Predicting Nods by using Dialogue Acts in Dialogue

Ryo Ishii, Ryuichiro Higashinaka, Junji Tomita
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, Japan
{ishii.ryo, higashinaka.ryuichiro, tomita.junji@lab.ntt.co.jp

Abstract

In addition to verbal behavior, nonverbal behavior is an important aspect for an embodied dialogue system to be able to conduct a smooth conversation with the user. Researchers have focused on automatically generating nonverbal behavior from speech and language information of dialogue systems. We propose a model to generate head nods accompanying utterance from natural language. To the best of our knowledge, previous studies generated nods from the final morphemes at the end of an utterance. In this study, we focused on dialog act information indicating the intention of an utterance and determined whether this information is effective for generating nods. First, we compiled a Japanese corpus of 24 dialogues including utterance and nod information. Next, using the corpus, we created a model that estimates whether a nod occurs during an utterance by using a morpheme at the end of a speech and dialog act. The results show that our estimation model incorporating dialog acts outperformed a model using morpheme information. The results suggest that dialog acts have the potential to be a strong predictor with which to generate nods automatically.

Keywords: nod, dialogue act, japanese dialogue

1. Introduction

Nonverbal behavior in human communication has important functions of transmitting emotions and intentions in addition to verbal behavior (BirdWhistell, 1970). This means that an embodied dialogue system should be able to express nonverbal behavior according to the utterance to communicate smoothly with the user (McBreen and Jack, 2001; Watanabe et al., 2003; Ishi et al., 2010). Against such a background, researchers have focused on constructing automatic generation models of nonverbal behavior from speech and linguistic information. Among nonverbal behaviors, nodding of the head is very important for emphasizing speech, giving and receiving speech authority, giving feedback, expressing conversational engagement, and intention of starting to speak (Senko Maynard, 1987; Senko Maynard, 1989; Ishii et al., 2015b; Ishii et al., 2017b; Ooko et al., 2011). It has been shown that nodding improves the naturalness of avatars and dialog systems and promotes conversation.

Nodding accompanying an utterance has the effect of strengthening the persuasive power of speech and making it easier for the conversational partner to understand the contents of the utterance (Lohse et al., 2014). Researchers have tried to generate nods during speaking from speech and natural language. In particular, they used several acoustic features, such as sound pressure and prosody, for generating nods (Yehia et al., 2002; KG et al., 2004; Graf et al., 2002; Busso et al., 2007; Beskow et al., 2006; Iwano et al., 1996; Ishi et al., 2010). However, it has been difficult to accurately generate nods at an appropriate time according to an utterance from only speech.

A few studies have tackled the problem of generating nods from natural language. These studies focused on the final morphemes in the phrase of an utterance and analyzed the co-occurrences with nods. They found that morphemes related to the interjections, feedback, and questionnaire and conjunctions appearing in turn-keeping (Ishi et al., 2006; Ishi et al., 2007) tend to co-occur with nods. On the basis of this information, a simple automatic nod-generation model was proposed (Ishi et al., 2010; Sakai et al., 2015). It was found that the behavior of humanoid robots and avatars that generated nods with the model gave a better impression of naturalness. It is thought that if a model that can generate nodding more accurately is constructed, it will lead to smoother communication between the dialog system and user. Therefore, a more accurate nod-generation model should be constructed by clarifying the relevance of more detailed language information and nodding. It is also known that the relevance of a speech feature to nodding and vice versa depends on the language; for instance this is weaker in Japanese (Yehia et al., 2002; Ishi et al., 2007). A detailed examination of a nod-generation model using language information is thus considered important.

In this research, we constructed a highly accurate head-nod-generation model using natural Japanese language by focusing on the dialogue acts of utterances, which has not been investigated. Dialogue acts are information indicating the intention of the speaker throughout the utterance, and it is considered that the occurrences of nods change according to the intention. Therefore, in contrast to the features of the morphemes in the final phrase that were taken into account in previous studies (Ishi et al., 2006; Ishi et al., 2007), the dialogue acts we handle include the information intended by the speaker throughout the utterance. We constructed our nod-generation model using dialogue act information of utterances and determined whether dialogue acts are useful for generating nods.

We collected a corpus consisting of 24 Japanese dialogues including utterances and head-nod information. Next, we used the corpus to create our model that estimates whether a nod occurs during an utterance by using the morpheme at the end of the speech and dialogue act. In an experiment, we found that our estimation model using dialogue act information outperformed that using morpheme information alone. We also found that a model using both dialog act information and morpheme information did not perform any

Figure 1: Photograph of two participants having dialogue

better than the model using only dialogue acts. We thus concluded that dialog acts have the potential to be a strong predictor that can be used to generate nods automatically.

2. Corpus

To collect a Japanese conversation corpus including verbal and nonverbal behaviors for generating nods in dialogue, we recorded 24 face-to-face two-person conversations (12 groups of two different people). The participants were Japanese males and females in their 20s to 50s who had never met before. They sat facing each other (Fig. 1). To gather more data on nodding accompanying utterances, we adopted the explanation of an animation participants have not seen as the conversational content. Before the dialogue, they watched a famous popular cartoon animation called "'Tom & Jerry" in which the characters do not speak. In each dialogue, one participant explained the content of the animation to the conversational partner within ten minutes. At any time during this period, the partner could freely ask questions about the content.

We recorded the participants' voices with a pin microphone attached to the chest and videoed the entire discussion. We also took bust (chest, shoulders, and head) shots of each participant (recorded at 30 Hz). In each dialogue, the data on the utterances and nodding behaviors of the person explaining the animation were collected in the first half of the ten-minute period (120 minutes in total) as follows.

- Utterances: We built an utterance unit using the inter-pausal unit (IPU) (Koiso et al., 1998). The IPU was used in the same manner as in previous studies dealing with morpheme information (Ishi et al., 2006; Ishi et al., 2007). The utterance interval was manually extracted from the speech wave. A portion of an utterance followed by 200 ms of silence was used as the unit of one utterance. We collected 2965 IPUs in total.
- Head nod: A head nod is a gesture in which the head is tilted in alternating up and down arcs along the sagittal plane. A skilled annotator annotated the nods by using bust/head and overhead views in each frame of the videos. We regarded nodding continuously in time as one nod event.
- Gaze: The participants wore a glass-type eye tracker (Tobii Glass2). The gaze target of the participants and the pupil diameter were measured at 30 Hz.

Label	Dialogue Act	Label	Dialogue Act
DA0	Greeting	DA15	Question (habit)
DA1	Provision	DA16	Question (desire)
DA2	Self-disclosure (fact)	DA17	Question (plan)
DA3	Self-disclosure (experience)	DA18	Question (evaluation)
DA4	Self-disclosure (habit)	DA19	Question (other)
DA5	Self-disclosure (positive preference)	DA20	Question (Yourself)
DA6	Self-disclosure (negative preference)	DA21	Sympathy
DA7	Self-disclosure (neutral preference)	DA22	Non-sympathy
DA8	Self-disclosure (desire)	DA23	Confirmation
DA9	Self-disclosure (plan)	DA24	Proposal
DA10	Self-disclosure (other)	DA25	Repeat
DA11	Acknowledgment	DA26	Paraphrase
DA12	Question (information)	DA27	Approval
DA13	Question (fact)	DA28	Thanks
DA14	Question (experience)	DA29	Apology
		DA30	Filler
		DA31	Admiration
		DA32	Other

Table 1: Dialogue act labels

- Hand gesture and body posture: The participants' body movements, such as hand gestures, upper body, and leg movements, were measured with a motion capture device (Xsens MVN) at 240 Hz.

All verbal and nonverbal behavior data were integrated at 30 Hz for display using the ELAN viewer (Wittenburg et al., 2006). This viewer enabled us to annotate the multimodal data frame-by-frame and observe the data intuitively. In this research, we only handled utterance and head-nod data in the corpus we constructed. Nods occurred in 1601 out of the 2965 IPUs.

3. Head-Nod-Generation Model

The goal of our research was to demonstrate that the dialogue act of an utterance is useful for generating nods. We evaluated our proposed model for estimating nods from dialogue acts and the previously constructed estimation model using the final morphemes at the end of utterance (Ishi et al., 2006; Ishi et al., 2007). As a primitive feature value, we created a feature value related to the length of an utterance and decided to treat the estimation model using this feature value as a baseline We constructed another estimation model using all information including dialogue act, final morphemes at the end of utterance, and length of utterance to evaluate the effectiveness of fusion (All model). The feature values for each IPU were as follows.

- Length of utterance (LU): Number of characters in an IPU. The feature value is a one-dimensional vector.
- Final morphemes (FM): This is binary feature as to whether morpheme injunctions related to feed-

Used Feature values	Precision	Recall	F-score
Chance level	0.500	0.500	0.500
LU	0.410	0.640	0.500
FM	0.412	0.646	0.521
DA	0.662	0.670	**0.666**
LU+FM	0.423	0.651	0.513
LU+**DA**	0.666	0.672	**0.669**
FM+**DA**	0.662	0.670	**0.666**
LU+FM+**DA**	0.666	0.672	**0.669**

Table 2: Partial excerpt of attribute weight rank in 19 features in All model.

back (e.g., "'en', "ee", "'aa", "hi", etc.) and particles related to questioning and turn-keeping (e.g., "de", "kara", "kedo", "'kana", "janai", etc.) co-occurring with the nod (as indicated in the previous studies (Ishi et al., 2006; Ishi et al., 2007) are included in the last morpheme of the IPU. Morphological analysis was conducted for this purpose. We used J-tag (Fuchi and Takagi, 1998), a general morphological analysis tool for Japanese. Based on the output result, it was judged whether the last morpheme is related to nods. The feature value is a one-dimensional vector.

- Dialogue act (DA): The dialogue act was extracted using an estimation technique for Japanese (Meguro et al., 2010; Higashinaka et al., 2014). The technique can estimate the dialogue act using the word N-grams, semantic categories (obtained from a Japanese thesaurus Goi-Taikei), and character N-grams. The dialog acts and number of IPUs are listed in Table 1. There was very little data for DA4, 10, 12, 13, 14-19, 22, 24, and 26-29, so they were excluded from feature values. The label of each dialogue act was expressed as a binary value as to whether the dialogue act appeared. Therefore, the feature value was a 17-dimensional vector.

We constructed the estimation models by using SMOreg (Keerthi et al., 2001), which implements a support vector machine (SVM) in Weka (Bouckaert et al., 2010) and evaluated the accuracy of the models and the effectiveness of each feature. The settings of the SVM — the polynomial kernel, C (cost parameter), and γ (hyper parameter of the kernel)—were determined using a grid search technique. The class was a binary value as to whether a nod occurred. We used 24-fold cross validation using a leave-one-person-out technique with the data for the 24 participants. We evaluated how well a participant's nods could be estimated with an estimator generated only from data of other people. As shown in Table 1, DA (dialogue act) had the highest performance among the models using only LU, FM, or DA, with an F-score of 0.666 ($t(23) = -9.46$, $p < .01$ in LU vs. DA; $t(23) = -9.46$, $p < .01$ in FM vs. DA). In addition, no differences in performance were found between any of the models in which other features were added to DA and that using only DA. These results suggest that DA is more useful for nod estimation than morpheme information at the end of speech. The results also suggest that adding morpheme information to DA does not lead to any performance improvement.

Rank	Feature	Attribute weight
1	DA3: Self-disclosure (experience)	-1.1044
2	DA20: Question (self)	1.0430
3	DA0: Greeting	1.0223
4	DA32: Confirmation	1.0217
5	DA30: Filler	-0.9358
6	DA25: Repeat	0.9255
7	LC (Length of utterance)	0.9123
8	DA23: Confirmation	0.9628
9	DA11: Acknowledgment	-0.9589
10	DA31: Admiration	-0.9583
11	DA21: Sympathy	0.9228
12	DA5: Self-disclosure (positive preference)	0.8653
13	DA6: Self-disclosure (negative preference)	0.8646
14	DA1: Provision	0.8458
15	DA9: Self-disclosure (plan)	0.8266
16	DA2: Self-disclosure (fact)	0.7872
17	FM (Final morphemes)	0.0330
18	DA7: Self-disclosure (neutral preference)	0.0000
19	DA3: Self-disclosure (experience)	0.0000
20	DA32: Other	0.0000
21	DA0: Greeting	0.0000
22	DA24: Proposal	0.0000
23	DA8: Self-disclosure (desire)	0.0000
24	DA13: Question (fact)	0.0000

Table 3: Excerpt of attribute weight ranking of 24 features in All model.

4. Discussion

The experimental results suggest that dialogue act information is most useful for estimating nods. Many of the spoken words in the IPUs used in this study are considerably collapsed compared with the written language, and it is considered that the accuracy of estimation using dialogue acts is not high. Nonetheless, the estimation model using dialogue acts performed well, so dialogue acts seem to have potential for nod estimation. The attribute weight of each feature value in the All model was calculated. The results are shown in Table 3 and suggest that DA20 (Question (self)), DA0 (Greeting), DA32 (Confirmation), DA25 (Repeat), DA23 (Confirmation), DA21 (Sympathy), DA5 (Self-disclosure (positive preference)), DA6 (Self-disclosure (negative preference)), DA1 (Provision), DA9 (Self-disclosure (plan)), and DA2 (Self-disclosure (fact)) contribute to accurate estimation of when a nod will occur. It can also be seen that DA3 (Self-disclosure (experience)), DA30 (Filler), DA11 (Acknowledgment), and DA31 (Admiration) contribute to estimating when a nod will not occur. The evaluation also showed that FM, which was used in previous research, is not very effective. The previous study focused on free dialogue, whereas in this research, one of the participants was instructed to explain something to the other party; hence, the results are dependent on the content of the dialogue. It would be interesting to see whether the effectiveness of feature values changes

depending on the dialogue scene.
In this research, we used an IPU as a unit of utterance and tried to determine whether nodding occurs in IPUs. We did not consider the detailed timing of occurrences or the number of nods in an utterance. We plan to focus on linguistic information other than dialogue acts and clarify co-occurrence relations with nods. Furthermore, we would like to work on constructing a model that can generate the occurrence timing and occurrence frequency within an utterance. The results of this study suggest that it is also possible that dialogue acts may be effective for nonverbal behaviors other than nods.

5. Conclusion

We constructed a highly accurate head-nod-generation model using natural Japanese language. In this research, we focused on using dialogue acts, which indicate the intention of utterances, for estimating nodding behavior accompanying utterance and demonstrated that they are effective information for generating nods. In an experiment, we found that our estimation model using dialogue acts outperformed those using morpheme information alone. We also found that a model using both dialogue act information and morpheme information showed no difference in performance from that using only dialogue acts. These results suggest that dialog act information has the potential to be a strong predictor that can be used to generate nods automatically.
In the future, we will focus on linguistic information other than dialogue acts and demonstrate the co-occurrence relation with nodding accompanying an utterance. Furthermore, we plan to construct a model for generating the occurrence timing of nods within an utterance and a model for generating other nonverbal behaviors such as gaze, which is important for turn management (Ishii et al., 2013b; Ishii et al., 2014; Ishii et al., 2015a; Ishii et al., 2015b; Ishii et al., 2016a; Ishii et al., 2016b; Ishii et al., 2017a) and expression of conversational engagement (Ishii et al., 2006; Ishii and Nakano, 2008; Ishii and Nakano, 2010; Nakano and Ishii, 2010; Ishii et al., 2011; Ishii et al., 2013a) and body posture.

6. Bibliographical References

Beskow, J., Granstrom, B., and House, D. (2006). Visual correlates to prominence in several expressive modes. In *INTERSPEECH*.

BirdWhistell, R. L. (1970). *Kinesics and context*. University of Pennsylvania Press.

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2010). WEKA–experiences with a java open-source project. *J. Machine Learning Research*, 11:2533–2541.

Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. (2007). Rigid head motion in expressive speech animation: Analysis and synthesis. In *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1075–1086.

Fuchi, T. and Takagi, S. (1998). Japanese morphological analyzer using word cooccurrence -jtag. In *International conference on Computational linguistics*, pages 409–413.

Graf, H. P., Cosatto, E., Strom, V., and Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 381–386.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). Towards an open-domain conversational system fully based on natural language processing. In *International conference on Computational linguistics*, pages 928–939.

Ishi, C. T., Ishiguro, H., and Hagita, N. (2006). Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. In *International Conference of Speech and Language*, pages 2006–2009.

Ishi, C. T., Haas, J., Wilbers, F. P., Ishiguro, H., and Hagita, N. (2007). Analysis of head motions and speech, and head motion control in an android. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 548–553.

Ishi, C. T., Ishiguro, H., and Hagita, N. (2010). Head motion during dialogue speech and nod timing control in humanoid robots. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 293–300.

Ishii, R. and Nakano, Y. (2008). Estimating user ' s conversational engagement based on gaze behaviors, 09.

Ishii, R. and Nakano, Y. I. (2010). An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication. In *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction*, EGIHMI '10, pages 33–40, New York, NY, USA. ACM.

Ishii, R., Miyajima, T., Fujita, K., and Nakano, Y. I. (2006). Avatar's gaze control to facilitate conversational turn-taking in virtual-space multi-user voice chat system. In *Intelligent Virtual Agents, 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006, Proceedings*, page 458.

Ishii, R., Shinohara, Y., Nakano, Y., and Nishida, T. (2011). Combining multiple types of eye-gaze information to predict user's conversational engagement. 02.

Ishii, R., Nakano, Y. I., and Nishida, T. (2013a). Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Trans. Interact. Intell. Syst.*, 3(2):11:1–11:25, August.

Ishii, R., Otsuka, K., Kumano, S., Matsuda, M., and Yamato, J. (2013b). Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the International Conference on Multimodal Interaction*, pages 79–86.

Ishii, R., Otsuka, K., Kumano, S., and Yamato, J. (2014). Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 694–698.

Ishii, R., Kumano, S., and Otsuka, K. (2015a). Multimodal fusion using respiration and gaze behavior for predicting next speaker in multi-party meetings. In *ICMI*, pages 99–106.

Ishii, R., Kumano, S., and Otsuka, K. (2015b). Predicting

next speaker using head movement in multi-party meetings. In *ICASSP*, pages 2319–2323.

Ishii, R., Otsuka, K., Kumano, S., and Yamamoto, J. (2016a). Predicting of who will be the next speaker and when using gaze behavior in multiparty meetings. *The ACM Transactions on Interactive Intelligent Systems*, 6(1):4.

Ishii, R., Otsuka, K., Kumano, S., and Yamamoto, J. (2016b). Using respiration to predict who will speak next and when in multiparty meetings. *The ACM Transactions on Interactive Intelligent Systems*, 6(2):20.

Ishii, R., Kumano, S., and Otsuka, K. (2017a). Analyzing gaze behavior during turn-taking for estimating empathy skill level. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 365–373, New York, NY, USA. ACM.

Ishii, R., Kumano, S., and Otsuka, K. (2017b). Prediction of next-utterance timing using head movement in multi-party meetings. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pages 181–187, New York, NY, USA. ACM.

Iwano, Y., Kageyama, S., Morikawa, E., Nakazato, S., and Shirai, K. (1996). Analysis of head movements and its role in spoken dialogue. In *International Conference on spoken language*, pages 2167–2170.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

KG, M., JA, J., DE, C., T, K., and E, V.-B. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. 15(2):133–7.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. In *Language and Speech*, volume 41, pages 295–321.

Lohse, M., Rothuis, R., Gallego-Pérez, J., Karreman, D. E., and Evers, V. (2014). Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1459–1466, New York, NY, USA. ACM.

McBreen, H. and Jack, M. (2001). Evaluating humanoid synthetic agents in e-retail applications. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 31:5.

Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K. (2010). Controlling listening-oriented dialogue using partially observable markov decision processes. In *International conference on computational linguistics*, pages 761–769.

Nakano, Y. I. and Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 139–148, New York, NY, USA. ACM.

Ooko, R., Ishii, R., and Nakano, Y. I. (2011). Estimating a user's conversational engagement based on head pose information. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'11, pages 262–268, Berlin, Heidelberg. Springer-Verlag.

Sakai, K., Ishi, C. T., Minato, T., and Ishiguro, H. (2015). Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *IEEE International Symposium on Robot and Human Interactive Communication*, pages 529–534.

Senko Maynard. (1987). Interactional functions of a nonverbal sign: Head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, 11:589–606.

Senko Maynard. (1989). Japanese conversation: Self-contextualization through structure and interactional management. *Norwood, New Jersey: Ablex Publishing Corporation*.

Watanabe, T., Danbara, R., and Okubo, M. (2003). Effects of a speech-driven embodied interactive actor interactor on talker's speech characteristics. In *IEEE International Workshop on Robot-Human Interactive Communication*, pages 211–216.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan a professional framework for multimodality research. In *International Conference on Language Resources and Evaluation*.

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. 30(3):555–568.