EECS 225d: Midterm 1, March 7, 1997

Version with correct answers (not necessarily unique)

For multiple choice questions, also write a brief (1-2 sentence) explanation of why the answer is correct. The parenthesized numbers after each question number give the relative point value of each question.

1. (2) The excitation model for the Voder was

   a) a pulse generator

   b) a collection of sin generators

   c) wide band noise

   d) (some other answer - specify your choice).

   ⇒ **d) A combination of a a) a pulse generator (for voiced sounds) and c) a wide band noise source (for unvoiced sounds) were used in the Voder**

2. (2) A range of different vowels can be synthetically produced by

   a) exciting a single uniform tube with periodic pressure pulses.

   b) exciting a multi-tube structure with the sum of sine waves.

   c) exceptionally smart single-cell animals

   d) (some other answer - specify your choice).

   ⇒ **b) works since a correctly weighted sum of sine waves can generate any excitation function — OR d) exciting a multi-tube structure with a periodic pulse train.**

3. (2) An acoustic tube closed at both ends and excited at its midpoint will resonate at frequencies

   a) higher than

   b) lower than

   c) equal to

   that of an acoustic tube open at one end and excited at the other end.

   ⇒ **a) since the boundary conditions of the first tube force the standing wave to be equal at both ends so that the lowest mode is a wave length of twice the tube length. I.e. the volume velocity is maximal at the center and 0 at both ends and hence a half-wave can fit in the tube. The second tube's wave length is four times the length of the tube since the pressure is maximal at the driven end and zero at the open end. (I.e. a quarter-wave can fit in the tube.) Thus since the wave length in the first case is half the wave length of the second case, the resonant frequencies of the first case are higher than those of the second case.**

4. (2) In response to pure sinusoidal tones, an auditory nerve will spike at different rates, depending on the frequency of the stimulus. These rates are predominantly determined by

   a) the properties of the hair cell stereocilia

   b) the basilar membrane vibration in the vicinity of the hair cell

   c) the specific parameters of the neuron

   d) the strength of the dollar in European markets

   ⇒ **b) is correct since the basilar membrane will achieve maximum vibration at some point, depending on the frequency of the tone. Hair cells in that vicinity will react and cause attached auditory nerves to spike.**

5. (2) Ignoring air absorption, frequency dependencies, nonlinearities, or spatial dependencies, what is the effect on the reverberation time (e.g., RT60) of doubling the absorption coefficient of a room's surfaces?

   a) increase the reverberation time for quiet sounds and decrease it for loud sounds

   b) double the reverberation time

   c) halve the the reverberation time

   d) reduce the reverberation time by a factor of 4.

   ⇒ **c) is correct since reverberation time is inversely proportional to absorption, so doubling the absorption coefficient reduces the reverb time by a factor of 2.**

6. (2) Three tones (100 Hz, 2000 Hz, and 7000 Hz) are presented monaurally over wideband headphones (40Hz -16 kHz $\pm$ 1 dB) to a young adult subject with normal hearing. In each case, the Sound Pressure Level (SPL) at the subject's ear is 40 dB. What would be the expected order of loudness for the tones, going from the loudest to the least loud?

   a) 100, 2000, 7000

   b) 2000, 100, 7000

   c) 2000, 7000, 100

   d) All would have roughly the same loudness.

   ⇒ **c) as was determined by the Fletcher-Munson curves. The main point is that the subjective sensation of "loudness", is a function of both frequency and intensity.**

7. (2) Dynamic programming has been applied to speech recognition for many years. One major advantage to this approach (as it has been commonly used) is

   a) The effects of different vocal tract lengths are normalized

   b) The effects of different durations for the same sounds are normalized

   c) The effects of different loudnesses for the same sounds are normalized

   d) The effects of post-nasal drip are normalized

   ⇒ **b) Dynamic programming has been applied to speech recognition primarily for the purpose of time normalization. It is used to find the best warping of the match between the input speech frames and the reference frames.**

8. (2) A classifier is trained on a set of examples that are labeled for class membership. The resulting classifier is used to estimate the class of each example in the training set, and is correct 90% of the time. The same classifier is then tested on a new set of examples that were not used in the training. For real-world data, the most likely result would be

   a) The test set performance would be at most 90%

   b) The test set performance would be at least 90%

   c) The test set performance would be close to zero

   d) The test set performance would depend on the eye of the beholder

   ⇒ **a) Performance on the training set is an optimistic estimate of the pattern classifier, since the training and test are done with the same examples. In some sense, the classifier is typically trained to be optimal for that group of examples. Examples from outside of the training set are likely to differ due to sampling variability, so the system trained on the first group is unlikely to do as well on the second. Thus, since the accuracy on the training set was 90%, the performance on the second is likely to be lower. It was fair to note that in any individual example it is possible to get "lucky" and have a higher test set accuracy than this, but this is not "the most likely result". Similarly, it is also possible to get very "unlucky" and get nearly a zero accuracy, but the system trained on the training set is in fact your current best guess of pattern characteristics, so this would not be the most likely result either.**

9. (6) Consider a digital filter consisting of two cascaded sections:

Section 1 is defined by the equation

$$v(n) = x(n) - x(n-K)$$

Section 2 is defined by the equation

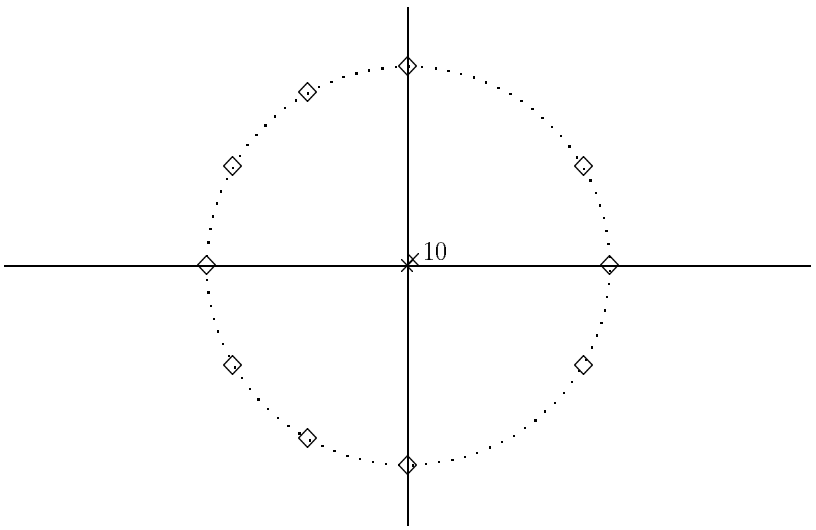
$$y(n) = 2cos(\theta)y(n-1) - y(n-2) + v(n)$$

a) Sketch the filter's poles and zeros for $K = 12$ and $\theta = 60$ degrees

$\Rightarrow$

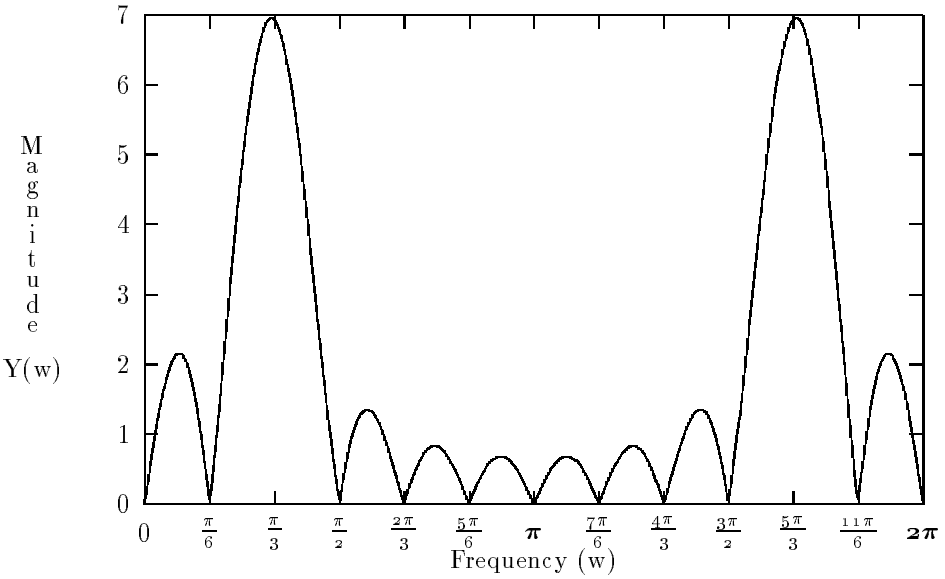
$$\begin{aligned} \boldsymbol{V(z)} &= \boldsymbol{X(z)(1 - z^{-K})} \\ \boldsymbol{Y(z)} &= \boldsymbol{\frac{1 - z^{-K}}{1 - 2\cos(\theta)z^{-1} + z^{-2}} X(z)} \end{aligned}$$

The numerator (Section 1) contributes K=12 zeros evenly spaced about the unit circle since these are the 12th complex roots of 1. It also contributes 12 poles at z=0 since when z=0, z^{-K} goes to ∞.
The denominator, from section 2, can be factored into $(1 - e^{-j\theta}z^{-1})(1 - e^{j\theta}z^{-1})$ and thus Y(z) has poles at $z = e^{-j\theta}$ and $e^{j\theta}$ and two zeros at z=0.
With θ = 60 degrees, the two poles in the denominator cancel the two zeros of the numerator located at $\pm$ 60 degrees and the two zeros in the denominator cancel two of the poles at the origin. Thus we are left with:



b) Sketch the frequency response, i.e. the amplitude of Y(z) evaluated on the unit circle, vs. ωT.

⇒ **Working from the pole-zero plot, as we trace around the unit circle, when we hit a zero, the magnitude spectrum goes to 0, and as we get further from the zeros or closer to poles, the magnitude spectrum increases. This will give a fine sketch of:**



If you wanted to plot it exactly (which you probably do not have time for in an exam), you can plug $z = e^{j\omega}$ into the eq of $Y(z)$ to get the amplitude response $\left|\frac{Y(e^{j\omega})}{X(e^{j\omega})}\right|$:

$$\frac{\boldsymbol{Y}(e^{j\omega})}{\boldsymbol{X}(e^{j\omega})} = \frac{1 - e^{-j\omega K}}{1 - 2\cos(\theta)e^{-j\omega} + e^{-2j\omega}} = \frac{1 - e^{-j\omega K}}{(1 - e^{-j\theta}e^{-j\omega})(1 - e^{j\theta}e^{-j\omega})}$$

$$= \frac{1 - e^{-j\omega K}}{(1 - e^{-j(\theta+\omega)})(1 - e^{j(\theta-\omega)})}$$

$$\left|\frac{\boldsymbol{Y}(e^{j\omega})}{\boldsymbol{X}(e^{j\omega})}\right| = \sqrt{\frac{\boldsymbol{Y}(e^{j\omega})}{\boldsymbol{X}(e^{j\omega})}\left(\frac{\boldsymbol{Y}(e^{j\omega})}{\boldsymbol{X}(e^{j\omega})}\right)^*}$$

$$= \sqrt{\frac{1 - e^{-j\omega K}}{(1 - e^{-j(\theta+\omega)})(1 - e^{j(\theta-\omega)})}\frac{1 - e^{j\omega K}}{(1 - e^{j(\theta+\omega)})(1 - e^{-j(\theta-\omega)})}}$$

$$= \sqrt{\frac{2 - 2\cos(K\omega)}{(1 - e^{-j(\theta+\omega)})(1 - e^{j(\theta-\omega)})(1 - e^{j(\theta+\omega)})(1 - e^{-j(\theta-\omega)})}}$$

$$= \sqrt{\frac{2 - 2\cos(K\omega)}{4 - 2e^{j(\theta-\omega)} - 2e^{-j(\theta-\omega)} - 2e^{j(\theta+\omega)} + e^{2j\theta} + e^{2j\omega} - 2e^{-j(\theta+\omega)} + e^{-2j\theta} + e^{-2j\omega}}}$$

$$= \sqrt{\frac{2 - 2\cos(K\omega)}{(4 + 2\cos 2\theta) + (8\cos\theta)\cos\omega + (2)\cos 2\omega}}$$

$$= \sqrt{\frac{2 - 2\cos(12\omega)}{3 - 4\cos\omega + 2\cos 2\omega}}$$

10. (4) Define the following terms (1 sentence per definition):

a) interval histogram

⇒ **An interval histogram is a histogram of the times between successive neural spikes for a given neuron. For example, a pure or complex tone is repeated until a reasonable number of spikes are obtained.**

b) efferent

⇒ **Efferent is used to describe neurons projecting from deeper in the brain towards lower levels — a form of feedback.**

c) stapes

⇒ **The stapes is the third bone in the middle ear that strikes the oval window of the inner ear to excite the inner ear.**

d) oval window

⇒ **The oval window is the outer membrane of the cochlea that, when struck by the stapes, sets the cochlea to vibrating.**

11. (4) Show that the constraint u(x,t) = w(x)v(t) can be applied to the 1-D wave equation, and that exponentials are a solution to the resulting form. Recall that the 1-D wave equation is

$$c^2\left(\frac{\partial^2 u}{\partial x^2}\right) = \frac{\partial^2 u}{\partial t^2}$$

$\Rightarrow$ **This can be shown as follows:**

$$\frac{\partial u}{\partial x} = \frac{\partial w(x)}{\partial x}v(t) + \frac{\partial v(t)}{\partial x}w(x) = w'(x)v(t) + 0$$

$$\frac{\partial^2 u}{\partial x^2} = w''(x)v(t)$$

$$\frac{\partial u}{\partial t} = \frac{\partial w(x)}{\partial t}v(t) + \frac{\partial v(t)}{\partial t}w(x) = 0 + w(x)v'(t)$$

$$\frac{\partial^2 u}{\partial t^2} = w(x)v''(t)$$

$$\textbf{So} \quad c^2 w''(x)v(t) = w(x)v''(t)$$

$$\textbf{Now take} \quad w(x) = Ae^x$$
$$\textbf{and} \quad v(t) = Be^{ct}$$

$$\textbf{then} \quad w''(x) = Ae^x$$
$$\textbf{and} \quad v''(t) = c^2 Be^{ct}$$

And:

$$c^2 Ae^x Be^{ct} = Ae^x c^2 Be^{ct}$$

12. (4) Give two dimensions of difficulty for a speech recognition task. For each dimension, describe an easier and a harder example, and explain (in one sentence) the difference in difficulty.

⇒ **Obviously there are many "dimensions of difficulty" for ASR, and anything reasonable was accepted. Some of the common ones are:**

Speaker-dependent vs speaker independent - the latter is harder, as it refers to training on some speakers and testing on others; The recognizer must be trained to ignore speaker-specific characteristics that are irrelevant to the message.

Vocabulary size - small is easier, large is harder; for more words, acoustically confusable words are more likely.

Perplexity - low perplexity is easier, high perplexity is harder - a very constrained task may have low perplexity (uncertainty about the next word) even if the vocabulary is large, and in that case the acoustic classifier does not have to be as accurate to get the right answer.

Isolated vs. continuous - words separated by pauses are easier to recognize than (for the same task) words spoken continuously; in the latter case, there is coarticulation across words and the pronunciations and spectral content are changed.

Speaking style - read vs conversational; speech that is read from a written text is easier to recognize than speech from a natural conversation. Speech in the latter case tends to have many phone deletions, slurred sounds, varied pronunciations, and more rapid speech with increased coarticulation effects.

Environmental noise - low noise is easier, high noise is harder. Particularly if the same noise is not present during training, the speech features used for recognition are a worse match to the training set; even if noisy data are available during training, more noise tends to blur the boundaries between potentially confusable sounds.

Reverberation - low reverberation time (and energy) is easier, high reverberation is harder. Longer reverberation times cause smearing between neighboring phonetic elements.

Microphone type - a close microphone is easier and a distant microphone is harder, due to increased noise and reverberation for the latter case. It is also easier if the training and testing microphones are the same.

13. (6) Let $\mathbf{x}$ be a discrete random variable, and let a and b be two classes that examples corresponding to $\mathbf{x}$ can belong to. Further, let the 2 class-conditional densities be known:

$$p(\mathbf{x}|a) = \begin{cases} \frac{1}{8} & \text{for } 1 \leq \mathbf{x} \leq 8 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$p(\mathbf{x}|b) = \begin{cases} \frac{1}{4} & \text{for } 7 \leq \mathbf{x} \leq 10 \\ 0 & \text{elsewhere} \end{cases}$$

Finally, the class priors are also known, and are $p(a) = \frac{3}{4}$ and $p(b) = \frac{1}{4}$.

Find the optimal decision rule to decide on class a or b given $\mathbf{x}$.

⇒ **Here the point was to use the Bayes Decision rule, in which you pick the class with the largest posterior (in this case, pick a if $p(a|x) > p(b|x)$, and pick b if $p(b|x) > p(a|x)$.) An equivalent formulation is based on the Bayes' transformation:**

$$p(a|x) = \frac{p(x|a)p(a)}{p(x)}$$

and looking at whether

$$p(x|a)p(a) > p(x|b)p(b)$$

is true (since p(x) is the same irregardless of class) or vice versa. Computing, we get:

$$p(x|a)p(a) = \frac{3}{32} \text{ for } 1 <= x <= 8 \text{ and}$$

$$p(x|b)p(b) = \frac{2}{32} \text{ for } 7 <= x <= 10$$

(Both products are zero outside of these ranges.)
So class a "wins" if $1 <= x <= 8$, and class b "wins" for $9 <= x <= 10$. This is the optimal (Bayes) decision.