Annotation of Utterances for Conversational Nonverbal Behaviors

Allison Funkhouser

CMU-RI-TR-16-25

Submitted in partial fulfillment of the

requirements for the degree of

Masters of Science in Robotics

May 2016

Masters Committee:

Reid Simmons, Chair

Illah Nourbakhsh

Heather Knight

Abstract

Nonverbal behaviors play an important role in communication for both humans and social robots. However, hiring trained roboticists and animators to individually animate every possible piece of dialogue is time consuming and does not scale well. This has motivated previous researchers to develop automated systems for inserting appropriate nonverbal behaviors into utterances based only on the text of the dialogue. Yet this automated strategy also has drawbacks, because there is basic semantic information that humans can easily identify that is not yet accurately captured by a purely automated system. Identifying the dominant emotion of a sentence, locating words that should be emphasized by beat gestures, and inferring the next speaker in a turn-taking scenario are all examples of data that would be useful when animating an utterance but which are difficult to determine automatically.

This work proposes a middle ground between hand-tuned animation and a purely text-based system. Instead, untrained human workers label relevant semantic information for an utterance. These labeled sentences are then used by an automated system to produce fully animated dialogue. In this way, the relevant human-identifiable context of a scenario is preserved without requiring workers to have deep expertise of the intricacies of nonverbal behavior. Because the semantic information is independent of the robotic platform, workers are also not required to have access to a simulation or physical robot. This makes parallelizing the task much more straightforward, and overall the amount of human work required is reduced.

In order to test this labeling strategy, untrained workers from the Amazon Mechanical Turk website were presented with small segments of conversations and asked to answer several questions about the semantic context of the last line of dialogue. Specifically, they

selected which emotion best matched the emotion of the sentence and which word should receive the most emphasis. This semantic information was input to an automated system which added animations to the particular utterance. Videos of a social robot performing the dialogue with animations were then presented to a second set of participants, who rated them on scales adapted from the Godspeed Questionnaire Series.

Results showed that 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.

Contents

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 hand created animation is time consuming. Even once a library of animations is created, adding in contextually appropriate gestures and expressions to every single line of dialogue – and hand tuning the timing of those gestures – still takes time and does not scale well as the number of possible utterances grows larger. This can cause designers to choose to limit the overall quantity of dialogue options, to only animate expressions on a select few dialogue lines, or to repeatedly reuse a small number of familiar animations.

Each of these strategies has downsides. Reducing the total dialogue means that people are more likely to encounter a situation where the robot does not have a specific response. Even in the subject areas where the robot *can* respond, it will be more likely to repeat itself and thus remind people of its limited capabilities and diminish the experience. If the majority of the robot's lines are performed without expression or gestures this forgoes the benefits of nonverbal behaviors. Repeated reuse of a small subset of available animations could lead to the same user awareness of the interaction's constructed nature as repeated dialogue. Each of these methods limit the full potential of robots as conversational agents.

One alternative to hand tuned animation is to create rule-based software which assigns animations automatically according to 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 (Cassell, Vilhjalmsson, & Bickmore, 2001), the Autonomous Speaker Agent (Smid, Pandzic, & Radman, 2004), and the automatic generator described in (Albrecht, Haber, & Seidel, 2002). 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 problem (Perikos & Jatzilygeroudis, 2016), 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 (Kiesler, 2005) 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 for individual lines and an automated pipeline with no humans involved. Instead, the author of the original dialogue could add labels specifying particular semantic information while they were composing the utterances, or a separate annotator could go through and add semantic labels later. This would allow the relevant human-identifiable context of a scenario to be preserved without requiring workers to have deep expertise of the intricacies of nonverbal behavior.

While there are currently existing markup languages that allow trained animators to quickly add gestures to utterances, these markups operate at a very low level. The Behavior Markup Language (Kopp, et al., 2006) specifies small details such as eyebrow movements, gaze shifts, and the synchronization points for the preparation, execution and relaxation of hand gestures. The average untrained worker would not immediately be able to use such a complex tool to its full potential because specialized knowledge of nonverbal behaviors is required. Roboticists have conducted studies of motion capture data and coded video recordings in order to extract precise relationships between nonverbal behaviors and speech content. Nodding at strong phrase boundaries (Isha, Liu, Ishiguro, & Hagita, 2010), gaze aversion as a turn-taking signal (Andrist, Tan, Cleicher, & Mutlu, 2014), and being more likely to blink on emphasized words (Zoric, Smid, & Pandzic, 2007) are not necessarily innately intuitive concepts. Taking advantage of this knowledge would require some level of study and practice.

Our goal was to decouple the high level semantic information from these low level implementation details. In this way, untrained workers who would not be able to provide low level specifics on animation implementation could instead provide labels for high level concepts which are more intuitively understood. Then these labels would be used by an automated pipeline to assign the lower level details for gestures and expressions. The following hypotheses help test the viability of this semantic labeling based approach:

(1) Untrained workers are capable of providing reasonable labeling of semantic information in a presented utterance.

(2) Nonverbal behaviors assigned based on the labeling in (1) will be rated as more natural and anthropomorphic than control groups.

We performed two experiments to test these hypotheses and found there was strong evidence to support the first hypothesis and some evidence to support the second hypothesis.

2. Background

2.1 Emphasis Gestures

First it is useful to establish what is meant by emphasis gesture or beat gesture in contrast to other categories of gestures. According to McNeill's system for classifying gestures – summarized in (Wagner, Malisz, & Kopp, 2014) – deictic gestures are used to refer to locations in physical space, such as using an index finger to point at an object or to gesture in a particular direction. Iconic gestures resemble the object or action they refer to, including spreading the hands to convey a particular width or making a circular motion to represent rotation. Emblematic gestures have culturally agreed upon meanings without appearing physically similar to what they signify, such as a thumbs up conveying approval or a waving hand serving as a greeting.

Beat gestures are simple, quick movements that synchronize with events in speech (also referred to as *batons* in Ekman's classification in (Ekman & Friesen, 1972). Unlike deictic, iconic, and emblematic gestures, beat gestures do not have inherent meaning if separated from their speech utterance. Instead, beat gestures help convey what information is important in a sentence by accenting a specific word or phrase in the dialogue.

While both McNeill and Ekman focus on emphasis through hand movements, other gestures can also be used to emphasize specific spoken words. The following emphasis gestures are described in (Zoric, Smid, & Pandzic, 2007). Head motions such as nodding can

be used to accentuate what is being said, synchronized at the word level. Eyebrow movements act as conversational signals by accenting a word or word sequence, with eyebrow raises (brows moving up and then down) occurring with positive affirmations and eyebrow frowns (brows moving down and then up) corresponding to speaker distress or doubt. Furthermore, while periodic eye blinks serve the biologic need to wet the eyes, blinks will also occur that synchronize with a word or pause. These described head and facial movements can allow robots to make use of gestural emphasis even if the robot is capable of little or no arm movement.

Different placements for emphasis can be chosen to highlight specific information (Graf, Cosatto, Strom, & Huan, 2002). This means there can be many valid choices of emphasis placement for a particular sentence. Consider an example sentence "I never said she ate those cookies" and imagine it spoken with different choices of emphasis. "I never said she ate ***those*** cookies" with emphasis on "those" implies the subject *did* eat some cookies just not those particular ones. "I never said ***she*** ate those cookies" indicates the cookies were eaten by someone else. "***I*** never said she ate those cookies" suggests this information was originally said by someone else. "I never ***said*** she ate those cookies" means it was not said outright but instead implied. "I never said she ***ate*** those cookies" indicates the subject took the cookies without eating them. "I never said she ate those ***cookies***" insinuates the subject did eat something, but not the cookies.

It is clear that having only the raw text of this utterance does not provide enough context to determine which of these six possible placements for emphasis is best. Knowledge of context is necessary in order to determine which piece of information should be highlighted.

This makes defining an automated system for determining the correct placement of an emphasizing gesture challenging.

One instance of such an automated animation system is the Behavoir Expression Animation Toolkit (Cassell, Vilhjalmsson, & Bickmore, 2001). In the language tagging module of this pipeline, each incoming noun, verb, adjective, or adverb was tagged with a newness value depending on whether the word or a synonym was previously tagged in the conversation. This word newness value was then used to determine which words should receive emphasis through verbal stress, eyebrow movements, or hand movements.

In some simple cases such a system could correctly identify the new information presented in an utterance. In the example conversation snippet "Who is he?" "He is a student." the words *he* and *is* from the second utterance have already been observed in the previous dialogue turn. The word *a* is an article and therefore is not considered for emphasis in this system. This leaves *student* as the only possible word to be emphasized.

However, language is complicated. There is no guarantee that the previous turns of dialogue will contain clear matches for all of the already established information, even when making use of synonyms. Consider the simple conversation below:

John: Good morning Martin. How was your drive in?

Martin: The weather's awful right now.

In this case, we as humans have enough prior knowledge about car travel to understand that the weather is a variable that factors into the quality of one's travel experience. We also understand that John is greeting Martin, and thus Martin has only arrived at the current

location recently. Therefore the words *weather*, *right*, and *now* are not the most important new information – the information to emphasize is that the weather is *awful*. And yet the words *weather*, *right*, and *now* would likely have been marked as "new" if this conversation was put through the previously described pipeline, thus making it unclear which word should be emphasized, if any.

2.2 Emotive Expressions

Currently, many emotion recognition tasks focus on extracting features from audio recordings of people speaking (Bhaskar, Sruthi, & Nedungadi, 2015). These methods rely on acoustic features such as pitch and speaking rate (Bhaskar, Sruthi, & Nedungadi, 2015) and can employ a variety of machine learning classifiers (Callejas & Lopez-Cozar, 2008). The authors of (Cowie, et al., 2008) considered additional features by making use of video recordings of speakers' facial expressions in addition to the audio data. Emotion recognition from auditory or visual data has applications in affective computing, when it is useful for a computer to identify the emotion of the person interacting with it, such as identifying when a user of a telephone based dialogue tree is feeling frustrated with the system (Callejas & Lopez-Cozar, 2008). In these situations audio recordings of each spoken utterance already exist. However, when the goal is to *generate* robot behavior, recordings of a human performing each utterance don't necessarily exist. Therefore making use of strategies that rely on audio or visual features would require an additional step of orchestrating human performances of every line of dialogue.

Text based emotion classifiers exist, though they can be limited in scope. Sentiment analysis refers to the study of people's "opinions, attitudes and emotions toward an entity," and mainly considers users' opinions on consumer products, determined from product

reviews, in order to provide feedback for businesses (Medhat, Hassan, & Korashy, 2014). These machine learning classifiers focus on determining sentiment polarity, namely whether a review was positive, negative, or neutral.

The authors of a very recent study (Bhaskar, Sruthi, & Nedungadi, 2015) – where the dialogue classification approach considered features from both audio recordings and lexical analysis of dialogue transcripts – specifically pointed out that "if we classify speech solely on its textual component, we will not obtain a clear picture of the emotional content." Of the previously mentioned text-based automatic animation pipelines, none of them attempted identification or portrayal of emotion. This suggests there is still work to be done in this area.

Another significant issue with using these systems in a generative context is they only consider a small number of emotions for classification. In (Callejas & Lopez-Cozar, 2008), annotators labeled only angry, bored, doubtful, and neutral emotions. The dataset used in (Davletcharova, Sugathan, Abraham, & James, 2015) included three categories of emotion: neutral, angry, and joyful. In (Bhaskar, Sruthi, & Nedungadi, 2015), only the six basic emotions established by Ekman (Ekman & Friesen, 1969) were considered: happy, sad, fear, disgust, surprise and anger.

Such small sets of emotions might be acceptable for merely identifying reactions from human conversational partners. Perhaps there are only a few specific emotions that the robot is prepared to respond to, or else the robot's current context means that only a limited set of emotions are expected. Yet for generation of facial animations it would be preferable to display a wide variety of expressions, both to reduce the repetitiveness of the interaction and to showcase the full expressiveness of the robot platform. Even if there are many expressive

animations that could be selected for a particular emotion, constraining the robot to only three to six emotions limits the nuance able to be conveyed.

3. Related Work

Existing systems for streamlining the animation process can be divided into three categories: rule based pipelines using lexical analysis, statistics based pipelines which 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 (Cassell, Vilhjalmsson, & Bickmore, 2001) generates XML style tags which mark the clause boundaries, theme and rheme, word newness, and pairs of antonyms. These tags are used to suggest nonverbal behaviors which include hand and arm motions, gaze behaviors, and eyebrow movements. Beat gestures are suggested for new words occurring in the rheme. Iconic gestures are inserted when an action verb in the sentence rheme matches a keyword for an animation in the action knowledge base, such as an animation which mimes typing on a keyboard corresponding to the word type in the phrase "You just have to type in some text." Contrast gestures mark the distinction between pairs of antonyms, such as the words *good* and *bad* in "Are you a good witch or a bad witch." Robot gaze behaviors are based on general turn-taking patterns, such as glancing away at the start of a turn. Finally, a conflict resolution filters 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 (Smid, Pandzic, & Radman, 2004) 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.

The text-to-gesture system described in (Kim, Ha, Bien, & Park, 2012) was an extension of a previous work (Kim, Lee, Kim, Park, & Bien, 2007) where 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 (Kopp, et al., 2006) is an XML style language which 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 (including pointing, reaching, and signaling), 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.

4. 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. However, since the 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 assigned animations.

While there are many possible pieces of information that annotators could conceivably mark, in this work we limit the scope to *emphasis location* and *dominant emotion*. The envisioned implementation uses an XML tagging format, an example of which is shown below. An XML format has several advantages. It is easily extensible, permitting new labels to be introduced without invalidating already labeled utterances. It also could potentially be combined with existing or future automated pipelines, which could then handle the other components of behavior generation. In particular, the Speech Synthesis Markup Language also uses XML tagging and already provides a format for tagging emphasized words, so that text to speech programs can enact verbal accents. Since verbal and gestural emphasis are

often co-occurring (Graf, Cosatto, Strom, & Huan, 2002) these tags could potentially be combined.

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. The human face in particular is complex, so it makes sense that few robots are designed to enact all 46 identified facial action units (Ekman & Freisen, 1978). When text is supplemented using systems like the Behavior Markup Language to specify movement of certain degrees of freedom, the implementation is constrained to platforms which are capable of those specific motions. Choosing to label higher level concepts means that any robot, humaniod or not, could be programed to take advangage 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.

One of the main goals for this approach is for it to accommodate labeling by people who do not have a 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 were asked to read aloud before answering questions about the emotion and emphasis of a particular dialogue line. Workers from Mechanical Turk must be at least 18 years old or older and were required to 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.

Turkers were asked to read each 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 in the given context. Because of the correlation between verbal and gestural emphasis, 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. This is important because the subtler forms of gestural emphasis, such as blinks and eyebrow movements, are less likely to be consciously remembered.

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 Figure 1. Furhat operates by projecting an image of its face onto a semi-transparent plastic mold. This provides freedom of facial movement similar to what is achievable with a virtual character, while also providing physical depth and embodiment.

Figure 1: Furhat Robot with projected face

In phase two of the experiment, a subset of these animated expressions were viewed and rated by a separate set of Mechanical Turk workers. Videos of the robot performing an utterance were made which showed either an emotive expression *or* an emphasis gesture. This was done so that the effectiveness of the emotion and emphasis labels could be evaluated independently. Each participant viewed two videos of the robot performing the same utterance and was asked to compare the two 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 which received the majority of selections in phase one, a video with an emphasis gesture at an incorrect location (accenting a word which 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 which matched the consensus from phase one, or a video with an emotive expression which directly opposed the emotion from the phase one consensus. The order in which the two videos were shown was randomized for each participant, so that any biases due to first impressions from viewing the initial video were controlled.

While animated expressions for Furhat were created for all eleven emotions evaluated in the first phase, in the validation phase only two emotive expressions were used: Happy and Unhappy. This was done so that the videos showing the incorrect emotional expression could clearly be directly opposite the correct emotional expression. Also, limiting the number of expressions used for the validation phase reduced the chance that participant ratings on anthropomorphism would be biased by which expressions looked most humanlike, rather than which expressions were appropriate given the context of their corresponding utterance. The happy, neutral and unhappy facial expressions can be seen in Figure 2.

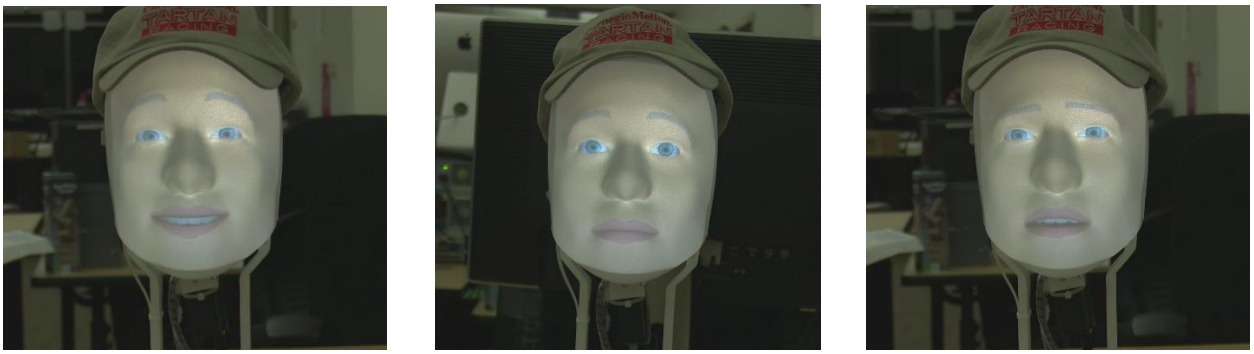


Figure 2: Furhat Expressions – Happy (left), Neutral (center), and Unhappy (right)

Eyebrow motions were chosen as the emphasis gesture because they were easier to precisely synchronize with a specific word compared to a full head nod and they were more noticeable than the more subtle eye blinks. Based on the observations from (Zoric, Smid, & Pandzic, 2007), 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, based on previous works mentioning that perfectly motionless robots can be seen as unnatural (Graf, Cosatto, Strom, & Huan, 2002). These movements included small upward or downward movements of the mouth corners, slight widening or narrowing of the eyes, and very small brow movements. This was done to prevent the control videos from being seen as arbitrarily less appealing due to lack of motion compared to the videos containing emphasis or emotive expressions.

5. Experiment – Phase One

5.1 Phase One Methodology

The first phase of the experiment tested whether untrained workers could come to a consensus about the dominant emotion and location of emphasis for a given utterance. In this study, twenty workers from Amazon's Mechanical Turk crowdsourcing website were each presented with short pieces of conversation. For each short conversation the participants answered questions about the appropriate emphasis and dominant emotion in the last line of dialogue. The participants were shown several turns of dialogue which occurred prior to the line they were to analyze in order to provide situational context. The conversations used in this phase are shown in Table 1, and the instructions and layout for the questions are presented in Figure 3.

Table 1: All Conversation Turns used in Phase One

	Conversations
1	**Alex:** Hey Martin, how's it going? **Martin:** Pretty good. You been busy lately? **Alex:** I'm swamped with work today.
2	**Jill:** Good morning. **Ben:** Good morning. **Jill:** How was your drive? **Ben:** The weather's awful right now.
3	**Steven:** Wow Greg, is that you? It's been a while! **Greg:** It's great to see you!
4	**Zoe:** The day's almost over. **John:** Yeah, thank goodness. **Zoe:** Got any plans after work?
5	**Claire**: Hey Evan, how're you? **Evan:** Okay I guess. **Claire:** Weren't you just assigned to that new project? How's that going? **Evan:** I'm really worried I won't get it all done.
6	**Lisa:** What time is tomorrow's seminar?

	Fred: This schedule says it's at twelve am. **Lisa:** Are you sure? That seems strange.
7	**Lucas:** Did you see my presentation? How do you think I did? **Martin:** You were fantastic.
8	**Peter:** What was your name again? **Susan:** I'm Susan. **Peter:** Nice to meet you Susan.

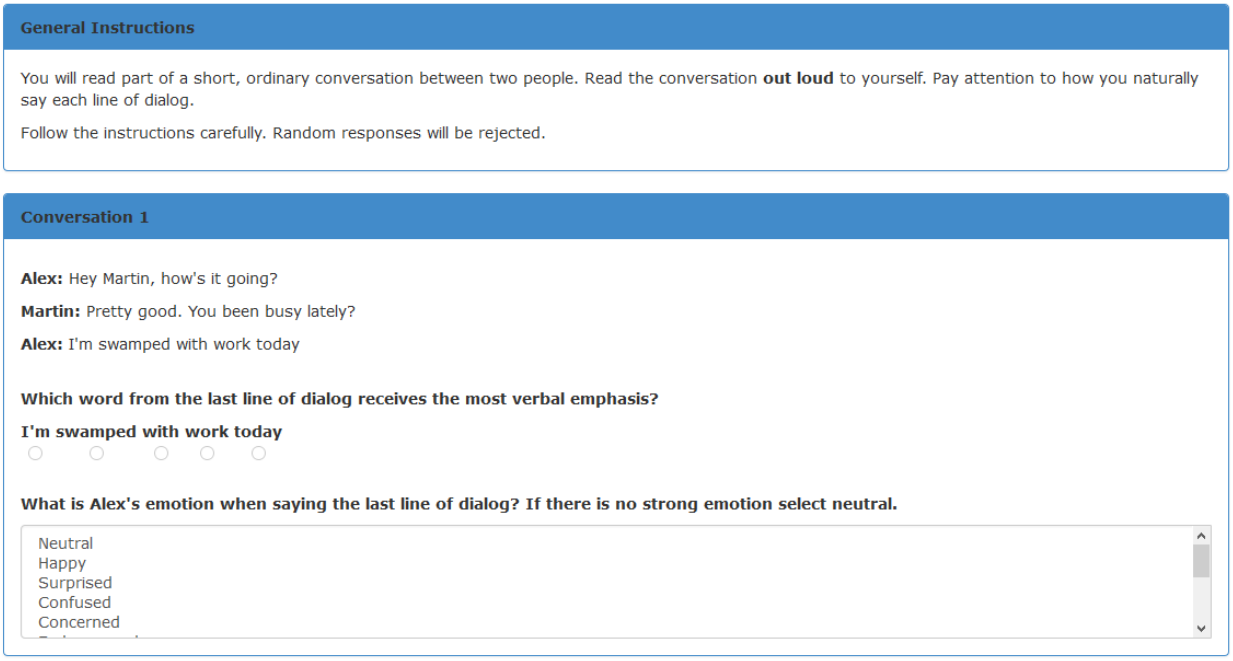


Figure 3: Phase one instructions and conversation layout

Note that in the interface given to the Turkers, they can select only a single word that receives the most verbal emphasis. While people are clearly capable of accenting multiple words within a particular line of dialogue, the question was formatted so that each participant could select only one emphasized word per utterance. This was done to prevent participants from misunderstanding the question and possibly selecting *every* word in a very short utterance, which was not the intent.

Limiting each participant to only one emphasis selection still allowed informative trends to emerge in the data as a whole. If two words in a longer utterance represented equally valid

placements for emphasis, then we would expect the bulk of selections to be split between those two words. If a line of dialogue had no logical place for strong vocal emphasis, then we would expect answers split evenly among the possible words, perhaps with some preference for words that are not prepositions or articles. Therefore the relevant information would still be visible in the data while also allowing the question format to be kept consistent across any length of utterance.

5.2 Phase One Results

The charts detailing the participant responses for the emphasis portion of the study can be seen in Figure 4. Out of the eight utterances presented, four of them – utterances 1, 2, 3, and 7 – contained words which received at least 75% of participant selections for that utterance. Three of these utterances had words which received 90-95% of the selection. This represents strong indication that these words should be emphasized.

In utterance 5, the longest utterance, participant answers clustered around the noun-adjective pair *really worried*. These two words together made up 90% of the selections for this utterance, which shows that there can still be a clustering of responses for a longer sentence. The words *really* and *worried* each received 45%, an even split between them, making it clear that both words receive emphasis. This shows that emphasized phrases were able to be identified even though each participant could select only a single word for each utterance.

In utterance 6, a multi-sentence utterance, the participant selection was split between two words, one from each sentence. This again shows that bimodal distributions will be visible in the data even though each individual may make only one selection.

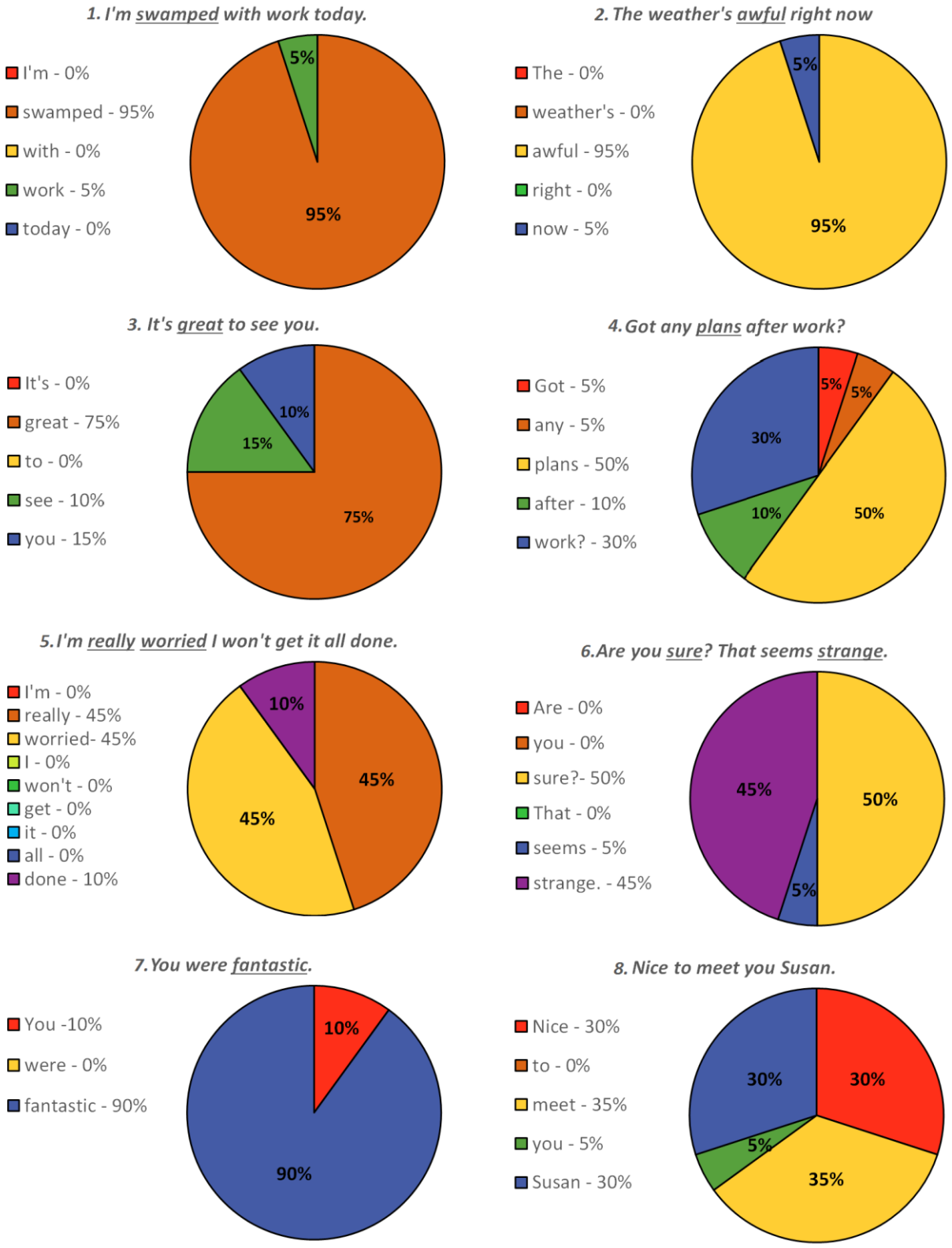


Figure 4: Emphasis percentages for each utterance

The two remaining utterances show that it is possible to determine when a particular utterance does *not* have a strong candidate for word level emphasis. The participant selections for utterance 8, "Nice to meet you Susan," were fairly evenly split between the words nice, meet, and Susan, receiving 30%, 35%, and 30% respectively. Utterance 4, "Got any plans after work?" received participant selections on every possible word. While the word *plans* received 50% of selections and *work* received 30%, these words were separated by the low scoring word after and thus do not form a continuous phrase.

Next consider the emotion data, shown in Figure 5. 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. This shows that the Turkers were being attentive to the task and accurately identifying key words.

For four of the other utterances, participant selections were split between two closely related emotions which together accounted for at least 70% of responses. In utterance 7, "You were fantastic," the split was 65% happy and 30% excited. In utterance 2, "The weather's awful right now," the distribution was 40% sad and 30% concerned. For utterance 3, "It's great to see you," 60% of selections were happy and 20% were excited. Utterance 6, "Are you sure? That seems strange," was 50% confused and 40% skeptical. In each of these cases 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.

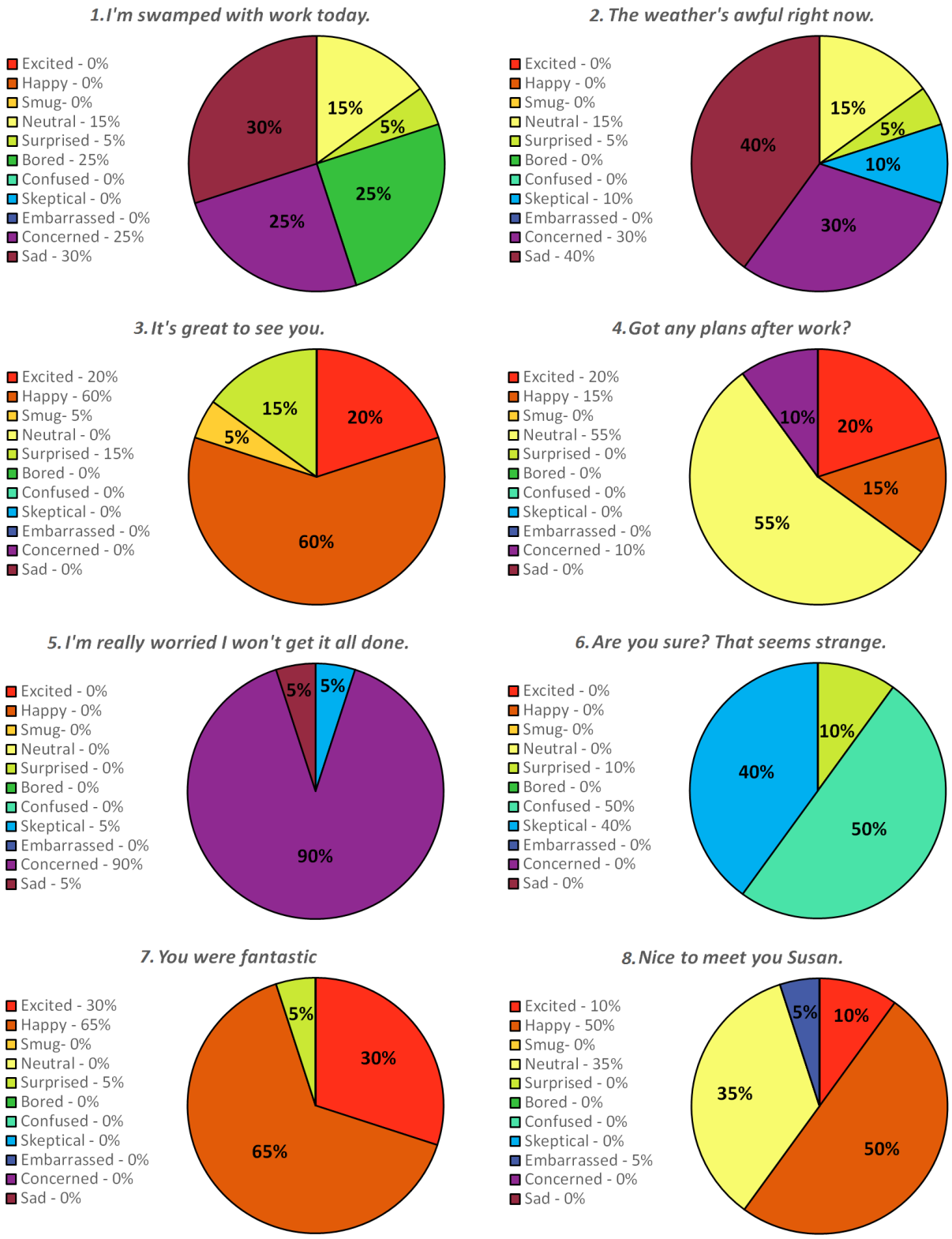


Figure 5: Emotion percentages for each utterance

Of the remaining utterances, “Got any plans after work?” 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. “Nice to meet you Susan” had 50% happy and 10% excited, and “I’m swamped with work today” was 30% sad and 25% concerned. While this gives some suggestion of possible emotions, it is not as strong of a consensus by comparison.

6. Experiment - Phase Two

6.1 Phase Two Methodology

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. These chosen utterances are shown in Table 2 and Table 3. The ones selected for emphasis included the four utterances which received 75% or more of their selections on a single word, and one which contained an emphasized noun-adjective phrase, in which the two adjacent words received a combined 90% of selections. The words chosen as incorrect placements of emphasis received 10% or less of the participant selections and were not adjacent to the word which was selected for emphasis. These selections for incorrect emphasis are also shown in Table 2.

Table 2: Emphasis locations for use in the correct and incorrect emphasis videos

N	Utterance	Correct Emphasis	Incorrect Emphasis
1	I'm swamped with work today.	swamped – 95%	today – 0%
2	The weather's awful right now.	awful – 95%	now – 5%
3	It's great to see you.	great – 75%	see – 10%
5	I'm really worried I won't get it all done.	really worried – 45%, 45%	get – 0%
7	You were fantastic	fantastic – 90%	you – 10%

Table 3: Emotions used in matched and mismatched emotion videos

N	Utterance	Correct Emotion	Incorrect Emotion
2	The weather's awful right now.	Unhappy – total 70%	Happy
3	It's great to see you.	Happy – total 80%	Unhappy
5	I'm really worried I won't get it all done.	Unhappy – 90%	Happy
6	Are you sure? That seems strange.	Unhappy – total 90%	Happy
7	You were fantastic	Happy – total 95%	Unhappy

This data from phase one was then used to animate the Furhat robot. A python script was created that read in the Mechanical Turk data, used the participant responses to select animations, and output tagged utterances which were performed by the Furhat robot. The animation selection process was as follows. If the most commonly chosen emotion received at least 70% of participant selections, add the expression corresponding to that emotion. Otherwise, if the two most commonly chosen emotions have a difference in valence of less than 0.3, and if the two emotions together received at least 70% of the participant selection, add the expression corresponding to the most chosen emotion. Otherwise use a neutral expression. Similarly, if the most commonly chosen emphasis placement received at least 75% of participant selections, add an emphasis gesture at that location. Otherwise, if the two most commonly chosen locations are adjacent and form a noun/adjective or verb/adverb pair, and if the two words together received at least 75% of the participant selection, add an

emphasis gesture at that location. Otherwise do not add emphasis. Each emotional expression was loosely assigned a numerical valence value in order to be able to roughly compare the distance between emotions. These values can be seen in Figure 6, and were loosely chosen based on the relative valences of similar emotions found in Russell's circumplex model of affect (Russell, 1980).

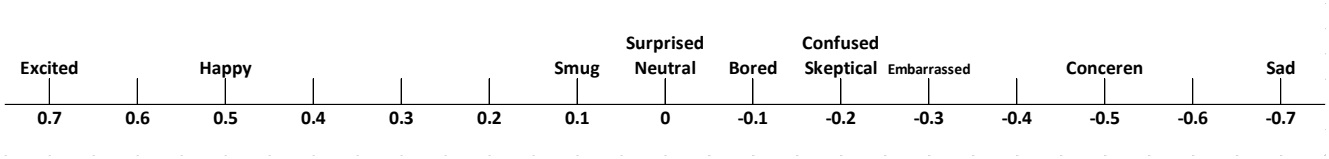


Figure 6: Relative Valence Values for Given Emotions

Videos of the robot's performance were shown to a new group of Mechanical Turk workers, which they rated on a several metrics. First, participants watched the two videos presenting different performances of the same utterance. Then, they compared the videos using the format shown in Figure 7, which allowed them to select 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 (Bartneck, Croft, Kulic, & Zoghbi, 2009). The believable and appropriate metrics were added in order to distinguish between cases where the expression appeared realistic in isolation but did not match with the context or dialogue.

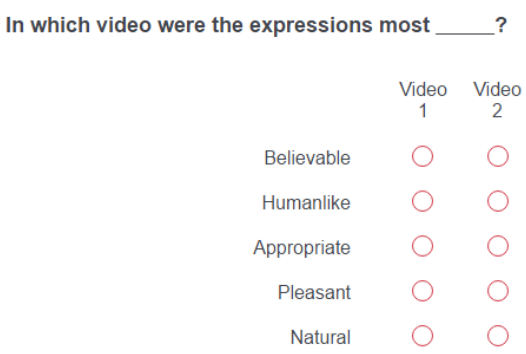


Figure 7: Rating metrics for video comparison

Next participants re-watched each of their videos in isolation and rated them on a subset of the scales from the Godspeed Questionnaire, which can be seen in Figure 8. For each metric, the participant rated the robot's performance on a scale of 1 to 5, with one and five representing the opposite ends of a defined dichotomy. The humanlike/machinelike, fake/natural, and artificial/lifelike dichotomies were from the Godspeed series measuring anthropomorphism. The stagnant/lively, mechanical/organic, and artificial/lifelike choices were from the series measuring animacy. The likable/unlikable and unpleasant/pleasant measures were adapted from the series for determining likability, and unintelligent/intelligent was from the series measuring perceived intelligence.

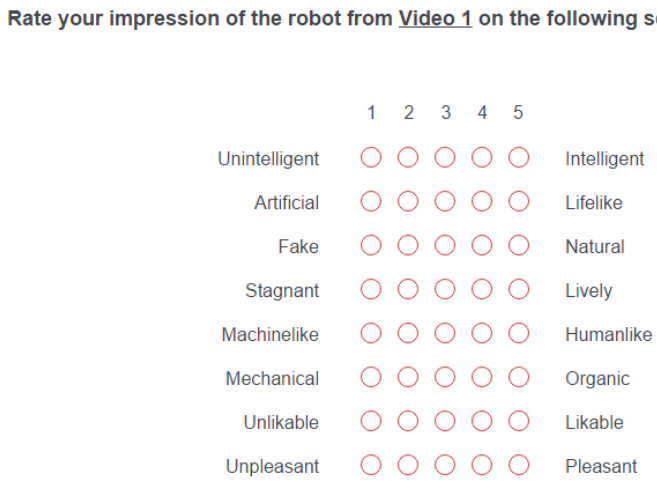
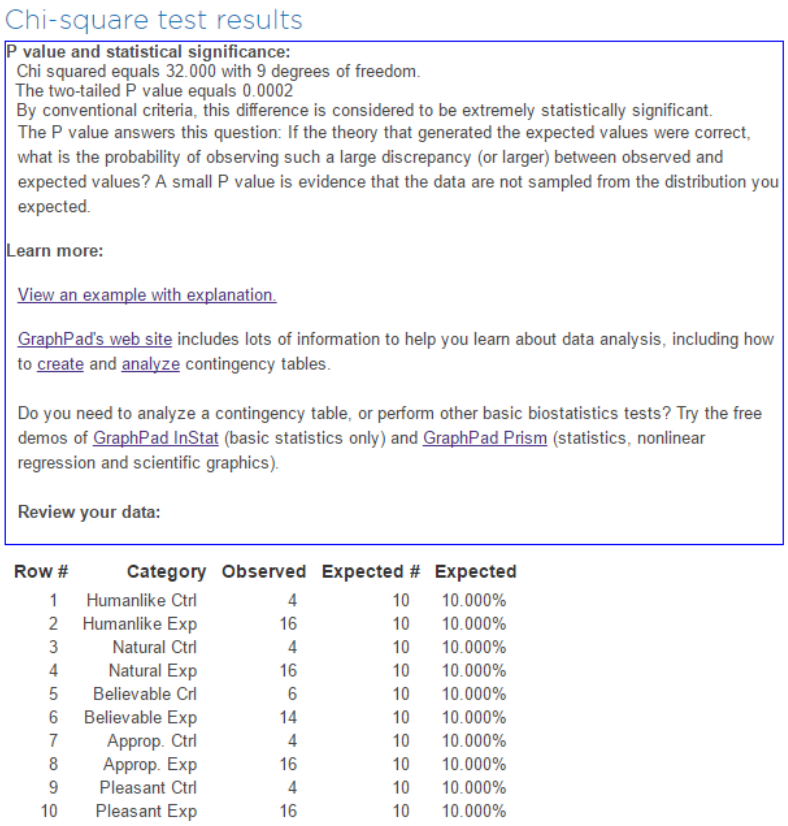


Figure 8: Rating metrics for individual videos

6.2 Phase Two Results

Tables 4 through 7 show the results of the direct video comparison survey questions. 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. Chi-square tests were performed using the GraphPad online calculation software; an example screenshot of feedback from the calculator is shown in Figure 9.



Chi-square test results

P value and statistical significance:
Chi squared equals 32.000 with 9 degrees of freedom.
The two-tailed P value equals 0.0002
By conventional criteria, this difference is considered to be extremely statistically significant.
The P value answers this question: If the theory that generated the expected values were correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values? A small P value is evidence that the data are not sampled from the distribution you expected.

Learn more:

View an example with explanation.

GraphPad's web site includes lots of information to help you learn about data analysis, including how to create and analyze contingency tables.

Do you need to analyze a contingency table, or perform other basic biostatistics tests? Try the free demos of GraphPad InStat (basic statistics only) and GraphPad Prism (statistics, nonlinear regression and scientific graphics).

Review your data:

Row #	Category	Observed	Expected #	Expected
1	Humanlike Ctrl	4	10	10.000%
2	Humanlike Exp	16	10	10.000%
3	Natural Ctrl	4	10	10.000%
4	Natural Exp	16	10	10.000%
5	Believable Ctrl	6	10	10.000%
6	Believable Exp	14	10	10.000%
7	Approp. Ctrl	4	10	10.000%
8	Approp. Exp	16	10	10.000%
9	Pleasant Ctrl	4	10	10.000%
10	Pleasant Exp	16	10	10.000%

Figure 9: GraphPad Chi-Square Calculations for Utterance 3 - Matched Emotion

Table 4 shows the results when utterances were performed with emotions that matched phase one consensus compared to the control video. The entries show the percentage of participants who selected the animated performance of an utterance for a particular category, with the remainder of participants selecting the control. Utterances with significant chi-squared results ($p < 0.05$) are marked with asterisks. Table 5 uses a similar format to show the surveys where the displayed emotion did not match the consensus from phase one.

The robot performances which 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 which are viewed as appropriate.

Table 4: 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 5: 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*

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.

The one video with a mismatched emotional expression which did perform significantly better than the control was utterance 7, “You were fantastic.” One possible explanation is that the robot’s delivery of the line was interpreted as sarcastic by the Mechanical Turk workers.

Sarcasm is a mismatch between the literal meaning of an utterance and its current context, and can be conveyed through tone of voice, body language, and facial expressions. The text-to-speech voice used by the robot is relatively emotionless, and was kept constant across all videos. Therefore phase two participants were presented with a *happy* sentence portrayed with an *unhappy* expression and a *monotone* voice. This could have lead phase two participants to view the delivery as sarcastic, even though phase one participants who were not influenced by an unexpressive text-to-speech voice treated the utterance as sincere.

Overall, the videos showing expressions which matched the phase one responses were rated as significantly more humanlike than the control videos. This supports the idea that emotion labels from untrained workers have the potential to improve viewers' perception of a social robot. Since only one of the five videos showing a mismatched expression was rated significantly more positively than the control, it would not be beneficial to arbitrarily add emotional expressions without consulting labeling. Merely displaying an expression with no regard for context would not create the same improvements as seen for the matched expressions.

The data from the emphasis surveys is less clear. Table 6 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 7. 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 6: 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 7: 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

After completing the direct comparison questions, participants rated each video on several metrics from the Godspeed Questionnaire, as described in Section 6.1. Ratings are on a scale of one to five, where Graphs of participants' average ratings for each emotion case can be seen in Figures 10 - 19. Participant's ratings of the control video were compared to the same group's ratings of each expressive video, using paired sample t tests with Bonferroni corrections in order to determine significance.

Of the ratings given to the videos showing expressions which matched phase one responses, four of the five utterance achieved statistical significance in three or more of the

eight metrics used. Notably, utterances 3 and 6 achieved significance in seven of the eight metrics and utterance 7 reached significance in 5 of the metrics used. Comparably, only one of the five videos showing incorrect expressions reached significance in three or more metrics. Utterance 5 did not reach significance on any of the metrics when shown with the incorrect expression. Furthermore, the metric which most commonly achieved significance in the incorrect expression group was the measurement of *liveliness*. Since the control videos only showed very small facial movements when compared to either the correct or incorrect expression cases, it's possible the significant differences in ratings of liveliness when compared to the control videos were because of the greater amount of movement, regardless of whether the movement was contextually appropriate.

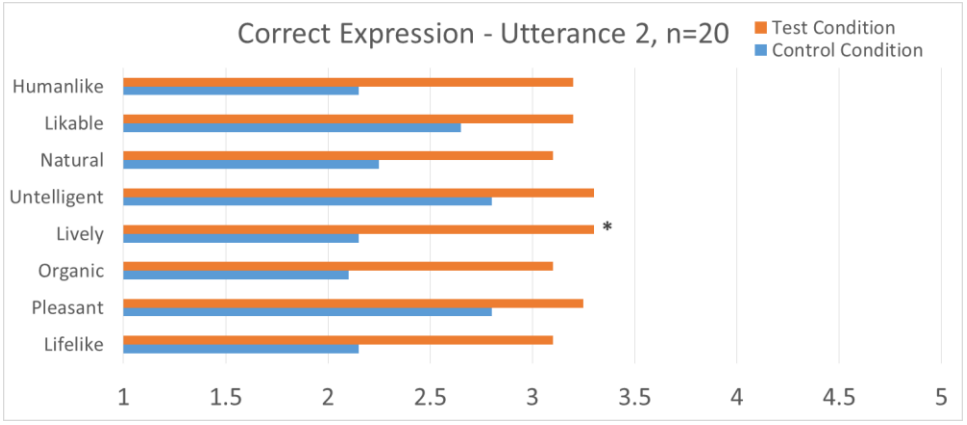


Figure 10: Correct Expression vs Control Video for Utterance 2

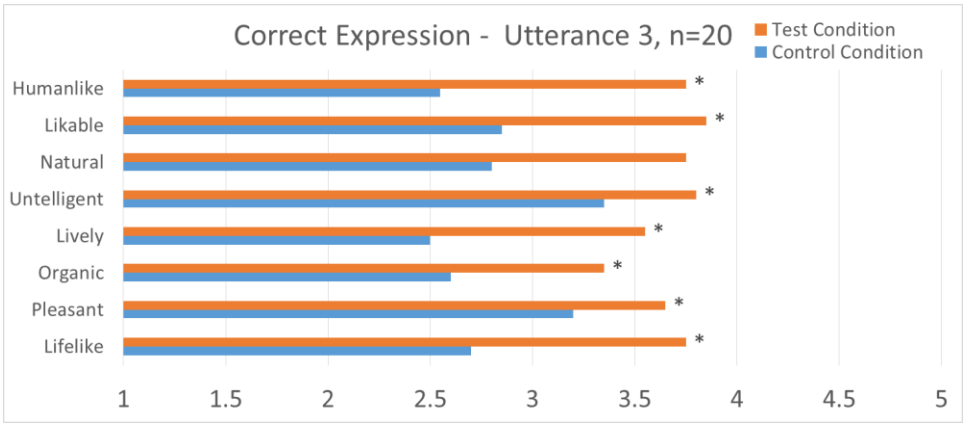


Figure 11: Correct Expression vs Control Video for Utterance 3

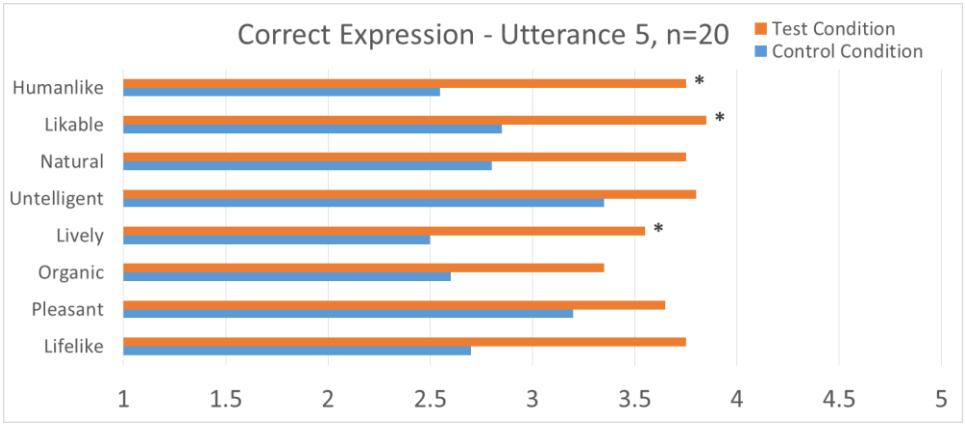


Figure 12: Correct Expression vs Control Video for Utterance 5

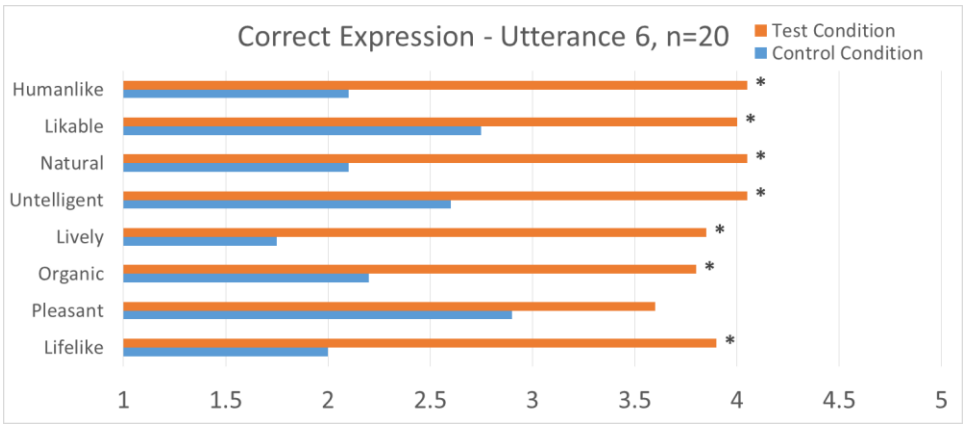


Figure 13: Correct Expression vs Control Video for Utterance 6

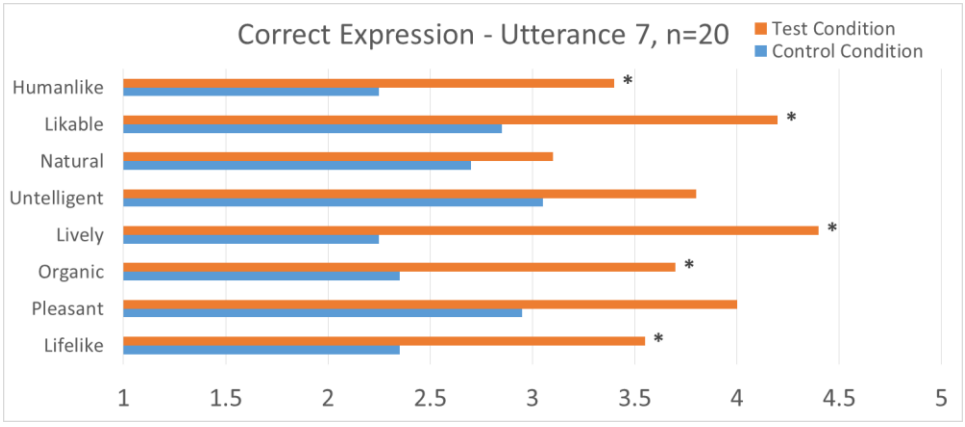


Figure 14: Correct Expression vs Control Video for Utterance 7

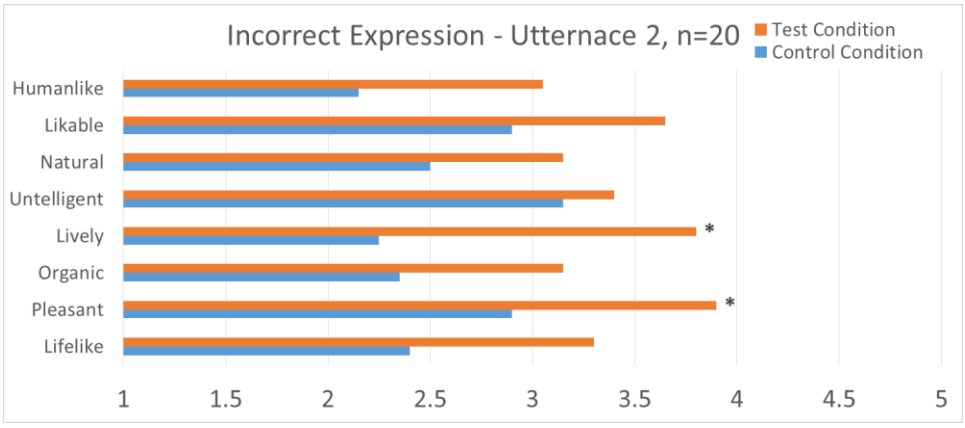


Figure 15: Incorrect Expression vs Control Video for Utterance 2

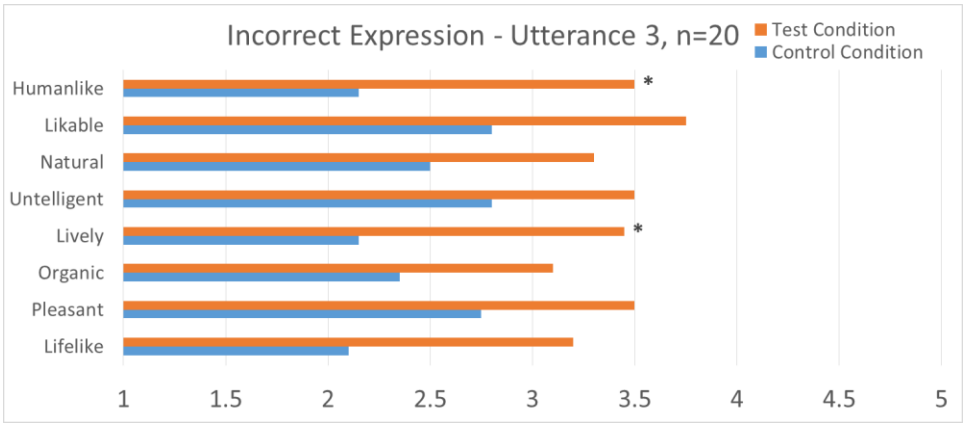


Figure 16: Incorrect Expression vs Control Video for Utterance 3

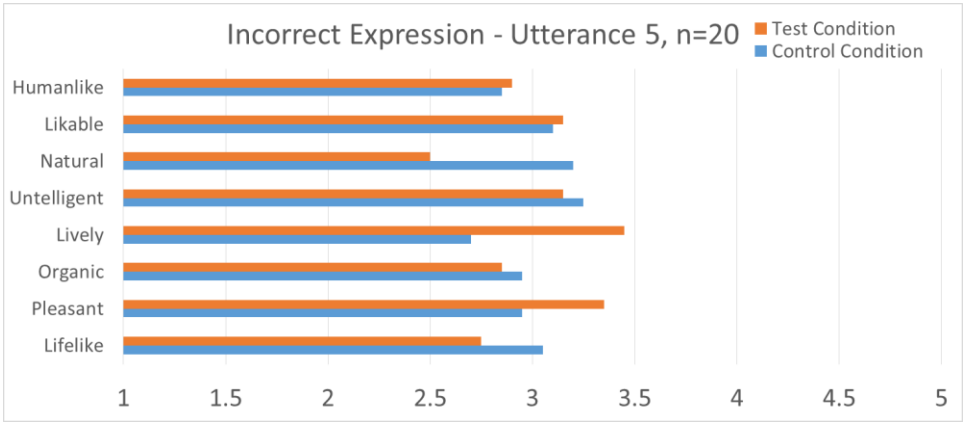


Figure 17: Incorrect Expression vs Control Video for Utterance 5

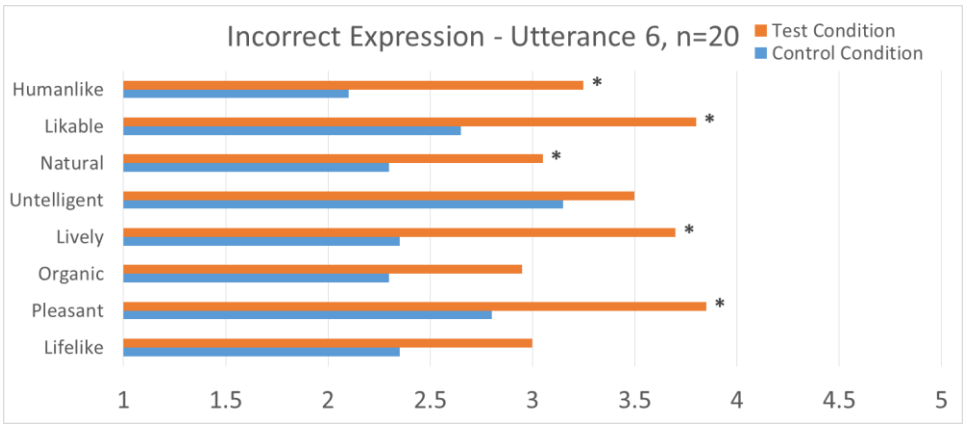


Figure 18: Incorrect Expression vs Control Video for Utterance 6

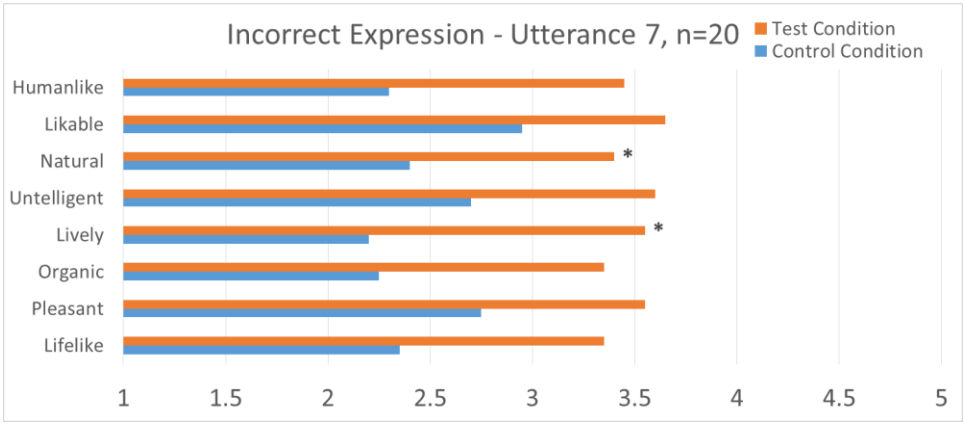


Figure 19: Incorrect Expression vs Control Video for Utterance 7

7. Discussion and Future Work

The results from phase one showed that untrained workers are able to come to a consensus on labels for emotion and emphasis. Cases with a clear location for emphasis and cases where no emphasis was needed were distinguishable based on the distribution of participant selections. Workers' responses for emotion labeling showed consideration for affect-laden keywords in the utterance. When participant's emotion selections were split between emotions, the emotions with the largest number of selections were often closely

related in terms of valence, showing that the participants had similar interpretations of the underlying sentiment of the utterance despite slight differences in final selection.

In phase two, videos displaying expressions that matched the phase one responses were rated as significantly more humanlike than controls. This supports the concept that basing emotive animations on emotion labels provided by untrained workers has a positive impact on perception of the robot. Furthermore, given that only one of the five utterances displayed with mismatched expressions was rated significantly more positively than the control, it would not be advisable to displaying an arbitrary expression.

Even though eyebrow movements do occur in conjunction with emphasized words, these movements appear to be too subtle to make a difference in people's ratings of the robot's performance, since overall the emphasis placements chosen in phase one did not perform better than alternate emphasis placements. Further exploration of this area would likely involve using a different robot platform capable of making obvious hand and arm based gestures, which would be more easily visible to participants. Clearer emphasis gestures, combined with verbal emphasis through a different text to speech engine, would be more likely to be noticed and thus could potentially show a more significant effect on the participants' ratings.

Other areas of future work include testing the emotion labeling using the full range of robot emotions – rather than just happy and unhappy expressions – and testing across multiple robot platforms. Also, more exploration of the effects of emphasis placement at different words in the same sentence, and whether different emphasis affects viewer's interpretation of the underlying implications of the utterance, would help clarify the role emphasis plays in our understanding of information.

8. Conclusions

Nonverbal behaviors play an important role in making interaction with conversational social robots believable and engaging. This work proposed an approach for reducing the time spent animating each utterance of a social robot while still making use of human knowledge of social contexts. Through this study, we found that 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.

9. 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.

10. References

Albrecht, I., Haber, J., & Seidel, H.-P. (2002). Automatic Generation of Non-Verbal Facial Expressions from Speech. In *Advances in Modelling, Animation and Rendering* (pp. pp 283-293). Springer London.

Andrist, S., Tan, X. Z., Cleicher, M., & Mutlu, B. (2014). Conversational Gaze Aversion for Humanlike Robots. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 25-32.

Bartneck, C., Croft, E., Kulic, D., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71-81.

Bhaskar, J., Sruthi, K., & Nedungadi, P. (2015). Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining. *Procedia Computer Science, 46*, 635-643.

Callejas, Z., & Lopez-Cozar, R. (2008, May). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication, 50*(5), 416-433.

Cassell, J., Vilhjalmsson, H., & Bickmore, T. (2001). BEAT: The Behavior Expression Animation Toolkit. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 477-486.

Cowie, R., Douglas-Cowie, E., Karpouzis, K., Caridakis, G., Wallace, M., & Kollias, S. (2008). Recognition of Emotional States in Natural Human-Computer Interaction. In D. Tzovaras (Ed.), *Multimodal User Interfaces* (pp. 119-153).

Davletcharova, A., Sugathan, S., Abraham, B., & James, A. P. (2015). Detection and Analysis of Emotion From Speech Signals. *Procedia Computer Science, 58*, 91-96.

Ekman, P., & Freisen, W. (1978). Facial Action Coding System - A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press*, 1-2.

Ekman, P., & Friesen, W. V. (1969). Pan-Cultural Elements in Facial Display of Emotions. *Science, 164*, 86-88.

Ekman, P., & Friesen, W. V. (1972). Hand Movements. *Journal of Communication, 22*, 353-374.

Graf, H. P., Cosatto, E., Strom, V., & Huan, F. J. (2002). Visual Prosody: Facial Movements Accompanying Speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 396-401.

Isha, C. T., Liu, C., Ishiguro, H., & Hagita, N. (2010). Head Motion during Dialogue Speech and Nod Timing Control in Humanoid Robots. *5th ACM/IEEE International Conference on Human-Robot Interaction*, 293-300.

Kiesler, S. C.-R. (2005). Fostering Common Ground in Human-Robot Interaction. *IEEE International Workshop on Robot and Human Interactive Communication*, 729-734.

Kim, H. H., Lee, H. E., Kim, Y. h., Park, K. H., & Bien, Z. Z. (2007). Automatic Generation of Conversational Robot Gestures for Human-friendly Steward Robot. *The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 1155-1160.

Kim, H.-H., Ha, Y.-S., Bien, Z., & Park, K.-H. (2012). Gesture encoding and reproduction for humanrobot interaction in text-to-gesture systems. *Industrial Robot: An International Journa, 39*(6), 551-563.

Kopp, S., Krenn, B., Marsella, S., Marshal, A. N., Pelachaud, C., Pirker, H., . . . Vilhjalmsson, H. (2006). Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. *Proceedings of the 6th international conference on Intelligent Virtual Agents*, 205-217.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113.

Perikos, I., & Jatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence, 51*, 191-201.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology, 30*(6), 1161-1178.

Smid, K., Pandzic, I. S., & Radman, V. (2004). Autonomous Speaker Agent. *Proceedings of Computer Animation and Social Agents Conference*.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gestrure and speech in interaction: an overview. *Speech Communication, 57*, 209-232.

Zoric, G., Smid, K., & Pandzic, I. S. (2007). Facial Gestures: Taxonomy and Application of Non-Verbal, Non-Emotional Facial Displays for Embodied Conversational Agents. In T. Nishida (Ed.), *Conversational Informatics: An Engineering Approach* (pp. 161-182). John Wiley & Sons, Ltd.