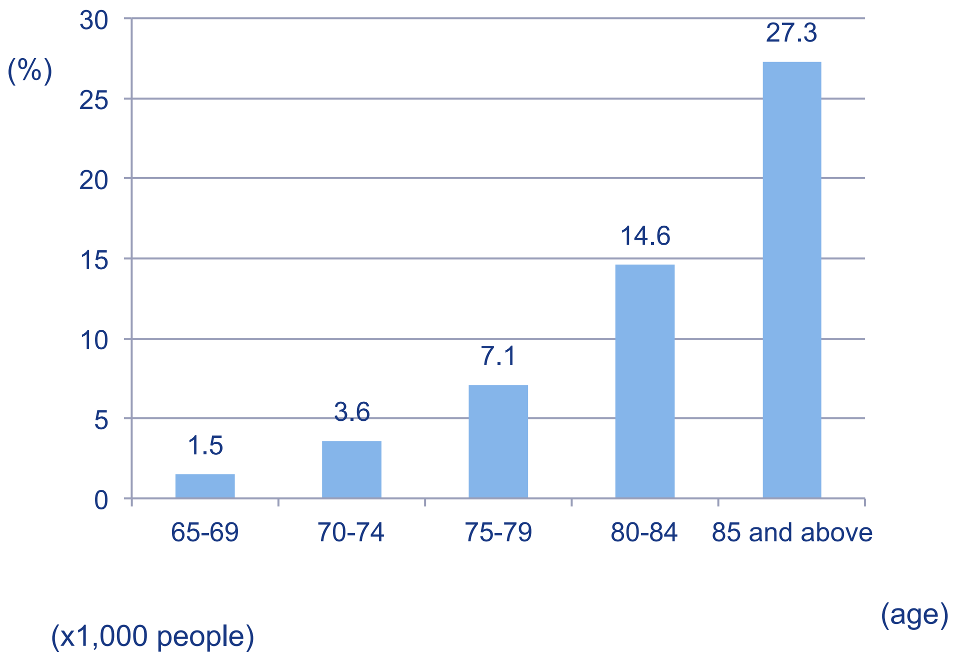
Exploring the Potentiality to Estimate Speaker's Attitude by Low-level Features in Active Listening Conversation

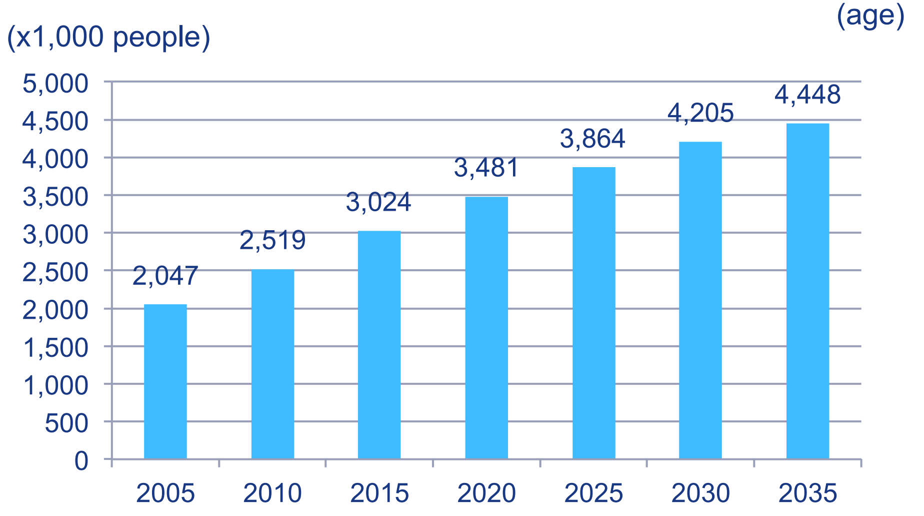
Hung-Hsuan Huang*, Kanehiro Kubushiro, Sayumi Shibusawa, and Kyoji Kawagoe

College of Information Science & Engineering,
Ritsumeikan University, Japan

Background

- Dementia
  - A loss of cognitive ability in a previously unimpaired person, beyond the degree expected from normal aging
- Rapidly increasing number of dementia patients in developed countries
  - Around 10% of above-65 population
  - Increasing to 4.4 million (4.1%) by 2035 in Japan
- Decreasing number of younger generations (Japan)
  - Total population: 127m -> 110m
  - 15-59 population: 56.1% -> 48.6%





[Kurita 2008]

(year)

Care and Support

- No effective treatment to heal yet
- Decay of cognitive ability can be slowed down
  - Reminiscence (photos, songs, etc.)
  - Life review
  - Active listening volunteers
  - Group talk of patients
  - Robots
- High cost of human caregivers
  - Laborious
  - Few volunteers

[Otake 2009]

[Kanoh 2010]

Communication at Higher Level

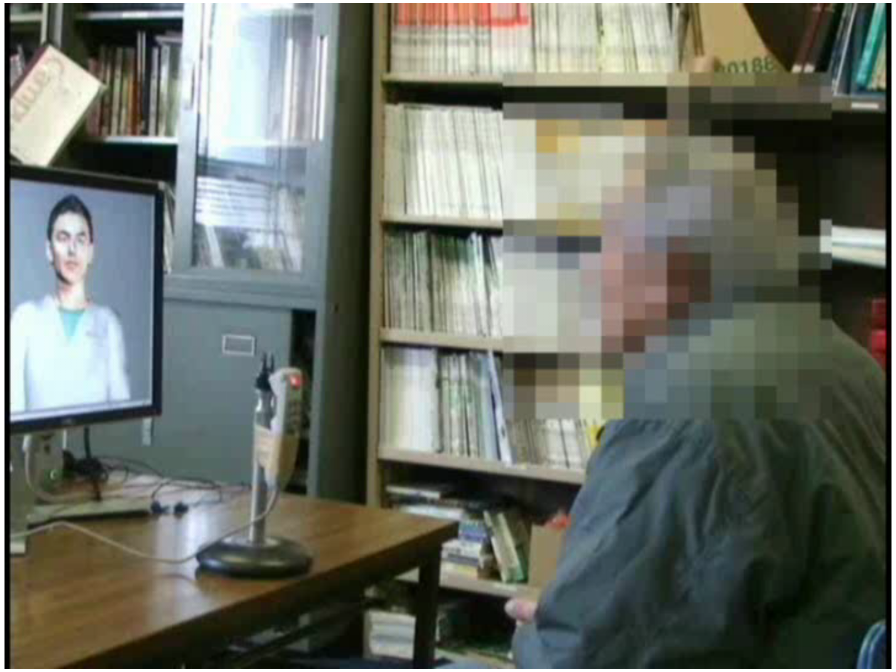
- Rapport agent [Huang 2011]
  - Build rapport with the subjects by low-level signals like nodding and smiles
  - Low-level signals to low-level signals with the rules based on literatures
- Companion agent [Vardoulakis 2012]
  - Long-term relationship
  - Field study with volunteers
  - Wizard-of-Oz experiment
  - The agent did not have interactive backchannel behaviors
- General issues
  - Based on the empirical results with Western subjects

[Huang 2011]

[Vardoulakis 2012]

Active Listener Agent

- Prototype
  - Agent-initiative dialogues
  - Backchannel feedback timings generated from acoustic features of user's voice
  - Natural language understanding based on matching with QA templates defined in advance
  - The dialogue is not personalized yet
  - Pilot test was encouraging. The patients were happy with the agent because they don't feel inferiority
- **Goal:**
  the elderly enjoys the talk with the agent and speak much

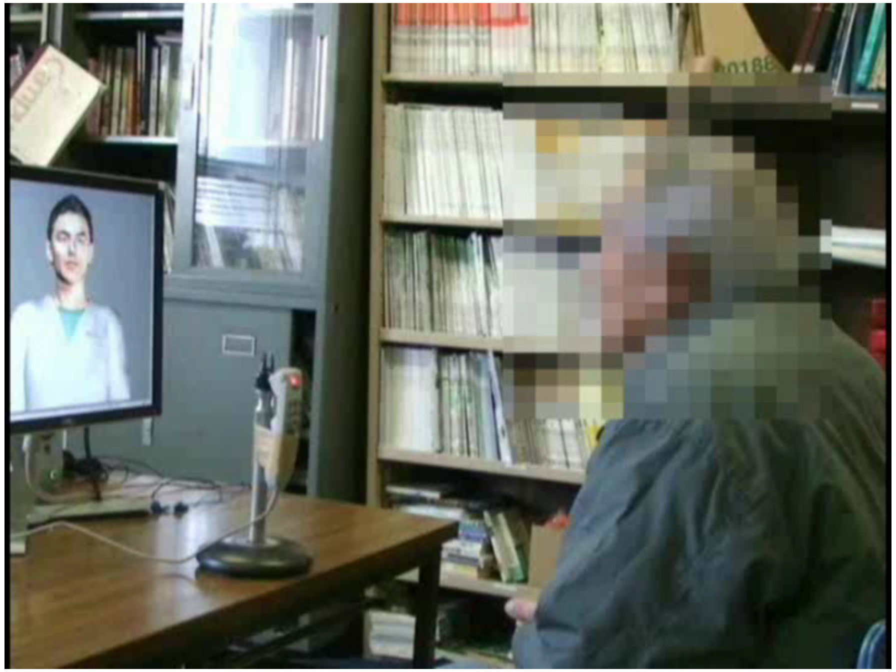


[Nonaka 2012]

- **Active listening:**
  actively follow the talk of speaker: show interests, ask questions, agreeing attitude, etc.

Active Listener Agent

- Prototype
  - Agent-initiative dialogues
  - Backchannel feedback timings generated from acoustic features of user’s voice
  - Natural language understanding based on matching with QA templates defined in advance
  - The dialogue is not personalized yet
  - Pilot test was encouraging. The patients were happy with the agent because they don’t feel inferiority

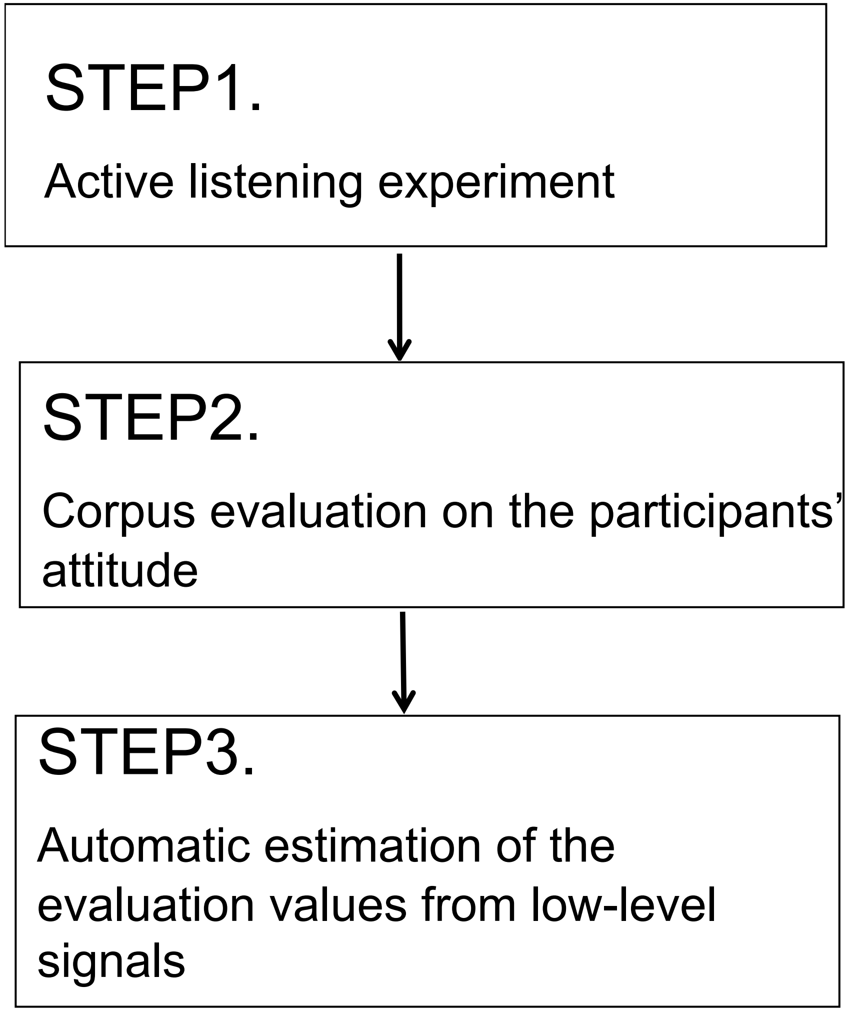


[Nonaka 2012]

- **Goal:**
  the elderly enjoys the talk with the agent and speak much

- **Active listening:**
  actively follow the talk of sp[illegible] interests, ask questions, agreeing attitude, etc.

Procedure



STEP1.

Active listening experiment

STEP2.

Corpus evaluation on the participants' attitude

STEP3.

Automatic estimation of the evaluation values from low-level signals

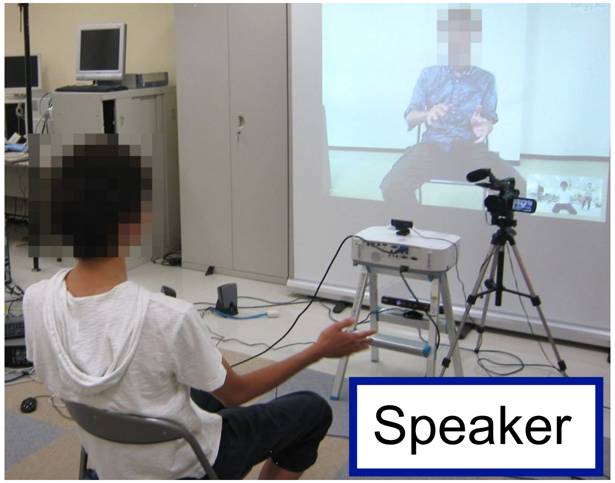
- **Experiment Setup**
  - Participant: 9 pairs of college students (5 male, 4 female)
  - Native Japanese speakers
  - Close friends
  - Average age: 22.1

STEP1. Active listening experiment

- Role: interchanged in the sessions
  - ◆Speaker: talk about his / her family
  - ◆Listener: active listening
- Participants: separated into two rooms
- Topic: pleasant / unpleasant experience with family

Session	Topic	Speaker	Listener
1	Pleasant experience with family	A	B
2		B	A
3	Unpleasant experience with family	A	B
4		B	A

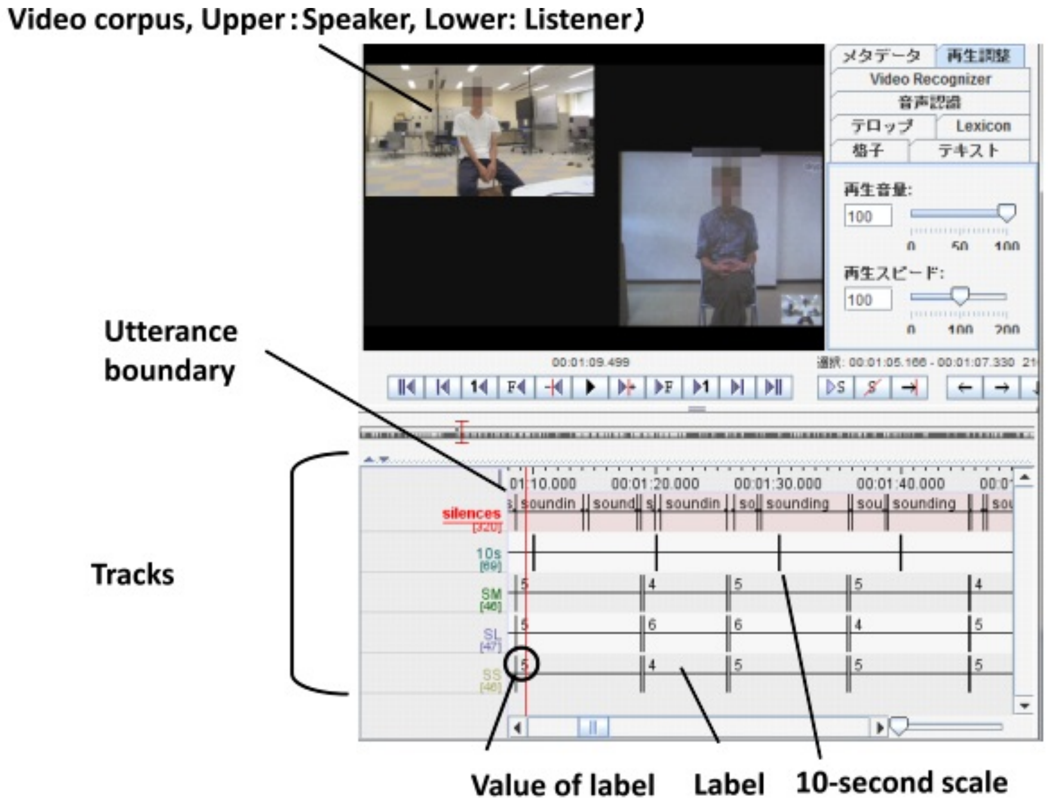
* 7 minutes each



STEP2. Corpus evaluation

- Third person (2 males and 2 females) evaluated the recorded video on the attitude of the speakers and the listeners
- The evaluators were not in either room, have no or little knowledge of the participants (shares similar abilities as the agent)
- Use annotation tool, ELAN, immediately after the experiment
- Subject evaluate with 7-scale measurement (1~7)

Annotation



- – Time lines filled without blanks
- – Label boundaries aligned to utterance boundaries
- – One label can cover multiple boundaries
- – Maximum length of labels is 10 sec.

Video Corpus

- Example of positive attitude

※S：Speaker, L：Listener

S：Maybe my mom overprotects me.

L：Maybe.

S：Must be overprotective.

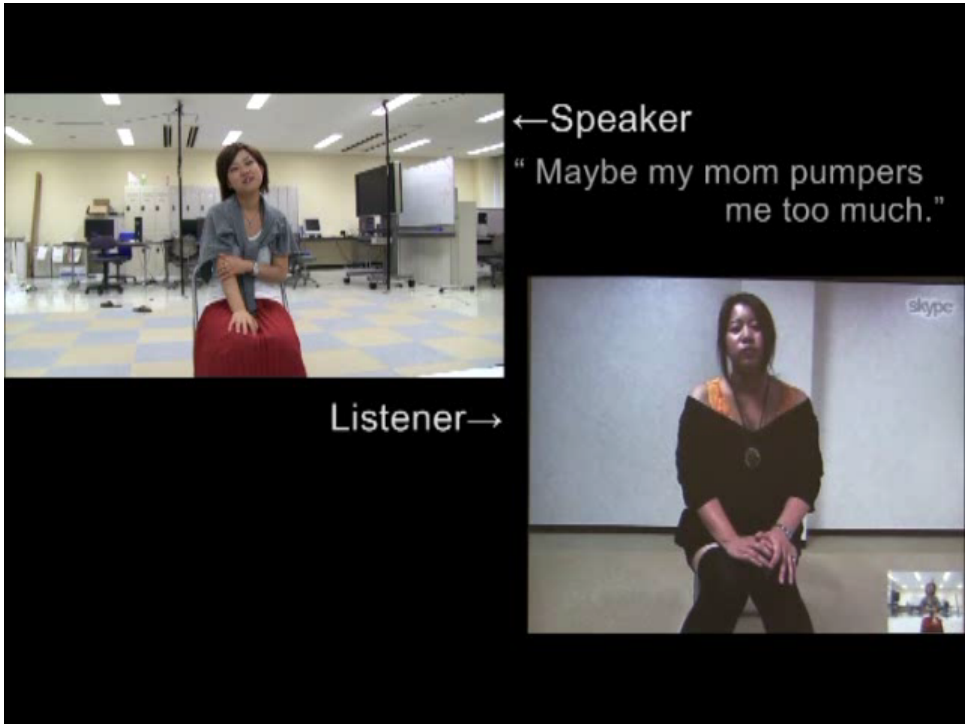
L：Uh-huh.

S: Well so…

L：Overprotective mother, right?

S：Yes, quite

L：The child should have hard life .



Video Corpus

■ Example of negative attitude

※S：Speaker, L：Listener

S：Do you remember that whether you ever rode my car?

L: I don’t think so.

S：Well, probably no.

L：Yes, I didn’t.

S：Maybe we were on a rental car.

L：Rental car. We rented a car when we went to travel.

S： O-oh, we are not talking about family.

L：What?

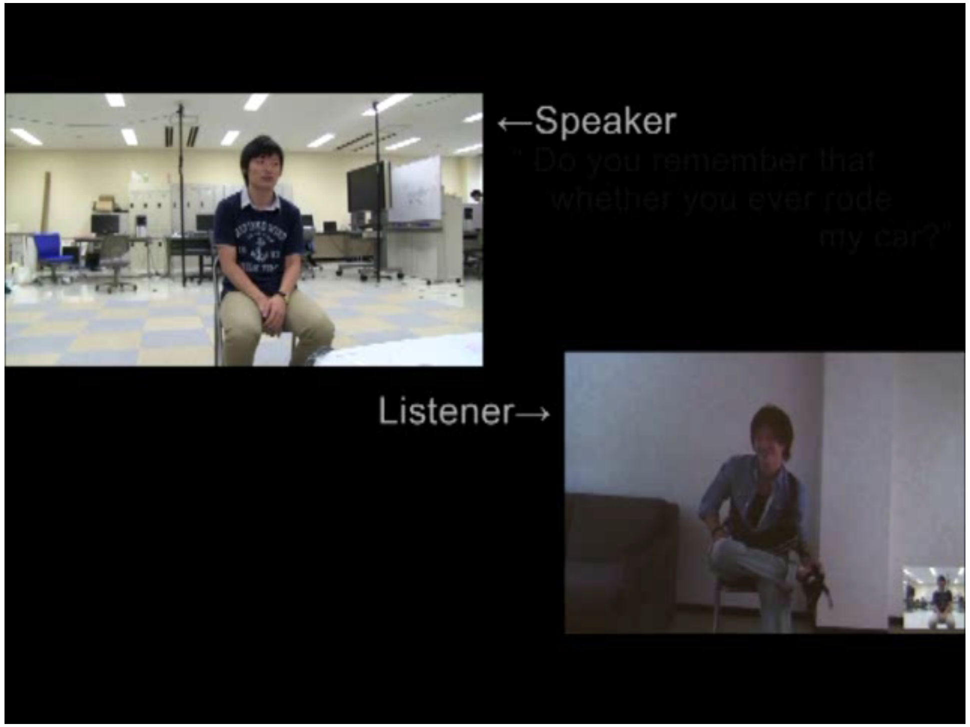
S: We should talk about my family.

L: Just because your talk was so boring.

S：Hmm….

L：Hey, give me more interesting stories.

S: Oh…well…let me think.



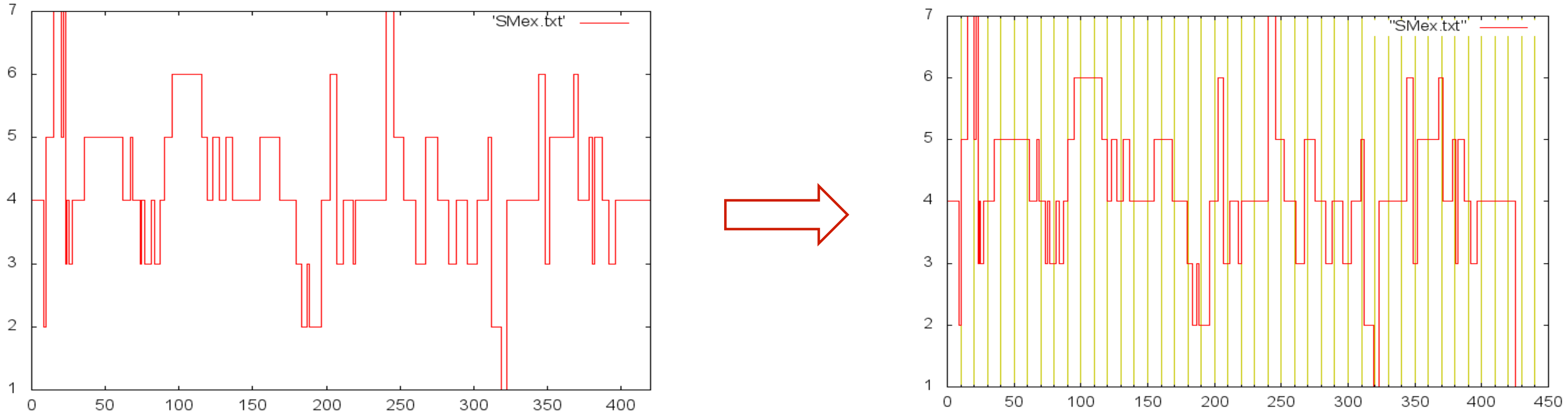
Normalization of label values

Evaluator	1	2	3	4	5	6	7	m	σ
A	15	39	80	235	253	88	46	4.48	1.27
B	26	50	92	188	345	55	23	4.33	1.25
C	37	98	193	413	276	78	34	4.03	1.27
D	22	54	113	302	284	138	67	4.48	1.33

Z Score

Evaluator	1	2	3	4	5	6	7
A	-0.781	-0.562	-0.343	-0.124	0.096	0.315	0.534
B	-0.802	-0.603	-0.405	-0.207	-0.009	0.190	0.388
C	-0.750	-0.499	-0.249	0.002	0.252	0.503	0.753
D	-0.772	-0.544	-0.316	-0.089	0.140	0.367	0.595

STEP3. Automatic estimation



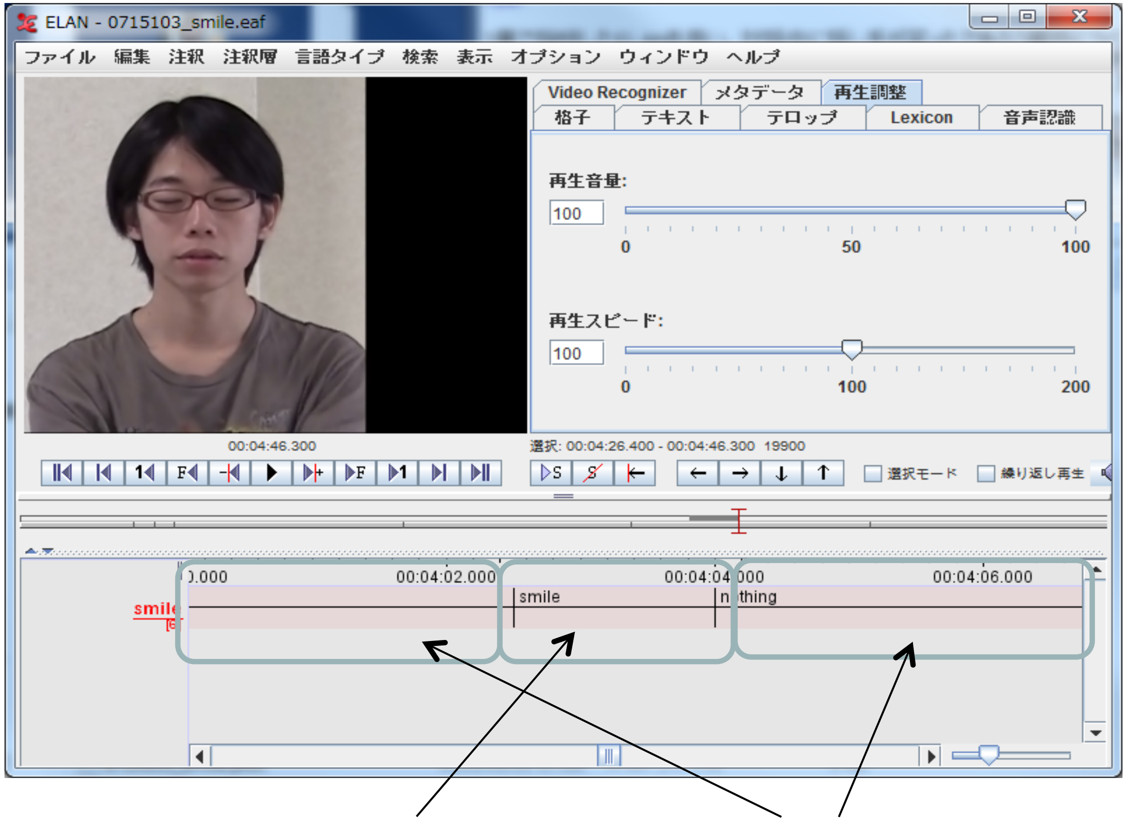
- Label: wave-form like data streams
- Sampling rate: 10Hz

  ※Shortest length of the label: 0.244 sec
- Data normalization: *z*-score

Low-level signals

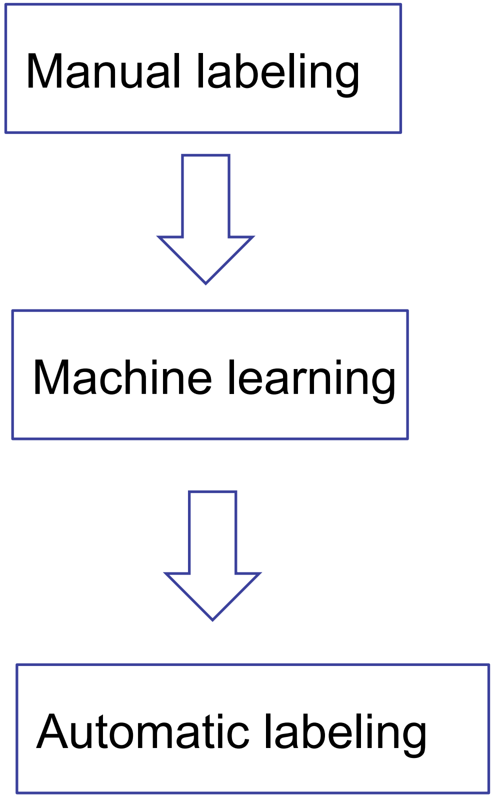
- Face activities
  - – Smile: may imply pleasant mood
  - – Nod: may imply agreement to or the willing to listen to the partner’s opinion
- Speaking frequency
  - – May imply the willing to talk to the partner

Extraction of face activity values



Positive instance

Negative instances



Manual labeling

Machine learning

Automatic labeling

* Software used: visage|SDK

Classification results

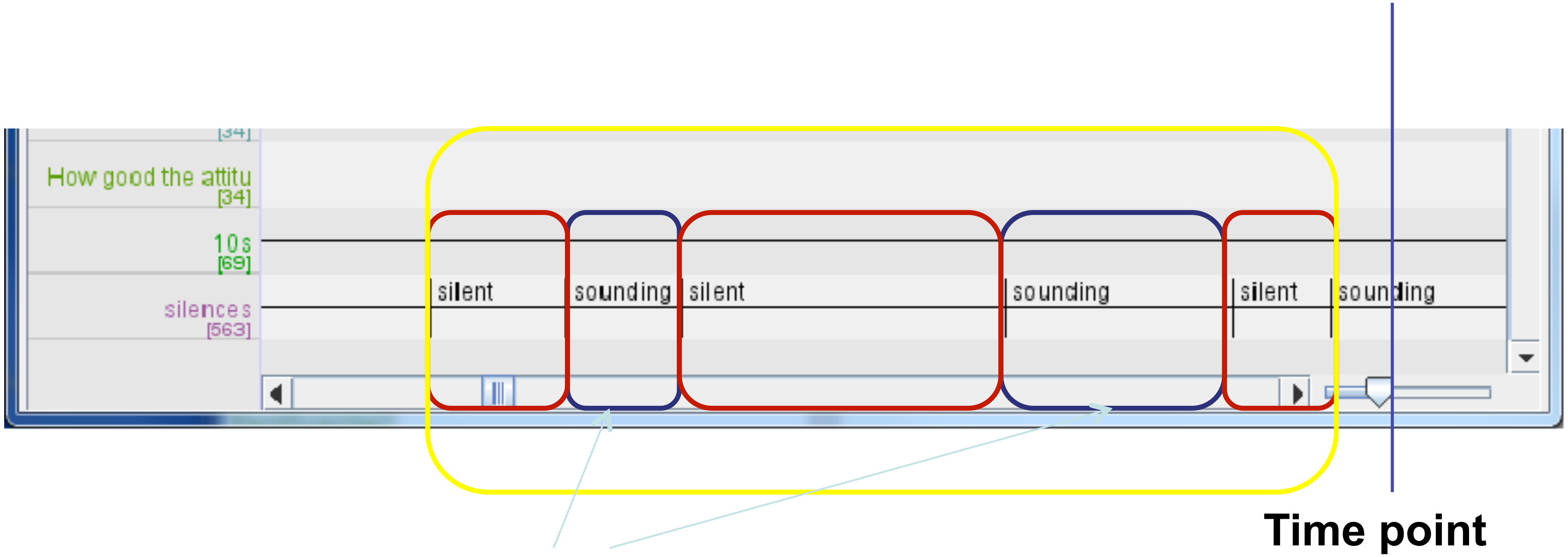
Smile: upper lip raising, lower lip raising,
lip corner raising, brow raising; C4.5 decision tree

	Precision	Recall	F Measure
smile	0.869	0.885	0.877
nothing	0.957	0.951	0.954
Overall	93.3%		

Nod: head position, head position difference,
head direction, head rotation difference; C4.5 decision tree

	Precision	Recall	F Measure
nod	0.853	0.842	0.847
nothing	0.933	0.921	0.930
Overall	90.2%		

Speaking frequency

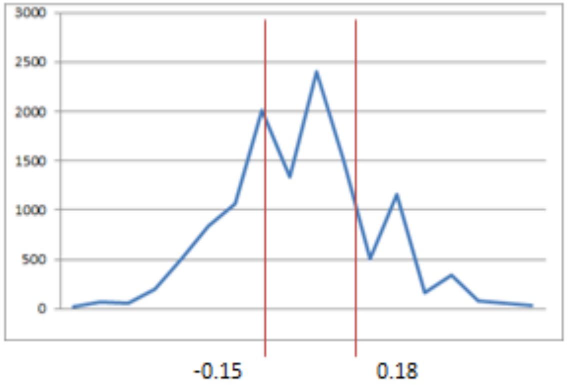


Last five segments

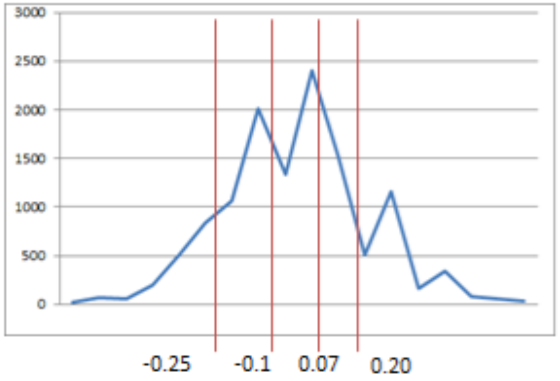
Time point of detection

* Software used: Praat

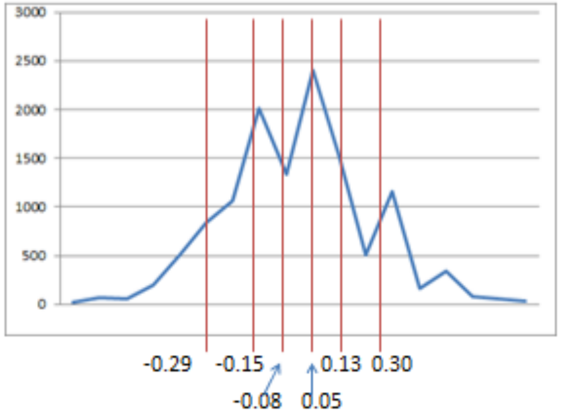
Classification targets



3 classes



5 classes



7 classes

Classification results (3 classes & 5 classes)

	Precision	Recall	F Measure
1/3	0.207	0.545	0.311
2/3	0.750	0.471	0.579
3/3	0.400	0.523	0.470
Overall	58.2%(33.3%)		

	Precision	Recall	F Measure
1/5	0.292	0.875	0.438
2/5	0.273	0.375	0.316
3/5	0.720	0.485	0.610
4/5	0.118	0.400	0.182
7/7	0.412	0.583	0.483
Overall	49.3%(20%)		

Classification results (7 classes)

	Precision	Recall	F Measure
1/7	0	0	0
2/7	0.400	0.421	0.410
3/7	0.091	0.118	0.103
4/7	0.471	0.400	0.432
5/7	0.333	0.250	0.286
6/7	0.375	0.286	0.324
7/7	0.100	0.167	0.125
Overall	29.4%(14.3%)		

- Conclusions
  - Evaluation method of participants' attitude (engagement) during active listening conversation
  - Automatic estimation method of above based on empirical results
  - Accuracy was moderate but showed the potential of this method
- Future works
  - Improvement of accuracy
    - Postures, acoustic and other non-verbal features
    - Verbal features
  - Development of the model of the listener agent's responses
  - Experiments with elderly subjects
  - Development of the fully working agent
  - Long-term evaluation of the agent

Discussion

- Other non-verbal cues?
- The way how we defined the automatic estimation targets?
- From inputs to outputs?
  - The rules?
  - The appearance of the avatar?
  - The communication style?

Thank you for your attention

contact: hhhuang@acm.org