

The Effects of Emotion of Voice in Synthesized and Recorded Speech

Clifford Nass

Department of Communication
Stanford University
Stanford, CA 94305-2050
+1 650-723-5499
nass@stanford.edu

Ulla G. Foehr

Department of Communication
Stanford University
Stanford, CA 94305-2050
+1 408-249-2444
ufoehr@stanford.edu

Michael Somoza

Department of Computer Science
Stanford University
Stanford, CA 94305
+1 650-625-1863
msomoza@cs.stanford.edu

ABSTRACT

This study examines whether emotion conveyed in recorded and synthesized voices affects perceptions of emotional valence of content, perceptions of suitability of content, liking of content, and credibility of content as well as whether recorded or synthesized speech influences perceptions differently. Participants heard two news stories, two movie descriptions and two health stories in a 2 (type of speech: recorded vs. synthesized) by 2 (consistency of voice emotion and content emotion: matched vs. mismatched) balanced, between subjects experiment. A happy voice, whether synthesized or recorded, made content seem happier and more suitable for extroverts, and a sad (synthesized or recorded) voice made content seem less happy and less interesting for extroverts. Participants reported liking content more when voice emotion and content emotion were matched, but rated information as more credible when voice emotion and content emotion were mismatched. Implications for design are discussed.

KEYWORDS

TTS (Text-to-Speech), CASA (Computers are social actors), Speech User Interfaces, emotion, consistency theory.

INTRODUCTION

As the Web evolves, it is rapidly moving toward voice-based interfaces. This evolution has been dictated by market demands for convenience (e.g., ability to do other things while surfing the Web; smaller devices lacking keyboards), safety (e.g., eyes-occupied situations, accessing web content while driving), and wider access for those with visual impairments and other disabilities. This raises the question: What characteristics of voices are important when presenting Web content?

When voice talents record content, they are careful to read content in a manner that is consistent with the material. That is, if they are reading happy or positive content, they speak in a cheerful tone, using higher pitch, faster speech and more pitch range than used in normal speech. Conversely, sad content is read more slowly and flatly. This matching of voice emotion with content is common (and necessary to be deemed emotionally intelligent) in everyday conversations [9, 16].

The psychological literature has confirmed that consistency is important. People expect consistency and prefer it to inconsistency. When they encounter inconsistencies, they enter a drive state and are motivated to adapt their perceptions in order to resolve the inconsistency [8].

The need for emotional consistency of recorded speech has been well understood in traditional media [7], but the situation is less clear in human-computer interaction. When computers present voices, the computer is cognized as the *source* of the voice, rather than as a medium [14, 18]; because computers are essentially emotionless [16], it is possible that emotion in the voice would be irrelevant in this case. The human-computer interaction literature has provided evidence that personality consistency [4, 10, 13] and consistency across modalities [12] are important, but there is no comparable research on emotional consistency in human-computer interactions [16].

If the question of importance of emotional consistency of *recorded* speech and content is open, then it seems even more questionable whether consistency in *synthetic* speech matters. It has been shown that users can identify emotion in synthetic speech [3], even if the speech sounds artificial and clearly machine-generated. However, identification is very different than effect. First of all, discovering that users can identify emotions *when asked to do so by an experimenter* does not mean that individuals think about the emotion of the computer *when they are not prompted to do so*. Furthermore, even if a user feels that a voice “seems” to sound happy or sad, that

does not imply that users will be *influenced* by that assessment. Synthesized speech, in comparison to recorded speech, provides an *additional* reminder that one is interacting with a machine rather than a person, and thus that emotion should be irrelevant to the interaction. On the other hand, research shows that individuals apply social expectations of consistency in unambiguous human-computer interaction in the same way that they do in human-human interaction [17].

The problem of consistency is made more complex because of the enormous scope and rapid growth of content on the Web. Content comes from so many sources so rapidly that it is impractical to label each sentence, let alone each part of a sentence, with the correct emotion. Automating the process is proving extremely difficult, as at present, there is no software that can reliably read text and determine its emotional tone. This leaves providers of content with a dilemma about how to convey information. If they choose one voice to communicate all content, what kind of a voice should be chosen? What are the implications if a voice's characteristics do not match the content being conveyed? What effects might mismatched voice and content have on people's perception of the information being conveyed?

To begin to address these questions, we present an experimental study that investigates: 1) whether voice emotion impacts perceptions of content emotion and suitability of content 2) the effect of consistency or inconsistency of voice emotion and content emotion on liking for and perceived credibility of the information presented, and 3) whether emotion in recorded speech influences perceptions of content differently than synthesized speech.

METHOD

Participants

Participants were 56 university students between the ages of 19 and 29 who were recruited for this study through an undergraduate communication course. Gender was approximately balanced across conditions. All participants were native English speakers, so as to reduce potential difficulties in understanding synthesized speech. Participants received course credit for their participation.

Procedure

The experiment was a 2 (consistency of voice and content: matched vs. mismatched) by 2 (type of speech: text-to-speech vs. recorded speech) between-participants, repeated measures design, with content type (news, movie descriptions, and health) as the repeated factor. Participants were randomly assigned to condition.

After being recruited by email, participants were asked to visit a website with the instructions for the experiment. Once they had read the instructions, participants called a phone number, punched in (via DTMF) a passcode to hear the appropriate content, and listened to six information segments, or "stories:" two news stories (one happy, one sad), two movie descriptions (one happy, one sad), and two

health stories (one happy, one sad). After each story, participants answered a set of questions on a website (<http://www.stanford.edu/~msomoza/study/main.fft>). At the end of the experiment, participants heard a sample of each of the voices via telephone, and answered questions about each voice ("Here is one of the voices you have heard. Please go to the website and indicate your opinions about this voice."). All instructions were delivered via the website, not the telephone, in order to avoid using a new voice for giving instructions.

Manipulation

Type of Voice Manipulation. The CSLU Toolkit¹ running on an NT machine was used to run the experiment; a Dialogics board answered the phone calls. The synthesized voices were based on the default male voice provided by the Toolkit. To create a happy voice, we used settings of F0=184 Hz, pitch range=55 Hz, and word rate=198 words per minute. The sad voice had F0=90 Hz, pitch range=5 Hz, and word rate=157 words per minute. These settings were based on the literature that discussed markers of emotion in speech [3]. For the recorded voice, we used a male graduate student whose voice was similar to the TTS voice in pitch.

Manipulation checks indicated that in both the TTS and recorded conditions, the happy voice was perceived as significantly happier than the sad voice.

Content Manipulation. Within each content category (news, movie descriptions and health), we created one "happy/positive" story and one "sad/negative" story. A manipulation check confirmed that the happy stories were perceived as significantly happier than the sad stories.

Within the text-to-speech (TTS) and recorded speech conditions, half of the participants heard each story read in a voice that matched the emotional valence of the story; the other half heard each story read in a voice that mismatched the emotion conveyed in the story.

Measures

All dependent measures were based on items from the Web-based questionnaire. Most questions used a ten-point Likert scale, with radio buttons for indicating user response.

Questions about perceptions of story content, suitability of the content for particular audiences, liking of the content, and credibility of the content were asked for each story. The first set of questions asked: "Please indicate how interested the following kinds of people would be in the story about..." followed by a list of types of people and a ten-point Likert scale anchored by "Very Uninterested" (=1) and "Very Interested" (=10). The second set of questions asked: "Please indicate how well

¹ The CSLU Toolkit is downloadable without cost at <http://cslu.cse.ogi.edu>.

the following adjectives describe the story about...” followed by a ten-point Likert scale anchored by “Describes Very Poorly” (=1) and “Describes Very Well” (=10). Questions assessing the two different voices were asked at the end of the questionnaire; they also employed ten-point Likert scales.

Several indices were created for each story to measure the concepts being tested. Indices were driven by theory and confirmed using factor analysis.

Perception of happiness of content was an index comprised of the adjectives happy and sad (reverse-coded).

Perception of suitability of content for extroverts was an index comprised of the categories of people: extroverts and introverts (reverse-coded).

Liking of the content was an index comprised of six adjectives: boring (reverse-coded), engaging, enjoyable, interesting, informative, and likable. The indices were highly reliable (average Cronbach’s *alpha* for the six indices was .91).

Credibility of the content was an index comprised of three adjectives: objective, tells the whole story, and unbiased. The indices were reliable (average *alpha* was .71).

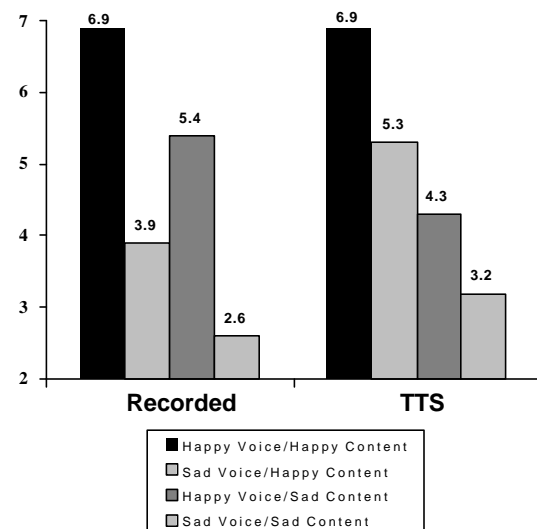
RESULTS

To examine the effect of emotion of voice on perception of content, for each dependent variable, we first analyzed the recorded speech and TTS conditions separately, looking at happy and sad stories independently. In this analysis, valence of voice emotion was the between-participants factor, with content as the repeated factor. We then pooled the recorded and TTS data and used that as a second factor in our analysis.

Perception of valence of content

Results indicate that the emotion in the voice significantly affected the perception of the valence of content for both recorded speech and TTS (see Figure 1).

Figure 1. Perceived Happiness of Content



In the recorded condition, participants who heard happy content read in a happy voice rated the stories as happier than those who heard the same content read in a sad voice, $F(1, 27)=51.2, p<.001, \eta^2=.66$. Similarly, those who heard sad content read in a sad voice rated the stories as less happy than those who heard it read in a happy voice, $F(1, 27)=29.8, p<.001, \eta^2=.53$.

The TTS conditions exhibited the same effects as recorded speech. Participants who heard happy content read in a happy voice rated the stories as happier than those who heard the same content read in a sad voice, $F(1, 27)=7.3, p<.01, \eta^2=.22$. Similarly, those who heard sad content read in a sad voice rated the stories as less happy than those who heard it read in a happy voice, $F(1, 27)=4.5, p<.04, \eta^2=.15$.

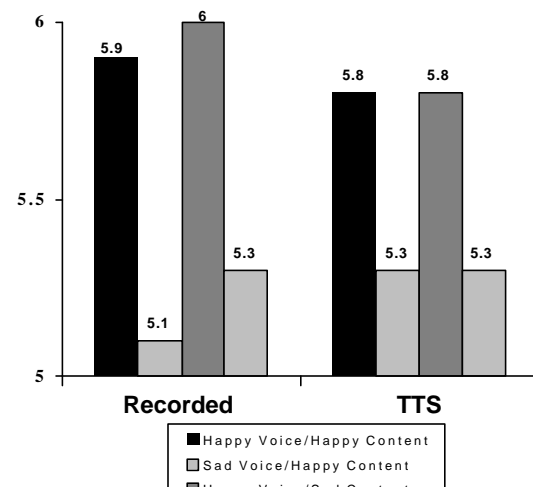
When TTS and recorded speech conditions were analyzed together, there was a significant interaction for both happy and sad content, indicating that the effect of voice emotion on content perception was stronger in recorded speech than in TTS. For happy content, there was a significantly greater difference between happy and sad voices for recorded speech than TTS (the two left bars in each cluster in Figure 1), $F(1,55)=4.0, p<.05, \eta^2=.07$. Similarly, for sad content, there was a significantly greater response to the recorded voices as compared to the TTS voices (the two right bars in each cluster in Figure 1), $F(1,55)=5.4, p<.02, \eta^2=.09$.

As a result of the significance for the two modalities independently, when the data from the two modalities are pooled, a happy story read in a happy voice was perceived as happier than the same story read in a sad voice, $F(1, 55)=40.7, \eta^2=.44, p<.001$, and a sad story read in a sad voice was perceived as less happy than the same story read in a happy voice, $F(1,55)=28.7, \eta^2=.36, p<.001$.

Perception of suitability of content

The effect of voice emotion on perception of suitability of content for particular audiences was assessed using the same procedures as used to examine the effect of emotion of voice on perception of valence of content. We first analyzed TTS and recorded conditions separately, looking at happy and sad stories independently. In this

Figure 2. Perceived Suitability of Content for Extroverts



analysis, valence of voice was the between-participants factor, with content as the repeated factor. We then pooled the recorded and TTS data and used that as a second factor.

Results show that the emotion conveyed by the voice significantly affected the perception of suitability of content for extroverts for both recorded speech and TTS (see Figure 2).

In the recorded condition, participants who heard happy content read in a happy voice rated the stories as more interesting for extroverts than those who heard the same content read in a sad voice, $F(1, 27)=6.2, p<.02, \eta^2=.19$. Similarly, those who heard sad content read in a sad voice rated the stories as less interesting for extroverts than those who heard it read in a happy voice, $F(1, 27)=6.9, p<.01, \eta^2=.21$.

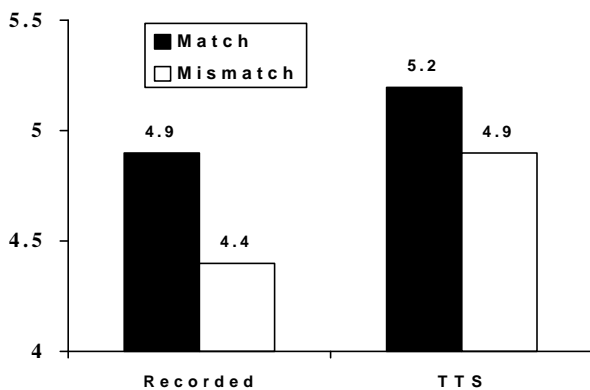
The TTS condition displayed the same effects as recorded speech. Participants who heard happy content read in a happy voice rated the stories as more interesting for extroverts than those who heard the same content read in a sad voice, $F(1, 27)=6.5, p<.02, \eta^2=.20$. Similarly, those who heard sad content read in a sad voice rated the stories as less interesting for extroverts than those who heard it read in a happy voice, $F(1, 27)=5.2, p<.03, \eta^2=.17$.

Consistent with the findings for each modality independently, when the data from the TTS and recorded speech conditions are pooled, a happy story read in a happy voice was perceived as more interesting for extroverts than the same story read in a sad voice, $F(1, 55)=11.8, p<.001, \eta^2=.19$, and a sad story read in a sad voice was perceived as less interesting for extroverts than the same story read in a happy voice, $F(1,55)=12.1, p<.001, \eta^2=.19$. There was no interaction for either happy content ($p>.39$) or sad content ($p>.68$), suggesting that participants were affected similarly by TTS and recorded speech.

Liking of the content

A repeated measures ANOVA was conducted to determine how consistency of voice emotion and content emotion influence liking of the content. This analysis used the six stories as the repeated factor, and consistency of voice and content (matched or mismatched) as the between-subjects

Figure 3. Liking of Content



factor.

In the recorded condition, respondents liked the content more when voice emotion and content emotion were matched than when they were mismatched, $F(1,27)=6.3, p<.02, \eta^2=.20$ (see Figure 3). Although the effects were in the same direction, there was no significant difference in liking for the TTS condition ($p>.34$).

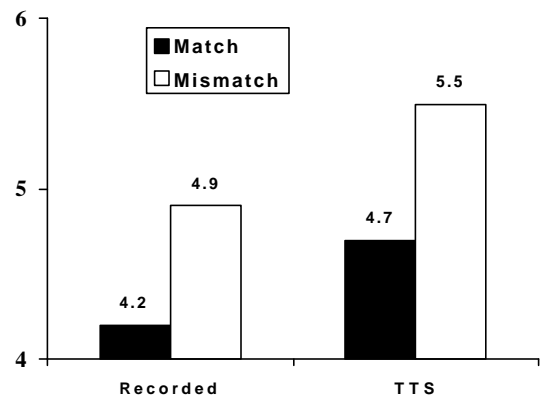
For the pooled data, there was a clear and significant preference for the matched content as compared to the mismatched content, $F(1,55)=6.32, p<.01, \eta^2=.08$. Content was also liked more when read in a TTS voice as opposed to recorded speech, $F(1,55)=6.32, p<.01, \eta^2=.09$. There was no interaction ($p>.65$).

Credibility of the content

Consistency also had an effect on credibility, but in a direction that would seem to be opposite to that of liking: Participants found the content to be more credible when voice emotion and content emotion were *mismatched* rather than matched.

In the recorded condition, respondents rated the content

Figure 4. Credibility of Content



more credible when voice emotion and content emotion were mismatched rather than matched, $F(1,27)=6.6, p<.02, \eta^2=.20$ (see Figure 4).

The same effect held true in the TTS condition, with respondents finding content more credible when voice characteristics and content were mismatched rather than matched, $F(1,27)=6.8, p<.02, \eta^2=.21$.

For the pooled results, main effects were found for both consistency and type of speech. Given the foregoing, content was rated more credible when voice emotion and content emotion were mismatched rather than matched, $F(1,55)=13.4, p<.001, \eta^2=.20$. TTS was considered more credible than recorded speech, $F(1,55)=6.09, p<.02, \eta^2=.11$. There was no interaction between type of speech and consistency of voice and content ($p>.83$).

DISCUSSION

Emotion detection is a finely-tuned skill in humans as well as other animals [16]. The present results demonstrate that humans use that skill when responding to voices on computers, whether those voices are recorded speech or synthesized speech. But users don't stop at detection: The assessment of emotional valence in voice affects their perception of, liking, and perceived credibility of what is being said. While these results for recorded speech are consistent with the findings in traditional media, the fact that the emotion of TTS voices influences users is remarkable.

Our first results demonstrate that the emotion manifest in a voice influences the perception of the content; specifically, happy voices make both happy content and sad content seem happier than the same content presented by a sad voice. This is a prediction of cognitive dissonance theory [15], which argues that inconsistent perceptions leads to a negative drive state that tends to reduce the inconsistency; in this case, the user reinterprets the content to make it seem more consonant with the presenting voice.

Not surprisingly, recorded speech seems to have a greater effect on users' perceptions than does synthesized speech. It is unclear whether this is because the emotion was less detectable or whether the emotion was partially dismissed because computers are not emotional. Future research should address this question, although the lack of an interaction for perceptions of suitability of content, liking and credibility suggest that the differences between TTS and recorded speech may be overrated.

Voice emotion also determined the perceived target audience for content. Content read in a happy voice was perceived as being more interesting to extroverts, while content read in a sad voice was perceived as being less interesting to extroverts. In sum, even synthesized speech can have strong effects on the perceived emotion and suitability of content.

Our other results focus on the match or mismatch between the emotion of the voice and the emotion of the content. Effects in these cases are striking because the previous results suggest that users, through their processing of the content, minimize the actual mismatch.

For recorded speech, match is clearly positively associated with liking, consistent with the predictions from consistency theory. Fortunately, achieving consistency in recorded speech is straightforward: Because readers, unlike computers, are extremely good at detecting the emotion reflected by content, and because humans tend to automatically match the emotion of their voice to content, emotionally-appropriate recorded speech is essentially automatic.

The results for TTS match and liking were not significant, but in the expected direction. Again, it is unclear whether weakness of the manipulation or dismissal was at play. These results, like those for perceived emotion in content,

suggest that emotion in TTS is not as powerful as emotion in recorded speech.

The results for credibility may seem paradoxical: Credibility is generally positively associated with liking [2]. We believe the explanation for this phenomenon is that in cases of voice/content mismatch, users draw on their experiences in human-human interaction to understand the discrepancy between voice and content [17]. In human-human interaction, voice and content are "appropriately" mismatched when valenced content is read in a neutral voice to reflect objectivity or neutrality. If mismatch in the present study suggested a sense of detachment, the mismatch condition would be perceived as more objective and hence credible. In support of this explanation, a study involving matched or mismatched personalities and bidding behavior [13] found greater persuasion in the mismatched conditions, while individuals liked the matched conditions significantly more. The fact that mismatch was more credible in both recorded speech and TTS in our present study lends support to this explanation.

One possible practical resolution to the conflict between liking and credibility (when both are important) is to have a neutral voice, or at least a voice that becomes more neutral as credibility becomes more important than liking. Future research should elaborate the tradeoffs between clearly marked emotion and effects on perceptions of content.

This study focused on the manifestation of emotion through both recorded and synthetic voice. It is likely, however, that the issue of emotion matching will be relevant to other modalities as well. For example, when people present themselves on the Web via video, their facial expressions generally match their emotion. However, manifestation of emotion in synthetic faces or embodied conversational agents [5, 6] require highly complex control over the face and body as well as a nuanced emotion model. Because humans' sensitivity to facial emotion is much more highly developed than even voice-based detection, the importance of appropriately manifesting emotions in these more anthropomorphic representations is even more critical.

There are also other opportunities for manifesting emotion. Many websites and software are associated with happy (e.g., party sites) or sad (e.g., funeral parlor sites) issues. Even graphical images or sound icons are assigned a valence [1, 11]. Designers must consider all of these potential emotional matches and mismatches.

At the highest level, the most remarkable result of the present study is that emotion markers in synthesized speech can dramatically influence perceptions of and responses to content. While mere discernment of emotion might have a straightforward cognitive processing explanation [3, 12], the influence of TTS emotion on responses to content suggests that users are treating the voice as an emotion-rich *source* of content [18],

independent of the emotionless machine. That is, even *synthetic* voices are treated as social actors [17].

The importance of emotion in synthetic speech represents a significant opportunity and problem for interface designers. As an opportunity, the present results suggest that an “emotion markup language” would be an enormous plus for systems, especially as voices become more human-like. The problem is that at present, there are no systems that can reliably automate the mark-up process, so emotional specification is a labor-intensive process.

One caveat in the present study is that we used a single recorded voice and a single synthesized voice. Future research should vary the characteristics of the voices, including gender, age, personality, etc..

Another open issue is emotional *adaptation*. In the present study, all of the voices adapted to the content, although for half of the subjects, the adaptation was always *mismatched*. However, many systems assign a single emotional tone to the voice, regardless of content, thereby sometimes matching and sometimes mismatching. Similarly, many e-commerce websites show a still smiling face which is referred to as the “helper” or “advisor,” even when the user is told that their desired item is out of stock or that the user has made a mistake. Future research should explore whether consistent emotional expression is discounted, or whether users nonetheless are influenced.

In sum, the present results suggest that careful attention to the emotion manifest in voice output, whether recorded or synthesized, is a critical issue for interface designers. Just as a bad actor can destroy a great play, a poorly directed or created voice can undermine an interface.

Reference

1. Ball, G. & Breese, J. Emotion and personality in a conversational agent. Pp. 189-219 in J. Cassells, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents*. Cambridge, MA: MIT Press.
2. Berlo, D., Lemert, J. & Mertz, R. (1970). Dimensions for evaluating the acceptability of message sources, *Public Opinion Quarterly*, 33, 563-576.
3. Cahn, J. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8: 1-19.
4. Cantor, N., & Mischel, W. (1979). Prototypes in person perception. *Advances in Experimental Social Psychology*, 12, 3-52.
5. Cassells, J. & Thorisson, K. (in press). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Journal of Applied Artificial Intelligence*.
6. Cassells, J., Sullivan, J., Prevost, S. & Churchill, E. (Eds.), *Embodied conversational agents*. Cambridge, MA: MIT Press.
7. Field, S. (1994). *Screenplay: The foundations of screenwriting*. New York: Bantam Doubleday Dell.
8. Fiske, S.T. & Taylor, S.E. (1991). *Social cognition*. New York: McGraw-Hill, Inc.
9. Goleman, D. (1995). *Emotional Intelligence: Why it can matter more than IQ*. New York: Bantam Books.
10. Isbister, K. & Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Interaction*, 53(1), 251-267.
11. Lang, P.J. (1993). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5), 372-385.
12. Nass, C. & Gong, L. (in press). Social aspects of speech interfaces from an evolutionary perspective: Experimental research and design implications. *Communications of the ACM*.
13. Nass, C., & Lee, K.M. (submitted). Does computer-generated speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency. *Journal of Experimental Psychology: Applied*.
14. Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research*, 19, 504-527.
15. Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Dubuque, Iowa: Wm. C. Brown Company Publishers.
16. Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.
17. Reeves, B. & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.
18. Sundar, S.S. & Nass, C. (in press). Source orientation in human-computer interaction: Programmer, networker, or independent social actor? *Communication Research*.