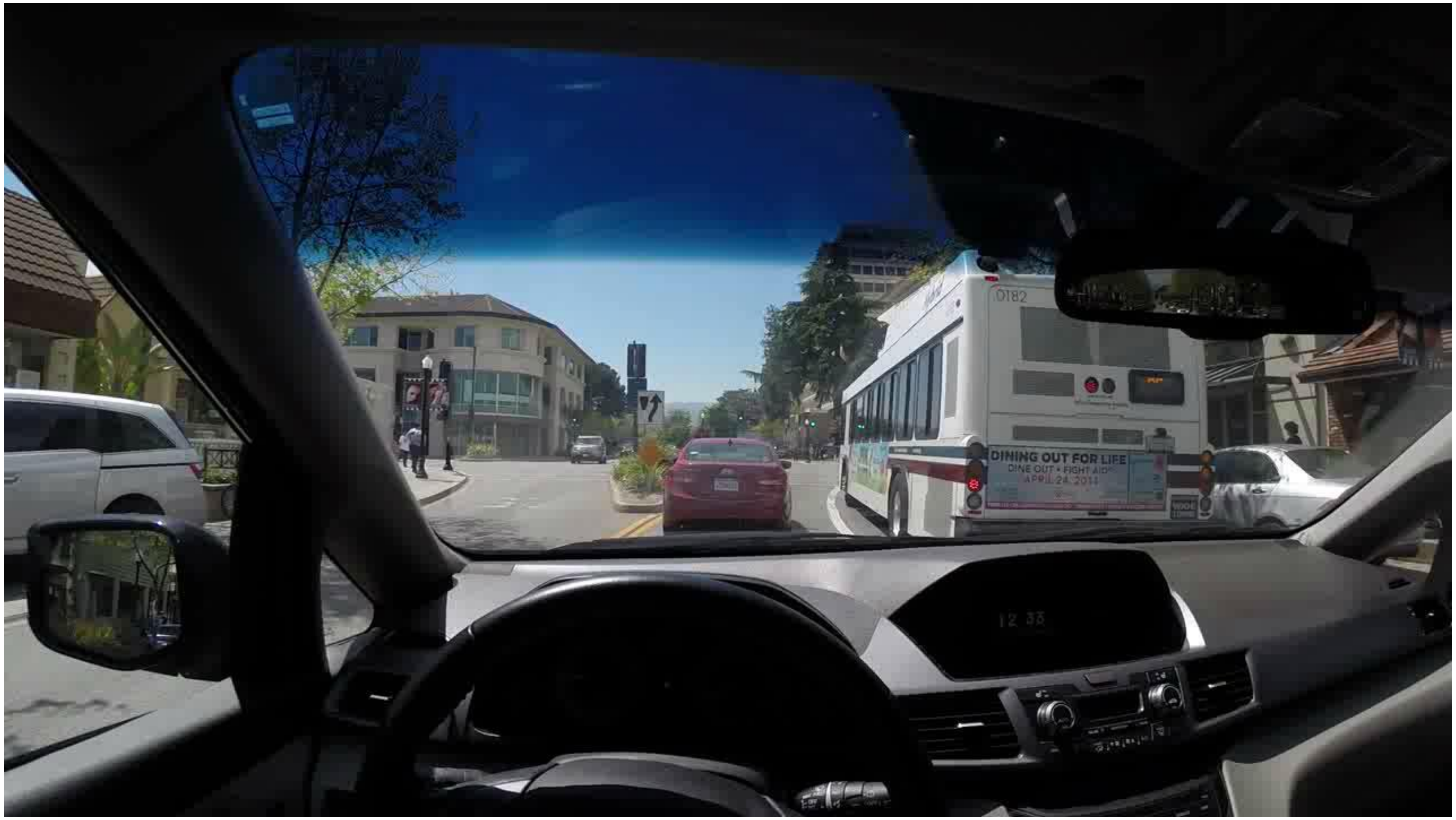
Situated Language Understanding at 25 Miles per Hour

Teruhisa Misu, Antoine Raux, Ian Lane, Rakesh Gupta
Honda Research Institute, USA

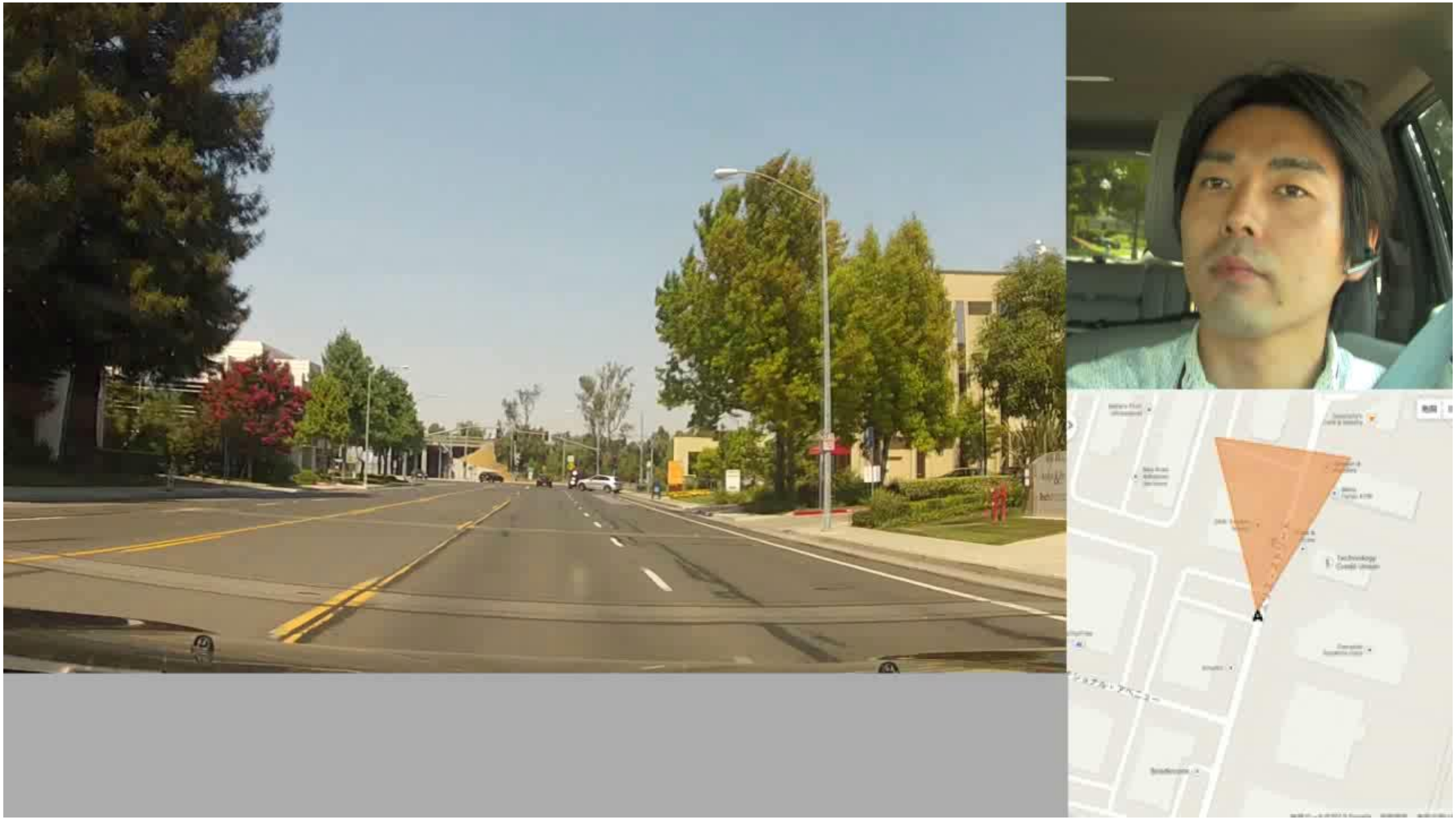
Our goal

(situated spoken interaction in a car)

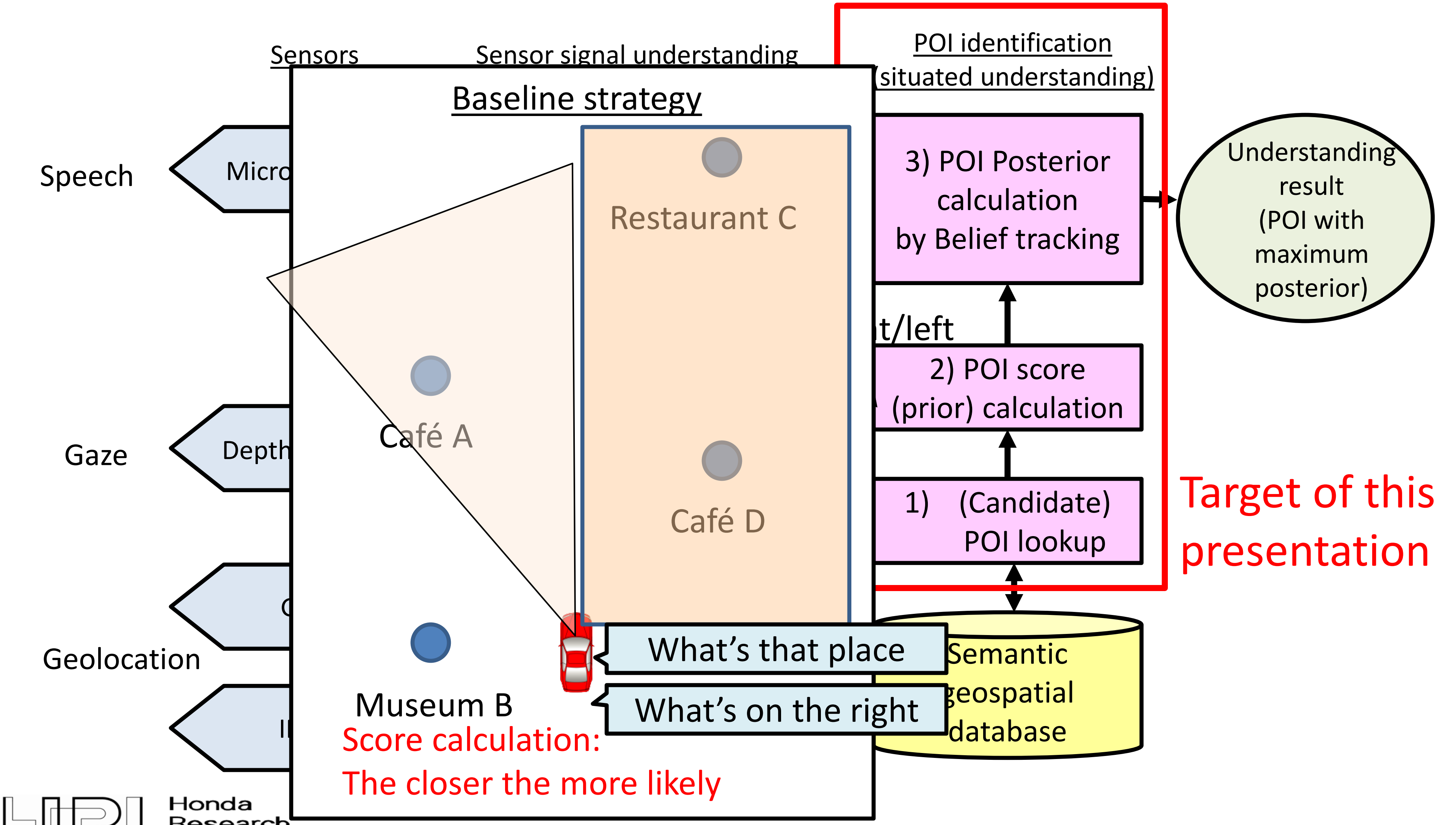


Motivation: “I’d like to know about the business (POI) that I see”

“Townsurfer” System video



System architecture of Townsurfer



Sensors
Sensor signal understanding
POI identification
(situated understanding)
Speech
Micro
Gaze
Depth
C
Geolocation
I
Baseline strategy
Restaurant C
Café A
Café D
Museum B
t/left
3) POI Posterior calculation by Belief tracking
2) POI score (prior) calculation
1) (Candidate) POI lookup
Understanding result (POI with maximum posterior)
Semantic geospatial database
What's that place
What's on the right

Score calculation:
The closer the more likely

Target of this presentation

FAQ:

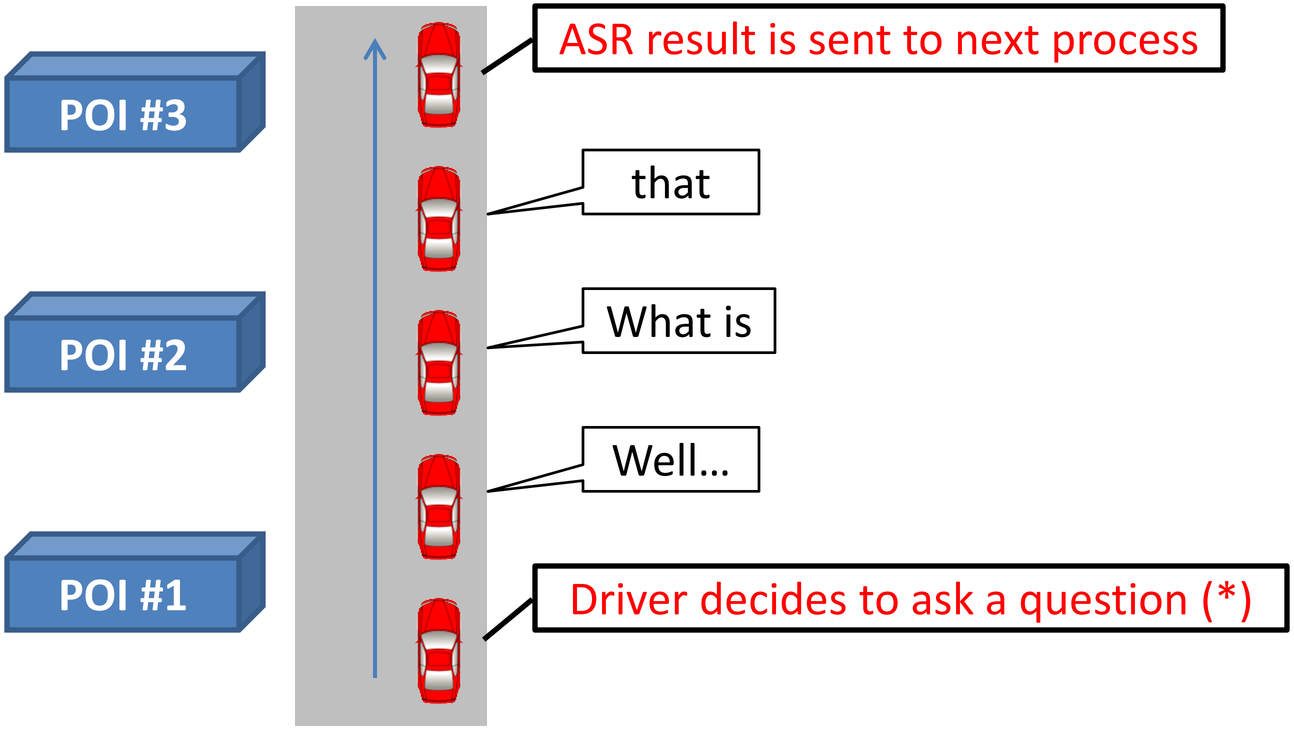
“I understand that the DEMO works well.”

“But, does the system really work for real users?”

Problem 1

1) Timing and spatial relationship

Environment changes very quickly (10m/s)



ASR result is sent to next process

that

What is

Well…

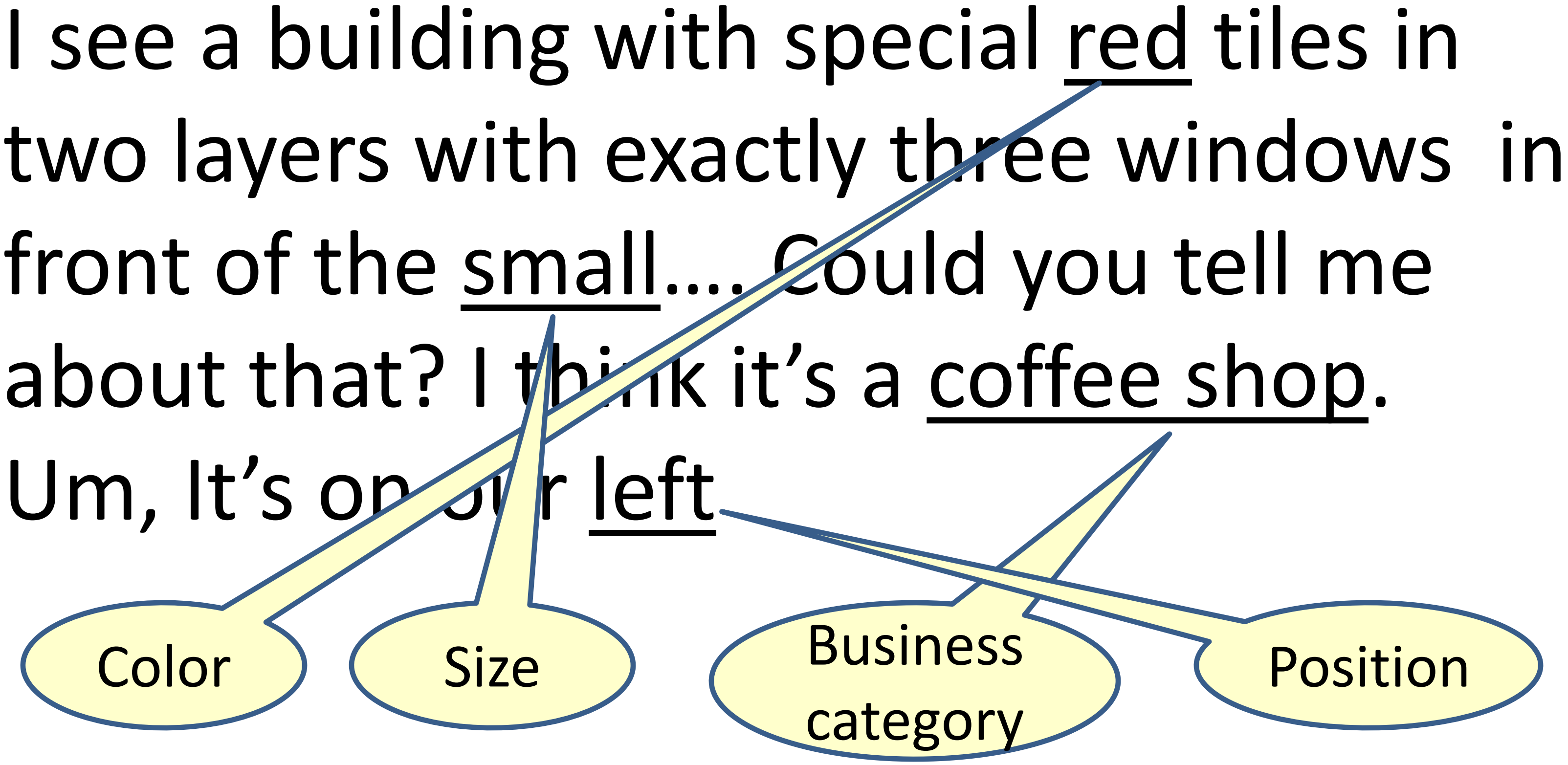
Driver decides to ask a question (*)

The timing of user queries, spatial relationships between the car and targets, head pose of the user

Problem 2

2) Linguistic cues

Linguistic cues are useful, however



I see a building with special red tiles in two layers with exactly three windows in front of the small…. Could you tell me about that? I think it's a coffee shop. Um, It's on our left

Color

Size

Business category

Position

What kinds of linguistic cues do drivers naturally provide?

Our focus issues

1. Timing

← Is timing a important factor? 1-2 sec makes difference?

2. Head pose and spatial distance

e.g. - Does head pose play an important role?

- Or spatial distance is enough?

3. Linguistic cues

← What kind of linguistic cues is useful for POI identification?

→To answer these questions, we need field data

Data collection

- System installed in Honda Pilot experimental car
- Data collection by 14 subjects
  - 399 utterances (w/ valid target) in total
  - Manually annotated user intended POI (business)
- Sites:

Residential area

< 3 POIs in FOV

Downtown (MV)

> 7 POIs in FOV

Data analysis

1. Timing

← Is timing a important factor?

2. Head pose and spatial distance

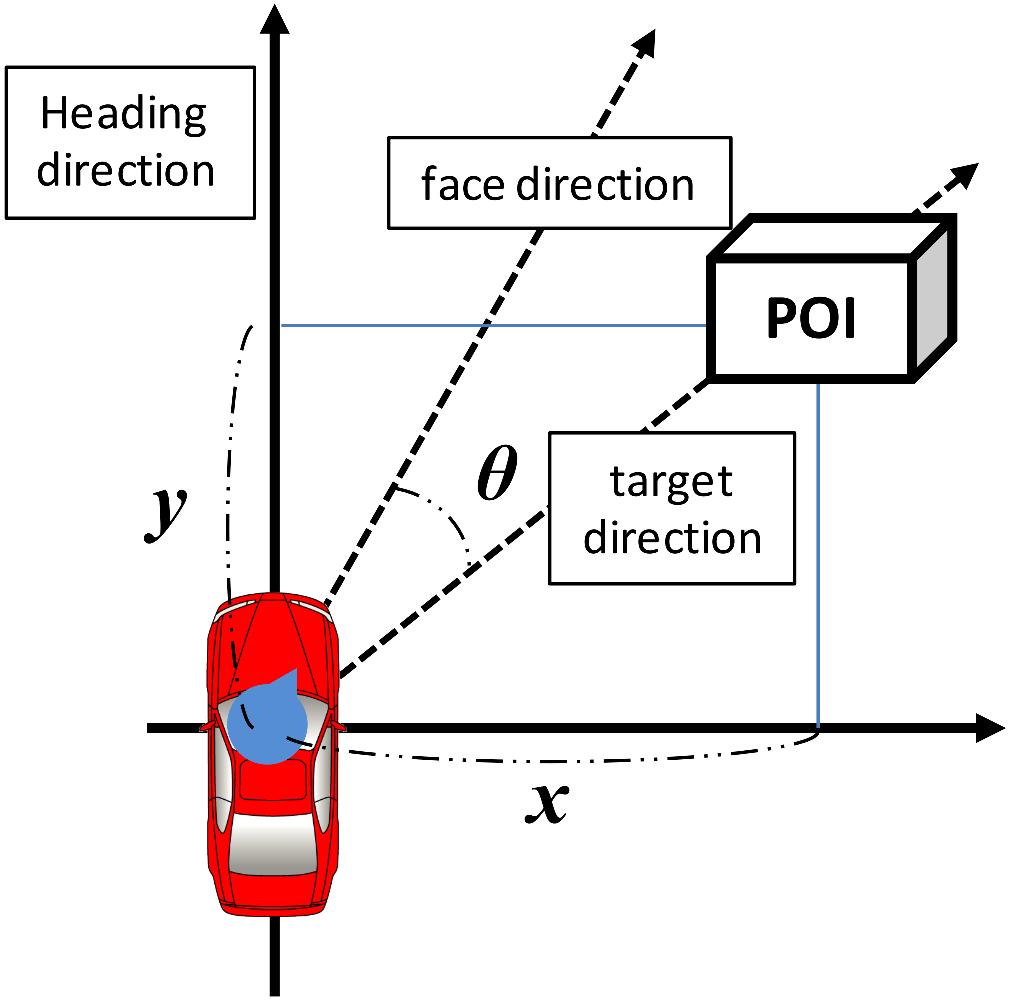
e.g. - Does "right" means "front right" or "side"?

- Does head pose play an important role?

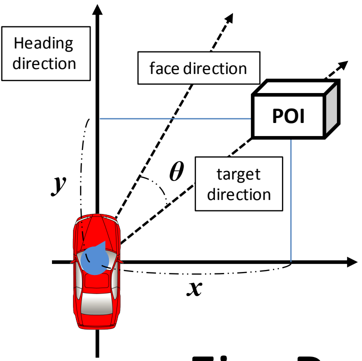
3. Linguistic cues

← What kind of linguistic cues do drivers naturally provide?

Parameters used for the analysis on relationship between car and target

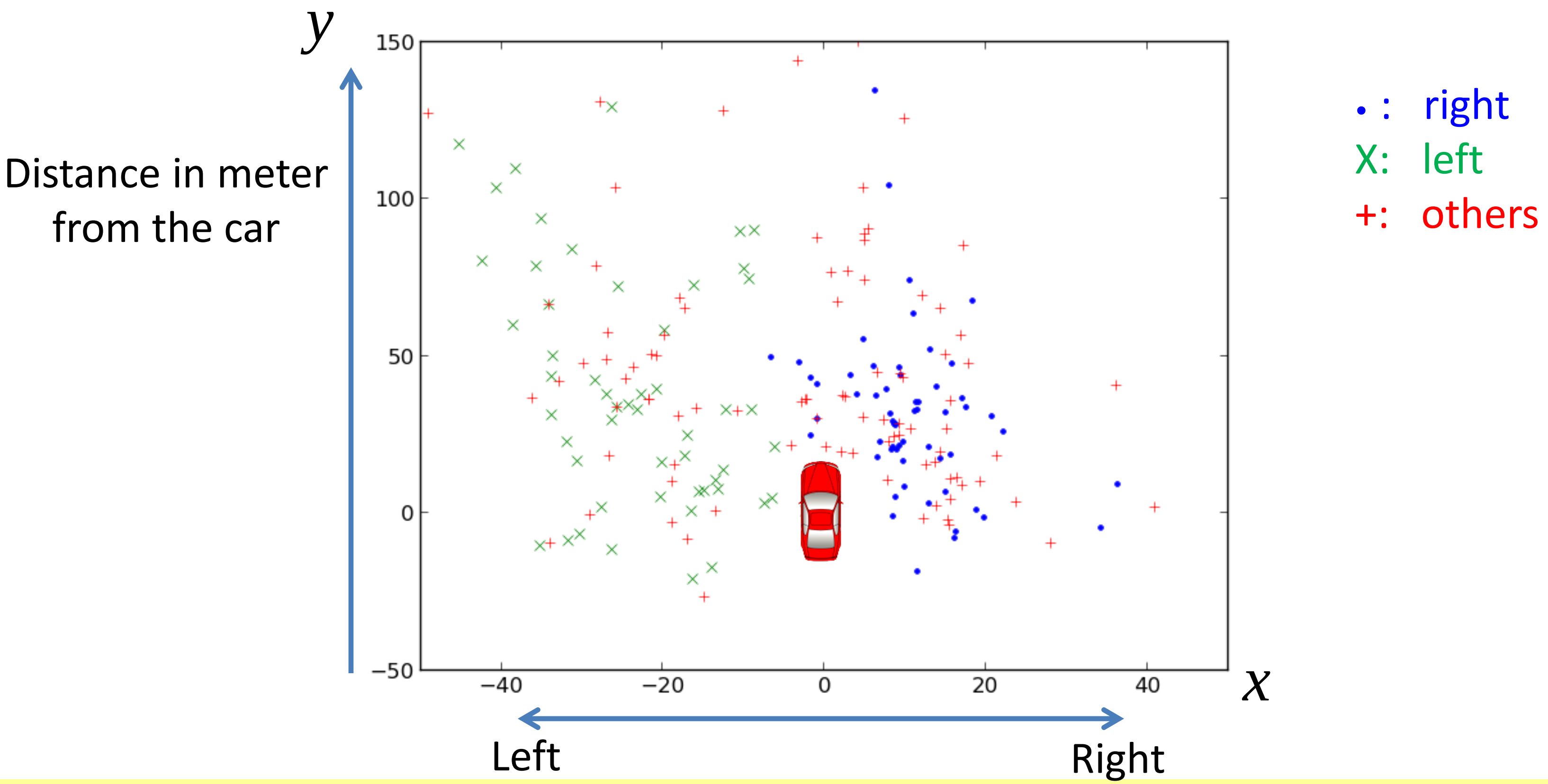


Honda Research Institute USA

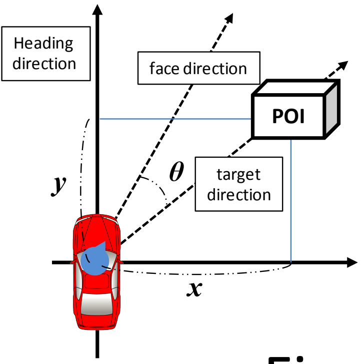


Analysis on POI position

Fig: Relation between target POI positions and position cues

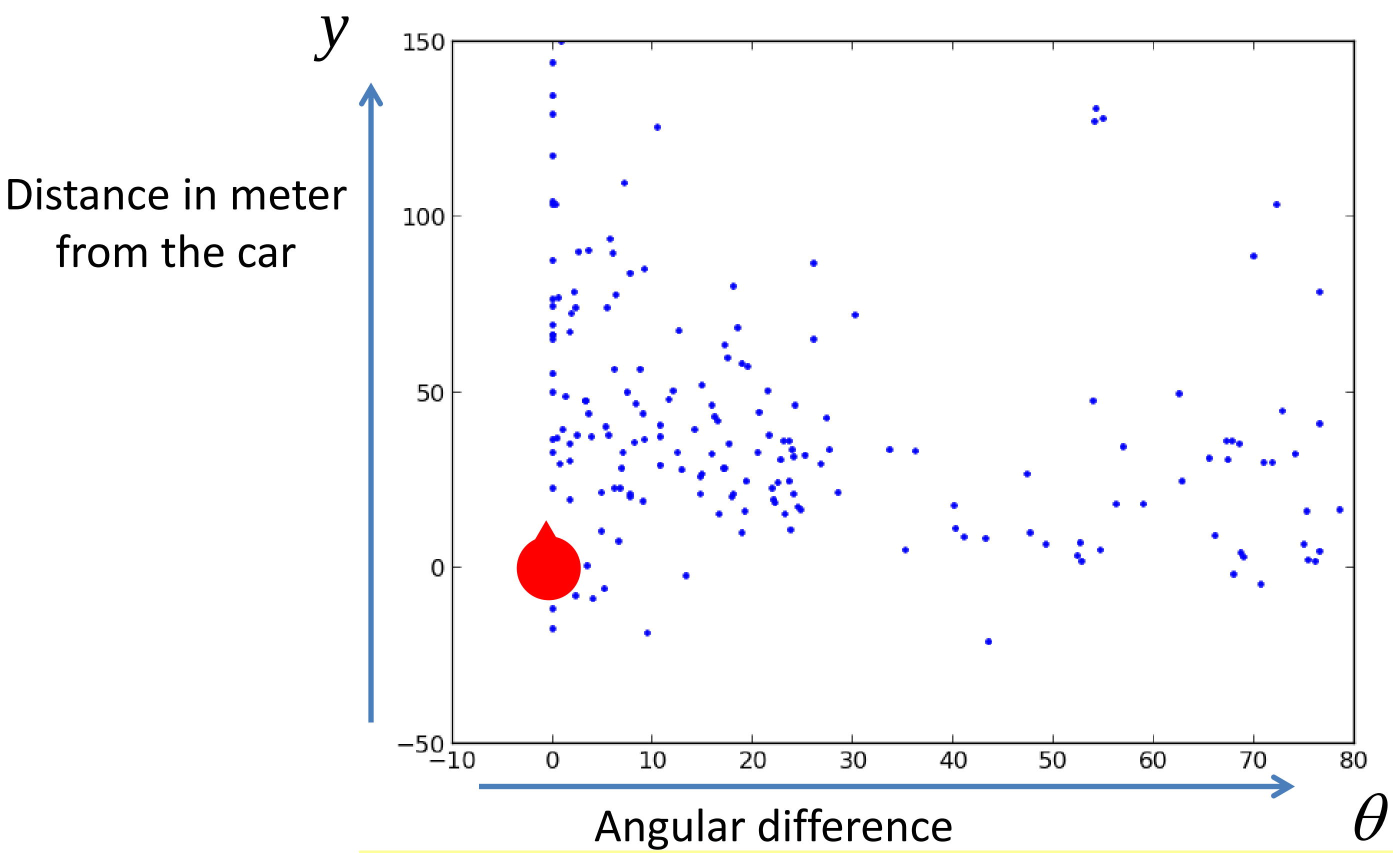


Right and left are powerful cues though distribution has large variance



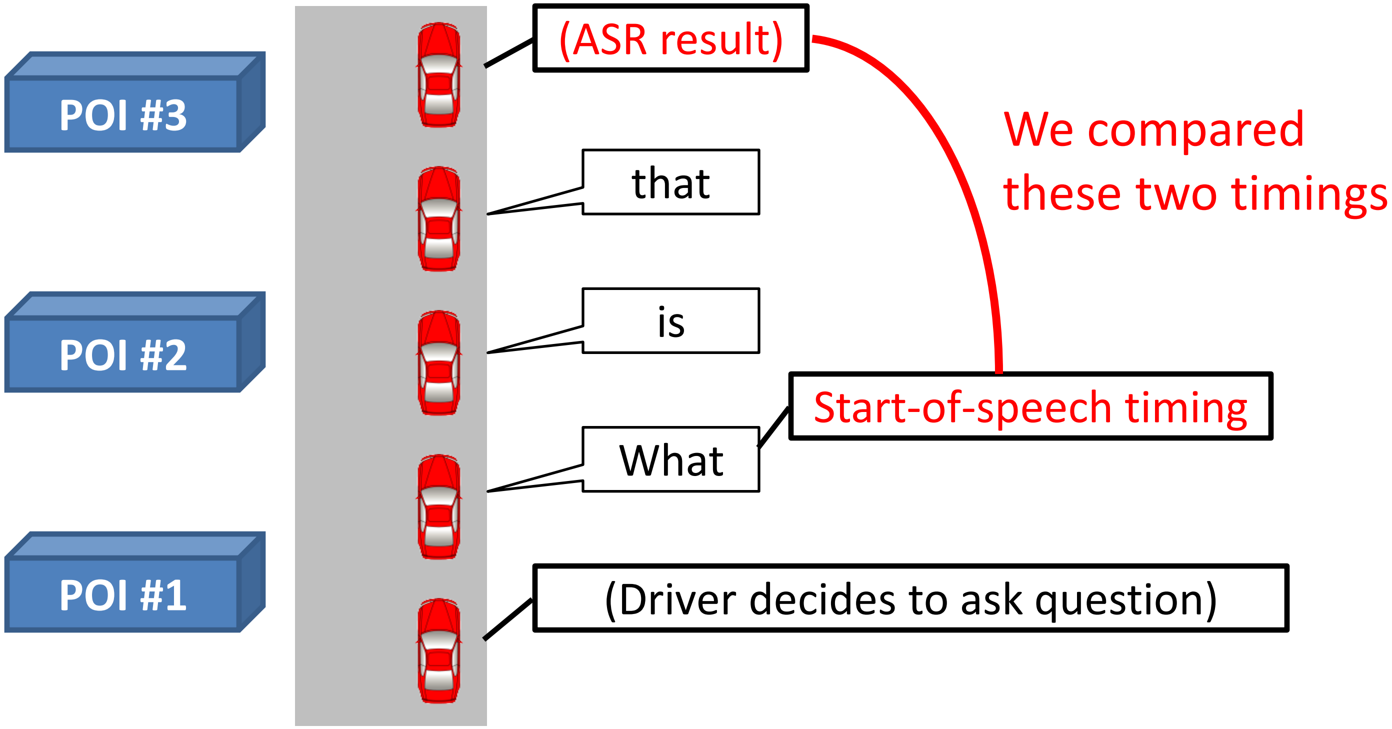
Analysis on effect of head-pose

Fig: Relation between distance and angular difference



- Angular differences for distant targets is often small
- Angular differences for close targets has large variance

Analysis on timing (focusing on axial distance *y*)



Comparison of average and STD of y-distance (in meter) of POI from the car

		ASR result timing		Start-of-speech timing	
Position	Site	Ave dist.	Std dist.	Ave dist	Std dist.
Right/left	Downtown	17.5	31.0	31.9	28.3
	Residential	22.0	36.3	45.2	36.5
No right/left cue	Downtown	17.4	27.8	31.1	26.5
	Residential	38.3	45.9	52.3	43.4

→Presence of a better POI likelihood function using the positions at the start-of-speech timing than using the ASR result timing

Data analysis

1. Timing

← Is timing a important factor?

2. Head pose and spatial distance

e.g. - Does "right" means "front right" or "side"?

- Does head pose play an important role?

3. Linguistic cues

← What kind of linguistic cues do drivers naturally provide?

Major linguistic cues

Analysis of linguistic cues included in the collected utterances (subjective cues are excluded)

Clue	Percentage used
Relative position to the car (right, left)	59.4 %
Category of the POI (e.g. restaurant, gas station)	32.8 %
Color of the POI (e.g. green, yellow)	12.8 %
Cuisine (e.g. Chinese, Japanese, Mexican)	8.3 %
Equipments (e.g. awning, outside seating, sign)	7.2 %
Relative position to the road (e.g. corner)	6.5 %

Position related to the car is most often provided, followed by category, color, cuisine

Comparison of number of linguistic cues user provided to the system

# category per utterance

In downtown: 1.51 categories/utterances

Residential: 1.03 categories/utterances

→Drivers provide cues considering environmental complexity

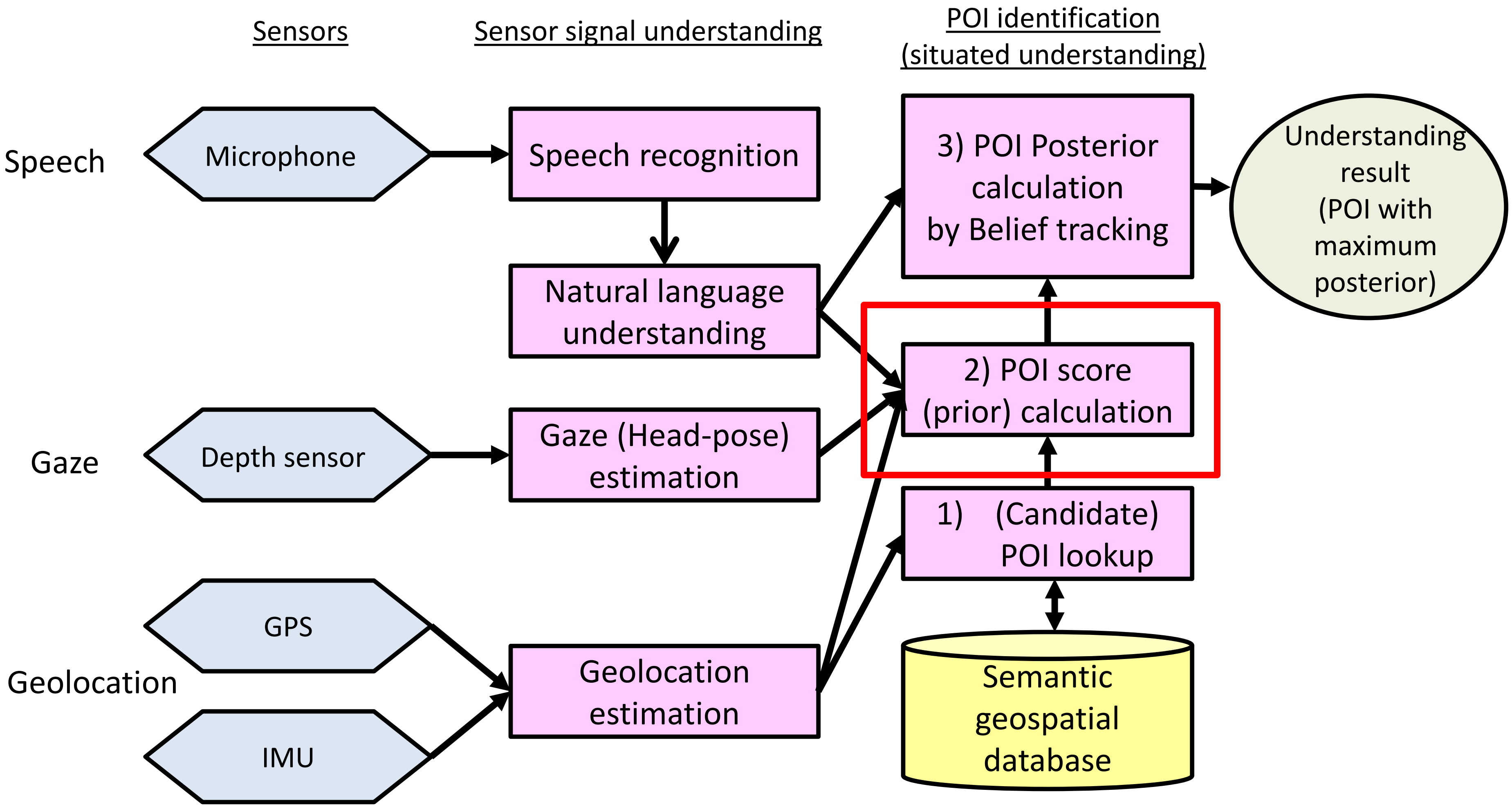
Methods to achieve better POI identification

1. Using start-of-speech timing for the POI likelihood calculation

2. Gaussian mixture model (GMM)-based POI probability calculation

3. Linguistic cues for POI selection

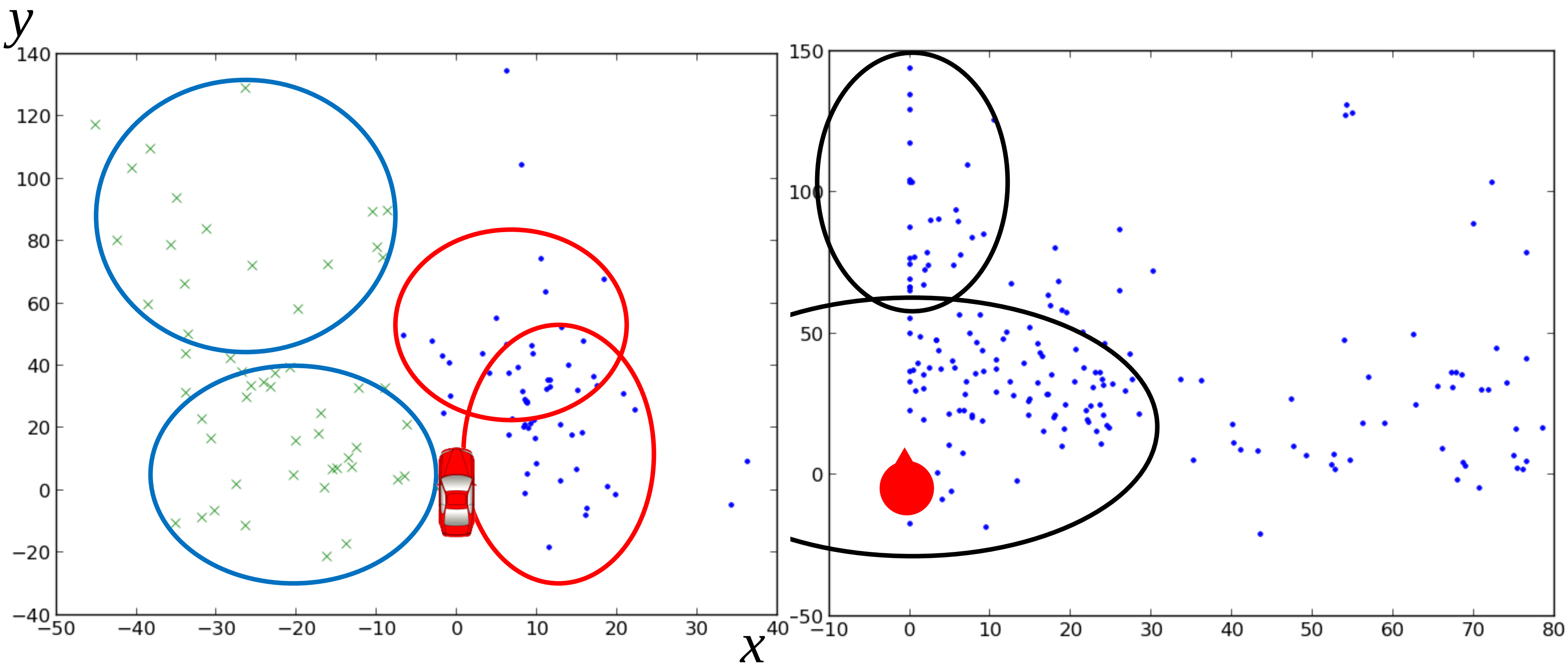
System architecture of Townsurfer


Sensors
Sensor signal understanding
POI identification
(situated understanding)
Speech
Microphone
Speech recognition
Natural language understanding
3) POI Posterior calculation by Belief tracking
Understanding result (POI with maximum posterior)
2) POI score (prior) calculation
Gaze
Depth sensor
Gaze (Head-pose) estimation
1) (Candidate) POI lookup
GPS
Geolocation
IMU
Geolocation estimation
Semantic geospatial database

Honda Research Institute USA

Method 2: GMM-based likelihood calculation

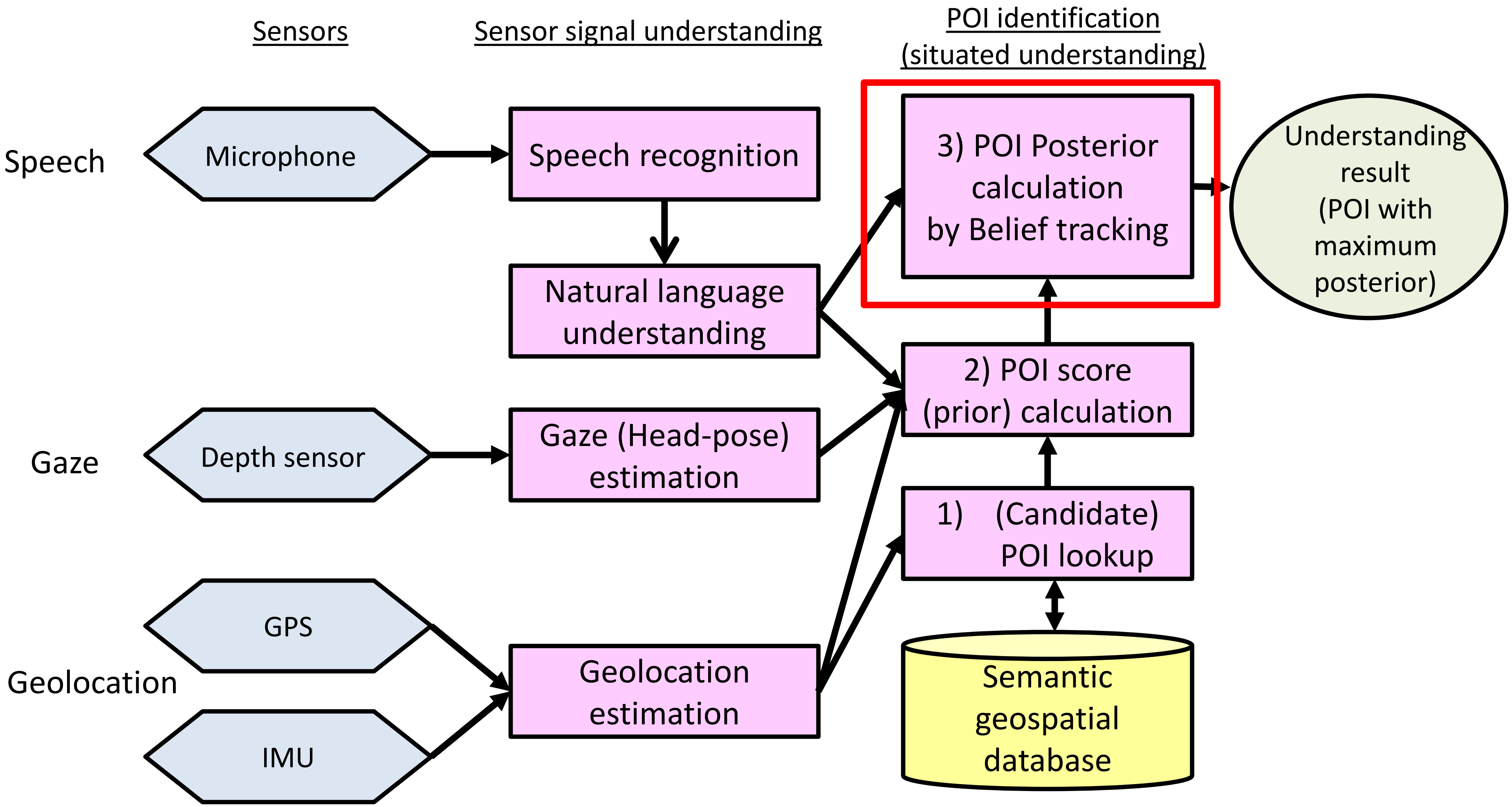
Gaussian mixture model for likelihood calculation



→ Optimized FOV and distance

Honda
Research
Institute USA

System architecture of Townsurfer


Sensors
Sensor signal understanding
POI identification
(situated understanding)
Speech
Microphone
Speech recognition
Natural language understanding
3) POI Posterior calculation by Belief tracking
Understanding result (POI with maximum posterior)
2) POI score (prior) calculation
Gaze
Depth sensor
Gaze (Head-pose) estimation
1) (Candidate) POI lookup
GPS
Geolocation
Geolocation estimation
IMU
Semantic geospatial database

Honda Research Institute USA

Method 3: Linguistic cue using belief tracking

- We use the linguistic likelihood of the following 5 categories
  - Category
  - Color
  - Cuisine
  - Equipments
  - Relative position

→ Remove candidate POIs that do not have the category values specified by the user

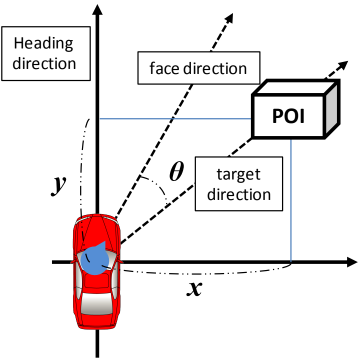
Experiment by simulation

- User-based cross validation
  - Data by 13 drivers for training GMM parameters, List of linguistic cues, the other for testing
- Evaluation based on POI identification rate
  - Task success = Likelihood of the target POI is the highest
- Chance rate is 10%

Evaluation in POI identification rate (chance rate is 10%)

Method	POI identification rate (%)
Right and left linguistic cues, the closer the more likely, ASR result timing **<<Baseline>>**	43.1 %
Baseline + (1) Start-of-speech timing	42.9 %
Baseline + (2) GMM-based likelihood	47.9 %
Baseline + (3) Linguistic cues for belief tracking	54.6 %
(1) + (2)	50.6 %
(1) + (3)	54.4 %
(2) + (3)	62.2 %
(1) + (2) + (3)	67.2 %

- Combination of timing and spatial distance optimizations is important
- Improvement by 24.1% absolute over the baseline method



Breakdown of effect of spatial/gaze information

Feature used as GMM parameter	Right/left	Others
x only	58.6	51.2
y only	59.5	53.7
gaze (θ) only	43.3	44.4
x + y	73.8	54.3
x + gaze (θ)	57.8	48.1
y + gaze (θ)	59.1	54.9
x + y + gaze (θ)	68.4	57.4

Contribution of head pose information is small
← Driver finished looking at the POI and returned the face to the front
→ Use of trajectory information would be important

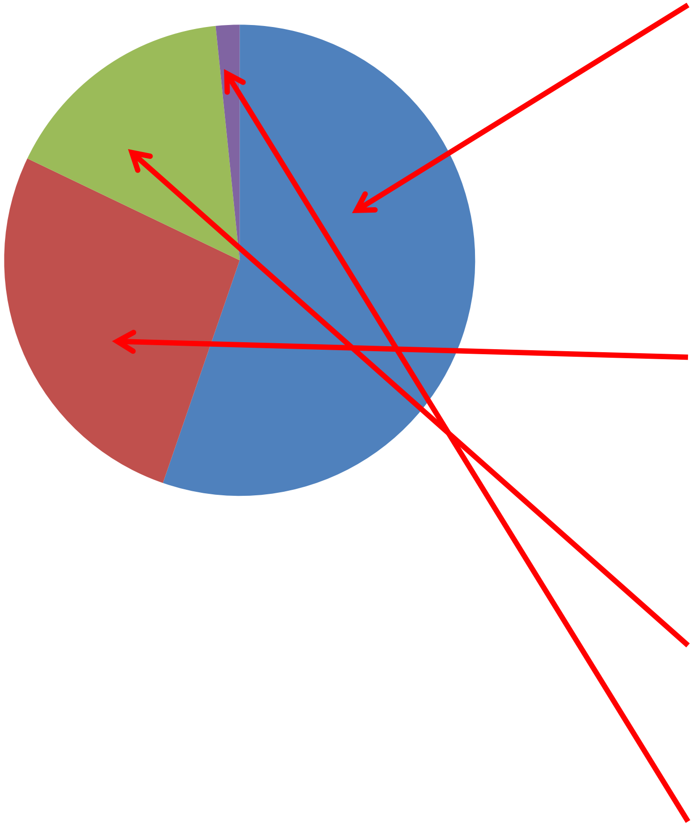
Breakdown of effect of linguistic cues

Category of linguistic cue	POI identification rate (%)
No linguistic cue (*)	50.6
(*) + business category (café, restaurant)	59.1
(*) + color of POI (green, white)	57.6
(*) + cuisine (Chinese, Japanese)	54.1
(*) + Equipment (awnings, outside seating)	53.9
(*) + Relative position (corner)	51.4
All	67.2

- Improvement is proportional to the rate used
- The contribution of the categories readily available is large
- Contribution of linguistic cues is larger in Downtown (20.0% vs. 14.4%)

Error analysis

- Main error causes



- Ambiguous reference:
  There are more than two POIs that corresponds to user query
  (e.g. two green place in row)

- Linguistic cue:
  Use dynamic object as linguistic cue
  (e.g. pedestrian in front)

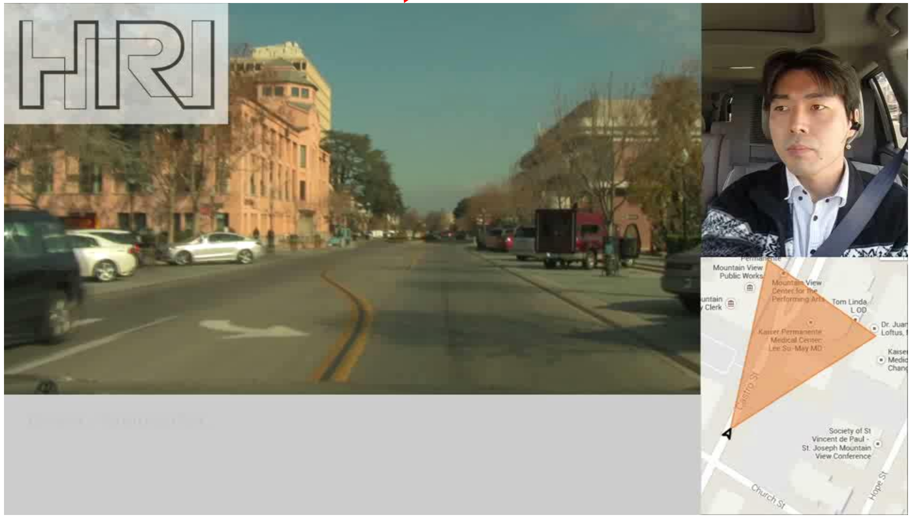
- Localization:
  Error by GPS/IMU

- User error:
  User confused POI’s equipment

Honda Research Institute USA

Summary

This video is attached in the USB proceedings


HRI
Mountain View Public Works
Mountain View Center for the Performing Arts
Tom Linda
Kaiser Permanente Medical Center
Castro St
Society of St Vincent de Paul - St. Joseph Mountain View Conference
Church St
Hope St

- Townsurfer feasibility is demonstrated through real world experiments
  - We collected and analyzed data by 14 users
  - We proposed methods to improve success rate focusing on timing, spatial distance, linguistic cues
  - Limitation of this work comes from small data we collected ←→ Methods we proposed are general
  - Please visit us at MV to see the demo! HRI is hiring!

Success rate per site

Site, Condition	Downtown	Residential area
Without linguistic cues	40.8%	57.5%
With Linguistic cues	60.8%	71.9%

Success rate vs # Gaussian component

# Gaussian component	Success rate
1	62.9 %
2	67.2 %
3	66.1 %
4	67.2 %
5	66.2 %

Possible solutions to enhance user experience

1) Clarification strategy
2) Eye tracker
3) POI identification using face direction trajectory
4) Feedback

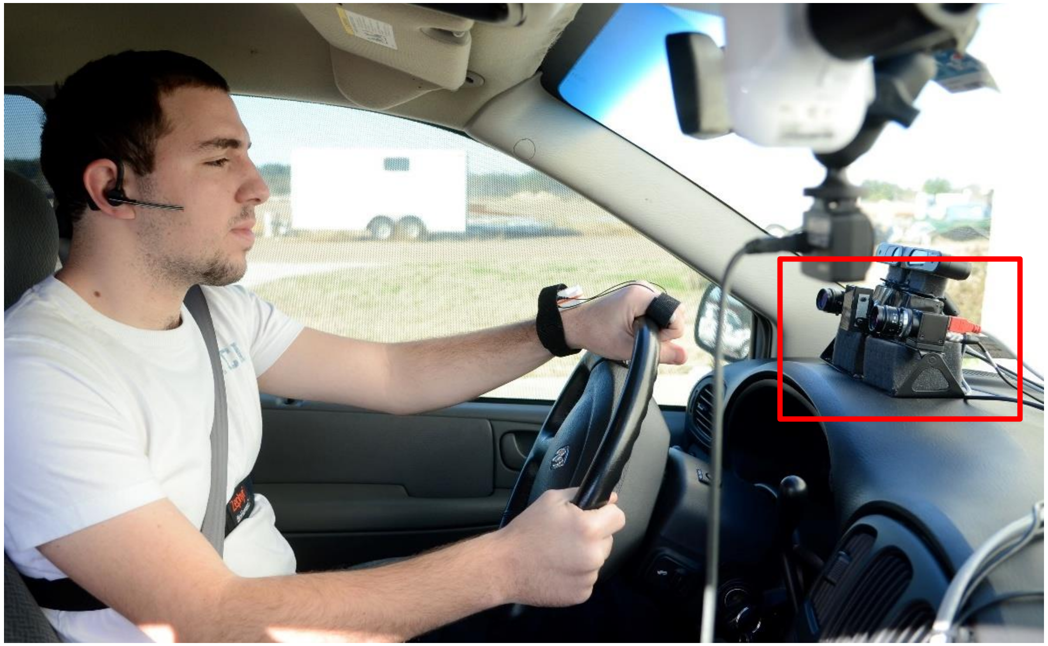
1) Clarification strategy

Most errors are ambiguous references

- Confirmation like human
  - “Did you mean the one in front or back?”
- Visual confirmation
  - “Please select from the followings restaurants.”

2) Eye tracker

- Issues: Eye tracking vs. Face direction
  - Performance in a car
  - Cost of the sensor



3) POI identification using face direction trajectory

- Our analysis showed that the use of face direction sometimes degrades the POI identification performance

- Using a trajectory of face direction will change the result

4) Visual feedback

Feedbacks might fundamentally change the story