated Korean word recognition [11]. Their method is an audio-visual ASR that combined both ASR and VSR. They implemented their method on the car navigation system to demonstrate the effectiveness in noisy environments.

ASR has been put to practical use in various systems such as smart device interfaces. However, with regard to lip reading, no practical system such as an interface that can be easily used by anyone has been developed.

3. LiP25w: Word-level Lip Reading Web Application

3.1. Overview

Although the lip reading application developed in this research is aimed at the next-generation interface as our final goal, it is not a highly practical application. One purpose is to introduce the lip reading technology to many people. Therefore, it is more desirable to develop applications for smart devices than to develop applications for PC. When assuming an application that available on smart devices, it is necessary to support various operating systems such as iOS, Android, and Windows. In this paper, we develop a web application named LiP25w that is easy for everyone to try.

When developing an application, we need to consider both recognition accuracy and processing time. Conventional methods that do not use deep learning have the advantage of low computational cost and fast processing time. However, these methods have lower recognition accuracy than deep learning. By making a web application, processing time can be reduced by performing high cost process on the server. Thus, LiP25w adopted deep learning-based method.

An overview of LiP25w is shown in Fig. 1. The operation procedure of LiP25w is roughly as follows:

1. User accesses the website of the application shown in Fig. 2(a)(b) through the browser.
2. User inputs a registered user ID and selects a speech movie. At this time, it is possible not only to select a speech movie shot in advance but also to shoot a speech movie by accessing a camera of a smart device from a browser. Since our application is based on the lip reading technology, a speech movie without voice or a speech movie shot in noisy environment is accepted.
3. When the user presses the upload button at the bottom of website (Fig. 2(b)), the user ID and video file are transferred to the server.
4. The server executes the process described later in **3.3**, **3.4**, and **3.5** to the uploaded speech movie to estimate the utterance word.
5. The application automatically generates a Web page that reflects the processing result as shown in Fig. 2(c) using PHP, and presents it to the user.
6. User gets the recognition result.

The display of recognition results will be described in detail in **3.7**.

3.2. Speech movie

Figure 3 shows frame images actually uploaded to our server. Since the smart device and the shooting environment differ depending on each user, the image resolution and the size of the face appearing in the image also differ.

3.3. Feature Points Detection

Our method uses the motion-based features that represent temporal changes of facial feature points. Details of the motion feature will be described later in **3.4**. Here, the facial feature point detection is explained.

Various facial feature point detection methods have been proposed. Furthermore, SDKs and APIs such as Luxand FaceSDK and OpenPose [12] are also provided, and we can use them easily. We apply the method implemented in dlib [13] to

Figure 3: *Uploaded frame images.*

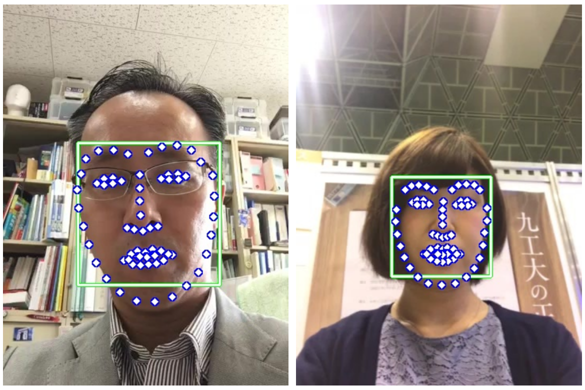


Figure 4: *Detected facial feature points.*

detect 68 facial feature points as shown in Fig. 4. In the figure, the green rectangle is the detected face, and the blue points are the detected face feature points.

3.4. Feature Extraction

Affine transformation for scale and rotation is applied to normalize the facial feature points to reduce differences in distance between camera and face, camera shake, and head movement, based on the detected facial feature points of both eyes.

In the previous research [2], two types of image-based feature and motion-based feature were defined, and after recognizing each feature independently, RNN-based late fusion approach is applied. Although SCAE can obtain high recognition accuracy, SCAE takes several time to process. On the other hand, it has been confirmed that the motion-based feature is simple, however there is no big difference in accuracy compared with the feature using SCAE. In this paper, we adopt an approach that uses only motion-based feature.

The motion-based feature is defined as follows. Among the 68 normalized feature points, the motion-based feature is calculated from the 20 feature points around the lip. A subtraction value between current frame and next frame at each feature point is defined as a motion-based feature by the following equation: $d_*(i,f) = P_*(i,f) - P_*(i,f+1)$, where i is a feature point number, f means a frame number, and $P_*(i,f)$ is a coordinate of i-th feature point. A symbol $*$ is either x or y. This feature is not a distance value; it has negative or positive value. The number of dimensions of this feature is 40.

Table 1: *25 Japanese words*

#	Japanese pronunciation	English
0	/ze-ro/	zero
1	/i-chi/	one
2	/ni/	two
3	/sa-N/	three
4	/yo-N/	four
5	/go/	five
6	/ro-ku/	six
7	/na-na/	seven
8	/ha-chi/	eight
9	/kyu/	nine
10	/a-ri-ga-to-u/	thank you
11	/i-i-e/	no
12	/o-ha-yo-u/	good morning
13	/o-me-de-to-u/	congratulation
14	/o-ya-su-mi/	good night
15	/go-me-N-na-sa-i/	I'm sorry
16	/ko-N-ni-chi-wa/	good afternoon
17	/ko-N-ba-N-wa/	good evening
18	/sa-yo-u-na-ra/	good bye
19	/su-mi-ma-se-N/	excuse me
20	/do-u-i-ta-shi-ma-shi-te/	you are welcome
21	/ha-i/	yes
22	/ha-ji-me-ma-shi-te/	nice to meet you
23	/ma-ta-ne/	see you
24	/mo-shi-mo-shi/	hello

3.5. Recognition

As for the recognition process, Gated Recurrent Unit (GRU) [14], which is a type of RNN, as in the previous research [2]. RNN is an extension of a conventional feedforward NN, which is able to handle a sequential data.

3.6. Database

We collected Japanese word utterance scenes taken with smart device, and built a database SSSD [1]. This database is publicly available database in our website [1]. Table 1 shows the target 25 Japanese words. The first ten words (#0 – #9) are digit words, and remaindering 15 words (#10 – #24) are greeting words.

The number of speakers in SSSD is 24 males and 24 females, for a total of 48 people. The number of utterance scenes used in this paper is (48 speakers) × (25 words) × (30 samples) = (36,000 samples).

3.7. Web application

Our lip reading method is implemented on Linux PC (OS: Ubuntu 16.04 LTS). Its CPU is Intel Core i7-8700K 3.7 GHz, and the GPU is NVIDIA GeForce GTX1080 Ti. We used TensorFlow to build and train the GRU model.

The operation procedure is described in **3.1**. Regarding the recognition result, two candidates, the first candidate and the second candidate, are displayed as shown in Fig. 2(c). Since

[1] http://www.slab.ces.kyutech.ac.jp/SSSD/index_en.html

Table 2: *Breakdown of trail.*

Type	# of trials
(1) Correct result	841
(2) Wrong result	375
(3) Detection failure	19
(4) Unknown word	8
(5) Too short video	14
(6) Too long video	3
(7) Wrong video format	4
Total	1,264

our application assumed to be used as an interface, Open JTalk [2] which is a Japanese text-to-speech system is used to generate audio data (wav-file) based on the recognition result, and the user can be played it on the Web. In order to use our application as the interface, it is not only to display the text of the recognition result on the screen to show it to the conversation partner, but also to convey it as an audio to the conversation partner. This is to make possible not only to show the recognition result on the screen but also to convey it as speech when communicating the recognition results to others.

At the bottom of the recognition result page as shown in Fig. 2(d), there is a form that asks the user whether the recognition result is correct or not. By this, it is possible to count the success or failure of the recognition result and to verify the usefulness of the application.

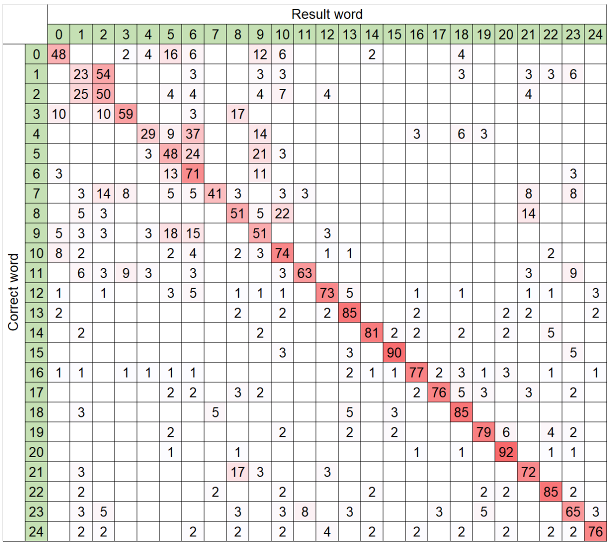
The developed application is accessible from the our laboratory website [3]. You can try using the user ID “AVSP2019”.

4. Evaluation Experiment

Since the recognition method implemented in the application has been evaluated in previous research [2], we evaluate the demonstration trials of the developed application.

The closed beta test (CBT) of our application was conducted from October 8, 2018 with our laboratory members. Furthermore, our application has been released to the public from December 7, 2018. 162 people tried our application for 200 days from the CBT start date to April 26, 2019. Thirteen of the 162 people are speakers included in SSSD speech scene. We introduced our application at several events, such as the domestic conference, and university information sessions for high school students. After releasing our application on last December, some companies and disabled people has registered to our application. Moreover, our application was also introduced in newspapers and radio through press releases from the university. As a result, the number of trials reached 1,264 times. Of all trials, 73.4% of users were male, 26.3% of users were female, and the rest were not speech videos. About half of the trials are by our laboratory members. Since the utterance content is Japanese, most of the people who tried our application are Japanese.

The devices used in many trials was our iPad Pro., however smartphones owned by the user were also used. The image size of the uploaded movie was four types: 360 × 480 [pixel], 480 × 640 [pixel], 540 × 960 [pixel], 720 × 1280 [pixel], and 1080 × 1920 [pixel].



		Result word																								
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Correct word	0	48			2	4	16	6			12	6				2				4						
	1		23	54				3			3	3								3			3	3	6	
	2		25	50			4	4			4	7		4									4			
	3	10		10	59			3		17																
	4					29	9	37			14							3		6	3					
	5					3	48	24			21	3														
	6	3					13	71			11														3	
	7		3	14	8		5	5	41	3		3	3										8		8	
	8		5	3						51	5	22											14			
	9	5	3	3		3	18	15			51			3												
	10	8	2				2	4		2	3	74		1	1									2		
	11		6	3	9	3		3				3	63										3		9	
	12	1		1			3	5		1	1	1		73	5			1		1			1	1		3
	13	2								2		2		2	85			2				2	2			2
	14		2								2					81	2	2		2		2		5		
	15											3			3		90								5	
	16	1	1		1	1	1	1							2	1	1	77	2	3	1	3		1		1
	17						2	2		3	2							2	76	5	3		3		2	
	18		3						5						5		3			85						
	19						2					2			2		2				79	6		4	2	
	20						1			1								1		1		92		1	1	
	21		3							17	3			3									72			
	22		2						2			2				2					2	2		85	2	
	23		3	5						3		3	8		3				3		5				65	3
	24		2	2				2		2		2		4		2		2		2		2		2	2	76

Figure 5: *Confusion matrix.*

As described above, our application inquires the user whether the recognition result is correct or not when displaying the recognition result. The answers to all the trials and the uploaded videos were checked visually. As a result, all trials were classified into seven items shown in Table 2. In the table, (3) is the case which face detection has failed. In our method, a face is detected first, and facial feature points are detected based on it. Therefore, when face detection fails, the feature points are not calculated and the recognition result cannot be obtained. This item includes not only scenes that failed to be detected despite the fact that the face is actually appeared, but also scenes that are not consider to be input, such as a scene in which the face is not appeared or a scene with an oblique face. (4) is the case in which a word not contained in the target 25 words was uttered. Our application does not analyze too short or too long videos to prevent accidental operation. These trials were classified into (5) or (6). (7) is the case which uploaded video file could not be analyzed because of incorrect video format. When only considering (1) and (2), that is, when the assumed speech movie was uploaded, the recognition accuracy was 69.2%.

To analyze recognition accuracy, we calculated a confusion matrix (CM). The calculated CM is shown in Fig. 5. The confusion matrix contains information about correct and result words down by the recognition task. The squares along the diagonal indicate the rate of correct recognition, whereas the squares off the diagonal indicate the rate of incorrect recognition. Since digit words (#0 – #9) have few frames and few sounds, the recognition accuracy of digit words tends to be lower than that of greeting words (#10 – #24). It can be found that #1 (/i-chi/) and #2 (/ni/) is easy to be confused each other.

Next, in order to analyze 375 trials of (2) wrong result in Table 2, all uploaded speech scenes were classified into six types. The results are shown in Table 3. In Table 3, with regard to (2b) to (2e), it is inferred that the recognition method failed because the speech scene was not included in the database used for training. Several scenes including a long waiting time before the start of speech or after the start of speech, and scenes not including a part of the scene at the start of speech are classified into (2f). It can be determined that there were some problems

[2] http://open-jtalk.sourceforge.net/

[3] https://demo.slab.ces.kyutech.ac.jp/VSR/index_en.html

Table 3: *Breakdown of wrong result trail.*

Type	# of trials
(2a) No problem	312
(2b) Large lip motion	15
(2c) Small lip motion	4
(2d) Slow speech	18
(2e) Fast speech	11
(2f) Other problems	15
Total	375

in shooting. Except for (2b) to (2f), 312 trials were recognition failure without problems in the speech scene.

5. Conclusion

In this paper, we developed a practical Web application LiP25w using motion-based feature using RNN that integrates the results of previous researches. Approximately 1200 trials were performed in 200 days, and a recognition accuracy of 73.4% was obtained. The subjects include not only laboratory members but also high school students and the ordinary people. Naturally, these subjects are not included in the speakers of the SSSD. Since our system is a web application, the user can try using smart devices anytime and anywhere. We confirmed that our application is highly practical.

Among the uploaded movies, there were some movies that the camera moving at the time of shooting and some that the face was not seen. One of the future tasks is to implement a mechanism to prompt the user to shoot an appropriate movie when an unexpected movie is uploaded.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16H03211.

7. References

[1] T. Saitoh and M. Kubokawa, “SSSD: Speech scene database by smart device for visual speech recognition,” in *Proc. of International Conference on Pattern Recognition (ICPR2018)*, 2018, pp. 3228–3232.

[2] M. Iwasaki, M. Kubokawa, and T. Saitoh, “Two features combination with gated recurrent unit for visual speech recognition,” in *IAPR Conference on Machine Vision Applications (MVA 2017)*, 2017, pp. 300–303.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” in *European Conference on Computer Vision*, no. 2, 1998, pp. 484–498.

[4] T. Saitoh, “Efficient face model for lip reading,” in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2013, pp. 227–232.

[5] H. L. Bear and R. Harvey, “Comparing heterogeneous visual gestures for measuring the diversity of visual speech signals,” *arXiv preprint arXiv:1805.02948v1*, 2018.

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

[7] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikainen, “Concatenated frame image based cnn for visual speech recognition,” in *ACCV 2016 Workshops, LNCS 10117*, 2017, pp. 277–289.

[8] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in *Asian Conference on Computer Vision (ACCV2016)*, 2016.

[9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6447–6456.

[10] T. Saitoh, “Development of communication support system using lip reading,” *IEEJ Trans. Electrical and Electronic Engineering*, vol. 8, no. 6, pp. 574–579, 2013.

[11] J. Shin, J. Lee, and D. Kim, “Real-time lip reading system for isolated Korean word recognition,” *Pattern Recognition*, vol. 44, no. 3, pp. 559–571, Mar. 2011.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.

[13] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.