

Social Multimodal Intelligence: From Human Understanding to Social Understanding and Decision Making

Japan Advanced Institute of Science and Technology

Division of Advanced Science and Technology,
AI and Intelligence Research Area

Shogo Okada

Agenda

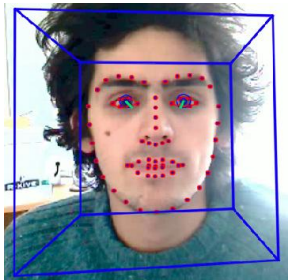
- What Is Multimodal Interaction (MI) and Social Signal Processing(SSP)?
- Research of MI and SSP
- Introduction to Applied Systems of MI and SSP
- Specific problems in multimodal ML for SSP ?

Multimodal Interaction

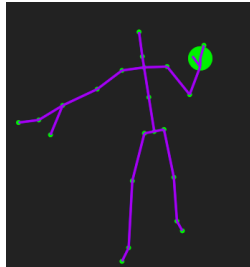
3

Vision

Examples: face, gestures



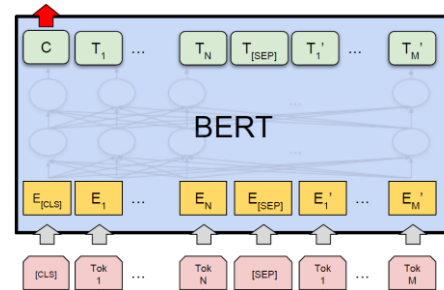
OpenFace



Kinect

Language

Example: Speech content

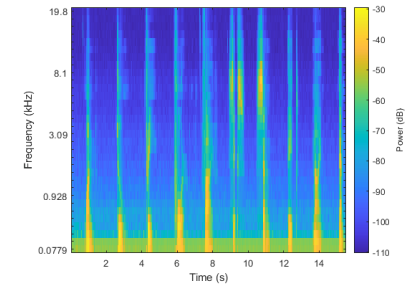


BERT



Speech

Examples: Volume, pitch



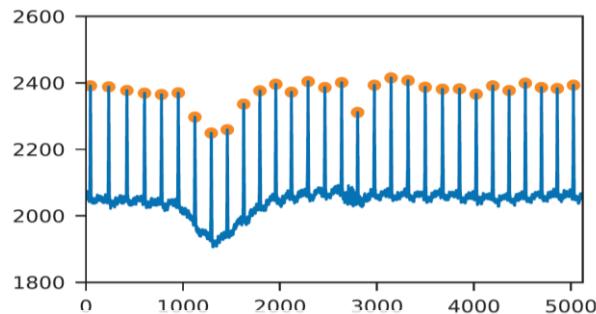
Mel Spectrogram

Potential difference (μV)

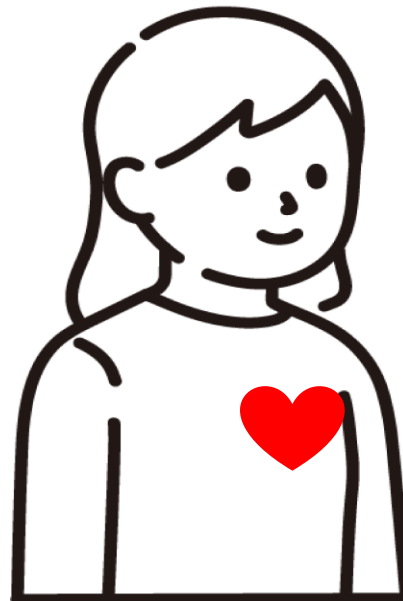
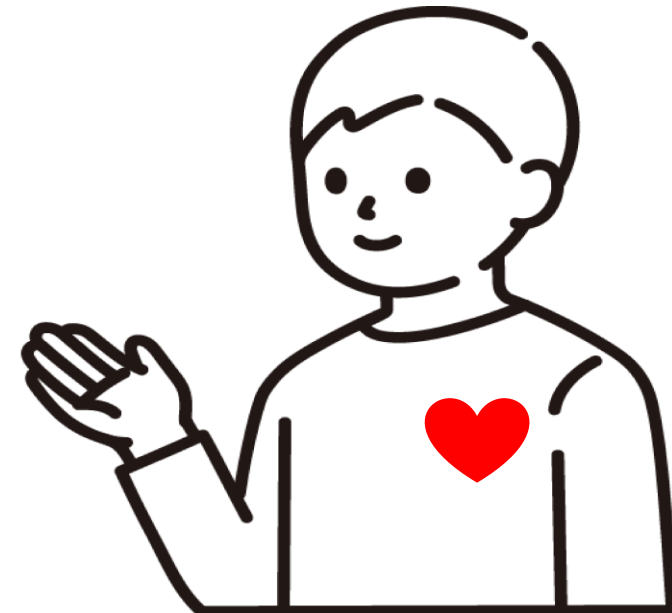


Biological Signals

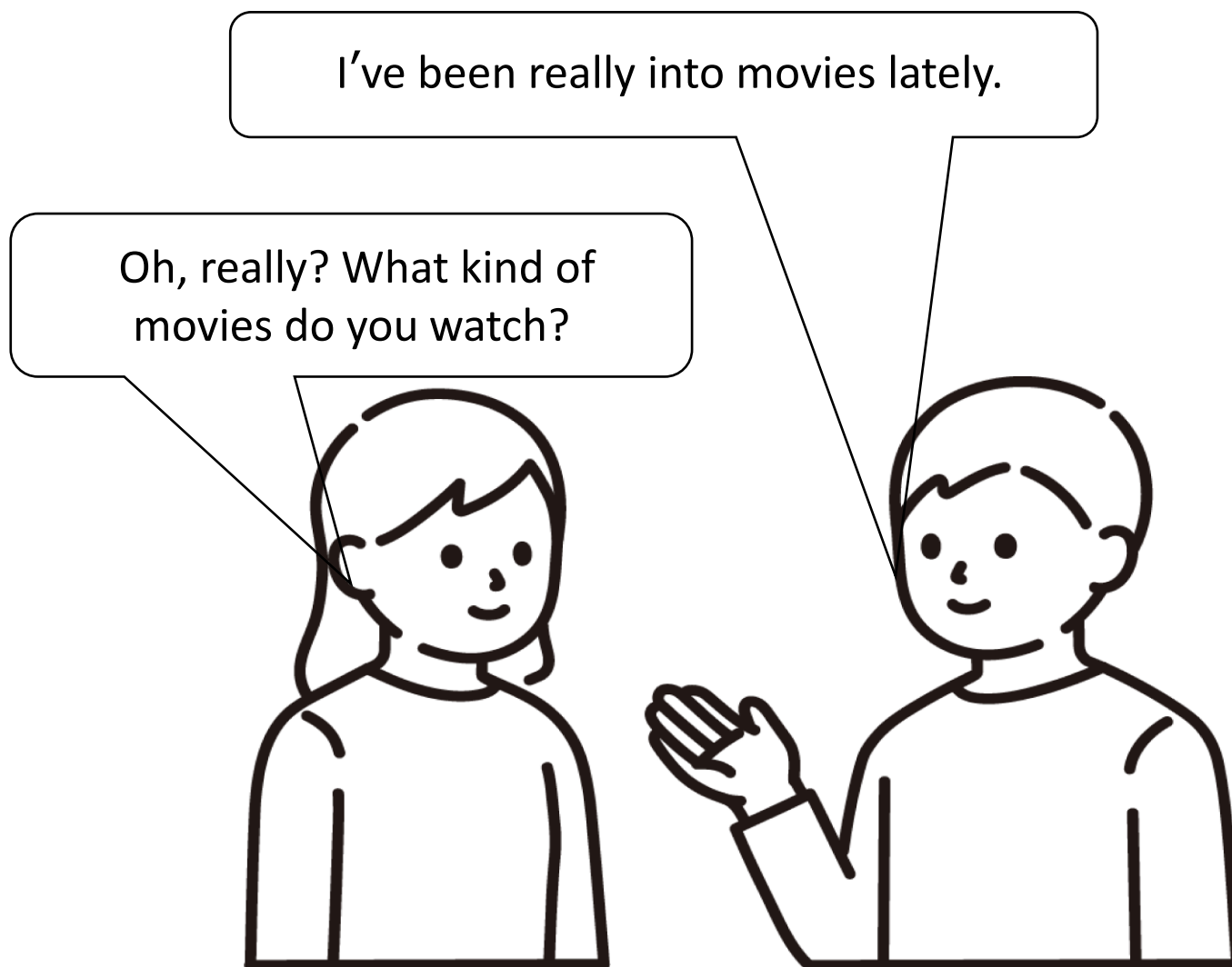
Examples: HR, EDA, EEG



Heart rate data



Inner states and personality traits that come through in conversation...



Emotions

(Mood, Sentiment)

Rapport, Engagement

Personality (BigFive, Skills)

Relationships

Cognitive State

What is Social Signal (SS)?

Important verbal and nonverbal information displaying intentions, inner-state attitudes, [Vinciarelli et al. 2009]



- Speech
 - Speech features: speech duration, number of utterances, turns
 - Prosodic features: Energy, pitch
- Visual
 - Facial expressions, gaze, posture, body movements, gestures
- Physiological indicators
 - Heart rate, EDA, Electromyography, EEG

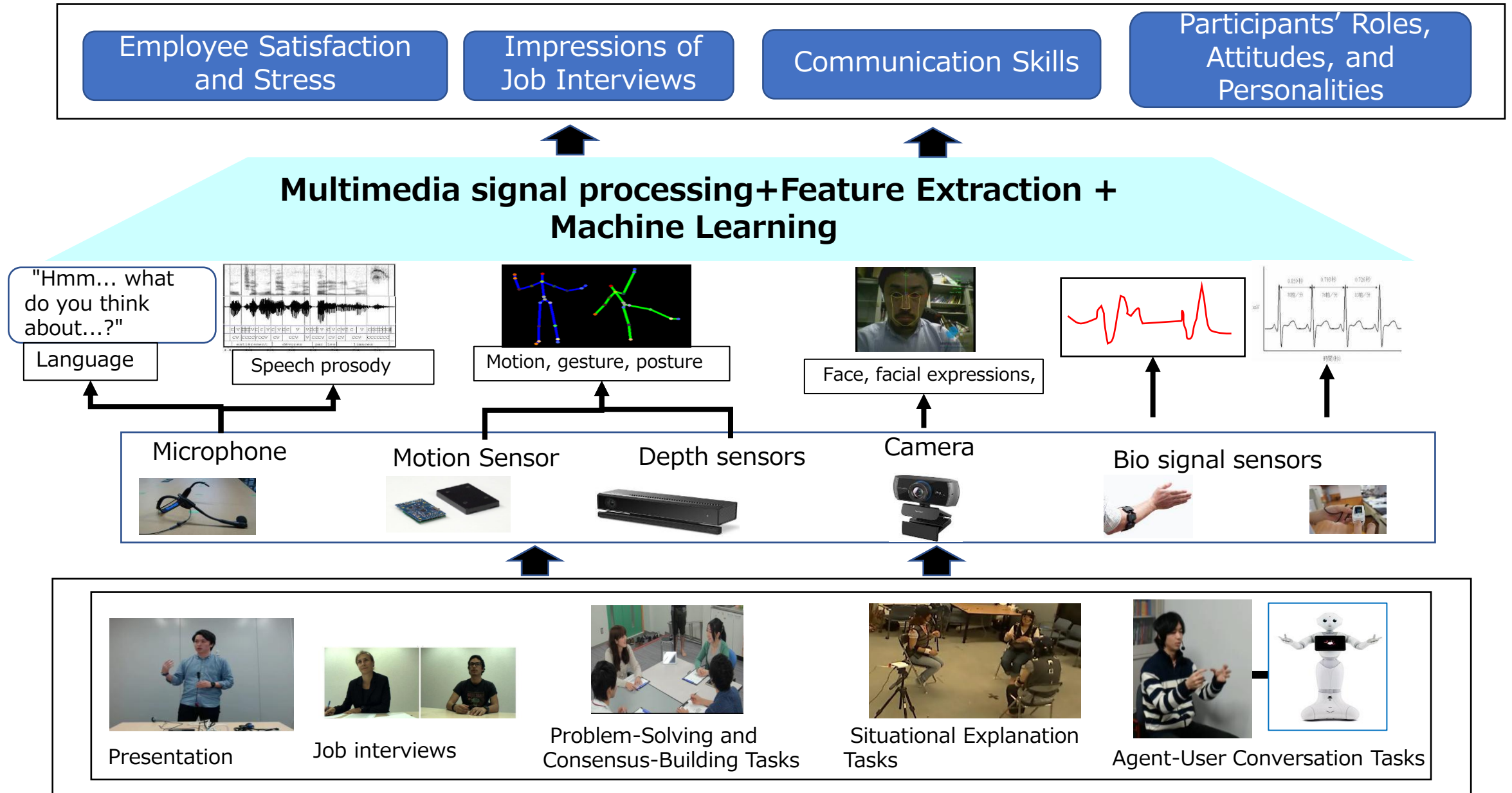


Individual: Agreement, Personality Traits, Leadership, Active Listening

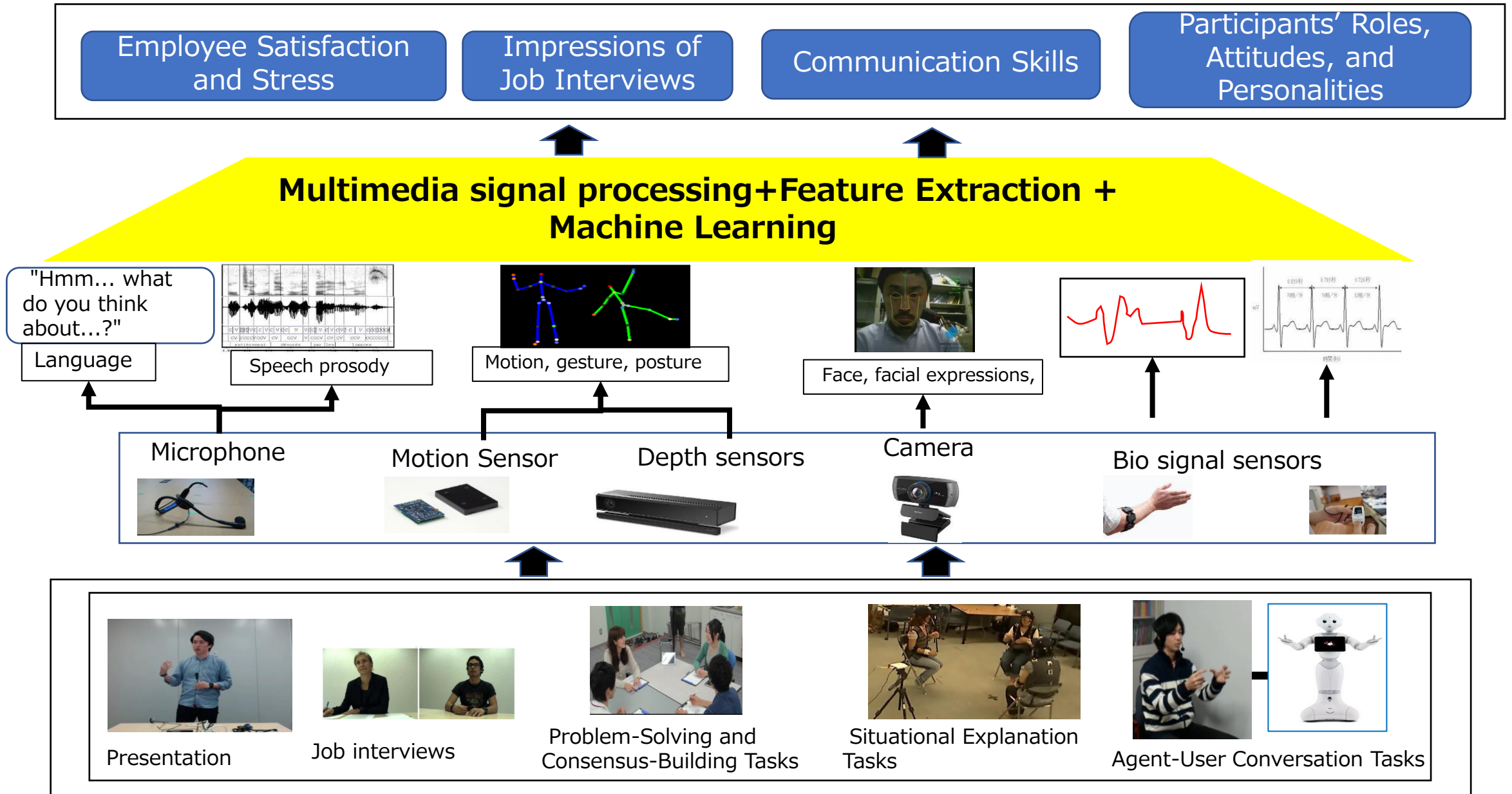
Group: Excitement, Stalemate, Conversation Dynamics, Group output

Understanding Communication Based on Social Signal Processing

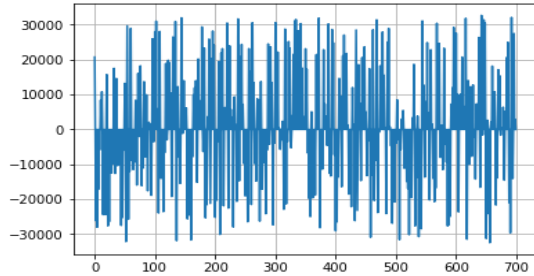
6



Understanding Communication Based on Social Signal Processing



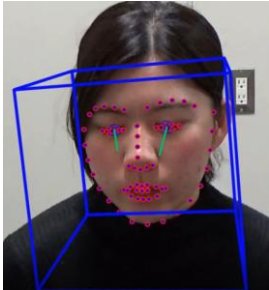
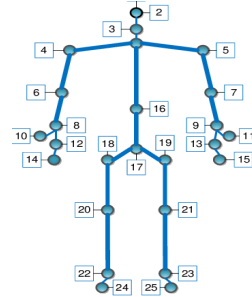
Extraction of multimodal features



OpenSMILE
FuBERT

• Auditory (Speech) Features

- Pitch, energy, speech rate



OpenFace
OpenPose
DeepFace

• Visual (Image) Features

- Face (facial expressions, gaze)

- Movement

• Linguistic Features

- Vocabulary of spoken words

P1: Do you like sweets?

P2: Sweets... well, not so much.



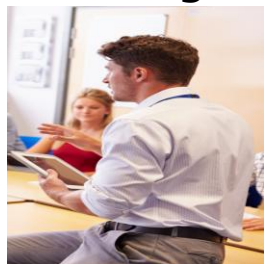
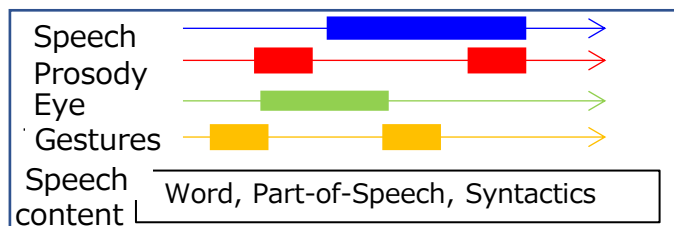
MeCab
Word2Vec
LLMs

S: System utterance

U: User utterance

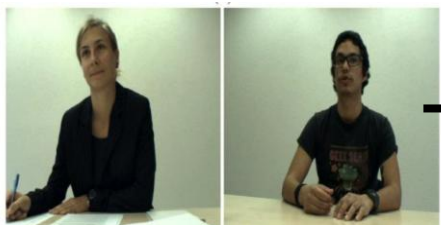
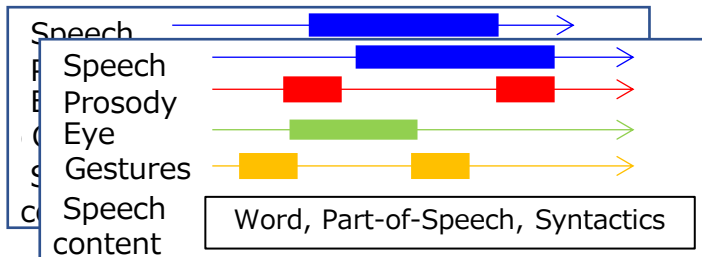
Number of Participants and Feature Design

Monologue



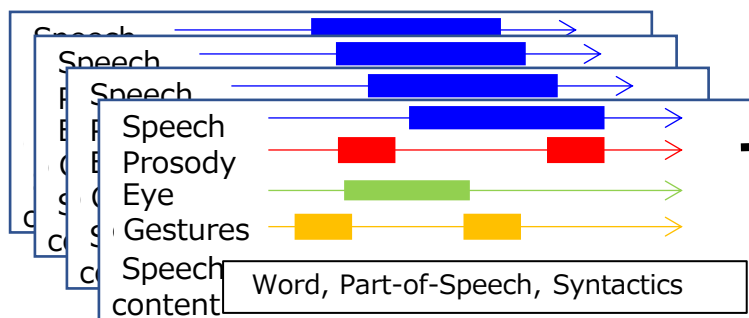
Individual Features

Dyadic dialogue



**Individual Features
Interaction features**

Group conversation



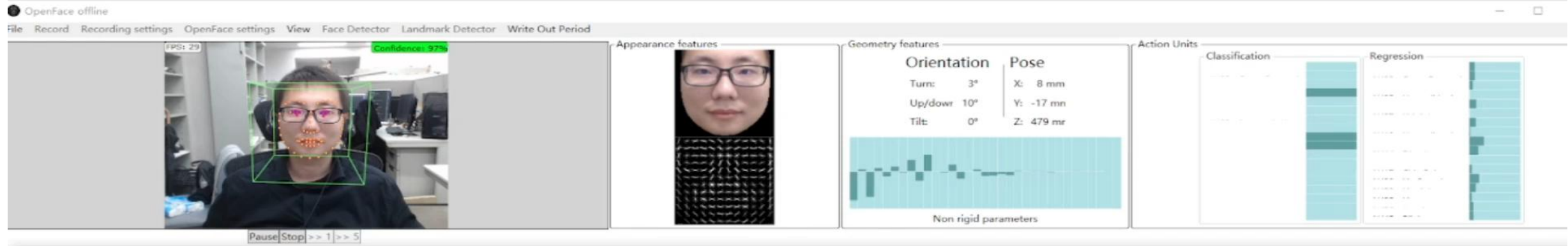
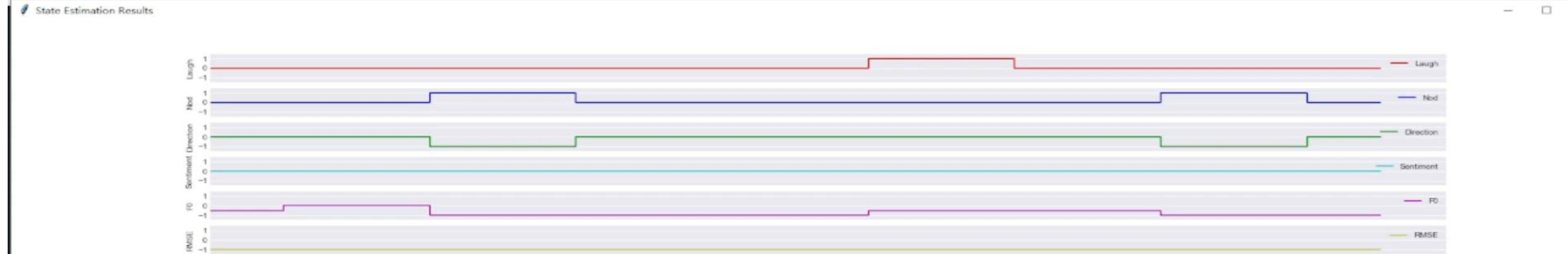
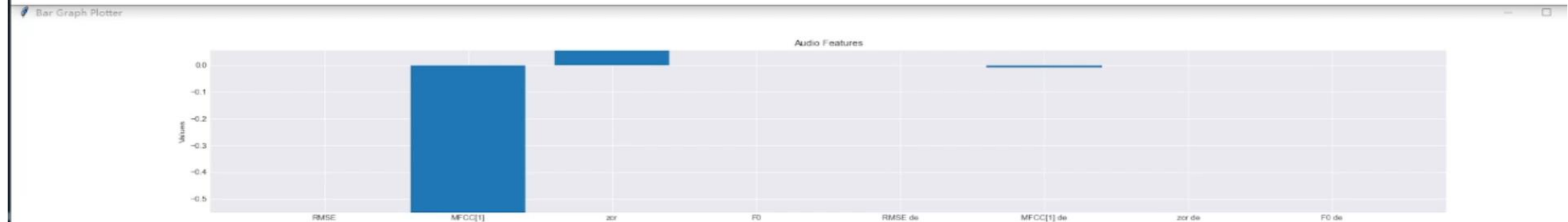
**Individual Features
Interaction Features
Group-level features**

classifier

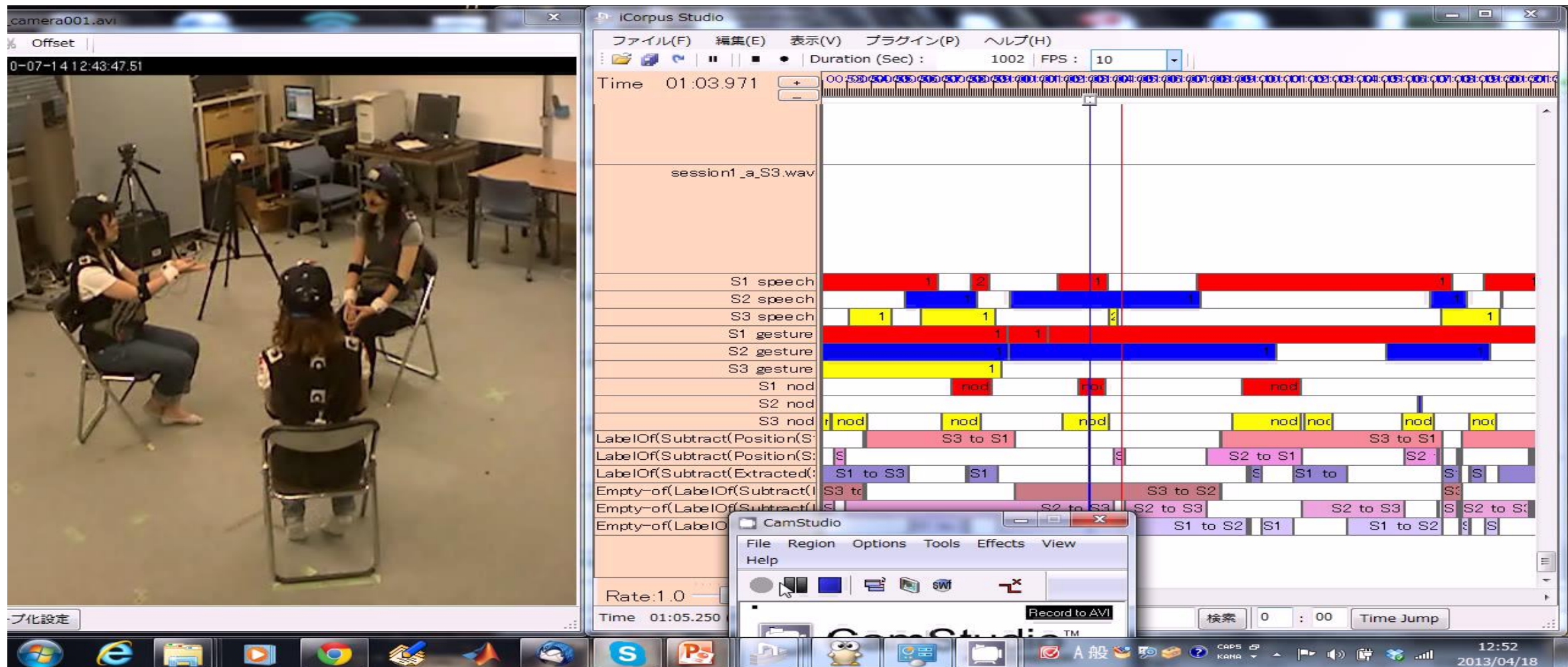
Feature design becomes more complex



Sensing Multimodal Nonverbal Information, Including Facial Expressions and Voice

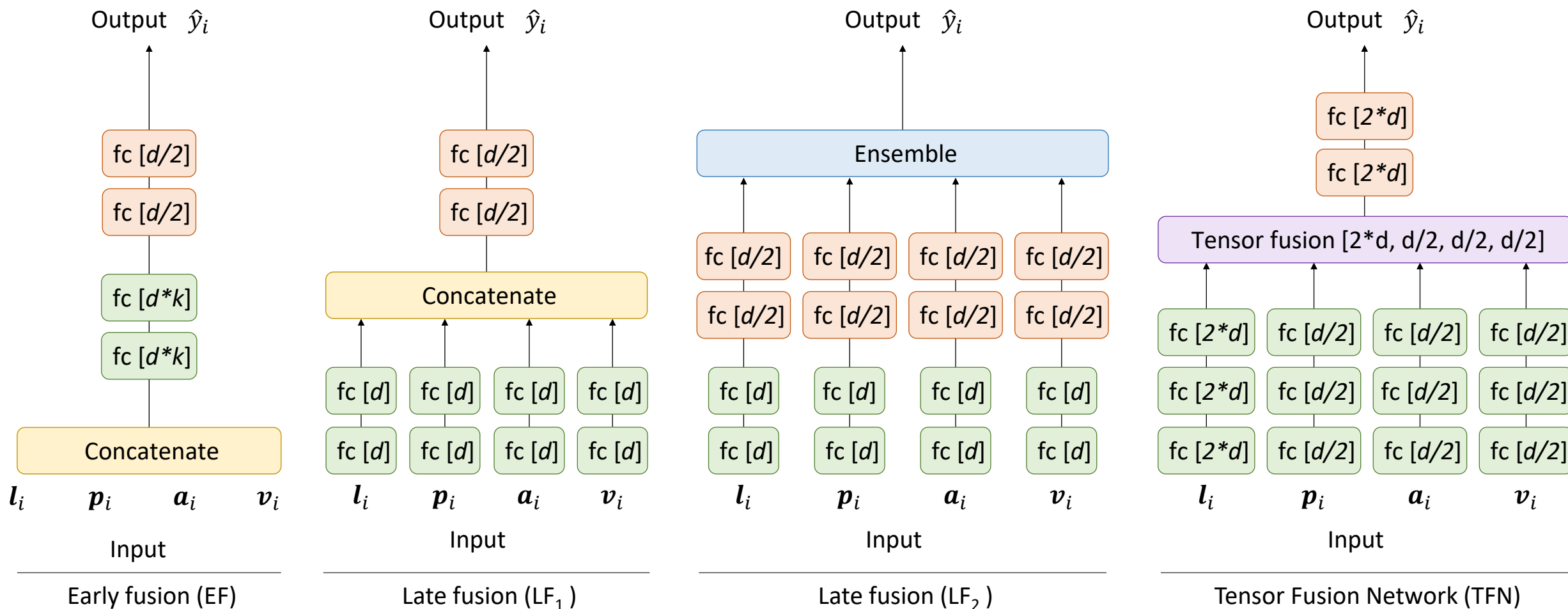


Group-Level Sensing of Multimodal Nonverbal Information



Who spoke and when? Who performed an action and when?

Who nodded and when? Who looked at whom and when?



Language : l_i

Physiological : p_i

Speech
(Acoustic) : a_i

Vision : v_i

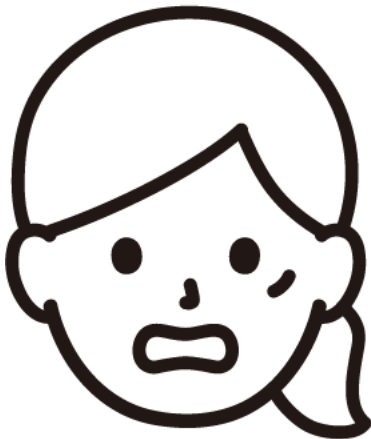
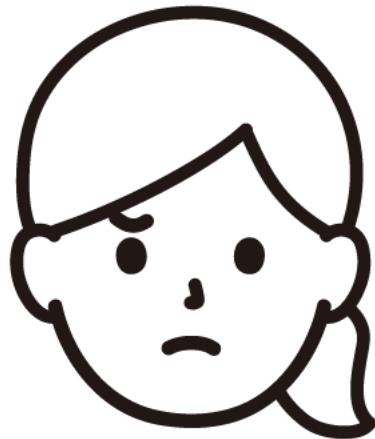
Agenda

- What Is Multimodal Interaction (MI) and Social Signal Processing(SSP)?
- **Research of MI and SSP**
- Introduction to Applied Systems of MI and SSP
- Specific problems in multimodal ML for SSP ?



Emotions and Sensibilities

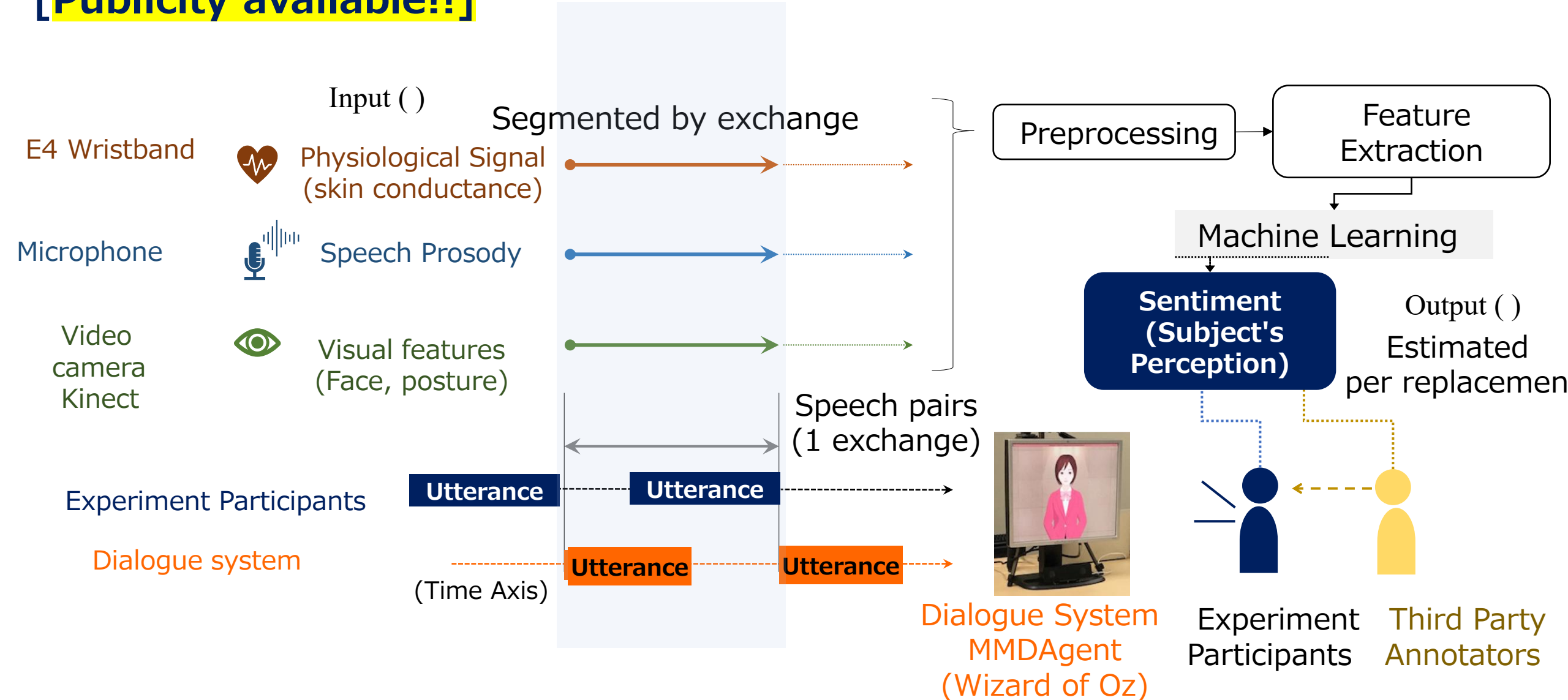
(Mood, Sentiment)



Task: Estimating the Speaker's Sentiment

Multimodal Dialogue Corpus: Hazumi1911 Dataset

[Publicity available!!]



Multimodal nonverbal feature extraction



• Biometric Signal Features

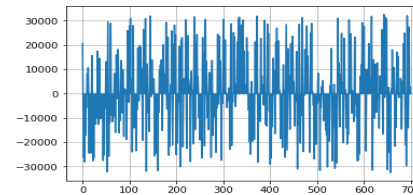
- Skin Conductance (GSR Count, Statistics)
- Heart Rate

Example of skin potential analysis during system interaction



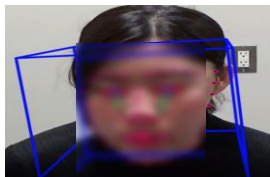
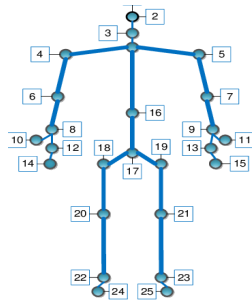
• Prosodic features

- Voice volume, pitch, etc.

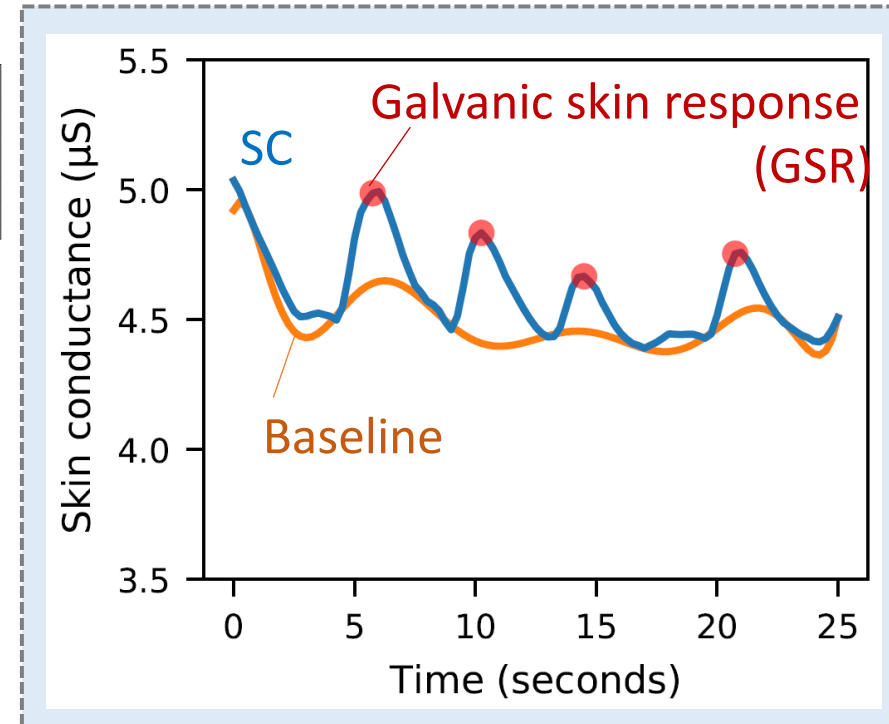


• Visual Features

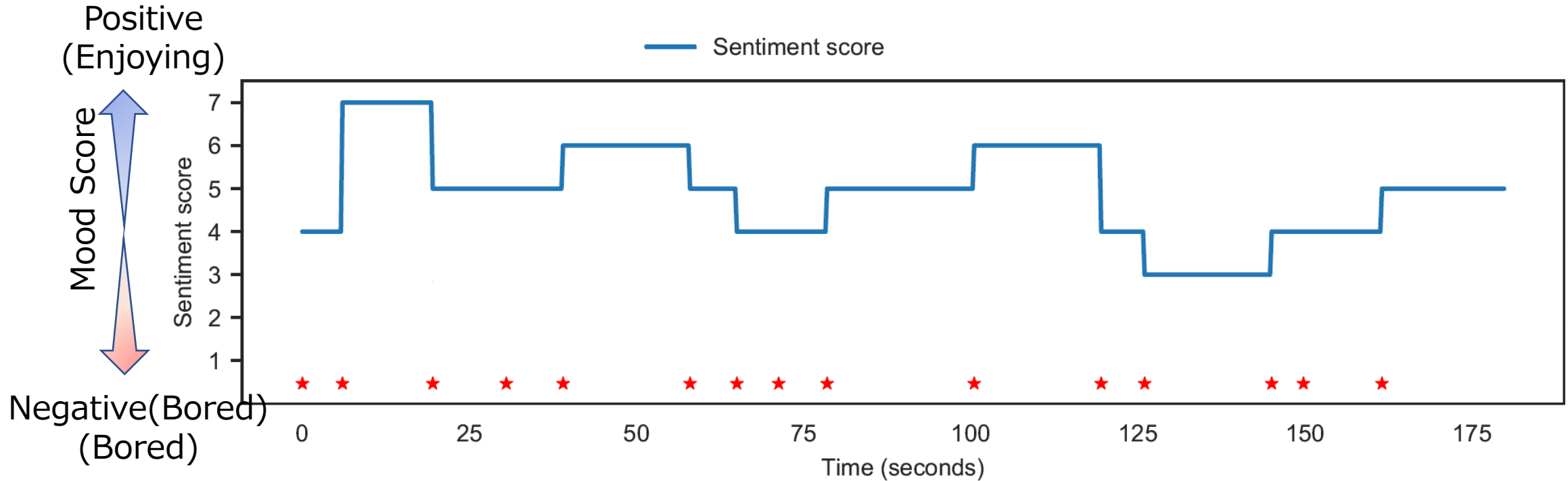
- OpenFace (facial expressions)
- Kinect (posture, gestures)



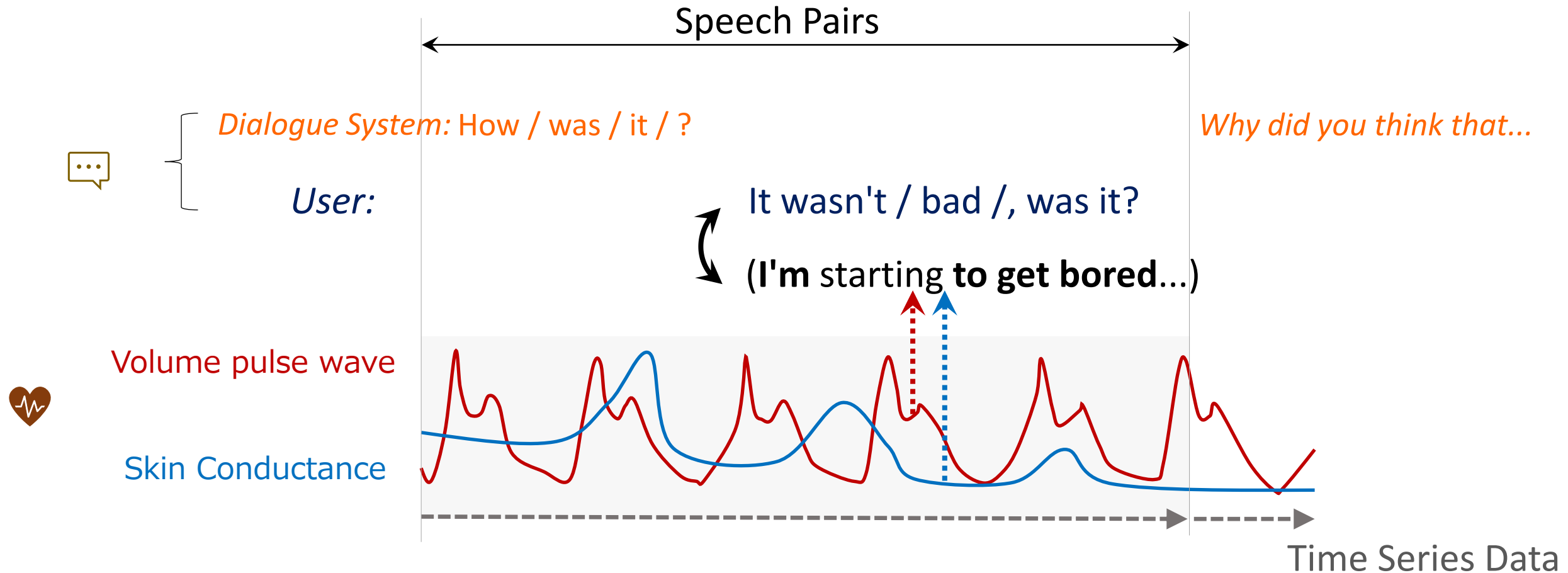
**Empatica
E4 Sensor**



Dynamic Changes in Experiment Participants' Perceptions (Enjoyment/Boredom) and GSR Readings

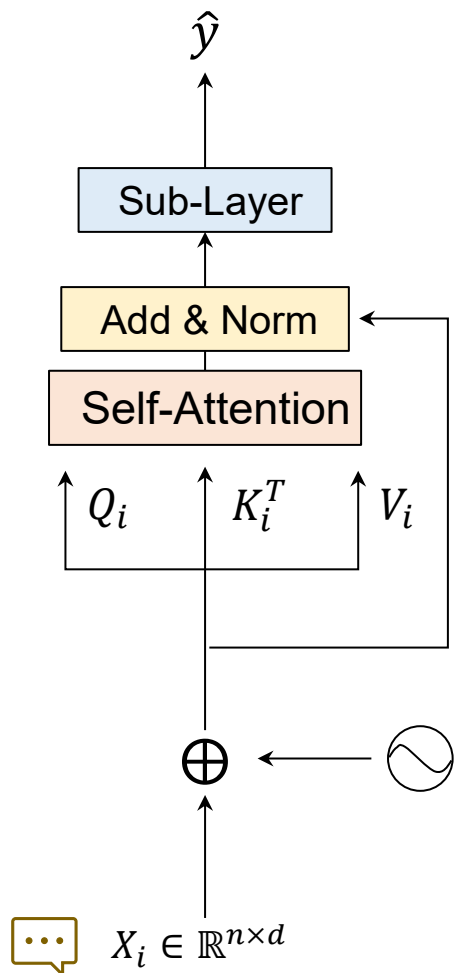


- It is necessary to capture such dynamic changes to build adaptive dialogue systems
- The co-occurrence of emotional scores and GSR readings may contribute to estimation performance



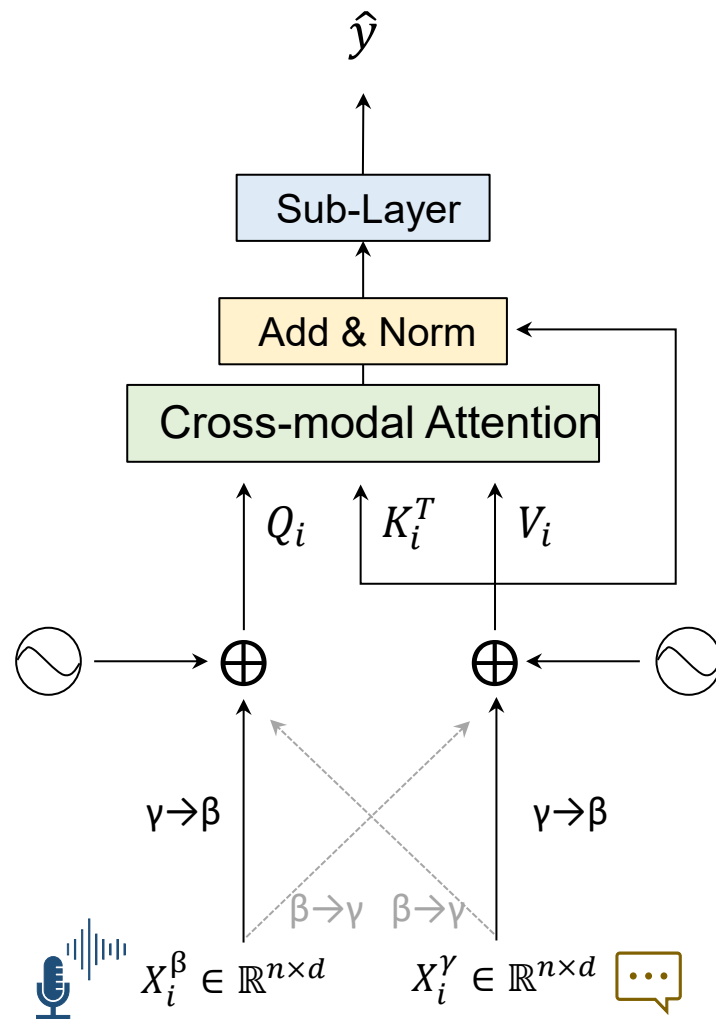
The potential for capturing changes in physiological time-series data that are not reflected in speech content

Time-Series Physiological Signals with Transformer



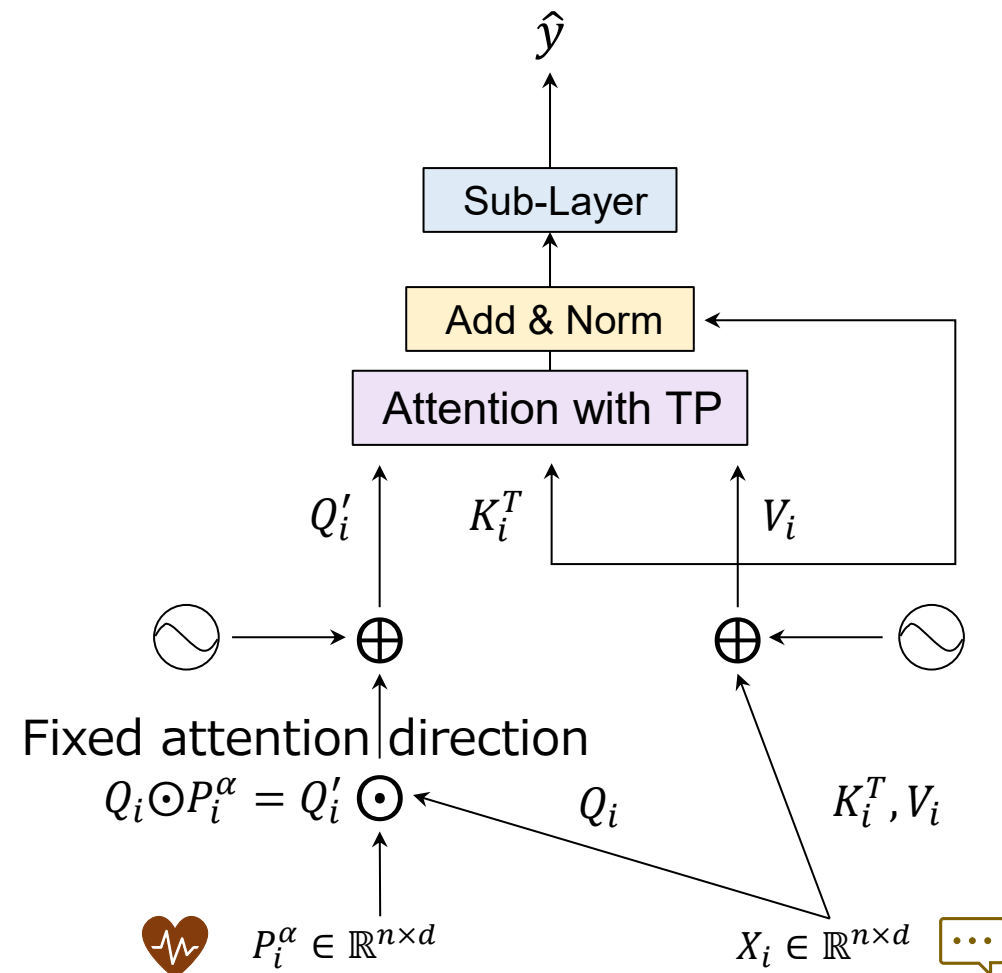
Transformer

(Vaswani et al., 2017, NIPS)



Crossmodal Transformer

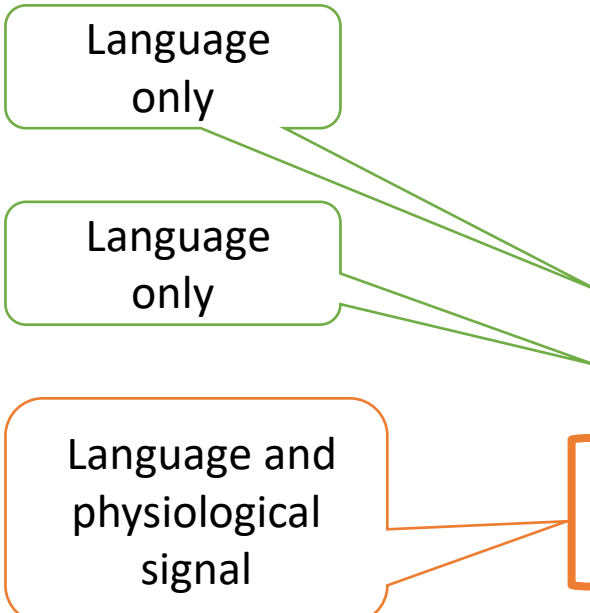
(Tsai et al., 2019, ACL)



TPTTr, Proposed Method
(Katada et al., 2022, ICMI)

Experimental Results

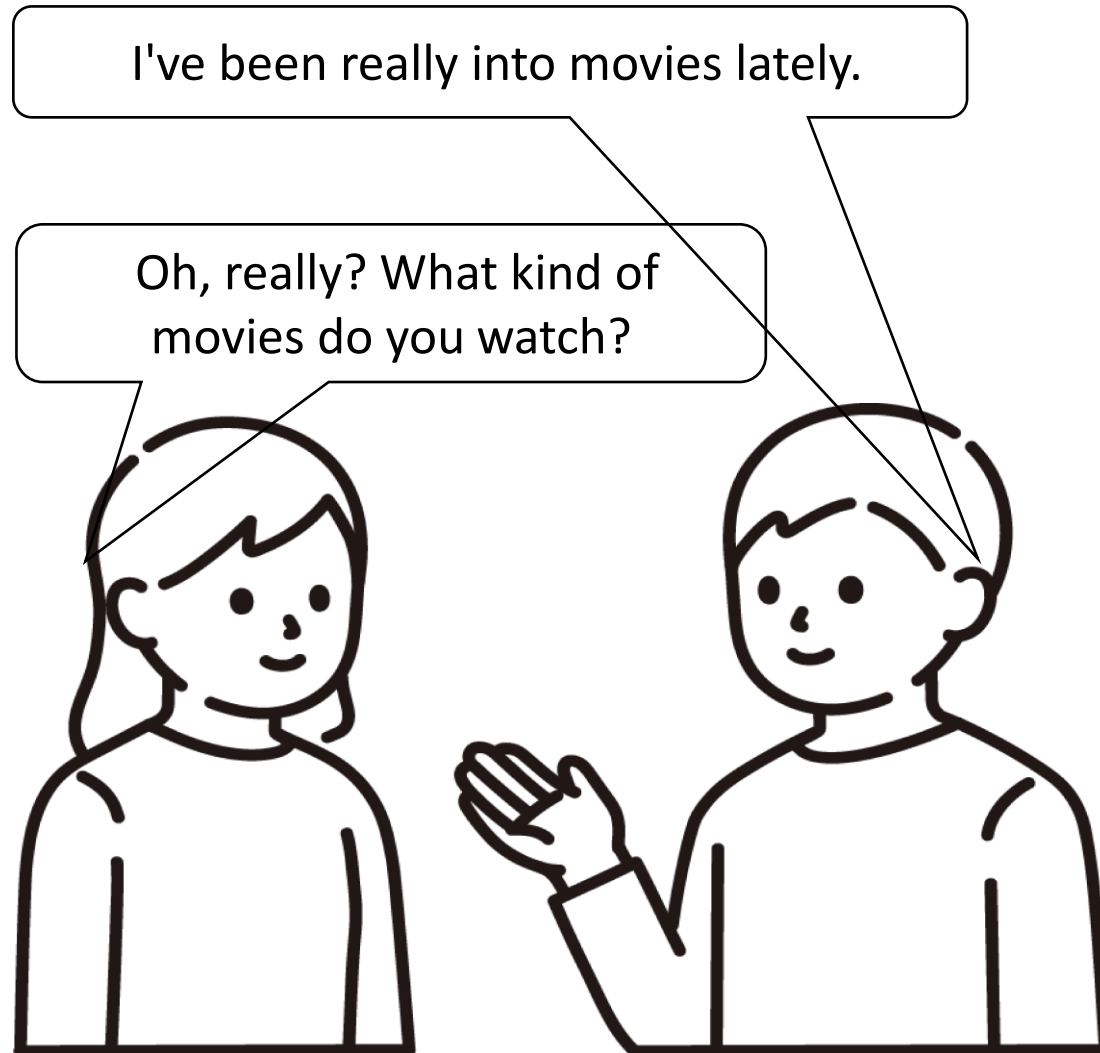
Self-Image Estimation Results by Each Model (Regression)



Model	Reference		
		MAE	Corr
Tr	NIPS, 2017	1.082	0.227
Tr×3		1.109	0.219
CMTr	ACL, 2019	1.083	0.190
TPTr	Proposed	1.114	0.228
TPTr×3	Proposed	1.068	0.232
Human		1.008	0.406

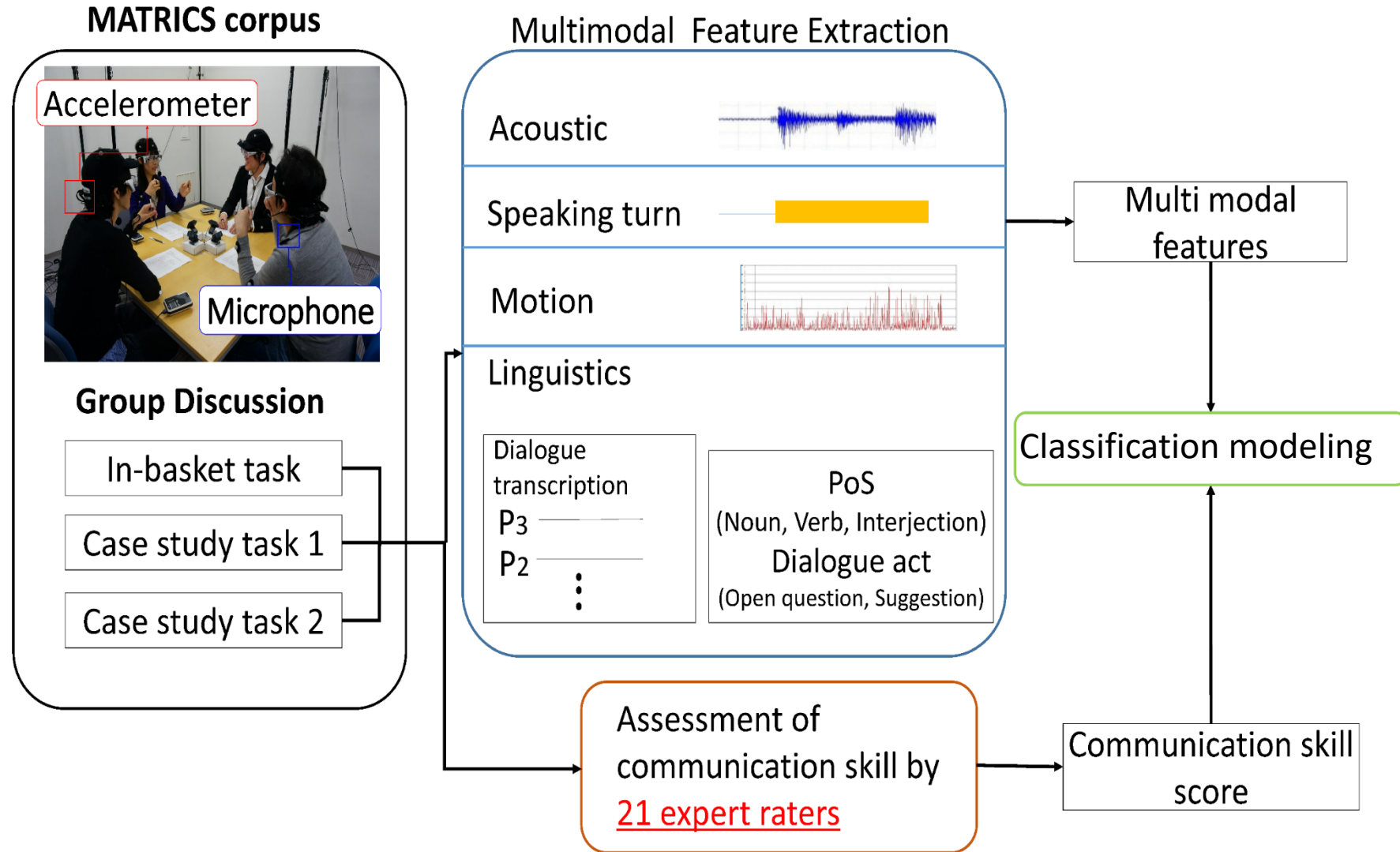
×3...Three Transformer blocks (networks) processed in parallel

Best performance among models using the proposed method



Communication Skills

Overview of modeling communication skills



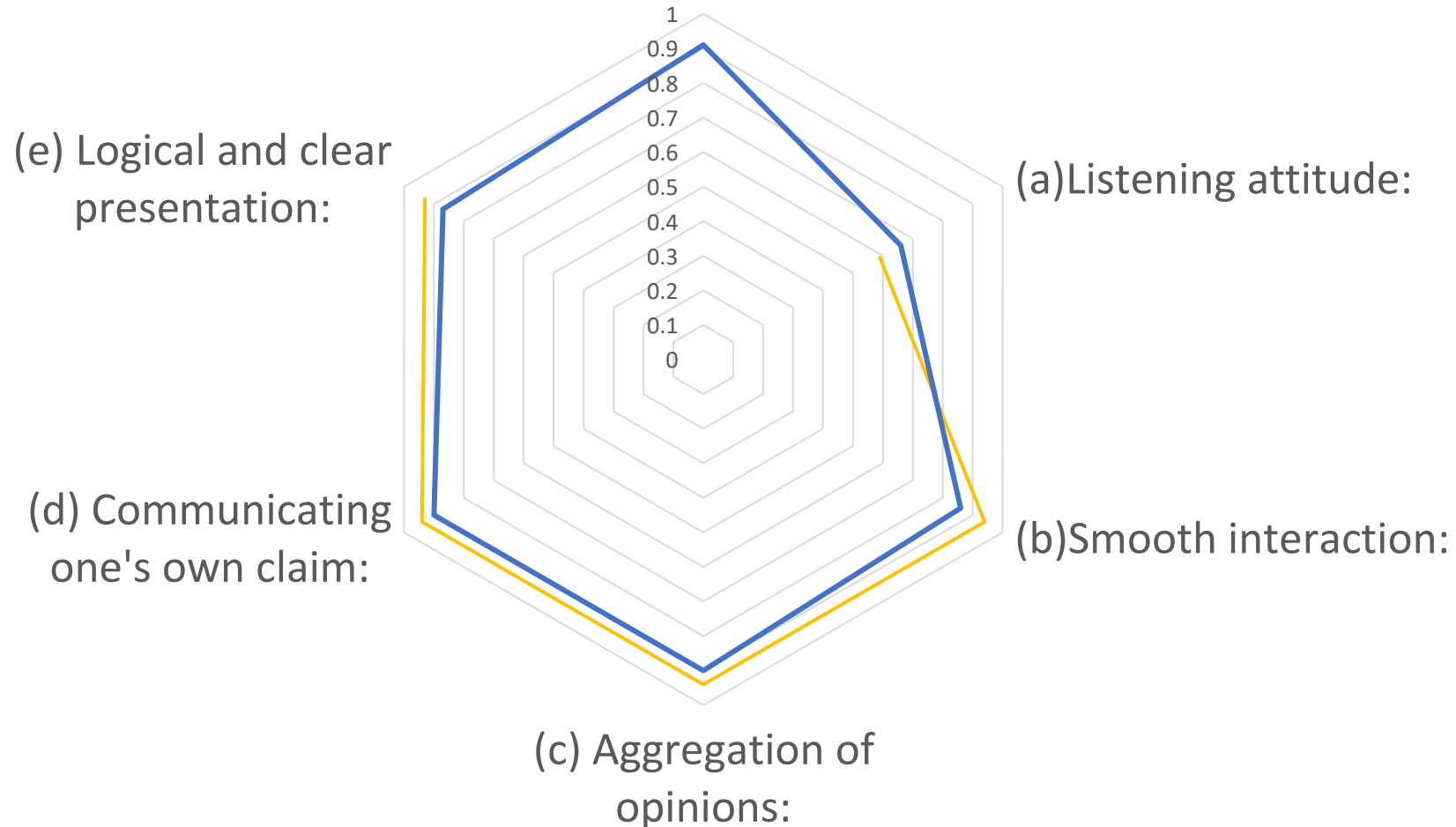
Dimensions of communication skills

Dimension	Description	Score
(a) Listening attitude:	Could participants listen to the opinions of the other members in a polite manner ?	[1-5]
(b) Smooth interaction:	Could they accept various opinions put forward by others?	
(c) Aggregation of opinions:	Could they organize and summarize opinions from others ?	
(d) Communicating own claim:	Could they communicate appropriate info. to the others ?	
(e) Logical and clear presentation:	Could they present opinions in a logical and clear manner.	
(f) Total communication skill:	The skill is defined as the total of the above five types of skill.	[1 – 10]

21 expert coders were assessing the skill scores based on the dimensions

Agreements of assessment

(f) Total communication skill



21 expert coders were assessing the skill scores based on the dimensions

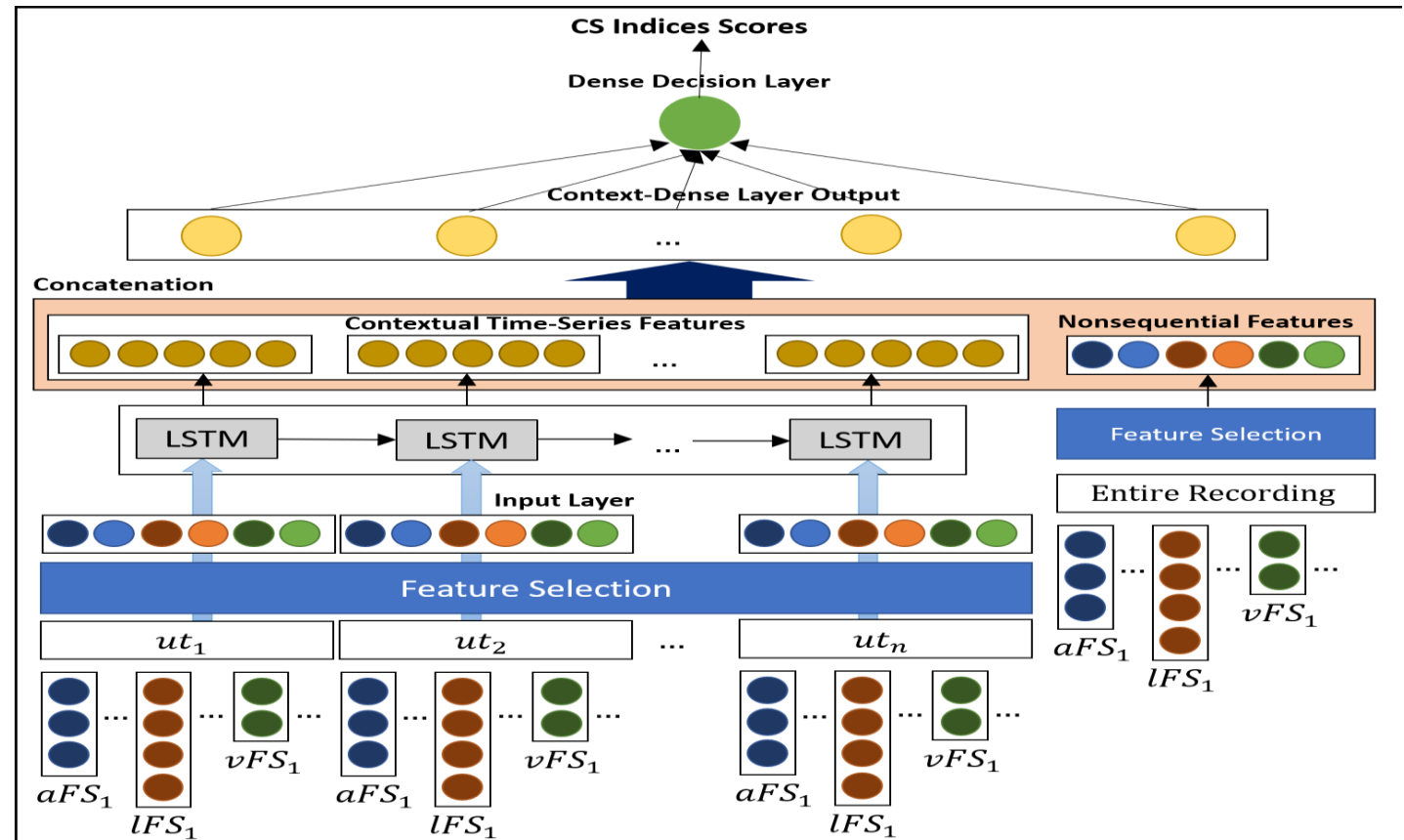
Automatic skill assessment results

1. Classification by Linear SVM

Feature set	Accuracy
(1) A : Prosody	0.68
(2) S : Turn	0.82
(3) L : Language	0.85
(4) M : Motion	0.64
(5) $A + S + L$:	0.81
(6) $A + S + M$:	0.82
(7) $A + L + M$:	0.93
(8) $S + L + M$:	0.81
(9) All :	0.90

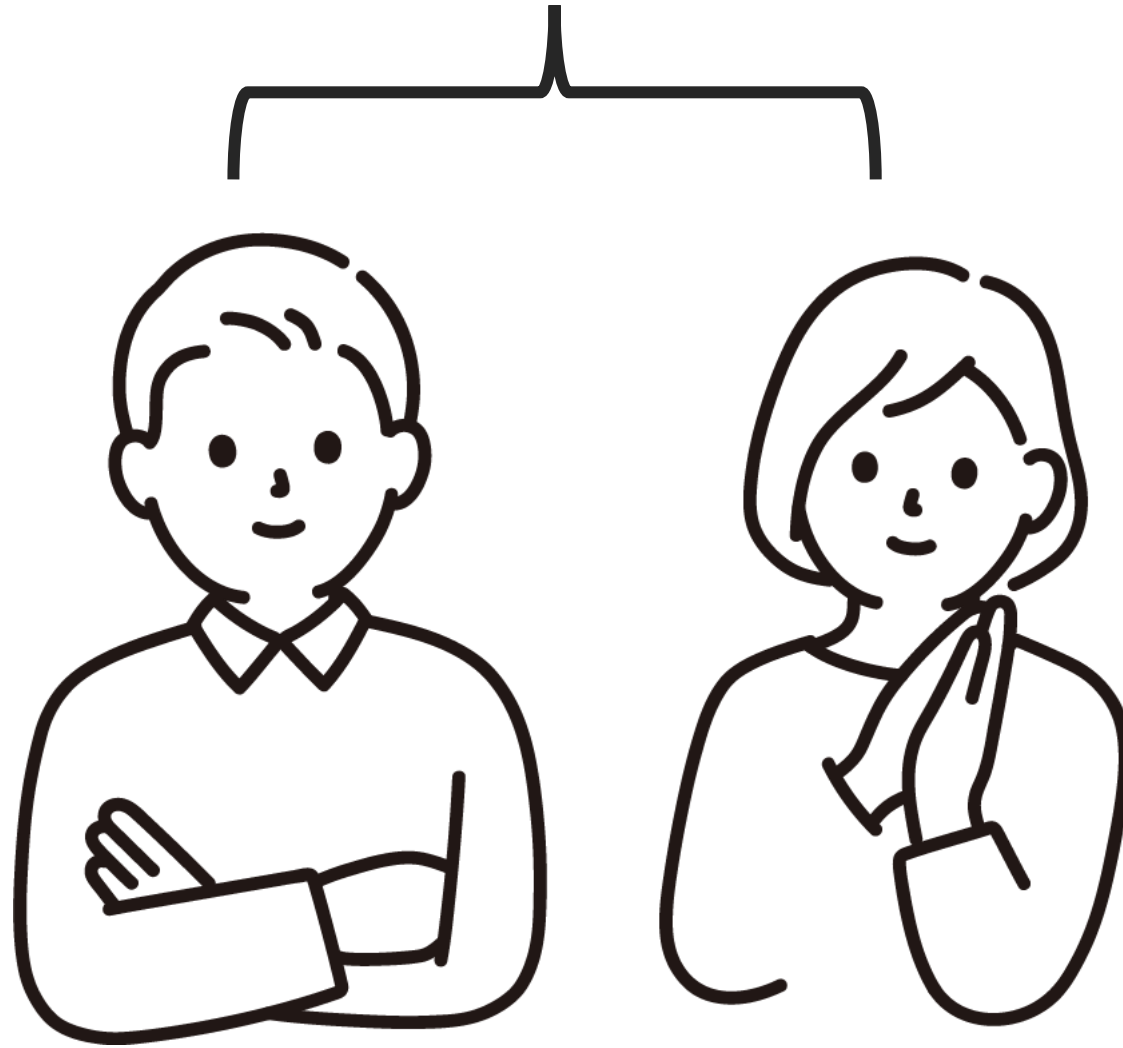
Random baseline is 63% ($p < 0.01$)

2. Regression by LSTM based Fusion model

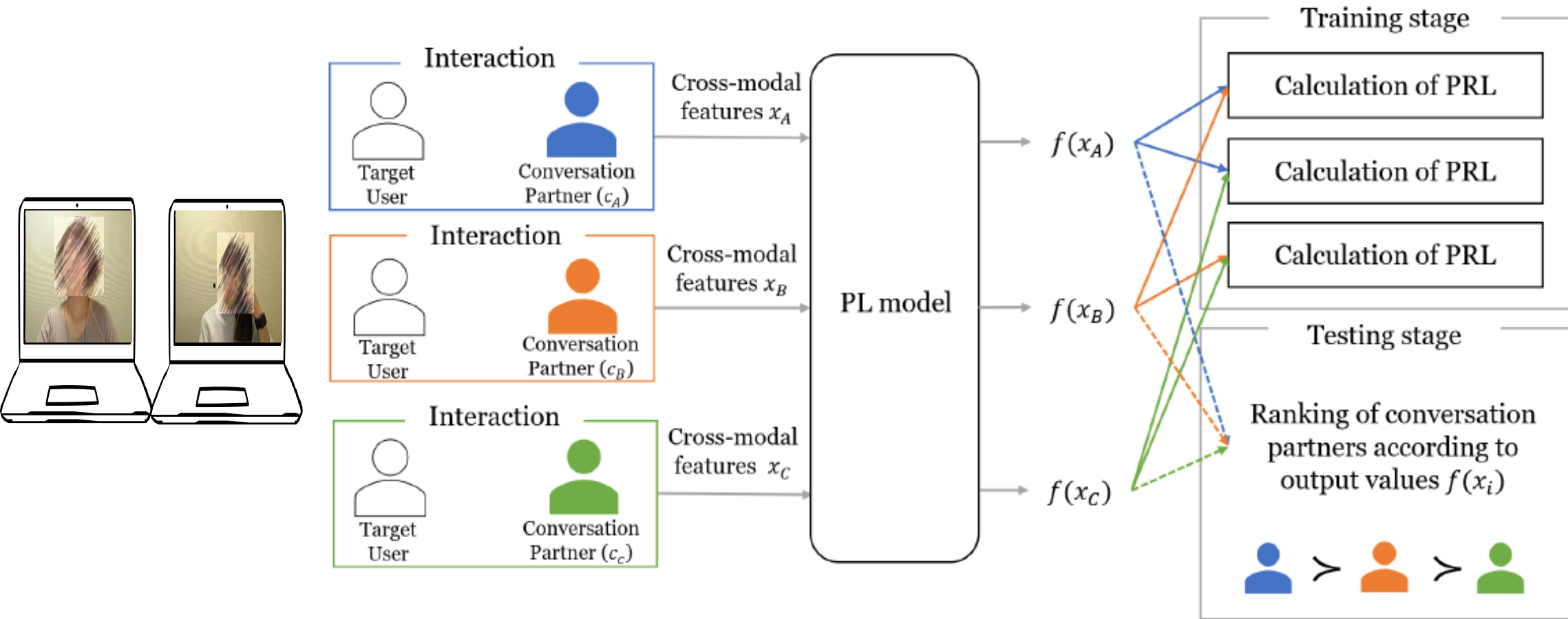


Task Depend. $R^2 = 0.8$, Independent $R^2 = 0.6$

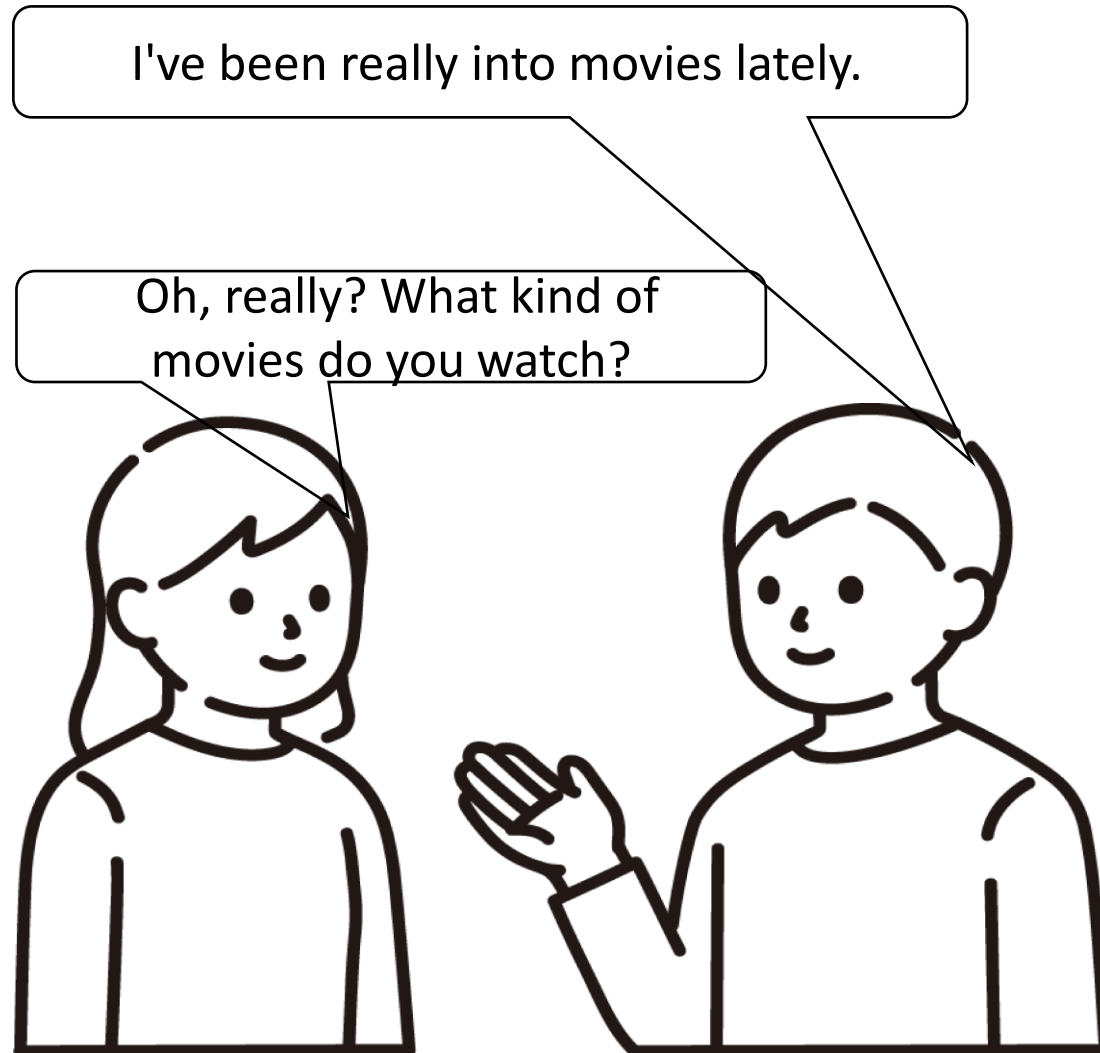
The Relationship Between the Two Parties



Estimating the Degree of Subjective Rapport Based on Rank Learning



Takato Hayashi, Candy Olivia Mawalim, Ryo Ishii, Akira Morikawa, Atsushi Fukayama, Takao Nakamura, Shogo Okada. "A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels," IEEE Access (2023).



Personality Traits

Predicting personality traits based on behavior

Prediction of personality traits



Target
Nonverbal
Features

+

Group
Nonverbal
Features

Co-occurring Features

Target Audience			
Target Body	ON		
Target head			
Target gaze	sp		
Group speech	ON		
Group body		ON	
Group head		ON	
Group gaze		S1	

Personality Traits

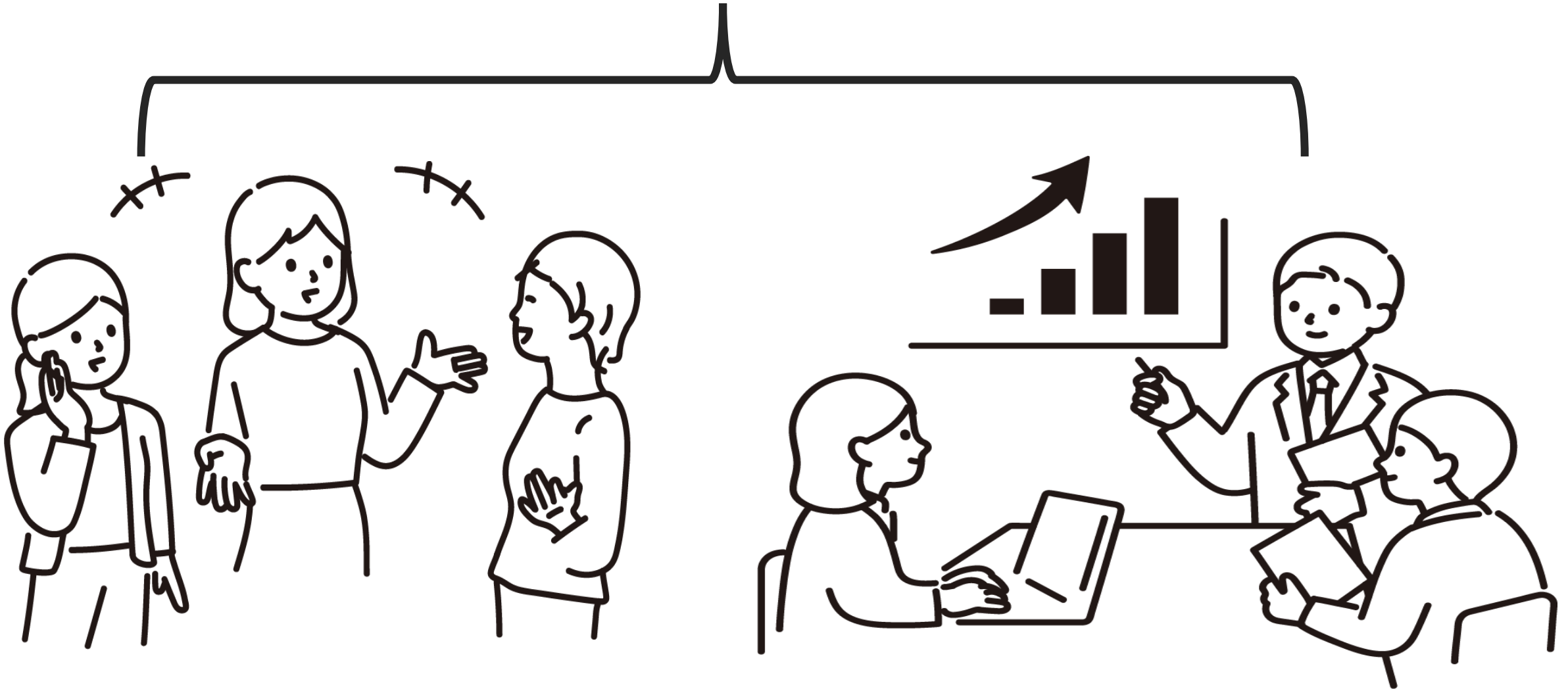
- Big Five
- Leadership
- Competence
- Dominance

- Binary classification with up to 75% accuracy based solely on nonverbal information
- Mining features as combinations of individual and group behaviors
⇒ 17% improvement over conventional methods

[Okada2019]

Okada, S., Nguyen, L. S., Aran, O., and Gatica-Perez, D. (2019). "Modeling Dyadic and Group Impressions with Inter-Modal and Inter-Person Features." ACM Transactions on Multimedia Computing, Communications, and Applications, **9**(4): Article 39.

The Quality of Group Discussions



Predicting the Quality of Group Discussions



Evaluation by
four raters

- Novelty of the idea
- Consistency between objectives and conclusions
- Clarity of content

Group output
quality score

Accelerometer

Microphone

A: Prosody
S: Speech turns
L: Language
M: Head movement

Machine learning
model

Predictive Accuracy of Group Discussion Quality (Regression)

Dimension index	A	S	L	M	AS	AL	AM	SL	SM	LM	ASL	ASM	ALM	SLM	All
(I) Action Orientation	0.31	-0.16	0.29	-0.11	0.44	0.35	0.40	0.31	0.09	0.38	0.44	0.36	0.40	0.39	0.43
(III) Originality	<u>0.49</u>	0.38	<u>0.65</u>	-0.74	<u>0.65</u>	<u>0.72</u>	<u>0.49</u>	<u>0.62</u>	0.38	<u>0.65</u>	0.76	<u>0.65</u>	<u>0.72</u>	<u>0.62</u>	0.76
(IV) Optimism	-0.06	-0.78	0.44	0.25	-0.18	0.25	-0.04	0.41	0.08	0.51	0.24	-0.11	0.29	0.40	0.30
(V) Quality of Presentation	0.19	-0.56	0.01	-0.25	0.12	0.12	0.24	-0.03	-0.39	-0.03	0.10	0.16	0.13	-0.09	0.10
(VI) Issue Involvement	0.25	0.37	0.37	-0.73	0.37	<u>0.52</u>	0.25	0.39	0.37	0.37	0.54	0.37	<u>0.52</u>	0.39	0.54

A: Prosody

S: Speech Turn

L: Language

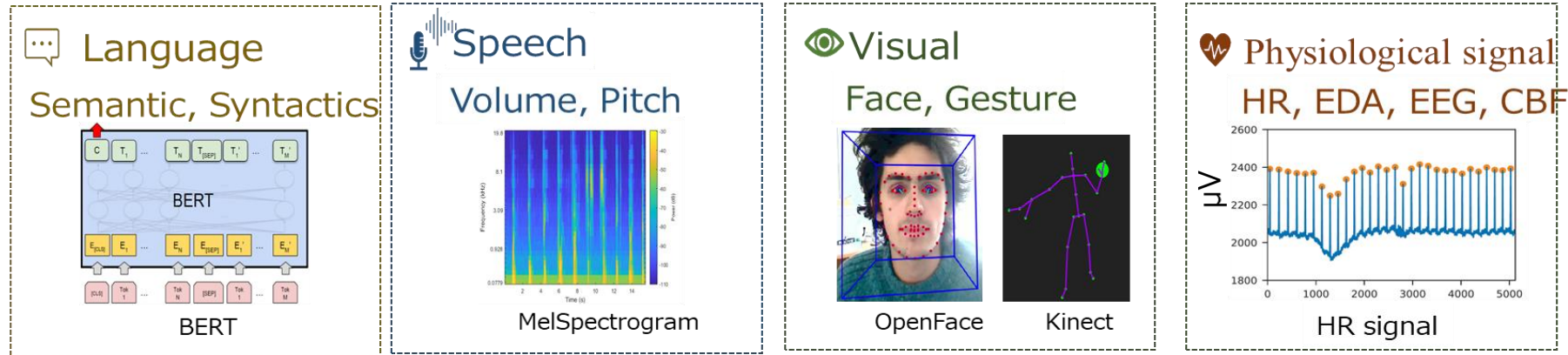
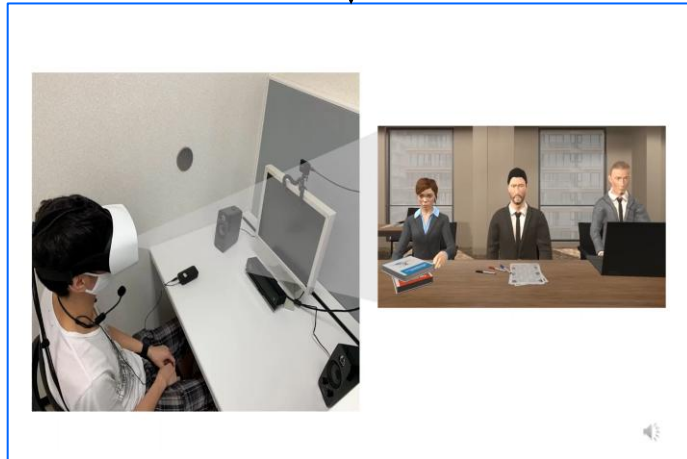
M: Head movements

- **The correlation coefficients** between the estimates from the language (L) feature model and the correct answers ranged from **0.01 to 0.65**
 - Adding prosody (A), speech turns (S), and head (M) improved the correlation to **0.24–0.76**
- ⇒ When predicting quality, “how it was spoken” is just as important as “what was spoken”

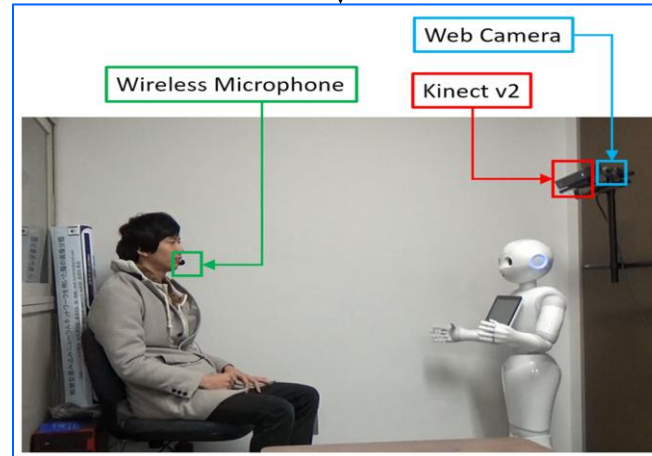
Agenda

- What Is Multimodal Interaction (MI) and Social Signal Processing(SSP)?
- Research of MI and SSP
- **Introduction to Applied Systems of MI and SSP**
- Specific problems in multimodal ML for SSP ?

Applications of MM fusion and Social Signal Processing



Presentation and Interview Training System



MM × Conversational/Physical AI

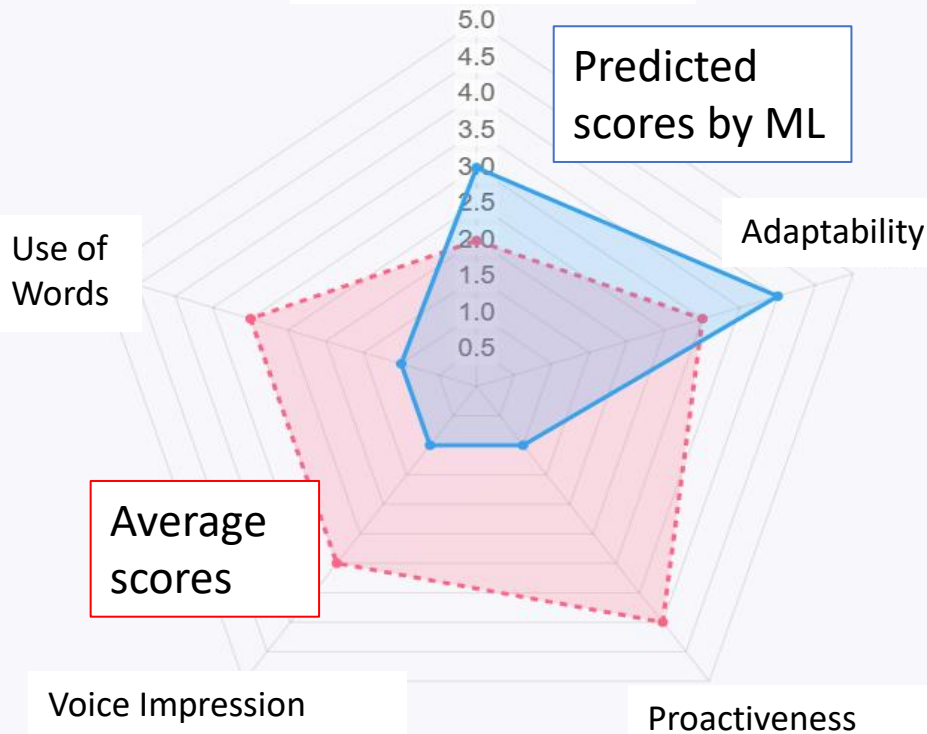


Analysis of Dementia Tendencies

Interview Training System

37

Overall Evaluation



Provides results predicting skills such as vocabulary, logical reasoning, and delivery

What Is the PREP Method?

The PREP method is a structured communication technique commonly taught in job interview training. By using the PREP method, you can communicate what you want to say to the interviewer in a clear and easy-to-understand manner.

P1 = Point (Conclusion)

This is the most important part of your response. By stating your main point first, you make it easier for the listener to understand what follows.

R = Reason

Give a concise reason that explains and supports your conclusion.

E = Example (Specific Example)

Provide numbers, data, facts, or personal experiences that support your reason.

P2 = Point (Concluding Point)

Summarize your response by restating the main point from the beginning.

<Example> What is your strength?

Point: My strength is perseverance.

Reason: This is because I can continue working on things that I find difficult or challenging until I achieve results.

Example: For example, I struggled with English when I was in high school. However, after entering university, I studied English for three hours every day and eventually achieved a TOEIC score of 800.

Concluding Point: Through this experience, I developed the ability to work persistently and make steady efforts even when facing difficulties.

Based on the structuring of interview responses by an LLM, the system suggests improvement information that should be added

Autonomous Dialogue Generation Based on Phase Boundary Detection Within Continuous Motion for Domestic Robots

Sixia Li, Tamon Miyake, Tetsuya Ogata, Shigeki Sugano, Shogo Okada

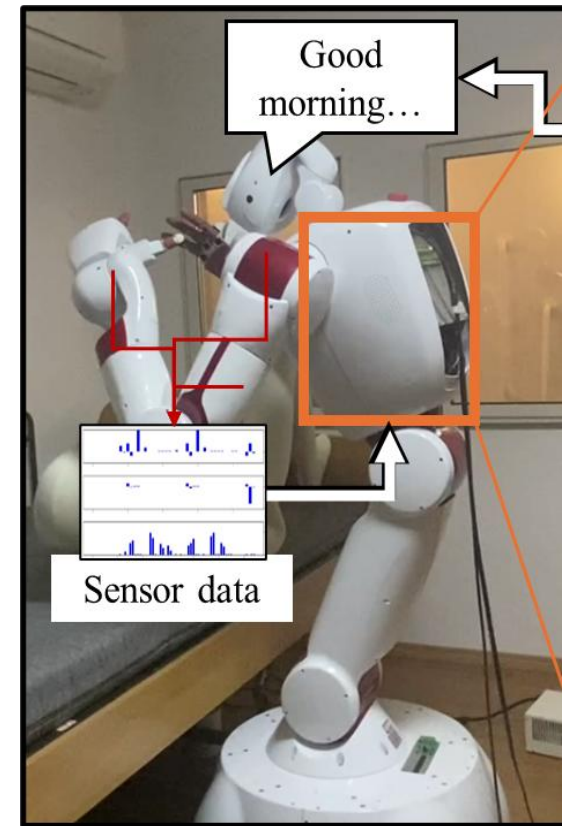
Japan Advanced Institute of Science and Technology
Waseda University



<https://xia-code.github.io/PhaseDialogue/>

Demo: Multimodal dialogue for supporting getting up

- Human robot dialogue interaction based on multimodal physical interaction



Introduction

- Since a continuous motion generally involves **physical and nonphysical phases**, providing dialogues when the phase changes is crucial for decreasing users' anxiety and guaranteeing safety.



Nonphysical phase



Physical phases

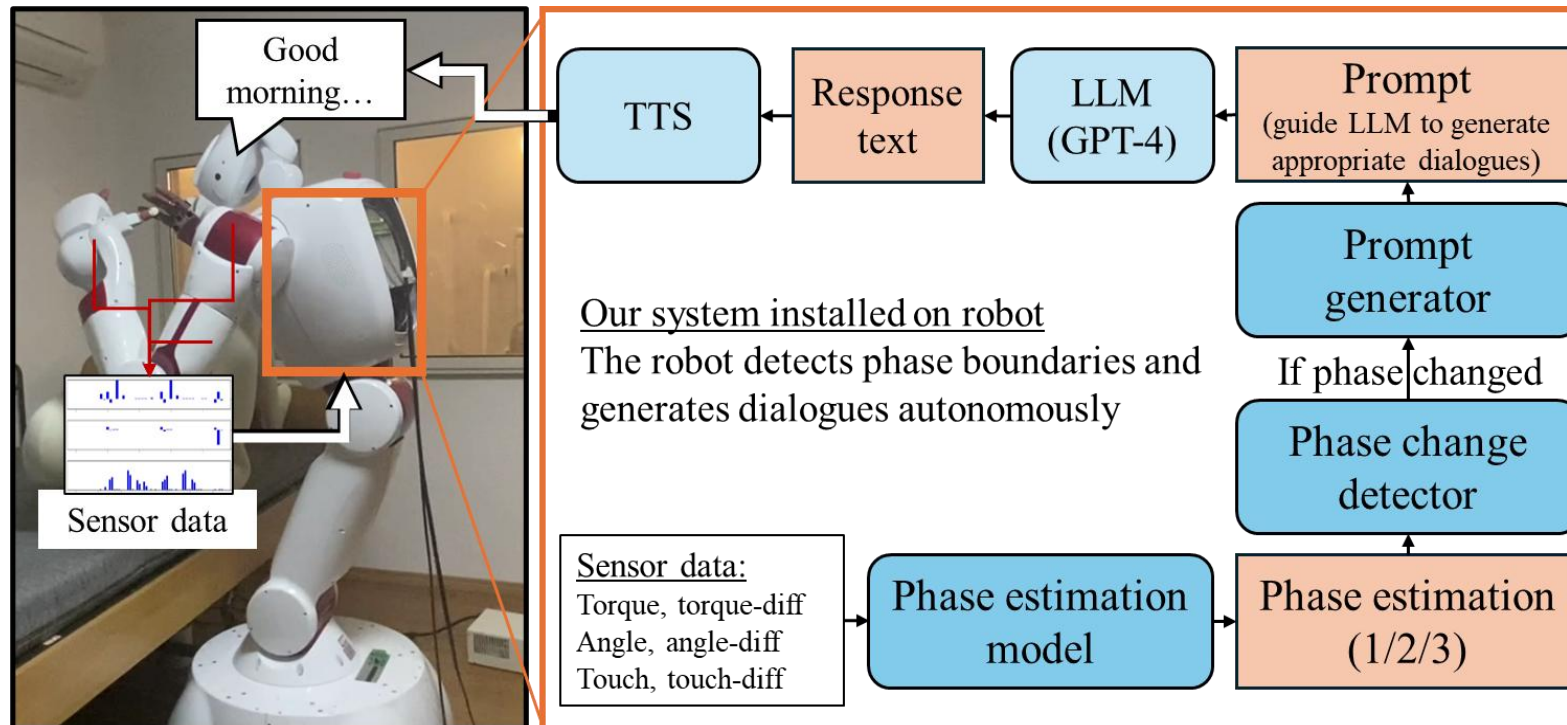


Good morning.
Today is good weather

Now, **I am lifting you up.**
Please let us know if you feel any pain in your body.

Dialogue system

- We build a system to enable autonomous dialogue generation to inform about phases and communicate to the user at phase boundaries



Agenda

- What Is Multimodal Interaction (MI) and Social Signal Processing(SSP)?
- Research of MI and SSP
- Introduction to Applied Systems of MI and SSP
- **Specific problems in multimodal ML for SSP ?**

Current Status and Challenges of MM Machine Learning for Internal State Estimation

1. Decreased accuracy in estimating internal states due to individual differences (AI bias)
 - (Example) If women speak in a brighter tone than men, the system may incorrectly classify only women as extroverted
2. It is unknown where in a conversation important nonverbal information is expressed
 - (Example) The conversation is most lively at the beginning or end...

Current Status and Challenges of MM Machine Learning for Internal State Estimation

1. Decreased accuracy in estimating internal states due to individual differences (AI bias)
 - (Example) If women speak in a brighter tone than men, the system may incorrectly classify only women as extroverted
2. It is unknown where in a conversation important nonverbal information is expressed
 - (Example) The conversation is most lively at the beginning or end...

Mitigating the Decline in Estimation Accuracy Due to Individual Differences

The recognition performance is reduced due to individual differences (AI Bias)

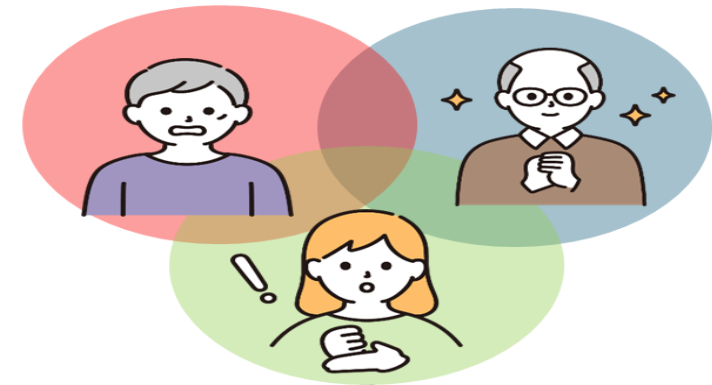
- (Example) If female tends to speak more cheerfully than male, the system may incorrectly classify only women as extroverted

⇒ **Train the system not to distinguish individual differences such as personality, gender, or age**



- **Face:** Laugh a lot, Poker face
- **Speech:** Intonation, pitch, volume
- **Language:** Catchphrase of individuals

X: Multimodal behaviors

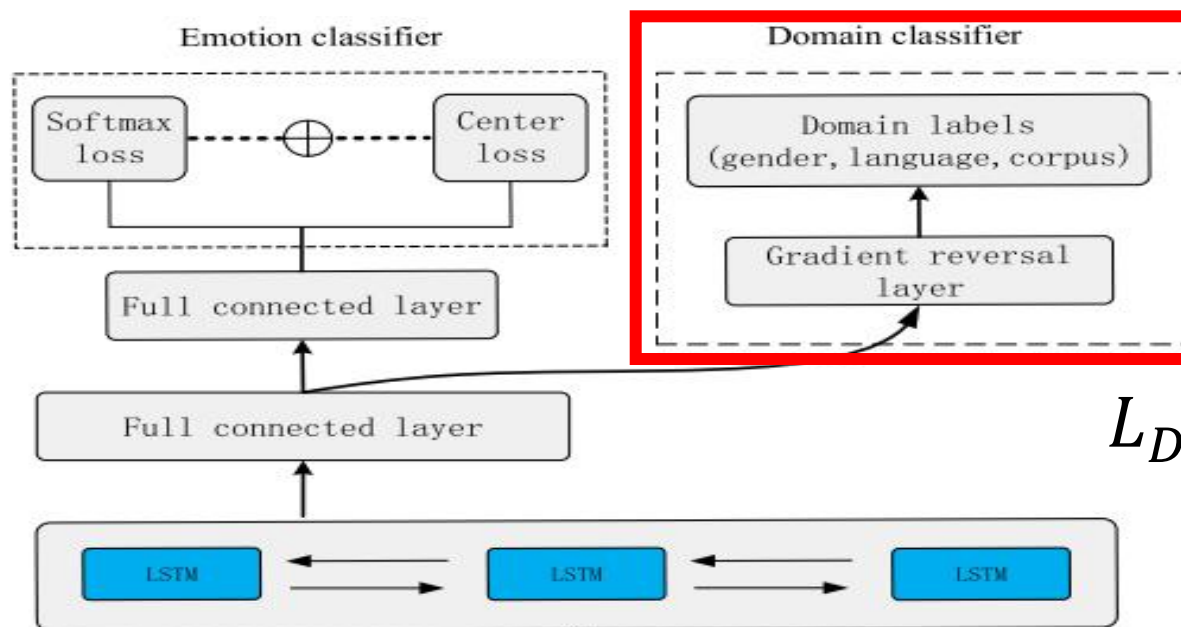


Machine learning model

Y: Social signal

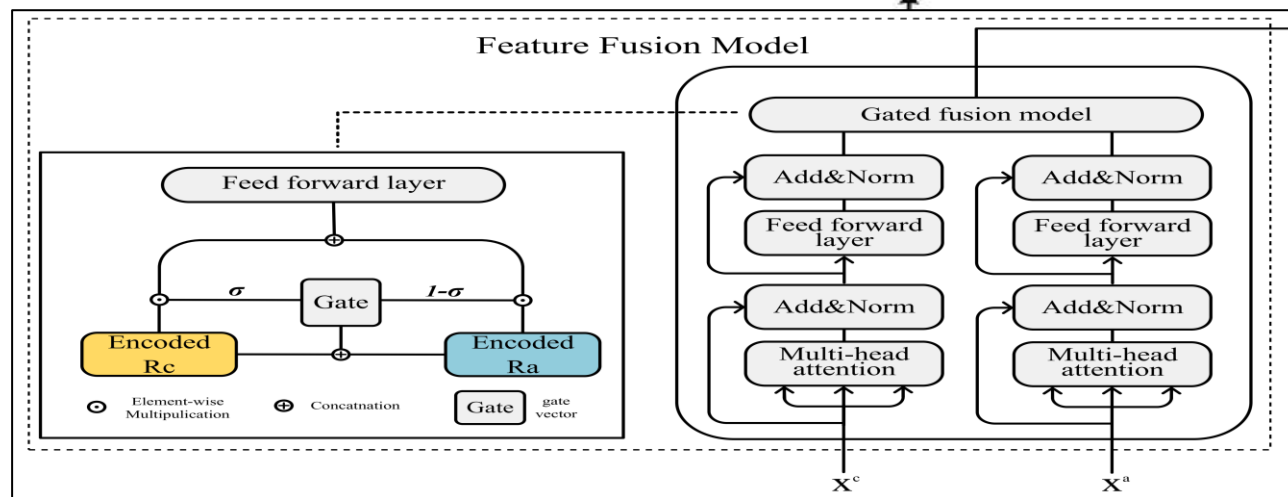
- Emotion
- Sentiments

Investigating the Effects of Non-Primary Dependent Variable in Speech Emotion Recognition Domain Adversarial Neural Network



To prevent to discriminate gender, corpus, and language by Controlling the Loss

$$L_D = L_e - \lambda(\beta L_s + (1 - \beta)L_c)$$

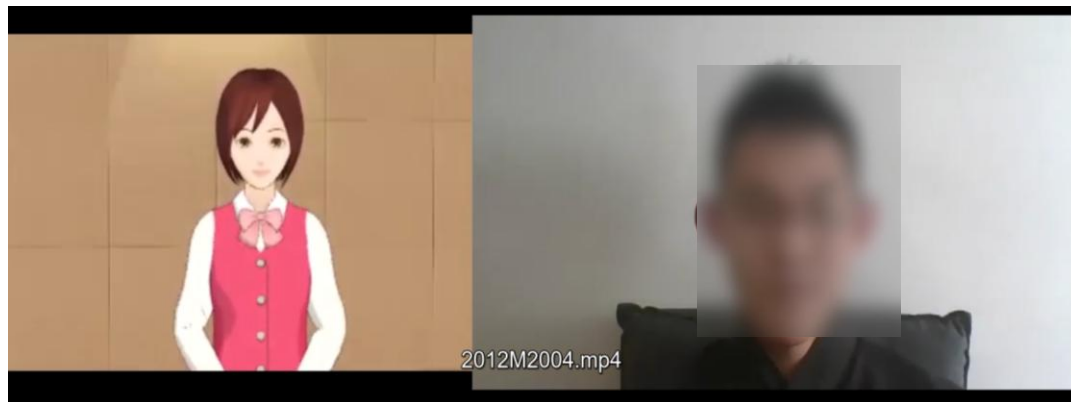
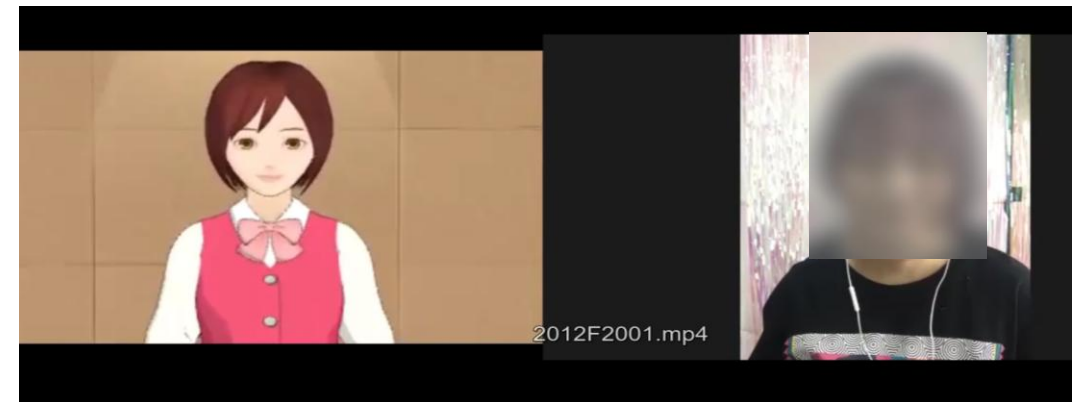


Yuan Gao, Shogo Okada, et al.
Domain-Invariant Feature Learning for Cross-Corpus Speech Emotion Recognition, IEEE ICASSP 2022

Yuan Gao, et al., Adversarial Domain Generalized Transformer for Cross-Corpus Speech Emotion Recognition, IEEE Transactions on Affective Computing 2024

Evaluating the Effectiveness of Adversarial Training (Age and Gender)

- Conducted a cross-validation experiment using data from Online Hazumi (122 participants)
- **Task:** Predict **conversation satisfaction of users** who interacted with the conversational AI



F-score		
Did the AI succeeded in adjusting the speech?	Human Judge	0.510
	LLM	0.531

Current Status and Challenges of MM Machine Learning for Internal State Estimation

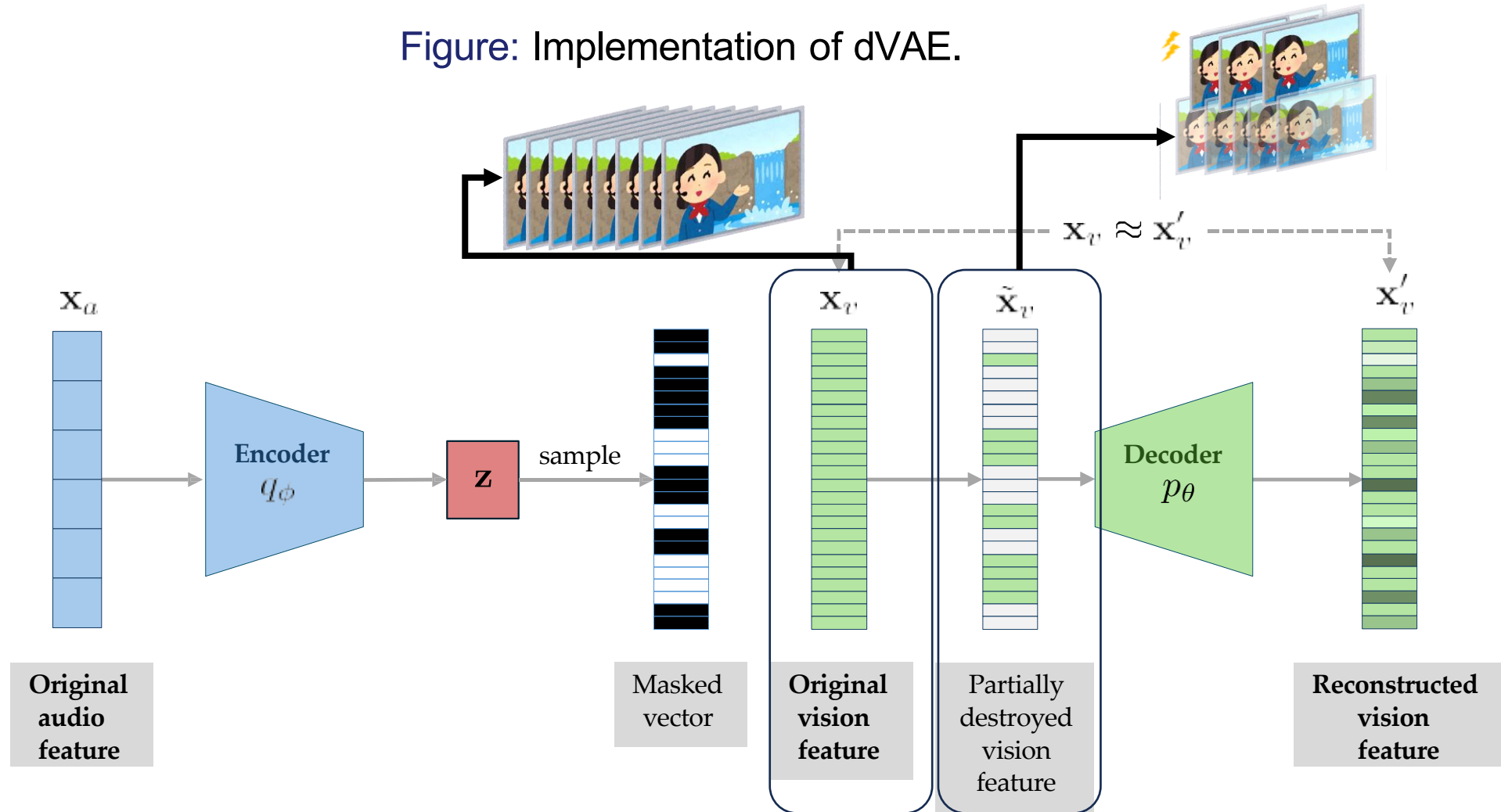
1. Decreased accuracy in estimating internal states due to individual differences (AI bias)
 - (Example) If women speak in a brighter tone than men, the system may incorrectly classify only women as extroverted
2. It is unknown where in a conversation important nonverbal information is expressed
 - (Example) The conversation is most lively at the beginning or end...

Salient Visual Frames Selection: via Discrete Variational Autoencoder⁷⁸



Hung Le

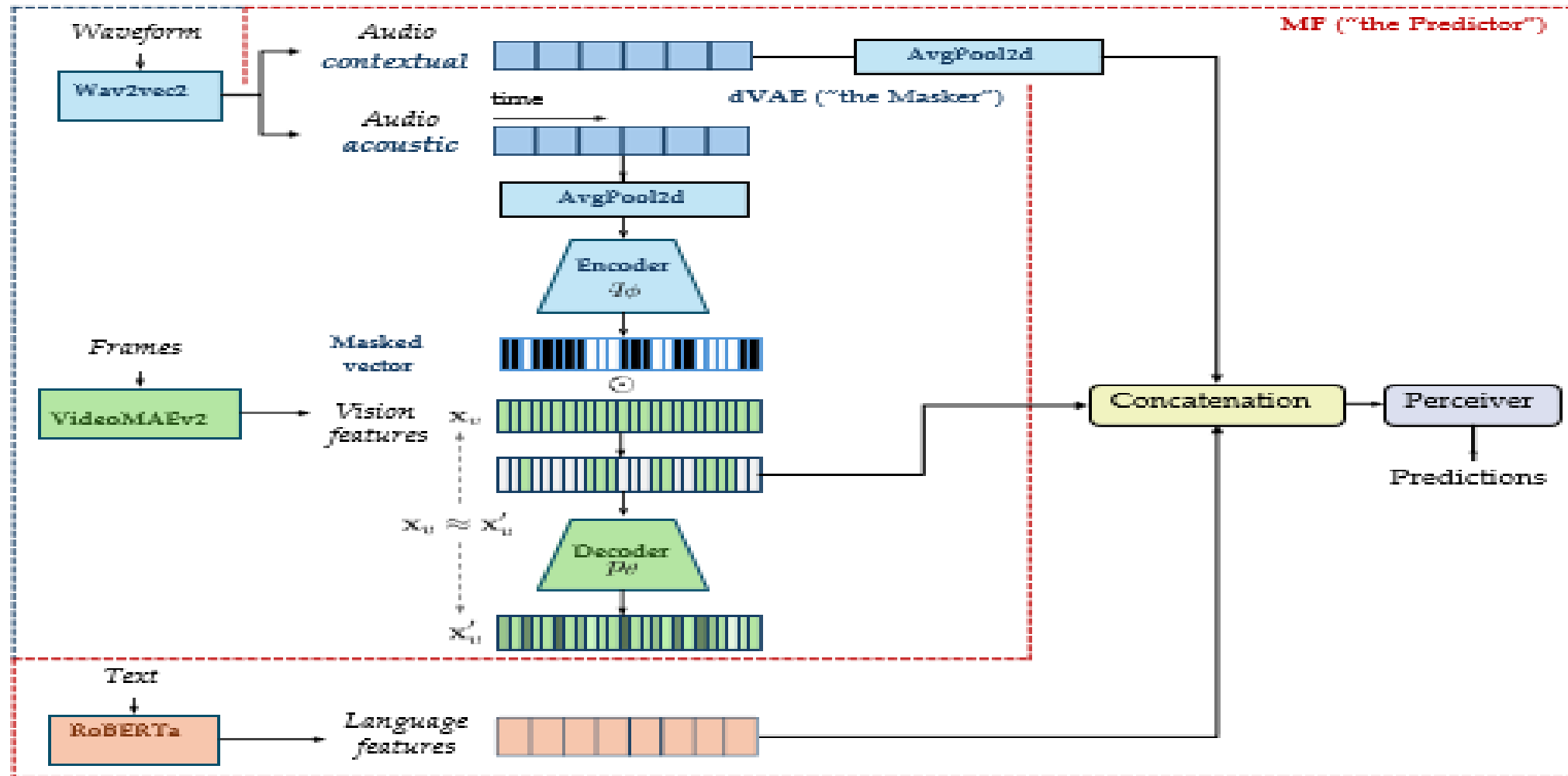
Figure: Implementation of dVAE.



Hung Le , Hung-Hsuan Huang , Candy Olivia Mawalim, Chee Wee Leong, Shogo Okada, Audio-guided visual selection for efficient multimodal fusion via a discrete variational autoencoder, *Information Fusion*, accepted

The Framework: EMF-dVAE

79



Hung Le , Hung-Hsuan Huang , Candy Olivia Mawalim, Chee Wee Leong, Shogo Okada, Audio-guided visual selection for efficient multimodal fusion via a discrete variational autoencoder, *Information Fusion*, accepted

Empirical Results: Sensitivity Curves

81

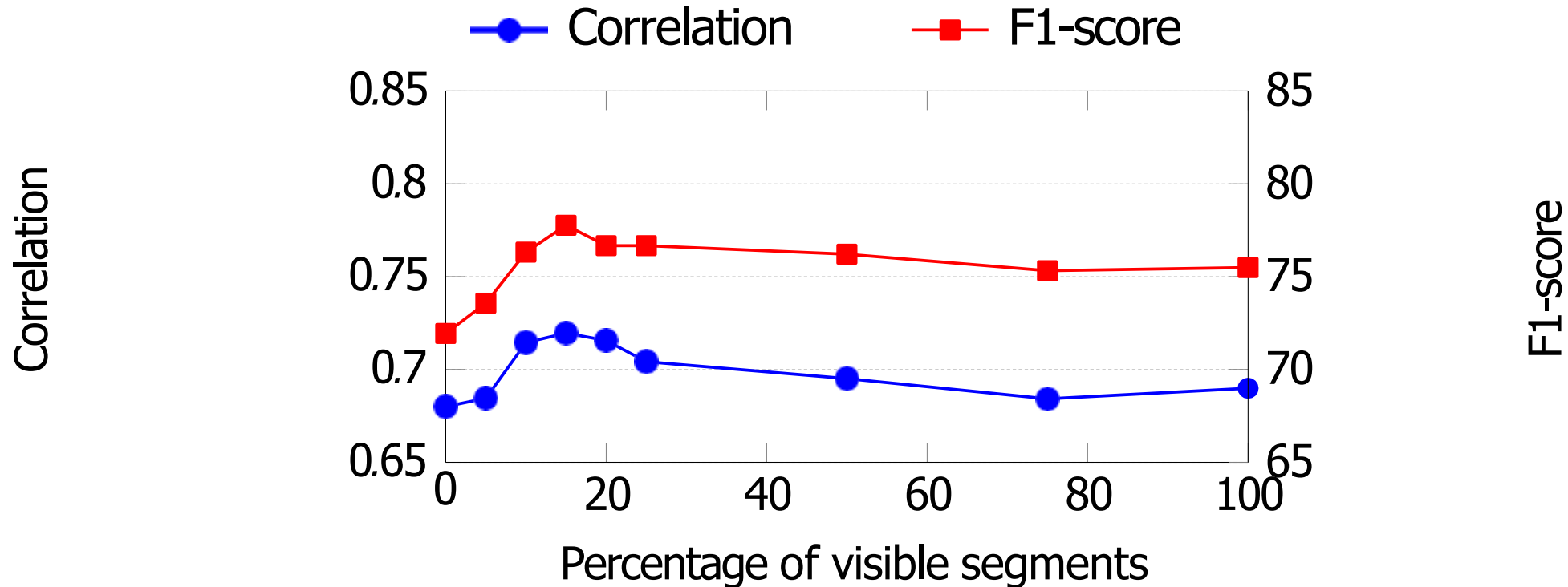


Figure: Performance of the Predictor model with different levels of masking ratio produced by the Masker.

Empirical Results: Performance

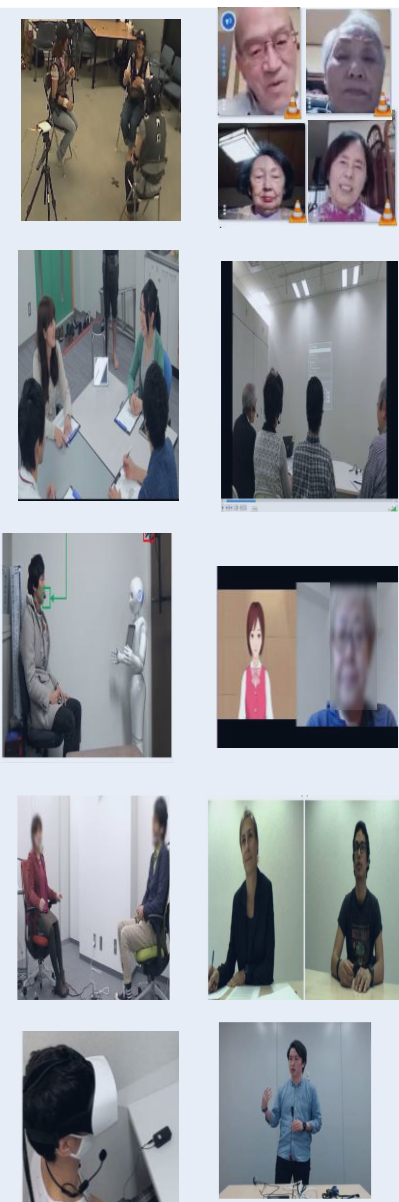
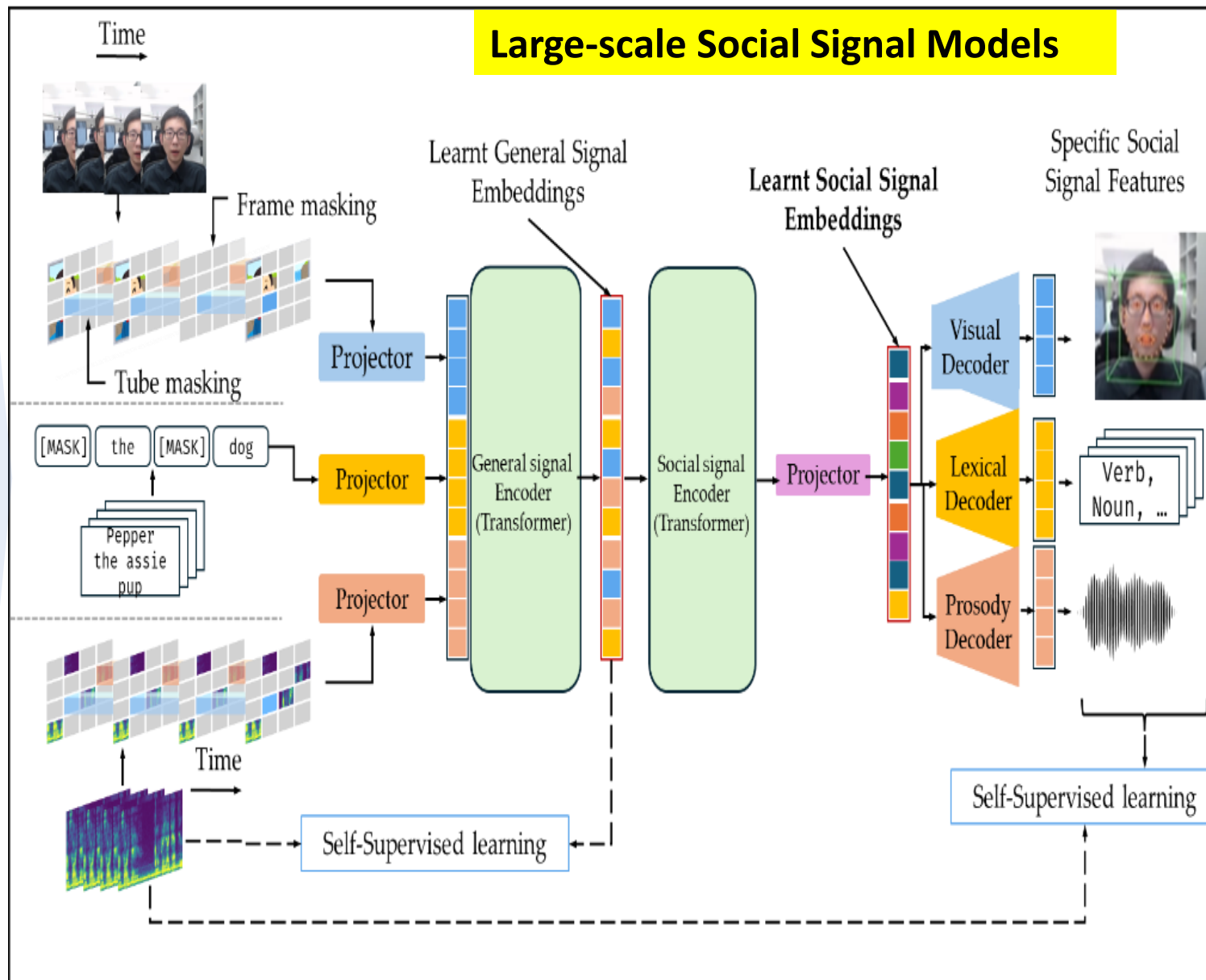
82

Table: Performance results on the ETS-Interview dataset.

Description	Vision	Corr $\uparrow$	F1 $\uparrow$	Acc2 $\uparrow$	Acc7 $\uparrow$	MAE $\downarrow$
ACII2017	-	-	66.38	66.40	-	-
HCII2023 (Chapter 3)	-	-	71.43	71.41	-	-
ICMI2024 (Chapter 4)	20%	0.6792	74.73	74.57	57.53	0.5201
MuT	100%	0.6174	67.83	68.1	52.24	0.5793
TFR-Net	100%	0.3860	61.95	64.07	51.79	0.6582
EMT-DLFR	100%	0.6816	72.24	72.58	55.91	0.5259
EMF-dVAE	15.42%	0.7196	76.64	76.77	59.28	0.4943

$\Rightarrow$ The model achieved agreement comparable to human raters (Corr = 0.70–0.79)

- Modeling Individual Differences and Corpus Differences
 - Formulation of Individual Differences and Adaptation
 - Implementation of a meta-transfer learning model across multiple different SSP tasks
- Applications of Internal State Estimation
 - Adaptation via Conversational AI
 - Feedback and Training
 - Modeling Animal-AI-Human Interaction
- Active Sensing
 - Exploring system-side interactions that elicit inner states
- Addressing Privacy and Fairness Issues



SSP downstream recognition tasks

Self efficacy, Willingness

Emotion, Mood

Cognitive performance

Relationship, Rapport

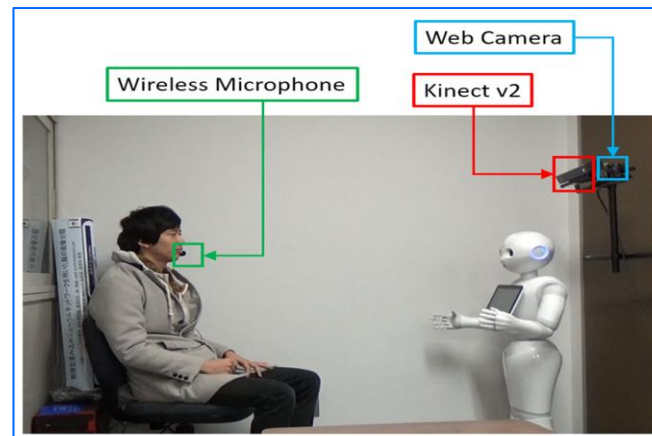
Personality

Communication skills

Group performance



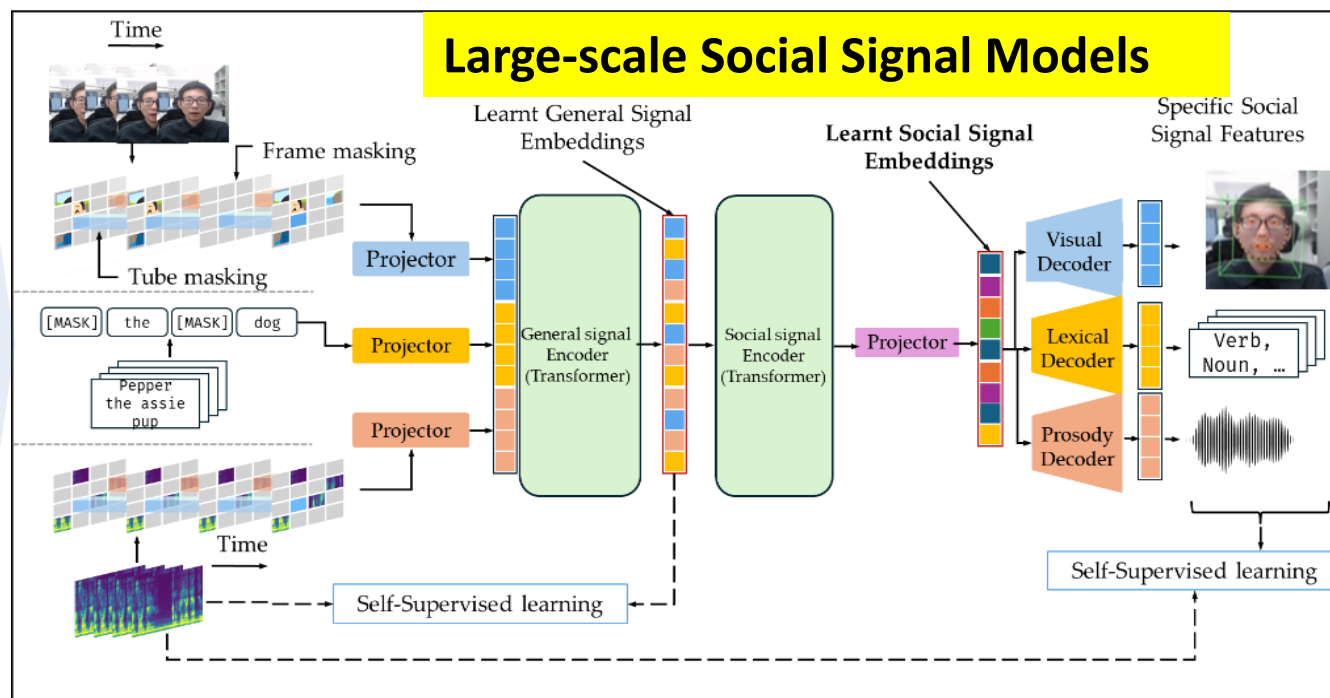
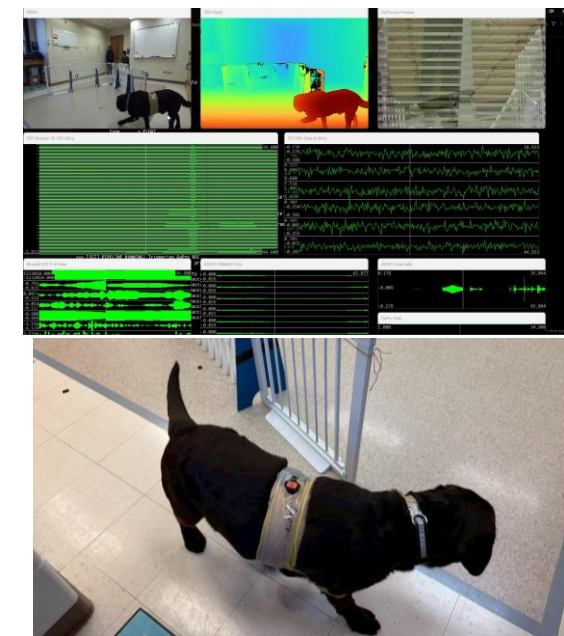
Social soft skill training



Behavior Control of Empathetic Robots • AI



Support for Monitoring Development, Cognition, and Mental State



Human Animal AI Interaction

L Acknowledgments

I would like to express my gratitude to all my collaborators for their support.

Lab members:

- Candy Olivia Mawalim, Li Sixia, Yuki Hirano, Shun Katata, Fu Yahui, Gao Yuan, Wei Weiquing, Fumiaki Nagasawa, Takato Hayashi, Le Hung, Ryusei Kimura.

Research collaborators:

- Professor Yukiko Nakano (Seikei U.), Professor Tetsuya Ogata (Waseda U.), Professor Toshihisa Tanaka (TUAT), Professor Kazunori Komatani (Osaka U.), Takahiro Tanaka (Nagoya U.), Mihoko Ohtake, TD (RIKEN), Leong, Chee Wee (ETS), Shiro Kumano (NTT), Ryo Ishii (NTT), Daniel Gatica Perez (EPFL-IDIAP), Elisabeth Andre (University of Augsburg)