Wired for Speech

How Voice Activates and Advances the Human-Computer Relationship

Clifford Nass and Scott Brave

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Preface ix
Acknowledgments xiii
A Note to Readers xix

1 | Wired for Speech: Activating the Human-Computer Relationship 1

2 | Gender of Voices: Making Interfaces Male or Female 9

3 | Gender Stereotyping of Voices: Sex Is Everywhere 19

4 | Personality of Voices: Similars Attract 33

5 | Personality of Voices and Words: Multiple Personalities Are Dangerous 47

6 | Accents, Race, and Ethnicity: It's Who You Are, Not What You Look Like 61

7 | User Emotion and Voice Emotion: Talking Cars Should Know Their Drivers 73

8 | Voice and Content Emotions: Why Voice Interfaces Need Acting Lessons 85

9 | When Are Many Voices Better Than One? People Differentiate Synthetic Voices 97

10 | Should Voice Interfaces Say "I"? Recorded and Synthetic Voice Interfaces' Claims to Humanity 113

11 | Synthetic versus Recorded Voices and Faces: Don't Look the Look If You Can't Talk the Talk 125

12 | Mixing Synthetic and Recorded Voices: When "Better" Is Worse 143

13 | Communication Contexts: The Effects of Type of Input on User Behaviors and Attitudes 157

14 | Misrecognition: To Err Is Interface; To Blame, Complex 171

15 | Conclusion: From *Listening to* and *Talking at* to *Speaking with* 183

Notes 185
Author Index 271
Subject Index 285

Preface

Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship represents the culmination of ten years of research into the psychology and design of voice interfaces by Clifford Nass and his coworkers at Stanford University. Scott Brave joined the laboratory as a Ph.D. student in 1999 and, since graduating in June 2003, has continued as a postdoctoral scholar and integral member of the team.

Until 1996—when Cliff's first book (coauthored with Byron Reeves), *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places,* appeared—the laboratory focused on textual and graphical user interfaces. Soon after, many companies and researchers turned their attention to interfaces that talked and listened. Voice-interface technology was not quite ready for prime time, but most agreed that it was "worth watching." Computers were routinely shipping with speakers; using interactive voice systems was proving to be more cost-effective than hiring additional workers for call centers; and efficient and inexpensive technologies for digital signal processing enabled voice interfaces to appear in cars, toys, and appliances.

Unfortunately, because little theoretical research or practical guidance was available on creating successful voice interfaces, most voice interactions with technology were frustrating and ineffective. The juxtaposition of the clear need for better interfaces and the opportunity for fundamental research in an intriguing and unexplored area was irresistible to our lab. Furthermore, the lab was in a unique position to make significant contributions: its focus has always been the *social* aspects of human-technology interactions, and nothing is more social than speech.

Unfortunately, building voice interfaces with which to test our theories was difficult. For the most part, technologies that could talk and listen required a tremendous amount of nurturance from engineers and were expensive and difficult to use. Even if we could assemble the requisite staff and funding, we could not craft new experiments with the necessary speed. The impediments seemed insurmountable until 2000,

1 Wired for Speech: Activating the Human-Computer Relationship

Speech is the fundamental means of human communication. Even when other forms of communication—such as writing, facial expressions, or sign language—would be equally expressive, (hearing) people in all cultures persuade, inform, and build relationships primarily through speech.[1]

Perhaps the greatest proof of the importance of speech comes from the deep and broad ways that humans have evolved to process and understand speech. Speech is such an integral part of being human that people with IQ scores as low as 50 or brains as small as 400 grams (one-third the size of a normal human brain) can speak.[2] Although scientists debate whether the brain contains a "speech organ,"[3] there is no question that speech implicates more parts of the brain than any other function.[4] Humans are so tuned to speech production and processing that from about the age of eighteen months, children on average learn eight to ten new words a day and typically retain that rate until adolescence.[5]

Humans are also the only species that is wired to understand speech fully. It has long been known that the left side of the brain (which corresponds to the right ear) shows a clear advantage in processing the hearer's native language, nonsense syllables, speech in foreign languages, and even speech played backwards, while the left ear attends to all other sounds.[6] New research suggests that even the ears themselves are different. In the right outer ear, hair cells, which amplify sounds that are transmitted to the acoustic nerve in the brain, react more to sounds that reflect speech than do hairs in the left outer ear.[7]

This specialization appears very early in human development. One-day-old infants respond differently to speech-like sounds than they do to any other sounds.[8] By four days after birth, babies' brains automatically distinguish and prefer the sounds of their native language over the sounds of other languages.[9] By the teens, humans can perceive speech at the phenomenally fast rate of up to forty to fifty phonemes (the

smallest distinguishable speech sound) per second, while other sounds become indistinguishable at twenty phonemes per second.[10]

Brains Are Voice Experts

Speech does more than simply transmit words from a speaker to a listener.[11] Humans have evolved voices and listening apparatuses that convey a wide range of socially relevant cues that human brains are wired to analyze and respond to.[12] Indeed, some scholars argue that language evolved primarily to exchange social information rather than information about the environment.[13] For example, because information concerning gender[14] is of great evolutionary importance, people rapidly categorize voices as male or female based on pitch, cadence, and other factors.[15] Even if people change their minds about whether they are listening to a male or a female, the gender they assign to the voice influences their interpretation of everything that is said.[16] Other parameters, such as speech rate and volume, convey more subtle human characteristics, such as personality, emotion, and hometown: extroverts,[17] excited people,[18] and people from New York City[19] speak much more rapidly and more loudly than average.

Humans are so attuned to vocal characteristics that they quickly and accurately distinguish one person's voice from another.[20] Even before birth, a fetus in the womb can distinguish its mother's voice from all other voices (demonstrated via increased heart rate for the mother's voice and decreased heart rate for strangers' voices).[21] Within a few days after birth, a newborn prefers his or her mother's voice over that of a stranger's[22] and can distinguish one unfamiliar voice from another.[23] By eight months, infants can tune in to a particular voice even when another voice is speaking.[24]

The words that people select also carry social information. Sentences that use the first person (such as "I made a mistake" or "I would like the following information") evoke different meanings and feelings than sentences that avoid the use of "I" (such as "Mistakes were made" or "The following information should be provided").[25] Similarly, a person can address a misunderstood comment by taking responsibility ("I'm sorry. I didn't understand that"), blaming the speaker ("Be more articulate"), or scapegoating ("The transmission tower is having problems. Could you please repeat that?"). These differing approaches have dramatic consequences.[26]

Of course, voices and spoken words are not independent.[27] Consider how odd it would seem if a deep, booming voice asked, "Could I possibly ask you if you wouldn't mind doing a tiny favor?" or a high-pitched, soft voice announced, "You had better help me right now!"[28]

As a result of human evolution, humans are automatic experts at extracting the social aspects of speech.[29] People do not receive formal training to discern social cues from voice. Even distinguishing between subtle wording differences[30] and deciphering words with multiple meanings[31] seem to be in-born human abilities.

The ability to extract social meaning from speech is not a parlor trick or the accidental outgrowth of the ability to process sound waves. Every society provides rules about *how* to use categories of voices and words to guide attitudes, thoughts, and behaviors.[32] These rules, whether evolved or culturally selected, provide systematic guidance for determining gender-, personality-, and emotion-specific actions. These rules also advise people whom to like, whom to trust, and with whom to do business.

Sensitivity to voice and language cues has played a critical role in interactions with other people for as long as humans have lived in social groups. Those in the past whose brains automatically and easily responded to social cues in voices had an enormous evolutionary advantage. People with social intelligence were better judges of whom to trust and with whom to mate and were better able to convince others to trust them and mate with them.[33] By contrast, those who spent time struggling with the identification of social cues could not maximize opportunities. The ability to quickly classify others and use those judgments to guide behavior predicts success in all aspects of life. The ancestors of current humans—those who won the evolutionary battle—were equipped to learn and apply the social rules of voice.

In sum, over the course of 200,000 years of evolution, humans have become *voice-activated* with brains that are wired to equate voices with people and to act quickly on that identification.[34] Talking, listening, and human society have elegantly coevolved into a remarkably interwoven, effective, and stable system.

New Media, New Rules?

Evolution resolved many of the complex problems of voice communication among humans. Recently, however, tricky new technologies have emerged that can produce and understand voices. People now routinely use voice-input and voice-output systems to check airline reservations, order stocks, control cars, navigate the Web, dictate memos into a word processor, entertain children, and perform a host of other tasks. Voice user interfaces complement and at times replace graphical user interfaces, freeing people from the constraints of "WIMP" (windows, icons, menus and pointers).[35] Handheld and mobile devices, televisions, and household appliances can be controlled by spoken commands. Ubiquitous computing[36]—access to all information for anyone, anywhere, at any time—relies on speech for those whose eyes or hands

are directed to other tasks (such as driving, holding, or building) or for those who cannot read or type (such as children, the blind, or the disabled).

At the same time, however, these interfaces create an interesting problem. Suddenly people's successful and stable perception of voices as intrinsically part of the social world is misguided because they are conversing with technologies as well as with people. *How will a voice-activated brain that associates voice with social relationships react when confronted with technologies that talk or listen?*

This book demonstrates that the conscious knowledge that speech can have a non-human origin is not enough for the brain to overcome the historically appropriate activation of social relationships by voice. As a number of experiments will show, the human brain rarely makes distinctions between speaking to a machine—even those machines with very poor speech understanding and low-quality speech production—and speaking to a person. In fact, humans use the same parts of the brain to interact with machines as they do to interact with humans.

Listeners and talkers cannot suppress their natural responses to speech, regardless of source. People draw conclusions about technology-based voices and determine appropriate behavior by applying the same rules and shortcuts that they use when interacting with people. These technologies, like the speech of other people, *activate* all parts of the brain that are associated with social interaction.

As a result of these automatic and unconscious social responses to voice technologies, the psychology of interface speech is the psychology of human speech: voice interfaces are intrinsically social interfaces. Designers must create voice interfaces for brains that are obsessed with extracting as much social information as possible from speech and with using that information to guide attitudes and behaviors.

Because humans will respond socially to voice interfaces, designers can tap into the automatic and powerful responses elicited by all voices, whether of human or machine origin, to increase liking, trust, efficiency, learning, and even buying.[37] Using insights from both traditional social science and new research on how people interact with technology, this book answers critical questions concerning the future of voice interfaces. For example, how will people respond to an e-commerce application with a female voice? (Learn from the research on human females who sell products; see chapter 3.) Should an automated call center apologize when it can't understand a spoken request? (Leverage the finding that modest people are interpreted to be likeable but unintelligent; see chapter 14.) Should a car's voice sound enthusiastic or subdued? (Voices, like people, are more effective when they match the emotion of the listener; see chapter 7.) And when would an interface benefit from multiple voices?

(People who are distinguished as specialists are perceived to be more knowledgeable than generalists; see chapter 9.)

How Dare You Claim That?

If people respond to technology-based voices in the same way that they respond to people, then prediction and design might seem to be trivial. Theoretical issues and design questions could be resolved in the social science section of the library: simply find a description of a similar situation among humans, and apply the results. Unfortunately, this approach has four limitations.

The first limitation of this approach is that psychological theories and design prescriptions compete sometimes. For example, will people like voice interfaces with personalities similar to their own (the "birds of a feather flock together" principle[38]), or will they prefer voice personalities that complement their own (the "opposites attract" principle[39])?[40] Experimentation provides a rigorous way to resolve such contradictory principles.

A second limitation of relying on previous research is that the scientific literature may be silent on important design or theory questions. For example, would a voice interface be more persuasive when using the first person ("I have a beautiful lamp for sale") or the third person ("There is a beautiful lamp for sale")? The previous research literature ignores this question (for the answer, see chapter 10). Furthermore, new technologies suggest questions that could not previously be addressed or even asked. For example, no humans exhibit the bizarre speech patterns that are ssociated with synthetic speech (chapter 2), nor would any normal humans consistently exhibit a mismatch between the emotions that are in their voice and the words that they are saying (chapter 8), problems that are uniquely associated with voice interfaces. The systematic experimentation described in this book addresses these questions.

A third reason for reaching conclusions from experiments rather than merely relying on previous research is that all theories have boundaries and limitations. Even if a theory is essentially correct, it will not apply in every situation. For example, accents that match the listener's accent lead to greater trust.[41] However, in some cases, using an accent that matches the user is the wrong strategy (chapter 6). Experimental evidence and interpretation of the evidence allows definitive statements to be made that specify when and why particular theories and designs work or fail.

The final reason for performing experiments is that they provide the necessary resistance to the temptations of anecdote. When watching a usability test, everyone notices and remembers the "cute" remark or the "ideal" behavior. These are seductive

and rhetorically convincing but can frequently mislead. To ensure that an idea is valid or a design is effective, results from numerous experimental participants must be systematically and rigorously analyzed.

In this book, common theories and assumptions are thoroughly tested via twenty experiments with thousands of participants. The results are frequently surprising and often turn common understandings and design practices on their heads. To ensure that the theories were rigorously tested and could be applied confidently across a range of products and services, the research included a wide variety of contexts and technologies. The participants in the research learned, bid in auctions, disclosed personal information, listened to news, creatively answered questions, received advice, and heard jokes. Some people worked directly with a computer; others used the telephone, the Web, a wireless array microphone, or a driving simulator. The interfaces sometimes only talked, sometimes only listened, and sometimes did both.

Moving Forward

By communicating via the method that humans have evolved to use, voice interfaces should represent an extraordinarily pleasant and effective way to interact with technology. They conveniently fit in with the user's environment, providing access to information products and services through ubiquitous technologies such as telephones. Interaction through voice also frees up users' hands, eyes, and legs, enabling them to concurrently perform other tasks, such as driving. Finally, voice interfaces provide significant ergonomic advantages over traditional interfaces, as they do not require users to sit in fixed positions or repeatedly perform unnatural physical actions, such as typing or mouse clicking.

Given the elegant ways that voice interfaces could fit into people's lives, how is it possible that they have become notorious for fostering frustration and failure rather than encouraging effectiveness and enjoyment? Designers often blame limitations of technology as the underlying cause: "Yes, speech recognition can be frustrating," they say, "but that's because it's just not good enough yet," or "People have such unrealistic expectations about voice user interfaces that they're doomed to be frustrated," or "Advanced interfaces are intrinsically complicated; people have to accept a learning curve before they become proficient." While there is certainly validity to these arguments, people are remarkably willing and able to interact with nonnative speakers or young children, chat via a noisy phone line, or listen to poor-quality audio speakers—situations that have the same, if not more, limitations than computer technologies present.

As this book demonstrates, voice interfaces can be significantly improved by a careful understanding and application of how people are built for speech. Each of the following chapters first describes in detail how an understanding of the wiring of the human brain from birth (and even before!) informs how people think, feel, and behave. The chapter then describes, via a participant's view, one or more experiments that provide grounding and nuance for the way people respond to speaking technology. (For the detail-oriented, all of the measures, data tables, and statistical analysis can be found in the notes for each chapter.) The results of the experiments are discussed in terms of both human psychology and their application to the creation of more likeable, effective, and engaging products and services, including automated call centers, personal computer software, e-commerce Web sites, vehicles, home appliances, and toys.

The fundamental insights obtained from the theories, experiments, and applications discussed in this book will help designers build better interfaces, scientists construct better theories, and everyone gain better understandings of the future of machines that speak with us.

2 Gender of Voices: Making Interfaces Male or Female

If you say to someone, "Please write a description of yourself for someone who knows nothing about you," the first sentence will likely be "I am female" or "I am male."[1] Information theorists might argue that this is not a wise choice: the responder would be most informative by mentioning his or her most unusual characteristic.[2] However, for biological, psychological, and cultural reasons, the definition of self begins with gender.[3] Indeed, in many languages, including English, it is almost impossible to talk about a person without indicating gender through the use of pronouns (such as *he* or *she*) or other gender markers.

From the age of two or three years old, children identify themselves and others as exclusively belonging to a group called "female" or a group called "male."[4] In all cultures, the majority of people spend the majority of their time with people of the same gender,[5] and even very young children show a striking tendency to segregate by gender when choosing with whom to play.[6] This sense of membership provides the lens through which every perception, of both observer and observed, is filtered.[7]

Generally, people have to know each other's gender before they can interact comfortably.[8] Individuals with unclear gender, such as "Pat" on the television show *Saturday Night Live*, become "evocative objects"[9] that create inner turmoil and disturb and fascinate through their ambiguity.

Given the prominence of gender and voice in everyday life, it is not surprising that voice plays a key role in identifying gender.[10] Six-month-old infants are able to categorize voices into female and male,[11] and between the ages of eleven and fourteen months, infants learn to recognize the association between the gender of voices (women have "higher voices")[12] and people shown in photographs, suggesting that gender rapidly crosses sensory modalities.

Gender is one of the first attributes that people identify in a human voice; it is generally discernible within seconds.[13] Although people might simply think of females as the "higher-pitched gender,"[14] the assignment of gender is based on a number of voice

characteristics,[15] all of which are analyzed in language-related areas of the brain.[16] Indeed, the brain is so cued to gender categorization that when people are asked how similar voices are, the most important criterion is whether the voice is male or female.[17] Remarkably, even when pitch (fundamental frequency) and the range of the natural voice spectra (characteristics of the sound waves) are equalized, people can nevertheless identify gender.[18]

A further combination of biology and culture (the relative extent of each is hotly debated) may then prescribe other gender characteristics that distinguish the speaking of females and males. For example, in the United States, women's voices tend to exhibit wider pitch range, more expressiveness, and more sentences with rising pitch than do men's voices.[19] In addition, female speech tends to include more questions and a greater proportion of social and relational information than male speech.[20]

The unrelenting identification of and prescriptions for gender in one's own life, in others' lives, and in the larger social structure make the determination of whether another person's gender is the same as or different than one's own the most powerful and basic judgment a person can make. *Social identification*—the decision of whether someone is "like me" or "not like me"[21]—is critical to a fundamental human principle: the more similar two people are, the more positively they will be disposed toward each other. Thus, social identification leads to greater trust, liking, perceived intelligence, and other positive judgments:[22] "socially identified with" means "socially favored by." The mere awareness of categories is sufficient to encourage people to "choose sides" and exhibit in-group bias.[23]

Because of its fundamental role in understanding human behavior, various social scientists have provided explanations for the link between social identification and positive attitudes and behaviors. Social psychologists point to predictability:[24] the more similar people are to you, the more easily you can predict their behavior, and hence the safer you feel with them.[25] Cognitive psychologists argue from human preferences for "cognitive economy":[26] the more similar people are to you, the more readily you can understand their thoughts and behaviors without contemplation. Anthropologists claim that throughout most of human evolution, humans lived in highly interdependent, homogeneous small groups that shared shelter, food, rituals, and language:[27] similarity meant social support. Sociologists note that societies tend to encourage interactions between those who are similar:[28] similarity implies familiarity, and familiarity leads to liking.[29] Although mating certainly requires opposite sexes, even evolutionary psychology can explain the general rule of "similar is positive." Reproductive success can be defined as the presence of one's genes in the next generation. Because similarity is associated with similar genes, a similar other's

reproductive success is almost as valuable as an individual's own reproductive success.[30]

The desirability and simplicity of same-gender interactions is one of the most consistent findings in the social science literature. Indeed, research demonstrates that children as young as three years of age like their own gender more than the other[31] and attribute more positive attributes to their own gender than the opposite gender.[32]

How can these discoveries transfer from the world of human-human interactions to human-technology interactions? People do not spend a majority of their time with *technologies* that are male or female, and use of most technologies does not involve any requirement to consider gender. Technologies do not have biology of male and female, although people sometimes refer to certain machines with gender labels (for example, boats as "she"). Technologies also do not think differently based on their "genders" and have not been socialized as female or male by spending time with others of that gender.

Gender and Technology-Generated Voices

Because of limitations of storage space (digital recordings are large), processing speed (finding and combining arbitrary utterances can be slow), bandwidth speed (sound files do not transmit gracefully over a 33 kilobyte phone line), dynamism of content (all of the Web's content cannot be spoken and recorded in real time), and other technical constraints, much of the speech that is and will be produced by computers, the Web, telephone interfaces, and wireless devices will be "synthesized" speech, also known as text to speech (TTS).[33] Thus, more and more interfaces will not simply present fully recorded words and phrases but will produce speech artificially.

While synthesized-voice systems effectively communicate content, even the best synthesized-speech systems do not yet match the quality and structure of natural human speech.[34] Speech systems tend to have inexplicable pauses, misplaced accents and word emphases, discontinuities across phonemes, and inconsistent structures that make them sound nonhuman. All of these cues remind users that they are not interacting with something that has an intrinsic gender, just as a broken film during a screening reminds people that "It's only a movie." In other words, synthetic-speech systems should make apparent that the vocal cues that suggest gender are artificial and therefore not worthy of interpretation.

On the other hand, the human brain evolved with a very liberal definition of *speech*.[35] People even process nonsense syllables and speech played backward as if they were normal speech.[36] All speech is regarded as a communicative act,[37] and people will

struggle through assigning meaning to sounds even when they are garbled or unclear. Thus, the wiring of the human brain suggests that despite the many reminders that gender has nothing to do with synthetic speech, people might forget that they are working with a machine. This could lead people to take seriously cues, such as pitch, that lead them to socially identify with the voice.

The question is whether, given all the ways in which a computer (or other technology) indicates gender neutrality—a clearly mechanical voice and a box instead of a body—can the automatic activation of gender assignment by voice, combined with a cultural emphasis on gender differentiation, overcome the clear reminders that a user is working with a technology rather than a person?

Confronting People with "Machine" Gender

To answer this question, a variant of a traditional social psychology experiment was conducted.[38] Of the forty-eight adults who participated in the study, half were female and half were male.

For this study, participants read descriptions of situations in which a person had to choose between two courses of action: the classic dilemma. Psychologists study choice dilemmas[39] because situations with two options are the easiest to analyze statistically.[40] The dilemmas for this experiment were written so that they would not have an obvious (that is, a clearly more desirable or socially appropriate) choice. This design made it possible for the voice to influence people to accept one option or the other.

Here is an example of a typical choice dilemma, which was presented in text for the participant to read:

> Amy and John are college students who have been living together in an apartment near campus. John's allowance buys food, and they are sharing the rent. Amy has told her parents that she is rooming with another girl, and now her parents are coming to visit their daughter. They have never seen the apartment. Should Amy ask John to move out for the time that her parents are in town?

The dilemma in this case is whether Amy should or should not ask John to move out. After reading the dilemma, the participant clicked on a button, which caused an obviously computer-synthesized voice to provide a recommendation about which course of action should be taken. The computer always made a forceful argument for one of the two possible options. For example, here is the computer voice's argument, presented to all participants, for the moving-out dilemma:

> She should ask him to leave for a while. If she tells the truth to her parents, it would cause a lot of unnecessary trouble. She can always confess her situation to her parents later when she feels that it is the right time.

Half of the male and half of the female participants heard a voice that was computer-generated but "female"; the other half heard a voice that was computer-generated but "male."[41] The primary distinction between the voices was fundamental frequency (pitch): the "female" voice had the average fundamental frequency of human females (210 Hz), and the "male" voice had the average fundamental frequency of human males (110 Hz).[42] The voices had similar characteristics (such as their volumes and frequency ranges) to control for the possibility of effects from personality, accents, and emotion.[43] The voices gave the identical advice for each scenario. The voices were selected to be clearly artificial and sounded nothing like actual people. They had bizarre accents, phrasings, and mispronunciations, expressed no emotion or understanding of the content, and had a "mechanical" sound. Nonetheless, all participants indicated that they could fully understand what the voice said.

The research strategy was to give half of the males and half of the females an opportunity to "socially identify" with the voice whose gender "matched" their own. If people do not socially identify with machine voices, then the gender of the voice would prove irrelevant. If, however, the gender of a voice—even a clearly machine-like voice—activates the brain's obsessive focus on gender, then a synthetic voice would elicit the same social identification effects that a human voice does. Thus, for females, a synthetic female voice should be perceived as more credible and likeable than a synthetic male voice, while the opposite should be true for males.

After listening to the computer's argument in support of one of the two possible positions, participants indicated their opinion about what the person confronted with the dilemma should do. Users' recommendations were given on an eight-point scale from "Definitely do option A" ("Ask John to leave") to "Definitely do option B" ("Not ask John to leave"). The scales displayed an even number of points so that people had to take a stand: there was no middle point. If people identified socially with these bizarre voices, then people would agree more with what the voice suggested when it "matched" their gender than when it did not match.

One way to study this phenomenon would be to ask people their opinion before hearing the voice and then ask again after they heard the voice. This would seem to be a more precise measure than assuming that prior to the experiment, all people had the same average opinion. Unfortunately, psychologists have discovered that this strategy is ineffective, perhaps because people often are reluctant to change their answers once they are "on record."[44] Hence, the average responses of people who heard a similarly gendered voice were compared to the average of those who heard a differently gendered voice.

Focusing on a single scenario is risky. A given scenario might be one for which women always opt for option A while men always opt for option B, thereby masking

the effects of voice. More generally, the scenario might be one that triggers strong personal reactions that would make people immune to persuasion. Conversely, researchers might be "lucky" and find the only scenario that elicits a particular phenomenon; basing their conclusions on that single scenario would be misleading. Thus, the norm in research is to choose a range of choice dilemmas (or any type of scenario) to make sure that none is problematic and then to average across the scenarios.[45] (In the current research, six different scenarios were averaged.)

When participants finished going through all six decisions—first hearing the computer's opinion and then making their own decision—they completed a questionnaire.[46] The questionnaire was designed to provide further evidence for social identification and to answer the most basic question of whether people noticed that the voice was "female" or "male." First, participants were asked to rate the trustworthiness of the voice that they heard.[47] If people social identify with synthetic voices, the participants would indicate more trust in the "matched" voice than the "mismatched" voice. Second, participants were asked to rate the likeability of the voice that they heard.[48] This was important both because social identification predicts liking and because likeability is often found to be highly correlated with user satisfaction and with likelihood of purchase and reuse. Finally, participants were asked to rate the masculinity or femininity of the computer voice that they heard.[49]

Does Machine Gender Equal Human Gender?

We found that users clearly distinguished voices based on gender cues,[50] rating the male synthesized voice as significantly more masculine than the female synthesized voice.[51] This demonstrates that the voices activated a categorization that is profoundly associated with living things: sex. The automatic assignment of gender to voices thus extends beyond the realm of humans to the technological world.

Despite all of the signals that indicated to the participants that the processes that underlie social identification are not applicable to the voices they heard, participants' choices conformed more to the suggestions made by the voices that "matched" their gender than did the choices of participants who heard voices that did not "match."[52] Female participants found the female voice to be more trustworthy than the male voice, while males felt the opposite,[53] another indicator that social identification can lead to persuasion.

The liking measure also revealed effects indicating social identification. Males liked the male voice better than the female voice, while female users liked the female voice better than the male.[54]

When we were presenting these results at a conference, a woman in the audience raised her hand and asked, "Women often have the experience that when they say something a meeting, people don't seem to listen to them, but when men say the same things, people seem to pay attention. Did that happen in your research?"[55] We *did* pay attention to her and rushed back to the data to see if her experience was corroborated. Sure enough, in addition to the social identification effects, we also found a simple effect from the gender of the voice: people conformed much more with the male voice than the female voice, even though the male and female voices made identical statements.[56] Listeners also found the male voice to be more trustworthy.[57] This meant that males listening to the male voice were the most convinced by the voice, and males listening to a female voice were the least convinced. Sadly, this tendency for people, especially males, to take males more seriously than females extended to the obviously synthetic world of technology. This was not a result of simple liking of one voice over another (the male voice was not more likable overall than the female voice) but was a reflection of learned social behaviors and assumptions.

Extending Gender in Voice Interfaces

Gender of voice is as prominent and definitive in the technological world as it is in the social world: users treat synthetic voices as if they reflected the biological and sociocultural realities of sex and gender. Just as any voice that sounds even remotely like human speech is processed as human speech, any voice that sounds remotely male or female is likely to be automatically classified as such. Even a hint of gender, as manifested through the pitch of a synthetic voice, is enough to activate the social-identification processes that bias people positively toward others of the same gender. This activation led people to be more conforming, trusting, and positively oriented toward synthetic voices whose "gender" matched their own.

Male and Female versus Masculine and Feminine

Although the roots of the words *feminine* and *masculine* come from *female* and *male*, respectively, the concepts of masculine and feminine are not grounded purely in biology. Each culture defines canonical behaviors for females and males,[58] and individuals of both genders can adhere to those prescriptions to varying degrees. Thus, unlike the concepts of female and male, feminine and masculine are personality traits and follow all the rules of personality design discussed in later chapters.[59] For example, feminine people tend to like feminine voices, while masculine people like masculine voices.

Many psychologists do not view femininity and masculinity as opposite sides of the same dimension.[60] Instead, they argue that a person can be *high* on *both* masculinity and femininity (androgynous), *low* on *both* masculinity and femininity (undifferentiated), or high on one and low on the other (clearly feminine or clearly masculine). In the United States, there are strong similarities between masculinity and the *dominance/submissiveness* aspect of personality and between femininity and the *friendly/unfriendly* dimension of personality.[61]

A voice—whether male or female, synthetic or recorded—can be made to sound more feminine and less masculine by cutting off the low frequencies and increasing the volume of the high frequencies.[62] Conversely, a voice can be made to sound more masculine and less feminine—regardless of perceived gender—by cutting off the high frequencies and increasing the volume of the low frequencies. The more feminine-sounding the voice is (regardless of whether it is a male or a female voice), the more female gender stereotypes will be attributed to and expected from it, and conversely, the more masculine-sounding the voice is, the more male stereotypes will be attributed to and expected from it.

Confusion in Gender of Voices: Is It Desirable?

If female voices sound too masculine or male voices sound too feminine, the listener may become puzzled as to which gender is speaking. This does not mean that people will view a gender-ambiguous synthetic or recorded voice as neither male nor female. There is clear evidence that although people can be uncertain about ambiguous voices and can even go back and forth between the determination that "It's a female" or "It's a male," people's brains are committed to assigning one and only one gender at any one moment.[63] Thus, an ambiguous voice is like a Necker cube,[64] the optical illusion in which people always perceive one orientation at a time but frequently switch from one orientation to the other.

Although people will assign a gender, a tremendous amount of evidence has been accumulated over the past thirty years that people strongly prefer clarity in classifying and categorizing people.[65] Anyone or anything with an ambiguous voice is classified as strange, dislikable, dishonest, and unintelligent.[66] Therefore, whether using a recorded or a computer-synthesized voice, clarity of gender (and other characteristics) is critical.

Coping with Gendered Interfaces

When the "gender" of a synthesized computer voice "matches" the gender of a user, people exhibit social-identification effects. Specifically, people conform more to synthetic voices whose gender matches their own and find matching voices more trustworthy. Furthermore, females like female voices more than male voices, even when they know that they are machine-generated, while males like male voices. These results suggest that people automatically assign gender, determine whether it matches their own, and follow the basic principle of "similar is better."

It is impossible to live in or even contemplate the social world without gender. These biological, social, and cultural constructs are automatically manifested by voices, both real and artificial, and automatically recognized by listeners' brains. A careful leveraging of gender identification can make users feel positive and supported, while ignorance of gender identification can unintentionally drive users away. No matter how unnatural or unclear the voice, voice gender is at the heart of interfaces that speak.

3 Gender Stereotyping of Voices: Sex Is Everywhere

The previous chapter demonstrates that people respond to machine-based voices as if they were "male" or "female." Furthermore, people see so much similarity between machine gender and their own gender that females identify with "females" and males identify with "males."

Since the work of Lawrence Kohlberg,[1] psychologists have understood that from childhood onward, people must figure out not only what gender identity they possess but also how that gender identity should guide their behavior. From a very young age, children are "gender detectives" who search for cues about which gender should or should not engage in a particular behavior,[2] primarily by spending time with and modeling others of the same gender.[3]

Children strive to come up with clear and unambiguous decision rules as to how females and males should and do act.[4] By the time children are five years old and continuing until the age of seven, they have extremely rigid rules about appropriate female and male behaviors.[5] For example, if a boy as young as three is told that a toy is "usually used by females,"[6] he will pay less attention to it and remember less about it than will girls; girls will pay less attention to and remember less about a "boyish toy."[7] Toys that have a hint of a female gender marker, such as a pink color, will be labeled as female, and therefore welcomed by girls and rejected by boys.[8] Toys with male indicators elicit the opposite reaction.

Although the absolutist notions of gender that are exhibited by five- to seven-year-olds tend to become less rigid as they grow older, the number and elaboration of social rules concerning gender only grow over time.[9] Adults have absorbed literally thousands of stereotypes about how gender should guide what to expect of each person, how to behave toward each person, and how each person is expected to behave in return.[10]

Even with the startling breadth and depth of guidelines, there seems to be an insatiable demand for further explanations of how women and men think and behave

differently. Best-sellers such as *You Just Don't Understand: Women and Men in Conversation*[11] and *Men Are from Mars, Women Are from Venus*[12] give specific guidelines for understanding how females and males behave differently, and two of the one hundred best American movies of all time (according to the American Film Institute)[13]—*Some Like It Hot* and *Tootsie*—focus on what can be learned when individuals from one gender try to behave like the other. Much of modern literature has at least one character struggling with "Who should I be?" with exploration of gender identity as the point of departure.

Of course, gender stereotypes, like all stereotypes, are misleading when applied to a particular individual. Nonetheless, virtually all people apply gender stereotypes frequently, even if those stereotypes are unfavorable to their own gender.[14] One reason for this stereotyping is that humans have limited processing capabilities.[15] People would find it extremely taxing if they could not filter decisions through a simple dichotomy whose two options provided wide-ranging guidance. For example, when people don't know what to do, they can simply think, "Let's see. I can be only one of two categories: female or male. It's easy to remember which. Knowing that, I can consult my 'rule book' to determine what to do without thinking very hard." The fewer the categories and the greater the homogeneity that are within each category, the smaller the burden of information gathering and analysis that is required to obtain a given level of certainty and the lower the likelihood of error.[16] In sum, there is a great deal of evolutionary pressure to rely on gender stereotypes.

Gender Stereotypes and Technology

The study described in the previous chapter gives the first hint that people might apply gender stereotypes as well as social identification to synthetic voices. Specifically, people seemed to follow the (regrettable) stereotype that females are less worthy of serious attention than males are.[17] Regardless of their own gender, participants found the male voice to be more trustworthy than the female voice. They also generally conformed more to the suggestions that were made by the male voice than those made by the female voice.[18]

Although stereotyping was not a focus of that study, the results suggest an approach that researchers could take to determine whether people routinely gender-stereotype technologies:

1. Select a situation in which one or more gender stereotypes provide misleading guidance about what a person should think or how a person should behave. That is, the gender stereotype should suggest a different approach or conclusion than would a purely rational and objective response.

2. Confront people with technologies that have male or female voices, and measure their responses with respect to the stereotypes. The simplest approach is to have some people use female-voiced technologies and other people use male-voiced technologies and compare the responses of the two groups. This "between-participants" approach[19] to research is the model that is used in chapter 2 and in the first experiment described below. A more subtle but sometimes more powerful approach, called the "within-participants approach,"[20] is to have people interact with *both* female and male voices doing the same task and determine whether the *same* person evaluates, feels, or acts differently depending on the voice they heard. The second study that is described in this chapter is a within-participants experiment.

Gender Stereotyping and Recorded Speech

The first study was inspired by a close friend. She was one of the first female professors at a well-known university's business school. A few (male) faculty members told her that she was likely to be evaluated more harshly by the students because she was a female teaching in a traditionally male discipline (a claim confirmed by the research literature).[21] Their advice was that she "dress and sound more like a man." There were two implications of this. The first—that any pernicious gender stereotyping was *her* problem to address—seemed beyond the realm of experimental research (although it did reveal how accepting her colleagues were of gender stereotyping in the first place). However, the second implication—that if she behaved "correctly," she would not suffer from gender stereotypes—led us to ask the following: *if a female removed all visual indicators of gender, said the same things that a male did, and clearly was not socialized as female, would she then be able to avoid being stereotyped?* Because this study would be impossible to do with an actual person, we used computers[22] with female or male voices.

This first experiment[23] employed computers with recorded female and male voices in a college teaching context. There has long been a concern about the effect of gender on teacher assessment in educational institutions. For example, professors are expected to perform according to gender stereotypes and are penalized when they do not.[24] Students expect female professors to be much more nurturing and supportive than male professors:[25] Female professors are evaluated more harshly when they make jokes about students' performance as compared to self-deprecating jokes, while male professors are

not penalized for making students look foolish.[26] Furthermore, female professors get less credit for making positive comments about students than do male professors, as the former are "doing what they are supposed to be doing"[27] while the latter are being unusually likeable and kind.[28]

Another important finding is that professors are evaluated differently depending on whether they are stereotypically knowledgeable about their disciplines.[29] For example, females who teach in technical areas are evaluated more harshly than females who teach in more stereotypically feminine areas, such as literature or women's studies.[30]

To determine whether these stereotypes would affect people's perceptions of voices on computers, forty college students—half men and half women—worked with a tutoring, testing, and tutor-evaluation system that a university was ostensibly considering for possible deployment. Specifically, the participants were told that they would use three networked computers. The first computer would provide an interactive tutoring session; the second computer would determine how much the student had learned from the tutoring computer; and the third computer would evaluate both the student's performance and the tutoring computer's performance. All three computers were physically identical and were located in three different corners of the room.

First, the "tutor" computer orally presented the participant with ten facts about a stereotypically male topic, technology (for example, "The more wires a computer has, the more slowly it runs"), and ten facts about a stereotypically female topic, love and relationships (for example, "More flowers are ordered for Mother's Day than for any other holiday"). Half of the participants (randomly selected) heard these facts spoken by a prerecorded female voice, and half heard the facts spoken by a prerecorded male voice. After receiving each fact, the participant indicated on a three-point scale ("Very familiar, Somewhat familiar, Not at all familiar") how much he or she knew about that fact. Although participants were told that the computer used this information to choose the next fact to present, thus making the participant feel that the computer was interactive, all participants actually received the same twenty facts.

Next, the participant was directed to the second computer to determine what he or she had learned from the tutoring session. The twelve-item, *text-based*, multiple-choice test[31] included questions that were relevant to but not directly answered by the tutoring session.

After completing this testing session, the participant was directed to a third, "evaluator" computer. The voice used by the evaluator computer was different than that used by the tutoring computer and was either female or male.[32] Thus, some people had a "female" tutor and a (different) "female" evaluator, some had a "male" tutor

and a (different) "male" evaluator, and some had a tutor with a voice of one gender and an evaluator of the opposite gender. Equal numbers of male and female participants were randomly assigned to each of the four conditions (two genders of tutor by two genders of evaluator).

For each question on the test, the evaluator computer began by indicating whether the person gave the wrong or right answer. The evaluator computer then evaluated the tutor computer's preparation of the participant. The evaluator's assessment of the tutor computer's performance was almost always positive (for example, "Your answer to this question was correct. The tutor computer chose useful facts for answering this question. Therefore, the tutor computer performed well") regardless of the participant's actual performance; two negative comments were included to make the evaluation plausible.

After the evaluation session, participants were asked to fill out a paper-and-pencil questionnaire. There were two sets of questions.[33] The first set asked participants for their assessment of the *tutoring* computer's overall competence,[34] the computer tutor's likeability,[35] and the informativeness of each round of the tutoring session.[36] If participants were more influenced by the male-voiced evaluator than the female-voice evaluator, then participants who heard a "male" evaluator praise the tutor would think that the tutor was more competent and likeable in general. Beyond this general effect, if gender stereotyping was applied to the tutoring computer, then the female-voiced computer would be perceived as a better teacher of love and relationships and a worse teacher of computing than would a male-voiced computer, even though they performed identically.

The second set of questions asked participants about the likeability of the evaluator computer.[37] If participants were influenced by gender stereotypes, then the male-voiced evaluator would be perceived as more likeable than the female-voiced evaluator, even though the two computers said the same thing.

Even Computers Are Gender Stereotyped

Participants treated the computer tutors with the same stereotypes as actual professors.[38] Consistent with the stereotype that praise from male professors is more credible than praise from female professors,[39] the tutor was perceived as significantly more competent when praised by the "male" computer,[40] even though the tutor computer's performance and the evaluator's comments were identical in all cases. The tutor was also perceived as significantly more likeable when praised by the "male" computer.[41] The female computer that praised benefited itself less than the male computer

that praised: consistent with the stereotype, the male evaluator was perceived as significantly more likable than the female evaluator.[42]

Gender stereotyping was so powerful that it led people to assess the *same computer* differently depending on the topic.[43] Thus, the female-voiced computer was seen as a better teacher of love and relationships and a worse teacher of technical subjects than was the male-voiced computer.[44]

As with the previous studies, individuals vehemently denied that they were influenced by the gender of the computer voices. They said that the voices were "arbitrary" and that it is obvious that computers are not "male" or "female." They denied that they harbored gender stereotypes with respect to people, let alone computers. The results of the study belie all of these claims.

What about social-identification effects? That is, did females feel more positively toward the female-voiced computers than the male-voiced computers (and conversely for males), as we found in the previous chapter? Unfortunately, the study did not have enough participants to allow us to do the statistical analysis that would address the question.[45] To remedy this deficiency and to explore a new context and additional stereotypes, a second study was run.

Stereotyping in E-Commerce

The second study[46] expanded on the findings of the first. This study used synthetic voices (similar to the ones used in chapter 2) to determine whether users would gender-stereotype even when the voices sounded distinctly machinelike and nonhuman. Instead of education, the experiment was grounded in e-commerce—specifically, an auction site that provided products that are stereotypically female (such as an encyclopedia of sewing) and stereotypically male (such as an encyclopedia of guns). This context invited different stereotypes to come into play. Finally, enough participants were included to determine whether social-identification effects *as well as* gender-stereotyping effects could be identified.

The first stereotype explored was whether some products would be consistently labeled as "feminine" while other products would be consistently labeled as "masculine." Identifying products as male or female does not terminate at childhood: adults tend to identify colors, clothes, food (*Real Men Don't Eat Quiche*[47]), entertainment ("chick flick"), and virtually every other product or service based on its appropriateness for a particular sex.

A second stereotype is that females (female voices) are believed to be more credible in their descriptions of female products, while males (male voices) are believed to be more credible in their description of male products.[48] In marketing, this is called the "match-up" hypothesis:[49] consistency between spokesperson and product characteristics is believed to create more effective, persuasive messages. Thus, because of stereotyping, a male spokesperson may more effectively and appropriately promote power tools (a stereotypically male product) than a woman, while a female spokesperson may more effectively and appropriately promote lingerie (a stereotypically female product) than a man,[50] even when consumers know that the spokesperson is simply reading a script that could be delivered by anyone.

Another stereotype is that whether more knowledgeable or not, females tend to talk more about female products while males tend to talk more about male products.[51] Thus, a voice (and the speaker) that discusses stereotypically female products will be perceived as more feminine than a voice that describes male products, regardless of the gender of the voice. This influence also works in the opposite direction: products that are described by a female voice will be perceived as more feminine than products described by a male voice, regardless of the gender stereotypically associated with the product.

Gender Stereotyping and Synthetic Speech

A total of eighty (forty male and forty female) adults participated in the experiment. Participants were directed to an online auction site that presented products from four different product categories—encyclopedias, matchbooks, watches, and cowboy boots. Each product category included a stereotypically male and stereotypically female instantiation, with all descriptions derived from eBay. Here is the description for the stereotypically female encyclopedia:

The Encyclopedia of Sewing. Doubleday, New York, 1987, 1st edition. 538 pages, hard cover in good condition, some edge/corner wear and shelf wear, corners bumped. All your sewing questions answered quickly and easily. Over 1,000 how-to entries arranged in alphabetical order with detailed step-by-step illustrations ranging from pillow to curtain making. A one-of-a-kind sewing guide you will not want to be without.

Here is the description of the stereotypically male encyclopedia:

The Complete Illustrated Encyclopedia of the World's Firearms by Ian V. Hogg. A. & W. Publishers, Incorporated, New York. 1978. 320 pages, hard cover, with over 500 illustrations, including 50 in color. An A to Z directory of 750 firearm makes and makers from 1830 to present. From Colt revolvers to Winchester rifles and shotguns, this volume is an indispensable reference for firearm owners and enthusiasts alike.

For the matchbook product categories, the examples were "Women's Rights: Susan B. Anthony" and "1934 Roy Parmelee" (a member of the New York Giants baseball team);

for the watches, a woman's classic and a man's retro pilot; and for the cowboy boots, MIU MIU outrageous and Men's Frye All Leather Western.[52]

Each participant heard a description of two randomly selected female products and two randomly selected male products, each from different product categories. A female synthesized voice (randomly) described one of the stereotypically female products and one of the stereotypically male products; a male synthesized voice did the same. The gender of the synthesized voice was created by manipulating vocal markers of gender (pitch and pitch range).

After listening to each product description, participants answered questions regarding the item itself, the item description, and the voice that read the description.[53] Participants were asked to judge both the femininity or masculinity of the voice[54] and the femininity or masculinity of the product.[55] This rating enabled the determination of whether the gender of the voice affected participants' perception of the gender of the product and conversely, whether the gender of the product affected participants' perception of the voice's gender. To test the stereotype that females are more knowledgeable about female products and males are more knowledgeable about male products, participants were asked to rate the credibility of the description that they heard.[56] To test directly whether people thought that females should be talking about female products and males should be talking about male products, participants were asked about the appropriateness of each particular voice-product combination.[57] Finally, to test social identification, we examined whether the match between the gender of the user and the gender of the voice would affect the perceived credibility of the descriptions.

Even Synthetic Voices Activate a Wide Range of Stereotypes

Participants knew which products were "female" or "male"[58]: the stereotypically female products were ranked as significantly more feminine than the stereotypically male products.[59]

The gender of the technology-based spokespersons was as influential as the gender of human spokespersons has been shown to be. Product descriptions were seen as more credible when the gender of the voice matched the gender of the product described.[60] Thus, even though participants knew that the voices were arbitrary and had no unique knowledge of any products, let alone gendered products, the participants were as swayed by synthetic voices as they are by a human spokesperson. Similarly, voices that matched the product gender were seen as more appropriate for describing the products than were mismatched voices.[61]

The gender of voice affected users' perceptions of the masculinity or femininity of the product. Female voices made products seem more feminine, and male voices made products seem more masculine.[62] Conversely, female voices were perceived as less feminine when describing a male product, and male voices were perceived as less masculine when describing a female product.[63] In other words, voice gender and product gender exerted a mutual influence on one another, consistent with stereotyping.

In addition to gender stereotyping, participants also relied on social identification. Females found the descriptions to be more credible when the female voice described the products as compared to the male voice, while the opposite was true for males.[64] This effect was in addition to the effect of the match between voice gender and product gender.

Manifesting Gender in Language

Gender is marked in speech by more than just the sound of a voice. Even within the confines of a given topic, men and women speak and write very differently.[65] In general, women's writing and speaking tend to be more "involved," while men's writing tends to be more "informational."[66] "Involved language" focuses on the interaction: It highlights interpersonal aspects and personal feelings more than specific, detailed information.[67] "Informational language," conversely, focuses on details about the things being mentioned. Thus, females tend to talk more about relationships than males, employing much greater use of "I" and "you," among other markers.[68] Females also express more concern for the listener than do males.[69] Thus, it is not surprising that females tend to use more compliments[70] and apologies[71] than men. Conversely, men tend to focus on objects and their characteristics and more frequently use the word "its."[72] Men also provide more details about place and time.[73]

Even the simplest words reflect differences between females and males. Females tend to use more personal pronouns—for example, *I, you, she, her, their, myself, herself*—than males, while males use a much larger number of determiners—*a, the, that, these*—and quantifiers—*one, two, some more*—than do females.[74] Thus, women's language tends to be about how things relate to each other, while men's language tends to focus on the object under scrutiny. Even the same simple word is used differently by females and males: Females strongly use "they" to refer to living things, while males more frequently use "they" for inanimate things.[75]

The above findings have been validated via a careful analysis of spoken and written dialogues. The ability of these small differences to very effectively (over 80%)[76] predict whether a particular piece of dialogue or text was produced by a male or a female

suggests that individuals could benefit from using these differences when determination of gender through language is important. Unfortunately, there is no research to determine whether individuals do in fact look for these cues. Nonetheless, given the negative consequences of consistency violation and the societal consequences of reinforcing stereotypes, producers of content should be very mindful of these differences and ensure that: (1) the content associated with a particular speaker should have a uniform style with respect to these differences, and (2) the style selected, especially when it involves gendered products, should be carefully considered.

Living and Designing in a World of Gender Stereotypes

The experiments that are described in this chapter suggest that gender stereotyping dominates people's views of the world. Gender stereotypes run so deep that people are often unaware that they harbor them. They are so readily elicited that even a newspaper article that provides biological explanations for gender differences can significantly increase people's endorsements of gender stereotypes.[77]

It was ironic that in both of the studies, participants insisted (when chatting with the experimenters after the experiment was completed) that applying gender stereotypes to computers or synthetic voices would be ludicrous. They also insisted that they were not guided by gender stereotypes in real life. According to the people in our study, they viewed male and female teachers equally and "knew" that the characteristics of spokespersons had nothing to do with their credibility assessments.

Gender stereotypes are pernicious because they discourage the treatment of each person as a unique and important individual and prevent people from realizing their full potential. Relevant morphological differences between the sexes are important for evolutionary success, but modern societies would benefit from a virtual elimination of dangerous overgeneralizations based on gender.

But stereotypes do not exist simply for their evolutionary convenience. Human brains are much too small and slow to process every fact about every individual. Regardless of mating, the simplicity of gender as a variable (only two categories) and its (generally) easy determinability make it a natural locus for limiting the efforts of a brain that is overmatched by a complex world. If the human brain fully eliminated gender stereotypes, it would need many more stereotypes, likely just as pernicious, to replace them.

We advocate a two-pronged approach to the issue of gender stereotypes in voice user interfaces, although it is not an optimal solution. At the macro (social) level, people need to be made more aware that *everyone* harbors significant and limiting

gender stereotypes, regardless of how enlightened they feel, and that such beliefs limit individuals and society by preventing everyone from realizing their potentials. At the micro (individual) level, designers should operate within the limitations of individuals' cognitive system while remaining respectful of all people. The following shows how this might play out in a few domains.

Gender in Teaching and Learning Software

Learners bring role expectations to the education context: teachers are perceived as better when they operate within disciplines in which they are stereotypically expert. For example, in the United States, hunting and fishing are stereotypically male, while cooking and sewing are stereotypically female. This tendency to stereotype then leads to a vicious cycle. As students perceive that certain disciplines are appropriate for one gender or the other, they use these expectations to guide their selections of disciplines. This selection process increasingly skews the gender distribution of those disciplines, thereby reinforcing the initial belief about appropriateness and reinforcing the pattern.[78]

Attempts to break the cycle run into the "pipeline problem." Because very few individuals of a particular gender go into fields that are stereotyped for the other gender, remedying the gender distribution that causes the perception in the first place is difficult. Because decisions about fields of interest often start in middle school, overturning stereotypes takes a very long time.

Technology provides a wonderful opportunity to overcome the "pipeline problem." Rapidly increasing the number of female teachers in stereotypically male disciplines (or vice versa) might be difficult, but it is easy to give female *voices* to all of the educational software for stereotypically male disciplines by simply hiring voice talents of the desired gender or manipulating the parameters of the synthetic voice. Just as people bring gender expectations *to* technology, they can draw gender expectations *from* technology. By "staffing" educational software with male or female voices that counter stereotypes, people are likely to draw the conclusion that people of both genders "belong" in all disciplines.[79] An even stronger strategy would be to have multiple voices in the software, with the vast majority of the voices reflecting the "atypical" gender. By having a limited number of exemplars from the dominant group and the majority from the less dominant group, the presence of the atypical gender will become even more salient.[80]

If these "unusual" assignments of gender might harm the initial credibility of the software, designers could make the female voices "masculine" and the male voices "feminine," whether recorded or synthetic, by using the technique discussed in

chapter 2.[81] This transitional step could be used until disciplinary stereotypes are weakened and voice gender can be selected based on other factors.

Gender in E-Commerce Software

In contexts where the credibility of the description is critical or the stereotype is unambiguous and not detrimental, designers should ensure that the voice matches the "gender" of the product or service. For example, a Web site that sells linen, cosmetics, or romance novels would benefit from a female voice, all other things being equal, while a site that sells hammers, oscilloscopes, or war novels would benefit from a male voice. Gender of voice can also distinguish a product category: a site that sells Barbie dolls would be more credible with a female voice, but a site that sells "action figures" would be more credible with a male voice.

When the product or service is not "gendered," a male voice tends to be perceived as more credible. But designers should also take social identification into account: females tend to believe female voices more than male voices, while males tend to believe male voices.

Credibility is not the only criterion in e-commerce, however. For example, when attention is the key consideration, there are two strategies. First, female voices that suggest that they are physiologically aroused (by using breathy speech, for example) will attract male (heterosexual) listeners, and vice versa. Second, people feel uncomfortable in situations where they are the only person of a particular gender.[82] Thus, a set of uniformly female voices will draw attention from female listeners and discourage male listeners, while the converse would attract males and discourage females.

In some cases, the particular task to be completed by the listener (and not the product or service category) is the most relevant factor in voice selection. For example, imagine a telephone-based call center that sells a stereotypically male product, such as football gear. While a male voice would be a logical choice for the initial sales function, the complaint line might be "staffed" by a female voice because women are perceived as more sensitive, emotionally responsive, people-oriented, understanding, cooperative, and kind.[83] However, if the call center has a rigid policy of "no refunds, no returns," the interface would benefit from a male voice, as females are harshly evaluated when they adopt a position of dominance.[84]

Gender in International Interfaces

All cultures have a wide range of gender stereotypes, but the specifics of the stereotypes differ from culture to culture. In the United States, there are not strong norms about who can provide directions to someone who is lost (although males stereotypically should not *ask* for directions). Thus, both female-voiced and male-voiced navigation systems can be found in cars. However, BMW in Germany had to have a product recall for a car navigation system that had a female voice: German male drivers (the vast majority of BMW drivers) do not take directions from females.[85] Cross-cultural gender stereotyping thus applies to voice systems as well. As another example, in the United States, all aspects of stock transactions are stereotypically handled by males. In Japan, conversely, females are considered more suited to researching and communicating stock information, while males are more suited to handling stock purchases. Hence, the design of a telephone-based stock brokerage system in Japan switched the user from listening to a female recorded voice (when the user obtained information) to a male recorded voice when the user made a purchase. Thus, internationalization involves not simply translating words from one language to another but also being aware of the gender stereotypes that are prevalent in that particular culture.

Gender Stereotypes: Friend or Foe?

Regardless of their appropriateness or accuracy, stereotypes serve a powerful and consequential role in all aspects of life, whether one interacts with people or with technologies. On the one hand, conforming to stereotypes seems to create more natural and effective interfaces—doing so simply acknowledges and leverages users' expectations. But at the same time, mindlessly designing interfaces to conform to every stereotype is often unjustified and even detrimental to society at large.[86] There are no easy answers to this dilemma, but designers must make conscious and considered decisions when choosing to follow, counter, or ignore gender stereotypes as they build computers that talk and listen.

4 Personality of Voices: Similars Attract

If gender is the most basic means for classifying people, personality is the richest. Such descriptions as extroverted or introverted, judging or intuiting, kind or unkind, and a host of other traits provide a powerful framework for understanding how other people think, feel, and behave.

People assign personality rapidly and automatically.[1] They make personality judgments in the first minutes of an encounter,[2] and those judgments strongly influence all subsequent interactions. The ability to rapidly assign personality is very useful: because it broadly predicts a wide range of attitudes and behaviors,[3] personality allows people to know how to behave toward others and what to expect in return.

What signals do people send that let others quickly make judgments about their personality? And how do people's personalities affect others' feelings about them?

Markers of Personality

People use a variety of cues to interpret personality.[4] For example, individuals observe that extroverts charge into situations, while introverts proceed cautiously.[5] Similarly, extroverts talk more than they listen and tend to use strong terms like *definitely* and *absolutely*, while introverts listen more than they talk and favor more neutral words like *perhaps* and *maybe*.[6] Other clues come from posture, with extroverts using many broad gestures, opening their arms wide, and holding their head erect.[7] Introverts, in contrast, use fewer, tighter gestures and hold their arms close to the side and their heads down. People even rely on body shape to guide their judgments:[8] tall and muscular people are perceived as more extroverted than short and frail people.

The sound of someone's voice also provides many important personality cues, which humans have evolved to identify and analyze.[9] Indeed, the term *personality* comes from the Latin *personare*, to "sound through," referring to the mouth opening in an actor's

mask.[10] Assessing personality from voice is an evolved skill.[11] Because you can often hear people before you see them, voice analysis allows a quick prediction of the speaker's attitudes and behaviors.

Four fundamental aspects of voices indicate personality.[12] The first is volume: compare the booming voice of a person who loves socializing to the soft voice of someone who prefers to read books.[13] Another is pitch, or how high or low a voice is:[14] the hyperactive comedian with a high voice suggests a different personality than the deep voiced news anchor.[15] A third is pitch range, or how much the voice ranges between highs and lows:[16] everyone has experienced the rich ups and downs of an animated story teller and the monotone delivery of a technical presenter. Similarly, rising intonation at the end of a statement is associated with a tentative stance, while falling intonation at the end of a statement is associated with confidence.[17] Finally, there is speech rate—the exuberant fast talker versus the calm articulator.[18] These four vocal indicators exert such a powerful influence on human attitudes that people may rely more on voice characteristics to make judgments about people's personalities than on the actual *content* of their speech.[19]

Each of these voice characteristics tells us something about a person, but in combination they become particularly influential.[20] For example, when people meet someone who speaks loudly and rapidly, in a high pitch, and with a wide voice range, they feel confident that they are dealing with the life of the party. Conversely, when people hear a soft, deep, monotone voice speaking slowly, they feel equally confident that this person is shy.

People do not use voice characteristics simply to identify a speaker's personality. They further use their decisions about voice personality to guide their feelings and behaviors toward the speaker.

Similars Attract

One of the most powerful and consistent behavioral patterns with respect to personality is "similarity attraction":[21] people are attracted to other people who are similar to themselves. When people encounter someone who seems to have a personality like their own, they tend to have positive feelings toward the person. People usually like the person and think that the person is intelligent, trustworthy, and friendly. In fact, people will attribute almost any positive characteristic to people who are similar to them.

Similarity attraction can best be thought of as an *overgeneralization* of social identification.[22] People who share similar genes or similar socialization experiences are likely

to share the same evolutionary goals—that is, survival of their community.[23] Thus, people are more positively oriented to family and clan members than to others. There are, then, evolutionary advantages in rapidly associating "like me" with "positive orientation." At the same time, however, people seem to overuse this heuristic and base it on attributes, such as personality, that very possibly have no relevant genetic or shared cultural basis.

What about the saying "Opposites attract"? Oddly enough, opposites who eventually like each other do so *because* of similarity attraction. Research shows that people like others whose personality mismatches them initially and then changes to be more similar to their own more than they like people who consistently match them from the beginning.[24] A third adage ties these seemingly conflicting proverbs together: "Imitation is flattery."[25]

Similarity Attraction and Technology

Will people further overgeneralize the notion of "similar is good" and apply it to voices on the Web, the telephone, and wireless devices? Although one of the ultimate goals of artificial intelligence researchers is to create technologies with a "personality," this goal has not seemed likely to happen in the foreseeable future. First, technologies do not seem to *have* personalities. Indeed, it is not a coincidence that the word *person* is imbedded in the word *personality*. (When people say that the Macintosh and Windows operating systems have different "personalities," they are referring to characteristics of appearance and design philosophy that are important but distinct from the psychological notions presented here.) Even if a computer could project the same vocal personality cues that a person can, and even if a listener would say, "That sounds like an extrovert," most people would draw a distinction between a computer "sounding like an extrovert" and a computer actually being extroverted.

Another reason to expect that similarity attraction would not carry over into human-computer voice interactions is that most speech that is presented via technologies is not and does not sound like human speech.[26] These incessant reminders that computer-generated voices are not human should logically discourage the assignment of personality.

On the other hand, as discussed in previous chapters, the human brain can detect paralinguistic cues in synthetic speech just as it does when listening to true human speech. As with gender,[27] determining another's personality may be so important that when people hear any voice, no matter how clearly not human, they automatically and unconsciously use their voice-analysis skills to assign a personality to the voice. Once they make this judgment, they may go even further and use the seeming

similarity or difference between their perception of the voice's personality and their own to draw conclusions about the system's intelligence, likeability, and trustworthiness, exactly as if they were interacting with another person. Essentially, although people know that synthetic speech systems do not really have personalities, the evolution-derived tendencies to assume that voices are social and respond accordingly may overwhelm that knowledge.

If the evolution-oriented view is correct, then designers should be able to manipulate the speech characteristics of any technology that can produce speech and thereby give it a "personality." For example, designers of Web sites, voice portals, or handheld devices could suggest an extroverted personality by creating a synthesized voice with a fast speech rate, loud volume level, high pitch, and wide pitch range, and "introverted" devices could be given the opposite type of voice. If these cues are successful, people will identify, perhaps subconsciously, the personality in the voice and will be influenced by their perceptions of the system and its content. And because artificial speech can be manipulated in real time, it provides opportunities for adapting to users that would not be available to recorded speech systems.

Do People Feel Like They Are Similar to Machine Voices?

To explore the relationship between synthesized voices and personality, an experiment[28] was conducted to determine whether varying the four key vocal cues (volume, pitch, pitch range, and speed rate) while presenting the same content will lead people to respond as if they encountered different personalities and whether matching users' personalities to the personalities of synthetic voices leads to similarity attraction.

The context for this study was a book-selling Web site.[29] The Web site presented five different books on the same screen, using a visual display similar to the one used by the online book retailer Amazon.com.[30] The screen displayed the title, author, and cover of each book but no text description. Instead, participants had to click on a link to an audio file to hear a description of the book.

For each book, participants heard the audio description and responded to Web-based questions[31] about the quality of the book, the likeability and quality of the description, and their likelihood of buying the book. Then participants answered Web-based questions that assessed their judgments of the Web site's voice, the descriptions of the books, and the reviewer.[32] Finally, because each participant's voice needed to be categorized as extroverted or introverted, all participants were recorded saying their name and describing their experiences in the experiment.

This study required some participants who were clearly extroverted and other participants who were clearly introverted. Using two Web-based standard personality tests,[33] we chose people who had the most extreme and consistent scores[34]—thirty-six extroverts and thirty-six introverts. To ensure that all combinations of participant personality and voice personality were included, half of the extroverts and half of the introverts heard the extroverted voice, while the other half heard the introverted voice.[35] To safeguard the results from the effects of the gender of the participants (some studies have found levels of extroversion to differ between men and women), every combination of participant personality and voice personality had approximately equal numbers of men and women.

Except for voice personality, all participants saw and heard the same visual layouts, instructions, book content descriptions, and questions. The synthesized voices were made to sound as realistic and compelling as possible[36] by simultaneously making all four fundamental voice features (volume, pitch, pitch range, and speech rate) extroverted or introverted.[37] To ensure that only the personality of the voices was manipulated, the voices were checked to ensure that they received similar ratings in categories such as gender, age, informativeness, and persuasiveness independent of personality.

We combined people's answers to the Web-based questions into eight concepts:[38] voice personality,[39] liking of the voice,[40] quality of the review,[41] credibility of the review,[42] whether users would purchase the book,[43] liking of the reviewer,[44] credibility of the reviewer,[45] and personality of the reviewer.[46] Because the book reviews were relatively similar, answers were combined across the reviews.[47]

Synthetic Personality Equals Human Personality

The participants in the book-description study had no problem indicating whether the voice they heard was extroverted or introverted:[48] the extrovert voice was rated as clearly more extroverted than the introvert voice.[49] This confirms that individuals will classify even synthesized voices by their personality.

A more important question is whether the personality of the voice influenced how people *liked* aspects of the Web-based book-selling system. Showing the power of similarity attraction, extroverts liked the extroverted voice more than the introverted voice, while introverts liked the introverted voice more than the extroverted voice.[50] The effects of the voice personality extended to user's feelings about the book descriptions, with extroverts liking the book reviews more when they were read by the

extrovert voice, and introverts liking the reviews more when they were read by the introvert voice, even though the content was identical.[51] Similarly, people found the book descriptions more credible when the voice speaking the descriptions matched their own personalities.[52]

The perceptions of the book descriptions even influenced people's intention to buy. Extroverted participants were more likely to buy the books when the descriptions were presented with an extroverted voice, while introverted participants were more likely to buy when the descriptions were read by an introverted voice,[53] showing the tremendous power of the brain's automatic assignment of personality.

Even though it was clear that the person who wrote the review was not actually speaking, participants liked the reviewer more when the reader's voice matched their own personality.[54] The reviewer also was perceived to be more credible when the reader's and participant's personalities matched.[55]

Even though all book reviews had identical content, the synthesized voice's personality influenced people's perceptions of the human reviewer's personality. People thought the human review writer was more extroverted when the extrovert computer voice narrated the review as compared to the introvert narrator.[56] This result reflects people's well-established tendency to gather personality cues from the closest source and apply that information to judgments about seemingly unrelated people and things.[57] Put another way, the messenger strongly influences people's perception of the message.

There was no instance in which the personality of the voice, *independent* of the user's personality, had a direct effect.[58]

Did You Think about This?

A compelling alternative explanation to the results that were obtained from this experiment is that people like *voices* that sound like their own voices; personality is irrelevant. That is, similarity is not grounded in comparison of the participant's personality and the voice's personality; it is instead based on volume, pitch, pitch range, and speech rate. This distinction is critical. If similarity refers to voice parameters rather than personality, then designers and researchers should be using pitch meters and vocal-analysis tools rather than personality questionnaires. In other words, should designers measure people's personalities via questionnaires (or other attitudinal and behavioral manifestations) and adapt voice interfaces to their personality? Or should they instead derive numeric assessments of people's four major voice characteristics and set the synthetic speech to have identical values?

Of course, if peoples' voices always mirror their personalities, then the question is moot: designers can measure people in whatever way is most convenient. To determine whether this was the case, two trained raters classified the tape recordings of each participant's voice as either extroverted or introverted.[59] Of the seventy-two participants, forty-one participants were classified as having an extroverted voice, and thirty were classified as having an introverted voice (one person declined to have her voice recorded).

The relationship between people's personalities as measured by the written personality tests and by the classification of their voices was remarkably low (although statistically significant).[60] This finding tells us that clear and possibly consequential differences exist between a person's personality and a person's voice. This somewhat puzzling and unexpected result may mean that while people *perceive* voices to be good indicators of personality and respond accordingly, voice characteristics may be deceptive.

We also tested whether the similarity between participants' voices and the computer's voice affected people's perceptions of the computer's voice. All of the statistical analyses were repeated using the participants' voice ratings rather than questionnaire ratings. The analyses showed *no* evidence that people liked voices that were similar to their own voices;[61,62] this effect, if it exists, must be very small or overwhelmed by the assignment of personality. Thus, the best approach to integrating personality and design seems to be to measure personality via a questionnaire (or other reliable methods) and then adapt voice interfaces to the user's responses.

Leveraging the Power of Synthetic Personality

People in the study interacted with obviously nonhuman voices that, because of their artificiality, constantly reminded participants of the nonsocial nature of the interaction. Surprisingly, people nonetheless assigned a fundamental, complex human property—personality—to the synthesized voice. In turn, that personality strongly influenced participants' attitudes, feelings, and intentions.

To date, designers have focused on making synthesized voices easier to understand. This effort has met with considerable success: current systems that use synthetic speech are generally quite comprehensible, even for nonnative speakers. Systems continue to progress in this regard, although the work is technically complex, labor- and computing-intensive, and accessible only to the coders of speech-synthesis systems.[63]

Tremendous opportunities also exist for rapid progress in the *social* aspects of design. Synthesized speech can be made to evoke different personalities quickly and easily simply by modifying a few parameters in real time. Thus, in contrast to the computer scientists who thought that personality would require high levels of artificial intelligence, graphics, and a host of other complex, if even attainable, technologies,[64] the creation of personality is immediately available to everyone.[65] This is particularly true because all of the principles, findings, and design ideas that are discussed in this chapter and throughout the book (with a few exceptions[66]) apply uniformly to both recorded and synthetic voices.

Overall, this experiment shows that the similarity-attraction principle applies to synthesized voices: people like voices that manifest personalities that are similar to their own. The positive feelings about the voice go beyond the voice itself to influence perceptions of the content and willingness to buy.

How can designers leverage these findings to build better interfaces? In traditional media, one of the key concerns is "casting"—the selection of a person for a particular role.[67] Clarity of speech is not a crucial quality. If it were, performers with nonstandard American accents (such as Sylvester Stallone, Arnold Schwarzenegger, Benicio del Toro, and Carol Channing) would not be stars. The producer's goal is to identify actors who can connect with the audience, regardless of paralinguistic characteristics.

When selecting a synthesized voice, the same idea applies: designers should select the voice that serves their goals for the product. Because virtually all voice interfaces want to increase some combination of liking, trust, perceived intelligence, and buying behavior, designers who discriminate among synthetic voices solely on the basis of minimal differences in comprehensibility are missing an enormous opportunity, especially now that virtually all synthetic voices have acceptable clarity. Because text-to-speech systems produce sounds instantaneously and can immediately conform to settings of volume, pitch, pitch range, and speech rate, it is inexpensive to make people feel that the system is "just like them" and hence worthy of affection, trust, and a range of other attributes.[68]

Learning about Users' Personalities: Some Strategies

Imagine that an interface were to say, "Before you can use this system, you must answer the following fifty personality questions about yourself. Once you do that, the system will select the voice that is right for you." This approach would be a mistake. First of all, most business models rightfully preclude any barriers between the system and the user: requiring people to fill out a form before buying, seeing advertisements,

or performing other tasks desired by the company is a poor strategy. Additionally, questions that ask about a person's personality and come from faceless companies or Web sites justifiably elicit feelings of suspicion and resistance because that information is personal. Finally, requiring answers rather than making them optional implies a level of presumptive control that can be disturbing.

Designers might argue that while these concerns are valid, users should be willing to do virtually anything that will make them happy in the long run. But the argument that designers are trying to "make" people happy reminds users that although the system can produce many different voices, the choice is *completely out of their hands*. Instead of presenting users with a series of voices and asking them which one they would like, the system paternalistically (and annoyingly) forces them to give answers and removes them from all decision making. Furthermore, the system doesn't say *how* the answers will affect the selection, inviting users to suspect that the company or interface is selecting voices that serve interests other than the users'.

If designers decide that a direct personality assessment would not be appropriate but do want information about a user's personality, then they can use a number of strategies. First, many interfaces ask users to fill out profile information. The profile could potentially include "personality" questions that do not look like personality questions. For example, a travel site can ask, "While you're on vacation, do you prefer spending time with others or taking time to read or walk alone?" and "Do you prefer visiting famous landmarks or spending time in museums and quieter places?" both of which are good measures of extroversion versus introversion.

There are even more indirect measures of personality. For example, certain occupations are associated with certain personality types; indeed, one of the standard measures for helping people determine the appropriate occupation asks personality questions rather than skill questions.[69] A more creative and whimsical example of personality assessment comes from a Web site that had a picture of a dog and asked, "Which would you give this dog for its birthday: a choke collar [controlling person] or a chewy toy [less controlling person]?" The system could then assign a voice personality based on the person's response.

Designers can also assess personality from user behavior. People who rapidly make choices, especially in new situations, are likely to be more impulsive than those who carefully ponder each option. In addition, sometimes users' personalities can be predicted from their presence at a particular Web site. For example, a site that sells party supplies is likely to be frequented by extroverts, while introverts may frequent a site that sells hobby items. Thus, the former site could confidently deploy only an

extroverted voice, while the latter could use only an introverted voice (this would also conform with the consistency principle discussed in the next chapter).

If the User's Personality Remains Unknown, Use an Extroverted Voice

What if the user's personality cannot be discovered? When the brand and role of the voice do not dictate the selection of a particular personality,[70] use an extroverted voice. Designers of traditional media have long understood this principle: the stars of virtually every situation comedy, the evening news anchors, spokespeople, and the most popular cartoon characters (including Bugs Bunny, Fred Flintstone, and Scooby-Doo) have extroverted personalities and voices.[71] These people and characters speak quickly, loudly, and with significant frequency range. The only exception to the extroversion rule is that when designers want to stress a system's competence, voices with deeper-than-average pitch are perceived as more effective.[72]

Why is extroversion more popular than introversion? The basic reason is that people like others who are expressive.[73] The evolutionary basis for this is that expressiveness is associated with predictability, and predictability reduces cognitive load and increases comfort. Extroverts are also friendlier and more attention getting.[74] Thus, while this selection will not prove optimal for all users, if only one voice can be deployed, an extroverted voice would be better than any other choice.

Attempting to placate all users by having one voice exhibit a variety of personality cues (that is, some cues consistent with one personality and other cues consistent with a conflicting personality) is not an effective solution. People do not respond to these voices as "somewhat like me." Instead, voices with mixed cues (such as faster than average speech, louder volume, higher frequency, but low frequency range), much like voices with ambiguous gender,[75] seem aberrant and therefore less intelligent and trustworthy.[76] People also report high levels of discomfort as they try to draw conclusions about a mixed voice's underlying personality.[77] Like the comic strip character Charlie Brown, hybrid voices get classified as "wishy-washy" rather than "half extroverted and half introverted" and thereby satisfy no one.

Changing the Voice's Personality

In some cases, the personality of the user emerges over time. Can a system significantly change its voice once it learns the user's personality? While gradual adaptation to the user is effective,[78] if a Web site, phone system, or wireless device drastically changes its voice, people will respond as if they are interacting with a different person.[79] This behavior creates a number of problems. First, wondering why a new

voice was introduced will distract people and prevent them from processing the content.[80] That is why the standard in radio is either to make an explicit transition from one voice to the next (for example, "I am now turning you over to Pamela Smith in Washington") or to introduce all of the speakers at the beginning ("With us today is Fred Jones." "Good to be here"). Designers should apply these same standards when using multiple voices (see chapter 9 for a further discussion).

A second problem with this transition from one voice to another is that all aspects of the relationship with the first voice will now have to be reestablished for the second voice: the sense of trust and loyalty built up with the initial voice will not automatically carry over. Indeed, the fact that the person was handed over to a new "person" may lead to psychological resistance. An explicit transition that provides an explanation of the consequences and *reasons* for the change is critical. A system could say, for example, "Because you have been such a wise investor [flattery is an effective strategy[81]], you will now work with our Tier 1 system." As was noted earlier, people do not respond well to the idea that systems are trying to match the voice to their personality. Hence, users should be presented with a different rationale for the change (although the ethics of this alternative explanation are questionable).

Other Personality Traits and Voice

Although voices do not convey all personality characteristics (for example, voices do not indicate whether a person is intuitive, relies on facts, plans ahead, or is spontaneous), there is a model that allows designers to make a more nuanced selection of voice personality than simply extroversion or introversion. This model, developed by the psychologist Jerry S. Wiggins, has two basic dimensions: friendly to unfriendly and dominant to submissive.[82]

The friendly to unfriendly dimension distinguishes people who are warm and sympathetic from people who are unpleasant and hard-hearted. Friendly voices are marked by high pitch (both male and female), wide frequency range, and rapid speech.[83] Classic TV moms of the 1950s are examples of extreme friendliness: they tended to have high, expressive voices and to speak rapidly. Conversely, TV hermits and misanthropes, who have contempt for others but don't actively attack or defend, tend to have deeper voices with little modulation and with measured speed.

The dominant to submissive dimension separates the assertive from the timid. Dominant voices are marked by low pitch, limited frequency range, and loudness.[84] Actors like James Earl Jones, John Wayne, and Clint Eastwood have the classic dominant

voice, and they tend to play characters that control or overcome their environments without being particularly kind or unkind. Submissives are buffeted by events and people and are essentially actionless: they usually don't have starring roles in media because they tend to reflect the plot rather than move the plot forward. Most Woody Allen roles and William Hurt in the title role in *The Accidental Tourist* are exceptions that are marked by relatively high and soft voices.

An added benefit of this dimensional scheme is that it allows the designer to mix these pure types in various ways without creating diffuse and unclear voices or personalities. For example, the mix of dominance with extreme unfriendliness creates the deep, slow, stolid voice of evil characters like Darth Vader in *Star Wars*. Superheroes, in contrast, combine the deepness of a dominant voice with wider frequency range and greater speed than evil doers, showing that they are the "good guys." Classic sidekicks—from Sancho Panza to Lou Costello to Robin the Boy Wonder—have submissive and friendly voices: soft, quick, high, and highly varying. Henchmen (submissive and unfriendly) also have varying voices, but they have deeper and slower voices than their friendly counterparts.

Casting Studies Are Quick and Valuable

The rules outlined above for manifesting personality in voice provide basic guidance for creating personality-rich voices. Nonetheless, the only way to ensure that designers have succeeded is to determine the feelings of a sample of prospective users of the system—that is, to do a casting study. In these studies, designers gather participants and have them listen to each voice in turn (ideally, different users hear the voices in different orders). The content should be relatively neutral. If each voice says something different or says something that strongly marks a personality, the content will influence perception of the voice.[85] The presentation need be only two or three minutes long because people rapidly and reliably assign personality. After hearing each voice, the user is presented with a questionnaire (paper-and-pencil or Web-based) and asked, "How well does each of these adjectives (or sentences) describe the voice?" Research with respect to casting suggests that the Wiggins adjectives are particularly useful for this assessment.

How many participants would a study like this require? The good news is that because assignment of personality is determined by the wiring of the brain, people tend to agree on the personality manifested by a particular voice. Hence, fifteen to twenty people are usually sufficient. Because human speech tends to be more rich and complex than its synthetic counterpart, verifying the selection of a voice or voices with a casting study is critical for recorded voices.

Personality Is Powerful

Voices are not merely a handy means to transmit information to the user. All voices—natural, recorded, or synthetic—activate automatic judgments about personality. These judgments influence people's expectations about the system and its behavior and determine whether they will like the system, trust it, and be willing to buy from it. Rarely does a change as inexpensive and simple as creating the appropriate volume, pitch, pitch rate, and speech rate initiate such large differences in how users will think, feel, and behave.

5 Personality of Voices and Words: Multiple Personalities Are Dangerous

For newborn infants, voice is the alpha and omega of social life. As noted in chapter 1, within days after birth, infants can distinguish their mother's voices from all other voices[1] and can use the sound of voices to distinguish people from other animals,[2] one person from another,[3] and people of their own culture from people of other cultures.[4] Hearing voices is also the first way in which babies identify other people's gender[5] and emotions.[6]

The extreme focus on the sounds of human speech has an even more important benefit: attending to speech is how children learn language.[7] The acquisition of language requires some deep insights:[8] speech sounds are tied to things (nouns), actions (verbs), and descriptions (adjectives and adverbs); words are the building blocks of sentences (all languages have sentences); the same sounds can have different meanings (homophones are present in almost every language); and the order of words is not arbitrary ("The man walked to the woman" conveys a different action than "The woman walked to the man").

It is remarkable that children can learn language in the way they do:[9] There is little systematic repetition in language learning.[10] Complex grammatical forms are introduced before simple ones are mastered, and people use different terms to refer to the same thing.[11] Speakers omit key words via a gesture or a glance. Different teachers (speakers) have different pronunciations and proclivities, people insert nonlinguistic sounds (such as "um" and "uh"), and many of the spoken sentences that the child hears are patently ungrammatical.[12] Some cultures do not even prescribe verbal interactions with babies until the infants demonstrate that they can understand.[13]

Children become fluent speakers in any one of an incredibly diverse set of languages by leveraging the voice activation of their brain and combining it with intensive, ingenious strategies for extracting syntax and semantics from the chaos of surrounding sounds. The key insight that justifies the child's brain's allocation of so much mental attention is that language is an *intentional* activity.[15] Thus, the brain is wired

to view every aspect of every human utterance as *meaningful*.[16] Any sound emanating from a person's mouth is viewed as speech unless proven otherwise. Because language is an intentional activity, the *particular choice of words* describes more than objective content; it also tells the listener about the speaker. Even children as young as two years old understand that various descriptors of their pet—such as *dog, doggy, dawg* (if the speaker is from New Jersey), *puppy, Spot, Spotty, silly dog, funny dog,* or *little fuzz-ball*—reflect the orientation of the speaker as much as the situation, as does the descriptiveness of language—"The woman throws the ball" conveys a different image than "The gray-haired, seventy-year-old woman throws the green, ovoid ball at blisteringly high speed."[17] More sophisticated listeners know that even subtle grammatical differences—such as "I broke the computer" and "The computer was broken by me"—can inform the listener about the speaker.[18] In sum, just as the voice of the speaker enables the brain to extract information that is socially relevant, so does the surface structure of every sentence.[19]

Personality is a primary predictor of how people will speak.[20] Extroverts, for example, tend to use highly expressive language, filling their speech with adjectives, adverbs, and more words overall.[21] Extroverts also tend to talk about objects and situations with extreme phrases such as "the most" and "the worst" and to describe the world in absolute terms.[22] They also tend to use facilitative tags (questions at the end of sentences that seem to invite the listener to respond), such as "I like this movie; don't you?"[23] Introverts, on the other hand, tend to state things more moderately, tentatively, and succinctly, and tend not to refer to the speaker.[24]

Personality in Textual Interfaces

An important transition in the world of technology occurred when computers and computer-based technologies shifted from producing only numbers to displaying, initially on teletypes and eventually on screens, actual words and sentences.[25] This was a tiny step compared to the leap envisioned by Alan Turing when he described a technology that used speech so effectively that it would be indistinguishable from a person,[26] but moving from zeros and ones to outputting something as simple as "Hello World"[27] dramatically changed the accessibility of computers.[28]

Just as every voice, no matter how "neutral," carries paralinguistic cues that can be potentially used to identify personality, every instance of content, no matter how "neutral," carries linguistic cues that can be potentially used to identify personality.[29] All e-mails, Web sites, teaching software, and other text-based materials present content that is influenced by the personality of the writer. Attempts to eradicate personality by implementing numerous guidelines for the creation of content simply

replace writer personality with brand personality or the personality of the rule maker. Although cues are available, will people attribute personality to content in an interface, just as they attribute personality to computer voices? If the analogy between content and oral cues is carried one step further, it also makes sense to inquire about similarity attraction: Will people prefer content whose "personality" matches their own, just as they prefer voice personalities that match their own?

When there is *only* textual content (that is, no voices), an extensive literature on textual personality on computers answers both of these questions with an unequivocal yes.[30] In these studies, the researcher keeps the core content identical while making the following changes: (1) the extroverted computer uses strong, assertive language during the task (such as "You should definitely do this"), and the introverted computer uses more equivocal language (such as "Perhaps you should do this"), and (2) the extroverted computer expresses high confidence in its actions during the task, and the introverted computer expresses low confidence.[31] In the first study of its kind, extroverted or introverted participants were paired with a computer that either matched or mismatched their personality.[32] Consistent with the principle of "similarity attraction" described in the previous chapter, extroverted participants were found to be more attracted to, to assign greater intelligence to, and to conform more with the extroverted computer compared to the introvert computer. Introverted participants reacted the same way to the introverted computer compared to the extroverted computer, despite the essentially identical content.

These initial results have been extended and replicated numerous times, making it one of the most robust findings in the human-technology interaction literature. For example, when people use a "matching" text-based computer, they are more willing to purchase.[33] Textual personality similarity can overcome the "self-serving bias":[34] when a computer's "personality" matches that of the user, individuals are more likely to give the computer credit for success and less likely to blame the computer for failure, compared to when there is a personality mismatch.[35] Personality cues can even be influential when the personality-infused content is not the user's primary interest: the same music, humor, and health advice were perceived as significantly better when the personality of a Web-based "News and Entertainment Guide" matched the user than when the guide did not match.[36]

Mixing Linguistic and Paralinguistic Personalities

If the human brain analyzed speech as simply linguistic cues (the words that are spoken) plus paralinguistic cues (the sounds of the voice), drawing conclusions about the effects of manifesting personality through speech would be easy. The observer

would (1) determine the effects of paralinguistic personality cues (voices) in isolation, (2) determine the effects of linguistic personality cues (content) in isolation, and (3) average the results of (1) and (2). However, the brain intimately links sounds and words from the moment the first syllable hits the ear.[37] The previous chapters provide suggestive evidence. For example, the personality of a synthetic voice influenced listeners' perceptions of book descriptions, as indicated by assessments of the reviewer.[38] Similarly, topics were perceived as more feminine when presented in a feminine rather than a masculine voice, and voices were perceived as more feminine when they were presenting feminine rather than masculine topics.[39] Voices and words are the warp and woof of speech; it is impossible to think about one without the other.

If words and sounds are intimately linked, then what happens when words and sounds manifest inconsistent personalities? Are people comfortable with an extroverted voice reading introverted content or with an introverted voice describing something using the expansive language of extroverts? In normal social interaction, this is rare: most extroverts both sound like extroverts and use words like extroverts, and similarly for introverts.[40] Even though creators of fiction can make anything happen, people expect the casting director to make sure that each character is portrayed by someone who manifests the matching personality. Personality matching is so ingrained that a reader probably would find it difficult to intone forcefully and rapidly with great verbal dynamics, "Perhaps you might like to do this?" or to whisper in a monotone, "This is absolutely the most important thing that you should do."[41]

Laboratory studies show that the brain struggles with inconsistency.[42] In J. Ridley Stroop's classic 1935 study, people were shown a word on a screen and asked to say the color of the ink in which the word was written.[43] People took much longer to identify red ink and say "red" when the word on the screen was *blue* than when the word on the screen was arbitrary, such as *ball* or *chair*. The inconsistent color word seemed to *interfere* with the processing of the ink color, an effect known as the "Stroop effect."[44] This discovery triggered hundreds of similar studies, most of which confirmed the Stroop effect.[45]

Similar studies have been conducted in the speech domain. In one 1972 study, the words *high* and *low* were spoken in either a high pitch (175 Hz) or a low pitch (110 Hz), and people were asked to repeat the word.[46] When the word was spoken with the inconsistent pitch, participants took much longer to say the word than when the word was consistent with the pitch. Similarly, it was more difficult to say which side of the room a word came from when the words *left* and *right* were spoken on the inconsistent rather than the consistent side of the room.[47] Finally, participants were slower to identify the speaker's gender when a male voice said the word *girl* and when a female voice said the word *boy* than when the voice said the word with the consistent gender.[48] Hence, inconsistency between the content of *what* is said, verbally, and *how* it is said, vocally, leads to difficulties in perception and processing.

Research has found that inconsistency in personality, in addition to requiring longer and more effortful processing,[49] leads to negative affect and dislike for the inconsistent person.[50] For the past twenty-five years, an enormous number of studies have demonstrated that participants report less liking of a hypothetical person when that person is described with inconsistent personality traits than when that person is described with consistent traits.[51] Without consistency of personality, people become confused, uncertain, and somewhat disturbed because they rely on a consistent personality to predict the future behaviors of another person.[52] In addition to similarity attraction, then, there is *consistency attraction*.[53]

Mixing Personalities in Interfaces: It Happens All the Time

Unfortunately, using words associated with one personality type and a voice associated with a different personality type in interfaces is a common occurrence. First, interfaces that are designed to read content written by others must contend with a wide range of personalities. For example, an e-mail reader must read missives from a stereotypical cheerleader, librarian, military general, and shoe salesman: how could a single voice personality appropriately manifest them all? Amazon has millions of book descriptions written by people representing every conceivable personality type: what voice personality could be consistent with all of these books? Second, unlike synthetic voice personality, in which creating one personality rather than another is straightforward, there is no simple, automated method to reliably "translate" and "homogenize" content from one personality to another.

If the foregoing theoretical framework is applicable to computer-synthesized voices, then there should be strong negative effects for inconsistency. However, as noted in the first chapter, sometimes psychological theories make conflicting predictions. In this case, strong evidence suggests that people "are not content simply to note inconsistencies or to let them sit where they are. [Inconsistency] is puzzling, and prompts [them] to look more deeply."[54] The most common strategies to resolve seeming inconsistencies are to attribute the inconsistent behavior to a unique aspect of the situation[55] or to ignore the inconsistent information.[56] With technology, users could potentially say to themselves: "That voice just doesn't sound right. Voices like that shouldn't say things like that. Now I see what the problem is. I fell into the silly idea

that computer voices have personality. I may have been lured into feeling that way when there was no inconsistency, such as when the content is neutral, but now that I think about it, I should simply ignore the personality of the voice." If this view is correct, then synthetic voice personality should have no effect at all. The critical question is whether people automatically integrate technology-based voice personality and content personality or whether they separate and possibly ignore the personalities.

Encounters with Confused Personalities: When the Voice Belies the Words

To address these questions, we needed a context in which it would be natural for the same core content to be presented with either an extroverted style or an introverted style.[57] Auction Web sites[58] seemed perfect: they are filled with content that manifests every imaginable personality because of the diversity of the people who submit objects for bids.

A few weeks before the experiment started, a personality test was given to approximately 120 people. Based on the results of that test, forty extroverts and forty introverts were selected to participate in the experiment (we did not tell them that the test was the basis for their selection). Gender was approximately balanced across conditions, and all participants were native English speakers.

Participants were directed to an online auction site, complete with an eBay-like interface. The site included the names and pictures of nine antique or collectible auction items: a 1963 stained-glass lamp, a 1995 limited-edition Marilyn Monroe watch, a 1920s radio, a 1968 Russian circus poster, an old (1910–1920s) church key, a 1916 map of the town of Oxford, a 1920 letter opener, a 1940s U.S. Treasury award medal, and a 1965 black rotary wall phone. The items were chosen so that they would not be of great interest to most participants: desire for a particular item could confuse the results.

For each item, two descriptions were created—one written by an extrovert and one written by an introvert. The descriptions were based on an actual description from eBay. The different personalities were created by modifying the length of the descriptions, word choice, and phrasing. The extroverted descriptions of the auction items were long, were filled with adjectives and adverbs, and used strong and descriptive language expressed in the form of confident assertions and exclamations. They also used self- and other references, such as "I am sure you will like this." In contrast, the introverted descriptions were relatively short, used fewer adjectives and adverbs, used more tentative and matter-of-fact language, and did not reference either the other person or themselves. For example, the extroverted description of the lamp read as follows:

> This is a reproduction of one of the most famous of the Tiffany stained-glass pieces. The colors are absolutely sensational! The first-class hand-made copper-foiled stained-glass shade is over six and one-half inches in diameter and over five inches tall. I am sure that this gorgeous lamp will accent any environment and bring a classic touch of the past to a stylish present. It is guaranteed to be in excellent condition! I would very highly recommend it.

Here is the introverted description of the lamp:

> This is a reproduction of a Tiffany stained-glass piece. The colors are quite rich. The hand-made copper-foiled stained-glass shade is about six and one-half inches in diameter and five inches tall.

The personality of the synthetic voice was created by using the four voice qualities discussed in the previous chapter—volume, pitch, pitch range, and speech rate. The introverted voice had a lower volume, lower pitch, smaller pitch range, and slow speech rate, and the extroverted voice had higher volume, higher pitch, wider pitch range, and rapid speed rate.

When participants clicked on a button next to each of the nine auction items, they heard either an extroverted or introverted description of the item via either an introverted or extroverted synthetic voice. That is, for all items, one-fourth of the participants heard extroverted descriptions spoken by an extroverted voice, one-fourth heard extroverted descriptions spoken by an introverted voice, one-fourth heard introverted descriptions spoken by an introverted voice, and one-fourth heard introverted descriptions spoken by an extroverted voice. Equal numbers of introverted and extroverted participants were randomly assigned to each of these four conditions.

After listening to the descriptions of all nine auction items, participants filled out a Web-based questionnaire.[59] These questions were asked to examine similarity attraction with respect to each of the modalities (voice and words) separately and to determine whether there was consistency attraction (preference for matching voices and words). Participants first indicated how much they liked the voice[60] and the content.[61] They were also asked how much they liked the description writers[62] and how much they trusted the description writers.[63]

To give further insights into whether people holistically process the content presented by synthetic voices, participants were also asked to assess the personality of the content[64] and the personality of the voice.[65] If the personality of one affects the perception of the other, this would provide evidence of a tight coupling.

Too Many Personalities Spoil the Auction

Despite the personality of the content, participants were significantly more positive about the auction when the voice's personality matched their own.[66] Users liked the voice more and liked the writers of the description more; there was no significant difference in how much they liked the descriptions or trusted the description writer, although the results were in the expected direction. In general, then, matching voice personality remains beneficial (although less powerful) even when the content being spoken exhibits a clear personality of its own.

With respect to the (spoken) content, similarity attraction was weaker than generally found with text, although the results paralleled the results for voice. Once again, participants liked the voice more and liked the writers of the descriptions significantly more when the content and user personalities were similar; there was no significant difference in how much participants liked the descriptions or trusted the description writer, although both were in the expected direction. Thus, voice personality and text personality were both seen as similarly diagnostic of the actual personality of the interface, even though the former was created by a computer and the latter by a person.

The most powerful effects, however, seemed to come from consistency. When the voice personality and content personality were consistent, people clearly liked the voice itself more and liked the content more. That is, consistency independently improved the perception of both modalities. People also clearly liked the writer more and found the writer to be more credible when the interface was consistent.

There was also strong evidence that voices and words are intimately linked. The personality of the voice affected users' perception of the personality of the content (as in the previous chapter): users who heard item descriptions spoken by an extroverted voice rated the descriptions as clearly more extroverted than the same descriptions spoken by an introverted voice.[67] This influence is particularly striking given that the personality of the text was unambiguously manifested in word choice, sentence length, ratio of nouns to adjectives, etc.

Consistency of Casting: The Case of Cars

Whenever possible, the best way to ensure consistency is to *start* with the personality of the interface and then have all of the content *emerge* from that personality. Ideally, an entire "back story" is created in which the designers describe the entire life story of the persona[68] of the interface, providing richness and nuance that can emerge over the course of the interaction. Even if the interface is a portal to content with a diverse set of personality markers, the interaction that provides access to the content must have a consistent personality.[69]

The design of the BMW car interface presents a wonderful example of this process. When BMW first introduced its in-car navigation system in Germany, it provided highly accurate information about the car's location and how to get to almost all city and street addresses. Unfortunately, a large number of drivers had a strong negative reaction to this model of engineering excellence and demanded a product recall. The German drivers felt uncomfortable with, and untrusting of, a "female" giving directions. In what has become known as the "gender fender bender,"[70] BMW acquiesced and agreed to provide a male voice. However, rather than simply exchanging the female voice for a male voice, the company used the same rigorous level of care in designing a voice interface as it did in every other aspect of its cars' design. A multidisciplinary team (which included the first author) was put together to decide on the voice interface's new persona.

In addition to providing driving directions, BMW's voice interface had to support a variety of other functions in the present and the future. Car computers have advanced from giving directions to providing mechanical warnings (such as "Your wiper fluid is low" and "Time to check your oil"), information ("There is an accident on Highway 280" and "You are driving 15 miles over the speed limit"), safety ("There is a pedestrian crossing in the middle of the road 0.5 miles ahead" and "There is black ice on the road"), and control ("Now playing the third song on your Beatles CD" and "This car is stopping for a school bus"). The BMW interface thus needed a voice whose personality and role was consistent with all of these potential functions.

BMW was also concerned with making sure that the personality of the voice was consistent with its branding, one of the greatest assets that a car company can have. Because cars are bought by both extroverts and introverts, the company was less concerned with matching the personality of the customer.

The first idea was to have the voice be the car itself; the model was KITT from the television series *Knight Rider*.[71] This idea was rejected because voice interfaces, especially in cars, make frequent mistakes,[72] which would undermine the car's perfect-design branding. This requirement of technical excellence also meant that the voice had to be recorded rather than synthetic.

The next thought was to use a loud, dominant, slightly unfriendly voice that suggested a German automotive engineer. Who could know more about cars? But the engineer was felt to be an overcompetent role that would lead drivers to feel under constant pressure to meet the expectations of the voice.

Another set of possibilities was to put the interface voice in charge of the car. For example, the voice could be that of a pilot (very dominant and neutral on friendliness) or a chauffeur (very submissive, neutral on friendliness, and a slightly British accent). These roles were rejected because people who buy this kind of car are "drivers" who like to control the machine (although luxury cars also appeal to elderly drivers who might benefit from a voice that suggests that the driver is being driven).

Roles such as "friend" or "golf buddy," which strongly implicate similarity attraction, would require the car to know the user's personality; this was beyond the capabilities of the system. Also, people are reluctant to issue commands to peers. "Sidekick" (very friendly and submissive) or the slightly crazy person who rides "shotgun" in the old Western (extreme friendliness, constant talking, enormous pitch variation) would highlight loyalty, which might lead to loyalty in return, but they were clearly problematic for manifesting competence. "Back-seat driver" and "in-law" were rejected immediately with roars of laughter.

A detailed analysis of the potential function of the voice interface and the car's branding suggested that the perfect voice would be a stereotypical copilot. Unlike a pilot, the copilot knows that the driver is in charge but is always ready to step in if the driver needs help. Copilots are extremely knowledgeable about the equipment but always defer to the judgment of the pilot (the driver). Copilots can also have a variety of responsibilities, so that the same voice could adopt additional responsibilities over time.

After choosing this role for the computer navigation system, we decided that the voice had to suggest a male who was very slightly dominant, somewhat friendly, and highly competent. This led to a voice that was medium in volume (with little volume range), was relatively deep in pitch, had moderate pitch range, and had a slightly faster than average speech rate. The interface's language was relatively terse and phrased as statements rather than commands (pilot) or questions (chauffeur). The voice did not say "I." Although our research demonstrates that is generally advisable for recorded voices to use "I,"[73] copilots try to place themselves in a subordinate role and thus avoid the use of "I."

But not every car voice interface should sound like a copilot. For example, a rounded, playful car such as the Volkswagen Beetle could have an extremely extroverted female voice. Cars driven by chauffeurs should have an extremely submissive voice, so that the drivers do not feel overwhelmed by too many directives. Companies that are not extremely concerned about branding of particular vehicles can consider installing the voices of famous people, thereby drawing orientation away from the car and toward the interface itself.

Making decisions about the persona of the voice is a critical part of the development of any product or service interface. Voices can be used to homogenize products within a brand (for example, by using the same type of voice across various product types) and to differentiate brands (for example, by choosing a different style of voice than those used by competitors). If the personality of the interface is selected in advance, each writer should be guided by a "manual of style" that dictates all of the characteristics that make up the personality. At the last stage, a writer should audit the content to ensure that *every sentence* and *every interaction* is consistent with the selected personality.

Consistency Versus Similarity: What Should Designers Choose?

This chapter's experiment has shown the importance of both the *similarity* between the personality of the voice and the personality of the user and the *consistency* between the personality of the voice and the words that it says. The similarity-attraction effects confirm the results of the previous chapter: people like voices whose personality matches their own. Similarity-attraction also extends to words: users respond positively to content that reflects their personality. The consistency-attraction results mirror the results for gender:[74] people like and trust interactions more when the voice "matches" the personality of the content. Finally, voice personality is so strong that it even shapes perceptions of the personality of the content.

In a perfect world, the personalities of the user, the voice, and the content would all match or be matchable—that is, voice or content could be easily changed when mismatches occurred. Unfortunately, in many cases, the designer does not have control over both voice personality and content personality. Sometimes a business has selected a particular voice (usually recorded) as the voice of the company: when that voice doesn't match the user, how should the content be sculpted? Conversely, an enormous amount of previously produced content must be presented by voice: when the content personality doesn't match the user, what voice should be adopted? The following sections address these questions.

Who Cares about Consistency? Who Cares about Similarity?

Everyone likes consistency, and everyone likes similarity, but different personality types place greater emphasis on one or the other. In general, people who are strongly oriented to social relationships and their place in the social world tend to be more concerned with similarity than consistency, while people who are inner-oriented focus on consistency more than similarity. In other words, extroverts tend to be more

concerned about relationships between themselves and others than with the internal workings of the other person. Hence, extroverts would likely be more concerned with similarity than consistency.[75] Conversely, introverts are less concerned with how another person is reacting to them than with their ability to predict the other person's behavior; thus they desire the comfort of consistency over similarity. With respect to the three other Myers-Briggs dimensions (information processing, decision making, and organization styles)[76] people who are labeled intuitive, feeling, and perceiving are likely to be more concerned with similarity, and sensing, thinking, and judging people are likely to be more concerned with consistency. Similarly, people who are at the extremes of the friendliness dimension of the Wiggins interpersonal circumplex[77] (highly friendly or highly unfriendly) tend to be influenced by similarity, and people toward the middle tend to be influenced by consistency because they are less socially oriented.

Other personality differences may also guide the tradeoff between consistency and similarity. For example, people with a high need for cognition (that is, people who enjoy thinking hard)[78] tend to be more concerned with consistency than similarity. Conversely, people with a low need for cognition tend to be more concerned with a readily discernible match between their own personality and that of the voice and less concerned about the complex interrelationships between words and sounds. Similarly, people with an external locus of control[79] (that is, people who believe that their fate is determined by others) tend to be concerned with how others orient toward them (similarity), and people with an internal locus of control tend to be concerned with the predictability (consistency) of others.

Adapting to the User and the Content

Adapting the interface to the user and content during long-term interactions begins with users' psychological preference for the usual over the unusual.[80] People are accustomed to encountering others who have clear and unambiguous personalities that are unlike their own but not to encountering people who display inconsistencies between words and voice. Hence, if the user's personality and the content personality do not match (and the content is fixed), designers should implement interfaces that initially present a voice that is consistent with the content rather than similar to the user. Over time, however, the voice could change to be more like the user, increasing similarity attraction.[81]

The key to the voice-adaptation strategy is to keep the adaptation gradual and natural. For example, when people talk to each other, they often adapt their speech rate to match the other's. A voice interface could also gradually do this without causing

major alarm. Similarly, a system could subtly alter its volume and frequency range to match the user's more closely because people naturally make these adjustments in everyday conversation. On the other hand, fundamental frequency (pitch) should never be changed because it is determined by biological or physical characteristics that people cannot easily or naturally change. In general, the principle is straightforward: if people can and do change a speech characteristic to match the person they are interacting with, then the system can as well.

Personality certainly can change over long periods of time. For example, in the various film versions of *A Christmas Carol*, the actor who plays Ebenezer Scrooge invariably starts the movie with a classic dominant, unfriendly voice—loud and deep with limited modulation. The content of his words is harsh and forceful. As Scrooge goes through his transformational experiences (meeting the ghosts of Christmas past, present, and future), his voice becomes friendly and warmer—softer and slightly higher, with much more lilt and pitch range—and his language indicates that while he is still dominant, he now believes in making people happy. This transitional strategy can be leveraged by product-recommendation software: begin with an authoritative voice and language that is purely dominant. As the customer moves from information gathering to purchase, the voice and language can become friendlier, although both linguistic and paralinguistic cues should retain an air of confidence.

If, on the other hand, the user personality and voice personality do not match (and the voice is fixed), the choice for content personality adaptation is simple: select content that matches the voice. Consistency is clearly a stronger effect than similarity attraction between the user and content, so very little is lost by changing the content to match the voice rather than the user.

Personality Is More Than Voice and Words

In a conversation, certain personality types can be diagnosed without information about the speaker's words *or* voice. For example, extroverts are much more likely to interrupt, occupy a larger percentage of "air time," and be less tolerant of silence in the conversation than are introverts. Thus, for example, the time that a voice-input system waits before it assumes that the user is done speaking should be shorter for an extroverted interface than for an introverted interface.

When the interface is embodied, for example, through an on-screen character, nonverbal cues are also important.[82] For example, extroverts tend to stand closer to other people, face them more directly, make larger and faster gestures, stand more upright,

and make more eye contact with others than do introverts;[83] people prefer characters whose postures conform to their speech.[84]

Finally, according to the theory of somatotypes, body types are also associated with personality.[85] For example, mesomorphs are people with muscular bodies and erect posture (such as Superman) and are perceived as energetic and assertive. Ectomorphs are tall, thin, and small-shouldered (such as Ichabod Crane and Woody Allen) and are viewed as fearful and introverted. Finally, endomorphs are people with round physiques and soft bodies (such as Santa Claus) and are viewed as gregarious and fun-loving. Every manifestation of personality in an interface must be consistent for maximum user liking and comfort.

Personality Is Greater Than the Sum of Its Parts

The human brain likes a neat and orderly world. Inconsistencies are uncomfortable, require significant processing, and lead to dislike and mistrust of the source of the confusion. Dissimilarity between the user and the voice or content is also frustrating and makes interfaces seem less desirable. In sum, because words and voices both provide so many different and unrelenting dimensions with which personality can be indicated, consistency and similarity must be considerations from the moment that design begins to avoid the extraordinary design or writing burden of repairing an inconsistent and ineffective voice interface.

6 Accents, Race, and Ethnicity: It's Who You Are, Not What You Look Like

When two strangers meet, the question "Where are you from?" usually is asked early in the conversation. One reason for asking this question is that the answer provides the strangers with an opportunity to find "common ground"[1]—that is, shared knowledge and beliefs.[2] When two people cannot find a topic that they both know something about, conversation, which is intrinsically cooperative,[3] simply breaks down.[4] Places are particularly fruitful bases for shared knowledge because even if one speaker has never been to the place, a characteristic of the location (a famous landmark or a friend who lived there, for example) can often provide a starting point for shared understanding.

A second reason for inquiring about a person's geographic origins is that place of origin can be as powerful as gender and personality in allowing someone to predict a conversational partner's attitudes and behaviors. Throughout most of human history, place of birth predicted culture, language, and family ties because cultures appeared within regions and there was limited mobility from one region (and hence culture) to another.[5] Thus, people who describe themselves as "Easterners," "Texans," or "Laplanders" provide much more than simply knowledge about natal locale. Like gender and personality, place of origin is one of the most critical traits that define a person.[6] Indeed, place of origin can be a greater predictor of people's attitudes and behaviors than gender or personality because most of the people that the developing child encounters during his or her formative years come from the same place and culture as the child does.[7]

Accents are the voice's way of manifesting hometown.[8] Newborn infants have the ability to produce any speech sound in any language in the world. Accents appear because over time babies learn which sounds are relevant and useful (the ones used by the people they hear) and which sounds are not important.[9] Children who emigrate under the age of seven rapidly acquire the accent of their new country.[10] However, children who change countries between the ages of eight and fifteen usually

retain some vocal remnants of their home country. By the age of seventeen, original accents seem to be fully solidified and unalterable.[11]

Because languages do not have an "official pronunciation," every speaker has an accent.[12] "Accent neutralization"[13]—an active topic of discussion today as telephone-based call centers move to countries with lower wages and different accents—is a misnomer. Although speakers can change their speech to reflect the most common paralinguistic cues in a particular locale, this simply replaces one accent with another.

Accents are manifested, for the most part, via different pronunciations of vowels. For example, in New York English, *Mary*, *marry*, and *merry* are pronounced differently. However, in parts of California and many other parts of the country, these words are pronounced in exactly the same way.[14] Less frequently, accents can also be marked by the inability to produce certain consonant sounds, as in the *shibboleth* example below.

Although it is unlikely that the brain evolved specifically for accent recognition, the brain's hyperattention to speech makes it easy for a person to distinguish whether someone else's speech patterns matches his or her own. For example, the word *shibboleth* (a password or marker of belonging) was a Hebrew word that could not be pronounced correctly by the fleeing Ephraimites. They pronounced it "sibboleth" and therefore were identified as enemies and killed by the Gileadites.[15] During World War II, Germans could not correctly pronounce the Dutch town name *Scheveningen*, which allowed the Dutch resistance to identify Nazi spies.

Dialects are differences in grammar and word choice that are associated with a region. Some parts of the United States refer to *soda*, while others refer to *pop*. New Yorkers use Yiddish phrases that are much less commonly used in the rest of the country, and Southerners use the phrase *you all* when they are talking to several people at once. Similarly, the differences between Australian (*G'day, mate*), British (*Cheerio*), and U.S. (*Goodbye*) English are often the basis of confusion and amusement. No speaker fails to use words that distinguish hometown: everyone sounds like they were brought up somewhere.[16]

Another socially relevant meaning of *place of origin* is where someone's ancestors came from and is usually referred to as *race*. Race is generally indicated by physical appearance,[17] and features like skin color and shape of the head frequently indicate that a person's progenitors once lived in a part of the world for which their body characteristics were optimized.[18] Hence, race is generally discernible by simply visually observing a person.

Because migration was very limited throughout most of human history, language and race were consistent. People who looked like a given race almost always belonged to the culture associated with people of that race and vice versa. Indeed, culture and race were so inextricably linked that the term *ethnicity* has come to be used interchangeably for both.[19] However, these two definitions of place of origin are not intrinsically related. Neither physical environments nor the genes that determine racial characteristics dictate speech or language characteristics (although speech characteristics are associated with geographic areas). All evidence suggests that any normal infant can learn any human language and will speak with the accent and words used in the geographic region in which he or she grows up. When children are raised in countries far from where their ancestors lived, their language and accent are almost wholly influenced by their adopted country. (Accents of family members are a limited constraint, as children's accents are determined much more by their peers than by their parents.)[20] Hence, race does not predict accents. Conversely, dialect has no effect on the appearance of a speaker or a speaker's children or grandchildren. Because evolution works slowly, multitudes of generations need to pass before the physical characteristics of immigrants come to match that of indigenous people[21] (although consistent intermarriage across ethnicities can result in physical changes much more rapidly). In sum, speech manifests cultural origins, and faces and skin manifest long-ago biological origins.[22]

One critical consequence of language and race independence is that people can "speak differently than they look." In one sense, this is a *non sequitur*: speech has nothing to do with physical appearance, so the word "differently" makes no sense. However, because most speakers of virtually any dialect tend to look like most other speakers of that dialect (mobility remains limited), people can be surprised and puzzled when an accent doesn't seem to "match" a voice. Although it is simple to explain how someone who looks Korean can speak with an Australian accent and use Australian phrases or vice versa, how someone who is racially South Indian can speak with a pure French accent, or how people with biological roots in Namibia can have an Irish brogue, encountering unexpected speech patterns is nonetheless surprising.

How do people socially identify with individuals who send these "mixed signals"—that is, how do people decide whether the individual is "similar to them" or "not similar to them"? This is a critical question because (as discussed in previous chapters) similarity is a strong predictor of a variety of positive responses, including liking, trust, and perception of intelligence.[23] One possible response to these mixed signals is to weigh the speech markers and the racial markers differently. That is, people might think, "I primarily care about speech, but race is not irrelevant to my assessments. Because speech counts 73 percent and race counts 27 percent, I will socially identify[24] with this person about three-quarters as much as I would if she or he were wholly like me." At the extremes, this would play out as, "This person exhibits

the same linguistic and paralinguistic cues as I do. Therefore, regardless of her race, I will clearly socially identify with her." (Speech and race could be reversed in the previous two sentences.)

Another possible response to these mixed signals is to integrate faces and voices and voices and word choice[25] in the same way. That is, people might think, "I prefer people who 'look like they talk.' Although I can logically understand why this type of inconsistency could happen, the novelty of it makes me uncomfortable.[26] I don't like or trust people who seem to be inconsistent,[27] no matter how explicable that inconsistency is."

In previous chapters, the psychological literature provided insights into (although not direct predictions of) how people would think and feel about voice interfaces. In this case, unfortunately, the psychological literature is silent on people's reactions to voices derived from one part of the world and faces derived from another. By addressing this question in the context of voice user interfaces, the current experiment can produce research results that are relevant to other domains by starting with technology and allowing others to apply the results to interactions between actual people.[28]

Adding a Face to the Interface

Early voice interfaces were restricted to audio only because the limited adoption of videophones effectively restricted telephone call centers to voice output. The large size of picture files relative to voice files prevented the inclusion of faces in applications for the desktop (hard-disk limitations) and the Web (transmission-speed limitations), and screens were much more expensive and power-consuming than speakers. Even though faces draw attention[29] and can improve speech understanding,[30] the technological constraints meant that voice user interfaces were much more like radio than television.

Thanks to Moore's law[31] (which predicts that computer components and hence computers will become inexorably faster and cheaper, doubling capacity about every eighteen months) faces can now be readily added to voice user interfaces. Cell phones now routinely come with screens that allow the high-quality presentation of at least still pictures, and video output to cell phones will soon be widely available. Computer hard disks have become so large that they easily store pictures and even video. Full-motion video (via streaming video) and video conferencing are now available to any computer with a high-speed Internet connection, and limited moving pictures via the Web can even be received with dial-up telephone lines. Screens have become so inexpensive that some household appliances have visual output as well as voices.

In these new technologies, voice user interfaces present combinations of speech and body. These "embodied conversational agents" or "embodied computer characters"[32] might be a still picture of a (possibly mythical) call-center representative, a training video with voice input, an animated conversational facilitator[33] or a synthetic character to teach deaf children to speak[34] and are becoming an increasingly popular interaction technique.[35]

Embodied conversational agents have been shown to make interactions between people and technology more natural and social.[36] Some dimensions of cultural variability that have been explored with animated characters include appearance, content and manner of speech, manner of gesturing, emotional dynamics, social-interaction patterns, and role dynamics.[37] Designers have recently begun to consider the possibility of localizing characters to the cultures where they will be deployed,[38] but they have yet to explore interactions with a character whose voice and appearance characteristics are not culturally consistent. That is, when voices and faces are shown together, do people integrate the two, even when they know that the two can be assigned independently? This question is addressed by examining social identification[39] and consistency-attraction[40] effects in interfaces that have clearly accented voices and clearly marked races.

Mixing and Matching Accent, Race, and Ethnicity

To test the effects on users of an agent's regionally marked speech and race, an experiment with an online e-commerce site was conducted.[41] A total of ninety-six male college students participated in the experiment. The participants were either Caucasian Americans (forty-eight participants) or first-generation Koreans (forty-eight participants). Participants were directed to an e-commerce Web site that displayed photographs of four different products—a backpack, a bicycle, an inflatable couch, and a desk lamp.

Half of the Korean American participants and half of the Caucasian American participants heard four product descriptions read by a voice with a Korean accent that occasionally used distinctly Korean phrases (such as "Anyonghaseyo," which is Korean for "hello"). The other half of the participants heard four product descriptions read by a voice with an Australian accent that occasionally used phrases associated with Australians (such as "G'day, mate").[42]

For a given participant, each of the four product descriptions was read by the same voice, and each description was accompanied by one of four full-length photographs of the same person in different poses (to give users a sense of animacy).

To hear the description of an item, participants clicked on the person's photograph. For participants who heard the Korean-accented voice, half were shown a photograph of a racially Korean male, and half were shown a photograph of a racially Caucasian Australian male. Similarly, for participants who heard the Australian-accented voice, half were shown a photograph of a racially Caucasian Australian male, and half were shown a photograph of a Korean male. Half of the participants in each of the four conditions were culturally and racially Korean, and half were culturally and racially Caucasian Americans.

To ensure that any effects were not attributable to a particular voice or face, the study used two different Korean voices, two different Australian voices, two different ethnically Korean faces, and two different ethnically Australian faces. Participants were randomly provided with one voice and one face that represented a particular combination of accent and race. The descriptions of the four products were identical except for a few stereotypically Australian words in the Australian-accented descriptions and a few stereotypically Korean words in the Korean-accented descriptions.

After hearing each product description, participants were asked to respond to a questionnaire that asked about the product's likeability[43] and the description's credibility.[44] After listening to the descriptions of all four products, participants were also asked to rate the agent's overall quality.[45] These questions allowed the effects of social identification and consistency on perceptions of the product and credibility of the agent to be assessed.

Agent's Accent, Agent's Race, or Consistency: What Is Most Important to Users?

Consistent with the results that have been obtained from studies of voice and gender,[46] people socially identified with the paralinguistic cue of accents (at least when coupled with verbal markers of hometown), regardless of the race of the agent.[47] The agents were rated much more positively when they spoke in an accent that matched the user's ethnicity.[48] Participants also evaluated the products[49] and product descriptions[50] more positively when they were spoken by an agent whose accent allowed the participant to socially identify with it. Hence, shared accents, which mark similar childhood cultures, clearly encouraged social identification.

There were no social-identification effects found for the *race* of the agent, however.[51] In other words, the match between the race of the user and the race of the agent (as manifested by the four photographs of the same person in four different poses) did not affect users' ratings of the agent or the products. Thus, participants' ancestral history did not seem to be an independent point of reference when evaluating the agent.

Although the paralinguistic cues of the voice and race of the agent are not logically linked, participants were disturbed when the agent did not "look the way it sounded." The photographic agents that had "consistent" voices and faces were perceived to be higher quality than those that were inconsistent, regardless of the ethnicity of the user.[52] Participants also found the products to be better[53] and the product descriptions to be more credible[54] when the two places of origin seemed to be consistent. Thus, people integrate voices and faces, even when the voice is clearly independent of the nonmoving mouth of the still photograph.[55] This integration also leads the race of the face to influence people's perceptions of the accent.[56]

Accent-uating Speech in Interfaces

Interfaces and software applications often have many places of origin. Each implementation task—design, programming, voice recording, photography, testing, burning to disk, placement on a server, warehousing—can be performed in a different region of the world. While the producers of voice interfaces must worry about all of these steps along the supply chain, consumers can ignore these diverse locales and simply define place of origin in terms of what is apparent and socially relevant to them—the voice, the face, and the content. The current research suggests that of the two places of origin—hometown and ancestral homelands—accent, which reflects the culture in which the person grew up, is critical for responses to interfaces, even when knowledge of an agent's (or person's) ancestry is readily apparent.

Users socially identify with agents based on the cultural background that is evident in the agent's voice but not based on the racial background evident in the agent's appearance. Voice may therefore more effectively localize a Web site to a particular region than the appearance of pictorial agents does. Multinational companies that attempt to make their Web sites cater to specific countries or populations should realize that the culture of the country, rather than its race, should be their key focus. Designers should worry about the language of the Web site, the accent of the voice, and other cultural references on the site much more than whether the skin color of an agent matches the most common race of the user population.

Creating and Identifying Accents

Because accents are simply a reflection of the way that the vowels and consonants of a language are pronounced, all synthetic and recorded voices manifest accents. Thus, both machine-generated and recorded voices lead to determinations of place of origin that automatically elicit assumptions about the speaker. Place of origin and its

accompanying assumptions can be further supported by having the voice use unique words that are associated with a dialect that matches the accent.[57]

There is currently no automated method for detecting accents in users. The most straightforward approach compares the phonemes of the speaker to phonemes derived from different accents, but the number of voices needed to represent each accent with a high degree of accuracy is currently prohibitively high, and a significant amount of speech is needed before an accent can be reliably discerned. The best available approach is to learn the location of the user (perhaps by using telephone caller ID or e-mail Internet protocol address) and then to select the dominant accent of that location. One elegant example of this comes from a Japanese company with an automated call center that determined the area code of the caller and automatically assigned the caller to a voice whose accent matched the region from which the call came. This strategy was particularly effective in Japan, where geographic mobility and hence accent dispersion is much lower than in the United States, but it should be effective anywhere in the world.

In addition to social identification, accents in voice interfaces are likely to elicit stereotypes, much as gender of voice does.[58] Particular accents are associated with particular characteristics, whether those stereotypes are grounded in socialization processes or not.[59] For example, in the United States, people with British accents are perceived as more intelligent than average, whereas people with Brooklyn accents are perceived as less intelligent than average, even though, as described by accent prestige theory,[60] there is no formal basis for these stereotypes. Similarly, people with Boston accents ("I pahked the cah in Hah-vahd Yahd" instead of "I parked the car in Harvard Yard") are perceived as elitist, while Southern U.S. accents are perceived as more egalitarian (although there are also racist stereotypes associated with Southern accents). As discussed earlier,[61] designers must decide when to conform to and thereby reinforce stereotypes (for example, by having history tutorials presented with a British accent) and when to challenge them (by using a person with a Brooklyn accent to explain brain surgery).[62]

Labels Are Powerful

All social identification and stereotyping is grounded in labels that make the labeled characteristic salient and blur other characteristics of the person.[63] Thus, once the brain is activated to focus on a personal category such as gender, place of origin, or personality, other attributes that might have been relevant, including whether a voice is synthetic or recorded, move to the background.

Labels serve a number of purposes. First, broad, sweeping generalizations about others based on labels allow the human brain to make quick assumptions and predictions about other people and help the brain cope with the complexities of social interactions.[64] Second, when a label is determined to be similar to the label that someone assigns to herself or himself, that person can apply the principle of similarity attraction and attribute a wide range of positive characteristics to the voice.[65] Finally, labels allow people to learn how to behave: when labels match, people can observe others to answer the question, "What would a person like me do in a case like this?"[66] Thus, labeling is a fundamental human process that dramatically decreases the need for decision making or information processing at the expense of nuance and accuracy.[67]

Age

Gender, accent, and personality are three of the strongest bases for labeling others. The most powerful category not yet discussed is age.[68] Individuals certainly define themselves by their age cohort, with people quickly announcing their age range in answer to the question "Who am I?" Although perceived age ranges differ as a function of culture and even age itself[69] (for example, an adult sees fewer age categories in children than children do),[70] age is almost as salient a characteristic as gender. Societies tend to be segregated by age:[71] individuals of an age group tend to have most of their interactions with others from the same age cohort. Age also enables people to make reasonable predictions about the physical and mental capabilities of people that they encounter.

As with gender, there are clear voice markers of age.[72] In general, voices of both males and females get deeper as they grow older. Very young children and the elderly both speak more slowly on average than people in their twenties.[73] Anecdotally, individuals have no problem in assigning an age range as well as a gender to machine-generated (as well as human) voices.

Race

Race has been intentionally left out of the list of important social categories, even though it is one of the best predictors of virtually any phenomenon in the sociological literature.[74] The current research suggests that race is not important as an independent effect, at least when an accompanying voice provides broader and deeper information than what can be determined from ancestral geography. The linguistic and paralinguistic cues are so informative that identifying a person's geographic progenitors is simply not valuable.

One caveat associated with the previous conclusion is that differences in responses to accent and race could be attributed to the contrast between a dynamic voice and a static face. This explanation seems dubious because participants are unlikely not to notice a face. First, whenever people see anything that looks even remotely like a face, they think of it as a face,[75] so it is highly improbable that people in our experiment were not drawn to the face. Second, we found large effects with respect to the mismatch between the face and the voice (if people ignored the face, there would be no effects for inconsistency). Nevertheless, the current study should be replicated with dynamic video or other moving images instead of still photographs of agents. This would likely produce an even tighter linkage between the agent's face and voice than still photographs did.

Other Social Categories

Other social categories that are manifested by speech—that is, by paralinguistic and language cues—include education and intelligence[76] (for example, educated people have greater fluency of speech, have larger vocabularies, and use more varied sentence structures) and class and status[77] (for example, in many cultures, certain accents and word selections are assigned as "high status"). In general, all categories that are associated with clear social and behavioral consequences tend to be recognizable through speech.[78]

Creating Social Identity When Users Are Unknown

What can designers do when they do not know the user's relevant social categories and hence cannot leverage social identification via those categories? The best method is to create *pseudocommunity*, an artificial bond between a diverse set of people.[79] For example, in all parts of the eBay Web site and in all its contacts with the customer, eBay refers to the "eBay community" and "eBay members like you." This is essentially the creation of a social category—eBay users—that had not been apparent previously. This category *a priori* matches the user and thereby encourages positive social identification. In a very different domain, an MP3 player oriented to the MTV audience could use a voice and a language that sound like a typical rock music lover; in that case, traditional categories of social identification would be almost invisible compared to this highly relevant category.

Children's television provides many examples of establishing social identity even when it might seem difficult to create. For example, Fred Rogers of *'Mister Rogers' Neighborhood* (a long-running children's U.S. public television series) has a famous line:

"Won't you be my neighbor?" Although Rogers eventually was old enough to be a grandfather to his fans, was Caucasian despite a diverse audience, and had a Pittsburgh accent, his invitation to children to join his community highlighted the similarities between him and his viewers. Similarly, Barney, the star of *Barney & Friends*, another popular children's public television series, is a six-foot-tall, purple dinosaur; nonetheless, he effectively reminds preschoolers that "We're a happy family."

Although social identification can be a powerful tool for designers, it can sometimes be a liability. For example, many voice interfaces use the term *we* to refer to the company ("We at XYZ company want you to be happy)" rather than to link the voice and the user. This implicitly tells the user that she or he should *not* identify with the voice, thereby increasing the probability that the user will become angry and frustrated during the interaction. Even more egregious are cases in which the user is given a choice of voices that represent the same social category, such as different female voices teaching cooking or nine-year-old voices teaching reading. Although these voices might conform to stereotypes or assumptions about the audience, respectively, the Hobson's choice[80] presented to people who do not match the relevant category (a man who is learning how to cook or an adult who is learning how to read) strongly reminds them how unusual they are.

Consistency in Agents

As discussed earlier, consistency is a fundamental human desire.[81] For example, a person who speaks in a happy voice while frowning is unlikely to inspire trust or liking.[82] Similarly, users would almost certainly prefer consistent visual and vocal cues of gender, age, and personality over inconsistent ones. Imagine the surprise of a person who discovers that a young, friendly looking female agent sounds like a surly old man. Inconsistency can make people feel uncomfortable well before this level of incompatibility.

Even though there was no logical or technological relationship between the voice and the body in the previous study, clear negative consequences were found when the two were mismatched. For social information in which the face and voice manifest redundant information, for example, emotion, consistency is even more important.

International companies have yet another dimension of consistency to consider—the country of origin of the product. For example, a Web site might sell Dresden glass, Chinese sculpture, or Iranian rugs. If the pictured spokesperson is of the race that is associated with the region of the product, the consistency of matching an accent with both the product and the spokesperson would likely overcome any benefits that might

be associated with matching the user. Situations in which the race of the product spokesperson does not match the region of the product are more complicated. Yet more complicated are cases in which both knowledge of the region and an understanding of the user are valuable: the most obvious category is tourism. In these cases, the picture and voice combination might be more trouble than help unless all components can be aligned.

People Are More Than Social Categories, But . . .

At a fundamental psychological level, users generally do not distinguish between voices coming from an actual human and voices coming from a computer.[83] Even though talking about a computer as having a gender, hometown, or personality makes no logical sense, if it has a voice, users will attribute all of these social categories to it. For best effect, designers must therefore consider all of these cues and their impact on stereotyping and identification when choosing a voice for an interface.

Everyone wants to be appreciated as a rich and unique person that "contains multitudes."[84] People insist on being "a name, not a number."[85] Similarly, the best interface designers carefully cast and script every aspect of a voice user interface, whether or not it is accompanied by pictorial representations.[86] But every voice activates certain responses in listeners—a rapid determination of relevant categories[87] and blindness to subtle and difficult-to-process categories—thereby enabling rapid social identification and social stereotyping. By attending to a limited number of categories, designers have a tremendous opportunity to optimize how users will feel and behave.

7 User Emotion and Voice Emotion: Talking Cars Should Know Their Drivers

The previous chapters describe how users respond to voices, both computer-generated and recorded, as if they manifest gender, accents, and personality. Each of these characteristics is an unvarying aspect of a person, known as a *trait*: sex (barring surgery), childhood socialization, and personality[1] last a lifetime. These unwavering qualities allow observers to predict a wide range of behaviors and attitudes and the general orientation that speakers will have toward themselves, other people, and the world around them. If humans were defined only by traits, life (and design) would be simple: Listen to people briefly, classify them, and from then on worry only about those paralinguistic cues (such as the generally rising tone at the end of a question) that affect interpretation of what they meant.

For better and worse, however, people are dramatically affected by their environment. Indeed, traits give us a broad sense of what a person will think and do, but predicting a person's behavior at any given moment in time also requires attention to the user's *state*—that is, the particular feelings, knowledge, and physical situation of the person. Extroverts are generally talkative, but they might be as silent as introverts in a library or even quieter when they bump into their secret crush. The most "feminine" female will exhibit a range of masculine characteristics when protecting a child. In a group, people often submerge their identity as they blend in and mimic the people around them.[2] Although traits provide the general trajectory of an individual's life, every specific attitude, behavior, and cognition also can be influenced by momentary states.

Of all the types of states that predict how a person will behave, the most powerful is *emotion*.[3] Rich emotions are a fundamental component of being human.[4] Emotion is so fundamental that a key part of the brain used in emotion (the amygdala) is also used to determine whether an image is a person or not.[5] Throughout any given day, affective states—whether short-lived emotions or longer-term moods—color almost everything people do and experience, from sending an e-mail to driving down the

highway. Emotion is not limited to the occasional outburst of fury at being insulted, excitement at winning the lottery, or frustration at being trapped in a traffic jam. It is now understood that a wide range of emotions plays a critical role in *every* goal-directed activity,[6] from asking for directions to asking someone on a date, from hurriedly eating a sandwich at one's desk to dining at a five-star restaurant, and from watching the Super Bowl to playing solitaire. Indeed, many psychologists now argue that it is impossible for a person to have a thought or perform an action without engaging, at least unconsciously, his or her emotional systems.[7]

Although emotion was one of the primary foci of the early field of psychology, the study of emotion has lain dormant for a long time.[8] The founder of the field of human-technology interaction, Frederick Winslow Taylor, focused on formalization, measurement, and "science,"[9] which, in his day, meant the exclusion of human feelings, an exclusion that extends to most present-day studies.[10] Until recently, it would have seemed ludicrous to even discuss whether machines could have or exhibit emotions.[11] This avoidance of the study of emotions has extended to the realm of voice interfaces.

Because of its critical importance and the neglect it has received, emotion is discussed in two chapters. This chapter discusses emotion from the *user's* point of view, addressing questions like "What is emotion?" and "How does user emotion affect interaction with voice interfaces?" The next chapter focuses on the expression of emotion from the *interface's* point of view, discussing "How do voices manifest emotion?" and "How does the manifestation of emotion in interfaces affect people?"

What Is Emotion?

What is emotion? Although the research literature offers a number of definitions,[12] two generally agreed-on aspects of emotion stand out:[13] (1) emotion is a reaction to events that are deemed relevant to the needs, goals, or concerns of an individual, and (2) emotion encompasses physiological, affective, behavioral, and cognitive components. Fear, for example, is a reaction to a situation that threatens (or seems to threaten, as a sudden loud noise does) an individual's physical well-being, resulting in a strong negative affective state, as well as physiological and cognitive preparation for action.[14] Joy, on the other hand, is generally a reaction to the fulfillment of goals and gives rise to "happy" behaviors such as smiling and engagement with the environment.[15] Emotions can be relatively short lived; when one or more are sustained they are called *moods*.

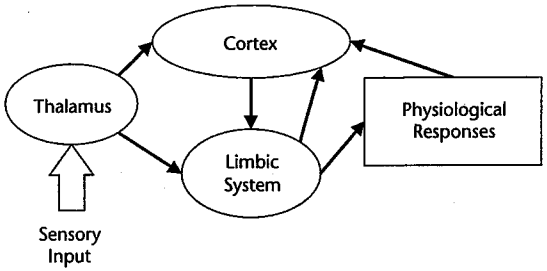

Figure 7.1
Neurological Structure of Emotion

Just as humans are built to process and produce speech, they are built to experience emotion (and mood).[16] Figure 7.1 presents a standard model[17] showing three key regions of the brain—the thalamus, the limbic system, and the cortex. Information comes through the senses and goes first to the thalamus, which does the basic processing of the input. The thalamus then sends information simultaneously to the cortex (for higher-level processing attention and other thought processes), and to the limbic system[18] (for evaluating the relationship between the information and the user's needs or goals). If relevance is determined, the limbic system (often called the "seat of emotion") sends messages to the various parts of the body (which can make the heart pump faster or the muscles contract, for example) and to the cortex.

The "primitive" or "primary" emotions, such as startle-based fear and innate aversions and attractions, originate from the direct link between the thalamus and the limbic system.[19] In the context of voice interfaces, unexpected sounds, such as a beep—instead of "You've got mail"[20]—have the potential to activate such primitive emotional responses, as do sounds that are innately disturbing or pleasing due to their evolutionary significance, such as screaming, crying, or laughing.[21]

Most of the emotions that play a role in the design of voice interfaces are the "secondary" emotions, such as frustration, pride, and satisfaction. These emotions require more complex analysis than the primitive emotions. Hence, when the thalamus receives the input, it sends a message to the cortex, which does extensive knowledge-based processing. Such cortical processing may or may not be conscious and can occur at various levels of complexity, from simple recognition (such as hearing a familiar voice) to intricate rational deliberation (such as evaluating the social and functional consequences of being verbally reprimanded). The cortex can even trigger emotion in reaction to a person's expectations or imaginings, such as thinking about

how difficult it will be to use a newly purchased voice-recognition system. The cortex then sends this information to the limbic system, which decides whether an emotion is warranted (based on the user's needs and goals) and how to manifest that emotion.

Finally, an emotion can result from a combination of both the thalamic-limbic and the cortico-limbic pathways. For example, an event causing an initial startle or fear reaction can be later recognized as harmless by more extensive, rational evaluation, such as when you realize that your car's sudden beep is just a ploy to draw your attention, ensuring that you turn off your headlights.

Although the limbic system is a critical locus for the *activation* of emotion, emotion is *experienced* only when the limbic system sends messages to the cortex and the body.

Are Emotions Innate or Learned?

The above discussion provides a useful framework for considering one of the classic debates in emotion theory: are emotions innate or learned? At one extreme, most evolutionary theorists argue that all emotions (including complex emotions such as regret and relief) are innate[22] and have evolved to address a specific environmental concern of humans' ancestors. These theories assume that the limbic system is complex because it distinguishes a wide range of emotions and triggers unique sets of body and brain (cortex) responses for each of these emotions.

At the other extreme, many emotion theorists argue that—with the exception of the startle response, the innate-affinity response, and the disgust response (which they consider preemotional)—emotions are almost entirely learned social constructions.[23] Such theories emphasize the role of the cortex in differentiating emotions and concede minimal, if any, specificity within the limbic system (and consequently, within physiological responses). For example, the limbic system may operate in an on or off manner[24] or at most differentiate along the orthogonal dimensions of valence (positive to negative) and arousal (excited to calm). From this perspective, emotions are likely to vary considerably across cultures, with any consistency arising from common social structure, not biology.

Between these two extremes lie those who believe that there are "basic emotions." Citing both cross-cultural universals and primate studies, these theorists contend that all humans share a small set of innate, basic emotions.[25] Which emotions qualify as basic is yet another debate, but the list typically includes fear, anger, sadness, joy, and disgust.[26] Other emotions are seen either as combinations of these basic emotions or as socially learned differentiations within the basic categories.[27] For example, agony, grief, guilt, and loneliness are socially constructed variations of sadness. In this view, the limbic system is prewired to recognize the basic categories of emotion, but social learning and the cortex still play a significant role in differentiation.

Effects of Affect

Attention

One of the most important effects of emotion lies in its ability to become completely absorbing. Emotions direct and focus people's attention on those objects and situations that have been appraised as important to people's needs and goals. Emotion-relevant thoughts then tend to dominate conscious processing: the more important the situation is, the higher the arousal and the more forceful the focus.[28] In a voice interface context, this attention-getting function can be used advantageously (as when a navigation system's "turn left immediately" is used to alert the user), or it can be distracting (as when a user is frustrated by poor voice recognition and can think about nothing else). More generally, people tend to pay more attention to any thoughts and stimuli that have some relevance to their current emotion.[29]

Just as emotions can direct users to aspects of an interface, emotions can also drive attention *away* from the stimulus that is eliciting the emotion.[30] For example, becoming angry with a voice-recognition system may be seen as silly or illogical (because it doesn't recognize your anger). An angered user may then actively try to avoid parts of an interface that rely on voice input, rendering the user's interaction less efficient or effective. In extreme cases, the user will simply avoid the system. If the emotion is too strong, however, the user will not be able to ignore the source,[31] potentially even resulting in "computer rage." Positive emotions may likewise require regulation at times, as when amusing content, such as a voice-based "joke of the day," leads to inappropriate laughter in a work environment.

Performance

Emotion has also been found to affect cognitive style and performance. The most striking finding is that even mildly positive feelings profoundly affect the flexibility and efficiency of thinking and problem solving.[32] In a well-known experiment, participants were induced into a good or bad mood and then asked to solve Duncker's candle task:[33] given only a box of thumbtacks, the participant must attach a lighted candle to the wall so that no wax drips on the floor. The solution requires participants to have the creative insight to thumbtack the box to the wall and then tack the candle to the box. Participants who were first put into a good mood were significantly more successful at solving this problem.[34] In another study, medical students were asked to diagnose

patients based on X-rays after first being put into a positive, negative, or neutral mood. Participants in the positive-affect condition reached the correct conclusion faster than did subjects in other conditions.[35] Conversely, positive affect has been shown to increase reliance on stereotypes and other simplifying rules of processing, which could lead happy users to make less nuanced judgments about a voice interface and to be more influenced by labels, such as the gender, accent, and personality of the voice.[36]

Judgment

Emotion has also been shown to influence judgment and decision making. As mentioned above, emotion tends to bias attention and thoughts in an emotion-consistent direction. One important consequence of this is that everything—even those people, things, and events that are unrelated to the current affective state—is judged through the filter of emotion.[37] This suggests, for example, that users would likely judge both a voice interface and the interface's message more positively if they were in a good mood than if they were in a negative or neutral mood. Recommendations, for example, will obtain a greater level of acceptance among happy people. Positive emotion also decreases risk taking. That is, although people in a positive mood are more risk-prone when making *hypothetical* decisions, they tend to be more cautious when presented with an actual risk situation.[38]

Emotion and Voices in Cars

Driving presents a context in which emotion can have enormous consequences. Attention, performance, and judgment are of paramount importance in automobile operation, with even the smallest disturbance potentially having grave repercussions. The road-rage phenomenon[39] provides one undeniable example of the impact that emotion can have on the safety of the roadways. Considering the above discussion of the effects of emotion (in particular, that positive affect leads to better performance and less risk taking), it is not surprising that research and experience demonstrate that happy drivers are better drivers.[40]

Now that car manufacturers are increasingly turning to voice as a promising strategy for safe, pleasant, and engaging interactions with everything from in-car navigation systems and environmental controls to road-aware, chatting copilots (the remarkably intelligent and speaking car, KITT, presented in the 1980s television series *Knight Rider*, may soon become at least a partial reality), it is critical to know how emotion expressed by an in-car voice interface interacts with a driver's emotion in

affecting attention, performance, and judgment. Might the emotional characteristics of the voice have as much impact on attention, performance, and judgment as the emotions of the driver? More specifically, what happens when the emotion of the voice and the emotion of the driver are mismatched (such as when an upset driver encounters an upbeat voice)?

To investigate this question,[41] a driving-simulator-based[42] experiment with twenty female and twenty male participants was conducted. The participants drove a virtual car down a simulated road (complete with other vehicles) using a gas pedal, a brake pedal, and a force-feedback steering wheel. All participants drove on the same simulated course[43] for approximately fifteen minutes.

To address the question of the effects of user emotion on driving performance, half of the participants had to be happy at the time of the experiment, and the other half had to be upset. In experiments concerning emotion, researchers can work with the emotions that participants bring into the lab or create an emotion after participants enter the laboratory. There are pros and cons to both approaches, but the difficulty of scheduling (in advance) equal numbers of happy and upset participants as well as the desire to ensure that the emotion experienced by participants be created in the same manner led us to *create* emotion rather than rely on recruitment. Therefore, either a happy state or an upset state was induced in each participant at the beginning of the experiment. This was accomplished by showing half of the participants (randomly selected) seven minutes of very happy video clips and showing the other half seven minutes of very upsetting video clips.[44] A questionnaire confirmed that individuals had the correct emotional state before they entered the driving simulator.[45]

While driving the course, participants interacted with a "virtual passenger," who was represented by a recorded female voice that made light conversation with the driver. The virtual passenger introduced "herself" by saying,

> Hi. My name is Chris, and I will be your virtual passenger for today. We are going to be driving today on a coastal road, one that I've traveled many times before but one that may be new to you. The trip shouldn't take too long: ten or fifteen minutes. Let's get going.

At thirty-six separate points along the course, the virtual passenger made a different remark.[46] For half of the happy participants and half of the upset participants (randomly selected), the voice was energetic and upbeat; for the other half, the voice was calm and subdued.[47]

In the case of driving, safety is of utmost importance. To assess the effects of the link between driver emotion and voice emotion on drivers' performance, the simulator automatically recorded the number of accidents that drivers were involved

in during their time in the simulator. Another way to assess whether the driver was paying attention to the road was to determine the driver's reaction time while driving. Specifically, participants were told to honk the car's horn as quickly as possible in response to randomly occurring honks. Because honks indicate that another driver felt the need to communicate about some aspect of driving, the quicker, on average, that participants honked back, the more they were paying attention to the road. The drivers were also asked, via a questionnaire, how attentive they thought they were.[48]

While everyone wants drivers to pay attention to the road, designers of car interfaces also want people to feel that the voice is an important part of the driving experience. The driver's engagement with the system was measured by the amount of time that drivers spent talking back to the voice in the car while driving down the simulated road.

Voice Emotion Influences Driving Performance

Matching the voice of the car to the drivers' emotions had enormous consequences.[49] Drivers who interacted with voices that matched their own emotional state (enthused voice for happy drivers and subdued voice for upset drivers) had *less than half* as many accidents on average as drivers who interacted with mismatched voices.[50] This magnitude of reduction is far greater than the effects of virtually any technological change in a car at dramatically less expense; influencing the driver is more important than influencing the car.

Drivers who heard voices whose emotions were suited to their own emotion also rated themselves as significantly more attentive while driving.[51] The mismatched voice distracted drivers, probably because they unconsciously tried to figure out why someone interacting with them was unresponsive to their own emotional state. Drivers paired with matched voices also communicated much more with the voice, even though the voice said exactly the same thing in all conditions.[52] This suggests that although drivers who heard emotion-matched voices spoke to the virtual passenger much more, they were nonetheless better able to avoid accidents.

The effects of matching emotion versus mismatching emotion were so powerful that neither driver emotion nor voice emotion by itself had a consistent effect on drivers. Although there was a slight tendency for happy drivers to be better drivers, even this effect was minimal compared to the effects of matching. In other words, finding the appropriate in-car voice for the driver's emotion stood out as the most critical factor in enabling a safe and engaging driving experience.

The foregoing suggests that people find it easier and more natural to attend to voice emotions that are consistent with their own.[53] Emotion-inconsistent voices are arguably more difficult to process and attend to than emotion-consistent ones; in other words, interacting with an inconsistent voice takes more cognitive effort.[54] It then follows that listening to or conversing with the virtual passenger was more distracting in the mismatched case, leading to poorer driving performance, less attention to the road, and a less comfortable experience.

Making Driving, and Voice Interfaces, Safer and Better

Auto manufacturers spend a great deal of time and money attempting to make cars safer. They install antilock brakes, adaptive steering and gas pedals, air bags, and various warning indicators. All of these design decisions require very long lead times to retool factories and often involve dramatic increases in cost.

The current research demonstrates that a simple, inexpensive, and fully controllable aspect of a car interface can also have a dramatic influence on driver safety. Something as small as changing the paralinguistic characteristics of a voice is sufficient to improve driving performance significantly. Even with the same words spoken at the same times by the same voice under the same road conditions, driver performance can be radically altered when the voice is simply changed from energetic and upbeat to subdued. Remarkably, changing the tone of a voice can strongly influence the number of accidents, the driver's perceived attention to the road, and the driver's engagement with the car.

A key finding here is that the same voice cannot be effective for all drivers. For both actual and perceived performance, upset drivers clearly benefited from a subdued voice, while happy drivers clearly benefited from an energetic voice. This suggests that voices in cars must adapt to their users. This presents two important questions: How can an interface detect driver emotion, and how can that information be used most effectively?

Detecting Driver's Emotion

How should cars assess the emotion of drivers? The basic emotions (fear, anger, sadness, joy, and disgust) are the most distinguishable and therefore measurable emotional states (both in emotion-recognition systems as well as in postinteraction evaluations). Further, the basic emotions are less likely to vary significantly from culture to culture, facilitating the accurate translation and generalizability of emotion detection.

Recent advances in emotion-recognition technology provide a number of possibilities for determining the emotional state of drivers and other users of interactive systems. Sensing technologies include cameras to assess facial expressions[55] (for example, to detect road rage or other emotions), skin conductance and blood pressure sensors attached to the steering wheel or mouse (particularly good for detecting arousal),[56] technologies to measure head nodding or other movement (to detect driver drowsiness or general relaxation), and eye-tracking equipment (for example, to detect pupil expansion, which is associated with positive feelings).[57]

In certain contexts, prediction offers another possibility for detecting users' emotions. For example, if a car's safety system recognizes a potentially dangerous situation, the virtual passenger can assume that the driver will soon be afraid or stressed and quickly adapt its voice characteristics (without any direct measurement of emotion). In less time-critical situations, such prediction could also be used in conjunction with direct measurement to improve emotion-recognition accuracy.[58]

Voice analysis is particularly useful for systems with voice input or for situations in which the user is frequently talking (such as when a passenger is in the car). Voice-analysis systems can use the same set of vocal cues to detect emotion that people use to detect emotion and that synthetic voices use to manifest emotion, as described in the next chapter.

The critical issue in all detection systems is accuracy. Although focusing on multiple indicators leads to better accuracy than assessments of only a single indicator of emotion, even the most complete, wide-ranging, and intelligent systems are not as good as the average person.[59] Thus, any reactions to the user's emotion must be tempered by the knowledge that the judgments are not perfect.

How Fast Should the Emotion of Voice Change?

If a car will make a determination of the driver's emotion, the designer must decide how the system should respond to that emotion. One useful strategy is to change the emotion of the voice in step with the user—that is, to exhibit empathy. Empathy greatly fosters relationship development because it communicates support, caring, and concern for the welfare of another.[60] A voice that expresses happiness in situations where the user is happy and sounds subdued or sad in situations where the user is upset would strongly increase the connection between the user and the voice.[61] Similarly, a voice telling users that an important file has just been corrupted or that the product they requested is no longer in stock should do so with a degree of expressed

regret. In these cases, the system leverages its ability to predict rather than measure the user's emotion.

Although a rapid response to the emotion (or predicted emotion) of the user can be effective, there are a number of dangers in this approach. In the human brain and body, emotion can change in seconds.[62] If someone tells a joke to a sad person, he or she may become momentarily happy but will fall back into a sad state relatively quickly. Conversely, happy drivers may become frustrated as they must slam the brakes after expecting to zoom through a yellow light, but their emotion may quickly switch back to feeling positively. If the voice in the car adapted immediately and forcefully to the user's emotions, drivers would experience such bizarre occurrences as the voice of the car changing its emotion in midsentence. People or voices that manifest emotions at that rate would dramatically increase cognitive load,[63] constantly activate new emotions in the user, and be perceived as psychotic. While this can be entertaining when performed by manic comedians like Robin Williams or Jim Carrey, it is psychologically exhausting and disturbing when encountered in daily life. Such a virtual passenger would be marked quickly as manic-depressive rather than empathetic.

To make the voice in the car an effective interaction partner, mood also must be taken into account. Moods tend to bias feelings, cognitive strategies, and processing over a longer term than emotions.[64] While moods can be influenced by emotions, they tend to be more stable and can even serve as a filter through which both internal and external events are appraised. A person in a good mood tends to view everything in a positive light, while a person in a bad mood does the opposite. Even when a negative event is viewed negatively, moods can limit the impact of that perception. Thus, when assessing people's responses to a voice, designers should consider the biasing effects of user mood. Users entering a car or using a voice-based product in a good mood, for instance, are more likely than users in a bad mood to experience positive emotion during an interaction. Thus, emotion in technology-based voices must balance responsiveness and inertia by adapting to both emotion and mood.

Even easier to assess than moods is temperament.[65] Temperament reflects the tendency of certain individuals to exhibit particular moods with great frequency.[66] Sales people and implementers can measure users' temperament before the adoption of the product or service, select a voice, and tune the emotion-recognition system to understand and respond based on the general affective tendency of the user.

Stabilizing the Unstable User

People are who they are. Grounding interface design in stable characteristics of users (called *personas*)[67] can simplify the design process and allow interfaces to be standardized. At the same time, humans are what they feel. Emotion—influenced by every encounter with the physical, psychological, and social world—plays a critical role in determining users' performance, knowledge, beliefs, and feelings. Every voice, synthetic or real, emanating from people or attached to interfaces, creates and interacts with an emotion-driven user who will think, feel, and succeed (or fail) based on the sensitivity of the design to the user's affective reality.

8 Voice and Content Emotions: Why Voice Interfaces Need Acting Lessons

The previous chapter demonstrated that emotions can have a tremendous influence on individuals. If people lived in isolation, each person's own emotions would be the only emotions that would be relevant. However, people live in a social world in which every person is surrounded and affected by every other person's emotions.

Given the central role that is played by emotion, a clear evolutionary advantage goes to those who can determine the affect (emotion or mood) of the individuals with whom they interact. This is reflected in the wiring of the brain: The left hemisphere recognizes positive experiences more rapidly, and the right hemisphere recognizes negative experiences more rapidly.[1] In addition to this biological optimization, training in emotion detection begins at three to six months after birth during facial play with mothers.[2] Knowing how a person feels about us at a particular moment gives us general information about whether to approach or avoid as well as a nuanced understanding of the interaction strategies that will be effective. Indeed, it is more important that people understand others' emotions than understand their own. Thus, it is not surprising that one of Charles Darwin's most influential books[3] is called *The* Expression *of the Emotions in Man and Animals*[4] rather than *The* Experience *of the Emotions in Man and Animals*. (The previous chapter discusses people's internal experiences of emotion.)

Given how valuable emotion detection is, the human brain has evolved to detect emotion not just in the words,[5] facial expressions,[6] and other behaviors of people but also in their voices. Emotion manifests through vocal properties such as pitch range, rhythm, and amplitude or duration changes.[7] A bored or sad person, for example, will typically exhibit slower, lower-pitched speech, with little high-frequency energy, while a person experiencing fear, anger, or joy will speak faster and louder, with strong high-frequency energy and more precise enunciation.[8] Emotion detection through voice is so important that it is localized in the brain: paralinguistic cues of emotion, regardless of content, activate the lateral temporal lobe (superior or middle temporal

gyri), primarily on the right side of the brain.[9] Similarly, the right hemisphere of the brain is primarily responsible for *producing* paralinguistic cues of emotion, while the left side controls the language content.[10]

From moment to moment, voice characteristics that manifest emotional information allow a listener to better predict a speaker's intentions and actions[11] and better understand the speaker's meaning.[12] Humans are quite good at detecting emotions from voice,[13] even when the voice is machine-generated.[14] Table 8.1.A (see page 95), created by Iain R. Murray and John L. Arnott, provides a detailed account of the vocal effects that are associated with several basic emotions.[15]

Detection of emotion tells us more than just the psychological orientation of the speaker. Emotional tone in human speech can also color what is said; in other words, voice emotion can determine how specific words and sentences are interpreted by the listener.[16] The phrase, "My voice recognition system is 95 percent accurate," for example, can either be praise (when said with joy) or criticism (when spoken with anger). "Oh, my goodness" means something very different when expressed fearfully as opposed to happily. Even sentences that seem intrinsically happy—such as "I'm really enjoying myself"—can radically alter meaning when said in a disgusted or sarcastic tone of voice. Emotion of voice thus critically affects a listener's understanding of spoken content.

When voice quality and emotional content are matched, the brain readily integrates the two. However, when the emotion in a voice is clearly inconsistent with the content being presented, a person is presented with an irreconcilable problem. This leads to problems from an early age. Mothers who exhibit inconsistency between verbal and vocal cues, such as criticizing in a syrupy voice, are much more likely to have children with behavioral and emotional problems.[17] Even within groups of disturbed children, the boys of the mothers who showed more inconsistent communication were reported to be more aggressive in school than the boys whose mothers showed less inconsistent communication.[18]

As people become adults, the primary effects of mismatches between verbal and vocal emotion involve attributions to the content and the speaker. Except in cases where sarcasm is intended and clear, people tend to like a person better when their emotional tone is consistent with the emotional information implicit in the words being said.[19] People generally view inconsistent emotion as an indication of insincerity, instability,[20] or deception[21]—all unappealing attributes—and thus anything said by such a person tends to be unlikeable. Inconsistency between the emotion in words and in voice also places a burden on the listener: more brain regions are used when voice and content are emotionally inconsistent than when they are consistent.[22]

Mixing Voice and Content Emotion in Interfaces

Emotions are part of what defines humans as "human."[23] Technologies, on the other hand, seem to be the quintessence of unemotional.[24] However, *all* voices activate the parts of the brain associated with the measurement of paralinguistic (vocal) cues that mark emotion. Could voice interfaces, then, trigger the same emotional processes in the brain as human voices?

When listening to humans, voice emotion affects interpretation. However, given that people know that machines don't have emotion or truly understand the emotional content of the words they speak, the emotional tone of a *machine-based* voice rationally should not influence perceptions of the content. Listeners should understand that technologies generally have a single, relatively invariant voice that is essentially decoupled from the words being spoken. However, given that people assign traits (such as gender, region, and personality) to machine-based voices even when those assignments seem irrational, might they also assign states, that is, emotions, to these voices?

To determine whether people assign emotion to machine-based voices and to determine whether those assignments influence their perceptions, a telephone-based experiment was conducted.[25] Fifty-six adults participated in the study. Gender was approximately balanced across conditions.

Participants were first directed to a Web page that gave instructions for the experiment and provided a call-in number. After they read the instructions, participants called the phone number, punched in a pass code to hear the appropriate content, and listened to two news stories (one happy, one sad), two movie descriptions (one happy, one sad), and two health stories (one happy, one sad).[26] For example, the happy news story described a new cure for cancer that was showing incredible promise, especially for children, while the sad story described a mystery of dead gray whales washing ashore on the San Francisco coast. All stories were read by the same male synthetic voice.

For half of the participants (randomly assigned), the emotion of the voice matched the emotion of the story being read, and for the other half of the participants, voice emotion and story emotion were mismatched. In other words, half of the participants heard happy content read in a happy voice and sad stories read in a

sad voice (consistent emotion), while the other half of the participants heard happy content read in a sad voice and sad stories read in a happy voice (inconsistent emotion).

After hearing each story, participants answered a set of questions on the Web site.[27] For each story, questions were asked to determine how happy or sad the story was[28] and how much participants liked the story.[29] Finally, participants were asked whether the stories would be more suitable for extroverts or introverts.[30]

Users Have Strong Feelings about Voice and Content Emotions

The emotion of the voice had a clear impact on participants' perceptions of the content. Participants who heard happy content read in a happy voice perceived the stories to be happier than those who heard the same content read in a sad voice.[31] Similarly, those who heard sad content read in a sad voice rated the stories as less happy than those who heard it read in a happy voice.[32] That is, although the content was identical, the voice influenced the *meaning* of the stories, even though the stories were clearly not created by the technology itself.

Participants liked the content significantly better when read by a voice whose emotion matched the emotion of the content than when voice and content were mismatched.[33] This demonstrates the importance of consistency, even when it is clear that the technology would not understand why consistency is important or know how to achieve it.

In sum, paralinguistic cues like speed, pitch, and volume are *integrated* with the perceived meaning of spoken words into a single message. Matching a voice to the content is just as important, and may even be more important, than matching the voice to the emotion of the user.

Technology for Creating Emotional Interfaces

Whenever a technology "speaks," the emotion expressed by the voice can have significant impact on users and their perceptions. Choosing the right emotional tone for a given piece of content can thus be as critical as choosing the right gender, accent, and personality for the voice—if not more so.[34] For example, a doll can say, "Be my friend" with an enthusiastic, apathetic, or angry voice—leading to perceptions of friendliness, insincerity, and contempt, respectively. A car can describe a problem with a battery with varying levels of urgency, leading to varying levels of user response. A music-recommendation system can sound confident or nervous, influencing people's

likelihood of accepting the suggestion. A performer can make listening to a book on tape as moving an experience as reading a print book by infusing words with the proper emotion at the proper moment, or she can ruin the experience by striking a false note with the match between words and sounds. A voice-based financial Web site could joyfully announce that a stock has gone up but must avoid this if the customer sold the stock just before its value rose. A recipe program can say "Now *lightly* whisk the two eggs" with a tone of disgust, suggesting suspicions that the cook will mangle the recipe.

In the instances above, designers have no excuse for inconsistency because the content is prerecorded. With even a reasonably skilled actor, achieving high degrees of emotional consistency is simple if not automatic. The only caveat is that when sentences are recorded *out of context* by a voice talent, there can be unintentional inconsistency: "Your team has twenty-one points" should be spoken very differently depending on whether the previous sentence was "Your rival has five points" or "Your rival has fifty points."

The problem becomes much harder when the amount and variety of content are too large or dynamic to have a person record all of it. E-mail readers must be prepared to read any sentence written by any person at any time: "This is the happiest day of my life" should be read in a cheerful voice instead of a monotone, unless the rest of the message reveals that the sentence was sarcastic. The Associated Press has an enormous number of stories coming into its offices from all over the world: it is impossible to predict what will come next, and virtually instantaneous broadcasting is the norm. eBay receives postings for thousands of new products from thousands of people every day, including everything from computers to Ming vases. In cases like these, a single voice, no matter how it sounds, will mismatch a significant fraction of the content, and randomly assigning multiple voices will not solve the problem.

The seeming solution when content clearly manifests a single emotion is to have the technology first determine the emotional content of the particular message and then alter the voice accordingly. With most synthetic-speech systems, manipulating the parameters of a voice (such as pitch, pitch range, word speed) that are associated with emotion manifestation is straightforward.[35] Unfortunately, the technology required to deduce emotion from text is not effective because of the complexity of the problem.[36]

At first glance, written language would seem to be so filled with emotional tone that emotion should be easy to identify. Emotion-modeling systems can also look for

explicit word choices and phrases that indicate emotion, such as "This is great" or "I feel awful." In more formal communication, the topic often presents a good indicator of emotion: wedding Web sites should have happier vocal indicators than should funeral Web sites.

If emotion were always indicated through emotion words (such as *happy* or *sad*) or through a topic, the problem of textual-emotion determination would be simple to resolve. To understand the difficulty, however, consider this sentence: "Another dead gray whale has washed ashore on the California Coast, deepening a mystery that has baffled scientists for years." This was the lead for the sad news story that was used in the study described above. Any human would know that this is a sad story, but think of all the knowledge that is imbedded in that conclusion. For example, although *dead* might be a clue that it was a sad story, think of a joyous song, such as "The Munchkin Land Song (Ding, Dong, the Witch Is Dead)" that has the word *dead* in it. The emotional-analysis system would have to know that the death of gray whales is sad but that the death of the Wicked Witch of the West in the film *The Wizard of Oz* was happy (although the death of Glinda, the Good Witch, would have been sad). "Baffling mysteries" can be frightening ("Was that really a ghost?"), angering ("Who took the last cookie?"), sad (the current example), happy ("Ah, the baffling mysteries of the universe!"), and a host of other emotions, depending on the tremendously varying characteristics of the individual and the situation. Remarkably, humans are able to grasp all of the subtleties of a situation to determine the appropriate emotional response.

The current best solution seems to be to put the responsibility for labeling the emotional tone of the content on the author. Authors can be asked to tag or describe their content at whatever level of granularity they prefer—the entire work (such as "a humorous piece"), a paragraph or sentence ("the 'dead gray whale' sentence is sad"), or even a word ("In this context, the word *hooray* should be read sardonically, reflecting sarcasm"). Perhaps the best example of a method for marking content for emotion is the use of *emoticons*, typographic symbols that serve as the equivalent of paralinguistic cues of emotion. Emoticons arose when it became apparent that emotional indicators were necessary to support interactions via e-mail and instant messaging systems. The word *emoticon* comes from the contraction of *emotion* and *icon*;[37] emoticons are icons because they are meant to capture the appearance of a face (lying on its side) that manifests the particular emotion. For example, the five basic emotions (and the extremely useful *irony*) can be represented as follows:

Fear	8-o
Happy	:-)
Sad	:-(
Angry	(:-&
Disgusted	:-|
Irony	;-)

Emoticons are one useful method for marking emotions, with the potential to represent a large number of different states.[38] Beyond these emoticons, computational linguists and computer scientists have developed more sophisticated and subtle mark-up schemes that can indicate context, mood, and emotion as well as specific vocal parameters.[39]

Mark-up systems to indicate emotion have proven to be useful, but they have a number of limitations. First, it is burdensome for a writer to go through an entire work marking appropriate emotions, potentially at multiple points within every sentence. (Picture Tolstoy working his way through *War and Peace*, tortured by the need to enter tens of thousands of emotional markers, each reflecting the precise and nuanced emotion that he was trying to capture.) Second, unskilled producers of content—such as people submitting letters to the editor, book reviews to Amazon.com, or descriptions of products to eBay—might find the requirement of emoticons to be a significant "barrier to entry" and not submit content. Finally, the vast majority of content in media is *preexisting*; it would be an impossible task to go back and label every book, advertisement, or newspaper article.

Though the benefits of consistent, appropriate voice emotion are well established, the challenges are great, ensuring that content-rich voice interfaces will suffer the problems of inconsistency for a long time to come.

What If an Interface Can Manifest Only One Emotion?

While pessimism may be warranted with respect to solving the problem of inconsistency in voice emotion, interfaces nonetheless must be built. When the emotion of the content or the user cannot be detected, and the entire interface does not have a general emotional tone, what voice should be selected? The best emotion to select is a slightly happy voice.[40] Humans are built with "hedonic preference"—the tendency to experience, express, and perceive more positive emotions than negative ones.[41] In social interactions, people like others who display positive emotions more than those who display negative emotions.[42] If a person shows positive emotion, he or she is perceived to be more attractive and appealing to work with than someone who shows negative emotion.[43] The effect of positive emotion is even found in the mass

media: commercials imbedded in or following a happy TV program are perceived to be more effective, are liked more, and are recalled better than ads following sad programs.[44]

When is it advantageous to select emotions other than happiness? Even though people volunteer positive events when asked to recall their lives—the so-called Pollyanna effect[45]—anger and sadness are more attention-getting and memorable than positive feelings.[46] In addition, the expression of submissive emotions (such as sadness and fear, for example) often increases trust in a relationship. Listeners view such emotional disclosure as a willingness to "open up" and be vulnerable,[47] which they often reciprocate.[48] Voice-based interfaces that need to obtain personal information from users—a health Web site or telephone-based surveys, for instance—could leverage voices (as well as content) that encourage trust.[49] We know of no advantage in defaulting to disgust, the fifth of the basic emotions, although this and all other emotions should be expressed in voice when dictated by a desire to match the content.

Manipulating More Than the Voice

Sometimes designers have control over both the content and the voice. The same rules that allow automated systems to begin to identify emotion in text can allow systems to *generate* text with emotion. By simultaneously manipulating the style of both voice and text, designers can make the problem of consistency easier by deriving both from the same emotional base.

In contexts where no author predetermines the exact text to communicate, such as when a system itself compiles and presents information to the user, an interface can simultaneously tune both the emotional tone of the voice and the emotional tone of the text. For example, a system could make the interface sad by choosing words with depressing connotations *and* using a voice that was somewhat slower, lower in pitch, and narrower in pitch range than a typical voice. Similarly, a travel site could simultaneously tune linguistic and paralinguistic cues to describe a hotel for business or pleasure, using more formal language and a relatively flat voice in the former case and enthusiastic language and a happy voice in the latter.

Music presents a further modality through which a voice interface—and any audio-based interface, in general—can set a mood. People typically associate minor scales with negative emotions and major scales with more positive and happy emotions, for example.[50]

Emotions as Categories versus Emotions as Dimensions

Throughout this chapter and the previous chapter, emotion has been referred to in terms of *categories*. The categorical approach assumes that the best way to describe and understand emotions is to talk about the presence or absence of particular emotions, each represented by a single word. An alternative is the *dimensional* approach to emotion. Dimensional approaches assume that researchers can draw a simple coordinate system created from a very small number of axes to capture virtually all emotions; any and all emotion words can then be located at a particular point in this space.[51]

The most useful dimensional scheme was developed by Peter Lang and his colleagues.[52] Lang argues that most emotions can be placed along two independent dimensions—valence (which has been the focus of this chapter and the last), or how happy or sad someone feels, and arousal, or how excited or calm someone is. (A third dimension, control, is sometimes included.)[53] Using these dimensions, for example, disgust is extremely low on valence and neutral on arousal, and anger and fear are quite low on valence and high on arousal. Most other emotion words fit under this scheme, although it doesn't perfectly differentiate emotions; two different emotions may land in the same place.[54]

The advantage of the dimensional approach is that voices can simply be positioned along the two dimensions. The farther a voice is along the positive side of valence (happy instead of sad), the more the voice has higher pitch, wider pitch range, greater intensity, and upward as opposed to downward inflections. The farther a voice is along the excited side of arousal, the more the voice has higher pitch, wider pitch range, wider volume range, and faster word speed, for example. To create a particular emotion in a voice by using paralinguistic cues, a designer simply finds the coordinates of the emotion along each axis and selects the parameters accordingly.

Are Emotions Contagious?

When voice emotion activates different hemispheres of the brain of the listener, the activated hemisphere encourages the listener to *feel* the emotion; that is, recognition of voice emotion inspires a listener's emotional *response*. A drug called amabarbitol sodium, which limits communication between the hemispheres of the brain, helps demonstrate this phenomenon.[55] When the drug is injected into the right hemisphere, the left hemisphere becomes dominant and people report euphoric moods.[56] Conversely, when the drug is injected into the left hemisphere, people feel depressed as

the right hemisphere becomes dominant. Thus, the activation of the associated hemispheres can lead to *emotion contagion*,[57] the "catching" of the emotions of others. Sometimes this social phenomenon seems logical, such as when a person becomes afraid when seeing another experience fear: there may be a true danger nearby. At other times, contagion seems completely illogical, such as when another person's laughter induces immediate, "inexplicable" amusement. Contagion offers a partial explanation for why people generally enjoy being around happy people (and happy voices).

Full contagion, however, does not always occur. If the speaker's expressed emotion is strong and positive but the listener's current mood or orientation toward the speaker is unfavorable, the listener may catch the speaker's emotional arousal (excitement) but not the speaker's emotional valence (positive or negative aspect). In such cases, the listener could end up feeling strong negative emotion (such as frustration or jealousy), even though the speaker is expressing strong happiness. In general, excitement is the aspect of emotion that is most reliably transferred.[58] Thus, extreme emotions, whether positive or negative, will lead people to feel their own emotions, regardless of what they are, more intensely. Hence, designers should be cautious about presenting voices with highly accentuated emotions.

Linkages between Perceived Emotion and Perceived Traits

The human voice does not neatly distinguish between states and traits. The voice markers for submissiveness, a trait, are similar to the voice markers for depression, a state, for example. In the long run, this is not a problem: Personality sets a baseline for the voice, and emotion is then evaluated relative to that baseline. However, in the short run, states and traits become confused. This is why extroverts are initially perceived as happier than introverts.[59] Indeed, in the current study, participants who heard happy content read in a happy voice rated the content as more interesting to extroverts than those who heard the same content read in a sad voice.[60] Similarly, those who heard sad content read in a sad voice rated the stories as less interesting for extroverts than those who heard it read in a happy voice.[61]

Similar confusion occurs with emotion and gender. Females tend to have voice characteristics that are associated with positive valence and neutral arousal, while males tend to have voices that are associated with high arousal and neutral valence. Similarly, the gruffness associated with older voices can lead people to perceive the elderly as being in a bad mood.[62] Even though individuals know that these characteristics are simply markers of traits, they automatically apply them to deductions about states.

For limited-term interactions, there is little that designers can do about this confusion beyond selecting voices with perception of both traits and states in mind.

Providing Consistency in an Inconsistent World

Voice interfaces may not be capable of feeling emotion, but they are certainly capable of expressing emotion. In fact, it is unavoidable. Users respond to even the most monotone synthetic voice as if it were manifesting extremely low arousal and neutral valence. The content that interfaces present to users will also be associated with emotion, no matter how "unemotional" an interface tries to be. Because the brain tightly integrates the emotion of voice and the emotion of words, consistency between them is critical, even when they are generated by a clearly synthetic, unemotional machine.

Appendix

Table 8.1.A
Vocal effects associated with several basic emotions

	Fear	Anger	Sadness	Happiness	Disgust
Speech rate	Much faster	Slightly faster	Slightly slower	Faster or slower	Very much slower
Pitch average	Very much higher	Very much higher	Slightly lower	Much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Normal	Higher	Lower	Higher	Lower
Voice quality	Irregular voicing	Breathy chest tone	Resonant	Breathy blaring	Grumbled chest tone
Pitch changes	Normal	Abrupt on stressed syllables	Downward inflections	Smooth upward inflections	Wide downward terminal inflections
Articulation	Precise	Tense	Slurring	Normal	Normal

Source: I. R. Murray and J. L. Arnott, Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal Acoustical Society of America* 93, no. 2 (1993): 1097–1108.

9 When Are Many Voices Better Than One? People Differentiate Synthetic Voices

Each human voice reaches the human ear with its own range of frequencies, harmonics, and distortions. Within less than one second, each of these characteristics may vary in volume, pitch, speed, or cadence; may be influenced by a host of personal and situational realities; and may be interfered with by other voices or sounds in the environment. If the human voice were a tuning fork producing a clear tone and a unique frequency for each person, the brain's ability to differentiate and process multiple voices would not amaze. But voice qualities and environmental influences are obstacles that make humans' ability to distinguish one voice from another and even one voices out of many a remarkable feat.

What enables people to distinguish the many voices coming out of a speaker phone? How does the brain keep track of a presidential primary debate on the radio when the various candidates incessantly interrupt each other? How can travelers simultaneously comprehend their laptop announcing "You've got mail," their MP3 player intoning "Batteries are low," and a voice calling out "Room service"? How do parents simultaneously listen to three kids speaking, a televised news report, and a telephone caller?

The answer to these questions is that the human brain has evolved to solve the extraordinarily difficult problem of distinguishing one voice from another: voice differentiation is located on both sides of the brain (although words are adduced primarily with the left side of the brain).[1] As mentioned earlier, even a fetus in the womb can distinguish its mother's voice from all other voices,[2] while it takes days after birth before infants can distinguish their mother from other people via either sight[3] or smell.[4] Soon after birth, newborns can also distinguish one unfamiliar voice from another.[5] By eight months old, infants' brains are tuned to those aspects of sound waves that occur in the range of speech (and music).[6] Infants can also focus on one voice even if multiple voices are presented simultaneously—the so-called cocktail party effect.[7]

Researchers debate whether the human brain is built specifically to discriminate voices or whether this ability is simply an outgrowth of the general activation of voice processing coupled with pattern-matching skills. On the one hand, no training is required to detect differences in pitch or other fundamental characteristics of voices; novices are as good as experts.[8] On the other hand, clear evidence shows that voice discrimination draws on generalized cognitive abilities. For example, when a person is confronted with a new voice, cognitive load is increased.[9] In fact, when a great deal of cognitive attention is allocated to a task, people can exhibit "change deafness"—or the inability to detect when one voice changes to another.[10]

Humans can distinguish voices along an enormous number of dimensions. The most obvious category is the basic pitch of the voice, used to identify gender as well as other characteristics.[11] Pitch range, volume, and volume range can also be used to distinguish people,[12] although these (as well as pitch) can also be altered by the situation[13] (for example, a frightening situation makes all voices get higher and exhibit a great deal of frequency range). The brain also performs a highly complex analysis of the timbre of a voice, which includes all aspects of the sound waves (frequency spectrum) associated with voices beyond pitch and volume.[14] Timbre enables voices to be described as "smoky," "nasal," "breathy," "raspy," or "shrill." Timbre in the human voice is determined by the diaphragm, lungs, trachea, pharynx, larynx, vocal cords, glottis, oral cavity, jaw, teeth, tongue, and lips.[15] Each of these is influenced in turn by factors such as genetics, endocrine system, health history, physical environment, and habits.[16] With all these sources of differentiation, every body produces a different voice. Finally, accents,[17] word choice, and all other aspects of the linguistic features of voice also enable differentiation.

Because each human voice is unique in how it combines these dimensions, the human brain has evolved to draw a one-to-one correspondence between a speaker and the speech characteristics of a particular voice. A given person (at least within a limited time period) has only one voice (excluding impersonators and ventriloquists), and a given voice is generally associated with only one person (although two people may have remarkably similar voices). Thus, voice has become an automatic and highly efficient marker of *identity*.

Identity is important because, unlike ant and bee colonies (in which virtually all members behave identically to all other members),[18] people treat one another differently. Many evolutionary theorists argue that the rapid determination of the identity of a person was and is one of the most important abilities for survival among humans.[19] Voice, as a virtually universal means of communication, and faces became

the most readily apparent cues for identifying an interaction partner. Voice was and is particularly valuable because it can identify individuals in situations such as hunting, when eyes are occupied with other tasks.

Understanding the coupling between voice, body, and identity is extremely valuable for the brain. Multiple voices—coming from an adjacent room, for example—immediately indicate that multiple people are present, a critical fact for a social animal. The same coupling enables us to make sense of and enjoy radio: people naturally imagine a different person behind each of the different voices.

The inherent correspondence of voice and identity has moved beyond individual brains to social institutions. Police now routinely use *audio* lineups in which *ear*witnesses attempt to identify a person they heard during the commission of a crime,[20] although evidence suggests that earwitnesses are much less reliable than eyewitnesses.[21] In the first nonmilitary case involving the admissibility of voice-identification evidence,[22] the New Jersey Supreme Court ruled in 1968 that "the physical properties of a person's voice are identifying characteristics"[23] when recorded and analyzed by a specialist. The 2003 Federal Rules of Evidence (Article 9, Rule 901b5) clearly permit the use of voice identification in court, although the identification must include both spectrographic analysis and expert interpretation.[24] An increasing number of businesses, including the Home Shopping Network, are using fully automated "voiceprint" analysis to validate a person's identity; consumer acceptance of this technology is growing.[25]

The equivalence of voice, body, and identity breaks down in the world of technology. Traditional media like telephones, radios, and televisions have always produced more than one human voice. As data-storage solutions continue to offer more plentiful space in smaller and less expensive hardware, prerecording a number of different voices to come out of a single box becomes possible. Thus a single "body" can have multiple voices.

From their inception, systems for creating synthetic voices allowed changes to be made in the fundamental characteristics of the voice.[26] With advances in software, digital signal processors, and other hardware, a host of parameters that affect the timbre of a synthetic voice can be changed:[27] it has become as easy as changing the name of a file on a hard drive.

The proliferation of voice technologies challenges the idea that voice is truly a mark of identity or personhood: the intimate link between a particular physical object (a body or a box) and a voice is now broken. Indeed, in the digital era, the number of voices may become completely arbitrary. While multiple voices playing over the radio

imply multiple actors speaking in a remote studio, a computer's synthetic voices have no humans behind the voice at all: multiple synthetic voices can all be created from a single computer.

Do Multiple Synthetic Voices Matter?

The previous chapters have demonstrated that people can recognize and respond to social characteristics of technology-based voices, even when those voices are clearly synthetic. Users have no problem deciding that a voice is female and not male, extroverted and not introverted, or happy and not sad. However, will people make the leap from voice differences to voice identities, an appropriate conclusion when hearing human voices but a faulty one when applied to technologies? Put another way, will the evolutionarily useful and logical tendency to equate voices and identities be automatically extended to synthetic voices on interfaces? If people do make this extension, can the confusion be eliminated by reminding people that equating synthetic voices with different identities is illogical?

The most direct way to answer these questions would be simply to ask people directly. Researchers could have people listen to different voices on a computer, a Web site, or a refrigerator and say, "Do you think that those voices represent either different identities or people or the same identity or person?" However, this method of ascertaining an answer to the question is not as straightforward as it might appear. When questions like this are asked, the first response is usually one of incredulity: "Why would someone ask something like that?" After deciding that it is not a trick question, people generally laugh and reply, "Of course not. Mechanical voices have nothing to do with people or identities." More thoughtful individuals might add, "I suppose that children might think that a computer's voice has an identity, but no adult would ever think that." From the studies discussed in previous chapters, people are known to deny that they are influenced by the gender or personality of a computer voice, although their behaviors indicate that they are indeed influenced. Hence, a more subtle approach is needed.

In this study, the paradigm that was adopted was inspired by the "multiple-source effect"[28] (also known as the "source-magnification effect"):[29] an individual is more convinced by an opinion or a statement when it is expressed by a variety of people than when it is expressed by a single person. One explanation for why groups exert more social influence than individuals is *normative social influence*:[30] people like to hold the opinions of the groups with whom they interact, even when those opinions are almost certainly wrong.[31] Another explanation is *informational social influence*:[32] when multiple people believe something, differences between individuals suggest that a variety of information and information-processing strategies led to the same conclusion, making the conclusion more convincing. Of course, unanimous support from a number of people gives better evidence for public opinion than does a single person's opinion. If these findings are relevant to interfaces, then individuals should be more influenced by multiple synthetic voices rather than by a single synthetic voice, even if the content that the voices are relating is identical.

In addition to the multiple-source effect, a second criterion for determining whether listeners assign distinct identities to different synthetic voices is whether the voices make a person feel that he or she is *among* a group of people. The more voices that a person hears when walking into a party, for example, the greater the sense of others. That is, when many voices are heard, the people become more *present*.[33]

Hearing Multiple Voices on the Web

In this study, the context for exploring the psychology of multiple voices was a simulated book-selling Web site patterned after Amazon.com.[34] The site presented a separate Web page for each of three books. Each Web page included the book's title, the author's name, a photo of the cover, and five actual customer reviews from Amazon.com. All of the reviews for all of the books were positive so that participants would feel unambiguous social and normative influences. The reviews were not presented in text form, as they appear on Amazon.com. Instead, the site presented a link to an audio file that played when clicked. The reviews were described as coming from "Customer A" up to "Customer E." Some participants heard a single voice read all five reviews, and others heard multiple voices read the reviews.

Here are three of the five reviews about the book *Plainsong* by Kent Haruf:

Customer A: The pace of the story mimics that of the small town it takes place in. The characters are richly drawn but not caricatures. Not a lot happens, but I believe that's the point. There is no urgency to the story, and I liked it that way. The author, like James Lee Burke, has an affection for the beauty of the land. Reading this book is like taking a stroll down the main street of Holt County, in which the story takes place. Highly recommended.

Customer B: I loved this book. I stayed up all night reading it because I cared so much about the people. I could not put it down. I haven't done that since I was a child. After a lifetime of reading, I'm aware how rare this kind of experience is. And the reason? Kent Haruf's honesty, skill, and compassion as a writer.

Customer C: I really enjoyed this book. It starts slowly with disconnected stories. As it builds character development and speed, the stories become intertwined. I like stories like this anyway, but this is better than most. The descriptions of the plains, weather, and other

sensations are real. The characters are people you want to take home with you. A good read.

The participants had not heard about or read any of the three books, as indicated by their questionnaire responses. The books and their authors were selected based on low sales to increase the likelihood that participants would not be familiar with the books. All books were fiction to avoid bias based on participants' general knowledge of various topics.

Forty adults participated in the study. Twenty participants heard all five customer reviews via a single voice that was randomly selected for each participant. The other twenty participants (randomly selected) heard the five reviews delivered via five different synthetic voices that varied by fundamental frequency, speech rate, and timbre.[35] The order of the voices was varied for each person.[36] To avoid the possible effects of voice gender preferences,[37] all synthesized voices were male.

After participants listened to the reviews for the three books, they were directed to another Web site, where they were asked a series of questions about their own feelings about the reviewed books[38] and their perceptions of how the general public would feel about the books.[39] By looking at how positively participants rated the books, the researcher could measure how persuaded they were by the reviews (which were all positive). If the results indicated that participants were more persuaded by reviews that were read by multiple voices rather than by a single voice, even when the review contents were identical, that would be a strong indication that participants perceived the five different voices as five distinct identities. Questions about how the general public would feel about the books were designed to determine if the diversity of voices made people feel that a broader spectrum of the public liked a book. Finally, participants were asked (using questions derived from previous studies)[40] about their feelings of how present the voices seemed.[41]

Multiple Voices Activated Thoughts of Multiple Identities

Just as in the human-only world, participants in this study were more persuaded by several synthetic voices than by a single synthetic voice,[42] even though the comments were identical and the ideas of normative and informational influence should not be relevant. Users evaluated the books significantly more positively when they heard multiple voices,[43] consistent with the multiple-source effect. Because hearing various voices triggered the sense that numerous people agreed, listeners also perceived more public support for the book than did participants who heard a single voice.[44]

The varied voices also led to a feeling that the potential book buyer was surrounded by people: significantly more social presence was experienced when participants heard several voices than when they heard a single voice.[45]

This study shows that users do, in fact, perceive multiple synthetic voices to be multiple distinct entities. Synthetic voices manifest individuality in the same way that human voices do. The evolutionary tendency to use voice characteristics as a mark of individual identity thus overcomes the fact that one computer generated several voices.

Would a Reminder of Unnaturalness Have an Impact?

Although this experiment provides compelling evidence that multiple synthetic voices are perceived as multiple entities, the experiment could not tell how robust this effect would be. To address this issue, the experiment was rerun with a different group of forty participants, this time strongly emphasizing to participants that the various voices that they were going to hear were arbitrarily and randomly generated by the same computer. First, participants read the following brief description of how a text-to-speech system works:

> In this experiment, you will hear *machine* voices synthesized by a TTS (text-to-speech) system. TTS is the creation of audible speech from computer-readable text. Many TTS systems are currently available. Most of them can generate more than one type of voice from any given text by altering speech rate, fundamental frequency (F0), loudness, and other vocal parameters.

This warning was important because people often behave "mindlessly"[46] without thinking about all aspects of a situation. By reminding participants that the computer's voices were arbitrary and meaningless, the tendency to ignore this critical aspect of the situation should be decreased or eliminated.

Then participants logged on to the text-to-speech demo site at AT&T Labs, typed a sentence, and listened as the computer produced multiple versions of their sentence using different synthetic voices. This immediate, hands-on experience helped participants to understand that Web sites can easily produce many different synthetic voices. After experimenting with various TTS parameters and voices, participants went to the same book-selling Web site that participants in the previous experiment visited, heard the same reviews about each of the same three books, and answered the same questions. Again, half of the participants heard reviews read by several different synthetic voices, and half heard reviews by a single synthetic voice.[47]

Even after being told and shown that multiple synthetic voices could be easily produced by the same computer, participants showed no decrement in their tendency to

treat multiple synthetic voices as if they represented multiple distinct entities: voice activation is powerful. Participants were once again more persuaded when reviews were read by various voices: they evaluated the books more positively[48] and perceived public opinion of the books to be more positive.[49] Finally, although they were reminded that the creation of multiple voices was as simple as creating a single voice, participants nonetheless felt much greater social presence when they heard several voices than when they heard a single voice.[50] Users' tendency to process computer-generated voices in the same way that they process human voices proved to be extremely robust.

Other Evidence for the Multiple-Voice Effect

It might be argued that the previous research is compelling but applies only to a limited domain. First, the studies were limited to a single, though important context: e-commerce. Second, the studies employed synthetic speech. Finally, the studies were Web-based: Web users know that when they click on a link (even an audio file link), they can be sent to a different source than that of the initial site.

Fortunately, two other studies remedy these deficiencies. One study examined politeness[51]—particularly the tendency for people to say nicer things to a person who asks about herself or himself than to a person who asks about another.[52] Participants in this study first did a task with a computer that spoke with a recorded, high-fidelity voice and then answered the computer's spoken questions about the computer's performance. Half of the participants were asked the evaluation questions by the voice that spoke during the task; the other half were asked the evaluation questions by a different recorded voice. Participants provided significantly more positive responses to the computer when it asked with its original voice than when it asked with a different voice,[53] which demonstrates the power of multiple voices to encourage both identity and social norms.

The other study looked at self-praise and other praise in the context of a computer tutoring application.[54] In this experiment, participants first learned from a computer with a recorded voice. Then half of the participants heard the computer praise its tutoring using the same voice as in the tutoring session, and half of participants heard the computer praise its tutoring using the same words but a different voice. Consistent with the notion that other praise is more believable than self-praise[55] and that different voices represent different identities, strong differences in response were observed between the people who heard praise coming from the voice that taught them and people who heard praise coming from a different voice. Although the words

used in the tutor's performance and in the praise of the tutor's performance were identical, participants perceived the different-voice evaluation to be significantly more accurate than the other participants perceived the same-voice evaluation.[56] Different-voice participants also perceived the evaluation session to be fairer than did same-voice participants.[57] That is, distinct voices lead to responses consistent with distinct identities.

Multiple Voices as Specialists, One Voice as a Generalist[58]

Given how attuned the human ear and brain are to detecting differences in voices, it is not surprising that different voices, even different synthetic voices, seem to evoke perceptions of different "people" or at least different "social actors." But it is striking that this differentiation of voices activates the cognitive apparatus that evolved to assign distinct identities to voices. Thus, changing voices within an interface is not the same as changing colors or fonts:[59] the latter is simply an artistic (and perhaps cognitive) decision; the former has social consequences. The implications for design are first discussed in terms of when to use multiple voices versus when to use a single voice. The chapter then turns to some issues that are unique to multiple voices.

The simplest way for designers to decide whether to use multiple voices or a single voice is to ask two questions: "When would users like to be working with multiple people, and when would they like to work with just a single person?" and "Do I want this device or service to feel like a collection of parts or a unified whole?" Here are some of the guidelines that the research literature provides for answering these questions.

Specialists Are a Plus

When people need help doing something complicated or important, they turn to a specialist. Experiments have shown that the products of specialists are perceived to be better than the products of generalists,[60] even when their contents are identical. A more surprising finding is that expertise can be inferred simply from a label;[61] demonstrated performance is unnecessary. At the end of the film *The Wizard of Oz*, the sham Wizard leverages this principle: declare someone intelligent (the Scarecrow) or brave (the Lion), and a perception of that characteristic will follow.

Using two computer voices, a designer can create a specialist simply by having one voice designate the other as expert (as noted above, self-labeling is not as compelling). Even though people will know that the voices are coming from the same computer,

the automatic tendency to separate voices should lead to enhanced perceptions of the "specialist's" performance.

Specialist voices can be used compellingly when a function is known to require significant expertise. For example, a car interface might have one voice for handling general commands from the driver but a specialist voice for mechanical problems and navigation, which require highly specific knowledge and understanding. An MP3 player might have one general voice for basic functionality and a separate voice for music recommendations, a highly personalized and complex activity. An e-commerce site might use one voice to excite people about the products and another voice to take people through the tedium of filling out the order form.

Our favorite example of a specialist voice uses the good cop/bad cop scenario in a prototype voice-mail system:

Voice 1 (warm and friendly): We at the XYZ Company look forward to helping you. Please use our new telephone system so that you can be routed to the right person.

Voice 2 (formal and unfriendly): Press 1 for Department A, 2 for Department B. . . .

After the user passed through a couple of voice-mail levels, he or she heard this message:

Voice 1: We at the XYZ Company are very sorry that this voice system is making it difficult for you to get through. Please bear with it.

Voice 2: Press 1 for Supervisor Q, 2 for Supervisor R. . . .

Voice 1: We at the XYZ Company hate this system as much as you do. We're doing everything we can go get your call to the right place.

Finally, after a few more rounds with the bureaucratic Voice 2, the caller's line was answered by a live person.

We then interviewed participants and asked how they felt about the voice-mail system. "Boy, I really hate that voice-mail system. But it's great how much XYZ cares about me!" The clever idea here was that the warm and friendly voice of XYZ *distanced itself* from the navigation voice. Indeed, all references to XYZ were made by the voice that was "on the side of the user." The voice of XYZ thus made the navigation voice a "scapegoat."[62] The power of voice differentiation led listeners to ignore the fact that it was XYZ that deployed the annoying system in the first place.

A final advantage of specialist voices is that a particular voice can be selected to be uniquely relevant to particular tasks. Because even synthetic voices can manifest gender, accent, and personality,[63] one voice might suggest maximal expertise, and another might best challenge stereotypes for each application or service. For example, the interface to a mathematical application might be male, middle-age, and introverted sounding, or it might break those stereotypes to encourage the social goal of inclusiveness.

Specialists Are a Minus

There are downsides to having specialists. Imagine that an interface has a general master of ceremonies (M.C.) voice that directs people to various "experts" with distinct voices. Although the experts will be seen as highly competent, the M.C. will be seen as unable to perform even the simplest task. If the M.C. becomes associated with the product or service, his or her incompetence may lead to a "negative halo effect"[64] in which weakness in one characteristic is imputed to all characteristics, including the product.

Another problem with employing specialist voices is that multiple voices exact a cognitive price.[65] If the system introduces new voices precisely at the times when the user is performing a complex activity (which is why the system employed a specialist in the first place), then the person's performance could actually be impeded as the person tries to cope with the new cognitive load.

In an elegant series of studies, Teenie Matlock, Paul Maglio, and colleagues have explored how a lack of specialization can have benefits when a caller speaks to a system. In their studies, they asked participants to speak to devices that perform various functions (such as a calendar, a trip planner, an address book, or an e-mail device). When people were asked to "talk to the system" (a single, generalist place), people took less time and were overall more efficient in their interactions (for example, they used more pronouns than full noun phrases) than when they were asked to "talk to each device" (several specialist places).[66] Furthermore, when speaking to the "system," users would feel free to look anywhere; when talking to the "device," they looked directly at it and referred to it by name (for example, "Address book, show me the address for . . ." versus "Show me the address for . . ").[67] Participants who worked with the talk-to-the-system condition judged it to be more natural and more like working with a colleague than did those who worked with the talk-to-the-device condition.[68]

A final cautionary note about interface specialists is that they should not be employed when a real-life specialist would *not* be appropriate. For example, in one prototype telephone-based system that involved a "virtual secretary," the user interacted with one voice for voice mail and a different voice for e-mail (which was read to the users), even though the structure of commands and replies was identical. Users found this puzzling and distracting and wondered why two voices were speaking to them when their functions were logically the same. Using voices in this way can also

make users feel that "no one" wants to take responsibility, much as people grow frustrated when they are bounced from one person to another to another and back to the first when navigating a bureaucracy. In sum, specialist voices can give extraordinary power and credibility to an interface, but caution must be exercised.

A Single Voice in Multiple Contexts

Once a particular voice is assigned to a particular identity, the human brain is very effective at recognizing the voice. Parents can easily discern their child's voice in the sobbing, virtually incomprehensible phone call from summer camp; the giggling description of a day at school; the static-filled cell-phone conversation; the hymn in church; the echoing, unlocatable voice running through a cave; and the recording from five years earlier. People can also distinguish the voices of others they have never met in person. A listener can automatically recognize that the same speaker, James Earl Jones, breathily intones, "Luke, I am your father,"[69] announces "This is CNN," and played the submissive servant in *"Master Harold" . . . and the Boys*.[70] The evocation of identity through voice even applies to fictional characters: it is impossible to imagine Barney or Tickle Me Elmo dolls, both enormous sellers in the United States, with a voice other than the one that the character uses on television.

The brain's talent for discerning identity from voice is even robust against sound cards and speaker quality. Humans can recognize the same voice whether it comes from a live person, a high-fidelity music player, a medium-fidelity television, a low-fidelity cell phone, or an even lower-fidelity plastic toy. The ability to be both thoroughly forgiving of variations in a single voice and acutely sensitive to distinguishing different voices is a remarkable combination. Part of the explanation for how these seemingly opposite abilities can reside in the same person is that they reside in different parts of the brain.[71] The ability to recognize a familiar voice is located in the inferior and lateral parietal regions of the right hemisphere of the brain (which is not the hemisphere associated with voice processing), while the ability to distinguish between two voices is located in the temporal lobes of both hemispheres of the brain.[72]

What happens when the same voice appears on multiple technologies? Research has shown that the strong link between personhood and voice means that the same voice from two different technologies will tend to blur the segmentation between the technologies and lead the user to perceive a single identity, just as various voices on a single machine will feel like multiple identities. For example, in one experiment, the same voice spoke from two computers and was responded to as if it represented a single person, while, as noted above, a second voice on the same computer was experienced as a different person.[73]

What should designers do when they have an option of using the same voice across product lines or different voices for each product? It depends on what the company wants to brand.[74] For example, if a company wants to suggest homogeneity across a set of media technologies—such as a television, a stereo, and a DVD player—the same voice should be used across products. However, if the company wants consumers to orient to a particular product, then having quite different voices across the various products would be worthwhile. Finally, if the company wants consumers to orient to the services provided around the product rather than the product itself (such as when most of the profit comes from monthly subscriptions rather than the initial purchase), then the interface to the product itself should be subdued, either via text or via a soft, terse voice.

Multiple Voices in a Single Context

Very often, designers require a single piece of technology to produce multiple voices. This is the case for telephones, radios, televisions, CD players, and video games. These technologies would not be effective if they homogenized all voices to sound alike. Because these technologies do not use a single voice to connect users with the product, the product becomes psychologically *invisible*: people focus solely on the voices and what they say.

This invisibility now extends beyond the psychological to the physical design of technologies. Televisions have transitioned from huge picture tubes to flat screens, cell phones have become credit-card size, and MP3 players and Walkmans are now no larger than the tiny disks or cassettes that they play. This is also the spirit of "ubiquitous computing,"[75] in which systems are designed to encourage people to forget about the technology and simply orient to the voice and task. The lack of a physical presence facilitates the sense that the product is a conduit or medium rather than an independent identity.

A Voice in the Crowd

When brains encounter groups of two or more people, they instantly begin to analyze how the group members feel about and interact with one another.[76] Are they lovers, friends, acquaintances, or enemies? Even when the voices do not speak directly to one another, if one voice is affected by another's comments, then a social dynamic emerges

among the voices. For example, a voice that always interrupts other voices would likely be seen as dominant or abusive, while a voice that never interrupts, even when such interruption would be justified, would be seen as submissive. In these cases, it is particularly important that the gender, accent, personality, and emotion of the voice match its behavior.[77] Of course, as the number of voices increases, the user will have to spend cognitive energy processing each of the voices and also keeping track of the social relationships among them, even if subconsciously. Hence, listening to too many voices is exhausting.

What if designers want one voice to stand out from the others? A few subtle but powerful options are available. One strategy that is used effectively in television advertising is to make one voice louder than the others. The human brain orients to the loudest voice,[78] probably because throughout evolutionary history, the loudness of a voice was associated with the proximity of the speaker, and physically close people represented a greater opportunity and a greater threat than more distant people.

A second strategy for making a voice memorable is to have people interact with it. People have a better memory for voices with which they have interacted than for voices that they have listened to passively.[79] By picking one voice for all interactions and restricting all other voices to information provision, the interactive voice will become most prominent.

A final strategy is to use a typical voice along with a set of unusual voices. Each unusual voice will be stored in memory as a deviant from a prototype,[80] that is, as "weird." Conversely, the typical voice will be stored as prototypical. Although the unusual voices might be arresting in the short run,[81] because standard, well-defined categories are easier to access from memory than fuzzy categories, the unusual voices will all blur together and thus be less prominent than the single typical voice.

Should Users Choose the Voice of Their Interface?

A halfway measure between requiring a single voice and having the designer employ a wide range of voices is to have the user select one voice from a set of voices. This could come at the time of purchase or could be an option throughout the life of the product. The booming market for cell-phone "ringtones" demonstrates the attractiveness of this idea.

Imagine, for example, that all cars were equipped with a voice interface. Should every car of a particular manufacturer speak using the same voice, or should each driver select his or her own unique car voice? Similarly, would people enjoy their interactions with customer support more if they could choose the telephone voice that provides the support?

The key point is that choice is psychologically powerful.[82] In general, because of cognitive dissonance, people will like the voices they select and will try to avoid information about the desirability of other choices[83] (although there is some evidence for buyer's remorse,[84] the tendency to regret choices soon after making them). There is no evidence that people are successful at selecting the voices they will like in the long run; however, the mere act of selection can be a positive experience. The feeling that the user hears a voice that virtually no other user will hear may create a perception of a unique interpersonal bond between the user and the voice, especially if the user interacts with the voice and does not merely listen to it.[85]

How should the choice be managed? The best approach is to have a second computer voice (that is, a voice other than the voice the user is currently using) propose the option of selecting a different voice. Because people are polite to voices, they might find it hard to "reject" a voice with whom they have been working.[86]

Another critical point is that too many options can be overwhelming. In a classic series of studies performed in both field and laboratory settings, Sheena S. Iyengar and Mark R. Lepper showed that people were more likely to purchase gourmet jams and to undertake an optional class essay assignment when offered only six choices than when they were offered twenty-four or thirty choices.[87] Furthermore, participants felt better about their choice and performed better when their selections were limited. Unfortunately, there are no clear guidelines that prescribe exactly how many choices of voices (or choices of any product or service) are too many. Assuming that the voices are quite different, the optimal number of choices might be approximately seven because that is approximately the number of items that can be kept in short-term memory.[88] If the voices can be categorized (for example, by gender), participants will likely exhibit a greater tolerance and desire for more options. All of the options should remain consistent with the brand.

Currently, the goal of cutting-edge voice-interface designers is to allow users to create a voice for the interface: better algorithms will soon make this a reality. Some people might even choose to have their own voices for their interface; research shows that this is not disturbing.[89] Others may choose the voice of a family member, a relative, a friend, or a famous celebrity. The danger of allowing this level of personalization is that the product or service may become invisible because of the presence of the voice, which thereby will undermine branding.

Other Challenges of Using Multiple Voices

The most significant practical issue that is associated with multiple voices is the maintenance problem. When designers add new content to a recorded voice system with

14 Misrecognition: To Err Is Interface; To Blame, Complex

When two people perform a task together rather than in parallel, they obtain significant benefits: dyads lead to better learning, better performance, and greater satisfaction than the same two people working independently can produce.[1] That is, two heads *are* better than one.[2]

How does being a member of a dyad affect each member's feelings of success? "One of the best established, most often replicated findings in social psychology"[3] is that people exhibit the self-serving bias.[4] That is, virtually all people—not only the inhabitants of Lake Wobegon[5]—believe that they are above average. Thus, in the context of a successful dyad, both participants claim primary responsibility for the team's accomplishments.[6]

The self-serving bias also leads individuals to assign primary responsibility to anyone and anything but themselves when they fail.[7] In a dyad, this often means that each person blames the other person in the dyad, although individuals also can blame the unfairness of the task or the difficulty of the environment.[8]

When a member of a dyad discovers a problem and cannot independently solve it, he or she must alert the other member of the dyad.[9] There are two basic methods for announcing a problem—by taking the blame or by assigning the blame to another. Sometimes the person who discovers the problem takes the blame for it. These modest responses have many desirable features,[10] especially for women.[11] First, they confirm the other person's assumption, grounded in the self-serving bias, that he or she was not responsible. This confirmation of the person's expectations reduces cognitive load and increases comfort with the interaction. In addition, when the person who discovers the problem admits a failure, this acceptance of responsibility suggests commitment to success and implicitly urges the other person to be committed in return.[12]

Unfortunately, modesty does have one serious problem: modest comments are believed to be true.[13] That is, when people continually blame themselves for errors, they seem incompetent (though likeable). The reason that modest remarks and

skepticism toward self-praise[14] are accepted is that there is a generalized belief that when people say things that they should not be motivated to say, they are likely telling the truth (this is why U.S. courts give "statements against interest" special validity).[15] This belief also explains why women seem to benefit more from self-praise than men:[16] modesty is a normatively positive trait for women,[17] so when women speak positively of themselves, the comments are given extra weight.

Modesty in Human-Computer Dyads

As with human dyads, human-technology dyads trigger feelings of success and failure. For the most part, these dyads succeed, in that people obtain or manipulate the information they need. In these instances, the interface can compliment users (kindness is an effective strategy),[18] thank the user for completing the task (politeness is an effective strategy),[19] or spread praise by saying "we"[20] (sharing credit is an effective strategy).[21] When users say, "Thank you," a remarkably frequent occurrence, the interface can simply accept the praise silently or mutter an occasional, "You're welcome."

Unfortunately, voice user interfaces must deal with the problem of failure on a regular basis. Unlike textual or graphical user interfaces (GUIs), the system does not always understand what the user requests. Even with word-recognition accuracy of 95 percent, which is extraordinarily high given current and even anticipated technology,[22] a voice interface will, on average, have one recognition error out of every twenty user commands or statements. Virtually everyone can recall an automated phone system, an in-car navigation system, or a computer-based dictation system that failed to understand something they said.

Because the interface does not know what to do next, it must alert the user to its lack of understanding. If the literature on human-human dyads can be applied to human-interface interaction, then two of the options that are available to people may also be available to interfaces. The approach almost always adopted by voice-interface designers is to have the system blame itself. That is, interfaces say, "Sorry, I didn't understand you," "Sorry, I didn't get that," "I'm sorry. I'm not finding a match," or "This system could not recognize what you said." These modest responses may relieve users from the burden of responsibility and may also lead them to feel that the interface is committed to making the interaction work.[23]

However, when the perceived competence of the system is a critical concern (as, for example, in cases of high consequence such as stock ordering or medical information),

a different strategy may be necessary. Because modesty is accepted as accurate, especially with male voices, these systems may benefit from blaming the user rather than themselves. That is, a voice interface could accuse the user of not speaking clearly enough or of not paying attention. Although these kinds of interfaces might annoy users, they also might have the advantage of being perceived as smarter than the typical modest interface.

"It's Not You; It's Me." "But on the Other Hand . . ."

Given the prominence of misrecognition in voice interfaces, an experiment[24] that investigated the effects of modesty seemed worthwhile. Thirty-six participants were directed to a Web site where they were told that they would be evaluating a new phone-based system for book buying. The Web site included a phone number and a list of tasks that participants needed to complete by using the telephone-based system,[25] including searching for particular books and listening to their descriptions, browsing the best-sellers' list, placing books on a wish list or in a shopping cart, and purchasing books using surrogate credit-card and address information. Users heard descriptions of four books, and the system employed a male voice.[26]

Participants were given a specific set of tasks to accomplish in a specific order. Hence, the system essentially knew what the user was going to say next and did not, in reality, have to perform speech recognition. Using this approach, there was no possibility of a misrecognition error. However, for purposes of the experiment, at ten preselected points in the interaction, the system would *say* that there was a misrecognition problem. For half of the participants, the system blamed itself for each of these (supposed) errors—that is, the system was modest. For the other half of the participants, the system blamed the user for each of the (supposed) errors.

The modest interface used traditional error messages. To avoid the complexities of using *I*,[27] the system referred to itself as "the system," although it did use the active voice to make clear that it was taking responsibility.[28] The self-blaming system said one of the following four sentences each time that it ostensibly could not understand the user:

The system did not understand the selection. Please repeat it.

The system could not process your selection. Please say it again.

The system has a problem and was unable to recognize that response. Please repeat your selection.

The system could not understand that selection. Please say it again.

Conversely, in the user-blame condition, the system clearly blamed the user for the misrecognition problem, using one of the following four messages:

You are speaking too quickly. Please repeat your response.
Please repeat your response. You should speak directly into the receiver.
You must speak more clearly. Please repeat what you said.
You are speaking too softly. Please repeat your selection in a louder voice.

After completing all of their assigned tasks, participants filled out an online questionnaire.[29] Participants were first asked to rate how likeable[30] the system was. Next, because user frustration is one of the main problems with misrecognition in current voice-based systems, participants were asked to rate their level of frustration[31] while using the system. Participants were also asked how likely they would be to purchase each of the books they heard described when they used the system.[32] If people feel happier and more comfortable with the modest interface, each of these should be positively influenced by modesty. To test the one characteristic that might be positively associated with criticism of the user, the experiment measured the perceived competence of the system.[33]

Blame, Blame Go Away

What happened?[34] Participants did not like getting blamed for speech-recognition problems. Participants rated the system that blamed the problem on them as less likeable than the self-blaming system. Participants who used the user-blaming system also found the interaction to be much more frustrating than the system that blamed itself. The criticism also hit the designer's bottom line: people were less likely to buy the books when they were criticized than when the system took the blame.

This all suggests that modesty in the interface's language makes users feel better and more comfortable with the interface but that "self"-criticism by an interface leads to the same problems for interfaces as it does for people: the interface is perceived as significantly less competent when it blames itself. That is, people take the interface at its word when it says that it performed poorly.

Is There an Option Other Than Blaming the System or Blaming the User?

The previous results do not provide an optimal solution to the problem of voice misrecognition. No one likes to get blamed (users included), but a system that blames itself suffers in perceived competence. Given that both self-blame and other-blame

have deficiencies, one approach might be to seem to blame no one. For example, the system could simply say, "There is a problem" or "Please say that again." Given the self-serving bias, people would likely interpret such statements as equivalent to the system blaming itself. However, a problem with this approach is that users may also interpret the remarks as an attempt by the system to avoid taking responsibility (as President Richard Nixon did when he said, "Mistakes were made").

A more promising approach, which might enable a voice interface to retain perceived competence while not frustrating the user, is suggested by the research literature: have the system identify a plausible scapegoat.[35] That is, rather than blame the user or itself, the system can indicate than an external impediment prevented understanding between the dyad. For example, the system could blame the noisy environment or problems with the wireless service, whether or not these actually caused the voice misrecognition.

In many situations, scapegoating can be just as ineffective as blaming no one. The root problem is the "fundamental attribution error,"[36] which suggests that people blame an individual's failures on her or his own faults rather than on problems in the external environment. That is, people view others' behaviors in exactly the opposite way that they view their own. Thus, scapegoating can seem like an attempt, both unbelievable and contemptible, to avoid responsibility.

However, when members of a dyad are relatively mistake-free, the psychology is very different.[37] Because both members of the dyad feel that they are part of a team and jointly responsible for the outcome, feelings of interdependence lead the two members to apply the self-serving bias to themselves and their partner.[38] That is, both members of the dyad perceive failure to be the responsibility of neither partner of the dyad. In these cases, then, blaming a scapegoat meets all of the team's needs.

A study that was performed in the context of a voice-based car interface provides support that scapegoating can be an effective strategy for human-voice-interface dyads.[39] This study does not directly address misrecognition, but the participants' reactions to scapegoating should be applicable to misrecognition as well as driving. Driving was selected because it represents a situation in which the self-serving bias is in full flower: 80 percent of drivers in accidents indicate that it was not their fault.[40]

In the study, twelve male and twelve female participants drove a car simulator[41] that simulated driving on country roads with various driving conditions (such as multiple cars and fog) for twenty minutes. The simulator included a gas pedal, a brake pedal, and a force-feedback steering wheel. All participants drove on the same simulated course.

After a nine-minute practice session, a male voice calling itself "Chris" presented verbal warnings regarding the current driving situation. Throughout the driving session, Chris made twenty-one comments indicating that either the driver was performing poorly or that road conditions were posing driving challenges. Specifically, for half of the participants, Chris blamed the driver, with such comments as these:

You're driving too fast.
Your steering is very erratic.
You are not braking fast enough.
You should slow down when taking the curves.
You should pay more attention to the road.
You should turn more carefully.

The other half of the participants heard Chris identify the same problems but blame something external. That is, neither the driver nor Chris was at fault for poor performance: something outside the dyad was causing the problem. Here is how Chris blamed the environment for the same problems:

This road is easy to handle at slow speeds.
Steering is difficult on this road.
This road requires frequent braking.
These curves require slow speed.
The road demands a lot of focus.
Turns on this road appear very quickly.
The conditions make it hard to maintain a constant speed.

To determine whether scapegoating would be as effective in human-car dyads as in human-human dyads, participants filled out an online questionnaire[42] after the driving session. The questionnaire first asked participants how well they drove.[43] The questionnaire then asked about their perceptions of the voice[44] and the car.[45] Drivers' attention to the road was measured by determining how quickly they could respond to eleven horn honks scattered throughout the driving session.[46]

Consistent with the results from traditional dyads,[47] people liked the voice more[48] when it attributed poor driving to external forces rather than to the dyad itself. They also thought that the car was better. Beyond positive feelings about the other member of the dyad, participants who were not directly blamed felt better about their own driving than those who were blamed felt. Most important of all, people who worked with a partner who scapegoated actually paid more attention to the road than those who were blamed—that is, blamed drivers showed less attention, even though they were more directly urged to drive better.

Avoiding Errors before They Happen

The previous experiments presented a voice-interface problem that the system needed to resolve, but clearly the larger goal is to avoid having problems arise in the first place. For systems that rely heavily on speech recognition, error rates will steadily decline with improvements in processor speed and algorithms, reductions in the costs of hard disks and memory, and larger databases for comparing user responses. Regardless of the quality of the underlying technology, designs that reflect the social aspects of speech can reduce errors even further.

Users Will Automatically Mirror the Interface

One common limitation in speech-recognition systems is that the words and phrases that they can recognize are more limited than those that the average native speaker can recognize. A solution that does not rely on advances in technology is *semantic alignment*,[49] the principle that people automatically and unconsciously use the same words that their interaction partner uses. For example, one voice user interface for e-mail was having trouble distinguishing between the phrases "read it" and "delete it," which was a serious problem. The solution was to have the system say, "Your next e-mail is from Fred Smith. Would you like me to read it, save it, or throw it away?" thereby discouraging users from saying, "Delete it."

Another way to encourage users to employ a particular phrase is through "air quotes"—that is, pauses between the key words that the system wants the user to say. Thus, in the previous example, the system would say, "Do you want me to [long pause] read it, [long pause] save it, or [long pause] throw it away?" While this strategy is necessary for new users, the principle of semantic alignment suggests that over a period of time, the user will *automatically* use the words they hear,[50] regardless of whether they are highlighted by "air quotes" or other methods, such as increased volume. Thus, air quotes can recede in prominence as the user naturally adapts to the system.

Newer speech-interface systems can recognize more than single words or phrases. Users can now use full sentences to speak to many systems, and some systems can even support open-ended responses to such utterances as, "How can I help you today?" In one sense, these systems are intrinsically more cooperative than the classical systems that require specific, one-word responses. The "controlling" one-word systems can be frustrating for the speaker, who can tend to feel that the system is passive-aggressively controlling the conversation. A user may feel like the falsely accused witness who is asked, "Answer yes or no: Were you drunk when you hit your wife?" Any possible response is intrinsically inaccurate.

The problem with open-ended systems is that their failure rate is much greater than that of the limited-option systems. Fortunately, research has shown that people will not only align semantically with technology; they will also align syntactically—that is, use the same grammatical structures that the system uses.[51] Hence, the continual use of the same easily recognizable grammatical constructions by the system will actually lead to a significant improvement in recognition through user conformance.

In addition to mirroring words and syntax, people will even mirror the system's accent, cadence, and style of speech—a process that is known as accommodation.[52] In most interfaces, the voice speaks much more rapidly than the average person; thus, the recognition engine must be prepared for rapid speech and resultant poorer enunciations from users. To reduce the burden on the recognition system, the voice should speak slowly and with very clear enunciation, which will lead the user to do the same.[53]

Using Incompetence to Improve Recognition

Another strategy for encouraging users to speak understandably is for the system to suggest limited competence. Some markers of bounded skills include synthetic speech,[54] slowness of speech,[55] and fidelity of speech[56] (to which the human brain is even more sensitive than visual fidelity).[57] As far as content, speaking nongrammatically,[58] avoiding multisyllabic words, using only simple sentence constructions, and blaming oneself for failure also indicate limited intelligence. Just as people speak more simply and clearly to a foreign speaker or a child than to a native adult,[59] users will simplify and homogenize their speech for a simplistic system, thereby increasing recognition rates and discouraging deviance from the stated options.

Conversely, markers of competence will encourage users to accommodate to the system by adopting a rich vocabulary.[60] However, this could lead to much poorer performance and hence more negative responses.[61]

Making Users Try Harder

Another way to improve the error rate of voice-recognition systems is to encourage users to adapt to the system's limitations. Users will work harder to make the system understand them when the system matches them on characteristics such as gender, personality, and accent.[62] Social identification leads people to feel more committed to the dyad's success[63] and thereby to search for the proper way to say things and the proper things to say.

Design practices that make users feel that the system is *trying* to understand them will encourage users to try harder to help the system understand in return.[64]

Reciprocity has been shown to be a powerful inducement for users who work with both people and technologies.[65] One mechanism that suggests that the system is paying full attention to the user is for the system to repeat the idiosyncratic words and phrases that the user says.[66] This approach partially explains the tremendous success of the computer therapist Eliza,[67] which could not understand what the user said but merely repeated back the user's words and phrases. A second strategy for suggesting attention and effort is to pause before giving a response to an important or complex user request. Even if the system could respond immediately, delaying for a second or two suggests that the system is taking the user's statement seriously.[68]

Confirming Understanding

At times, the system may be unsure about whether it has understood the user correctly. When misunderstanding is highly consequential (such as for stock orders or flight-arrival time checks), the system can adopt a standard approach that is also used in human-human communication—confirmation. For example, if the system is unsure whether the user said "IBM" or "Spy-BM," it can explicitly ask, "Did you say 'Spy-B-M'?" Such confirmation enables the system to address potential misrecognition problems before they become disasters.

A more subtle approach is to repeat the utterance as part of the answer, as in "The stock price of Spy-B-M is $10," "Flight 73 is leaving Houston at 10:57," or "Now selecting radio station 680 FM," rather than simply saying "$10," "10:57," and "Radio is set." The responsibility for flagging an error is then on the user. As discussed in an earlier chapter, the entire sentence should be spoken in the same voice, not the fixed part in recorded speech and the variable part in synthetic speech.[69]

Keeping the Conversation Going

Misrecognitions are not the only errors that can disrupt a voice interaction. Even if recognition were perfect, the failure of the system to follow conversational norms can frustrate and even infuriate users. Over the millennia, humans have developed some ingenious principles to ensure that conversations work effectively and do not break down.[70] These conversational *maxims*[71] have become so much a part of language that people are normally not aware of how important they are. Because the knowledge is tacit rather than explicit, designers often overlook these maxims, leading to widespread user dissatisfaction and early termination of interactions. The three maxims most often violated by voice interfaces are quantity, relevance, and clarity.

Quantity

One of the fundamental unspoken rules in conversation is that conversants should provide no more or less information than the conversation demands. This is known as the maxim of quantity.[72] Voice interfaces are notoriously bad at following this convention. For example, imagine a voice user interface that asks ten yes or no questions. The first question might ask, "Are you female? Please answer yes or no." The second question would then follow, "Are you a U.S. citizen? Please answer yes or no." By the time the third question, "Are you over twenty-one years of age?" is asked, it would be a violation of the maxim of quantity to once again say, "Please answer yes or no." Because voice activates the (social) human-computer relationship, users would interpret a third request to "Please answer yes or no" in one of two ways: either "The system does not care enough about me to think about what it is asking," or "The system is trying to make me feel stupid," even if they also viewed the interface as "machinelike."

The problem of too much quantity also often occurs when porting a graphical user interface from the Web or a PC to voice. Because voices are very different than text,[73] large menus are extremely difficult to maneuver through in voice. The traditional strategy has been to create multiple levels of hierarchy to avoid an overwhelming number of options. For example, a repair system may have to deal with four types of televisions, VCRs, and stereos each. To avoid listing twelve options at once, the system can say, "Do you need help with your TV, VCR, or stereo?" If the user says "stereo," the system can then say, "Which model of stereo: the 200, the 300, the 400, or the 500?" On the other hand, if the hierarchy is created arbitrarily simply to reduce the number of options that are presented at any one time, these groupings increase cognitive load and user errors. The better alternative is to say a small number of options, no more than seven,[74] with the last option being "Something else." If only a few options represent the majority of requests, the system should not provide even seven possibilities: it should simply state the critical ones and finish with "Other options."

Sometimes, designers cannot gauge, before the interaction begins, how much information is appropriate. *Barge-in* is a technique for solving the overspeaking problem. Under barge-in, users may speak at any time, which causes the system to stop speaking and begin to interpret their utterances. Allowing users to speak as soon as they know what to say allows the interaction to move along quickly and leads users to feel that the system is cooperating by not saying more than necessary. In car contexts, barge-in is not effective because of the noisy environment. *Push-to-talk* addresses this constraint: when the user pushes a button, the system stops speaking and listens to the user. If users barge in a great deal, the system should assume that it is saying too much: solutions include curtailing the lists of options and shortening utterances.

Too little quantity can also be a problem in voice interfaces. For example, systems sometimes fail to describe the options in enough detail for users to know what they should say. For example, if a computer repair company's interface says, "Please indicate whether you have a problem with your motherboard or not," many users might think, "If I knew what a motherboard was, I probably wouldn't have to call you!" When users simply pick an option in desperation, the call center incurs high expenses as an expert, paid to work on a problem within her or his scope of expertise, must explain to users that they have the wrong person.

Another example of an underuse of quantity is the simple question, "How can I help you?" Questions like this provide too little information unless the system has a recognition capacity that is sufficient to accept virtually any utterance. The alternative need not be as crude as "Please say one of these three options: A, B, or C." Instead, the user can be asked, "Do you need help with A, B, or C?"

Relevance

Relevance is a second maxim requirement for successful conversation. Conversants are expected to make comments and ask questions that advance the conversation.[75] Users become frustrated when they cannot figure out why voice interfaces request information before it seems relevant. For example, credit-card information should be requested only when the transaction is completed, much as in a bricks-and-mortar interaction: early requests seem like bullying by or nosiness of the system.

Relevance should not be interpreted strictly as "task specificity." As noted earlier, humor,[76] an occasional flattering remark ("You have shown excellent judgment in your selections"),[77] and polite comments (such as "Thank you")[78] are not directly relevant to the task, but they support the social expectations activated by all voice interfaces and so are "relevant" to making users feel more comfortable.

Just as text-based designers have discovered the value of hiding infrequently used menu options, the best voice-interface systems should adapt to their users by detecting the most commonly used options of individuals and bringing them to the front of the menus. Although highly experienced users can simply barge in before any of the options are stated, users who want a quick reminder will greatly appreciate that options are thoughtfully presented. The user should not be constantly reminded that the system is recording their responses (in order to adapt) because this would make the user feel uncomfortable.[79]

Clarity

Clarity, the final conversational maxim, is greatly facilitated when the designer conforms with the domain knowledge of users by establishing "common ground."[80] This goes beyond the user's knowledge of how to use the interface to include the frequently more important question of how the system should provide answers. For example, consider a voice interface that allows customers to configure and buy computers. For a novice user, stating that "This will make graphics appear more rapidly on your screen" is probably the right level of complexity. Conversely, this is not the right level of clarity for an expert user, who might wonder, "Is this happening on the motherboard, the graphics card, or the accelerator board? Is it leveraging RAM or processing power?," and so on.

The principle of clarity states that a speaker must adopt words and concepts that the listener can understand. Hence, both permitting users to specify their expertise and adapting to users by monitoring their use of the interface and the questions they ask can be very effective strategies to maintain clarity.

When questioning and monitoring are not possible, an understanding of a typical user's background can be sufficient. For example, a site that sells do-it-yourself items can assume that users have a greater knowledge of tools than can a site that sells prebuilt goods.

Conversation as Cooperation

Conversing is more than spewing words at one another in turn: it is an agreement between participants to work toward mutual understanding. Partners in conversation must make a concerted effort to speak understandably, to listen with attention and sensitivity,[81] and to respond appropriately to what the other conversant is saying and thinking.[82] Humans are not simply wired to be talkers and listeners; they are members of a cooperative species[83] that is built to integrate talking and listening into a smooth, joint interaction, much like a ballroom dance.[84]

When things go wrong, cooperation can fall apart, or partners can work together to overcome the problem as a team. Blaming your partner for mistakes—or even attempting to avoid the placement of blame altogether—can have disastrous consequences. An awareness of conversational principles can thus help systems to minimize errors and deliver the most effective responses to errors.

15 Conclusion: From *Listening to* and *Talking at* to *Speaking with*

Speaking is intrinsically social. Whether it comes from a person or a machine, speech activates a powerful and varied cognitive apparatus that is designed to express and recognize who a person is and what she or he is thinking and feeling. Although people have separate parts of the brain that are devoted to each social judgment and each aspect of speech production and understanding, people do not have separate parts of the brain for human speech and technology-generated speech. Even when voice interfaces exhibit all of the limitations associated with machines—including bizarre pronunciations, emotional ignorance, and chronic inconsistencies—they are not exempted from the social expectations that are activated by talking and listening. When technologies, regardless of quality, fail to conform to social norms, users experience confusion, frustration, and cognitive exhaustion and question the competence, utility, and enjoyability of the system. Socially inept interfaces suffer the same fate as socially inept individuals: they are ineffective, criticized, and shunned.[1]

Fortunately, careful design that is grounded in the social and cognitive sciences can dramatically improve a voice interface's social success. An understanding of the wiring of the human brain can enable voice interfaces to be as compelling, competent, and supportive as the most successful friends, teachers, and salespeople. For example, creating a feeling of similarity between the user and the interface through matching gender, personality, or accent fosters feelings of liking and trust, two fundamental ingredients for cooperation and success. At the same time, interfaces should present a consistent social "face" to users: voices that talk, behave, and look like they sound are much more desirable, intelligent, and comfortable interaction partners.

Sensitivity to the emotional aspects of social life is also critical for both people and speaking technologies, as emotion influences every perception and conclusion that the human brain makes. Although voice interfaces cannot truly experience emotion, a voice that appropriately matches the emotional state of the user and the emotional tone of what it says communicates empathy, support, and sincerity. Conversely, voice

interfaces that fail to express the appropriate emotion risk being labeled uncooperative, cognitively burdensome, or hypocritical.

The brain is wired to treat each voice in an interface as a separate individual, even when the person knows that the same technology is producing all of the voices. Thus, multiple voices enable designers to create a rich social landscape that can facilitate communication through the creation of community or overwhelm users through the complexity of too many interpersonal relationships.

Social rules also dictate that an interface should not be a "show off." This principle discourages the mix of nonhuman and human interface characteristics. It also suggests that interfaces do better when they speak uniformly in synthetic speech, even if the speech is of poor quality, rather than shifting back and forth between synthetic and higher-quality recorded speech.

Through thoughtful consideration of social factors, voice interfaces can even overcome situations that are highly stressful. The right microphone selection can limit speakers' feeling of being monitored, thereby enabling people to be more relaxed, creative, and disclosive. Similarly, when things go wrong, the strategic placement of blame can reassure users and improve system-user cooperation.

Successful voice interfaces require more than sophisticated algorithms and advanced hardware. They demand an appreciation of speech as a rich, mutually supportive, and inherently social interaction. When voice interfaces fully leverage how humans are wired for speech, users will not simply talk *at* and listen *to* computers, nor will computers simply talk *at* and listen *to* users. Instead, people and computers will cooperatively *speak with* one another.

Notes

Chapter 1

1. See, e.g., S. Pinker, *The language instinct* (New York: Morrow, 1994).

2. D. I. Slobin, *Psycholinguistics,* 2nd ed. (Glenview, Ill.: Scott, Foresman, 1979).

3. One can compare N. Chomsky, *Language and mind* (New York: Harcourt, Brace, & World, 1968), and Pinker, *The language instinct,* to C. Holden, How the brain understands music, *Science* 292 (2001): 623, and D. W. Massaro, *Perceiving talking faces: From speech perception to a behavioral principle* (Cambridge, Mass.: MIT Press, 1998).

4. Massaro, *Perceiving talking faces;* Slobin, *Psycholinguistics.*

5. Pinker, *The language instinct.*

6. Slobin, *Psycholinguistics.*

7. Y. S. Sininger and B. Cone-Wesson, Asymmetric cochlear processing mimics hemispheric specialization, *Science* 305, no. 5960 (2000): 1581.

8. C. Moon, P. Cooper, and W. P. Fifer, Two-day-olds prefer their native language, *Infant Behavior and Development* 16 (1993): 495–500. R. Naatanen, The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm), *Psychophysiology* 38 (2001): 1–21; Pinker, *The language instinct.*

9. M. Pena, A. Maki, D. Kovacic, G. Dehaene-Lambertz, H. Koizumi, F. Bouquet, and J. Mehler, Sounds and silence: An optical topography study of language recognition at birth, *Proceedings of the National Academy of Sciences,* 100 (2003): 11702–11705.

10. Slobin, *Psycholinguistics.*

11. Although the psychological literature frequently fails to distinguish between an "addressee" (a person to whom a comment is directed) and a "listener" (anyone who can hear the comment), there are important psychological differences between the two. H. H. Clark, *Using language* (New York: Cambridge University Press, 1996). Because the listeners who are discussed in this book are