Text-to-Speech Synthesis

Mahbubur R. Syed
Minnesota State University, USA

Shuvro Chakrobartty
Minnesota State University, USA

Robert J. Bignall
Monash University, Australia

INTRODUCTION TO SPEECH SYNTHESIS

Speech synthesis is the process of producing natural-sounding, highly intelligible synthetic speech simulated by a machine in such a way that it sounds as if it was produced by a human vocal system. A text-to-speech (TTS) synthesis system is a computer-based system where the input is text and the output is a simulated vocalization of that text. Before the 1970s, most speech synthesis was achieved with hardware, but this was costly and it proved impossible to properly simulate natural speech production. Since the 1970s, the use of computers has made the practical application of speech synthesis more feasible.

In principle, a TTS system is a two-step process (Figure 1) in which text is converted to its equivalent digital audio. A text and linguistic analysis module processes the input text to generate its phonetic equivalent and performs linguistic analysis to determine the prosodic characteristics of the text. A waveform generator then produces the synthesized speech (Carvalho, Trancoso, & Oliveira, 1998).

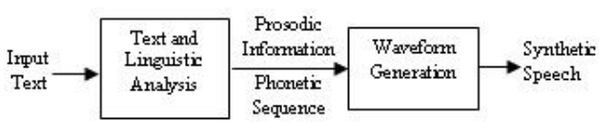
In the following sections, speech synthesis and its applications are described. Its historical development is outlined and the key challenges encountered by developers are summarized. A brief description is provided of some TTS synthesizers available at present.

APPLICATION OF SPEECH SYNTHESIS

Text-to-speech synthesis has potential applications in any domain in which speech is necessary or enhances communication with users. It has applications in education, telecommunications, consumer products, and a range of other areas. Imagine a TTS engine built into presentation software that not only shows a series of slides but also reads them to the audience. An important use of TTS synthesis is for the disabled, for example, to read an e-book or e-mail to a visually impaired person, or to read text typed by a deaf or vocally handicapped person. TTS synthesis becomes significant and useful if the audio information that needs to be communicated is too extensive to be stored, too expensive to prerecord, or if its recording is impossible because the application does not know ahead of time what output might be needed. For example, in a telecommunication application, a back-end TTS system may be able to read credit card information in real time over the phone to a customer. In such situations, TTS synthesis could offer a significant advantage over prerecorded audio sound if the information to be communicated is searched from a large database and cannot be anticipated in advance, or if the prerecording of audio information is not feasible.

Benefits of text-to-speech include reading dynamic text, conserving storage space, providing audible feedback, notifying a user of an event, proofreading documents, and so on. The usefulness of TTS synthesis in different application domains has at-

Figure 1. A two-step representation of a text-to-speech system



tracted researchers to take up the challenge of developing natural sounding and intelligible TTS systems in many languages around the world (Carvalho et al., 1998; Möbius, 1999; Mukherjee, Rajput, Subramaniam, & Verma, 2000; Syed, Chakrabartty, & Bignall, 2004).

A BRIEF HISTORY OF SPEECH SYNTHESIS

Speech synthesis has gone through a long development process from the mechanical to the electrical to the electronic era. According to Schafer and Markel, one of the first electrical synthesizers was the Voder (Voice Operation Demonstrator), which attempted to produce connected speech and followed the principles of source-tract separation. The electrical networks in the device could be selected by finger-actuated keys, whose resonances were similar to those of individual spoken sounds. This speaking machine was demonstrated by trained operators at the World Fairs of 1939 (New York) and 1940 (San Francisco). To produce speech, the operators could play the device as if it was an organ or a piano, but it required that they undergo a year or so of training (Schafer & Markel, 1979).

"Before the 1980s, speech synthesis research was limited to large laboratories that could afford to invest the necessary time and money for hardware. In the mid-1980s, more laboratories and universities started to join in as the cost of the hardware fell. By the late '80s, purely software-based synthesizers that not only produced reasonable quality speech but could do so in near real time became feasible" (Black & Lenzo, 2003). Up to the present time, several software companies and research groups have developed a variety of mono- to multilingual TTS systems on a range of platforms. Some are described in a later section. A historical archive of audio clips produced by different TTS systems can be found at http://www.cs.indiana.edu/rhythmsp/ASA/Contents.html.

METHODS OF SPEECH SYNTHESIS

- **Formant Synthesis:** "Formant synthesis is a source filter method of speech production in which the vocal tract filter is constructed from a number of resonances similar to the formants of natural speech" (Donovan, 1996).
Formant synthesis uses resonators and filters to emulate the human vocal system. Formant synthesis provides an infinite number of sounds, enabling maximum flexibility in voice customization and it requires three formants to produce intelligible speech and up to five formants to produce high-quality speech. Each formant is usually modeled with a two-pole resonator that enables both the pole-pair formant frequency and its bandwidth to be specified (Donovan, 1996). Rule-based formant synthesis uses a set of rules to determine the parameters necessary to synthesize a desired utterance using a formant synthesizer (Allen, Hunnicutt, Klatt, Armstrong, & Pisoni, 1987). Typically, a fundamental voicing frequency (F0), three other formant frequencies (F1, F2, F3) and three formant amplitudes (A1, A2, A3) are used in this synthesis (Lemmetty, 1999).
- **Concatenative Synthesis:** Concatenative synthesis is done by connecting prerecorded natural utterances to produce intelligible and natural-sounding, arbitrary synthetic speech. The natural speech segments are selected and stored in an acoustic inventory. According to Lemmety, the selection of the optimal speech-unit length is one of the most important decisions, requiring a trade-off between longer and shorter units. High naturalness, fewer concatenation points, and good control of coarticulation are achieved with longer units, but the number of required units and the amount of memory needed are increased. Less memory is needed with shorter units, but sample collecting and labeling procedures become more difficult and complex (Lemmetty, 1999). The prerecorded elements of natural utterances cannot be whole words because of their large number, different forms, and constant new additions to the language. "In 1958, Wang and Peterson proposed the 'diphone' (the acoustic chunk from the middle of one phoneme to the middle of the next phoneme) since coarticulatory influences tend to be minimal at the acoustic center of a phoneme. With this more satisfactory unit, they estimated that as many as 8,000 diphones may