Autonomous Systems Project

Spring 2020

NAOBEARD

Group 27

Songha Ban

Gökçe Kuşcu

Presentation: https://youtu.be/N0B8rYc_TaY

Word count: 2150

1. Introduction

Interpersonal communication is a whole made up of various integral components that ensure an effective interaction. As such, the verbal aspect of our interactions is significantly enhanced by the use of non-verbal aspects [3] including body language such as facial and body gestures. Some of these gestures may hold meaning on their own whereas others are used to accompany their correspondent in verbal communication. They can aid in the conveyance and comprehension of information, serve as a visual aid, regulate attention and much more [2]. As the use of gestures within speech comes naturally to most humans, it is an inevitable challenge for artificial agents to make human-like gestures along with speech in the field of human-robot interaction (HRI).

To ensure a harmonious combination of the gestures and speech, it is important to determine when gestures start and end and how they align with parts of the speech. Previous studies offer several possibilities to implement in gesture-speech coordination. These are aligning the prominent part of the gesture with: the focused word [1][8], the lexically stressed syllable [5], syllables with intonation peaks [7], or with no alignment [6]. In this project, however, the model used a virtual robot without control of the speech and was given a script annotated with gesture markups. Therefore, the focus was more on performing premade general gestures appropriately within the length of the speech in different situations rather than on where to start and end gestures.

In addition, understanding different types of gestures and using them in the appropriate semantic context are also significant. These different types are listed as deictic, beat, iconic and metaphoric gestures as in [2]. According to [2], deictic gestures bring attention to references in the environment, while beats mark significant parts of the speech, and iconics portray objects and events. Except for metaphoric gestures, we employed these three as the main points of reference to design our model.

The aim of this project is to design a generic model of robot that can tell a story multimodally. Based on the findings from the previous studies and our own analyses, we designed the model and created a short story to test the model. The remainder of the paper will elaborate how the model was designed, implemented and evaluated.

2. Design

In this project, the robot is given stories annotated by markup language. Therefore, the model has to read an annotated story file, segment the story into chunks, and execute a multimodal plan for each chunk. It was programmed in Python using NAOqi SDK and was executed on the virtual robot NAO using Choregraphe.

2.1. Annotated Story

Stories that the robot should tell are given as text explicitly annotated. The markups indicate which gesture should be accompanied by which part of the verbal phrase. The story we created was inspired by an excerpt from J.R.R Tolkien's *The Lord of the Rings: The Two Towers* [10]. It is originally in the form of a dialogue, but we edited it both to turn it into a story-like prose and to fit into the appropriate time period. It was chosen from among other options for its potential to support a variety of gestures to fully communicate our point. Then, we annotated the story with appropriate gesture markups. The details about gestures will be explained in the next section, and the annotated story is attached in the appendices.

Markups	Explanation
<g-beat head emphasis>	Move head
<g-beat hand emphasis>	Move hands and arms
<g-beat head down>	Look down
<g-beat head up>	Look up
<g-iconic hand scratch-chin>	Scratch chin
<g-iconic hand on-waist>	Put hands on the waist
<g-iconic hand big>	Make a big hand gesture
<g-iconic hand release>	Release the hands to sides
<g-iconic hand shrug>	Shrug
<g-deictic hand point-to-body>	Point itself
<g-deictic hand point-forward>	Point forward

Table 1: A list of gestures used in the model

2.2. Gestures

In order to find semantically appropriate gestures to align with the speech, we performed the speech ourselves and made a list of all possible gestures for the story. Then we polished the gestures to be generic enough to be applied to other stories based on theoretical foundations. For example, we changed some of our initial design choices that included animal imitations into subtler movements that spoke to a more general audience.

We used deictic gestures for introductions and emphases on the target to draw the audience's attention to the object we point at, such as the robot itself. Beat gestures were kept closer to the body and were used in places where some small natural moves seemed appropriate or an emphasis in speech was required. Iconic gestures were used to convey certain meanings or to put emphasis on certain points. Our specific gesture choices were rooted in further observations of daily interactions, both online and in real life. For instance, iconic gestures such as <g-iconic hand shrug>, <g-iconic hand scratch-chin>, and <g-iconic hand big> were based on commonly used emojis (🤷♀,🤔, and 💁) that have come to represent certain feelings.

The final list of gestures consists of 11 gestures (4 beats, 5 iconics, 2 deictics) as shown in Table 1. We created the gestures using Choregraphe's timeline and exported each gesture into a python code. The code containing a list of joint names, a list of timestamps, and a list of joints was saved as a python module.

2.3. Model

The model focused on synchronizing the timings of the premade gesture and the speech, as the semantics were already explored when the gestures were created. We followed Kendon's three phases [4] - preparation, nucleus, and retraction - which are commonly used in practice for temporal segmentation of the gestures. For this project, the model used only premade gestures and was given an annotated story which tells where exactly to perform each gesture. The model already knew that the start and end of the gesture should be the start and end of the annotated chunk of verbal phrase, and therefore, the preparation phase was not necessary. Instead, retraction also partially took a role of preparation for the next gesture as a transition between gestures.

First, the model reads an annotated story text given a name of the file and finds all the gesture tags. For each annotated gesture, it retrieves the angles and timestamps information that were saved earlier. If the model faces the closing tag such as *</g-beat>*, it goes into the three phases to execute the multimodal plan.

In the preparation phase, it counts the number of syllables of the utterance and scales the length of the gesture accordingly. To do the scaling, it multiplies the timestamps by the following weight (1).

$$weight \ = \ max \ (\ 1, \ \ \frac{((syllable\ count) * (duration\ per\ syllable))}{((gesture\ duration) + 1)} \ \ ) \tag{1}$$

The duration per syllable was set to 0.3 as the average amount of time it took per syllable for NAO to talk was 0.3 seconds. The number of syllables was computed from the story chunk, and the gesture duration was calculated as a maximum value of the timestamps. By setting the minimum weight as 1, it extends the duration of the gesture but not shrinks to guarantee the minimum duration of the gesture. Also, by adding 1 to the gesture duration, it doesn't let the weight get too big which may lead to very slow performance of the gesture. In case the gesture is nested in another one, the model simply appends the joints of the nested gesture to the joints of the bigger gesture. It marks the start of the nested gesture by counting the syllables of the chunk contained in the big gesture before the start of the tag, and the bigger gesture which has nested gesture information included in its joints goes through the same preparation process. With the adjusted timestamps and joints, the model executes the gesture and measures how long it actually took to perform the gesture.

After the performance (retraction phase), if the actual duration of the performance took less than expected, the model waits a bit until the next gesture so that the robot has time to finish the speech. Otherwise, it waits only for a preset interval which was set to 0.7 seconds as shown in the equation (2). In addition, to prevent abrupt start of the next gesture, if the syllable count is smaller than 5, which are usually one or two words, the model waits for extra 0.3 seconds.

$$wait \ = \ max \ (interval, \ ((syllable\ count) * (duration\ per\ syllable) - (actual\ duration) + interval)) \tag{2}$$

3. Evaluation

Following the example of [9], a survey was chosen as the method of evaluation. The questions were aimed to rate NAO's performance on two fronts, the speech and the gestures as well as the combination and coordination of these two. 14 naive raters who were not involved in the development of the model were recruited online by convenience sampling.

The participants were invited to watch a video demonstration of NAO's performance first and were asked to rate it on a 5-point scale. The questions were about the naturalness of the speech and gestures (1: artificial, 5: natural), liveliness and the engagement of the participant (1: not lively at all, 5: very lively), semantic and temporal synchrony and fluidity(1: not appropriate, 5: very appropriate), speed of the gestures (1: too slow, 5: too fast), and the quantity of the gestures (1: too few, 5: too many). At the end, an open ended question was added for users to leave their comments and additional thoughts they might have. No demographic information was requested from the potential participants. The percentages of their answers are given in Table 2.

Least Ideal 🟨🟨🟨🟧🟥 **Most Ideal**

Questions	1	2	3	4	5
How natural was NAO's speech?	0%	**57.1%**	28.6%	14.3%	0%
How natural were NAO's gestures?	0%	0%	**42.9%**	**42.9%**	14.3%
How lively did you find NAO?	0%	14.3%	21.4%	**50%**	14.3%
Did you find the speech and gesture were semantically matching?	0%	0%	21.4%	**57.1%**	21.4%
Did you find the speech and gesture were well synchronized?	0%	7.1%	7.1%	**50%**	35.7%
Did you find the hand and arm movements fluid?	0%	7.1%	**35.7%**	**35.7%**	21.4%
How did you think of the amount of gestures performed by NAO?	0%	0%	**50%**	**50%**	0%
What did you think about the speed of the gestures?	0%	14.3%	**78.6%**	7.1%	0%

Table 2: Responses from the survey for evaluation of the model

The results show that the majority of our participants finds NAO’s voice unnatural. Answers to our final open-ended question further emphasize this discontent with more feedback regarding the problems with the intonation of the speech, the constant pauses and the unclarity of the audio. On the other hand, our gestures seem to perform rather well with a higher percentage of people opting for the more neutral or optimal ratings.

4. Conclusion

The results from our participants show that, although our NAO is definitely not on-par with a human story-teller, it still performs fairly well. To improve its performance in future projects, we list here some of the further considerations that can be made based on the feedback from our participants and our own observations from the literature.

Firstly, we had chosen to use NAO’s own voice to preserve the authenticity of a robot story-teller. However, this gave us little freedom to make changes to the audio, resulting in many complaints about the intonation, fluency and understanding of the speech. Better speech performance by NAO might increase ratings for both speech and gesture - speech combination.

Secondly, our gestures are limited by the range of motion that NAO can perform. The foremost among these are the lack of facial expressions, and the limited movement of NAO’s finger, wrist and ankle movements. We believe that these joints play an important role in the execution of subtler movements that may be involved in communication such as finger or foot tapping, for which our trials went horribly wrong. The inclusion of these joints would not only make the use of a variety of new gestures possible, it could also improve the execution of some of the gestures that already exist within our project.

Moving on to changes within our control, building the more complex gestures by combining simpler gestures might be a worthy alternative. Although for this project, the combination of such gestures make the annotated story too complicated, and thus is not optimal, for future projects that include more gestures, it might prove more efficient.

Lastly, establishing a more concrete framework for the gesture-speech correspondence might improve performance. Actually implementing the gesture coordination methods discussed in the introduction into the design process might be a start. However, in the absence of human speech, this is rather difficult to put to test as NAO's voice does not always hold the natural stresses or intonations of human speech. If we come to a point where artificial agents can produce natural human speech, further consideration should be given to such theories to improve performance.

References

[1] Butterworth, B., & Beattie, G. (1978). Gesture and Silence as Indicators of Planning in Speech.

[2] Huang, Chien-Ming & Mutlu, Bilge. (2013). Modeling and Evaluating Narrative Gestures for Humanlike Robots. 10.15607/RSS.2013.IX.026.

[3] Jokinen, K., & Wilcock, G. (2014). Multimodal Open-Domain Conversations with the Nao Robot. *Natural Interaction with Robots, Knowbots and Smartphones, Putting Spoken Dialog Systems into Practice*.

[4] Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511807572

[5] Loehr, D. P. (2007). Aspects of rhythm in gesture and speech. Gesture, 7, 179–214.

[6] McClave, E. (1994). Gestural beats: The rhythm hypothesis*. Journal of Psycholinguistic Research,* 23, 45–66*.*

[7] Nobe, S. (1996). Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network threshold model of gesture production.

[8] Roustan, B., & Dohen, M. (2010). Co-Production of Contrastive Prosodic Focus and Manual Gestures : Temporal Coordination and Effects on the Acoustic and Articulatory Correlates of Focus.

[9] Salem, M., Kopp, S., Wachsmuth, I. et al. Generation and Evaluation of Communicative Robot Gesture. *Int J of Soc Robotics 4, 201–217* (2012). https://doi.org/10.1007/s12369-011-0124-9

[10] Tolkien, J. R. R. (1955). The Two Towers. London: George Allen & Unwin.

Appendices

Appendix A: the Annotated Story

<g-deictic hand point-to-body>I am an Ent,</g-deictic>
<g-beat head emphasis>or that's what they call me.</g-beat>
<g-deictic hand point-forward>What are you,</g-deictic>
<g-iconic hand shrug>I wonder? </g-iconic>
<g-iconic hand on-waist>You are in my country </g-iconic>
<g-beat hand emphasis>but you do not seem to come in the old lists that I learned when I was young.</g-beat>
<g-iconic hand scratch-chin>Let me see! <g-beat head emphasis> How did it go?</g-beat></g-iconic>
<g-beat hand emphasis>Beaver the builder, buck the leaper,Bear bee-hunter, boar the fighter;</g-beat>
<g-beat head down>Hound is hungry</g-beat> <g-beat head up> hare is fearful</g-beat>
<g-iconic hand scratch-chin>Hoom, hm; It was a long list. </g-iconic>
<g-iconic hand on-waist>But anyway <g-beat head emphasis>you do not seem to fit in anywhere!</g-beat></g-iconic>
<g-iconic hand big>Why not make a new line?</g-iconic>
<g-iconic hand release>Half-grown hobbits, the hole-dwellers.</g-iconic>