3Dp1-7L
(Late-News Paper)

Development of Spoken Language Identification System using Directional Volumetric Display

Mitsuru Baba[1], Naoto Hoshikawa[2], Hirotaka Nakayama[3], Tomoyoshi Ito[1], Atsushi Shiraki[1]

mbaba41045@chiba-u.jp
[1] Graduate School of Engineering, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan
[2] Department of Innovative Electrical and Electronic Engineering, National Institute of Technology, Oyama College, 771 Nakakuki, Oyama, Tochigi 323-0806, Japan
[3] Center for Computational Astrophysics, National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
Keywords: Volumetric displays, Digital signage, Machine learning.

ABSTRACT

In this study, we develop the spoken language identification system by a directional volumetric display. We proposed the identification model using CNN for four languages, English, French, German, and Spanish. The identification accuracy was 93.4%, and became possible to display an image in the observer's direction.

1 INTRODUCTION

A three-dimensional data visualization technology for displaying images in space is being researched with the advancement of augmented reality and virtual reality, and so on technology[1-3].

Our research group has developed a directional volumetric display that can display different images depending on the viewing direction by using the technology of recording multiple images in the same space[4,5]. A directional volumetric display can display images in the direction specified by the program in advance and has the concealment that meaningful information cannot be obtained from the display surface other than the specified direction. Therefore, if the image can be displayed according to the observer's position, it can be expected to be used as a practical and highly entertaining display. Overview of a directional volumetric display is shown in Fig.1.

In this study, to apply a directional volumetric display to multilingual digital signage, we develop and evaluate a spoken language identification system using a directional volumetric display that displays images according to the observer's speaking language.

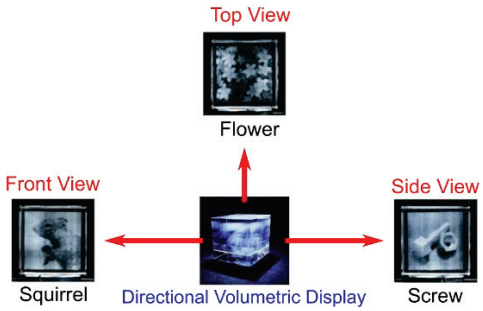


Fig. 1 Overview of a directional volumetric display

2 METHOD

In this section, to describe the development of a spoken language identification system using machine learning to display images according to the observer's speaking language.

2.1 Spoken Language Identification using CNN

With the advent of unmanned assistant systems like a smart speaker, interest in speech identification systems using deep learning has increased in recent years. As a spoken language identification model using Convolutional Neural Network (CNN), S. Mukherjee et al., using five layers of CNNs, have an accuracy of 92.7% for three languages of German, English, and Spanish[6]. Additionally, H. Mukherjee et al., using two layers of CNNs, have an accuracy of 95.5% for seven languages of Bengali, Marathi, Telugu, Tamil, Malayalam, Kannada, and Hindi[7]. In Recurrent Neural Network (RNN), C. Bartz al., using five CNN blocks and Long Short-Term Memory, have an accuracy of 91.0% for four languages of English, French, German, and Spanish[8]. In addition, S. Jauhari et al., using four CNN blocks and Bi-directional Gated Recurrent Unit, have an accuracy of 95.4% for six languages of English, French, German, Spanish, Russian, and Italian[9]. In this study, we propose four language identification models: English, French, German, and Spanish, using CNN, which can be learned quickly compared to RNN.

We propose a spoken language identification model consisting of four CNN blocks and a fully connected layer, including the output layer, concerning the configuration of a spoken language identification model proposed in reference[6,7]. One CNN block consists of four processes: a convolutional layer that extracts features from an image, a batch normalization, an activation layer by Rectified Linear Unit, and a maximum pooling layer that reduces the feature map. A dropout layer is introduced in a fully connected layer to improve generalization performance. Classification is performed based on the features extracted through four CNN blocks, and the classification probability is output via the softmax

function. CNN model architecture is shown in Fig.2.

For the dataset, we use the wav format voice samples, which are provided in VoxForge[10] sampled at 8 kHz, 22000 voice samples per language are randomly extracted, and a dataset is divided into 16000 training data, 3000 validation data, and 3000 test data. The voice samples used for learning have an average playback time of 5.7 seconds, so data are processed so that playback time is 6.0 seconds and converted to a log Mel spectrogram.

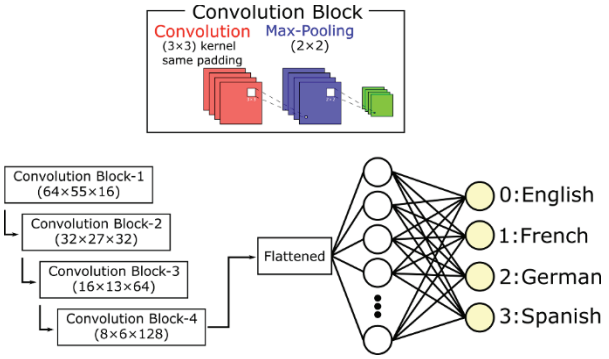


Fig.2 CNN model architecture

2.2 A Directional Volumetric Display

In this study, we adopt a directional volumetric display with a thread and a projector as a video presentation system, displaying a color video and enabling the system to be upsized. Since a directional volumetric display shows three-dimensional data represented by voxels, it is necessary to calculate the voxel value according to the display direction. In three-dimensional coordinates in the right-handed coordinate system, if the observer's position angle for the front direction is, the coordinates can be expressed by equation (1) from the rotation of the y-axis.

$$\begin{bmatrix} i' \\ j' \\ k' \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} i \\ j \\ k \end{bmatrix} = \begin{bmatrix} i\cos\theta + k\sin\theta \\ j \\ -i\sin\theta + k\cos\theta \end{bmatrix} \quad (1)$$

Hence the image in the front direction is, the image in the observer direction θ is, the voxel value is, and the normalization constant is, the voxel value is expressed by equation (2). Creation of projection image is shown algorithm in Fig.3.

$$V_{x,y,z} = \lambda * I_A(i,j,0) * I_B(i\cos\theta + k\sin\theta, j, 0) \quad (2)$$

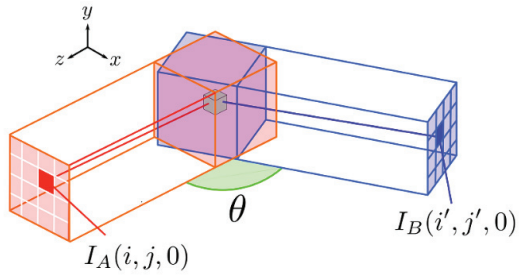


Fig. 3 Creation of projection image algorithm

2.3 System Configuration

In the system to be developed, Kinect V2 is used to locate and acquire the voice to identify the spoken language and display the result of the language identification toward the observer. Besides, voice samples of each language that are not used for learning are sent from the speaker system and let it as the identification target voice.

The size of a directional volumetric display to be constructed is 0.95[m] in-depth, 0.95[m] in width, and 1.87[m] in height; a total of 211 threads are arranged on the whiteboard is installed on the upper part of the frame.

3 EXPERIMENT

We evaluate the development system by performing spoken language identification of the sound emitted from the speaker system within the range of the front and side directions and projecting an image that results as a character string of identification language toward the speaker system's direction. As an initial state of the development system, the English letters "NONE" are displayed on the display surface in the front direction and the directional volumetric display's side direction. By operating the Graphical User Interface software on the host computer (Host PC), the sound produced by the speaker system within the measurement range is acquired, and the projection image according to the spoken language identification result is displayed in the direction of the speaker system.

4 RESULTS

The proposed CNN model was trained under the conditions of batch sizes 64, epochs were 30, and using Adam, an optimized algorithm. The proposed model's accuracy using test data was 92.6% for English, 87.4% for French, 96.8% for German, and 96.5% for Spanish. In the proposed model, the probability of predicting English when the true label data was French was 6.47%, and there was a tendency for English and French to be confused slightly. The accuracy of the four languages was 93.4%. The confusion matrix for the classification of four languages is shown in Fig.4.

The overview of the development system is shown in Fig.5. The projection image and original image in each direction are shown in Fig.6 and Fig,7. The frame rate of the projection images was 10 frame per seconds. Fig. 6(a) and Fig.7(a) were the display results when the English speech was spoken in the front direction, and the letters "ENGLISH" were projected in the front direction($\theta = 0$). Fig. 6(b) and Fig.7(b) were the display results when the Spanish speech was spoken in the side direction, and the letters "SPANISH" were projected in the side direction($\theta = 45$). Fig. 6(c) and Fig.7(c) were the display results when the French speech was spoken in the side direction, and the letters "FRENCH" were projected in the side direction($\theta = 90$). Fig. 6(d) and Fig.7(d) were the display results when the German

speech was spoken in the side direction, and the letters "GERMAN" were projected in the side direction($\theta = 90$). A meaningful image could not be confirmed from the direction other than the projection display surface in the results.

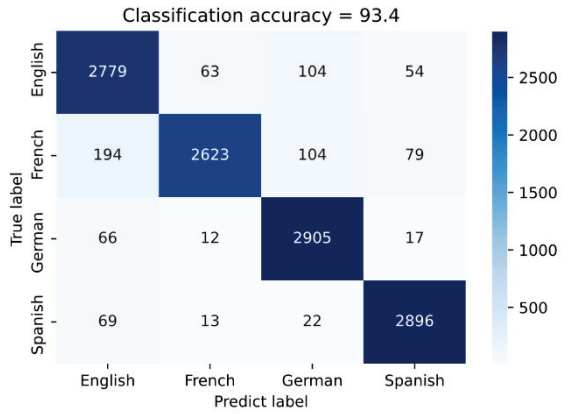


Fig. 4 The confusion matrix for the classification of four languages

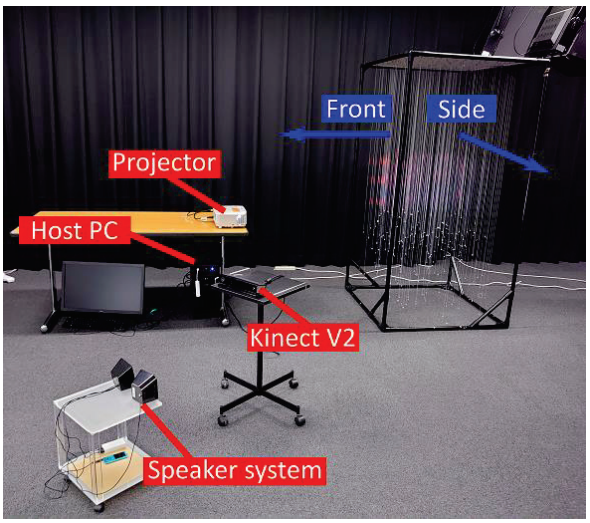
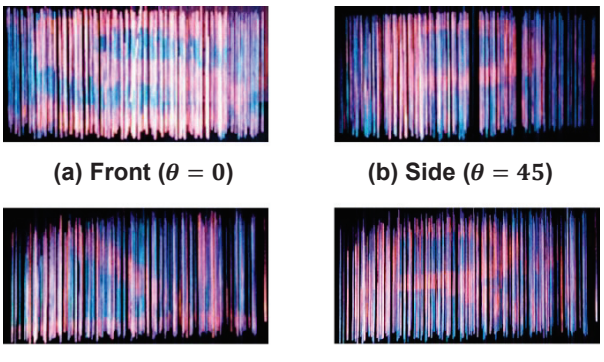


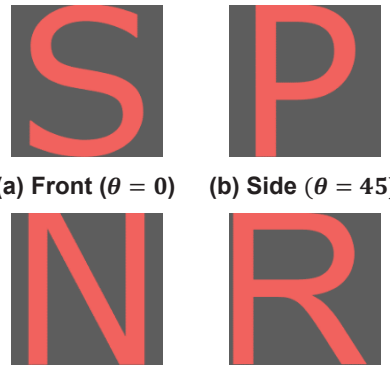
Fig. 5 Overview of development system



(a) Front ($\theta = 0$) **(b) Side ($\theta = 45$)**

(c) Side ($\theta = 90$) **(d) Side ($\theta = 90$)**

Fig. 6 Projection image



(a) Front ($\theta = 0$) **(b) Side ($\theta = 45$)**

(c) Side ($\theta = 90$) **(d) Side ($\theta = 90$)**

Fig. 7 Original image

5 DISCUSSION

It was performed spoken language identification according to the voice emitted from the speaker system and confirmed that a directional volumetric display displayed the spoken language identification result as an image in the direction. However, while the result of spoken language identification was displayed as a projected image, there was a difference in the identification probability output by the proposed CNN model between the original voice sample and the voice sample emitted by the speaker system. This is due to white noise, which is not seen in the voice samples provided by VoxForge because the measurement environment includes operating sounds of servers and outdoor units for business in addition to voice samples. Therefore, to improve the development system's robustness, datasets include the white noise of the measurement environment to learn the spoken language identification model. Also, necessary to take measures such as removing white noise from the input voice sample.

6 CONCLUSIONS

In this study, to apply the directional volumetric display to multilingual digital signage, we developed and evaluated the spoken language identification system by a directional volumetric display, displaying images with directionality according to the speaking language of the observer. To identify the observer's spoken language, we proposed a language identification model using CNN for English, French, German, and Spanish, and obtained 91.9% accuracy for the test data. Besides, using the proposed spoken language identification model and directional volumetric display, we developed multilingual digital signage that displays images according to the language emitted by the observer, and it became possible to display images with directionality. However, there are problems that the identification rate of the spoken language identification model decreases due to the influence of white noise that is not confirmed in the data set. Additionally, the frame rate of the projection image is low because the algorithm processing of voxel

calculation is not optimized.

Further studies are required to propose a multilingual spoken identification model with white noise in the measurement environment and consider the projection image's processing method for the high resolution of a directional volumetric display. Also, apply the image quality improvement algorithm in the directional volumetric display proposed in reference[11] to the projection image.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 18K11599.

REFERENCES

[1] D. E. Smalley, E. Nygaard, K. Squire, J. V. Wagoner, J. Rasmussen, S. Gneiting, K. Qaderi, J. Goodsell, W. Rogers, M. Lindsey, K. Costner, A. Monk, M. Pearson, B. Haymore, and J. Peatross, “A photophoretic-trap volumetric display”, Nature 553, pp.486–490, 2018.

[2] K. Rathinavel, H. Wang, A. Blate, and H. Fuchs, “An Extended Depth-at-Field Volumetric NearEye Augmented Reality Display”, in IEEE Transactions on Visualization and Computer Graphics, Vol.24, No.11, pp.2857-2866, Nov.2018.

[3] M. D. Medeiros, J. Nascimento, J. Henriques, S. Barrao, A. F. Fonseca, N. A. Silva, N. M. Coelho, and V. Agoas, “Three-Dimensional Head-Mounted Display System for Ophthalmic Surgical Procedures”, Retina, Vol. 37, Issue 7, pp.1411-1414, July.2017.

[4] H. Nakayama, A. Shiraki, R. Hirayama, N. Masuda, T. Shimobaba, and T. Ito, “Three-dimensional volume containing multiple two-dimensional information patterns”, Scientific Reports, Vol.3, No.1931, pp.1-5, 2013.

[5] A. Shiraki, M. Ikeda, H. Nakayama, R. Hirayama, T. Kakue, T. Shimobaba, and T. Ito, “Efficient method for fabricating a directional volumetric display using strings displaying multiple images”, Applied Optics, Vol.57, No.1, pp.A33-38, 2018.

[6] S. Mukherjee, N. Shivam, A. Gangwal, L. Khaitan. and A. J. Das, “Spoken Language Recognition using CNN”, 2019 International Conference on Information Technology (ICIT), pp.37-41, Dec. 2019.

[7] H. Mukherjee, S. Ghosh, S. Sen, O. M. Sk, K. C. Santosh, S. Phadikar, and K. Roy, “Deep learning for spoken language identification: Can we visualize speech signal patterns?”, Neural Comput & Applic 31, pp.8483–8501, Sep.2019.

[8] C. Bartz, T. Herold, H. Yang, and C. Meinel, “Langugage Identification Using Deep Convolutional Recurrent Neural Networks”, In International Conference on Neural Information Processing, pp.880-889, Aug.2017.

[9] S. Jauhari, S. Shukla, and G. Mittal, “Spoken Language Identification using ConvNets”, European Conference on Ambient Intelligence, pp.252-265, 2019.

[10] voxforge.org. Free speech recognition (linux, windows and mac)-voxforge.org, “https://www.voxforge.org/”, accessed on 23 July 2020.

[11] A. Shiraki, D. Matsumoto, R. Hirayama, H. Nakayama, T. Kakue, T. Shimobaba, and T. Ito, "Improvement of an algorithm for displaying multiple images in one space," Applied Optics, Vol. 58, Issue. 5, pp. A1-A6, 2019.