Predicting When People will Speak to a Humanoid Robot

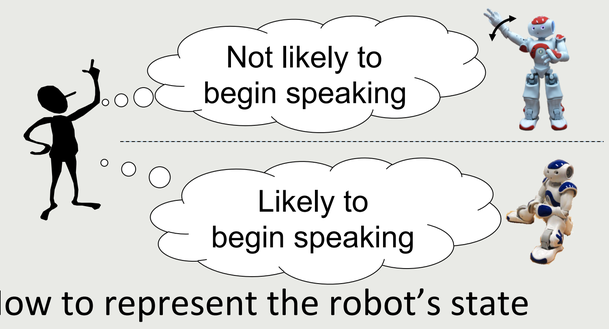
Takaaki Sugiyama, Kazunori Komatani, Satoshi Sato (Nagoya Univ.)

NAGOYA UNIVERSITY

Summary

Goal Predicting whether the user is likely to begin speaking

Approach Machine learning based on the robot's state

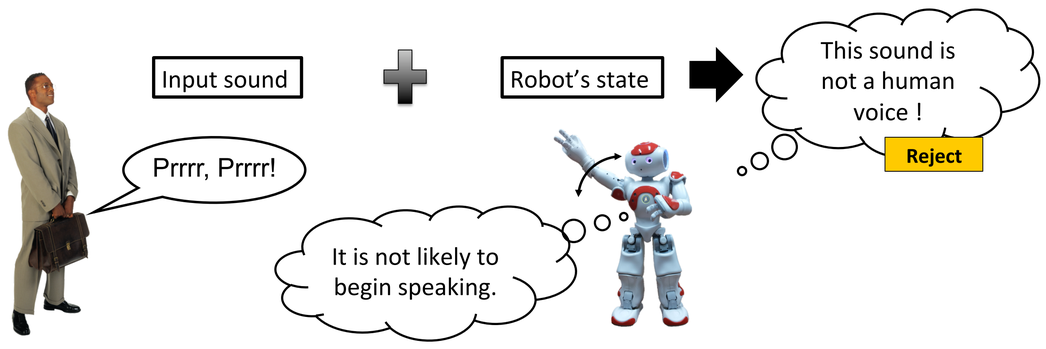


Issue

1. How to represent the robot's state
2. How to prepare the training data
   (Timing when users begin speaking may depend on individuals)

What can "whether user is likely to begin speaking" be used for ?

- **The model can be used to distinguish user utterances from noises**
  - **Prior probability on whether cooperative users begin speaking or not**
    - **Users do not utter in arbitrary timing**
    - **Especially when the robot is anthropomorphized**



Proposed Method

Input 9 features

Utterance

(1)	Speech interval
(2)	Utterance pattern
(3)	Prosody

Motion

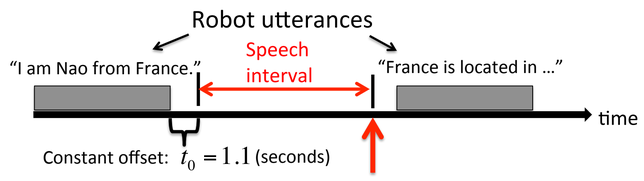
(4)	Head
(5)	Left arm
(6)	Right arm
(7)	Legs

Head/eye direction

(8)	Horizontal
(9)	Vertical

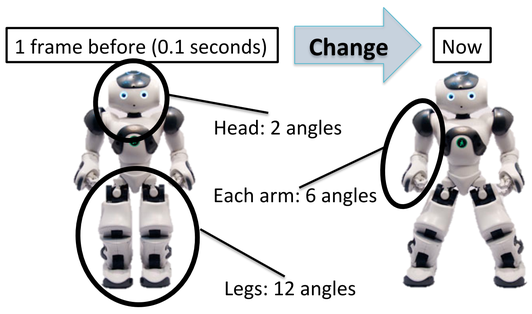
(1)Speech Interval

- elapsed time from the end of the previous robot's utterance



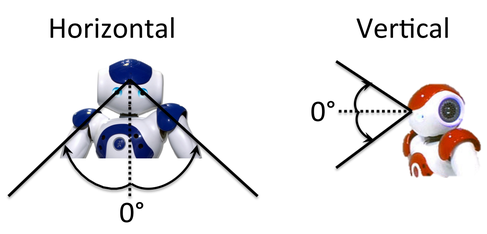
(4)-(7) Motion

- The changes in joint angles of the robot
- The robot has 26 joint angles
  - Summing up the difference of each part



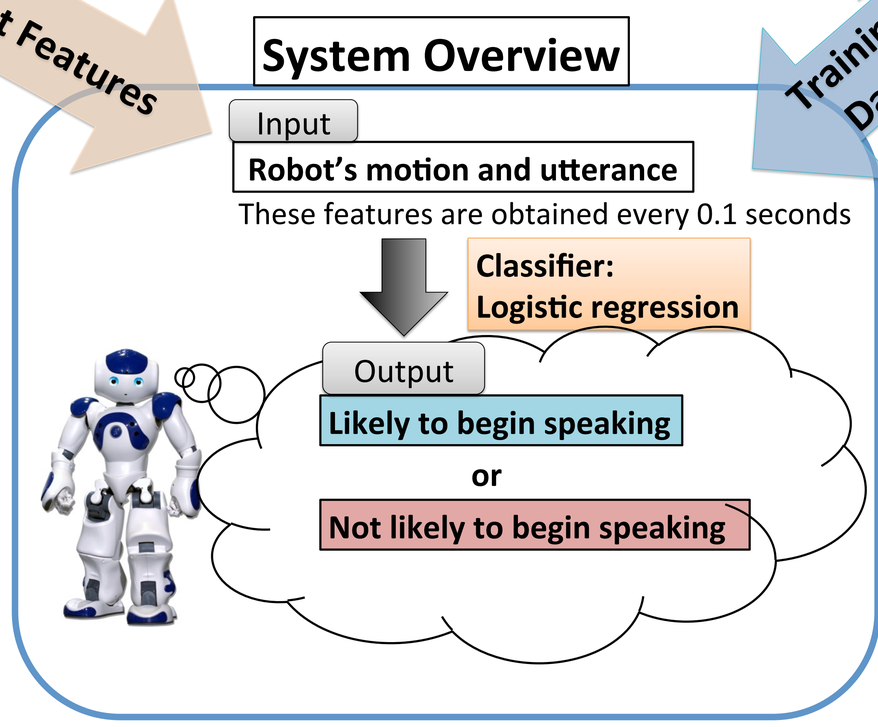
(8)(9) Head/Eye Direction of Robot

- Representing how much the robot heads toward the user

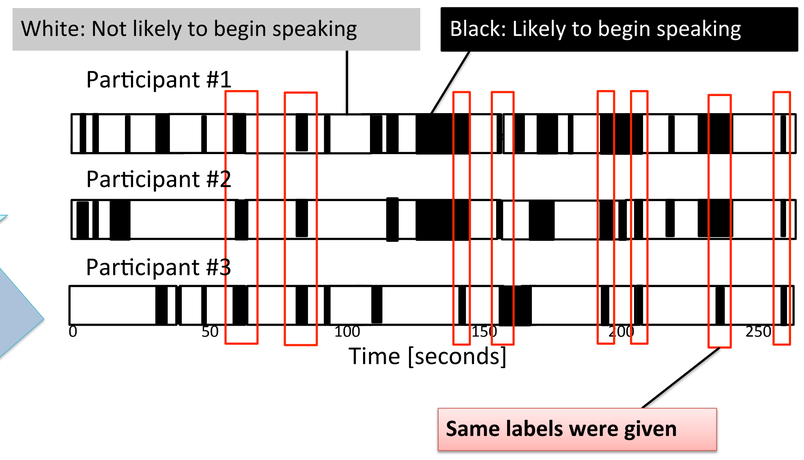


- The user position: in front of the robot

System Overview



1. Labels by 3 participants 2. Using common parts



1. Three participants gave the labels for the robot's behavior

- Using a GUI shown on a computer display:
  - "Likely to begin speaking" : continue to push
  - "Not likely to begin speaking" : release

2. Using parts to which they gave the same labels

- To eliminate the effect of individuality

Sequence A 150.0 sec.

Sequence B 259.3 sec.

Sampled by 0.1 seconds

	A	B
Number of labels	1500	2593
Same labels were given	**1350**	**1430**
Different labels were given	150	1163

Training and test data	A	B
Likely to begin speaking	142	161
Not likely to begin speaking	1208	1269

gave weights to "Likely to begin speaking" according to the ratio of the two labels

Training Data A

Training Data B

Evaluation

- Target data A
  - Prediction accuracy: $\frac{\text{number of correctly predicted labels}}{\text{number of all agreed labels among three participants}}$
  - Evaluate the model for data A

Trained model by	Relationship with target	Prediction acc. (#: 1350)
Data set A	10-fold cross validation	87.4%
Data set B	**Open**	**88.5%**

Almost equivalent performance

does not depend on its training data

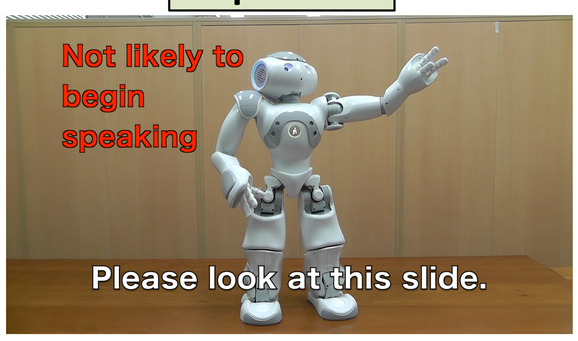
[Demo] Predicting from Robot's Actions (The model was trained on data B)

Our model can predict whether people feel possible to begin speaking for all sequences

Sequence A

- Less motion
- We made this first
- Open test

Sequence B



- We added more motions
- The model was trained on this data (closed)

Unknown