IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 12, No. 1, March 2023, pp. 12∼22
ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i1.pp12-22

Information-gathering dialog system using acoustic features and user's motivation

Ryota Togai, Takashi Tsunakawa, Masafumi Nishida, Masafumi Nishimura
Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University, Hamamatsu, Japan

Article Info

Article history:

Received Nov 14, 2021
Revised Sep 3, 2022
Accepted Sep 22, 2022

Keywords:

Chat dialog system
Information-gathering dialog system
Nonverbal acoustic features
Talking motivation
Topic induction

ABSTRACT

Recently, as society continues to age, automation of watching elderly people who live apart from their families has been gradually expected. However, we must prevent them from losing their purpose in life, which declines due to lack of communication. Thus, a chat dialog system has attracted widespread attention as a method that achieves both problems: keeping their purpose in life and watching their daily lives. Unlike a task-oriented dialog system, a chat dialog system has explicitly no task to accomplish and makes a conversation to continue communication with the users. Keeping a conversation is essential for elderly people who are mostly unfamiliar with digital devices. Moreover, conversing daily on the chat dialog system provides the opportunity to collect information for their care. This study realizes an information-gathering dialog system, a chat dialog system that collects healthcare information of elderly people. Furthermore, we use the nonverbal acoustic features from their speech, since automatic speech recognition is not necessarily accurate in current systems. This paper illustrates the effectiveness of two important elements, topic change for keeping the talking user motivated with the dialog system and motivation estimation, for attaining an information-gathering dialog system using nonverbal acoustic features.

Corresponding Author:

Takashi Tsunakawa
Department of Computer Science, Faculty of Informatics, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011 Japan
Email: tuna@inf.shizuoka.ac.jp

1. INTRODUCTION

Recently, society of many developed countries has noticed an increased aging population, leading to the automation of watching elderly people who live apart from their families [1], [2]. These elderly people living away from their families lack communication and eventually lose something to live for. To solve this problem, a chat dialog system has attracted widespread attention as a method for keeping their purpose in life and watching their daily lives [3], [4]. By having the elderly people talk with the chat dialog system daily, the system helps increase chances of communication and obtain the necessary information for monitoring the elderly people by asking important and timely questions. Studies over the years have reported methods for gathering information from users through dialog systems. In information-gathering dialog systems, Kobayashi *et al.* [3] highlighted the difficulty for users to answer questions if the dialog system sequentially asks questions to be answered without considering the context of the dialog. They proposed a method using the chain structure of dialogs, which gradually shifts dialog topics to follow the dialog context and ask the questions to be answered. Such a topic shifting is called topic induction. Nagasaka *et al.* [5] used WordNet to build a topic induction model that shifts a current topic to the specified one in chat dialogs, to automate questions on

dementia in chat dialogs with elderly people. Yoshito *et al.* [6] aimed to build an active information-gathering dialog system, and created a model that determines the user's intention to end the dialog from nonverbal acoustic information to avoid the system asking persistent questions. Ishihara *et al.* [7] performed the interviewee's dialog willingness estimation to calculate questioning strategies in real-time for an information-gathering dialog robot. Previously, studies have considered dialog management, including topic induction algorithms, using features based mainly on linguistic information. Meguro *et al.* [8] have employed partially observable markov decision process (POMDP), which is a statistical dialog management method that sets rewards for actions in a probabilistically determined state transition structure, and executes actions that maximize the rewards that can be obtained in the future. Lison [9] have developed a dialog system that combines statistical dialog management and rule-based dialog management, which is available as an open source software [10]. However, it has not been sufficiently studied on appropriate timing and destination of topic induction considering nonverbal acoustic information.

However, simply talking without any consideration does not automate monitoring. For example, asking many questions to collect a lot of information ends up like a questionnaire session with a chat dialog system, which would not be appreciated by the elderly. On the other hand, by pursuing only the naturalness of the dialog, the system fails to ask the questions required of a monitoring chat dialog system. Additionally, one major challenge is topic transition. In information-gathering dialogs, the problem is reflected in how to efficiently move from the current topic to a new topic for the system to talk about. Humans recognize mental distances between topics in conversations, and feel uncomfortable when conversation suddenly moves to a distant topic or stays on near topics for so long. To solve these problems, estimating the user's talking motivation on the current topic is essential. By understanding the user's talking motivation, the system decides when to ask appropriate questions and which topic to move to. To make users continue talking with dialog systems daily, appropriately switching topics to talk is necessary. To achieve this, the dialog system judges whether it changes the current topic according to user's talking motivation or topic interest. Yokoyama *et al.* [11] developed a chat dialog system that switches the system's role to "listener" and "speaker" depending on the user's interest.

Previous studies on estimating users' talking motivation have used facial images, voice, and linguistic information of the user. Schuller *et al.* [12] studied to estimate the user's interest in current topics from multimodal information of facial expressions, nonverbal acoustic information, and verbal information obtained from a single speech of the user. Chiba *et al.* [13] automatically estimated talking motivation from multimodal information to build an interview dialog system. Saito *et al.* [14] estimated users' attitudes toward dialog from multimodal information in dialog data with dementia patients. Many previous studies have estimated the user's talking motivation using multimodal information. However, when one uses the dialog system, simultaneously capturing the user's facial expressions with cameras or performing complete speech recognition to acquire linguistic information is difficult. Since the dialog state changes gradually through multiple turns, efficiently learning the information of multiple turns is necessary. In dialog-state tracking challenge (DSTC) [15], a shared task that analyzes dialog using information from multiple turns, methods using recurrent neural network (RNN) has shown high performance [16]. In these methods, using the linguistic information of the user's speech as input, the probability distribution of tasks, user's requests, and so on are estimated as dialog states. A dialog-state tracking method using long short term memory (LSTM), which improves the drawback of RNNs with difficulty storing long-term information, has been proposed [17]. We consider that the dialog-state tracking is very similar to the task of measuring user's talking motivation, since the motivation can be regarded as a kind of dialog states. We apply this dialog-state tracking method to track the user's talking motivation using RNN with nonverbal acoustic information as the user input.

In this study, we experiment by measuring the degree of user satisfaction when the Wizard of Oz system [18], [19] switches topics according to the user's estimated talking motivation with the current topic. In addition, we focus on introducing nonverbal acoustic information for estimating the talking motivation. In human-human dialog, various nonverbal information such as prosody and facial expressions is also frequently used. Hence, such information has been considered important as an input to the dialog system [20], [21]. We analyze the relationship between nonverbal acoustic information and the talking motivation to be estimated.

2. THE PROPOSED METHOD

To collect information from users by asking questions during a chat dialog, question timing and topic transition must be adjusted appropriately. This section proposes two hypotheses about topic induction from the

topic space model. This section also verifies them using the Wizard of Oz.

2.1. Modeling of the topic space

We describe the topic space model proposed for realizing a dialog system with topic induction. Suppose the dialog system suddenly switches from the current topic to a mentally distant topic, the user will feel that the system skips from topic to topic. However, if the system repeatedly talks about similar topics, the user gets bored and the satisfaction of the dialog decreases. For a user to enjoy a chatting dialog system for a long time, the system must switch to distant or near topics at the right time. Therefore, it is important to model the topic space representing the mental distance among topics.

We model the topic space with a two-dimensional undirected graph structure reflecting mental distance among topics referring to Nagasaka *et al.*'s work [5]. Figure 1 shows an example of modeling the topic space. Each node represents a topic, and topics connected by an edge are mutually transitable. The length of an edge represents the mental distance between the topics. For users to feel naturally induced by these topics, gradually moving from the current topic to the goal topic in the topic space is essential.

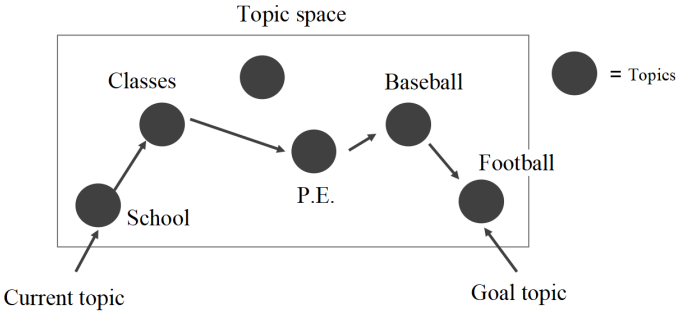


Figure 1. Example of modeling a topic space

Here, we model the topic space using WordNet [22] and Word2vec [23]. WordNet is a tree representation of the conceptual structure of words. Furthermore, the WordNet-based distance between concepts can be obtained by following the shortest path between nodes in the tree. suppose the distance between concepts is roughly consistent with the human mental scale. Then, we can apply this to the topic space model. Word2vec is a model representing words as vectors and obtains the similarity between two words by calculating the cosine of their vectors. The cosine similarity is negatively correlated with the mental distance between two words and can be used for a topic space. The Word2vec-based distance is the value of subtracting the Word2vec similarity between the keywords that indicate the topic from 1. We used Japanese Wikipedia as a training corpus to obtain word vectors by Word2vec.

2.2. Topic induction and user satisfaction

The straightforward way to obtain the required information is to directly ask questions about the information. However, as Kobayashi *et al.* [3] stated, this reduces user satisfaction. Therefore, to simultaneously maximize both user satisfaction and the amount of information obtained by the dialog system, we find the time at which user satisfaction does not decrease even if the topic is changed to ones the user asks a question about once. We formulated the following hypothesis about topic induction, referring to human interaction.

Hypothesis 1 *When the user's talking motivation with the current topic is low, switching to a distant topic does not decrease the user's satisfaction.*

In human-human conversation, if the person we talk to seems to enjoy the current topic, we delve deeper into the topic, otherwise, we change the topic to a different one to explore the person's talking motivation. If the same dialog strategy can be used for dialog systems, it would be possible to continue dialog without lowering user motivation by choosing topics close in the conceptual distance when the user's motivation for dialog is high and switching to farther topics otherwise. Also, we consider another hypothesis:

Hypothesis 2 *The user's talking motivation is correlated with features of nonverbal acoustic information, such as loudness and length of the user's speech.*

This is also supported by human-human conversation; loud voice and/or long speech of the person can indicate more motivation to talk about the current topic, whereas smaller voice and/or shorter replies show little motivation. Also, we estimate the user's talking motivation from features of nonverbal acoustic information.

3. METHOD

Our proposed model is based on two hypotheses described in the previous section. Here, we verify these hypotheses and the effectiveness of our topic induction strategy for collecting information by two experiments. One is estimating user's talking motivation, and the other is topic switching Wizard of Oz experiment for analyzing the talking motivation, topic distance, and user satisfaction.

3.1. Experiment 1: estimating user's talking motivation

First, we focus on showing the appropriateness of the hypothesis 2. To analyze and estimate users' talking motivation, we collected spoken dialog data with recorded talking motivation at each turn. The user talks with the dialog system through the microphone of the smartphone. The voice during the dialog is recorded using the microphone of a smartphone with a sampling frequency of 16 kHz and a quantization bit of 16 bits. Following the previous study [3], the system talks only one topic in one session and takes 20 turns as either a listener or a speaker. The system as a listener only asks questions to the user, and the system as a speaker only discloses itself to the user. The system employed the use of fixed scenarios based on fixed topics for speech, and no questions from the user were allowed. The user records his/her current talking motivation on a 7-point scale from -3 to 3 for each turn during the dialog. The first turn of the dialog is set to 3 because we assume that the user actively begins to talk with the dialog system. Five-topic scenarios were prepared for the system to talk about including computers, cooking, fashion, travel, and music. This follows the literature in [13] so that the level of users' interests would be distributed. The change in talking motivation depends on the level of interest in the topic. Therefore, the user's level of interest was recorded in each topic on a 5-point scale from -2 to 2 upon completing the dialog. To conduct each session independently, one session was held per day, and six subjects were asked to talk with the dialog system at home for ten days. From this experiment, audio data were obtained from 60 sessions with six subjects acting as listeners and speakers, respectively, for five topics.

3.2. Experiment 2: talking motivation and user satisfaction

To analyze the relationship between the user's talking motivation and the conceptual distance between topics, we conducted a Wizard of Oz dialog experiment with the topic switched according to the user's motivation for dialog. During the dialog, subjects inputted their motivation to talk about the current topic at each turn of the dialog in 11 levels: 0, 10, 20, ..., 100. The greater value showed higher motivation. The Wizard switched the topic every four turns according to users' talking motivation. Thus, for the two dialog sessions, each with a 10 min duration, the experiment for each of the 10 subjects is as:

- Session A: a session in which the system chooses a distant topic when the user's talking interest is 50 or more, and a closer topic otherwise.
- Session B :a session in which the system chooses a closer topic when the user's talking interest is 50 or more, and a distant topic otherwise.

The distance between topics is measured using the Wizard's mental scale. After the dialog, the subjects rated their satisfaction on a 7-point scale from -3 to 3. A higher value showed a higher level of satisfaction.

4. RESULTS AND DISCUSSION

4.1. Experiment 1: estimating user's talking motivation

4.1.1. User's interest in the topic

We analyze the effects of "user's interest in the topic" and "the role of the system as a listener or a speaker" among the factors considered influencing the user's talking motivation. Figure 2 shows a scatter plot of the slope of the change in user's talking motivation and the user's level of interest in the topic. Here, the slope of the change in the user's talking motivation is obtained from the slope of the linear regression calculated for the series of user's talking motivation for 20 turns. The distribution of the plots in the scatter plot is right-shouldered, and the slope of the regression line is positive, indicating that higher level of user's interest in the current topic positively affects, increases, or keeping user's talking motivation.

4.1.2. Role of the dialog system

Next, we analyzed the transition of a user's talking motivation depending on whether the system plays the role of a listener or a speaker. Figure 3 shows the average slope of the change in user's talking motivation for each user and system role. The error bars represent the standard deviation. From the figure, the slope of the change in users' talking motivation is negative for nearly all users, indicating that their talking motivation decreased as the dialog progressed. Thus, the role of the system as a listener or a talker had no significant effect on the user's talking motivation.

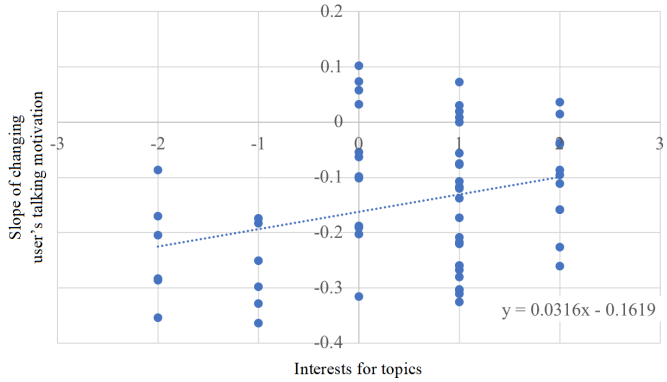


Figure 2. Scatter plot of interests for topics and slope of changing user's talking motivation

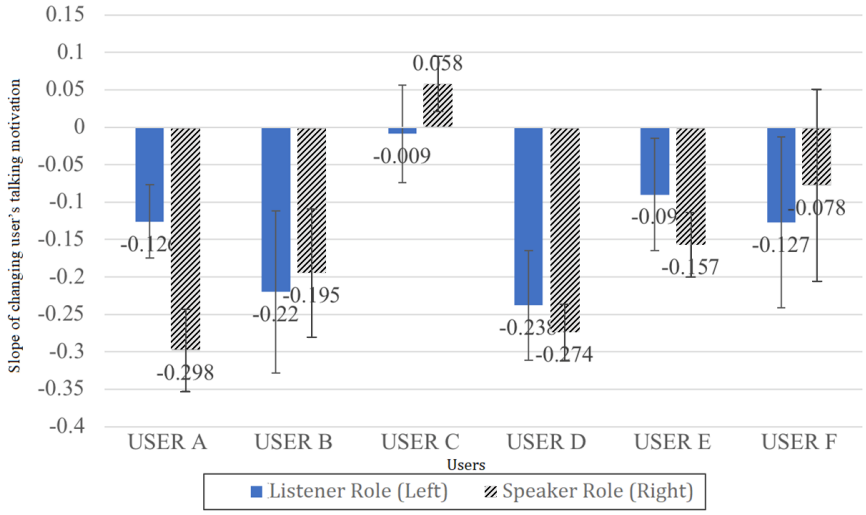


Figure 3. System roles and slope of changing user's motivation of each user

4.1.3. Nonverbal acoustic information

Estimating the user's talking motivation from factors, such as the role of the dialog system is difficult. Even by collecting each user's interesting topics, it is still difficult to estimate the current user's talking motivation due to the low correlation in Figure 2. To more directly estimate the user's talking motivation, we employ nonverbal acoustic information obtained from the user's speech. First, we deleted the silence before and after each turn of the audio data obtained from the user's dialog. Next, we extracted 384 features that can be extracted using openSMILE [24] IS09 emotion challenge configuration [25], which adds features of speech length, articulation rate, and delay. The delay feature encompasses the time from the end of the system speech to the beginning of user speech.

The nonverbal acoustic information extracted from the user speech during a dialog is highly dependent on the content of the speech, which significantly changes in a single turn. However, since the user's talking motivation does not change significantly from one turn to the next, the features for estimating the user's talking motivation become values that gradually change. Therefore, the extracted nonverbal acoustic information was smoothed by taking a five-point moving average in the turn direction for each session. This implies that the feature value of a given turn was the average of five turns, including both turns before and after the corresponding nonverbal acoustic information.

Also, we analyzed the correlation between the extracted nonverbal acoustic information and the user's talking motivation to investigate which nonverbal acoustic information is effective for the estimation [26]. Table 1 shows the top 10 features in the absolute value of the correlation coefficient. Among the nonverbal acoustic information, the correlation coefficient for the most strongly correlated feature was 0.311. No acoustic feature with a strong correlation was applied to all users. Furthermore, results showed that many mel-frequency cepstral coefficients (MFCC) ranges appeared in the top 10 features. Since the correlation coefficient is positive, the range of MFCCs became smaller as the user's talking motivation decreased.

Table 1. Top 10 features correlating with user's motivation

Feature	Correlation coefficient
(cf.) Turn number in session	-0.628
Voice rate	0.311
MFCC 8-dim. stddev	0.277
MFCC 8-dim. linregQ	0.260
Prob. of voice amean	0.257
Volume amean	0.248
MFCC 6-dim. range	0.236
MFCC 1-dim. range	0.227
MFCC 9-dim. stddev	0.227
MFCC 9-dim. linregQ	0.223

Figure 4 shows the maximum absolute value of the correlation coefficient calculated for each user. The maximum correlation coefficient exceeded 0.5 for many users, indicating that a correlation between nonverbal acoustic information and users' talking motivation exists. This result shows that there are individual differences in the relationship between nonverbal acoustic information and the user's talking motivation. Furthermore, it indicates that we can create a model with high accuracy by creating an individual estimation model for each user.

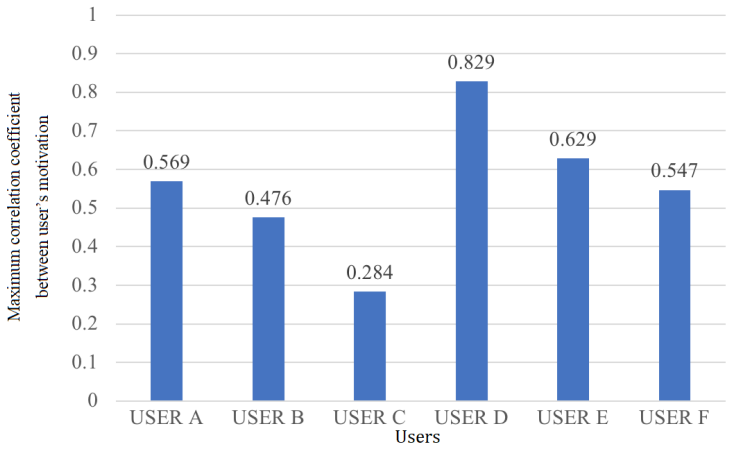


Figure 4. Maximum correlation coefficient between user's motivation for each user and acoustic features

4.1.4. Estimating user's talking motivation

This section compares the following three methods for estimating dialog motivation using nonverbal acoustic information of multiple turns:

– NN1 :a neural network (NN) that performs estimation using only features from one turn.
– NN3 :a NN that performs estimation using information from three previous turns.
– LSTM3 :an LSTM with a window size of three previous turns.

Each of them is evaluated using the mean absolute error (MAE) with a 10-point cross-validation.

Figure 5 shows the estimation accuracy of NN when only turn information is used as features and when nonverbal acoustic features are combined. The nonverbal acoustic features used were those of the top 20 correlations with users' talking motivation. The estimation error with nonverbal acoustic information was 0.451 lesser in MAE than that without nonverbal acoustic information, indicating that nonverbal acoustic information is an effective feature in estimating the user's talking motivation.

Figure 6 compares the error in estimating the user's talking motivation among the three estimation methods (here, the turn information not included in the nonverbal acoustic information is not used). The blue and orange bars show the results for the top 20 and top 300 correlated features, respectively. From Figure 6, for both the top 20 and top 300 features, the estimation error for using multiple turns of information was smaller than using only one turn of information. Also, the error was smallest when using LSTM. This indicates that the information from multiple turns is effective for the estimation and that LSTM reduces the estimation error.

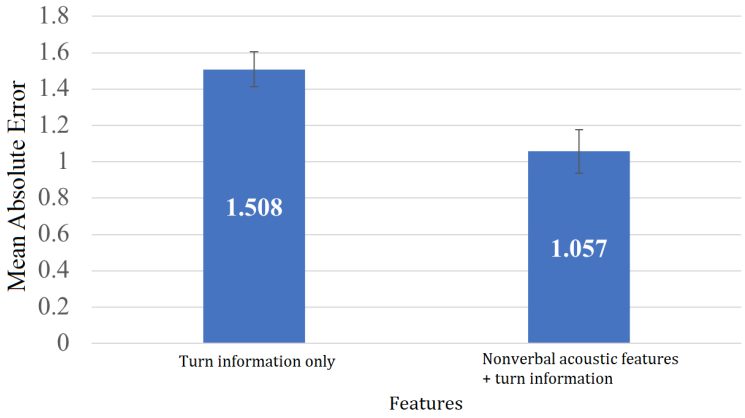


Figure 5. Comparison of estimation errors with and without nonverbal acoustic features

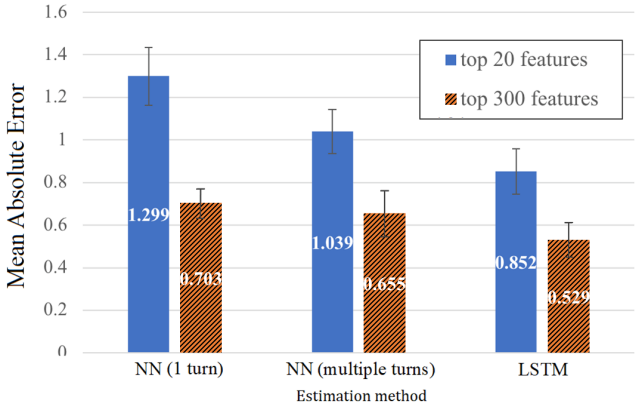


Figure 6. Comparison of estimation errors among estimation methods

4.2. Experiment 2: talking motivation and user satisfaction

4.2.1. Topic distance

This section checks whether the wizard chooses between distant and near topics according to the human mental scale. The relationship between the subject's talking motivation at the timing of topic switching and the conceptual distance between the previous and following topics is shown in Figure 7. In session A, the higher the subject's talking motivation, the more distant the wizard chose the topic, thereby creating a right-shouldered regression line. However, session B has the opposite strategy and has seen a steady increase. Therefore, sessions A and B have data that conformed to the conditions for topic selection, as shown in hypothesis 1. However, the slope of the regression line is not large, confirming the gap between the conceptual distance of WordNet and the human mental scale.

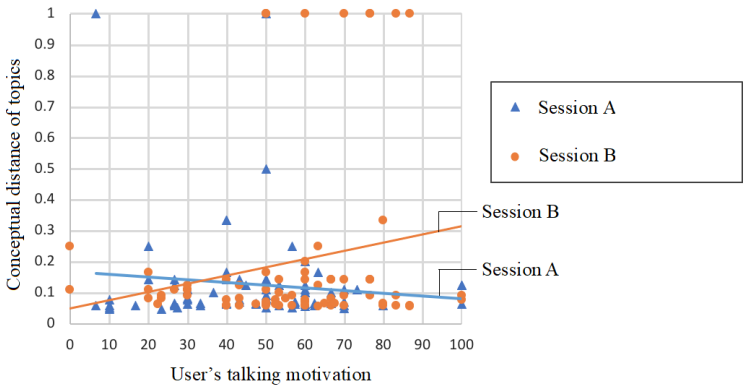


Figure 7. Correlation between user's talking motivation and conceptual distance of topics (calculated using WordNet)

Figure 8 shows the scatterplot relationship between the talking motivation and conceptual distance between topics. Here, the conceptual distance is calculated using Word2vec trained from Japanese Wikipedia. The results showed that the slope was larger than that of Figure 7 and that the conceptual distance when modeling the topic space can be modeled closer to the human mental scale using Word2vec.

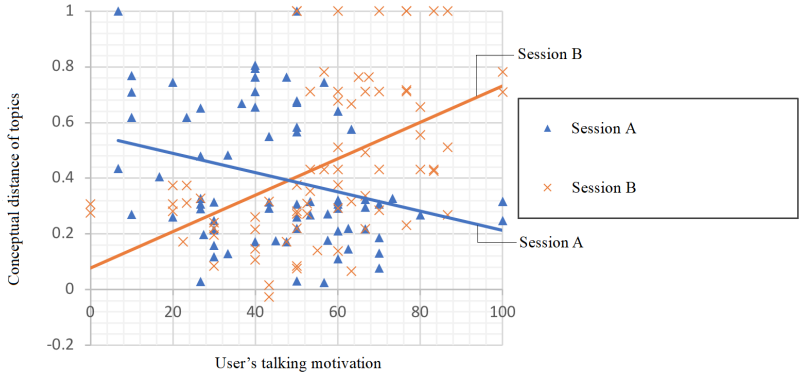


Figure 8. Correlation between users' motivation and concept distance of topics (calculated using Word2vec)

4.2.2. User satisfaction

The results of user satisfaction are shown in Figure 9. The average score was 1.0 higher in Session B than in Session A. Despite variations in the scale for each user's satisfaction, the results for each user show that most users were more satisfied in Session B.

4.2.3. Effectiveness of nonverbal acoustic information

To verify hypothesis 2, we analyzed the relationship between users' talking motivation and nonverbal acoustic information. Simple nonverbal acoustic features were extracted from user speech during dialog and the correlation between the average value of nonverbal acoustic features for four turns before the topic switched and the user's talking motivation on the topic switch was calculated. The correlation values between the nonverbal acoustic features and the user's talking motivation are shown in Table 2. A certain degree of correlation was confirmed for speech length and fundamental frequency, demonstrating hypothesis 2. However, this information is still insufficient to control the timing of switching topics. In the future, we will consider methods, such as combining multiple features to make decisions of switching topics.

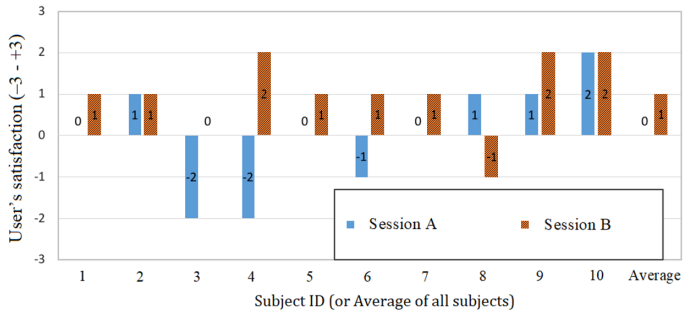


Figure 9. Evaluation results of user's satisfaction

Table 2. Correlation between acoustic features and user's motivation

Features	Correlation
Time from ending of the system's speech to beginning user's speech	0.036
Average volume of the speech interval	-0.044
Speech length	0.29
Tone ratio	-0.14
Fundamental frequency	0.37

5. CONCLUSION

In this paper, we proposed a topic induction method using users' talking motivation to automatically estimate the user's talking motivation from nonverbal acoustic information, to improve the efficiency of gathering information by a chat dialog system. In the automatic estimation of the user's talking motivation, results showed that the user's talking motivation varied depending on the interest level in the current topic, correlating to several nonverbal acoustic information. Additionally, we compared the estimation error among several estimation methods and confirmed the error reduction using the information of multiple turns. In the proposed topic induction method, the user's talking motivation is used as input, and a dialog experiment with a dialog system that transitions from the current topic to either a near or far topic is conducted using the Wizard of Oz method. Thus, the system that transitions to a topic close to the current topic when the user's talking motivation is high, and a far topic otherwise, recorded higher user satisfaction. Furthermore, the user's talking motivation was weakly correlated with the nonverbal acoustic information obtained from the user's speech. In the future, it will be necessary to automatically estimate the user's talking motivation using nonverbal acoustic information from multiple turns and to verify such estimation using an automated system that switches topics toward high user satisfaction.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP16K01543.

REFERENCES

[1] I. Azimi, A. M. Rahmani, P. Liljeberg, and H. Tenhunen, "Internet of things for remote elderly monitoring: a study from user-centered perspective," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, pp. 273–289, 2017. doi: 10.1007/s12652-016-0387-y.

[2] S. Majumder, E. Aghayi, M. Noferesti, H. Memarzadeh-Tehran, T. Mondal, Z. Pang, and M. J. Deen, "Smart homes for elderly healthcarerecent advances and research challenges," *Sensors*, vol. 17, no. 11, p. 2496, 2017. doi: 10.3390/s17112496.

[3] Y. Kobayashi, D. Yamamoto, T. Koga, S. Yokoyama, and M. Doi, "Design targeting voice interface robot capable of active listening," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 161–162. doi: 10.1109/HRI.2010.5453214.

[4] Y. Sakai, Y. Nonaka, K. Yasuda, and Y. I. Nakano, "Listener agent for elderly people with dementia," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI)*, 2012, pp. 199–200. doi: 10.1145/2157689.2157754.

[5] H. Nagasaka, H. Kawanaka, K. Yamamoto, K. Suzuki, H. Takase, and S. Tsuruoka, "A study on topic control method for robot-assisted therapies dementia evaluation using simple conversation with robots ," *Transactions of Japanese Society for Medical and Biological Engineering*, vol. 51 Supplement, pp. R–262, 2013. doi: 10.11239/jsmbe.51.R-262.

[6] O. Yoshito, M. Makoto, and K. Shirai, "Construction of decision model for the spoken dialogue system to close communication," *IPSJ SIG Technical Report*, 2011-HCI-142(2), pp. 1–8, 2011.

[7] T. Ishihara, K. Nitta, F. Nagasawa, and S. Okada, "Estimating interviewee's willingness in multimodal human robot interview interaction," in *Proceedings of the 20th International Conference on Multimodal Interaction (ICMI)*, 2018, pp. 1–6. doi: 10.1145/3281151.3281153.

[8] T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka, "Controlling listening-oriented dialogue using partially observable markov decision processes," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2010, vol. 2, pp. 761–769. doi: 10.1145/2513145.

[9] P. Lison, "A hybrid approach to dialogue management based on probabilistic rules," *Computer Speech & Language*, vol. 34, no. 1, pp. 232–255, 2015. doi: 10.1016/j.csl.2015.01.001.

[10] P. Lison and C. Kennington, "OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules," in *54th Annual Meeting of the Association for Computational Linguistics (ACL) 2016 - System Demonstrations*, 2016, pp. 67–72. doi: 10.18653/v1/P16-4012.

[11] S. Yokoyama, D. Yamamoto, Y. Kobayashi, and M. Doi, "Development of dialogue interface for elderly people-switching the topic presenting mode and the attentive listening mode to keep chatting," in *IPSJ SIG Technical Report*, vol. 2010-SLP-80, no. 4, pp. 1–6, 2010.

[12] B. Schuller, R. Müller, and B. Hörnler, A. Höethker, H. Konosu, G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, 2007, pp. 30–37. doi: 10.1145/1322192.1322201.

[13] Y. Chiba, T. Nose, and A. Ito, "Analysis of efficient multimodal features for estimating user's willingness to talk: comparison of human-machine and human-human dialog," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 428–431. doi: doi: 10.1109/APSIPA.2017.8282069.

[14] N. Saito, S. Okada, K. Nitta, Y. I. Nakano, and Y. Hayashi, "Estimating user's attitude in multimodal conversational system for elderly people with dementia," in *2015 AAAI spring symposium series*, 2015, vol. SS-15-07, pp. 100–103.

[15] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *SIGDIAL 2014 - 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2014, pp. 263–272. doi: 10.3115/v1/W14-4337.

[16] M. Henderson, B. Thomson, and S. Young, "Word-based dialog state tracking with recurrent neural network," in *SIGDIAL 2014 - 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2014, pp. 292–299. doi: 10.3115/v1/W14-4340.

[17] K. Yoshino, T. Hiraoka, G. Neubig, and S. Nakamura, "Dialogue state tracking using long short term memory neural networks," in *Proceedings of Seventh International Workshop on Spoken Dialog Systems*, 2016, pp.1-8.

[18] N. M. Fraser and N. Gilbert, "Simulating speech systems," *Computer Speech & Language*, vol. 5, no. 1, pp. 81–99, 1991. doi: 10.1016/0885-2308(91)90019-M.

[19] M. Okamoto, Y. Yang, and T. Ishida, "Wizard of Oz method for learning dialog agents," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2001, vol. 2180, pp. 20–25. doi: 10.1007/3-540-44799-7_3.

[20] S. Fujie, D. Yagi, Y. Matsusaka, H. Kikuchi, and T. Kobayashi, "Spoken dialogue system using prosody as paralinguistic information," in *Proceedings of Speech Prosody*, 2004, pp. 387–390.

[21] T. Ohsuga, M. Nishida, Y. Horiuchi, and A. Ichikawa, "Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue," in *Proceedings of Interspeech*, 2005, pp. 33–36. doi: 10.21437/Interspeech.2005-32.

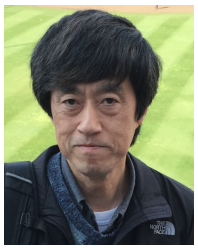
[22] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, "Development of Japanese WordNet," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.
[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.
[24] 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*, 2010, pp. 1459–1462. doi: 10.1145/1873951.1874246.
[25] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of Interspeech*, 2009, pp. 312–315. doi: 10.21437/Interspeech.2009-103.
[26] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comp. Sci., University of Waikato, Hamilton, NewZealand, 1999.

BIOGRAPHIES OF AUTHORS

Ryota Togai was a master-course student at department of informatics, Graduate School of Integrated Science and Technology, Shizuoka University. He obtained bachelor's degree in informatics from Shizuoka University in 2016, and master's degree in Informatics from Shizuoka University in 2018. He can be contacted at email: togairyota@gmail.com.

Takashi Tsunakawa is a lecturer at the faculty of informatics, Shizuoka University from 2019. He obtained a master's degree and Ph.D in information science and technology at the University of Tokyo in 2005 and 2010, respectively. He was a project researcher at the University of Tokyo, a scientific researcher, and an assistant professor at Shizuoka University. His researches are in the fields of natural language processing, especially in machine translation, dialog systems, and assistance in education. He is affiliated with ACL, the association for natural language processing in Japan, the information processing society of Japan, and the Japanese society for artificial intelligence. Besides, he is also involved in some groups including a SIG for patent translation. Further info on his homepage: https://wwp.shizuoka.ac.jp/tsunakawa/. He can be contacted at email: tuna@inf.shizuoka.ac.jp.

Masafumi Nishida is an associate professor at the faculty of informatics since 2015, Shizuoka University. He completed a Ph.D. in engineering at Ryukoku University in 2002. He was an assistant professor at Chiba University, an associate professor at Doshisha University, and a designated associate professor at Nagoya University. His researches are in the fields of speech information processing, behavior signal processing, and well-being information technology. He is affiliated with the information processing society of Japan, human interface society, the institute of electronics, information and communication engineers in Japan, the acoustical society of Japan, and the Japanese society for artificial intelligence. Further info on his homepage: https://lab.inf.shizuoka.ac.jp/nisimura/Nishida.html. He can be contacted at email: nishida@inf.shizuoka.ac.jp.



Masafumi Nishimura is a professor at the faculty of informatics, Shizuoka University since 2014. He obtained a master's degree at the graduate school of engineering science, Osaka University in 1983. He obtained Ph.D. in engineering. He was engaged in research on speech-language information processing at IBM research Tokyo. He received the Yamashita memorial research award from the information society of Japan in 1998, and the technical development award from the acoustical society of Japan in 1999. His researches are in fields of speech information processing, human augmentation using sound information, assistance for the elderly and the disabled. He is affiliated with IEEE, the information processing society of Japan, the institute of electronics, information and communication engineers in Japan, the acoustical society of Japan, and the Japanese society for artificial intelligence. Further info on his homepage:https://lab.inf.shizuoka.ac.jp/nisimura/. He can be contacted at email: nisimura@inf.shizuoka.ac.jp.