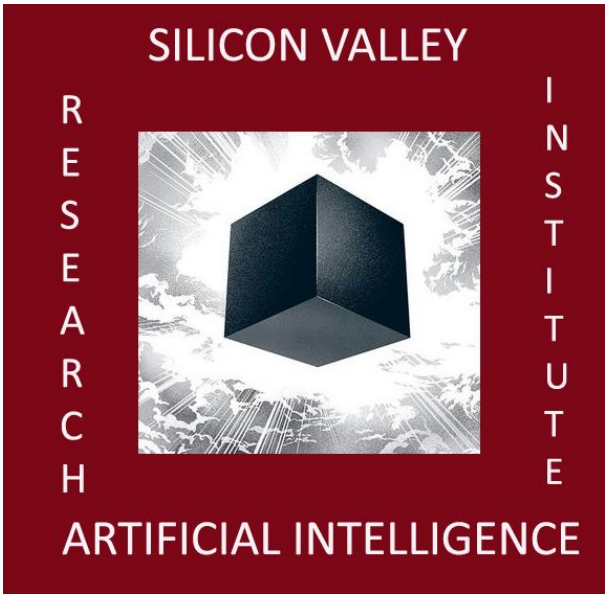

Speech Synthesis: a brief introduction

Speech synthesis predates A.I. by at least two decades. The vocoder (from Voice Operated reCOrDER), capable of synthesizing a human voice, was invented in 1940 by Homer Dudley at Bell Labs and used during the war to scramble the phone conversations between British prime minister Winston Churchill and US president Franklin Roosevelt. The first computer to sing was an IBM 7094 programmed in 1961 at Bell Labs by computer-music pioneer Max Mathews and neuropsychologist Louis Gerstman. It sang Harry Dacre's pop song "Daisy Bell" of 1892. This artificial song was later included in the soundtrack of Stanley Kubrick's "2001 A Space Odyssey" (1968). The vocoder became popular in electronic music after Robert Moog built one in 1968, most famously used to sing the "Ode to Joy" melody of Beethoven's "Ninth Symphony" in Stanley Kubrick's film "A Clockwork Orange" (1971), and later used on Kraftwerk's album "Autobahn" (1974) and Giorgio Moroder's album "From Here to Eternity" (1977). Meanwhile, the first full text-to-speech system had been built in Japan by Hitachi after being designed at Japan's Electrotechnical Laboratory by the psychologist Ryunen Teranishi and the linguist Noriko Umeda (first reported in ETL News #197, 1966). Then Umeda moved to Bell Labs to work on a "talking computer" with Cecil Coker. Their speech synthesizer was first reported in a New York Times article of 1972 ("Where Science Grows Miracles"). Trivia: Coker engineered early electronic music installations such as John Cage's "Variations V" (1965) and Robert Rauschenberg's "Linoleum" (1966). In 1979 Dennis Klatt at the MIT unveiled the MITalk system for text to speech conversion that was turned into a product by DEC in 1984 (DECtalk) and used to create the electronic voices in Robert Zemeckis' film "Back to the Future II" (1985). The era of concatenative speech synthesis began with Eric Moulines and Francis Charpentier at the National Center of Telecommunications in France ("Text-to-speech algorithms based on FFT synthesis", 1988), a technique popularized by Andrew Hunt and Alan Black at ATR Labs in Japan as part of the CHATR system ("Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", 1996). The generation of Apple's Siri, Microsoft's

Cortana, Amazon's Alexa and the Google Assistant all the way to Google's WaveNet used some kind of synthesis-by-concatenation method. The competing approach was statistical speech synthesis, based on hidden Markov models, for example the HTS system demonstrated in 1995 by Keiichi Tokuda and others at the Nagoya Institute of Technology in Japan ("Speech Synthesis from HMMs Using Dynamic Features," 1996), which in 2002 became a popular open-source toolkit.