Identification of Lip Shape during Japanese Pronunciation Using Deep Learning in Point Cloud Video

Shou Shimizu[1], Ryo Sato[1], Koki Nakamura[1], Akira Taguchi[1], Yue Bao[1]

[1] Division of Informatics, Graduate School of Integrative Science and Engineering, Tokyo City University, Tokyo 158-8557, Japan -g2481428@tcu.ac.jp

Keywords: deep learminmg, point cloud, silent speech, image processing, LiDAR, face recognition

Abstract

People with speech impediments and hearing impairments,whether congenital or acquired, often encounter difficulty in speaking. there is a strong need to actually communicate using vocalization. This aproach needs to pratice and speak correctly. To extract features from data containing individual differences, deep learning methods are utilized. However Prior works extracts features based on point cloud data for lip moving change in three dimensional space, it lacks consideration for temporal sequences.In this work we first identify temporal depth sequences as a new unique sensory information of Japanese pronunciation. We utilized P4Transformer as temporal-spacial model with point clouds. In this study, we performed identification of Japanese pronunciation using point clouds video by machine learning. The accuracy of vowel and consonant identification was estimated to be 96.0% and 33.2% based on the results obtained from experiments.The estimation of vowels was 10% improvement.

1. INTRODUCTION

People with speech impediments and hearing impairments,whether congenital or acquired, often encounter difficulty in speaking (Kariyasu, M. 2016). Speaking is a means of communication. Although it is possible to communicate using sign language,it does not work for everyone.Thus there is a strong need to actually communicate using vocalization. This approach needs to practice and speak correctly.(Hoshina 1978) The hearing impaired however,cannot hear their own voice. It is necessary to interpret mechanism of Japanese pronunciation. Differences in Japanese pronunciation are attributed to variations in air friction caused by differences in the position of the lips, teeth, and tongue.Impediments and hearing impairments also communicate by using lip reading and interpret pronunciation based on lip movements.

Machine lip reading, an automated speech recognition system, based on geometric features such as mouth height, width, area, and perimeter to identify pronunciation, has been actively conducted. For example, sensing modalities include accelerometers(Kwon 2023) and EMG(Deng 2023) acquire lip movements from muscle movements with sensor on the face.This method However depends on specialized equipment,recent advancements in deep learning and the availability of large scale datasets have enabled the development of more sophisticated and accurate techniques for visual feature extraction of pronunciation. These methods employed multiple modalities, the most common one of which is RGB videos(Kondo 2024). Despite usually using RGB video,lip moving change in three dimensional space. In recent years, depth sensing has garnered significant attention duo to have ability to capture 3D information and become more accessible as they are now incorporated into common devices such as the iPhone 12 mini. Prior research on identification of pronunciation(Xue 2024) capture visual data with depth sensor and recognize pronunciation.

Sato(Sato 2022) proposed recognition Japanese pronunciation with point cloud. While this approach extracts features based on point cloud data. However Japanese pronunciation involves movements in the articulators, such as the mouth , it lacks consideration for temporal sequences.In this work we first identify temporal depth sequences as a new unique sensory information of Japanese pronunciation.

In contrast to the alphabet, Japanese, as a 50-sylable language, combine vowel and consonant sounds. In Japanese pronunciation, vowels are typically articulated following consonants, and unique lip shapes and articulatory features are observed in each consonants and vowels.In this work, we evaluate the approach through classification of vowel and consonant classes.

2. PRIOR WORKS

2.1 Image Processing-based Classification Method

Identification of Japanese pronunciation with lip reading is more complicated.The variability such as mouth opening, speed, shape, and detail movement among individuals presents a significant challenge for machine lip reading in visual feature. To extract features from data containing individual differences, deep learning methods are utilized.(Berkol 2024) In particular,in the field of image processing, to owning ability to capture spatial hierarchies in data, convolutional neural network(CNN) based approach are commonly employed. In alphabet visual speech recognition, Lipnet was the first to improve the accuracy of the recognition at the high end-to-end sentence level(Yannis 2016). In this method, With RGB video data,Spatio-temporal Convolutional Neural Networks (STCNN) also known as 3-dimensional CNN(3DCNN) is used to extract visual and temporal sequence feature for lip reading in RGB video. In Japanese pronunciation, Saitoh(Saitoh 2007) proposed a method for pronunciation estimation using RGB images. This method's accuracy is low at 59% due to be indistinguishable from similar lip shapes such as 'a' and 'e' with the only 2D visual information.

2.2 Point Cloud-based Classification Method

Depth sensors have improved drastically in precision and resolution, changing them into a popular sensing solution for a wide array of interactive systems.Specifically,point cloud can represent high density information about geometric details of object's

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-419-2024

shape.Unlike images, point cloud retain three dimensional information and can capture movements around the mouth that cannot be captured in two dimensions.
Sato (Sato 2022) proposed a method using PointNet,representative deep learning model with point clouds to Identification Japanese vowel. PointNet which is trained on 3D point cloud data in the lip region as input, and has shown an accuracy of 85.20 % in the estimation of Japanese vowels.In contrast to grid representations such as 3DCNN, while maintaining invariance to permutations of the input data, involves extracting features of lip shape from inputs that closely represent raw point cloud data. However, this method did not consider temporal feature of the lip moving. In order to handle point clouds video in machine learning, it is necessary to combine spatial feature extractors with models capable of capturing temporal dependencies.

2.3 TimeSeriesProcesssing

Lip shape during Japanese pronunciation affects not only the 3D visual appearance but also exhibits temporal changes over time. For instance, Japanese character 'sa' is compose of vowel 'a', visual appearance of the mouth open and the tongue positioned low in the mouth after consonant "s",visual appearance of closing teeth.
Time series models of machine learning, such as recurrent neural network (RNN) and Long Short Term Memory (LSTM) network, are effective for extracting temporal sequential features. Especially, this model extract continuous changes from the past to the future within time series data. Recently, Bi-RNN and Bi-LSTM, extract features in both directions of past and future have been proposed. Compared with RNN and LSTM, Transformer have substantial advantages in long term processing.
By utilizing attention mechanism including Transformer, the model extract features between non contiguous elements in a sequence. As a result, the attention mechanism allows for flexible feature extraction within the sequence, enabling the model to dynamically focus on relevant elements regardless of their positions In time series processing of visual data, features are initially extracted using CNN layers and subsequently passed to a time series model for sequential analysis. Lip-Interact (Sun 2018) , English 20 command recognition technique with RGB videos as input achieving an accuracy of 96.18% This model composed of CNN and Bi-GRU takes 20 feature points of a user's lip as input. In time series processing of point cloud, features are initially extracted using CNN or PointNet subsequently passed to a time series model such as LSTM, GRU for sequential analysis commonly used, similar to models in RGB image as input. P4Transformer(Fan 2021) composed of 4D Convolutional neural network with point cloud and Transformer achieve acuracy of 90% of classification on 3D recognition tasks such as action recognition and scene flow estimation. Wang(Wang 2024) achieve CER and WER by 4.13% and 8.06%.This method consisted of 4DCNN and GRU and Transformer classification 30 command used in every conversation.
Despite the advancements in prior research, several challenges remain limitations, particularly limitation of previous studies is the inability to effectively distinguish between lip movings during Japanese pronunciation with three dimensional or temporal differences. These limitations indicate a need for improved approaches that can overcome these shortcomings.

3. PROPOSED METHOD

We propose a new identification framework of Japanese pronunciation by learning temporal-spacial feature of lip moving in point cloud video. As shown in Figure 1, the proposed approach consists of several stages, including Data acquisition , Data Preprocessing, and Model Training.

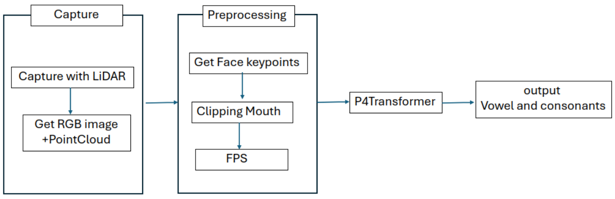

Figure 1. Proposal method Flow

3.1 Data Acquisition

Lip movements are essential component of Japanese pronunciation, such as the opening and closing of the lip, spatial position of the tongue and teeth. we utilized depth sensing as low cost sensing to capture high spatial resolution of depth data in the form of point clouds to reconstruct user pronunciation. In depth sensing, Time-of-Flight (TOF) and stereo cameras are two distinct methods. TOF sensors calculate depth by measuring the time it takes for light to bounce back from an object. In contrast, stereo cameras measure depth using parallax with two or more lenses. In this work,TOF camera are utilized for providing precise distance measurements even in low-light conditions such as oral cavity,ambient brightness. Specifically, We utilize Light Detection and Ranging (LiDAR) , known for its cost-effectiveness and high performance, similar to the one integrated in iPhone 12 mini.(Foix 2011)

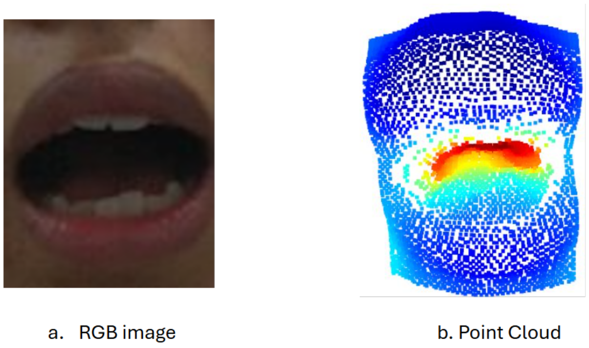
a. RGB image
b. Point Cloud

Figure 2. Capture data

3.2 Data Preprocessing

Identification of Lip Shape during Japanese Pronunciation relies on the fact that human faces have distinctive shape changes resulting from movements of lips, tongue, teeth, and jaw during speech. To prevent factors that may hinder learning, the region of interest (ROI) around the mouth is defined. To prevent factors that may hinder learning, we define the region of interest (ROI) around the mouth and perform lip segmentation.In conventional approaches that utilize the ROI of lips is determined using pretrained machine learning based face detectors, such as Dlib. However, Using a feature extractor for ROI extraction results in a lack of flexibility in adapting to varying conditions,for instance too big face. In this work,we utilize Face Recognition as python library Built using dlib's state of the art face recognition built with deep learning. With this method, we detect face and annotate using the 68 landmark frontal face markup scheme.(Deng 2019).In ROI of English Pronunciation, conventional approaches demonstrated that the region around the mouth, which includes the lips but excludes the jaw and nose, yielded the highest accuracy.(Xue 2024) We extract region around the mouth indicated in the Figure2. Since clipping only the mouth region results in an excessive number of points, downsampling technique is applied. We utilize Farthest Point Sampling (FPS). This approach is particularly effective for capturing the structural shape of objects because it selects points that are maximally distant from the nearest previously selected points, thereby promoting greater spatial separation within the point clouds. Consequently input data are preserved the shape of the tongue, lips, and teeth

3.3 Model Training

Identification of lip shape during Japanese pronunciation include the variability such as mouth opening, speed, shape, and detail movement among individuals presents a significant challenge for machine lip-reading in visual feature,we utilize deep learning. Especially. there are very few deep learning models that take into account both point cloud and temporal sequences. As a result, deep learning model with point clouds usually are combined with some architecture. Due to the limitations observed with CNN-GRU methods in achieving high accuracy solely with visual features, In this work, we utilized P4Transformer as 4DCNN-Transformer based approach with point clouds. The input data is tensor(3 × L × N) as point clouds (x,y,z) and video frame number 'L', N samples par frame with mouth parts extracted by preprocessing. In this work, We conduct experiment with N=4096 sample and L=26 frame. 4DCNN extract local spatio-temporal features with Input transfered grid data and self-attension help the model focus on meaningful spatio-temporal correlations across frames. Finally, multi layer perceptron(MLP) classify vowel or consonant with extracted feature.

4. EXPERIMENTS AND DISCUSSIONS

4.1 Enviroment

The experiment was conducted in the environment illustrated. Data acquisition was performed using an Intel RealSense LiDAR Camera L515. The distance between the subject ' s face and the camera was approximately 30 cm, close to the camera ' s minimum effective range of 25 cm, to maximize detail capture. Capture was conducted in a well lit room without specialized ambient lighting or backgrounds. Given the LiDAR sensor's capability and the targeted extraction of the mouth region during preprocessing, these environmental conditions are not expected to affect the accuracy of the estimation.

4.2 Datasets

To ensure the accuracy and reliability of our proposal method performance, we collected Japanese Pronunciation dataset in our data collection process. Speaker is non-impediments and Japanese without any discernible accents. In conventional dataset, the number of frames per character is limited, which hinders the ability to fully capture the visual dynamics of each articulation. Consequently, the model may inadvertently learn linguistic patterns, including word occurrence probabilities, rather than focusing solely on visual features, thereby deviating from the primary objectives of this study purpose, vocal practice for hearing-impaired people with interpret pronunciation based on lip movements. In addition, The dataset is composed of individual recordings of each of the Japanese character as 50-sylable language to avoid including irrelevant information of natural language such as complex grammar. This datasets include RGB images and point cloud data captured at 30 fps with a resolution of 640 × 480. To establish consistent timing during recording, a metronome was employed to guide the timing of each capture. Especially, participants were instructed to speak in sync with a metronome set to 70 bpmas 0.857 second. As shown in Figure. 4, Japanese character as 50-sylable language compose of 5 vowel as 'a','i','u','e', and 'o', and 9 consonants as 'k', 's', 't', 'n', 'h', 'm', 'y', 'r' and 'w'. However, Some individuals fail to differentiate between the sounds of 'wo' and 'o' in speech and 'nn' Vocalized with closing lip, this work evaluate classification performance using 5 vowel and 8 consonants without 'w'. The data was partitioned into training, validation, and test sets with a distribution of 60%, 20%, and 20%, respectively.

Figure 3. Intel RealSense LiDAR Camera L515

consonant

wa	ra	ya	ma	ha	na	ta	sa	ka	a	Alphabet
わ	ら	や	ま	は	な	た	さ	か	あ	Japanese
	ri		mi	hi	ni	ti	si	ki	i	
	り		み	ひ	に	ち	し	き	い	
wo	ru	yu	mu	hu	nu	tu	su	ku	u	vowel
を	る	ゆ	む	ふ	ぬ	つ	す	く	う	
	re		me	he	ne	te	se	ke	e	
	れ		め	へ	ね	て	せ	け	え	
nn	ro	yo	mo	ho	no	to	so	ko	o	
ん	ろ	よ	も	ほ	の	と	そ	こ	お	

Figure 4. Japanese character as 50-sylable language

4.3 Result

The learning environment used in this study consists of loss function as stochastic gradient descent (SGD), batch size was 8, learning rete=0.01, Python and PyTorch. and Nvidia's CUDA toolkit

	Accuracy(%)	Precision(%)	Recall(%)	F-Measure(%)
vowel	96.0	96.0	96.0	96.0
consonant	33.2	37.2	32.9	34.9

Table 1. Result.

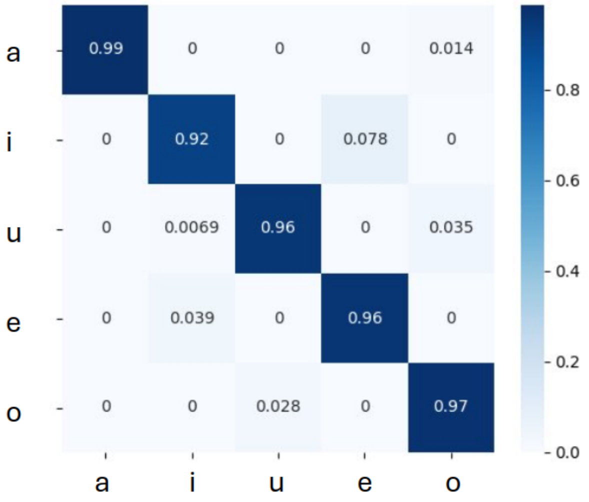

Figure 5. The confusion matrix of vowel

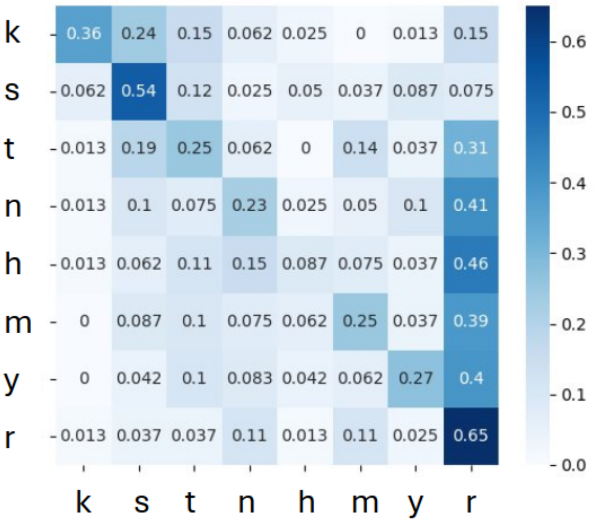

Figure 6. the confusion matrix of consonant

4.4 Discussion

The results of the experiment to verify the accuracy of the Japanese identification are shown in Table 1. The results show that the accuracy of the Japanese vowel is increased by more than 10 percent compared to the previous accuracy of 85.2%. The confusion matrix shown Figure .5 described that the highest accuracy when classifying the vowel ’a’, followed closely by the vowel ’o’. These sounds are characterized by the tongue contacting the base of the mouth, which positions it in a way that allows for greater visibility into the back of the oral cavity. As a result, distinct depth based differences appeared in the depth data, contributing to the model’s ability to identify effectively. Furthermore, the lowest identification was found for ’i’ with 92 % accuracy, which was mistaken for ’e’ with an error rate of about 8 %. Especially, This result frequently misclassified the sound ’ri’ as ’e’ likely due to the similarity in tongue movements for both sounds. Specifically, the ‘ri’ involves the tongue briefly contacting the upper part of the mouth, while the production of ‘e’ similarly requires the tongue to be raised toward the roof of the mouth. This resemblance in articulation

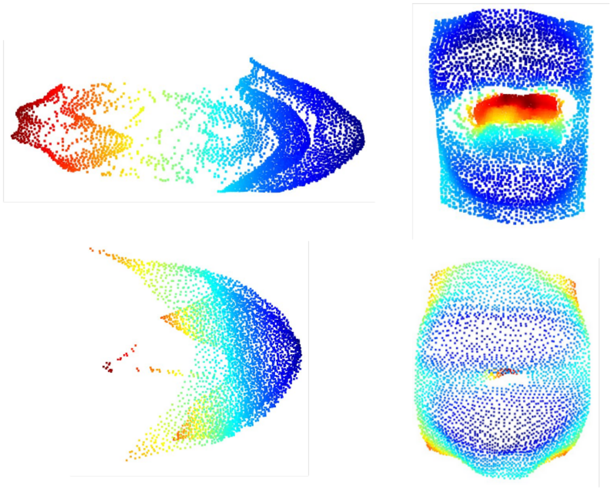

Figure 7. point cloud of vowel ’a’(up),’o’(down). Top-down view(left), front view(right)

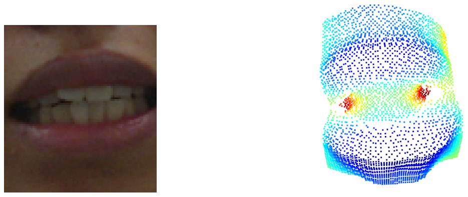

Figure 8. ”i” 20 frame

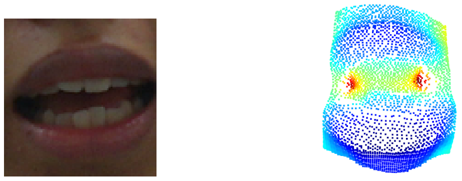

Figure 9. ”e” 20 frame

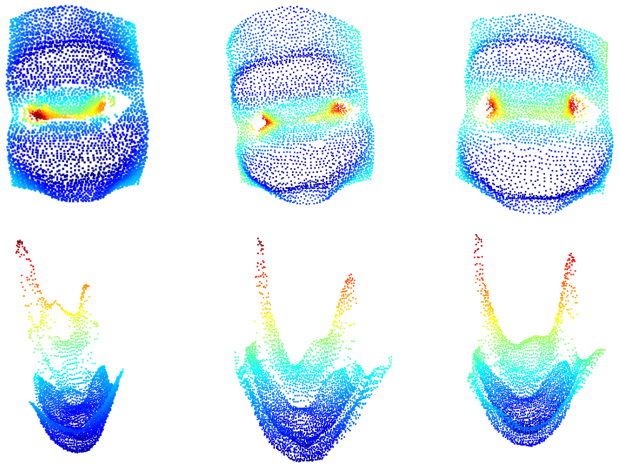

Figure 10. point cloud of frame start of ’ri’ as ’r’(left),’frame end of ’ri’ as ’i’ (center), vowel ’e’(right), Top-down view(down), front view(up)

can create visual ambiguities, leading to challenges in accurately distinguishing between the two sounds.

In the results of consonant's identification, due to the lack of existing studies focusing on the estimation of Japanese consonants, a direct comparison with prior work is not feasible. Therefore, we report the achieved accuracy of our model to provide a baseline for future research in this area. Our approach yielded an accuracy of 33.2%, demonstrating the potential effectiveness of this model for Japanese consonant identification and offering a preliminary benchmark for similar tasks. The confusion matrix of Japanese consonant shown Figure .6 described that the highest accuracy when classifying the consonant 'r', followed closely by the consonant 's', The high accuracy observed for the consonant 'r' can be attributed, in part, to the model ' s tendency to predict 'r' more frequently than other consonants. This prediction bias may inflate accuracy for 'r,' as the model is more likely to classify ambiguous cases as 'r,' which increases the overall count of correct 'r' predictions. This trend suggests that the model may favor certain classes, impacting the balance of recognition across different consonants. The production of the 's' involves a close positioning of the teeth, creating a narrow passage for airflow to pass through. This positioning generates a characteristic frictional noise as the air is forced between the teeth, which gradually transitions into the vocalization of a following vowel. This friction driven articulation makes 's' acoustically distinct and can be challenging for models to accurately capture, especially as it requires precise tracking of both the airflow and the subsequent vowel transition. Moreover, the consonant 'h' is lowest accuracy. 'h' as a voiceless fricative is produced with minimal visible movement in surroundings of the mouth, as it primarily involves the friction of breath against the inner surfaces of the oral cavity rather than articulated movements of the tongue, lips, or teeth. Due to the lack of distinctive oral movements, 'h' is challenging to capture visually, as its production does not create prominent visual features that models can easily detect.

Overall, classes with pronounced visual changes achieved higher accuracy, indicating that the model more effectively recognizes distinctive visual cues. Conversely, subtle depth variations were not consistently captured, suggesting limitations in the model's ability to distinguish finer depth-based distinctions.

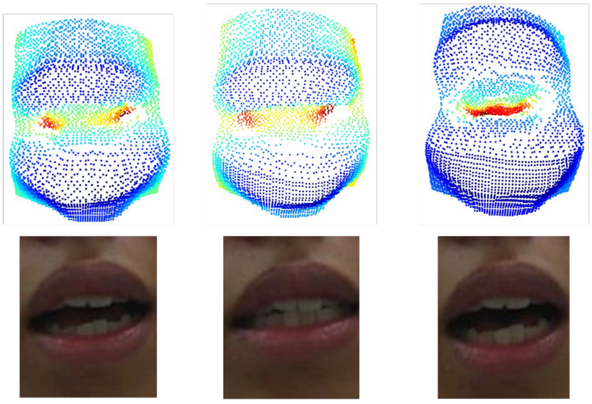

Figure 11. "sa"

5. Conclusion

We present the first identification approach of lip shape during Japanese pronunciation using deep learning in point cloud Video. In this work, we identified depth sensing thinking about time series as the new advantageous information source for Identification of Japanese pronunciation dataset for Japanese pronunciation and evaluated. The identification of vowels showed an accuracy 96.0% and consonants showed a 33.2%.The estimation of vowels is 10% improvement compared to the previous accuracy of 85.2%.

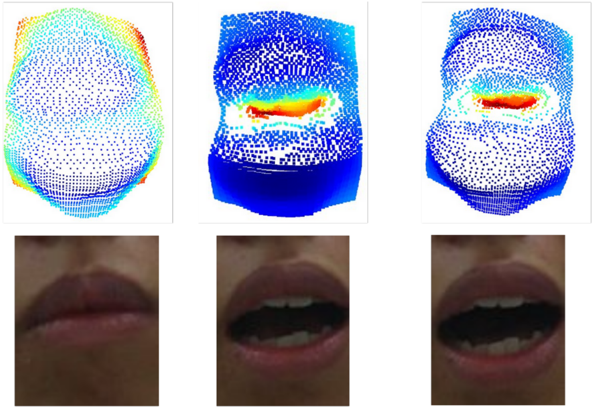

Figure 12. "ha"

Below we list our key contributions:

- identified depth sensing thinking about time series as the new advantageous information source for Identification of Japanese pronunciation
- dataset for Japanese pronunciation and evaluated to classify vowel and consonant
- approach to identification of Japanese pronunciation through visual comprehension

References

Kariyasu, M.; Toyama, M.; Matsuhira, Y. Epidemiology of Communication Disorders—Prevalence and Estimates.*Bull. Fac. Health Med. Sci.* 2016, *1*, 1–12. https://doi.org/10.20558/00001201

Hoshina, N. A Study of Phonetic Functions of Hearing Impaired Students. *Niigata Med. J.* 1987, *101*, 577–593. Available online: https://hdl.handle.net/10191/36694 (accessed on 4 September 2024).

Kwon, Jinuk, et al. "Novel three-axis accelerometer-based silent speech interface using deep neural network." *Engineering Applications of Artificial Intelligence 120* (2023):105909. https://doi.org/10.1016/j.engappai.2023.105909

Deng, Zhihang, et al. "Silent speech recognition based on surface electromyography using a few electrode sites under the guidance from high-density electrode arrays." *IEEE Transactions on Instrumentation and Measurement 72* (2023): *1-11*. doi:10.1109/TIM.2023.3244849

Kondo, Fumiya, and Satoshi Tamura. "Inter-language Transfer Learning for Visual Speech Recognition toward Under-resourced Environments." *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING* 2024. https://aclanthology.org/2024.sigul-1.19

Xue Wang, Zixiong Su, Jun Rekimoto, Yang Zhang ”Watch Your Mouth: Silent Speech Recognition with Depth Sensing” *CHI ’24: Proceedings of the CHI Conference on Human Factors in Computing Systems Article No.: 323, Pages 1* -(2024). https://doi.org/10.1145/3613904.364209

Sato, Yoshihiro, and Yue Bao. ”Identification of 3D Lip Shape during Japanese Vowel Pronunciation Using Deep Learning.” *Applied Sciences 12.9* (2022): 4632. https://doi.org/10.3390/app12094632

Berkol, Ali, Talya Tümer Sivri, and Hamit Erdem. ”Lip Reading Using Various Deep Learning Models with Visual Turkish Data.” *Gazi University Journal of Science* (2024): *1-1.* https://doi.org/10.35378/gujs.1239207

Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. *Lipnet: Sentence-level lipreading. arXiv preprint arXiv:1611.01599 2, 4*(2016). https://doi.org/10.48550/arXiv.1611.01599

Saitoh, T.; Hisagi, M.; Konishi, R. ”Analysis of Features for Efficient Japanese Vowel Recognition” *IEICE Trans. D* 2007, 90.1889–1891.doi:10.1093/ietisy/e90-d.11.1889

K. Sun, C. Yu, W. Shi, L. Liu and Y. Shi, ”Lip-Interact: Improving mobile device interaction with silent speech commands”,*Proc. 31st Annu. ACM Symp. User Interface Softw. Technol., pp. 581-593*, 2018. https://doi.org/10.1145/3242587.3242599

H. Fan, Y. Yang and M. Kankanhalli, ”Point 4D transformer networks for spatio-temporal modeling in point cloud videos”, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 14199-14208, 2021. https://doi.org/10.1109/cvpr46437.2021.01398

Foix, Sergi, Guillem Alenya, and Carme Torras. ”Lock-in time-of-flight (ToF) cameras: A survey.” *IEEE Sensors Journal 11.9* (2011): 1917-1926.doi:10.1109/JSEN.2010.2101060

Deng, Jiankang, et al. ”The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking.” *International Journal of Computer Vision 127* (2019): 599-624. https://doi.org/10.1007/s11263-018-1134-y