Relationships between Gender, Personality Traits and Features of Multi-Modal Data to Responses to Spoken Dialog Systems Breakdown

Kazuya Tsubokura[1], *Yurie Iribe*[1], *Norihide Kitaoka*[2]

[1]Aichi Prefectural University, Japan
[2]Toyohashi University of Technology, Japan

im212008@cis.aichi-pu.ac.jp, iribe@ist.aichi-pu.ac.jp, kitaoka@tut.jp

Abstract

Automated dialog systems are currently being used in various applications, but it is unclear if they will ever be able to converse as naturally as humans do. One challenge is avoiding breakdowns during conversations due to inappropriate system utterances. Although many studies have focused on dialog breakdown detection, the influence of differences among individual users on dialog breakdowns and breakdown detection has not been sufficiently examined. In this study, we focus on individual differences thought to be related to emotional responses after breakdowns, specifically language, acoustic, and facial features, as well as gender and BigFive personality traits, to analyze differences in user responses to breakdowns. Our results suggest that gender and personality traits influence user responses to dialog breakdowns. For example, users with low Openness scores were more likely to express anger, while women were less likely to do so.

Index Terms: spoken dialog system, dialog breakdown, individual difference, personality traits

1. Introduction

Recent advances in natural language processing and speech recognition technology have led to the practical use of automated dialog systems such as smart speakers and dialog robots. Ideally, these dialog systems should always be able to respond appropriately to the utterances of the user, but “dialog breakdowns”, in which the dialog system responds to the user with an improper utterance, are still a frequent occurrence [1]. Therefore, further research is necessary to develop technology that can detect and repair dialog breakdowns, using language information included in the utterances of users and the dialog system, as well as non-verbal cues displayed by the user. If potential breakdowns in conversations can be predicted, or actual breakdowns detected, it may be possible to initiate breakdown avoidance or recovery strategies that will increase the user’s willingness to continue the dialog.

Since the ability to detect dialog breakdowns can improve the naturalness of conversations with dialog systems, the Dialog Breakdown Detection Challenge was convened to advance research aimed at detecting dialog breaks in text chats [1]. Although interaction breakdown detection using multimodal information, such as user voice features, facial expressions and gestures, in addition to user utterances, has been studied [2, 3, 4], individual differences between users have yet to be considered in conventional breakdown detection methods using such multimodal information. Previous studies have reported individual differences in the non-verbal features of users during dialog breakdowns however [2]. If differences in the multimodal information obtained from users can be linked to differences in individual responses to dialog breakdowns, this could dramatically improve breakdown detection.

In this study, we first conducted dialog breakdown experiments, and then analyzed individual differences in the multimodal information collected from users during their responses to breakdowns. Since users respond differently when dialog systems break down, expressing anger, confusion, or amusement, for example, we also focused on gender and personality traits, which are considered to be associated with emotional responses, as factors of user individuality.

In Section 2, we discuss the data we collected from participants during our experiment, and in Section 3 we describe how we extracted multimodal features from this data. In Section 4, we report the relationships observed between these multimodal features and the individual user characteristics of gender and personality traits. We then summarize the findings of this paper in Section 5.

2. Collected Data

This section describes the multimodal dialog data collected for our experiment. We recorded audio of all user and dialog system utterances, as well as continuous video of both the animated dialog agent and the area above each participant’s neck, using the Zoom app. Participants in the dialog experiment consisted of 33 university students (19 men and 14 women). A total of 99 chat dialog sessions were recorded (three sessions per participant). Participants were instructed to utter more than ten responses per session. The dialogs started with a system utterance, and ended when participants said “Goodbye”, when they wanted to end the dialog. This study has been approved by the research ethics committee of the organization to which the first author belonged. Informed consent has been obtained from all participants included in this study. All participants of this study did not agree for their data to be shared publicly, so data set generated and analyzed in this study is not available.

2.1. Dialog System Used in the Experiment

An outline of the architecture of the spoken dialog system used in our experiment is shown in Fig. 1. The voice of the experimental participant is input into the system remotely using the online conferencing system Zoom, and then extracted by a virtual mixer. The Google Cloud Speech-to-Text API then performs speech recognition, and the speech recognition results for the user are used as the input for the dialog system’s response generation model, “Japanese-dialog-transformers” [5], a Japanese-language, Transformer-based, encoder-decoder dialog model. After the response generation model generates response candidates, filtering is performed to avoid repetition of the same responses. Specifically, we computed the Jaccard coefficients

10.21437/Interspeech.2023-1267

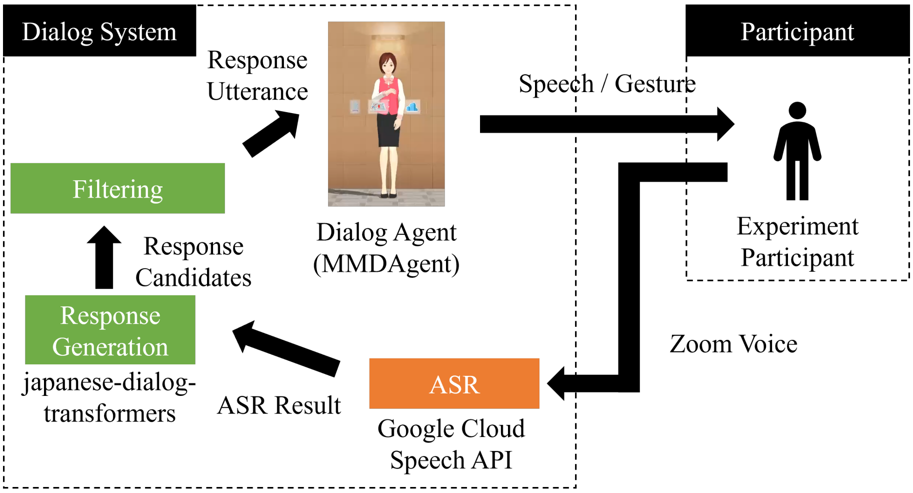

Figure 1: *Remote spoken dialog system used for experiment.*

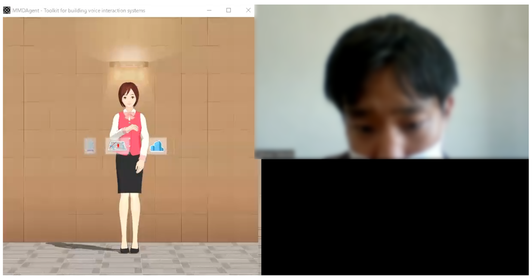

Figure 2: *Video data collected during experiment's dialog sessions (dialog agent on the left, experimental participant on the right).*

of the words included in the response candidates with the words contained in previous system utterances, and excluded candidates whose maximum similarity scores were greater than 0.20. The generated reply utterance text was then sent to the interactive agent (MMDAgent [6]), and the dialog system's response was delivered as synthesized speech output. Zoom shared the dialog agent's movements and speech with the experimental participants via the participant's monitor screen.

2.2. Recording of Dialog

The dialog between the experiment participant and the dialog system was recorded during the dialog sessions using the recording function of Zoom (Fig. 2). We also recorded the text of the speech recognition results for the participants' utterances, the system's utterances, and the time of each utterance as dialog logs. In addition, after each dialog session was completed, the participants labeled each system utterance as either normal ("non-breakdown") or abnormal ("breakdown"). The annotation standard for a "breakdown" was if the participant felt uncomfortable, or if they felt the system's utterance was inappropriate. Otherwise, responses were labeled "non-breakdown".

During the experiment, we collected 1,085 pairs of system utterances and user response utterances. The average number of utterances pairs per dialog session was 11.0. The total breakdown rate was 24.3% (264 utterance pairs), while the average breakdown rate per participant ranged from 0.0% for the user with the fewest breakdowns to 63.0% for the user with the most breakdowns, thus there was significant variation in the breakdown rate among the participants. In other words, for each dialog session in which one experiment participant did not identify any breakdowns, another experiment participant felt there had been breakdowns more than half the time, indicating that sensitivity to breakdowns, and thus the likelihood of experiencing a breakdown, differed greatly among participants.

2.3. Questionnaire to Assess Personality Characteristics

After completing their three dialog sessions, a questionnaire was administered to assess the personality traits of each of the experiment's participants. The BigFive personality traits [7][1] were used to represent the personalities of the participants in this study. We utilized the Japanese version of the Ten Item Personality Inventory (TIPI-J) [8] questionnaire, which measures the properties of BigFive personality traits using 10 self-assessment statements. For each personality trait, participants were given a score on a scale from 2 to 14, which was then used to place them in the upper or lower group for that trait.

3. Extracting Features

In order to analyze participant reactions to dialog breakdowns, we extracted multimodal features from the recorded dialog session data after dialog breakdowns. These features included language features, acoustic features, and visual features. The timing of this data extraction is shown in Fig. 3.

3.1. Language Features

Language features were extracted from the user utterance that immediately followed a "broken" utterance made by the dialog system. First, the user's utterance was decomposed into morphologies by the Japanese morphological analyzer MeCab[2], and ratios of the use of each part of speech were obtained. Because interjections and conjunctions are frequently used when a speaker attempts to switch topics or expresses emotional reac-

[1] BigFive captures the holistic architecture of personality in five dimensions; Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.

[2] https://taku910.github.io/mecab/

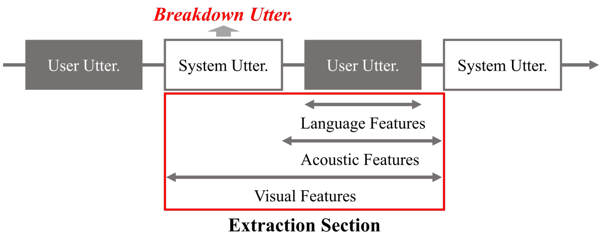

Figure 3: *Features extraction timing.*

tions after a dialog breakdown, there may be an increase in the number of conjunctions or interjections to the breakdown. Thus we selected interjections and conjunctions as part-of-speech features.

3.2. Acoustic Features

Acoustic features were extracted from participant speech segments during the interval from the end of the dialog system's breakdown utterance to the beginning of the system's following utterance, using OpenSMILE [9]. Speech characteristics were selected from the emobase2010 set, and loudness, frequency, jitter, and shimmer were used. We also selected the mean of each feature for statistical purposes. Our objective was to capture changes in prosody representing user emotion (anger, surprise, amusement, etc.) in response to encountering a dialog breakdown.

3.3. Visual Features

Visual features of each participant's facial expressions and head movements were extracted from the recorded videos using Zoom functions. Video was extracted from the start of the system's breakdown utterance to the start of the following system utterance. OpenFace [10] was used to extract the visual features. ActionUnit (AU) features and head movement features were extracted for each video frame. AU is a unit of action used for describing facial expressions, a feature adopted in the Facial Action Coding System (FACS) proposed by P. Ekman and W.V. Friesen [11]. In this study, we used AU2 (raising the outer eyebrows), AU4 (lowering the eyebrows), AU6 (raising the cheeks), and AU12 (raising the lips and corners of the mouth). The previous study shows that these AUs effectively detect dialog breakdown [3]. AU2 was found to be correlated with surprise, AU4 with anger and disgust, and AU6 and AU12 with joy. In addition, the standard deviation of the pitch and yow of the head of the entire facial frame were obtained as features representing movement of the head. This was done in reference to a previous study [2] in which head movement was used to detect dialog breakdowns.

4. Result and Discussion

In this section we analyze associations between the multimodal features described in Section 3 and the individual traits of participants described in Section 2.3 during dialog breakdowns. Note that in this study, we treat both gender and personality traits as individual characteristics possibly associated with differences in user responses when encountering dialog breakdowns.

First, standardization was carried out for each participant for all data, including data collected under the non-breakdown condition (i.e., normal dialog). This was done to include individual differences in the selected features which occurred during normal dialog (i.e., dialog without discomfort), in order to more accurately identify differences which only occurred during breakdowns. Based on the results of their personality assessments, participants were divided into upper and lower groups for each personality trait (e.g., more Open or less Open), which along with gender were scored in relation to the multimodal features exhibited after breakdowns in the dialog, using the Mann-Whitney U-test. We used the 50th percentile for the division of the two groups.

Table 1 shows our U-test results, while Table 2 shows the mean values of the multimodal data features for each group (upper group or lower group). Table 2 only show features in which significant differences and significant trends are noted.

4.1. Language Features

There were significant differences between most of the individual differences (gender and all of the BigFive personality traits except for Openness) and the use of conjunctions. Conjunctions are often used for paraphrasing or switching topics. Thus, it is inferred that when the system broke down, the experimental participant with certain personality traits coped with the breakdown by switching to another topic using the conjunction. This result suggests that the dialog strategy chosen by the experiment's participants after dialog system breakdowns differed depending on the participant's individual traits. Furthermore, significance differences or significance tendencies were confirmed between the use of interjections and gender, Agreeableness, and Openness, with men tending to use interjections more frequently than women. Similarly, the lower Agreeableness group used interjections more frequently than the upper Agreeableness group, and the upper Openness group tended to use interjections more often than the lower Openness group. Women and people with high Agreeableness tended not to use emotional language in response to system breakdowns. We can also infer that people with high Openness tend to respond to system breakdowns by using emotional vocabulary, possibly because they are being playful.

4.2. Acoustic Features

Regarding acoustic features, significant trends were identified between shimmer and Neuroticism. We found that the upper Neuroticism group had higher shimmer scores. Previous studies have reported that vocal shimmer becomes elevated when speakers are experiencing stress [12]. Since individuals with a tendency to neuroticism often have angry characters, it can be inferred that they find dialog system breakdowns stressful, resulting in an elevation in vocal shimmer. Individuals with a high propensity for neuroticism may also be more prone to anxiety and discomfort when the system does not understand their speech. Therefore, dialog systems should apologize to highly neurotic users, bow, etc., to reduce user stress when dialog breakdowns occur during conversations.

4.3. Visual Features

Significant trends were identified between ActionUnit trait AU4, representing anger and disgust, and Openness. Therefore, curious people may be less likely feel discomfort after dialog breakdowns, since the upper Openness group had less AU4 presentation. Significant differences were also identified between AU6, expressing pleasure, and gender, since women

Table 1: *P-values of the U-test between individual characteristics and multimodal features. P-values below 0.00004 are indicated as .0000.*

Features		Gender		Extroversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
Language	Conjunctions	**.0000**	***	**.0000**	***	**.0000**	***	**.0000**	***	**.0009**	***	.7937	n.s.
	Interjections	**.0037**	**	.2558	n.s.	**.0592**	+	.1026	n.s.	.1801	n.s.	**.0068**	**
Acoustic	loudness	.4503	n.s.	.9197	n.s.	.5395	n.s.	.6081	n.s.	.6816	n.s.	.8581	n.s.
	F0	.3437	n.s.	.5521	n.s.	.4353	n.s.	.6289	n.s.	.4134	n.s.	.6666	n.s.
	Jitter	.8420	n.s.	.6358	n.s.	.5734	n.s.	.8652	n.s.	.2042	n.s.	.6362	n.s.
	Shimmer	.4237	n.s.	.8307	n.s.	.4090	n.s.	.7676	n.s.	**.0955**	+	.1133	n.s.
Visual	AU2	.8557	n.s.	.3047	n.s.	.8637	n.s.	.3837	n.s.	.8444	n.s.	.6160	n.s.
	AU4	.5728	n.s.	.4687	n.s.	.4447	n.s.	.6575	n.s.	.1258	n.s.	**.0945**	+
	AU6	**.0405**	*	.7616	n.s.	.5762	n.s.	.8864	n.s.	.9461	n.s.	.1744	n.s.
	AU12	.1164	n.s.	.7506	n.s.	.8020	n.s.	.5561	n.s.	.7098	n.s.	.1585	n.s.
	Head pitch	.9004	n.s.	.7481	n.s.	.6302	n.s.	.1320	n.s.	.9461	n.s.	.6149	n.s.
	Head yaw	.4359	n.s.	.5910	n.s.	.5901	n.s.	.9447	n.s.	.7338	n.s.	.1032	n.s.

+: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, n.s.: No significant difference

Table 2: *Mean values of the multimodal features for each group.*

Features	Individual Characteristics	Upper/ Women mean values	Lower/ Men mean values
Conjunctions	Gender	-0.051	0.080
	Extroversion	0.051	0.013
	Agreeableness	0.021	0.050
	Conscientiousness	-0.034	0.079
	Neuroticism	0.066	-0.034
Interjections	Gender	0.028	0.125
	Agreeableness	0.065	0.134
	Openness	0.109	0.074
Shimmer	Neuroticism	0.175	-0.017
AU4	Openness	-0.060	0.125
AU6	Gender	0.356	0.116

tended to show more amusement than men when encountering system breakdowns.

No significant differences were observed between features related to head movement and any of the individual characteristics examined, although individual-specific variations in the frequency of head movements were reported during breakdown detections by older adults in another study [2]. While our study only used college students as experiment participants, the age of dialog system users may be another individual characteristic responsible for differences in user responses to breakdowns.

5. Conclusions

In this study we have analyzed the responses of experimental participants to dialog breakdowns from the viewpoint of their individual characteristics. As a result of our analysis of the relationships between selected multimodal features (language usage, speech acoustics and visual expressions) and the selected individual characteristics (gender and BigFive personality traits), it became clear which individual characteristics were responsible for the individual differences we observed in user responses to dialog breakdowns, namely gender, Openness, Neuroticism, and Agreeability.

In this study, only university students are participating in the experiment. Therefore, in the future, we will need to extend the object of the experiment participants and conduct similar verification for ages other than university students. Moreover, we would like to explore methods of improving dialog breakdown detection accuracy by taking into consideration the individual characteristics of system users.

6. References

[1] R. Higashinaka, L. F. D' Haro, B. Abu Shawar, R. E. Banchs, K. Funakoshi, M. Inaba, Y. Tsunomori, T. Takahashi, and J. Sedoc, "Overview of the Dialogue Breakdown Detection Challenge 4," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, ser. Lecture Notes in Electrical Engineering, E. Marchi, S. M. Siniscalchi, S. Cumani, V. M. Salerno, and H. Li, Eds. Singapore: Springer, 2021, pp. 403–417.

[2] A. Kawamoto, K. Wada, T. Kitamura, and K. Kuroki, "Preliminary Study on Detection of Behavioral Features at Conversation Breakdown in Human-Robot Interactions," in *2019 IEEE/SICE International Symposium on System Integration (SII)*. Paris, France: IEEE, Jan. 2019, pp. 375–378.

[3] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures," in *Proceedings of the 2021 International Conference on Multimodal Interaction*. Montréal

QC Canada: ACM, Oct. 2021, pp. 112–120. [Online]. Available: https://dl.acm.org/doi/10.1145/3462244.3479887

[4] K. Tsubokura, Y. Iribe, and N. Kitaoka, “Dialog breakdown detection using multimodal features for non-task-oriented dialog systems,,” in *Proceedings of the IEEE GCCE 2022*, Oct. 2022.

[5] H. Sugiyama, M. Mizukami, T. Arimoto, H. Narimatsu, Y. Chiba, H. Nakajima, and T. Meguro, “Empirical analysis of training strategies of transformer-based japanese chit-chat systems,” *arXiv:2109.05217*, 2021.

[6] A. Lee, K. Oura, and K. Tokuda, “Mmdagent—a fully open-source toolkit for voice interaction systems,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8382–8385.

[7] L. R. Goldberg, “An alternative” description of personality”: the big-five factor structure.” *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[8] A. Oshio, S. Abe, and P. Cutrone, “Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j) (in japanese),” *The Japanese Journal of Personality*, vol. 21, no. 1, pp. 40–52, 2012.

[9] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM ’10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[11] P. Ekman and W. V. Friesen, “Facial action coding system,” *Environmental Psychology & Nonverbal Behavior*, 1978.

[12] E. Mendoza and G. Carballo, “Acoustic analysis of induced vocal stress by means of cognitive workload tasks.” *Journal of voice : official journal of the Voice Foundation*, vol. 12 3, pp. 263–73, 1998.