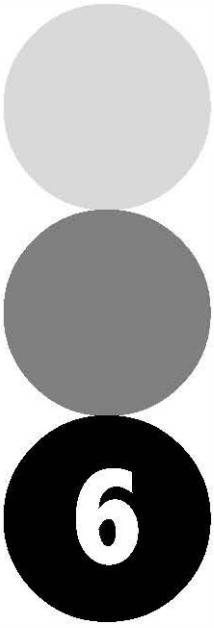

6 Spectral Modeling and Additive Synthesis

6.0 Introduction

In prior chapters we discovered (over and again) how important sine waves are for modeling physical systems (Chapter 4; modes), and for mathematically representing signals and shapes (Chapter 5; Fourier analysis). We finished in Chapter 5 with the notion of a frequency spectrum, computed from a signal by means of the Fourier transform. In this chapter we will look at different types of spectra, noting certain properties and perceptual attributes. We will then revisit the additive synthesis algorithm with an eye toward improving the representation and parametric flexibility for synthesis.

6.1 Types of Spectra

6.1.1 Harmonic

In Chapter 5 we motivated the notion of the Fourier series for representing periodic signals. We saw that periodic (or quasi-periodic) signals exhibit a harmonic spectrum. In general, periodic signals with harmonic spectra are perceived with a strong sense of pitch (at least between 100 and 4000 Hz fundamental frequency). When we think of sounds that give us a strong perception of pitch, we often think of musical tones. The sounds produced by

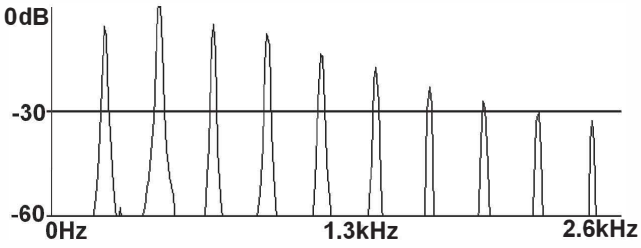

Figure 6.1. Spectrum of trumpet tone steady-state (same tone as shown in Figure 2.1).

trumpets, violins, clarinets, etc., are examples of such harmonic spectra. Figure 6.1 shows the steady-state spectrum of a trumpet tone (from the trumpet tone of Figure 2.1).

Another important set of periodic sounds that have harmonic spectra are the voiced sounds of speech. Figure 6.2 shows the spectrum of the voiced vowel sound "ahh" (as in father), and Figure 6.3 shows the spectrum of the voiced vowel sound "eee" (as in beet).

((22))

Speech spectra have important features called *formants* which are the three–five gross peaks in the spectral shape located between 200 and 4000 Hz. These correspond to the resonances of the acoustic tube of the vocal tract. We will discuss formants further in Chapter 8 (Subtractive Synthesis). For now, we will note that the formant locations for the "ahh" vowel are radically different from those for the "eee" vowel, even though the harmonic spacing (and thus, the perceived pitch) is the same for the two vowels. We know this, because a singer can sing the same pitch on many vowels (different spectral shapes, but same harmonic spacings), or the same vowel on many pitches (same spectral shape and formant locations, but different harmonic spacings).

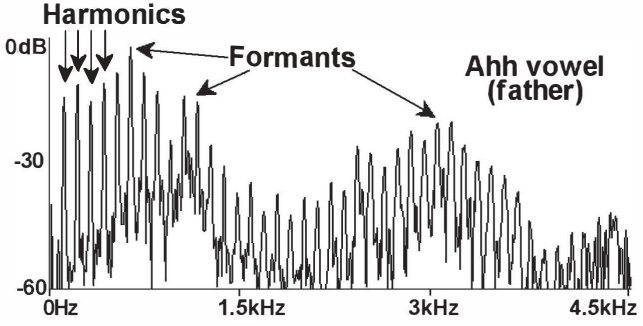

Figure 6.2. Spectrum of voiced vowel sound "ahh" (as in father).

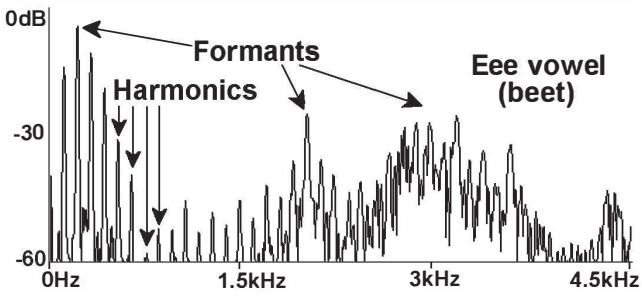

Figure 6.3. Spectrum of voiced vowel "eee" (as in beet).

The property of audio related to spectral shape is called *timbre*. Timbre is often defined as *those properties that allow us to differentiate two sounds which have the same pitch and loudness*. So an "ahh" versus an "eee" sound at the same pitch and loudness would be said to have different timbres. There are other properties related to timbre, such as the attack time of a sound, whether it is sustained or not, harmonicity versus inharmonicity, and the amount of noise contained in the signal. In the next sections, we will look at these components, and how they evolve in time.

6.1.2 Inharmonic

Many systems exhibit strong sinusoidal modes, but these modes are not restricted to any specific harmonic series. Such systems include even relatively simple shapes such as circular drum heads, square plates, and cylindrical water glasses. For example, a square metal plate would exhibit modes that are spaced related to the roots of integers (irrational numbers, so clearly inharmonic),

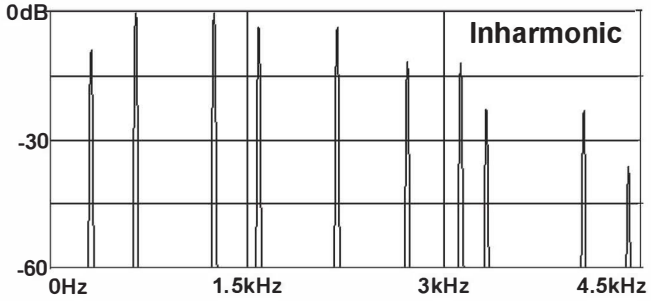

Figure 6.4. The inharmonic spectrum of a metal chime.

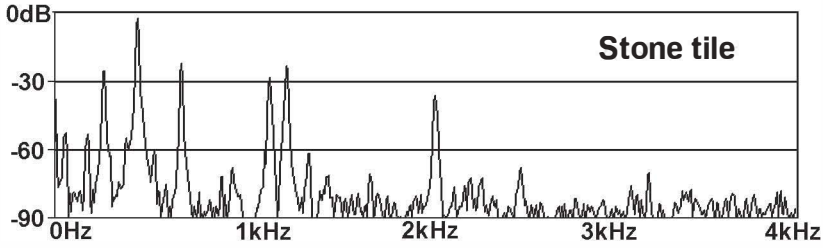

Figure 6.5. Spectrum of a struck square stone tile.

while a circular drum head would exhibit modes related to the roots of Bessel functions (also irrational, and clearly inharmonic). Figure 6.4 shows the inharmonic spectrum of a metal chime tone.

Figure 6.5 shows the spectrum of a struck square stone tile, and Figure 6.6 shows the spectrum of a struck circular drum (an African djembe). All of these spectra show clear sinusoidal modes, but the modes are not harmonically related.

23

6.1.3 Noise

There are many sounds that do not exhibit any clear sinusoidal modes. Whispered speech, wind blowing through trees, the crunches of our feet on gravel as we walk, and the sound of an airplane taking off are all examples of sounds that do not have clear sinusoidal modes. We can still represent these sounds using sums of sinusoids via the Discrete Fourier Transform, although it might not be a particularly efficient or revealing representation. Such sounds are generally classified as *noise*. Figure 6.7 shows the spectrum of a *white* (flat spectrum) noise signal. Another type of *pure* noise is called *pink* noise, which has a spectrum that rolls off linearly with log frequency.

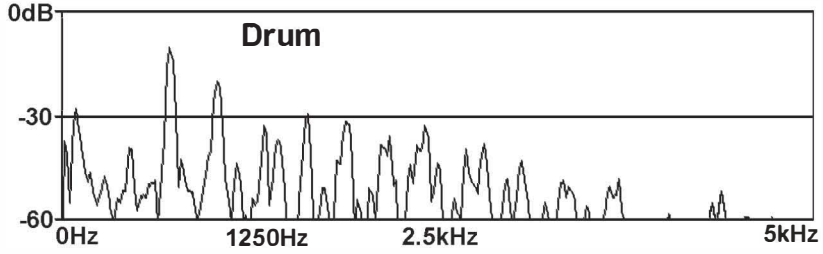

Figure 6.6. Spectrum of a circular drum head.

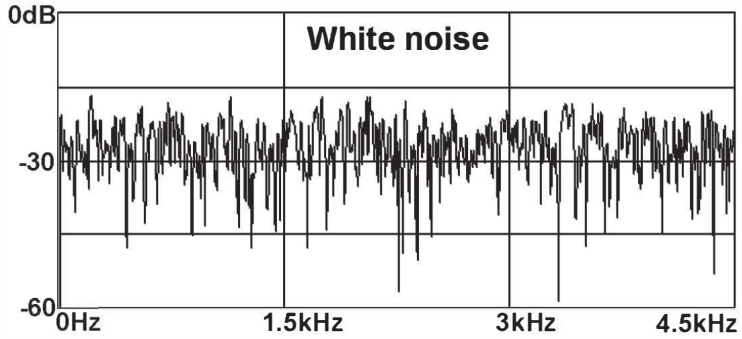

Figure 6.7. Spectrum of white noise signal.

Since we know we can whisper speech as well as voice it, we might get some clues as to what is going on perceptually from looking at the spectra of whispered vowels. We can whisper an "ahh" or "eee," and there will be no particular pitch to the sound, but the perception of each vowel remains. Figure 6.8 shows the spectrum of a whispered ahh, and Figure 6.9 shows the spectrum of a whispered eee. Note that the spectrum has no clear sinusoids (at least not spaced widely enough to be perceived or seen as individual sinusoidal modes), but still exhibits the formant peaks that are characteristic of those vowels.

24

6.1.4 Mixtures

Of course, most of the sounds we hear in day-to-day life are mixtures of other sounds. Even isolated single-source sounds will usually exhibit a mixture of spectral properties. For example, even our voiced speech vowels still have

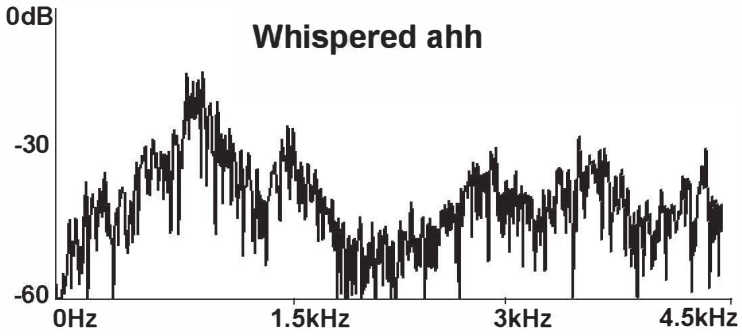

Figure 6.8. Spectrum of a whispered "ahh" sound.

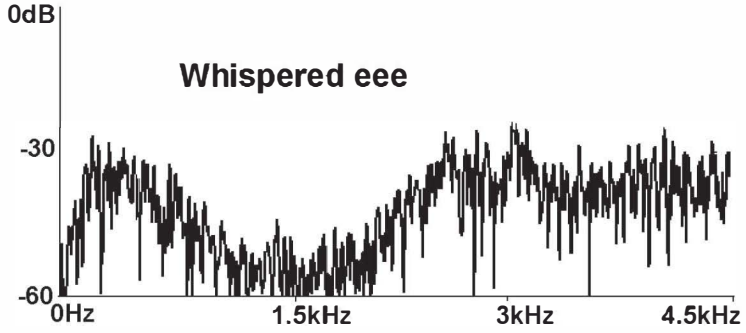

Figure 6.9. Spectrum of a whispered "eee" sound.

some noise components mixed in. A close inspection of Figures 6.2 and 6.3 will reveal that there is some amount of "fuzz" between the harmonics, corresponding to a breathy component in the voice. Another example can be found in percussive sounds, which might display clear sinusoidal modes, but also have lots of noise during the excitation segment of the sound. Figure 6.10 shows the spectrum of a struck wooden bar. Note that there are only three clear modes. The other spectral information is better described as noise.

6.2 Spectral Modeling

The notion that some components of sounds are well-modeled by sinusoids, while other components are better modeled by spectrally shaped noise, further motivates the residual-excited refinement to the purely sinusoidal additive synthesis model presented in Chapter 4 (Figure 4.4). Using the Fourier transform, we can inspect the spectrum of a sound and determine which

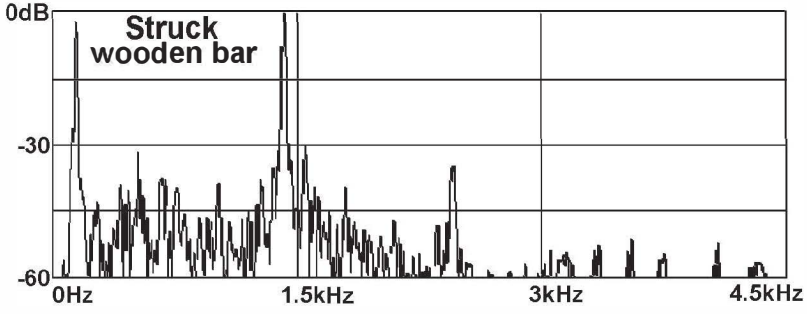

Figure 6.10. Spectrum of a struck wooden bar.

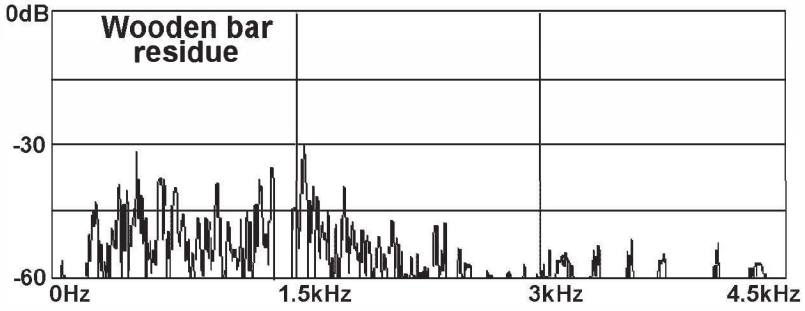

Figure 6.11. Spectrum of noise-only component of wooden bar strike.

components would be well-modeled by sinusoids. With careful signal processing, we can then subtract those components from the original signal. Ideally, what is left after subtracting the sinusoids will have no sinusoidal components remaining, and will thus be only noise component. Figure 6.11 shows the spectrum of the noise residual of the wooden bar strike after removing the sinusoidal modes.

25

The notion of "sines plus noise" modeling was posed and implemented by Xavier Serra and Julius Smith in the *Spectral Modeling Synthesis* (SMS) system. They called the sinusoidal components the *deterministic* component of the signal, and the leftover noise part the *residual* or *stochastic* component. Figure 6.12 shows the decomposition of a sung "ahh" sound into deterministic (harmonic sinusoidal) and stochastic (noise residue) components.

6.2.1 Spectral Modeling to Improve Additive Synthesis

Informed by the technique of sines plus noise spectral modeling, we can now improve our sinusoidal additive synthesis model significantly by simply adding a filtered noise source as shown in Figure 6.13.

The beauty of this type of model is that it recognizes the dominant sinusoidal nature of many sounds while still recognizing the noisy components that might be also present. More efficient and parametric representations—and many interesting modifications—can be made to the signal on resynthesis. For example, removing the harmonics from voiced speech, followed by resynthesizing with a scaled version of the noise residual, can result in the synthesis of whispered speech. Pitch and time shifting can be accomplished, but extra information is available to do the shifts more intelligently. For example, if the signal is speech and a segment seems to have no sinusoidal components, it is likely a consonant. Time or pitch shifting might be performed differently, or not at all, in such segments.

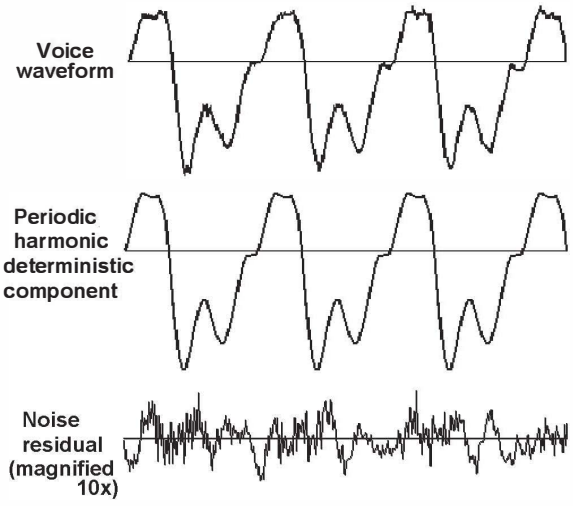

Figure 6.12. Top: Original sung "ahh" waveform. Middle: Sinusoidal components. Bottom: Residual components.

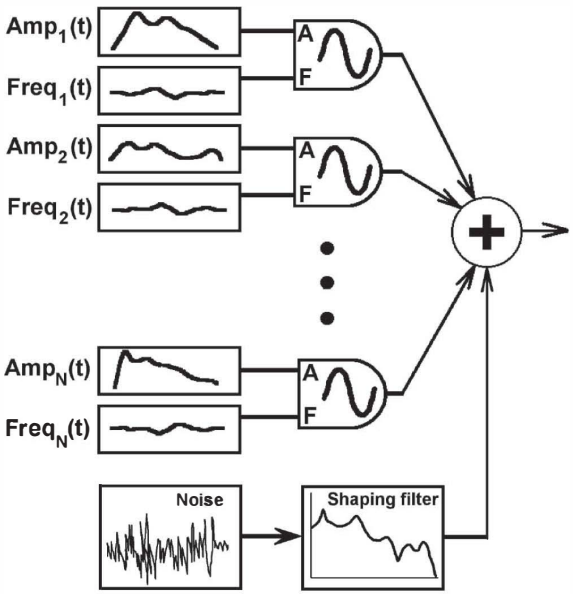

Figure 6.13. Sinusoidal additive model with filtered noise added for spectral modeling synthesis.

6.3 Sines Plus Noise Plus Transients

One further improvement to spectral modeling is the recognition (by Verma and Meng) that there are often brief (impulsive) moments in sounds that are really too short in time to be adequately analyzed by spectrum analysis. Further, such moments in the signal usually corrupt the sinusoidal/noise analysis process. Such events, called transients, can be modeled in other ways (often by simply keeping the stored PCM for that segment). Figure 6.14 shows the process of decomposing a signal into a transient, then sinusoidal components, then a noise residual.

((26))

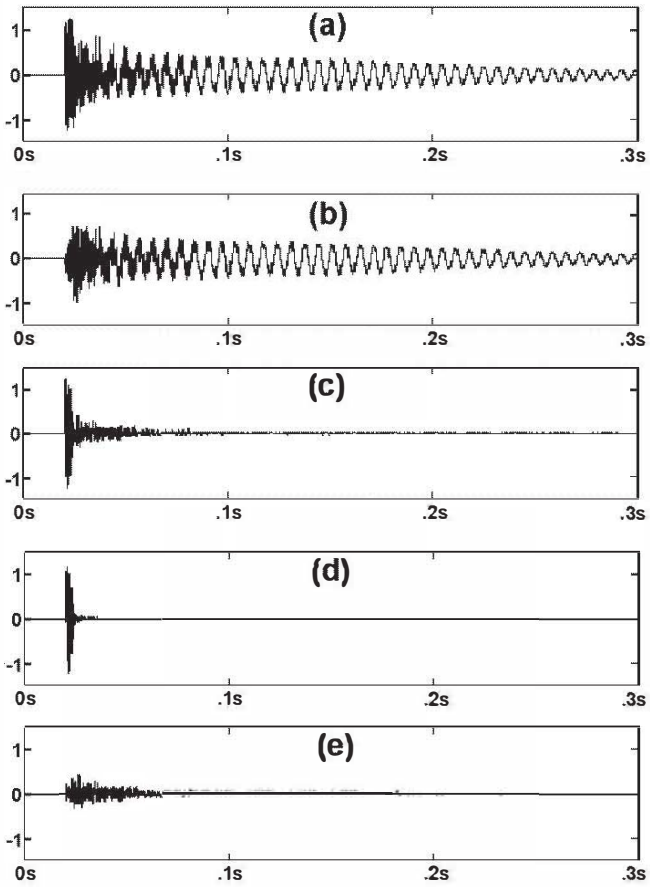

Figure 6.14. a) Original signal; b) without transient; c) transient with noise (sines removed); d) transient alone; e) noise alone (see *Verma and Meng*).

6.4 Spectra in Time

Of course, one of the most important aspects of sound is that it takes place in, and evolves over time. The basis of many of our time-domain manipulations in Chapter 2, and of short-time Fourier analysis as discussed in Chapter 5, is the breaking up of long sounds into shorter windows to more closely match our perception of frequency-domain aspects (relatively constant within a window) versus time-domain aspects (things that change slowly—remember the 30 Hz shift). Thus, it is often useful to plot spectra as they evolve in time. The two most common ways to do that are called the *spectrogram* (or *sonogram*), and the *waterfall plot*. In the spectrogram (sonogram), time runs left to right on the abcissa (x-axis), frequency runs bottom to top, and intensity is plotted as grayscale or color. In the waterfall plot, frequency runs left to right, intensity is plotted as height, and time runs diagonally upward plotted as a pseudo-three-dimensional depth. Each waterfall time "slice" shown is a traditional frequency spectrum.

Figure 6.15 shows a spectrogram plot of a plucked string tone, and Figure 6.16 shows a waterfall plot of the same plucked string tone. Note the harmonics and the way that high frequencies decay faster than the low frequencies. Also note that the first and second modes are strong, the third is weaker, and the fourth is nearly absent. The fifth is strong, and the sixth and seventh are weak. The eighth and ninth are stronger, and the tenth–twelfth are weak. These observations, combined with the plucked string boundary condition discussion in Chapter 4, indicate that the string was plucked at somewhere between the 1/3 and 1/4 position along the string.

To present an additional example of the usefulness of the sonogram, Figure 6.17 shows a sonogram of the "synthesize" utterance introduced in

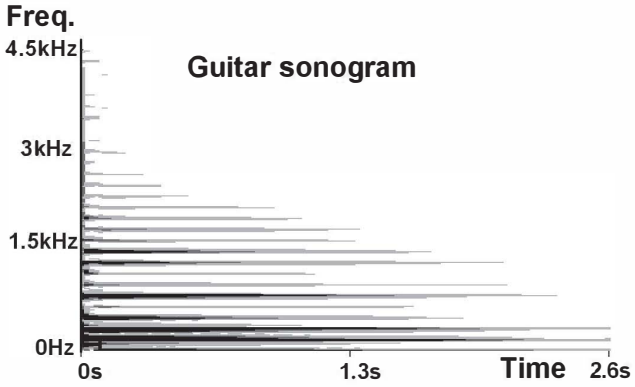

Figure 6.15. Sonogram of a plucked string.

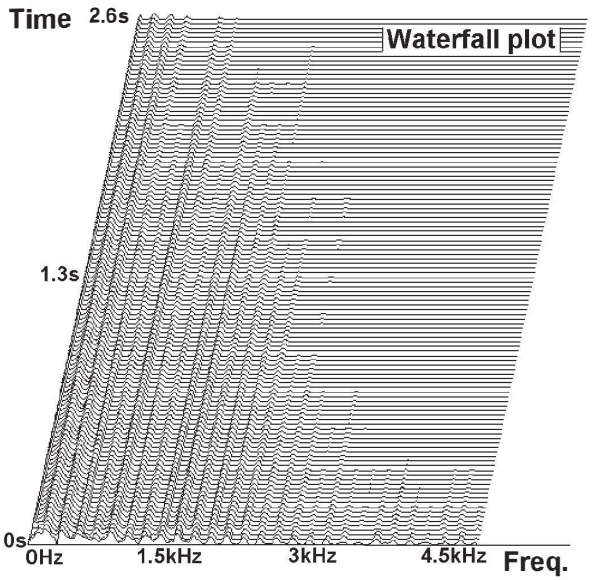

Figure 6.16. Waterfall plot of a plucked string.

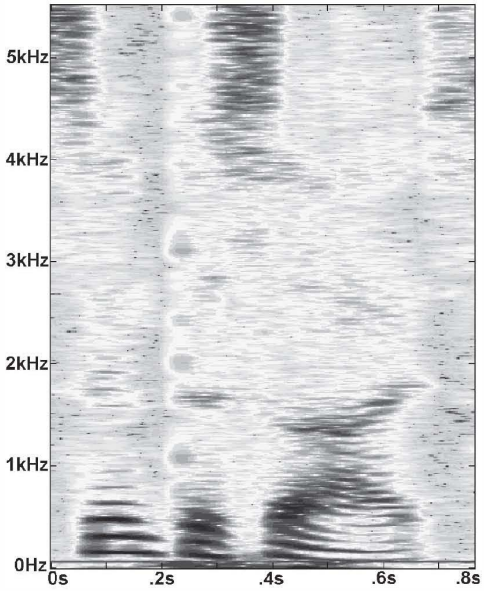

Figure 6.17. Sonogram of the uttered word "synthesize."

Chapter 2. Inspecting the spectrogram allows us to see the individual phonemes. Vowels /I/, /U/, /a/, etc., exhibit harmonics (parallel horizontal stripes), while consonants /s/, /th/, and /z/ show as vertical clouds of high-frequency fuzz.

6.5 Conclusion

In this chapter (and in the prior chapter on the Fourier transform) we've seen that lots of interesting information can be gleaned by transforming waveforms into the frequency domain. The appearance of a spectral plot can often be directly related to the sonic quality of the auditory experience of hearing that sound. In future chapters, we will return to the spectrum often to observe characteristics of sounds and the vibrations of physical systems.

Reading:

Robert J. McAulay and Thomas Quatieri. "Speech Analysis/Synthesis Based on a Sinusoidal Representation." *IEEE Transactions Acoustics, Speech, and Signal Processing ASSP* (34): 744–754 (August 1986).

Xavier Serra and Julius O. Smith. "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition." *Computer Music Journal* 14(4):12–24 (1990).

Tony Verma and Teresa Meng. "An Analysis/Synthesis Tool for Transient Signals that Allows a Flexible Sines + Transients + Noise Model for Audio." IEEE ICASSP-98. (1998).

Code:

fft.c

peaksfft.c

notchfft.c

Sounds:

[Track 22] Harmonic Sounds: Trumpet, "ahh," and "eee."

[Track 23] Inharmonic Sounds: Synth Bell, Stone Tile, Djembe Drum.

[Track 24] White Noise, Pink Noise, Whispered "ahh" and "eee."

[Track 25] Marimba Strike and Residue, Voice Original, Periodic, Residue.

[Track 26] Transient Extraction/Modeling (from *Verma and Meng*).