Empirical Analysis of Individual Differences Based on Sentiment Estimation Performance Toward Speaker Adaptation for Social Signal Processing

Sixia Li and Shogo Okada(✉)

Graduate School of Advanced Science and Technology,
Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
{lisixia,okada-s}@jaist.ac.jp

Abstract. Understanding user's internal state is indispensable for human-robot interaction in social signal processing. To mitigate the bias of sentiments observed by third-party annotators, the importance of self-reported by users themselves was pointed out recently. However, the self-reported internal state is not displayed as similar multimodal behaviors among different individuals, this leads to performance gap between self-reported and third-party sentiment estimations. Speaker adaptation for social signal processing (SASSP) is necessary to learn individual social signal characteristics to mitigate the individual differences. Towards effective adaptation for speakers with different characteristics, clarifying influence of individual differences in internal state estimation is necessary but has not been clarified. To address this problem, this study conducted empirical analysis by training and testing models on multimodal data of a group of speakers. Then, we analyze the relationships between the best model's performance and speaker's characteristics including age, gender, personalities, and speaker's expectation before human-robot interaction experiment. The results showed that these aspects all have influence on estimation performance in SASSP due to expression differences. This study provides suggestions and directions on setting SASSP policies for self-reported internal state estimation.

Keywords: Sentiment estimation · speaker adaptation for social signal processing · machine learning

1 Introduction

Understanding user's internal state is indispensable for human-robot interaction in social signal processing, by which the machine can adjust actions to satisfy users. In practical, internal states are usually represented by user's sentiment [1,2] and emotions [3,4]. Multimodal signal cues are widely used for estimating internal states since user's state can be reflected from linguistic, acoustic, visual, and physiological aspects [5]. In modeling internal state estimation, deep learning methods are effective on capturing relationships between internal state and multimodal signals through dialogues [1].

A. Coman and S. Vasilache (Eds.): HCII 2024, LNCS 14703, pp. 359–371, 2024.
https://doi.org/10.1007/978-3-031-61281-7_26

Although internal state labels were mainly annotated by third parties in many previous studies, the importance of self-reported by users themselves was pointed out recently [2,6]. The self-reported internal state reflects user's real state more accurately compared to third-party annotations. Recent studies [7] also pointed out that self-reported sentiment differs from third-party sentiment in many aspects since the real internal state can be different from what is observed from outside. Although self-reported annotations reflect a person's real internal state, the internal state is not necessarily displayed as multimodal behaviors. As a result, estimation performance depends on the degree to which an individual person displays key-nonverbal behaviors related to the specific inner state type. To address this problem, besides using nonverbal features such as speech and facial expressions including individual difference, previous studies tried to utilize physiological signals for speaker-independent modeling to improve estimations performances [2]. But the performance on self-reported internal state still has gap to the performance on third-party internal state due to individual differences.

Speaker adaptation for social signal processing (SASSP) is necessary to learn individual social signal characteristics to mitigate the individual differences. Similar to speaker adaptation in speech processing area, SASSP aims to train the model with specific speakers' data to correlate individual multimodal behaviors and internal states. In this way, the influence of expression differences can be mitigated. Towards effective adaptation for speakers with different characteristics, clarifying influence of individual differences in internal state estimation is necessary. However, such analysis was not considered in previous studies, this leads to difficulties in setting adaptation policies.

To address this problem, this study analyzes individual differences in SASSP in an empirical way. Specifically, we train models by using multimodal data of a group of speakers. Then we test the model performance by the data of same users with guaranteeing that the training and testing data are different. After that, we analyze the relationships between the best model's performance and speaker's characteristics, including age, gender, personalities, and speaker's expectation before human-robot interaction experiment. We try to clarify what kind of speakers can be adapted easily or difficultly. By this analysis, this study provides suggestions and directions on setting SASSP policies for self-reported internal state estimation.

The contributions of this study can be summarized as follows:

1. This study analyzed the individual differences in SASSP for self-reported internal state estimation. To our best knowledge, we are the first to conduct such analysis. This can provide suggestions for adaptation policy settings.
2. We regarded the individual differences in the degree to expressions as differences in estimation accuracy. The differences in estimation performance are analyzed in detail by comparing them with personal information including gender, age, personality traits, and expectations before human-robot interaction. The results showed that these aspects all have influence on estimation performance in SASSP due to expression differences.

2 Methodology

For our purpose, we train self-reported internal state estimation models on multimodal data of a group of speakers. Specifically, we use self-reported sentiment as the self-reported internal state since the sentiment can generally represent the internal state. The sentiment is generally represented as high or low labels to show whether the speaker has a positive or negative internal state. Speakers in experiments cover various characteristics including different genders, ages, etc. So the data can be considered as a 'general data.' Then we test model performances on these speakers' data by guaranteeing the data is individual in training and test. This process can be considered as adapting a model to each specific speaker.

After that, we analyze the performance regarding to four characteristics to clarify the influence of these characteristics in SASSP. For categorized characteristics such as gender, we compare the speaker-averaged performance among categories for analysis. For scored characteristics such as personality traits, we compute the correlation coefficients between speaker performances and characteristic scores to clarify whether a common score leads to a common performance.

2.1 Characteristics for Analysis

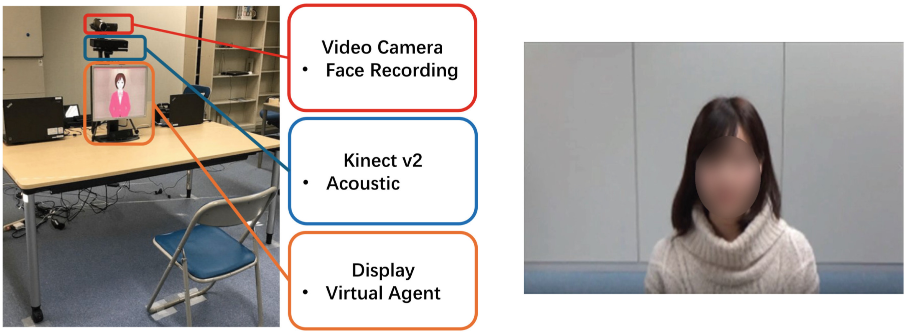
We analyze the influence of gender, age, personality traits, and expectations before human-robot interaction experiment.

Gender gender is a general characteristic that bring differences in many tasks [8–10]. In this experiment, the speakers were separated into male and female categories. We compare the average performance between genders to show the influence of gender in SASSP.

Age age is another typical characteristic that has influence on many tasks [11, 12]. We separate age categories by every 10-years-old period. As a consequence, we have age categories such as 20, 30, etc. We compare average performances among ages to show the influence of age in SASSP.

Personality traits the influence of personality traits in human-robot interaction have attracted attentions recently [13–16]. One's personality has influence the interaction process from multiple aspects, including the expressions and the change of internal states. Therefore, considering personalities is necessary for adapting to specific speaker. In this study, we use the Big-five personality traits [17] as the metric for measuring one's personalities. The big-five is a grouping of five unique personality characteristics including openness, conscientiousness, extraversion, agreeableness, and neuroticism. We compute the Pearson correlation coefficients between model performances and big-five score to analyze the influence of each trait in SASSP.

Expectations Before Human-Robot Interaction Experiments recent study [18] pointed out that one's prior thought and expectations towards the interaction with robot would influence the performance during the interaction.



Data collection environment

Face video recording

Fig. 1. Data collection environment and video recording of Hazumi1911 dataset

Therefore, the expectations can be speculated to have influence on adaptations. In this study, we use questionnaires about expectations of the human-robot interaction to obtain expectation scores from various aspects. And We compute the Pearson correlation coefficients between model performances and the score of expectations to analyze the influence of each expectation aspect in SASSP.

3 Experiment

3.1 Dataset

We use the Hazumi1911 dataset [6,27–29] for conducting experiments. This dataset is a Japanese dataset of human-robot interaction dialogues. The dataset was collected by using a virtual agent to make dialogues with speakers. Video recordings and text transcripts are available in the dataset. So we can use the linguistic, acoustic, and facial modalities. Figure 1 shows the data collection environment and the video recording screenshot of the dataset.

In Hazumi1911 dataset, the self-reported sentiments were annotated by speakers themselves as their internal states. The internal state were originally annotated by 7-point Likert scale, from 1 to 7. Previous studies [2,5] usually compressed the 7-point sentiment into low sentiment and high sentiment for practice use. The threshold was 4, which means when the point is less or equal than 4, the sentiment is treated as low; while when the point is greater or equal than 4, the sentiment is treated as high. We follow this setting in this study to make our analysis to be useful for practice use.

This dataset contains speaker information of gender, age, personality traits, and questionnaires of expectations to human-robot interaction. The gender contains male and female categories. The ages of speakers are categorised into 20, 30, 40, 50, 60, and 70-yeas-old. The personality traits were originally measured by Japanese ten item personality inventory (TIPI-J) with 10 items. We referred to the manual [25] to transform TIPI-J to big-five scores. The questionnaires contains 18 aspects that describe expectations towards human-robot interaction.

Table 1. Characteristics and items for analysis of SASSP

Characteristic aspect	items
Gender	Male, Female
Age	20, 30, 40, 50,60, 70-years-old level
Big Five personality traits	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
Expectations before human-robot interaction	well-coordinated, boring, cooperative, harmonious, unsatisfying, uncomfortably paced,cold, awkward, engrossing, unfocused, involving, intense, friendly, active, positive, dull, worthwhile, slow

Answers of questionnaires are in 7-point Likert scale form, which represents how the speaker expect the interaction in specific aspect. Table 1 summarizes the individual characteristics and items for analysis.

For the experiment, we first conducted data cleaning to remove missing data for experiments. After data cleaning, Hazumi1911 contains 26 human-robot dialogues by different speakers. The speakers contain 12 male and 14 female. The age categories of 20, 30, 40, 50, 60, and 70 contains 5, 3, 5, 4, 6, and 3 speakers, respectively. The total utterance is 2468, including 1359 low sentiment utterances and 1109 high sentiment utterances. Based on the labels we use, we conduct experiments as a binary classification task to predict whether an utterance is low sentiment or high sentiment.

For each dialogue, we split 80% utterances from the beginning as training set, and the remaining 20% utterances of each dialogue were used for testing. In the training set, we further split 80% utterances for training the model, and the remaining 20% utterances in the training set is used as validation set for choosing the best parameters. As a consequence, we have 1566 utterances for training, 398 utterances for validation, and 504 utterances for test.

3.2 Multimodal Features

We use multiple modalities and their combinations, including linguistic (L), acoustic (A), and facial (F) features.

Linguistic Features. Bert [19] is a powerful model to obtain effective linguistic representations in various tasks [20]. We use the mean embedding of the last hidden layer of Bert as the linguistic feature. In particular, we use the cl-tohoku/bert-base-japanese model to extract features. As a consequence, we obtain a 768-dimension vector of the linguistic feature for each utterance.

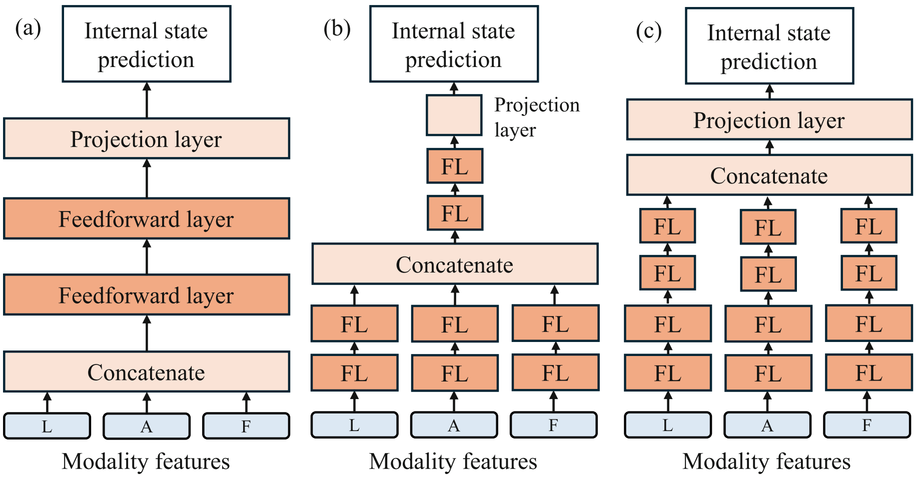


Fig. 2. Model structures

Acoustic Features. We use the Interspeech2009 feature set (IS09) [21] as the acoustic feature. The IS09 contains 16 low-level descriptions including F0, energy, zero cross rate, and MFCCs, and 12 functional description including mean, variance, etc. This feature set was shown effective in emotion recognition and sentiment analysis tasks [21], so it is suitable for modeling sentiment-related internal state. We use OpenSmile [22] tool to extract IS09 features. As a consequence, we obtain a 386-dimension vector of the acoustic feature for each utterance.

Facial Features. We use action units (AUs) from facial action coding system (FACS) [23] to represent facial information. In particular, we use OpenFace [23] tool to extract 18 discrete AUs for each frame of utterances. Each AU is represented as value 1 if that AU is estimated to be appeared, or value 0 if that AU is estimated as not to be appeared. After that, we compute the mean value of each AU among all frames of an utterance as the facial feature. As a consequence, we obtain a 18-dimension vector of the facial feature for each utterance.

3.3 Models

We use Three DNN-based models that were widely used in many studies and especially be shown effective [2] for modeling self-reported internal state. These models include an early fusion model (named DNN Early hereafter), two late fusion models with different fusion strategies (named DNN Late 1 and DNN Late 2 hereafter). The model structure is shown in Fig. 2.

DNN Early. Figure 2 (a) shows the structure of DNN Early model. This model is consisted of 2 feedforward layers with 256 units. The Relu function is used for activation functions of each layer. The modality feature vectors are first concatenated before inputting into the model. After feedforward layers, a feedforward layer with softmax function is used to project middle tensors to label probabilities. Finally, the label with the highest probability is chosen as the prediction.

DNN Late 1. Figure 2 (b) shows the structure of DNN Late 1 model. In the model processing, each modality is first inputted into separated modality feedforward layers. Each modality feedforward layer contains 2 layer with 64 units of each layer. Then the output of all modality feedforward layers are concatenated into one vector. After that, the concatenated vector is inputted into fusion feedforward layers. Fusion feedforward layers contains 2 layers with 32 units of each layer. After feedforward layers, a projection layer with softmax function is used to obtain label probabilities. The Relu function is used for activation functions of each layer.

DNN Late 2. Figure 2 (c) shows the structure of DNN Late 2 model. In the model processing, each modality is embedded with separated modality feedforward layers. Each modality feedforward layer contains 4 layers. The units of layers are 64, 64, 32, and 32 from the input side to the output side. After separated embedding, the output of modality layers is concatenated into one vector. Then the concatenated vector is inputted into a feedforward layer. After the feedforward layer, a projection layer with softmax function is used to obtain label probabilities. The Relu function is used for activation functions of each layer.

3.4 Experiment Setting

We use macro F1 as the evaluation metric, and we conduct analysis of variance (ANOVA) test to clarify significant difference between different groups. All training and testing were conducted three times, the average performances were used for evaluation to reduce the influence of random initialization.

4 Result and Discussion

Table 2 lists averaged macro F1 of models using each modality combinations. Underlined numbers indicate the best performance of each model, bold number indicates the best performance among all models. As seen in the table, using A+F modality is the best performance of each model. The best performances of DNN Early, DNN Late 1, and DNN Late 2 are 0.725, 0.732, and 0.739 of macro F1 score, respectively. We found that the best modality A+F is inconsistent to the best modality in speaker-independent studies using Hazumi1911 [6], in which the best modality combination was usually consisted with linguistic modality. We speculate that the A+F is the best combination is related to the intrinsic characteristic of self-reported internal state. The internal states of different speakers are affected by individual differences and could differ from similar outside expressions. Linguistic modality can somehow be effective cue on common situations among speakers, where speakers' specific contents reflecting specific internal states. So the linguistic modality is effective in speaker-independent modeling. On the other hand, this study trains the model by using

Table 2. Macro F1 of different models using different modality combinations

	Model		
Modality	DNN Early	DNN Late 1	DNN Late 2
L	0.621	0.628	0.606
L+A	0.682	0.707	0.711
L+F	0.657	0.690	0.687
L+A+F	0.696	0.728	0.737
A	0.718	0.711	0.720
F	0.703	0.685	0.688
A+F	0.725	0.732	**0.739**

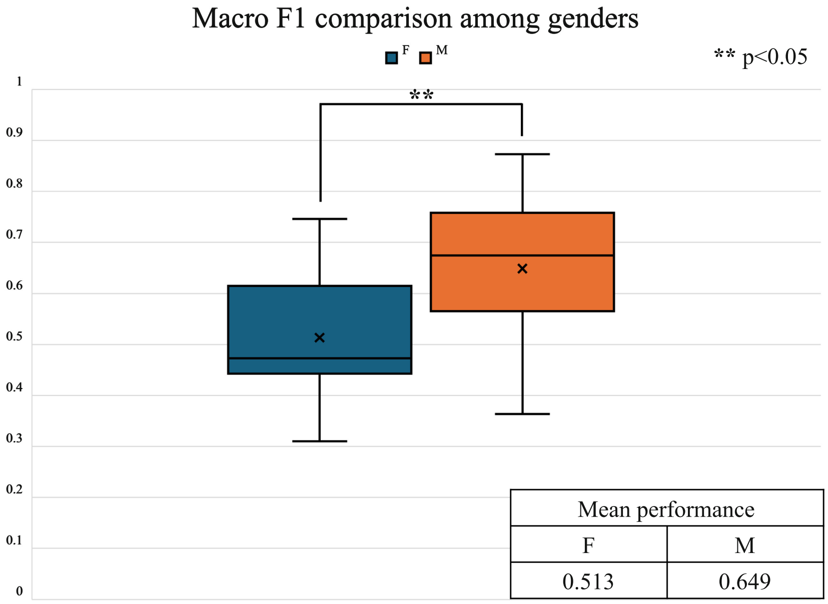
speaker's data to adapt to each speaker. The model can better capture the relationship between individual multimodal behaviors and internal states. Therefore, the results demonstrates that acoustic and facial modalities are more related to internal states than linguistic modality for adapting individual situations. So they can be better cues than linguistic in SASSP.

By comparing the best performance of three models, one can see that DNN Late 2 using A+F modality has the best performance among all models. Therefore, we treat DNN Late 2 using A+F modality as the representative model for analysis.

Next, we analyze the influence of individual differences in SASSP by the way described in Sect. 2 based on the result of DNN Late 2 using A+F modality.

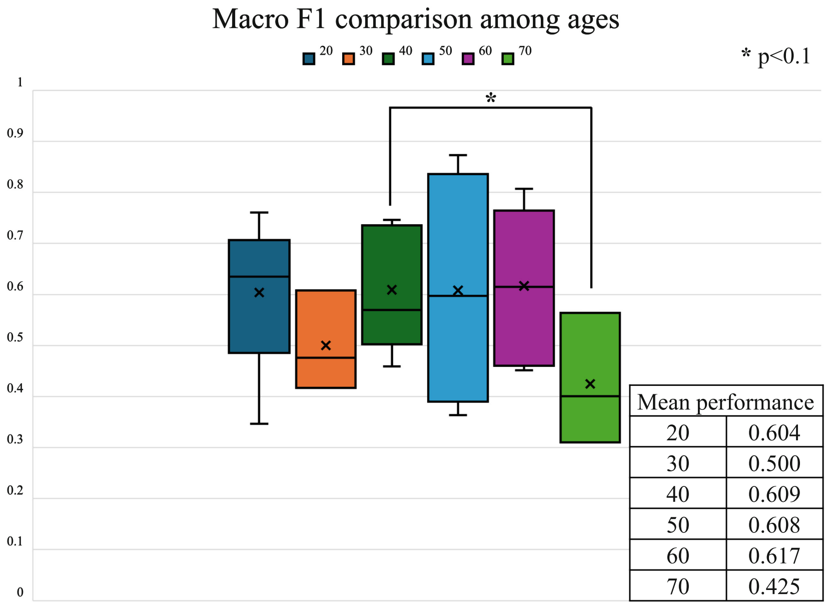
Figure 3 shows the performance comparison among gender categories. As seen in the figure, the mean performance of female speakers is 0.513, while the mean performance of male performance is 0.649. The performances of two genders are significant different based on ANOVA result. We speculate that female speakers tend take more care to match the rhythm of the robot to make the dialogue not to be awkward. Thus, external performances corresponding to high or low sentiment are not consistent through the whole dialogue. So it became difficult for the model to use the relationship learned from a part of dialogue to adapt the remaining dialogue for female speakers.

Figure 4 shows the performance comparison among age categories. As seen in the figure, the mean performance from 20-years-old to 70-years-old are 0.604, 0.500, 0.609, 0.608, 0.617, and 0.425, respectively. By comparing the performance of each age category, one can see that the mean performance of groups except 30-years-old and 70-years-old are close. While 70-years-old and 30-years-old group has a lower mean performance than other groups. But the significant difference only exists between 70-years-old and 40-years-old group. We speculate the reason of significant low performance of 70-years-old group is related to that facial expression patterns of elder people is not as obvious as young people [26]. Meanwhile elder people may not control their expressions as well as young people. So their internal state is difficult to be estimated from facial aspect,



Mean performance	
F	M
0.513	0.649

Fig. 3. Macro F1 of gender categories



Mean performance	
20	0.604
30	0.500
40	0.609
50	0.608
60	0.617
70	0.425

Fig. 4. Macro F1 of age categories

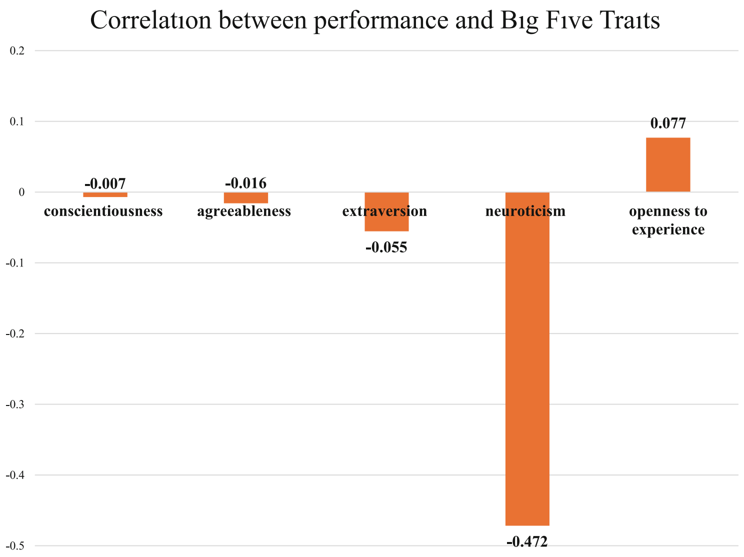


Fig. 5. Pearson correlation coefficient between personality traits and speaker performance

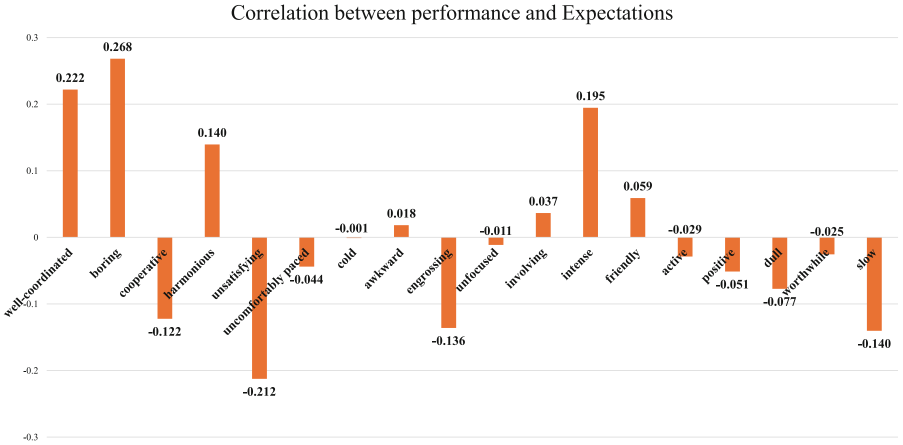


Fig. 6. Pearson correlation coefficient between expectation before human-robot interaction and speaker performance

meanwhile they may not have consistent external performance related to their internal state. Therefore, the model learned from a part of dialogue is hard to adapt to handle remaining dialogues for old people in 70-years-old.

Figure 5 shows the Pearson correlation coefficient between personality traits and speaker performances. As seen in the figure, among five personality traits, only neuroticism trait has a significant negative correlation to the model performance. While other traits rarely related to performances. We speculate the reason is that people with high neuroticism tend to hide their real thought, so their external performances are not related to their internal state consistently. This leads to difficulties on adaptation for high neuroticism people.

Figure 6 shows the Pearson correlation coefficient between expectation before human-robot interaction and speaker performances. As seen in the figure, cooperation and boring items have weak positive correlations 0.222 and 0.268, respectively; unsatisfied item has weak negative correlation −0.212. The results show that expectations have influence on speaker's internal state-related external performances through dialogue, thus influence the adaptation.

5 Conclusion

This study conducted empirical analysis to clarify the influence of individual differences in speaker adaptation for social signal processing on internal state estimation. The results demonstrated that acoustic and facial modalities are more effective than the linguistic modality in adaptation. The results showed that various individual differences including gender, age, personality traits, and expectations before human-robot interaction all have influence on adaptations, leading to relationships between speakers' outside expressions and their internal states to be not consistent during dialogues. Therefore, capturing and mitigating such inconsistencies can be considered a direction for better adaptation.

Acknowledgement. This work was also partially supported by JSPS KAKENHI (22K21304, 22H04860, 22H00536, 23H03506), JST AIP Trilateral AI Research, Japan (JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237).

References

1. Gandhi, A., Adhvaryu, K., Poria, S., et al.: Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf. Fusion **91**, 424–444 (2023)
2. Katada, S., Okada, S., Hirano, Y., Komatani, K.: Is she truly enjoying the conversation? Analysis of physiological signals toward adaptive dialogue systems. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 315–323 (2020)
3. Mittal, T., Bhattacharya, U., Chandra, R., et al.: M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 02, pp. 1359–1367 (2020)
4. Busso, C., Bulut, M., Lee, C.C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
5. Katada, S., Okada, S., Komatani, K.: Effects of physiological signals in different types of multimodal sentiment estimation. IEEE Trans. Affect. Comput. (2022)
6. Komatani, K., Okada, S.: Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE (2021)

7. Komatani, K., Takeda, R., Okada, S.: Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In: Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pp. 104–113 (2023)
8. Usart, M., Grimalt-Álvaro, C., Iglesias-Estradé, A.M.: Gender-sensitive sentiment analysis for estimating the emotional climate in online teacher education. Learn. Environ. Res. **26**(1), 77–96 (2023)
9. Volkova, S., Wilson, T., Yarowsky, D.: Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing 2013, pp. 1815–1827 (2013)
10. Abbruzzese, L., Magnani, N., Robertson, I.H., et al.: Age and gender differences in emotion recognition. Front. Psychol. **10**, 2371 (2019)
11. Bailey, P.E., Brady, B., Ebner, N.C., et al.: Effects of age on emotion regulation, emotional empathy, and prosocial behavior. J. Gerontol. Ser. B **75**(4), 802–810 (2020)
12. Kim, E., Bryant, D.A., Srikanth, D., et al.: Age bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 638–644 (2021)
13. Le, H., Li, S., Mawalim, C.O., et al.: Investigating the effect of linguistic features on personality and job performance predictions. In: Coman, A., Vasilache, S. (eds.) HCII 2023. LNCS, vol. 14025, pp. 370–383. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35915-6_27
14. Yan, D., Chen, L.: The influence of personality traits on user interaction with recommendation interfaces. ACM Trans. Interact. Intell. Syst. **13**(1), 1–39 (2023)
15. Böckle, M., Yeboah-Antwi, K., Kouris, I.: Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. In: Degen, H., Ntoa, S. (eds.) HCII 2021. LNCS (LNAI), vol. 12797, pp. 3–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77772-2_1
16. Alves, T., Natálio, J., Henriques-Calado, J., et al.: Incorporating personality in user interface design: a review. Personal. Individ. Differ. **155**, 109709 (2020)
17. Gosling, S.D., Rentfrow, P.J., Swann, W.B., Jr.: A very brief measure of the Big-Five personality domains. J. Res. Pers. **37**(6), 504–528 (2003)
18. Blut, M., Wang, C., Wünderlich, N.V., et al.: Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. J. Acad. Mark. Sci. **49**, 632–658 (2021)
19. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
20. Liu, P., Yuan, W., Fu, J., et al.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
21. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge (2009)
22. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the MUNICH versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)
23. Ekman, P., Friesen, W.V.: Facial action coding system. Environ. Psychol. Nonverbal Behav. (1978)

24. Baltrusaitis, T., Zadeh, A., Lim, Y.C., et al.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
25. https://jspp.gr.jp/doc/manual_TIPI-J.pdf
26. Lopes, N., Silva, A., Khanal, S.R., et al.: Facial emotion recognition in the elderly using a SVM classifier. In: 2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW), pp. 1–5. IEEE (2018)
27. Hirano, Y., Okada, S., Komatani, K.: Recognizing social signals with weakly supervised multitask learning for multimodal dialogue systems. In: Proceedings of the International Conference on Multimodal Interaction, pp. 141–149 (2021)
28. Hirano, Y., Okada, S., Nishimoto, H., et al.: Multitask prediction of exchange-level annotations for multimodal dialogue systems. In: 2019 International Conference on Multimodal Interaction, pp. 85–94 (2019)
29. Wei, W., Li, S., Okada, S.: Investigating the relationship between dialogue and exchange-level impression. In: Proceedings of the International Conference on Multimodal Interaction, pp. 359–367 (2022)