Speech Signal Analysis

By:
Don Johnson

Speech Signal Analysis

By:
Don Johnson

Online:
< http://cnx.org/content/col10476/1.3/ >

CONNEXIONS

Rice University, Houston, Texas

Collection structure revised: December 5, 2007
PDF generated: November 22, 2011
For copyright and attribution information for the modules contained in this collection, see p. 18.

Table of Contents

Chapter 1

Modeling the Speech Signal[1]

Vocal Tract

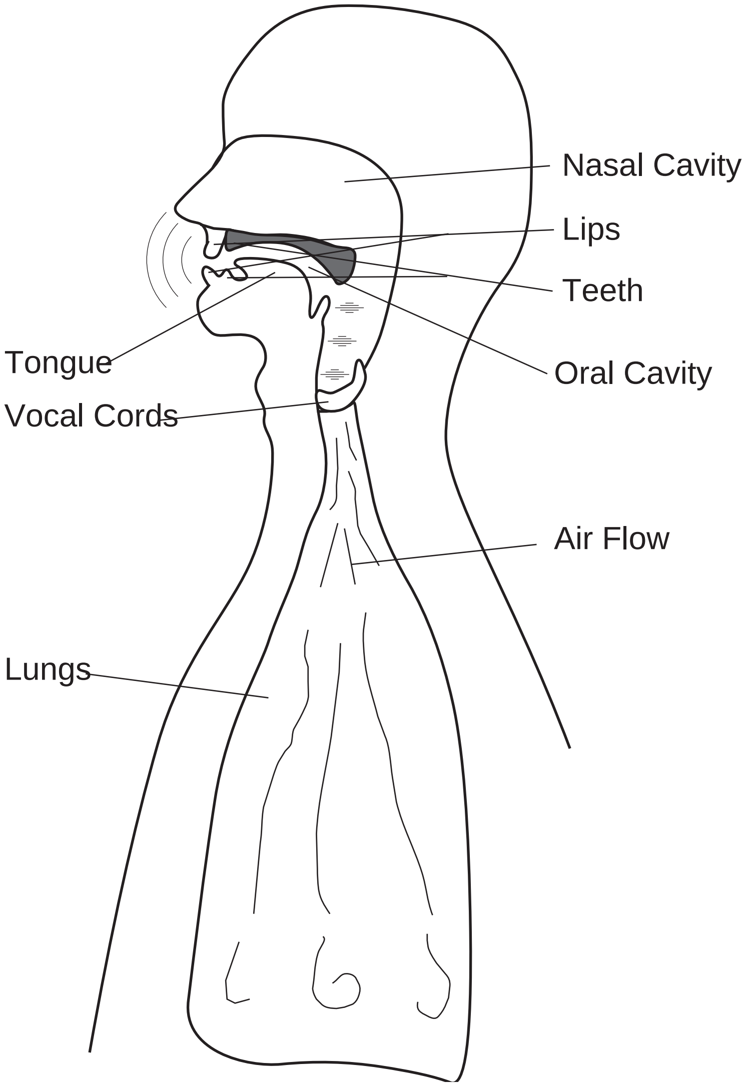


Figure 1.1: The vocal tract is shown in cross-section. Air pressure produced by the lungs forces air through the vocal cords that, when under tension, produce puffs of air that excite resonances in the vocal and nasal cavities. What are not shown are the brain and the musculature that control the entire speech production process.

Model of the Vocal Tract

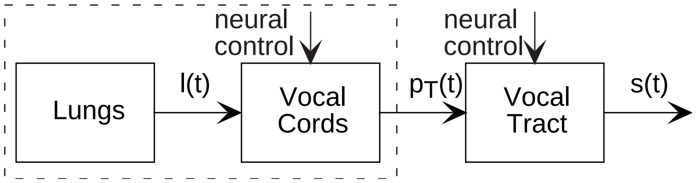


Figure 1.2: The systems model for the vocal tract. The signals $l(t)$, $p_T(t)$, and $s(t)$ are the air pressure provided by the lungs, the periodic pulse output provided by the vocal cords, and the speech output respectively. Control signals from the brain are shown as entering the systems from the top. Clearly, these come from the same source, but for modeling purposes we describe them separately since they control different aspects of the speech signal.

The information contained in the spoken word is conveyed by the speech signal. Because we shall analyze several speech transmission and processing schemes, we need to understand the speech signal's structure – what's special about the speech signal – and how we can describe and **model** speech production. This modeling effort consists of finding a system's description of how relatively unstructured signals, arising from simple sources, are given structure by passing them through an interconnection of systems to yield speech. For speech and for many other situations, system choice is governed by the physics underlying the actual production process. Because the fundamental equation of acoustics – the wave equation – applies here and is linear, we can use linear systems in our model with a fair amount of accuracy. The naturalness of linear system models for speech does not extend to other situations. In many cases, the underlying mathematics governed by the physics, biology, and/or chemistry of the problem are nonlinear, leaving linear systems models as approximations. Nonlinear models are far more difficult at the current state of knowledge to understand, and information engineers frequently prefer linear models because they provide a greater level of comfort, but not necessarily a sufficient level of accuracy.

Figure 1.1 (Vocal Tract) shows the actual speech production system and Figure 1.2 (Model of the Vocal Tract) shows the model speech production system. The characteristics of the model depends on whether you are saying a vowel or a consonant. We concentrate first on the vowel production mechanism. When the vocal cords are placed under tension by the surrounding musculature, air pressure from the lungs causes the vocal cords to vibrate. To visualize this effect, take a rubber band and hold it in front of your lips. If held open when you blow through it, the air passes through more or less freely; this situation corresponds to "breathing mode". If held tautly and close together, blowing through the opening causes the sides of the rubber band to vibrate. This effect works best with a wide rubber band. You can imagine what the airflow is like on the opposite side of the rubber band or the vocal cords. Your lung power is the simple source referred to earlier; it can be modeled as a constant supply of air pressure. The vocal cords respond to this input by vibrating, which means the output of this system is some periodic function.

Exercise 1.1 ***(Solution on p. 10.)***

Note that the vocal cord system takes a constant input and produces a periodic airflow that corresponds to its output signal. Is this system linear or nonlinear? Justify your answer.

Singers modify vocal cord tension to change the pitch to produce the desired musical note. Vocal cord tension is governed by a control input to the musculature; in system's models we represent control inputs as signals coming into the top or bottom of the system. Certainly in the case of speech and in many other cases as well, it is the control input that carries information, impressing it on the system's output. The change of signal structure resulting from varying the control input enables information to be conveyed by the signal, a process generically known as **modulation**. In singing, musicality is largely conveyed by pitch; in western speech, pitch is much less important. A sentence can be read in a monotone fashion without completely destroying the information expressed by the sentence. However, the difference between a statement and a question is frequently expressed by pitch changes. For example, note the sound differences between "Let's go to the park." and "Let's go to the park?";

For some consonants, the vocal cords vibrate just as in vowels. For example, the so-called nasal sounds "n" and "m" have this property. For others, the vocal cords do not produce a periodic output. Going back to mechanism, when consonants such as "f" are produced, the vocal cords are placed under much less tension, which results in turbulent flow. The resulting output airflow is quite erratic, so much so that we describe it as being **noise**. We define noise carefully later when we delve into communication problems.

The vocal cords' periodic output can be well described by the periodic pulse train $p_T(t)$ as shown in the periodic pulse signal[2], with T denoting the pitch period. The spectrum of this signal[3] contains harmonics of the frequency $\frac{1}{T}$, what is known as the **pitch frequency** or the **fundamental frequency** F0. The primary difference between adult male and female/prepubescent speech is pitch. Before puberty, pitch frequency for normal speech ranges between 150-400 Hz for both males and females. After puberty, the vocal cords of males undergo a physical change, which has the effect of lowering their pitch frequency to the range 80-160 Hz. If we could examine the vocal cord output, we could probably discern whether the speaker was male or female. This difference is also readily apparent in the speech signal itself.

To simplify our speech modeling effort, we shall assume that the pitch period is constant. With this simplification, we collapse the vocal-cord-lung system as a simple source that produces the periodic pulse signal (Figure 1.2 (Model of the Vocal Tract)). The sound pressure signal thus produced enters the mouth behind the tongue, creates acoustic disturbances, and exits primarily through the lips and to some extent through the nose. Speech specialists tend to name the mouth, tongue, teeth, lips, and nasal cavity the **vocal tract**. The physics governing the sound disturbances produced in the vocal tract and those of an organ pipe are quite similar. Whereas the organ pipe has the simple physical structure of a straight tube, the cross-section of the vocal tract "tube" varies along its length because of the positions of the tongue, teeth, and lips. It is these positions that are controlled by the brain to produce the vowel sounds. Spreading the lips, bringing the teeth together, and bringing the tongue toward the front portion of the roof of the mouth produces the sound "ee." Rounding the lips, spreading the teeth, and positioning the tongue toward the back of the oral cavity produces the sound "oh." These variations result in a linear, time-invariant system that has a frequency response typified by several peaks, as shown in Figure 1.3 (Speech Spectrum).

[2] "Complex Fourier Series", Figure 1 <http://cnx.org/content/m0042/latest/#pps>
[3] "Complex Fourier Series", (9) <http://cnx.org/content/m0042/latest/#pulsespec>

Speech Spectrum

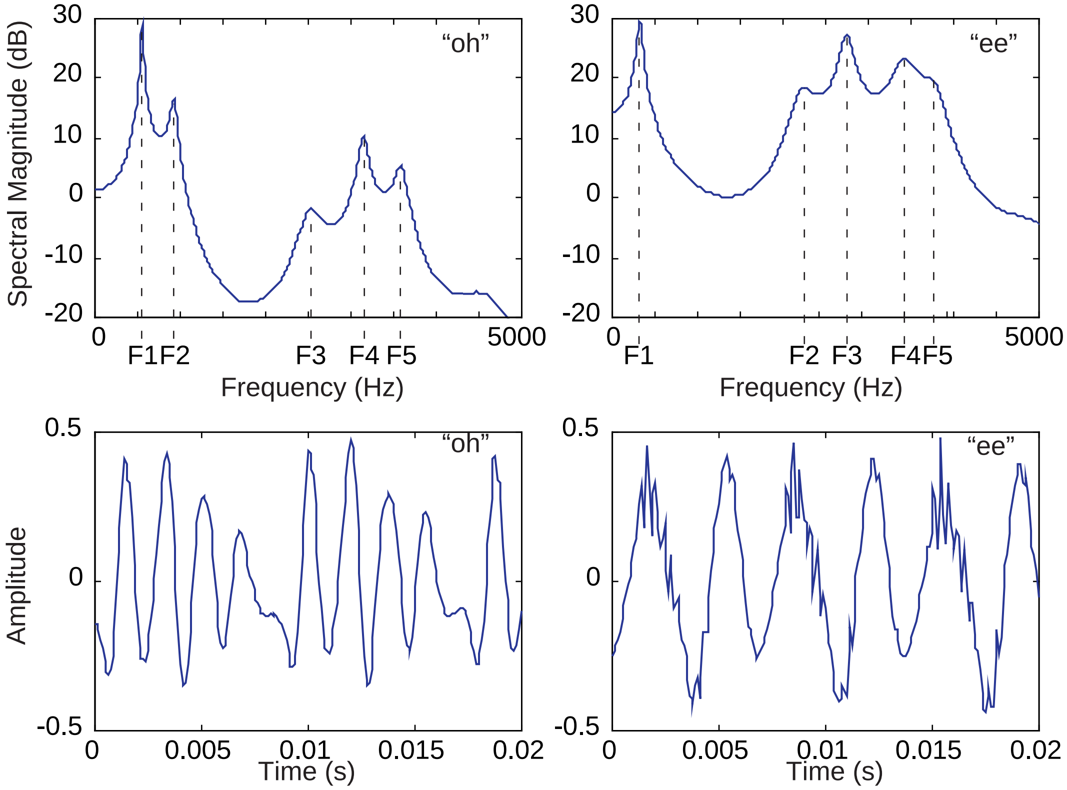


Figure 1.3: The ideal frequency response of the vocal tract as it produces the sounds "oh" and "ee" are shown on the top left and top right, respectively. The spectral peaks are known as formants, and are numbered consecutively from low to high frequency. The bottom plots show speech waveforms corresponding to these sounds.

These peaks are known as **formants**. Thus, speech signal processors would say that the sound "oh" has a higher first formant frequency than the sound "ee," with F2 being much higher during "ee." F2 and F3 (the second and third formants) have more energy in "ee" than in "oh." Rather than serving as a filter, rejecting high or low frequencies, the vocal tract serves to **shape** the spectrum of the vocal cords. In the time domain, we have a periodic signal, the pitch, serving as the input to a linear system. We know that the output—the speech signal we utter and that is heard by others and ourselves—will also be periodic. Example time-domain speech signals are shown in Figure 1.3 (Speech Spectrum), where the periodicity is quite apparent.

Exercise 1.2 ***(Solution on p. 10.)***

From the waveform plots shown in Figure 1.3 (Speech Spectrum), determine the pitch period and the pitch frequency.

Since speech signals are periodic, speech has a Fourier series representation given by a linear circuit's response to a periodic signal[4]. Because the acoustics of the vocal tract are linear, we know that the spectrum of the

[4] "Filtering Periodic Signals", (1) <http://cnx.org/content/m0044/latest/#output>

output equals the product of the pitch signal's spectrum and the vocal tract's frequency response. We thus obtain the **fundamental model of speech production**.

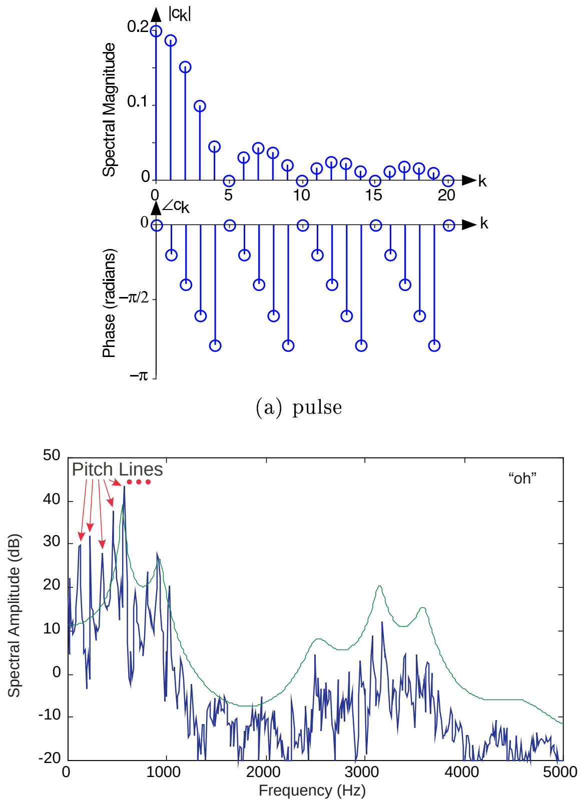
$$S(f) = P_T(f)\, H_V(f) \tag{1.1}$$

Here, $H_V(f)$ is the transfer function of the vocal tract system. The Fourier series for the vocal cords' output, derived in this equation[5], is

$$c_k = Ae^{-\frac{j\pi k\Delta}{T}} \frac{\sin\left(\frac{\pi k\Delta}{T}\right)}{\pi k} \tag{1.2}$$

and is plotted on the top in Figure 1.4 (voice spectrum). If we had, for example, a male speaker with about a 110 Hz pitch ($T \simeq 9.1$ms) saying the vowel "oh", the spectrum of his speech **predicted by our model** is shown in Figure 1.4(b) (voice spectrum).

voice spectrum



(a) pulse

(b) voice spectrum

Figure 1.4: The vocal tract's transfer function, shown as the thin, smooth line, is superimposed on the spectrum of actual male speech corresponding to the sound "oh." The pitch lines corresponding to harmonics of the pitch frequency are indicated. (a) The vocal cords' output spectrum $P_T(f)$. (b) The vocal tract's transfer function, $H_V(f)$ and the speech spectrum.

[5] "Complex Fourier Series" <http://cnx.org/content/m0042/latest/#expar1>

The model spectrum idealizes the measured spectrum, and captures all the important features. The measured spectrum certainly demonstrates what are known as **pitch lines**, and we realize from our model that they are due to the vocal cord's periodic excitation of the vocal tract. The vocal tract's shaping of the line spectrum is clearly evident, but difficult to discern exactly, especially at the higher frequencies. The model transfer function for the vocal tract makes the formants much more readily evident.

Exercise 1.3 ***(Solution on p. 10.)***

The Fourier series coefficients for speech are related to the vocal tract's transfer function only at the frequencies $\frac{k}{T}$, $k \in \{1, 2, \dots\}$; see previous result[6]. Would male or female speech tend to have a more clearly identifiable formant structure when its spectrum is computed? Consider, for example, how the spectrum shown on the right in Figure 1.4 (voice spectrum) would change if the pitch were twice as high ($\approx (300)\,\mathrm{Hz}$).

When we speak, pitch and the vocal tract's transfer function are not static; they change according to their control signals to produce speech. Engineers typically display how the speech spectrum changes over time with what is known as a spectrogram (Chapter 2) Figure 1.5 (spectrogram). Note how the line spectrum, which indicates how the pitch changes, is visible during the vowels, but not during the consonants (like the **ce** in "Rice").

[6] "Complex Fourier Series", (9) <http://cnx.org/content/m0042/latest/#pulsespec>

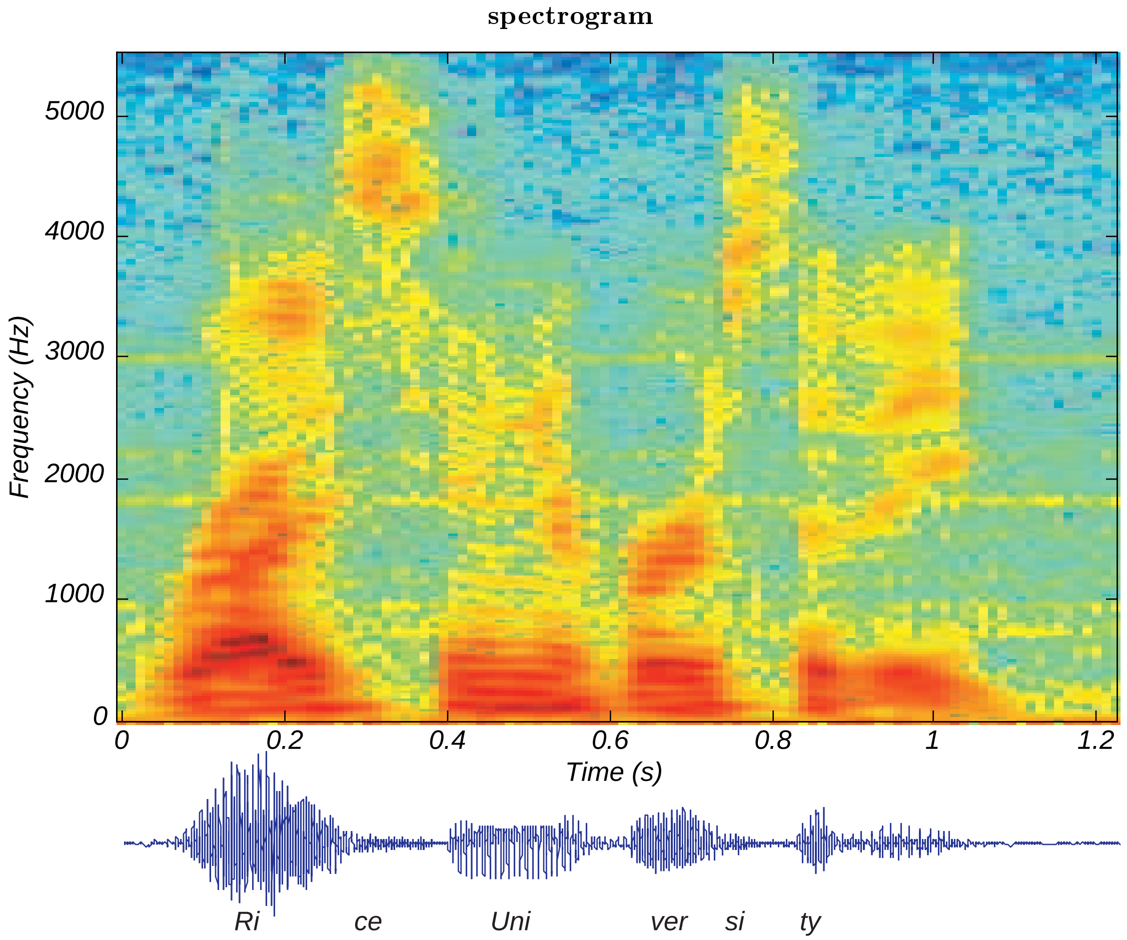


Figure 1.5: Displayed is the spectrogram of the author saying "Rice University." Blue indicates low energy portion of the spectrum, with red indicating the most energetic portions. Below the spectrogram is the time-domain speech signal, where the periodicities can be seen.

The fundamental model for speech indicates how engineers use the physics underlying the signal generation process and exploit its structure to produce a systems model that suppresses the physics while emphasizing how the signal is "constructed." From everyday life, we know that speech contains a wealth of information. We want to determine how to transmit and receive it. Efficient and effective speech transmission requires us to know the signal's properties and its structure (as expressed by the fundamental model of speech production). We see from Figure 1.5 (spectrogram), for example, that speech contains significant energy from zero frequency up to around 5 kHz.

Effective speech transmission systems must be able to cope with signals having this bandwidth. It is interesting that one system that does **not** support this 5 kHz bandwidth is the telephone: Telephone systems act like a **bandpass filter** passing energy between about 200 Hz and 3.2 kHz. The most important consequence of this filtering is the removal of high frequency energy. In our sample utterance, the "ce" sound in "Rice"" contains most of its energy above 3.2 kHz; this filtering effect is why it is extremely difficult to distinguish the sounds "s" and "f" over the telephone. Try this yourself: Call a friend and determine if they can distinguish between the words "six" and "fix". If you say these words in isolation so that no context provides a hint about which word you are saying, your friend will not be able to tell them apart. Radio does support this bandwidth (see more about AM and FM radio systems[7]).

Efficient speech transmission systems exploit the speech signal's special structure: What makes speech speech? You can conjure many signals that span the same frequencies as speech—car engine sounds, violin music, dog barks—but don't sound at all like speech. We shall learn later that transmission of **any** 5 kHz bandwidth signal requires about 80 kbps (thousands of bits per second) to transmit digitally. **Speech** signals can be transmitted using less than 1 kbps because of its special structure. To reduce the "digital bandwidth" so drastically means that engineers spent many years to develop signal processing and coding methods that could capture the special characteristics of speech without destroying how it sounds. If you used a speech transmission system to send a violin sound, it would arrive horribly distorted; speech transmitted the same way would sound fine.

Exploiting the special structure of speech requires going beyond the capabilities of analog signal processing systems. Many speech transmission systems work by finding the speaker's pitch and the formant frequencies. Fundamentally, we need to do more than filtering to determine the speech signal's structure; we need to manipulate signals in more ways than are possible with analog systems. Such flexibility is achievable (but not without some loss) with programmable **digital** systems.

[7] "Modulated Communication" <http://cnx.org/content/m0518/latest/>

Solutions to Exercises in Chapter 1

Solution to Exercise 1.1 (p. 3)

If the glottis were linear, a constant input (a zero-frequency sinusoid) should yield a constant output. The periodic output indicates nonlinear behavior.

Solution to Exercise 1.2 (p. 5)

In the bottom-left panel, the period is about 0.009 s, which equals a frequency of 111 Hz. The bottom-right panel has a period of about 0.0065 s, a frequency of 154 Hz.

Solution to Exercise 1.3 (p. 7)

Because males have a lower pitch frequency, the spacing between spectral lines is smaller. This closer spacing more accurately reveals the formant structure. Doubling the pitch frequency to 300 Hz for Figure 1.4 (voice spectrum) would amount to removing every other spectral line.

Chapter 2

Spectrograms[1]

We know how to acquire analog signals for digital processing (pre-filtering[2], sampling[3], and A/D conversion[4]) and to compute spectra of discrete-time signals (using the FFT algorithm[5]), let's put these various components together to learn how the spectrogram shown in Figure 2.1 (Speech Spectrogram), which is used to analyze speech (Chapter 1), is calculated. The speech was sampled at a rate of 11.025 kHz and passed through a 16-bit A/D converter.

> POINT OF INTEREST: Music compact discs (CDs) encode their signals at a sampling rate of 44.1 kHz. We'll learn the rationale for this number later. The 11.025 kHz sampling rate for the speech is 1/4 of the CD sampling rate, and was the lowest available sampling rate commensurate with speech signal bandwidths available on my computer.

Exercise 2.1 ***(Solution on p. 16.)***

Looking at Figure 2.1 (Speech Spectrogram) the signal lasted a little over 1.2 seconds. How long was the sampled signal (in terms of samples)? What was the datarate during the sampling process in bps (bits per second)? Assuming the computer storage is organized in terms of bytes (8-bit quantities), how many bytes of computer memory does the speech consume?

[1] This content is available online at <http://cnx.org/content/m0505/2.20/>.

[2] "The Sampling Theorem" <http://cnx.org/content/m0050/latest/>

[3] "The Sampling Theorem" <http://cnx.org/content/m0050/latest/>

[4] "Amplitude Quantization" <http://cnx.org/content/m0051/latest/>

[5] "Fast Fourier Transform (FFT)" <http://cnx.org/content/m10250/latest/>

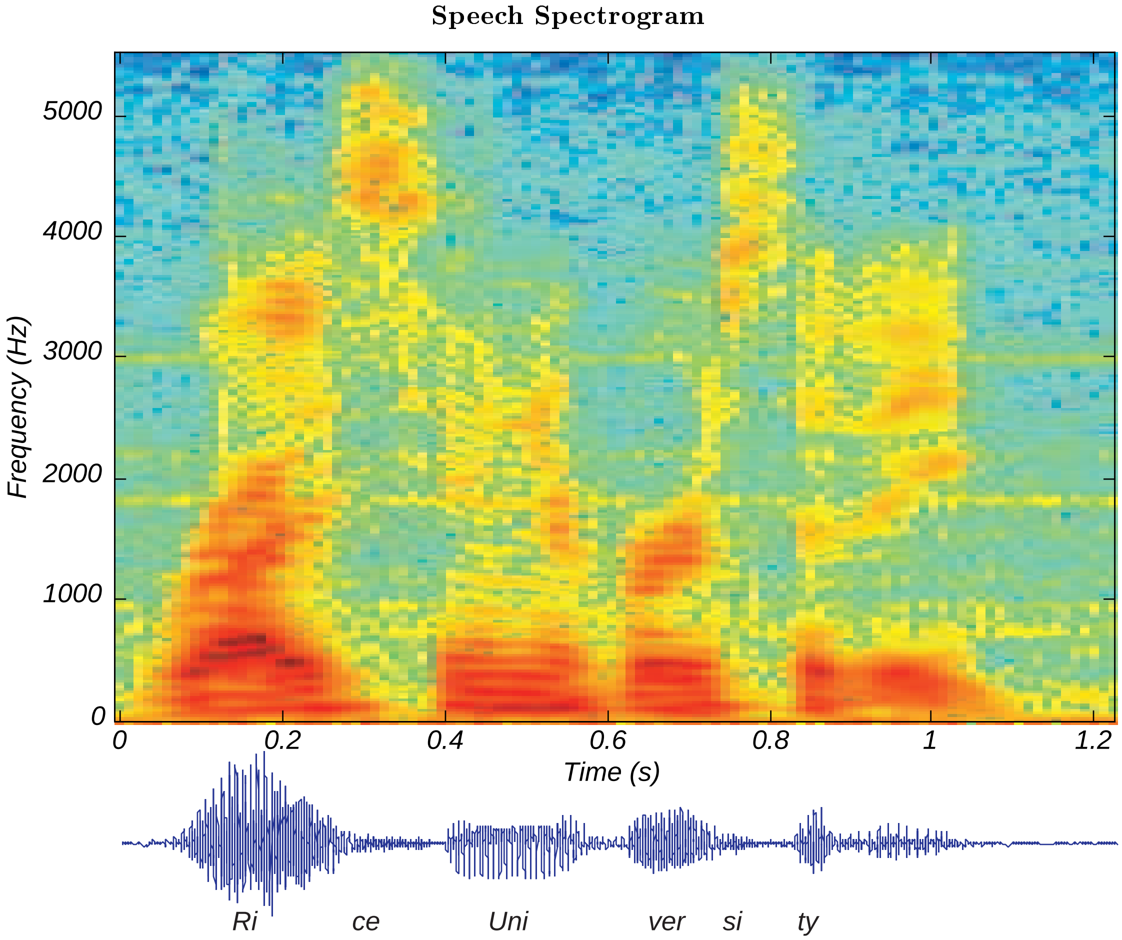


Figure 2.1

The resulting discrete-time signal, shown in the bottom of Figure 2.1 (Speech Spectrogram), clearly changes its character with time. To display these spectral changes, the long signal was sectioned into **frames**: comparatively short, contiguous groups of samples. Conceptually, a Fourier transform of each frame is calculated using the FFT. Each frame is not so long that significant signal variations are retained within a frame, but not so short that we lose the signal's spectral character. Roughly speaking, the speech signal's spectrum is evaluated over successive time segments and stacked side by side so that the x-axis corresponds to time and the y-axis frequency, with color indicating the spectral amplitude.

An important detail emerges when we examine each framed signal (Figure 2.2 (Spectrogram Hanning vs. Rectangular)).

Spectrogram Hanning vs. Rectangular

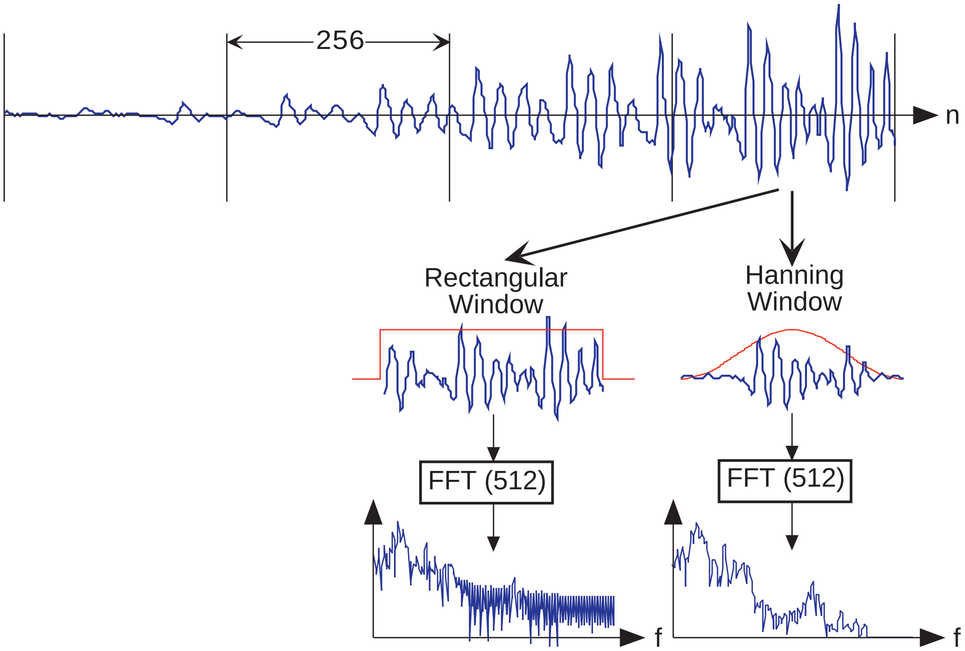


Figure 2.2: The top waveform is a segment 1024 samples long taken from the beginning of the "Rice University" phrase. Computing Figure 2.1 (Speech Spectrogram) involved creating frames, here demarked by the vertical lines, that were 256 samples long and finding the spectrum of each. If a rectangular window is applied (corresponding to extracting a frame from the signal), oscillations appear in the spectrum (middle of bottom row). Applying a Hanning window gracefully tapers the signal toward frame edges, thereby yielding a more accurate computation of the signal's spectrum at that moment of time.

At the frame's edges, the signal may change very abruptly, a feature not present in the original signal. A transform of such a segment reveals a curious oscillation in the spectrum, an artifact directly related to this sharp amplitude change. A better way to frame signals for spectrograms is to apply a **window**: Shape the signal values within a frame so that the signal decays gracefully as it nears the edges. This shaping is accomplished by multiplying the framed signal by the sequence $w(n)$. In sectioning the signal, we essentially applied a rectangular window: $w(n) = 1$, $0 \leq n \leq N-1$. A much more graceful window is the **Hanning window**; it has the cosine shape $w(n) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N}\right)\right)$. As shown in Figure 2.2 (Spectrogram Hanning vs. Rectangular), this shaping greatly reduces spurious oscillations in each frame's spectrum. Considering the spectrum of the Hanning windowed frame, we find that the oscillations resulting from applying the rectangular window obscured a formant (the one located at a little more than half the Nyquist frequency).

Exercise 2.2 ***(Solution on p. 16.)***

What might be the source of these oscillations? To gain some insight, what is the length- $2N$ discrete Fourier transform of a length-N pulse? The pulse emulates the rectangular window, and certainly has edges. Compare your answer with the length- $2N$ transform of a length- N Hanning window.

Non-overlapping windows

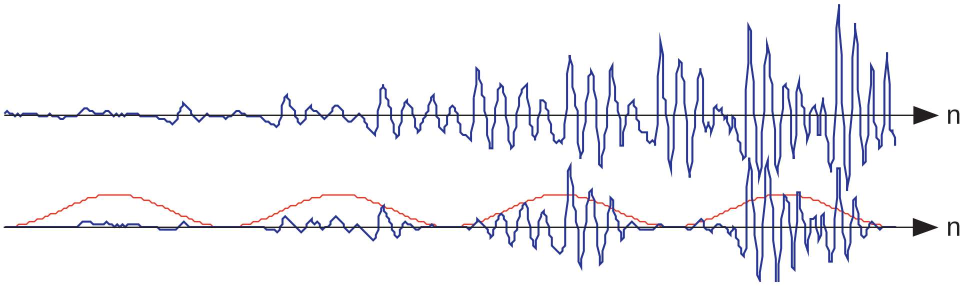


Figure 2.3: In comparison with the original speech segment shown in the upper plot, the non-overlapped Hanning windowed version shown below it is very ragged. Clearly, spectral information extracted from the bottom plot could well miss important features present in the original.

If you examine the windowed signal sections in sequence to examine windowing's affect on signal amplitude, we see that we have managed to amplitude-modulate the signal with the periodically repeated window (Figure 2.3 (Non-overlapping windows)). To alleviate this problem, frames are overlapped (typically by half a frame duration). This solution requires more Fourier transform calculations than needed by rectangular windowing, but the spectra are much better behaved and spectral changes are much better captured.

The speech signal, such as shown in the speech spectrogram (Figure 2.1: Speech Spectrogram), is sectioned into overlapping, equal-length frames, with a Hanning window applied to each frame. The spectra of each of these is calculated, and displayed in spectrograms with frequency extending vertically, window time location running horizontally, and spectral magnitude color-coded. Figure 2.4 (Overlapping windows for computing spectrograms) illustrates these computations.

Overlapping windows for computing spectrograms

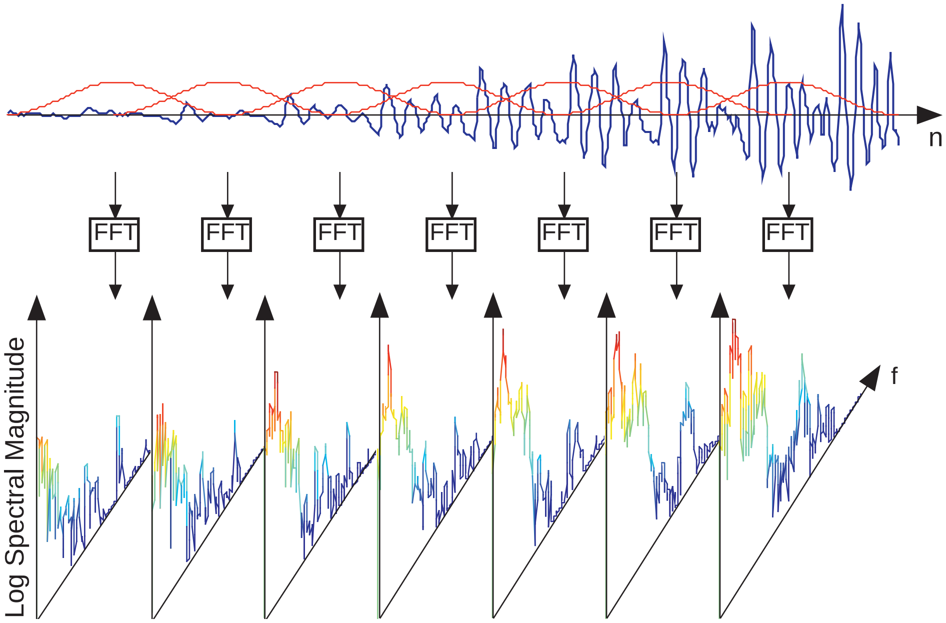


Figure 2.4: The original speech segment and the sequence of overlapping Hanning windows applied to it are shown in the upper portion. Frames were 256 samples long and a Hanning window was applied with a half-frame overlap. A length-512 FFT of each frame was computed, with the magnitude of the first 257 FFT values displayed vertically, with spectral amplitude values color-coded.

Exercise 2.3 ***(Solution on p. 16.)***

Why the specific values of 256 for N and 512 for K? Another issue is how was the length-512 transform of each length-256 windowed frame computed?

Solutions to Exercises in Chapter 2

Solution to Exercise 2.1 (p. 11)

Number of samples equals $1.2 \times 11025 = 13230$. The datarate is $11025 \times 16 = 176.4$ kbps. The storage required would be 26460 bytes.

Solution to Exercise 2.2 (p. 13)

The oscillations are due to the boxcar window's Fourier transform, which equals the sinc function.

Solution to Exercise 2.3 (p. 15)

These numbers are powers-of-two, and the FFT algorithm can be exploited with these lengths. To compute a longer transform than the input signal's duration, we simply zero-pad the signal.

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Speech Signal Analysis*
Edited by: Don Johnson
URL: http://cnx.org/content/col10476/1.3/
License: http://creativecommons.org/licenses/by/2.0/

Module: "Modeling the Speech Signal"
By: Don Johnson
URL: http://cnx.org/content/m0049/2.29/
Pages: 2-10
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/3.0/

Module: "Spectrograms"
By: Don Johnson
URL: http://cnx.org/content/m0505/2.20/
Pages: 11-16
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Speech Signal Analysis
An overview of the speech maodel and spectrogram analysis of speech signals.