History of Speech Synthesis in Phonetics Research

Brad H. Story
Department of Speech, Language, and Hearing Sciences
University of Arizona
Tucson, AZ

Citation: An invited chapter for *The Routledge Handbook of Phonetics*, pp., W. Katz and P. Assmann, Eds., Routledge (2018).

1 Introduction

For the past two centuries or more, a variety of devices capable of generating artificial or synthetic speech have been developed and used to investigate phonetic phenomena. The aim of this chapter is to provide a brief history of synthetic speech systems, including mechanical, electrical, and digital types. The primary goal, however, is not to reiterate the details of constructing specific synthesizers but rather to focus on the motivations for developing various synthesis paradigms and illustrate how they have facilitated research in phonetics.

2 The mechanical and electro-mechanical era

On the morning of December 20, 1845, in Philadelphia, Pennsylvania, a prominent American scientist attended a private exhibition of what he would later refer to as a "wonderful invention." The scientist was Joseph Henry, an expert on electromagnetic induction and the first Secretary of the Smithsonian Institution. The "wonderful invention" was a *machine that could talk*, meticulously crafted by a disheveled sixty-year-old tinkerer from Freiburg, Germany, named Joseph Faber. Their unlikely meeting, which was encouraged by an acquaintance of Henry from the American Philosophical Society, might have occurred more than a year earlier had Faber not destroyed a previous version of his talking machine in a bout of depression and intoxication. Although he spent some twenty years perfecting the first device, Faber was able to reconstruct a second version of equal quality in a year's time (Patterson, 1845).

The layout of the talking machine, described in a letter from Henry to his colleague H. M. Alexander, was like that of a small chamber organ whose keyboard was connected via strings and levers to mechanical constructions of the speech organs. A carved wooden face was fitted with a hinged jaw, and behind it was

an ivory tongue that was moveable enough to modulate the shape of the cavity in which it was housed. A foot-operated bellows supplied air to a rubber glottis whose vibration provided the raw sound that could be shaped into speech by pressing various sequences or combinations of 16 keys available on a keyboard. Each key was marked with a symbol representing an "elementary" sound which, through their linkages to the artificial organs, imposed time-varying changes to the air cavity appropriate for generating apparently convincing renditions of connected speech. Several years earlier Henry had been shown a talking machine built by the English scientist Charles Wheatstone, but he noted that Faber's machine was far superior because instead of uttering just a few words, it was "capable of speaking whole sentences composed of any words what ever" (qtd. in Rothenberg et al. 1992, p. 362).

In the same letter, Henry mused about the possibility of placing two or more of Faber's talking machines at various locations and connecting them via telegraph lines. He thought that with "little contrivance" a spoken message could be coded as keystrokes in one location which, through electromagnetic means, would set into action another of the machines to "speak" the message to an audience at a distant location. Another thirty years would pass before Alexander Graham Bell demonstrated his invention of the telephone, yet Henry had already conceived of the notion while witnessing Faber's machine talk. Further, unlike Bell's telephone which transmitted an electrical analog of the speech pressure wave, Henry's description alluded to representing speech in *compressed* form based on slowly-varying movements of the operator's hands, fingers, and feet as they formed the keystroke sequences required to produce an utterance, a signal processing technique that would not be implemented into telephone transmission systems for nearly another century.

It is remarkable that, at this moment in history, a talking machine had been constructed that was capable of transforming a type of phonetic representation into a simulation of speech production, resulting in an acoustic output heard clearly as intelligible speech - and this same talking machine had also inspired the idea of electrical transmission of low-bandwidth speech. But the moment is also ironic considering that no one seized either as an opportunity for scientific or technological advancement. Henry understandably continued on with his own scientific pursuits, leaving his idea to one short paragraph in an obscure letter to a colleague. In need of funds, Faber signed on with the entertainment entrepreneur P. T. Barnum in 1846 to exhibit his talking machine for a several months run at the Egyptian Hall in London. In his autobiography, Barnum (1886) noted that a repeat visitor to the exhibition was the Duke of Wellington, who Faber eventually taught to speak both English and German phrases with the machine (Barnum, 1886, p.134). In the exhibitor's autograph book, the Duke wrote that Faber's "Automaton Speaker" was an

"extraordinary production of mechanical genius." Other observers also noted the ingenuity in the design of the talking machine (e.g., "The Speaking Automaton", 1846; Athenaeum, 1846), but to Barnum's puzzlement it was not successful in drawing public interest or revenue. Faber and his machine were eventually relegated to a traveling exhibit that toured the villages and towns of the English countryside; it was supposedly here that Faber ended his life by suicide, although there is no definitive account of the circumstances of his death (Altick, 1978). In any case, Joseph Faber disappeared from the public record, although his talking machine continued to make sideshow-like appearances in Europe and North America over the next thirty years; it seems a relative (perhaps a niece or nephew) may have inherited the machine and performed with it to generate an income ("Talking Machine", 1880; Altick, 1978).

Although the talking machine caught the serious attention of those who understood the significance of such a device, the overall muted interest may have been related to Faber's lack of showmanship, the German accent that was present in the machine's speech regardless of the language spoken, and perhaps the fact that Faber never published any written account of how the machine was designed or built. Or - maybe a mechanical talking machine, however ingenious its construction, was, by 1846, simply considered passé. Decades earlier, others had already developed talking machines that had impressed both scientists and the public. Most notable were Christian Gottlieb Kratzenstein and Wolfgang von Kempelen, who had independently developed mechanical speaking devices in the late 18th century.

Inspired by a competition sponsored by the Imperial Academy of Sciences at St. Petersburg in 1780, Kratzenstein submitted a report that detailed the design of five organ pipe-like resonators that, when excited with the vibration of a reed, produced the vowels /a,e,i,o,u/ (Kratzenstein, 1781). Although their shape bore little resemblance to human vocal tract configurations, and they could produce only sustained sounds, the construction of these resonators won the prize and marked a shift toward scientific investigation of human sound production. Kratzenstein, who at the time was a professor of physics at the University of Copenhagen, had shared a long-term interest in studying the physical nature of speaking with his former colleague at St. Petersburg, Leonhard Euler who likely proposed the competition. Well known for his contributions to mathematics, physics, and engineering, Euler wrote in 1761 that "all the skill of man has not hitherto been capable of producing a piece of mechanism that could imitate [speech]..." and further noted that "The construction of a machine capable of expressing sounds, with all the articulations, would no doubt be a very important discovery." (Euler, 1761). He envisioned such a device to be used in assistance of those "whose voice is either too weak or disagreeable..." (Euler, 1761, p. 79).

During the same time period, von Kempelen, a Hungarian engineer, industrialist, and government official, used his spare time and mechanical skills to build a talking machine far more advanced than the five vowel resonators demonstrated by Kratzenstein. The final version of his machine was to some degree a mechanical simulation of human speech production. It included a bellows as a "respiratory" source of air pressure and air flow, a wooden "wind" box that emulated the trachea, a reed system to generate the voice source, and a rubber funnel that served as the vocal tract. There was an additional chamber used for nasal sounds, and other control levers that were needed for particular consonants. Although housed in a large box, the machine itself was small enough that it could have been easily held in the hands. Speech was produced by depressing the bellows which caused the "voice" reed to vibrate. The operator then manipulated the rubber vocal tract into time-varying configurations which, along with controlling other ports and levers, produced speech at the word level, but could not generate full sentences due to the limitations of air supply and perhaps the complexity of controlling the various parts of the machine with only two hands. The sound quality was child-like, presumably due to the high fundamental frequency of the reed and the relatively short rubber funnel serving as the vocal tract. In an historical analysis of von Kempelen's talking machine, Dudley and Tarnoczy (1950) note that this quality was probably deliberate because a child's voice was less likely to be criticized when demonstrating the function of the machine. Kempelen may have been particularly sensitive to criticism considering that he had earlier constructed and publicly demonstrated a chess-playing automaton that was in fact a hoax (cf., Carroll, 1975). Many observers initially assumed that his talking machine was merely a fake as well.

Kempelen's lasting contribution to phonetics is his prodigious written account of not only the design of his talking machine, but also the nature of speech and language in general (von Kempelen, 1791). Called "On the Mechanism of Human Speech" [English translation], the experiments he describes that consumed more than twenty years clearly showed the significance of using models of speech production and sound generation to study and analyze human speech. This work motivated much subsequent research on speech production, and to this day still guides the construction of replicas of his talking machine for pedagogical purposes (cf., Trouvain and Brackhane, 2011).

One person who was particularly inspired by von Kempelen's work was, in fact, Joseph Faber. According to a biographical sketch (Wurzbach, 1856), while recovering from a serious illness in about 1815 Faber happened onto a copy of "On the Mechanism of Human Speech" and became consumed with the idea of building a talking machine. Of course, he built not a replica of von Kempelen's machine, but one with a

significantly advanced system of controlling the mechanical simulation of speech production. As remarkable as Faber's machine seems to have been regarded by some observers, Faber was indeed late to the party, so to speak, for the science of voice and speech had by the early 1800's already shifted into the realm of physical acoustics. Robert Willis, a professor of mechanics at Cambridge University, was dismayed by both Kratzenstein's and von Kempelen's reliance on trial and error methods in building their talking machines, rather than acoustic theory. He took them to task, along with most others working in phonetics at the time, in his 1829 essay titled "On the Vowel Sounds, and on Reed Organ-Pipes." The essay begins:

> *"The generality of writers who have treated on the vowel sounds appear never to have looked beyond the vocal organs for their origin. Apparently assuming the actual forms of these organs to be essential to their production, they have contented themselves with describing with minute precision the relative positions of the tongue, palate and teeth peculiar to each vowel, or with giving accurate measurements of the corresponding separation of the lips, and of the tongue and uvula, considering vowels in fact more in the light of physiological functions of the human body than as a branch of acoustics."* -Willis, 1829, p. 231

Willis laid out a set of experiments in which he would investigate vowel production by deliberately neglecting the organs of speech. He built reed-driven organ pipes whose lengths could be increased or decreased with a telescopic mechanism, and then determined that an entire series of vowels could be generated with changes in tube length and reeds with different vibrational frequencies. Wheatstone (1837) later pointed out that Willis had essentially devised an acoustic system that, by altering tube length, and hence the frequencies of the tube resonances, allowed for selective enhancement of harmonic components of the vibrating reed. Wheatstone further noted that multiple resonances are exactly what is produced by the "cavity of the mouth," and so the same effect occurs during speech production but with a nonuniformly shaped tube.

Understanding speech as a pattern of spectral components became a major focus of acousticians studying speech communication for much of the 19th century and the very early part of the 20th century. As a result, developments of machines to produce speech sounds were also largely based on some form of spectral addition, with little or no reference to the human speech organs. For example, in 1859 the German scientist Hermann Helmholtz devised an electromagnetic system for maintaining the vibration of a set of eight or more tuning forks, each variably coupled to a resonating chamber to control amplitude

(Helmholtz, 1859). With careful choice of frequencies and amplitude settings he demonstrated the artificial generation of five different vowels. Rudolph Koenig, a well-known acoustical instrument maker in 1800's, improved on Helmholtz's design and produced commercial versions that were sold to interested clients (Pantalony, 2004). Koenig was also a key figure in emerging technology that allowed for recording and visualization of sound waves. His invention of the phonoautograph with Edouard-Léon Scott in 1859 transformed sound via a receiving cone, diaphragm and stylus into a pressure waveform etched on smoked paper rotating about a cylinder. A few years later he introduced an alternative instrument in which a flame would flicker in response to a sound, and the movements of flame were captured on a rotating mirror, again producing a visualization of the sound as a waveform (Koenig, 1873).

These approaches were precursors to a device called the "phonodeik" that would be later developed at the Case School of Applied Science by Dayton Miller (1909) who eventually used it to study waveforms of sounds produced by musical instruments and human vowels. In a publication documenting several lectures given at the Lowell Institute in 1914, Miller (1916) describes both the analysis of sound based on photographic representations of waveforms produced by the phonodeik, as well as intricate machines that could generate complex waveforms by adding together sinusoidal components and display the final product graphically so that it might be compared to those waveforms captured with the phonodeik. Miller referred to this latter process as harmonic synthesis, a term commonly used to refer to building complex waveforms from basic sinusoidal elements. It is, however, the first instance of the word *synthesis* in the present chapter. This was deliberate to remain true to the original references. Nowhere in the literature on Kratzenstein, von Kempelen, Wheatstone, Faber, Willis, or Helmholtz does "synthesis" or "speech synthesis" appear. Their devices were variously referred to as talking machines, automatons, or simply systems that generated artificial speech. Miller's use of *synthesis* in relation to human vowels seems to have had the effect of labeling any future system that produces artificial speech, regardless of the theory on which it is based, a *speech synthesizer*.

Interestingly, the waveform synthesis described by Miller was not actually synthesis of sound, but rather synthesis of graphical representations of waveforms. To produce synthetic sounds, Miller utilized a bank of organ pipes, each of which, by design, possessed a different set of resonant frequencies. By controlling the amplitude of the sound produced by each pipe, he could effectively produce a set of nearly pure tones that were summed together as they radiated into free space. The composite waveform could then be captured with the phonodeik device and compared to the graphical synthesis of the same vowel. These were primarily vowel synthesizers, where production of each vowel required a different collection of

pipes. There was little ability to dynamically change any aspect of the system except for interrupting the excitation of the pipes themselves; Miller did suggest such an approach to forming some basic words.

At this point in time, about a decade and a half into the 20th century, the mechanical and electro-mechanical era of speech synthesis was coming to a close. The elaborate talking machines of von Kempelen and Faber that simulated human speech production were distant memories, having been more recently replaced by studies of vowels using electro-mechanical devices that produced the spectral components of speech waveforms. Although there was much debate and disagreement about many details on the production of speech, primarily vowels, the ideas generated in this era were fundamental to the development of phonetics. It had become firmly established by now (but not universally accepted) that the underlying acoustic principle of speech production was that resonances formed by a given configuration of an air cavity enhanced or accentuated the spectral components of a sound source (Rayleigh, 1878). The enhanced portions of the spectrum eventually came to be known as "formants," a term that seems to have been first used by Ludimar Hermann in his studies of vowel production using phonograph technology (Hermann, 1894, 1895). Thus, the stage had been set to usher in the next era of speech synthesis.

3 The electrical and electronic era

A shift from using mechanical and electro-mechanical devices to generate artificial speech to purely electrical systems had its beginnings in 1922. It was then that John Q. Stewart, a young physicist from Princeton published an article in the journal *Nature* titled "An Electrical Analogue of the Vocal Organs" (Stewart, 1922). After military service in World War I during which he was the chief instructor of "sound ranging" at the Army Engineering School, Stewart had spent two years as research engineer in the laboratories of the American Telephone and Telegraph Company and the Western Electric Company (Princeton Library). His article was a report of research he had completed during that time. In it he presents a diagram of a simple electrical circuit containing an "interrupter" or buzzer and two resonant branches comprised of variable resistors, capacitors, and inductors. Noting past research of Helmholtz, Miller, and Scripture, Stewart commented that "it seems hitherto to have been overlooked that a functional copy of the vocal organs can be devised... [with] audio-frequency oscillations in electrical circuits." He demonstrated that a wide range of artificial vowels could be generated by adjusting the circuit elements in the resonant branches. Because of the ease and speed with which these adjustments could be made (e.g., turning knobs, moving sliders, etc), Stewart also reported success in generating diphthongs by rapidly shifting the resonance frequencies from one vowel to another. Although the title of

the article suggests otherwise, the circuit was not really an electrical analog of the vocal organs, but rather a means of emulating the acoustic resonances they produced. The design was essentially the first electrical formant synthesizer; interestingly, however, Stewart did not refer to his system as a *synthesizer*, but rather as an electrical analogue of the vocal system.

Stewart moved on to a long productive career at Princeton as an astrophysicist and did not further develop his speech synthesizer. He did, however, leave an insightful statement at the end of his article that foreshadowed the bane of developing artificial speech systems for decades to come, and still holds today. He noted that:

> "*The really difficult problem involved in the artificial production of speech sounds is not the making of the device which shall produce sounds which, in their fundamental physical basis, resemble those of speech, but in the manipulation of the apparatus to imitate the manifold variations in tone which are so important in securing naturalness.*" -Stewart, 1922, p. 312

Perhaps by "naturalness" it can be assumed he was referring to the goal of achieving natural human sound quality as well as intelligibility. In any case, he was clearly aware of the need to establish "rules" for constructing speech, and that simply building a device with the appropriate physical characteristics would not in itself advance artificial speech as a useful technology or tool for research.

A few years later, in 1928, a communications engineer named Homer Dudley, also working at the Western Electric Company (later to become Bell Telephone Laboratories), envisioned a system that could be used to transmit speech across the transatlantic *telegraph* cable (Schroeder, 1981). Because it was designed for telegraph signals, however, the cable had a limited bandwidth of only 100 Hz. In contrast, transmission of the spectral content of speech requires a minimum bandwidth of about 3000 Hz, and so the telegraph cable was clearly insufficient for carrying an electrical analog of the speech waveform. The bandwidth limitation, however, motivated Dudley to view speech production and radio transmission analogously. Just as the information content carried by a radio signal is embedded in the relatively slow modulation of a carrier wave, phonetic information produced by movements of the lips, tongue, jaw, and velum, could be considered to similarly modulate the sound wave produced by the voice source. That is, the speech articulators move at inaudible syllabic rates which are well below the 100 Hz bandwidth of the telegraph cable, whereas the voice source or carrier makes the signal audible but also creates the need for the much larger bandwidth. Understanding the difficulties of tracking actual articulatory movements,

Dudley instead designed a circuit that could extract low frequency spectral information from an acoustic speech signal via a bank of filters, transmit that information along the low-bandwidth cable, and use it to modulate a locally-supplied carrier signal on the receiving end to reconstruct the speech. This was the first *analysis-synthesis* system in which some set of parameters determined by analysis of the original signal could be sent to another location, or perhaps stored for later retrieval, and used to synthesize a new version of the original speech. Dudley had achieved almost exactly that which Joseph Henry had imagined in that letter he wrote long ago about linking together several of Faber's talking machines to communicate across distance.

Dudley's invention became known as the VOCODER, an acronym derived from the two words *VOice CODER* (to avoid the repetition of capital letters and to reflect its addition to our lexicon, "Vocoder" will be used in the remainder of the chapter). The Vocoder was demonstrated publicly for the first time on September 11, 1936 at the Harvard Tercentary Conference in Cambridge, Massachusetts (Dudley, 1936). During an address given by F. B. Jewitt, President of Bell Telephone Laboratories, Dudley was called on to demonstrate the Vocoder to the audience (Jewitt, 1936) and showed its capabilities for analysis and subsequent synthesis of speech and singing. Dudley could also already see the potential of using the Vocoder for entertainment purposes (Dudley, 1939a). He noted that once the low frequency spectral modulation envelopes had been obtained from speech or song, any signal with sufficiently wide bandwidth could be substituted as the carrier in the synthesis stage. For example, instrumental music or the sound of a train locomotive could be modulated with the spectral-phonetic information present in a sentence, producing a bizarre but entirely intelligible synthetic version of the original speech utterance (Dudley, 1940). Ironically, due to the international political events of the late 1930's and early 1940's, the first major application of the Vocoder was not to amuse audiences, but rather to provide secure, scrambled speech signals between government and military officials during World War II, particularly the conversations of Winston Churchill in London and Franklin D. Roosevelt in Washington, D.C.

One of the difficulties that prevented wide acceptance of Vocoder technology for general telephone transmission was the problem of accurately extracting pitch (fundamental frequency) from an incoming speech signal (Schroeder, 1993). Transmitting pitch variations along with the other modulation envelopes was essential for reconstructing natural-sounding speech. It was not, however, necessary for transmitting *intelligible* speech, and hence could be acceptably used when the security of a conservation was more important than the naturalness of the sound quality. Even so, both Churchill and Roosevelt complained that the Vocoder made their speech sound silly (Tompkins, 2010), certainly an undesirable quality for

world leaders. Eventually the pitch extraction problem was solved, other aspects were improved, and Vocoder technology became a viable means of processing and compressing speech for telephone transmission.

With the capability of isolating various aspects of speech, Dudley also envisioned the Vocoder as a tool for research in phonetics and speech science. In 1939, he and colleagues wrote,

"After one believes he has a good understanding of the physical nature of speech, there comes the acid test of whether he understands the construction of speech well enough to fashion it from suitably chosen elements." - Dudley et al., 1939a, p. 740

Perhaps Dudley realized, much as Stewart (1922) had warned, that building a device to decompose a speech signal and reconstruct it synthetically was relatively "easy" in comparison to understanding how the fundamental elements of speech, whatever form they may take, can actually be generated sequentially by a physical representation of the speech production system, and result in natural, intelligible speech. With this goal in mind, he and colleagues modified the Vocoder such that the speech analysis stage was replaced with manual controls consisting of a keyboard, wrist bar, and foot pedal (Dudley, Riesz, and Watkins, 1939a). The foot pedal controlled the pitch of a relaxation oscillator that provided a periodic voice source to be used for the voiced components of speech; a random noise source supplied the "electrical turbulence" needed for the unvoiced speech sounds. Each of the ten primary keys controlled the amplitude of the periodic or noise-like sources within a specific frequency band, which together spanned a range from 0-7500 Hz. By depressing combinations of keys and modulating the foot pedal, an operator of the device could learn to generate speech.

This new synthetic speaker was called the "VODER" or "Voder," a new acronym that comprised the capitalized letters in "Voice Operation DEmonstratoR" (Dudley et al., 1939b). In a publication of the Bell Laboratories Record (1939), the machine's original moniker was "Pedro the Voder," where the first name was a nod to Dom Pedro II, a former Emperor of Brazil who famously exclaimed "My God, it talks!" after witnessing a demonstration of Bell's invention of the telephone in Philadelphia in 1876. The Bell publication ("Pedro the Voder", 1939) pointed out that the telephone did not actually talk, but rather transmitted talk over distance. In contrast, the Voder *did* talk and was demonstrated with some fanfare at the 1939 World's Fair in New York and at the Golden Gate Exposition in San Francisco the same year. It is interesting that this publication also states "It is the first machine in the world to do this [i.e., talk]"

(Bell Labs Pubs, 1939, p. 170). If this was a reference to synthetic speech produced by an *electronic* artificial talker, it is likely correct. But clearly Joseph Faber had achieved the same goal by mechanical means almost a century earlier. In fact, the description of the Voder on the same page as a "...little old-fashioned organ with a small keyboard and a pedal" could have easily been used to describe Faber's machine. In many ways, Dudley and colleagues at Bell Labs were cycling back through history with a new form of technology that would now allow for insights into the construction of speech that the machines of previous eras would not reveal to their makers.

One of the more interesting aspects of the Voder development, at least from the perspective of phonetics, was how people learned to speak with it. Stanley S. A. Watkins, the third author on the Dudley et al. (1939a) article describing the Voder design, was charged with prescribing a training program for a group of people who would become "operators." He first studied the ways in which speech sounds were characterized across the ten filter bands (or channels) of Voder. Although this was found to be useful information regarding speech, it was simply too complex to be useful in deriving a technique for talking with the Voder. Various other methods of training were attempted, including templates to guide the fingers and various visual indicators, but eventually it was determined that the most productive method was for the operator to search for a desired speech sound by "playing" with the controls as guided by their ear. Twenty-four people, drawn from telephone operator pools, were trained to operate the Voder for the exhibitions at both sites of the 1939 World's Fair. Typically, about one year was required to develop the ability to produce intelligible speech with it. In fact, Dudley et al. wrote "... the first half [of the year of training was] spent in acquiring the ability to form any and all sounds, the second half being devoted to improving naturalness and intelligibility" (Dudley et al., 1939b, p. 763). Once learned, the ability to "speak" with the Voder was apparently retained for years afterward, even without continued practice. On the occasion of Homer Dudley's retirement in 1961, one of the original trained operators was invited back to Bell Labs for an "encore performance" with a restored version of the talking machine. As recalled by James Flanagan, a Bell Labs engineer and speech scientist, "She sat down and gave a virtuoso performance on the Voder" (qtd. in Pieraccini, 2012, p. 55).

In his article "The carrier nature of speech," Dudley (1940) made a compelling analogy of the Voder structure to the human speech production system. But the Voder was really a spectrum shaping synthesizer; the cutoff frequencies and bandwidths of the ten filters associated with the keyboard were stationary, and so control was imposed by allowing the key presses to modulate the signal amplitude within each filter band. In effect, this provided the operator a means of continuously enhancing or

suppressing the ten discrete divisions of the spectrum in some selective pattern such that an approximation of time-varying formants were generated. It can be noted that Faber's mechanical talking machine from a century earlier presented an operator with essentially the same type of interface as the Voder (i.e., keyboard, foot pedal), but it was the shape of cavities analogous to the human vocal tract that were controlled rather than the speech spectrum itself. In either case, and like a human acquiring the ability to speak, the operators of the devices learned and internalized a set of rules for generating speech by modulating a relatively high-frequency carrier signal (i.e., vocal fold vibration, turbulence) with slowly-varying, and otherwise inaudible, "message waves" (Dudley, 1940). Although the ability of a human operator to acquire such rules is highly desirable for performance-driven artificial speech, it would eventually become a major goal for researchers in speech synthesis to explicate such rules in an attempt to understand phonology and motor planning in speech production, as well as to develop algorithms for transforming symbolic phonetic representations into speech via synthetic methods.

The research at Bell Labs that contributed to the Vocoder and Voder occurred in parallel with development of the "sound spectrograph" (Potter, 1945), a device that could graphically represent the *time-varying record of the spectrum* of a sound rather than the waveform. The output of the device, called a "spectrogram," was arranged such that time and frequency were on the *x* and *y* axes, respectively, and intensity was coded by varying shades of gray. It could be set to display either the *narrowband* harmonic structure of a sound, or the *wideband* formant patterns. Although development of the spectrograph had been initiated by Ralph Potter and colleagues just prior to the United States involvement in World War II, it was given "official rating as a war project" (Potter, 1945, p. 463) because of its potential to facilitate military communications and message decoding. During the war, the spectrograph design was refined and used extensively to study the temporo-spectral patterns of speech based on the "spectrograms" that it generated. It wasn't until several months after the war ended, however, that the existence of the spectrograph was disclosed to the public. On November 9, 1945, Potter published an article in *Science* titled "Visible Patterns of Sound" in which he gave a brief description of the device and explained its potential application as a tool for studying phonetics, philology, and music. He also suggested its use as an aid for persons who are hearing impaired; the idea was that transforming speech from the auditory to the visual domain would allow a trained user to "read" speech. Other publications regarding the spectrograph soon followed with more detailed descriptions concerning its design (Koenig, Dunn, and Lacy, 1946; Koenig and Ruppel, 1948) and use (Kopp and Green, 1946; Steinberg and French, 1946; Potter and Peterson, 1948; Potter, 1949).

Just as instrumentation that allowed researchers to *see* acoustic speech waveforms had motivated earlier methods of synthesis (e.g., Miller, 1916), the spectrographic visualization of speech would rapidly inspire new ways of synthesizing speech, and new reasons for doing so. Following World War II, Frank Cooper and Alvin Liberman, researchers at Haskins Laboratories in New York City, had begun extensive analyses of speech using a spectrograph based on the Bell Labs design. Their goals, which were initially concerned with building a reading machine for the blind, had been diverted to investigations of the acoustic structure of speech, and how they were perceived and decoded by listeners. They realized quickly, however, that many of their questions could not be answered simply by inspection of spectrograms. What was needed was a means of modifying some aspect of the visual representation of speech provided by the spectrogram, and transforming it back into sound so that it could be presented to a listener as a stimulus. The responses to the stimuli would indicate whether or not the spectral modification was perceptually relevant.

In 1951, Cooper, Liberman, and Borst reported on the design of a device that would allow the user to literally draw a spectrographic representation of a speech utterance on a film transparency and transform it into a sound wave via a system including a light source, tone wheel, photocell, and amplifier. The tone wheel contained 50 circular sound tracks that, when turned by a motor at 1800 rpm, would modulate light to generate harmonic frequencies from 120-6000 Hz, roughly covering the speech spectrum. The photocell would receive only the portions of spectrum corresponding to the pattern that had been drawn on the film, and convert them to an electrical signal which could be amplified and played through a loudspeaker. The "drawn" spectrographic pattern could be either a copy or modification of an actual spectrogram, and hence the device came to be known as the "Pattern Playback." It was used to generate stimuli for numerous experiments on speech perception and contributed greatly to knowledge and theoretical views on how speech is decoded (cf., Liberman et al., 1952, 1954, 1957; Liberman, 1957; Harris et al., 1958; Liberman et al., 1967). The Pattern Playback was the first speech synthesizer used for large scale systematic experimentation concerning the structure of speech, and proved to be most useful for investigations concerning isolated acoustic cues such as formant transitions at the onset and offset of consonants (Delattre et al., 1952; Borst, 1956).

The usefulness of speech synthesizers as research tools was summarized in a review article by Cooper (1961) in which he wrote:

"*The essential point here, as in all of science, is that we must simplify Nature if we are to understand her. More than that: we must somehow choose a particular set of simplifying assumptions from the many sets that are possible. The great virtue of speech synthesizers is that they can help us make this choice."* - Cooper, 1961, p. 4

And so, the Pattern Playback served the purpose of "simplifying Nature" by making the spectro-temporal characteristics of speech accessible and manipulable to the researcher. When used in this manner, a speech synthesizer becomes an experimenter's "versatile informant" that allows for testing hypotheses about the significance of various spectral features (Cooper, 1961, pp. 4-5). One of advantages of the Pattern Playback was that virtually anything could be drawn (or painted) on the film transparency regardless of the complexity or simplicity, *and* it could be heard. For example, all of the detail observed for a speech utterance in a spectrogram could be reconstructed, or something as simple as a sinusoid could be drawn as a straight line over time. The disadvantage was that the only means of generating an utterance, regardless of the accuracy of the prescribed rules, was for someone to actually draw it by hand.

The users of the Pattern Playback became quite good at drawing spectrographic patterns that generated intelligible speech, even when they had not previously seen an actual spectrogram of the utterance to be synthesized (Delattre et al., 1952; Liberman et al., 1959). Much like the operators of the speaking machines that preceded them, they had, through practice, acquired or internalized a set of rules for generating speech. Delattre et al., (1952) did attempt to characterize some speech sounds with regard to how they might be drawn spectrographically, but it was Frances Ingemann who formalized rules for generating utterances with the Pattern Playback (Ingemann, 1957; Liberman et al., 1959). The rules were laid out according to place, manner, and voicing, and could be presumably used by a novice to draw and generate a given utterance. Although the process would have been extremely time consuming and tedious, Mattingly (1974) notes that this was the *first* time that explicit rules for generating speech with a synthesizer had been formally documented.

Other types of synthesizers were also developed during this period that facilitated production of artificial speech based on acoustic characteristics observed in a spectrogram, but were based on different principles than the Pattern Playback. In 1953, Walter Lawrence, a researcher for the Signals Research and Development Establishment in Christchurch, England, introduced a speech synthesizer whose design consisted of an electrical circuit with a source function generator and three parallel resonant branches. The frequency of each resonance could be controlled by the user, as could the frequency and amplitude of

the source function. Together, the source and resonant branches produced a waveform with a time-varying spectrum that could be compared to a spectrogram, or modified for purposes of determining the perceptual relevance of an acoustic cue. Because the parameters of the circuit (i.e., resonance frequencies, source fundamental frequency, etc.) were under direct control, Lawrence's synthesizer became known as the "Parametric Artificial Talker" or "PAT" for short. PAT was used by Peter Ladefoged and David Broadbent to provide acoustic stimuli for their well-known study of the effects of acoustic context on vowel perception (Ladefoged and Broadbent, 1957).

At about the same time, Gunnar Fant was also experimenting with resonant circuits for speech synthesis at the Royal Institute of Technology (KTH) in Stockholm. Instead of placing electrical resonators in parallel as Lawrence did in building PAT, Fant configured them in a series or "cascade" arrangement. Fant's first cascade synthesizer, called "OVE I," an acronym based on the words "Orator Verbis Electris," was primarily a vowel synthesizer which had the unique feature of a mechanical stylus that could be moved in a two-dimensional plane for control of the lowest two resonance frequencies. A user could then generate speech (vowels and vowel transitions) by moving the stylus in the vowel space defined by the first two formant frequencies, a system that may have had great value for teaching and learning the phonetics of vowels. It may have had some entertainment value as well. Fant (2005) reminisced that one of the three "opponents" at his doctoral dissertation defense in 1958 was, in fact, Walter Lawrence who had brought his PAT synthesizer with him to Stockholm. At one point during the defense proceedings Fant and Lawrence demonstrated a synthesizer dialogue between PAT and OVE I. Eventually, Fant developed a second version of the cascade-type synthesizer called "OVE II' (Fant and Martony, 1962). The main enhancements were additional subsystems to allow for production of nasals, stops, and fricatives, as well as a conductive ink device for providing time-varying parameter values to the synthesizer.

The development of PAT and OVE set the stage for a category of artificial speech that would eventually be referred to as *formant* synthesis, because they provided for essentially direct control of the formants observed in a spectrogram. In a strict sense, however, they are *resonance* synthesizers because the parameters control, among other things, the frequencies of the electrical (or later, digital) resonators themselves. In most cases though, these frequencies are aligned with the center frequency of a formant, and hence resonance frequency and formant frequency become synonymous. Although it may seem like a minor technological detail, the question of whether such synthesizers should be designed with parallel or cascaded resonators would be debated for years to come. A parallel system offers the user the largest

amount control over the spectrum because both the resonator frequencies and amplitudes are set with parameters. In contrast, in a cascade system the resonance frequencies are set by a user, while their amplitudes are an effect of the superposition of multiple resonances, much as is the case for the human vocal tract (Flanagan, 1957). Thus, the cascade approach could potentially produce more natural sounding speech, but with somewhat of a sacrifice in control. Eventually, Lawrence reconfigured PAT with a cascade arrangement of resonators, but after many years of experimentation with both parallel and cascade systems, John Holmes of the Joint Speech Research Unit in the UK later made a strong case for a parallel arrangement (Holmes, 1983). He noted that replication of natural speech is considerably more accurate with user control of both formant frequencies and their amplitudes.

Simultaneous with the development of formant synthesizers in the 1950s, was another type of synthesis approach, also based on electrical circuits, but intended to serve as a model of the shape of the vocal tract so that the relation of articulatory configuration to sound production could be more effectively studied. The first of this type was described in 1950 by H. K. Dunn, another Bell Labs engineer. Instead of building resonant circuits to replicate formants, Dunn (1950) designed an electrical transmission line in which consecutive (and coupled) "T-sections," made up of capacitors, inductors, and resistors, were used as analogs of the pharyngeal and oral air cavities within the vocal tract. The values of the circuit elements within each T-section were directly related to the cross-sectional area and length of the various cavities, and thus the user now had parametric control of the vocal tract shape. Although this was an advance, the vocal tract configurations that could be effectively simulated with Dunn's circuit were fairly crude representations of the human system.

Stevens, Kasowski, and Fant (1953), in their article "An Electrical Analog of the Vocal Tract," describe a variation on Dunn's design using a similar transmission line approach but they were able to represent the vocal tract shape as a concatenation of 35 cylindrical sections, where each section was 0.5 cm in length. The purpose in pursuing a more detailed representation of the vocal tract shape was to be able to "... study in detail the mechanism of speech production and to investigate correlations between the acoustic and articulatory aspects of speech" and noted that "a speech synthesizer would be required to simulate more closely the actual dimensions of the vocal tract" (p. 735). Fant also began work on his own version of a Line Electrical Analog (LEA) that he used for studies of speech sounds. Both Stevens and House (1955) and Fant (1960) used these very similar synthesizers to better understand vowel articulation by first developing a three parameter model of the vocal tract shape in which the location and radius of the primary vowel constriction were specified, along with the ratio of lip termination area to lip tube length.

Their synthesizers allowed for a systematic exploration of the parametric space and resulted in nomographic displays that demonstrated the importance of the location (place) and cross-sectional area of the primary vocal tract constriction in vowels. Collectively, this work significantly altered the view of vowel production.

A limitation of both the Stevens et al. (1953) and Fant (1960) line analogs synthesizers was that they could not generate time-varying speech sounds because they accommodated only static vocal tract configurations; i.e., they couldn't actually talk. Using a more complex line analog circuit system and a bank of switches, George Rosen, a doctoral student at MIT, devised a means of transitioning from one vocal tract configuration to another (Rosen, 1958). This new system, known as "DAVO" for "**d**ynamic **a**nalog of the **vo**cal tract," could generate fairly clear diphthongs and consonant-vowel (CV) syllables, but was not capable of sentence-level speech.

It can be noted that the parametric models of the vocal tract shape developed by Stevens and House (1955) and Fant (1960) were independent of the transmission line analogs that were used to produce the actual synthetic speech. The limitation of only static vowels, or CVs in the case of DAVO, was entirely due to the need for a complicated electrical circuit to simulate the propagation of acoustic waves in the vocal tract. The vocal tract models themselves could have easily been used to generate time-dependent configurations over the time course of a phrase or sentence, but a system for producing the corresponding synthetic speech waveform with such temporal variation simply did not yet exist, nor did the knowledge of how to specify the time-dependence of the vocal tract parameters.

A third type of speech synthesizer was also under development during the 1950s. With significant improvements in the state of audio recording technology, particularly those related to storing speech waveforms on magnetic tape, it was now possible to consider synthesis, perhaps in the purest sense of the word, based on splicing together small segments of prerecorded natural speech. Harris (1953a) designed a system in which segments of tape were isolated that contained many instances (allophones) of each consonant and vowel. Then, with a recording drum, tape loop, and timing and selector circuits (Harris, 1953b), synthetic speech could be generated by piecing together a sequence of segments deemed to match well with regard to formant frequencies and harmonics. The speech produced was found to be fairly intelligible but quite unnatural, presumably because of the discontinuities created at the segment boundaries.

Rather than focusing on vowel and consonant segments, Peterson, Wang, and Sivertson (1958) experimented with alternative segmentation techniques and determined that a more useful unit for synthesizing speech could be obtained from segments extending in time from the steady state location of one phoneme to the next. Referring to this unit as a "dyad," they suggested that it preserved the acoustic dynamics of the transitions between phonemes, precisely the information lost when the segmentation unit is the phoneme itself. The potential of this method to generate intelligible speech was demonstrated by Wang and Peterson (1958) where they constructed a sentence from more than 40 dyad segments extracted from previously recorded utterances. Much care was required in selecting the segments, however, in order to maintain continuity of pitch, intensity, tempo, and vocal quality. The range of phonetic characteristics that can be generated in synthetic speech by concatenating segments is, of course, limited by the segment inventory that is available. Sivertson (1961) conducted an extensive study of the size of inventory needed relative to the type of segmentation unit chosen. She considered various segments with a wide range of temporal extent that included phonemes, phoneme dyads, syllable nuclei, half syllables, syllables, syllable dyads, and words, and found that, in general, "the size of the inventory increases with the length of the segment." That is, a few small units can be combined in many ways to generate hundreds or thousands of increasingly larger units, but if the starting point is a large temporal unit, an enormous number is needed because the possibilities for recombining them are severely limited.

This approach to synthesis clearly has played a major role in technological applications such as modern text-to-speech systems utilizing unit selection techniques (cf., Moulines and Charpentier, 1990; Sagasaka et al., 1992; Hunt and Black, 1996), but Sivertson (1961) also made a strong case for the use of segment concatenation methods as a research tool. In particular, she noted that using stored segments of various lengths can be used for evaluating some theories of linguistic structure, as well as for investigating segmentation of speech signals in general. In fact, Sivertson suggested that essentially all speech synthesis methods could be categorized relative to how the speech continuum is segmented. If the segmentation is conceived as "simultaneous components" then speech can be synthesized by controlling various parametric representations "independently and simultaneously." These may be physiological parameters such as vocal tract shape, location and degree of constriction, nasal coupling, and laryngeal activity, or acoustical parameters such as formant frequencies, formant bandwidths, fundamental frequency, voice spectrum, and amplitude. If, instead, the speech continuum in segmented in time, synthetic speech can be accomplished by sequencing successive "building blocks," which may be extracted from recorded natural speech or even generated electronically.

The advent of digital computing in the early 1960s would dramatically change the implementation of speech synthesizers, and the means by which they may be controlled. The underlying principles of the various synthesis methods, however, are often the same or least similar to those that motivated development of mechanical, electrical, or electronic talking devices. Thus, delineation of synthesizer type relative to the segmentation of the speech continuum is particularly useful for understanding the differences and potential uses of the wide range of synthetic speech systems that had so far been advanced at the time, and also for those yet to be developed.

4 The digital and computational era

As Stewart (1922) had suggested in the early days of electrical circuit-based synthesis, building a device capable of producing sounds that resemble speech is far less difficult than knowing how to impose the proper control on its parameters to make the device actually talk. Although much progress had been made in development of various types of systems, controlling electronic speech synthesizers by manipulating circuit parameters, whether they were vocal tract or terminal analog types, was cumbersome and tedious. This could now be mitigated to some degree, however, by engaging a digital computer to control speech synthesizers that were, themselves, still realized as electronic circuits. That is, commands typed on a keyboard could be transformed into control voltages that imposed parameter changes in the synthesis circuitry. In effect, this allowed the "computational load" for generating the speech waveform to remain in the analog circuit, while transferring the control of the system to a user via a computer interface. It would not be long, however, before the hardware synthesizers were replaced with software realizations of the same circuit elements, offering far greater flexibility and ease with which synthetic speech could be generated.

Digital control facilitated development of “speech synthesis by rule” in which an orthographic representation of an utterance could be transformed into artificial speech. Based on a set of “rules” embedded in a computer program, a series of symbols representing the phonetic elements of a word or phrase were converted to temporal variations of the parameters of a specific type of synthesizer. For example, Holmes et al. (1964) described the rules and associated computer program that calculated the time course of the parameters of a parallel resonance (formant) synthesizer. These included, among other variables, frequencies and amplitudes of three resonances, and fundamental frequency. A word such as “you (/ju/) might might be produced with a simple interpolation of the second formant frequency, F_2, from a high value, say 2200 Hz, down to a much lower value, perhaps 400 Hz, while other parameters

could be held constant. The interpolated F_2 would then be used to alter the settings of circuit elements over a particular period of time, resulting in a speech waveform resembling "you."

A similar goal of producing "synthetic speech from an input consisting only of the names of phonemes and a minimum of pitch and timing information..." was pursued by Kelly and Lochbaum (1962), but in their system a digital lattice filter, entirely realized as a computer algorithm, was used to calculate the effective propagation of acoustic waves in an analog of the vocal tract configuration. Control of the system required specification of 21 time-varying cross-sectional areas representing the vocal tract shape along the axis extending from the glottis to lips, as well as nasal coupling, fundamental frequency, aspiration, and affrication. Each phoneme was assigned a vocal tract shape (i.e., 21 cross-sectional areas) read from lookup table; change in tract shape was accomplished by linearly interpolating, over time, each cross-sectional area specified for one phoneme to those of the next phoneme. This design functioned essentially as a digital version of a synthesizer like George Rosen's DAVO; but because it was software rather than hardware, it allowed for more precise specification of the vocal tract shape and almost endless possibilities for experimentation with interpolation types and associated rules.

Kelly and Lochbaum expressed disappointment in the sound quality of the speech generated by their system, but attributed the problem to inadequate knowledge of the cross-sectional areas that were used as the vocal tract shapes corresponding to phoneme targets. Although based on Fant's (1960) well-known collection of vocal tract data obtained from X-ray images, it would not be until the 1990's that imaging methods would allow for three-dimensional reconstructions of vocal tract shapes produced by human talkers (cf., Baer et al., 1991, Story et al., 1996, 1998), and hence, improve this aspect of analog vocal tract synthesis. The time-varying spatial characteristics of a linearly interpolated vocal tract shape were, however, also potential contributors to the undesirable quality of the synthesis. A more complex set of rules for control of a vocal tract analog was described a few years later by Nakata and Mitsuoka (1965), and resulted in intelligible speech with "fairly good naturalness." Nonetheless, they, along with others believed that significant improvements in vocal tract analog synthesis required more detailed knowledge of realistic articulatory movement from which better timing rules could be derived.

By the 1960s, x-ray cineradiography technology had developed to a point where the spatial and temporal resolution were suitable for studying the articulatory movements of speech in a sagittal projection image. Motion picture X-ray films collected for various speech utterances could be analyzed frame by frame to tract the changing positions of articulators and the time-varying configuration of the vocal tract outline.

Just as the instrumentation that allowed scientists to *see waveforms and spectrograms* had motivated earlier forms of synthetic speech, the ability to now *see the movement of the articulators* motivated development of a new type of synthesis paradigm called "articulatory synthesis."

In 1967, Cecil Coker of Bell Laboratories demonstrated a synthesis system based on a computational model of the speech articulators. Simplified positions of the tongue, jaw, lips, velum, and larynx were represented in the midsagittal plane, where each could be specified to move with a particular timing function. The result was a time-varying configuration of the midsagittal vocal tract outline. To produce the speech waveform, the distances across the vocal tract airspace from glottis to lips at each time sample first needed to be converted to cross-sectional areas to form the area function (cf., Heinz and Stevens, 1964). These were then used in a vocal tract analog model like that of Kelly and Lochbaum (1962) to calculate wave propagation through the system. The resonance frequencies could also be calculated directly from the time-varying area function and used to control a formant synthesizer (Coker and Fujimura, 1966). Similar articulatory models were developed by Lindblom and Sundberg (1971) and Mermelstein (1973), but incorporated somewhat more complexity in the articulatory geometry. In any case, the temporal characteristics of the synthesized articulatory movement could be compared to, and refined with, data extracted from midsagittal cineradiography films (e.g., Truby, 1965). An articulatory synthesizer developed at Haskins Laboratories called ASY (Rubin, et al., 1981), extended the Mermelstein model, incorporating several additional sub-models and an approach based on key frame animation for synthesizing utterances derived from control of the movement over time of the vocal tract. This was one of the earliest articulatory synthesis tools used for large scale laboratory phonetic experiments (e.g. Abramson, et al., 1981). It was later enhanced to provide more accurate representations of the underlying vocal tract parameters and flexibility in their control by a user (CASY: The Haskins Configurable Articulatory Synthesizer; see Rubin, et al., 1996).

Articulatory synthesis held much promise because it was assumed that the rules required to generate speech would be closer to those used by a human talker than rules developed for controlling acoustic parameters such as formant frequencies. While that may ultimately be the case, such rules have been difficult to define in such a way that natural sounding, intelligible speech is consistently generated (Klatt, 1987). Articulatory synthesizers have become important tools for research, however, because they can serve as a *model* of speech production in which the acoustic consequences of parametric variation of an articulator can be investigated. Using the ASY synthesizer (Rubin et al., 1981) to produce speech output, Browman et al. (1984), Browman and Goldstein (e.g., 1985, 1991), Saltzman (1986, 1991), and others at

Haskins Laboratories embarked on research to understand articulatory control. Guiding this work was the hypothesis that phonetic structure could be characterized explicitly as articulatory movement patterns, or "gestures." Their system allowed for specification of an utterance as a temporal schedule of "tract variables," such as location and degree of a constriction formed by the tongue tip or tongue body, lip aperture and protrusion, as well as states of the velum and glottis. These were then transformed into a task-dynamic system that accounted for the coordination and dynamic linkages among articulators required to carry out a specified gesture. Over the years, techniques for estimating the time course of gesture specification have continued to be enhanced. Recently, for example, Nam et al. (2012) reported a method for estimating gestural "scores" from the acoustic signal based on an iterative analysis-by-synthesis approach.

Some developers of articulatory synthesis systems focused their efforts on particular subsystems such as the voice source. For example, Flanagan and Landgraf (1968) proposed a simple mass-spring-damper model of the vocal folds that demonstrated, with a computational model, the self-oscillating nature of the human vocal folds. Shortly thereafter, a more complex two-mass version was described (Ishizaka and Matsudaira, 1972; Ishizaka and Flanagan, 1972) that clearly showed the importance of the vertical phase difference (mucosal wave) in facilitating vocal fold oscillation. Additional degrees of freedom were added to the anterior-posterior dimension of the vocal folds by Titze (1973, 1974) with a 16 mass model. Although the eventual goal of subsystem modeling would be integration into a full speech synthesis system (cf., Flanagan et al., 1975; Sondhi and Schroeter, 1987), much of their value is as a tool for understanding the characteristics of the subsystem itself. Maeda (1988, 1990), Dang and Honda (2004), and Birkholz (2013) are all examples of more recent attempts to integrate models of subsystems into an articulatory speech synthesizer, whereas Guenther (cf., 1994) and Kröger et al. (2010) have augmented such synthesizers with auditory feedback and learning algorithms. The main use of these systems has been to study some aspect of speech production, but not necessarily the conversion of a symbolic representation of an utterance into speech. It can also be noted that a natural extension of articulatory synthesis is the inclusion of facial motion that coincides with speaking, and has led to development audio-visual synthetic speech systems that can be used explore multi-modal nature of both speech production and perception (cf., Yehia et al., 1998; Massaro, 1998; Vatikiotis-Bateson et al., 2000). This area will be covered in the chapter entitled "New horizons in clinical phonetics."

Other researchers focused on enhancing models of the vocal tract analogs *without* adding the complexity of articulatory components. Strube (1982), Liljencrants (1985), and Story (1995) all refined the digital lattice filter approach of Kelly and Lochbaum (1962) to better account for the acoustic properties of time-varying vocal tract shapes and various types of energy losses. Based on the relation of small perturbations of tubular configuration to changes in the acoustic resonance frequencies, Mrayati et al. (1988) proposed that speech could be produced by controlling the time-varying cross-sectional area of eight distinct regions of the vocal tract. The idea was that expansion or constriction of these particular regions would maximize the change in resonance frequencies, thus providing an efficient means of controlling the vocal tract to generate a predictable acoustic output. Some years later a set of rules was developed by Hill et al. (1995) that specified the transformation of text input into region parameters and were used to build a vocal tract-based text-to-speech synthesizer. In a similar vein, Story (2005,2013) has described an "airway modulation model" of speech production (called "TubeTalker") in which an array of functions can be activated over time to deform the vocal tract, nasal tract, and glottal airspaces to produce speech. Although this model produces highly intelligible speech, its primary use is for studying the relation of structure and movement to the acoustic characteristics produced, and the perceptual response of listeners (cf., Story and Bunton, 2010).

In parallel with development of articulatory-type synthesizers was the enhancement of resonance or formant-based synthesis systems. Along with numerous colleagues, Dennis Klatt's research on digital resonators as well as his studies on the acoustic characteristics of nearly all aspects of speech, led to a comprehensive system of rule-based formant synthesis. With various names such as "Klattalk," "MITalk," "DecTalk," and later "KLSYN88," this type of text-to-speech system has become well known to the public, particularly because of its collection of standard voices (cf., Klatt, 1982, 1987; Klatt and Klatt, 1990). Perhaps best known today is "Perfect Paul," the voice that was synonymous with the late British physicist Stephen Hawking who used the synthesizer as an augmentative speaking device. Formant synthesis can also be combined with articulatory methods. "HLSyn," developed by Helen Hanson and Ken Stevens (2003), is a system designed to superimpose high level (HL) articulatory control on the Klatt formant synthesizer. This approach simplified the control scheme by mapping 13 physiologically-based HL parameters to the 40-50 acoustic parameters that control the formant synthesis. The advantage is that the articulatory parameters constrain the output so that physiologically unrealistic combinations of the voice source and vocal tract filter cannot occur. This type of synthesizer can serve as another tool for studying speech production with regard to both research and educational purposes.

Throughout this chapter, it has been presumed that regardless of the reasons for developing a particular type of synthesizer, at some level, the goal was to generate high-quality, intelligible speech. Some synthesizers have been developed, however, for the explicit purpose of degrading or modifying natural, recorded speech. Such synthesizers are useful for investigating speech perception because they allow researchers to systematically remove many of the acoustic characteristics present in the signal while preserving only those portions hypothesized to be essential cues. Remez et al. (1981) described a synthesis technique in which the first three formant frequencies tracked over the duration of a sentence, were replaced by the summation of three tones whose frequencies were swept upward and downward to match the temporal variation of the formants. Although the quality of the synthesized sound was highly artificial (perhaps otherworldly), listeners were able to identify the sentences as long as the tones were played simultaneously, and not in isolation of one another, revealing the power of speech cues that are embedded in the dynamic spectral patterns of the vocal tract resonances. Shannon et al. (1995) showed that intelligible speech could alternatively be synthesized by preserving temporal cues, while virtually eliminating spectral information. Their approach was essentially the same as Dudley's Vocoder (1939a) in which the speech signal was first filtered into a set of frequency bands, and time-varying amplitude envelopes were extracted from each band over the duration of the recorded speech. The difference was that the number of bands ranged from only 1-4, and the amplitude envelopes were used modulate a noise signal rather than an estimation of the voice source. Shannon et al. showed that listeners were adept at decoding sentence level speech with only three bands of modulated noise. Similarly designed synthesizers (e.g., Loizou et al., 1999) have been used simulate the signal processing algorithms in cochlear implant devices for purposes of investigating speech perception abilities of listeners under these conditions. Yet another variation on this type of synthesis was reported by Smith et al. (2002) who developed a technique to combine the spectral fine structure of one type of sound with the temporal variation of another to generate "auditory chimeras." These have been shown to be useful for investigating aspects of auditory perception.

Many other types of speech synthesis methods have been developed in the digital era whose primary purpose is to generate high quality speech for automated messaging or be embodied in a digital assistant that converses with a user. These systems typically make use of synthesis techniques that build speech signals from information available in a database containing many hours of recordings of one or more voice professionals who produced a wide range of spoken content and vocal qualities. The "unit selection" technique, also referred to as "concatenative synthesis," is essentially the digital realization of the tape splicing method of Harris (1953b) and Peterson, Wang, and Sivertson (1958), but now involves a

set of algorithms that efficiently search the database for small sound segments, typically at the level of diphones, that can be stacked serially in time to generate a spoken message. A different technique, called "parametric synthesis," relies on an extensive analysis of the spectral characteristics of speech recordings in a database to establish parametric representations that can later be used to reconstruct a segment of speech (Zen et al., 2009). Unit selection typically produces more natural sounding speech but is limited by the quality and size of the original database. Parametric synthesis allows for greater flexibility with regard to modification of voice characteristics, speaking style, and emotional content, but generally is of lower overall quality. Both techniques have been augmented with implementation of deep learning algorithms that improve the efficiency and accuracy of constructing a spoken utterance, as well as increasing the naturalness and intelligibility of the synthetic speech (Zen et al., 2013; Capes et al., 2017). More recently, a new approach called direct waveform modeling has been introduced that utilizes a deep neural network (DNN) to generate new speech signals based on learned features of recorded speech (cf., van den Oord et al., 2016; Arik et al., 2017). This method has the potential to significantly enhance the quality and naturalness of synthetic speech over current systems, even though it is currently computationally expensive. It can be noted, however, that because unit selection, parametric, and direct waveform synthesizers construct speech signals based on underlying principles that are not specifically related to the ways in which a human forms speech, they are perhaps less useful as a tool for testing hypotheses about speech production and perception than many of the other techniques discussed in this chapter.

5 Summary

For centuries, past to present, humans have been motivated to build machines that talk. Other than the novelty, what is the purpose of speech synthesis, and what can be done with it? Certainly, technological applications have resulted from development of these devices, many of them having a major impact on how humans communicate with each other. Mattingly (1974) seems to have hit it just about right when he suggested that the "traditional motivation for research in speech synthesis" has been simply the desire to explain the mystery of how we humans successfully use our vocal tracts to produce connected speech. In other words, the primary means of scientifically investigating speech production has been based on building artificial talking systems and collecting relevant data with which to refine them. Mattingly (1974) also points out that, regardless of the underlying principles of the synthetic speech system built, the scientific questions are almost always concerned with deriving the "rules" that govern production of intelligible speech. Such rules may be elaborate and explicitly stated algorithms for transforming a string

of text into speech based on a particular type of synthesizer, or more subtly implied as general movements of structures or acoustic characteristics. In any case, achieving an understanding of the rules and all their variations, can be regarded as synonymous with understanding many aspects of speech production and perception.

As in any area of science, the goal in studying speech has been to first determine the important facts about the system. Artificial talkers and speech synthesizers embody these "facts," but they typically capture just the essential aspects of speech. As a result, synthesis often presents itself as an aural caricature that can be perceived as an unnatural, and sometimes amusing rendition of a desired utterance or speech sound. It is particularly unique to phonetics and speech science that the models used as tools to understand the scientific aspects of a complex system produce a signal intended to be heard as if it were a human. As such, the quality of speech synthesis can be rather harshly judged because the model on which it is based has not accounted for the myriad of subtle variations and details that combine in natural human speech. Thus, we should keep in mind that the degree to which we can produce convincing artificial speech is a measure of the degree to which we understand human speech production.

6 Acknowledgments

The research was supported in part by NIH R01 DC011275 and NSF BCS-1145011.

7 References

Abramson, A. S., Nye, P. W., Henderson, J. B., and Marshall, C. W., 1981. Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America*, 70, pp. 329-339.

Altick, R. D., 1978. *The shows of London*. Cambridge, MA and London: Belknap Press of Harvard University Press, 355-356.

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., and Shoeybi, M., 2017. Deep Voice: real-time neural text-to-speech. arXiv:1702.07825 [Accessed 5 March 2018].

Athenaeum, Jul 25, 1846, *British Periodicals*, 978, p. 765.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W., 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, *Journal of the Acoustical Society of America*, 90, pp. 799-828.

Baldwin Jewett, F., 1936. The Social Implications of Scientific Research in Electrical Communication. *The Scientific Monthly*, 43, pp. 466-476.

Barnum, P. T., 1886. *Life of P. T. Barnum Written By Himself*, Buffalo, NY: The Courier Company Printers, p. 134.

Birkholz, P., 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis, *PloS one*, 84), e60603, pp 1-17. https://doi.org/10.1371/journal.pone.0060603.

Browman, C. P., Goldstein, L., Kelso, J. A. S., Rubin, P. E., and Saltzman, E., 1984. Articulatory synthesis from underlying dynamics, *Journal of the Acoustical Society of America*, 75, S22.

Browman, C. P. and Goldstein, L. M., 1985. Dynamic modeling of phonetic structure, *Phonetic Linguistics*, pp. 35-53.

Borst, J.M., 1956. The use of spectrograms for speech analysis and synthesis, *Journal of the Audio Engineering Society*, 4, pp. 14-23.

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., Prahallad, K, Raitio, T., Rasipuram, R., Townsend, G., Williamson, B., Winarsky, D., Wu, X. and Zhang, H., 2017. Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System. *Proceedings of Interspeech 2017*, pp. 4011-4015.

Carroll, C. M., 1975. *The great chess automaton*, New York:Dover Publications.

Coker, C. H., and Fujimura, O., 1966. Model for specification of the vocal tract area function. *Journal of the Acoustical Society of America*, 40(5), pp. 1271-1271.

Coker, C. H., 1967. Synthesis by rule from articulatory parameters, *Proceedings of the 1967 Conference on Speech Communication Processes*, Cambridge, MA: IEEE, A9, pp. 52-53.

Cooper, F. S., 1961. Speech synthesizers. *Proceedings of the 4th International Congress of Phonetic Sciences (ICPhS'61)*, pp. 3-13.

Cooper, F. S., Liberman, A. M., and Borst, J. M., 1951. The interconversion of audible and visible patterns as a basis for research in the perception of speech, *Proceedings of the National Academy of Sciences*, 37(5), pp. 318-325.

Dang, J., and Honda, K., 2004. Construction and control of a physiological articulatory model, *Journal of the Acoustical Society of America*, 115, pp. 853-870.

Delattre, P., Cooper, F. S., Liberman, A. M., and Gerstman, L. J., 1952. Speech synthesis as a research technique. *Proceedings of the Vllth International Congress of Linguists*, London, pp. 543-561.

Dudley, H., 1936. Synthesizing speech, *Bell Laboratories Record*, 15, pp. 98-102.

Dudley, H., 1939a. The vocoder, *Bell Laboratories Record*, 18(4), pp. 122-126.

Dudley, H., Riesz, R. R., and Watkins, S. S. A., 1939b. A synthetic speaker, *Journal of the Franklin Institute*, 227(6), pp. 739-764.

Dudley, H., 1940. The carrier nature of speech, *Bell System Technical Journal*, 19(4), pp. 495-515.

Dudley, H. and Tarnoczy, T. H., 1950. The speaking machine of Wolfgang von Kempelen, *Journal of the Acoustical Society of America*, 22(2), pp. 151-166.

Dunn, H. K., 1950. The calculation of vowel resonances and an electrical vocal tract. *Journal of the Acoustical Society of America*, 22(6), pp. 740-753.

Euler, L., 1761. The wonders of the human voice, *Letters of Euler on Different Subjects in Natural Philosophy addressed to German Princess*, David Brewster, ed., New York: J.J. Harper (publisher in 1833), pp. 76-79.

Fant, G., 1960. *Acoustic theory of speech production*, The Hague:Mouton.

Fant, G., and Martony, J., 1962. Speech synthesis instrumentation for parametric synthesis (OVE II). *Speech Transmission Laboratory Quarterly Progress and Status Report (KTH)*, 2, pp. 18-24.

Fant, G., 2005. Historical notes, *Speech Transmission Laboratory Quarterly Progress and Status Report (KTH)*, 47(1), pp. 9-19.

Flanagan, J. L., 1957. Note on the design of "terminal-analog" speech synthesizers. *Journal of the Acoustical Society of America*, 29(2), pp. 306-310.

Flanagan, J., and Landgraf, L., 1968. Self-oscillating source for vocal-tract synthesizers, *IEEE Transactions on Audio and Electroacoustics*, 16(1), pp. 57-64.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L., 1975. Synthesis of speech from a dynamic model of the vocal cords and vocal tract, *Bell System Technical Journal*, 54(3), 485-506.

Guenther, F. H., 1994. A neural network model of speech acquisition and motor equivalent speech production, *Biological cybernetics*, 72(1), 43-53.

Hanson, H. M., and Stevens, K. N., 2002. A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn, *Journal of the Acoustical Society of America*, 112(3), pp. 1158-1182.

Harris, C. M., 1953a. A study of the building blocks in speech, *Journal of the Acoustical Society of America*, 25(5), pp. 962-969.

Harris, C. M., 1953b. A speech synthesizer, *Journal of the Acoustical Society of America*, 25(5), pp. 970-975.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper, F. S., 1958. Effect of third formant transitions on the perception of the voiced stop consonants. *Journal of the Acoustical Society of America,* 30(2), pp. 122-126.

Heinz, J. M., and Stevens, K. N., 1964. On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *Journal of the Acoustical Society of America*, 36(5), pp. 1037-1038.

Helmholtz, H., 1859. Ueber die Klangfarbe der Vocale [On the timbre of vowels], *Annalen der Physik*, 123, pp. 527-528.

Helmholtz, H., 1875. *On the sensations of tone as a physiological basis for the theory of music*, London: Longmans, Green, and Co.

Hermann, L., 1894. Nachtrag zur Untersuchung der Vocalcurven, *Pflügers Archiv European Journal of Physiology*, 58, pp. 264-279.

Hermann, L., 1895. Weitere Untersuchungen über das Wesen der Vocale, *Pflügers Archiv European Journal of Physiology*, 61(4), pp. 169-204.

Hill, D., Manzara, L., and Schock, C., 1995. Real-time articulatory speech-synthesis-by-rules, *Proceedings of AVIOS*, 95, pp. 27-44.

Holmes, J. N., Mattingly, I. G., and Shearme, J. N., 1964. Speech synthesis by rule, *Language and Speech*, 7(3), pp. 127-143.

Holmes, J. N., 1983. Formant synthesizers: cascade or parallel?, *Speech Communication*, 2(4), pp. 251-273.

Hunt, A. J. and Black, A. W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database, In *Acoustics, Speech, and Signal Processing, ICASSP-96, Conference Proceedings,* 1, pp. 373-376.

Ingemann, F., 1957. Speech synthesis by rule. *Journal of the Acoustical Society of America*, 29(11), pp. 1255-1255.

Ishizaka, K., and Matsudaira, M., 1972. Fluid mechanical considerations of vocal cord vibration. *Speech Communications Research Laboratory Monograph*.

Ishizaka, K., and Flanagan, J. L., 1972. Synthesis of voiced sounds from a two?mass model of the vocal cords, *Bell System Technical Journal*, 51(6), pp. 1233-1268.

Kelly, J. L., and Lochbaum, C. C., 1962. Speech synthesis, *Proceedings of the Fourth International Congress of Acoustics*, G42, 1-4.

Kempelen, W. R. von, 1791. *Mechanismus der menschlichen Sprache*, Degen.

Klatt, D., 1982. The Klattalk text-to-speech conversion system, In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82,* 7, pp. 1589-1592.

Klatt, D. H., 1987. Review of text?to?speech conversion for English, *Journal of the Acoustical Society of America*, 82(3), pp. 737-793.

Klatt, D. H., and Klatt, L. C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America*, 87(2), pp. 820-857.

Koenig, R., 1873. I. On manometric flames, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 45(297), pp. 1-18.

Koenig, W., Dunn, H. K., and Lacy, L. Y., 1946. The sound spectrograph, *Journal of the Acoustical Society of America*, 18(1), pp. 19-49.

Koenig, W., and Ruppel, A. E., 1948. Quantitative amplitude representation in sound spectrograms. *Journal of the Acoustical Society of America*, 20(6), pp. 787-795.

Kopp, G. A., and Green, H. C., 1946. Basic phonetic principles of visible speech. *Journal of the Acoustical Society of America*, 18(1), pp. 244-245.

Kratzenstein, C.G., 1781. *Tentamen Resolvendi Problema ab Academia Scientiarum Imperiali Petroplitana ad annum 1780 Publice Problema*, Petropoli: Typis Academiae Scientiarium.

Kröger, B. J., Birkholz, P., Lowit, A., and Neuschaefer-Rube, C., 2010. Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production, *Speech motor control: New developments in basic and applied research*, Maassen, B. and Van Lieshout, P. eds., New York: Oxford University Press, pp. 23-36.

Ladefoged, P., and Broadbent, D. E., 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1), pp. 98-104.

Lawrence, W., 1953. The synthesis of speech from signals which have a low information rate, *Communication Theory*, pp. 460-469.

Liberman, A. M., Delattre, P., and Cooper, F. S., 1952. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 65, pp. 497-516. http://dx.doi.org/10.2307/1418032

Liberman, A. M., Delattre, P., Cooper, F. S., and Gerstman, L., 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants, *Psychological Monographs: General and Applied*, 68, pp. 1-13. http:/dx.doi.org/10.1037/h0093673

Liberman, A. M., 1957. Some results of research on speech perception. *Journal of the Acoustical Society of America*, 29, 117-123. http://dx.doi.org/10.1121/1.1908635

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C., 1957. The discrimination of speech sounds within and across phoneme boundaries, *Journal of Experimental Psychology*, 54, pp. 358-368. http://dx.doi.org/10.1037/h0044417

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. S., 1959. Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America*, 31(11), pp. 1490-1499.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M., 1967. Perception of the speech code, *Psychological Review*, 74, pp. 431-461. http://dx.doi.org/10.1037/h0020279

Liljencrants, J., 1985. *Speech Synthesis with a Reflection-Type Line Analog*. DS Dissertation, Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech., Stockholm, Sweden.

Lindblom, B. E., and Sundberg, J. E., 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50(4B), 1166-1179.

Loizou, P. C., Dorman, M., and Tu, Z., 1999. On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, 106(4), pp. 2097-2103.

Maeda, S., 1988. Improved articulatory model. *Journal of the Acoustical Society of America,* 84, Sup. 1, S146.

Maeda, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model, *Speech production and speech modeling*, W.L. Hardcastle, W.L. and Marcha, A., eds. Dordrecht: Kluwer Academic Publishers, pp. 131-149.

Massaro, D. W., 1998. *Perceiving talking faces: From speech perception to a behavioral principle*, Cambridge, MA: MIT Press.

Mattingly, I. G., 1974. Speech synthesis for phonetic and phonological models, *Current trends in linguistics*, Sebeok, T.A., ed., The Hague: Mouton, pp. 2451-2487.

Mermelstein, P., 1973. Articulatory model for the study of speech production, *Journal of the Acoustical Society of America*, 53(4), pp. 1070-1082.

Miller, D.C., 1909. The phonodeik, *Physical Review*, 28, p. 151.

Miller, D.C., 1916. *The Lowell lectures: The science of musical sounds*, New York: MacMillan and Co., pp. 215-262.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform pro- cessing techniques for text-to-speech synthesis using diphones, *Speech Communication*, 9, pp. 453-467.

Mrayati, M., Carré, R., and Guérin, B., 1988. Distinctive regions and modes: A new theory of speech production, *Speech Communication*, 7, pp. 257-286.

Nakata, K., and Mitsuoka, T., 1965. Phonemic transformation and control aspects of synthesis of connected speech, *Journal of the Radio Research Laboratory*, 12, pp. 171-186.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., and Goldstein, L., 2012. A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America*, 132(6), pp. 3980-3989.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. [Accessed 5 March 2018].

Pantalony, D., 2004. Seeing a voice: Rudolph Koenig's instruments for studying vowel sounds. *American Journal of Psychology*, pp. 425-442.

Patterson, R., 1845. *Proceedings of the American Philosophical Society*, 4(34), p. 222.

Pedro the Voder: A Machine That Talks, 1939. *Bell Laboratories Record*, 17(6), pp. 170-171. (no author listed).

Peterson, G. E., Wang, W. S. Y., and Sivertsen, E., 1958. Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America*, 30(8), pp. 739-742.

Pieraccini, R., 2012. *The voice in the machine*, Cambridge, MA: MIT Press.

Potter, R. K., 1945. Visible patterns of sound, *Science*, 102(2654), pp. 463-470.

Potter, R. K., and Peterson, G. E., 1948. The representation of vowels and their movements. *Journal of the Acoustical Society of America*, 20(4), pp. 528-535.

Potter, R. K., 1949. Objectives for sound portrayal. *Journal of the Acoustical Society of America*, 21(1), pp. 1-5.

Rayleigh, J. W. S., 1878. *The theory of sound, vol. II*, New York: MacMillan and Co., pp. 469-478.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D., 1981. Speech perception without traditional speech cues, *Science*, 212, pp. 947-950.

Remez, R. E., 2005. Perceptual organization of speech. *The Handbook of Speech Perception*, Pisoni, D. B., and Remez, R. E., eds., Oxford, UK: Wiley-Blackwell, 28-50.

Rosen, G., 1958. Dynamic analog speech synthesizer. *Journal of the Acoustical Society of America*, 30(3), pp. 201-209.

Rothenberg, M., Dorman, K. W., Rumm, J. C., and Theerman, P. H., 1992. 1846 letter from Joseph Henry to Henry M. Alexander, *The papers of Joseph Henry: Volume 6*, Washington, DC: Smithsonian Institution Press, pp. 359-364,

Rubin, P., Baer, T., and Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70, pp. 321-328.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R, Tiede, M., and Browman, C., 1996. CASY and extensions to the task-dynamic model, *Proceedings of the 4th Speech Production Seminar*, Grenoble, France, pp. 125-128.

Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K., 1992. ATR m-TALK speech synthesis system, *Proceedings of the International Conference on Spoken Language Processing*, pp. 483-486.

Saltzman, E., 1986. Task dynamic coordination of the speech articulators: A preliminary model, *Experimental Brain Research Series*, 15, pp. 129-144.

Saltzman, E., 1991. The task dynamic model in speech production. *Speech motor control and stuttering*, Peters, Hulstijn, and Starkweather, eds., pp. 37-52.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M., 1995. Speech recognition with primarily temporal cues, *Science*, 270, pp. 303-304.

Schroeder, M. R., 1981. Homer W. Dudley: A Tribute, *Journal of the Acoustical Society of America*, 69(4), p. 1222.

Schroeder, M. R., 1993. A brief history of synthetic speech. *Speech Communication*, 13(1-2), 231-237.

Scott, E-L., 1859. The phonoautograph, *Cosmos*, 14, p. 314.

Sivertsen, E., 1961. Segment inventories for speech synthesis, *Language and Speech*, 4(1), pp. 27-90.

Smith, Z. M., Delgutte, B., and Oxenham, A. J., 2002. Chimaeric sounds reveal dichotomies in auditory perception, *Nature*, 416, pp. 87-90.

Sondhi, M., and Schroeter, J., 1987. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7), pp. 955-967.

Steinberg, J. C., and French, N. R., 1946. The portrayal of visible speech. *Journal of the Acoustical Society of America*, 18(1), pp. 4-18.

Stevens, K. N., Kasowski, S., and Fant, C. G. M., 1953. An electrical analog of the vocal tract. *Journal of the Acoustical Society of America*, 25(4), pp. 734-742.

Stevens, K. N., and House, A. S., 1955. Development of a quantitative description of vowel articulation, *Journal of the Acoustical Society of America*, 27(3), pp. 484-493.

Stewart, J. Q., 1922. An electrical analogue of the vocal organs, *Nature*, 110(2757), pp. 311-312.

Story, B. H., 1995. *Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract*. Ph.D. Dissertation, University of Iowa.

Story, B. H., Titze, I. R., and Hoffman, E. A., 1996. Vocal tract area functions from magnetic resonance imaging, *Journal of the Acoustical Society of America*, 100(1), pp. 537-554.

Story, B.H., Titze, I.R., and Hoffman, E. A., 1998. Vocal tract area functions for an adult female speaker based on volumetric imaging, *Journal of the Acoustical Society of America*, 104(1), 471-487.

Story, B. H., 2005. A parametric model of the vocal tract area function for vowel and consonant simulation, *Journal of the Acoustical Society of America*, 117(5), pp. 3231-3254.

Story, B. H., and Bunton, K., 2010. Relation of vocal tract shape, formant transitions, and stop consonant identification, *Journal of Speech, Language, and Hearing Research*, 53, pp. 1514-1528.

Story, B.H., 2013. Phrase-level speech simulation with an airway modulation model of speech production, *Computer Speech and Language*, 27(4), pp. 989-1010.

Strube, H., 1982. Time-varying wave digital filters and vocal-tract models. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82,* 7, pp. 923-926.

Talking Machine, 1880. *London Times*, February 12. Retrieved from http://www.phonozoic.net/library.html

The Speaking Automaton, 1846. *London Times*, August 12. p. 3. Retrieved from http://www.phonozoic.net/library.html

Titze, I. R., 1973. The human vocal cords: a mathematical model. *Phonetica*, 28(3-4), pp. 129-170.

Titze, I. R., 1974. The human vocal cords: a mathematical model. *Phonetica*, 29(1-2), pp. 1-21.

Tompkins, D., 2010. How to wreck a nice beach: *The Vocoder from World War II to Hip-Hop, the machine speaks*. Brooklyn, NY: Melville House.

Trouvain, J., and Brackhane, F. 2011. Wolfgang von Kempelen's speaking machine as a an instrument for demonstration and research, *Proceedings of the 17th International Conference of Phonetic Sciences*, pp. 164-167.

Truby, H. M., 1965. Matching Speech Cineradiography with Pseudospeech. *Journal of the Acoustical Society of America*, 37(6), p. 1187.

Vatikiotis-Bateson, E., Kuratate, T., Munhall, K. G., and Yehia, H. C., 2000. The production and perception of a realistic talking face. *Proceedings of LP'98: Item order in language and speech*, Fujimura, O., Joseph, B. D., and Palek, B., eds., Prague: Karolinum Press (Charles University, 2, pp. 439-460.

Wang, W. S. Y., and Peterson, G. E., 1958. Segment inventory for speech synthesis, *Journal of the Acoustical Society of America*, 30(8), pp. 743-746.

Wheatstone, C., 1837. Reed organ-pipes, speaking machines, etc. *The scientific papers of Sir Charles Wheatstone* (1879). London: Taylor and Francis, pp. 348-367. Originally published in the London and Westminster Review, No. xi and liv., Oct. 1837.

Willis, R., 1829. On the vowel sounds, and on reed organ pipes, *Transactions of the Cambridge Philosophical Society*, 3, pp. 231-268.

Wurzbach, C., *Biographies of the Empire Austria: Containing life sketches of memorable people, who lived from 1750 to 1850 in the imperial state and in its crown lands*. 60, pp. 1856-1891.

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E., 1998. Quantitative association of vocal-tract and facial behavior, *Speech Communication*, 26(1-2), pp. 23-43.

Zen, H., Tokuda, K., and Black, A. W., 2009. Statistical parametric speech synthesis, *Speech Communication*, 51(11), pp. 1039-1064.

Zen, H., Senior, A., and Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. *Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7962-7966.