UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Department of Electrical and Computer Engineering
Instructor: Mark Hasegawa-Johnson

ECE 537 Speech Processing

Problem Set 4

Fall 2009

Issued: Wed Sep. 23, 2009 **Due:** Wed Oct. 14, 2009

Reading for problem set 4: Flanagan, Allen & Hasegawa-Johnson 3.1-5

This problem set contains one problem with seven parts (parts (a)-(g)). I will give up to one point for each part, if you get it mostly right. You may turn this homework in late if you wish. Notice that I am distributing solutions along with the homework; therefore, in order to get full credit, you must show me your code, as well as your solution figures!

Problem 4.1

In this problem you are to use a frequency-domain multi-tube model of the vocal tract to synthesize the four so-called "point vowels" and the schwa: /i/ as in "beat," /æ/ as in "bat," /ɑ/ as in "bought," /u/ as in "boot," and /ə/ as in the first syllable of "about" (Table 1).

Vowel	Section	Length [cm]	Area [cm^2]
/i/	1	9	8
	2	8.5	1
/æ/	1	4	1
	2	13.5	8
/ɑ/	1	9	1
	2	8.5	7
/u/	1	16.5	7
	2	1	1
/ə/	1	17.5	6
	2	0	6

Table 1: Table of lengths for the various vowel sounds.

(a) In this homework, we will assume the glottis is a volume velocity $U_S(t)$ with infinite impedance, meaning that as far as the vocal tract transfer function is concerned, $Z_G = \infty$.

To generate the $U_S(t)$, let's assume the fundamental frequency $f_0(t)$ is slowly rising from 150 to 200 Hz, and has a slight vibrato to it. (If you don't know the word "vibrato," look it up

in the dictionary.) Compute an $f_0(t)$ "pitch" signal over the duration of the speech sample. Lets assume that the speech will be D seconds long. How long will your array need to be that defines $f_0(t)$?

Compute this with a vibrato of 8 Hz, having a 0.2 Hz deviation, and with some variation in it, as you might use when you speak, namely

$$f_0(t) = 150 + 100 * t/\text{Duration} + 25 * \sin(2 * \pi * t)/\text{Duration} + 0.2 * \sin(2 * \pi * 20 * t);$$

Your speech sample should be at least D = 1.5 [sec].

Plot $f_0(t)$ on linear-linear coordinates, and label your plot. Add this figure to your report, with a figure caption describing what it is.

Now, compute the glottal area as a function of time, using a reasonable sampling frequency, e.g., $F_S = 11025$samples/second. Assume a glottal area that opens slowly, then closes rapidly. We will use the KLGLOTT88 model, which basically says that the width of the glottis, in centimeters, is $w_G(t) = (t/T_a)^2 - (t/T_a)^3$. Assume that the vibrating part of the glottis is 1.5cm long. Assume, also, that there is a fixed "chink" between the arytenoid cartilages, with an area of 3mm^2. All together, we have that

$$A_g(t) = \begin{cases} 3 \times 10^{-6} + (1.5 \times 10^{-4}) \left((t/T_a)^2 - (t/T_a)^3\right) \text{m}^2 & 0 \le t \le T_a \\ 3 \times 10^{-6} \text{m}^2 & T_a \le t \le T_0 \end{cases}$$

Assume that $T_a \approx 0.6 T_0$, i.e., T_a should change in proportion to T_0 as T_0 changes, during the course of your synthesized speech.

Use the Bernoulli formula, $P_s = 0.875 \rho_0 U_S^2 / 2A_g^2$, to find the glottal volume velocity, $U_S(t)$. Plot $U_S(t)$ for two pitch periods (make sure your abscissa is labeled in milliseconds). Verify that the period starts out with a duration of 6.6 ms (150 Hz), given your sample rate (do this by plotting $U_S(t)$, and look to see that the period is correct).

Also create a log-log plot of the magnitude spectrum of one glottal pulse, e.g., using something like this:

```
USS=abs(fft(US(1:(T0*FS)),N));
loglog([0:(N/2-1)]*FS/N,USS(1:(N/2)));
```

Note that you may need to use the `axis` command to adjust the axes, so that you can see necessary information in the plot.

What is the slope of the glottal spectrum at low frequencies (below 1000Hz), in decibels/decade (think: how do decibels relate to the magnitudes shown in your log-log plot)? What is the slope at high frequencies (above 1000Hz), in decibels/decade?

Solution: The glottal source should look roughly like this. Actually, it's been pointed out to me that there are two problems with this plot: (1) I bungled the time axis; the first pitch period should be 6.6ms, and (2) the spectrum shows the spectrum of 10 consecutive pitch periods, rather than a single pitch period. Rather than fix these problems, I'm telling you about them, so that your graph can be better.

(b) Next, you need to find the two-port matrix (the ABCD matrix) for each tube section. Use the formulas that you derived in HW3. As in HW3, ignore any terms that are third-order polynomials, or higher, in $s = j\omega$.

There are two different methods that DSP engineers usually use in order to transfer from continuous time to discrete time. One method is the "impulse-invariant transform," i.e., $h[n] = h(nT)$, which is equivalent to substituting $z = e^{s/F_s}$, where $z = e^{j\omega}$ is a unit advance, and $s = j\Omega$ is radians/second. The problem with this method is that it causes aliasing: any part of $H(j\Omega)$ for $|\Omega| > \pi/F_s$ gets aliased to lower frequencies. The second standard method is the "bilinear transform," i.e., $s = 2F_s(1 - z^{-1})/(1 + z^{-1})$. This method works very well for most filter design tasks, but works remarkably badly for lumped-element simulations of a transmission line like the vocal tract, because the lumped-element models only work at relatively low frequencies.

Because neither of the standard methods works well, I recommend that you use the oversampling hack. The oversampling hack includes two steps. The first step is the impulse-invariant transform with a sampling frequency at least four times higher than you eventually intend to use. For example, if you eventually intend to synthesize speech at 11025 samples/second, you would compute the two-port matrices using a 44100Hz sampling frequency. Thus, for the kth frequency bin ($0 \le k \le N/2 - 1$, I used $N = 1024$), define the digital two-port matrix $ABCD[k]$ in terms of the continuous-time ABCD matrix $ABCD(s)$ as

$$\begin{bmatrix} A[k] & B[k] \\ C[k] & D[k] \end{bmatrix} = \begin{bmatrix} A(s = j2\pi kF_s/N) & B(s = j2\pi kF_s/N) \\ C(s = j2\pi kF_s/N) & D(s = j2\pi kF_s/N) \end{bmatrix}$$

If your vocal tract model has M tube sections, then there will be M ABCD-matrices: one for each tube section (there is no separate tube section for the glottis, because we will assume that the glottis has infinite impedance). Each of these matrices includes up to four frequency response vectors, $A[k]$, $B[k]$, $C[k]$, and $D[k]$. In addition to these quantities, there is one more transfer function that you need to compute using a bilinear transform: the radiation

impedance, $Z_R[k]$. Thus, a complete M-section vocal tract model includes $4M + 1$ transfer function vectors.

Write a function `[H,TF,TPV,TPS,ZR]=twoport(A,N,FS);` that accepts the area function, A[m] (where m denotes the tube number), the size of the FFT, N, and the sampling rate, FS, and computes the following outputs:

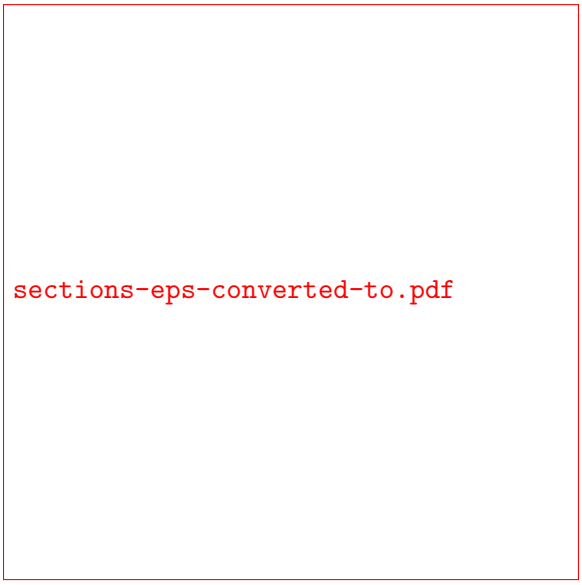
- ZR[k]=PL[k]/UL[k] is the radiation impedance
- TPS[i,k,m] is the k'th frequency component of the i'th element in the two-port matrix for the m'th tube section. There are four elements in each two-port matrix: $A_m[k]$, $B_m[k]$, $C_m[k]$, and $D_m[k]$. If you wish to re-order this array into some other indexing order, feel free.
- TPV[i,k] is the k'th frequency component of the i'th element in the two-port matrix for the entire vocal tract, computed as the product of M component two-port matrices.
- TF[k]=UL[k]/US[k] is the frequency response from glottis to lips.
- H[k]=Pmic[k]/US[k] is the frequency response from glottis to microphone; $H[k] = j\rho_0 f TF[k]$.

In order to make this function easy to write, I recommend that you put all of the physical constants into a separate function, called from inside your code. For this purpose you can download and use my function, `constants.m`, which gives constants the values that they have in footnote 4 of chapter 3 in the Flanagan text:

```
function [c,rho0,z0,mu,lambda,cp,eta,PS] = constants()
% [c,rho0,z0,mu,lambda,cp,eta,PS] = constants()
% Generate constants that are useful for acoustic simulations
  c = 350; % m/s (speed of sound, air at body temperature)
  rho0 = 1.14; % kg/m^3 (density of moist air at body temperature)
  z0 = rho0*c; % Rayls (characteristic impedance of air)
  mu = 1.86e-5; % Pa-s (coefficient of viscosity)
  lambda = 5.5e-8; % cal/kg-s-degree (coefficient of heat conduction)
  cp = 2.4e-4; % cal/kg (specific heat)
  eta = 1.4; % (coefficient of adiabatic expansion)
  PS = 80; % Pa (typical lung pressure)
```

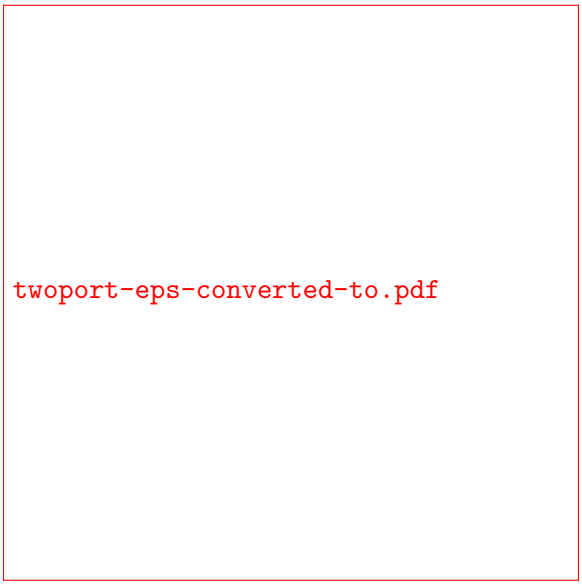
Choose two different tube cross-sectional areas, and for each of those two areas, plot the tube section $20\log_{10}|A_m(e^{j\omega})|$, $20\log_{10}|B_m(e^{j\omega})|$, $20\log_{10}|C_m(e^{j\omega})|$, and $20\log_{10}|D_m(e^{j\omega})|$ in four subfigures, for frequencies ranging from 0Hz to 22050Hz. Comment on the effect of area on the transfer function of a simple tube. This plot only needs to be computed for one of your five vowels.

Solution: Increasing area by a factor of eight results in a 20dB decrease in $B[k]$, a 20dB decrease in $C[k]$, and no change to $A[k]$ or $D[k]$, as shown here:



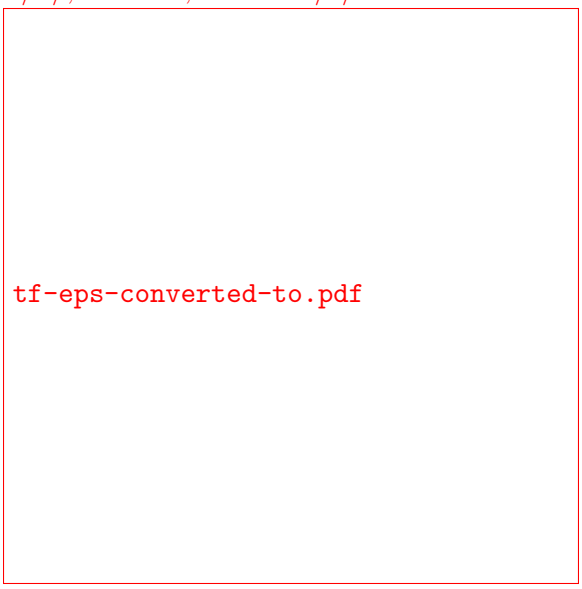
(c) Plot the two-port matrix for the complete vocal tract, from glottis to lips. On four separate axes, plot $20\log_{10}|A_m(e^{j\omega})|$, $20\log_{10}|B_m(e^{j\omega})|$, $20\log_{10}|C_m(e^{j\omega})|$, and $20\log_{10}|D_m(e^{j\omega})|$. This plot will be most informative if you only plot frequencies between 0Hz and 5000Hz, because the components above 5kHz are distorted by impulse-invariant aliasing. This plot only needs to be computed for one of your five vowels.

Solution: My two-port matrix has lots of sharp zeros, as shown here. This is the kind of two-port matrix that you should get if you modeled the two-tube vowel model using lots of short tubes, as we did in the time-domain simulation of homework 2. If you really just used two ABCD matrices—one for the back tube, one for the front tube—then you won't get a plot that looks anything like this one, but your transfer function plots (next section) should look pretty similar to mine.



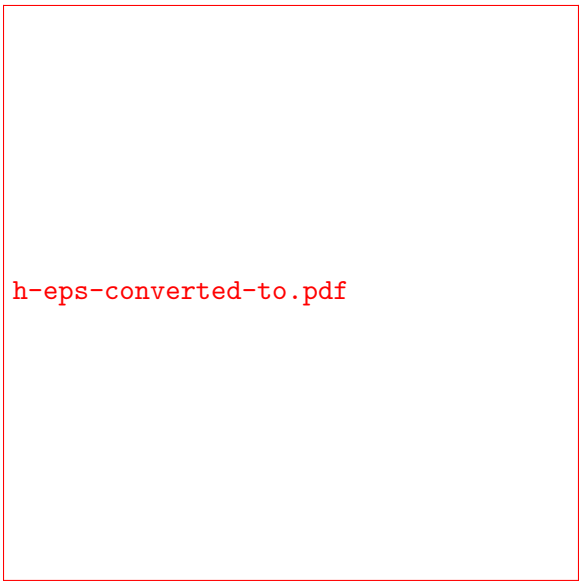
(d) Use the vocal tract ABCD matrix, together with $Z_R(e^{j\omega})$, to find the vocal tract transfer function $TF(e^{j\omega})$. On five separate axes, plot $20\log_{10}|TF(e^{j\omega})|$ for all five vowels, in the frequency range 0 to 5kHz (don't plot the higher frequencies). What are the first two formant frequencies of each vowel? Do they match the formant frequencies shown in Flanagan, Fig. 3.30?

Solution: The formant frequencies seem to be 200, 1800 for /i/, 600, 1700 for /æ/, 700, 1200 for /ɑ/, 400, 1200 for /u/, and 500, 1500 for /ə/.



(e) Compute $H(e^{j\omega}) = j\rho_0 f TF(e^{j\omega})$. Plot $H(e^{j\omega})$ for all five vowels, in decibels, for frequencies in the 0-5kHz range. Make sure that $20\log_{10}|H(e^{j\omega})| = 20\log_{10}|TF(e^{j\omega})| + 20\log_{10}\rho_0 f$.

Solution:



(f) Part (b) of this problem introduced a two-step "oversampling hack," but then only told you

step 1 of the hack. If you noticed that there was a missing step, go back to your write-up of part (b), and write "the two-step hack is missing a step!", and I will put an entertaining sticker on your homework assignment when I grade it.

Step 2 of the oversampling hack is to assume that the components of $TF(e^{j\omega})$ and $H(e^{j\omega})$ above about $\omega = \pi/4$ are wrong (aliased), and to therefore throw them away. Truncate both of these transfer functions at $\omega = \pi/8$, effectively downsampling by a factor of four. If TF[vowel,k] and H[vowel,k] are 5x1024 matrices containing the 1024-point frequency responses of the five vowels, then you can implement the oversampling hack as follows:

```
for vowel=1:5,
  TF2(vowel,:) = [TF(vowel,1:128),0,conj(fliplr(TF(vowel,2:128)))];
  H2(vowel,:) = [H(vowel,1:128),0,conj(fliplr(H(vowel,2:128)))];
  tf(vowel,:) = ifft(TF2(vowel,:));
  h(vowel,:) = ifft(H2(vowel,:));
end
```

Notice that the negative-frequency components of each transfer function have been generated by flipping the transfer function left-to-right, and taking its complex conjugate. The component at $\omega = \pi$ is assumed to be zero.

Inverse Fourier transform $TF(z)$ and $H(z)$ to get the vocal tract impulse response functions $tf(t)$ and $h(t)$. Find $U_L(t) = tf(t) * U_s(t)$ and $p_{mic}(t) = h(t) * U_S(t)$. In five separate axes, plot several pitch periods (about ten pitch periods) from each of the five vowels.

Solution: Here are some waveforms! They look quite different from each other — more different from each other, even, than natural vowels would look:

(g) Listen to each of the five synthesized vowels. Discuss the following points.

- Are the vowels intelligible? If a robot from Battlestar Galactica were to come up to you in a dark alley and play you one of these vowels, would you know what it was saying?

- Are the vowels natural-sounding? If not, in what way do they sound unnatural?

Solution: My vowels sounded intelligible, but certainly not natural. There was some reverberation, which is an artifact often caused by time-domain wraparound in the impulse response, often the result of inverse Fourier transforming a desired frequency response. The glottal source is also too mechanical; it needs aspiration noise, and perhaps a bit of jitter and shimmer (variation in T0 and in amplitude).