ES 157/257: Speech and Audio Processing
Spring 2009 Midterm 1

1 hour 23 minutes

Answer ALL THREE questions. All have equal weight of 25 points apiece; the numbers in [ ] indicate the relative weights given to parts of questions.

SOLUTIONS
1. Acoustic Phonetics and the Source-Filter Model of Speech

The figure below shows a narrowband spectrogram of a speech utterance:

(a) Explain specifically what is meant by “the source-filter model of speech,” being sure to define the roles of pitch, voicing, glottal flow, and formants. For which classes of phoneme is the source-filter model a good one? For which classes does it exhibit clear shortcomings, and why? [7 marks]

Solution: The source filter model of speech production describes two independent operations, the generation of a forcing function at the glottis or at some point along the vocal tract and the time varying filtering of this source by the vocal tract itself. In voiced speech, the source is characterized by the regular vibration of the vocal folds in response to airflow from the lungs. This may be modeled as a pulse train convolved with a shaping function; this is the periodic glottal flow with a period called the “pitch period.” On the other hand, in unvoiced speech, these vibrations are no longer regular; constrictions in the vocal tract cause turbulence in the airflow. The source is then shaped or colored according to a filter whose resonant frequencies, or formants, are determined by the configuration of the vocal tract. The model works well for voiced phonemes and unvoiced strident fricatives. However, the model lacks a description of the physics of the nasal cavity and does a poor job in modeling sounds produced by tongue constrictions; such configurations are not well approximated by the tube model. Consequently, nasals and laterals are poorly modeled by the source-filter approach. Stops and other phonemes exhibiting transient behavior are also poorly modeled.

(b) In the figure above, identify and label an example of each of the following:
- A voiced speech sound
- An unvoiced speech sound
- A speech segment where time-varying pitch is visible
- A speech segment where time-varying formants are visible [8 marks]

Solution:

TURN OVER
(c) In the source-filter model, we often work with the derivative of the glottal flow waveform, rather than the waveform itself directly. Describe carefully why this is done, and explain under what assumptions this substitution is a reasonable one to make. Note: you may wish to use diagrams to illustrate your explanation. [4 marks]

Solution: The volume velocity output of the vocal tract is modified by a lip radiation impedance. This impedance is well modeled as a differentiator. By the commutivity of linear systems, we may apply this differentiator to the glottal flow input.

(d) The source-filter model is fundamentally linear, and hence cannot capture nonlinear effects. List three such sources of nonlinearity known to be present in the speech production mechanism, and explain briefly their effects (in time and/or frequency) on the resultant speech waveform. [6 marks]

Solution:

- Source tract coupling
- Viscous and Thermal Loss
- Vibrating Walls
- Radiation Loss
2. Acoustic Tube Models of the Vocal Tract

Consider a lossless two-tube model of the vocal tract in which tube cross-sections and lengths are given by

\[ A_1 = 2 \text{ cm}^2, \quad l_1 = 7 \text{ cm}; \]
\[ A_2 = 4 \text{ cm}^2, \quad l_2 = l - 7 \text{ cm}, \]

with \( l \), the total vocal tract length, unknown.

(a) Draw and label the complete signal flow diagram for this lossless two-tube model, including glottal and radiation effects. [7 marks]

Solution:

(b) Calculate all reflection coefficients and delay elements corresponding to this model and your flow diagram in part (a) above. Leave your answers in terms of \( l \) when necessary. [6 marks]

Solution:

\[ \tau_1 = \frac{l_1}{c} = \frac{0.07}{350} \text{ sec} \quad (1) \]
\[ \tau_2 = \frac{l_2}{c} = \frac{l - 0.07}{350} \text{ sec} \quad (2) \]
\[ r_1 = \frac{A_2 - A_1}{A_2 + A_1} = 4 - 2 \cdot \frac{1}{4 + 2} = \frac{1}{3} \quad (3) \]

(c) Now consider instead a single lossless uniform tube of length \( l = l_1 + l_2 \), with the glottis assumed closed. State the equation for volume velocity \( u(x, t) \) in this single-tube case, and solve for an equation that will yield the resonant frequencies of the tube in terms of \( l \). [6 marks]

Solution:

\[ u(x, t) = \frac{\cos(\Omega(l - x)/c)}{\cos(\Omega l/c)} U_g(\Omega) e^{j\Omega t} \quad (4) \]

Turn Over
Resonant frequencies:

\[
\cos(\Omega_k l/c) = 0 \Rightarrow \Omega_k = \frac{k\pi c}{2l} \quad \text{for } k = 1, 3, 5, \ldots
\]  \hspace{1cm} (5)

(d) Now suppose that you are told that the first formant resonance \( F_1 \) is equal to 500 Hz in the single-tube case of Part 2c above. Use this information to solve for the overall tube length \( l \), and sketch the spectral magnitude response (up to 2500 Hz) that would result if this tube were driven by an idealized glottal source consisting of a pulse train with pitch period 5 ms. \hspace{1cm} [6 marks]

**Solution:**

\[
\Omega_1 = 2\pi F_1 = 2\pi \cdot 500 = \frac{c\pi}{400} \Rightarrow l = \frac{c}{2000} = 17.5 \text{cm}
\]  \hspace{1cm} (6)

The magnitude response is given by a pulse train with spacing \( \frac{1}{5\text{ms}} = 200\text{Hz} \) sampling some envelope function with \( F_1 = 500\text{Hz} \).
3. Linear Predictive and Homomorphic Approaches to Deconvolution

(a) Compare and contrast the linear predictive and cepstral approaches for smoothing as a means of estimating the vocal tract transfer function from speech waveform data. What are the principal pros and cons of each approach? [6 marks]

Solution:
Cepstrum:
- **Pros:**
  - Handles both poles and zeros
  - Can produce mixed phase estimates
  - Reconstructions are more “natural”
  - Proper regularization may overcome the high pitch problem
- **Cons:**
  - May yield wider spurious resonances consistent with the spectral smoothing approach
  - Phase sensitivity to improper window alignment

LPC:
- **Pros:**
  - Yields sharp smooth resonances
- **Cons:**
  - Does not model zeros
  - Estimate must be minimum phase
  - Model based approach may be oversimplified, producing “mechanical” sounding reconstructions; Proper model order selection is key

(b) Consider the autoregressive model formulation for a speech signal $s[n]$ in terms of coefficients $\alpha_k$, model order $p$, and residual $e[n]$. Write down this model, and use the $z$-transform to explain how $p$ is related to the lossless concatenated tube model of the vocal tract. If $s[n]$ is known to be sampled at 10 KHz, what range of $p$ would be needed in order to achieve a good model for a voiced sound, and why? [5 marks]

Solution: We may write:

$$s[n] = \sum_{k=1}^{p} \alpha_k s[n - k] + Au_g[n]$$

The residual is given as:

$$e[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n - k]$$

Taking the $Z$-transform:

$$S(z) = \frac{AU_g(z)}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}$$

**TURN OVER**
From page 147 in the book, the transfer function for a \( p \) tube model is given by:

\[
V(z) = \frac{Az^{-N/2}}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}
\]

We observe \( V(z) = z^{-N/2} \frac{S(z)}{U_g(z)} \). For the second part follow Example 4.4 and 4.5 in the book with \( f_c = 5000 \text{Hz} \).

(c) Consider a rectangular analysis window of length \( N_w = 2M + 1 \) applied to a speech sequence over the interval \([n - M - p, n + M] \). If \( s_n[m] \) denotes the resultant short-time segment, show that minimizing the mean-square prediction error via the autocorrelation method of linear prediction leads to the following form of the normal equations:

\[
\Phi_n[i, k] = \sum_{m=0}^{N_w+p-1} s_n[m - i]s_n[m - k], \quad i = 1, 2, \ldots, p, \quad k = 0, 1, \ldots, p
\]

\[
\Phi_n[i, 0] = \sum_{k=1}^{p} \alpha_k \Phi_n[i, k], \quad i = 1, 2, \ldots, p
\]

[9 marks]

**Solution:** See solutions to homework 5.1

(d) It is well known that high-pitched speakers present a particularly difficult deconvolution when vocal tract estimates are sought. Give:

- Examples of speaker types likely to have high pitch
- A time-domain argument for why higher pitch will manifest itself through aliasing in the short-time autocorrelation function \( r_n[i - k] = \Phi_n[i, k] \) of Part 3c above
- A frequency-domain argument for why this same issue will also affect cepstral “liftering” to recover the vocal tract transfer function

[5 marks]

**Solution:**

- Woman and children
- Refer to homework solution 5.7
- Higher pitched speakers will produce spectra whose envelope is “undersampled” by the spectral pulse train representing the glottal source. Thus vocal tract response recovery in the cepstral domain will be hindered by this form of aliasing.