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 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 nonhuman 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.