A Speech-driven Hand Gesture Generation Method and Evaluation in Android Robots

Carlos T. Ishi, Daichi Machiyashiki, Ryusuke Mikata, Hiroshi Ishiguro

***Abstract*— Hand gestures commonly occur in daily dialogue interactions, and have important functions in communication. We first analyzed a multimodal human-human dialogue data and found relations between the occurrence of hand gestures and dialogue act categories. We also conducted clustering analysis on gesture motion data, and associated text information with the gesture motion clusters through gesture function categories. Using the analysis results, we proposed a speech-driven gesture generation method by taking text, prosody, and dialogue act information into account. We then implemented a hand motion control to an android robot, and evaluated the effectiveness of the proposed gesture generation method through subjective experiments. The gesture motions generated by the proposed method were judged to be relatively natural even under the robot hardware constraints.**

***Index Terms*—Android robots, Emotion, Hand Gesture, Motion generation, Speech-driven.**

I. INTRODUCTION

THE background of this work is the generation of natural motions in humanoid robots, matched with the speech utterances. Android robots have a highly human-like appearance, which gives them the ability to achieve natural communication with humans through several types of non-verbal information, such as facial expressions and gestures. So far, we have investigated the relations between speech and several modalities including facial, head and torso movements, accounting for dialogue act functions, laughing speech and surprise utterances [1-3], and proposed several methods for generating natural motions in humanoid robots matched with the speech utterances [4-8].

Besides facial and head movements, hand gestures also commonly occur in dialogue interactions, having important functions in human-human communication [9-12]. Although there are controversies on whether hand gestures are speaker-directed or listener-directed, we consider that in either of the cases they are important for expressing human-likeness in human-robot interactions. Thus, in this study, we focus on analysis and generation of hand gestures.

Several studies have been conducted on text-based gesture synthesis in CG animated agents [13-17]. For example, lexicon-based approaches have been proposed for generating iconic gestures in [13]. An imagistic description tree was proposed for representing the semantics of shape-related expressions in [14]. A framework that combines data-driven with model-based techniques to model the generation of iconic gestures with Bayesian decision networks was proposed in [15]. In [16], gesture and speech features are associated and modeled by HMMs, and directional (pointing) gestures are generated. In [17], a system that converts text into an animated agent by synchronizing gestures and speech was implemented. It is reported that lexical and syntactic information are strongly correlated with gesture occurrences, and that syntactic structures are useful for judging gesture occurrences.

Prosodic information has also been exploited when generating hand gestures, mainly by considering relations between prosodic focus (emphasis) and beat gestures. For example, relationship between gestures and intonation has been investigated in [18]. It is reported that apexes of gestural strokes and pitch accents aligned consistently, and gestural phrases and intermediate phrases aligned quite often. In [19], a prosody-based approach has been proposed for synthesis of body language, by associating motion and speech streams. It is reported that realistic and compelling body language could be produced.

From the above cited past studies, we can say that iconic gestures can be associated with the lexical contents, while beat gestures can be associated with the prosodic features of the speech utterances. However, the past studies remain unclear when and to what extent gestures should be generated. Hand gestures do not occur in every utterance, so that some criterion is necessary to decide when to generate or not a gesture.

In this study, we first analyze the relations of speech and gesture occurrence with special focus on dialogue acts. We then take previous studies on gesture generation into account, and propose a speech-driven gesture generation method based on text, prosody, and dialogue act information. We implemented a hand motion control in an android robot, and evaluated the effectiveness of the proposed gesture generation method through subjective experiments.

Manuscript received: February, 24th, 2018; Revised April, 25th, 2018; Accepted June, 1st, 2018. This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by JST, ERATO, Grant Number JPMJER1401.

C. T. Ishi is with the ATR Hiroshi Ishiguro Labs., Kyoto, Japan (e-mail: carlos@atr.jp).

D. Machiyashiki is with the ATR Hiroshi Ishiguro Labs., Kyoto, Japan (e-mail: machiyashiki.daichi@irl.sys.es.osaka-u.ac.jp).

R. Mikata is with the ATR Hiroshi Ishiguro Labs., Kyoto, Japan (e-mail: mikata.ryusuke@irl.sys.es.osaka-u.ac.jp).

H. Ishiguro is with the ATR Hiroshi Ishiguro Labs., Kyoto, Japan (e-mail: ishiguro@sys.es.osaka-u.ac.jp).

Digital Object Identifier (DOI) : see top of this page.

The main contributions of the proposed method are as follows: 1) we propose a text-based gesture generation that associates input text with gesture motion clusters through probabilistic functions; 2) we integrate the text-based gesture generation with a prosody-based gesture generation, for accounting for hand gestures that are dependent on both content and speaking style; 3) we introduce a function that constrains the generation of a gesture, based on analysis results on relationship between gesture and dialogue act categories.

II. Hand Gesture Analysis and Modeling

In this study, we collected a multimodal three-party dialogue data, and first analyzed the relations of speech and gesture occurrence focusing on dialogue act categories (Sections II.A ~ II.C). Motion features are extracted and clustered in order to reduce dimensionality and increase the representability of the gesture motions (Section II.D). Probabilistic models are then created for associating text information with motion clusters, considering word concepts and gesture function categories (Section II.E). Duration features are also analyzed for different gesture phases (Section II.F).

A. Analysis data

For analysis, we use a dataset of the multimodal three-party conversational speech database collected at our research institute (ATR). The database contains multiple sessions of face-to-face conversations among three speakers. Audio, video and motion data are captured by headset microphones (DPA4060), RGB-D Kinect sensors (Microsoft Kinect-V2), and two IMU sensors at the head top and at the back (Intersense InertiaCube4), for each speaker. The dialogue participants had a seat in chairs around a table, with a distance of about 2 meters between each other. The table is 60cm height, so that the Kinect sensors can detect hand motions around the speakers' knee area. Fig. 1 shows the setup for three-party dialogue data collection. The right picture shows an example of the image captured by one of the Kinect sensors.

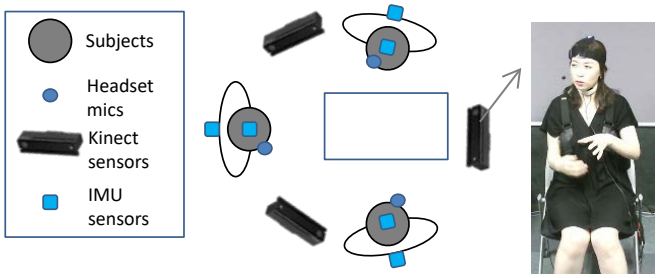

Figure 1. Setup for three-party dialogue data collection, and an example of the image captured by one of the Kinect sensors.

Each dialogue session comprises about 20 to 30 minutes of random topic conversations. The speech utterances were segmented in phrase units (accentual phrases) and text-transcribed by a native speaker. For the present analysis, a dataset of 8 speakers (5 female and 3 male speakers) extracted from 4 Japanese dialogue sessions were used. Some of the speakers participated in multiple dialogue sessions. The ages of all participants are within 20s to 30s.

B. Annotation data

Dialogue act categories were annotated for each phrase, according to the label set of Table I, which is based on previous studies [1].

TABLE I. Dialogue Act Label Set

Dialogue act	Description of the dialogue act category
Interjectional backchannels (**bc**)	Feedback responses like "*un*", "*hai*" (equivalent to "*uhm*", "*uh-huh*", "*yes*" in English)
Non-interjectional backchannels (**bc2**)	Feedback responses like "*hontodesuka*", "*sugoine*" ("*really?*", "*great!*"); also includes repetition of dialogue partner's words/phrases.
Turn-giving statements (**g**)	End of sentences where the turn is released to the dialogue partner
Turn-giving questions (**q**)	End of sentences requesting an answer from the dialogue partner.
Turn-giving/ keeping (**gk**)	Utterances ambiguous between turn-giving and turn-keeping
Turn-keeping with strong boundary (**k**)	Middle of sentences with strong prosodic boundaries (intonational phrases); accompanied by a short pause or a clear pitch reset.
Turn-keeping with weak boundary (**k2**)	Middle of sentences with weak prosodic boundaries (accentual phrases).
Fillers (**f**)	Utterances like "*eetoo*", "*anooo*" (equivalent to "*uhmmm*", "*I mean…*" in English)

The category "bc2" was newly introduced in the present study, since non-interjectional backchannel utterances were partly mixed with the turn-giving category "g" in the previous studies. A research assistant (native speaker of Japanese with previous experience in dialogue act categorization) annotated the labels above, by listening to the speech utterances.

The hand gestures were segmented in gesture phases according to Fig. 2 [9, 10]. The stroke is the meaningful part of a gesture. It is preceded by a preparation phase, where the hands move from a rest position into a gesture space where it can begin the stroke. In the retraction phase, the hands return to the rest position (not always the same position as at the start). There may not be a retraction phase if the speaker immediately moves into a new stroke. Hold phases are temporary cessations of motion either before or after the stroke motion. Holds ensure that the meaningful part of the gesture (the stroke) remains semantically active during the co-expressive speech.

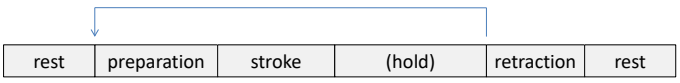

Figure 2. Gesture phases.

Hand gesture function categories were annotated for the stroke phases of the hand gestures according to Table II [11,12].

The gesture phases were segmented by one researcher and later checked by a research assistant, by looking at the videos and listening to the speech contents. In the present research, a separate segmentation layer was conducted for beat gestures, since those can co-occur with the hold phases of other gestures. For example, beats can occur during a deictic gesture or a metaphoric gesture.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2018.2856281, IEEE Robotics and Automation Letters

TABLE II. GESTURE FUNCTION CATEGORIES

Gesture function	Description of the gesture function categories
Iconic	Gestures presenting images of concrete entities and/or actions. The gesture, as a referential symbol, functioning via its formal and structural resemblance to event or objects.
Metaphoric	Gestures not limited to depictions of concrete events. They also picture abstract content, in effect, imagining the unimageable. In metaphoric gestures, an abstract meaning is presented as if it had form and/or occupied space.
Deictic (pointing)	The prototypical deictic gesture is an extended 'index' finger, but almost any extensible body part or held object can be used.
Beat	So called because the hand appears to beating time. Beats are mere flicks of the hand(s) up and down or back and forth, zeroing in rhythmically on the prosodic peaks of speech.
Emblem	Conventionalized signs, such as thumbs-up or the ring (first finger and thumb tips touching, other fingers extended) for "OK". Emblems are culturally specific, have standard forms and significances, and vary from place to place.
Adapter	Movements that often involve self-touch, such as scratching or touching the hairs. Adapters happen almost entirely unaware, and may or not be accompanied by speech. Adapters can reflect the perceived emotional stability [20].

For each gesture stroke segment, the gesture categories were annotated by two research assistants (inter-rater agreement κ = 0.47), also by looking at the video and listening to the speech contents.

C. *Analysis of hand gestures and dialogue acts*

The total number of hand gesture occurrences found in the dataset was 49 emblems, 211 iconic, 208 deictic, 357 metaphoric, 861 beats and 343 adapters. Among the gesture categories, beat gestures occurred with the highest frequency. Part of it occurred along with other gesture categories, while the other part occurred individually.

The number of overlapping incidents between hand gesture intervals and speech intervals were counted in each dialogue act categories. The adapter (self-touch) motion events were removed from the present analysis, since they are less dependent on the speech contents. Fig. 3 shows the distributions of hand gesture occurrences in different dialogue act categories, for all speakers. The number of occurrences in each dialogue act category is shown within brackets. The distributions are shown for all hand motion intervals ("all"), and for each gesture phase intervals (stroke, preparation, hold and retraction).

From the results in Fig. 3, it can firstly be observed that for all speakers, the occurrence rates of hand gestures during both interjectional and non-interjectional backchannel utterances ("bc" and "bc2") show very low occurrence of hand gestures. Higher occurrence rates can be observed for turn-keeping ("k" and "k2"), fillers ("f") and "gk" (ambiguous category between turn-giving and turn-keeping). This suggests that when the speaker is in a listening mode, where backchannel utterances are predominant, the occurrence rate of gestures is low, while when the speaker is in speaking mode, the occurrence rate of gestures becomes higher.

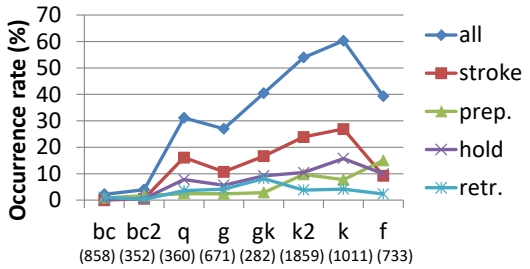

Figure 3. Occurrence rates of hand gestures in different dialogue act categories, on the whole gesture intervals ("all") and on each gesture phase. The total number of utterances in each dialogue act is shown within brackets.

In questions and turn-giving phrases ("q" and "g"), the occurrence rates of gesture were intermediate between backchannels ("bc") and turn-keeping ("k") categories. This indicates that gestures occur with higher frequency in the middle of long sentences (where phrases with turn-keeping occur) rather than at the end of the sentences (where "g" and "q" phrases occur).

Regarding the fillers ("f"), relatively high occurrence rates are observed on the whole gesture intervals ("all"), but relatively lower occurrence rates are observed for the stroke intervals only ("stroke"), in comparison to the others. It is also observed that preparation and hold phases are predominant during fillers ("prep." and "hold").

D. *Motion feature extraction and hand motion clustering*

3D skeleton data was extracted using the Kinect SDK, and 2D skeleton data was extracted from the video data of the Kinect sensor using the Openpose software [21]. Before matching the Kinect and Openpose data, the image distortions caused by the camera lens were compensated according to the procedure in [22].

The 3D skeleton extracted by the Kinect SDK is depth-based, so that estimation errors often occurred when the hands and arms are in contact with other parts of the body and face. Also, hand shape information is not provided by the Kinect SDK. On the other hand, the skeleton extracted by the Openpose is more robust and also includes estimation of hand and fingers, but it is 2D and given in pixels. The skeleton information from Kinect and Openpose were then combined to estimate 3D positions of the shoulder, elbow, wrist, and hands. Fig. 4 shows the skeleton joint positions extracted by Openpose and Kinect (left panels), and the merged joints accompanied by the joint numbers used in this study (right panels).

The joints 0, 1, 2, 3 and 6 are directly obtained from Kinect 3D data. Then, the elbow and wrist joints (4, 5, 7 and 8) are obtained by combining Openpose and Kinect data, as follows. To associate the Kinect 3D data to the Openpose 2D data, the shoulder positions were used as reference, since they are relatively robust in both skeleton estimation data. Then, candidates for the elbow position are generated over a sphere centered in the shoulder with radius equal to the upper arm length. These candidates are projected to the 2D pixel space, and the one closest to the 2D elbow pixel is selected to provide the 3D position of the elbow. Similar procedure is conducted to estimate the 3D positions of the wrist and hands.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2018.2856281, IEEE Robotics and Automation Letters

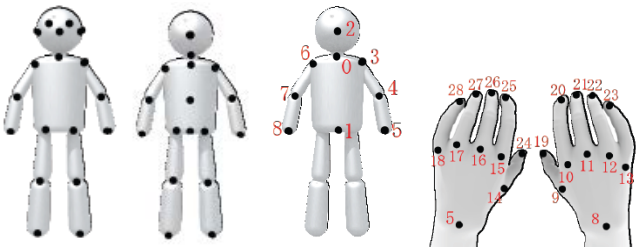

Figure 4. Skeleton joints extracted by Openpose and Kinect (left panels), and the merged joints (right panels).

The hand and finger joints (9 to 28) shown in the rightmost image of Fig. 4 correspond to a simplified version of the joints extracted by the Openpose data. Since the extraction of the hand shapes is less robust by the Openpose software, and there is also a hardware constraint for hand shapes in the current android, we focused on the production of hand positions and orientations, disregarding detailed hand shapes.

The hand motion data will then be expressed as a time series of the hand, wrist, elbow and shoulder position vectors in the 3D space, for each of the left and right hands. The time series data is 30 fps (the same frame rate of the collected video data). In order to normalize the 3D coordinate positions among speakers, the estimated 3D points are scaled by the distance between the left and right shoulder points (joints 3 and 6 in Fig. 4) for the horizontal axis, by the distance between the neck and the torso points for the vertical axis (joints 0 and 1 in Fig. 4), and by the arm length (forearm + upper arm: joints 3 to 5 in Fig. 4) for the depth direction.

As the hand motions are complex and include a variety of patterns, we conducted a clustering analysis on the observed input hand motion data, for the stroke phases. A k-means method was adopted for the clustering analysis. Although the Euclidean distance is commonly used as a distance measure between two vectors in the k-means clustering computation, it is not appropriate for the present problem, since the input data is a time series of the hand position trajectories. To account for this issue, a dynamic time warping (DTW) based approach was used to allow non-linear time expansion/contraction of the input time series. The DTW barycenter averaging (DBA) approach [23] was used to estimate the average trajectory data.

In order to take the motion size into account, the hand position trajectory distance within the gesture interval was computed. The input data was then split in three size categories (small, medium and large) according to the computed trajectory distance. The thresholds between the size categories were set in a way to balance the data distribution among categories. Twenty clusters were set for each size category. Fig. 5 shows examples of trajectories for both hands in one of the large size clusters. In this example, large vertical movements of the left hand are grouped within the cluster.

The distributions of the gesture functions between the clusters were also analyzed. Although most of clusters included occurrence of multiple gesture functions, some of the clusters clearly showed predominance of specific gesture functions. These distributions will be used to associate text to gesture, as described in the next sub-section.

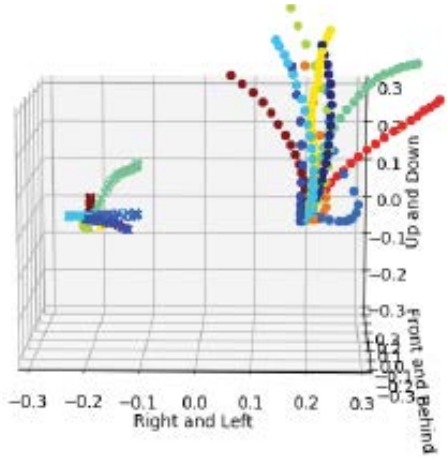

Figure 5. Examples of hand position trajectories for one of the clusters. Left panels are side views and right panels are back views. All axes are in meters.

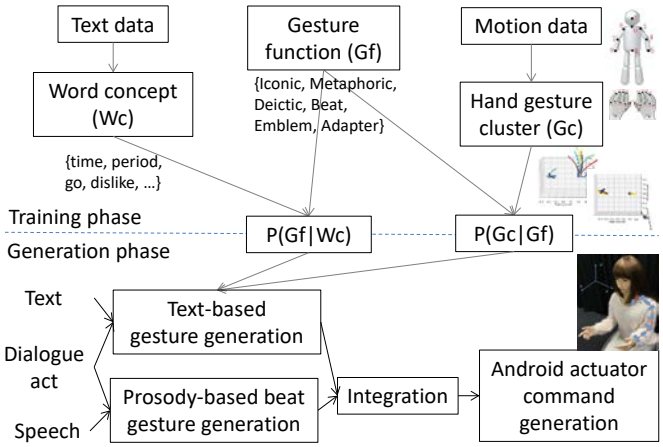

Figure 6. Block diagram of the training and generation phases for the proposed hand gesture generation method.

E. Association of text information and gesture and gesture modeling

Fig. 6 shows a block diagram of the proposed method for gesture generation. In this section, we explain the training phase, where gesture motions are modeled from the text information and gesture functions.

In order to create a generalized model for associating text to their accompanying gesture functions and motions, we make use of WordNet, which is a large lexical database where the words are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept [24]. For example, the hypernym “pleasant” is associated with the following hyponyms: “good”, “festive”, “happy”, “enjoyable”, “delightful”, “entertaining”. The idea is based on the assumption that words with similar meanings would be accompanied by similar gestures [25].

In the proposed approach for the text-based gesture generation, we associate words to concepts, concepts to gesture functions, and gesture functions to gesture motions. Conditional probabilities are estimated to model the relationships between concepts and gesture functions, and between gesture functions and gesture motion clusters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2018.2856281, IEEE Robotics and Automation Letters

Conditional probabilities between word concepts (Wc) and gesture functions (Gf) are computed for each concept using the whole dataset of the six speakers. The input text is firstly classified in part-of-speech units through morphological analysis using MeCab [26]. Then, the concepts are obtained using the Japanese WordNet provided by NPU/NICT [27], for the words classified as nouns, verbs, adverbs and adjectives.

Next, conditional probabilities between gesture functions (Gf) and gesture motion clusters (Gc) are computed for each speaker. This model is made speaker-dependent, considering that there is individuality in the manner for expressing a gesture. The computation is done by using the gesture function annotation data and the motion clusters extracted from the database. The gesture functions were restricted to iconic, metaphoric, and deictic, which are closely related to the text information. These models will be used in the gesture generation phase described in Section III.A.

F. Analysis of motion phases during hand gestures

Fig. 7 shows the distributions of the durations during different gesture phases (preparation, stroke, retraction and hold) for all speakers in the dataset.

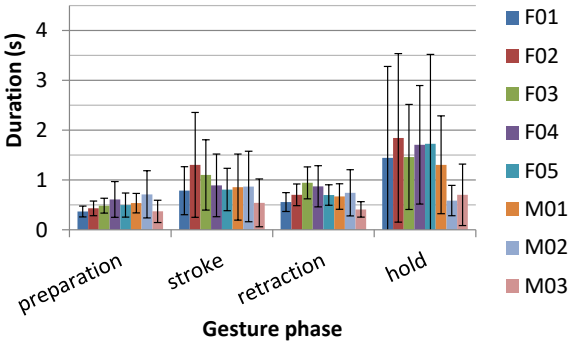

Figure 7. Distributions of durations of different phases during hand gestures.

It can be observed that the average durations for all speakers do not vary largely within a gesture phase. The preparation phase is the shortest (~0.5s on average), followed by retraction (~0.7s), stroke (~0.9s) and hold (~1.4s on average). These distributions will be used in the gesture generation timing control.

III. PROPOSED HAND GESTURE GENERATION AND EVALUATION IN AN ANDROID ROBOT

In this Section, we describe our proposed hand gesture generation method, which uses the analysis results presented in Section II.

A. Hand gesture generation method

We propose two gesture generation methods, one text-based and the other prosody-based, constrained by dialogue act information, as shown in the bottom part of Fig. 6. The prosody-based method is to account for beat gestures, which usually occur in prosodic peaks (pitch accents) and are less dependent on the speech contents [18].

Text-based gesture generation: the text transcriptions of the dialogue segments extracted from the database are used as input for the gesture generation. The concepts are extracted from the words in the input text, and the gesture function is sampled from the gesture function conditional probabilities P(Gf|Wc). A gesture will be generated if the sampled category is iconic, metaphoric or deictic. For the selected gesture function, the gesture cluster is then sampled from the gesture cluster conditional probabilities P(Gc|Gf).

From the sampled cluster, one gesture is randomly selected and the stroke phase is generated. The stroke phase is synchronized with the beginning of the word that triggered the gesture. Then, a gesture preparation phase is generated to move the hands from the rest position to the start point of the gesture stroke. A hold period of 1 to 2 seconds is set after the stroke phase, based on the distributions of Section II.F. If another gesture stroke is to be generated during the hold period, the hands are moved directly to the start position of the next gesture stroke without moving back to the rest position. If either the hold period or the utterance interval finishes, the gesture retraction phase is generated by moving the hands back to the rest position. The motions in the preparation and retraction phases are implemented by sigmoid functions with durations of 0.8s and 1.4s respectively. These values were set slightly longer than the distributions in Fig. 7, to avoid jerky motions by the android.

Prosody-based beat gesture generation: beat gestures consistently appear in prosodic peaks of focused words and phrases [18]. First, F0 (fundamental frequency) values are extracted from the input speech signal every 10ms. In the proposed method, prosodic peaks are detected in strong pitch accents, by firstly searching the position of F0 peak within a phrase. Then it is checked if the F0 falls down by more than 3 semitones within an interval of 500ms from the F0 peak, and if the F0 peak is higher than the speaker's average F0. The threshold for F0 fall detection is based on previous studies on tone detection [28].

If a prosodic peak is detected, a beat gesture is generated at that position, by moving the hands down and up by about 15 centimeters from the current hand position, so that a clear beat motion can be perceived. The beat duration was set to 0.8s, based on the analysis results in Fig. 7. This way, the beat gestures will be superimposed to the text-based gestures. If the hands are in the rest position, the hands are firstly moved to a gesture space (by sampling from the beat motion clusters) before making the down-up motion.

Dialogue act constraint: In addition to both text and prosody-based methods describe above, we also impose another constraint for gesture generation, based on the dialogue act analysis results in Section II.C. If the dialogue act category is interjectional utterance ("bc", "bc2") or filler ("f"), a gesture stroke will not be generated. In this way, gesture strokes that would be produced by word concepts or prosodic peaks will be filtered out in interjectional utterances and fillers.

B. Motion control method of an android robot

A female-type android robot, called ERICA, was used to evaluate the proposed motion-generation method. The current version of the android has a total of 44 air actuators, 13 DOFs

for the face, 3 DOFs for the head motion, 1 DOF for the base of the neck, 3 DOFs for the torso motion, and 12 DOFs for each hand/arm (2 for the shoulder, 3 for the upper arm, 2 for the forearm, 2 for the wrist and 1 for the thumb, 1 for the index finger, and 1 for the remaining three fingers). Fig. 8 shows the external appearance of the android and the actuator positions from the chest to the wrist. The proposed method was evaluated in the android ERICA, but the method can be generalized to any robot arm.

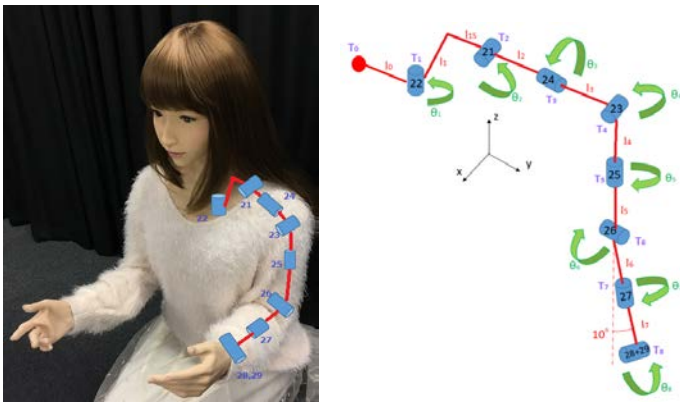

Figure 8. Actuators involved in the hand position control for the female-type android ERICA.

The actuators excluding the hand/arm ones are controlled according to the methods proposed in previous studies for lip control [4], laughter motion control [7], and surprise motion control [8]. In this paper, we explain the method implemented for hand motion control.

Firstly, the normalized 3D coordinate data are converted to the ERICA's coordinate space based on ERICA's geometry (inter-shoulder distance for horizontal axis, neck to torso distance for the vertical axis, and arm length for the depth axis.) The hand orientation is estimated from the plane formed by the joints 8, 10 and 13 in Fig. 4.

Then, the DH (Denavit-Hartenberg) convention [29] is adopted to attach individual coordinate frames to each actuator. Transformation matrices between the 3D coordinates of adjacent actuators are pre-computed given their relative positioning and orientations. The 3D coordinates of each actuator can then be estimated in sequence (starting from the chest actuator) from the rotation angles of the previous actuator.

In a conventional approach, the rotation angles would be estimated in sequence, in order to match the 3D positions of the shoulder, elbow, wrist and hands. However, due to differences in the lengths of the upper arm and forearm between the humans and the android, the resulting 3D positions will deviate from their target positions. Considering that the positions of the hands are more important than the positions of the elbows for the expression of a gesture, we fixed the target positions of the hands and shoulders and re-estimated the elbow position in such a way to fit the android upper arm and forearm lengths. In order to keep the arm form as close as possible from the target one, the elbow position was constrained to be over the plane formed by the target positions of the shoulder, elbow and wrist. The newly estimated elbow position is then used to estimate the rotation angles following the DH convention.

C. *Evaluation of the proposed motion-generation method*

The data of one of the female speakers who showed the largest gesture frequency in the three-party dialogue database was used for evaluation. Five dialogue segments with durations of 30 to 60 seconds, including chunks the subject is talking predominantly, were randomly selected from the dataset of the target female speaker. The text transcriptions and utterance intervals are used to generate the text-based gesture sequences, while the speech data and utterance intervals are used to generate the prosody-based beat gesture sequences, according to the proposed methods described in Section III.A.

For evaluating the proposed method, two motion types were generated, one based on text-only, and another based on text and prosody. For comparison, another two motion types were generated, one without hand motion, and the other by mapping the human motion directly to the robot motion control. In the direct mapping, the 3D skeleton data (shoulder, elbow, wrist and hand positions) extracted from the dataset were used as input targets, and the actuator command values were computed through the same procedure described in Section III.B. The four motion types generated for the evaluation experiments are listed in Table III.

Preliminary evaluation revealed that it was clearly unnatural that the arms and hands moved, while the head and torso only slightly moved in the vertical direction. In order to reduce the unnaturalness caused by the lack of head and torso motions, we mapped the head and torso angles measured from the IMU sensors of the speaker directly to the three (pitch, yaw and roll) actuators of the head and torso, for all motion types.

TABLE III. MOTION TYPES GENERATED FOR EVALUATION

Motion id	Description of the motion generation methods
0	Without hand motion
1	Direct mapping of human data
2	Text-based gesture generation
3	Text-based generation + prosody-based beat generation

The different motion types shown in Table III were generated for the same speech contents, so that twenty video clips were recorded in total (5 dialogue segments x 4 motion types) to use in the subjective experiments. In the experiment, participants watched the 20 videos and graded perceptual subjective scores after watching each video. The order of the videos was randomized, and participants were allowed to play them at most two times each.

The perceptual subjective scores were graded according to the questionnaire shown below. The numbers within parentheses were used to quantify the perceptual scores. The participants were also asked to write down the reasons of perceived motion unnaturalness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2018.2856281, IEEE Robotics and Automation Letters

Q1. What was the overall degree of naturalness (human-likeness) of the android's motion? (7 point scale: Very unnatural (-3), Unnatural (-2), Slightly unnatural (-1), Difficult to decide (0), Slightly natural (1), Natural (2), Very natural (3))
Q2. Were the gestures suitable with the speech contents? (7 point scale: very unsuitable (-3) ~ difficult to decide (0) ~ very suitable (3))
Q3. Were the hand movements natural? (7-point scale: very unnatural (-3) ~ difficult to decide (0) ~ very natural (3))
Q4. Was the gesture frequency suitable? (7-point scale: too few (-3) ~ suitable (0) ~ too much (3))
Q5. Was the gesture timing suitable? (7-point scale: too late (-3) ~ suitable (0) ~ too early (3))

Twenty subjects (male and female, aged from 20s to 60s) participated in the evaluation experiments.

D. *Evaluation results*

Fig. 9 shows the average subjective scores for the items that resulted in statistically significant differences between some of the motion types. Statistical significance tests were conducted through t-tests (* $p<0.05$, ** $p<0.01$).

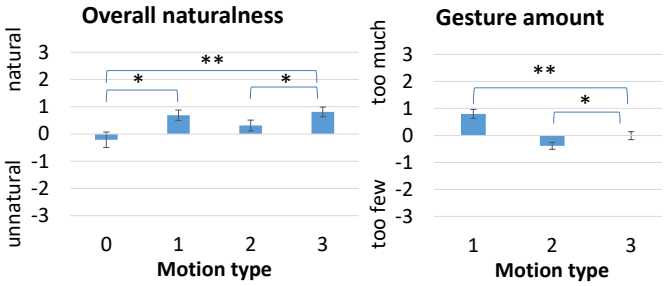

Figure 9. Subjective perceptual scores for each motion type (mean and standard errors) for Q1 (left panel) and Q4 (right panel).

The results in the left panel of Fig. 9 indicate that the absence of hand gestures (motion 0) and the text-only generation method (motion 2) received average naturalness scores close to 0 ("difficult to decide"), while the direct motion mapping (motion 1) and the proposed gesture generation method based on text and prosody (motion 3) received average naturalness scores close to 1 ("slightly natural").

Regarding the appropriateness of gesture amount (right panel of Fig. 9), the direct mapping (motion 1) received scores close to 1 ("slightly excessive"), the text-only method (motion 2) received slightly negative scores ("slightly insufficient"), and the proposed method based on text and prosody (motion 3) received scores around 0 ("suitable"). The introduction of the prosody-based beat generation was effective to compensate the lack of gestures by the text-based generation.

Regarding the subjective scores for the other questions, gestures were judged to be "slightly suitable" with speech contents, hand movements were judged to be "slightly natural" and gesture timing was judged to be "suitable" for the proposed method (motion 3). No statistically significant differences were found between the other motion types.

IV. Discussion

The evaluation results indicated that the proposed method could generate hand gestures with overall naturalness scores compatible to the direct motion mapping. However, the reasons for unnaturalness judged by the participants were different for different motion types.

Most of the unnaturalness reasons can be attributed to the current constraints of the robot hardware. These include motion speed limitations (especially when the whole arms have to move against the gravity), lack of DOFs for hand shaping (only 3 DOF for the control of all five fingers), limited angle range for the hand and forearm bending and arm turning for the outside direction. In particular, the lack of DOFs for hand shaping leads to insufficient expression of emblems and deictic gestures, while the limitations in the angle range for the forearm bending leads to insufficient expression of (self-touch) adapters (since the android hands cannot touch her face). The limitations on motion speed and range and the lack of DOFs for the hand shaping are source of motion unnaturalness in all motion types, while the insufficient expression of adapters was a strong reason for perceived unnaturalness in the direct motion mapping (motion 1). The judgements on "excessive" gesture amount in the direct motion mapping may be because the meaning of an adapter motion has not been understood due to insufficient motion range.

On one side, improvements on hardware would lead to higher naturalness scores. On the other side, given the hardware constraints, the robot actuators should be controlled to avoid gestures that cannot be properly expressed and/or to convert to motions that can be expressed. Further, it would be also be worth to check motion generation in CG avatars, which are free from hardware constraints. These are topics for future investigations.

Regarding head and torso motions, these were directly mapped from the speaker's motion in the present evaluation, in order to reduce the unnaturalness caused by the lack of body movements during hand gestures. For automation of motion generation, the coordination of head and torso movements accompanied with the hand gestures is another topic to be investigated.

In this study, we conducted video-based evaluation instead of face-to-face evaluation, since the speech contents are fixed for all conditions and participants do not interact with the robot, so that the influence of other factors such as eye gazing can be avoided from the evaluation. In future studies, we intend to conduct face-to-face evaluation after solving eye gazing control issues. Nonetheless, we can expect similar results to the video-based evaluation in this study, given that in a previous study on head motion control, subjective scores did not change between video-based and face-to-face interaction [5].

Finally, the present analysis data is limited in size since we used Kinect sensors to capture 3D joint data. Although estimation accuracies might decrease, other alternatives to increase the data would be by estimating 3D poses from 2D video data, for example by using PnP (Perspective-n-Points) techniques.

V. Conclusion and Final Remarks

In this study, we first analyzed a multimodal database of

human-human dialogue interactions, and investigated the relations between the occurrence of hand gestures and speech in different dialogue act categories. The analysis results indicated that hand gestures occur with highest frequency in turn-keeping phrases, and seldom occur in backchannel-type feedback utterances.

Clustering analysis was conducted on gesture motion data, and conditional probabilities were trained to associate text information and gesture motion clusters, through word concepts and gesture functions. We then proposed a data-driven gesture generation method based on text, prosody, and dialogue act information, by taking the dialogue act analysis results and gesture motion models into account.

We implemented a hand motion control method from the target hand positions in the android ERICA, and evaluated the effectiveness of the proposed gesture generation method through subjective experiments. The overall motion was evaluated as "slightly natural" by the proposed method, even under the robot hardware constraints. Compared to the direct mapping of hand motion and the text-only based method, the proposed method using text and prosody generated appropriate amount of gestures.

Future works include coordination of head and torso movements with hand gestures, and motion generation strategies accounting for the robot hardware constraints.

ACKNOWLEDGMENT

We thank Taeko Murase, Miki Okuno, Megumi Taniguchi, and Kyoko Nakanishi for their contributions to the speech and motion data analyses. We also thank Chaoran Liu for discussions regarding gesture modeling, and Takashi Minato for contributions to the android calibration.

REFERENCES

[1] C. Ishi, H. Ishiguro, and N. Hagita. Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication* 57, No.2014, 233–243, June 2013.

[2] C. Ishi, H. Hatano, and H. Ishiguro. "Audiovisual analysis of relations between laughter types and laughter motions," Proc. of *the 8th international conference on Speech Prosody*, pp. 806-810, May, 2016.

[3] C. Ishi, T. Minato, and H. Ishiguro. "Motion analysis in vocalized surprise expressions," Proc. *Interspeech 2017*, pp. 874-878, Aug. 2017.

[4] C. Ishi, C. Liu, H. Ishiguro, N. Hagita. "Evaluation of formant-based lip motion generation in tele-operated humanoid robots," Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura, Portugal, pp. 2377-2382, October, 2012.

[5] C. Liu, C. Ishi, H. Ishiguro, and N. Hagita. Generation of nodding, head tilting and gazing for human-robot speech interaction. *International Journal of Humanoid Robotics*, vol. 10, no. 1, Jan. 2013.

[6] S. Kurima, T. Minato, C. Ishi, and H. Ishiguro. (2017) Novel speech motion generation by modelling dynamics of human speech production. *Frontiers in Robotics and AI*, Vol.4, Art.49, 1-14, Oct. 2017.

[7] C. Ishi, T. Funayama, T. Minato, and H. Ishiguro. "Motion generation in android robots during laughing speech," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3327-3332, Oct., 2016.

[8] C.T. Ishi, T. Minato, and H. Ishiguro. Motion analysis in vocalized surprise expressions and motion generation in android robots. IEEE Robotics and Automation Letters, Vol.2, No.3, 1748 - 1754, July 2017.

[9] A. Kendon. "Gesticulation and speech: two aspects of the process of utterance," In M. R. Key (ed), *The Relationship of Verbal and Nonverbal Communication*, The Hague: Mouton and Co, pp.207-227, 1980.

[10] D. McNeill. Hand and Mind: What Gestures Reveal about Thought, Chicago and London: The University of Chicago Press, 1992.

[11] S. Kita. How representational gestures help speaking. In D. McNeill (ed.), *Gesture and Language*, pp. 162-185. Cambridge: Cambridge University Press, 2000.

[12] D. McNeill. Gesture: a psycholinguistic approach, in *The Encyclopedia of Language and Linguistics*, eds Brown E., Anderson A. (Amsterdam; Boston: Elsevier;), 58–66, 2006.

[13] J. Cassell, M. Stone, and H. Yan. (2000) "Coordination and context-dependence in the generation of embodied conversation," In Proc. of *the First International Conference on Natural Language Generation*, 2000.

[14] T. Sowa and I. Wachsmuth. "A model for the representation and processing of shape in coverbal iconic gestures." In Proc. *KogWis05*, pp. 183–188, 2005.

[15] K. Bergmann, and S. Kopp. "GNetIc - Using Bayesian Decision Networks for Iconic Gesture Generation" in Proc. of *the 9th International Conference on Intelligent Virtual Agents*, vol. 5773, (Berlin/Heidelberg, Germany: Springer), pp.76-89, 2009.

[16] M.E. Sargin, O. Aran, A. Karpov, F. Ofli, Y.Yasinnik, S. Wilson, E. Erzin, Y. Yemez, A.M. Tekalp. "Combined gesture–speech analysis and speech driven gesture synthesis," In Proc. of *IEEE International Conference on Multimedia*, 2006.

[17] Y.I. Nakano, M. Okamoto, D. Kawahara, Q. Li, T. Nishida. "Converting Text into Agent Animations: Assigning Gestures to Text," In Proc. *Human Language Technology Conference of the North American Association for Computational Linguistics*, pp. 153–156, 2004.

[18] D. Loehr. Gesture and Intonation. Washington DC: Georgetown University, PhD Dissertation, 2004.

[19] S. Levine, C. Theobalt, V. Koltun. "Real-Time Prosody-Driven Synthesis of Body Language," In *SIGGRAPH Asia*, 2009.

[20] M. Neff, N. Toothman, R. Bowmani, J.E. Fox Tree, and M. Walker. "Don't Scratch! Self-adaptors Reflect Emotional Stability," in Proc. *10th International Conference on Intelligent Virtual Agents* (IVA 2011), Sep. 2011, pp.398-411.

[21] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[22] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330-1334, 2000.

[23] F. Petitjean, A. Ketterlin, and P. Gancarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, Vol. 44, No. 3, pp. 678–693, 2011.

[24] https://wordnet.princeton.edu/

[25] Y. Kadono, Y. Takase, Y. Nakano. "Analyzing Metaphoric Gestures towards Automatic Gesture Generation," *The 29th Annual Conference of the Japanese Society for Artificial Intelligence*, 2015. (in Japanese)

[26] http://taku910.github.io/mecab/

[27] http://compling.hss.ntu.edu.sg/wnja/

[28] C. Ishi, H. Ishiguro, and N. Hagita. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50(6), 531-543, June 2008.

[29] J. Denavit and R. S. Hartenberg. A kinematic notation for lower-pair mechanisms based on matrices. *Trans. ASME, J. Appl. Mech.*, Vol. 22, No. 2, pp. 215 – 221, 1965.