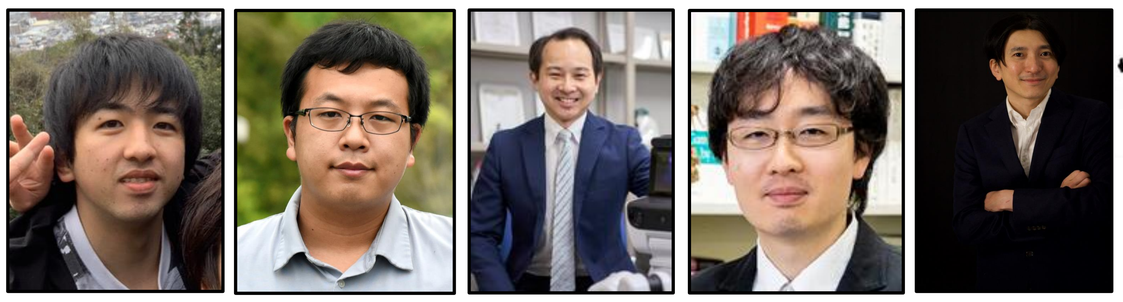


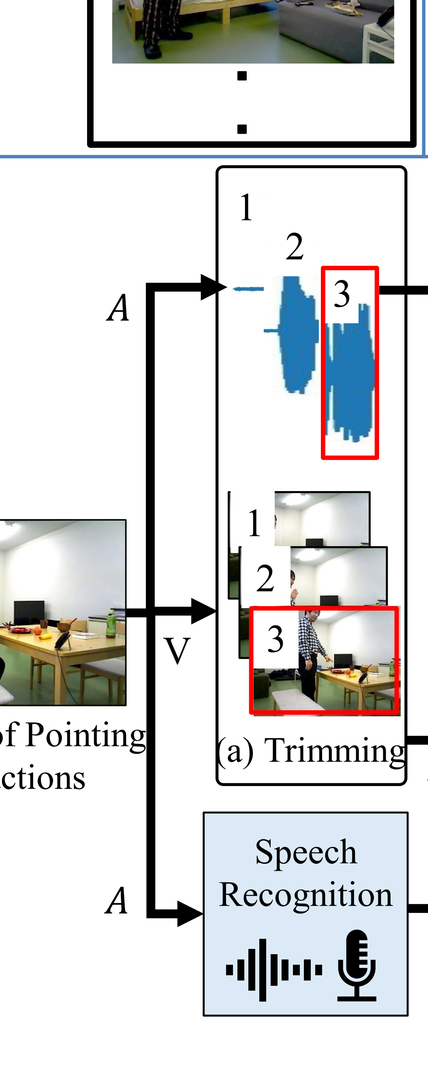
H. Nakagawa (Ritsumeikan Univ. Japan), S. Hasegawa* (Ritsumeikan Univ. Japan), Y. Hagiwara (Soka Univ. / Ritsumeikan Univ. Japan), A. Taniguchi (Ritsumeikan Univ. Japan), T. Taniguchi (Kyoto Univ. / Ritsumeikan Univ. Japan)

Pointing Frame Estimation with Audio-Visual Time Series Data for Daily Life Service Robots

INTRODUCTION

People often give instructions with pointing. Pointing is important information for identifying objects [1], **but it is difficult to know when pointing is given.**

Therefore, the robot needs to capture the timing when the person points (**pointing frame**).



PREVIOUS RESEARCH

Embodied Reference Understanding Framework [2]

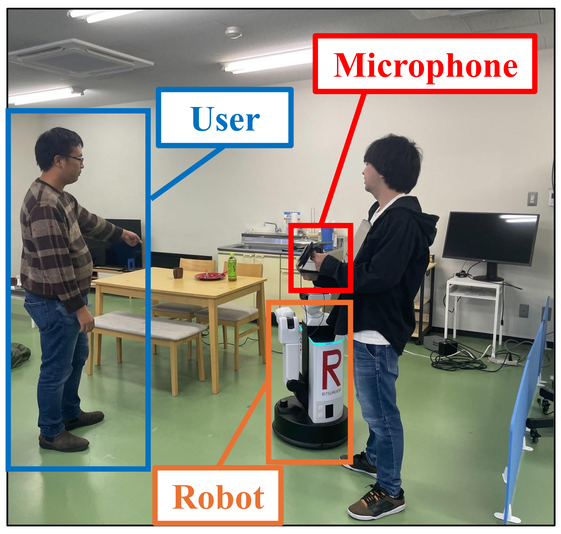
Gestures unrelated to user instructions are not addressed, which may reduce estimation accuracy.

We address this issue by improving the pointing frame estimator in previous research.

PURPOSE

To verify the extent to which **the performance of pointing frame estimations can be enhanced through the integration of speech data** during human-robot interactions involving gestures and language.

PROPOSED METHOD



(b) Performing Data Collection

EXPERIMENT AND RESULT

We tested how well a model with voice information improves pointing frame estimation in a scenario that simulates actual human-robot communication.

Evaluation Items

$$F_1 score = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Condition					
	User Posture	Upright Posture		✓	
		No upright posture and behavior other than pointing	✓		✓
	Placement of Reference Object	Inside of an Image	✓		
		Outside of an Image		✓	✓
Pointing Frame Estimator [4]			0.800 ± 0.029	0.857 ± 0.026	0.797 ± 0.020
Proposed Model	Speech Segment	Transformer	0.809 ± 0.024	0.853 ± 0.019	0.794 ± 0.006
		Bi-LSTM	0.702 ± 0.034	0.814 ± 0.014	0.761 ± 0.019
	MFCC	Transformer	0.847 ± 0.015	0.878 ± 0.017	0.822 ± 0.015
		Bi-LSTM	0.719 ± 0.007	0.817 ± 0.004	0.758 ± 0.009
	Mel-Spectrogram	Transformer	$\mathbf{0.866 \pm 0.014}$	$\mathbf{0.886 \pm 0.003}$	$\mathbf{0.832 \pm 0.011}$
		Bi-LSTM	0.748 ± 0.006	0.824 ± 0.004	0.766 ± 0.010

CONCLUSION AND FUTURE WORK

- Since the speech data was complex, higher-order features such as Mel-Spectrogram and MFCC may have been more effective.
- Incorporate methods for estimating the location of objects in the environment using pointing frames [3] and robot action planning [4].

REFERENCES & ACKNOWLEDGEMENTS

[1] N. Kotani et al. "**Point Anywhere: Directed Object Estimation from Omnidirectional Images**." ACM SIGGRAPH, 2023.
[2] Y. Chen et al. "**YouRefIt: Embodied Reference Understanding with Language and Gesture**." ICCV, 2021.
[3] A. Oyama et al. "**Exophora Resolution of Linguistic Instructions with a Demonstrative based on Real-World Multimodal Information**." IEEE RO-MAN, 2023.
[4] S. Hasegawa et al. "**Integrating Probabilistic Logic and Multimodal Spatial Concepts for Efficient Robotic Object Search in Home Environments**." SICE JCMSI, 2023.

This work was supported by JSPS KAKENHI Grants-in-Aid for Scientific Research (Grant Numbers JP23K16975 and JP22K12212), JST Moonshot Research & Development Program (Grant Number JPMJMS2011), "Society for the Advancement of Science and Technology at Ritsumeikan", and JST SPRING, Grant Number JPMJSP2101.