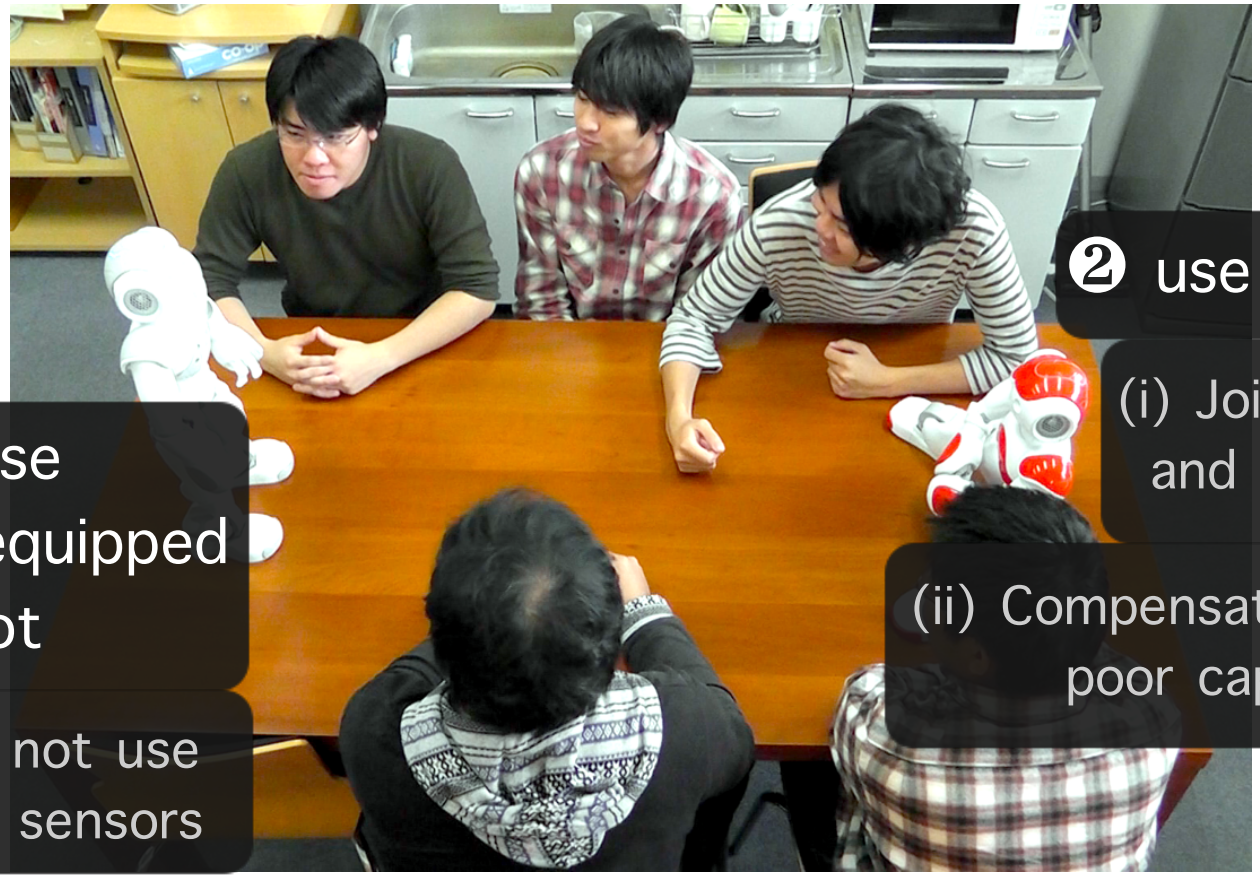
Integration of Multiple Sound Source Localization Results for Speaker Identification in Multi-party Dialogue System

Graduate School of Engineering, Nagoya University
Taichi Nakashima, Kazunori Komatani, Satoshi Sato

Goal "Implementing multi-party dialogue system"

interacts with more than two users



❶ only use sensors equipped on a robot

i.e., Do not use special sensors

❷ use two robots

(i) Join conversation and keep it going

(ii) Compensate each other's poor capabilities.

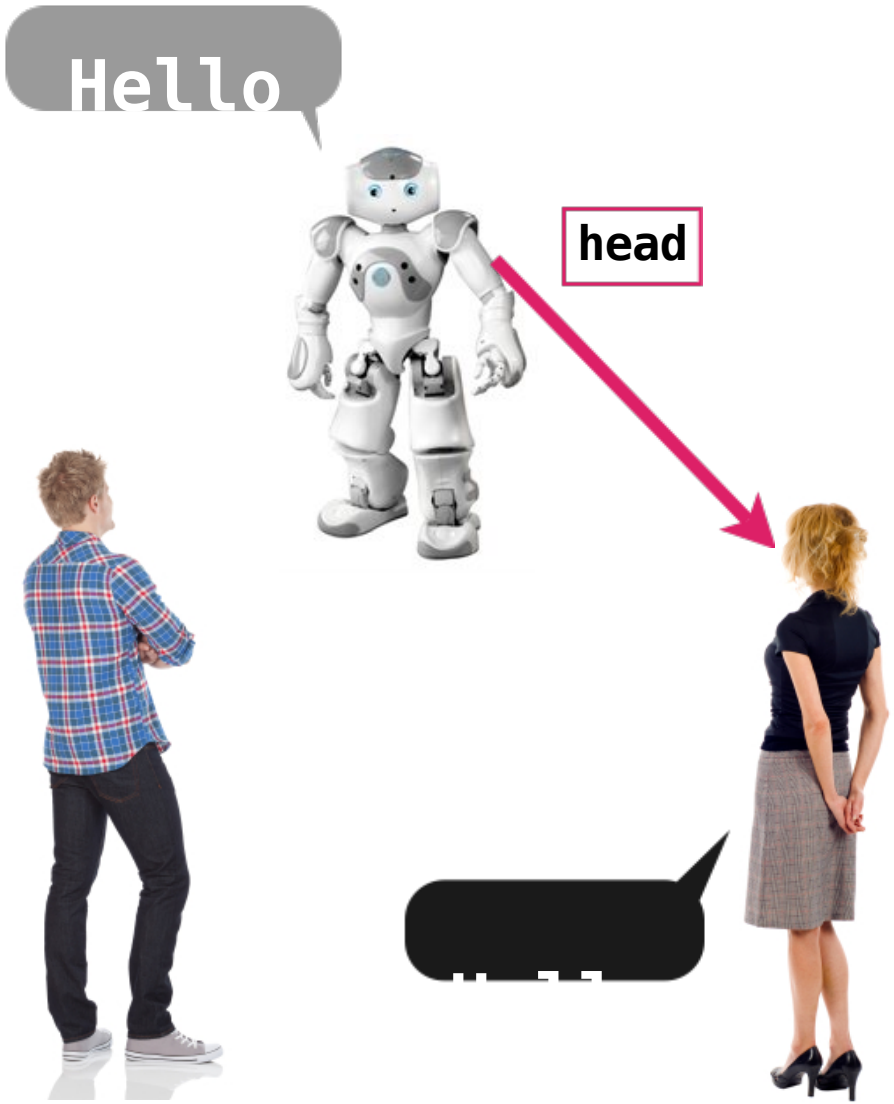
Multiple users sit around a table

→ Simplify the problem to decide the positions of users

i.e., the positions of users are naturally narrowed down

Speaker Identification

identifying where speaker is



Heading toward the user
to answer his/her questions

This behavior enables users

→ to understand role of addressee [Mutlu, 2009]

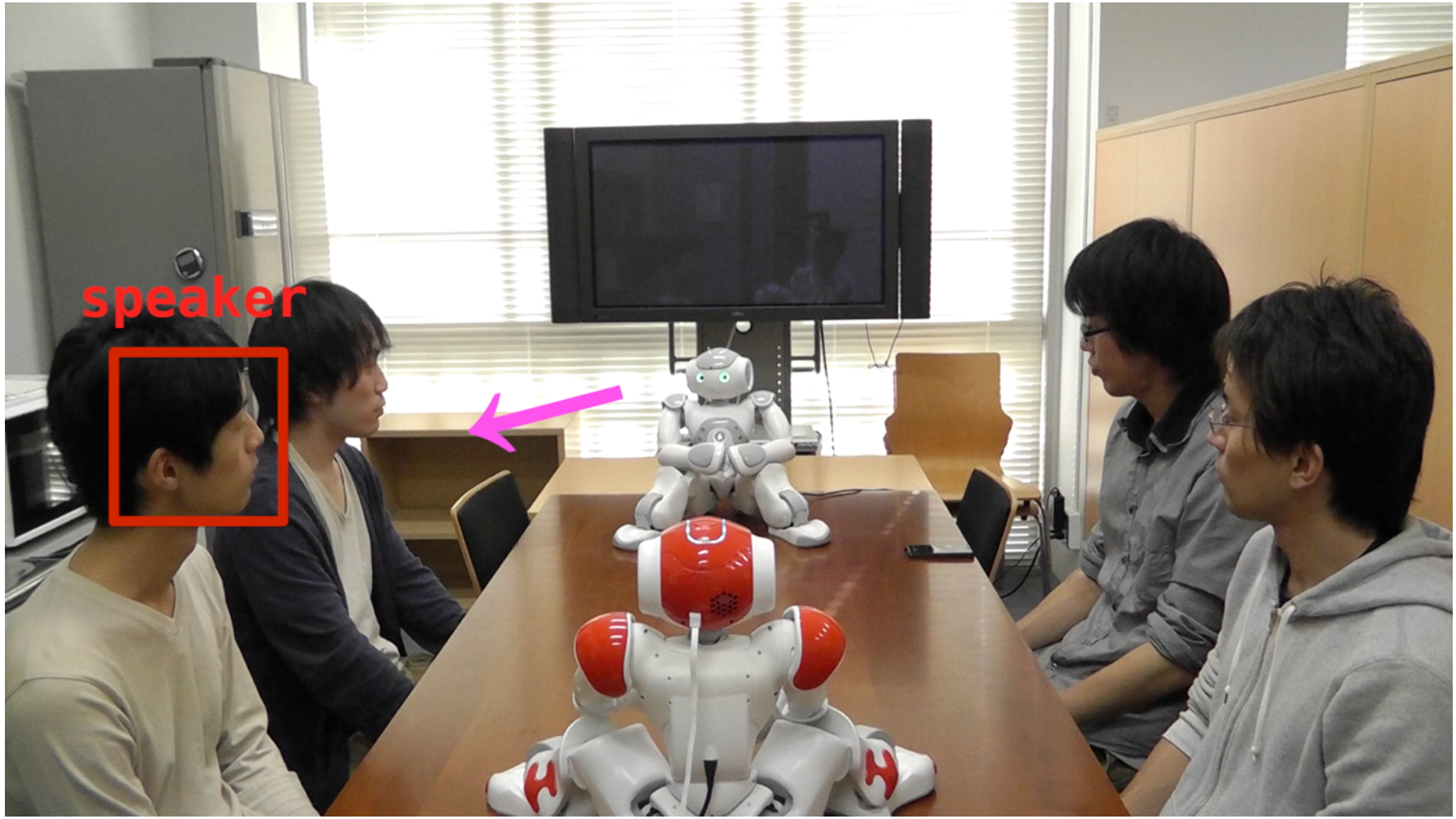
→ to feel involved in conversation [Bennewitz, 2005]

▼

We use sound source localization results
to identify a speaker

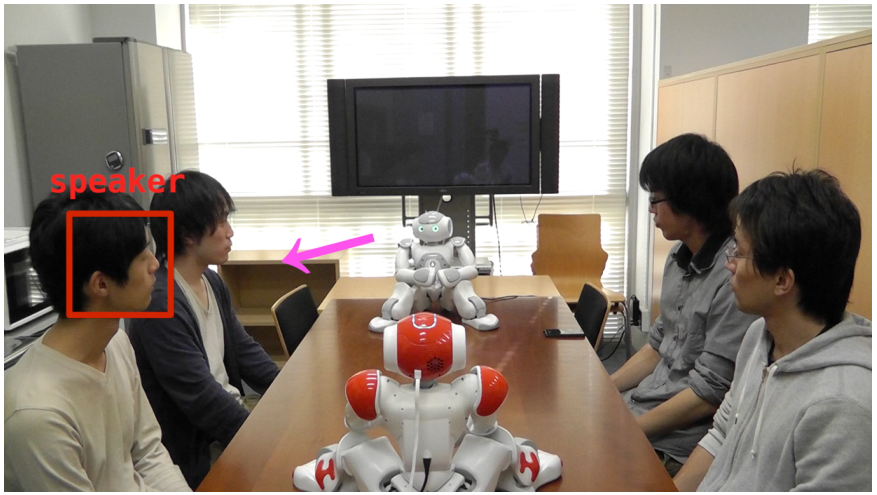
Construction of Demo System

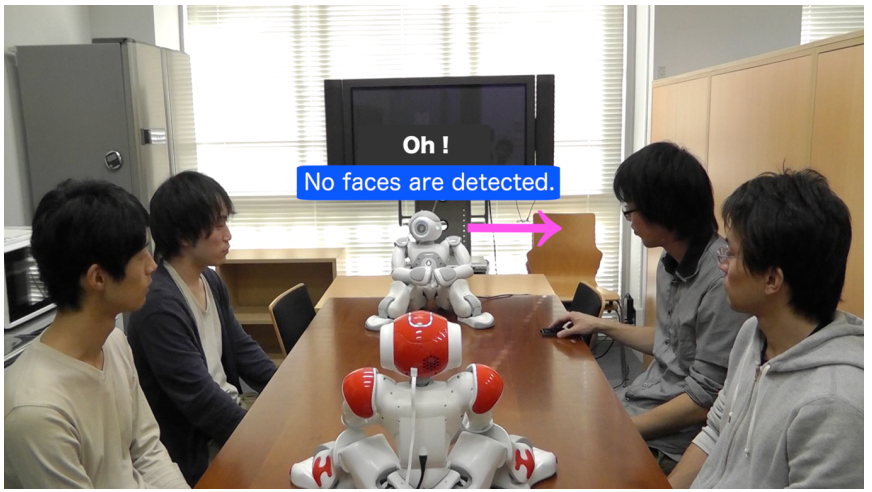
Demo system identifies a speaker



Construction of Demo System

Demo system identifies a speaker





1. Identifying a speaker and heading toward to answer his/her question

   → **integrating multiple sound source localization results**

2. Executing face detection to check whether a speaker exists

   → **using power as a confidence measure of localization results**

3. Two robots talk with each other when users stop talking

Outline

1. Background
2. Demo System
3. Related Work of Speaker Identification
4. Problems of Sound Source Localization
5. Solutions
   - 5-1. Inputs and Outputs of Our Method
   - 5-2. Integration of Multiple Sound Source Localization Results
6. Evaluation Experiments
   - 6-1. Results of identifying loudspeakers - Using only one robot / Integration -
   - 6-2. Localization results by Power
7. Conclusion & Future Work

Outline

1. Background
2. Demo System
3. Related Work of Speaker Identification
4. Problems of Sound Source Localization
5. Solutions
   - 5-1. Inputs and Outputs of Our Method
   - 5-2. Integration of Multiple Sound Source Localization Results
6. Evaluation Experiments
   - 6-1. Results of identifying loudspeakers - Using only one robot / Integration -
   - 6-2. Localization results by Power
7. Conclusion & Future Work

3. Related Work -Speaker Identification-

Using visual information

Detecting lip movements [Faish, 2012]

Recognizing gestures [Bohus, 2009]

→ **It's difficult to identify speakers when they are out of the field of the system camera**

Using sound source localization

In our situation

- It's difficult to keep track of users in the field of the robot's camera
  (the angle of robot's camera is narrow)
- The robot cannot always look around
  (the robot is a participant in the conversation)

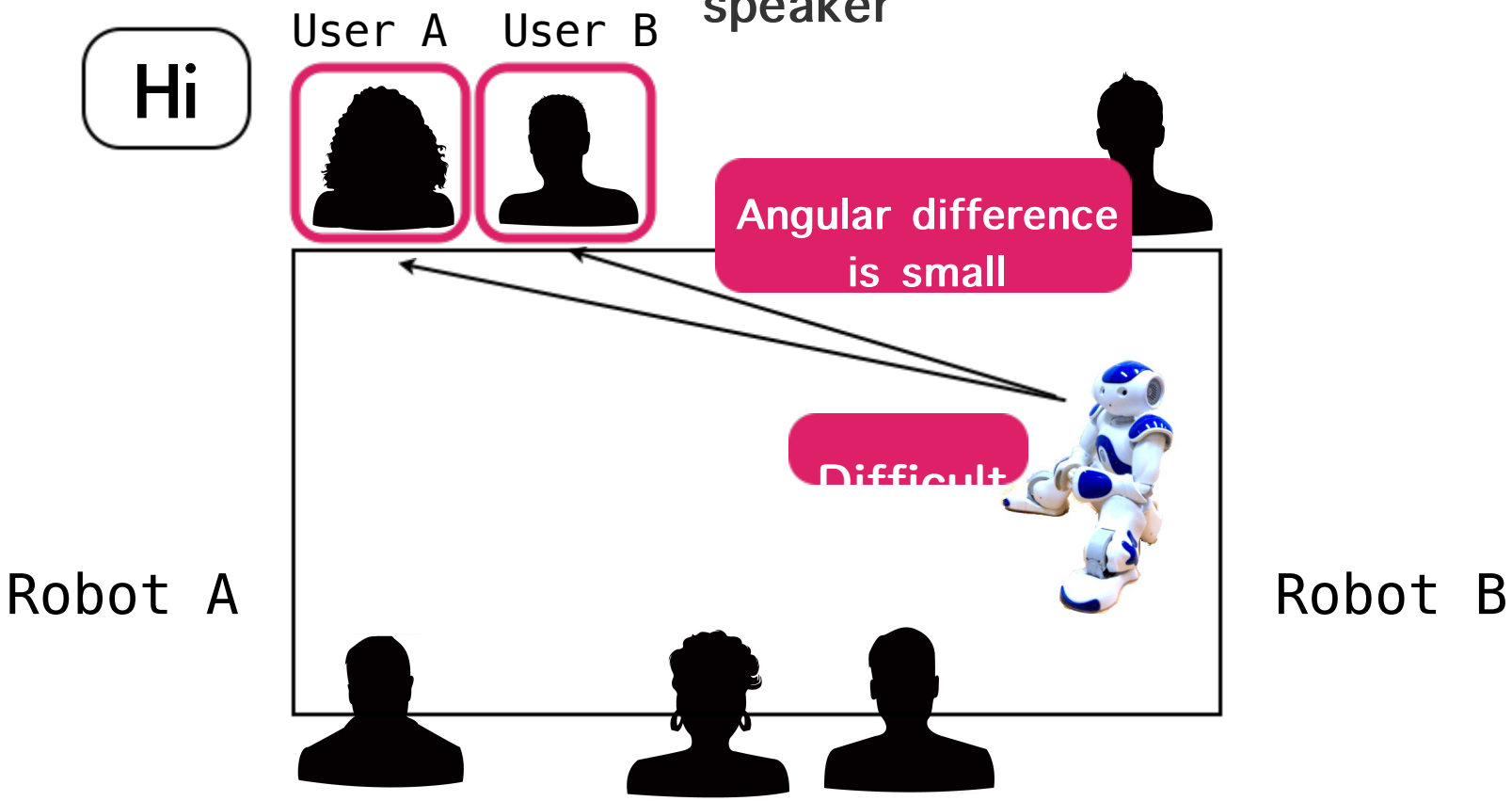
→ **Using localization results enable us to identify speakers who are out of the field of the robot's camera**

Outline

1. Background
2. Demo System
3. Related Work of Speaker Identification
4. Problems of Sound Source Localization
5. Solutions
   - 5-1. Inputs and Outputs of Our Method
   - 5-2. Integration of Multiple Sound Source Localization Results
6. Evaluation Experiments
   - 6-1. Results of identifying loudspeakers - Using only one robot / Integration -
   - 6-2. Localization results by Power
7. Conclusion & Future Work

4. Problems of Sound Source Localization

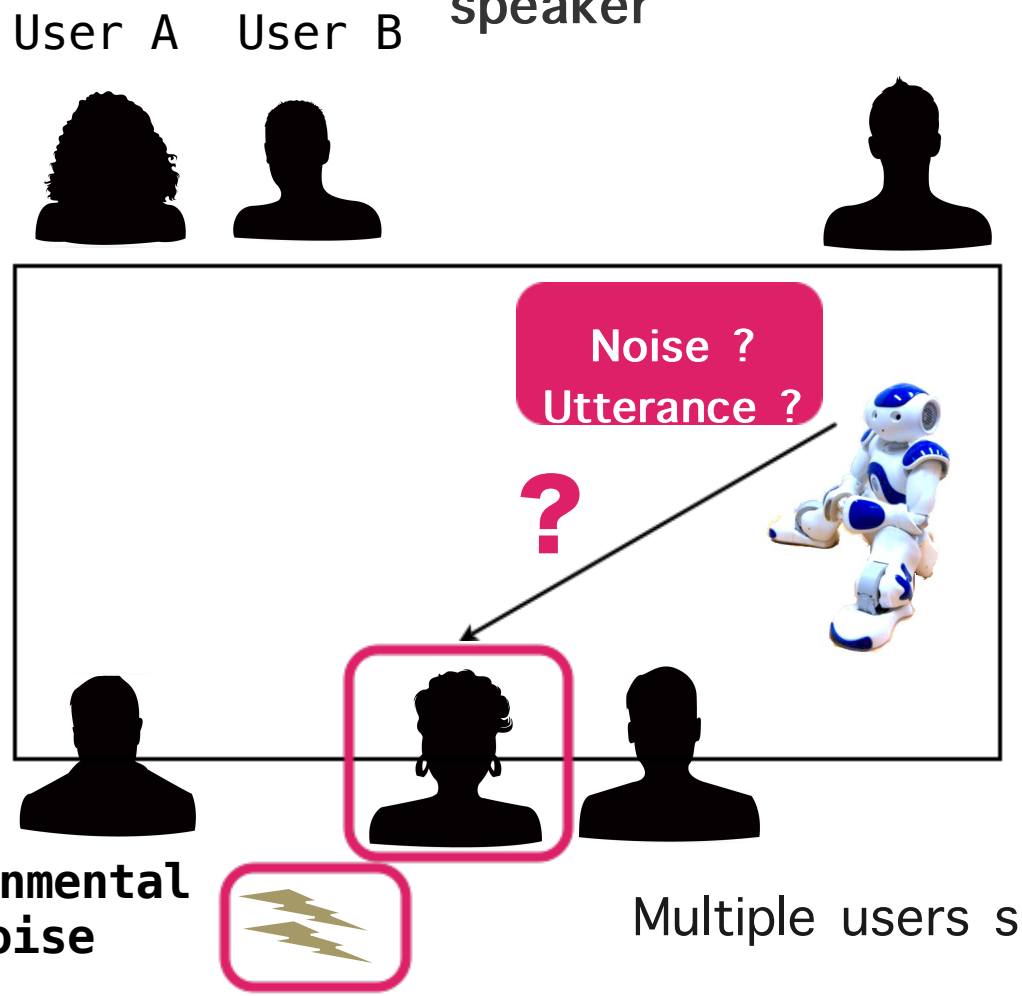
1. Some positions of users are difficult to localize
2. Environmental noise may cause incorrect localization
   → **Localization results do not always indicate the direction of speaker**



Multiple users sit around a table

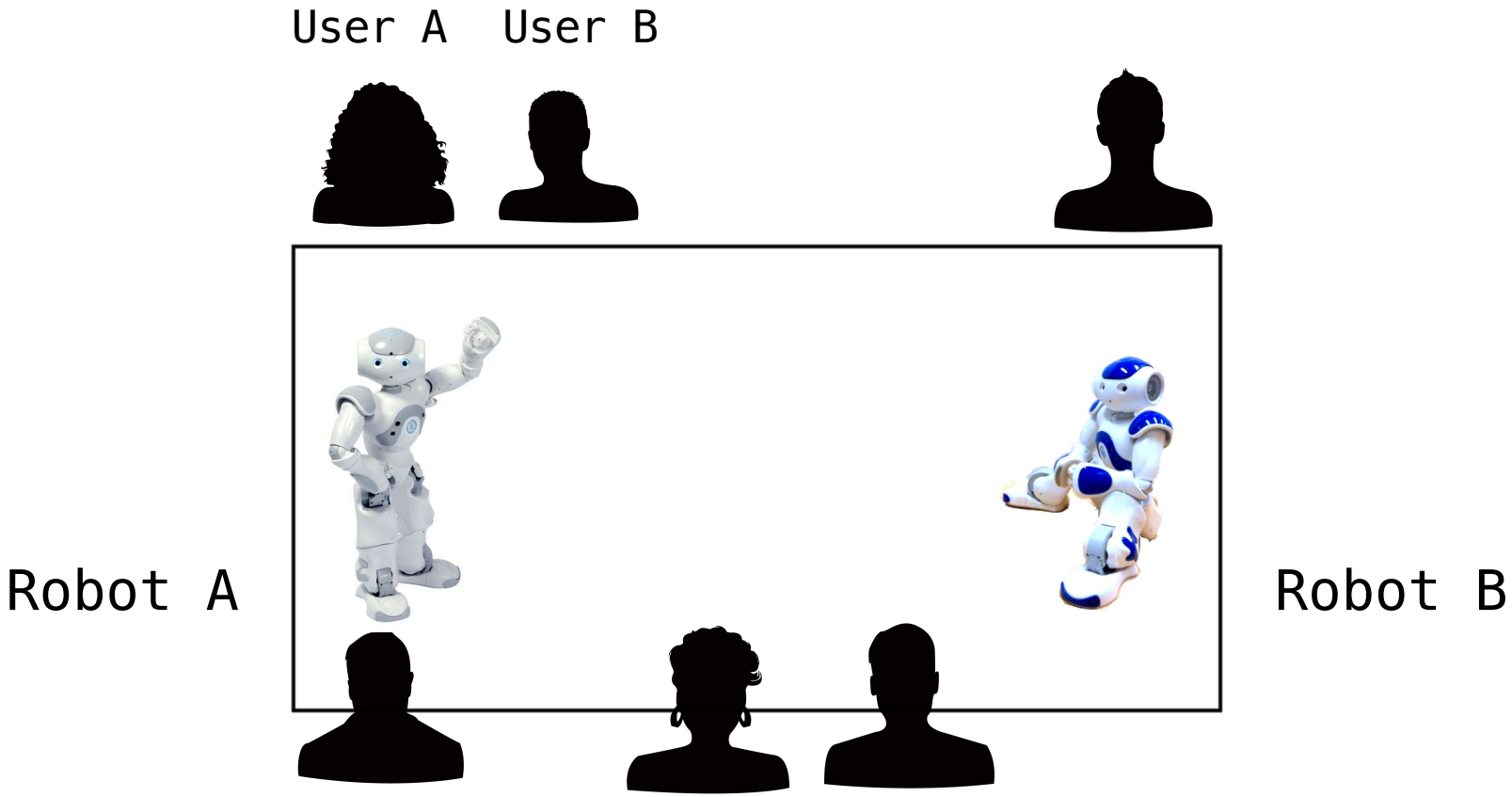
4. Problems of Sound Source Localization

1. Some positions of users are difficult to localize
2. Environmental noise may cause incorrect localization
   → **Localization results do not always indicate the direction of speaker**



5. Solutions -Overview-

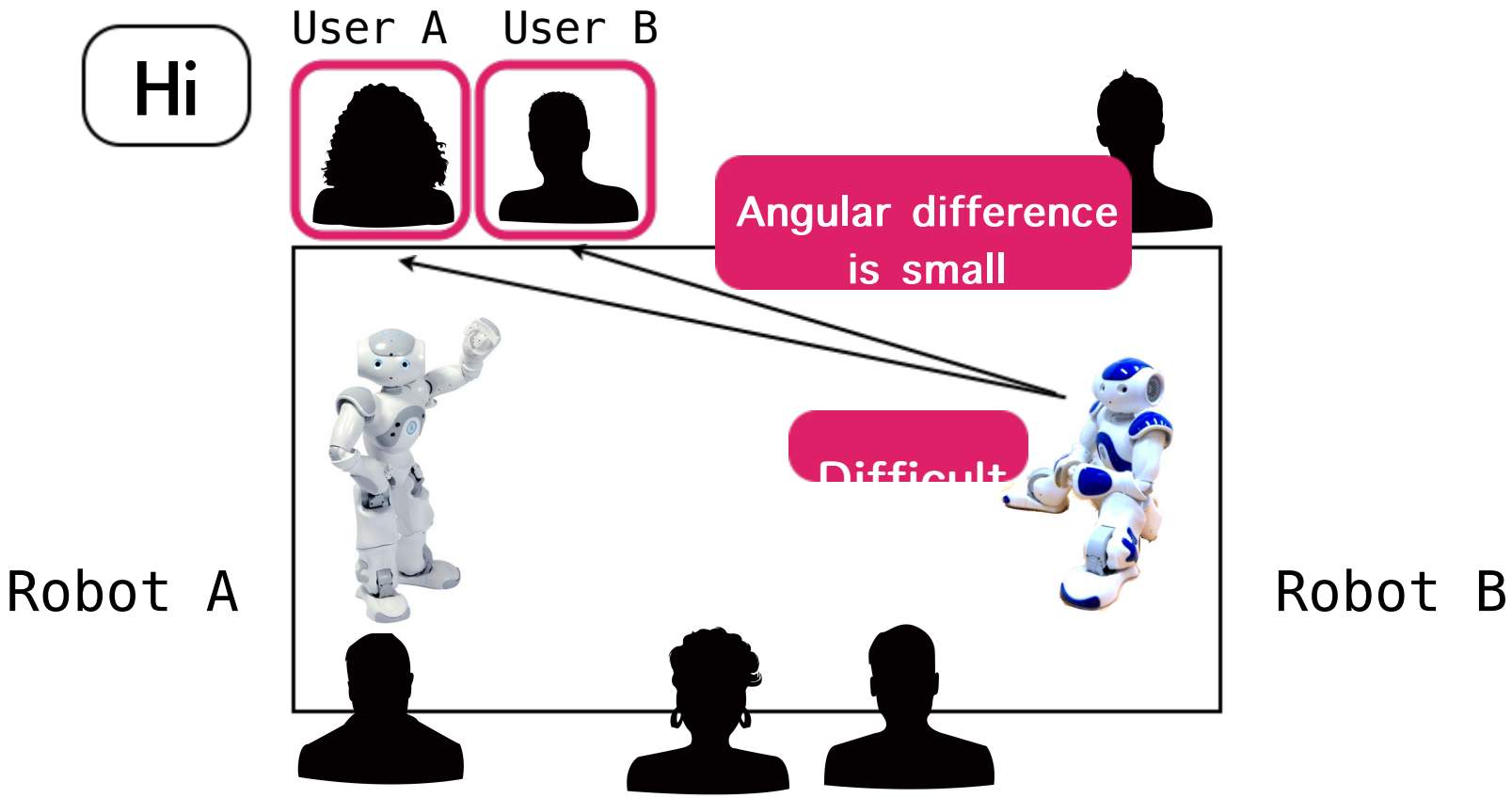
1. Placing robots on a table to opposite each other
   so as to compensate each other's capabilities
2. Integrating sound source localization results from the robots



Multiple users sit around a table

5. Solutions -Overview-

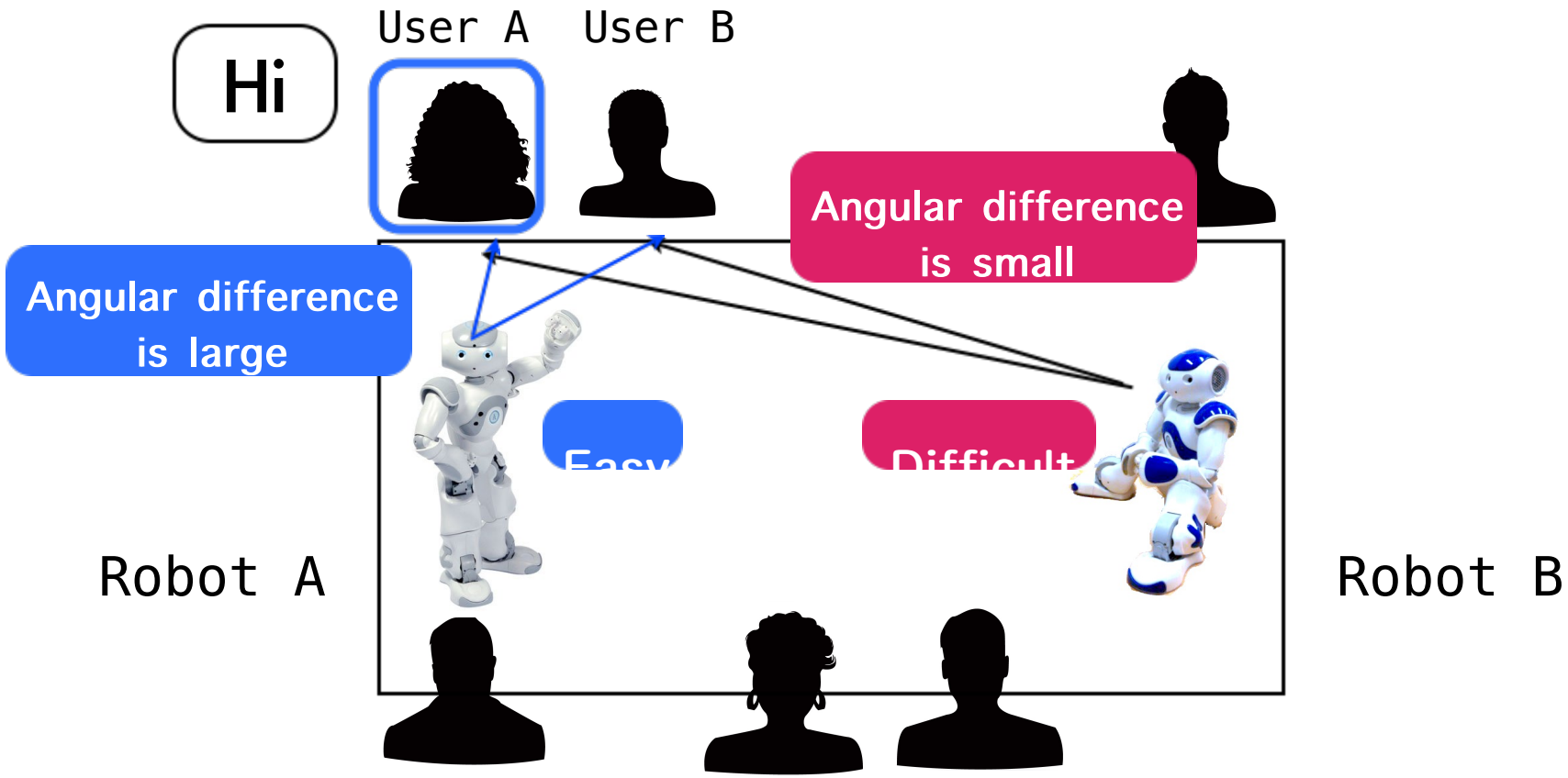
1. Placing robots on a table to opposite each other
   so as to compensate each other's capabilities
2. Integrating sound source localization results from the robots



Multiple users sit around a table

5. Solutions -Overview-

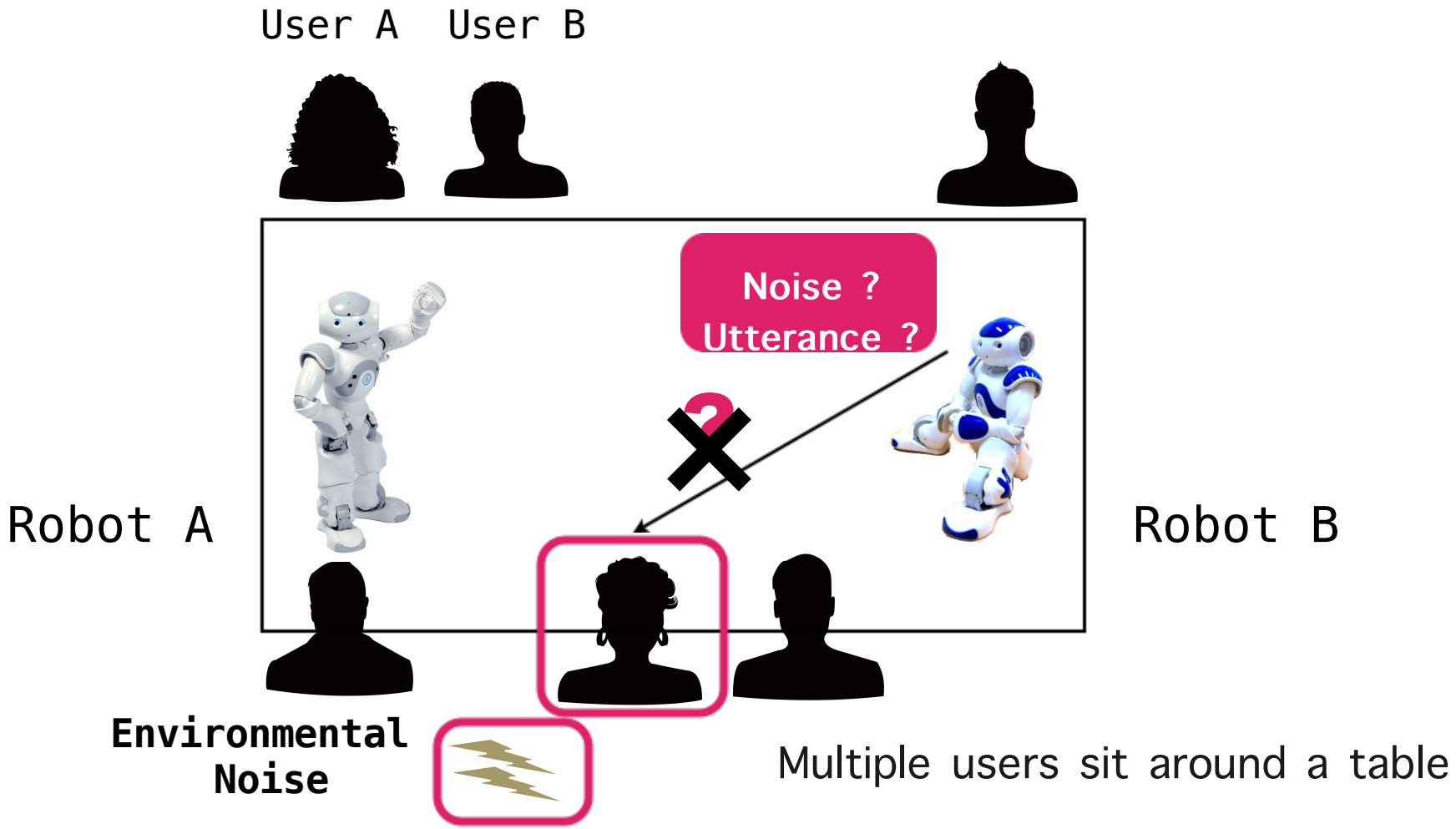
1. Placing robots on a table to opposite each other
   so as to compensate each other's capabilities
2. Integrating sound source localization results from the robots



Multiple users sit around a table

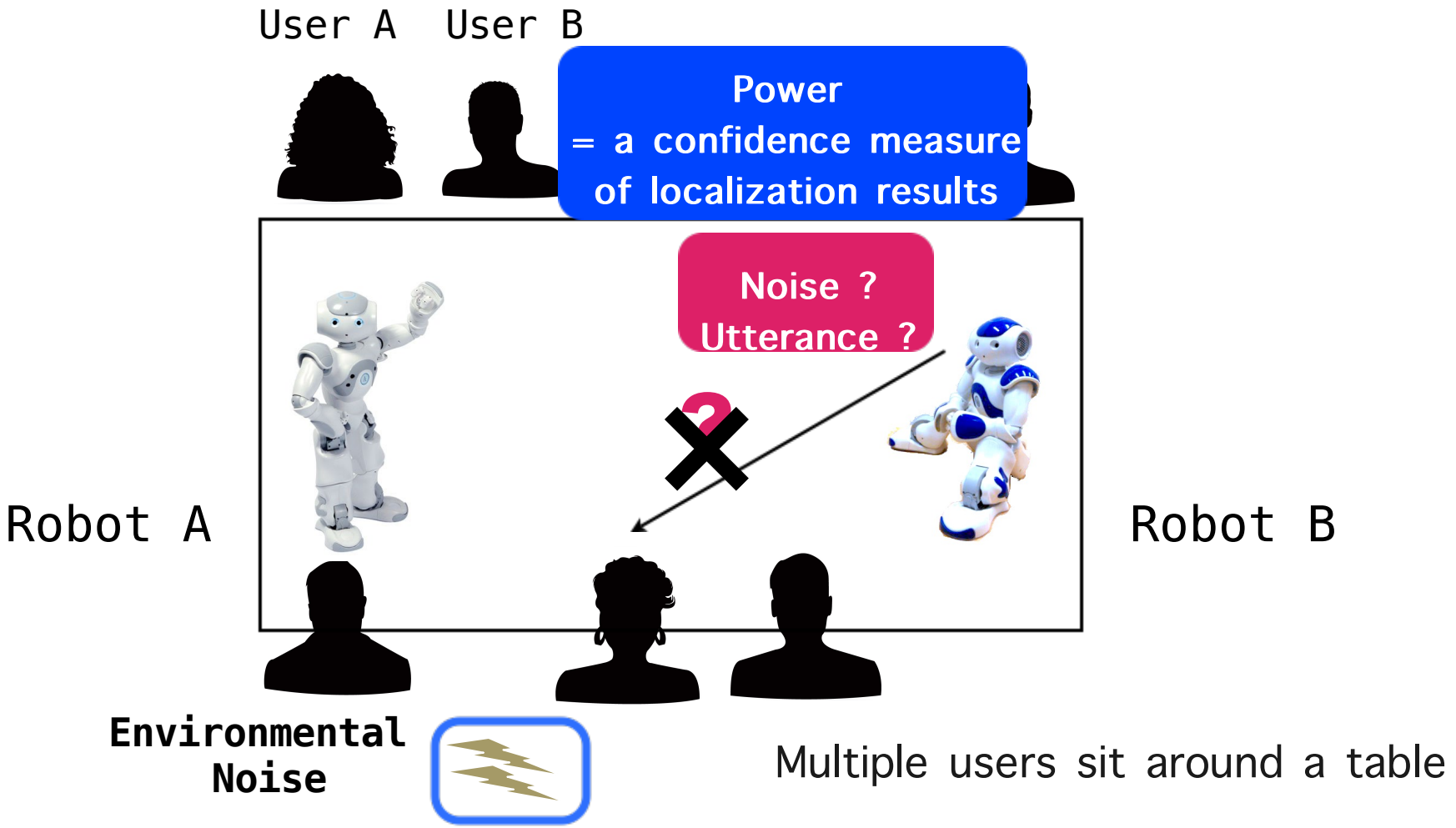
5. Solutions -Overview-

1. Placing robots on a table to opposite each other
   so as to compensate each other's capabilities
2. Integrating sound source localization results from the robots



5. Solutions -Overview-

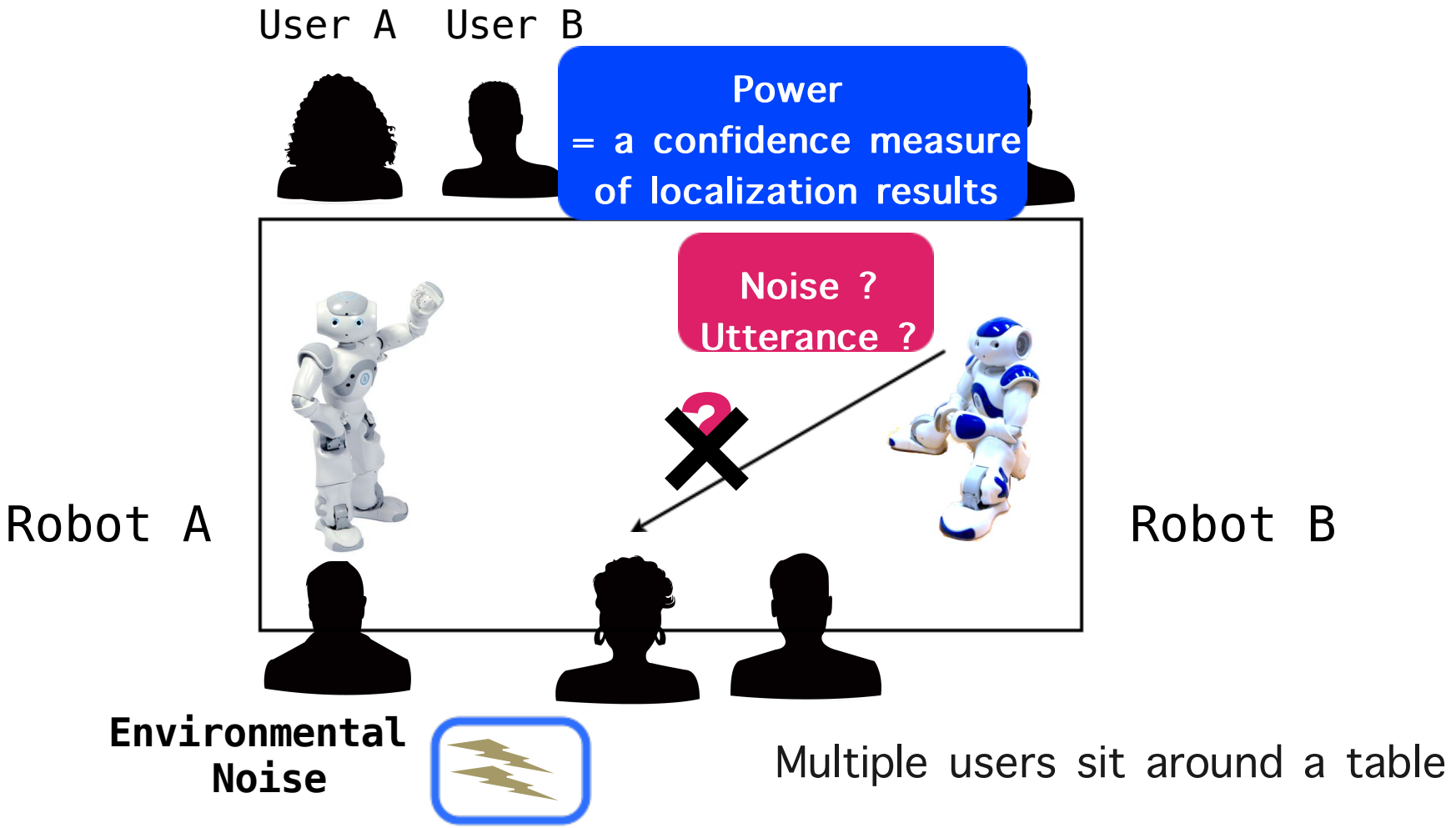
1. Placing robots on a table to opposite each other
   so as to compensate each other's capabilities
2. Integrating sound source localization results from the robots



Multiple users sit around a table

5. Solutions -Overview-

1. Placing robots on a table to opposite each other
   so as to compensate each other's capabilities
2. Integrating sound source localization results from the robots



Outline

1. Background
2. Demo System
3. Related Work of Speaker Identification
4. Problems of Sound Source Localization
5. Solutions
   - 5-1. Inputs and Outputs of Our Method
   - 5-2. Integration of Multiple Sound Source Localization Results
6. Evaluation Experiments
   - 6-1. Results of identifying loudspeakers - Using only one robot / Integration -
   - 6-2. Localization results by Power
7. Conclusion & Future Work

5-1. Inputs and Outputs of Our Method (1/2)

Settings

- Sound source localization

Robot audition software

HARK **outputs** { Localization results: θ [deg] ; Power: p [dB] } every 1 frame (=0.01 second)

developed in Kyoto Univ.

based on MUltiple SIgnal Classification (MUSIC) method

- Microphones

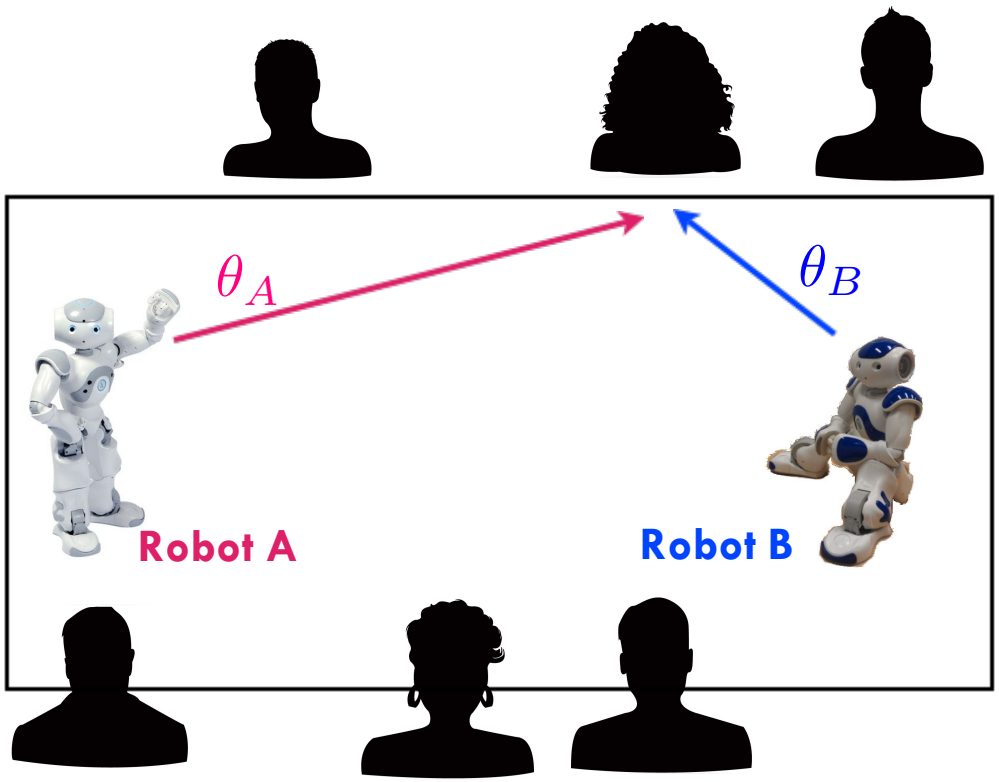
four microphones in head.

Impulse response for calculating the transfer function

→recorded at 36 points, at intervals of 10 [deg]

Angular resolution = 10 [deg]

5-1. Inputs and Outputs of Our Method (2/2)



Inputs

Localization results : θ_A

Power: p_A

Inputs

Localization results : θ_B

Power: p_B

Outputs

Integrated { **Localization results** : θ_{mix}

Power: p_{mix}

5-2. Integration of Multiple Localization Results (1/2)

When we obtain localization result θ_r[deg] and its power p_r[dB] at one frame ....

1. Define probability density function from θ_r

   Assumption: the ambiguity of localization results follows a normal distribution

$$f_r(\theta) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp\left(-\frac{(\theta - \theta_r)^2}{2\sigma_r^2}\right)$$

2. Define the maximum probability is proportioned to p_r

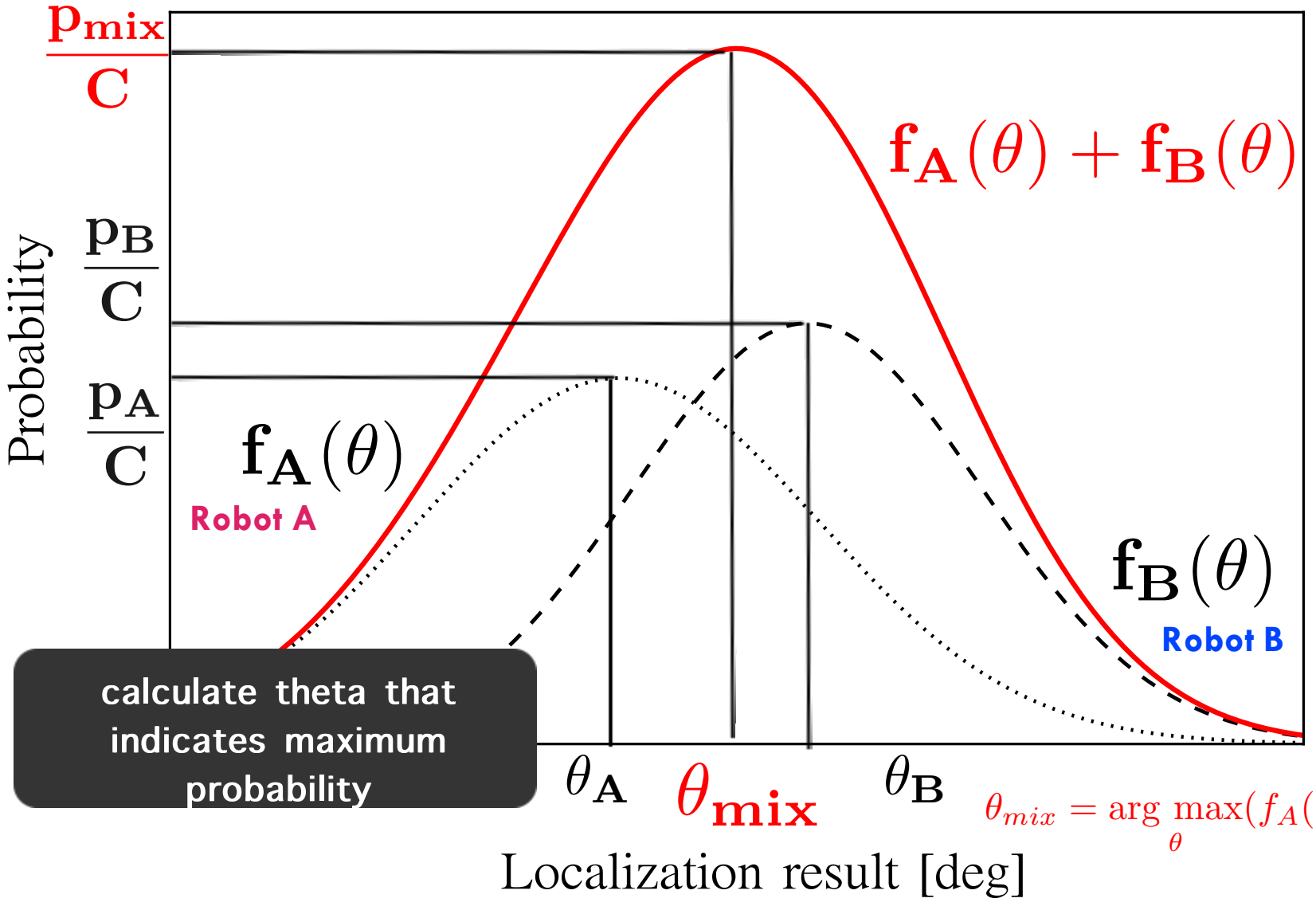
   Assumption: the power of localization results caused by noise is low

$$f_r(\theta_r) = \frac{1}{\sqrt{2\pi\sigma_r^2}} = \frac{1}{C} p_r$$

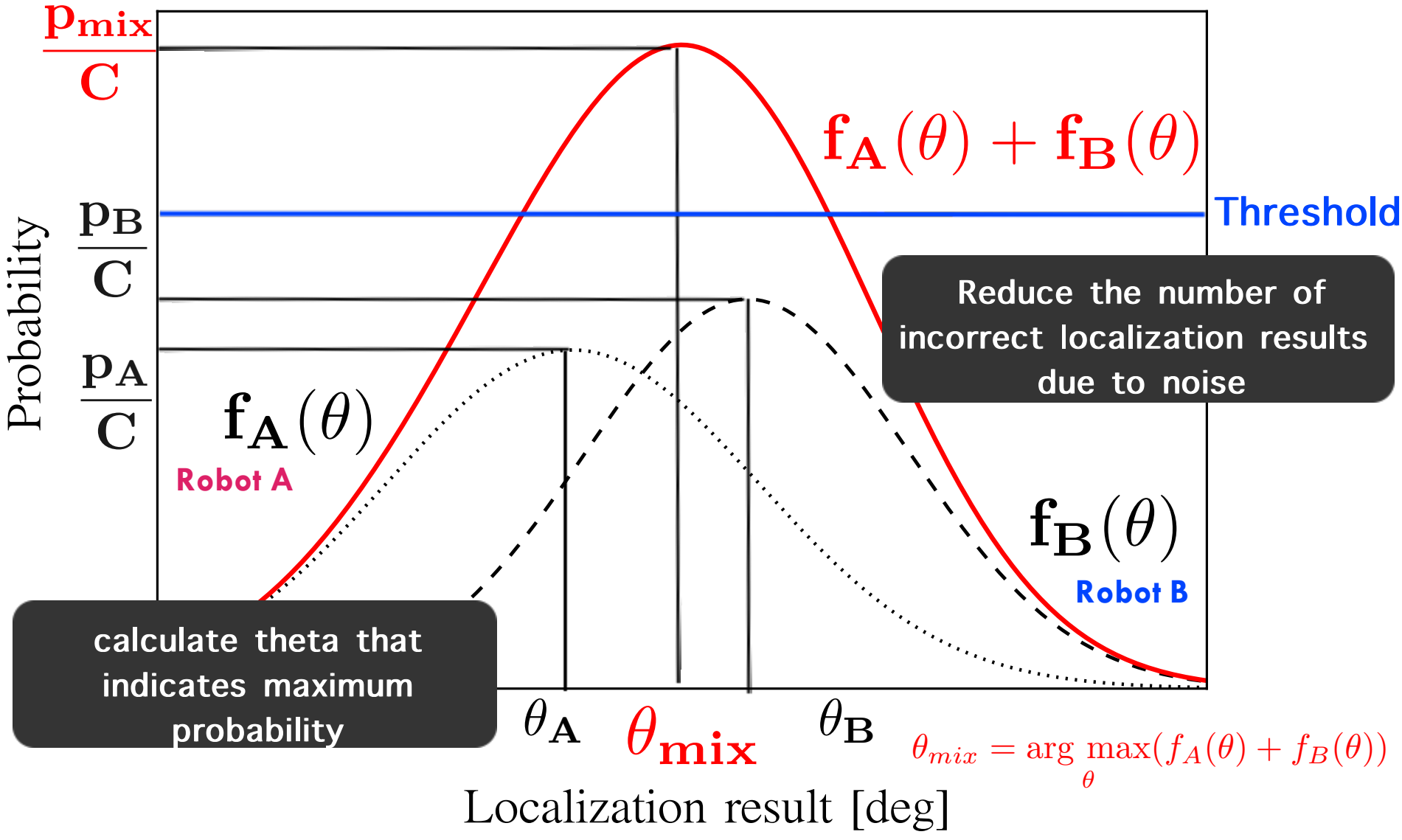
The larger power indicates the higher probability

C is a constant value and determined empirically

5-2. Integration of Multiple Localization Results (2/2)



5-2. Integration of Multiple Localization Results (2/2)

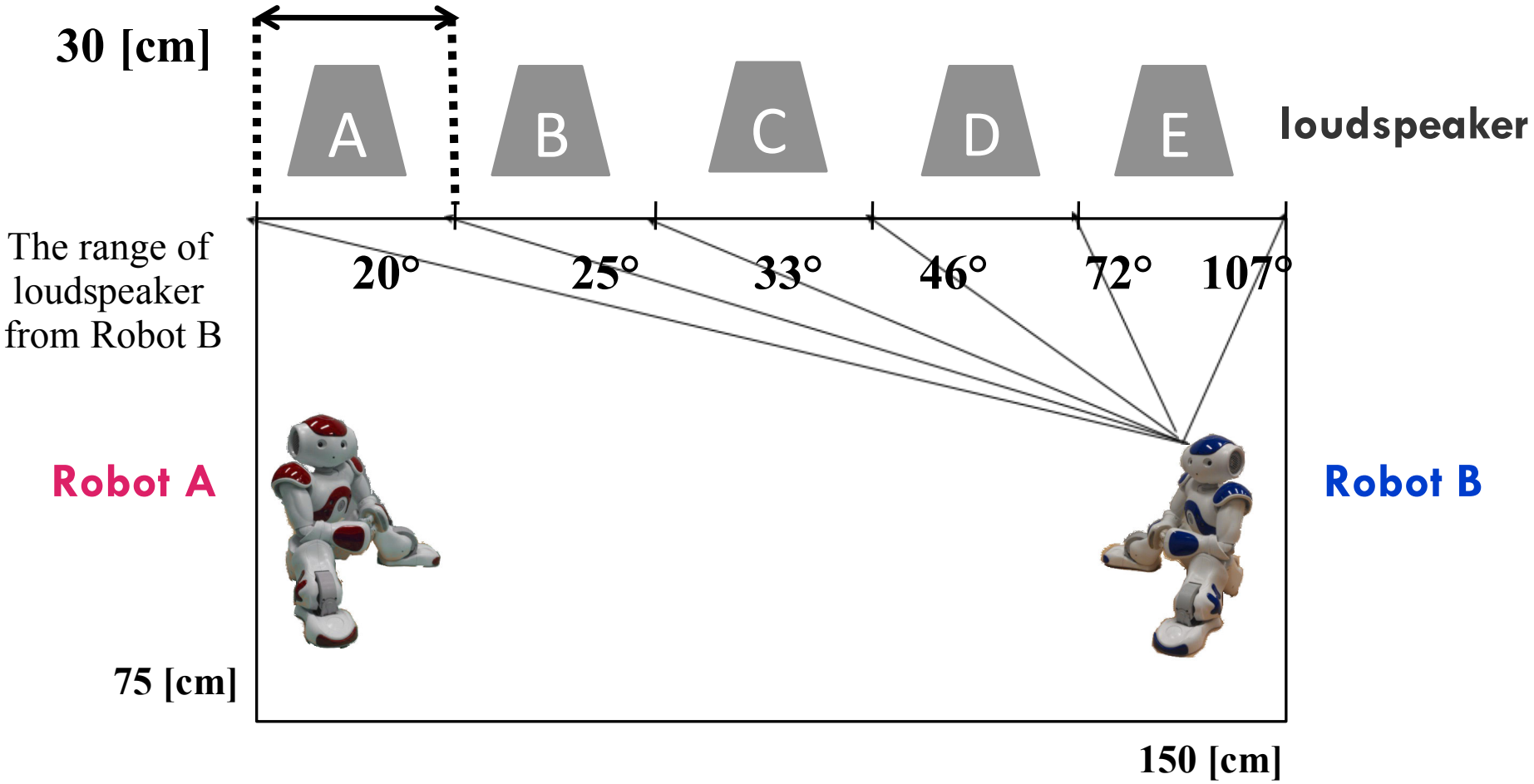


Outline

1. Background
2. Demo System
3. Related Work of Speaker Identification
4. Problems of Sound Source Localization
5. Solutions
   - 5-1. Inputs and Outputs of Our Method
   - 5-2. Integration of Multiple Sound Source Localization Results
6. Evaluation Experiments
   - 6-1. Results of identifying loudspeakers - Using only one robot / Integration -
   - 6-2. Localization results by Power
7. Conclusion & Future Work

6. Evaluation Experiments -Settings (1/2)-

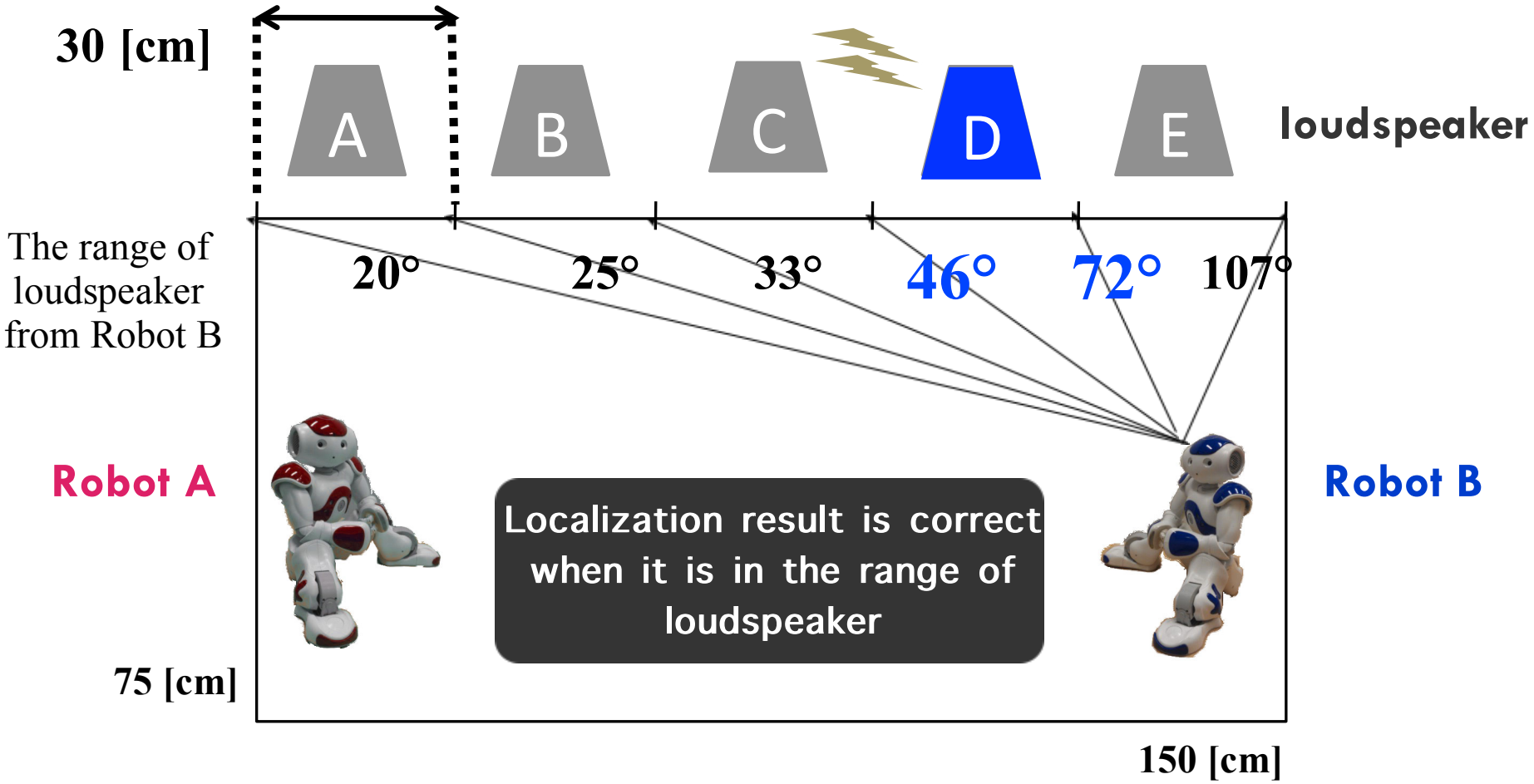
Evaluated whether using two robots improved speaker identification



1. Placing loudspeakers where users may sit
2. Playing speech sounds from loudspeakers and identifying them

6. Evaluation Experiments -Settings (1/2)-

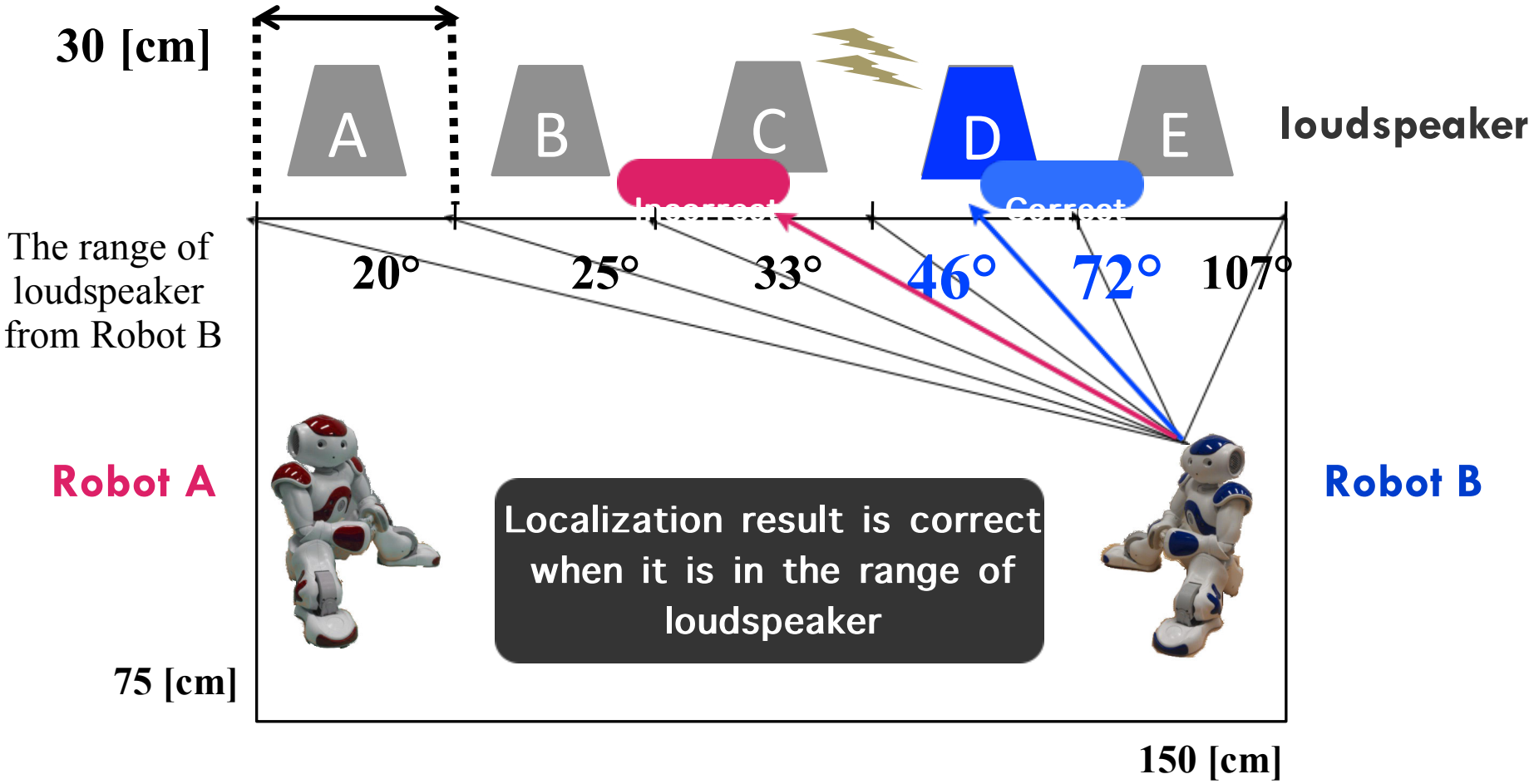
Evaluated whether using two robots improved speaker identification



1. Placing loudspeakers where users may sit
2. Playing speech sounds from loudspeakers and identifying them

6. Evaluation Experiments -Settings (1/2)-

Evaluated whether using two robots improved speaker identification



1. Placing loudspeakers where users may sit
2. Playing speech sounds from loudspeakers and identifying them

6. Evaluation Experiments -Settings (2/2)-

- Data

5 utterances × 5 points × 4 speakers = 100 data

One audio file includes one utterance whose duration is 1.0 second

- Evaluation Measure

$$Precision = \frac{\text{Number of frames when localization result was correct}}{\text{Number of all detected frames}}$$

$$Recall = \frac{\text{Number of frames when localization result was correct}}{\text{Number of speech frames}}$$

$$F = 2\left(\frac{1}{Precision} + \frac{1}{Recall}\right)^{-1}$$

Outline

1. Background
2. Demo System
3. Related Work of Speaker Identification
4. Problems of Sound Source Localization
5. Solutions
   - 5-1. Inputs and Outputs of Our Method
   - 5-2. Integration of Multiple Sound Source Localization Results
6. Evaluation Experiments
   - 6-1. Results of identifying loudspeakers - Using only one robot / Integration -
   - 6-2. Localization results by Power
7. Conclusion & Future Work

6-1. Results of identifying loudspeakers -Using only one robot-

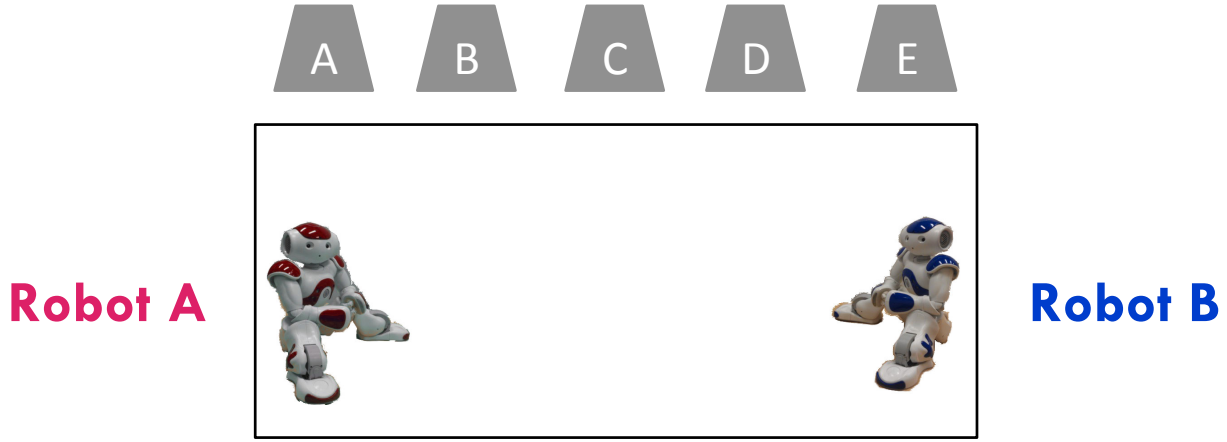
Evaluated whether using two robots improved speaker identification

Robot A

SPK	precision	recall	F
A	0,59	0,86	0,69
B	0,45	0,60	0,50
C	0,00	0,00	-
D	0,02	0,02	0,02
E	0,03	0,05	0,04
ALL	0,22	0,31	0,25

Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,05	0,03	0,04
C	0,14	0,19	0,16
D	0,56	0,84	0,67
E	0,44	0,64	0,52
ALL	0,24	0,34	0,28



6-1. Results of identifying loudspeakers -Using only one robot-

Evaluated whether using two robots improved speaker identification

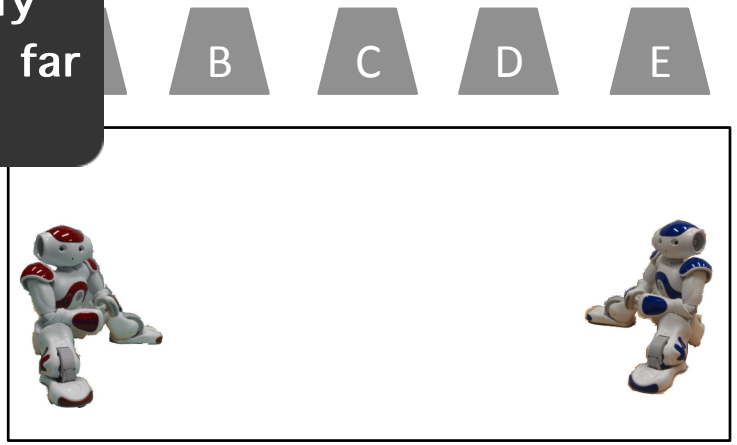
Robot A

SPK	precision	recall	F
A	0,59	0,86	0,69
B	0,45	0,60	0,50
C	0,00	0,00	-
D	0,02	0,02	0,02
E	0,03	0,05	0,04
ALL	0,22	0,31	0,25

Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,05	0,03	0,04
C	0,14	0,19	0,16
D	0,56	0,84	0,67
E	0,44	0,64	0,52
ALL	0,24	0,34	0,28

It's difficult to identify loudspeakers that were far from the robots.



6-1. Results of identifying loudspeakers -Using only one robot-

Evaluated whether using two robots improved speaker identification

Robot A

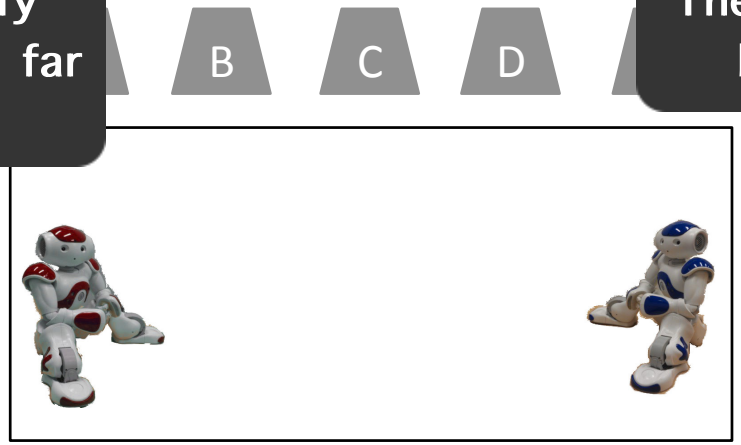
SPK	precision	recall	F
A	0,59	0,86	0,69
B	0,45	0,60	0,50
C	0,00	0,00	-
D	0,02	0,02	0,02
E	0,03	0,05	0,04
ALL	0,22	0,31	0,25

Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,05	0,03	0,04
C	0,14	0,19	0,16
D	0,56	0,84	0,67
E	0,44	0,64	0,52
ALL	0,24	0,34	0,28

It's difficult to identify loudspeakers that were far from the robots.

The performances differed between two robots.



6-2. Results of identifying loudspeakers -Integration-

Evaluated whether using two robots improved speaker identification

$$C = 800$$
$$thresh = \frac{25.5}{800}$$

Integration

SPK	precision	recall	F
A	0,57	0,85	0,68
B	0,40	0,50	0,45
C	0,38	0,49	0,43
D	0,48	0,67	0,56
E	0,39	0,61	0,48
ALL	0,45	0,62	0,52

Robot A

SPK	precision	recall	F
A	0,56	0,89	0,69
B	0,49	0,65	0,56
C	0,00	0,00	-
D	0,06	0,03	0,04
E	0,09	0,03	0,05
ALL	0,33	0,32	0,33

Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,00	0,00	-
C	0,13	0,13	0,13
D	0,63	0,83	0,72
E	0,50	0,69	0,58
ALL	0,39	0,33	0,36

6-2. Results of identifying loudspeakers -Integration-

Evaluated whether using two robots improved speaker identification

System can identify the areas that only one robot cannot

$$C = 800$$
$$thresh = \frac{25.5}{800}$$

Integration

SPK	precision	recall	F
A	0,57	0,85	0,68
B	0,40	0,50	0,45
C	0,38	0,49	0,43
D	0,48	0,67	0,56
E	0,39	0,61	0,48
ALL	0,45	0,62	0,52

Robot A

SPK	precision	recall	F
A	0,56	0,89	0,69
B	0,49	0,65	0,56
C	0,00	0,00	-
D	0,06	0,03	0,04
E	0,09	0,03	0,05
ALL	0,33	0,32	0,33

Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,00	0,00	-
C	0,13	0,13	0,13
D	0,63	0,83	0,72
E	0,50	0,69	0,58
ALL	0,39	0,33	0,36

6-2. Results of identifying loudspeakers -Integration-

Evaluated whether using two robots improved speaker identification

Integration

System can identify the areas that only one robot cannot

SPK	precision	recall	F
A	0,57	0,85	0,68
B	0,40	0,50	0,45
C	0,38	0,49	0,43
D	0,48	0,67	0,56
E	0,39	0,61	0,48
ALL	0,45	0,62	0,52

In particular, the loudspeaker at C get correctly identified, for which neither robots cannot

$$C = 800$$

$$thresh = \frac{25.5}{800}$$

Robot A

SPK	precision	recall	F
A	0,56	0,89	0,69
B	0,49	0,65	0,56
C	0,00	0,00	-
D	0,06	0,03	0,04
E	0,09	0,03	0,05
ALL	0,33	0,32	0,33

Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,00	0,00	-
C	0,13	0,13	0,13
D	0,63	0,83	0,72
E	0,50	0,69	0,58
ALL	0,39	0,33	0,36

6-2. Results of identifying loudspeakers -Integration-

Evaluated whether using two robots improved speaker identification

Integration

System can identify the areas that only one robot cannot

$$C = 800$$
$$thresh = \frac{25.5}{800}$$

SPK	precision	recall	F
A	0,57	0,85	0,68
B	0,40	0,50	0,45
C	0,38	0,49	0,43
D	0,48	0,67	0,56
E	0,39	0,61	0,48
ALL	0,45	0,62	0,52

In particular, the loudspeaker at C get correctly identified, for which neither robots cannot

Integration improved performance

Robot A

SPK	precision	recall	F
A	0,56	0,89	0,69
B	0,49	0,65	0,56
C	0,00	0,00	-
D	0,06	0,03	0,04
E	0,09	0,03	0,05
ALL	0,33	0,32	0,33

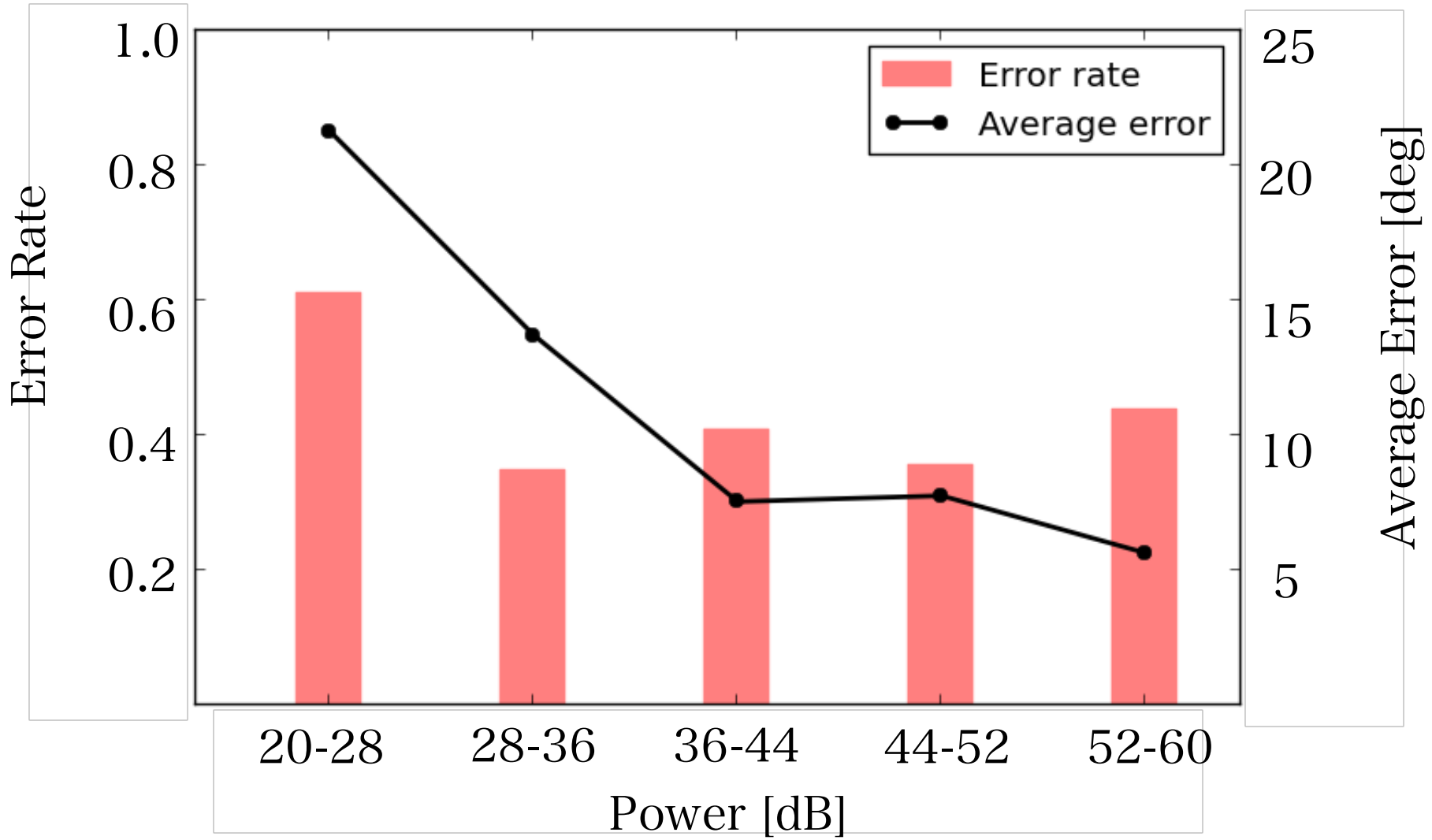
Robot B

SPK	precision	recall	F
A	0,00	0,00	-
B	0,00	0,00	-
C	0,13	0,13	0,13
D	0,63	0,83	0,72
E	0,50	0,69	0,58
ALL	0,39	0,33	0,36

Outline

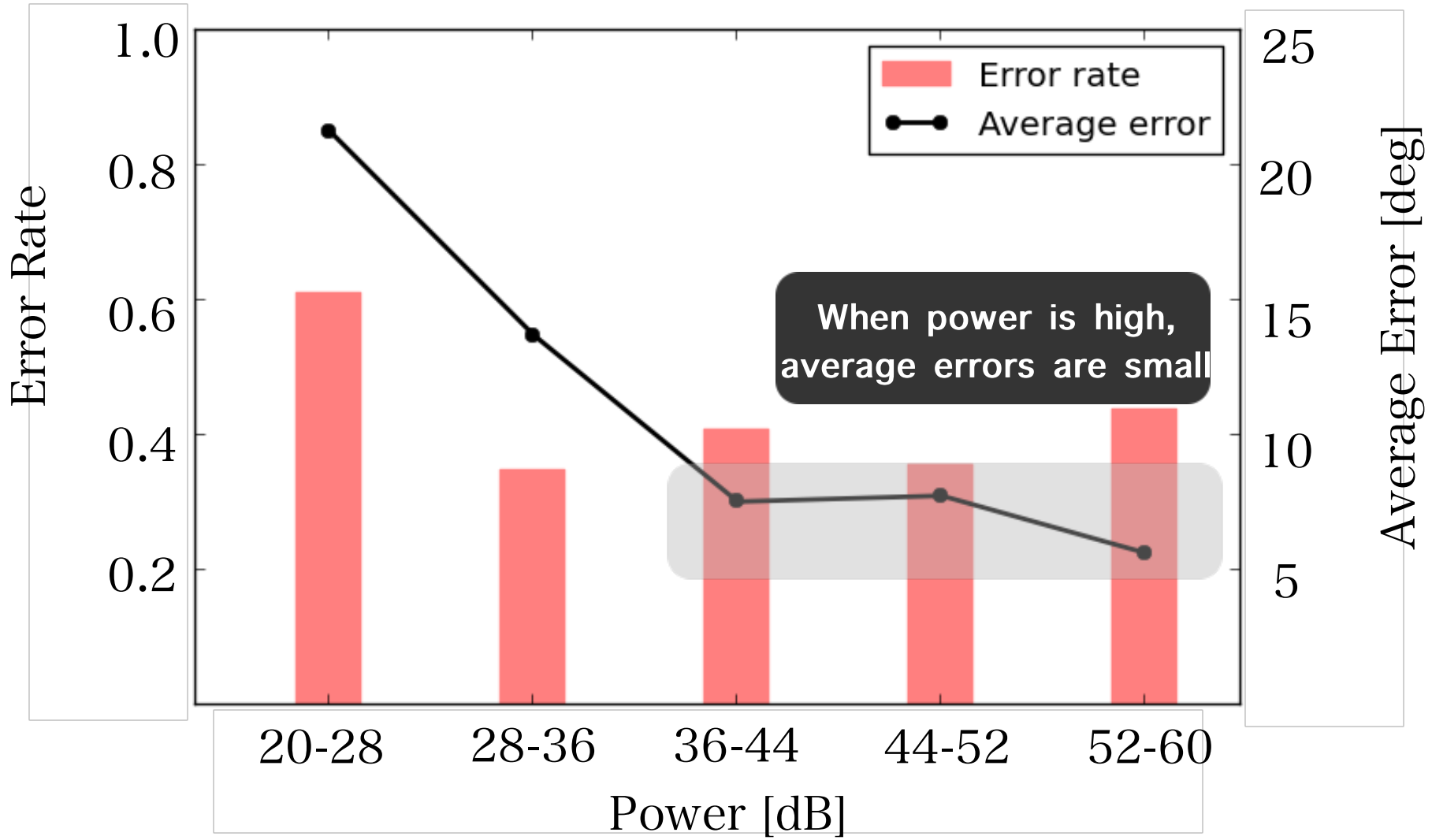
6-2. Localization Results by Power (1/2)

Evaluated whether integrated power was valid as a confidence measure



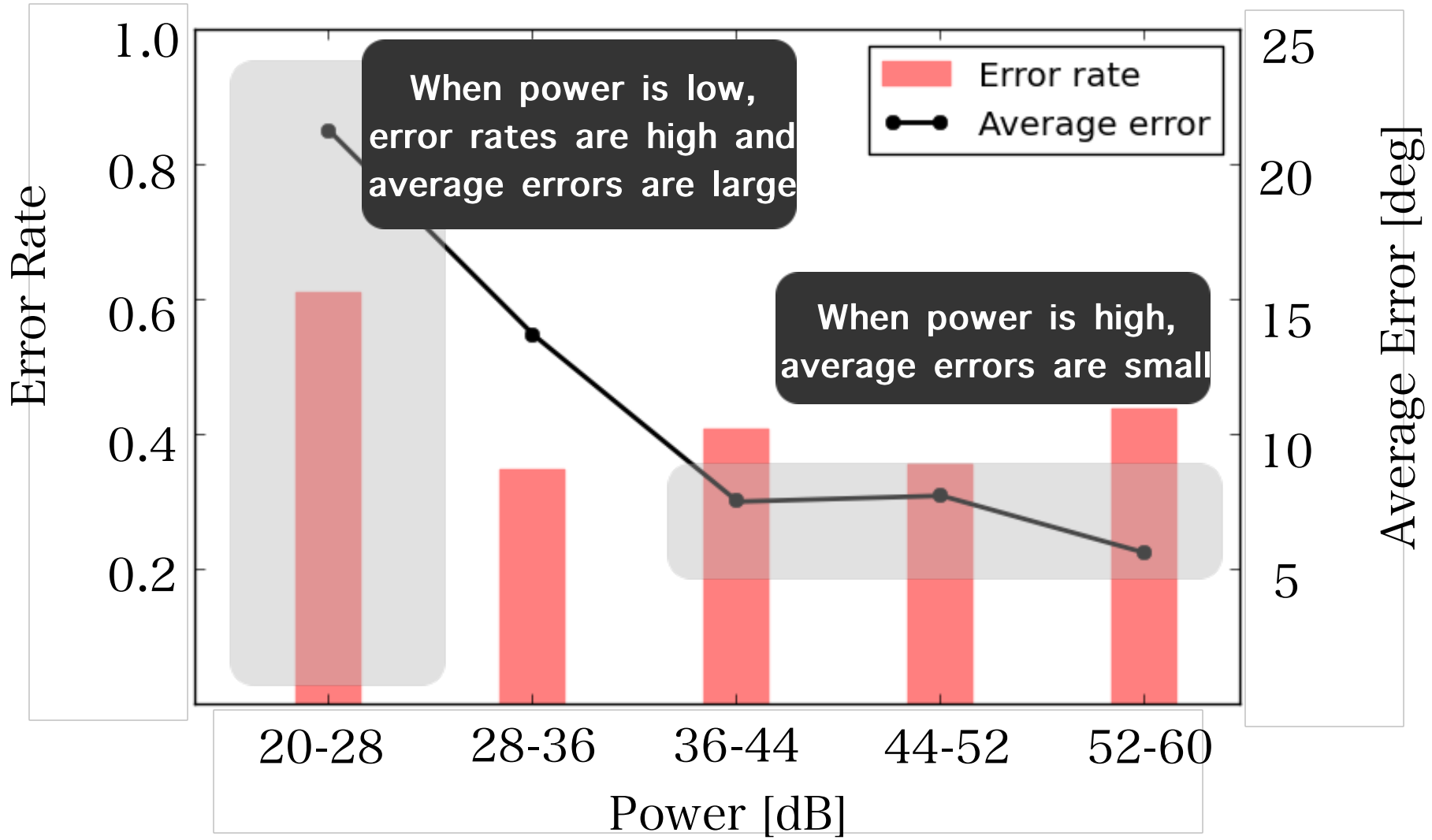
6-2. Localization Results by Power (1/2)

Evaluated whether integrated power was valid as a confidence measure



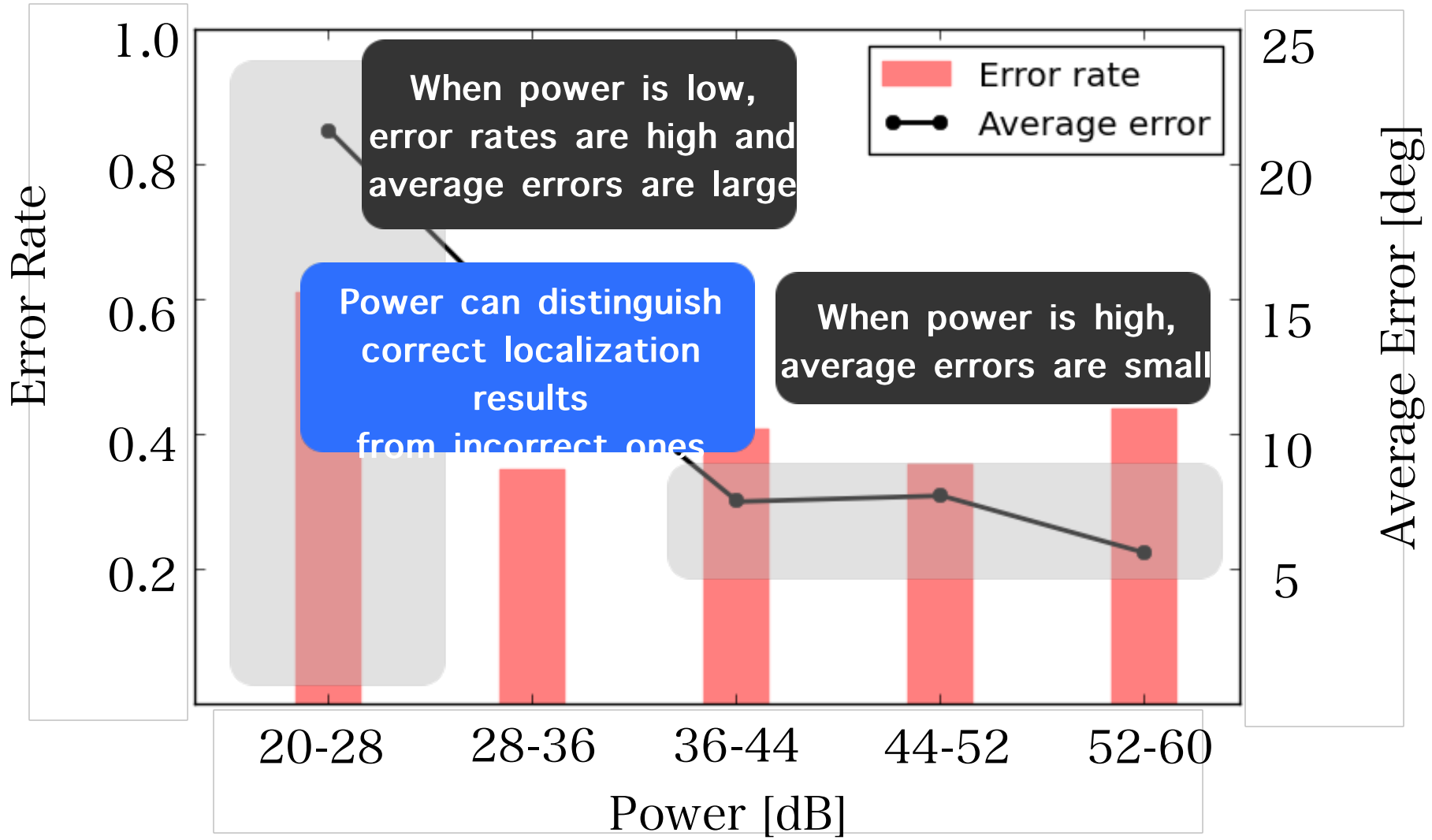
6-2. Localization Results by Power (1/2)

Evaluated whether integrated power was valid as a confidence measure



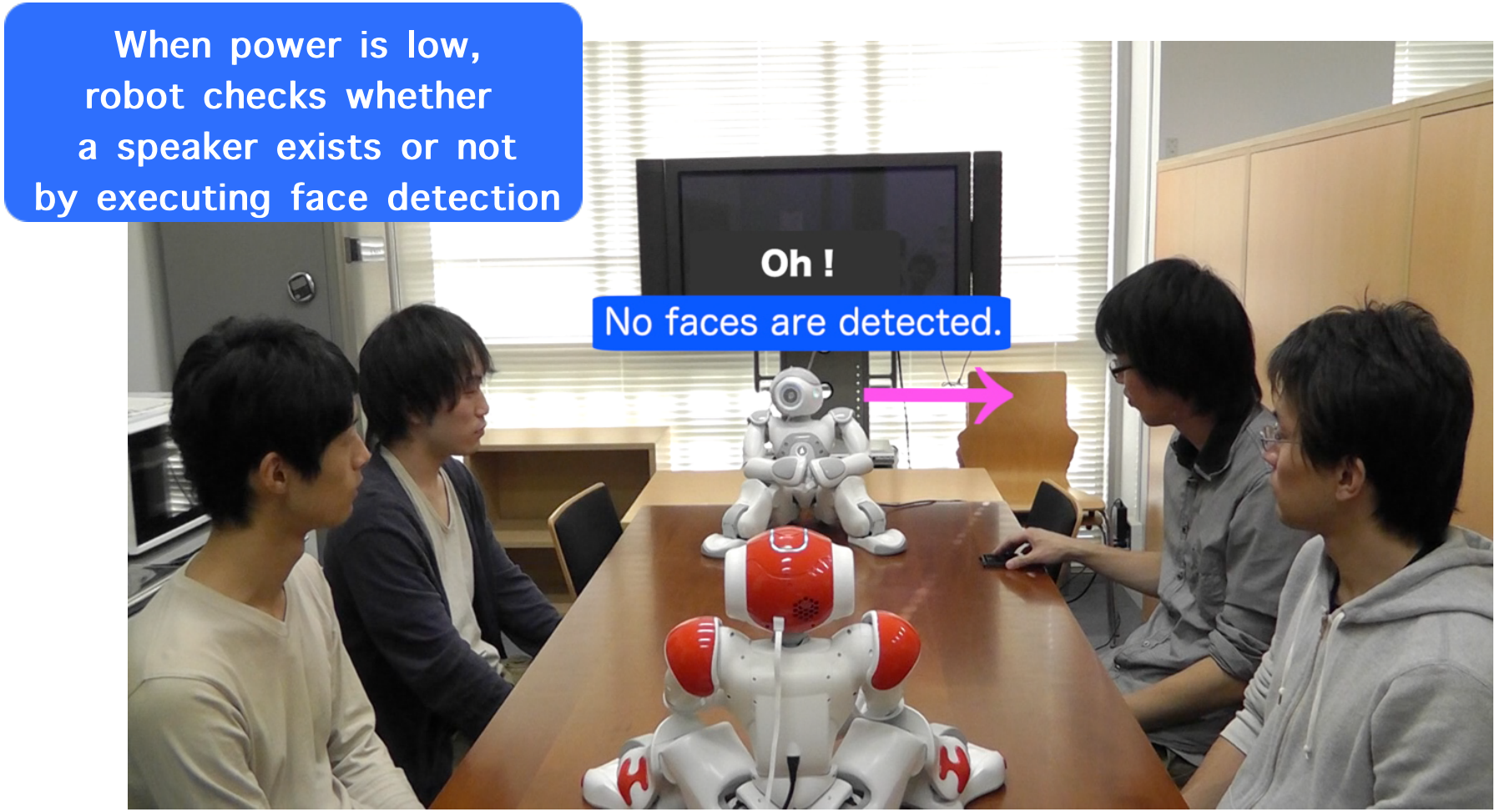
6-2. Localization Results by Power (1/2)

Evaluated whether integrated power was valid as a confidence measure



6-2. Localization Results by Power (2/2)

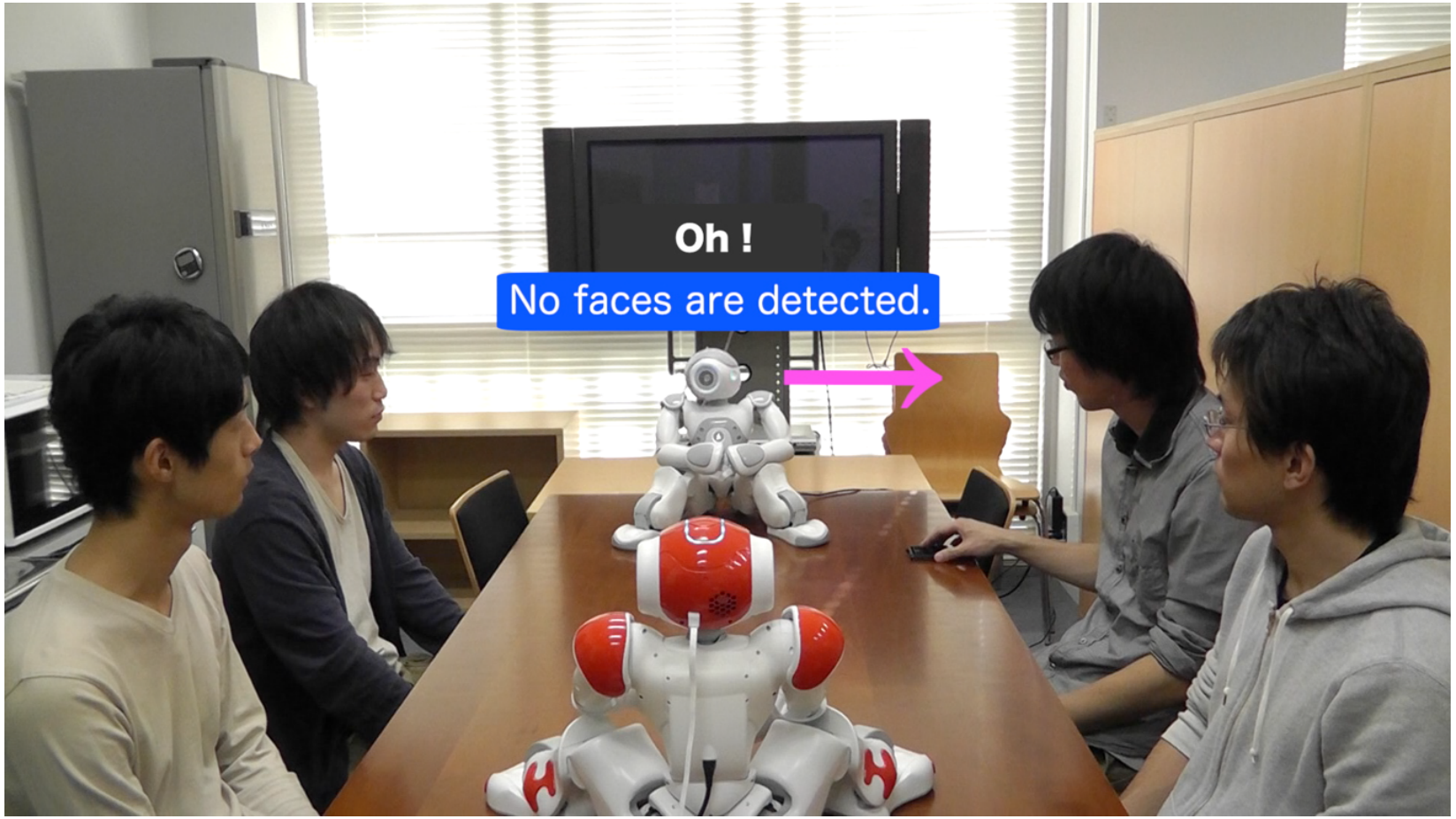
Evaluated whether integrated power was valid as a confidence measure



How to use power

6-2. Localization Results by Power (2/2)

Evaluated whether integrated power was valid as a confidence measure



How to use power

7. Conclusion

Integrate multiple sound source localization results

→ improve performance compared with using only one robot

Implement demo system

→ identifying a speaker and heading toward to answer

→ executing face detection to check whether a speaker exists on the basis of power

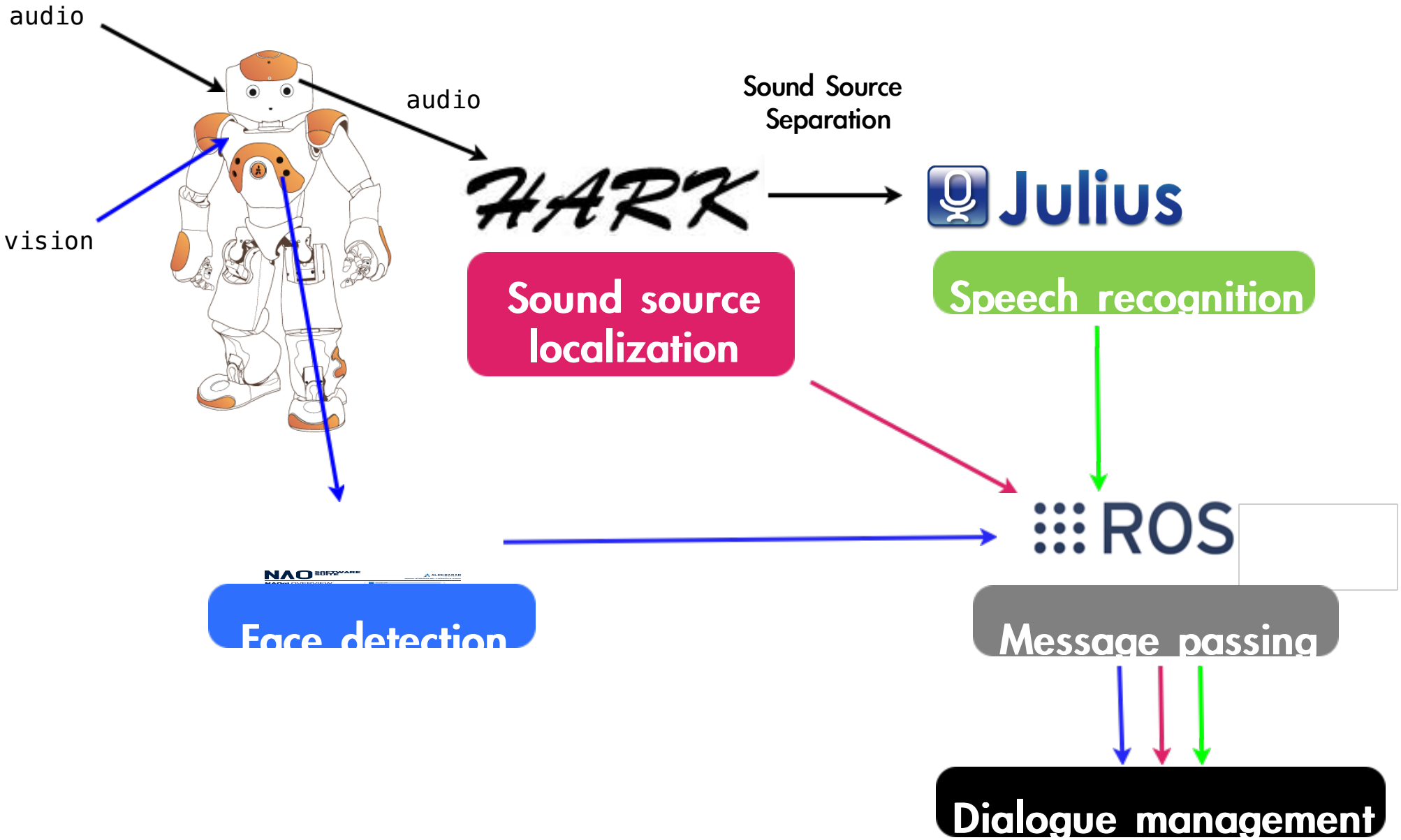
Future Works

Use other evidence of speaker's existence

e.g. image processing

→ improve performance of speaker identification

System Overview



Speech Recognition

We use 

- Language model : Grammar model

(vocabulary size is 20 words)

- Performance : We had no large experiment

In my impression, word correctness is 0.5

Problems of Sound Source Localization

1. Some positions of users are difficult to localize.