Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5759–5767
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Collection and Analysis of Travel Agency Task Dialogues with Age-Diverse Speakers

Michimasa Inaba[1], Yuya Chiba[2], Ryuichiro Higashinaka[3], Kazunori Komatani[4], Yusuke Miyao[5], Takayuki Nagai[6]
[1]Artificial Intelligence Exploration Research Center, The University of Electro-Communications
[2]NTT Communication Science Laboratories, NTT Corporation
[3]Graduate School of Informatics, Nagoya University
[4]SANKEN, Osaka University
[5]Graduate School of Information Science and Technology, The University of Tokyo
[6]Graduate School of Engineering Science, Osaka University
m-inaba@uec.ac.jp, yuuya.chiba.ax@hco.ntt.co.jp, higashinaka@i.nagoya-u.ac.jp,
komatani@sanken.osaka-u.ac.jp, yusuke@is.s.u-tokyo.ac.jp, nagai@sys.es.osaka-u.ac.jp

Abstract

When individuals communicate with each other, they use different vocabulary, speaking speed, facial expressions, and body language depending on the people they talk to. This paper focuses on the speaker's age as a factor that affects the change in communication. We collected a multimodal dialogue corpus with a wide range of speaker ages. As a dialogue task, we focus on travel, which interests people of all ages, and we set up a task based on a tourism consultation between an operator and a customer at a travel agency. This paper provides details of the dialogue task, the collection procedure and annotations, and the analysis on the characteristics of the dialogues and facial expressions focusing on the age of the speakers. Results of the analysis suggest that the adult speakers have more independent opinions, the older speakers more frequently express their opinions frequently compared with other age groups, and the operators expressed a smile more frequently to the minor speakers.

Keywords: multimodal dialogue, dialogue corpus, dialogue act, facial expression

1. Introduction

Task-oriented dialogue systems have long been an active area of research in dialogue systems (Bobrow et al., 1977; Zue et al., 1991; Raux et al., 2005; Budzianowski et al., 2018). Recently, deep neural networks have been successfully applied to response generation (Wen et al., 2017; Chen et al., 2019; Zhang et al., 2020) and dialogue state tracking (Mrkšić et al., 2017; Zhong et al., 2018; Chen et al., 2020). These studies mainly focus on generating an appropriate response to users' inputs.

When individuals communicate with others, they use different vocabulary, speaking speed, facial expressions, and body language depending on the people they communicate with. For example, with children, speakers may use simpler vocabulary and speak with more emotion, while with elderly individuals, speakers may speak more slowly. In contrast, current dialogue systems rarely change their speaking style or dialogue strategy according to the user. We believe that dialogue systems should change their dialogue strategies according to the user to accomplish their tasks more efficiently and increase user satisfaction.

A complex of factors such as gender, social relationships, and roles significantly affect communication. However, we focused on the speaker ' s age as a factor that dramatically affects the change in communication because speakers of various ages are relatively easy to recruit and age is one of the most important factors among them. We collected a multimodal dialogue corpus with a wide range of speaker ages from children to the elderly (see Figure 1).

Figure 1: Multimodal dialogue corpus with a wide range of speaker ages (left: operator, right: customer). Customers are minors (upper right), adults (middle right), and older adults (lower right).

As a dialogue task, we focus on travel, which interests people of all ages, and we set up a task based on

a tourism consultation between an operator and a customer at a travel agency. The operator was able to use the tourist information retrieval system to obtain information about tourist spots during the dialogue. We also collected the system's log, such as queries, retrieval results, and corresponding timestamps. As with the Wizard of Wikipedia dataset (Dinan et al., 2019), our data may help construct dialogue systems that access external resources during interactions.
Features of our corpus are following:

- Wide range of age speakers, from 7 to 72 years old
- Over 115 hours of large multimodal dialogue data in Japanese
- Contains queries by the operator and outputs by the system, which are associated with the dialogue
- Manually transcribed with the subset of the ISO 24617-2 dialogue acts annotated

This paper describes the details of the dialogue task, the methods and results of the corpus collection, and the results of the analysis of the collected corpus regarding the dialogue phase transition and the customers' facial expressions to obtain effective interaction strategy according to the user's age for constructing dialogue systems.

2. Related Work

Several multimodal dialogue corpora between two speakers have been collected to analyze human interactions, facial expressions, emotions, and gestures. The Cardiff Conversation Database (CCDb) (Aubrey et al., 2013) contains audio-visual natural conversation with no role (listener or speaker) and no scenario. Some of the data were annotated with dialogue acts such as Backchannel and Agree, emotions such as Surprise and Happy, and head movements such as Head Nodding and Head Tilt. Each conversation lasted five minutes and includes a total of 300 minutes of dialogue, with participants ranging in age from 25 to 57 years old. The Emotional Dyadic Motion CAPture (IEMOCAP) dataset (Busso et al., 2008) has been collected for communication and gesture analysis. The actors were wearing markers on their faces, heads, and hands, and two types of dialogues were collected, performing improvisations and scripted scenarios. The utterances are annotated with emotion labels. The total recording time is approximately 12 hours. The NoXi corpus (Cafaro et al., 2017) contains dialogues mainly in English, French, and German, annotated with head movements, smiles, gazes, engagement, etc. The total recording time was approximately 25 hours, and the age of the participants ranged from 21 to 50 years old. In this study, we collected a total of over 115 hours of data, which is larger than all the two-party multimodal dialogue corpus mentioned above. The age range of the speakers in our data is also wider than in previous studies.
Multimodal corpora containing conversations between multiple people have also been collected. Belfast storytelling dataset (McKeown et al., 2015), the AMI meeting corpus (Carletta, 2007), ICSI meeting corpus (Janin et al., 2003), Computers in Human Interaction Loop (CHIL) (Waibel et al., 2005) and Video Analysis and Content Extraction (VACE) (Chen et al., 2005) are well-known examples. The above corpus of both two-person and multi-person dialogues mainly consists of speakers in their 20s to 50s, however, and does not include children or older speakers.
Some monologue corpora include speakers of a broader range of ages. The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) (Zadeh et al., 2018) is one of the largest corpora, containing 23,500 YouTube videos by 1,000 people, each utterance annotated with an emotion label. In addition, several monologue corpora have been shared in the research community, for example, the Multimodal Corpus of Sentiment Intensity (CMU-MOSI) (Zadeh et al., 2016), the ICT Multi-Modal Movie Opinion (ICT-MMMO) (Wöllmer et al., 2013), and the Multimodal Opinion Utterances Dataset (MOUD) (Pérez-Rosas et al., 2013). Because these datasets are relatively large, they include minor and older speakers; nevertheless, they are not dialogue corpora.

3. Travel Agency Task

We collected dialogues between two speakers, one playing the role of an operator and the other playing the role of a customer, simulating a tourist consultation at a travel agency. The two speakers make a video call using Zoom[1], and each conversation lasts 20 minutes. The customer conceives a travel situation before the dialogue (described in section 3.2) and then consults with the operator to decide tourist spots based on the situation. The operator elicits requests from the customer and recommends tourist spots by using information obtained from the tourist information retrieval system (described in section 3.1).

3.1. Tourist Information Retrieval System

We have developed a tourist information retrieval system using the Rurubu data[2] provided by JTB Publishing Corporation. Rurubu is one of the most famous travel guidebook series in Japan. Rurubu data contain approximately 45,000 Japanese sightseeing spots.
A screenshot of the system is shown in Figure 2. The operator can specify search queries on the left side of the system screen, such as region and area, free keyword search, genre-based search (e.g., "See – Buildings – Historical Sites – Historical Buildings", "Eat – Foreign Cuisine – French Cuisine"), and budget. The

[1] https://zoom.us/
[2] https://solution.jtbpublishing.co.jp/service/domestic/

Figure 2: Tourist information retrieval system. The operator retrieves tourist spot information from the system and provides it to the customer.

right side of the screen shows the search results, including descriptions, maps, images, addresses, and access information.

To make the operator speakers use the information obtained from the information retrieval system, we instructed them not to provide information on tourist spots based on their memory as much as possible but instead to use information using the system's search results. We collected time stamped input query logs and output data from the system during dialogues.

3.2. Dialogue Scenario

Before the dialogue, the customers specify some situations of the trip they are planning, keeping in mind their personal relationships and the destinations they actually want to visit.

We adopted two types of dialogue scenarios with different types of situations. In Dialogue Scenario 1, the customer defines a specific situation, and the speaker then determines the destination (prefecture or region in Japan), season (spring, summer, fall, winter), number of people, and relationships (friends, family, etc.) as the situation. The customer decides three destinations they want to visit during the dialogue. In Scenario 2, the customer describes briefly what kind of trip they want to take, providing a short description of activities or outings they are looking for. Examples of the situation descriptions include "I want to relax in a hot spring resort," or "I want to visit shrines and temples during the day and eat local specialties at night." During the dialogue, the customer decides at least one destination.

4. Data Collection

4.1. Recording

We recorded dialogues from November 10, 2020, to February 25, 2021 using Zoom's local recording function. We collected an mp4 video file, an m4a audio file, and two m4a separate audio files for the operator and customer per interaction. The data collection procedure has been approved by the ethics committee for experiments on human subjects, the University of Electro-Communications (No. 19061(2)).

No.	Screen mode	Scenario type
1	Gallery view	1
2	Gallery view	1
3	Gallery view	2
4	Screen sharing	1
5	Screen sharing	1
6	Screen sharing	2

Table 1: Dialogue order, screen mode, and scenario in the recording

The speakers interacted in two ways: by looking at each other's faces and by using screen sharing. In the former setting, the speaker's faces are displayed on the left and right sides of the screen (gallery view), and the speakers interact by looking at the other person's face (Figure 1). In this condition, only the operator speaker can see the screen of the tourist information retrieval system. In the latter dialogues using screen sharing, both customer and operator speaker can see the system screen using the screen sharing function of Zoom. There are two reasons for having a shared screen: (1) screen sharing has become common in recent years in video calls and (2) the speakers can communicate smoothly by viewing the same screen.

Each customer interacted three times through gallery view and three times through screen-sharing for a total of six dialogues, and Table 1 shows the order of these dialogues. Dialogue Scenario 1 is used twice, and Dialogue Scenario 2 is used once for each screen setting.

4.2. Speakers

For the customers, we employed 55 people: 20 minors, 25 adults, and 10 older adults and have provided a breakdown of customers per age in Figure 3. Minors participated in this data collection with the consent of their parents. As described in the previous section, each speaker had six dialogues, so the number of dialogues collected was 330 (55×6).

For the operator speakers, we employed five people, three of whom had experience working at a travel agency (age and gender are as follows: 36/male, 41/female and 57/female), and the remaining two had experience in customer service (35/male and 27/male). The three travel agency experienced speakers handled 78.2% of the total dialogues (258 out of 330).

4.3. Annotation

4.3.1. Dialogue Act Tags

The collected dialogues were manually transcribed and annotated with dialogue act (DA) tags via crowdsourcing, and a subset of the ISO 24617-2 annotation scheme (the first edition) (Bunt et al., 2017) was used as the DA

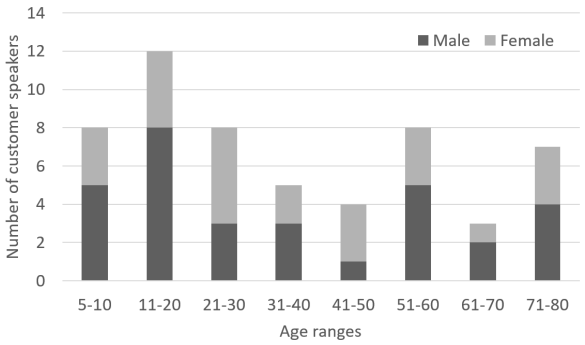

Figure 3: Age and gender distribution of customers

Inform	AddressSuggest
Agreement	AcceptSuggest
Disagreement	DeclineSuggest
Correction	Instruct
Answer	AutoPositive
Confirm	AutoNegative
Disconfirm	AlloPositive
Question	AlloNegative
SetQuestion	FeedbackElicitation
PropositionalQuestion	Stalling
ChoiceQuestion	Pausing
CheckQuestion	InitGreeting
Offer	ReturnGreeting
AddressOffer	InitSelfIntroduction
AcceptOffer	ReturnSelfIntroduction
DeclineOffer	Apology
Promise	AcceptApology
Request	Thanking
AddressRequest	AcceptThanking
AcceptRequest	InitGoodbye
DeclineRequest	ReturnGoodbye
Suggest	Other

Table 2: Dialogue act tags

tag set. In the ISO 24617-2 annotation scheme, tags are classified into nine dimensions, and tags of different dimensions can be annotated in duplicate. To allow only one tag to be assigned to a segment, we excluded tags in the following dimensions: Turn Management (six tags), Own Communication Management (three tags), Partner Communication Management (two tags), and Discourse Structuring (one tag). These tags can be annotated in duplicate for a single interval if these dimensions are different, and are rarely annotated independently. Table 2 shows the 44 tags we used. The "Other" tag in the table is not an ISO 24617-2 tag and is mainly assigned to inaudible sections during transcription. Note that our data can be compliant with ISO 24617-2 if we additionally annotate tags in the excluded dimensions.

4.3.2. Annotation Procedure

The DA tag in ISO 24617-2 assigns tags not to entire utterances (turns) but to the functional segments. Functional segments are defined as the minimal stretches of communicative behavior that have a communicative function (Bunt et al., 2017).

We adopted a two-stage annotation approach: (1) segmenting the utterances into functional segments and (2) assigning DA tags to the functional segments; this approach allows us to evaluate the consistency of the segmentation and tagging, respectively.

4.3.3. Preliminary Experiments for Consistency

To confirm the consistency of the segmentation and tagging, we conducted experiments that two annotators individually performed the annotation on randomly selected dialogues.

We conducted a functional segmentation experiment on 2,109 utterances from 10 dialogues and calculated the perfect matching rate and partial matching rate. The partial matching occurs when two segmentation results are perfectly matched if either one is divided into smaller segments. For example, we have two segmentation results, "AA BB / CC DD" and "AA / BB / CC DD" (where "/" means the division separator). If we split "AA BB" into "AA / BB" in the first segmentation, it will match the second segmentation and thus be a partial match. Conversely, "AA / BB CC / DD" and "AA BB / CC DD" are not a partial match because even if only one of them is split into smaller segments, it will not match the other. The results of the segmentation experiment showed that the perfect matching rate was 0.833 and that the partial matching rate was 0.935, indicating a high degree of agreement.

Two workers annotated the DA tags to 9,220 functional segments from 10 dialogues, and the experimental results showed that Cohen's κ was 0.632, indicating good agreement. Note that (Bunt et al., 2017) reported the agreement of the ISO 24617-2 tag annotations. They only calculated the κ for each dimension, however, so an equal comparison is difficult. For reference, the kappa values of the dimensions corresponding to the tags used in their paper are 0.21 to 0.58 (the lowest is the Auto-Feedback dimension consisting of AutoPositive and AutoNegative, and the highest in the Time Management dimension consisting of Staling and Pausing), which are lower than our results.

Because the above results showed that segmentation and tagging were consistently annotated, we assigned one person per dialogue to conduct the segmentation and tagging for all the remaining dialogues.

4.4. Data Statistics

Table 3 shows statistics of our corpus. Although we set the duration of each dialogue as 20 minutes, because the dialogues were not automatically terminated by time, their duration is approximately 5% longer than $6,600 (= 330 \times 20)$ minutes.

Dialogues	330
- Minor customers (7 to 18 years old)	120
- Adult customers (20 to 60 years old)	150
- Older customers (65 to 72 years old)	60
Duration (minutes)	6,948
Utterances (turns)	111,771
- Operator utterances	66,594
- Customer utterances	45,177
Tags (functional segments)	246,316

Table 3: Corpus statistics

An example of dialogue and annotations for the child customer is shown in Table 4, and the older customer is shown in Table 5. From the tables, we can confirm that the operator's speech style (for non-Japanese speakers: honorific expressions are used for the older customer, while not for the child customer), length of each turn, and topic selection differ greatly depending on the age of the customer.

5. Analysis of Dialogue Act Sequence

In the following sections, we analyze the verbal and non-verbal information to see how the speaker's behavior differs between age groups in travel agency task dialogue. One of the most popular methods to visualize the time structure of sequential data is the Hidden Markov Model (HMM) (Meguro et al., 2009). For the verbal information, we employ HMM using sequences of DA, and discuss the difference of dialogue transition among age groups. Additionally, we focus on facial expressions for the analysis of non-verbal information.

5.1. Experimental Conditions

We used the 330 whole dialogues with DA tags for the analysis of the verbal information. The experimental data contains 120 dialogues of minor customers, 150 dialogues of adult customers, and 60 dialogues of older customers. We used the `hmmlearn` package[3] of Python to train the HMM. The DA tags of both speakers were treated as different labels, and the labels of "Stalling," "Pausing," and "Other" were excluded because these labels have little impact on the dialogue content. The number of DA tags used for the analysis was 82 (two speakers $\times$ 41 tags). By our observation, typical travel agency task dialogues consist of three phases as below:

1) The initial phase, which includes greetings and asking requirements

2) The middle phase where an operator gives information and a customer gives feedback

3) The final phase, which includes confirmation of the travel plan and farewell greetings

In this study, the structure of HMM is designed to capture the transition of these phases. The initial and final phases were modeled by a single state with a self-loop. In contrast, we constructed the ergodic HMM that has transitions from any one state to any other for the middle phase because this phase is assumed to have complex transitions that are different by age group. We consider that complex exchanges of the DA can be modeled to some extent in the second state although all dialogues do not always follow the assumed three phases. We will provide a more detailed analysis of DA transitions in a future study. For training HMMs, the number of states in the middle phase was changed from one to five, and the definitive model was decided based on the minimum description length. We trained the HMMs for three customer groups: minor, adult, and older customers; the model was a discrete HMM with an output probability distribution of the multinomial distribution.

[3]https://hmmlearn.readthedocs.io/en/latest/

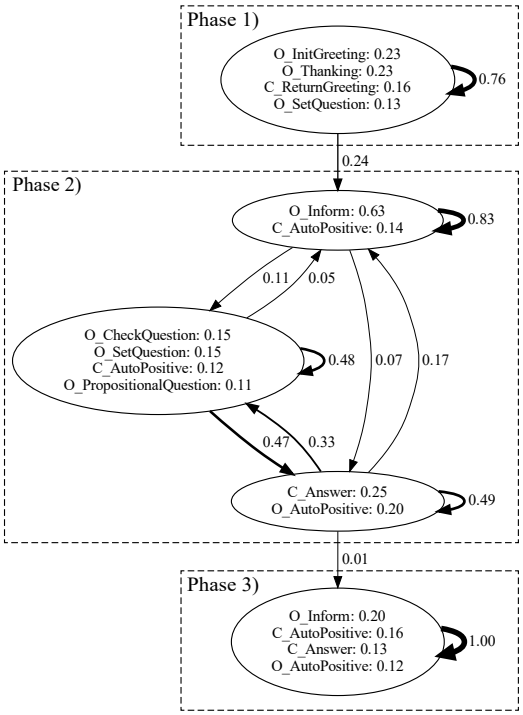

Figure 4: HMM for dialogue with minor customers: The alphabet before the tag name represents speakers. "O" is the operator, and "C" is the customer.

5.2. Experimental Results

Figures 4, 5, and 6 show the HMMs obtained by fitting the DA sequences, and in the figures, the alphabet before the tag name represents speakers. "O" is the operator, and "C" is the customer. To clarify the characteristics of the models, we only show the tag names of the output probability greater than 0.10. The thickness of the edge is proportional to the value of the transition probability. From the figures, we find that the model is slightly different among the age groups.

Speaker	Start	End	Functional Segments	Dialogue act tag
Operator	02:47.80	02:56.26	んとねー、(Well,)	Stalling
			市場の周りとかにー、お寿司とか食べられるお店がたくさんあるんだけどー。 (there are a lot of places around the market where you can eat sushi.)	Inform
Customer	02:57.01	02:57.30	はい。 (Yes.)	AutoPositive
Operator	02:58.09	03:00.58	どんなお寿司が食べたいとかってある？ (What kind of sushi do you want to eat?)	PropositionalQuestion
Customer	03:01.53	03:05.80	えっと、(Uh,)	Stalling
			いくら丼みたいにいくらがたっぷーりあるお寿司屋さん。(a sushi restaurant has a lots of salmon roe, like a bowl of salmon roe.)	Answer
Operator	03:06.34	03:08.70	あ、(Oh,)	Stalling
			いくらがたくさん乗ってるお寿司屋さん？ (a sushi restaurant has a rice bowl with lots of salmon roe on it, don't you?)	CheckQuestion
Customer	03:09.33	03:09.69	はい。 (Yes.)	Answer
Operator	03:10.00	03:11.78	うん、(Okay,)	AutoPositive
			じゃ、それを探してみるね。(I'll try to find one.)	Inform
Customer	03:12.52	03:12.80	はい。 (Yes.)	AutoPositive

Table 4: Dialogue Example (Child Customer)

Speaker	Start	End	Functional Segments	Dialogue act tag
Operator	10:19.08	10:21.41	やっぱり歩くのはつらいですかね。 (Is it hard for you to walk?)	PropositionalQuestion
Customer	10:22.16	10:29.58	うーん、 (Yeah,)	AutoPositive
			やっぱり歳と共にちょっと膝もきてるのであまり、若いころは山に登ったりするのも好きだったけど。 (I used to love climbing mountains when I was younger, but my knees are getting worse with age.)	Answer
Operator	10:29.20	10:47.56	あー、(Ah,)	Stalling
			色々やはり (there are various)	CheckQuestion
			あの、 (uh,)	Stalling
			歌舞伎屋根の家とかやっぱあるんですけども、いかがですかね、(types of houses with kabuki roofs, what do you think?)	CheckQuestion
			なんか (And)	Stalling
			ミュージアムみたいなところもあるんですけども、そういった古い (there is also a place like a museum,)	Inform
			あの (uh,)	Stalling
			家並みを展示しているみたいなんですけども。 (it seems to exhibit these old houses.)	Inform
Customer	10:33.74	10:48.87	ええ、ええ、(Yes, yes.)	AutoPositive
			ミュージアム、今みんな近くなんでしょ、そういったの。(The museum, they're all close by, aren't they?)	CheckQuestion
Operator	10:48.93	10:49.47	はい。 (Yes.)	Answer
Customer	10:51.16	10:55.29	近いんならもう少し位、足を伸ばしても大丈夫かもしれないですね。(If it's close to there, I might be able to go a little further.)	Inform

Table 5: Dialogue Example (Older Customer)

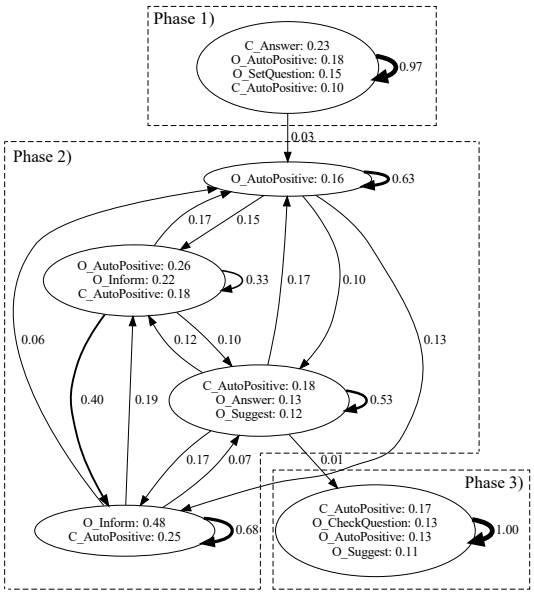

Figure 5: HMM for dialogue with adult customers: The alphabet before the tag name represents speakers. "O" is the operator, and "C" is the customer.

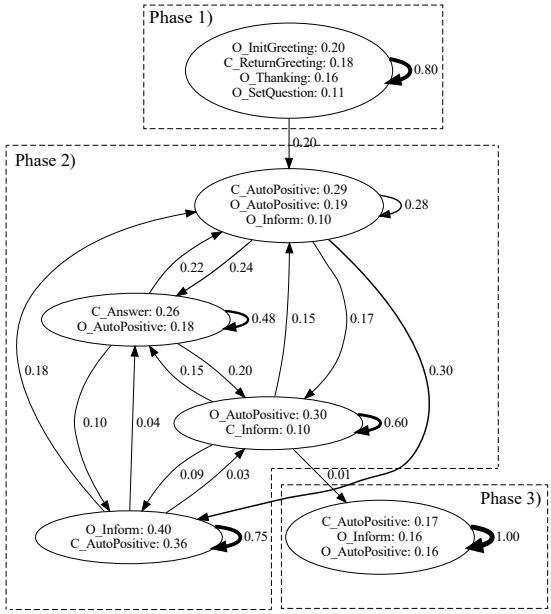

Figure 6: HMM for dialogue with older customers: The alphabet before the tag name represents speakers. "O" is the operator, and "C" is the customer.

In the middle phase of the dialogue with minor customers, the structure is simple: The operator gives information and the customer responds to it with AutoPositive. Additionally, the operator attempts to get a more detailed opinion from the customer by asking CheckQuestion and PropositionalQuestion.

The model of the dialogue with adult customers is slightly complex, but two states in the middle phase represent a typical dialogue transition in the target task; that is, the operator gives the information ("O_Inform"), and the customer responds to it ("C_AutoPositive"). One of the features is a state with a high output probability of the operator's suggestion ("O_Suggest"). This result suggests that adult customers have more independent opinions than minor customers, and the operators provide them more constructive guidance.

In the model of dialogue with older customers, two states in the middle phase have the high probability of "O_Inform" and "C_AutoPositive"; it is the same with the model for dialogue with adult customers. In contrast, it is characteristic that there is a state in which the customer gives information ("C_Inform"), and the operator responds to it ("O_AutoPositive"). This result indicates that the older customers tend to talk about their motivations and wishes for the trip compared with other age groups.

The above analysis suggests that the characteristics of the dialogues differ depending on the age of the customer. Therefore, a travel guidance dialogue system should change its dialogue strategy according to the user's age to obtain high user satisfaction.

AU	Comparison	Diff.	p-value
AU06	Minor−Adult	0.382	$< 0.001^{***}$
	Minor−Older	0.350	0.011^{*}
	Adult−Older	−0.031	1.000
AU12	Minor−Adult	0.507	$< 0.001^{***}$
	Minor−Older	0.423	0.023^{*}
	Adult−Older	−0.084	1.000

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 6: Results of multiple comparison tests for operator's facial expression.

6. Analysis of Facial Expression

We next investigated the difference in the operator's facial expression between customers' age groups; we cropped regions of the operator's face from the video images and extracted facial action units (AUs) using OpenFace (Baltrušaitis et al., 2016). The facial regions were too small to extract AUs in half of the data because the dialogues were conducted while sharing the screen, so we used 165 dialogues recorded with gallery view for the analysis.

The AUs were obtained frame by frame, and they were averaged over the dialogue to compare between age groups. We conducted a one-way layout ANOVA factoring the age group and performed multiple comparison tests for AUs that showed significant differences. Table 6 shows the results of multiple comparison tests for AU06 (Cheek Riser) and AU12 (Lip Corner Puller), which exhibited a significant difference from ANOVA.

These AUs become high when the speaker expresses a smile.

As shown in the table, significant differences were observed between the minor group and the adult group for AU06 and AU12. For both AUs, the values were larger in the minor group. These results reflect that the operator expressed a smile more frequently to the minor customer. The system should therefore adapt not only verbal but also non-verbal dialogue behavior to the user's age to provide natural travel guidance.

7. Conclusion

This paper describes our multimodal dialogue corpus with a wide range of speakers' ages from children to the elderly. This corpus consists of 330 dialogues of 20 minutes each, recorded using Zoom and is one of the largest corpora of two-person dialogues. The dialogue task is based on the consultation of tourist spots at a travel agency between a speaker playing the role of an operator and a customer. The dialogues were manually transcribed, and the DAs were annotated using a subset of the ISO 24617-2 annotation scheme.

Using the corpus, we analyzed the sequences of dialogue acts and the facial expressions focusing on the age of the speakers. The experimental results indicated that the characteristics of dialogue and facial expressions differed depending on the age of the speakers. In a future study, we will investigate the influence of the customer's age on the operator's speech activity.

Currently, our corpus is only available to the research organizations participating in our joint research project, the Communicative Intelligent Systems Towards a Human-Machine Symbiotic Society[4]; however, we are working on making the corpus accessible to the public.

8. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19H05692. The authors would like to thank Koki Washio for useful discussions.

9. Bibliographical References

Aubrey, A. J., Marshall, D., Rosin, P. L., Vendeventer, J., Cunningham, D. W., and Wallraven, C. (2013). Cardiff conversation database (CCDb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pages 1–10.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

[4]https://www.commu-ai.org/

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M. (2018). Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. (2017). Dialogue act annotation with the iso 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., André, E., and Valstar, M. (2017). The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M., et al. (2005). Vace multimodal meeting corpus. In *Proceedings of the 2005 International Workshop on Machine Learning for Multimodal Interaction*, pages 40–51.

Chen, W., Chen, J., Qin, P., Yan, X., and Wang, W. Y. (2019). Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709.

Chen, L., Lv, B., Wang, C., Zhu, S., Tan, B., and Yu, K. (2020). Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7521–7528.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–364–I–367.

McKeown, G., Curran, W., Wagner, J., Lingenfelser, F., and André, E. (2015). The belfast storytelling database: A spontaneous social interaction database with laughter focused annotation. In *Proceedings*

of the 2015 International Conference on Affective Computing and Intelligent Interaction, pages 166–172.

Meguro, T., Higashinaka, R., Dohsaka, K., Minami, Y., and Isozaki, H. (2009). Analysis of listening-oriented dialogue for building listening agents. In *Proceedings of SIGDIAL*, pages 124–127.

Mrkšić, N., Séaghdha, D. Ó., Wen, T.-H., Thomson, B., and Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. (2013). Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Raux, A., Langner, B., Bohus, D., Black, A. W., and Eskenazi, M. (2005). Let's go public! taking a spoken dialog system to the real world. In *Proceedings of the Interspeech*.

Waibel, A., Steusloff, H., and Stiefelhagen, R. (2005). CHIL: Computers in the human interaction loop. In *Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services*.

Wen, T., Vandyke, D., Mrkšíc, N., Gašíc, M., Rojas-Barahona, L., Su, P., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 438–449.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. (2013). Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. (2016). MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Zhang, Y., Ou, Z., and Yu, Z. (2020). Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Zhong, V., Xiong, C., and Socher, R. (2018). Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467.

Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. (1991). Integration of speech recognition and natural language processing in the mit voyager system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 713–716.