His Master's Voice

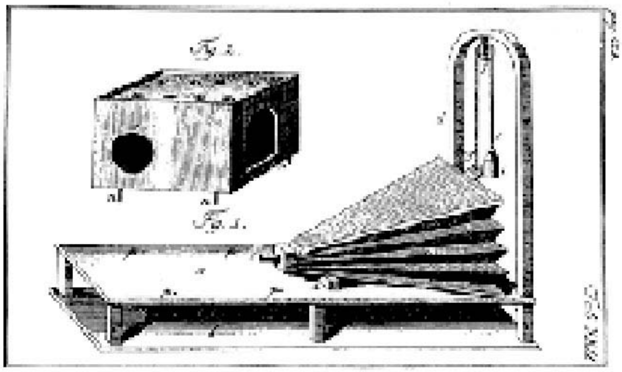
Manfred Kroboth
Mexikoring 15
D - 22297 Hamburg
0049 40 29888354
kroko@foni.net

It is hard to imagine what astonishment was caused, when the first time a human voice, produced by a machine, was heard from human beings. The picture of the dog listening to the funnel of a record player because it hears "his master's voice" coming out, provides from this astonishment a miraculous impression.

It is an early dream of mankind to create afficial life. And therefore you need not only a movable body; it must also be able to communicate, to speak.

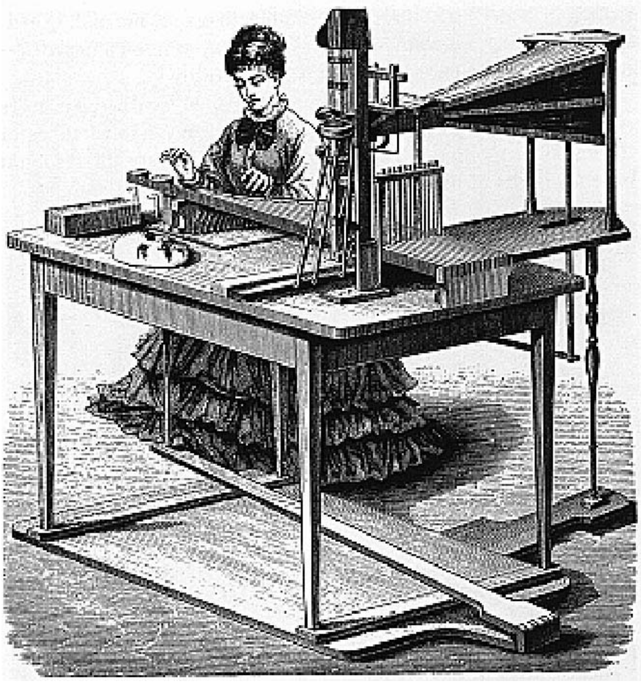
Long before the first apparatus that was able to record speech or sound in general was invaded, there were successfully attempts to produce artificial speech. Gerbert von Aurillac (1003) built a "speaking head" out of bronze that could say yes and no. Ch. G. Kratzenstein, professor of physiology in Copenhagen, succeeded in producing 1779 a device that spoke five long vowels (a, e, i, o and u). Wolfgang von Kempelen developed shortly after that a speaking machine. In his book "*Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine"* (1791) he included a detailed description - in order for others to reconstruct it and make it more perfect. Von Kempelen was an ingenious person in the service of empress Maria Theresa in Vienna. While he became known for various additional feats (i.e. the Chess-playing Turk), his main concern was the study of human speech production, with therapeutic applications in mind. He has been called the first experimental phonetician.

Von Kempelen's machine was the first that allowed producing not only some speech sounds, but also whole words and short sentences. According to von Kempelen, it is possible to acquire an admirable facility in playing the machine `within three weeks, especially if one chooses the Latin, French, or Italian language, since German is much more difficult because of its many closed syllables and consonant clusters´.



The final version of von Kempelen's machine is preserved to this day. It was kept at the k. k. Konservatorium für Musik in Vienna until 1906, when it was donated to the Deutsches Museum (von Meisterwerken der Naturwissenschaft und Technik) in Munich that had been founded three years before. There, it is exhibited in the department of musical instruments.

In the 19th century, some additional machines of similar kind were constructed, but there were no really fundamental innovations in the field of speech synthesis. Only a machine called "Euphonia", constructed by Joseph Faber in 1835, can be said to represent some progress in that its speech production mechanism included a model of the tongue and a pharyngeal cavity whose shape could be controlled. It was also suited for the synthesis of singing. Its bellows was operated via a pedal, and otherwise it was controlled via a key board.



At the beginning of the 20th century, the progress in electrical engineering made it possible to synthesize speech sounds by electrical means. The first device of this kind that attracted the attention of a wider public was the "VODER", developed by Homer Dudley in the Bell Labs and presented at the World Fair in New York in 1939.

The first computer based speech synthesis systems were developed in the late 1950th. The first complete text-to-speech-system was realized in 1968. The physicist John Larry Kelly, Jr created in the year 1961 in the Bell Labs a speech synthesis with an IBM 704 that could sing the song "Daisy Bell". The director Stanley Kubrick was so impressed that he used it in his film "2001 A Space Odyssey".

Fundamentally you can divide between two models of producing speech signals. The signal can completely created in the computer with the so-called physiological modelling or you use pre-recorded speech (samples) with signal modelling.

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.

With a system that uses diphones it is theoretically possible to sample the voice of a person (if there is enough recorded material), to produce a data bank and to generate the voice with a computer [1]. This possibility is theoretically because it would cause a incredible amount of work to produce in this way a spoken text, which sounds not only like its origin - but also sound really "human".

In the field of film, you have a long tradition of cartoon. Since a few years it is possible to produce films, which looks relatively realistic, but are completely computer generated. In all these productions real actors spoke the text.

The question is: why you should create images or speech should be produced artificially, if it is easier (or cheaper) to produce it in a traditional way. From an artistic point of view, it makes only sense, it the images or the speech is not to produce with conventional means. Concerning the speech it means that synthesised speech makes artistically only sense, if it produces something acoustically, which is impossible to produce with a human voice. As an impulse for the reader, try to read this:
rznmklwsdf gnytrpwzfsjkbdrwxtgzmfvpwrzlghtnwscgkfmt gnsmzifthwirrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrtzmst

m2k2 (a copy of the author)

[1] I made this, in cooperation with the TCTS Lab of the Faculté Polytechnique de Mons (Belgium), with my one voice. The software and the database are free to use.
http://tcts.fpms.ac.be/synthesis/mbrola/mbrcopybin.html