Annotation of Utterances for Conversational Nonverbal Behaviors

Allison Funkhouser, Reid Simmons

Robotics Institute, Carnegie Mellon University

Abstract. Nonverbal behaviors play an important role in communication for both humans and social robots. However, adding contextually appropriate animations by hand is time consuming and does not scale well. Previous researchers have developed automated systems for inserting animations based on utterance text, yet these systems lack human understanding of social context and are still being improved. This work proposes a middle ground where untrained human workers label semantic information, which is input to an automatic system to produce appropriate gestures. To test this approach, untrained workers from Mechanical Turk labeled semantic information, specifically emotion and emphasis, for each utterance, which was used to automatically add animations. Videos of a robot performing the animated dialogue were rated by a second set of participants. Results showed untrained workers are capable of providing reasonable labeling of semantic information and that emotional expressions derived from the labels were rated more highly than control videos. More study is needed to determine the effects of emphasis labels.

1 Introduction

Nonverbal behaviors are an important part of communication for both humans and social robots. Gestures and expressions have the ability to convey engagement, to clarify meaning, and to highlight important information. Thus the animation of nonverbal behaviors is an important part of creating engaging interactions with social robots. Yet adding contextually appropriate animations by hand is time consuming and does not scale well as the number of utterances grows larger.

One alternative is to create rule-based software that assigns animations automatically based on the text of the dialogue. Such pipelines have the benefit that, once implemented, much left effort is needed in order to add new utterances to a database of dialogue. Examples of such automated systems include the Behavior Expression Animation Toolkit [1], the Autonomous Speaker Agent [2], and the automatic generator described in [3]. These systems use lexical analysis to determine parts-of-speech, phrase boundaries, word newness, and keyword recognition. This information is then used to place gestures such as head rotations, hand movements, and eyebrow raises.

However, these automated pipelines also have drawbacks. Because there is no longer a human in the loop, the entire system depends only on the information that can be automatically extracted from raw text. While there have been great strides forward

in natural language understanding, there is still progress to be made. Specifically, classification of emotions based on text is a difficult [4], and current methods would constrain the number of emotions classified and thus limit the robot's expressivity. Also, determining the placement of emphasis gestures currently relies on word newness – whether a word or one of its synonyms was present in previous utterances in the same conversation. The complexity of language and speakers' reliance on common ground [5] create situations where implied information is not necessarily explicitly stated in previous sentences, which makes this form of emphasis selection less robust.

This work considers a potential middle ground between hand tuning animation and an automated pipeline with no humans involved. Instead, an annotator could add labels specifying particular semantic information, such as emphasis and emotion, which would then be input to an automatic system to produce appropriate gestures. This strategy allows the relevant human-identifiable context of a scenario to be preserved without requiring workers to have deep expertise of the intricacies of nonverbal behavior.

In order to test this labeling strategy, untrained workers from Amazon Mechanical Turk read small segments of conversations and answered questions about the semantic context of a particular line of dialogue. This semantic information was input to an automated system which added animations to the utterance. Videos of a Furhat robot performing the dialogue with animations were rated by the phase two participants. Results showed that untrained workers were capable of providing reasonable labeling of semantic information. When these labels were used to select animations, the selected emotive expressions were rated as more natural and anthropomorphic than control groups. More study is needed to determine the effect of the labeled emphasis gestures on perception of robot performance.

2 Related Works

Existing systems for streamlining the animation process can be divided into three categories: rule based pipelines using lexical analysis, statistics based pipelines that draw on videos of human behavior, and markup languages using tags from expert users. Examples of each of these strategies are discussed below.

The Behavior Expression Animation Toolkit [1] generates XML style tags that mark the clause boundaries, theme and rheme, word newness, and pairs of antonyms. These tags are used to suggest nonverbal behaviors including hand and arm motions, gaze behaviors, and eyebrow movements. Beat gestures are suggested for new words occurring in the rheme, the part of the clause presenting new information. Iconic gestures are inserted when an action verb in the sentence rheme matches a keyword for an animation, such as an animation that mimes typing on a keyboard corresponding to the word *typing*. Contrast gestures mark the distinction between pairs of antonyms. Robot gaze behaviors are based on general turn-taking patterns. Finally, a conflict resolution filter processes the suggested nonverbal behaviors, identifies conflicts where simultaneous gestures use the same degrees of freedom, and removes the lower priority gestures in these conflicts.

The Autonomous Speaker Agent [2] uses a phrase tagger to determine morphological, syntactic, and part-of-speech information about the given text. Similar to the Behavior Expression Animation Toolkit, the Autonomous Speaker Agent also records word newness based on previously mentioned words in a given utterance. This lexical data is used to assign head movements, eyebrow raising, eyes movements, and blinks for a virtual character through the use of a statistical model of facial gestures. To build this statistical model, videos depicting Swedish newscasters were hand labeled with blinks, brow raises, and head movements.

In the text-to-gesture system described in [6], the hand gestures of speakers on TV talk shows were manually labeled as belonging to one of six side views and one of five top views. A morphological analyzer was used to label the parts of speech for the words in the spoken Korean utterances, and these labels were correlated to speaker gestures. Specifically, certain combinations of content words and function words were indicative of either deictic, illustrative, or emphasizing gestures. This mapping data was used to select movements from a library of learned gestures.

The Behavior Markup Language [7] is an XML style language that allows specific behaviors for virtual characters to be defined and synchronized with text. Behavior elements include movements of the head (such as nodding, shaking, tossing, and orientation), movements of the face, (including eyebrows, mouth, and eyelids), and arm and hand movements, to name a few. Because the original design was for virtual characters with humanlike appearances, it assumes that the character's possible motions include these humanoid style degrees of freedom. A robot that lacked these degrees of freedom – such as not being capable of certain facial movements, head movements, or arm motions – would not have a way of realizing all possible labeled motions. Furthermore, a nonhumanoid robot could potential have many other degrees of freedom not covered by this humanoid-centric markup. Using the Behavior Markup Language to command such a robot would lead to these potentially expressive motions not being used. The low level nature of the highly specific action commands makes this markup language less suitable for use across a wide variety of diverse robot platforms.

3 Approach

The goal of this work is to explore streamlining the robot animation process by having untrained workers label specific semantic information for each utterance, which is then used to determine appropriate nonverbal behaviors. Like the automated pipelines, this approach helps reduces the amount of human labor required to add animations when compared to animating each utterance by hand. The in-depth, low level knowledge of animation and the precise timing and types of gestures can be handled by an automatic pipeline. This speeds up the human work by reducing the task to merely labeling sentences, as opposed to meticulously tuning each set of degrees of freedom. However, because the labeling process still involves human input, it still allows for some of the subtleties gained from a human knowledge of interactions that is present in hand done animations.

While there are many possible pieces of information that annotators could conceivably mark, in this work we limit the scope to emphasis location – the word in the sentence that receives the most verbal stress – and dominant emotion. The envisioned implementation uses an XML tagging format, shown in the example below. The XML format is easily extensible and could potentially be combined with existing or future automated pipelines, such as the Speech Synthesis Markup Language.

Raw Text: Oh really? I didn't know that.
Annotated Text: <emotion=surprised> Oh really? </emotion> <emotion=embarrassed> I didn't <emphasis>know</emphasis> that. </emotion>

Another benefit of this overall approach is the independence from any specific robot platform. While tags specify what emotion is expressed, they do not dictate how this should be shown. Robotic platforms can be quite diverse, and even humanoid robots will not all be capable of the same degrees of freedom. When lower level specifications are used to define movement of certain degrees of freedom, the implementation is constrained to platforms that are capable of those specific motions. Choosing to label higher level concepts means that any robot, humanoid or not, could be programed to take advantage of these tags – it would only need to have some behavior that conveyed emphasis or expressed emotion.

These higher level labels can also be used to create greater variability in a robot's behavior. If a robot's animation library contains multiple animations that convey the same emotion, or multiple types of gestural emphasis, then the robot could select different animations each time it says an utterance while maintaining the original meaning. Thus, even if a robot is forced to repeat a particular dialogue line multiple times, different animations could be used so that the movements and expressions would not be identical. This could make the repetition less noticeable, since the performance would not be exactly the same.

Furthermore, while the current proposition is for these labels to be assigned by people, it would be preferable if eventually a machine learning algorithm was able to do this process instead. Having people create a large number of annotated utterances for robot performances thus serves a secondary purpose of creating labeled training and testing data that could be used for future machine learning.

4 Experiment

One of the main goals for this approach was to accommodate labeling by people who have no background in robotics or animation. To test this we performed an on-line experiment. In the first phase of the experiment, a group of Amazon Mechanical Turk workers were presented with transcripts of several short conversations, which they used to answer questions about the emotion and emphasis of a particular dialogue line. In phase two, videos of a robot performing the animated dialogue were rated by a second set of participants. The workers from Mechanical Turk must be at least 18 years old be able to accept payments through a U.S. bank account. No other restrictions were placed

on participants, which meant the participants could be of any education level, and would not necessarily have any prior experience with robots or animation of behaviors.

In phase one, Turkers were asked to read each short conversation out loud to themselves before answering the questions, paying specific attention to how they naturally said each line of dialogue. This was intended to help participants determine the location of verbal emphasis by having them consider how they would naturally say the sentence. Because of the correlation between verbal and gestural emphasis [8], it was possible to specifically ask participants about their verbal emphasis while speaking the sentence without needing them to consider what gestures they might make while talking. Participants also selected the emotion most associated with the utterance from a list of possible emotions: Excited, Happy, Smug, Surprised, Bored, Confused, Skeptical, Embarrassed, Concerned, Sad, and Neutral. This list was specifically made to be more extensive than the previously mentioned classification algorithms in order to more fully explore the amount of nuance that people could distinguish, especially since, ideally, social robots should eventually be capable of expressing a wide range of emotions.

Once this data was collected, it was used to animate the utterances, which were then performed by the Furhat robot shown in **Fig. 1**. A script read in the Mechanical Turk data, used the participant responses to select animations based on the consensus of Turker selections, and output tagged utterances that were performed by the robot. Based on the data collected in phase one, five utterances were chosen which showed the best consensus on emphasis location, and another five utterances were selected which showed the best consensus on dominant emotion. The ones selected for emphasis received either 75% or more of their selections on a single word, or a pair of adjacent words received a combined of more than 75%. The chosen utterances for dominant emotion received either more than 70% of selections or the selections for two similar emotions (sad/concerned, happy/excited, or confused/skeptical) received a combined percentage of more than 70%.

Eyebrow motions were chosen as the emphasis gesture because they were easier to precisely synchronize with a specific word compared to longer motions such as nodding. Based on the observations from [9], eyebrow raises were used for positive emotional utterances and eyebrow frowns were used for the negative emotional utterances. Small facial movements were added to each of the control group performances to prevent the control videos from being seen as arbitrarily less appealing due to lack of motion.

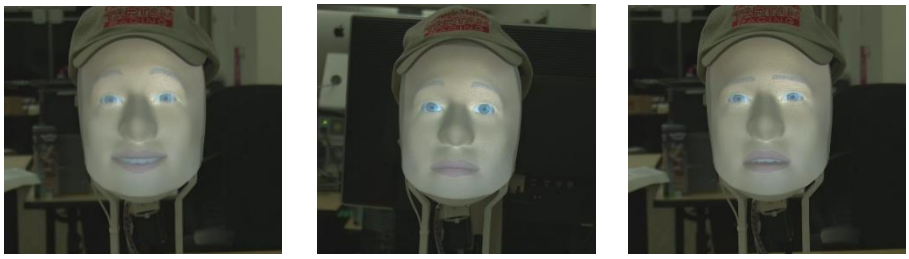

Fig. 1. Furhat Expressions – Happy (left), Neutral (center), and Unhappy (right)

In phase two of the experiment, a subset of these animated expressions were viewed and rated by a separate set of turkers. Videos of the robot showed either an emotive expression or an emphasis gesture. This was done so emotion and emphasis could be evaluated independently. While animated expressions for Furhat were created for all eleven emotions from phase, in the validation phase only two emotive expressions were used: Happy and Unhappy. This was so the videos showing the incorrect emotional expression could clearly be directly opposite the correct emotional expression. Each participant viewed two videos of the robot performing the same utterance and compared the videos on several scales. One video was a control video showing no emphasis and a neutral expression. The other video would represent one of four categories: a video with an emphasis gesture accenting the word that received the majority of selections in phase one, a video with an emphasis gesture at an incorrect location (accenting a word that received 10% or less of the phase one selections and was not adjacent to the word chosen by consensus), a video with an emotive expression that matched the consensus from phase one, or a video with an emotive expression that opposed the emotion from the phase one consensus.

Phase two participants rated which of the two videos they viewed was most believable, humanlike, appropriate, pleasant, and natural. The metrics humanlike, natural, and pleasant were taken from the Godspeed Questionnaire Series [10]. The believable and appropriate metrics were added in order to distinguish between cases where the expression appeared realistic in isolation but did not match the dialogue. Each pair of videos was rated by twenty participants.

5 Results

The charts detailing the phase one participant responses concerning word emphasis can be seen in Figure 2. Out of the eight utterances presented, four contained words that received at least 75% of participant selections for that utterance. Three of these utterances had words that received 90-95% of the selection. This represents strong indication that these words should be emphasized. In the remaining utterances, one (utterance 5) showed participant answers clustered around the noun-adjective pair “really worried”. These two words together made up 90% of the selections for this utterance, with an even split of selections between the two words. This shows that emphasized phrases were able to be identified. When presented a multi-sentence utterance – utterance 6 – the participant selection was split between two words, one from each sentence. This again shows that bimodal distributions will be visible in the data. The two remaining utterances (4 and 8) show that it is possible to determine when a particular utterance does *not* have a strong candidate for word level emphasis, displaying a wider spread of participant selection.

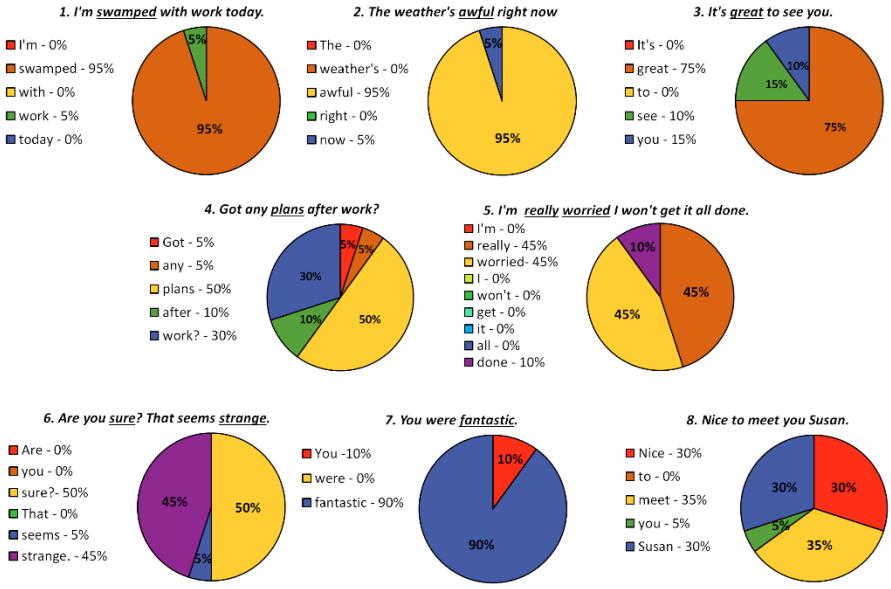

Fig. 2. Emphasis percentages for each utterance

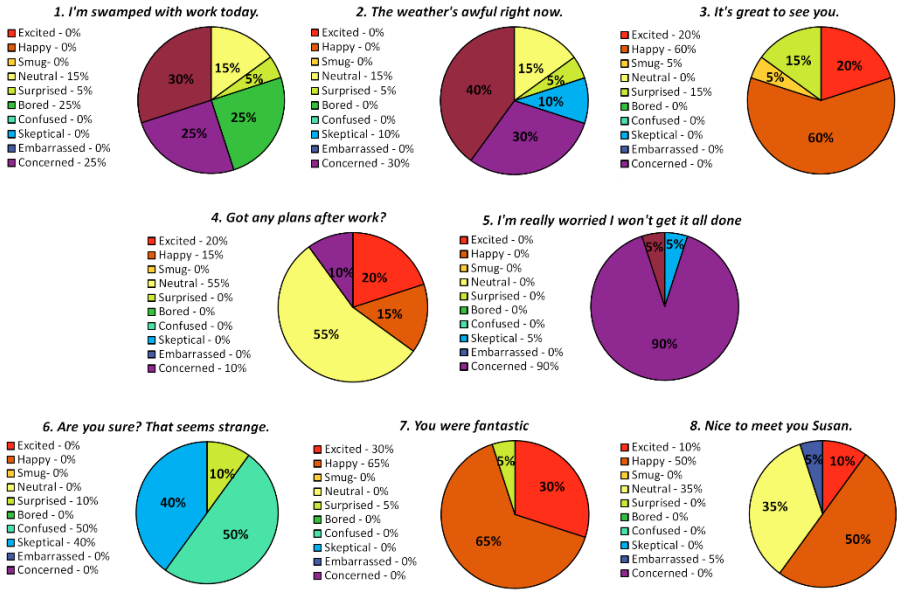

Fig. 3. Emotion percentages for each utterance

The charts detailing phase one responses for selected emotions can be seen in Figure 3. In utterance 5, "I'm really worried I won't get it all done," 90% of participants selected the concerned emotion. Such clear consensus is likely because the phrase "I'm

really worried" specifically calls out the speaker's emotion, and so responses cluster around the nearest related emotion, concern. For four of the other utterances, participant selections were split between two closely related emotions that together accounted for at least 70% of responses. In each of these cases – happy/excited, sad/concerned, and confused/skeptical – the two most chosen emotions expressed similar emotions with relatively close valence values. This shows a significant number of the participants were interpreting the utterances in similar manners, even if they chose slightly different emotions. Of the remaining utterances, utterance 4 had 55% selection for neutral, with the other selections divided fairly evenly between three other emotion options. This suggests that there is no strong emotion associated with this sentence, and the expression should be left neutral. Utterance 8 had 50% happy and 10% excited, and utterance 1 was 30% sad and 25% concerned. While this gives some suggestion of possible emotions, it is not as strong of a consensus by comparison.

Tables 1 through 4 show the results of the direct video comparison survey questions from phase two. Chi-square tests were used to evaluate the significance of the data. The chi-square test is a statistical method assessing the goodness of fit between observed values and those expected if the null hypothesis was true. In this case the null hypothesis would mean no difference between the animated video and the control video, therefore producing an even split of 10 participants selecting the control video for every 10 that selected the experimental video. In order to reduce the risk of Type 1 errors all five metrics – humanlike, natural, believable, appropriate, and pleasant – were evaluated as a part of the same chi-square group for each utterance.

The robot performances that used the emotion selected by consensus in phase one were consistently rated more highly by participants when compared to the neutral control videos. All five test utterances received significant chi-square results, with p values ranging from 0.0001 to 0.0329. This confirms that people can assign emotions that are viewed as appropriate. Furthermore, of the videos shown where the emotion opposed the one chosen in phase one, four of the five received statistically insignificant results when compared to the control videos, with p-values ranging from 0.2757 to 0.7399. For these four utterances, adding a mismatched emotional expression performed no better than a neutral face. Overall, the videos showing expressions that matched the phase one responses were rated as significantly more humanlike than the control videos.

Table 1. Percent of Participants that chose the Matched Emotion

	*Utterance 2**	*Utterance 3**	*Utterance 5**	*Utterance 6**	*Utterance 7**
Humanlike	80%	80%	75%	90%	70%
Natural	85%	80%	65%	85%	60%
Believable	75%	70%	70%	90%	75%
Appropriate	80%	80%	75%	85%	85%
Pleasant	60%	80%	70%	65%	70%
Chi-Squared	30.000	32.000	18.200	47.000	26.000
p-value	0.0004*	0.0002*	0.0329*	0.0001*	0.002*

Table 2. Percent of Participants that chose the Mismatched Emotion

	Utterance 2	*Utterance 3*	*Utterance 5*	*Utterance 6*	*Utterance 7**
Humanlike	65%	65%	55%	60%	80%
Natural	60%	65%	50%	50%	75%
Believable	55%	65%	45%	45%	70%
Appropriate	55%	65%	50%	45%	70%
Pleasant	70%	60%	75%	85%	65%
Chi-Squared	6.200	8.000	6.000	11.000	20.400
p-value	0.7197	0.5341	0.7399	0.2757	0.0156*

The data from the emphasis surveys is less clear. Table 3 shows that for three of the five utterances, the videos showing correct emphasis were selected significantly more than their control video counterparts. However, the remaining two utterances resulted in very high p-values. Furthermore, three of the videos showing incorrect emphasis also yielded statistically significance, as shown in Table 4. Thus the videos showing emphasis locations selected in phase one did not appear more realistic or believable overall compared to emphasis at other locations. This could indicate that even with the small random motions added to the neutral expression in the control video, the more obvious motion of the eyebrow raises and frowns was appealing for the sake of being more animated, regardless of the location of the emphasis.

Table 3. Percent of Participants that chose the Correct Emphasis

	Utterance 1	*Utterance 2**	*Utterance 3**	*Utterance 5*	*Utterance 7**
Humanlike	55%	80%	75%	45%	80%
Natural	55%	80%	75%	50%	80%
Believable	55%	80%	70%	50%	80%
Appropriate	60%	85%	75%	50%	85%
Pleasant	50%	75%	75%	60%	90%
Chi-Squared	1.400	36.400	23.200	1.000	44.200
p-value	0.9978	0.0001*	0.0058*	0.9994	0.0001*

Table 4. Percent of Participants that chose the Incorrect Emphasis

	*Utterance 1**	*Utterance 2**	*Utterance 3**	*Utterance 5*	*Utterance 7*
Humanlike	75%	80%	80%	60%	70%
Natural	75%	80%	80%	60%	70%
Believable	75%	80%	75%	55%	70%
Appropriate	75%	75%	75%	50%	70%
Pleasant	65%	70%	85%	60%	70%
Chi-Squared	21.8000	29.800	34.200	2.600	16.000
p-value	0.0095*	0.0005*	0.0001*	0.9781	0.0669

6 Conclusions

This work proposed an approach for reducing the time spent animating each utterance of a social robot. We found untrained workers were capable of providing reasonable labeling of semantic information in a presented utterance. When these labels were used to select animations for a social robot, the selected emotive expressions were rated as more natural and anthropomorphic than control groups. More study is needed to determine the effect of the labeled emphasis gestures on perception of robot performance.

Acknowledgements. We are thankful to Disney Research and The Walt Disney Corporation for support of this research effort. This material is based upon research supported by (while Dr. Simmons was serving at) the National Science Foundation.

References

1. Cassell, J., Vilhjalmsson, H., & Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 477-486. (2001)
2. Smid, K., Pandzic, I. S., & Radman, V.: Autonomous Speaker Agent. In: Proceedings of Computer Animation and Social Agents Conference. (2004)
3. Albrecht, I., Haber, J., & Seidel, H. P.: Automatic Generation of Non-Verbal Facial Expressions from Speech. In: Advances in Modelling, Animation and Rendering, pp 283-293. Springer London. (2002)
4. Perikos, I., & Jatzilygeroudis, I.: Recognizing emotions in text using ensemble of classifiers. In: Engineering Applications of Artificial Intelligence, vol. 51, pp. 191-201. (2016)
5. Kiesler, S. C.R.: Fostering Common Ground in Human-Robot Interaction. In: IEEE International Workshop on Robot and Human Interactive Communication, pp. 729-734. (2005)
6. Kim, H. H., Lee, H. E., Kim, Y. H., Park, K. H., & Bien, Z. Z.: Automatic Generation of Conversational Robot Gestures for Human-friendly Steward Robot. In: The 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 1155-1160. (2007)
7. Kopp, S., Krenn, B., Marsella, S., Marshal, A. N., Pelachaud, C., Pirker, H., Thorisson, K. R., Vilhjalmsson, H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In: Proceedings of the 6th International Conference on Intelligent Virtual Agents, pp. 205-217. (2006)
8. Graf, H. P., Cosatto, E., Strom, V., & Huan, F. J.: Visual Prosody: Facial Movements Accompanying Speech. In: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 396-401. (2002)
9. Zoric, G., Smid, K., & Pandzic, I. S.: Facial Gestures: Taxonomy and Application of Non-Verbal, Non-Emotional Facial Displays for Embodied Conversational Agents. In: Conversational Informatics: An Engineering Approach, pp. 161-182, John Wiley & Sons, Ltd. (2007)
10. Bartneck, C., Croft, E., Kulic, D., & Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. In: International Journal of Social Robotics, vol. 1, no. 1, pp. 71-81. (2009)