The Secrets of Speech Synthesis: From Military Communications to Apple's Siri

Mark Allen-Piccolo

Writer's Comment: In winter 2019 I was enrolled in a course on Digital Signal Processing in the Department of Electrical and Computer Engineering. For the course project, I developed a voice synthesizer app that allowed a user to select a prerecorded voice or record their own voice, spoken or sung; the app would then produce a synthesized version of that same phrase. After the project was finished, I thought I would gain a better understanding of what I had created if I tried to explain it in words.

The subject of vocal synthesis seemed an intriguing topic to write about for my translating knowledge essay for UWP 101 with the help of Professor MacArthur. Engineering is often difficult to discuss because it requires a background in math and other specialized knowledge. When I enter the workforce, however, I imagine I will need to discuss this technical work with non-engineers. Writing this essay gave me the chance to practice this essential skill. I hope that by giving ample definitions and illustrative examples, the essay will be accessible to an average reader. I also hope to demystify some of the technology that has become ubiquitous in today's society.

Instructor's Comment: One reason I love teaching UWP 101 is the variety of majors. Mark earned a B.A. in music from UC Berkeley, worked for a while, and enrolled at UC Davis to earn a B.S. in electrical engineering. When he chose early vocoders as the topic for his Translating Knowledge essay, I was mostly delighted. This topic connects to my own interdisciplinary research—the study of perfor-

mative speech, which involves sound visualization and quantitative methods—and I thought other readers would also be interested, given the increasing ubiquity of virtual agents like Alexa and Siri. But I also know how dauntingly complex speech production and audio signal analysis are, and worried a bit that Mark would have too much technical information to translate. I needn't have worried. Discussing his strong draft in office hours, I encouraged him to use some visualization and add more examples of things Siri might say to clarify the process of speech synthesis. The result is a compelling explanation of both the human voice and the basis for a technology we rely on more and more.

—Marit J. MacArthur, University Writing Program

Have you ever asked Apple's virtual assistant Siri who she is? She will respond by saying, "I'm just a humble virtual assistant." We know that she is artificial intelligence developed by the SRI International Artificial Intelligence Center and Nuance Communications, and that her voice is based on the actor Susan Bennett. She communicates using speech synthesis, a voice made from software engines that work to create the myriad words and phrases that she speaks. Siri has two main engines: prosody selection (the rhythmic and intonational aspect of speech) and unit selection (look-up table for phonemes, the distinct sounds in speech). Siri is a fairly believable artificial intelligence, the result of years of research and development. Her speech engine is quite sophisticated, but how did we get to this point? What are the elements that make up speech, and how could we possibly recreate such a complex mechanism as human speech? How can we teach artificial intelligence to recognize and interpret phrases, or answer questions? To answer these questions, we can start at the beginning, when speech synthesis was a rudimentary technology called an LPC vocoder (linear predictive coding vocoder).

The LPC vocoder was invented at Bell Laboratories in 1966. It was designed to be used in military applications as a means of communication. Imagine a voice command sent from base to an airplane, or vice versa. At that time, computing power was limited and there was a need to constrain the amount of data used in communication. Engineers thought

that if they could strip a human voice down to some essential elements, they could reduce the computation needed while still keeping the speech understandable. After all, if simply relaying information, there is no need for a signal to be high fidelity. What matters is that the command is intelligible.

The human voice is timbrally rich, containing frequencies ranging from one hundred hertz up to fifteen kilohertz (the hertz is the unit of measurement for frequency, or cycles per second—low sounds have a low hertz value—like 120 Hz—while high sounds will have a high hertz value—like 10 kHz). Our voice is made up of many overlapping frequencies, and the lowest of these frequencies—which is the most prominent and characterizes our speaking register—is called the fundamental frequency, produced by the vibration of the vocal cords. Fricative sounds—like "s", "th", "f", "k"—are made up of high frequencies. Engineers found that for a voice to be intelligible, it need only contain the fundamental and fricatives. In speech synthesis, these two components are called voiced (essentially vowels) and unvoiced (essentially consonants). In order to make an intelligible synthesizer, a voice can be limited to these two components: voiced and unvoiced. To better understand how to recreate a synthesized human voice, the engineers at Bell began by studying the physiological process of human speech.

The mechanism that produces human speech is complex but it can be simplified into some basic components. The trachea, or windpipe, carries air from our lungs to the larynx, or voice box, which holds the vocal cords (Figure 1). Air is trapped behind the closed vocal cords, building pressure. The chords open and then close again allowing puffs of air to pass. By unconsciously controlling the shape and tautness of the vocal cords, we can control pitch. The more the vocal cords vibrate—the higher the frequency of vibration—the higher the pitch. By pushing more air from our lungs, we can control loudness. The oscillations produced by the vocal cords travel into the mouth, nose cavities (sinuous), and chest where they resonate and are projected. With subtle changes to the shape of our mouth and position of our tongue (these shapes are also called formants), we can create distinct sounds, which are called phonemes. (Every word in any language is characterized by phonemes—it's like a universal standard for pronunciation).

Engineers set out to reproduce this physiological process by first

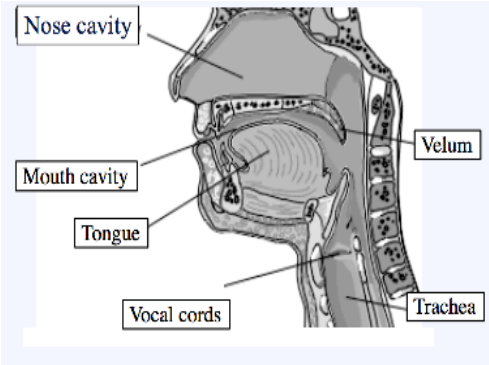

Figure 1. Vocal Apparatus. (Lathi and Ding, Modern Digital and Analog Communication Systems*, Oxford UP; 5th edition, Feb. 9, 2018.)*

creating a mathematical model. The windpipe and voice box (trachea and larynx) were modeled by a pulse generator, an electrical device that produces an electric pulse that occurs at regular intervals. It is analogous to the puff of air that is produced by the vocal cords. Like the vocal cords, the pulse generator is responsible for controlling pitch. It does this by varying frequency or periodicity, the time rate between pulses (again, measured as hertz). Increasing the amplitude of the pulse will make the sound louder. The formants (the shape the mouth makes) are modelled by a frequency selective filter, which allows some frequencies to pass while attenuating others. With this simplified mathematical model, the engineers at Bell Labs set out to construct the system that would reproduce this physiology.

The LPC vocoder is divided into two sections: analysis and synthesis. The analysis stage is where a human voice—usually pre-recorded and stored as data—is analyzed for pitch, loudness, formants, and the voiced/unvoiced decision (discussed below). Each of these components can be measured and stored as data. This data is then used in the synthesis stage to recreate a voice.

In the analysis stage, the recorded voice is separated into several categories that, as a whole, define the character of a voice: pitch, loudness, voiced/unvoiced, and formants. Each one of these components is analyzed using a unique technique. For example, cepstrum analysis is a series of mathematical steps that, when applied to a voice recording, will identify the pitch. The voiced/unvoiced decision uses a technique called zero-crossing that will determine whether the speaker uses a voiced or unvoiced sound. With the overview in mind, it can be enlightening to dive deeper into the specifics of each analysis technique.

There are currently many techniques that will determine pitch,

but the first—called cepstrum analysis—was developed in 1967 by Michael Noll at Bell Laboratories in New York. Cepstrum analysis is a good illustration of how engineers use the frequency domain to reveal otherwise obscure results. The frequency domain is a measurement of a particular frequency or group of frequencies with respect to that frequency's (or several frequencies') strength. The voice is made of myriad overlapping frequencies. By looking at the voice in the frequency domain, we can see the strength of every frequency all on one graph (Figure 2). (Another useful way to measure signals is by using a spectrograph, which combines signal strength, frequency, and time).

In speech, the vocal parameters—formants, pitch, loudness—are

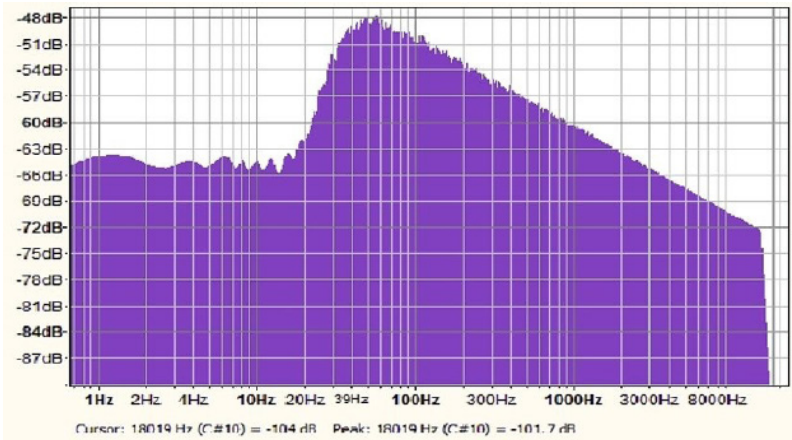

Figure 2. Measuring sound pressure level of loudspeakers (BBC Academy).

constantly shifting as we utter a sentence. For example, when Siri says, "I'm just a humble virtual assistant," the section "I'm just a humble . . ." takes about half a second to complete, but the pitch, formants, and loudness change quite drastically in that time. These changes in time require that speech analysis measures these parameters as time progresses. To do this, engineers use what is called a Short Time Fourier Transform (STFT), which is a frequency domain analysis of a 10-millisecond slice of time. In his seminal paper, "Cepstrum Pitch Determination" (1967), author Michael Noll discusses how one can use the STFT to analyze the logarithm of the frequency spectrum to garner pertinent information: "The effect of the vocal tract is to produce a low frequency ripple in the logarithmic spectrum, while the periodicity [recurring at regular intervals]

of the vocal source manifests itself as a high frequency ripple in the logarithmic spectrum." He goes on pointing out that ". . . the spectrum of the logarithmic power spectrum has a sharp peak corresponding to the high frequency source ripples in the logarithmic spectrum and a broader peak corresponding to the lower frequency formant structure in the logarithmic spectrum." In other words, using the logarithmic frequency domain, we can detect the fundamental frequency of any recorded speech; i.e., we can determine the pitch (Figure 3).

Analyzing the zero crossing is a technique that can determine when there is a voiced or unvoiced sound in the recording. A single frequency

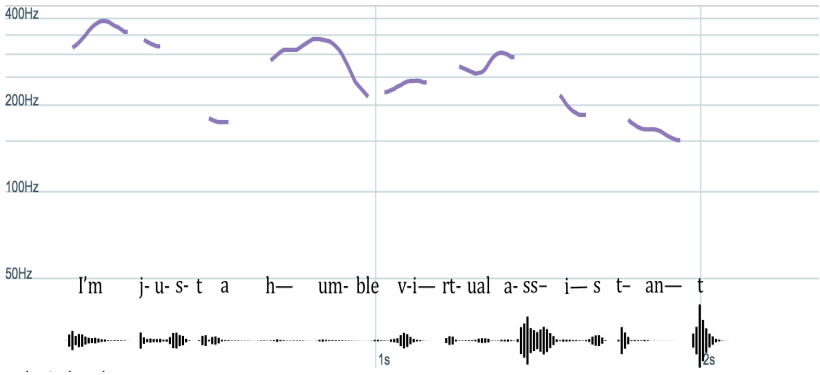

Figure 3. Pitch Contour in Drift (http://drift3.lowerquality.com/)

can be represented as a sine wave, and viewing it as such can illustrate the zero-crossing (Figure 4). Every time a signal changes directions, it crosses the "zero" axis. A zero-crossing analysis essentially counts how many times the signal crosses the "zero" in a specific amount of time. A high number of crossings corresponds to an unvoiced sound, i.e. a fricative. A low number of crossings corresponds to a voiced sound, i.e. pitched vowels. When Siri says "I'm just a humble virtual assistant," the section "I'm ju—" will show up as a lower rate of zero crossings; whereas, "-st" (of "just") will show up as a higher rate of zero-crossings. The next part of the phrase ". . . a humble virtual a—" will again have a lower number of zero-crossings, while "—ssist—" will have a higher number; and so on.

Each analysis technique produces a set of values, called coefficients, that are used to weight the prominence of vocal parameters: pitch, loudness, formants, and voiced/unvoiced decision. These coefficients are then passed to the synthesis stage where they are used to resynthesize

the voice. For example, when Siri says, "I'm just a humble virtual assistant," the contraction "I'm" changes in pitch from about 300 hertz to 400 hertz and then back down again (Figure 3). Each incremental change of pitch in this word is represented as a value between 0 and 1. When the pitch reaches 400 hertz it will acquire a high value, like 0.89. When the pitch decreases back to 300 hertz, it will acquire a lower value like 0.70. These values can then be used in the synthesis stage to reconstruct the pitch. The voiced/unvoiced decision is stored as a binary; a "0" for voiced and "1" for unvoiced. When these values are passed to the synthesis stage, they are used like a switch that can either select a voiced sound or the unvoiced sound. With each component of the voice characterized by a set of values, we can synthesize the voice from scratch.

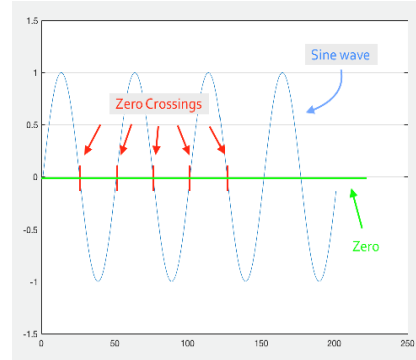

Figure 4. Zero Crossings in Matlab (created by the author).

The unvoiced sounds (the fricatives) are reproduced by a burst of white noise. White noise is the combination of all frequencies sounding simultaneously. White noise is equal energy at every frequency. To the human ear, it is not discernable as a pitch; it just sounds like static (the sound of a terrestrial radio between stations). White noise reproduces the fricative sounds particularly well.

The voiced sounds, on the other hand, are produced by a pulse generator. The coefficients that were produced when pitch, loudness, and formants were analyzed are combined with a pulse train (a series of pulses emanating from the pulse generator that occur at regular intervals—for example 16,000 pulses per second). Together, they regenerate the pitch, loudness, and formants of the original recording.

While these early synthesized voices were intelligible, they would never be mistaken for a human speaking. They are low-fidelity. Since the time the LPC vocoder was first created, technology has improved drastically. Initially, low computing power limited the fidelity of the synthesized voice. Since then, advancements in data storage and processing have allowed for refinement of the vocoder process, leading to higher fidelity and, ultimately, a more believable voice, like Siri's. The original

technologies are still used, however, but researchers found that vocoders work best when the best elements of each technique are cherry-picked and incorporated with newer technologies. In a paper published in 2018, researchers at the Department of Signal Processing and Acoustics at Aalto University tested various combinations of vocoder techniques, analyzing which combinations produced the best results. The authors found that "the choice of the voice has a profound impact on the overall quality of the vocoder-generated voice, and the best vocoder for each voice can vary case by case. . . . [In] future research, the integration of these approaches could be beneficial."

Another study published in 2015, "Incremental Syllable-Context Phonetic Vocoding" explores novel approaches to the ways in which humans understand words and phrases. The authors theorized that "Humans . . . 'encode' speech [in] real-time and in an incremental fashion, i.e., encoded speech depends only on current and past/already-uttered speech and not future/to-be-uttered speech (similar to causality in digital signal processing theory)." They were theorizing that the cognitive process of speech is similar to attributes of DSP (digital signal processing). Digital signal processing is a subfield in engineering and computer science devoted to processing digital information. A digital signal is essentially a signal that has been sampled and stored in computer memory. Cernak is explaining how certain properties in digital signal processing, like causality, are very similar to the way that the human brain recognizes meaning in words. Causality characterizes a type of signal whose information depends solely on current or past information, not future information. For example, when you ask Siri, "Who are you?", her speech recognition software can only process "Who," then "Who are," then "Who are you?" She has no idea what word will come after "Who" until it happens. This is similar to humans (although, truth be told, humans do have the ability to anticipate what might be said before it is said). Since the 1960s when the vocoder was first developed, digital technology has unlocked otherwise impossible analysis tools. Harnessing advancements like modern day computing power, memory, and sophisticated digital systems has brought new possibilities for vocoders. Incorporating these ideas into technology is now possible with neural networks and machine learning.

The advent of machine learning, under the umbrella of artificial

intelligence, is the new frontier for vocoder designs. It is what makes Siri work. Machine learning is a process where a machine is fed data, and through feedback by the designer, the machine can be trained to listen to a command and carry out a specific task. For example, SIRI can learn to process the phrase "What is the weather today?" as a series of sounds that, when combined, instruct the program to search for the daily weather conditions. This process is often carried out using a neural network. A neural network, named such because it is reminiscent of the network of human synapses, is a complex array of paths that allow sophisticated processing of various system inputs. The system is programmed (either by a programmer or through machine learning) to assign the input a fractional number between zero and one. This number indicates how much the input is similar to a particular desired output. For example, the phrase "Siri, what is the weather today?" may score a 0.99 towards the action of obtaining daily weather conditions. On the other hand, the command "Siri, play the band Weather Report" may score 0.6. The program will ultimately obtain the weather because it has the higher score of 0.99. In other words, a neural network uses weighting to carry out some task.

In the 2018 paper "Speech Sound Classification and Estimation of Optimal Order of LPC Using Neural Networks," the authors explain how the advent of neural networks offer exciting possibilities with vocoders: "Formulating an adequate mathematical expression to closely estimate the optimal order hence poses a huge challenge. Neural networks offer a solution to this problem. Since the LPC order has high correlation to the formants . . . using the spectral data would be adequate to train the network to predict the LPC order with high accuracy." The order corresponds to how much bandwidth—or frequency range—the vocoder has; a higher order is beneficial because the vocoder will be higher fidelity, but at the cost of higher computational power. Thus, they posit that neural networks can be trained to find an optimal order, striking a balance of fidelity and computational power.

Using neural networks and machine learning to refine speech synthesis has paved the way for many useful tools, like Siri. This technology is still young, and will only get better as computer scientists, linguists, and neuroscientists come up with novel approaches. Although Siri's speech is quite believable, there are times when she still stumbles on either pronunciation or speech recognition. These are the anomalies

that remind us that, yes, she still has a way to go. For example, there is a lag between when you ask Siri a question and the time it takes her to respond. Although short, the pause gives away that she is processing data in her speech recognition engine. As algorithms get faster, perhaps with the aid of sophisticated neural networks, this delay could become non-existent. Another problem is that Siri answers queries in a cycle of phrases: if you ask her a question, she will answer; if you ask her the same question, she will provide a different answer; if you ask her again, she may repeat the first answer again. In other words, she has limited available responses to the same question. However, with subtle shifts to her pitch, she could, for example, start to express frustration that we keep asking her the same question. Whether that is something customers want in a virtual assistant is a different matter.

Works Cited

Airaksinen, M., et al. "A Comparison Between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658-1670, Sept. 2018. DOI: 10.1109/TASLP.2018.2835720

Arun Sankar, M. S. Aiswariya, M., et al. "Speech Sound Classification and Estimation of Optimal Order of LPC Using Neural Networks." ICVISP 2018 Proceedings of the 2nd International Conference on Vision, Image and Signal Processing, Article No. 35, (no pages). DOI: 10.1145/3271553.3271611

Cernak, M., et al. "Incremental Syllable-Context Phonetic Vocoding." IEEE/ACM *Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, June 2015, pp. 1019-1030. DOI: 10.1109/TASLP.2015.2418577

Noll, A.M. "Cepstrum Pitch Determination." *The Journal of the Acoustical Society of America* 41, 293-309 (1967); View Table of Contents: https://asa.scitation.org/toc/jas/41/2. DOI: 10.1121/1.1910339