Toward a Social-Cognitive Psychology of Speech Technology:
Affective Responses to Speech-Based e-Service

by

Melanie Diane Polkosky

A dissertation in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Judith Becker Bryant, Ph.D.
Michael Brannick, Ph.D.
Mark Pezzo, Ph.D.
Sandra Schneider-Wright, Ph.D.
Paul Spector, Ph.D.

Date of Approval:
February 28, 2005

Keywords: Human computer interaction, dialogue design, customer satisfaction, e-service, speech user interface design

© Copyright 2005, Melanie D. Polkosky
Dedication

For the users - those who have ever had their day ruined by a poorly designed speech system and those who rely on communication technologies and know that conversation is a most precious gift the majority of us take for granted. May we continue to find ways to use technology to make life better.

And for Dr. Tom Tighe, who showed a tremendous love of knowledge and concern for his students during the short time I knew him – this paper would have greatly benefited from his guidance.
Acknowledgements

There are so many people to thank for all the support and assistance I received while working on my degree and this paper. First and most importantly, my wonderful husband Steve, who believes in me when I don’t, makes me grateful for everything I have been blessed with, and gives me a reason to smile every day. To Lyndsey, my niece, friend, and biggest cheerleader, who brings joy to my heart.

To the teachers who helped me see a future when I needed direction and strength - Marianne Sullivan, Dr. Steve Miller, and Dr. Jim Jenkins. Special additional thanks to Dr. Judy Bryant for patiently picking me up, dusting me off, and sending me on my way through some of the most trying times.

Maria Rosa Brea-Spahn, Laurie Diamond, and Lin Smith, great role models and friends, for fighting the fight to the dastardly dissertation and daunting degree with me. John Pruitt, who suggested that I would make a better human factors practitioner than computational linguist.

And to all my colleagues at IBM, who challenge me and help me understand what I do better than I ever could alone. Jim Lewis, the yin to my yang on this particular speech technology topic, who got me thinking about it in the first place. Tom Capone, who helped me recover my lost proposal draft after my computer’s hard drive finally refused to cooperate for the last time.

Vanessa Michelini, Brent Davis, Tami Harris, and Marv Silver, who have known the right thing to say at the right moment to make me keep believing.

I am truly blessed.
# Table of Contents

List of Tables

List of Figures

Abstract

Introduction

Review of Literature
  Social-Cognitive Theoretical Framework
    Interpersonal Communication is Goal-Directed Social Behavior
    Speech and Language Behavior are Social Cues
    Humans Automatically Infer Traits from a Partner’s Social Cues
    Service Delivery is a Special Case of Goal-Directed Interpersonal Interaction

Human Users Have Social Cognitions about Interactive Technologies

Toward a Social-Cognitive Psychology for Speech Based e-Service
  Service Quality
  Customer Satisfaction
  User Satisfaction
  Antecedents of Customer Satisfaction
    Individual Differences
    Point of View
    Personality and Behavior of Service Providers
    Trust
    Comfort
    Ease of Use and Usability
  Consequences of Customer Satisfaction
  Summary of Potential Speech-Based e-Service Constructs

Overview of the Proposed Study
  Rationale
  Research Questions

Study 1: Speech Interface Expert Ratings
  Method
    Participants
    Measures
    Stimuli
Appendices

Appendix A: Audio Recording Transcripts 120
Appendix B: Participant Instructions 121
Appendix C: Participant Demographic Questions 140
Appendix D: Dabholkar and Bagozzi (2002) Inherent Novelty Seeking Scale 141
Appendix E: Dabholkar and Bagozzi (2002) Need for Interaction with a Service Employee Scale 142
Appendix G: Items for Usability Scale 144
Appendix I: Eroglu and Machleit (1990) Customer Satisfaction Scale 146

About the Author
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Summary of speech-based e-service articles in leading journals 1990-2003</td>
<td>4</td>
</tr>
<tr>
<td>Table 2</td>
<td>Speech interface stimuli</td>
<td>45</td>
</tr>
<tr>
<td>Table 3</td>
<td>Experts’ overall ratings of quality for 16 speech interfaces</td>
<td>48</td>
</tr>
<tr>
<td>Table 4</td>
<td>Open ended and multiple choice manipulation check responses</td>
<td>58</td>
</tr>
<tr>
<td>Table 5</td>
<td>Means, standard deviations and correlations for hypothesized Customer Service Expectations factor</td>
<td>60</td>
</tr>
<tr>
<td>Table 6</td>
<td>Means, standard deviations and correlations for hypothesized Pragmatics factor</td>
<td>60</td>
</tr>
<tr>
<td>Table 7</td>
<td>Means, standard deviations and correlations for hypothesized Recognition factor</td>
<td>61</td>
</tr>
<tr>
<td>Table 8</td>
<td>Means, standard deviations and correlations for hypothesized Affective Response factor</td>
<td>61</td>
</tr>
<tr>
<td>Table 9</td>
<td>Means, standard deviations and correlations for hypothesized User Goal Orientation factor</td>
<td>61</td>
</tr>
<tr>
<td>Table 10</td>
<td>Means, standard deviations and correlations for hypothesized Accuracy factor</td>
<td>62</td>
</tr>
<tr>
<td>Table 11</td>
<td>Means, standard deviations and correlations for hypothesized Prompt Wording factor</td>
<td>62</td>
</tr>
<tr>
<td>Table 12</td>
<td>Means, standard deviations and correlations for hypothesized Usefulness factor</td>
<td>62</td>
</tr>
<tr>
<td>Table 13</td>
<td>Means, standard deviations and correlations for hypothesized Speech Impression factor</td>
<td>63</td>
</tr>
<tr>
<td>Table 14</td>
<td>Summary of reliability analysis for usability scale</td>
<td>65</td>
</tr>
<tr>
<td>Table 15</td>
<td>Initial factor analysis loadings for usability scale</td>
<td>67</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Table 16</td>
<td>Second factor analysis loadings for usability scale</td>
<td>68</td>
</tr>
<tr>
<td>Table 17</td>
<td>Post-hoc results for interface quality</td>
<td>70</td>
</tr>
<tr>
<td>Table 18</td>
<td>Descriptive statistics and bivariate correlations for individual differences, usability factors, comfort, and customer satisfaction</td>
<td>72</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Speech user interface quality and customer satisfaction</td>
<td>59</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Scree plot for usability scale factor analysis</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Usability factor scores for the different interfaces</td>
<td>71</td>
</tr>
</tbody>
</table>
Toward a Social-Cognitive Psychology of Speech Technology: Affective Responses to Speech-Based e-Service

Melanie Diane Polkosky

ABSTRACT

Speech technologies, or technologies that recognize and respond to human speech, have recently emerged as a ubiquitous and cost-effective form of customer self-service (e-service). Although customer satisfaction is regarded as an important outcome of e-service interactions, little is known about users’ affective responses to conversational interactions with technology. Using a theoretical foundation derived from research in social cognition, interpersonal communication, psycholinguistics, human factors, and services marketing, two studies develop items for a speech interface usability scale, which is then used to examine interrelationships among individual differences (e.g., self-monitoring, need for interaction with a service provider, inherent novelty seeking), usability, comfort, and customer satisfaction. In the first study, speech and language experts’ ratings of sixteen speech interfaces are used to identify six stimuli for the second study that represent high, average, and low quality. In the second study, participants listen to the six interfaces and provide ratings for the primary variables of interest. Results indicated that speech interface usability consists of four factors (User Goal Orientation, Customer Service Behavior, Verbosity, and Speech Characteristics). Usability items and individual differences predict affective responses to speech-based e-service. Implications of these findings for psychological and communication research and applied speech technology are described.
Toward a Social-Cognitive Psychology of Speech Technology:

Affective Responses to Speech-Based e-Service

In their influential text, Sanders and McCormick (1993) assert that the two major objectives of the field of human factors are to: (1) “increase the effectiveness and efficiency with which work and other activities are carried out” and (2) “enhance certain desirable human values, including improved safety, reduced fatigue and stress, increased comfort, greater user acceptance, increased job satisfaction, and improved quality of life” (p. 4). Because human factors has been a problem-focused field, the scope and boundaries of it continually change in response to technological advances (Sanders & McCormick, 1993; Wickens, Gordon, & Liu, 1998). Sanders and McCormick (1993) recognized that part of the advancement in human factors beyond the 20th century would probably involve variables rarely considered by past human factors experts: “We hope that in the future human factors will become more involved and recognized for its contribution to the quality of life and work, contributions that go beyond issues of productivity and safety and embrace more intangible criteria such as satisfaction, happiness, and dignity” (p. 9).

In the early 21st century, the emergence of technologies that provide customer service via electronic networks, known as e-service, marks a paradigm shift in traditional notions of customer service and human interaction with technology. Rust and Kannan (2003) describe the emerging e-service paradigm: “This service-focused paradigm uses two-way dialogue to build customized service offerings, counting on knowledge about the customer to build strong customer relationships. Profitability is based on revenue expansion more than cost reduction, with revenues driven by enhanced service and higher levels of customer satisfaction…. One of the natural outcomes of the network-based economy is the emergence of customers as the focal point of all businesses” (p. 36).
Speech user interfaces are a means of human-computer interaction in which both the input to the system (user response) and the output from the system (machine presentation) use speech (Balentine & Morgan, 1999). Speech interfaces are rapidly becoming an important part of e-service, often replacing or supplementing customer call centers, because they allow highly personalized, easy to use, and cost-effective interaction between an enterprise and its customers (Balentine & Morgan, 1999; Kotelly, 2003; Rust & Kannan, 2002). This form of e-service requires the customer (user) to converse with a self-service technology system as the primary mechanism of customer service.

Humans’ interaction with technology has received some limited empirical attention over the past several years but the applied literature has only focused minimally on the interactive aspects of speech technology. Instead, the majority of empirical work has been conducted with communication technologies that serve as the medium of communication; they enable human partners to converse. Technologies that allow human-human communication include the now-familiar forms of email, chat, videoconferencing, instant messaging, telephone, and augmentative and alternative communication for individuals with communication disabilities (Bedrosian, Hoag, Calculator, & Molineux, 1992; Coleman, Paternite, & Sherman, 1999; Fussell & Benimoff, 1995; Savicki, Kelley, & Oesterreich, 1999; Storck & Sproull, 1995; Walther, 1996, 1997; Wellens, 1993). These forms of technology disrupt or obscure most forms of nonverbal and extralinguistic communication, which has been the focus of applied research and social-psychological theory development. Because the interaction occurs between human partners, the application of social-cognitive and interpersonal communication theory to computer-mediated communication is clearly reasonable (for a review, see Spears, Lea, & Postmes, 2001): both partners are human and can be assumed to engage intentionally in their communicative interactions. In addition, humans have a wide range of social-cognitive behaviors, including attributions,
judgments, and inferences about their communicative exchanges (Fiske & Taylor, 1994; Holtgraves, 2002), which are thought to be markedly different in computer-mediated contexts (Spears, Lea, & Postmes, 2001).

In contrast, speech user interface systems replace a human partner in conversational exchanges. When speech interfaces are used in customer service, they generally replace a human customer service representative or operator. Speech interfaces may use speech recognition to understand a human user’s utterances and synthetic speech to respond, or they may use speech recognition for comprehension and the recorded utterances of a professional voice talent to respond to the user (Balentine & Morgan, 1999; Kotelly, 2003). Unlike e-service via Internet webpages, speech interfaces generally do not provide significant visual input to the user, but they allow ubiquitous access between business and its customers via the common telephone or other wireless technologies (Balentine & Morgan, 1999; Rust & Kannan, 2002). Because the mode of interaction between speech interface and a user is almost exclusively auditory, user skills and needs may dictate numerous aspects of design in a fundamentally different way than do visual interfaces (Sanders & McCormick, 1993). In addition, human social-communicative skills and conversational expectations form an important part of the usefulness and ease of use of speech-based e-service interfaces. Functions that are currently handled by speech user interfaces include banking and financial transactions, information retrieval, airline reservations, stock and mutual fund inquiries, directory assistance, and other relatively simple, predictable, or constrained customer service transactions (Balentine & Morgan, 1999).

Perhaps due to the recent emergence of both speech technologies and e-service, the empirical literature to date has been fragmented. A number of disparate disciplines have handled aspects of interpersonal communication, social interaction, customer service delivery, and speech technology, but these literatures have had minimal influence on the characteristics of speech
technology in the customer service arena. As shown in Table 1, a review of the articles published in several major and specialized journals in marketing, communication, human factors, and social psychology revealed no articles that simultaneously used the broad keywords “speech.”

Table 1. Summary of Speech-Based e-Service Articles in Leading Journals, 1990-2004

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Keyword(s)</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Marketing</td>
<td>Customer service or e-service</td>
<td>10 9 14</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0 0 1</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>0 2 0</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>2 1 1</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>5 2 1</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Journal of Service Research</td>
<td>Customer service or e-service</td>
<td>* 11 39</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>* 0 0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>* 0 2</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>* 0 0</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>* 1 1</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>* 0 1</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>* 0 0</td>
</tr>
<tr>
<td>Human Communication Research</td>
<td>Customer service or e-service</td>
<td>1 0 0</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>2 4 3</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>41 31 20</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>9 8 17</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Journal of Applied Psychology</td>
<td>Customer service or e-service</td>
<td>6 2 3</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>0 1 0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>2 1 1</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>2 2 1</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>3 7 14</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Journal of Personality and Social Psychology</td>
<td>Customer service or e-service</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0 0 0</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>0 0 1</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>19 18 12</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>146 110 86</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Journal Name</td>
<td>Customer service or e-service</td>
<td>Speech technology</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Personality and Social Psychology Bulletin</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0</td>
</tr>
<tr>
<td>Human Factors</td>
<td>Customer service or e-service</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0</td>
</tr>
<tr>
<td>International Journal of Human Computer Interaction</td>
<td>Customer service or e-service</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>0</td>
</tr>
<tr>
<td>International Journal of Speech Technology</td>
<td>Customer service or e-service</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Speech technology</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Self-service technology</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Human-computer interaction</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Interpersonal communication</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Social interaction</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Speech AND technology AND service</td>
<td>*</td>
</tr>
</tbody>
</table>


“technology,” and “service” in their titles or abstracts during the past 13 years. Searches for other common (and broad) keywords showed that these journals remain devoted to discipline-specific aspects of speech-based e-service, consistent with their traditional scope. For example, the Journal of Marketing and more specialized Journal of Service Research focused on customer service (as indicated by publication of 83 articles over the 13 year period) but were very limited in their treatment of technology and communication. Similarly, social psychology and communication journals also excluded extensive treatment of technology, in preference to articles concerned with human-human communication or broader social interaction topics. As an example of the limited treatment of communication technologies in these disciplines, all 10 articles using
“human-computer interaction” as a keyword in *Human Communication Research* were concerned with computer-mediated communication between humans. Neither the *Journal of Personality and Social Psychology* nor *Personality and Social Psychology Bulletin* published any applied articles concerned with technologies for social interaction.

Speech technology has also had very limited impact in applied psychology. *Human Factors* and the *International Journal of Human-Computer Interaction* have only tangentially included speech technology over the past decade. The specialized *International Journal of Speech Technology* was initiated in 1997, but it largely excluded topics closely related to user-system interaction and aspects of the technology as a customer service mechanism. Its published articles focused on the technology itself, as opposed to user issues related to speech interface design. Overall, this brief review of journal contents supports the contention that interpersonal communication between e-service speech user interfaces and human customers is largely ignored by leading journals in each of the relevant disciplines.

As a consequence of the limited research and applied nature of speech-based e-service, human factors experts who work in this field have a limited theoretical basis and few empirically-supported guidelines for designing speech user interfaces. This situation is inconsistent with the fundamental principles that define human factors as a distinct field. Sanders and McCormick (1993) identified six doctrines of human factors:

- “Commitment to the idea that things, machines, etc. are built to serve humans and must always be designed with the user in mind
- Recognition of individual differences in human capabilities and limitations and an appreciation for their design implications
- Conviction that the design of things, procedures, etc. influences human behavior and well-being
- Emphasis on empirical data and evaluation in the design process
- Reliance on the scientific method and the use of objective data to test hypotheses and generate basic data about human behavior
- Commitment to a systems orientation and a recognition that things, procedures, environments, and people do not exist in isolation” (p. 5).

Although these principles have provided grounding for the discipline of human factors, they have not guided speech interface research and design.

At the present time, speech technology design for customer service requires a firmer theoretical basis and a comprehensive understanding of the ways that these interfaces impact their human users. In addition, better understanding of the interactive effects of human social-communicative skills and speech interface characteristics is needed. Although the empirical research has not specifically filled this void to date, several research streams have considered aspects of the problem. First, the social psychology and interpersonal communication fields have demonstrated that communication is a goal-oriented social behavior that influences person perception, impression formation, expectations, and partner behavior within a normatively-structured conversation (Baron, Byrne, & Johnson, 1998; Fiske & Taylor, 1991; Holtgraves, 2002). In addition, a speaker’s speech and language behavior provides a powerful form of social information that can influence a communicative partner (Holtgraves, 2002; Neumann & Strack, 2001), even when the speaker is a non-animate technology. Social interaction results in automatic trait inference, or the assumption of speaker traits in the presence of social information (Fiske & Taylor, 1991), which suggests that a speaking technology interface may causally influence users’ cognitions. Finally, researchers in service marketing have identified customer service as a specialized form of interpersonal social interaction (Solomon, Surpreant, Czepiel, & Gutman, 1985), linking findings from social-cognitive psychology to the applied setting
of customer service delivery. A missing piece of this theoretical puzzle is an understanding of the interrelationships among perceptual and affective variables when customer service is provided by a speech user interface.

In this introduction, I review a broad, multidisciplinary literature that indicates that human users of speech recognition systems perceive these systems and interact with them much as they would a human conversational partner. The speech and language cues provided by a speech interface activate affective responses that would appear to be reserved only for human partners. The primary goal of the current study was to identify and elucidate relationships among affective outcomes of speech-based e-service. To establish a theoretical framework, I review the social psychological literature concerned with interpersonal communication and social cognition (specifically, impression formation and person perception), followed by a summary of this literature’s impact on customer service research. Then, I review variables that have been considered in the interpersonal communication, human factors, social psychology, and service marketing disciplines to identify an initial set of constructs that may illuminate users’ social-cognitive responses to interactive speech systems. From this foundation, several research questions are derived for an initial foray into an applied social-cognitive psychology of speech technology.

Review of Literature

A controversy exists in the field of speech technology: some researchers assume that human social behaviors are not relevant to the design of interactive dialogue systems. In such systems, the “speaker” is an abstraction and does not appear to warrant the same social consideration as a human conversational partner. The system does not have intentionality, as humans are presumed to have. Balentine and Morgan (1999) argue that
“humans have no precedent for verbal interaction with non-sentient devices that are not self-aware. Indeed, much human speech is dependent on a powerful set of social techniques that derive from a fundamental assumption – when a human being talks to another, that other human presumably has some stake in the outcome of the interaction. This assumption collapses when the partner is a machine. The result of this expectation of social awareness… is that structured and goal-oriented protocols become necessary to steer the user away from social speech behaviors and toward work-oriented, task-oriented interactions” (p. 12).

This perspective suggests that a speech-based interface can be designed so as not to be social, which will subsequently lead to greater efficiency in reaching user goals. Efficiency and sociability are seen as opposites: an increase in one leads to a decrease in the other. Balentine and Morgan (1999) identify anthropomorphism, “the attribution of human characteristics to non-human beings or things” (p. 218), as the underlying reason for social responses to speech interfaces. They suggest that a designer may choose to personify an interface, exploiting users’ tendency toward anthropomorphism. Their discussion indicates that they view personification negatively and caution designers to make informed choices in design: “The reasons for designing a personified interface may be marketing, design, or aesthetic in origin. They are not likely to be based on productivity arguments or ergonomic principles…. In effect, choosing a personified design represents a philosophical statement about the role of machines in human society” (p. 217).

In contrast to views that broadly reject the human model as a foundation for intelligent technologies (Russell & Norvig, 1995), other researchers have embraced the application of social psychology to human-computer interaction. One of the most developed programs of research in this realm has been undertaken by Nass and his colleagues. Nass and Moon (2000) describe humans as mindlessly applying social rules to computers, despite their explicit knowledge that
computers are not humans. Further, they argue that users are aware that social behaviors such as
politeness and reciprocity should not be used in these interactions, even though they are frequently
explanation of the behaviors exhibited by participants in their research, while rejecting
anthropomorphism, experimenter demand, and personification of the computer programmer as
explanations of their findings. For example, Nass and Lee (2001) examined whether listeners
would identify acoustic voice cues in synthetic text-to-speech and be attracted to voices that
seemed similar to their own measured personality characteristics (introvert or extrovert). The
authors interpreted their results as indicating that users prefer synthetic voices consistent with their
own personality and recommended that designers measure user personality and provide matching
synthetic voices in their applications. Other empirical work in this line has been concerned with
interaction between prompt style and interface organization (Vanhoucke, Neeley, Mortati, Sloan,
& Nass, 2001) and foreign accent perception (Dahlback, Swamy, Nass, Arvidsson, & Skageby,
2001).

Although Nass and colleagues rely on Langer’s mindlessness theory as a basis for their
work, it appears to be a limited theoretical foundation for social human-computer interaction. Fiske
and Taylor (1991) assert that mindlessness makes individuals less responsive to their environment
(as compared to mindfulness) and is a cognitive state appropriate for routinized, overlearned tasks.
In contrast, more broadly accepted theories of social cognition, such as schema theory, may offer a
more intuitively reasonable explanation of Nass’ findings (Abelson, 1981; Fiske & Taylor, 1991;

At the present time, the few empirical studies that apply social psychological theory to
speech interfaces have fueled the controversy over the validity of a social approach to speech
technology. On one side of the debate, researchers suggest that social behaviors and cognitions have no relevance to interaction with a speech interface. On the other side, individuals suggest that social-cognitive behaviors are relevant but the theoretical support for this position is somewhat weak. Does the previous literature offer a more robust theoretical foundation for a social-cognitive approach to speech-based e-service design? Indeed it does.

**Social-Cognitive Theoretical Framework**

To date, a social psychology of speech technology and interface design guidelines have emerged in a piecemeal fashion, typically without being informed by the literatures that offer significant theoretical insight into the social-communicative skills of human users. Social psychological findings have had relatively limited application in human factors and human-computer interaction research, except in the areas of workplace or environmental design and team interaction (Sanders & McCormick, 1993; Wickens, Gordon, & Liu, 1999).

Why might social-cognitive psychology be important to the advancement of speech technology theory and interface design? Distinct streams of research in interpersonal communication, service marketing and management, and social-cognitive psychology suggest that speech and language skills are intimately associated with social interaction. Further, humans automatically infer personality traits from speech and language cues contained in any interaction, even one with a speaking and listening technology. The premise that a speaking, interactive technology elicits social behavior and cognition from its human users is based on four streams of research that demonstrate: a) interpersonal communication is goal-directed social behavior, b) humans use speech and language behaviors to obtain social information, c) humans automatically infer personality traits from social information, and d) customer service is a special case of interpersonal communication. I now summarize each of these literatures in turn.
Interpersonal Communication is Goal-Directed Social Behavior

Although some speech technology researchers suggest that goal-directed and social behavior are mutually exclusive, theoretical approaches to interpersonal communication assert that goal-directed communication is social behavior:

“language can also be viewed as a tool, a tool that is used for accomplishing particular ends. To use language is to perform an action, and it is a meaningful action, with consequences for the speaker, hearer, and the conversation of which it is a part. This is a very different view of language. To understand meaning there must be a speaker. And context is critical. What a speaker means with an utterance (what he intends to accomplish) can only be derived with some reference to a context” (Holtgraves, 2002, p. 5).

In this sense, language is the interpersonal means for accomplishing a particular goal in a defined social context, and is inseparable from social behavior. Language use necessarily implies social intent (Austin, 1962; Searle, 1969; Holtgraves, 2002).

Pragmatics may be defined as “rules for how literal meaning can be changed by social context” (Whitney, 1998, p. 34). Pragmatics and speech act theory conceptualizes linguistic meaning as use of language: meaning is derived from the purpose an utterance is put to (Holtgraves, 2002). Krauss and Fussell (1996) describe this perspective as the intentionalist model of language, in which meaning resides in speaker intentions. However, listeners have a distinct role in language usage: they construct the speaker’s meaning and goal through inference. In this sense, then, conversational interactants accomplish their social goals, and derive their partners’ goals, through the use of language. This perspective is based primarily on the theoretical work of philosophers Austin, Searle, and Grice.

Austin’s (1962) speech act theory contradicted the prevailing view of language known as logical positivism, which claimed that all utterances were to be evaluated based on their truth value.
Instead, Austin (1962) defined an utterance as an action or performative (e.g., “I’m sorry”), which he contrasted with a constative utterance, or message that conveys truth (e.g., “I stepped on your foot”). He observed that performatives may or may not be successful (e.g., a hearer may or may not accept an apology), but these utterances do not contain truth value (Holtgraves, 2002). Austin (1962) also defined three dimensions of speech acts: 1) locutionary, or meaning of an utterance (propositional content); 2) illocutionary, or the speaker’s intent; and 3) perlocutionary, or the achieved consequences of an utterance (effect on the hearer). In addition, Austin (1962) described felicity conditions for performing a speech act: these conditions include a conventional procedure executed correctly and completely by an intentional speaker, an appropriate hearer in the appropriate context, and a conventional effect. He indicated that violations of the felicity conditions would result in a misfire or misapplication of the message. Thus, Austin’s contribution included the initial view of language as an action and early specification of the unique roles of the speaker and hearer to communicative interaction, contributions which Searle (1969) elucidated further.

Speech act theorists recognized that there were multiple ways of expressing a single meaning and that the hearer’s recognition of the speaker’s intention occurs through inferencing. Grice (1975) proposed a theory of conversational implicature that was concerned with how communicative partners infer speakers’ intended meanings. He observed that both speakers and hearers abide by a cooperative principle: “make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged” (p. 45). He refined this cooperative principle with four maxims: 1) Quantity – the contribution should be as informative, but not more informative, than necessary, 2) Quality – the contribution should be true (do not say false utterances, or those that lack evidence), 3) Manner – the contribution should be clear, unambiguous, brief, and orderly, and 4) Relation – the
contribution should be relevant.

Grice’s maxims and the cooperative principle are the normative expectations conversational partners rely on during conversation. However, speakers can violate or flout these norms, usually to convey a particular meaning (e.g., metaphor, understatement, overstatement). Speakers then expect the hearer to infer the intended meaning. Grice (1969) provided an example of flouting the Quantity maxim: a professor might write a very short recommendation letter when a student applies for employment. The individual receiving the letter, probably an interviewer, assumes that the professor is being cooperative by agreeing to write the letter (instead of refusing). However, because the letter is short, she may reason that the professor is unlikely to have so little knowledge of her own pupil. Through inference, the interviewer might further conclude that the recommendation includes information that is not explicitly written – that the student should not be considered for employment.

According to speech act theory, language is used to convey meaning; listeners have a vital and collaborative role in creating this meaning. To infer meaning, listeners must assume that speakers are intentional. In addition, they must use the context of a conversation to piece this meaning together. Empirical work confirms that speakers’ speech and language characteristics serve as social cues for listeners (and vice versa) during communicative interaction.

Speech and Language Behaviors are Social Cues

Pragmatic research demonstrates that humans make assumptions about the intentionality of their conversational partners and actively use observed behaviors in an interaction to determine the partner’s goals and meaning. The previous literature has identified speech and language as a major source of social information.

It is well known in the social psychological and communication literatures that linguistic (e.g., lexical choice, language style), nonverbal (e.g., body language, eye gaze, facial expression),
and extralinguistic (e.g., tone of voice, intonation, emphasis, rate of speech) communication variables are sources of information about speakers that listeners use to make inferences about intended meaning and speaker traits (DePaulo, 1992; Holtgraves, 2002). Holtgraves (2002) notes that, in human-human conversation, these cues are context dependent, ambiguous, and variable for any speaker, under even the most ideal interaction conditions. The previous literature has examined the effects of both speech and linguistic behaviors on social perception.

Speech is the physical, acoustic representation of language as conveyed by individual speech sounds (e.g., vowels and consonants), intonation, stress, loudness, pausing, pitch, pitch variability, and other extralinguistic behaviors (Cairns, 1999). In human-human interaction, research has indicated that speech and extralinguistic behavior are some of the mechanisms that speakers manipulate strategically to control their listeners’ impressions of them (Bradac, Cargile, & Hallett, 2002; DePaulo, 1992; Holtgraves, 2002). In addition, listeners automatically infer characteristics of the speaker from these vocal cues, and may even be unaware of the source of their inferences (Bargh & Ferguson, 2000; Holtgraves, 2002; Neumann & Strack, 2000).

DePaulo (1992) reviewed the literature on the role of nonverbal behavior in self-presentation (which can be interpreted to include extralinguistic aspects of speech) and asserted that nonverbal behavior is irrepressible, linked to emotion, and less accessible to actors than observers. Krauss, Chen, and Chawla (1996) showed that nonverbal behaviors minimally assist in conveying a speaker’s message, but instead reveal information about a speaker’s intrapersonal cognitive processes to the listener, such as attitude toward the topic, amount of stress, or current emotional state. Dialectal variants and nonstandard accents also have social consequences, are usually linked to the status (competence, power) and solidarity (friendliness, generosity) of the speaker (Holtgraves, 2002), and can even account for unique variance in listeners’ attributions about a speaker’s behavior (Dixon, Mahoney, & Cocks, 2002). In general,
the voice is used to make judgments of personal characteristics such as gender, age, socioeconomic background, ethnic group, personality, and status (Aronovitch, 1976; Kappas, Hess, & Scherer, 1991; Murray & Arnott, 1993).

Another vein of research has explored the relationship between voice characteristics and emotional effects, or the role of speech in conveying affect (for reviews, see Kappas, Hess & Scherer, 1991; Murray & Arnott, 1993). This research suggests that the voice is intimately associated with the affective state of the speaker, and specific acoustic characteristics are associated with such affective states as indifference, boredom, joy, depression, and stress (Kappas, Hess, & Scherer, 1991). Cosmides (1983) demonstrated that different speakers produce similar acoustic configurations to convey emotion, which listeners use to detect their affective state. Apple and Hecht (1982) showed that listeners generally could recognize anger, sadness, surprise, and happiness in content-filtered recordings with above-chance accuracy, although accuracy depended on the specific emotion being expressed (sadness best recognized, happiness least recognized). In an especially provocative series of experiments, Neumann and Strack (2000) provided an empirical demonstration of mood contagion, in which simply listening to emotional speech automatically induced a similar mood state in listeners, even when they were under cognitive load, unaware that vocal cues may have influenced their mood, and unmotivated to share the speaker’s mood state.

In their study of the language-based aspects of impressions, Wyer, Swan, and Gruenfeld (1995) found that linguistic behavior (question reciprocation, elaboration) was particularly important for impressions of partners in mixed-sex conversations (e.g., liking, aloofness, self-centeredness, interest in partner, liking for partner). Two experiments revealed gender effects, indicating that more positive impressions occurred when males used elaboration and females used question reciprocation. Additional research on normative expectations of communication suggest
that males’ and females’ violations of their expected linguistic style can have positive (e.g., when males use a nurturing style) or negative (e.g., when females use assertive/powerful style) effects on impressions (Burgoon, 1990; Burgoon & Klingle, 1998). Berry, Pennebaker, Mueller, and Hiller (1997) found that categories of emotion words, self-reference, unique lexical items, and verb tense accounted for more variance in listener perception than ratings of nonverbal expressiveness, attractiveness, and facial maturity. Speakers and listeners also use speech and language cues in their partners’ behavior to make attributions about the conversational partner’s knowledge, which is important to establishing mutual knowledge and constructing understandable messages (Bavelas, Coates, & Johnson, 2000; Clark, 1996; Fussell & Krauss, 1992; Krauss & Fussell, 1991).

In general, the theoretical models of social interaction clearly implicate speech and pragmatic language variables as a major source of speaker information during interaction (Bradac, Cargile, & Hallett, 2001; Cargile, Giles, Ryan, & Bradac, 1994; Patterson, 1996, 2001). They further suggest that a speaker’s speech and pragmatic variables influence a hearer’s information processing, responses, affective state, attitudes, attention, and social judgments.

Humans Automatically Infer Personality Traits from a Partner’s Social Cues

Recent researchers have posited that automaticity of frequently used social processes improves cognitive efficiency (Fiske, 1993). Bargh (1989) described goal-dependent automaticity, in which an individual’s ability to identify an activating stimulus and the effects of the stimulus depend on his perceptual goal, implying some level of cognitive control (as cited by Bargh & Ferguson, 2000). He contrasted this type with preconscious automaticity, in which one is not aware of the stimulus or one’s ensuing cognitive processes, and postconscious automaticity, in which one is aware of the stimulus but not its effects on cognitions. Bargh and Ferguson (2000) reviewed the literature on automaticity and asserted “social behaviors in the external environment often if not usually access their corresponding mental representations in an immediate and direct manner,
without conscious and effortful processes of categorization and interpretation being necessary” (p. 929). Perception of social cues results in immediate evaluative judgments of the stimulus (within 250 msec of the cue), and the resulting judgments also affect subsequent social behavior (Bargh & Ferguson, 2000). Thus, current research in automaticity supports the notion that social information is processed rapidly and automatically, often outside the individual’s conscious awareness, and affects subsequent cognition and behavior.

In addition to the notion that social perception and judgment are immediate and beyond conscious awareness, several researchers have shown that these processes may lead to counterintuitive effects. Spontaneous trait inference is the unconscious inference of personality traits when an individual is exposed to behavior (Brown & Bassili, 2002). This form of automaticity is thought to help individuals make sense of each other and promote cognitive efficiency (Fiske & Taylor, 1993). Several researchers have demonstrated that spontaneous trait inference occurs for speakers communicating about others (Skowronski, Carlston, Mae, & Crawford, 1998) as well as inanimate objects (Bassili, 1976; Brown & Bassili, 2002). Thus, following this line of reasoning, because any communication behavior has trait-implying properties, it is likely that users would infer traits from a conversational speech interface.

The interpersonal communication literature has also recognized a perception-behavior link in conversational exchanges, in which automatic inferences about a partner facilitate smooth and successful interaction. The research suggests that conversational partners make rapid attributions about a partner’s knowledge in order to construct appropriate messages (Holtgraves, 2002). Clark (1996) postulated that these attributions are based on heuristics, suggesting that speakers assume that the common ground of a conversation consists of items in the immediate physical environment, information that has been previously introduced in the conversation, and inferences about a partner’s social memberships (e.g., gender, race, ethnicity). A series of experiments has
demonstrated that individuals construct messages differently for self, friends, and strangers in a referential communication task (Fussell & Krauss, 1992; Krauss & Fussell, 1991, 1992), demonstrating that assumptions about a partner do impact subsequent conversational behavior. Krauss and Fussell (1991) suggest that assumptions about mutual knowledge are tenuous, probabilistic hypotheses that are continuously modified throughout interaction and are also subject to systematic errors and biases. Clark and Shaefer (1989) postulated that conversational progress occurs via utterance pairs, or a speaker’s initiation followed by the listener’s acceptance. Although speakers do appear to modify their messages based on knowledge attributions about listeners, listener feedback and co-construction of messages remains an essential part of successful communication (Bavelas, Coates, & Johnson, 2000; Clark, 1996).

Thus, the literature indicates that speech and language information contribute to the automatic evaluations of speakers. In addition, perceptions based on this information subsequently effect interactants’ conversational behavior, leading to successful, smooth communication.

*Service Delivery is a Special Case of Goal-Directed Interpersonal Interaction*

Solomon, Surprenant, Czepiel, and Gutman (1985) used a foundation of social-cognitive theory to argue that service encounters are a specialized form of interpersonal interaction. They pointed out that a service encounter is dyadic and both the customer and provider must adopt appropriate, coordinating roles for it to proceed smoothly. The interactants follow a ritualized script to increase the probability of attaining their goals. Both providers and customers learn these scripts, which form the expectations for each service encounter. The authors argue that customer satisfaction is determined by the amount of congruence between expectations of the service encounter and the actual encounter itself, a definition that has been operationalized as the most common measure of service quality (Parasuraman, Zeithaml, & Berry, 1985, 1988, 1994). A body of empirical work suggests that a service provider’s behavior and personality influence customer
satisfaction, providing support for the idea that provider behavior is an integral part of the service role (Baydoun, Rose, & Emperado, 2001; Cran, 1994; Holland & Baird, 1968; Humphreys, 1996; Mount, Barrick, & Stewart, 1998; Yagil, 2002). Mohr and Bitner (1991) presented a theoretical model of mutual understanding in service encounters, in which several characteristics of the service provider and customer (e.g., background similarity, interaction frequency, script strength, number of subscripts, experience with complementary role, goal compatibility) affect their role taking accuracy, or the degree to which expectations and actual role behaviors are similar. Role taking accuracy and cognitive similarity influence mutual understanding, which in turn influences customer satisfaction.

Scripts associated with service delivery would not be altered simply because a speech technology system replaces a human service provider. Users would continue to assume that the system is in place to support their goals and that the service encounter should proceed smoothly and pleasantly, in a behavioral sequence consistent with a human service encounter. Because speech-based e-service makes use of the most common, natural, and socially-rich means of interaction, it seems reasonable that the expectations of human service providers would also apply to speech user interfaces. However, user perception of e-service encounters and other types of human-computer social interaction has had limited treatment in the empirical literature to date. I now turn to a review of the few studies that have examined the social aspects of human interaction with speech technology.

**Human Users Have Social Cognitions about Interactive Technologies**

The notion that human users might perceive social characteristics in a conversational computer is not new: Turing (1950) proposed a test in which human and computer interactants converse as an evaluative method for the “humanness” of computer partners. The Turing Test has remained a gauge by which technological progress is still measured (Saygin & Cicekli, 2002). A
growing body of current work suggests that social psychological findings about interpersonal speech and language in humans may generalize to human interactions with speech technology.

There is general recognition that synthetic speech should sound humanlike to be perceived more positively by users (Henton, 1999, 2002). Interestingly, much of the research in voice and human emotion has used synthetic speech, because it allows “complete control over every acoustic parameter” (Kappas, Hess, & Scherer, 1991, p. 220) and thus provides better internal validity for experimental work. Although this choice may present some generalization problems relative to spontaneously produced human speech, it strongly supports the applicability of vocal affective research to synthetic speech. In commercial speech applications, concatenative text-to-speech (TTS) is the most commonly used form of synthetic speech (Henton, 2002). This type of speech is generated from units of recorded human speech; thus, even “synthetic” speech in the majority of deployed applications is human speech. Alternatively, speech interface designers may use recorded responses of a professional voice talent as the system’s contributions to a dialogue. In both cases, the use of either type of speech will elicit trait associations about the speaker (system itself).

Consistent with a social-cognitive perspective, listeners’ negative perceptions of a system’s speech may have implications for users’ overall judgment of an interface, even when their performance is unaffected. Gong and Lai (2001) measured the effect of synthetic speech only (TTS) and mixed human and synthetic speech (mixed) prompts on user performance, perception, and attitude. They found that the TTS condition resulted in improved user performance as compared with the mixed-voice condition. However, TTS also resulted in more repetition of prompts and more negative perceptions of performance, clarity, liking, and effort than did mixed voice prompts. Thus, synthetic speech can differentially affect users’ actual performance, perceived performance, and attitude, all of which may affect users’ preference and future use of a
system.

Although the pragmatics of conversation does not appear to have broadly influenced applied practitioners in interface design, a few researchers have used Gricean theory to gain insight into human-computer interactions. Bernsen, Dybkjaer, and Dybkjaer (1996) developed 14 guidelines for man-machine interaction based on analysis of dialogues from a prototype flight reservation system. The authors later observed the similarity between their principles and Grice’s theory, although Bernsen, Dybkjaer, and Dybkjaer (1996) claim to have developed their model independently. They concluded that the Gricean maxims “have been empirically validated for task-oriented, spoken human-machine dialogue…. When performing shared-goal dialogue, people, just like machines, should communicate their communication deficiencies, take background knowledge into account, and initiate repair and clarification metacommunication when needed. Failure to do so detracts, sometimes seriously, from the rationality of the shared-goal dialogue” (pp. 234-235).

In further support of applied pragmatics, Saygin and Cicekli (2002) found that use of Grice’s maxims differentially affected perceptions of a computer conversational partner. They provided participants with 14 conversation excerpts from annual Turing test competitions, in which human interlocutors interact with a human or machine conversational partner. Participants read dialogue transcripts, then rated their agreement with statements based on Gricean maxims (e.g., “A’s contribution to the conversation is irrelevant,” “A provides more or less information than required”), and also rated whether the speaker was a person or computer. Results indicated that violation of the Manner maxim caused participants to perceive the computer partner as more humanlike, while violations of Relation and Quantity revealed that the partner was a computer. Saygin and Cicekli (2002) suggest that empirical work in pragmatics and conversational interfaces is in its early stages and observe that “pragmatics is a crucial component of linguistic
communication. Although this is a widely acknowledged fact among pragmatics researchers, other scientists and writers, notably computer scientists, sometimes tend to underestimate how hard the problem of pragmatic modeling can be” (p. 255). Like Bernsen, Dybkjaer, and Dybkjaer (1996), Saygin and Cicekli (2002) acknowledge that Grice’s maxims may require some modification to account fully for human-computer interaction, but this research indicates the general applicability of pragmatic theory to human-computer interaction.

Social-cognition research provides a broad explanatory foundation for the interpersonal impact of speech-based e-service on human users. It seems reasonable that individuals would hold a script or event schema for customer service encounters that specifies that the service provider should be friendly, helpful, and polite. Individuals probably evaluate their interaction based on this internalized script by comparing the behaviors of the system to the prototypical behaviors expected of a human customer service provider. System characteristics, if they are sufficiently similar to the internal script, will activate the “service” knowledge structure and engage heuristic processing during the encounter. If the speech system responds as expected, the service encounter will proceed smoothly and result in positive affect. However, if the system does not take its expected role, negative affect and controlled processing to understand the discrepancy may result.

The literatures concerned with various aspects of this type of encounter have identified a number of affective variables that may be applied to speech-based e-service. I now review these variables.

*Toward a Social-Cognitive Psychology for Speech-Based e-Service*

Various literatures have implicated a number of social-cognitive variables as important to an understanding of speech-based e-service outcomes for users. To identify the relevant variables for speech-based e-service adequately, a broad review of several literatures (e.g., interpersonal communication, social psychology, service marketing, and human factors) is needed. Although the
literatures are distinct and largely parallel, they converge on the notion of satisfaction as a vitally important outcome of any customer service interaction (Gilmore, 2003) and an outcome of technology use (Gilmore, 2003; Wickens, Gordon, & Liu, 1998). In this light, I summarize the literature on customer satisfaction and its related constructs, its antecedents, and consequences as a basis for a social psychology of speech technology e-service.

*Service Quality*

Perhaps one of the most important, controversial, and widely-researched measures in the services marketing literature is service quality, a construct that has been used interchangeably with customer satisfaction. Service quality is defined as the congruence between customers’ normative expectations of service and their perception of the actual service received (Gilmore, 2003; Parasuraman, Zeithaml, & Berry, 1985). Parasuraman and colleagues developed a SERVQUAL scale, which has been widely adopted as a service measure in industry and has been a centerpiece of the service quality literature until the present time (Gilmore, 2003; Parasuraman, Zeithaml, & Berry, 1994). However, SERVQUAL has been challenged on a number of grounds, including its measurement of customer expectations (Cronin & Taylor, 1992, 1994), use of difference scores, factor structure, and psychometric properties (Gilmore, 2003). Parasuraman and colleagues have continued to refine and evaluate their scale, asserting its psychometric quality but acknowledging that there may be “considerable interdimensional overlap, especially among responsiveness, assurance, and empathy” (Parasuraman, Zeithaml, & Berry, 1994, p. 221). More recently, Zeithaml, Parasuraman, and Malhotra (2000) modified the SERVQUAL measure to create a conceptual framework of e-service quality and identified 11 dimensions of perceived e-service quality.

Although service quality is an important variable in services marketing, it appears to measure specific aspects of service provision by an enterprise, rather than an affective customer
response to service. Therefore, customer satisfaction may be a better operationalization of the customer variable of interest in the present study.

Customer Satisfaction

Customer satisfaction, a construct similar to service quality, has also received extensive attention in the services literature. Wirtz and Lee (2003) indicated that satisfaction, like service quality, is customers’ evaluation of a product or service relative to their expectations. However, the relationship between satisfaction and service quality is ambiguous:

“There has been considerable debate about the concepts of customer satisfaction and service quality. Each has its own research background and theory development. Service quality is defined in the literature as the ability of an organization to meet or exceed customer expectations. Customer expectations have been defined as the desires or wants of consumers or what they feel a service provider should offer. This definition of expectations differs from the way the term is used in the service quality management and the consumer satisfaction literature. Expectations in the satisfaction literature have been operationalized as predictions of service performance while expectations in the service quality literature are interpreted as what the service provider should offer. Different authors use different meanings for the concept of expectations both between and within satisfaction and service quality studies” (Gilmore, 2003, p. 23).

The nature of satisfaction also has come into question. A number of different theoretical and operational definitions have been proposed (Babin & Griffin, 1998; Nicholls, Gilbert, & Roslow, 1998). Wirtz and Lee (2003) reviewed nine commonly used satisfaction measures and found that the distinction between affective and cognitive measures was not supported empirically. Instead, they showed that the highest quality scales (as demonstrated by their high satisfaction loadings, high reliability, and low error variance) comparably measured satisfaction for services
considered primarily hedonic (e.g., ice cream restaurant) and utilitarian (e.g., ATM machine).

Recently, customer satisfaction in call centers has begun to receive increased empirical attention. This research is of particular relevance to speech-based self-service technologies, because such systems often replace human call center operators. Feinberg et al. (2000) examined operational antecedents of customer satisfaction with 514 callers to call centers representing 15 industries. The researchers found that center responsiveness to caller goals (service level) was positively correlated with customer satisfaction, and several variables (speed of answer, goal achievement with first contact, average call abandonment, work time or amount of time to resolve caller problem, and percentage of blocked calls) were negatively correlated with customer satisfaction. However, stepwise regression analysis indicated that only percentage of goal achievement with first contact and average call abandonment rate predicted customer satisfaction, accounting for minimal variance in customer satisfaction ($R^2=0.05$). Bennington, Cummane, and Conn (2000) used stepwise regression to determine predictors for customer satisfaction. In their study of an Australian call center, callers’ quality expectations, rating of staff, and satisfaction with the service center were significant predictors of satisfaction with the call center. In general, callers indicated they were less satisfied with call center service than face-to-face service delivery, and 43% of the sample indicated they would use a different service if it became available. Although the authors suggest that customer loyalty was associated with satisfaction, they did not provide data to support this assertion.

**User Satisfaction**

A similar construct, user satisfaction, has been a measure of affective response to computer technologies in the human factors literature. Nielson (1993) proposed that user satisfaction is one of five factors that define usability: learnability or how rapidly a user learns to use a system, efficiency or the extent to which a system supports user performance, memorability or the extent to
which appropriate system use can be recalled, errors or incorrect actions performed during system use, and the subjective factor of satisfaction. Like customer satisfaction, user satisfaction is typically measured using rating scales that ask users for their overall assessment of the extent to which a system pleased them (Nielsen, 1993; Wickens, Gordon, & Liu, 1999).

The operational definitions of user satisfaction and customer satisfaction appear to have much in common, although they seem to be conceptualized somewhat differently. In services marketing, researchers have asserted that perceived satisfaction is the only relevant construct (Gilmore, 2003). Conversely, Nielsen (1993) conceptualized subjective satisfaction as only one factor in a broader construct of usability, and others suggest that satisfaction (like the remaining usability categories) may range on a continuum from greater to lower priority, depending on the task and user (Wickens, Gordon, & Liu, 1999). In keeping with these theoretical orientations, satisfaction has received extensive theoretical and empirical measurement attention in services marketing (Gilmore, 2003), but the construct has not been consistently defined and measured in the human factors literature (Gray & Salzman, 1998a, 1998b; Olson & Moran, 1998). In general, satisfaction has cross-disciplinary support as a central variable in various types of interaction, making it clearly relevant to the affective outcomes of speech user interfaces in e-service.

Antecedents of Customer Satisfaction

In the scant but emerging literature, a number of variables have been identified that impact customer satisfaction with self-service technology. The bulk of this literature is largely based on research efforts in the services marketing and management literatures, as well as their newest incarnation, e-service (Rust & Kannan, 2002). From a marketing perspective, understanding of the user is vital to the effective positioning of new goods and services as they enter the marketplace. A less developed stream of research is concerned with characteristics of an e-service interface that influences users’ subsequent attitudes and behaviors. Although most of this research has been
conducted with visual interfaces (e.g., Internet webpages), previous empirical efforts illuminate a path for research in speech or auditory interface design as well.

**Individual Differences of Users.** Individual difference characteristics that impact speech technology use have not been systematically explored by researchers to date. However, previous research efforts point to a number of personality variables that may impact use of speech interfaces.

One of the most informative empirical sources of personality variables is e-service research. Dabholkar and Bagozzi (2002) present and evaluate an attitudinal model of self-service technology in which consumer traits (individual differences) and situational factors, including perceived waiting time and social anxiety, serve as moderators. They argue that demographic variables (e.g., gender, age, education, socioeconomic status) are the least relevant consumer traits because most individuals are familiar with basic technologies regardless of group membership. In contrast, personality variables such as self-efficacy, inherent novelty seeking, need for interaction with a service employee, and self-consciousness are “of greater interest than demographic or psychographic factors because such variation is at the heart of consumer attitude formation and behavior intentions” (p. 187). Results of their structural equation modeling indicated that self-efficacy, inherent novelty seeking, and self-consciousness moderated relationships between ease of use, fun, performance, and participant attitude.

Other researchers have also examined technology-based personality variables. Colby (2002) describes technology readiness as “the propensity of an individual to adopt and embrace new technology for personal use or at work. The technology in question would be defined as anything that is (1) cutting-edge and (2) removes a significant part of the human element from a product or service it replaces” (p. 27). Parasuraman (2000) developed the Technology Readiness Index (TRI), a 36-item scale that measures two drivers (Optimism, Innovativeness) and two inhibitors (Insecurity, Discomfort) of technology adoption. Subsequent research with the scale
demonstrated that its scores effectively differentiated consumers groups (Colby, 2002) and their rates of ownership of technology-based products and services, usage rates of technology-based services, and perceived desirability of future technology-based services (Parasuraman, 2000). However, this research was mostly concerned with Internet and home-based technologies (e.g., cable television, answering machine, caller ID, cellular phone, desktop computer) and did not involve speech-based e-service interfaces.

Beyond these user characteristics thought to relate specifically to technology use, Holtgraves (2002) argues that personality may also influence individuals’ ability to take the perspective of a communicative partner into account, which influences the success of a conversation. Because speech interface use mimics human-human conversation, a user’s skill in picking up subtle social cues provided by the system’s intonation, emphasis, or lexical choices may influence how he or she responds to the system. In this sense, users’ social skills may also differentially impact how an individual perceives an interactive system. A well-studied variable known as self-monitoring refers to the tendency of individuals to respond to their social environment and plan their own behavior accordingly (Snyder & Cantor, 1980). High self-monitors respond to social norms and have been found to adapt quickly to social situations, communicate a wider range of emotions, initiate conversation, seek out attractive partners, and point to situational factors to explain their own behavior more than low self-monitors (Fiske & Taylor, 1993; Snyder, 1974, 1979).

More recently, the consumer marketing field has started to investigate the role of self-monitoring in responses to product brands, advertising, and other forms of consumptive behavior (Allen, 2002; Auty & Elliott, 1999; Czellar, 2003; DeBono, Leavitt, & Backus, 2003; O’Cass, 2000; Snyder & Gangestad, 1986). Lehman and Winer (1997) argue that a general personality approach has relevance for market analysis and planning: “Given the relatively limited
predictive power of demographic and socioeconomic variables, the fact that many people in marketing are trained in psychology, and the natural desire to find a general basis for segmenting consumers that will be useful over many situations, it is not surprising that an attempt has been made to use personality traits as a basis for [market] segmentation” (p. 101).

As Snyder and Gangestad (1986) point out, high self monitors are “highly responsive to social and interpersonal cues of situationally appropriate performances” (p. 125). Applying this characteristic to interaction with a speech interface, high self monitors may be more sensitive than low self monitors to an interface designed to imitate prototypical service behavior. In addition, because self-monitoring predicts social behavior, high self monitors might be more successful with a speech interface because they quickly adapt their interpersonal behavior to the interface’s interaction style. Thus, self-monitoring may impact perceptions of a speech interface and the enterprise for which it provides service, and even be associated with specific user behaviors with the interface.

*Point of View.* A common current practice in speech interface design is to provide audio taped samples of a speech interface before it is designed, to give stakeholders and other observers a better understanding of how the final system will sound to the intended user group (e.g., an audio clip of a simulated interaction may be played for an executive who is interested in implementing speech technology). This practice may be problematic because observers may not perceive a speech interface in the same way users do. This issue has not been addressed in applied work. However, the previous psychological literature indicates that conversational interactants and listeners observers differ in their perceptions of, cognitions about, and memory for conversational interactions, as shown by a number of studies that use interactants and observers as a between-subjects independent variable.

Conversational interactants (speakers) and observers may perceive incoming information
about target persons differently. Burgoon and Newton (1991) showed only limited agreement between participants and observers in judgments of nonverbal cues. Lassiter, Geers, and Apple (2002) examined perception by individuals expecting to transmit or receive information in four studies using a behavior unitization technique. Unitization is a method of parsing observed behavior into units of action. Variations in rate, pattern or both indicate perceptual effects. The researchers demonstrated that individuals anticipating a speaker role parsed observed behavior into fewer units than individuals without this expectation. Further, speakers also showed decreased confidence in their trait impressions of targets, poorer memory, and decreased affect as compared with control and observer groups. Lassiter, Geers, and Apple (2002) explained that potential speakers adopt a processing strategy that allows a clear, concise summary of information for transmission to others. Indeed, a variety of studies have suggested that conversational interactants’ cognitions (Burgoon & Newton, 1991; Gilovich, 1987; Guerin & Innes, 1989; Inman, Reichle & Baron, 1993; Lerner & Tetlock, 1999; Monahan, 1995) and memory (Benoit, Benoit & Wilke, 1996; Frank & Gilovich, 1989; Miller & deWinstanley, 2002) are different than observers’. However, observers may have a more accurate perception of a speech-based e-service system because they have more cognitive resources with which to make the required social judgments than do users (Patterson, 1996).

In service research, this issue has not been recognized as a potential methodological problem. Researchers have assumed that observer judgments are similar to those of actual consumers and have used written narrative scenarios to evaluate models of e-service attitudes (Dabholkar, 1994, 1996; Dabholkar & Bagozzi, 2002). Similarly, in previous speech interface research, audiotaped user-system interactions have been used as stimuli (Polkosky, 2003). The use of an audiotaped, simulated interaction would appear to be more similar to an actual interaction with a speech system than a written narrative because listeners hear the actual voice of the system,
its prompts, and a user’s response to those prompts. In interpersonal communication studies, researchers have also used audiotaped telephone conversations and other indirect, non-interacting observational methods of evaluating partners in conversation (Cargile, Giles, Ryan, & Bradac, 1994; Kreuz & Roberts, 1993; Sacks, Schlegoff, & Jefferson, 1974; Schlegoff, 1968; Schlegoff, Jefferson, & Sacks, 1977; Schlegoff & Sacks, 1973; Trafimow & Wyer, 1983; White & Carlston, 1983). Although an observer method has been used in the disciplines that inform the present study, the impact of point of view warrants empirical attention.

**Personality and Behavior of Service Providers.** Speech interfaces usually replace human service providers for constrained or repetitive interactions. Because consumers have expectations about the service provider role and these expectations have implications for the success of the interaction (Solomon, Surprenant, Czepiel, & Gutman, 1985) and customer satisfaction (Parasuraman, Zeithamal, & Berry, 1985, 1988, 1994), findings on desirable personality and behavior characteristics of service providers provide a model for designing speech user interfaces. In the speech industry, finding the correct “persona” of an interface has been one of the most important design issues in the adoption of speech technology by enterprises (Cohen, Giangola & Balogh, 2004; Kotelly, 2002). Indeed, the literature suggests that discrepancies between users’ internalized script for customer service (including the service provider’s behavior) and their actual interaction may influence cognitive processing and affective responses.

Past research has analyzed aspects of human service provider personality and behavior that subsequently influence customers’ attitudes and behaviors. Thus, with human service providers, employee selection is a critical factor in customer satisfaction with service delivery. In their meta-analysis of performance in jobs requiring interpersonal interaction, Mount, Barrick, and Stewart (1998) found that Conscientiousness was most strongly related to performance for dyadic service providers \( (r = .29) \), followed by Openness to Experience \( (r = .17) \), Agreeableness \( (r = .13) \),
Emotional Stability ($r=.12$), and Extraversion ($r=.07$) when these personality traits were measured by the Personal Characteristics Inventory, a measure of the dimensions of the Five-Factor Model of personality.

In contrast to broad measures of personality, other researchers have more narrowly defined variables thought to impact customer satisfaction. Cran (1994) defined customer service orientation as a “set of basic individual predispositions and an inclination to provide service, to be courteous and helpful in dealing with customers and associates” (p. 36). Interpersonal skills, or a service provider’s “acquired ability for effective interaction” (Holland & Baird, 1968, p. 503) have been shown to impact customer satisfaction positively (Humphreys, 1996; Solomon, Surprenant, Czepiel, & Gutman, 1985). To investigate this variable further, Alge et al. (2002) measured personality characteristics and performance of 115 bus operators in a metropolitan transit authority. They found that interpersonal skills were significantly correlated with extraversion ($r=.65$), general disposition ($r=.39$), and self-reported performance ($r=.31$), but were not associated with supervisor ratings or other performance indicators. Despite relatively weak empirical support for the relationship between interpersonal competence and performance, interpersonal skills may be more strongly associated with customer attitude and performance in jobs that require interaction as a primary activity. Baydoun, Rose, and Emperado (2001) note that the general role and expected behavior is relatively clear: “In general, the existing measures of customer service orientation include similar subscales (e.g., interpersonal skills such as friendliness, politeness, and helpfulness) and most tend to emphasize an individual’s overall ‘energy’ level” (p. 618).

Yagil (2001) examined service providers’ use of ingratiation and assertiveness (influence tactics) and found that ingratiation increased customer satisfaction but assertiveness decreased satisfaction. The study was concerned with repeated service encounters, or service relationships, in 115 customer-provider dyads in several industries (e.g., banking, insurance, government). Yagil
(2002) speculated that the relationship between service providers’ influence tactics and customer satisfaction was mediated by customers’ self-esteem: “A service provider’s assertive behavior might be interpreted by the customer as reflecting a lack of respect; it may convey a degradation of the customer’s status and thus lead to a general sense of dissatisfaction with the service. On the other hand, the ingratiatory behavior of the service provider, which is deliberately designed to please the customer, grants the customer a respectable status, enhances his or her self-esteem, and consequently results in satisfaction” (p. 350). This interpretation is consistent with a broad literature on power and politeness, which are conveyed linguistically (Brown & Levinson, 1987; Holtgraves, 2002).

Price, Arnould, and Tierney (1995) proposed a framework that describes service encounters in terms of their duration, affective content, and spatial proximity. The authors found that these three predictors influenced the service providers’ behavior, as well as customers’ affective response and service satisfaction. However, the specific type of service encounter Price, Arnould, and Tierney (1995) studied was not the type of encounter that would be appropriate for speech technology or self-service in general (e.g., extended, affectively charged tour guide-traveler relationships). Despite this, the variables they identify show some similarity to constructs in other domains, most notably the emphasis on affective variables throughout the model and the use of interpersonal distance (proxemics) as a moderator of the service provider-customer interaction.

Mattila and Enz (2002) used Price, Arnould, and Tierney’s (1995) framework to evaluate affective response to and customer evaluation of service delivery in brief encounters at hotel front desks. They found that ratings of displayed emotions were positively correlated with customers’ evaluation of the service encounter and their overall assessment of the hotel. In addition, whereas demographic (ethnicity, gender) and duration of the interaction failed to add incremental variance in a hierarchical regression analysis, customer mood and rating of displayed emotions did account
for unique variance in the service encounter evaluation and overall hotel evaluation. This work
draws from a research base suggesting the influential role of mood in service encounters, and
further indicates the relevance of mood to even brief, everyday service interactions.

**Trust.** In a recent review of the literature, Corritore, Kracher, and Weidenbeck (2003)
define trust in an online environment as “an attitude of confident expectation in an online situation
of risk that one’s vulnerabilities will not be exploited” (p. 740). Trust functions as a cognitive
heuristic to simplify complex, uncertain environments (Grabner-Krauter & Kaluscha, 2003). This
multi-dimensional construct occurs in the presence of several common conditions (e.g., risk,
vulnerability, expectation, confidence, and exploitation), especially in informational or
transactional websites. The online environment is similar to more typical human social situations
in which individuals exhibit trust, including interpersonal interaction (Corritore, Kracher, &
Weidenbeck, 2003). Researchers identify trust as a complex, multifaceted construct that has been
operationally defined in numerous ways (Corritore, Kracher, & Weidenbeck, 2003;

Corritore, Krasher, and Wiedenbeck’s (2003) online trust model does not specify the
measurement of the various constructs and is not empirically validated. Their definition of trust
appears to overlap at least partially with Grice’s (1969) maxims, particularly the Quality and
Relation maxims. These two maxims require that conversational partners provide only true and
relevant information. Trustworthiness is also conveyed in interpersonal interaction through speech
and vocal cues (Polkosky & Lewis, 2003), which points to at least one aspect of a speech system’s
voice that may impact a user’s trust in the system.

**Comfort.** Spake, Beatty, Brockman, and Crutchfield (2003) argued that the psychological
construct of comfort, or “a psychological state wherein a customer’s anxiety concerning a service
has been eased, and he or she enjoys peace of mind and is calm and worry free concerning service
encounters with this provider” (p. 317), aids in the selection, development, and maintenance of service relationships. They developed an eight-item 10-point semantic differential scale and demonstrated that comfort is positively correlated with satisfaction, trust, commitment, and ease of complaining when problems occur. They also evaluated three structural models of affective variables and demonstrated that a model with comfort provided the best fit and explained more variance than alternate models, adding incremental understanding to the more common constructs studied in the marketing literature.

In addition to the empirical support for comfort, the construct appears to be a specific example of Austin’s (1962) perlocutionary dimension. According to Austin, a speech act consists of meaning, speaker intention, and a consequence on the hearer (perlocutionary dimension). In a service encounter, comfort may be viewed as the desired psychological effect of service on a customer. When service is provided by a speech interface, it seems reasonable that comfort may also be the desired effect on the user. In both cases, Spake, Beatty, Brockman and Crutchfield’s (2003) study suggests that increased comfort will be associated with increased customer satisfaction.

Ease of Use and Usability. Several researchers, especially in the service marketing literature, have defined constructs relating to the usability of a self-service technology and incorporated them into models of service delivery (Carritore, Krasher, & Wiedenbeck, 2003; Dabholkar & Bagozzi, 2002). Dabholkar (1994) developed a six-item measure of ease of use and four-item measure of performance, which tapped consumers’ perceptions of how complicated, confusing, effortful, time-consuming, and reliable a self-service technology was. The scales were again used in Dabholkar and Bagozzi (2002). Zeithaml, Parasuraman, and Malhotra (2002) identified this construct using two dimensions of e-service quality, which they referred to as Ease of Navigation and Efficiency. In contrast, Nielsen (1993) argued that usability is composed of
several components, including ease of learning, efficiency, ease of recalling the interface, and few user errors. Wickens, Gordon, and Liu (1998) assert that “usability is… one of the greatest concerns for those working on software interface design” (p. 453). Even in the human factors literature, definitions and measures of usability have historically been problematic and controversial (Bevan, 1995; Gray & Salzman, 1992, 1998a, 1998b; Hartson, Andre, & Williges, 2001; Hassenzahl, 2001; Hertzum & Jacobsen, 2001; Olson & Moran, 1998) and a universally acceptable definition and measure of usability continue to be elusive.

The concepts of ease of use and usability bear some similarity to the Gricean (1975) maxim that contributions should be brief, orderly, and to the point. They are also similar to Bernsen, Dybkjaer, and Dybkjaer's (1999) cooperative principles to “avoid superfluous or redundant interactions” (p. 122) and be clear and brief. A general ease or cooperative construct appears to have broad theoretical support across disciplines.

Consequences of Customer Satisfaction

The consequences of customer satisfaction and perceived service quality have been investigated in the previous literature. They are positively associated with behaviors such as repeat use of self-service technologies and intent to purchase (Gilmore, 2003).

An important consequence of customer satisfaction is customer loyalty or “the behavioral outcome of a customer’s preference for a particular brand or a selection of similar brands, over a period of time” (Gilmore, 2003, p. 24). Most researchers have conceptualized loyalty as a developmental process that evolves from cognitive loyalty to affective loyalty and finally to behavioral intention and action-based loyalty (Gilmore, 2003). Rust and Kannan (2002) point out that, although the e-service environment may hinder customer loyalty due to the ease of comparison shopping and lack of human interaction, it is these features that may also permit greater customer loyalty when they are part of an appropriate customer service strategy. They assert that
customer equity “manifests itself in e-loyalty outcomes, both attitudinal (repeat purchase intention, willingness to share information, and positive word of mouth) and behavioral (repeat-purchase behavior, frequency of visits, cumulative sales, and e-tailer share of customer spending)” (pp. 14-15). Harrison-Walker (2000) has also recently highlighted word of mouth communication as a distinct variable that influences customer expectations, attitudes, and behaviors. A positive association between service quality and word of mouth has been demonstrated empirically (Boulding, Kalra, Staelin, & Zeithaml, 1993).

Summary of Potential Speech-Based e-Service Constructs

Taken together, several literatures suggest that customer satisfaction is one of the major affective outcomes of interaction with a speech-based e-service. There are several important but uninvestigated antecedents that may influence a user’s satisfaction when customer service is provided by a speech interface, including perceptions of the interface derived from script-based knowledge of human service and conversation, individual differences, and point of view. Customer satisfaction and other affective responses to this modern customer service experience may influence longer-term attitudes and behaviors toward businesses, as well as have implications for the behavior of customers and other potential customers with whom they interact. Although many variables are implicated in this emerging field, the empirical work has been extremely limited, creating a need for research to further elucidate the interrelationships among the relevant constructs.

Overview of the Current Study

The purpose of the current research was to begin to examine the relationships among user characteristics and perceptual and affective outcomes of speech-based e-service. A review of the literature indicates that several disciplines have been concerned with aspects of speech user interfaces when they are used to provide customer service. However, these literatures have rarely
interacted, have generated a broad range of theoretical variables, and disparate findings have not
been systematically applied to speech technology. Despite discipline-specific definitions and
empirical treatment, many of the implicated fields are concerned with similar constructs, which
makes them potential candidates for investigation in applied speech technology.

Therefore, the current research uses two studies to address three main categories of
antecedents to customer satisfaction in speech-based e-service. In the first study, expert judgments
are used to assess the quality of several speech interfaces and select stimuli for the second study.
The second study uses student observers to collect data on, individual difference variables (need for
interaction with a service provider, inherent novelty seeking, age, gender, and self-monitoring).
These variables were selected for further investigation because they each have been implicated by
several studies and even across disciplines that inform the present research. In addition, a broad set
of interface antecedents are measured, based on the literature concerned with behavior and
personality of a service provider, trust, and ease of use. These variables were evaluated as potential
items for a usability scale for speech interfaces, with usability broadly defined as a multi-factor
construct (Nielsen, 1993). Finally, comfort and customer satisfaction also are measured in the
second study to determine their relationships to the individual and interface perceptual variables.

Little is known about the relationships between individual differences, interface usability,
comfort, and customer satisfaction with automated systems. It is generally acknowledged that
e-service interfaces may have broad implications for customers’ overall perceptions of service
from an organization (Ostrom, Bitner, & Meuter, 2002; Rust & Kannan, 2002). However, most of
this research has been concerned with service delivery via the Internet. Therefore, the current study
identified a subset of perceptual and affective interface variables that are related to and predict
customer satisfaction for speech technology interactions.
Rationale

There were several practical, theoretical, and methodological reasons for conducting the current studies. From a practical perspective, an understanding of the speech and linguistic perceptual variables that impact customers’ social cognitions and affect during an e-service encounter helps designers predict the success of their interfaces before they are commercially deployed. Such knowledge is likely to promote more usable and satisfactory designs in the emerging and highly competitive field of speech technology. In addition, by predicting customer satisfaction from interface characteristics, enterprises that use a speech user interface for their customer service may have clearer expectations about the impact of this technology on their customers.

On a theoretical level, this research begins to integrate separate literatures that are simultaneously addressing similar issues from different perspectives. Currently, the interpersonal communication, e-service, and speech technology literatures are advancing in parallel, with relatively little interaction among them. However, e-service researchers do acknowledge the importance of speech technology as a mechanism for future e-service delivery (Rust & Kannan, 2002). Similarly, the findings from interpersonal communication and social-cognitive psychology have had limited penetration into the applied fields of speech technology, e-service, and human-computer interaction. Thus, the current studies begin to integrate a range of variables in a meaningful way. Brennen (1998) asserts

“that [the] domain of human-computer interaction is a particularly relevant application for social and cognitive psychologists who study psycholinguistics for two reasons. First, experimental research has illuminated general principles about processing, representation, and interaction that can be applied directly to explaining, predicting, and improving human-computer interaction. Without such underlying principles, progress in interface
design will be ad hoc at best, especially for multimodal, ‘intelligent’ systems that use speech and language. At the same time, human-computer interaction provides an ideal testbed for demonstrating and testing models and principles… Transporting models from social and cognitive psychology to electronic communication and embodying such models in software has the potential to bring additional clarity and pragmatism to these fields” (p. 222).

Preliminary cross-fertilization of these disparate literatures will help promote a more unified perspective, which may sharpen theoretical perspectives and encourage transdisciplinary applied research.

Finally, this research also presents several methodological advances in using correlation and regression for analyzing human-computer interaction via speech. Although such analyses have been effectively utilized in education, industrial-organizational psychology, and other applied fields to address social-cognitive questions (Pedhazur, 1997), they have had relatively limited use in human factors or interpersonal communication research. This restricted range of methods applied to communication research has left relevant questions unanswered by the empirical literature. Thus, the current research uses variables and statistical analyses borrowed from other applied fields to promote a stronger methodological approach to theory-building in speech technology.

Finally, this research offers the potential for another methodological advance: the development of a scale for assessing the usability of speech user interfaces. No empirically-validated scales currently exist for speech interface evaluation, which presents a significant measurement problem that limits both theory development and practical understanding of these technologies.
Research Questions

Based on the previous literature, the primary issue of interest in the current studies is the interrelationships among individual difference, usability, comfort and customer satisfaction. The present research focuses on several individual difference variables: self-monitoring, need for interaction with a service employee, and inherent novelty seeking. These variables have been identified in the services marketing and psychological literatures as influencing individuals’ decisions about product and service consumption. In addition, a number of interface perceptual variables were considered, including aspects of voice and speech production, pragmatics, orientation toward users’ goals, customer service behaviors, and trust of speech interfaces. Finally, customer satisfaction and comfort, a psychological variable that is defined similar to perlocutionary force in the pragmatics literature, were also considered. Extending analyses targeting associations among these variables, the potential interaction between the individual difference and perceived usability variables was also analyzed using analysis of covariance.

The second major issue of interest is the prediction of affective outcomes such as customer satisfaction and comfort from perceived usability variables. Because customer satisfaction has been a central variable in the customer service literature, it seems that this variable will continue to be a practical focus as service increasingly is provided by technology.

The previous literature does not provide adequate background to allow hypotheses generation for the current research. Therefore, several research questions were developed:

Question 1: What items should be measured in assessing the usability of speech interfaces?

Question 2: How are individual difference variables (e.g., demographics, self-monitoring, need for interaction, inherent novelty seeking) related to comfort and customer satisfaction?

Question 3: What speech user interface characteristics best predict customer satisfaction?

Question 4: What speech user interface characteristics best predict comfort?
Question 5: Do individual difference variables interact with speech user interface characteristics to create customer satisfaction?

Study 1: Speech Interface Expert Ratings

The first study used experts in speech and language to evaluate the quality of several speech user interfaces and prepare stimuli for Study 2. The use of expert judges with background in psycholinguistics is due to several issues in the speech technology field. First, there is some acknowledgment in industrial settings that experts in communication behavior have a necessary set of skills for the design of effective speech interfaces. Second, engineers who have extensive knowledge of the underlying technology may hold biased perceptions of a speech interface, due to their own personality characteristics (Capretz, 2003), a lack of specific knowledge about human communication, or potential biases related to the constraints imposed by the languages used to code these interfaces (e.g., VoiceXML). Third, there is no universally accepted or empirically validated measure of usability that specifically applies to speech technologies. In a pilot study, two scales were developed and used to evaluate the usability of speech interfaces, but both scales demonstrated several weaknesses, specifically, limitations in scope and small effect sizes (Polkosky, 2003). Finally, a variety of guidelines exist for interface development (Balentine & Morgan, 1999; Cohen, Giangola, & Balogh, 2004; Kotelly, 2003), which may promote variability in quality across speech applications.

The first study uses expert judgments of several speech interface-user interactions to determine their relative quality. Based on their ratings of speech and language characteristics, six interfaces representing the best, average, and worst quality interfaces will be selected for use as stimuli in Study 2.
Method

Participants

The participant sample consisted of six subject-matter experts in human speech and language, recruited from the Psychology and Communication Sciences and Disorders programs at the University of South Florida. The participants included two Ph.D. level psycholinguistics researchers, three Ph.D. level speech science or speech-language pathology researchers, and one Ph.D. psychology student who is also a speech-language pathologist. None had specific expertise in speech technology or speech interface design. Although invitations were emailed to a known expert sample, data were collected by a web survey tool that does not allow identification of participants. Therefore, the data could not be specifically matched to their provider.

Measures

Fifteen items were used to evaluate each speech interface based on the review of literature and speech and language characteristics that appeared to vary among the interfaces. The items included: politeness, appropriate use of indirect vs. direct requests, use of pausing, use of emphasis, voice naturalness, use of wide intonation variation, effective response to communicative breakdowns, contingent, specific responses to user, pragmatic appropriateness, consistency with conversational expectations, deference to user, liking of interface, ease of conversation with this interface, overall quality of interface, and similarity to human-human conversation. Each item was presented in random order for each stimulus and for each participant. Participants rated each item on a five-point scale, ranging from 1 (poor) to 5 (excellent).
Table 2. Speech Interface Stimuli

<table>
<thead>
<tr>
<th>Type of System-User Interaction</th>
<th>Length of Interaction (min:sec)</th>
<th>System/Caller Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis Scoreboard</td>
<td>0:38</td>
<td>Male/male</td>
</tr>
<tr>
<td>Available Flights</td>
<td>2:27</td>
<td>Female/female</td>
</tr>
<tr>
<td>Retail Order</td>
<td>1:59</td>
<td>Male/male</td>
</tr>
<tr>
<td>Retail Purchase</td>
<td>2:09</td>
<td>Male/female</td>
</tr>
<tr>
<td>Credit Dispute</td>
<td>1:42</td>
<td>Male/male</td>
</tr>
<tr>
<td>Package Delivery</td>
<td>2:26</td>
<td>Female/male</td>
</tr>
<tr>
<td>Flight Status</td>
<td>2:36</td>
<td>Male/male</td>
</tr>
<tr>
<td>Phone Bill</td>
<td>3:41</td>
<td>Female/male</td>
</tr>
<tr>
<td>Car Rental</td>
<td>2:36</td>
<td>Male/female</td>
</tr>
<tr>
<td>Directory Dialer</td>
<td>0:34</td>
<td>Male/female</td>
</tr>
<tr>
<td>Patient Claim</td>
<td>1:14</td>
<td>Female/male</td>
</tr>
<tr>
<td>Movies</td>
<td>3:22</td>
<td>Female/male</td>
</tr>
<tr>
<td>Report Electrical Outage</td>
<td>0:44</td>
<td>Male/male</td>
</tr>
<tr>
<td>Prescription Refill</td>
<td>0:38</td>
<td>Female/male</td>
</tr>
<tr>
<td>Financial Services</td>
<td>3:48</td>
<td>Female/male</td>
</tr>
<tr>
<td>Trade Stock</td>
<td>0:51</td>
<td>Male/male</td>
</tr>
</tbody>
</table>

*Stimuli*

The stimuli consisted of 16 digitally-recorded audio files of speech interface-user interactions created at IBM and its competitors (Scansoft, Nuance). The interfaces were selected to
represent a diverse range of development stages (e.g., prototype and commercially deployed systems), durations of interactions, system voices, and industries. Table 2 summarizes the general characteristics of the audio clips; complete transcripts of each user-system interaction appear in Appendix A. The stimuli were presented in two orders to control for order effects.

Audio files of simulated telephone calls by a hypothesized user to a speech system (known as vision clips because they provide a “vision” of what the completed system will sound like) to demonstrate a speech user interface are frequently used to sell speech technology solutions or to validate an interface design with its financial sponsors in the early stages of development. Vision clips are typically created by editing together recordings of a professional voice presenting prompts with the simulated responses of a recorded “caller.” This technique allows stakeholders to hear a sample system-user dialogue before a speech system is actually developed. To the observer, vision clips cannot be differentiated from recordings of an actual interaction between a real user and a deployed system because both depict user-system dialogue. Because vision clips have a very prominent role in the speech technology industry and have received minimal empirical attention, the current studies used vision clips as a source of stimuli, in addition to actual recordings of user calls to working systems.

The stimuli for each participant consisted of a series of 16 recordings, each depicting a single telephone call to a synthetic speech system. Participants heard the operator’s prompts and a caller’s responses as if they were listening to the call on another telephone line.

Procedure

Participants were recruited from the Psychology and Communication Sciences and Disorders Departments at the University of South Florida. Each participant received an email invitation, which briefly explained the study and provided a link to a webpage. The participants were randomly divided into two groups to receive one of two webpages (each webpage presented
the stimuli in different orders). Participants clicked on the link in the email and then used an Internet browser to access the audio stimuli and questionnaire items. After completing the rating items, participants’ data were automatically uploaded to an IBM-based server using a web survey and data collection tool (WebSurveyor 4.1).

Results

A statistical software package (SPSS 11.5.1) was used for all analyses. The purpose of Study 1 was to obtain expert quality ratings of the 16 potential stimuli for Study 2. Analyses consisted of descriptive statistics (e.g., central tendency, dispersion, type of distribution) and a reliability analysis on the composite mean of the 15 items. 

Descriptive statistics

Visual inspection of the boxplots for each rating item indicated normal distributions for two items (Quality, Liking) and approximately normal, slightly skewed distributions for 12 items (Directness, Naturalness, Deference, Politeness, Contingency, Pragmatics, Intonation, Humanlike Quality, Consistency, Emphasis, Pausing, Ease of Use). The exception to this observation occurred for the Communication Breakdown item, which was characterized by minimal variability and neutral ratings ($M = 2.99$, $SD = .79$, $95\% CI = 2.82-3.16$). Participant comments consistently referred to the lack of breakdowns shown in the stimuli, further suggesting inadequate variability in the stimuli for this item. The items ratings were used to calculate the descriptive analyses for each interface.

The composite means and standard deviations for each interface are shown in Table 3. Overall means for each interface suggest that the Financial Services interface was most positively rated and Directory Dialer was most negatively rated.

Reliability

Reliability of the expert ratings was analyzed using all items and composite means with a
two-way random effects model intraclass correlation coefficient (Shrout & Fleiss, 1979). This model is used when judges are considered a random effect and belong to a larger population of judges (Shrout & Fleiss, 1979). The results indicated that the single rater intraclass correlation coefficient was poor, $ICC (2, 1) = .37$, suggesting that individual raters had minimal reliability.

Table 3. Experts’ Overall Ratings of Quality for 16 Speech Interfaces

<table>
<thead>
<tr>
<th>Type of System-User</th>
<th>$M$</th>
<th>$SD$</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Services</td>
<td>4.22</td>
<td>0.83</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Prescription Refill</td>
<td>3.86</td>
<td>0.83</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Credit Dispute</td>
<td>3.79</td>
<td>0.92</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Retail Order</td>
<td>3.78</td>
<td>0.92</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Retail Purchase</td>
<td>3.73</td>
<td>0.81</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Trade Stock</td>
<td>3.70</td>
<td>0.99</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Patient Claim</td>
<td>3.55</td>
<td>0.69</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Report Electrical Outage</td>
<td>3.39</td>
<td>1.11</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Tennis Scoreboard</td>
<td>3.37</td>
<td>0.88</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Movies</td>
<td>2.99</td>
<td>1.17</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Package Delivery</td>
<td>2.94</td>
<td>0.69</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Available Flights</td>
<td>2.79</td>
<td>0.90</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Phone Bill</td>
<td>2.78</td>
<td>0.86</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Car Rental</td>
<td>2.72</td>
<td>0.97</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Flight Status</td>
<td>2.66</td>
<td>0.86</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Directory Dialer</td>
<td>2.40</td>
<td>1.00</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
across items. However, the average rater statistic was acceptable, ICC (2,6)=.78, suggesting that group means were more reliable than individual ratings. As a result, only group mean ratings (as shown in Table 3) were considered in selecting Study 2 stimuli.

Stimuli Selection

The goal of Study 1 was to evaluate the quality of a sample of speech user interfaces according to expert judges and identify stimuli for Study 2. Although the data showed poor reliability within individuals, the mean group ratings were notably more reliable. In addition, overall mean ratings for each interface showed that the experts’ perception of interface quality could be quantified along a positive to negative continuum. Therefore, six interfaces were selected for inclusion in Study 2 to represent the best (Financial Services, Prescription Refill), worst (Flight Status, Directory Dialer), and average (Tennis Scoreboard, Movies) interfaces.

Study 2: Observers’ Affective Responses to Speech Interfaces

As the main focus of the current research, Study 2 is concerned with observers’ perceptual and affective responses to interactive speech interfaces. In applied industrial settings, observers (such as enterprise decision-makers and other stakeholders) often judge an interface and determine the merits of the design based on their perceptions of a recorded system-user dialogue. Therefore, this study generated data using a common “observer” approach to interface evaluation in the speech technology industry, increasing its applicability and importance for human factors work in this field.

The study employed a correlational design, using observers who listened to audiotaped recordings of user-system dialogues and complete rating scales for a range of affective and perceptual variables. Students were recruited to listen to one of six user-speech system interactions, then provided individual difference data and rated their
perceptions of the speech system, as well as their comfort and customer satisfaction with the speech system.

This study was used to address three main issues. First, factor and reliability analyses were used to evaluate items for a usability scale. Next, correlation and regression analyses were used to evaluate associations among individual difference variables (inherent novelty seeking, self-monitoring), usability, comfort, and customer satisfaction. Finally, quality categories derived from Study 1 were used to evaluate potential interactions between observer variables and interface perceptions.

Method

Participants

Participants in this study consisted of 862 students recruited using the University of South Florida (USF) Psychology Department’s participant pool. Participation in the study was voluntary, and the student participants received extra course credit.

The sample included 688 females and 161 males, with a mean age of 20.61 years (SD=3.78). The sample sizes for the six interface stimuli were approximately equivalent (Tennis Scoreboard, N=133; Directory Dialer, N=137; Flights, N=120; Movie, N=150; Financial Services, N=115) but the Prescription Refill group was largest (N=206). The majority of participants reported normal hearing (97%, N=840) and English as their native language (95%, N=818). They represented a very broad range of job roles and educational backgrounds but most were full-time students (72%, N=619) and psychology majors (56%, N=479). Participants indicated that they used speech interface systems occasionally (N=308), seldom (N=285), or never (N=156), and only 13% of the sample indicated they were “expert” users (N=114) of such systems. Several analyses were conducted on the various demographic variables to look at the distribution of participants across the six interface (between subjects) groups. These analysis suggested that representatives of
different genders, ages, educational backgrounds, native languages, job types, and frequency of previous speech technology usage.

Several analyses were conducted to determine whether participants’ native language (English vs. other languages) and hearing status effected their ratings of the primary variables of interest in this study. The sample included 816 native English speakers and 44 native speakers of other languages. Independent t-tests demonstrated that the two native language groups were not significantly different ($p>0.11$) on ratings of customer satisfaction, Inherent Novelty Seeking, Self-Monitoring Sensitivity, Self-Monitoring Ability, Comfort, or four subscales of the Usability scale (Speech Characteristics, User Goal Orientation, Hedonic Quality, Pragmatics). However, statistically significant difference between the groups occurred for Need for Interaction with a Service Provider ($p=0.03$) and the Customer Service Behavior ratings on the Usability scale ($p=0.03$). For these variables, non-native English speakers reported they had less need for interacting with a human ($M=4.86, SD=1.47$) but rated speech interfaces as having more negative service behavior ($M=5.08, SD=1.24$) than native English speakers (Need for Interaction $M=5.30, SD=1.26$; Customer Service Behavior $M=5.43, SD=1.08$). Because the non-native English speakers are more likely to choose technology-based self-service and are part of the general user population for speech technology, their data were retained in the sample. A similar independent $t$-test analysis was conducted to determine differences between participants with normal hearing ($N=840$) and a hearing loss or disability ($N=22$). No statistically significant differences between the groups were observed for any of the primary variables of interest ($p>0.31$) and their data was also retained in the sample.

Procedure

The participants were recruited using an automated research tracking tool (ExperimenTrak), which displayed a link to the first page of the online study to eligible students.
The ExperimenTrak interface was hosted on a USF-based server. When students clicked the survey link, they were directed to a website hosted on an IBM-based server. The website instructions (Appendix B), questionnaire, and visual display were identical for all participants, only differing in the audio file presented to each of six groups. After accessing the study start page, participants clicked an audio file link, which accessed a third-party internet service provider server, and listened to the recorded user-system interaction via .mp3 audio files. The participant then completed the questionnaire items, which required approximately 15 to 20 minutes. When the participant clicked a “Submit” button, his or her responses were uploaded to the IBM server and the participant was directed back to the USF ExperimenTrak system to receive credit for participation. To prevent individuals from participating multiple times, the ExperimenTrak software was configured to allow participation in only one of the six groups and WebSurveyor was configured to eliminate multiple responses from the same IP address. To guard against history effects and maintain approximately equal sample sizes, the data for all six interfaces were collected throughout the period of August to November 2004. The ExperimenTrak software automatically assigned participants based on internal criteria, which included number of completed surveys and students’ personal criteria established by the Psychology department.

**Measures**

*Demographic Questions.* Information regarding participants’ gender, age, native language, and previous experience with speech technology was collected (see Appendix C). In addition, a question regarding hearing acuity was used to screen individuals for hearing loss. Participants also identified their job type and educational background to allow the assessment of the generalizability of the data obtained.

*Inherent Novelty Seeking.* This variable was measured using the Dabholkar and Bagozzi (2002) four item scale (see Appendix D). Individuals high in inherent novelty seeking “tend to
look favorably on technology and the use of technology based products, have stronger intrinsic motivation to use such products, and enjoy the stimulation of trying new ways to approach old problems” (Dabholkar & Bagozzi, p. 188). Confirmatory factor analysis indicated that the scales formed a distinct factor and moderated relationships among other user variables (performance, intention, fun, ease of use, and attitude), providing some empirical support for the scale’s validity. Dabholkar and Bagozzi (2002) reported a Chronbach’s $\alpha$ of 0.72 for their scale. The current results indicated a Chronbach’s $\alpha$ of 0.85 for this scale.

*Need for Interaction with a Service Employee.* Dabholkar (1996) developed a four-item scale to measure this construct, which was further examined by Dabholkar and Bagozzi (2002) in their attitudinal model. Confirmatory factor analysis indicated that the scales formed a distinct factor and moderated relationships between ease of use and attitude and between fun and attitude, which provided some empirical support for the scale’s validity. Three items make up the scale (see Appendix E). Dabholkar and Bagozzi (2002) report a Chronbach’s alpha of 0.83 for this scale.

In the present study, reliability for the Dabholkar and Bagozzi (2002) scale was borderline (coefficient $\alpha = 0.68$) due to a lower item-to-total correlation for item 3 ($r=0.40$) than the other two items (item 1 $r=0.55$, item 2 $r=0.60$). However, because this was a previously published scale, the item was retained in the scale score for subsequent analyses.

*Self-Monitoring.* An aspect of personality known as self-monitoring refers to the extent to which individuals attend to the expressive behavior of others and thereby control their own expression (Snyder, 1974; Snyder & Cantor, 1980). Snyder (1974) introduced the Self-Monitoring Scale but this measure has drawn criticism based on its construct validity, factor structure, and psychometric qualities, leading to controversy and revised versions of the scale (Briggs & Cheek, 1986; Briggs, Cheek, & Buss, 1980; Lennox & Wolfe, 1984; O’Cass, 2003; Snyder & Gangestad, 1986).
To respond to difficulties in the factor structure of the original scale and the operationalization of the self-monitoring construct, Lennox and Wolfe (1984) presented a 13-item, two-factor scale. Unlike the original true/false response format of the Snyder (1974) scale, the Revised Self-Monitoring Scale (RSMS) uses six point scales anchored by ‘certainly always false’ to ‘certainly always true.’ Lennox and Wolfe (1984) argue that their narrow operationalization of self-monitoring provides better construct validity and is more consistent with the original theoretical definition (Snyder, 1974). O’Cass (2003) evaluated the psychometric qualities of the RSMS and confirmed a two-factor structure (Self-Monitoring Ability, Self-Monitoring Sensitivity) and high internal consistency reliabilities (Cronbach’s alphas exceeding 0.85 for factors and full scale score). O’Cass (2003) also modified the scale anchors to read ‘strongly agree’ and ‘strongly disagree’ due to participant confusion and misinterpretation of the original anchors during pilot testing.

The O’Cass (2003) RSMS was used to measure self-monitoring. Although Snyder and Gangestad (1986) argue in favor of a single factor construct, in this study, one or both RSMS factors may have been more strongly associated with experienced individuals’ perception of user responses to speech interfaces than an overall scale score. Therefore, the two factor RSMS appears to operationalize better the variable of interest than the original Snyder (1974) scale. The 12 items of the O’Cass (2003) RSMS appear in Appendix F.

In the current study, reliability for the Self-Monitoring Ability and Self-Monitoring Sensitivity factors was acceptable (Chronbach’s $\alpha$ 0.85 and 0.83, respectively).

*Speech Interface Usability.* Currently, no measures exist that adequately operationalize usability of speech interfaces. A previous pragmatic scale demonstrated very small effect sizes (Polkosky, 2003) and was restricted in the range of items included, requiring the development of a new measure. I generated 76 items based on a broad review of the literature, including pragmatic...
theory (Bernsen, Dybkjaer, & Dybkjaer, 1996; Grice, 1975) and measures of customer service and e-service (Dabholkar & Bagozzi, 2002; Liu & Arnett, 2000; Parasuraman, Zeithaml, & Barry, 1994; Zeithaml, Parasuraman, & Malhotra, 2000). In addition, items were included based on results of previous usability ratings and participant comments about speech interfaces under development at IBM (Polkosky, 2002a, 2002b, 2003, 2004; Polkosky & Lewis, 2003). Ten factors assumed to be associated with speech interface usability were used to comprehensively measure the broad range of variables suggested by the previous literature (see Appendix G). To complete the scale, participants indicated their agreement with each item by rating it on a scale from 1 (strongly disagree) to 7 (strongly agree).

Comfort. Comfort was measured using a modified version of Spake, Beatty, Brockman, and Crutchfield’s (2003) eight-item scale. These researchers define comfort as “a psychological state wherein a customer’s anxiety concerning a service has been eased, and he or she enjoys peace of mind and is calm and worry free concerning service encounters with this provider” (p. 317). Although it was defined specifically as related to service encounters, this variable identifies a specific case of a perlocutionary outcome of conversational interaction (Austin, 1962) during a service encounter and may be an important affective response to speech technology interactions.

Spake, Beatty, Brockman, and Crutchfield (2003) established item content validity using expert judgments on the items included in the final scale. Construct validity was statistically confirmed by comparing comfort to related constructs (satisfaction, trust, commitment, active voice) and by using structural equation modeling, which indicated “the useful addition of comfort to the more traditional satisfaction-trust-commitment paradigm” (Spake, Beatty, Brockman, & Crutchfield, 2003, p. 327). The anchors for seven of the 10-point semantic differential items are: ‘uncomfortable’-‘comfortable,’ ‘very uneasy’-‘very much at ease,’ ‘very tense’-‘very relaxed,’ ‘insecure’-‘secure,’ ‘worried’-‘worry free,’ ‘distressed’-‘calm,’ ‘turbulent’-‘serene.’ I modified
the anchors for the final item ‘troublesome’–’peace of mind’ to read ‘troubled’–’peaceful’ so it would be grammatically parallel to the rest of the scale (Appendix H).

The scale was reported to have a composite reliability (coefficient $\alpha$) of 0.99, with item reliabilities ranging from 0.79 to 0.91 (Spake, Beatty, Brockman, & Crutchfield, 2003). Current reliability of the scale was also acceptable (Crohnbach’s $\alpha = 0.96$).

*Customer Satisfaction.* Satisfaction has been defined as both an affective and cognitive construct (Crosby, Evans, & Cowles, 1990; Wirtz & Lee, 2003). It has central importance in the services marketing literature. The scale provided by Eroglu and Machleit (1990) includes the following items: 1) ‘satisfied’ to ‘dissatisfied,’ 2) ‘favorable’ to ‘unfavorable,’ 3) ‘pleasant’ to ‘unpleasant,’ and 4) ‘I like it very much’ to ‘I didn’t like it at all.’ Participants respond by indicating the strength of their response along a seven-point scale (Appendix I).

In their comparison of nine frequently-used measures of satisfaction, Wirtz and Lee (2003) reported that Eroglu and Machleit’s (1990) four-item seven-point bipolar scale of satisfaction showed the second-highest loadings on satisfaction (0.94, 0.83) and item reliabilities (0.88, 0.69), lowest error variances (0.16, 0.29), and positive correlations with an affective (0.52) and cognitive (0.52) factor across two different service contexts. They summarized their evaluation by saying “it seems these multi-item scales achieve finely grained measurement by tapping into satisfaction from different angles” (Wirtz & Lee, 2002, p. 353). Spake, Beatty, Brockman, and Crutchfield (2003) reported a composite reliability of 0.97 for this scale, which was comparable to its reliability in the present study (Chronbach’s $\alpha = 0.93$).

*Stimuli*

Stimuli for this study were selected based on expert ratings in Study 1. The two best (Financial Services, Prescription Refill), two worst (Flight Status, Directory Dialer), and two average (Tennis Scoreboard, Movies) interfaces were included in Study 2 to represent the range of...
quality in current implementations in the speech technology industry.

Results

In Study 2, data was automatically collected using WebSurveyor 4.1, a web survey and data collection tool, then transferred to a statistical software package (SPSS 11.5) for descriptive and inferential analyses. The manipulation checks and results of data analyses for each of the five research questions are presented in turn.

Manipulation Checks

Participants in each group responded to four questions about the interface they heard to determine whether they listened to the recording. Participants were scored on their responses to three factual questions about the user-system interaction. Correct answers were scored and summed to create a number correct; if the manipulation check score was two or greater, the participant’s data was retained in the sample. Based on this result, responses from 41 participants were omitted from all subsequent analyses, resulting in a sample size of 821.

In the first manipulation check question, participants were asked whether the system voice was male or female. The majority of participants correctly identified a female voice for the Directory Dialer (N=126, 98%), Flight (N=115, 100%), Prescription Refill (N=196, 98%), and Financial Services (N=107, 97%) interfaces. Male voices were correctly identified for Australian Open (N=124, 95%) and Movies (N=118, 86%) interfaces. The Movies system voice was less frequently identified as male, possibly due to an extensive main menu selection in a female voice, followed by the male voice once the user proceeded to the movie listings submenu (primacy effect).

For each interface, participants responded to an additional open-ended question and multiple choice question (see Table 4). As expected, participants provided correct answers more often on the multiple choice questions. In general, the results of the participant listening questions indicated that participants were generally accurate in their recall of the interactions.
Table 4. Open Ended and Multiple Choice Manipulation Check Responses

<table>
<thead>
<tr>
<th>Interface</th>
<th>Question</th>
<th>N% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis Scoreboard</td>
<td>The system provided scoring for what tennis match?</td>
<td>111 (87%)</td>
</tr>
<tr>
<td></td>
<td>Which of the following players did the user request scores for?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pete Sampras</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Roger Federer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Andre Agassi</td>
<td></td>
</tr>
<tr>
<td>Directory Dialer</td>
<td>What location did the user request?</td>
<td>119 (92%)</td>
</tr>
<tr>
<td></td>
<td>Which of the following commands did the user give?</td>
<td>106 (92%)</td>
</tr>
<tr>
<td></td>
<td>Fax number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Connect</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell phone</td>
<td></td>
</tr>
<tr>
<td>Flight Status</td>
<td>What airline did the user request information about?</td>
<td>108 (92%)</td>
</tr>
<tr>
<td></td>
<td>What was the arrival airport for the flight?</td>
<td>106 (92%)</td>
</tr>
<tr>
<td></td>
<td>JFK</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O'Hare</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LAX</td>
<td></td>
</tr>
<tr>
<td>Movies</td>
<td>The system provided movie listings for what location?</td>
<td>115 (84%)</td>
</tr>
<tr>
<td></td>
<td>What type of movie did the user request information about?</td>
<td>137 (94%)</td>
</tr>
<tr>
<td></td>
<td>Drama</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comedy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Suspense</td>
<td></td>
</tr>
<tr>
<td>Prescription Refill</td>
<td>What was the name of the pharmaceutical company?</td>
<td>159 (80%)</td>
</tr>
<tr>
<td></td>
<td>How did the user obtain his prescription?</td>
<td>193 (97%)</td>
</tr>
<tr>
<td></td>
<td>Delivery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pickup</td>
<td></td>
</tr>
<tr>
<td>Financial Services</td>
<td>What stock did the user buy?</td>
<td>101 (94%)</td>
</tr>
<tr>
<td></td>
<td>What kind of bill was the user alerted about?</td>
<td>100 (93%)</td>
</tr>
<tr>
<td></td>
<td>Phone</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Electric</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alimony</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mortgage</td>
<td></td>
</tr>
</tbody>
</table>

In addition to the manipulation checks for participants, I conducted an analysis to assess whether the quality categories influenced customer satisfaction ratings. This step was necessary
due to reliability problems with expert ratings of interface quality in Study 1 and provided another indicator of stimuli variability. A univariate ANOVA indicated that quality of stimulus (good, average, poor) had a significant influence on satisfaction ratings, $F(2,814)=82.51$, $MSe=1.97$, $p<0.0001$. Differences among all three quality categories were also supported by a post-hoc LSD test (average-poor, $p=.002$; average-good and poor-good, $p<0.0001$), as illustrated in Figure 1.

Figure 1. Speech User Interface Quality and Customer Satisfaction

![Bar chart showing mean customer satisfaction ratings for high, average, and low interface quality.]

**Question 1:** *What items should be measured in assessing the usability of speech interfaces?*

Data were collected on 76 items thought to measure aspects of usability in speech interfaces. Univariate analyses and visual inspection of the boxplots for the 76 items revealed that most items were characterized by a normal or near-normal, slightly skewed distribution. The only exceptions to this observation were five items that showed a positive skew (items 13, 45, 58, 70, and 75) and three items that had a negative skew (items 12, 34, and 61). The sample size, means, standard deviations, and correlations for the ten assumed factors appear in Tables 5 through 13.
Table 5. Means, Standard Deviations, and Correlations for Hypothesized Customer Service Expectations Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>10</th>
<th>19</th>
<th>34</th>
<th>60</th>
<th>64</th>
<th>66</th>
<th>74</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.25</td>
<td>1.63</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4.48</td>
<td>1.57</td>
<td>.56**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4.28</td>
<td>1.96</td>
<td>.38**</td>
<td>.49**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>5.60</td>
<td>1.32</td>
<td>.35**</td>
<td>.44**</td>
<td>.32**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>4.51</td>
<td>1.61</td>
<td>.55**</td>
<td>.72**</td>
<td>.45**</td>
<td>.42**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>5.45</td>
<td>1.33</td>
<td>.35**</td>
<td>.45**</td>
<td>.32**</td>
<td>.67**</td>
<td>.42**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>5.14</td>
<td>1.42</td>
<td>.48**</td>
<td>.61**</td>
<td>.36**</td>
<td>.44**</td>
<td>.65**</td>
<td>.47**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>5.35</td>
<td>1.43</td>
<td>.38**</td>
<td>.47**</td>
<td>.37**</td>
<td>.53**</td>
<td>.52**</td>
<td>.56**</td>
<td>.55**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Correlation is significant at 0.01 level (2-tailed)**

Table 6. Means, Standard Deviations, and Correlations for Hypothesized Pragmatics Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>5</th>
<th>8</th>
<th>15</th>
<th>26</th>
<th>41</th>
<th>42</th>
<th>50</th>
<th>65</th>
<th>76</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3.10</td>
<td>1.61</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3.62</td>
<td>1.61</td>
<td>.14**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>5.01</td>
<td>1.52</td>
<td>-.09**</td>
<td>-.13**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>3.51</td>
<td>1.62</td>
<td>.25**</td>
<td>.30**</td>
<td>-.05</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>5.04</td>
<td>1.55</td>
<td>-.09**</td>
<td>-.05</td>
<td>.55**</td>
<td>-.03</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>3.10</td>
<td>1.61</td>
<td>.24**</td>
<td>.35**</td>
<td>-.25**</td>
<td>.35**</td>
<td>-.17**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>3.37</td>
<td>1.71</td>
<td>.20**</td>
<td>.20**</td>
<td>-.24**</td>
<td>.20**</td>
<td>-.23**</td>
<td>.31**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>3.73</td>
<td>1.79</td>
<td>.13**</td>
<td>.35**</td>
<td>-.26**</td>
<td>.35**</td>
<td>-.22**</td>
<td>.47**</td>
<td>.34**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>4.17</td>
<td>1.52</td>
<td>.12**</td>
<td>.04</td>
<td>.30**</td>
<td>.05</td>
<td>.29**</td>
<td>.003</td>
<td>-.09**</td>
<td>-.03</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed)**

*Correlation is significant at the 0.05 level (2-tailed)
Table 7. Means, Standard Deviations, and Correlations for Hypothesized Recognition Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>13</th>
<th>27</th>
<th>29</th>
<th>30</th>
<th>33</th>
<th>35</th>
<th>71</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>2.61</td>
<td>1.54</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>4.19</td>
<td>1.41</td>
<td>-27**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>3.82</td>
<td>1.72</td>
<td>.26**</td>
<td>-.08*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>3.93</td>
<td>1.70</td>
<td>-.13**</td>
<td>.41**</td>
<td>-.08*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>4.95</td>
<td>1.48</td>
<td>-.24**</td>
<td>.38**</td>
<td>-.05</td>
<td>.26**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>2.97</td>
<td>1.59</td>
<td>.27**</td>
<td>-.11**</td>
<td>.24**</td>
<td>-.04</td>
<td>-.08*</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>4.54</td>
<td>1.66</td>
<td>-.41**</td>
<td>.35**</td>
<td>-.28**</td>
<td>.24**</td>
<td>.37**</td>
<td>-.26**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**correlation is significant at the 0.01 level (2-tailed)
* correlation is significant at the 0.05 level (2-tailed)

Table 8. Means, Standard Deviations, and Correlations for Hypothesized Affective Response Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>17</th>
<th>25</th>
<th>37</th>
<th>46</th>
<th>48</th>
<th>55</th>
<th>73</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>4.56</td>
<td>1.76</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>4.06</td>
<td>1.85</td>
<td>.63**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>4.42</td>
<td>1.58</td>
<td>.50**</td>
<td>.39**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>4.85</td>
<td>1.61</td>
<td>.59**</td>
<td>.45**</td>
<td>.62**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>2.65</td>
<td>1.56</td>
<td>-.27**</td>
<td>-.19**</td>
<td>-.20**</td>
<td>-.31**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>3.85</td>
<td>1.72</td>
<td>.67**</td>
<td>.68**</td>
<td>.45**</td>
<td>.48**</td>
<td>-.19**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>4.44</td>
<td>1.72</td>
<td>.67**</td>
<td>.58**</td>
<td>.46**</td>
<td>.53**</td>
<td>-.22**</td>
<td>.60**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**correlation is significant at the 0.01 level (2-tailed)

Table 9. Means, Standard Deviations, and Correlations for Hypothesized User Goal Orientation Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>14</th>
<th>23</th>
<th>28</th>
<th>32</th>
<th>51</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>4.86</td>
<td>1.45</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>4.46</td>
<td>1.57</td>
<td>.53**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>4.64</td>
<td>1.74</td>
<td>.49**</td>
<td>.48**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>4.40</td>
<td>1.54</td>
<td>.53**</td>
<td>.59**</td>
<td>.48**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>4.39</td>
<td>1.62</td>
<td>.56**</td>
<td>.53**</td>
<td>.53**</td>
<td>.49**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**correlation is significant at the 0.01 level (2-tailed)
Table 10. Means, Standard Deviations, and Correlations for Hypothesized Accuracy Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>47</th>
<th>57</th>
<th>63</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>5.42</td>
<td>1.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>3.09</td>
<td>1.52</td>
<td>.32**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>4.48</td>
<td>1.60</td>
<td>.32**</td>
<td>-.30**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>4.94</td>
<td>1.47</td>
<td>.50**</td>
<td>-.40**</td>
<td>.42**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**correlation is significant at the 0.01 level (2-tailed)**

Table 11. Means, Standard Deviations, and Correlations for Hypothesized Prompt Wording Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>12</th>
<th>20</th>
<th>58</th>
<th>59</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>5.59</td>
<td>1.49</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>5.53</td>
<td>1.33</td>
<td>.53*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>2.87</td>
<td>1.69</td>
<td>.05</td>
<td>.09*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>5.00</td>
<td>1.62</td>
<td>.45**</td>
<td>.47**</td>
<td>.08*</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>2.45</td>
<td>1.52</td>
<td>-.38**</td>
<td>-.38**</td>
<td>.10**</td>
<td>-.41**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**correlation is significant at the 0.01 level (2-tailed)**

* correlation is significant at the 0.05 level (2-tailed)

Table 12. Means, Standard Deviations, and Correlations for Hypothesized Usefulness Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>9</th>
<th>16</th>
<th>21</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4.54</td>
<td>1.70</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5.14</td>
<td>1.39</td>
<td>.58**</td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4.46</td>
<td>1.62</td>
<td>.30**</td>
<td>.23**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.60</td>
<td>1.84</td>
<td>-.53**</td>
<td>-.36**</td>
<td>.24**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3.44</td>
<td>2.03</td>
<td>.48**</td>
<td>.40**</td>
<td>.16**</td>
<td>-.39**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>4.55</td>
<td>1.74</td>
<td>.13**</td>
<td>.19**</td>
<td>.10**</td>
<td>.01</td>
<td>.06</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>4.72</td>
<td>1.81</td>
<td>.62**</td>
<td>.63**</td>
<td>.20**</td>
<td>-.38**</td>
<td>.45**</td>
<td>.16**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**correlation is significant at the 0.01 level (2-tailed)**
### Table 13. Means, Standard Deviations, and Correlations for Hypothesized Speech Impression Factor

<table>
<thead>
<tr>
<th>Item#</th>
<th>M</th>
<th>SD</th>
<th>2</th>
<th>11</th>
<th>24</th>
<th>31</th>
<th>39</th>
<th>40</th>
<th>44</th>
<th>45</th>
<th>52</th>
<th>54</th>
<th>62</th>
<th>68</th>
<th>69</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.04</td>
<td>1.68</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5.37</td>
<td>1.41</td>
<td>-0.24**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>3.55</td>
<td>1.93</td>
<td>0.41**</td>
<td>-0.35**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>5.29</td>
<td>1.55</td>
<td>-0.20**</td>
<td>0.47**</td>
<td>-0.38**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>4.96</td>
<td>1.72</td>
<td>-0.27**</td>
<td>0.61**</td>
<td>-0.52**</td>
<td>0.49**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>4.65</td>
<td>1.49</td>
<td>-0.19**</td>
<td>0.41**</td>
<td>-0.34**</td>
<td>0.45**</td>
<td>0.46**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>3.86</td>
<td>1.88</td>
<td>-0.23**</td>
<td>0.38**</td>
<td>-0.50**</td>
<td>0.33**</td>
<td>0.50**</td>
<td>0.34**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>2.78</td>
<td>1.86</td>
<td>0.24**</td>
<td>-0.29**</td>
<td>0.46**</td>
<td>-0.44**</td>
<td>-0.40**</td>
<td>-0.33**</td>
<td>-0.32**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>4.29</td>
<td>2.01</td>
<td>0.36**</td>
<td>-0.30**</td>
<td>0.60**</td>
<td>-0.30**</td>
<td>-0.48**</td>
<td>-0.28**</td>
<td>-0.53**</td>
<td>0.36**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>3.85</td>
<td>1.95</td>
<td>-0.28**</td>
<td>0.43**</td>
<td>-0.59**</td>
<td>0.43**</td>
<td>0.61**</td>
<td>0.41**</td>
<td>0.61**</td>
<td>-0.39**</td>
<td>-0.69**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>3.83</td>
<td>1.92</td>
<td>-0.32**</td>
<td>0.41**</td>
<td>-0.63**</td>
<td>0.41**</td>
<td>0.61**</td>
<td>0.37**</td>
<td>0.58**</td>
<td>-0.39**</td>
<td>-0.70**</td>
<td>0.78**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>5.30</td>
<td>1.47</td>
<td>-0.21**</td>
<td>0.61**</td>
<td>-0.36**</td>
<td>0.48**</td>
<td>0.60**</td>
<td>0.44**</td>
<td>0.37**</td>
<td>-0.28**</td>
<td>-0.32**</td>
<td>0.42**</td>
<td>0.43**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>4.96</td>
<td>1.88</td>
<td>-0.24**</td>
<td>0.50**</td>
<td>-0.51**</td>
<td>0.63**</td>
<td>0.60**</td>
<td>0.47**</td>
<td>0.47**</td>
<td>-0.64**</td>
<td>-0.45**</td>
<td>0.56**</td>
<td>0.55**</td>
<td>0.48**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>3.97</td>
<td>1.76</td>
<td>-0.21**</td>
<td>0.44**</td>
<td>-0.40**</td>
<td>0.25**</td>
<td>0.52**</td>
<td>0.34**</td>
<td>0.46**</td>
<td>-0.26**</td>
<td>-0.42**</td>
<td>0.51**</td>
<td>0.52**</td>
<td>0.43**</td>
<td>0.39**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Correlation is significant at the 0.01 level (2-tailed)*
The overall goal of analyses for this research question was to create a usability scale for practical usage in applied settings; therefore, the desired end result was a reliable scale with 25 or fewer items. To decrease the number of items from the original 76 items, reliability analysis and factor analysis were used to eliminate problematic items in the scale and identify items that clustered in subscales. First, a reliability analysis was conducted on items for each of the ten assumed subscales. Cronbach’s $\alpha$ was calculated for each factor and each item’s corrected item-to-total correlation and item deleted $\alpha$ were reviewed. When an item demonstrated a relatively low item-to-total correlation and an improvement in $\alpha$, it was deleted from the subscale factor and Cronbach’s $\alpha$ was recalculated. This procedure was repeated until Cronbach’s $\alpha$ reached or exceeded the criterion value of 0.70 (Nunnally & Bernstein, 1994). As shown in Table 14, this procedure resulted in elimination of 29 items (including two entire subscales, Recognition and Accuracy) due to reliability problems.

Next, a Principal Component analysis with varimax rotation was conducted to determine whether the subscale structure of the scale could be empirically supported. In contrast to the a priori assumption, visual inspection of the scree plot indicated a six-factor solution was most likely. This solution accounted for 60.86% of the variance in the data. However, only a single item loaded on one factor in this initial solution and four items did not load on any factor. Therefore, the four items (40, 50, 59, 61) were removed and a second factor analysis was conducted with the reduced item pool. This analysis again suggested a five-factor solution was most likely (see Figure 2) but no items loaded on Factor 5. This solution accounted for 61.26% of the variance in the data. Loadings for the rotated five-factor solution appear in Table 15. As shown, the majority of items loaded on Factor 1. The originally hypothesized Pragmatic and Speech factors were supported with four items (Factor 4 = items 8, 26, 42, and 65) and six items (Factor 3 = items 39, 44, 54, 62, 69, 72). Two items (5 and 36) did not load on any factor and were removed for the subsequent analysis.
Table 14. Summary of Reliability Analysis for Usability Scale

<table>
<thead>
<tr>
<th>Assumed Factor</th>
<th>Items</th>
<th>α</th>
<th>Items Removed</th>
<th>Final α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Service</td>
<td>1, 10, 19, 34, 60, 64, 66, 74</td>
<td>0.87</td>
<td>19</td>
<td>0.88</td>
</tr>
<tr>
<td>Pragmatics</td>
<td>5, 8, 15, 26, 41, 42, 50, 65, 76</td>
<td>0.51</td>
<td>15, 41, 76</td>
<td>0.70</td>
</tr>
<tr>
<td>Recognition</td>
<td>13, 27, 29, 30, 33, 35, 71</td>
<td>0.29</td>
<td>13, 29, 35</td>
<td>0.61 (factor deleted)</td>
</tr>
<tr>
<td>Affective Response</td>
<td>17, 25, 37, 46, 48, 55, 73</td>
<td>0.76</td>
<td>48</td>
<td>0.86</td>
</tr>
<tr>
<td>User Goal Orientation</td>
<td>14, 23, 28, 32, 51</td>
<td>0.84</td>
<td>None</td>
<td>0.84</td>
</tr>
<tr>
<td>Accuracy</td>
<td>47, 57, 63, 43</td>
<td>0.11</td>
<td>57</td>
<td>.67 (factor deleted)</td>
</tr>
<tr>
<td>Prompt Wording</td>
<td>12, 20, 58, 59, 70</td>
<td>0.21</td>
<td>70, 58</td>
<td>0.74</td>
</tr>
<tr>
<td>Usefulness</td>
<td>3, 4, 6, 9, 16, 21, 38</td>
<td>0.49</td>
<td>6, 9, 21</td>
<td>.80</td>
</tr>
<tr>
<td>Speech Impression</td>
<td>2, 11, 24, 31, 39, 40, 44, 45, 52, 54, 62, 68, 69, 72</td>
<td>0.91</td>
<td>2, 24, 45, 52</td>
<td></td>
</tr>
<tr>
<td>Mental Model/SUI</td>
<td>7, 18, 22, 49, 53</td>
<td>0.69</td>
<td>49, 53, 56, 75</td>
<td>0.85</td>
</tr>
<tr>
<td>Organization</td>
<td>61, 67, 75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A third Principal Component analysis (varimax rotation) was conducted with the reduced item pool and constrained to a four-factor solution, which accounted for 58.67% of the variance.

Factor loadings appear in Table 16. Similar to the previous solution, the majority of items loaded on the first factor, the proposed Pragmatics construct was supported by four items (Factor 5), and the Speech Characteristics factor was supported by five items (Factor 3). However, the Pragmatics...
factor included items that specifically related to the “talkativeness” of the interface, making the originally hypothesized factor name too general to provide an adequate description. Therefore, a more specific label, Verbosity, was adopted for this factor. Factors 1 and 2 included a broad variety of items from several of the hypothesized factors.

Additional analyses were used to reduce the number of items in the scale to make it suitable for use in applied settings. To reduce the size of the scale, items with the lowest loadings were eliminated from Factors 1 and 2 until they each included eight items. This process resulted in removal of 13 items from Factor 1 and one item from Factor 2. Additional analyses indicated that three of the factors had acceptable reliability: Factor 1, User Goal Orientation ($\alpha=0.92$) with eight items; Factor 2, Customer Service Behavior ($\alpha=0.89$) with eight items; and Factor 3, Speech Characteristics ($\alpha=0.87$) was acceptable with three items. Factor 4 (Verbosity) was marginally acceptable with four items ($\alpha=0.69$).
Table 15.

Initial Factor Loadings for Usability Scale

<table>
<thead>
<tr>
<th>Item#</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.679</td>
<td>.191</td>
<td>.154</td>
<td>.012</td>
<td>-.067</td>
</tr>
<tr>
<td>10</td>
<td>.703</td>
<td>.254</td>
<td>.317</td>
<td>.099</td>
<td>-.005</td>
</tr>
<tr>
<td>34</td>
<td>.228</td>
<td>.780</td>
<td>.114</td>
<td>.098</td>
<td>-.025</td>
</tr>
<tr>
<td>60</td>
<td>.729</td>
<td>.250</td>
<td>.328</td>
<td>.138</td>
<td>.122</td>
</tr>
<tr>
<td>64</td>
<td>.234</td>
<td>.756</td>
<td>.207</td>
<td>.066</td>
<td>-.041</td>
</tr>
<tr>
<td>66</td>
<td>.679</td>
<td>.390</td>
<td>.130</td>
<td>.057</td>
<td>.223</td>
</tr>
<tr>
<td>74</td>
<td>.343</td>
<td>.608</td>
<td>.252</td>
<td>.136</td>
<td>.146</td>
</tr>
<tr>
<td>5</td>
<td>.030</td>
<td>-.089</td>
<td>.138</td>
<td>-.414</td>
<td>-.421</td>
</tr>
<tr>
<td>8</td>
<td>-.076</td>
<td>-.022</td>
<td>-.118</td>
<td>-.630</td>
<td>.010</td>
</tr>
<tr>
<td>26</td>
<td>.069</td>
<td>-.050</td>
<td>-.025</td>
<td>-.708</td>
<td>-.162</td>
</tr>
<tr>
<td>42</td>
<td>-.304</td>
<td>-.061</td>
<td>.035</td>
<td>-.702</td>
<td>.008</td>
</tr>
<tr>
<td>65</td>
<td>-.278</td>
<td>-.032</td>
<td>-.141</td>
<td>-.678</td>
<td>.070</td>
</tr>
<tr>
<td>17</td>
<td>.671</td>
<td>.286</td>
<td>.379</td>
<td>.106</td>
<td>-.062</td>
</tr>
<tr>
<td>25</td>
<td>.553</td>
<td>.141</td>
<td>.430</td>
<td>.135</td>
<td>-.258</td>
</tr>
<tr>
<td>37</td>
<td>.672</td>
<td>.232</td>
<td>.097</td>
<td>.046</td>
<td>.176</td>
</tr>
<tr>
<td>46</td>
<td>.712</td>
<td>.248</td>
<td>.169</td>
<td>.061</td>
<td>.257</td>
</tr>
<tr>
<td>55</td>
<td>.625</td>
<td>.138</td>
<td>.373</td>
<td>.214</td>
<td>-.173</td>
</tr>
<tr>
<td>73</td>
<td>.611</td>
<td>.245</td>
<td>.407</td>
<td>.050</td>
<td>-.162</td>
</tr>
<tr>
<td>14</td>
<td>.640</td>
<td>.382</td>
<td>.099</td>
<td>.021</td>
<td>-.023</td>
</tr>
<tr>
<td>23</td>
<td>.670</td>
<td>.272</td>
<td>.188</td>
<td>.079</td>
<td>-.025</td>
</tr>
<tr>
<td>28</td>
<td>.562</td>
<td>.309</td>
<td>.233</td>
<td>-.078</td>
<td>-.129</td>
</tr>
<tr>
<td>32</td>
<td>.690</td>
<td>.181</td>
<td>.223</td>
<td>.184</td>
<td>.010</td>
</tr>
<tr>
<td>51</td>
<td>.570</td>
<td>.301</td>
<td>.252</td>
<td>.020</td>
<td>-.013</td>
</tr>
<tr>
<td>12</td>
<td>.238</td>
<td>.590</td>
<td>.071</td>
<td>-.042</td>
<td>.244</td>
</tr>
<tr>
<td>20</td>
<td>.277</td>
<td>.647</td>
<td>.086</td>
<td>-.004</td>
<td>.220</td>
</tr>
<tr>
<td>3</td>
<td>.636</td>
<td>.191</td>
<td>.185</td>
<td>.294</td>
<td>.074</td>
</tr>
<tr>
<td>4</td>
<td>.642</td>
<td>.442</td>
<td>.144</td>
<td>.130</td>
<td>.105</td>
</tr>
<tr>
<td>16</td>
<td>.555</td>
<td>-.040</td>
<td>.211</td>
<td>.260</td>
<td>-.024</td>
</tr>
<tr>
<td>38</td>
<td>.724</td>
<td>.238</td>
<td>.211</td>
<td>.184</td>
<td>.182</td>
</tr>
<tr>
<td>11</td>
<td>.266</td>
<td>.647</td>
<td>.346</td>
<td>.036</td>
<td>-.045</td>
</tr>
<tr>
<td>31</td>
<td>.312</td>
<td>.469</td>
<td>.294</td>
<td>.081</td>
<td>.446</td>
</tr>
<tr>
<td>36</td>
<td>.448</td>
<td>.287</td>
<td>.482</td>
<td>.087</td>
<td>.421</td>
</tr>
<tr>
<td>39</td>
<td>.383</td>
<td>.472</td>
<td>.543</td>
<td>.166</td>
<td>.003</td>
</tr>
<tr>
<td>44</td>
<td>.206</td>
<td>.161</td>
<td>.771</td>
<td>-.051</td>
<td>.097</td>
</tr>
<tr>
<td>54</td>
<td>.267</td>
<td>.177</td>
<td>.801</td>
<td>.095</td>
<td>.084</td>
</tr>
<tr>
<td>62</td>
<td>.287</td>
<td>.184</td>
<td>.768</td>
<td>.148</td>
<td>.098</td>
</tr>
<tr>
<td>69</td>
<td>.429</td>
<td>.291</td>
<td>.533</td>
<td>.086</td>
<td>.448</td>
</tr>
<tr>
<td>72</td>
<td>.233</td>
<td>.285</td>
<td>.602</td>
<td>.001</td>
<td>-.223</td>
</tr>
<tr>
<td>68</td>
<td>.360</td>
<td>.666</td>
<td>.278</td>
<td>.046</td>
<td>-.091</td>
</tr>
<tr>
<td>7</td>
<td>.620</td>
<td>.341</td>
<td>.110</td>
<td>.037</td>
<td>.247</td>
</tr>
<tr>
<td>18</td>
<td>.682</td>
<td>.271</td>
<td>.108</td>
<td>.046</td>
<td>.281</td>
</tr>
<tr>
<td>22</td>
<td>.497</td>
<td>.544</td>
<td>.176</td>
<td>.101</td>
<td>.187</td>
</tr>
<tr>
<td>67</td>
<td>.628</td>
<td>.365</td>
<td>.132</td>
<td>.073</td>
<td>.294</td>
</tr>
</tbody>
</table>
Table 16.

Second Factor Analysis Loadings for Usability Scale

<table>
<thead>
<tr>
<th>Item#</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.672</td>
<td>.168</td>
<td>.216</td>
<td>-.015</td>
</tr>
<tr>
<td>10</td>
<td>.687</td>
<td>.258</td>
<td>.350</td>
<td>.096</td>
</tr>
<tr>
<td>34</td>
<td>.198</td>
<td>.728</td>
<td>.184</td>
<td>.057</td>
</tr>
<tr>
<td>60</td>
<td>.712</td>
<td>.303</td>
<td>.323</td>
<td>.161</td>
</tr>
<tr>
<td>64</td>
<td>.206</td>
<td>.699</td>
<td>.276</td>
<td>.023</td>
</tr>
<tr>
<td>66</td>
<td>.664</td>
<td>.464</td>
<td>.116</td>
<td>.080</td>
</tr>
<tr>
<td>74</td>
<td>.315</td>
<td>.639</td>
<td>.254</td>
<td>.146</td>
</tr>
<tr>
<td>8</td>
<td>-.059</td>
<td>-.030</td>
<td>-.106</td>
<td>-.647</td>
</tr>
<tr>
<td>26</td>
<td>.082</td>
<td>-.106</td>
<td>.022</td>
<td>-.731</td>
</tr>
<tr>
<td>42</td>
<td>-.295</td>
<td>-.059</td>
<td>.023</td>
<td>-.692</td>
</tr>
<tr>
<td>65</td>
<td>-.259</td>
<td>-.025</td>
<td>-.150</td>
<td>-.687</td>
</tr>
<tr>
<td>17</td>
<td>.649</td>
<td>.276</td>
<td>.410</td>
<td>.114</td>
</tr>
<tr>
<td>25</td>
<td>.532</td>
<td>.072</td>
<td>.494</td>
<td>.126</td>
</tr>
<tr>
<td>37</td>
<td>.667</td>
<td>.290</td>
<td>.101</td>
<td>.047</td>
</tr>
<tr>
<td>46</td>
<td>.701</td>
<td>.345</td>
<td>.135</td>
<td>.097</td>
</tr>
<tr>
<td>55</td>
<td>.610</td>
<td>.089</td>
<td>.433</td>
<td>.195</td>
</tr>
<tr>
<td>73</td>
<td>.588</td>
<td>.204</td>
<td>.462</td>
<td>.049</td>
</tr>
<tr>
<td>14</td>
<td>.627</td>
<td>.363</td>
<td>.157</td>
<td>-.005</td>
</tr>
<tr>
<td>23</td>
<td>.653</td>
<td>.271</td>
<td>.224</td>
<td>.081</td>
</tr>
<tr>
<td>28</td>
<td>.543</td>
<td>.270</td>
<td>.284</td>
<td>-.082</td>
</tr>
<tr>
<td>32</td>
<td>.675</td>
<td>.198</td>
<td>.245</td>
<td>.192</td>
</tr>
<tr>
<td>51</td>
<td>.551</td>
<td>.301</td>
<td>.283</td>
<td>.024</td>
</tr>
<tr>
<td>12</td>
<td>.215</td>
<td>.654</td>
<td>.049</td>
<td>-.017</td>
</tr>
<tr>
<td>20</td>
<td>.248</td>
<td>.706</td>
<td>.065</td>
<td>.027</td>
</tr>
<tr>
<td>3</td>
<td>.618</td>
<td>.233</td>
<td>.184</td>
<td>.319</td>
</tr>
<tr>
<td>4</td>
<td>.624</td>
<td>.470</td>
<td>.163</td>
<td>.131</td>
</tr>
<tr>
<td>16</td>
<td>.548</td>
<td>-.028</td>
<td>.222</td>
<td>.267</td>
</tr>
<tr>
<td>38</td>
<td>.711</td>
<td>.308</td>
<td>.200</td>
<td>.205</td>
</tr>
<tr>
<td>11</td>
<td>.236</td>
<td>.604</td>
<td>.399</td>
<td>.013</td>
</tr>
<tr>
<td>31</td>
<td>.291</td>
<td>.612</td>
<td>.205</td>
<td>.141</td>
</tr>
<tr>
<td>39</td>
<td>.352</td>
<td>.466</td>
<td>.569</td>
<td>.167</td>
</tr>
<tr>
<td>44</td>
<td>.191</td>
<td>.192</td>
<td>.708</td>
<td>-.043</td>
</tr>
<tr>
<td>54</td>
<td>.239</td>
<td>.222</td>
<td>.775</td>
<td>.134</td>
</tr>
<tr>
<td>62</td>
<td>.253</td>
<td>.226</td>
<td>.748</td>
<td>.175</td>
</tr>
<tr>
<td>69</td>
<td>.411</td>
<td>.451</td>
<td>.415</td>
<td>.167</td>
</tr>
<tr>
<td>72</td>
<td>.208</td>
<td>.205</td>
<td>.671</td>
<td>-.023</td>
</tr>
<tr>
<td>68</td>
<td>.331</td>
<td>.606</td>
<td>.353</td>
<td>.009</td>
</tr>
<tr>
<td>7</td>
<td>.605</td>
<td>.428</td>
<td>.086</td>
<td>.069</td>
</tr>
<tr>
<td>18</td>
<td>.672</td>
<td>.372</td>
<td>.079</td>
<td>.077</td>
</tr>
<tr>
<td>22</td>
<td>.472</td>
<td>.597</td>
<td>.168</td>
<td>.121</td>
</tr>
<tr>
<td>67</td>
<td>.613</td>
<td>.469</td>
<td>.089</td>
<td>.115</td>
</tr>
</tbody>
</table>
The final usability scale consisted of four factors, named User Goal Orientation (items 1, 10, 18, 32, 37, 38, 46, 60), Customer Service Behavior (items 12, 20, 22, 31, 34, 64, 68, 74), Speech Characteristics (items 39, 44, 54, 62, 72), and Verbosity (items 8, 26, 42, 65). This result indicated that 25 items assess the usability of speech interfaces. User Goal Orientation included items that related to the system’s efficiency, user trust and confidence in the system, and clarity of a speech interface based on user needs. Customer Service Behavior included items that were related to the friendliness and politeness of the system, its speaking pace, and its use of familiar terms. The Verbosity factor was made up of items that measured the talkativeness and repetitiveness of the system and Speech Characteristics included items related to the naturalness and enthusiasm of the system voice.

To determine whether the final factors discriminated among the sets of interfaces, a series of four univariate ANOVAs was conducted, with interface as the independent variable and each of the usability factors as the dependent variables. The sets of interfaces were significantly different on ratings of User Goal Orientation ($F(5,814)=53.54$, $MSe=1.36$, $p<0.0001$, $\eta^2=0.196$), Customer Service Behavior ($F(5,814)=38.27$, $MSe=0.92$, $p<0.0001$, $\eta^2=0.191$), Speech Characteristics ($F(5,814)=96.15$, $MSe=1.41$, $p<0.0001$, $\eta^2=0.373$), and Verbosity ($F(5,814)=36.42$, $MSe=1.18$, $p<0.0001$, $\eta^2=0.184$). The mean ratings appear in Figure 3 (Verbosity is shown reverse-scored).

Post-hoc LSD tests also supported differences among the factor means, as shown in Table 17 ($p<0.05$). For example, the Financial Services interface was similar to the Refill interface in terms of Customer Service Behavior and User Goal Orientation but the two differed on perceptions of Speech Characteristics and Verbosity. Similarly, the Movie interface was perceived as similar to the Tennis Scoreboard and Flight interfaces on User Goal Orientation but was rated more positively on Speech Characteristics and Customer Service Behavior but more negatively on Verbosity. In general, the pattern of results indicated that the four factors of the usability scale did differentiate
among the interfaces.

Table 17. Post-Hoc Results for Interface Quality

<table>
<thead>
<tr>
<th>Quality Behavior</th>
<th>Customer Service Behavior</th>
<th>User Goal Orientation</th>
<th>Speech Characteristics</th>
<th>Verbosity (reverse-scored)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>Refill (5.92)</td>
<td>Refill (5.12)</td>
<td>Financial</td>
<td>Refill (5.22)</td>
</tr>
<tr>
<td></td>
<td>Financial (6.05)</td>
<td>Financial (5.36)</td>
<td>(5.52)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Refill (5.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Financial (4.69)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Flights (4.63)</td>
</tr>
<tr>
<td>Movie (5.46)</td>
<td>Tennis (4.20)</td>
<td>Movie (4.01)</td>
<td>Dialer (4.30)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Flights (4.27)</td>
<td></td>
<td>Tennis (4.28)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Movie (4.36)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worst</td>
<td>Dialer (4.82)</td>
<td>Dialer (3.66)</td>
<td>Tennis (3.33)</td>
<td>Movie (3.67)</td>
</tr>
<tr>
<td></td>
<td>Tennis (4.95)</td>
<td></td>
<td>Dialer (3.23)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Flights (5.14)</td>
<td></td>
<td>Flights (3.07)</td>
<td></td>
</tr>
</tbody>
</table>

(non-significantly different means are shown in each cell, significantly different means in different cells)
Question 2: How are individual difference variables (e.g., self-monitoring, need for interaction, inherent novelty seeking) related to comfort and customer satisfaction?

Descriptive statistics and Pearson product moment correlations were calculated for each of the individual difference variables, comfort, and customer satisfaction. For exploratory and comparative purposes, the four factor scores from the speech interface usability scale were also included in the analyses.

Visual inspection of the boxplots indicated approximately normal distributions for all variables. Means, standard deviations, and correlations appear in Table 18. Pooled within-group correlations were calculated (interface defined as an independent variable) to adjust for the effect of different interfaces on the relationships among the correlated variables.

Results showed that Inherent Novelty Seeking, Self-Monitoring Ability, and Self-Monitoring Sensitivity were weakly positively associated with both Comfort and Customer Satisfaction. Need for Interaction with a Service Provider (NISP) was weakly negatively
<table>
<thead>
<tr>
<th>Item</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Customer Satisfaction</td>
<td>5.08</td>
<td>1.54</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Comfort</td>
<td>7.24</td>
<td>1.85</td>
<td>.75**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Need for Interaction with a Service Provider</td>
<td>5.27</td>
<td>1.28</td>
<td>-.17**</td>
<td>-.11**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Inherent Novelty Seeking</td>
<td>5.07</td>
<td>1.19</td>
<td>.11**</td>
<td>.16**</td>
<td>.17**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Self-Monitoring Sensitivity</td>
<td>4.55</td>
<td>0.88</td>
<td>.07*</td>
<td>.08**</td>
<td>.18**</td>
<td>.26**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Self-Monitoring Ability</td>
<td>4.61</td>
<td>0.91</td>
<td>.12**</td>
<td>.13**</td>
<td>.10**</td>
<td>.25**</td>
<td>.42**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. User Goal Orientation</td>
<td>4.53</td>
<td>1.30</td>
<td>.71**</td>
<td>.65**</td>
<td>-.13**</td>
<td>.16**</td>
<td>.14**</td>
<td>.12**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Speech Characteristics</td>
<td>4.09</td>
<td>1.50</td>
<td>.43**</td>
<td>.40**</td>
<td>-.05*</td>
<td>.02</td>
<td>.09**</td>
<td>.08*</td>
<td>.53**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Customer Service Behavior</td>
<td>5.42</td>
<td>1.06</td>
<td>.40**</td>
<td>.40**</td>
<td>-.002</td>
<td>.11**</td>
<td>.20**</td>
<td>.21**</td>
<td>.64**</td>
<td>.49**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>10. Verbosity</td>
<td>3.49</td>
<td>1.20</td>
<td>-.26**</td>
<td>-.25**</td>
<td>.14**</td>
<td>.03</td>
<td>.06</td>
<td>.004</td>
<td>-.28**</td>
<td>-.15**</td>
<td>-.17**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Correlation is significant at the 0.05 level (2-tailed)
**Correlation is significant at the 0.01 level (2-tailed)
associated with these two variables. In contrast, the interface variables (User Goal Orientation, Customer Service Behavior, Speech) were strongly positively associated with Comfort and Customer Satisfaction. However, Verbosity showed a moderate negative association.

**Question 3: What speech user interface characteristics best predict customer satisfaction?**

The analyses for this question consisted of simultaneous multiple regressions to determine whether the usability scale factors and items predicted unique variance in customer satisfaction. These analyses were followed by stepwise multiple regressions to identify the best unique predictors of customer satisfaction. The item analysis was conducted to provide a shorter, more efficient set of items that predict customer satisfaction in applied settings. The multiple regression using factors as the predictors indicated a significant model \(F(4, 813)=302.79, MSe=0.95, p<0.0001\) with an \(R^2\) of 0.60. Diagnostic statistics indicated no evidence of collinearity. Significant \(b\) weights were observed for all four factor scores: Customer Service Behavior \((b=-0.15, t=-3.22, p=0.001)\), Speech Characteristics \((b=0.13, t=4.36, p<0.0001)\), Verbosity \((b=-0.12, t=-3.83, p<0.0001)\), and User Goal Orientation \((b=0.86, t=21.10, p<0.0001)\).

The stepwise analysis indicated a model with one predictor was significant \(F(1, 813)=1121.39, MSe=0.99, p<0.0001\) with an \(R^2\) of 0.58 and a resulting equation of Customer Satisfaction = 0.99 + 0.90(User Goal Orientation).

The analysis was repeated using the usability scale items, and indicated a significant model \(F(25, 800)=52.28, MSe=0.91, p<0.0001\) with an \(R^2\) of 0.63. Significant \(b\) weights were observed for nine items: item 1 \((b=0.07, t=2.51, p=0.01)\), item 10 \((b=0.15, t=4.14, p<0.0001)\), item 18 \((b=0.06, t=2.04, p=0.04)\), item 20 \((b=-0.08, t=-2.33, p=0.02)\), item 34 \((b=-0.08, t=-198, p=0.05)\), item 38 \((b=0.19, t=5.96, p<0.0001)\), item 46 \((b=0.12, t=3.62, p<0.001)\), item 60 \((b=0.18, t=4.78, p<0.0001)\), item 65 \((b=-0.05, t=-2.18, p=0.03)\).
The stepwise analysis showed that a model with four items was significant ($F(4, 800)=292.25, MSe=0.96, p<0.0001$) with an $R^2$ of 0.60. The resulting standardized regression equation ($p<.0001$) is Customer Satisfaction = 0.119 + 0.036(item 60) + 0.030(item 38) + 0.034(item 10) + 0.033(item 46).

In general, this analysis suggested that 58% to 63% of the variance in customer satisfaction can be accounted for by the factors or items from the Usability scale. A more efficient set of four usability items accounts for slightly less customer satisfaction variance (60%). The resulting stepwise regression equation shows that customer satisfaction ratings may be predicted from the User Goal Orientation score or ratings of general quality (item 60 and 10), expected repeat usage (item 38), and user confidence (item 46).

**Question 4: What speech user interface characteristics best predict comfort?**

As with question three, analysis for question four consisted of a simultaneous multiple regression and a stepwise multiple regression to identify the best unique predictors for the criterion comfort. These analyses were conducted with both the factor scores and the item ratings.

For the factors, the multiple regression indicated a significant model ($F(4, 813)=216.98, MSe=1.65, p<0.0001$) with an $R^2$ of 0.52. Diagnostic statistics indicated no evidence of collinearity. Significant $b$ weights were observed for three factors: User Goal Orientation ($b=0.90, t=16.80, p<0.0001$), Speech Characteristics ($b=0.15, t=3.57, p<0.0001$), and Verbosity ($b=-0.12, t=-3.07, p=0.002$). The $b$ weight for Customer Service Behavior was not statistically significant ($p=0.37$).

The stepwise analysis showed that a model with one factor was significant ($F(1, 812)=824.06, MSe=1.69, p<0.0001$) with an $R^2$ of 0.50. The resulting standardized regression equation ($p<0.0001$) is Comfort = 0.17 + 0.035(User Goal Orientation).

The analyses were again repeated with the items to determine an efficient set that may be used in lieu of a complete scale or subscale in applied settings. The stepwise analysis showed that
the model was significant \( (F(25, 800)=37.25, \text{MSe}=1.59, p<0.0001) \) with an \( R^2 \) of 0.54. Significant \( b \) weights were observed for nine items: item 10 \( (b=0.16, t=3.37, p=0.001) \), item 18 \( (b=0.14, t=3.45, p=0.001) \), item 31 \( (b=0.11, t=2.81, p=0.005) \), item 37 \( (b=0.14, t=3.35, p=0.001) \), item 39 \( (b=0.12, t=2.68, p=0.007) \), item 46 \( (b=0.20, t=4.45, p<0.0001) \), item 60 \( (b=0.10, t=1.93, p=0.05) \), item 62 \( (b=0.09, t=2.29, p=0.02) \), and item 64 \( (b=-0.10, t=-2.03, p=0.04) \).

The stepwise analysis showed that a model with four items was significant \( (F(4, 800)=198.47, \text{MSe}=1.72, p<0.0001) \) with an \( R^2 \) of 0.50. The resulting standardized regression equation \( (p<.0001) \) is Comfort = 0.163 + 0.043(item 60) + 0.041(item 46) + 0.039(item 18) + 0.028(item 62).

As with the previous research question, this analysis suggested that 50 to 54% of the variance in comfort can be accounted for by one factor or up to nine items from the Usability scale. However, only four items account for 50% of the variance in the criterion. The resulting stepwise regression equation shows that comfort ratings may be predicted from the User Goal Orientation subscale score or ratings of general quality (item 60), user confidence (item 46), navigation ease (item 18), and naturalness of the system voice (item 62).

**Question 5:** Do individual difference variables interact with speech user interface quality to create customer satisfaction?

This analysis consisted of an ANCOVA with two categorical independent (predictor) variables (interface quality, gender) and five continuous independent (predictor) variables (age, Need for Interaction with a Service Provider, Inherent Novelty Seeking, Self-Monitoring Ability, Self-Monitoring Sensitivity). The categorical variables were both effect coded for the analysis and vectors were constructed to represent the two-way interactions among all seven variables to complete the analysis. Customer satisfaction served as the dependent (criterion) variable.

Results of the first analysis with only main effect predictors indicated a significant model,
The present research integrates research in interpersonal communication, psycholinguistics, human factors, and services marketing to create the beginnings of a theoretical and methodological foundation for future applied work in speech technology. Although speech technology has only recently emerged as a mainstream means of e-service delivery, it is thought to have huge market potential: ABI Research predicts that the speech technology market will grow to over $5 billion by 2008 as it is used to replace or supplement call centers (Allied Business Intelligence Research, 2003) and analysts acknowledge that the quality of speech technology has improved to acceptable levels for broad user acceptance (Harris Interactive, 2003; Popova, 2004).
However, despite the substantial business interest in this technology and broad acceptance of the critical need for well-designed interfaces, there is limited research that applies basic social-communicative findings to speech interface design, and users’ affective responses to these interactions have been only minimally explored. The present research drew from a broad literature to develop a scale of usability for speech interfaces and examine interrelationships among this construct, individual differences, comfort and customer satisfaction. These results have implications for both the methodology and theory of speech technology, as well as the disciplines that inform the current study.

**Methodological Implications**

Although the development of a measurement scale formed a substantial part of the present research, its general methodology has implications for applied e-service research. The Usability Scale for Speech Interfaces advances practitioners’ ability to measure usability in industrial settings. The use of speech and language experts, recorded clips of interfaces, and third-party observers provides an efficient means of understanding affective responses to speech technology. Nonetheless, critical examination of the current methodology suggests a number of issues that should be empirically addressed in future applied research.

**The Usability Scale for Speech Interfaces**

The first research goal was to develop a scale of usability for speech interfaces. The development of a usability scale marks a significant methodological advance in speech technology, because no empirically-validated scales for measurement of this construct currently exist. In addition, although usability is widely regarded as an important consideration in designing interfaces, little has been known about it in the context of speech technology.

Seventy-six items for the usability scale were generated based on a broad review of the
literature in social psychology, communication and services marketing. The resulting usability scale contained 25 items and four factors, known as User Goal Orientation, Customer Service Behavior, Verbosity, and Speech Characteristics. In general, the Usability Scale for Speech Interfaces appeared to be psychometrically sound, compared favorably with its predecessor and related scales, and is consistent with the theoretical research that underlies its development. Each of these issues will be strengths will be discussed in turn.

**Psychometric Quality of the Usability Scale.** In general, the Usability Scale for Speech Interfaces appeared to be psychometrically sound with acceptable levels of reliability for four of the five factors. The new measure also appears to have construct validity because it indicates that usability for speech interfaces is a multifaceted, complex variable that incorporates technological features such as ease of use, communicative features such as voice naturalness, and human customer service behaviors. In this sense, it measures speech system usability with greater face validity than more narrowly operationalized scales concerned with only speech or pragmatics (Polkosky, 2003). In addition, the Usability Scale accounted for a substantial amount of variance in customer satisfaction and comfort, suggesting its predictive validity for e-service outcomes. The predictive capability of the scale is especially important to its usefulness as speech system usability measure because it may be used to predict user outcomes proactively during development. Finally, the new scale appears to have content validity because it includes items that were previously found to be important to observers’ liking and expected future use of a speech interface, such as voice naturalness (Polkosky, 2003), and friendliness of service providers (Baydoun, Rose & Emperado, 2001).

Despite the apparent strengths of the Usability Scale, the Verbosity factor demonstrated somewhat weaker reliability than the other four factors. Nonetheless, there are several reasons for retaining this factor and its associated items in the Usability Scale. First, the Verbosity score did
discriminate among the six Study 2 interfaces in a different pattern than the other factors, so it is likely that the combination of the five factors will provide designers with a comprehensive understanding of a speech system’s usability. Second, it seems intuitively reasonable that items like perceived “talkativeness,” amount of detail, and repetitiveness should be included in a usability scale, because participant comments during informal usability testing often include these perceptions of speech systems. Third, a more general version of this scale (Pragmatics) had extremely small effect sizes in previous research (Polkosky, 2003) and pragmatic perceptions are likely to be automatic and beyond the conscious awareness of observers (Bargh & Ferguson, 2000), which may result in reduced measurement reliability. Finally, the application of pragmatics to speech interface design is consistent with previous approaches (Bernsen, Dybkjaer & Dybkjaer, 1996; Brennen, 1998; Saygin and Cicekli, 2002), although the current scale measures a specific aspect of pragmatic behavior that appears to be intuitively relevant to customer service contexts. Therefore, despite the psychometric weaknesses of this factor, it should be retained in the Usability Scale.

Another potential weakness of the scale was the small effect size for each Usability factor, based on the Cohen’s (1969) interpretation guidelines. A possible interpretation of this finding is that the effects are too small for continued investigation. However, in the context of current trends in speech technology, even small effects may result in large differences in business-related outcomes. For example, consider an enterprise that receives 100,000 calls in to a speech system each month, at a savings of $5.85 per call as compared to human handling of that same call. If 10% of the callers opt out of the speech system (i.e., hang up or transfer to an operator) due to a negative perception of the system itself, the additional financial cost to the enterprise is approximately $702,000 yearly. There is also the potential loss of the customers who are dissatisfied to the extent that they take their business elsewhere. A seemingly small change to the interface, such as
substituting a more pleasant or friendly-sounding system voice, may effect large numbers of users, resulting in not only improved customer satisfaction, but increased accuracy of system use, increased preference for e-service, retention of the user in a speech system without transfer to a human operator, operational savings, and faster return on investment. Therefore, understanding even small effects may have a substantial impact on the use of speech technology usage in applied settings.

Comparison between the Usability Scale and Other Measures of Speech Technology and e-Service. The current scale suggests that usability is a multi-faceted construct that incorporates a broad variety of items. In contrast to previous measures for speech technology, the Usability Scale for Speech Interfaces is significantly broader in its scope. The present results of scale development indicate that speech system usability is a related but unique variant of visual and human forms of customer service.

Results of this study reveal the limitations of previous attempts to develop a usability scale for speech interfaces. Polkosky and Lewis (2003) developed the Mean Opinion Scale – Expanded (MOS-X) to measure speech characteristics of synthetic speech. The MOS-X scale, which measured Intelligibility, Naturalness, Prosody and Social Impressions, subsequently was used by Polkosky (2003) to measure observer perceptions of speech interfaces. Although the MOS-X expanded the scope of its predecessor (the Mean Opinion Scale or MOS), its singular focus on speech characteristics appeared to be too limited for measurement of interactive conversational systems (Polkosky, 2003). This outcome was expected, because this scale was developed to discriminate among different synthetic voices speaking the same text, not recorded human voices nor interactive dialogue usability. The Usability Scale includes speech items measuring the naturalness of a system voices; similarity between a system voice and radio, television or regular human voices; and the enthusiasm or energy of a system voice. This result is consistent with
previous MOS-X research, which showed that Naturalness and Social Impression of system speech were most strongly correlated with Liking of a system, but refines social impression ratings to include those items most relevant to customer service. Interestingly, the inclusion of items that measure the similarity between a speech system’s voice and radio or television voices suggests observers’ expectations for interactive systems may be defined by vocal standards in other forms of popular media or technology. In contrast to the Mean Opinion Scale-Expanded or MOS-X (Polkosky & Lewis, 2003), the speech items included in the Usability Scale appear to suggest that speech perceptions play a prominent role in e-service judgments.

The current Usability Scale more broadly operationalizes usability than the previous Pragmatics Scale for Dialogues, which was focused on the interactive and social-communicative aspects of conversational interaction (Polkosky, 2003). Results with the Pragmatics Scale for Dialogues indicated its potential benefit for applied e-service measurement, it indicated it had very low effect sizes and correlations between its factors and affective outcomes (Polkosky, 2003). In contrast to the previous work, the current Verbosity factor included items concerned with repetitiveness and “talkativeness” of a system, a more targeted set of items that address frequently-heard comments about speech systems. The items included in the Usability scale appear to be those that are more specifically related to appropriate customer service and conversational behavior, providing a potential indication that expectations for interactive dialogues are formed by human service scripts (Solomon, Surpreant, Czepiel & Gutman, 1985).

Comparison of the Usability Scale for Speech Interfaces to measures of other forms of e-service is also important, because they contributed to the current scale development effort. The usability scale revealed a number of similarities and differences with the Zeithaml, Parasuraman and Malhotra (2000) measure of e-service quality (e-SQ), which provides 11 dimensions of service quality, specifically targeting self-service via Internet websites. Similar to their scale, the User
Goal Orientation factor included items that measure several dimensions of e-SQ: Ease of Navigation, Reliability, Efficiency, and Customization/Personalization. The Customer Service Behavior scale is similar to their Assurance/Trust dimension. In contrast to the e-SQ, which included items related to the visual design of a website, the present scale includes items related to auditory presentation (Speech Characteristics) and expectations of appropriate conversational behavior (Verbosity), which is more similar to expectations of human conversation (Holtgraves, 2002). Another contrast with the e-SQ measure was its inclusion of a Price Knowledge dimension, which was not included in the speech usability scale, and may reflect the popularity of comparison-shopping on commercial Internet websites (Rust & Kannan, 2002).

A final comparison may be made between measures of human service provider personality and the Usability Scale for Speech Interfaces. The Usability Scale includes items that map to items measured in human service delivery. For example, Usability scale items measuring the confidence and enthusiasm of a speech system are related to findings of associations between extraversion and interpersonal skills (Alge et al., 2002) and performance in service providers (Mount, Barrick & Stewart, 1998). Mount, Barrick and Stewart (1998) found that Conscientiousness was most strongly correlated with job performance. Following from this finding, it is intuitively reasonable that a conscientious service provider would provide efficient service, behave with the customer’s best interests in mind, do their work correctly, and infer customer needs, all items included in the Usability Scale. The scale also includes several items specifically identified by Baydoun, Rose, and Emperado (2001) as being part of a customer service orientation (i.e., friendly, polite, and helpful).

In summary, comparison of the Usability Scale for Speech Interfaces with its predecessor and related e-service scales shows a number of similarities with other measures, as well as a number of differences. The scale refines specific factor scores (i.e., Speech Characteristics,
Verbosity) for the customer service context more than do previous scales, which defined these factors in a more open-ended manner. In addition, the scale includes items from both human and Internet service quality measures. Thus, although speech-based e-service has much in common with other forms of service, the usability of this new form of customer service is uniquely operationalized.

**Use of Experts and Students to Measure Speech System Usability**

The current research used both experts in human speech and language behavior and college students to measure usability. Generally, both groups demonstrated relatively consistent impressions of six speech interfaces, suggesting the potential of using both groups in applied measurement.

Both experts and students rated the Financial Services interfaces as highest in quality. The relative ranking of the six interfaces was similar for both groups (e.g., Financial Services followed by Prescription Refill, Flight Status and Directory Dialer poorest). A comparison of the overall Usability Scale score with mean expert ratings indicates that experts rated all interfaces more negatively than did observers on the seven-point scale. This difference was potentially due to their greater or more conscious understanding of conversational norms, which caused them to rate the interfaces more negatively when these norms were violated.

A group difference in ratings occurred for the Tennis Scoreboard and Movies interfaces: experts rated Tennis (expert $M=3.37$, observer $M=4.01$) as superior to Movies and observers rated Movies (expert $M=2.99$, observer $M=4.33$) as superior to Tennis. There are several potential reasons for this discrepancy. First, the most obvious difference between these interfaces was that the Tennis interface used a British English dialect as the system voice. Experts may have recognized the dialect without judging it but this speech pattern may have reduced the intelligibility of prompts for students, leading to their more negative impressions of the interface. Second, the
experts, who by definition have deep expertise in speech and language, may perceive interface characteristics differently than students. A general difference between these groups is likely because experts’ perceptions may be subject to controlled processing and extensive knowledge about underlying cognitive and physiological factors involved in speech production (i.e., metalinguistic skills), instead of heuristic processing. The third reason that the student and expert ratings differed may be due to differences in the interface characteristics measured by each group. For example, experts rated contingency of system responses and use of emphasis and intonation, items that were not specifically measured with the Usability Scale. Expert ratings were limited to aspects of speech and language, consistent with their expertise. In contrast, students rated a more extensive set of questions that included speech and language, but included a broad array of items concerned with technology usage and customer service as well. Finally, the students may have had a more positive affective response to the Movies interface because they are generally interested in and frequent this form of entertainment. Thus, the interface itself may have higher perceived usefulness to the students. In general, considering the differences in expertise and measurement of the two samples in these studies, both expert ratings and students demonstrated agreement about the relative quality of the six example interfaces, suggesting the concurrent validity of the Usability Scale.

Although students do comprise a large proportion of speech-based e-service users and are likely to be the majority of future users of this technology, the participant sample in this study was restricted. Nonetheless, age was a significant predictor of customer satisfaction for speech-based e-service. This result contradicts Dabholkar and Bagozzi’s (2001) assertion that basic demographic variables do not have an influential role in affective outcomes to e-service. However, future research required to more fully elucidate the potential impact of a broad age range on e-service outcomes. Much of the human factors literature has suggested that elderly individuals
require special considerations based on changing cognitive, motor, and other skills related to aging (Rogers, 1997). These skills are likely to impact their ability to interact with speech systems as well, suggesting that the present study should be replicated with a broader range of potential e-service participants.

The current results indicate that both speech and language experts and naive observers have a number of advantages for applied measurement. Because experts appeared to be highly critical of and sensitive to speech and language behavior, their impressions may be an efficient indication of a speech system’s use of conversational norms. Used early in development, expert ratings may be used to identify and refine problematic aspects of an interface design before the user population is given access to an e-service system. Given the significant business implications of poor interface design, expert ratings may prevent significant financial loss if they are used as part of an iterative process of evaluation and design during development. A potential weakness of expert ratings is that they have poor reliability, suggesting that they will change if they evaluate the same system multiple times. Use of students (and other non-expert potential users) also has an advantage, in that they are more numerous and possibly easier to recruit than speech and language experts. Ideally, a representative sample may be selected from the user population and their ratings may be used to better assess the post-deployment outcomes of a speech system. However, an adequate sample size will be more expensive and time-consuming to obtain, which will limit their practicality of sampling users in industrial settings.

Use of Third-Party Observers to Measure Affective Outcomes

One of the most important potential limitations of the present research was the use of observers instead of actual interface users. Findings from social cognition highlight this issue for not only applied speech technology research, but also marketing and interpersonal communication studies, which frequently use observers to generate data on conversational and service interactions.
In contrast, findings from the social-cognitive literature warn that interactants and observers may have different affective outcomes. Thus, the present results are limited to observers of speech interface usage and do not necessarily apply to users themselves. This methodological problem has important implications because the use of observers is an efficient and practical means of conducting applied research. In addition, vision clips have a central role in the proliferation of speech technology because they are the primary means of demonstrating speech technology functions to stakeholders before a system is designed. It should be a central goal of future research efforts that potential differences in user and observer affective responses be explored.

In spite of this weakness, the validity of the present research should not be underestimated. Because vision clips and an observer point of view are involved in the success of speech technology, the current research is a critical first step to understanding affective responses to this form of e-service. Continued advancement in this field requires that future studies cast point of view as an independent variable to better elucidate potential differences in affective outcomes of speech technology. Studies that have yoked conversational interactants and observers to compare their perceptions provide an excellent methodology that may be utilized in future studies (Gilovich, 1987). Increased variability in stimuli provides an empirical challenge even with yoked user-observer pairs, in that some users may proceed through an interface easily and quickly or have significant problems, requiring not only more time but more system messages to resolve communicative difficulties. In some cases, users may not be able to use a speech system at all. To resolve this issue, researchers may rate specific aspects of user-system behavioral differences (e.g., number of turns, average time for turn exchange) and statistically control behavioral variability across user-technology interactions. Thus, although the user-technology interactions will differ across the user and yoked observer pairs, methods of statistical control will remove variance due to interaction differences, allowing the researcher to measure the effects of point of view more
effectively.

**Theoretical Implications**

In addition to their methodological implications, these studies also have several theoretical implications. In particular, the present results provide empirical support for pragmatic and social-cognitive theories, extending their relevance to a new applied discipline. The current findings highlight the relevance of conversational and service expectations for affective responses to e-service but underscore the weaknesses of a number of construct definitions for applied work.

**The Explanatory Role of Pragmatic and Social-Cognitive Theory**

The results of the present research reinforce the notion that pragmatic and social-cognitive theory provides useful explanations of observations and empirical findings in speech technology. As pragmatic theory suggests, social goals are realized and inferred through the use of language. Perhaps the strongest evidence that this view holds for speaking technologies is the Usability Scale’s inclusion of a number of items that refer to user goals and priorities for speech e-service usage (e.g., I could complete my business quickly using this system, The system seemed efficient, The system allowed me to do things that are important to me, The system seemed to know my needs). In addition, the Usability Scale also included items that refer to the power relationship implied by a speech system’s use of language (e.g., The system made me feel like I was in control) and appropriate adherence to conversational norms (e.g., The messages were repetitive, The system gave me more details than I needed). As with human-human dialogue, to be perceived positively, a speech system’s language usage should be in line with the context and expected conversational behavior. Thus, the content of the Usability Scale itself appears to add to the existing empirical support for the applicability of pragmatic theory to the design of high-quality conversational technologies (Bernsen, Dybkjaer, & Dybkjaer, 1996; Saygin & Cicekli, 2002).

Also consistent with the existing research, the Usability Scale included a number of items
concerned with impressions of speech and language usage. For example, items that measured the
naturalness, confidence, professionalism, and enthusiasm of the system voice reinforce the
importance of speech behavior in social impression formation (Holtgraves, 2002; Cargile, Giles,
Ryan & Bradac, 1994; Patterson, 1996). Although this finding has been demonstrated with human
speakers, including those with various types of disordered, accented, dialectal, and typical speech
(Clopper & Pisoni, 2004; Collins & Missing, 2003; Feinberg, Jones, Little, Burt & Perrett, 2005;
Flege, 1998; Flege & Fletcher, 1992; Munro and Derwing, 1995, 1998; St. Louis, Myers, Faragasso,
Townsend, & Gallaher, 2004), the current results extend these findings to non-human speakers.
Interestingly, the Speech Characteristics factor had the largest effect on customer satisfaction of the
four Usability Scale factors. This finding provides some empirical support for the potential
importance of speech characteristics over other aspects of interface behavior and is consistent with
previous results (Nass & Lee, 2001; Polkosky, 2003).

Although the current research did not directly examine automaticity, the automatic trait
inference literature would seem to provide a more intuitively satisfying and theory-based
explanation for the non-volitional perception of a speaking technology’s humanlike traits than
other alternative explanations. Nass and Lee (2001) acknowledged that individuals recognize that
their trait inferences about speaking technologies are inappropriate. This behavior may be an
example of postconscious automaticity, because individuals are aware that the speech interface is
the basis of their perceptions, although they cannot explain having humanlike perceptions about
technology (Bargh & Ferguson, 2000). Understanding the causal links between specific speech and
language behaviors and their resulting perceptions is an important priority for not only for basic
research, but also interface design. Ideally, interface designers should have an empirical basis for
designing specific speech and language variables and should be able to articulate the desired
affective response in the user. Conversely, it would seem that speech interface design offers an

88
excellent method for modeling human conversational behavior and controlling speech and language variables to experimentally examine social perception. Speech technology modeling may also be combined with new methods of cognitive research, such as functional magnetic resonance imaging, to more fully examine the impact of auditory, speech, and linguistic variables on brain activation and automatic trait inference.

Items excluded from the Usability Scale may also be explained by the social-cognitive and pragmatic literature. For example, the elimination of perceived recognition accuracy items (e.g., I felt frustrated because the system didn’t understand me; The system understood more than I thought it would) due to low reliability was a somewhat surprising outcome of usability scale development analyses. Anecdotal evidence in the speech industry has suggested that recognition accuracy is a frequent concern of engineers. It is asserted that users often reject poorly-designed speech interfaces due to poor recognition (e.g., “It doesn’t understand what I say”), even when recognition accuracy rivals that of human conversational partners. Despite this attribution of poor “comprehension” by speech systems, the current results suggest that the present measurement approach did not yield reliable ratings, possibly due to inaccurate human perception of others’ comprehension (Keysar, Barr & Horton, 1998; Keysar & Henley, 2002) or other misattributions (Hilton, 1995).

Several research streams suggest that individuals may have a number of inaccurate perceptions about and explanations for communication behavior. Keysar and Henley (2002) showed that speakers tend to overestimate listener comprehension during conversation. Keysar, Barr and Horton (1998) argue that both speakers and listeners process conversation egocentrically, assuming the clarity of communication and listener comprehension, until overt errors require adjustment to the stream of conversation. In a broader consideration of communicative impacts on social perception, Hilton (1995) argued that conversational norms are an independent variable that
affects attributional processes. Hilton’s attributional model of conversational implicature draws on Grice (1975) to assert that partners enter conversations with expectations about the interaction. When a speaker violates Grice’s (1975) conversational maxims, partners engage in “mental calculus” to arrive at a causal explanation of the speaker’s intention, and these attributions form the foundation of social perception and judgment in interpersonal settings. Accordingly, because the social cues provided by speech systems may be designed ambiguously or inappropriately, users’ attributions about the cause of communicative difficulties are likely to involve the system’s poor comprehension because its behavior does not appear to be contingent on the user’s utterance.

Interestingly, misattribution about comprehension also occurs in non-human communication. In non-human animal research, despite empirical demonstration of animal comprehension of language comparable to human children (Kaminski, Call & Fischer, 2004) and human perception of animal comprehension (Pongracz, Miklosi, & Csanyi, 2001; Sims & Chin, 2002), the attribution of mental states to animals remains a controversial and frequently rejected explanation of findings (Schilhab, 2002; Wynne, 2004). In human-computer interaction research, Nass and Moon (2000) also reject anthropomorphism as an explanation of their demonstrations of human social responses to computers. Applying social-cognitive and pragmatic theory to human-computer interaction means that users will expect technology to be cooperative and engage in appropriate conversational behavior. When these expectations are violated, it would be reasonable to assume that attributions about the cause of atypical behavior and negative affective responses would result.

*Expectations and Social Perception of Speech-Based e-Service*

Expectations seem to underlie many of the results observed in this research. As previously suggested, conversational norms such as the Gricean (1975) maxims are one important source of expectations about speech system behavior. Glover’s (1995) discussion of various types of
conversations and their associated rules suggests that all conversations are subject to a set of expectations that include role schema about the relative power relationship of the self and conversational partners. In addition, conversation includes a set of expected behaviors such as greeting, turntaking, closing, and repair of communicative breakdowns (Goffman, 1976; Sacks, Schegloff, & Jefferson, 1974; Schegloff, 1968; Schegloff, Jefferson, & Sacks, 1977; Schegloff & Sacks, 1973). Behaviors within a conversational script influence both participants’ social cognitive processing and subsequent behavior, a perspective that is broadly supported by the literature (Abelson, 1981; Holtgraves, 2002; Shank & Abelson, 1977). The current findings lend empirical support to the association between expectations and affective outcomes and extend these findings to the realm of human-computer conversation.

In addition to expectations about the structure and content of conversation, items included in the Usability Scale for Speech Interfaces also suggest that expectations for these technologies may originate from a variety of sources. For example, the behavior of human service providers seems to figure prominently in the Usability Scale (Alge et al., 2002; Cran, 1994; Holland & Baird, 1968; Yagil, 2001). Items that were included in the Customer Service Behavior factor included the friendliness, politeness, enthusiasm, and courteousness of the system, as well as its use of familiar terminology. However, in addition to these general expectations, users seem to require speech systems to be efficient. The general behavior of humans may also be a source of expectations, because the usability scale included items that measure the naturalness of the system voice and the extent to which it sounds like a regular person. Finally, the dominance of professional voice talent recordings for system voices in the speech industry seems to be in line with the Usability Scale’s inclusion of an item that measures the similarity between a system voice and those on radio and television.

In the services marketing literature, the role of expectations has been identified as an
important determinant of affective responses such as customer satisfaction. Researchers in this field argue that service quality is determined by the gap between expected and actual service (Parasuraman, Zeithaml & Berry, 1985, 1988, 1994). Several researchers have suggested that favorable affective outcomes of service occur when the provider and customer adopt coordinating roles in a ritualized script (Mohr & Bitner, 1991; Solomon, Surprenant, Czepiel & Gutman, 1985). As with conversational expectations, service expectations seem to also be important to the usability and affective outcomes of speech-based e-service.

These findings are particularly important for speech technology practitioners, because they provide parameters for interface design. One of the predominant strains of work in applied speech technology has focused on designing a unique personality into a speech system (Cohen, Giangola & Bargh, 2004; Kotelly, 2003), using an upbeat, energetic prompt presentation by a professional voice talent. This design strategy has been adopted based on Nass’ assertion that speech interfaces should be designed to have personality characteristics (such as introversion or extroversion) consistent with the personality of the user (Nass & Lee, 2001). In contrast to this approach, the current research indicates that systems that incorporate a constrained set of behaviors (those associated with the service provider role, such as helpfulness, friendliness, awareness of user needs) and are designed to adhere to a service script (Solomon, Surprenant, Czepiel & Gutman, 1985) are more effective than unconstrained personality (persona) design. Given the current state of the technology, it is impractical for a speech interface to “infer” or explicitly ask for users’ individual differences and then adapt itself to each unique individual in a user population. However, this behavior underlies the relatively simplistic but common design of many current systems that ask users to identify their preferred language for interaction (e.g., “To continue in English, say English”), a behavior that is often mentioned by individuals as a disliked aspect of e-service. Continued research in sources of expectations and direct examination of expectations will reveal
design strategies that appropriately accommodate users and result in desirable affective outcomes.

Refining Constructs to Accommodate New Applications of Theory

Although the development of an applied usability measurement tool was an important goal of this research, equally important was evaluating the potential variables that may have implications for speech technology research. A broad review of relevant literatures indicated that customer satisfaction was an important potential outcome of speech technology usage. It also identified a number of potential antecedents and related constructs. The present findings provide a new perspective on these variables and their applicability to speech technology.

The literature review also suggested that a number of variables are theoretically defined in very similar ways, including usability, service quality, user satisfaction, customer satisfaction, and perlocutionary force. The current study did little to resolve this issue but it did indicate that usability is strongly associated with customer satisfaction, an important assumption in speech technology that was not previously based on empirical findings. In addition, comfort was included in the present study because of its apparent similarity to perlocutionary force, but it was predicted by virtually the same set of usability items as customer satisfaction. Unfortunately, comfort did not appear to contribute a unique or additional perspective on affective responses to speech technology, except to highlight the importance of trust (Corritore, Kracher & Weidenbeck, 2003; Grabner-Krauter & Kaluscha, 2003). Due to the emphasis on customer satisfaction in industrial settings, research efforts should probably use this dependent variable rather than comfort for future work. To resolve unclear relationships among constructs, a useful extension of the current research would be to use advanced statistical techniques such as path analysis or structural equation modeling to compare several hypothesized models of antecedents and affective outcomes to e-service. An empirical approach to these construct interrelationships would be useful in addressing the lack of clarity in their theoretical boundaries.
One issue in defining the relevant constructs was clarified in these studies. The current work did not attempt to measure service quality. However, the Usability Scale for Speech Interfaces includes items that measure service expectations and has notable similarity to other service quality scales such as SERVQUAL (Parasuraman, Zeithamal & Berry, 1994) and e-SQ (Zeithaml, Parasuraman & Malhotra, 2000). These findings suggest that service quality and usability may be operationalized similarly for speech-based e-service. This perspective suggests that both usability and service quality are flexible, abstract, and contextually-defined constructs that depend largely on expectations of the user or receiver of service, as well as the type of service provided (e.g., Internet e-service, speech-based e-service, human provider).

Another question that guided this research was to determine whether customer satisfaction was more closely associated with individual differences or interface perceptions. The results indicated that need for interaction with a service provider, inherent novelty seeking, and self-monitoring are weakly positively correlated with customer satisfaction. In contrast, perceived usability variables showed strong positive correlations with customer satisfaction. Interestingly, the related disciplines seem to handle these antecedents in different ways: many marketing researchers seem to examine individual differences and human factors researchers appear to be more concerned with interface perceptions. The observed discrepancy in the relative associations among these variables is important for development of speech systems because it suggests that perceptions of speech interfaces should be a more important and practical consideration for interface designers than catering to user personality characteristics, contradicting some design recommendations (Nass & Lee, 2001).

The present research was also concerned with individual difference and interface perception as predictors of customer satisfaction. Wickens, Gordon and Liu (1998) identify prediction as a primary characteristic of human factors as a scientific discipline. They state that “it
is important to be able to predict that solutions that are envisioned to create good human factors will actually succeed when put into practice” (p. 7). A regression equation made up of largely individual difference predictors accounted for less variance in customer satisfaction than a similar equation with interface perceptual (usability) interface items. This finding also reinforces the relative importance of interface perceptions to speech technology affective outcomes. A practical outcome of this work is that the regression equations will provide expected customer satisfaction scores when only usability scores are known. This information will add to the value of conducting usability evaluations during development, already considered a critical best practice task for speech applications (Balentine & Morgan, 1999; Cohen, Giangola & Balogh, 2004; Kotelly, 2002).

Nonetheless, the weak correlations between individual differences and customer satisfaction and lack of interactive effects for quality and individual differences were somewhat unexpected. The methodology used to respond to research question five may have been partially responsible, however. According to Pedhazur (1997), this type of methodology, known as Attribute-Treatment Interaction (ATI) research, has a “paucity of findings of interactions” due to a wide variety of design problems, including weak effects (p. 585). It is likely that the interactive effects are the result of weak effects in the current research.

In spite of this problem, the potential interaction between user characteristics and interface design characteristics should not be dismissed due to the limitations of the current methodology. Interactive effects do appear to have practical significance because they may mean the differences such as customer satisfaction, repeat usage of speech-based e-service, and ultimately cost savings for enterprises that implement speech technology. In accordance with their doctrines of human factors, Sanders and McCormick (1995) suggest that individual differences should be considered in interface design, thus implying that user and system characteristics interact. Although the current results provide one approach to examining potential user-system variable interactions, research on
this topic (like ATI research in general) is still in its early stages (Pedhazur, 1997). Therefore, because the current research has developed a theoretical and empirical foundation for what to measure, future research efforts should begin to focus on experimental methodologies to explore these possible user-system interactions, as well as their impact on usability and customer satisfaction. Although this method has been attempted previously (Polkosky, 2003), its primary limitation was that there were no scales for measuring usability, nor a comprehensive understanding of important affective outcomes. The current work is an important step in this line of applied research.

**Future Research**

Although the present research began to establish a social-cognitive psychology by developing a basic measurement tool and illustrating the explanatory value of theory, it also highlights the significant amount of work that remains. An important future goal should be to expand the symbiotic relationships among basic research and theory in social psychology, psycholinguistics, and interpersonal communication; applied research in services marketing and human factors; and applied research in speech technology and conversational design. As Brennan (1998) observed, applied work with speech technologies will promote questions that challenge and strengthen theoretical models and basic theory will provide a foundation for applied design. Past research has seen each of the relevant fields progress largely in parallel but it is imperative that more cross-disciplinary research be undertaken to provide a robust knowledge base for both researchers and practitioners. The current studies have suggested several topics that could facilitate more sophisticated research in all of these fields and exploit cross-disciplinary findings.

In addition, the current research provided an empirical and theoretical basis for positing several important affective outcomes of speech technology, but its design did not allow for interpretation of causal relationships. Because these studies were the first to integrate variables
from different disciplines, it was a necessary step to identify which variables should be measured and to examine associations among them. Future researchers will continue to advance the field by manipulating speech, language, and social aspects of speech interfaces and measuring their impact on affective outcomes (e.g., customer satisfaction). For example, applied researchers may manipulate syntactic forms, lexical choices, gender of system voice, and messages provided by the system when a miscommunication occurs (Berger, 2001; Holtgraves, 2002) to determine their impact on customer satisfaction. Research designs that allow causal inferences will promote better user-system conversations and facilitate the next generation of naturalistic human-computer dialogues. Experimental studies following the examples provided by basic social psychology and interpersonal communication researchers are important because designers should have control over linguistic and speech variables in creating an interactive e-service and should be able to manipulate them to achieve desired user responses. As an example, the present study did not provide adequate variability in dialogue error, but this variable has considerable potential for future research. A researcher may manipulate types of errors depicted in auditory clips and evaluate their impact on observers, similar to Kreuz and Roberts’ (1993) methodology in which they manipulated articulation, syntactic, and pragmatic errors and measured attributions about which conversational partner caused the error.

Applying findings from the present research back to basic research is also an important future task. For example, speech interfaces provide an ideal medium for modeling social-communicative interactions and controlling them in ways that are not possible, difficult, or highly unusual with human interactants. Such modeling will also provide a means for testing theoretical models of interpersonal communication and cognition (Bradac, Cargile & Hallett, 2001; Patterson, 1996), a task that should be undertaken to better explain social impressions about speech systems. A paucity of research that empirically demonstrates causal relationships among specific

97
speech and language behaviors, observer (or interactant) perceptions, and affective outcomes exists in basic literature. Although it is generally accepted that speech and language variables are important to social perception and affective responses (Holtgraves, 2002), there are very few findings that help interface designers understand specifically how to combine interface variables to facilitate desired outcomes.

Another area challenged by the present research concerns the selection of interactants for the majority of interpersonal communication research. Much of this work is concerned with long-acquainted pairs or couples who are romantically linked (Beebe, Beebe, & Redmond, 2002), making a substantial portion of this research difficult to generalize to more casual, goal-oriented interactions. By contrast, service delivery appears to provide a constrained yet highly social and goal-oriented domain for investigating affective and cognitive impacts of both typical and atypical speech, language, and social behavior. Increased focus on a broad variety of conversational interactions will provide more robust findings for applied work and theory. Finally, applied work in speech technology underscores the importance and relevance of basic research in a number of areas that have had limited exploration to date, including communication failure (Berger, 2001) and automatic trait inferences (Bassili, 1976; Brown & Bassili, 2002; Ham & Vonk, 2003; Skowronski, Carlston, Mae, & Crawford, 1998).

As Sanders and McCormick (1993) observed over a decade ago, human factors is a dynamic, ever-expanding area of research and applied work. As is the case with other, more traditional areas of human factors, speech technology design will also benefit from application of basic knowledge about human social, cognitive and communicative behaviors to develop increasingly usable, intuitive, and pleasurable conversational interactions with technologies. Perhaps the best argument for continued research at the intersection of social cognition, communication, psycholinguistics, services marketing, and human factors was advanced by
researchers who may not have even foreseen this field: “The ‘brave new world’ of the future should indeed be developed with people -- all of us -- in mind. Thus, the human factors discipline must be at the cutting edge of future developments to ensure that such developments will, in reality, contribute to the improvement of the quality of working life and of life in general” (Sanders & McCormick, 1993, p. 754). It is to this larger goal that speech technology practitioners and researchers should aspire.
References


Inman, M. L., Reichl, A. J., & Baron, R. S. (1993). Do we tell less than we know or hear
less than we are told? Exploring the teller-listener extremity effect. Journal of Experimental Social
Psychology, 29, 528-550.

‘fast mapping.’ Science, 304, 1682-1683.

Rime (Eds.), Fundamentals of nonverbal behavior: Studies in emotion and social interaction. (pp.
200-238). New York: Cambridge UP.

River, NJ: Prentice Hall.


Psychological Science, 13(3), 207-212.

Boston: Addison Wesley.

communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), Advances in


communication. In E. T. Higgins & A. Kruglanski (Eds.), Social psychology: A handbook of basic
principles (pp. 655-701). New York: Guilford.


Appendices
Appendix A. Audio Recording Transcripts

Tennis Scoreboard (Interface A)

Operator: From <company name>… this is the Australian Open On-Demand Scoreboard. Which player would you like to hear about?

User: Andre Agassi.

Operator: The match is currently in progress. Andre Agassi is losing to Marad Saffin. Marad Saffin leads 7-6, 7-6, 5-6, with Marad Saffin serving. Andre Agassi leads love-15 in the current game. Would you like to hear about another player?

User: Goodbye.

Operator: Thank you for calling the <company name> on-demand Australian Open Scoreboard. Goodbye.

Available Flights (Interface B)

System: Thank you for calling the <company name> automated travel service system. You can say repeat or help at any time. Select flight status, available flights…

User: (interrupts) Available flights.

System: Do you have a preferred airline?

User: No.

System: Departure city?


System: Do you want Kennedy, LaGuardia, or Newark?

User: LaGuardia.

System: Approximate departure time?

User: 6:45pm

System: Was that 6:45pm?

User: Yes.
Appendix A (Continued)

System: Arrival city?
User: Fort Lauderdale.
System: Was that Fort Lauderdale Hollywood International Airport?
User: Yes.
System: Which day?
User: November 21st.
System: Was that November 21st?
User: Yes.
System: Please wait. There are 16 flights that match your criteria. The closest match is American flight 2431, departing at 6:45pm. It arrives in Fort Lauderdale Hollywood International Airport at 9:52pm. Say repeat, list previous flights, list following flights...
User: (interrupts) List previous flights.
User: (interrupting) Goodbye.
System: Are you sure you want to end this call?
User: Yes.
System: Thank you for calling the <company name> automated travel service system. Goodbye.

Retail Order (Interface C)

System: Thank you for calling <company name>. What’s your team number?
User: 447108.
System: Thanks! Welcome back. I see you have 3 items in your wish list and 2 items in your shopping cart. What would you like to do today?
Appendix A (Continued)

User: Go to shopping cart.

System: Great! Your shopping cart contains 2 items: a Florida Bobcats #26 jersey in extra large and a pair of Nike soccer shoes in size 11. Would you like to place an order?

User: Yes.

System: Alright. We have 10056 Ridgway Avenue as your shipping address and a Visa card with last 4 numbers 5714 and expiration date May 2006 on file. Would you like to use this shipping address and Visa card?

User: No

System: Ok, what would you like to change?

User: Visa card.

System: What type of credit card would you like to use?

User: Mastercard.

System: What’s the number?

User: 555501010022089640950

System: Expiration date?

User: September 2007

System: Ok, so your credit card is Mastercard with last 4 numbers 0950 and expiration date September 2007. Is this correct?

User: Yes.

System: Placing your order…. Your order is $284.85 with standard shipping. Your package should arrive within 10 business days. Would you like a receipt emailed to you?

User: Yes.

System: Is your email address still gil@hotmail.com?

User: Yes.

System: Great. Anything else we can do for you today?

User: Yes.

System: Is your email address still gil@hotmail.com?

User: Yes.

System: Great. Anything else we can do for you today?
Appendix A (Continued)

User: No, thanks.

System: Thanks for calling <company name>. We’ll look forward to serving you again. Bye!

Retail Purchase (Interface D)

System: Welcome to <company name> by phone. If you ever need help on any question I ask you just say help and I’ll provide additional information. Let’s get started. Please say “place an order, check order status, request a catalogue or search for an item.”

User: Place an order.

System: Say the product code or the product name.


System: The Krups Bravo Espresso Machine, the price is $49.95. Say details to hear...

User: Buy it.

System: Would you like to order another item?

User: Yes.

System: Say the product code or the product name.

User: Commuter Cup.

System: The 18 oz. Commuter Cup. The price is $14.95. Say details to hear more about this product.

User: Buy it.

System: Would you like to order another item?

User: No.

System: We have several popular items that compliment your purchase. Would you like to hear them?

User: Yes.
Great, I have a list of the most popular items that compliment your purchase. If you hear an item that interests you say buy it to add it to your shopping cart. You can also say details to get more information about the product, next to go to the next item or checkout to complete your purchase. Here we go. I have five items for you. The first item is stainless steel cappuccino espresso cups and saucers.

User: Detail

System: Enjoy coffee bar chic at home with these gourmet coffee cups and saucers, Euro styled double walled stainless steel handled cups keep cappuccino and espresso hot. Dishwasher safe, price is $16.95 per set. Each set…

User: Buy it.

System: I’ve added the stainless steel cappuccino espresso cups and saucers to your shopping cart. The second item is oval…

User: Check out.

System: The total cost of your purchases is $41.90. Will you be paying for this with a <company name> credit card, Visa, MasterCard…<fade out>

Credit Dispute (Interface E)

System: Thank you for calling <company name>. I can help you receive a copy of your credit report, dispute items on your credit report, or check the status of items you have disputed. Which of these are you calling for today?

User: Place a dispute.

System: Next, I need to get your social security number. What is your social security number?

User: 154-678-4932

System: Ok, now tell me your first and last name.

User: Kent Montgomery.

System: Finally, I need to collect your address, starting with your zip code. Please tell me your zip code now.

User: 75008

System: And now the street address?

125
User: 1601 Bethany Road.

System: Ok, now we’ll take a look at your report. Using the item number on your report, tell me the first item number you’d like to dispute.

User: Four.

System: We show that item number to be a listing for Dayton-Hutson. Is this correct?

User: Yes.

System: Ok, what’s the problem with this item?

User: It’s not mine.

System: Ok, I’ve sent notice to that creditor that you believe this account is not yours. Are there any more items you’d like to dispute?

User: Yes.

System: What’s the next item number you’d like to dispute?

User: Seven.

System: We show that item number to be a listing for Sears. Is this correct?

User: Yes.

System: Ok, what’s the problem with this item?

User: It looks fraudulent.

System: Ok, I’ve sent notice to that creditor that you believe this account is fraudulent. Are there any more items that you would like to dispute?

User: No.

System: Ok, the creditors receiving disputes have up to 30 days to respond. Once we have all the responses, we will mail the results to you. Thank you for calling <company name>.
Appendix A (Continued)

Package Delivery (Interface F)

System: For faster service, please choose from the following options: to schedule a pickup using your <company name> account number say or press one, to track a package using your tracking number say or press two, for international shipping say or press three, to find package drop off locations say or press four, for <company name> express US rates say or press five…

User: (interrupting) Five

System: You can interrupt the rating system by speaking at any time. Please say or enter your <company name> account number. If you don’t know your account number, say I don’t know and we will move along.

User: I don’t know.

System: Ok, I’ll give you a quote based on standard rates, but you can always talk to a customer representative who can look into this for you. Which package type are you sending? You can say envelope, pack, box…

User: (interrupting) box.

System: What is the approximate weight of your package? Please round up to the next pound.

User: Two pounds.

System: What 5 digit zip code are you sending the package from?

User: 33435

System: And what 5 digit zip code are you sending the package to?

User: 37411

System: Please hold while I calculate the rate for shipping your package without any special handling to a business address. The basic cost for <company name> priority overnight scheduled to arrive tomorrow by 10:30 am is $26.14. You can say read that again or rate another package, to find out about other delivery times or special handling say lets go on, for other FedEx shipping services say menu and if you are done say good bye.

User: Good bye

System: Thanks for calling <company name> rate finder, good bye.
Appendix A (Continued)

*Flight Status (Interface G)*

System: Thank you for calling the <company name> automated travel service system. You can say repeat or help at any time. Select flight status, available flights, My Flights, My Hotel, or exit.

User: Flight status.

System: Which airline?

User: American.

System: Flight number?

User: 153

System: Was that 153?

User: Yes.

System: Which day?

User: Tomorrow.

System: Please wait while Pat gets the information. For Thursday, November 7, 2002, American flight 153 departs Miami International from gate 2c at 7:52am. It arrives in O’Hare International airport gate A19 at 10:03am. American flight 153 departs O’Hare International airport gate A16 at 11am. It arrives in gate [unintelligible] at 3:30pm. Say repeat…

User: Goodbye.

System: Are you sure you want to end this call?

User: Yes.

System: Thank you for calling the <company name> automated travel service system. Goodbye.

*Phone Bill (Interface H)*

System: Thank you for calling <company name>, also on the web at <website>. To ensure quality service, your call may be monitored or recorded. To continue in English, press or say 1 now. [Spanish language prompt]

User: One
Appendix A (Continued)

System: This center is using a system that can understand you when you speak. If you are currently a customer, please say or key in the number you are calling about, beginning with the area code.

User: 972-539-2443

System: One moment while I access your account. Thank you. Please tell me briefly the reason for your call today.

User: I need to make arrangements for some payment.

System: Thank you. To verify your account, please say or key in your 6-digit customer ID, which follows your telephone number at the top of your bill.

User: (presses 6 digits on telephone keypad)

System: You have agreed to pay $102.31 by October 2, 2003. Would you like to hear your agreement again?

User: No, I need to change it.

System: Was that a yes?

User: No.

System: Do you want to change this agreement?

User: Yes.

System: $102.31 is due on your account. The last date I can give you to pay is October 2, 2003. Can you make payment by this date?

User: Yes.

System: What would you like to do? Pay each week, pay every other week, or pay the full amount in one payment?

User: (no response)

System: I’m sorry, I missed your response. If you can pay weekly, say weekly. If you can pay every other week, say biweekly. If you can pay full amount in one payment, say one payment.

User: One payment.
Appendix A (Continued)

System: You’ve agreed to pay full amount of $102.31 by October 2, 2003. Is this arrangement acceptable?

User: Yes, it is.

System: Please hold while I update your account. I’ll be right back with your confirmation number. Please note the following: Information service disconnection, reconnect fee, and a security deposit may apply if your payment is not received by October 2, 2003. Your confirmation number is 090410. Keep in mind that although an agreement has been made, you may continue to receive notices. When mailing your payment, please allow enough time for it to reach our office by the agreement date. Any new bills you receive while this agreement is in effect must be paid by the due date on the bill. If you exceed your credit limit at any time, your long distance service could be disconnected. Do you want to hear this information again?

User: No thanks.

System: If there is anything else I can help you with today, please state it so I may direct your call. Or you may hang up to end this call.

User: Goodbye.

System: You can say something like check my balance or billing questions or I’d like to order voice mail. What is the main reason for your call today? (call end)

Car Rental (Interface I)

System: Welcome to <company name> automated reservations. I can help you make, change or cancel reservations. What would you like to do?

User: Change reservations

System: Please say or enter your member number.

User: (enters on telephone keypad)

System: Member’s last name?

User: Lewis.

System: Changing reservation. What is your reservation number?

User: B as in boy, 56212, t as in tack, 4567.

System: I’m getting that reservation information, please wait. Is that the reservation you made to pick up at the Dallas-Fort Worth airport May 30 3:40pm?
User: Yes.

System: What would you like to change first?

User: The drop off date, June 8.

System: For example, you could say a different date like May 3 or a day like Saturday, today, or tomorrow.

User: June 8.

System: Change anything else?

User: The drop off time, 1pm.

System: Drop off at what time?

User: 1pm

System: Change anything else?

User: No.

System: Let’s review to be sure I heard you correctly. Are you planning to pick up the car at the Dallas-Fort Worth airport May 30 3:40pm and returning to the Dallas Fort-Worth airport June 8 1pm using your current profile?

User: Yes.

System: After applying all passes, fees, and discounts, the total charge of this reservation will be $390 with unlimited milage. Would you like to hear a breakdown of the charges?

User: No

System: Would you like me to make this reservation?

User: Yes.

System: The reservation number is B as in boy, 56212, C as in cat, 4567. Would you like to hear it again?

User: No.
Appendix A (Continued)

System: Is there anything else I can do for you today?
User: No, thank you.
System: Thank you for using the <company name> automated reservation system. Goodbye.
User: Thank you, goodbye.

Directory Dialer (Interface J)

System: <company name> directory dialing. What location and name?
User: West Palm Beach, Brent Davis.
System: West Palm Beach, Brent Davis. Tieline 975-2177. Other information?
User: No.
System: Choose outside line, tieline, fax number, pager number, pin code, department number, location, or connect. You can say new name, new location, or main menu to select another listing. Other information?
User: Connect.
(phone rings)

Patient Claim (Interface K)

System: Thank you for calling the <company name> Provider Support System. Please say the patient’s nine digit I.D. Number.
User: 119001234.
System: Coverage is in effect for this patient only. Would you like to check claim status, hear plan details or hear claim history?
User: Check claim status.
System: This patient has twelve claims to review. Would you like to check a specific claim or hear all claims?
Appendix A (Continued)

User:  Specific claim.

System:  What’s the date of service?


System:  This claim for $249.80 has been processed. It was paid on April 30th, 2003. Would you like to hear additional details for this claim?

User:  No.

System:  Would you like to check another claim status, hear plan details or hear claim history? You can hang up to end the call.

Movies (Interface L)

System:  Good morning, welcome to 1-800-555-tell. Get movie tickets before you find out it’s sold out! Say tell me more. Main menu. Here are all the categories you can choose from: stock quote, news center, sports, entertainment…

User:  Entertainment

System:  Okay, entertainment. Here are the choices in entertainment: movies, horoscopes…

User:  Movies

System:  Movies. Now you can buy tickets over the phone, powered by <company name>. Do you want to look for movies in or around Boca Raton, Florida?

User:  Yes

System:  Okay Boca Raton, Florida. Now looking for movies in and around that area. Remember you can always say go back to back up a step or start over to start your search again. If you know the name of the movie you want, just say it; to hear about new releases or other movie categories, say movie listings, for show times at your local theaters…

User:  Movie listings

System:  What kind of movie do you want to see? For a list of movie categories, say tell me my choices.

User:  Tell me my choices
Appendix A (Continued)

System: Here is a list of movie categories. When you hear the one you want, just say it: new releases, up-coming movies, drama, comedy, romance…

User: Up-coming movies

System: Here are the up-coming movies. When you hear the one you want, just say it: Signs starring Mel Gibson and Jauquin Phoenix, rated pg13. To hear the next movie, say next, to end this list say stop.

User: Next

System: Martin Lawrence Live rated R, full frontal…

User: Go back

System: What kind of movie do you want to see? For example, say comedy.

User: Comedy

System: Here are the comedies. When you hear the one you want, just say it: Austin Powers in Goldmember, starring Mike Meyers, rated PG13; Men in Black II starring Tommy Lee Jones…

User: Men in Black II

System: Here are show times for Men in Black II: it’s playing at Regal Shadow Wood 16, 9889 Glades Road in Boca Raton. You can see it today at 12:30, 2:45, 5 o’clock, 7:10 and 9:20. Advanced tickets are available at this theater. To buy tickets, say buy tickets; to hear show times for the next theater, say next. It is also playing at Regal Delray Beach 18, 1660 South Federal Highway, Plaza of Delray, in Delray Beach. You can see it today at 12:35, 2:50, 5 o’clock, 7:50 and 10:10. Advanced tickets are available at this theater.

User: Good bye

System: I heard you say good bye. If you don’t mean to hang up, say go back; otherwise, thanks for connecting, good bye.

Report Electrical Outage (Interface M)

Operator: Welcome to <company name>. Before I can help you I need to find out who you are. Are you calling from home today?

Customer: No.

Operator: Tell me your ten digit home phone number.
Appendix A (Continued)

Customer: 2018186600.
Operator: Great, we found your information. Let’s go to the main menu. You can report a power outage, check your balance or enter your meter reading. Which one would you like?
Customer: I have a utility outage at my house.
Operator: You can report a complete outage or flickering light. Which type of outage would you like to report?
Customer: My power’s out.
Operator: Thank you for reporting the outage and we will have the problem resolved within two hours. Thank you for calling <company name>, good bye.

Prescription Refill (Interface N)

Operator: Welcome to <company name> prescription refill service. Please say your name to access our service.
Customer: Tom Smith.
Operator: Thank you. We have accessed your record. You are due for a refill on Imitrex. Would you like to refill your Imitrex prescription?
Customer: Yes.
Operator: Is this for pick up or delivery?
Customer: Delivery.
Operator: Are you still at 5671 Round Tree Avenue?
Customer: Yes.
Operator: Thank you. Your Imitrex prescription will be delivered to your home address within three to five business days. Would you like another refill?
Customer: No.
Operator: Thank you for using <company name>, good bye.
Appendix A (Continued)

Financial Services (Interface O)

Operator: Welcome to <company name>. I’ll help you get solutions for all your financial services. All you have to do is talk to me. Let’s get started. So I can tell who’s calling, say your account number.

Customer: 94837273

Operator: What’s your password?

Customer: My voice is my password.

Operator: Your voice print has been verified. Hi, John. <company name> menu, tell me which service you would like to use? You can say banking, investments or mortgages. You can also say “my <company name>” to access the custom information you created on our web site.

Customer: Banking.

Operator: Banking. What would you like? You can say account balances, credit card information or say make a payment.

Customer: What’s the balance in my checking account?

Operator: Your balance as of today is $8257.33. For more information say last five transactions, last five checks…

Customer: Last five transactions.

Operator: Your last five transactions are all ATM transactions. I’ll list them; on January first a withdrawal in the amount of $200, on January third a withdrawal in the amount of $500, on January sixth a withdrawal in the amount of $100…

Customer: What’s the balance on my credit card?

Operator: As of January sixth your credit card balance is $3455.32. What next? You can say…

Customer: Make a payment to my credit card.

Operator: Pay how much?

Customer: All of it.
Appendix A (Continued)

Operator: Should I take the money from you checking account?
Customer: Yes.

Operator: Okay, $3455.32 has been paid from checking to your credit card. Your credit card balance is now zero dollars. These changes will show up immediately on your on-line account. What next?
Customer: Investments.

Operator: Investments. You can say account balance, stock quotes or place a trade.
Customer: Stock quotes.

Operator: Say the name of the security or its ticker symbol.
Customer: SunTrust.

Operator: SunTrust International, symbol STI is trading at 62 and 7/16 up 1 and ¼.
Customer: Buy two hundred shares of SunTrust from my checking account.

Operator: I think you wanted to buy two hundred shares of SunTrust International at a share price of 62 and 7/16 with funds from your checking account. Is that correct?
Customer: Yes.

Operator: Okay, I’ve made the purchase. Your confirmation number is 14432. I’ll send a summary of this transaction to your home address and your on-line account will reflect the change immediately. What next? You can say…
Customer: My <company name>.

Operator: I’ve got your personal <company name> page; you can say news, traffic, weather, sports, stocks, vacations or alerts.
Customer: Weather.

Operator: Today’s weather in Atlanta is sunny with a high of 62 degrees and a low of 54. For more information say “get details”.
Customer: Sports.
Operator: Sports in Atlanta today. The Braves just signed Chipper Jones to a five year one hundred twenty million dollar contract. You can say get details or say another category from your <company name> page.

Customer: Alerts.

Operator: You have one bill alert. Here it is; your alimony payment in the amount of $2117.34 is due today. Would you like to pay this bill now?

Customer: Yes.

Operator: From you checking account?

Customer: No, my savings account.

Operator: Okay, $2117.34 was paid from your savings account to Mandy Smith. Your on-line account has been updated. What next? You can say news, traffic, weather…

Customer: I need a vacation.

Operator: Vacation. Delta airlines has several special exclusive getaways for <company name> subscribers. Each week there’s one unique vacation in the United States and one at an international destination. Today’s US destination is Vale, CO…

Trade Stock (Interface P)

Operator: Welcome to <company name>. To get started, just tell me your account number.

Customer: 193601701.

Operator: And your last name is?

Customer: Johnson.

Operator: Great, I have accessed your account. What can I do for you today? Do you want to buy, sell, transfer or get prices?

Customer: What’s the price of America Online?

Operator: The current price of America Online is $12.53. What else can I do for you?

Customer: I would like to purchase five hundred shares please.
Appendix A (Continued)

Operator: Confirming the purchase of five hundred shares of America Online at $12.53. Please say yes to confirm the purchase.

Customer: Yes.

Operator: Is there anything else I can help you with today?

Customer: Nope.

Operator: Thank you for calling <company name>, good bye.
Appendix B. Participant Instructions

I invite you to participate in a study for my doctoral dissertation in psychology on how people perceive and evaluate speech user interfaces. The information you provide is strictly anonymous. Therefore, you will not need to provide your name or other identifying information.

In this study, you will listen to an interaction between a speech interface and a user, similar to listening to a telephone conversation on another line. After you listen, you will be asked to rate various aspects of the speech system. The questionnaire also asks you to provide information about yourself, such as your age, gender, job type, preferences for customer service, and social interaction style, so that I can better understand how different people view speech interfaces.

To hear the user’s interaction with the speech system, please double click on the following audio file:

[Click here to hear user-speech system interaction]

Participation in this study is strictly voluntary. Again, your answers will be completely anonymous. If you have any questions regarding this study, please do not hesitate to let me know at any of the points of contact below.

Thank you,

Melanie D. Polkosky
Advisory Human Factors Engineer,
IBM Pervasive Computing
8051 Congress Ave, Suite 2207
Boca Raton, FL 33487
internet: polkosky@us.ibm.com
phone: 561-862-2037 (TL: 975-2037)
fax: 561-862-2988 (TL: 975-2988)
Appendix C. Participant Demographic Questions

1. Age: ________
2. Gender: Male Female
3. Native language:
   - English
   - Spanish
   - Other (please describe);
4. How frequently have you used speech technology in your daily life?
   - Never
   - Seldom
   - Occasionally
   - Often
   - Frequently
5. Which of the following best describes you as a speech technology user?
   - Novice user
   - Expert user
6. Which of the following best describes your hearing ability?
   - Normal
   - Hearing loss or disability
7. What is your current job function?
   - Management
   - Administrative Staff
   - Engineer
   - Researcher
   - Other technical professional
   - Non-technical professional
   - Marketing
   - Sales
   - Other (please describe):
8. Highest educational level achieved:
   - High school diploma
   - Technical degree
   - Bachelor degree
   - Masters degree
   - Ph.D.
   - Other (please describe):
Appendix D. Dabholkar and Bagozzi (2002) Inherent Novelty Seeking Scale

<table>
<thead>
<tr>
<th></th>
<th>1 Strongly Disagree</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I am always seeking new ideas and experiences.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>When things get boring, I like to find some new and unfamiliar experience.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>I like to continually change activities.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>I like to experience novelty and change in my daily routine.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix E. Dabholkar and Bagozzi (2002) Need for Interaction with a Service Employee Scale

| 1. Human contact in providing services makes the process enjoyable for the customer. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. I like interacting with the person who provides the service. | | | | | | | |
| 3. It bothers me to use a machine when I could talk to a person instead. | | | | | | | |

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. In social situations, I have the ability to alter my behavior if I feel that something else is called for.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2. I have the ability to control the way I come across to people, depending on the impression I wish to give them.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. When I feel that the image I am portraying isn’t working, I can readily change it to something else that does.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I have trouble changing my behavior to suit different people and different situations.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I have found that I can adjust my behavior to meet the requirements of any situation in which I find myself.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Once I know what a situation calls for, it’s easy for me to regulate my actions accordingly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. I am often able to read people’s true emotions correctly (through their eyes).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. In conversations, I am sensitive to even the slightest change in the facial expression of the person with whom I am conversing.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. My powers of intuition are quite good when it comes to understanding the emotions and motivations of others.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I can usually tell when others consider a joke to be in bad taste, even though they may laugh convincingly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. I can usually tell when I’ve said something inappropriate by reading it in the listener’s eyes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. If someone is lying to me, I usually know it at once from that person’s manner of expression.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix G. Items for Usability Scale

<table>
<thead>
<tr>
<th>Item</th>
<th>1 Strongly Disagree</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The system made me feel like I was in control.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. When speaking to me, the system paused or hesitated unnaturally.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I could complete my business quickly using this system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. The system seemed efficient.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Some of the system’s responses were unexpected or surprised me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. The system said things in a short, brief manner.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. It was easy to figure out how to do what I needed with this system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. The messages were repetitive.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. It would be too time consuming to use this system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. The system gave me a good feeling about being a customer of this business.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. The system’s voice seemed confident.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. The system used terms I am familiar with.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. I felt frustrated because the system didn’t understand me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. The system was designed in a way that seemed to have my best interests in mind.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. The system was easy to interact with.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. I would rather use this system than a webpage.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. This system impressed me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. I could find what I needed without any difficulty.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. The system reminded me of a human operator or customer service representative.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20. This system used everyday words.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. This system reminded me of other technologies,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>such as touch tone phone systems or webpages.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. The system was organized and logical.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23. The system allowed me to do things that are important to me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24. The system’s voice sounded unusual.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25. I thought this system was interesting to listen to.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26. The system gave me more details than I needed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. If the system misunderstood me, it was easy to fix the problem.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. The system provided personal, customized information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

145
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>29.</td>
<td>I had to carefully consider what I should say.</td>
</tr>
<tr>
<td>30.</td>
<td>The messages I heard gave me help if I made a mistake.</td>
</tr>
<tr>
<td>31.</td>
<td>The system spoke at a pace that was easy to follow.</td>
</tr>
<tr>
<td>32.</td>
<td>The system would help me be productive.</td>
</tr>
<tr>
<td>33.</td>
<td>The system understood more than I thought it would.</td>
</tr>
<tr>
<td>34.</td>
<td>The system seemed polite.</td>
</tr>
<tr>
<td>35.</td>
<td>I felt like I couldn’t say anything this system might understand.</td>
</tr>
<tr>
<td>36.</td>
<td>I was able to easily understand everything this system said.</td>
</tr>
<tr>
<td>37.</td>
<td>I could trust this system to work correctly.</td>
</tr>
<tr>
<td>38.</td>
<td>I would be likely to use this system again.</td>
</tr>
<tr>
<td>39.</td>
<td>The system’s voice was pleasant.</td>
</tr>
<tr>
<td>40.</td>
<td>The system emphasized words in a way that helped me know what I was supposed to say.</td>
</tr>
<tr>
<td>41.</td>
<td>The system provided all the information I needed.</td>
</tr>
<tr>
<td>42.</td>
<td>The system was too talkative.</td>
</tr>
<tr>
<td>43.</td>
<td>This system seemed reliable.</td>
</tr>
<tr>
<td>44.</td>
<td>The system’s voice sounded like people I hear on the radio or television.</td>
</tr>
<tr>
<td>45.</td>
<td>I felt frustrated because I couldn’t understand what the system said.</td>
</tr>
<tr>
<td>46.</td>
<td>I felt confident using this system.</td>
</tr>
<tr>
<td>47.</td>
<td>The system gave me accurate information.</td>
</tr>
<tr>
<td>48.</td>
<td>Using this system required a lot of work or effort on my part.</td>
</tr>
<tr>
<td>49.</td>
<td>I could predict what the system would say before it said it.</td>
</tr>
<tr>
<td>50.</td>
<td>I felt rushed when I was thinking of what to say.</td>
</tr>
<tr>
<td>51.</td>
<td>The system seemed to know my needs.</td>
</tr>
<tr>
<td>52.</td>
<td>The system’s voice sounded mechanical or artificial.</td>
</tr>
<tr>
<td>53.</td>
<td>I knew what this system could and couldn’t do as soon as I started interacting with it.</td>
</tr>
<tr>
<td>54.</td>
<td>The system’s voice sounded like a regular person.</td>
</tr>
<tr>
<td>55.</td>
<td>Using this system was fun.</td>
</tr>
<tr>
<td>56.</td>
<td>I felt like I was talking to a person.</td>
</tr>
<tr>
<td>57.</td>
<td>Using this system would result in errors that would prevent me from getting what I want.</td>
</tr>
</tbody>
</table>
Appendix G (Continued)

58. The system used slang, common sayings, or other language I hear in my everyday conversations.
59. The system provided clear instructions on what I needed to do.
60. The quality of this system made me want to remain a customer of this business.
61. It would be easy to remember how to use this system if I didn’t use it for several days.
62. The system’s voice sounded natural.
63. The system seemed private and secure.
64. The system seemed courteous.
65. I felt like I had to wait too long for the system to stop talking so I could say something.
66. The system was helpful.
67. The system was easy to use.
68. The system seemed friendly.
69. The system’s speech was easy to understand.
70. The system said things in a vague, ambiguous, or unclear way.
71. I felt confident that the things I said were understood correctly.
72. The system’s voice sounded enthusiastic or full of energy.
73. The system pleasantly surprised me.
74. The system seemed professional in its speaking style.
75. This system seemed complicated to use.
76. The messages I heard helped me when I felt unsure about what to do.
### Appendix H. Modified Spake, Beatty, Brockman, & Crutchfield (2003) Comfort Scale

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uncomfortable</strong></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td><strong>Very uneasy</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Very much at ease</td>
</tr>
<tr>
<td><strong>Very tense</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Very relaxed</td>
</tr>
<tr>
<td><strong>Insecure</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Secure</td>
</tr>
<tr>
<td><strong>Worried</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Worry free</td>
</tr>
<tr>
<td><strong>Distressed</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Calm</td>
</tr>
<tr>
<td><strong>Turbulent</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Serene</td>
</tr>
<tr>
<td><strong>Troubled</strong></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>Peaceful</td>
</tr>
</tbody>
</table>
Appendix I. Eroglu and Machleit (1990) Customer Satisfaction Scale

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissatisfied</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Unpleasant</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I like it very</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>much</td>
<td></td>
<td>I didn’t like it at all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I didn’t like it</td>
<td></td>
<td>at all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
About the Author

Melanie Diane Polkosky received a Bachelor of Science in Education in 1992 from Millersville University of Pennsylvania and a Master of Science in Communication Disorders from The Pennsylvania State University in 1995. At Penn State, she specialized in augmentative and alternative communication for individuals with severe communication disorders. She completed her Clinical Fellowship in Speech-Language Pathology at All Children’s Hospital (St. Petersburg, Florida) and received the Certificate of Clinical Competence in 1996. She then served on a multidisciplinary pre-kindergarten diagnostic team for Pinellas County Schools and attended art school until entering the USF Psychology Ph.D. program in 1999.

In 2001, she began a year-long internship at IBM Voice Systems, which lead to full time employment in 2002. At IBM, she consults on human factors issues and designs speech technology systems for clients in a variety of industries. Melanie lives in Collierville, Tennessee with her husband, dog, and four cats.