1 Lecture Outline

Reading: Chapters 2 and 7.

We will continue our quick review of DSP

- Quantization
- Windowing in Spectral Analysis
- Short-Time Fourier Transform
- Spectrographic Analysis of Speech

2 (Scalar) Quantization

Sampling and quantization are necessary prerequisites for DSP. Each measured sample $x[n]$ is held for at most T_s seconds during which time the A/D converts it to a quantized sample, $\hat{x}[n]$ which is represented by B (finite) bits.

Figure 1: Analog-to-Digital Conversion

The quantized sample being represented by B bits can take on only one of 2^B possible values. The ADC is characterized by a peak amplitude $x_{\max}$, which is divided equally into 2^B quantization levels. The spacing between levels, called the quantization width or quantization resolution is given by

$$\Delta = \frac{2x_{\max}}{2^B}. \tag{1}$$

In the analysis of quantization noise, we assume an additive noise model

$$\hat{x}[n] = x[n] + e[n] \tag{2}$$

where $e[n]$ is the difference between the actual and quantized sample value or error sequence and

$$-\Delta/2 \leq e[n] \leq \Delta/2. \tag{3}$$

In a first-order analysis, we make the following assumptions

Figure 2: Signal quantization

- $x[n]$, $e[n]$ are stationary random processes
- the probability density function (pdf) of $x[n]$ is uniform over $[-x_{\max}, x_{\max}]$ therefore $\mu_x = 0$ and $\sigma_x^2 = (2x_{\max})^2/12 = x_{\max}^2/3$
- the pdf of $e[n]$ is uniform over $[-\Delta/2, \Delta/2]$ therefore $\mu_e = 0$ and $\sigma_e^2 = \Delta^2/12 = x_{\max}^2/(3 \cdot 2^{2B})$
- $e[n]$ is white process, i.e. $E\{e[n]e[m]\} = \sigma_e^2 \delta[n-m]$
- $e[n]$ is uncorrelated with $x[n]$

Note that the second assumption is not true (and several others are weak) if $x[n]$ is speech. The signal-to-quantization noise ratio (SNR) is the ratio of signal power (variance) to quantization noise variance:

$$\begin{aligned} \text{SNR(dB)} &= 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \\ &= 10\log_{10}\left(\frac{x_{\max}^2}{3} \cdot \frac{3 \cdot 2^{2B}}{x_{\max}^2}\right) \\ &= 10\log_{10}\left(2^{2B}\right) \\ &= 20B\log_{10}(2) \\ &\approx 6B. \end{aligned} \tag{4}$$

Thus for each bit in our sample word length, we increase the SNR or dynamic range of the quantizer by 6dB "6dB per bit rule."

Example: 8 bit audio, 16 bit audio demos.

Example: The CD standard specifies 16 bit samples. Therefore, the dynamic range or SNR of the CD standard is $6 \times 16 = 96$dB. Since the human ear has a dynamic range of about 91dB the quantization noise is barely at the threshold of hearing. This is the reason why "CD quality" digital audio requires at least 16-bit quantization.

3 Windowing in Spectral Analysis

Windows play an important role in the spectral analysis of finite-duration signals. To begin consider the windowing theorem.

Theorem 1 *Let $x[n]$ be a signal and $w[n]$ be a "window" with the following DTFTs*

$$\begin{aligned} x[n] &\leftrightarrow X(\omega) \\ w[n] &\leftrightarrow W(\omega). \end{aligned}$$

Then $y[n] = x[n]w[n]$ is a "windowed"signal with the following DTFT

$$\begin{aligned} y[n] = x[n]w[n] &\leftrightarrow Y(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\theta)\, W(\omega - \theta)\, d\theta \\ &\leftrightarrow Y(\omega) = X(\omega) * W(\omega) \end{aligned} \tag{5}$$

where $*$ denotes convolution (periodic).

Theorem 2 *Let $x[n]$ be a signal and $w[n]$ be a "window" with the following DFTs*

$$\begin{aligned} x[n] &\leftrightarrow X[k] \\ w[n] &\leftrightarrow W[k]. \end{aligned}$$

Then $y[n] = x[n]w[n]$ is a "windowed"signal with the following DFT

$$\begin{aligned} y[n] = x[n]w[n] &\leftrightarrow Y[k] = \frac{1}{N}\sum_{m=0}^{N-1} X[m]\, W[(k-m)_N] \\ &\leftrightarrow Y[k] = X[k] \otimes W[k] \end{aligned} \tag{6}$$

where $(\cdot)_N$ denotes $(\cdot)$ modulo N and $\otimes$ denotes circular convolution, i.e. convolution of two periodic signals over one period.

The following three examples will illustrate the theorem and the tradeoffs.

Example 1: Consider the infinite-duration tone. Assuming a frequency, ω_0 the DTFT pair is given by

$$x[n] = \cos[\omega_0 n] \quad \leftrightarrow \quad X(\omega) = \pi\left[\delta(\omega - \ \omega_0) + \delta(\omega + \omega_0)\right]. \tag{7}$$

Figure 3: DTFT of infinitely-long tone

Example 2: Consider the infinite-duration wave (tone), $x[n]$ and a length-L rectangular window, $w[n]$

$$w[n] \quad = \quad \begin{cases} 1, & 0 \le n \le L-1 \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Multiplying sample-by-sample, the window with the tone, i.e., "windowing" produces a finite-duration tone, whose DTFT is

$$y[n] = x[n]w[n] \quad \leftrightarrow \quad Y(\omega) = X(\omega) \otimes W(\omega). \tag{9}$$

Figure 4: DTFT of rectangularly-windowed tone

and illustrated below.

The effect of windowing is to introduce abrupt on-off transients in the time-domain which result in spurious frequencies (sidelobes). Note that the level of these sidelobes is quite high (13dB down from the main lobe). Also note that the smearing effect of the frequency-domain convolution is to turn the spectral lines of the infinite-duration tone into main lobes. For the rectangular window, the main lobe width is $2\pi/L$.

Example 3: Consider once again the infinite-duration wave (tone), $x[n]$ and a length-L Hamming window, $w[n]$

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right), & 0 \leq n \leq L-1 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The Hamming window produces a finite-duration signal which is tapered at the edges. The tapered edges reduce the abrupt on-off transients in the time-domain which results in reduced spectral sidelobes. The resulting DTFT is illustrated below. Note that the level of these sidelobes is now 40dB down from the

Figure 5: DTFT of rectangularly-windowed tone

main lobe. The tradeoff, however, is that the main lobe is now twice the width of the main lobe from the rectangular window thus reducing the resolution.

In the presence of two of more closely-spaced tones, the main lobes will "blur" together thus limiting the resolution ability.

4 Short-Time Fourier Transform

Since speech is a non-stationary process, a single DTFT over the entire signal (called the periodogram) would not reveal the finer details of time-varying frequency content. Thus we are motivated to take DTFTs

over short "windows" of the signal in order to capture the time-varying spectral behavior. This leads to the idea of the time-dependent Fourier transform or Short-Time Fourier Transform (STFT).

The Discrete-Time Short-Time Fourier Transform (DT-STFT) of a signal is given by

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\omega m} \tag{11}$$

where $w[n]$ is the window sequence. This equation is interpreted as the DTFT of the shifted signal $x[n+m]$, as viewed through the window $w[m]$. The window has a stationary origin, and as n changes, the signal slides past the window so that at each value of n, a different portion of the signal is extracted by the window for Fourier analysis.

Alternately, we could reverse signal and window roles

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n+m]e^{-j\omega m} \tag{12}$$

and interpret as the DTFT of the signal $x[m]$ as viewed through the shifted window $w[n+m]$. The signal has a stationary origin, and as n changes, the window slides past the signal so that at each value of n, a different portion of the signal is extracted by the window for Fourier analysis.

Let us denote a sliding window, $w[n+m]$ where n is the index into the window and m denotes the position of the leading edge of the window. The figure below illustrates the idea.

Figure 6: Sliding window

The Discrete STFT of a signal is given by

$$X[n,k] = \sum_{m=-\infty}^{\infty} x[m]w[n+m]e^{-j\omega_k m} \tag{13}$$

where $x[m]$ is the signal, $w[n+m]$ is the analysis window positioned with its leading edge at $m=n$ and "slides" from left to right through the signal. The frequency $\omega_k = 2\pi k/N$ is the k-th DFT frequency. As m is increased in the summation, the analysis window $w[n+m]$ slides from left to right "through" $x[m]$. For each value of n, the DFT (spectrum) of $x[n]$, inside the window, is computed.

5 Spectrographic Analysis of Speech

The collection of DFTs (one at each point in time, n) is usually visualized as a *spectrogram*. In the spectrogram, we plot the power (magnitude-squared) spectrum, $S(n,\omega) = |X(n,\omega)|^2$ vs. n with magnitude values mapped to a color.

Instead of computing the STFT at each time instant, we often slide or advance the window by more than one sample. This is equivalent to computing the STFT every R samples or decimating. Usually the advance is specified in samples, R or as a % overlap of the window, i.e. 50% overlap.

MATLAB Code to Computer and Visualize a Spectrogram:

```
% [S,F,T] = spectrogram(x,window,noverlap,F,fs)
[S,f,t] = spectrogram(x,256,128,256,8000,'yaxis');
surf(t,f,10*log10(abs(S)),'EdgeColor','none');
axis tight;view(0,90);
xlabel('Time (s)');ylabel('Frequency (Hz)');
```

Example: Consider a chirp or police siren, i.e. tone whose frequency is linearly increasing

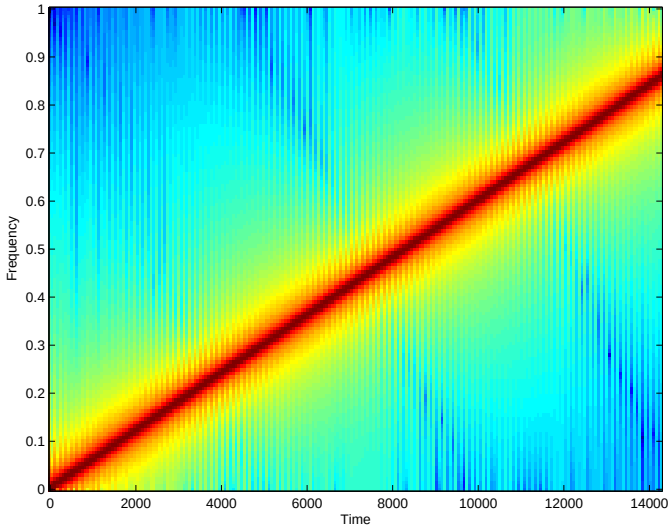
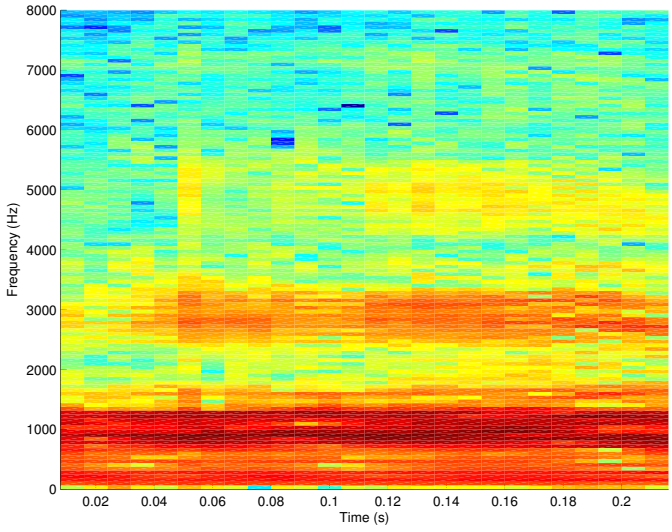
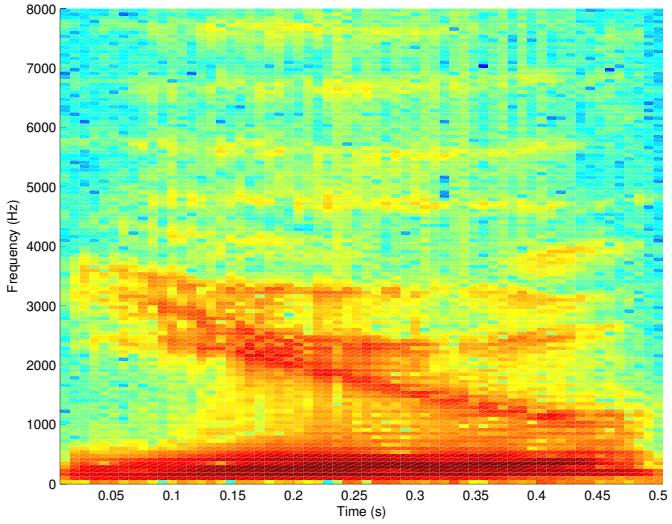


Figure 7: Spectrogram (wideband) of a chirp

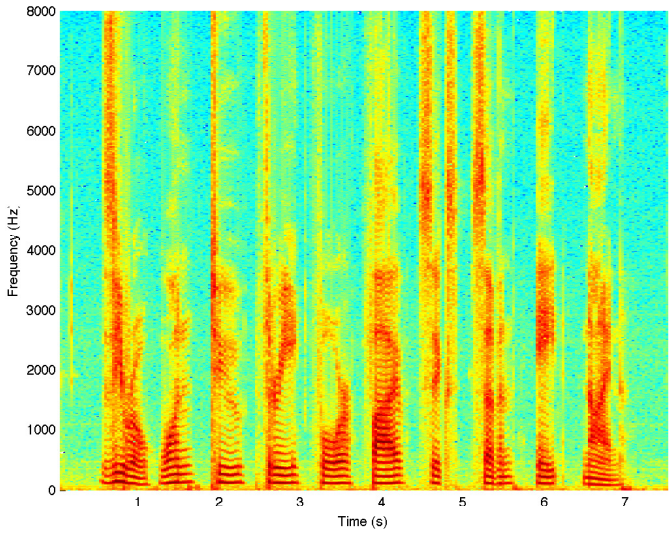
Example: Consider the speech sound (phoneme) 'AA' as in "B$\underline{o}$b"



Example: Consider the speech sound (phoneme) 'Y' as in "you"



Example: Consider the speech signal (utterance) "0-1-2-3-4-5-6-7-8-9"



There are two kinds of spectrograms: *narrowband* and *wideband*. The narrowband gives good spectral resolution and the wideband gives good temporal resolution. In the narrowband spectrogram, harmonic or spectral lines are resolved; these harmonic lines are seen as horizontal striations in the time-frequency plane of the spectrogram. In the wideband spectrogram, we similarly have vertical striations. The difference between the two is the length of the window: for narrowband spectrograms we use a "long" window; for wideband spectrograms we use a "short" window. Due to the uncertainty principle, we cannot have both good spectral and temporal resolution simultaneously.

6 Time-Frequency Resolution Tradeoffs

The design conflict in spectral analysis is the compromise between a long window for good frequency resolution and a short window for good temporal resolution. For a nonstationary signal such as speech, we have a

clear problem in simultaneously obtaining good resolution in time and frequency. For speech signal analysis, the window duration is typically set at about 20–30ms or a few pitch periods.

7 STFT Decimation or Overlapped Windows

In the STFT definition, for each point in time, we compute the DTFT of the windowed signal. Depending on the window length, advancing the window by one sample at a time results in a very redundant STFT. We can decimate the STFT in time retaining every R-th DTFT. This is equivalent to advancing the window by R samples before we compute the DTFT. If the window advance R is less than the window length N, the windows overlap. For speech signal analysis, we typically overlap the windows by 25%–75% of the window length.

Figure 8: Overlapping window