vowel	F1	F2	F3	example
iy	270	2290	3010	beet
ih	390	1990	2550	bit
eh	530	1840	2480	bet
ae	660	1720	2410	bat
ah	520	1190	2390	but
aa	730	1090	2240	hot
ao	570	840	2410	bought
uh	440	1020	2240	foot
uw	300	870	2240	boot
er	490	1350	1690	bird

Table 1: Formant frequencies of vowels

1 Short-Time Analysis

Please be aware that the mathematical notation can vary across text-books, also within the same field. In the course book (Vary & Martin, 2006), the *predictor filter* is usually denoted as

$$A(z) = a_1 z^{-1} + \ldots + a_p z^{-p}$$

and then the complete speech *analysis filter* is $\tilde{A}(z) = 1 - A(z)$, and the complete speech *synthesis filter* is

$$H(z) = \frac{1}{\tilde{A}(z)} = \frac{1}{1 - A(z)}$$

However, in this exercise collection, the complete *analysis filter* is sometimes called $A(z)$ instead of $\tilde{A}(z)$, and the *synthesis filter* is then denoted as

$$H(z) = \frac{1}{A(z)}$$

Please be aware of this notational inconsistency.

1.1 In this exercise we will exploit our knowledge of acoustic phonetics (see chapter 2.2 in the textbook) to recognize isolated digits. A male adult talker spoke each of the 11 digits in random sequence, and their spectrograms are shown in Figure 1.1. Try to identify each of the digits based on its acoustic properties. Use table 1.1 giving the typical formant values (F1,F2,F3) for vowels in english. Are the spectrograms shown in figure 1.1 wideband or narrowband spectrograms?

1.2 In this problem we study the spectra of vowels. Based on the formant frequencies for the vowels /iy/ and /aa/, see Table 1.1,

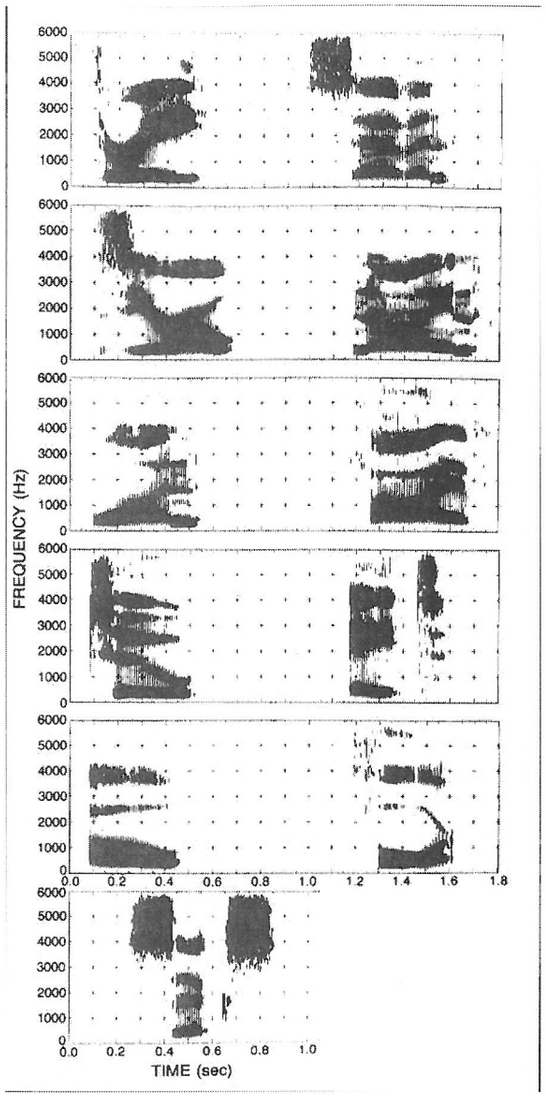

Figure 1: Spectrograms of 11 isolated digits, 0 through 9 plus 'oh' in random sequence. [1] (Black color denotes high intensity.)

1. Sketch the spectral envelopes (the magnitude transfer function of the vocal tract filter).
2. Sketch the pole positions in the complex plane.
3. Indicate the tongue position for each vowel.
4. Synthesize the vowels artificially using the following piece of MATLAB code

```
Fs = 8000;
F1 = 270; F2 = 2290; F3 = 3010; % /iy/
f = [F1 F2 F3] / Fs;
r = 0.95; % Put poles close to unit circle to create distinct resonance
z = [r*exp(i*2*pi*f) r*exp(-i*2*pi*f)]; % Complex conjugate poles!
a = poly(z);
F0 = 80; % Pitch frequency in Hz
T0 = floor(Fs/F0) % Pitch period in samples
e = zeros(16000,1); e(1:T0:end) = 1; % Excitation signal;
x = filter(1, a, e);
play16(x*8000, 8000) % Or your favorite playback command
```

1.3 Calculate the filter polynomial that corresponds to a 2nd order vocal tract filter with poles at a radius 0.9 and at angles ± 0.63 radians.

1.4 Voiced and unvoiced sounds. For the phoneme /ih/, assuming a pitch of 150 Hz, sketch

1. the magnitude of the DtFT
2. the poles of the vocal tract filter
3. Repeat 1. and 2. for the phoneme /f/

1.5 Consider an infinite periodic pulse train with period P (samples):

$$x(n) = \sum_{i=-\infty}^{\infty} \delta(n - iP)$$

Here $\delta(n)$ is the *Kronecker* delta, defining a *unit-sample signal* with integer-valued (time-discrete) argument n. The unit-sample signal $\delta(n)$ contains a single non-zero discrete-time sample of amplitude 1 at $n = 0$. To process the signal in a computer we need to window it, i.e., we extract the signal $x_m(n) = x(n)w(n - m)$, where $w(n)$ is a window function, for example the rectangular window. The length of the window is N samples, $w(n) \neq 0, n = 0, \ldots N - 1$. Here we study how windowing affects the spectrum.

1.5.a Sketch $x(n)$, and $x_m(n)$ for $m = P/3$. Assume $N = 4P$.

1.5.b Calculate the discrete time Fourier transform (DtFT) of $x_m(n)$, i.e., calculate the short time Fourier transform of $x(n)$ using a window $w(n)$. Assume $m = 0$ for simplicity.

1.5.c Calculate the N-point DFT.

1.5.d Sketch the magnitude of the DtFT in 2 (assuming for example a rectangular window). See problem 1.16 for the DtFT of $w(n)$ in that case.

Hints:

Poisson's summation formula may be useful:

$$\sum_{i=-\infty}^{\infty} e^{-j2\pi fiP} = \frac{1}{P} \sum_{i=-\infty}^{\infty} \delta(f - i/P),$$

where $\delta(u)$ is Dirac's delta function, defined for continuous-valued argument u.

1.6 The air volume velocity $x(n)$ at the glottis is periodic for voiced sounds and is depicted in Figure 2 (time along the horizontal axis). We wish to calculate the short time Fourier transform when using a window $w(n)$ of length N samples, $w(n) \neq 0, n = 0, \ldots N-1$. Repeat tasks 2-4 from problem 1.5 but for the glottal pulse train.

Figure 2: Glottal pulse train.

Hints:

Use the fact that a glottal pulse can be modeled by an impulse filtered by a filter $G(z)$ with an impulse response equal to the glottal pulse shape.

1.7 For voiced sounds, the speech signal $x(n)$ can be modeled like in Figure 3, where $1/A(z)$ is the vocal tract filter and $1 - z^{-1}$ is the lip radiation. In this problem we study how windowing affects the speech signal when we do frequency analysis.

1. Repeat tasks 2-4 from problem 1.5. When sketching the magnitude of the DtFT, you have to insert specific DtFTs for the vocal tract filter and the glottal pulse shape. For the glottal pulse, you can use problem 1.8. For the vocal tract, see table 1.1 for typical shapes. Also don't be too fussy about the phase of the Fourier transforms, just assume zero phase for simplicity when sketching.

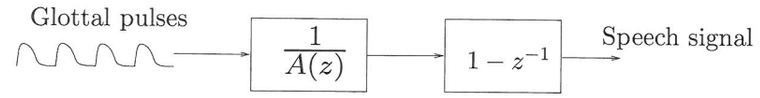

Figure 3: Source-filter model of speech.

2. What is the distance between harmonics? What is the width of the main lobe of the window's Fourier transform (assuming rectangular window)? How do you choose the frame length N?

3. Which window function, rectangular or Hanning, would you prefer and why? See also problem 1.16.

1.8 A glottal pulse can be modeled by

$$g(n) = \begin{cases} 1 + \cos(\pi \frac{n}{T}) & n = 0, \ldots, T-1 \\ 0 & \text{otherwise} \end{cases}$$

1. Sketch $g(n)$.

2. Calculate the DtFT of $g(n)$.

1.9 Show that the short time ACF

$$R_m(k) = \sum_{n=-\infty}^{\infty} x_m(n) x_m(n+k),$$

where $x_m(n) = x(n)w(n-m)$, can be written like

$$R_m(k) = \sum_{n=0}^{N-1-|k|} x(n+m)x(n+m+k).$$

Here $w(n)$ is a window function which is non-zero for $n = 0, \ldots, N-1$.

1.10 Here we study a popular method for speech time-scale modification: waveform similarity over-lap and add (WSOLA). In WSOLA, frames of length N samples are extracted from the original speech waveform and then these frames are overlapped 50% and added to form the output. Here we study a special case of WSOLA. To formalize we index each frame with i, and denote the frame length N. If we extract non-overlapping frames the extraction point (index of the first sample in the frame) is iN. The extracted frames are overlapped 50 % and added, see Figure 4.

1. Calculate (sketch) the output for the input signal in Figure 5. Use a frame length $N = 8$ and calculate the output for 6 extracted frames.

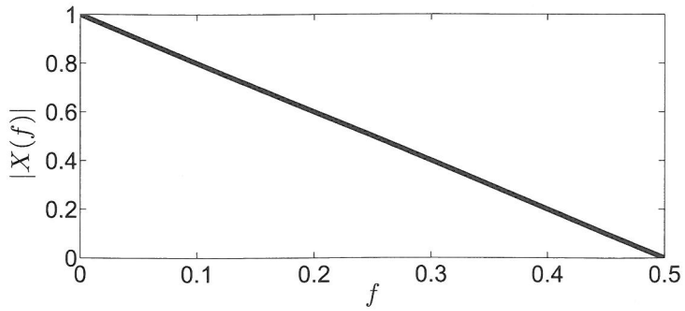

Figure 6: Spectrum of original signal in problem 1.12.

1.13 Consider a speech processing system which first performs a factor 2 upsampling, then WSOLA (see problem 1.10) with 50% overlap. Characterize the speech that is produced (we play back the output at the same rate as the original)!

1.14 The speech signal is noise-like for unvoiced sounds. Assuming that the speech has a white spectral shape (an over-simplification), what is the ACF? Assume stationary stochastic signals.

1.15 Fig. 7 represents the magnitude of the discrete-time Fourier transform of a steady-state vowel segment which has been extracted using a rectangular window. The envelope of the spectral magnitude, $|H(\omega)|$, is sketched with a dashed line. Note that three formants are shown, and that only the main lobe of the window Fourier transform is depicted. Suppose the sampling rate is 6000 samples/s and was set to meet the Nyquist rate.

1. What is the pitch period in milliseconds?
2. Compare to table 1.1. Which vowel corresponds to the spectrum depicted in figure 7.
3. How long is the rectangular window in milliseconds?
4. (*) If $F_1 = 750$ Hz and the vocal tract is considered to be a single acoustic tube, what is the length of the vocal tract? Assume zero pressure drop at the lips, an ideal volume velocity source, and speed of sound $c = 350$ m/s.
5. (*) If the length of the vocal tract were shortened, how would this affect the spacing of the window main lobes that make up the discrete-time magnitude spectrum of the signal? Explain your answer.

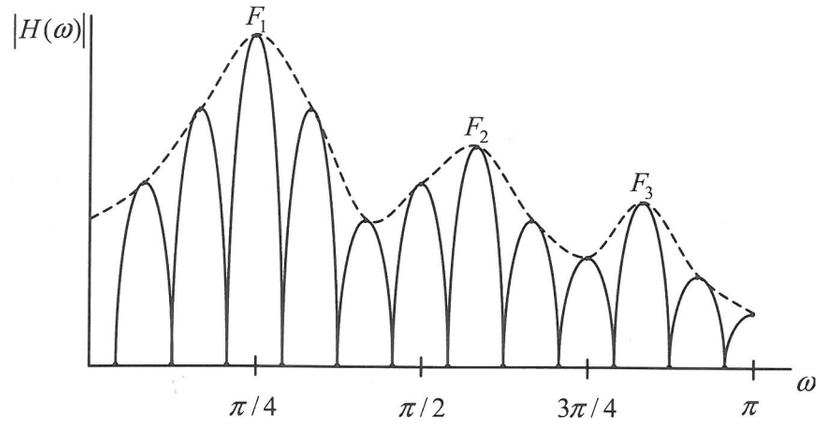

Figure 7: $|H(\omega)|$ of steady-state vowel segment.

1.16 For short time signal analysis we cut out pieces of the signal using a windowing function of limited time support (limited duration). This problem illustrates the different behavior of different windows in the frequency domain.

The rectangular window is defined as

$$w_R(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} .$$

The Hamming window is defined as

$$w_H(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} .$$

1. Show that the Fourier transform of the rectangular window is
$$W_R(e^{j\omega}) = \frac{\sin(\omega N/2)}{\sin(\omega/2)} e^{-j\omega\frac{(N-1)}{2}} .$$
2. Sketch $W_R(e^{j\omega})$ as a function of ω. (Disregard the linear phase factor $e^{-j\omega(N-1)/2}$.)
3. Derive an expression for $W_H(e^{j\omega})$.

   Hints: Express $w_H(n)$ in terms of $w_R(n)$.
4. Sketch $W_H(e^{j\omega})$. (Disregard the linear phase factor $e^{-j\omega(N-1)/2}$.) Your sketch should illustrate how the Hamming window trades frequency resolution for increased suppression of higher frequencies. What are the advantages/disadvantages of using the rectangular or Hamming window?