Speech Synthesis and Recognition: Development and its Future

Kerry Robinson

ENGR 005: Engineering Methodologies

Eric Cheever

17 December 2012

 After about a decade of stagnation, the technology of speech synthesis, or the ability of a machine to produce artificial human speech, has once again ascended into the public eye. As technology shares an increasingly closer relationship with humans, speech synthesis has become a vital means of communication between human and machine. From its beginnings as a parlor trick and possible reading aid for the blind, the usability of speech synthesis has expanded to customer service, reading and writing aid, educational tools for children, and general assistance. This paper will focus on the history, the process, and the future of speech synthesis; however, since it is difficult to mention one without the other, speech recognition, or the ability of a machine to identify words and translate them into a readable format, will also be featured.

 The history of speech synthesis begins in 1791 with the invention of the manual speaking machine by Wolfgang von Kempelen. While interest in such a machine has persisted long before this time and other models had been invented, von Kempelen’s was the first to produce phrases and sentences. A diagram and a detailed description are included below:



 “The machine consisted of a bellows that simulated the lungs and was to be operated with the right forearm (rightmost drawing). A counterweight provided for inhalation. The middle and lower drawings show the 'wind box' that was provided with some levers to be actuated with the fingers of the right hand, the 'mouth', made of rubber, and the 'nose' of the machine. The two nostrils had to be covered with two fingers unless a nasal was to be produced. The whole speech production mechanism was enclosed in a box with holes for the hands and additional holes in its cover…”(Traunmüller).

 By the early 20th century, technology had advanced to the point the producing artificial speech by electrical means became a reality. The first example of this is Homer Dudley’s VODER (Voice Operating Demonstrator), which debuted at the 1939 World Fair in New York. Similar to von Kempelen, Dudley took into account factors such as resonance and the vocal cords. Dudley went further by setting up a keyboard system to select the resonance and frequency. A wrist bar selected the voice source and the source was routed through filters controlled by the keyboard to determine the frequency (Lemmetty). An illustration of the process is below:



 The VODER was mostly for entertainment purposes. In addition, the voice quality was rather poor and the machine required months of training to master. Nevertheless, the VODER sparked interest and proved that speech synthesis had potential use. In fact, the basic structure of VODER and the filter-source model of speech were heavily used in later incarnations, as well as in present day models. Other electric models in the 1950s and 1960s also used a filter to pass the electric source along and emulated the resonance of the vocal tract. Two main methods of articulation include transmission line analog, which uses numerous circuits the represent sections of the vocal tract, and terminal analog, which unlike the former simulates the vocals regardless of the shape of the circuits (Traunmüller).

 From the 1970s onward, automatic text-to-speech systems effectively commercialized speech synthesis. The first such system was created by Noriko Umeda and his team in Japan in 1968 (Lemmetty). Furthermore, in text-to-speech systems, “the simplistic idea of concatenating stored words or various shorter segments in order to produce speech was also realized. However, single speech sounds (phones) can not be successfully concatenated into words and sentences, since the acoustic properties of these minimal distinctive segments of speech vary as a function of their context, and this variation is necessary for intelligibility and naturalness. Better success has been achieved with so called "diphones", which consist of the second half of one speech sound and the first half of the subsequent. This results in a large number of elements, which have to be carefully selected” (Traunmüller). Early attempts to incorporate this technology into consumer products were rather cheaply manufactured and the voice quality suffered as a result (Lemmetty). However, speech synthesis has mainly kept up with advancements in computers. Even though the method of speech synthesis has remained generally the same since the 1980s, the quality of the voice and the accuracy of the pronunciation have continued to improve.

 Around the time automatic speech-to-text became commercialized, speech recognition had also dramatically improved; “Over the next decade, thanks to new approaches to understanding what people say, speech recognition vocabulary jumped from about a few hundred words to several thousand words, and had the potential to recognize an unlimited number of words” (Pinola). This change was due mostly in part to the transition from synthesis-by-rule to the hidden Markov model. Despite the flexibility synthesis-by-rule offered for speech synthesis, it considerably restricted speech recognition to several hundred words. “The hidden Markov model [HMM], which is a doubly stochastic process, models the intrinsic variability of the speech signal (and the resulting spectral features) as well as the structure of spoken language in an integrated and consistent statistical modeling framework... The resulting model is then used to provide an indication of the likelihood (probability) that an unknown utterance is indeed a realization of the word (or words) represented by the model” (Juang, and Rabiner). Indeed, HHM has been the standard for speech recognition for decades; this description from a 2004 Microsoft speech program demonstrates how closely present technology resembles past methods:

“TTS engines are generally based on either a formant or concatenative approach. Formant systems generate audio from text using an entirely algorithmic design to create the speech. Concatenative systems draw from a database of speech recorded from a voice talent. Sound segments are then joined (concatenated), to form words…functions TTS engines perform in preparing streaming text for output include: text normalization (identifying words or symbols appropriate to a phrase or a sentence), linguistic analysis (determining the overall flow of control using phrase breaks), and prosody generation (identifying the relative word stress levels and pitch accents within each phrase)” (Staff).

In addition to a concatenated system, current models also feature neural networking, or ‘outsourcing’ the process of speech recognition. An example includes Apple’s Siri (2010), which performs speech recognition on an external cloud rather than in the device itself. In terms of speech synthesis however, Siri and other current examples tend to use the same process as used by their decades-old counterparts.

While it is likely that the general process of speech synthesis will not change unless to keep pace with current technologies, manipulating the voice created by speech synthesis has gained importance. The emphasis is now on emotion, personality, and other uniquely human attributes; “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 to everyone” (Nass and Brave 40). Because text-to-speech systems produce sounds instantaneously and can immediately conform to settings of volume, pitch, pitch range, and speech rate, it is very inexpensive to make people feel that the system is “just like them” and hence worthy of affection, trust, and a range of other attributes (Nass and Brave 40). Meanwhile, future speech recognition development is focused on the ability of machines to accept everyday speech, seamless expansion into other languages, developing a more cognitive-based model of speech recognition, and real comprehension of spoken language (Baker, Li, Khudanpur, and et al). Therefore, speech synthesis, as well as speech recognition, are advancing with the rest of technology and its products are becoming integral aspects of the relationship between human and computer.

Works Cited

Baker, Janet, Deng Li, Sanjeev Khudanpur, and et al.*Historical Development and Future Directions in Speech Recognition and Understanding*. National Institute of Standards & Technology, n.d. Web. 14 Dec 2012. http://www.itl.nist.gov/iaui/894.02/MINDS/FINAL/speech.web.pdf.

Juang, B.H., and Lawrence R. Rabiner. *Automatic Speech Recognition – A Brief History of the Technology Development*. University of California at Santa Barbara, 08 2004. Web. 14 Dec 2012. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354\_LALI-ASRHistory-final-10-8.pdf.

Lemmetty, Sami. *History and Development of Speech Synthesis*. Helsinki University Of Technology, 06 1999. Web. 14 Dec 2012. http://www.acoustics.hut.fi/publications/files/theses/lemmetty\_mst/chap2.html.

Nass, Clifford, and Scott Brave. *Wired for Speech: How Voice Activiates and Advances the Human-Computer Relationship*. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2005. 40. Print.

Pinola, Melanie. *Speech Recognition Through the Decades: How We Ended Up With Siri*. PCWorld, 02 2011. Web. 14 Dec 2012. http://www.pcworld.com/article/243060/speech\_recognition\_through\_the\_decades\_how\_we\_ended\_up\_with\_siri.html.

Staff, *Text-to-Speech and the Microsoft Speech Technologies Platform*. Microsoft Developer Network, n.d. Web. 14 Dec 2012. http://msdn.microsoft.com/en-us/library/ms994644.aspx.

Traunmüller, Hartmut. "Wolfgang von Kempelen's Speaking Machine and its Successors." . Stockholms universitet, n.d. Web. 14 Dec 2012. http://www2.ling.su.se/staff/hartmut/kemplne.htm.