Synchronized Gesture and Speech Production for Humanoid Robots

Victor Ng-Thow-Hing
Honda Research Institute USA, Inc.
Mountain View, CA, USA
vng@honda-ri.com

Pengcheng Luo
University of California, Davis
Davis, CA, USA
pcluo@ucdavis.edu

Sandra Okita
Columbia University
New York, NY, USA
okita@tc.columbia.edu

Abstract— **We present a model that is capable of synchronizing expressive gestures with speech. The model, implemented on a Honda humanoid robot, can generate a full range of gesture types, such as emblems, iconic and metaphoric gestures, deictic pointing and beat gestures. Arbitrary input text is analyzed with a part-of-speech tagger and a text-to-speech engine for timing information of spoken words. In addition, style tags can be optionally added to specify the level of excitement or topic changes. The text, combined with any tags, is then processed by several grammars, one for each gesture type to produce several candidate gestures for each word of the text. The model then selects probabilistically amongst the gesture types based on the desired degree of expressivity. Once a gesture type is selected, it coincides with a particular gesture template, consisting of trajectory curves that define the gesture. Speech timing patterns and style parameters are used to modulate the shape of the curve before it sent to the whole body control system on the robot. Evaluation of the model's parameters were performed, demonstrating the ability of observers to differentiate varying levels of expressiveness, excitement and speech synchronization. Modification of gesture speed for trajectory tracking found that positive associations like happiness and excitement accompanied faster speeds, with negative associations like sadness or tiredness occurred at slower speeds.**

I. Introduction

Many people feel a natural affinity for humanoid robots because their appearance and features are similar to our own. Beyond appearance, the expectations for the level of behavior and functionality of these humanoid robots are raised for the same reason that if a robot looks like us, it should behave and communicate like us. With this in mind, we decided to develop a model for synchronized gesture and speech communication for humanoid robots.

The phenomenon of gesture as a communication modality has been investigated for many years, dating back to at least the nineteenth century[7]. More recently, there has been growing evidence that gesture and speech are simultaneously generated from a common thought source, the hypothesized *growth point*[13]. The combination of symbolic characteristics of human spoken language with the imagery of gesture complete the expression of human thought.

Gesture itself has been categorized into different types[13]:

- *Emblems* are self-contained gestures whose meaning can be understood without spoken words. They can be culturally-specific and tend to be more constrained in their expression. Mainly, there are specific ways one can act out the gesture before the meaning becomes confused or lost. For example, waving the hand to say goodbye or gesturing someone to come closer are emblems.
- *Iconics* refer to concrete things and actions when used in conjunction with words. Tracing out a trajectory of a path or specifying how big something is with your hands spaced apart are examples.
- *Metaphorics* provide imagery of the abstract. This is a very useful function for gesture in that it can help people visualize difficult concepts that are entirely imaginative. For example, a person may refer to the different sides of an argument by appearing to be holding invisible items in her left and right hands.
- *Deictics* utilize parts of the body to point out both concrete and abstract things during conversation. Typically, one uses an arm with the index finger extended at the target of interest (real or imaginary).
- *Beats* are rhythmic hand motions that move up and down in synchrony with the cadences of speech. They can be one handed or two-handed, and can vary in the hand shapes used. Although beats have little semantic content, the manner in which they are performed can convey emphasis, emotion and personality.

Our main contribution is a model for generating all of these types of gestures in a humanoid robot from arbitrary input text. Unlike other approaches, we do not require the text to be annotated with the semantic structure of the sentences to be spoken or explicit mark-up of gesture directives. Instead, automatic methods are used for analyzing text. However, in cases where stylistic preferences cannot be extracted from text alone (e.g., saying something in a calm or excited way), we provide various tags that can be added to the text. The model has a gesture selection component that analyzes the text to determine appropriate gestures with the goal that even in the presence of unrecognized word meanings, appropriate default gesticulation can be produced. The gesture modification component adjusts trajectories to synchronize with the timing of speech as well as adjustments to infuse emotion and style variation. Probabilistic elements incorporated in gesture selection and modification ensure the gesture sequences produced will not appear unnaturally deterministic for multiple instances of the same text input.

Our approach allows gesture to be added as a fast post-process to spoken text and does not require careful linguistic analysis and annotation of text that would involve a significant level of expertise and manual effort. For example, gestures are automatically added to pre-existing text of speeches to create multi-modal presentations on a Honda humanoid robot (Figure 6). The robot's lack of visible facial expression underscores the need for more expressive bodily communication. Although the model focuses on parameters for designing the trajectories of the arms and hands, the physical expression of gesture in our model also influences torso and head orientation.

Section II reviews previous research in robotics and virtual embodied conversational agents to produce expressive body motion. We provide an overview of the model in Section III. Section IV describes the design and development of our gesture model and how we address the problems of gesture selection and gesture modification. Section V describes our implementation and examples of gesture in our model. We evaluate some parameters of our model in Section VI and conclude with discussion and future work in Section VII.

II. Related Work

Although there has been research in the area of gesture recognition and analysis for humanoid robots[1], limited gesture phenomena have been modeled for expression on humanoid robots. Deictic gestures were used in [22] to establish common object referents between humans and robots. Emblems for displaying emotional states were implemented on the WE-4RII robot[6]. These systems focused on one particular type of gesture and do not attempt to model a general framework for handling all types of gesture. The study in [14] found relationships between the physical properties of robot gesture and human emotional perception of those gestures. While not specifically gesture, arm movements synchronized to music for dance were implemented on an HRP-2 robot platform[19]. Our model is also concerned with synchronizing arm movements to an external data stream, mainly input text.

In the area of embodied conversational agents, there has been active work in developing complex gesture models for animating virtual characters. The main goals are producing meaningful and synchronized arm motions to match either synthetic or recorded speech. This work can be compared on the basis of several design choices in their respective algorithms: input to the gesture planner, gesture selection and gesture modification.

Gesture systems for animation often utilize motion-captured data and recorded dialog of a human performance. In [20], both motion capture data and speech segments are recombined under constraints obtained analyzing gesture structure to create expressive hand motions for dialog. In [11] and [10], motion-captured gesture clips are matched up to prosody features from live speech using probabilistic models to generate real-time gestures. Although the use of prosody can effectively express emphasis and emotional cues, semantic meanings of words cannot be conveyed. In cases where the input is text, some systems assume that the text has been annotated to indicate what type of gesture and their parameters to use [4], [9]. In [15], higher level semantic tags for the theme, rheme and focus of a sentence are manually provided and a probabilistic model derived from training data is used to gesture in a given speaker's style. Although manual annotation offers direct control of how gestures can be coordinated with speech, it requires significant effort to linguistically analyze the text and assign gesture parameters. It is also susceptible to deterministic gesture behavior if only one type of gesture is defined for a particular text sequence. In our model, a single expressivity parameter can be used to control selection from a range of possible gesture interpretations for a text sequence.

Several gesture models focus on the role of gesture in conversation in conjunction with automated dialog managers [24] that provide communicative goals that are passed onto a gesture planner for eventual expression with speech. With additional information provided by the communication goal, text can be annotated precisely with specific iconic representations parameterized by form features like hand locations and palm orientation [9], [18]. However, some robot applications use manual user input (Wizard-of-Oz methods) or prescripted dialog to generate speech. Therefore, we attempt to reconstruct the communicative intent through text and parts-of-speech analysis to select appropriate gestures.

The BEAT gesture system[2] is closest in philosophy to our own model in that their system generates synchronized gesture with synthesized speech. They feature an extensible rule set for suggesting what types of gestures to perform and establish beat gestures as a default gesture when no other gesture types are suggested. Our model extends this idea by allowing simultaneous analysis of the text using multiple grammars designed for each type of gesture (emblems, iconics, etc.) and probabilistically selects amongst the various candidate gestures based on different factors.

Once the gesture types have been selected, the trajectories can be modified for different reasons. The trajectories for the resulting arm motion can be modified for the purpose of synchronization to speech[26]. Additionally, trajectories can be parameterized along different expressive axes[3], [4], [17].

There are several unique challenges to implementing a gesture system on a physical robot rather than a virtual character. Current humanoid robots tend to have less degrees of freedom than a virtual character, potentially decreasing the expressiveness of the gestures produced. We also desire the gesture planning time to be as fast as possible in interactive applications where a robot must respond quickly to interactive queries from its human partner. Finally, the motions produced by the robot must be dynamically safe as well as collision-free. From an application standpoint, even if the robot has no advance knowledge of the text to be spoken, reasonable gesture behavior should be shown.

III. Model Overview

Our gesture system takes as input arbitrary text sentences and outputs synchronized synthetic speech and gestures, included coordinated head and torso movements. Figure 1 describes the process starting from the top left where the

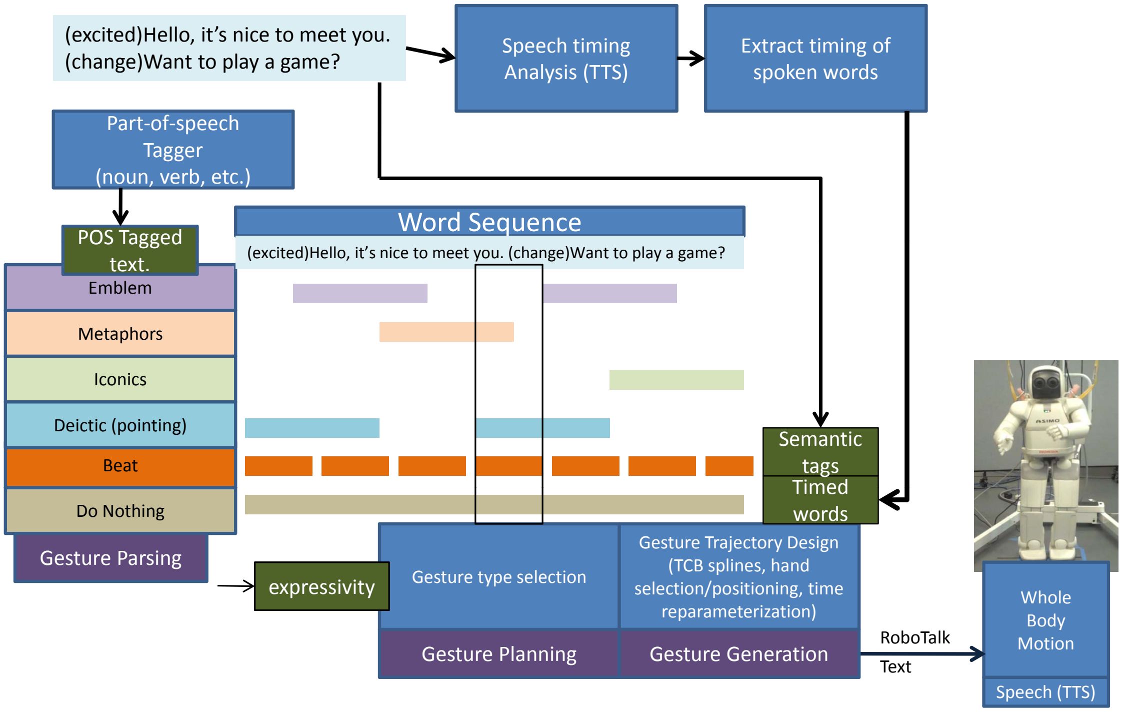

(excited)Hello, it's nice to meet you.
(change)Want to play a game?
Speech timing Analysis (TTS)
Extract timing of spoken words
Part-of-speech Tagger (noun, verb, etc.)
POS Tagged text.
Word Sequence
(excited)Hello, it's nice to meet you. (change)Want to play a game?
Emblem
Metaphors
Iconics
Deictic (pointing)
Beat
Do Nothing
Gesture Parsing
Semantic tags
Timed words
expressivity
Gesture type selection
Gesture Trajectory Design (TCB splines, hand selection/positioning, time reparameterization)
Gesture Planning
Gesture Generation
RoboTalk
Text
Whole Body Motion
Speech (TTS)

Fig. 1. Overview of our gesture model.

sample text is initially processed separately in parallel by a part-of-speech tagger and a text-to-speech (TTS) engine used to determine word timings. The model is initially configured by providing grammars for the different gesture types and deciding on a parameter value for expressivity, which defines the personality of the robot presenter. Once the system is configured, arbitrary text can be provided to the gesture model to generate co-expressed speech and gesture.

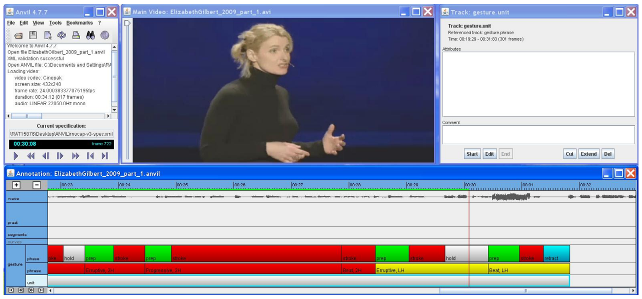
Once the text is tagged by part-of-speech, it is simultaneously processed by five different grammars, each designed to identify appropriate candidates for each gesture type: emblem, metaphoric, iconic, deictic and beat gestures. There is also an option to do nothing to prevent excessive gesture. In our system, all gesture strokes (the portion of gesture associated with meaning) begin on word boundaries. As a word usually is the minimal encapsulation of a thought, we believe this is a valid assumption to make as the growth point hypothesis suggests that gesture types will not change in the middle of a word. For each word in the input sentence, there can be up to six different possibilities of gesture that can be expressed. Section IV-B describes how we select which gesture to use. Once both processes for selecting gesture types and timing information have been completed, this information is combined in a process that selects the basic trajectory shapes by gesture type and then modulates the shape of the trajectory using timing information to start gesture strokes on certain words. In addition, optional tags can be added to the input text to provide contextual hints such as *change* for indicating a change of topic and *excited* or *calm* for controlling the degree of excitement in the speaker. These hints are usually difficult if not impossible to pick up just from text alone. The final gesture plan, consisting of the continuous stitched sequence of gesture trajectories for both arms is then sent to a gesture generation module to be expressed on the robot. Speech is generated at the same time in parallel with the gestures.

IV. Model Design

In designing the requirements for our model, we envisioned the gesture system to be a reusable component not tied to a particular application. We wanted existing applications in our robot like the memory game described in [16] to benefit immediately from the greater expressivity of gestures without requiring substantial rewriting or modification of all the dialog text. If an application designer had to annotate gestures for all text, we felt that the manual labor involved would be a disincentive to use gesture. Instead, the gesture system attempts to analyze and process the text and can add gesture independently of the application. However, we do not preclude other modules from annotating the speech text to provide more guidance over gesture selection and modification and make use of this extra contextual or stylistic information when available.

A. Gesture Analysis

Using the approach and gesture lexicon (*lexemes*) of [15], we studied several videos of dynamic presenters (Elizabeth Gilbert, J.J. Abrams, Isaac Mizrahi and Tony Robbins) from the TED conference series [23] and analyzed them with the ANVIL video annotation tool. A sample annotation is shown in Figure 2. The lexemes and gesture phases, denoting the



```
Welcome to Anvil 4.7.7
Open file ElizabethGilbert_2009_part_1.anvil
XML validation successful
Open ANVIL file: C:\Documents and Settings\RA
Loading video:
    video codec: Cinepak
    screen size: 432x240
    frame rate: 24.000383377075195fps
    duration: 00:34:12 (817 frames)
    audio: LINEAR 22050.0Hz mono
```

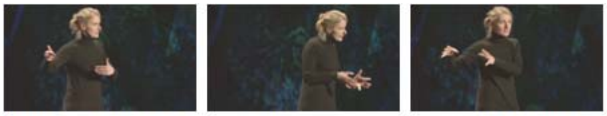
Current specification:
\RAT15876\Desktop\ANVIL\mocap-v3-spec.xml
00:30:08 frame 722

Track: gesture.unit
Referenced track: gesture.phrase
Time: 00:19:29 - 00:31:83 (301 frames)

Fig. 2. Anvil Annotation

start, main stroke and retraction of gestures over time were annotated manually for these videos. It was important to study more than one speaker to look for general patterns of gesture formation across speakers and not build in rules that may be specific to only one person. Although the individual mannerisms and frequency of different gesture were very different for each speaker, we were able to identify several trends.

- There are limited sets of gestures to be represented which reduce the complexity of modeling. Defining the gesture lexicon and defining appropriate parameterizations to their trajectories produce a more compact and less data-intensive model than resorting to many examples of motion-captured gesture data.
- Specific gestures tend to be associated with certain words, phrases or part-of-speech. For example, the *erupt* gesture where the arms gesture outward typically occurs on verbs.
- *Catchments*[13] were observed where the style of gesture beats tends to stay the same within a contextual topic, and can change as the topic changes. This motivated defining a *change* tag for the input text. An example of catchments is shown in Figure 3 where the speaker's pose and hand beats change when different phrases are spoken.
- The same words may map to several different gestures, influenced by many contextual factors as well as adjacent word use.



.. have a safe distance ... .. Between me ... .. and my natural anxiety ...

Fig. 3. Sample gesture sequence indicating context change.

These insights influenced our gesture selection model during gesture planning.

B. Gesture Selection

To minimize the number of parameters that need to be specified for gesture selection, the different gesture types were placed in a hierarchy of increasing expressivity (no gesture, beats, iconic gestures, metaphoric gestures, deictics and emblems) whose members make up the set G. Intuitively, gestures with increasing expressivity convey more imagery of the content of the accompanying speech. For example, beats serve mainly to emphasize words, whereas iconic gestures provide specific imagery to help describe items being spoken about. This concept is similar to level-based language analysis [12], where our hierarchy of gesture types correspond to the language elements in different layers. It was also important to offer the possibility of not gesturing at all as gesturing without pause can appear overly active and tends to muddle the overall communication.

The distribution of gesture occurrence for each gesture type i, $i \in G$, is represented by a normal distribution centered over different values of an expressivity parameter x (Figure 4). Given the mean μ_i and variance σ_i^2 for a given gesture type i, the weight function $w_i(x)$ for selecting a gesture type i is modeled as a Gaussian over expressivity values:

$$w_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(x-\mu_i)^2/(2\sigma_i^2)} \qquad (1)$$

where $x \in [0,1]$ represents expressivity. We set emblem gestures with a high mean and relatively low variance while iconic, metaphoric and deictic gestures have their means centered at intermediate values of x. The distribution for no gestures and beats are set fairly wide over the entire range of expressivity.

Different types of gestures can be generated for the same sequence of words. We use the expressivity parameter to model the relative probability that certain gesture types will be selected (if at all). The input text (that may contain style or part-of-speech tags) is parsed independently through multiple grammars designed for each gesture type to produce several candidate gestures for each word of the text. With this scheme, it is possible that a gesture type can be assigned to an individual word or a sequence of words. Consequently, each word in the input text can be labeled with one or more gesture type candidates. We defined the grammar rules based on video analysis of speakers as described in Section IV-A. Selecting an expressivity value for x determines the relative weighting $w_i(x)$ of each gesture type i. All candidate gesture types for a given word are collected in the set $C \subseteq G$, the relative weights for each gesture type are re-weighted by the total sum of weights of the set of available candidate gestures, C, to produce a probability of selecting that gesture type i:

$$p_i(x) = \frac{w_i(x)}{\displaystyle\sum_{j \in C} w_j(x)}. \qquad (2)$$

The computed probabilities for $p_i(x), i \in C$ are used to select the final gesture expressed. Once a gesture type is selected, it is assigned to all the words it spans.

C. Grammar Parsing

The input text is first automatically tagged by the Stanford Log-linear Part-Of-Speech Tagger[25] to assign part of speech (e.g., noun, verb, adjective) to each word. The

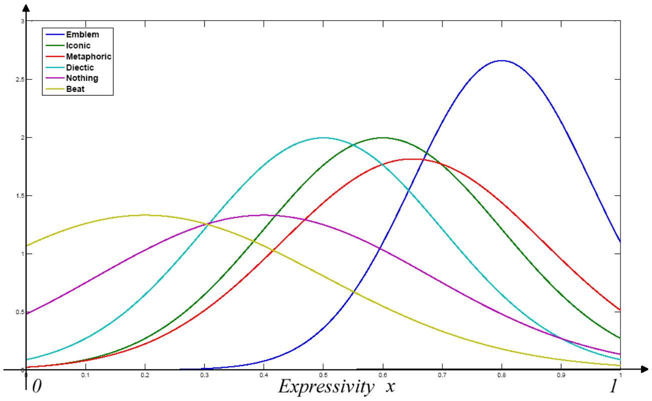


Fig. 4. Normal distribution for determining weights of gesture types over the range of expressivity.

tagged text is processed by several context-free grammars defined for each gesture type (emblem, metaphoric gestures, iconic gestures, etc.). Within a gesture type's grammar, rules are defined to choose which subtype of gesture to use. For example, within an emblem type, finding the text pattern "goodbye" will cause a hand-waving gesture to be selected. In addition to words, the grammar may act on part of speech, such as specifying eruptive-type gestures (outward arm motions) for verbs or cyclic arm motions for verbs in progressive tense (ending in *ing*). Grammars can also extract higher level patterns. For example, the grammar for finding iconic gestures may look for phrases that match the pattern: *between ... and ...*, to direct body orientation changes on certain words. The grammars we use have the following sample form:

$$\begin{array}{rcl} Sentence & \rightarrow & KeyWord_1 \\ Sentence & \rightarrow & Sentence\ KeyWord_2 \\ & & Sentence\ KeyWord_3 \\ Sentence & \rightarrow & KeyWord_4 \\ KeyWord_1 & \rightarrow & Gesture_1 \\ KeyWord_2 & \rightarrow & Gesture_2 \\ KeyWord_3 & \rightarrow & Gesture_3 \\ KeyWord_4 & \rightarrow & Gesture_4, \end{array}$$

where $KeyWord_i$ represents different sets of candidate words and/or tags and $Gesture_i$ defines the corresponding gesture lexemes to use. For example, in our iconic gesture, $Keyword_{size}$ contains the words "large" and "big" and maps to $Gesture_{distance}$ where both hands are held apart.

D. Gesture Modification

The selection of gesture lexemes corresponds to the selection of a gesture template consisting of the basic trajectory shape for the hand positions of the stroke portion of the gesture over time as well as time trajectories for wrist rotation and hand shape. Hand shape is controlled by one parameter ranging from 0 (open hand) to 1 (closed fist). The trajectories are stored as a set of key points for each parameter value. Kochanek-Bartels (TCB) cubic splines[8] are used to define trajectory curves by interpolating over the key points. TCB (tension-continuity-bias) splines have useful shape parameters that can control how smoothly or tightly the trajectories follow the key points, offering expressive variability (see Figure 5). Since gesture strokes are tied together with continuous spline trajectories, gestures are blended together smoothly and preparatory motion leading to a gesture stroke can occur prior to a word utterance.

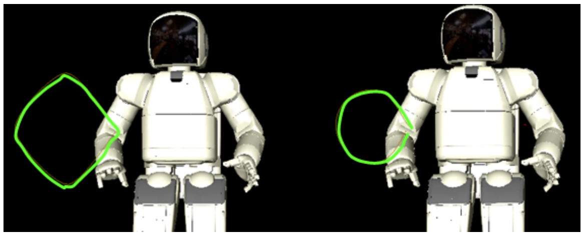


Fig. 5. The same hand trajectory with different values of tension in the TCB-spline curves.

1) Style Parametrization: The trajectory curves are functions of style parameters such as *excitement* and these parameter values can change with context or catchment changes. We define parameters for controlling the styles of gestures,

$$S = \{A, F, T_i, C_i, B_i, t_i | i = 0 \cdots n\}, \tag{3}$$

where n is the number of key frames for the current gesture lexeme, A is Amplitude, F is frequency for this gesture lexeme since some gestures have a repetitive sequence, T_i is tension,C_i is continuity,B_i is bias, and t_i is time for keyframe i which is normalized to be from 0 to 1.

A scalar value α is randomly generated within a numeric range based on the style tag value ($[0.6, 1]$ for $excited$, $[0.3, 0.7]$ for $neutral$ and $[0.0, 0.4]$ for $calm$). This parameter α is used to perturb the starting positions of gesture strokes by adding offsets to the defined template position or modifying the shape of the trajectory by defining the parameters A, F, B_i and T_i in Equation 3 as linear transformations of α. More sophisticated functions of α and style parameters can be defined, perhaps derived from empirical data or based on new style tags.

The number of hands involved in the gesture is also considered. In our model, there are two mechanisms where hands can change. The first is a change of topic and the second is the prescribing of specific hands for specific gestures. Head motions are defined by directing the robot to look at the centroid of the active hand positions involved in a gesture. If a gesture is one-handed, the robot only looks at the active hand. This has the natural effect of the robot's head following its own hand motions while expressing itself. Alternatively, the control of head motion can be given over to another process such as an attention-guided vision system.

V. Results

We implemented our gesture model on a Honda humanoid robot[5]. For controlling the robot, we used real-time collision avoidance whole body motion control described in

[21]. The controller performs task-based control to try its best to match the end-effectors of the arms to the targeted trajectory curves, while keeping constraints such as balance and velocity limits. The degrees of freedom for the arm and torso are both recruited to match the trajectory constraints, allowing changes in body pose. This allowed our model to focus on the shapes of hand and arm trajectories without too much concern for self-collisions as the collision avoidance system would automatically adjust trajectories smoothly to avoid collisions or in the worst case stop motion just before collision. The controller has adjustable parameters for maximum velocity of end effectors and time constants for how tightly the trajectory is followed over time. A large time constant can create very smooth arm motions, but fine details of the trajectory can be lost. Alternatively, a small time constant can capture many trajectory details but may make the arms appear jerky if the trajectories change shape rapidly.

Figure 6 features still frames from three different gesture sequences. The top (A) sequence demonstrates context change as described in Figure 3. The middle (B) sequence demonstrates an emblem gesture of waving goodbye for the phrase ”bye-bye”. The last sequence shows an metaphoric gesture of tracing a circle as the robot speaks the phrase ”circle around”. We have integrated the gesture module into our robot architecture to add gesture behavior to interactive applications like the memory game we use as one of our research platforms (Figure 7). Videos of our gesture sequences can be viewed via `http://www.honda-ri.com/HRI_Us/Projects`.

VI. Evaluation

We performed four studies designed to evaluate various parameters of our model. The first three studies were conducted from a sample of 29 adults. All gesture sequences were generated from text excerpts of two speeches from the late Soichiro Honda: ”What Mistakes Teach Us (1965)” (speech A) and ”First, You Work for Yourself (1969)” (speech B).

In Study 1, we tested the ability of the model to generate synchronized gesture and speech. Subjects were shown in series two identical videos of a portion of the gesture sequence generated with our model from speech B. However, one video had the correct audio track from speech B, while the other had audio from speech A. Respondents were asked to identify which video was better synchronized with the audio as well as which one seemed the most natural. The majority of respondents (83%) correctly identified the properly synchronized video with 76% describing it as the most natural.

In Study 2, we sought to evaluate how effective the style tags *excited* and *calm* were at modifying gesture sequences from the same identical input text. Two adjacent videos of gesturing robots (labeled Robot A and Robot B) generated from speech A were used with Robot A set with the *excited* tag and Robot B with the *calm* tag. Subjects were asked to identify which robot appeared more excited, calm and confident. From Figure 8, a large majority of subjects were

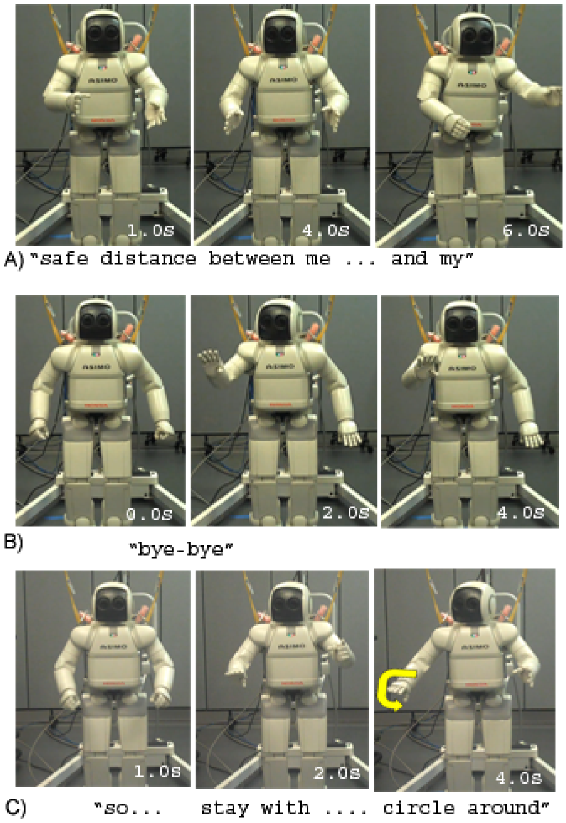


Fig. 6. Gestures on humanoid robot

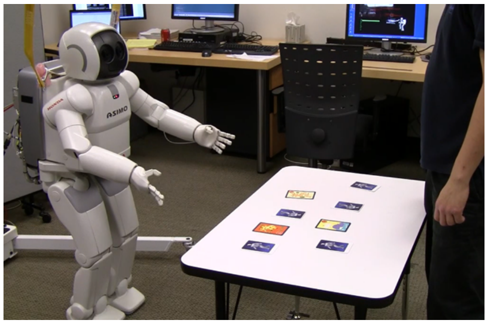


Fig. 7. Gesture integrated into the memory game.

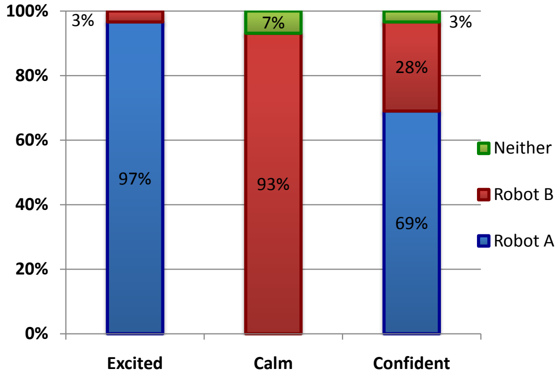


Fig. 8. Study 2: Excited versus calm style parameters

able to correctly identify the excited (97%) and calm (93%) settings. There was less agreement over which robot seemed the most confident, with a smaller majority (69%) associating confidence with the excited robot. When asked, subjects seemed to employ contrasting rationale, with both fast and slow motions being associated with confidence by different people. This suggests the importance of choosing style labels for motion parameters not subject to different interpersonal interpretation.

For study 3, we generated two groups of three gesture sequences from speech B with different expressivity settings: low ($x = 0.05$), medium ($x = 0.5$) and high ($x = 0.95$). Group 1 videos had audio and Group 2 videos had no audio track. The three unidentified videos for each group were placed from left to right with a medium-low-high order for Group 1 and high-low-medium for Group 2. Subjects were then asked questions rating which video was the best presenter, most graphic, least passionate and least expressive. Reviewing Figure 9, a majority of subjects were able to associate the qualities of least passionate and least expressive with the low-expressivity gesture setting. Medium and high expressivity settings were associated with "best presenter" and "most graphic" qualities. Subjects appeared to have more difficulty differentiating medium and high expressivity settings. This is probably due to the medium and high expressivity settings producing gesture type distributions which were too similar to perceive differences. On the other hand, the low expressivity settings feature a higher chance of the robot doing nothing or using beat gestures which may be more noticeable to a viewer. Subjects had a more difficult time distinguishing different levels of expressivity in the absence of sound. This may be partially explained by the fact that the interpretation of expressive gesture types like iconics and metaphorics are only meaningful with the accompanying spoken words.

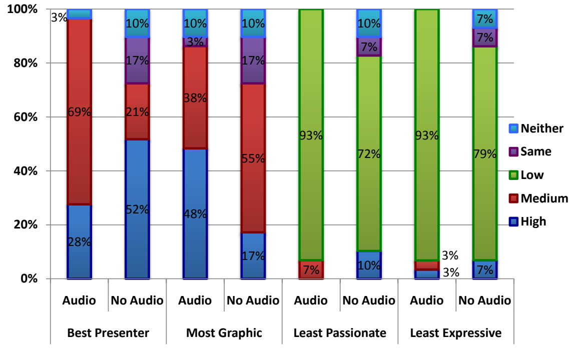
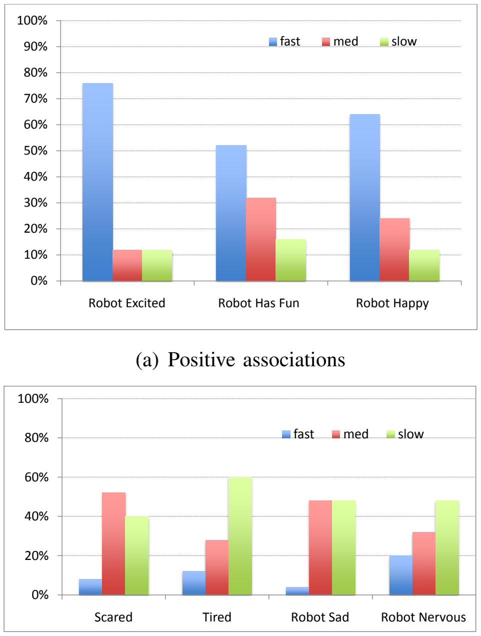


Fig. 9. Expressivity settings with and without sound

In the final Study 4, we wished to determine the best speed settings to use for arm motions. Twenty-five participants between the ages of twenty-three to forty years-old watched three video clips of the humanoid robot gesturing at different speeds and were asked for their qualitative impressions. The video consisted of the robot giving a short story about geckos. The speed settings were adjusted by changing the time constant of a second order system that controls how quickly the robot's end effectors converge to the target set point of the commanded trajectory. The participants saw one robot at slow speed, another robot at medium speed, and the third robot at fast speed. The three video clips were shown side-by-side, one at a time in two, thirty-five second increments. The videos were shown in the random order of medium, slow, fast, slow, medium, fast. Questions included measures for positive impressions (e.g. Which robot seemed to have the most fun?) and negative impressions (e.g. Which robot seemed most nervous?). Each video clip was labeled separately as A, B, or C with the speed label hidden.

The broader effect seemed to be that speed is associated with positive and negative feelings. As speed level increased, the robot's behavior was associated more with positive impressions. As speed level decreased, the robot's behavior was associated with negative impressions. For example, the participants felt that the robot was excited (over 75%) and happy (over 65%) when the speed level was fast (See Figure 10(a)). As the robots speed level decreased, more negative impressions are associated with the robot's behavior. For example, participants felt that the robot was tired when its speed was slow (60%) compared to medium (30%) and fast (12%) speeds. See Figure 10(b).



(a) Positive associations

(b) Negative associations

Fig. 10. Gesture impressions study

VII. Discussion

We have shown that our gesture model can produce synchronized gesture motions with arbitrary text input that can demonstrate many different gesture types: emblems, deictics, metaphoric and iconic gestures and beats. Evaluation studies demonstrate the effectiveness of gesture and speech synchronization and the ability of style tags and the expressivity parameter to alter an observer's perception of gesture style. Speed tests show an association of faster gestures with positive impressions and slower gestures with negative impressions. These findings are consistent with the findings in [14] who also observed similar associations with end-effector speeds.

A. Limitations and Future Work

The prototypical trajectories we designed for our model were all hand-crafted and done with relatively little key points (averaging about three). More realistic trajectories could be created to enhance the realism of gestures and the expressiveness of the robot. The robot we used only has 5 degrees of freedom in each arm, restricting the range of motions we can perform. A robot with higher degrees of freedom can use our gesture model with potentially greater gesture expressiveness, especially with a more flexible wrist and dextrous hands.

Our current gesture system expresses gesture motions and speech simultaneously once the gesture plan is finalized, but in an open-loop fashion. As others have done[9], we intend to re-design the system to allow closed-loop feedback to account for small system delays and re-adjust the timings of gesture with speech. This will allow speech to be paused to give more time for complicated and expressive gestures to complete. In the current system, speech dictates the timing of gestures completely, but gestures cannot affect speech patterns.

The grammars used in our model contain relatively few rules for each type of gesture. We can easily add more rules to increase the number of successful mappings between input text and appropriate gestures. New grammars can also be added to the model to handle other non-text cues, such as visual cues from a person. This would allow more appropriate gestures to be generated in conversational settings and provide greater awareness for turn-taking cues between the robot and human partner. We would also like to improve the number of stylistic parameters that can act on the gesture model simultaneously. We believe that use of gesture during communication enhances a person's overall experience when working with a humanoid robot due to the enhanced imagery gesture provides in addition to speech.

Acknowledgments

We would like to thank Deepak Ramachandran for his helpful discussions on probabilistic models for gesture selection.

References

[1] Andrew G. Brooks. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *In Proceedings of Human-Robot Interaction*, pages 297–304, 2006.

[2] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *SIGGRAPH 2001: Proceedings of ACM SIGGRAPH*, pages 477–486, New York, NY, USA, 2001. ACM.

[3] D. Chi, M. Costa, L. Zhao, and N. Badler. The emote model for effort and shape. In *SIGGRAPH 2000: Proceedings of ACM SIGGRAPH*, pages 173–182, 2000.

[4] B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *In Gesture in Human-Computer Interaction and Simulation*, volume 3881, pages 188–199. Springer, 2006.

[5] Honda Motor Co., Ltd. Asimo year 2000 model. http://world.honda.com/ASIMO/technology/spec.html, 2000.

[6] K. Itoh, H. Matsumoto, M. Zecca, H. Takanobu, S. Roccella, M.C. Carrozza, P. Dario, and A. Takanishi. Various emotional expressions with emotion expression humanoid robot we-4rii. In *IEEE Conference on Robotics and Automation 2004 TExCRA Technical Exhibition Based*, pages 35–36, 2004.

[7] A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[8] D.H.U. Kochanek and R.H. Bartels. Interpolating splines with local tension, continuity, and bias control. *ACM SIGGRAPH Computer Graphics*, 18(3):41, 1984.

[9] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Comp. Anim. Virtual Worlds*, 15(1):39–52, 2004.

[10] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Trans. Graph.*, 29(4), 2010.

[11] S. Levine, C. Theobalt, and V. Koltun. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):1–10, 2009.

[12] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 2002.

[13] D. McNeill. *Gesture and Thought*. University of Chicago Press, 2005.

[14] Hisayuki Narahara and Takashi Maeno. Factors of gestures of robots for smooth communication with humans. In *RoboComm '07: Proceedings of the 1st international conference on Robot communication and coordination*, pages 1–4, 2007.

[15] M. Neff, M. Kipp, I. Albrecht, and H. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, 27(1):5:1–24, 2008.

[16] V. Ng-Thow-Hing, J. Lim, J. Wormer, R. Sarvadevabhatla, C. Rocha, K. Fujimura, and Y. Sakagami. The memory game: Creating a human-robot interactive scenario for asimo. In *International Conference on Intelligent Robots and Systems (IROS 2008)*, pages 779–786, 2008.

[17] C. Rose, B. Bodenheimer, and M. Cohen. Verbs and adverbs: Multidimensional motion interpolation using radial basis functions. *IEEE Computer Graphics and Applications*, 18(5):32–40, 1998.

[18] M. Salem, S. Kopp, I. Wachsmuth, and F. Joublin. Towards meaningful robot gesture. In *Human Centered Robot Systems*, volume 6, pages 173–182, 2009.

[19] T. Shiratori and K. Ikeuchi. Synthesis of dance performance based on analyses of human motion and music. *IPSJ Online Transactions*, 1:80–93, 2008.

[20] Matthew Stone, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler. Speaking with hands: creating animated conversational characters from recordings of human performance. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*, pages 506–513, New York, NY, USA, 2004. ACM.

[21] H. Sugiura, M. Gienger, H. Janssen, and C. Goerick. Real-time collision avoidance with whole body motion control for humanoid robots. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2053–2058, 2007.

[22] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. Natural deictic communication with humanoid robots. In *International Conference on Intelligent Robots and Systems (IROS 2007)*, pages 1441–1448, 2007.

[23] TED. Ted (technology, entertainment, design) conference series. www.ted.com. http://www.ted.com/.

[24] P. Tepper, S. Kopp, and J. Cassell. Content in context: Generating language and iconic gesture without a gestionary. In *Proceedings of the Workshop on Balanced Perception and Action in ECAs at AAMAS '04*, 2004.

[25] K. Toutanova and C. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70, 2000.

[26] I. Wachsmuth and S. Kopp. Lifelike gesture synthesis and timing for conversational agents. In *Gesture and Sign Language in Human-Computer Interaction*, pages 225–235. Springer, 2002.