Chapter LIX
Computer Mediated Speech Technology:
Perceptions of Synthetic Speech and Attitudes Toward Disabled Users

John W. Mullennix
University of Pittsburgh, USA

Steven E. Stern
University of Pittsburgh, USA

ABSTRACT

A frequently overlooked form of CMC is computer synthesized speech (CSS). Although the first CSS systems were rather crude and unintelligible, newer systems are fairly intelligible and are widely used for a number of applications, most importantly as aids for the speaking or visually disabled. In this chapter, we briefly review the development of CSS technology and discuss the work on perception and comprehension of CSS. Then, we examine how CSS use influences interactions between disabled people and nondisabled people. We conclude by emphasizing that the development of CSS systems should take into account various social psychological factors rooted in prejudice and stigma of the disabled.

INTRODUCTION

While most of the social and behavioral literature on Computer Mediated Communication (CMC) focuses on Internet-based and related applications of technology, one overlooked technology is that of computer-synthesized speech (CSS[1]). CSS is an important element of CMC as it is a frequently used computer-based interface. CSS is used for a variety of applications, such as talking computer terminals, training devices, warning and alarm systems and information databases (Syrdal, 1995). Most importantly, CSS is a valuable assistive technology for the speaking-disabled and the visually-impaired. Other forms of CMC such as the Internet are made more accessible to disabled people through the use of CSS.

CSS Systems: An Overview

The development of artificial voices and mechanical aids for speaking has been a topic of interest

for a long time, dating back to at least the Renaissance period (Flanagan, 1972). With steady advancements in electronics and acoustic models of human voice and speech production, in the 1930s and 1940s research on synthesized speech from machine first began to take off. One of the first electrical-based speech synthesizers was the Voder (Voice Operation Demonstrator) developed by Dudley and colleagues (Dudley, 1940; Dudley, Riesz, & Watkins, 1939). This analog-based device consisted of valve and gas discharge tube sound sources routed through a network of electrical bandpass filters and potentiometers that produced somewhat intelligible speech. The system was unwieldy and had to be "played" by a trained operator, with the operator manipulating hand and foot controls to produce varying types of speech utterances. Its later cousin the Vocoder (Voice Operated Recorder) combined speech analysis with speech output (Dudley, 1939).

In the 1950s, further progress in the acoustic analysis of speech resulted in a device developed at Haskins Laboratories called the Pattern Playback synthesizer (Cooper, Liberman, & Borst, 1951). The Pattern Playback was a device that took stylized hand-painted spectrograms (i.e., voiceprints) of speech, "read" them, and produced short intelligible utterances. Later developments updated the internal workings of this device (i.e., the "Voback," Borst & Cooper, 1957), but ultimately this invention was abandoned in lieu of later analog and digital formant synthesizers that more closely resemble the systems available today.

Currently, the technologies used to produce intelligible speech by machine are based on software and hardware systems on computer. There are two primary methods for producing computer-synthesized speech (CSS): Formant synthesis by rule and concatenation of stored speech segments (see Olive, 1997; O'Shaughnessy, 1995; Syrdal, 1995 for specific explanations). These synthesis methods are derived from research focusing on the acoustic-phonetics of speech as produced by the human vocal tract. The speech synthesis is usually performed via software programs that take advantage of the sound-related hardware found in personal computers. These programs can be bundled into application products that have CSS as their output. In general, most CSS systems currently available allow the user to select among different male and female voices and to modify different aspects of the output such as loudness and speaking rate.

One specific application where CSS is found is in text-to-speech conversion systems (see Klatt, 1987; Syrdal, 1995). In these systems, the user types text into an application and the software converts the text into CSS output. Perhaps where text-to-speech conversion has the most impact is as a communication aid for the speech disabled, as illustrated by the text-to-speech CSS system used by the famous astrophysicist Stephen Hawking. These systems allow the speech disabled to communicate with others in either short utterances or in long spoken passages.

Perception and Comprehension of CSS

When thinking about the practical applications of CSS, one must first consider whether the synthesized speech is intelligible, natural-sounding, and comprehensible. To this end, over the years a number of researchers have conducted experiments to evaluate CSS in terms of whether the speech information conveyed is understandable to the average listener (see Koul, 2003, for a review). The initial studies on the evaluation of these systems focused on the intelligibility of individual sound segments (such as consonants and vowels) and how perception of speech sounds output from these systems compared to human speech (see Fucci, Reynolds, Bettagere, & Gonzales, 1995; Hoover, Reichle, VanTassell, & Cole, 1987; Kangas & Allen, 1990; Koul & Allen, 1993; Logan, Greene, & Pisoni, 1989; Mirenda & Beukelman, 1987; 1990; Pisoni, Nusbaum, & Greene, 1985; Venkatagiri, 1991).