Dialog with Robots: Papers from the AAAI Fall Symposium (FS-10-05)

Meta-Analysis of User Age and Service Robot Configuration Effects on Human-Robot Interaction in a Healthcare Application

Manida Swangnetr, Biwen Zhu, David Kaber, Kinley Taylor

Edward P. Fitts Department of Industrial & Systems Engineering, North Carolina State University
Raleigh, NC 27695-7906
{mswangn, bzhu, dbkaber, kbtaylor}@ncsu.edu

Abstract

Future service robots applications in healthcare may require systems to be adaptable in terms of verbal and non-verbal behaviors to ensure patient perceptions of quality healthcare. Adaptation of robot behaviors should account for patient emotional states. Related to this, there is a need for a reliable method by which to classify patient emotions in real-time during patient-robot interaction (PRI). Accurate emotion classification could facilitate appropriate robot adaptation and effective healthcare operations (e.g., medicine delivery). We conducted and compared two simulated robot medicine delivery experiments with different participant age groups and robot configurations. A meta-analysis of the data from these experiments was to identify a robust approach for emotional state classification across age groups and robot configurations. Results revealed age differences as well as multiple robot humanoid feature manipulations to cause inaccuracy in emotion classification using statistical and machine learning methods. Younger adults tend to have higher emotional variability than elderly. Combinations of robot features were also found to induce emotional uncertainty and extreme responses. These findings were largely reflected in terms of physiological responses rather than subjective reports of emotions.

Introduction

Service robots have been developed to assist nurses in routine patient services including healthcare related materials delivery. Although delivery robots have been implemented in many hospitals and nursing homes, current commercially available units are not capable of delivering medicines directly to patients. Existing robot designs do not support verbal and non-verbal interaction with patients. Prior research identified three key aspects of humanoid robot design (i.e., humanoid features) facilitating social interaction, including: face and/or head features; and voice and interactivity capabilities (Zhang et al. 2008). Empirical studies have been conducted to examine the effect of these features on user perceptions of robot humanness and emotional responses in simulated medicine delivery tasks. The overarching goal of these studies is to provide a basis for robot expression adaptation according to current user emotional states, especially in hospital environments. Physiological signals from patients can be monitored in real-time in a hospital in order to determine patient status. Other research has been conducted to define methods by which to extract physiological responses and facilitate emotional states classification in real-time during PRI. Such systems would ensure that robots not only successfully perform tasks, but also provide positive emotional experiences for patients.

Empirical Work

In our first study (Zhang et al. 2010), we recruited 24 senior subjects from two local retirement centers (mean age=80.5, SD=8.8) to interact with a mobile robot prototype (PeopleBot) and receive medication. Repeated trials were used to present two general types of robot configurations (abstract vs. humanlike) by manipulating face, voice and interactivity features (see Figure 1; Table 1). All trials were completed at the senior centers. The facial feature was either an abstract face with two camera "eyes", or a humanoid face with a smooth mask. The voice was either synthesized or digitized audio messages. Interactivity referred to user actions required by either reading verbal instructions from a tablet PC screen onboard the robot (i.e., message only) or pressing a button on the touchscreen to confirm delivery. The settings of each feature were presented in the absence of any other features. There was also a control condition in which none of the humanoid features were presented.

Subject emotional responses were evaluated using the self-assessment manikin (SAM; [Bradley and Lang], 1994) questionnaire or through physiological measures, including heart rate (HR) and galvanic skin response (GSR). The SAM questionnaire evaluated subject post-test valence (how happy they were) and arousal (how excited they felt) using rating scales with graphic characters as anchors. After each test trial, participants were instructed to recall their feelings when the robot opened its gripper (a medicine bag was attached to the gripper) and to select the character on a SAM scale best representing their emotions.

HR and GSR were collected in real-time from sensors attached to a subject's chest and two non-dominant fingers.

Results revealed all interface manipulations lead to stronger positive emotional responses, as compared with the control condition. However, facial appearance appeared to have the greatest effect on the perception of robot humanness and was recommended for use in future service robot design for medicine delivery applications.

Our follow-on study (Swangnetr et al. 2010) investigated the effects of combinations of robot humanoid features on younger population emotional responses. Thirty-two subjects (mean age=23.16, SD=3.12) were recruited to interact with the medicine delivery robot in a lab environment. Similar to the first study, there were two general types of robot configuration (machinelike and humanlike) for each interface feature (i.e., face, voice and interactivity). In designing the specific robot conditions, we considered the face feature as the most important for perceptions of humanness. An abstract face was always coupled with a synthesized voice to represent robot conditions with lower "humanness"; whereas, a human face was always coupled with a digitized voice to represent higher humanness. Such manipulations were used to avoid inconsistency in robot design relative to user expectations. Integrating the interactivity feature, seven conditions were presented in this second study representing increasing levels of humanness (see Table 1). Physiological responses (HR, GSR) and SAM data were also collected in real-time. Subject instructions for both studies were comparable.

Figure 1. Basic PeopleBot platform (left). Additional humanoid features (right), clockwise from top left: abstract face; humanoid face; confirmation; and message only.

Results showed that a robot with higher degrees of humanness lead to higher arousal and valence ratings and HR responses. We also found that additional humanoid features lead to higher GSR ratings, but the trend was not strictly linear.

Emotional State Classification Algorithm

A three-stage algorithm was developed for real-time emotional state classification based on the physiological and subjective rating data collected in the two experimental studies. Stages included: (1) physiological feature extraction and noise elimination; (2) statistical-based feature selection; and (3) machine learning modeling of emotional states. Analysis of physiological measures for emotion identification is generally conducted on an event basis, using brief time windows of data (4 seconds) after an event (Ekman 1984). In both of our previous studies, the robot opening its gripper to release the medicine to participants was selected as the event for analysis. This was a key event in the robot service as it requires participants to perceive the robot interface cues and understand how to interact with the robot. It also provided the greatest degree of emotion discrimination among robot configurations (Swangnetr et al. 2009).

Statistical features of HR, including mean HR and standard deviation of HR (SDHR), were examined. However, GSR signals were non-stationary and noisy and were further processed using wavelet analysis. Daubechies 3 wavelet (db3) was selected as a mother wavelet due to the orthogonality property and shape resembling the GSR signal. Wavelet technology was also used for signal noise elimination. Based on decomposition of the GSR signal using the db3 wavelet, the coefficients representing the high frequencies (>0.5 Hz) of the signal (noise) were set to zero. A wavelet soft threshold shrinkage algorithm attenuated noise, which overlapped GSR frequency. Since the distribution of wavelet detail coefficients is better represented by a zero-mean Laplace distribution (Lam 2004), we proposed the threshold of wavelet shrinkage to be $4.18\sigma_L$. This threshold provides 99.73% confidence that noise will be eliminated from the signal. The σ_L of noise was estimated based on data collected individually during a rest period. As a result, a set of 24 wavelet coefficients was available to represent an entire 4-s GSR signal and to reveal time, amplitude and frequency features.

Table 1. Robot configurations investigated in two prior studies.

Condition	1st study			2nd study		
	Face	Voice	Interactivity	Face	Voice	Interactivity
1	No	No	No	No	No	No
2	Abstract	No	No	Abstract	Synthesized	No
3	Human	No	No	Abstract	Synthesized	Visual message
4	No	Synthesized	No	Abstract	Synthesized	Confirmation
5	No	Digitized	No	Human	Digitized	No
6	No	No	Visual message	Human	Digitized	Visual message
7	No	No	Confirmation	Human	Digitized	Confirmation

Note: The ascending number of robot condition was not considered to represent an increase in the level of "humanness" in the first study.

A stepwise regression procedure was used to reduce classification model complexity by selecting statistical and wavelet features with significant relationships with each class of subjective emotional state. Significant physiological features were then used as inputs in machine learning models (artificial neural networks) for arousal and valence state classification. The arousal and valence scores from the SAM ratings were converted to z-scores in order to address individual differences. The normalized ratings were then categorized as representing low, medium or high levels of valence/arousal (lower, mid and upper 33% of the normal distribution) and used as desired outputs. Backpropagation neural network models were constructed with a single hidden layer. The number of hidden layer nodes was optimized to achieve the highest percentage of correct classifications (PCC) in validation (from 20% of samples) for predicting subject emotional responses.

Analysis on the first study revealed overall PCCs in validation of the ANN for predicting arousal and valence to be 82% and 73%, respectively. However, the emotional state classification models for the second study produced PCCs for arousal and valence of 50% and 49%, respectively. The confusion matrix revealed a lack of prediction capability for the medium level of arousal and valence states. The PCCs for medium emotional responses were 22% for both arousal and valence. Consequently, we reconstructed models to predict only low and high level emotional states. Results revealed the overall PCC in validation to increase to 70% for arousal and 63% for valence. The overall PCCs for the best arousal and valence classification networks from both studies are presented in Table 2. All classification models were constructed with 5-7 hidden nodes based on a set of 9-15 physiological inputs.

Table 2. Summary of overall PCCs for arousal and valence classification networks.

Classification Model	Arousal	Valence
1st study with 3 emotion levels	82%	73%
2nd study with 3 emotion levels	50%	49%
2nd study with 2 emotion levels	70%	63%

Motivation and Implications

Reliability in patient emotional state classification is necessary in order to provide a basis for effective adaptation of service robot behaviors and expressions in direct interaction with patients. A challenge in developing human-robot dialog is the design of robots with the capability to adaptively service according to human expectations in specific cultural contexts or task environments (Breazeal 2003). With respect to non-verbal interaction, for example, Dautenhahn et al. (2006) investigated how a robot should approach a human in an object delivery task. Results showed that people preferred robots to approach from either left or right, but not the front, with approaching distances in a range comparable to human-human social distance (0.45-3.6m). Adaptive verbal communication between robots and humans also represents a potential application for a patient emotional state classification algorithm. Zhu and Kaber (2010) found that people expect different linguistic etiquette strategies during interaction with a service robot in a medicine delivery task, depending upon environment and system configuration. If a patient's emotional state can be effectively classified as negative or positive, the robot may be able to use a linguistic strategy that is most effective for the user's condition. Failure in emotion classification may lead to robot behaviors that do not conform with human expectations. This, in turn, could cause lower rates of compliance with robot requests during tasks (Cialdini 2001). This is a critical issue in a healthcare context, as patient omission of medicine dosages can cause fatality (see Barker et al. 2002).

In general, high accuracy of user state classification models is important to ensure successful robot adaptation during HRI. However, results from our previous studies showed a substantial reduction in classification model (neural network) accuracy in the second study for predicting moderate user arousal and valence responses. We speculated the reduced accuracy might have been due to high emotional variability in younger participants and/or emotional uncertainty and intensity induced by the combination of robot humanoid features. Understanding underlying sources of user emotion variability is fundamental in designing adaptive robot dialog systems. The objective of the present study was to conduct a meta-analysis on the results of the two studies described above to define a robust classification algorithm across age groups and robot configurations.

Detailed Hypotheses

It was hypothesized that the age difference between the two participant groups might have led to emotional differences, reflected in both subjective and physiological responses (Hypothesis (H)1). Dissimilarities in the emotional patterns of age groups could cause failures in model classification of moderate emotional states. Beyond this, subjective ratings from younger participants in the second study were expected to have greater variability, as compared with senior ratings in the first study (H1.1). Prior research has found that emotional states in older adults tend to be more controlled and less variable (e.g., Gross et al. 1997; Lawton et al. 1992). This literature suggests older adults regulate emotions differently than younger adults in a way that promotes well-being. In addition, Bradley and Lang (1994) stated that there were differences in perceptions of arousal and valence among different age groups. For example, Backs et al. (2005) found younger adults rated greater arousal and pleasantness than older adults when presented with pleasant-aroused affective pictures. In this research, there was no difference in the way that younger and older participants used the SAM when rating pictorial stimuli.

Physiological responses from younger participants were also expected to have greater variation, as compared with older participations (H1.2). Related to Hypothesis 1.1, regulation and consistency of emotions in elderly adults was expected to lead to low emotional variation in terms of physiological responses. In general, the magnitude of change in physiological measures is smaller in older adults (Levenson et al. 1991). Maximum HR is strongly correlated with age; when age increases, maximum HR decreases (Tanaka, Monahan, and Seals 2001). However, the relationship between GSR and age groups has not been found to be consistent across studies. Drory and Korczyn (1993) reported a significant decrease in GSR amplitude in the elderly. Opposite to this, Baba et al. (1988) observed no age-dependent significant decrease in GSR amplitude.

It was also expected that use of different robot configurations in the two studies could have lead to problems in moderate emotional state classification (H2). ANOVA results on SAM ratings for the various robot configurations in the second study (Swangnetr et al. 2010) revealed particular sets of configurations (C_i) to yield different levels of arousal (referred to as "Arousal-based Groupings") and different levels of valence (referred to as "Valence-based Groupings") (see Table 3). On this basis, the set of robot conditions yielding moderate emotional responses was expected to produce higher variations in terms of subjective ratings and physiological responses, as compared with groups of robot conditions yielding low or high emotional responses (H2.1).

One motivation for this hypothesis was that there were several robot conditions, including multiple design features, which induced moderate emotional responses. Because of the combination of features, participants may have been uncertain about whether they felt aroused or happy. Such uncertainty could have produced high variations in both ratings and physiological responses. It is also possible that participants felt compelled by the rating scales to rate emotions in one direction or another.

Robot configurations in the second study may have generated more extreme emotional responses when compared with robot conditions in the first study (H2.2). In the first study, we examined user reactions to individual design feature manipulations. Low, medium and high emotional responses may have been dependent upon feature type. However, in the second study, conditions might have generated only low and high responses as a result of feature combinations.

Analytical Methodology and Results

Age Effects

In order to test the hypothesis of age differences in emotional regulation and variation, an unequal variance analysis (Bartlett's test) was conducted on the subjective and physiological response data from the two studies. (Only data on the control condition were included in this analysis, as the robot configuration was constant across investigations.) Arousal and valence ratings from the SAM questionnaire were converted to z-scores. A 4-s window of HR and GSR data, recorded after the robot opened the gripper, was used for analysis. The data were extracted and normalized with respect to baseline. Maximum GSR and average HR were calculated for statistical analysis.

Contrary to our hypothesis (H1.1), results from Bartlett's test revealed there to be no difference in the variance in SAM responses across studies (arousal $\chi^2_{(1)}$=0.0071, p=0.93; valence $\chi^2_{(1)}$=0.5905, p=0.44). This suggested the young and old participants were consistent in the emotional ratings.

Opposite to H1.2, the GSR variances for the senior participants were significantly higher than those for the younger participants ($\chi^2_{(1)}$=78.08, p<0.0001). This could have been due to other reasons than age differences (such as subject conditioning; we say more about this in the discussion section). However, we did not consider this result as a basis for rejecting the general hypothesis on age effects (H1). Related to this, the variances in HR in the second study were significantly higher than those in the first study ($\chi^2_{(1)}$=80.55, p<0.0001). Therefore, high emotional variability in terms of HR from younger participants could lead to classification inaccuracy (supporting H1.2).

Emotional Uncertainty

To evaluate the hypothesis on unequal emotion variability induced by robot configurations (H2.1), the seven robot conditions were grouped into three levels, as presented in Table 3. A series of Bartlett tests was conducted to evaluate variance differences in SAM ratings and physiological responses across the levels. Results revealed robot conditions yielding low, medium or high emotional responses produced comparable variations in terms of SAM ratings (see Table 4). This suggested robot stimuli did not generate inconsistency in self-reported emotional ratings. In support of H2.1, significant differences were found for physiological response variances across three levels of stimuli (see Table 4). Specifically, HR induced by low-valence robot conditions and GSR induced by high-arousal and high-valence robot conditions had the largest variances, as compared to other groups. The variability reduced the degree of emotional state discrimination in classification.

Table 3. Sets of robot conditions from the second study induced different degrees of emotion.

	Low	Medium	High
Arousal-based Groupings	C_1	C_2,C_3,C_4,C_5,C_6	C_7
Valence-based Groupings	C_1	C_2,C_3,C_4,C_5	C_6,C_7

Table 4. Bartlett test results on SAM ratings and physiological responses across three levels of robot conditions.

Robot Condition	Arousal	Valence	HR	GSR
Arousal-based Groupings	$\chi^2_{(2)}$=3.36 p=0.19		$\chi^2_{(2)}$=5.03 p=0.08	$\chi^2_{(2)}$=9.2 p=0.01
Valence-based Groupings		$\chi^2_{(2)}$=0.95 p=0.62	$\chi^2_{(2)}$=7.36 p=0.025	$\chi^2_{(2)}$=19.79 p <0.0001

Emotion Intensity

We also hypothesized that the robot configurations in the second study would generate more extreme responses (H2.2). To test this, the ranges of SAM ratings as well as physiological responses were calculated for each subject in both studies. These ranges were compared across studies using standard t-tests because of small sample sizes.

Results revealed that the range of SAM responses in the second study was significantly greater than in the first study (arousal $t(37)=-2.1544$, $p=0.0377$; valence $t(35)=-2.09$, $p=0.0439$). The range of HR responses in the second study was also significantly higher ($t(53)=-6.65$, $p<0.0001$). These results were consistent with our hypothesis that robot configurations in the second study generated more extreme emotions. However, the ranges for GSR were found to be comparable across studies ($t(30)=0.428$, $p=0.6714$).

To further test H2.2, a series of ANOVAs were conducted to determine the effects of robot conditions on the SAM ratings and HR responses in both studies. To make a comparable comparison across studies, the robot conditions in the first study were also grouped into three levels. The control condition was considered as a low emotional stimulus. Robot conditions with less human-like features (#2, 4 and 6 from Table 1) were categorized as medium emotional stimuli. Highly human-like robot conditions (#3, 5 and 7) were grouped into the high emotional stimuli category. Post-hoc analysis using Tukey's test was used to compare groupings of the SAM ratings across studies, while a non-parametric version of the ANOVA (Wilcoxon Ranks Test) was used to compare groupings of physiological responses (see Table 5).

Table 5. Tukey groupings across studies.

Robot condition		Arousal	Valence	HR
First study	Low	A	A	A
	Medium	B	B	A
	High	C	C	B
Second study (Arousal-based grouping)	Low	A	A	A
	Medium	B	B	B
	High	C	C	B
Second study (Valence-based grouping)	Low	A	A	A
	Medium	B	B	B
	High	C	C	C

Results revealed three distinct Tukey groups for the SAM ratings across the three levels of robot stimuli in both studies. Although previous analysis showed the range of SAM ratings for the second study to be significantly greater, it did not provide evidence of which robot configurations generated more extreme emotions. Results on the HR data revealed low and moderate emotional states in the first study to be comparable in physiological response. However, in the second study, HR responses for low emotional states were significantly different from moderate and high emotional states. Since the low level of emotion was induced by the same robot configuration (control condition) for both studies, it served as a common base for comparison. It can be concluded that any combination of humanoid robot features (i.e., Conditions 2 to 7 from the second study) yields more extreme HR responses, compared to the control condition, as well as individual machine-like features (i.e., Conditions 2, 4 and 6 from the first study).

Discussion

On the basis of the meta-analysis, age appears to be a significant factor influencing accuracy in moderate emotional state classification in PRI. In support of our hypothesis, HR responses from a younger population had greater variation compared with an elderly population when interacting with the same robot configuration. However, GSR results were opposite to our expectation. We found greater variation in GSR from senior participants. This is possibly due to a training session included in the second study. Subjects were exposed to a robot without humanlike features before formal test trials. Participant familiarization with the control condition might have led to smaller GSR variation.

Contrary to hypothesis, no significant differences were found in the variability of subjective responses by younger versus older participants. However, this finding was in line with consistency in the use of the SAM, as reported in Backs et al. (2005) study. Both age groups were consistent in emotional ratings and such consistency was comparable across studies.

Robot condition, investigated in the second study, was also found to be a significant factor in emotional state classification accuracy. Although robot stimuli generated comparable variations in terms of SAM ratings, unequal variability was found between the sets of conditions yielding different levels of emotional responses in terms of physiological signals. Substantially higher physiological response variability induced by particular types of robot configurations may degrade the predictive utility of HR and GSR inputs in an emotional state classification model. The highest variability in HR was found when participants interacted with the robot without humanoid features; whereas, the highest variability in GSR was found when participants interacted with a robot integrating human face, voice and interactivity features. This indicated that extremely low or high degrees of humanness in a service robot can lead to inconsistency in emotional response. High emotion variability might have occurred with the very low degree of humanness robot (Condition 1) because some participants felt uncomfortable and/or disappointed with the lack of humanoid features, as compared to multiple feature robots. High emotion variability with the high degree of humanness robots (Condition 6 and 7) may have been due to user expectations regarding visual message and confirmation interactivity, based on prior HCI task experience. Participants might have felt these forms of interaction did not fit with human physical appearance and voices of the robot.

Conforming with hypothesis, robot configurations in the second study also generated more extreme emotional responses in terms of SAM ratings and HR responses, as compared with conditions in the first study. Results from Tukey's test revealed combinations of robot features to yield extremely high HR responses, compared to the control condition as well as individual feature manipulations, including an abstract face, synthesized voice and visual message (in the first study).

The neural network models we constructed included different sets of input and output features and were capable of representing non-linear relationships. Although Tukey's test is applied to data that can be represented using a linear model, results revealed highly separable levels of emotion (in terms of both subjective and physiological variables) for the second study. A linear classification model including mean HR and maximum GSR as inputs may provide good predictive utility for three categorical levels of arousal and valence responses, especially for a younger population.

Conclusion

In order to achieve healthcare service robot behavior adaptation in real-time when interacting with patients, a reliable algorithm is needed for classification of patient emotional states. Any classification approach should be robust across age groups and robot configurations. Based on our results, emotional state classification inaccuracy may stem from high emotion variability in younger participants and emotion uncertainty and intensity induced by multiple humanoid features in robot design. However, such extreme variability in user responses is only reflected through physiological signals. The SAM ratings were found to be consistent and comparable across age groups.

To account for age differences, the classification method should be customizable for individual patients. An alternative is to develop specific emotional state classification models for particular age groups. Similarity of physiological response characteristics should be considered as a basis for selecting demographic groups for data collection and model construction.

To account for robot configuration differences, research needs to indentify a set of robot features that yields specific emotional states. Caution should be used when designing service robots with extremely low or high degrees of humanness since particular robot feature combinations may induce different responses in different user groups. Future research should also design service robots to accommodate a wide range of users in terms of machine expressions in order to work effectively in public healthcare environments.

References

Baba, M., Watahiki, Y. Matsunaga, M., and Takebe, K. 1988. Sympathetic skin response in healthy man. *Electroencephalogr Clin Neurophysiol* 28: 277-283.

Backs, R.W., da Silva, S.P., and Han, K. 2005. A comparison of younger and older adults' self-assessment manikin ratings of affective pictures. *Exp Aging Res* 31(4): 421-440.

Barker, K.N., Flynn, E.A., Pepper, G.A., Bates, D.W., and Mikeal, R.L. 2002. Medication errors observed in 36 health care facilities. *Arch Int Med* 162(16): 1897-1903.

Bradley, M., and Lang, P. 1994. Measuring emotion: the self-assessment Manikin and the semantic differential. *J Behav Ther Exp Psychiatr* 25(1): 49-59.

Breazeal, C. 2003. Emotion and sociable humanoid robots. *Int J Hum Comput Stud* 59: 119-155.

Cialdini, R.B. 2001. *Influence: Science and practice* 4th edn. Needham Heights, MA.: Allyn & Bacon.

Dautenhahn, K., Walters, M.L., Woods, S.N., Koay, K.L., Nehaniv, C. L., Sisbot, A., Alami, R., and Siméon, T. 2006. How may I serve you? a robot companion approaching a seated person in a helping context. In *Proc ACM SIGCHI/SIGART Conf HRI*. Salt Lake City, UT.

Drory, V.E., and Korczyn, A.D. 1993. Sympathetic skin response: age effect, *Neurology* 43: 1818-1820.

Ekman, P. 1984. Expression and the nature of emotion. in *Approaches to Emotion*, eds. K.S. Scherer & P. Ekman, Hillsdale, NJ.: Erlbaum.

Gross, J.J., Carstensen, L.L., Pasupathi, M., Tsai, J., Skorpen, C.G., and Hsu, A.Y.C. 1997. Emotion and aging: Experience expression, and control. *Psychol Aging* 12(4): 590-599.

Lam, E.Y. 2004. Statistical modelling of the wavelet coefficients with different bases and decomposition levels. *IEE Proc Vis Image Signal Process* 151(3): 203-206.

Lawton, M.P., Kleban, M.H., Rajagopal, D., and Dean, J. 1992. Dimensions of affective experience in three age groups. *Psychology and aging* 7: 171-184.

Levenson, R.W., Carstensen, L.L., Friesen, W.V., and Ekman, P. 1991. Emotion, physiology, and expression in old age. *Psychol Aging* 6(1): 28-35.

Swangnetr, M., Kaber, D., and Lee, Y-S. 2009. Human emotional state classification based on wavelet analysis and statistical feature selection. In *Proc 17th IEA2009*. Beijing, China.

Swangnetr, M., Zhu, B., Taylor, K., and Kaber, D. 2010. Assessing the effects of humanoid robot features on patient emotion during a medicine delivery task. In *Proc 54th HFES*. San Francisco, CA.

Tanaka, H., Monahan, K.D., and Seals, D.R. 2001. Age-predicted maximal heart rate revisited. *J Am Coll Cardiol* 37:153-156.

Zhang, T., Kaber, D., Zhu, B., Swangnetr, M., Mosaly, P., and Hodge, L. 2010. Service robot design effects on user perceptions and emotional responses. *J Intell Serv Robotics* 3(2): 73-88.

Zhang, T., Zhu, B., Lee, L., and Kaber, D.B. 2008. Service robot anthropomorphism and interface design for emotion in human-robot interaction. In *Proc IEEE Int Conf Autom Sci Eng*, 674-679.

Zhu, B., and Kaber, D.B. 2010. Linguistic etiquette for service robots in a patient medicine delivery task. In *Proc AHFE*. Miami, FL.