LANDMARK-BASED SPEECH RECOGNITION

Mark Hasegawa-Johnson

Lab 1

Issued: Monday, October 11, 2004 **Optionally Due:** Monday, October 18

Reading

- Gordon E. Peterson and Harold L. Barney, "Control Methods Used in a Study of Vowels." Journal of the Acoustical Society of America 24(2):175-184, 1952.
- René Carré and Maria Mody, "Prediction of Vowel and Consonant Place of Articulation." Technical Report, CNRS, 1997.
- Pierre C. Delattre and Alvin M. Liberman and Franklin S. Cooper, "Acoustic loci and transitional cues for consonants," Journal of the Acoustical Society of America, 27(4):769-773, 1955.
- The International Phonetic Alphabet, http://www.arts.gla.ac.uk/IPA/ipachart.html.

Mathematical Exercises

Problem 1.1

The acoustic pressure and particle velocity in a hard-walled tube are denoted $p(x,t)$ and $u(x,t)$ respectively[1]; their Fourier transforms are $P(x,j\Omega)$ and $U(x,j\Omega)$, meaning that

$$P(x,j\Omega) = \int_{-\infty}^{\infty} p(x,t)e^{-j\Omega t}dt \quad (1.1\text{-}1)$$

$$U(x,j\Omega) = \int_{-\infty}^{\infty} u(x,t)e^{-j\Omega t}dt \quad (1.1\text{-}2)$$

In the general case, $P(x,j\Omega)$ can be an arbitrary two-dimensional function of x and Ω. In the special case when the tube has constant area ($A(x) = A_0$ for all x), however, $P(x,j\Omega)$ and $U(x,j\Omega)$ are completely determined by the forward-going wave function $P_+(j\Omega)$ and backward-going wave function $P_-(j\Omega)$ as follows:

$$P(x,j\Omega) = P_+(j\Omega)e^{-j\Omega x/c} + P_-(j\Omega)e^{j\Omega x/c} \quad (1.1\text{-}3)$$

$$U(x,\Omega) = \frac{1}{\rho c}\left(P_+(j\Omega)e^{-j\Omega x/c} - P_-(j\Omega)e^{j\Omega x/c}\right) \quad (1.1\text{-}4)$$

where Ω is temporal frequency in radians/second, c is the speed of sound at human body temperature, and ρ is the density of air.

(a) In order to find $p(x,t)$ and $u(x,t)$ for all x and t, it suffices to find two unknowns: $P_+(j\Omega)$ and $P_-(j\Omega)$. In order to find two unknowns, you need two equations. Usually, these two equations are given by the boundary conditions. For example, if the glottis is closed, then air flow at the glottal end of the tube is zero, i.e.,

$$U(x=0,j\Omega) = 0 \quad (1.1\text{-}5)$$

[1] The total pressure at position x is $p(x,t) + P_{\text{atm}}$. P_{atm}, the atmospheric pressure, is usually much larger than $|p(x,t)|$, but because it is constant, it can be ignored.

Likewise, if the lips are wide open to the air, then air pressure at the lips must equal atmospheric pressure, so that

$$P(x = L, j\Omega) = 0 \tag{1.1-6}$$

Solve Eqs. 1.1-5 and 1.1-6 to find $P_+(j\Omega)$ and $P_-(j\Omega)$. You should find that $P_+(j\Omega)$ can only be nonzero at a countably infinite number of resonant frequencies, $\pm\Omega_n$, for $1 \leq n < \infty$. Find Ω_n in terms of L and c.

(b) Suppose that $P_+(j\Omega)$ is given by:

$$P_+(j\Omega) = \pi \sum_{n=1}^{\infty} P_{+,n}(j\Omega) \left(\delta(\Omega - \Omega_n) + \delta(\Omega + \Omega_n)\right)$$

Under these circumstances, $P(x, j\Omega)$ and $U(x, j\Omega)$ can also be written as

$$P(x, j\Omega) = \sum_{n=1}^{\infty} \pi P_n(x) \left(\delta(\Omega - \Omega_n) + \delta(\Omega + \Omega_n)\right) \tag{1.1-7}$$

$$U(x, j\Omega) = \sum_{n=1}^{\infty} \pi U_n(x) \left(\delta(\Omega - \Omega_n) + \delta(\Omega + \Omega_n)\right) \tag{1.1-8}$$

and the time-domain waveforms can be written as

$$p(x, t) = \sum_{n=1}^{\infty} p_n(x, t) \tag{1.1-9}$$

$$u(x, t) = \sum_{n=1}^{\infty} u_n(x, t) \tag{1.1-10}$$

Find $P_n(x)$, $U_n(x)$, $p_n(x, t)$, and $u_n(x, t)$ in terms of $P_{+,n}(j\Omega)$. Under the assumption that $P_{+,n}(j\Omega) = 1$ for all $n \leq 3$, plot the standing wave patterns $P_1(x)$, $U_1(x)$, $P_2(x)$, $U_2(x)$, $P_3(x)$, and $U_3(x)$.

(c) A uniform tube is a good model for the English vowel /AH/ (as in "tug;" this vowel is close to the Chinese vowel "e," as in the particle "de"). Estimate the formant frequences $F_n = \Omega_n/2\pi$ of the vowel /AH/, assuming that $L = 17.7$ cm, and assuming that $c = 354$m/s at body temperature.

(d) Suppose that $A(x)$ is "perturbed" by a small amount, so that

$$A(x) = A_0 + \alpha(x), \quad |\alpha(x)| \ll A_0 \tag{1.1-11}$$

Given non-constant $A(x)$, Eqs. 1.1-3 and 1.1-4 are no longer true, therefore it is not possible to use these two equations to compute the resonant frequencies of the vocal tract. Instead, Chiba and Kajiyama proposed the following perturbation method. Let $\Omega_{n,0}$ be the natural frequencies of the uniform tube, and let

$$\Omega_n = \Omega_{n,0} + \delta_n, \quad |\delta_n| \ll \Omega_{n,0} \tag{1.1-12}$$

be the natural frequencies of the perturbed vocal tract. When $A(x)$ is perturbed, the kinetic and potential energies of the tube are als perturbed. In order to keep them in balance, the resonant frequency of the tube must change by the following amount:

$$\delta_n \approx \frac{\pi c}{2L} \int_0^L \frac{\alpha(x)}{A_0} \left(|\rho c U_n(x)|^2 - |P_n(x)|^2\right) dx \tag{1.1-13}$$

where $P_n(x)$ and $U_n(x)$ are the standing wave patterns of the unperturbed tube.

Perturbations at different places lead to different changes in the resonant frequencies. Assume that $\alpha(x)$ is an extremely local perturbation at the location $x = \xi$, i.e.

$$\alpha(x) = \alpha_\xi \delta(x - \xi) \tag{1.1-14}$$

Define the perturbation sensitivity function $S_n(\xi)$ to be the partial derivative of Ω_n with respect to $A(x)/A_0$, assuming that $\alpha(x) = \alpha_\xi \delta(x - \xi)$, thus:

$$S_n(\xi) = \frac{\delta_n}{\alpha_\xi / A_0}$$

Find and sketch $S_1(\xi)$, $S_2(\xi)$, and $S_3(\xi)$.

(e) A /y/ or /i/ is created by constricting the tongue tip at $\xi \approx 3L/4$, thus $\alpha(x) \approx -0.5A_0\delta(x - 3L/4)$. Estimate F_1, F_2, and F_3 of /y/ and /i/, assuming that $L \approx 17.7$cm.

(f) A /w/ or /u/ is created by constricting the lips at $\xi \approx L$, thus $\alpha(x) \approx -0.5A_0\delta(x - L)$. Estimate F_1 and F_2 of /w/ or /u/, assuming that $L \approx 17.7$cm.

Problem 1.2

(a) A /g/ has a constriction of about 1cm in length, along the hard palate. The back cavity has a length of about 10cm; the front cavity has a length of about 5cm; assume that both have a cross-sectional area of $A_0 = 5\text{cm}^2$. Draw a three-tube model of /g/, just after the moment of release, so that the area of the constriction is $A_c = 0.5\text{cm}^2$. Use the three-tube approximation to estimate the formant frequencies of /g/, assuming that the tubes are completely decoupled.

(b) Assume that the constriction area, in cm^2, is given by $A_c(t) = 0.1t$ for $0 \le t \le 50$ ms. Assume that the transient and frication last for 10ms, then voicing begins. Sketch the spectrogram. Show the front cavity resonance peak in the frication spectrum. Show the formant frequencies in the voiced transition region. Assume that formant frequencies change in a straight line between $t = 0$ and $t = 0.05$.

Laboratory Exercise

Problem 1.3

In this problem, you will use matlab to plot wideband and narrowband spectrograms.

(a) Open matlab. If you have never used matlab before, first read through the matlab tutorial handed out in class.

(b) Use the `wavrecord` function to record your own voice, or use `wavread` to read in a short waveform. Call the waveform vector something like `x`. If the length of `x` is less than one second, append zeros to make it one second in length; if it is longer than one second, truncate to one second. Type `figure(1)` to get a figure window, then use `plot(t,x)` to plot `x` as a function of `t`. `t` should be a vector containing the times at which each sample of `x` was taken; if `FS` is the sampling frequency, `t` can be created as `t=[1:length(x)]/FS;`. Use the `zoom` function to zoom in on particular regions of `x`. Zoom in on the first vowel region. Can you estimate its pitch frequency?

(c) Use the `enframe` function (part of the voicebox toolkit, available at

`http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`)

to chop the waveform into about 1000 overlapping frames. Each frame should be windowed by a Hamming window 25 or 30 ms in length (use the function `hamming`, if available, otherwise use

```
w=0.54-0.46*cos(2*pi*[0:N]/N)
```

for some `N`). Each frame should begin only 1 ms after the beginning of the previous frame. Use the `subplot` and `plot` commands to plot five consecutive frames from the same vowel, in five subplots, on the same figure. Can you see several pitch periods in each window? How does the Hamming window affect the pitch periods?

(d) Each column of the spectrogram is one half of the log magnitude Fourier transform of one frame of speech. If the rows of matrix `X` contain frames of speech, then you can create the spectrogram `S` as

```
S = 20*log10(abs(fft(X,1024,2))); S=S(:,1:512)';
```

In order to make sure that you understand this, type `help fft`, and read about the FFT function.

Use `image` to plot the spectrogram. Type `h=image(T,F,a*S+b); axes xy; set(h,'Units','pixels'); set(h,'Position',[20 20 size(S)]);` for some constants `a` and `b`, and for vectors `T` and `F` that specify the time of each frame and the frequency of each spectral sample. Start out with `a=1` and `b=1`. You can set `T=[1:1000]` (if you have 1000 frames), and `F=[0:512]*FS/512` (if `S` contains 512 spectral samples from a 1024-point FFT).

The constants `a` and `b` adjust the brightness and contrast of the image. What is the best setting of these constants? How is the optimum setting related to `max(max(S))-min(min(S))`? How is it related to `size(colormap)`? Type `help colormap` to find out more about the colormap. Try some of the other colormaps available, such as `colormap gray` or `colormap hsv`.

What happens if you don't type `axes xy`?

The variable `h` contains a "handle" to the image plot. Characteristics of the image plot can be observed using the `get` function, or changed using the `set` function. The two `set` functions specified above will force the image to be plotted at one pixel per spectral sample. Changing the resolution in this way is necessary in order to observe all of the detailed of the spectrogram.

The spectrogram you have just plotted is called a "narrowband" spectrogram. The term "narrowband" refers to the bandwidth of the transform of the window function, $B = 2/D$, where D is the window length in seconds. Thus, for example, if $D = 0.025$ seconds, then $B = 80$Hz.

80Hz bandwidth is usually narrow enough to show the harmonics of the fundamental frequency as horizontal lines on the spectrogram. Pick any two vowels in the spectrogram, and estimate their pitch frequencies based on the horizontal striations.

(e) Repeat all previous sections, but using a "wideband" spectrogram instead of the "narrowband" spectrogram. Set $D = 0.006$ seconds. All other parameters (1ms inter-window skip, 1024-point FFT) should be the same.

In the wideband spectrogram, the pitch pulses should show up in time, rather than in frequency. Use the vertical pitch striations to estimate the pitch frequency in a few different vowels.

Problem 1.4

In this problem, you will use the Praat program to transcribe a spectrogram using either arpabet or pinyin. If Praat is not already available on your computer, you can download it from http://www.fon.hum.uva.nl/praat/.

(a) Construct a sentence of no more than 20 syllables. The sentence should include at least two different syllable-initial nasal consonants (e.g., /n/ and /m/), at least two different syllable-initial fricatives (e.g., /s/ and /sh/), and at least two different syllable-initial stops (e.g., /t/ and /p/). Record the sentence using any program.

(b) Start Praat. If you have not used it before, read through Sidney Wood's beginners manual:

`http://www.ling.lu.se/persons/Sidney/praate/frames.html`

especially the parts about viewing and editing sound files.

(c) Read in your waveform file. Create a transcription with two tiers; the second tier (larger units) is called "words," and the first tier (smaller units) is called "pinyin" or "arpabet" (depending on which phonetic transcription system you want to use). You can create a transcription by selecting the waveform name in the "objects" window, then pressing the button "Label & Segment," and choosing "To IntervalTier" from the popup menu.

(d) Select both the waveform and the transcription in the objects window, and press "Edit" to get an editing window. Zoom in until you can see the spectrogram. Place the cursor at the beginning of the first word, then hit CTRL-1 to enter a boundary in the first tier. To enter text in any interval, select the interval, and type text at top of the window. In this way, transcribe all words, all syllable onsets ("initials"), and all syllable rhymes ("finals").

(e) Measure the formant frequencies of any two vowels. Can you explain the formant frequencies (within about 300Hz), using either perturbation theory or a three-tube or two-tube vowel model?

(f) Describe the difference in the spectrogram between sonorants (including vowels and nasals) and obstruents (stops and fricatives).

(g) Describe the difference in the spectrogram between nasal consonants and vowels. What happens at the "landmark" between a nasal consonant and a vowel?

(h) Describe the difference in the spectrogram between fricatives and stop consonants.

(i) What happens at the "landmark" between a stop consonant and a vowel?

(j) Measure the front cavity resonances of both fricatives, and of both stop consonants. Based on your measurement, estimate the length of the front cavity, in centimeters.

(k) Measure the formant frequencies immediately after release of the stop consonants, and immediately after release of the fricative consonants. Can you explain these formant frequencies, using either perturbation analysis, or using a three-tube model?

Appendix A Some Phonemes

The international phonetic alphabet (IPA) is the international one-character standard for phonetic transcription. The table of IPA characters is available at ...

ARPABET is a standard for phonemic transcription of American English, on computes without an IPA font.

	IPA	ARPABET	IPA	ARPABET
Vowels				
	Front		Back	
High Rounded	y	ux	u	uw
High Unrounded	i	iy		
Mid Unrounded	e	ey		
Mid Rounded			o	ow
Low Unrounded	æ	ae	a	aa
Low Rounded				ao
Reduced (Schwa)		ix		ax
Consonants				
	Unvoiced		Voiced	
Labial Fricative	f	f	v	v
Alveolar Fricative	s	s	z	z
Palatal Fricative		sh		zh
Labial Stop	p	p	b	b
Alveolar Stop	t	t	d	d
Velar Stop	k	k	g	g
Palatal Affricate		ch		jh
Labial Nasal			m	m
Alveolar Nasal			n	n
Velar Nasal				ng
Labial Glide			w	w
Palatal Glide			y	y
Alveolar Liquid			l	l
Retroflex Liquid			r	r
h	h	hh	h	hv