Generating Body Motions using Spoken Language in Dialogue

Ryo Ishii
NTT Media Intelligence Laboratories, NTT Corporation
Yokosuka-shi, Kanagawa, Japan
ishii.ryo@lab.ntt.co.jp

Taichi Katayama
NTT Media Intelligence Laboratories, NTT Corporation
Yokosuka-shi, Kanagawa, Japan
katayama.taichi@lab.ntt.co.jp

Ryuichiro Higashinaka
NTT Media Intelligence Laboratories, NTT Corporation
Yokosuka-shi, Kanagawa, Japan
higashinaka.ryuichiro@lab.ntt.co.jp

Junji Tomita
NTT Media Intelligence Laboratories, NTT Corporation
Yokosuka-shi, Kanagawa, Japan
tomita.junji@lab.ntt.co.jp

ABSTRACT

We propose a model to automatically generate whole body motions accompanying utterances at appropriate times, similar to humans, by using various types of natural-language-analysis information obtained from spoken language. Specifically, we focus on the co-occurrence relationship between various types of natural-language-analysis information such as words included in the spoken language, parts of speech, a thesaurus, word positions, dialogue acts of the spoken language, and human motions. Our model automatically generates nods, head postures, facial expressions, hand gestures, and upper-body posture using such information. We first recorded a two-person dialogue and constructed a multimodal corpus including utterance and whole body motion information. Next, using the constructed corpus, we constructed our model for generating a motion for each phrase unit using machine learning and using words, parts of speech, a thesaurus, word positions, and speech acts of the entire spoken language as inputs. These types of natural-language-analysis information were useful for motion generation. The effectiveness of our model was verified through a subjective experiment using a virtual conversational agent. As a result, the agent's body motions and impressions regarding naturalness of motion, degree of coincidence between utterance and motion, humanness of the agent, and likability of the agent improved with our model.

CCS CONCEPTS

• **Human-centered computing → Collaborative interaction**; **HCI theory, concepts and models**; **Collaborative and social computing theory, concepts and paradigms**;

KEYWORDS

body motion, generation method, natural language, language, dialogue, virtual agent

ACM Reference Format:
Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. 2018. Generating Body Motions using Spoken Language in Dialogue. In *IVA '18: International Conference on Intelligent Virtual Agents (IVA '18), November 5–8, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3267851.3267866

IVA '18, November 5–8, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6013-5/18/11…$15.00
https://doi.org/10.1145/3267851.3267866

1 INTRODUCTION

In human communication, body motion, in addition to spoken language, is important to convey emotion and intention [3]. Therefore, a virtual agent or humanoid robot in a dialogue system should express appropriate body motions according to utterances and effectively communicate with a user. Communication robots and virtual agents have also been drawing attention in industry, and are applied in various services such as communication opponents, reception, and Q&A. One of the main problems in handling actual services is that it takes a huge amount of time to manually create a motion scenario for each utterance for a virtual agent or humanoid robot to generate body motions. Also, what kind of motion is to be generated and the timing of its appearance are situations in which the creator's subjectivity is taken into consideration, and an appropriate motion is necessarily generated. As dialogue technology advances, it will become unrealistic to manually provide appropriate motions to all speech for systems capable of automatically responding to various utterances.

In response to these problems, we are working on enabling humanoid robots and virtual agents to automatically generate motions based on the content of utterances at the proper time, similar to humans. It is believed that the generation cost of motion will significantly decrease or be eliminated with such technology. We propose a model for generating whole body motions such as nodding, head pose, facial expressions, hand gestures, and upper-body posture for each phrase using various types of natural-language-analysis information obtained from spoken language as input. Words, parts of speech, a thesaurus in the spoken language, word positions, and all utterance actions of the spoken language are used as various types of natural-language-analysis information. This is the first attempt to study the relevance of such information and

motions. An advantage of generating motions using spoken language information is that spoken language and body motions are simultaneously generated and output to the brain and deep connection. That is, based on such a co-occurrence relation, spoken language information is considered effective for motion generation.

Initially, a two-person dialogue is recorded, and a multimodal corpus including utterance, head-movement, facial-expression, hand-gesture, and body-posture information is constructed. Next, using the constructed corpus, a model for generating a motion for each phrase unit is constructed using machine learning and using words, parts of speech, a thesaurus, word positions, and speech acts of the entire spoken language as inputs. As a result, such information was useful for motion generation in our research. The effectiveness of the motion generated with our model was verified through a subjective experiment using a virtual agent. As a result, the generated motions improved impressions regarding naturalness of motion, degree of coincidence between utterance and motion, humanness of the agent, and likeability of the agent.

Our generation model outputs whole body motions such as nodding, head pose, facial expressions, hand gestures, and upper-body posture for each phrase using various types of language-analysis information obtained from spoken language as input. The relevance of such language information and motions has not been studied, and this is the first attempt.

An advantage of generating motions using spoken language information is that spoken language and body motions are simultaneously generated and output in the brain and deep connection. That is, based on such co-occurrence relation, spoken language information is considered effective for motion generation.

2 RELATED WORK

In human communication, body motions, such as nodding, head posture, facial expressions, hand gestures, and upper-body posture, as well as speech, are known to transmit emotion and intention [3, 24–26]. Therefore, it has been shown that enabling a humanoid virtual agent or robot appropriate body motions not only improves their natural appearance but also promotes conversation. For example, a motion accompanying an utterance has the effect of strengthening the persuasive power of that utterance, making it easier for the other party to understand the content of the utterance [18]. Against such a background, many studies have been conducted to enable virtual agents to generate appropriate motions. Cassel et al. proposed a system framework for generating motion that suits the uttered voice based on text information [5]. In this framework, they proposed to generate representative nonverbal behavior, such as beat gesture and gaze direction, based on a heuristic rule by using word information of speech.

Attempts have been made to generate motion during speaking, especially using speech-sound information [2, 4, 6, 9, 12–14, 16, 26, 28, 29], which is used frequently as sound pressure and prosodic features. However, it was difficult to accurately generate body motions from the speech-sound of utterances. It is known that the co-occurrence relationship between speech-sound features and nodding is weak in Japanese [11, 29]. Therefore, if a model capable of generating body motions at a more appropriate time also uses information other than speech-sound information, more effective communication between a dialog system and user may be possible.

Figure 1: Example of dialogue scene

In conventional studies on motion generation from natural language information, small motions, such as presence or absence of nodding and limited hand gestures [6, 11, 12, 15, 21], are created mainly using word information, which is very simple information. Some studies also involved generating limited hand gestures using some natural language information. They also addressed generating hand gestures based on speech and natural language information. However, our focus was generating whole body motions from detailed natural-language-analysis information.

There are also several advantages of using natural-language-analysis information. A dialog agent using speech synthesizes speech sound from synthetic sound processing, such as feature-amount extraction, when generating motions using speech information after acquiring text information from which the dialogue agents speaks. This increases processing time. As a result, the problem with motion generation using voice information is that the response time to the user is delayed. On the other hand, when natural-language-analysis information is used, such processing is unnecessary and responsiveness is ensured, which is important for real-time communication. Since the input of speech synthesis is based on the spoken language, motion-generation accuracy does not change when using only spoken languages but may decrease. It is also expected that this technology will be applied more widely, such as to text chatting without using speech, by using only natural-language-analysis information.

3 CORPUS DATA

To collect a Japanese conversation corpus including verbal and nonverbal behaviors for generating nods in a 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 a cartoon that participants had not seen as the conversational content. Before the dialogue, they watched a famous cartoon called "'Tom & Jerry" in which the characters do not speak. In each dialogue, one participant explained the content of the cartoon to the conversational partner within ten minutes. At any time during this period, the partner could freely ask questions about the content.

Generation target	Number of classes	Class details
Number of nods	6	0, 1, 2, 3, 4, more than 5
Depth of nod	4	micro, small, medium, large
Head rotation (yaw)	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
Head rotation (roll)	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
Head rotation (pitch)	7	front, up-micro, up-small, up-medium, up-large, up-micro, up-small, up-medium, up-large
Facial expression	8	happiness, sadness, surprise, fear, anger, disgust, contempt, normal
Hand gesture	9	none, iconic, metaphoric, beat, deictic, feedback, compellation, hesitate, others
Upper-body posture	7	center, forward-small, forward-medium, forward-large, forward-small, forward-medium, forward-large

Table 1: List of generated motion parameters

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 cartoon 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) [17]. 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. We also used J-tag [8], which is a general morphological analysis tool for Japanese, to divide an IPU into phrases. We collected 11877 phrases in total.
- Head direction: Using the facial-image processing tool OpenFace [23] for the front image of the participant obtained from the video camera installed in front of the participant, three-dimensional face orientation information such as the angles of yaw, roll, and pitch, were acquired. When each of these angles is 10 degrees or less, it is micro, 20 degrees or less, small, 30 degrees or less, medium, and 45 degrees, large.
- 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. The frequency (number) of nods was also manually labeled as 1, 2, 3, 4, or 5 or more. The change in the rotation angle of the head when nodding occurred was measured using OpenFace, which is head-tracking software that uses image processing [1]. The difference between the direction angle of the head at the beginning of nodding and that when the head is oriented furthest downward was obtained as nodding-depth information. Furthermore, the nodding depth was classified into the following four stages.
  - Micro: Depth less than 5 degrees
  - Small: Depth greater than 5 degrees and less than or equal to 10 degrees
  - Medium: Depth greater than 10 degrees and less than or equal to 20 degrees
  - Large: Depth greater than 20 degrees
- Facial expression: We used OpenFace to extract the strength of an action unit (AU) [22]. We extracted seven facial expressions (joy, anger, disgust, sadness, fear, surprise, normal) using the related combinations of AU strengths. Specifically, joy is related to AU 6 and AU 12; anger to AU 4, AU 7, and AU 24; disgust to AU 4, AU 7, and AU 25; sadness to AU 1, AU 4, AU 7, AU 15, and AU 17; fear to AU1, AU2, AU5, and AU26; and surprise to AU1, AU2, AU20, and AU43. We calculated the average strength of each AU related to each facial expression as that of that facial expression. We set this mean value as the strength x_f of an expression f. The average value μ_f and standard deviation σ_f of x_f were also calculated for each facial expression using the intensity value of each facial expression obtained from the entire dialogue corpus. We calculated the Z score Z_f using these values and normalized x_f. We selected the facial-expression classes that had the largest Z_f of each facial expression as the current facial expressions. When the Z_f of all facial expressions was less than 0.3, these expressions were assigned to the "normal" class .
- Hand gesture: manual annotation was carried out on the state in which the hand gesture is being performed. The series of actions of this gesture was classified into the following four states.
  - Prep: Raise hands to gesture from home position
  - Hold: State of holding hand in the air (standby time to start gesture)
  - Stroke: Perform gesture
  - Return: Return hand to home position

  However, a series of actions from Prep to Return was considered as one gesture event for convenience. Furthermore, hand gestures were classified into the following eight types based on McNeil's hand-gesture classification [19].
  - Iconic: Gesture used to express scene depiction and motion.
  - Metaphoric: Similar to Iconic, it is a painterly and graphical gesture, but contents instructed are abstract, for example, the flow of time.
  - Beat: Tone of utterance and emphasizing remarks. Vibrate hands and waving them according to speech.

Target	Chance level	Proposed model
Number of nods	0.226	0.428
Depth of nod	0.304	0.475
Head rotation (yaw)	0.232	0.331
Head rotation (roll)	0.297	0.397
Head rotation (pitch)	0.261	0.379
Facial expression	0.175	0.294
Hand gesture	0.156	0.305
Upper-body posture	0.183	0.313

Table 2: Evaluation results of our model and chance level

  - Deictic: A gesture that points directly to directions, places, and things, such as finger pointing.
  - Feedback: Gesture to synchronize, agree with, and respond to the utterances of others. A gesture accompanying speaking in response to a previous utterance/gesture of another person; also, imitating the opponent's gesture.
  - Compellation: Gesture to call the other party.
  - Hesitate: A gesture when it is time to say something.
  - Others: Gestures other than the above.
- Upper-body posture: participants were seated in this dialogue scene, and there was no significant change in the seated position. For this reason, we extracted the posture behind the upper body based on the three-dimensional position of the head. Specifically, we obtained the difference between the center position and coordinate position in the front-back direction of the head position obtained using OpenFace. Based on this positional information, the change in posture angle of the upper body was calculated. When it was 10 degrees or less, it was classified as micro, 20 degrees or less, small, 30 degrees or less, medium, and over 45 degrees, large.

All verbal and nonverbal behavior data were integrated at 30 Hz for display using the ELAN viewer [27]. This viewer enabled us to annotate the multimodal data frame-by-frame and observe the data intuitively.

4 BODY-MOTION GENERATION

We constructed our model using the conditional random field (CRF). The model generated one motion class for each phrase in each of the eight motions parameters listed in Table 1 using the constructed corpus including words, parts of speech, a thesaurus, word positions, and speech acts of the entire spoken language as input. That is, eight motion labels were generated for each clause. We used the following natural-language-analysis features.

- Length of phrase (LP): Number of characters in a phrase.
- Word position (WP): Word position in a sentence.
- Bag of words (BW): Other studies focused on a limited number of words to generate head nods. To handle more generic word information, we examined the number of occurrences of all words, not some words. We used J-tag [8], a general morphological analysis tool for Japanese.
- Dialogue act (DA): A DA was extracted using an estimation technique for Japanese [10, 20]. The technique can estimate a DA using the word N-grams, semantic categories (obtained from a Japanese thesaurus Goi-Taikei), and character N-grams. There are 32 types of DAs.
- Part of speech (PS): Number of occurrences of the PSs of words in a phrase. We used J-tag [8] to extract PS information.
- Large-scale Japanese thesaurus (LT): The LT is a large lexical database of Japanese. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

Figure 2: Virtual agent in experiment

We constructed our motion-generation model using the eight motion parameters listed in Table 1. We used 24-fold cross validation using a leave-one-person-out technique with the data for the same 24 participants mentioned above. We evaluated how effectively the body motions of a participant could be estimated with an estimator generated only from the data of other people.

Table 2 shows the average F-measure. The chance level is a performance F-measure of when all classes with the most correct answers are output. The F-measure significantly improved over the chance level in all motion-generation parameters (the result of the corresponding t-test: $p < .05$). The proposed model using words, PSs, the LT, WPs, and DAs of the entire spoken language was effective in generating whole body motions.

5 SYSTEM CONSTRUCTION AND EVALUATION EXPERIMENT

The proposed motion-generation model was assessed through subjective experiments as to whether it is effective in generating motions when controlling the body motions of a virtual agent.

An experimental system was configured for our model to generate the eight motion parameters for each phrase. A speech engine (FutureVoice) [7] using a DNN created the synthesized sound from the spoken language. The virtual agent shown in Fig. 4 was created on UNITY, and correspondences to the eight motion parameters were created. Animation of body motion was generated on UNITY according to the occurrence of sound based on the reproduction of speech sound and motion-schedule information. At this time, the eight motion-generation parameters could be used independently.

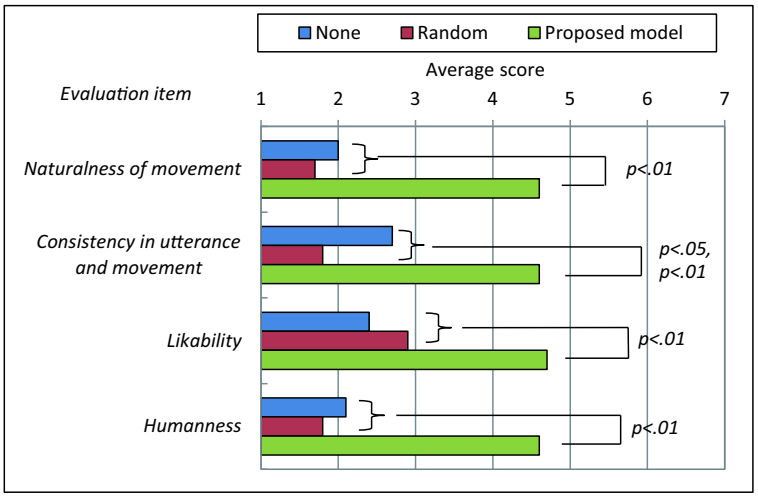

Figure 3: Results of subjective evaluation

For head movement, all parameters were mixed and generated, and lipsync was implemented so that simple lip motion matching the voice could be generated.

Two control conditions, i.e., not performing a motion that matches an utterance (unconditional) and random motion generation (random), were applied for comparison with motions generated with the proposed model. The same five utterances of the agent's self-introduction were prepared for each condition. After observing the behavior of the agent under each condition, subjective evaluation on a 7-point Likert scale was carried out. The evaluation items were "naturalness of movement", "consistency in utterance and movement", "likability", and "humanness". Ten participants (men and women) took part in the evaluation. Each model was randomly presented for each participant to offset the order effect.

The average scores are shown in Fig. 3. A one-way analysis of variance was conducted within each evaluation item, and it was verified whether the effect of each experimental condition was significant. Significant differences were found for all items (naturalness of movement: $F(2, 27) = 37.12, p < .01$, consistency in utterance and movement: $F(2, 27) = 18.34, p < .01$, likability: $F(2, 27) = 7.80, p < .01$, and humanness: $F(2, 27) = 33.76, p < .01$). There were significant differences between the proposed model's conditions and the two control conditions from the corresponding t test. Therefore, all items seemed to have improved with the proposed model.

6 DISCUSSION

The accuracy of our model was from 0.294 to 0.475, which is not necessarily high. Regarding this, in the first place, human motion is not necessarily of a nature that must occur with a specific utterance, that is, it is originally generated naive in the first place. Therefore, it is difficult to perfectly reproduce actual human motions, but this is not essential. We examined how accurately a virtual agent could reproduce human motions by conducting a subjective evaluation. The agent's motions generated with the proposed model were considered natural, consistent in motions and utterances, likable, and human-like, demonstrating the effectiveness of our model.

There are several issues with our model. First, estimation accuracy needs to be improved. Previous studies have shown that hand gestures can be generated using deep learning. Second, it is possible to generate more types of hand gestures. There are currently several types of hand gestures. In reality, however, hand gestures have many types of movement. Third, generation of facial expressions from text information is not necessarily accurate. Since speaking the same text does not necessarily have the same emotion, it is difficult to generate expressions only with text information. We are planning to solve these problems in the future.

7 CONCLUSION AND FUTURE WORK

We focused on the co-occurrence relation between various types of natural-language-analysis information such as words included in the spoken language, parts of speech, a thesaurus, word positions, speech acts of the entire spoken language, and human motions. We constructed a model that automatically generates nods, head pose, facial expressions, hand gestures, and upper-body posture. First, we recorded a two-party dialogue and constructed a multimodal corpus including utterance, head-movement, facial-expression, hand-gesture, and body-posture information. Next, using the constructed corpus, we constructed our model for generating a motion for each phrase by using machine learning and various natural-language-analysis information as input. These types of information were useful for motion generation. To promote the wide use of this model, we constructed a demonstration system that can easily generate motions by using the proposed generation model and automatically control virtual agents created on UNITY from the spoken language only. In this system, when an

arbitrary spoken language is input, synthesized sound and motion information are acquired from the speech synthesizer and motion-generation API, and an utterance of UNITY and animated motion are generated. By conducting a user subjective evaluation using this demonstration system, it was possible to evaluate the naturalness of motion, consistency between utterance and motion, likability, and humanness of the agent. We confirmed that the impression of these items improved with our model.

This motion-generation model can be widely applied not only to interactive systems using virtual agents and humanoid robots but also to avatars on text-chat applications and CG animation.

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[2] Jonas Beskow, Bjorn Granstrom, and David House. 2006. Visual correlates to prominence in several expressive modes. In *INTERSPEECH*.

[3] Ray L. BirdWhistell. 1970. *Kinesics and context.* University of Pennsylvania Press.

[4] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. 2007. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. In *IEEE Transactions on Audio, Speech, and Language Processing*. 1075–1086.

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: The Behavior Expression Animation Toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. ACM, New York, NY, USA, 477–486.

[6] C.-C Chiu, L.-P Morency, and Stacy Marsella. 2015. Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach. , 152-166 pages.

[7] NTT TechnoCross Corporation. 2018. FutureVoice. http://www.v-series.jp/futurevoice/

[8] Takeshi Fuchi and Shinichiro Takagi. 1998. Japanese Morphological Analyzer using Word Cooccurrence -JTAG. In *International conference on Computational linguistics*. 409–413.

[9] Hans Peter Graf, Eric Cosatto, Volker Strom, and Fu Jie Huang. 2002. Visual Prosody: Facial Movements Accompanying Speech. In *IEEE International Conference on Automatic Face and Gesture Recognition*. 381–386.

[10] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *International conference on Computational linguistics*. 928–939.

[11] Carlos T. Ishi, Judith Haas, Freerk P. Wilbers, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Analysis of head motions and speech, and head motion control in an android. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 548–553.

[12] Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2010. Head motion during dialogue speech and nod timing control in humanoid robots. In *ACM/IEEE International Conference on Human-Robot Interaction*. 293–300.

[13] Ryo Ishii, Toshimitsu Miyajima, Kinya Fujita, and Yukiko I. Nakano. 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'06)*. 458.

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

[15] Yuki Kadono, Yutaka Takase, and Yukiko I. Nakano. 2016. Generating Iconic Gestures Based on Graphic Data Analysis and Clustering. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, Piscataway, NJ, USA, 447–448. http://dl.acm.org/citation.cfm?id=2906831.2906920

[16] Munhall KG, Jones JA, Callan DE, Kuratate T, and Vatikiotis-Bateson E. 2004. Visual prosody and speech intelligibility: head movement improves auditory speech perception. 15, 2 (2004), 133–7.

[17] H Koiso, Y Horiuchi, S Tutiya, A Ichikawa, and Y Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. In *Language and Speech*, Vol. 41. 295–321.

[18] Manja Lohse, Reinier Rothuis, Jorge Gallego-Pérez, Daphne E. Karreman, and Vanessa Evers. 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)*. ACM, New York, NY, USA, 1459–1466. https://doi.org/10.1145/2556288.2557274

[19] David McNeill. 1996. *Hand and Mind: What Gestures Reveal About Thought.* University Of Chicago Press.

[20] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling listening-oriented dialogue using partially observable Markov decision processes. In *International conference on computational linguistics*. 761–769.

[21] Yukiko I. Nakano, Masashi Okamoto, Daisuke Kawahara, Qing Li, and Toyoaki Nishida. 2004. Converting Text into Agent Animations: Assigning Gestures to Text. , 4 pages.

[22] Ekman Paul, Friesen Wallace, and Hager Joseph. 2002. *Facial action coding system: a technique for the measurement of facial movement.* Consulting Psychologists Press.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *CoRR* abs/1503.03832 (2015). arXiv:1503.03832 http://arxiv.org/abs/1503.03832

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

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

[26] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. 57 (2014), 209–232.

[27] P. Wittenburg, H. Brugman, A. Russel, A Klassmann, and H Sloetjes. 2006. ELAN a Professional Framework for Multimodality Research. In *International Conference on Language Resources and Evaluation*.

[28] Michiya Yamamoto and Tomio Watanabe. 2007. Development of an Embodied Image Telecasting Method Via a Robot with Speech-Driven Nodding Response. In *HCI*, Vol. (8). 1017–1025.

[29] Hani Camille Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. 2002. Linking facial animation, head motion and speech acoustics. 30, 3 (2002), 555–568.