Robotics and Autonomous Systems 155 (2022) 104154

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Generation of co-speech gestures of robot based on morphemic analysis[☆]

Yu-Jung Chae[a,e], Changjoo Nam[b], Daseul Yang[c], HunSeob Sin[d], ChangHwan Kim[e,*], Sung-Kee Park[e]

[a] HCI & Robotics Division of Nano & Information Technology, University of Science and Technology, Daejeon, Republic of Korea
[b] Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea
[c] Department of medical robot UX, Kohyoung Technology, Seoul, Republic of Korea
[d] R&D Center, SFA Engineering, Gyeonggi-do, Republic of Korea
[e] Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 14 February 2020
Received in revised form 6 February 2022
Accepted 21 May 2022
Available online 3 June 2022

Keywords:
Human–robot interaction
Co-speech gesture generation
Machine learning
Morphemic analysis
Word embedding

ABSTRACT

We propose a methodology for a robot to automatically generate felicitous co-speech gestures corresponding to robot utterances. First, the proposed method determines the part of a given robot utterance, where the robot makes a gesture by doing a morphemic analysis on the sentence of utterance. The part is herein called an *expression unit*. The method then predicts a gesture type to characterize the expression unit in the sense of conveying thoughts and feelings. The gesture type is selected from the four types of iconic, metaphoric, beat, and deictic categorized by McNeill by performing morphemic analysis on the sentence. A gesture proper to the gesture type is retrieved from a database of motion primitives that are built with predefined a limited number of words. For retrieving, Word2Vec is applied to estimate word similarity between the predefined words in the database and words in the expression unit such that the method can deal with an arbitrary sentence and generate an appropriate gesture for similar words in meaning.

The proposed method showed 83% accuracy in determining expression units and gesture types for a set of sentences in Korean. Furthermore, a user study on feasibility has been performed with a humanoid, NAO, and received positive evaluations in terms of anthropomorphism for the robot.

1. Introduction

People engage in many actions if they participate in interaction with others. Among such actions, gestures are a visible activity to convey their thoughts and feelings appropriately for interaction while speaking [1]. We call these gestures *co-speech gestures* [2] or *co-verbal gestures* [3]. Positive effects of co-gestures in human–robot interaction (HRI), especially in non-verbal communication, have been studied. In this regard, multimodal communication involving speech and gesture enhances comprehension compared to unimodal communication [3]. Robot gestures could make it easier to perform a difficult task than using verbal behavior (i.e., speech) only [4]. It was noticed that robot gestures enhance the human ability to recall more information corresponding to its utterances. Based on these early studies, it becomes essential and fundamental in HRI areas how co-speech gestures of a robot are generated during interaction with a person.

Most common methods produce gestures to a spoken dialogue of a robot or a virtual human by matching with the texts predefined by a user [5–8]. In these methods, the duration (i.e., start and end time of a gesture during a speech) and specification (e.g., palm orientation, hand location) of gestures were assigned at the specific part of the dialogue. Kim et al. [9] selected a single word that is matched with predefined five patterns of Parts-of-speech (PoS) like 'noun and post position'. Gestures were generated from a word-gesture database that retrieved a gesture from a predefined single word in the selected words. Huang et al. [10] coordinated robot behaviors (i.e., speech, gaze, gestures) by learning alignment parameters of human behaviors for the particular narration scenario based on *Dynamic Bayesian Network (DBN)*. They used a limited number of patterns of words as speech features that were employed by a theory of gesture types (details on gesture types are described in Section 2). Ng-Thow-Hing et al. [11] generated gestures of a humanoid according to a style tag (e.g., excited) with a sentence. In this method, candidate gestures were generated for each gesture type by matching

[☆] This work was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIP) (No. CRC-20-04-KIST.)

[*] Corresponding author.
E-mail addresses: 19cyj88@gmail.com (Y. Chae), ckim@kist.re.kr (C. Kim).

https://doi.org/10.1016/j.robot.2022.104154
0921-8890/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

predefined word patterns (e.g., a word pattern 'between A and B' was defined as a gesture type 'iconic'). According to predefined relations between a style tag and a gesture type, gestures were selected among the candidates and then modified. Even though these methods might be useful for generating gestures corresponding to a limited number of robot spoken dialogues, it still needs to predefine all the necessary information such as specification or words manually to retrieve gestures, which could be costly and time-consuming.

Kim et al. [12] generated gestures for a given sentence by using three punctuation marks (i.e., period(.), interrogation(?), exclamation(!)). The database of the method has two gestures according to each mark and a gesture is randomly generated by one of the two gestures depending on the mark in the sentence. Ferstl et al. [13] extracted prosodic features for points of pitch peaks on robot speech. Gesture parameters were then obtained by a predefined map having the relations between the prosodic features and parameters (i.e., velocity, initial acceleration, size, arm swivel, and hand opening). A gesture that had similar parameters was retrieved from a database. Although these methods might have benefits in some applications, they could not be expressive enough to deliver a text's meaning properly. Pérez-Mayos et al. [14] suggested three different approaches and conducted individual evaluations for each approach. The first one generated gestures based on predefined rules using keywords with PoS attached; The second one analyzed a pitch curve on robot speech and generated gestures that have the most similar curves from a database involving relations between the pith curve and gesture; The last one was combined by the first and second approaches. Beat and the rest of the gestures were generated based on the pitch curves and keywords with PoS, respectively. In this experiment (i.e., user study), the first one obtained the highest score. We assume that generating gestures by analyzing the robot's synthesized voice may have difficulty conveying a sufficient meaning to the user.

Ahn et al. [15] trained relations between language and human actions based on a generative adversarial network (GAN) using 29770 pairs of human action and sentence annotation. Yoon et al. [16] also trained RNN (Recurrent Neural Network) by using 52 h of TED talks to train the relations between language and human actions. These models could generate human gestures from a new text, and then human gestures were converted for a humanoid robot. However, these models need a large number of pairs of human actions and texts as a training dataset to generate complex gestures. In practice, it is difficult to collect such a large dataset, which regards detecting the start and end of every single gesture clearly from original video clips.

According to the studies mentioned above, the generation of co-speech gestures seems to be highly dependent on predefined patterns of words or morphological patterns of texts. In addition, those methods could not generate gestures corresponding to new sentences containing undefined patterns of words. Other methods have focused on encoding a text into a human action, which could be applied only to a humanoid robot. It was observed that a large number of training data from humans were necessary to generate robot gestures. Such methods may have difficulties in being applied to robots in various shapes and geometries. To minimize the dependency of word or morphological patterns and apply for various types of robots, we propose a methodology of generating co-speech gestures by determining parts of executing gestures and generating gestures for new sentences.

Details on all the procedures above are presented in the rest of the paper. Section 2 introduces the conceptual mechanism of co-speech gestures of human based on McNeill's theories. In Section 3, the main contributions of this paper for the algorithmic procedure of generating robot gestures followed. Experiments and results will be discussed in Sections 4 and 5. Discussion and limitations for our method will be dealt with in Section 6. Lastly, conclusion is described in Section 7.

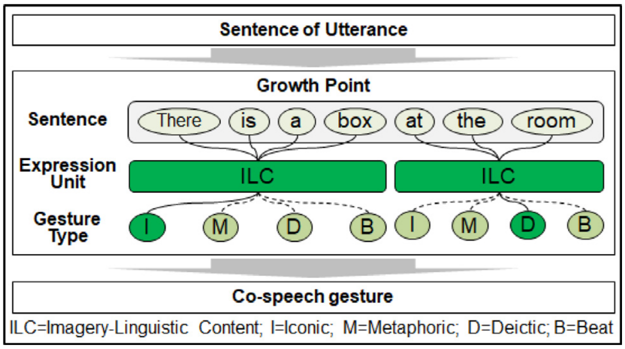


Fig. 1. Conceptual framework for co-speech gesture generation.

2. Concept of co-speech gestures of human

2.1. Conceptual framework of gesture generation

We employ McNeill's *growth point* theory about human speech-synchronized gestures [17–19] to present a conceptual framework for the generation of co-speech gestures. The growth point is the smallest and initial idea unit of thinking-for-speaking, where it combines imagery and linguistic categorial contents and initiates cognitive events like co-speech gestures. In the theory, unpacking the growth point into grammatical structures could make the combination of the imagery-language code. Gestures could be formulated in coordination with contents of utterance from unpacking the growth point [17]. Such gestures could be classified into four categories [18]: *Iconic*, *Metaphoric*, *Deictic*, and *Beat*; Iconic type expresses a concrete image (e.g., shape of an object) and action (e.g., holding); Metaphoric type indicates an abstract state (e.g., upper) and action (e.g., drop); Deictic type describes a movement to point to or indicate an object; Beat type contains repetitions of gestures to emphasize some part of speech.

Based on the theory above, we aim to realize a conceptual mechanism for the generation of co-speech gestures in Fig. 1. The growth point can provide imagery-linguistic contents (ILC) in the sentence of utterance while speaking. Such imagery-linguistic contents may be comprised by words or phrases[1] in a grammatical point of view. Each imagery-linguistic content is semiotically characterized with a gesture type (one of the four types) as McNeill proposed. Gesticulation for the imagery-linguistic content can ensure the meaning, mood, and better understanding. In our proposed framework, an imagery-linguistic content is called an *expression unit*, which is a minimal unit to bring up imagery for gesticulation. A single gesture can be generated into an expression unit. To characterize an expression unit with a semiotic manner, we determine one of such four gesture types as *Iconic*, *Metaphoric*, *Deictic*, and *Beat* for each expression unit. When the same word is expressed as a gesture in a sentence, it is expressed as a different gesture depending on the determined gesture type.

For generating co-speech gestures of a robot, our method segments a sentence of robot utterance into expression units, and then the gesture types for the expression units are obtained. Some of the expression units are then selected, and gestures in a database are retrieved by considering keywords and gesture types

[1] A phrase is a group of words (e.g., to play the guitar.) [20] and is categorized into noun (e.g., the sports car), verb (e.g., might enjoy a message), gerund (e.g., walking in the rain), infinitive (e.g., to see the stage), appositive (e.g., $\cdots$,my love of my life, $\cdots$), participial (e.g., washed with my clothes), prepositional (e.g., on the table), and absolute phrase (e.g., weather permitting) [21].

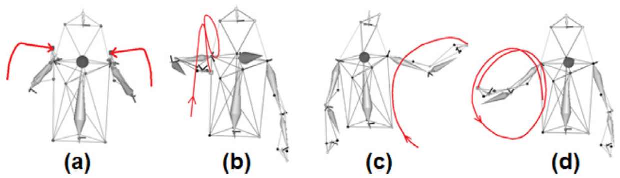


Fig. 2. Snapshots of gesture four gesture types of human motion capture data (front-side view): (a) Iconic (e.g., small) (b) Metaphoric (e.g., eat) (c) Deictic (e.g., pointing to the left) (d) Beat (e.g., so much).

in each expression unit. The gestures are finally synchronized with the sound file of robot utterance and executed through the robot.

2.2. Redefinition of gesture types

We refine the definitions of the four gesture types[2] introduced by McNeill [17] prior to describing the procedures, since his definitions may have ambiguities to be applied to a robot. In the refinement, iconic gestures are used to depict images of objects or spaces (e.g., square box, small village, big house) or static states involving feelings or mood (e.g., sunny day, positive thought) in small and slow motions. Metaphoric gestures express actions (e.g., eating food, lifting heavy weights) or dynamic states including movements of an object (e.g., fall out of my pocket, drive up the price) in large and fast motions. For example, an expression unit having the word 'bird' can be expressed with different gestures according to the gesture type. In the case of the iconic type, a gesture of drawing a small circle may be created to describe the bird's shape. On the other hand, a gesture representing the bird's movement may be generated in the metaphoric type. Deictic gestures mean pointing gestures to indicate spots or directions of an object (e.g., over there, the box). Finally, beat gestures can strongly highlight important parts in an utterance of a robot (e.g., very, so much, so beautiful). These gestures are also generated in the parts of an utterance including repeated words (e.g., one by one, step by step). Fig. 2 shows snapshots of each gesture type that we collect human gestures based on our definition using motion capture cameras to make robot gestures.

3. A method for generation of gestures

The overview of the proposed method is given in Fig. 3. Morphological analysis is first performed to obtain such grammatical information as morphemes of each one of words and its PoS, and constituents of the words.[3] Second, we conduct gesture planning by using patterns of PoS and constituents. Specifically, a sentence is segmented into expression units based on the phrase[1], and a gesture type is then predicted for each expression unit.

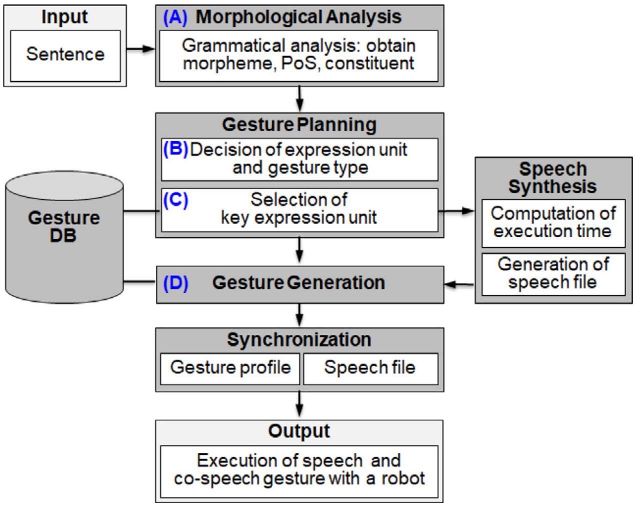
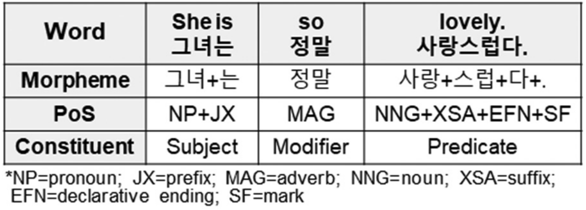

Input
Sentence
(A) Morphological Analysis
Grammatical analysis: obtain morpheme, PoS, constituent
Gesture Planning
(B) Decision of expression unit and gesture type
(C) Selection of key expression unit
Gesture DB
Speech Synthesis
Computation of execution time
Generation of speech file
(D) Gesture Generation
Synchronization
Gesture profile
Speech file
Output
Execution of speech and co-speech gesture with a robot

Fig. 3. Procedure of co-speech gesture generation.

Table 1
Example of results from KKMA.



Word	She is 그녀는	so 정말	lovely. 사랑스럽다.
Morpheme	그녀+는	정말	사랑+스럽+다+.
PoS	NP+JX	MAG	NNG+XSA+EFN+SF
Constituent	Subject	Modifier	Predicate

*NP=pronoun; JX=prefix; MAG=adverb; NNG=noun; XSA=suffix; EFN=declarative ending; SF=mark

Key expression units, where gestures will be executed with the utterance, are selected. Gestures are retrieved from the gesture DB considering gesture types and morphemes for selected key expression units. Lastly, each gesture is adjusted by matching with the execution time of a key expression unit. All gestures are interpolated and simultaneously played with the speech file through actuators and speakers. More details on the procedures mentioned above are described in the following subsections.

3.1. Morphological analysis

For a given sentence of robot utterance to reduce the dependency of words, grammatical information[3] such as morphemes, PoS, and constituents is used for gesture planning. Such tags for representing morphemes, PoS, and constituents are obtained from the morphological analysis. For the case of a sentence in Korean, *KKMA*[4] [26] morphological analyzer is used. As seen in Table 1, *KKMA* morphological analyzer generates morphemes and provides the PoS for the morphemes. In addition, the constituent for each one of the words in the sentence is obtained respectively. In the example of "She is so lovely". in the table, the example consists of three words in Korean, which are analyzed with 7 morphemes, including the mark of period(.), and 7 corresponding PoS. Three constituents for the tree words are obtained. For a sentence that consists of N words, N constituents and R PoS ($R \geq N$) are obtained.

For a sentence in English, few morphological analyzers have been developed [27–29]. Once morphemes and PoS are obtained for an English text, the same procedures as followed in the coming subsections could be applied to generate co-speech gestures. This is considered as our future work.

[2] Gesture types are defined by McNeill as we mentioned in Section 2 [22]. He reported the density of space usage on hands for each gesture type [17]. Iconic gestures express concrete images or actions, and the hand positions mainly move in front of the torso. Metaphoric gestures present abstract meaning or occupied space. When using these gestures, most of the hands use the lower part of the body. Deictic gestures are extended index fingers to point a reference such as objects or locations. Beat gestures are rhythmical actions such as the gesture 'hands up and down' to highlight some parts of speech.

[3] A word in a sentence has an immediate constituent [23] such as subject and predicate. Subsequently, a word can be divided into morphemes that are the smallest unit of language. A morpheme [24] has its own meaning such as a word or a part of a word. For example, the sentence "He likes cats". can be divided into 5 morphemes "he", "like", "-s", "cat", and "-s". Lastly, each morpheme has a PoS (Part-of-Speech), which is one of the grammatical groups such as noun, verb, and adjective [25].

[4] KKMA [26] has tags of 59 PoS and 15 constituents. This analyzer has 75% accuracy, measured using the 5670 sentences from news, blogs, and reviews.

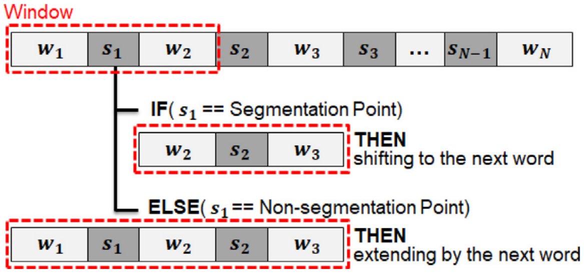

Window
w_1 s_1 w_2 s_2 w_3 s_3 … s_{N-1} w_N
IF(s_1 == Segmentation Point)
w_2 s_2 w_3
THEN
shifting to the next word
ELSE(s_1 == Non-segmentation Point)
w_1 s_1 w_2 s_2 w_3
THEN
extending by the next word

Fig. 4. Procedure of creating windows to determine expression units for a sentence that contains N words, where the w_i and s_i denote the ith word and space between words, respectively.

3.2. Determination of expression units and gesture types

By using PoS and constituents from results of morphological analysis, a sentence is segmented into one or more expression units, where an expression unit is matched with a phrase or word, and a gesture type is then predicted for each of the expression units. To deal with various grammatical structures of sentences, expression units and their gesture types are determined by using a learning scheme with patterns of PoS and constituent. To implement the learning scheme, we built tree models based on *Random Forest*. It is known that Random Forest has a small variation in accuracy for untrained data since it is an ensemble learning method comprised of tree-structured classifiers [30]. For applying Random Forest, multiple windows and feature vectors are required to make input data. This data is used for two Random Forests (RF): (1) determining expression units (RF^U) and (2) predicting gesture types (RF^T). More details on all the procedures follow.

Window. Feature vectors such as PoS and constituents are defined from windows that are to include several words and spaces between words in a sentence. For each space between words, Random Forest for expression units, RF^U, determines whether a breaking point (so-called a segmentation point) or not. Fig. 4 shows windows to deal with these two cases. To create windows, (1) a window is initially defined to cover the first two words, w_1 and w_2, and space s_1 between words. (2-A) If the space s_1 between words is expected to be a segmentation point (i.e., this case means that the words w_1 and w_2 can not be a phrase.), the word w_1 before the breaking point s_1 becomes an expression unit. A new window is shifted to the next word w_3 including the remaining word w_2 in the previous window. (2-B) If the space s_1 between words is determined not to be a segmentation point (i.e., this case means that the words w_1 and w_2 can be a phrase.), it extends to include the next word w_3 such that the window has three words, w_1, w_2, and w_3, and two spaces, s_1 and s_2. (3) next space s_2 between words is checked to determine segmentation or non-segmentation. These procedures, such as (2) and (3), are sequentially repeated for all spaces between words. In Random Forest to determine gesture types RF^T, a window is matched with an expression unit since a gesture type is predicted for a single expression unit.

Feature representation. To make input data for RF^U and RF^T, a feature vector is defined using PoS and constituents from a window. The feature vector $\mathbf{x}$ is given for a window as follows,

$$\mathbf{x} = [\mathbf{v_p}, \mathbf{v_c}, N_p, N_c]^T \tag{1}$$

where a sub-vector $\mathbf{v_p}$ contains Q components that each component has a zero value or one of PoS. Another sub-vector $\mathbf{v_c}$ includes P components that each component has a zero value or one of constituents. Scalars N_p and N_c indicate the number of PoS and that of constituents in the window, respectively. Dimension of a feature vector, M, is fixed. Therefore, when the number of features is less than Q or P, zeros added to the remaining a size of feature vector $\mathbf{x}$. For example, consider the first window that has 'She is' and 'so' as shown in Table 1. There are two words (in Korean) but three morphemes in the window such that $N_p = 3$ and $N_c = 2$. When the dimensions of $\mathbf{v_p}$ and $\mathbf{v_c}$ are larger enough, the first three components of $\mathbf{v_p} = [\text{NP}, \text{JX}, \text{MAG}, 0, \dots, 0]^T$ are the three PoS and the remaining are all zeros. In the same way, $\mathbf{v_c} = [\text{Subject}, \text{Modifier}, \text{Predicate}, 0, \dots, 0]^T$.

The number of words in a window can vary as the window shifts or extends. This affects the dimension of a feature vector for the window as well, which makes difficulties in defining the size of feature vector. To resolve this, we analyze a training dataset and set the dimension of a feature vector large enough to cover all the words in a window. In this work, the dimensions of $\mathbf{v_p}$ and $\mathbf{v_c}$ are 20 and 10 to build the tree model for expression units such that the dimension of $\mathbf{x}$ is 32 ($M = 32$). We assume that the number of words of a window, which is equal to that of constituents, is 10 and each word has two morphemes. For another tree model to expect gesture types, the dimensions of $\mathbf{v_p}$ and $\mathbf{v_c}$ are set 10 and 5 so that the dimension of $\mathbf{x}$ is 17 ($M = 17$).

Random Forest. For determining expression units and their gesture types, Random Forest that is described below is applied to RF^U and RF^T (i.e., the same procedure based on Random Forest is applied.).

For training, a training data set is given as $\mathcal{L} = \{(\mathbf{x}, y)\}$, where $\mathbf{x} = [x_1, x_2, \dots, x_M]^T$ denotes a feature vector, where y indicate a class label from a set of class labels Y ($y \in Y$). In RF^U, a set of class labels $Y = \{\text{Segmentation}, \text{Non-seg- mentation}\}$ is defined to represent breaking points between words in a sentence. RF^T has a set of class labels $Y = \{\text{Iconic}, \text{Metaphoric}, \text{Deictic}, \text{Beat}\}$ to indicate gesture types.

Random Forest has multitude of tree-structured classifiers. To train each classifier, a sub-data set from the training data set is used. The process of creating sub-data sets is as follow. The W sub-data sets $\bar{\mathcal{L}}_k = \{(\bar{\mathbf{x}}, y)\}$ for $(k = 1 \sim W) : \bar{\mathcal{L}}_k \subset \mathcal{L}$ are built from the training data set $\mathcal{L}$ by sampling with replacement. Dimensions of the random feature vector $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_Z]^T$ is reduced ($M > Z$) and comprised of randomly selected components from $\mathbf{x}$. Each class label y is identical to those of the training data set $\mathcal{L}$. Each tree-structured classifier is built based on Decision Tree using a sub-data set [31]. Random Forest $H = \{h(\bar{\mathcal{L}}_k),\ k = 1 \sim W\}$ has W tree-structured classifiers, where $h(\bar{\mathcal{L}}_k)$ denotes a kth tree-structured classifier that is built by $\bar{\mathcal{L}}_k$.

In this work, we built 50 and 40 tree-structured classifiers for $RF^U(W = 50)$ and $RF^T(W = 40)$, respectively. In RF^U, each classifier was trained by 6 dimensional random vector $\bar{\mathbf{x}}$. In RF^T, 4 dimensional random vector was applied. RF^T uses less number of classifiers and dimension of feature vectors than RF^U, since RF^U has less dimension of feature vectors than RF^T. After training, all classifiers predict a result y, which is one of the class labels for a feature vector $\mathbf{x}$. The most popular class label is then finally selected like a majority vote.

3.3. Selection of key expression units

It should be possible to generate gestures even when a sentence containing words not defined in a robot is inputted. In addition, key expression units should be selected since executing gestures for all the expression units are not practical because motions need to take much longer than speech. Therefore, we score in each expression unit based on semantic similarity. Key expression units are selected in the order of the highest rank

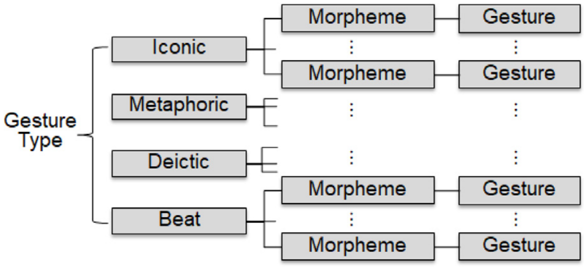


Fig. 5. Structure of gesture DB.

while considering whether or not the execution time conflicts with selected expression units (i.e., expression units are able to be selected in the order of the highest rank if there is no collision with the previously selected expression units.). Lastly, appropriate gestures are assigned for each key expression unit by using semantic similarity. Details are given below.

Gesture DB. Gesture DB is comprised of four sections according to gesture types such as iconic, metaphoric, deictic, and beat, as seen in Fig. 5. In each section, relations between a morpheme and gesture are predefined. A given gesture type can retrieve all morphemes. Moreover, a given gesture type and morpheme extract a particular gesture.

Gestures would be different depending on the geometry of a robot. In a non-humanoid robot, gestures are able to be made by some tools like 3dMax. In this work, gestures are recorded using motion capture cameras and converted to a humanoid robot NAO. To collect human gestures, we provided an actor with sentences that indicated where gestures were needed.

Semantic similarity. To obtain a score of semantic similarity and assign an appropriate gesture for each expression unit, we apply a Word2Vec model based on a continuous skip-gram. This model is based on a neural network that learns context words for a center word for a given sentence [32]. After training this model, words are converted from a word to a vector. In addition, words that have a similar meaning (i.e., topic group) are clustered in the vector space (e.g., a word 'apple' is closer to a word 'fruit' than a word 'car' in the vector space.).

In this study, we use morphemes instead of words because it is not essential to consider the tenses such as 'work' and 'worked'. In addition, to cover various topics such as travel, health, and daily conversation, a corpus is collected from sentences by newspaper articles, lectures, interviews, and daily conversation records, which the National Institute of Korean Language provides. It is to make the proposed method applicable to various service domains, which require the use of a large amount of vocabulary. As a result, this corpus is built by 69,243 morphemes.

Fig. 6 shows a procedure to train the Word2Vec model. The input and output layers have 69,243 nodes that are mat- ched with the number of morphemes in a corpus. The hidden layer is defined by 50 nodes. A relation between input and output morphemes is learned, and then a weight matrix between the input and hidden layers becomes a Word2Vec model, which works like a lookup table. After training, this model can convert a morpheme into a vector and estimate semantic similarity between two morphemes represented as vectors. The semantic similarity is estimated using cosine similarity, which is $\{(\mathbf{a} \cdot \mathbf{b}) \setminus (\|\mathbf{a}\| \|\mathbf{b}\|)\}$ by given vectors **a** and **b**.

To score expression units, all morphemes, connected with a gesture type of each expression unit in the gesture DB, are first retrieved. A score of semantic similarity is computed between morphemes from the gesture DB and morphemes of an expression unit. Gestures, matched with morphemes that have the highest similarity in the gesture DB, would be assigned for each key expression unit.

4. Accuracy comparison on determining expression units and gesture types

To verify the performance of our method, we conducted an accuracy comparison of the tree models for determining expression units and gesture types. In this section, a dataset and parameters for training Random Forests are also described.

Training dataset. In the models of determining expression units and predicting gesture types, grammatical elements su- ch as PoS and constituents are used as features. In this experiment, 857 sentences from a textbook for Korean elementary schools [33] containing various grammatical structures were collected. A Korean literature expert working in the field for the past ten years made annotations on the expression units of sentences. Herein a text is manually separated by a delimiter such as period, comma, and question mark except for some parts (e.g., time 2:30, degree 36.5) since delimiters help segment the text into meaningful units pragmatically. Then, the separated parts are segmented once again based on the phrase to deal with long texts. In addition, another expert who has studied human gestures attached annotations to each expression unit. As a result, 3014 class labeled data was corrected for the determination of expression units and gesture types. Using these data, we automatically created and used 2096 non-segmentation data for training a model of determining expression units (if there is no segmentation label in a space between words, it would be non-segmentation data.).

Compared methods. In addition to the Random Forest, three more methods such as Naive Bayes [34], SVM (Support Vector Machine) [35] and Decision Tree [36] were implemented to compare the performance. These four methods are implemented using WEKA library [37], which is a collection of machine learning algorithms. We applied k-fold cross validation[5] with $k = 10$ to more accurately measure the performance of each method [38]. To show the performance of each method, F-measure[6] and the *Receiver Operating Characteristic* (ROC)[7] area are applied.

Result Analysis. The performance of the four methods is shown in Table 2. Naive Bayes is shown to be unsuitable for our experiment due to the lack of independence or high dimensional of features [41]. The accuracy of SVM decreases when many irrelevant features are included in the dataset [42]. SVM might classify very well using selected features through a feature extraction method. Even though the tree-based Decision Tree showed high performance, it seemed to have an over-fitting problem [43]. In our experiment, Random Forest outperformed the others. The reason for this result is possible that the Random Forest enhances accuracy and reduces the over-fitting problem since multiple tree-structured classifiers, which were independently built by sub-datasets and randomly selected feature vectors, were applied [44].

[5] In the k-fold cross-validation, the dataset is randomly divided into k sets, each method is trained from $k - 1$ sets, and the rest of the 1 set is used for the test. This procedure is repeated k times, and the average of the evaluated values is given.

[6] F-measure [39] is a harmonic means of precision and recall, and this measure shows how well our model consistently classifies the data, such as the expression units and gesture types.

[7] ROC area is defined by Specificity and Precision as x and y axes respectively. This area is used to evaluate the discrimination capacity to distinguish between class labels. There is a performance verification criteria for the ROC area: ROC area = 0.5 means a useless model, while ROC area = 1.0 indicate an excellent model [40].

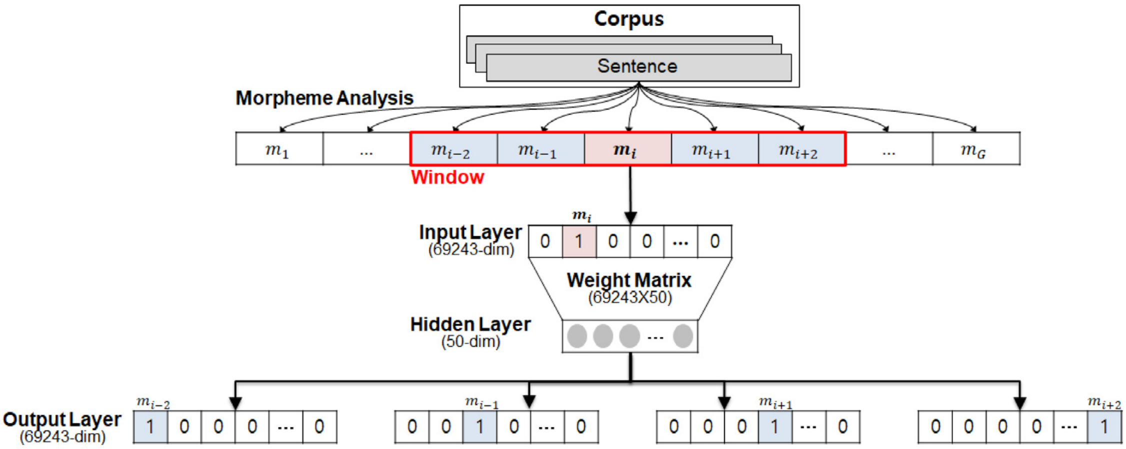


Fig. 6. For the training Word2Vec model, a sentence is divided into morphemes through morpheme analysis, and the ith morpheme is denoted as m_i. The G windows for the Word2Vec model (it is not related to windows for determining expression units and gesture types.) are created based on m_i for ($i = 1 \sim G$). The size of a window is set to five to include left and right two morphemes m_{i+j} for ($j = -2 \sim 2$) based on the morpheme m_i. The window is used to conduct feed-forward four times for this network such as $\{(m_i, m_{i\pm j})\}$, where m_i is input data, where $m_{i\pm j}$ are output data. However, if there is no morpheme left or right based on m_i, a set of training data is constructed using only the morphemes in the window. For example, an initial window based on m_i makes training data, such as (m_1, m_2) and (m_1, m_3), and the network conduct feed-forward two times. In our work, we used 874,573 sentences as the training dataset, and it took 8 min and 2.5 GB of memory. The training environment is as follows: CPU: i5-6600 K, GPU: Geforce GTX 960.

Table 2
Comparison of learning models.

	Expression units		Gesture types	
	F-measure	ROC area	Accuracy	ROC area
Naive Bayes	0.744	0.831	0.783	0.912
SVM	0.730	0.705	0.761	0.754
Decision Tree	0.818	0.874	0.820	0.887
Random Forest	**0.830**	**0.905**	**0.835**	**0.934**

5. Comparisons of effectiveness in HRI

5.1. Experimental design

A user study was conducted to verify the performance of the proposed method and compare the effects of human–robot interaction with other methods. The Godspeed questionnaire was applied to evaluate the effect of the human–robot interaction in 5-scale [45]. In this questionnaire, four items (i.e., anthropomorphism, likeability, perceived intelligence, animacy, perceived safe) were measured. For the comparison, two different methods were implemented that were applied by a principle of selecting and synchronizing a gesture from (1) the *Pattern* [11] and (2) *Word* [9] (these methods are mentioned in more detail in Section 1). In the compared methods, the execution time (i.e., expression unit) and gestures are determined according to matched a word pattern or a single word, respectively. For *Pattern* example, when there is a predefined pattern 'clear sky' in the gesture DB, a gesture is executed to express 'clear sky' while speaking the pattern. In *Word* example, when there is a predefined word 'sky' in the gesture DB, a gesture expressing the sky is executed while speaking that single word. A gesture was randomly generated when there was nothing matched with a pattern or word in a sentence. To compare the methodological differences rather than evaluating the design of the gestures, the gestures of the same design were used. Scenarios were designed into two types: (I) the weather forecast and (II) the donation. In order to evaluate the ability to deliver information according to the applied methods, we selected the weather forecast scenario. The donation scenario was taken to assess the robot's intentions to convey and persuade the person who interacts with the robot. We implemented each method using ROS (Robot Operating System) for real-time execution. The experiment was done in a home environment shown in Fig. 7.

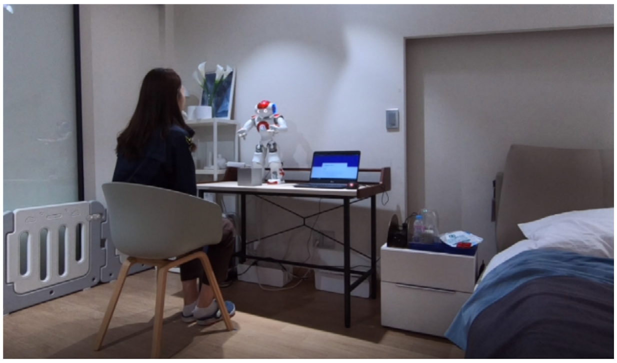


Fig. 7. Experimental setting: a robot provides an information related weather.

Scenario-I: weather forecast. A script consisted of 403 sentences from articles, where 28 newspaper articles in Korean were selected randomly. We searched them using keywords that are spring, summer, autumn, and winter. These sentences consisted of an average of 8.5 words and a total of 7247 morphemes. Each participant observed and evaluated gestures with utterances generated using 10 randomly selected sentences from the script. To make gestures for this experiment, we automatically extracted 10 most used words (i.e., increase, descend, continue, stop, little, warn, warm, most, again, and sky) based on counting word frequencies using the library *KoNLPy* [46] in the script. The gestures matched with the selected words were made in the gesture DB for *ours* and *Word*. In the case of Pattern, the most used pattern for the extracted words was additionally extracted from each word such as 'clear+sky'. Fig. 8 shows three gestures of them.

Scenario-II: donation. It was prepared by summarizing the text of a donation website into 13 sentences, where the scenarios consisted of the need for donations to plant trees in desertification areas [47]. In the same way as the weather forecast scenario, four gestures were made. The reason we made the four gestures was that the same words repeatedly appeared in each sentence. Participants are asked to watch a 20-min video clip as a donation of their time, where the clip is about a campaign for explaining why trees are needed in desertification areas. We informed that

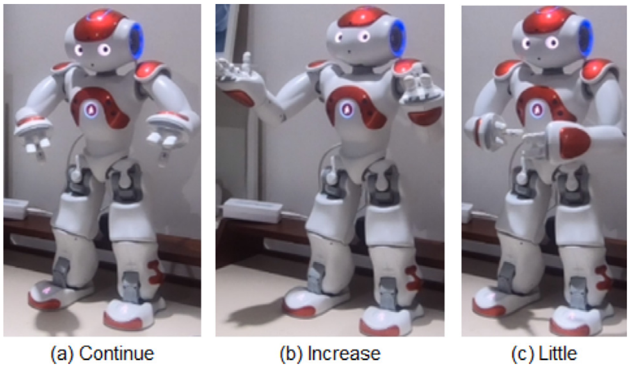


Fig. 8. Examples of gestures in the gesture DB: (a) metaphoric type of gestures that represent the continuous flow of an object (b) metaphoric type of gestures that express an increasing direction of an object (c) iconic type of gestures that indicates the small amount of an object.

advertising fees would be donated to NGO, if they watch the clip. Any participant who wanted to donate was asked to put the advertisement watching agreement in the donation box and was then requested to sit in a chair in front of a TV in a living room. If a participant sits down, this was considered a success case for donation.

Participants. 30 participants (age: $M = 26.63, SD = 2.57$ and gender: 19 female, 11 male) participated in the experiment, where M indicates a mean value and SD denotes a standard deviation. All participants were native speakers of Korean who had no experience with robots and were recruited at the Korea Institute of Science and Technology. They received $12.35 for the experiment lasting one hour. The participants were equally divided into three groups such as *Ours*, *Pattern*, or *Word* method. Each participant had time to interact with the robot without knowing state what was used.

Procedure. The experiment was carried out in the following order. (1) We informed participants that the goal of the experiment was to evaluate the suitability between the content of the speech and the gesture. The age, gender, and personality (i.e., extroversion, introversion) information of each participant was collected for analysis of results. Their personality was measured based on the Big Five personality model using a questionnaire consisting of 44 questions [48]. They were also asked to sit in front of a desk where the robot was located. (2) The Weather forecast scenario was held for 5 min. (3) A participant was asked to complete the questionnaire for 10 min. (4) Donation scenario was done for 5 min. (5) A participant decided whether to donate and filled out the questionnaire regarding the donation scenario for 10 min. (6) A post-interview was conducted with the participant for 10 min.

5.2. Results of user study

Results of scenario-I: weather forecast. Reliability analysis (Cronbach's α) was conducted to measure the overall consistency of results. In the scenario of weather forecast, anthropomorphism ($\alpha = 0.83$), likeability ($\alpha = 0.85$), perceived intelligence ($\alpha = 0.80$), and animacy ($\alpha = 0.77$) were within the acceptable range ($\alpha \geq 0.70$). However, perceived safe ($\alpha = -0.57$) was the unacceptable. This item might have irrelevance to the safety assessment as the participants observe the robot. Second, we analyzed the average score of each item with ANOVA (analysis of variance) and Cohen's d (effect size η) with a 95% confidence. Fig. 9 shows the average score for each item. In anthropomorphism ($p = 0.04$, $\eta^2 = 0.21$), the participants assessed that the robot with our method ($M = 2.62, SD = 0.71$) looked like more human than others (Pattern: $M = 2.1, SD = 0.62$ and Word: $M = 1.92$, $SD = 0.47$). In likeability ($p = 0.38$, $\eta^2 = 0.07$), our method ($M = 4.08, SD = 0.53$) also received a higher score than Pattern ($M = 3.88, SD = 0.61$) and Word ($M = 3.67, SD = 0.74$). On the other hand, other items of all three methods were similar, and no meaningful difference was found.

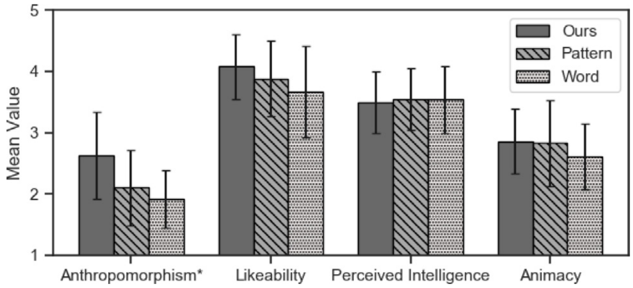


Fig. 9. Average scores in Scenario-I, weather forecast. A star($*$) denotes statistically significant difference (i.e., $p < 0.05$).

For a more detailed analysis, we analyzed sub-items that comprise each item. For each item, anthropomorphism comprises 5 sub-items 'natural, humanlike, conscious, lifelike, and moving elegantly'. Likeability has 5 sub-items 'like, friendly, kind, pleasant, and nice'. Perceived intelligence has 5 sub-items 'competent, knowledgeable, responsible, intelligent, and sensible'. Animacy includes 6 sub-items 'alive, lively, organic, lifelike, interactive, and responsive'. Among the 21 sub-items, 13 sub-items have the effect size above the medium level as seen in Fig. 10. The robot with our method achieved higher scores for 12 sub-items such as 'natural, humanlike, conscious, lifelike, moving elegant, like, friendly, pleasant, competent, responsible, and sensible', where all Sub-items were included to measure anthropomorphism.

Results of scenario-II: donation. In the scenario of donation, anthropomorphism ($\alpha = 0.89$), likeability ($\alpha = 0.91$), perceived intelligence ($\alpha = 0.86$), and animacy ($\alpha = 0.86$) were accepted by the Cronbach's α reliability test. However, the average of scores for all items was similar as seen in Fig. 11. Their difference was not statistically significant (ANOVA with a 95% confidence). The number of successful donations also did not differ significantly by each method, where a robot that was applied by Our, Pattern, and Word methods received 4, 4, and 2 donations respectively.

6. Discussion and limitation

For other items in the scenario of the weather forecast, significant results were not found statistically. The cause of the result is analyzed as follows. First, it may not be enough to assess the perceived intelligence in the scenarios of weath- er forecast and donation that unilaterally spoke to the participants. This reason can be supported from a result of the meta-analysis of the Godspeed questionnaire for the item of the received intelligence [49]. In the meta-analysis, researchers mentioned that perceived intelligence is more affected by the interaction scenario. Second, assessments of likeability and animacy could be affected by an imbalance in gender ratio for each experimental group (e.g., *Pattern* method was assessed from 9 female and 1 male). For 30 participants (19 female, 11 male), we found that female participants (likeability: $M = 4.09$, animacy: $M = 2.93$) gave higher scores than males (likeability: $M = 3.50$, animacy: $M = 2.48$) to likeability ($p = 0.02$, $\eta^2 = 1.01$) and animacy ($p = 0.02$, $\eta^2 = 0.81$).

In the donation scenario, all items received similar scores compared to other methods. The cause of this result was reasoned

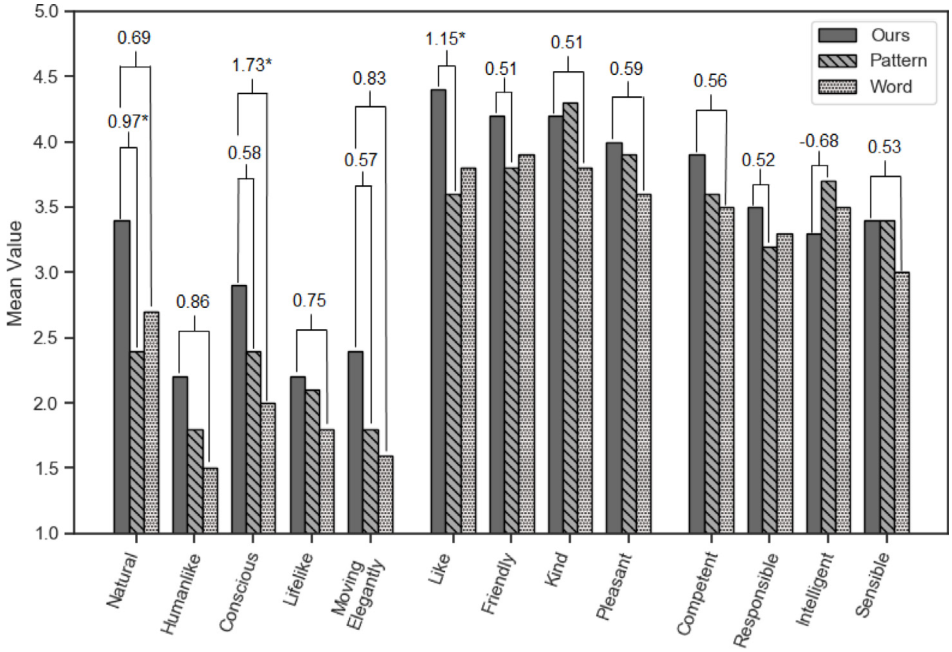


Fig. 10. Sub-items that have Cohen's d (effect size η) above the medium level. A star($*$) denotes statistically significant difference (i.e., $p < 0.05$).

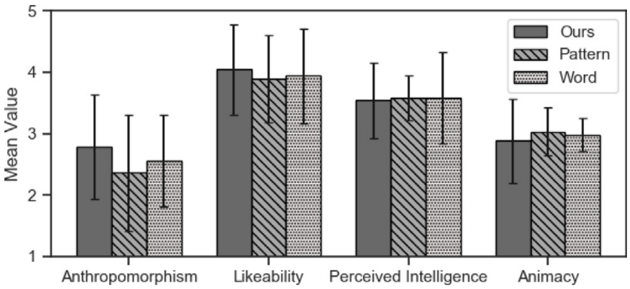


Fig. 11. Average scores in Scenario-II, donation.

by post-interview with participants. In the donation scenario, 19 participants mentioned that the content of the robot speech was not so persuasive. 7 participants indicated that the voice tone of a robot seemed to have no emotion. Although we asked for an evaluation based on robot gestures, the participants mentioned that they focused because they had to concentrate on what the robot said. Goal-oriented scenarios such as donations might not have been appropriate to evaluate robot gestures. In addition, we found that the personality of the participants (15 extroverts, 15 introverts) influenced the gesture evaluation. Extroverts gave higher scores of all items than introverts in the donation scenario. In the anthropomorphism ($p = 0.04$, $\eta^2 = 0.75$), there was statistically significant that extroverts ($M = 2.86$) gave higher scores to robots than introverts ($M = 2.26$).

We assume that assessments except for anthropomorphi-sm in the scenario of weather forecast became obscure due to the reasons that we mentioned above. In this regard, even though several researchers have studied analyzing the interaction effect according to the gender [50,51] and personality [52,53] of participants, there is rarely found that the perceived gesture of the participants can be changed depending on the service purpose. Therefore, our experiment can provide insight into which factors to consider in the study of robot gestures.

There are some limitations in this work. First, our method could generate gestures considering input sentences. It is not easy to deal with environments requiring different gestures for the same sentence since context information is not used. Second, inappropriate gestures could be generated for words (e.g., eat, peel) that have similar meanings but need to express different actions. To overcome this limitation, a user needs to define some corresponding relations between a morpheme and gesture. Third, our method determined expression units and gesture types with 83% accuracy. It still needs to improve the accuracy above 90%. Lastly, although our method can be applied to non-humanoid robots as we change the gesture DB, we could not conduct experiments for various types of robots. These experiments are in progress.

7. Conclusion

We proposed a method to generate co-speech gestures for a robot. We focused on automatically generating gestures even for a sentence containing words that the robot does not know. To generate arbitrary sentences, we used Random Forest to determine expression units and gesture types using patterns of PoS and constituents. We used Word2Vec to generate appropriate gestures from the gesture DB considering a semantic similarity with gesture types. To verify our method, we conducted an accuracy comparison for determining expression units and gesture types. In this experiment, Random Forest showed higher accuracy than other methods. In addition, we performed a user study to compare effectiveness in terms of human–robot interaction by using two scenarios such as weather forecast and donation. Our method could generate gestures that are more humanlike (Anthropomorphism) better than other methods. This result may indicate that our method could select appropriate gestures for a given sentence like a human since the same robot and gesture DB were used for our and other methods. Furthermore, this result is encouraged because users could feel a companionship to a robot that looks similar to humans [54].

In future works, our method should be able to generate appropriate gestures by considering robot sentences with context. If gestures can be created considering situations, the domain to which the proposed method can be applied will become larger. In addition, our method for English would be implemented and verified. We assume that our method can be worked for English since we use grammatical information instead of words. In our experiments, we could not allocate the participants by considering the gender, personality, and type of scenario in this experiment. Furthermore, experiments with the improved method would be conducted for various scenarios to verify sub-classes and described in the next publish.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Kendon, Gesture: Visible Action As Utterance, Cambridge University Press, 2004.

[2] P. Wagner, Z. Malisz, S. Kopp, Gesture and speech in interaction: An overview, Speech Commun. 57 (2014) 209–232.

[3] P. Bremner, U. Leonards, Efficiency of speech and iconic gesture integration for robotic and human communicators-a direct comparison, in: Proc. IEEE Int. Conf. Robotics and Automation, 2015, pp. 1999–2006.

[4] M. Lohse, R. Rothuis, J. Gallego-Pérez, D.E. Karreman, V. Evers, Robot gestures make difficult tasks easier: the impact of gestures on perceived workload and task performance, in: Proc. ACM Int. Conf. Human Factors in Computing Systems, 2014, pp. 1459–1466.

[5] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, F. Joublin, Generation and evaluation of communicative robot gesture, Soc. Robot. 4 (2) (2012) 201–217.

[6] Q.A. Le, C. Pelachaud, Generating co-speech gestures for the humanoid robot NAO through BML, in: Proc. Int. Gesture Workshop, Springer, 2011, pp. 228–237.

[7] C.-M. Huang, B. Mutlu, Robot behavior toolkit: generating effective social behaviors for robots, in: Proc. ACM/IEEE Int. Conf. Human-Robot Interaction, 2012, pp. 25–32.

[8] I. Mlakar, Z. Kačič, M. Rojc, TTS-driven synthetic behaviour-generation model for artificial bodies, Int. J. Adv. Robot. Syst. 10 (10) (2013) 344.

[9] H.-H. Kim, Y.-S. Ha, Z. Bien, K.-H. Park, Gesture encoding and reproduction for human-robot interaction in text-to-gesture systems, Ind. Robot 39 (6) (2012) 551–563.

[10] C.-M. Huang, B. Mutlu, Learning-based modeling of multimodal behaviors for humanlike robots, in: ACM/IEEE Int. Conf. Human-Robot Interaction, 2014, pp. 57–64.

[11] V. Ng-Thow-Hing, P. Luo, S. Okita, Synchronized gesture and speech production for humanoid robots, in: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, 2010, pp. 4617–4624.

[12] J. Kim, W.H. Kim, W.H. Lee, J.-H. Seo, M.J. Chung, D.-S. Kwon, Automated robot speech gesture generation system based on dialog sentence punctuation mark extraction, in: Proc. IEEE Int. Symp. System Integration, 2012, pp. 645–647.

[13] Y. Ferstl, M. Neff, R. McDonnell, ExpressGesture: Expressive gesture generation from speech through database matching, Comput. Animat. Virtual Worlds (2021) e2016.

[14] L. Pérez-Mayos, M. Farrús, J. Adell, Part-of-speech and prosody-based approaches for robot speech and gesture synchronization, J. Intell. Robot. Syst. (2019) 1–11.

[15] H. Ahn, T. Ha, Y. Choi, H. Yoo, S. Oh, Text2Action: Generative adversarial synthesis from language to action, in: Proc. IEEE Int. Conf. Robotics and Automation, 2018, pp. 1–5.

[16] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, G. Lee, Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 4303–4309.

[17] D. McNeill, Hand and Mind: What Gestures Reveal About Thought, University of Chicago Press, 1992.

[18] D. McNeill, Gesture and Thought, University of Chicago Press, 2008.

[19] D. McNeill, S.D. Duncan, J. Cole, S. Gallagher, B. Bertenthal, Growth points from the very beginning, Interact. Stud. 9 (1) (2008) 117–132.

[20] C.U. Press, Cambridge Dictionary: Meaning of Phrase in English, Cambridge University Press, 1999, URL: https://dictionary.cambridge.org/dictionary/english/phrase.

[21] L. Media, Your Dictionary: Phrase Examples, LoveToKnow Media, URL: https://examples.yourdictionary.com/phrase-examples.html.

[22] D. McNeill, Gesture: a psycholinguistic approach, Encycl. Lang. Linguist. (2006) 58–66.

[23] H.M.H.P. Company, Dictionary of the English Language, Houghton Mifflin Harcourt Publishing Company, 2016, URL: https://www.thefreedictionary.com/immediate+constituent.

[24] C.U. Press, Cambridge Dictionary: Meaning of Morpheme in English, Cambridge University Press, 1999, URL: https://dictionary.cambridge.org/dictionary/english/morpheme.

[25] C.U. Press, Cambridge Dictionary: Meaning of Part of Speech in English, Cambridge University Press, 1999, URL: https://dictionary.cambridge.org/dictionary/english/part-of-speech.

[26] D.-J. Lee, J.-H. Yeon, I.-B. Hwang, S.-G. Lee, KKMA: a tool for utilizing sejong corpus based on relational database, Korean Inst. Inf. Sci. Eng. Comput. Pract. Lett. 16 (11) (2010) 1046–1050.

[27] K. Toutanova, D. Klein, C.D. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, in: Proc. Int. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 173–180.

[28] M.P. Marcus, M.A. Marcinkiewicz, B. Santorini, Building a large annotated corpus of english: The Penn Treebank, Comput. Linguist. 19 (2) (1993) 313–330.

[29] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, 2009.

[30] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Vol. 1, in: Springer series in statistics, (10) 2001.

[31] W. Du, Z. Zhan, Building decision tree classifier on private data, in: Proc. IEEE Int. Conf. Privacy, Security and Data Mining, 2002, pp. 1–8.

[32] C. McCormick, Word2vec tutorial-the skip-gram model, 2016.

[33] M. of Education, Elementary Korean Textbooks, Ministry of Education, 2016.

[34] K.P. Murphy, Naive Bayes Classifiers, 18, University of British Columbia, 2006, p. 60.

[35] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proc. ACM Int. Workshop on Computational Learning Theory, 2003, pp. 144–152.

[36] Y. Freund, L. Mason, The alternating decision tree learning algorithm, in: Proc. Int. Conf. Machine Learning, 99, 1999, pp. 124–133.

[37] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[38] J.D. Rodriguez, A. Perez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2009) 569–575.

[39] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, Mach. Learn. Technol. (2011).

[40] M.J. Pencina, R.B. D'agostino, K.M. Pencina, A.C.J. Janssens, P. Greenland, Interpreting incremental value of markers added to risk prediction models, Am. J. Epidemiol. 176 (6) (2012) 473–481.

[41] I. Rish, et al., An empirical study of the naive Bayes classifier, in: Proc. Int. Workshop on Empirical Methods in Artificial Intelligence, 3, (22) 2001, pp. 41–46.

[42] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, in: Proc. Int. Conf. Neural Information Processing Systems, 2001, pp. 668–674.

[43] N. Horning, Introduction to decision trees and random forests, in: American Museum of Natural History's Center for Biodiversity and Conservation, 2013.

[44] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[45] C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots, Soc. Robot. 1 (1) (2009) 71–81.

[46] E.L. Park, S. Cho, KoNLPy: Korean natural language processing in Python, in: Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea, 2014.

[47] G. Asia, Happy bean for donation, 2019, URL: https://happybean.naver.com/donations/H000000154724.

[48] R.R. McCrae, P.T. Costa, Validation of the five-factor model of personality across instruments and observers, Personal. Soc. Psychol. 52 (1) (1987) 81.

[49] A. Weiss, C. Bartneck, Meta analysis of the usage of the godspeed questionnaire series, in: Proc. IEEE Int. Conf. Robot and Human Interactive Communication, 2015, pp. 381–388.

[50] C.-M. Huang, B. Mutlu, Modeling and evaluating narrative gestures for Humanlike robots, in: Proc. Int. Conf. Robotics: Science and Systems, 2013, pp. 57–64.

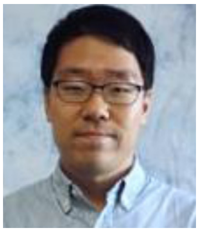
[51] B. Mutlu, Designing embodied cues for dialog with robots, AI Mag. 32 (4) (2011) 17–30.

[52] A. Aly, A. Tapus, A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction, in: Proc. ACM/IEEE Int. Conf. Human-Robot Interaction, 2013, pp. 325–332.

[53] A. Tapus, C. Ţăpuş, M.J. Matarić, User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy, Intell. Serv. Robot. 1 (2) (2008) 169.

[54] K. Dautenhahn, S. Woods, C. Kaouri, M.L. Walters, K.L. Koay, I. Werry, What is a robot companion-friend, assistant or butler? in: Proc. Int. Conf. Intelligent Robots and Systems, 2005, pp. 1192–1197.

Yu-Jung Chae is a Ph.D student at the University of Science and Technology (UST). She received the B.S degree in Electrical Engineering from Kangwon National University. Her current research interests are human–robot interaction and behavior expression (gesture, facial expression, speech) for a robot.

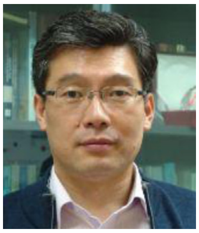


Changjoo Nam received the Ph.D. degree in computer science from Texas A&M University, College Station, TX, USA, and the M.S. and B.S. degree in electrical engineering from Korea University, Seoul, South Korea. He is currently a Senior Research Scientist with the Robotics and Media Institute, Korea Institute of Science and Technology (KIST), Seoul, South Korea. He was with the Robotics Institute, Carnegie Mellon University as a Postdoctoral Fellow. His research interest includes task planning for multirobot coordination, robotic manipulation, and human–robot interaction.

Daseul Yang received the B.S. degree in Computer Science and Industrial Engineering from Hongik University in 2016 and the M.S. degree in Human–Computer Interaction and Accessibility from Sungkyunkwan University in 2018. He is currently a researcher at Korea Institute of Science and Technology and is conducting a Human-Robot Interaction study. His current research interests include: HRI(Human–Robot Interaction), UX design for various devices.

HunSeob Sin received the B.S. degree in electric engineering from the Korea polytechnic university in 2015. Received the M.S. degree in electric engineering from Korea University in 2018. He is currently an assistant researcher at Korea institute of robot and convergence.

ChangHwan Kim received B.S. and M.S. degrees in mechanical engineering and machine design engineering from Hanyang University, Seoul, Korea, in 1993 and in 1995. He received the Ph.D. degree in mechanical engineering from the University of Iowa, IA, U.S.A., in 2002. He was a Research Associate at the Robotics and Automation Laboratory in the University of Notre Dame, IN, U.S.A. from 2002 to 2004. Since 2004, he had worked in robotics research groups at the Korea Institute of Science and Technology (KIST). His research interests include human motion imitation, motion generation of a humanoid, task and motion planning, cooperation of multiple robots, and rehabilitation robots.



Sung-Kee Park received the B.S. and M.S. degrees in mechanical design and production engineering from Seoul National University, Seoul, Korea, in 1987 and 1989, respectively. He received the Ph.D. degree from the Korea Advanced Institute of Science and Technology, Seoul, in 2000, in the area of computer vision.

He is a principal Research Scientist for Korea Institute of Science and Technology (KIST), Seoul. He has been working with Center for Robotics Research at KIST. During his period at KIST, he held a visiting position at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 2005, where he did research on object recognition. His current research interests include cognitive visual processing, object recognition, visual navigation, and human–robot interaction.