11-751 Speech Recognition and Understanding
Homework 1 Solution

Released: Monday, Sep 16, 2013
Due: Monday, Sep 30, 2013 before class

Problem 1. General Spoken Language Processing, Speech Production (3 * 9 + 5 * 4 = 47 points)

1. Give three examples in which speech recognition (SR) and/or speech understanding are useful to apply. Motivate why this is the case?

   1) In hand/eye busy situations, such as driving.
   2) In situations where the keyboard or mouse is not a suitable input device, such as PDA.
   3) Speech translation systems for visitors in foreign countries.

2. Show three challenges for speech recognition, and explain why.

   1) Large vocabulary and continuous speech;
   2) Variability across speakers and within speaker;
   3) Noise and channel effects;
   4) Ambiguity of speech; etc.

3. Explain the difference between a phone and a phoneme. Give an example of two phones that are considered the same phoneme in English.

A phone is an actual sound that people make. A phoneme is the smallest speech unit that differentiates meaning in a given language. Phonemes are language dependent, while phones are not. Phones are realizations of phonemes. A phoneme may have multiple realizations (depending on the context and the speaker) that do not differentiate meaning; phones that realize the same phoneme are called “allophones” of the phoneme.

The aspirated voiceless [t^h] (as in “top”) and the unaspirated voiceless [t] (as in “stop”) are different phones, but they are allophones of the same phoneme /t/ in English.

Note: Phones are usually written in [brackets], while phonemes are usually written in /slashes/.

4. Why are vowels important in speech recognition?

Vowels are generally long in duration and are spectrally well defined. As such, they can be more easily and reliably recognized.

5. Consonants are usually very short and hard to recognize. They can be classified by the manner and place of articulation. Give 1~2 examples for each category listed below: plosive, fricative, affricate, nasal, lateral, retroflex, and glide.

Below are the English consonants divided into the categories in the question, in Arpabet notation:

Plosive: P B T D K G
Fricative: F V TH DH S Z SH ZH HH (TH as in “thin”, DH as in “these”, ZH as in “leisure”)
Affricate: CH JH (JH as in “jeep”)
Nasal: M N NG
Lateral: L
Retroflex: none
Glide: Y W

(Technically speaking, the English R sound /   / is an “alveolar approximant” and not a retroflex. But I’ll also accept it if you classify it as a retroflex.)

6. Thinking intuitively, what information do we use, as humans, to help ourselves interpret speech? Two examples would be knowledge about the phonemes that are allowed to follow each other in the language (e.g. /pf/ is illegal in English), and knowledge of a speaker’s accent or dialect. Please name at least three other factors.

1) Knowledge of the speaker;
2) Lip movements and body language;
3) Grammar of the language;
4) Semantic context;
5) Situation context;
6) Common sense or word knowledge; etc.

7. Give a short answer to the following questions:

1) What is the fundamental frequency (F_0)?
2) What are formants?
3) Is there any relationship between the fundamental frequency and the formants? Why or why not?
4) What kind of voice characteristics does F_0 reflect?

1) The fundamental frequency is the number of open-close cycles of the vocal cords per second.
2) Formants are the resonant frequencies of the vocal tract.
3) No. Because F_0 and formants are determined by the vocal cords and the vocal tract respectively, they are independent.
4) F_0 reflects the pitch of the voice.

8. When spelling in English over a telephone line, which pairs of letters are the most confusable? Why is that? (Hint: Consider the frequency response of a telephone line, and the frequency range of the sounds used to spell letters)

The telephone line is band-pass filter that only lets through frequency components between 300 and 3400 Hz. Many of the fricatives of English have most of their energy in the frequency range above 3400 Hz, and letters spelled with these fricatives will be confusable. Examples include F /ef/ – S /es/, and C /si:/ – Z /zi:/.

Other answers, such as M /em/ – N /en/, B /bi:/ – P /pi:/, D /di:/ – T /ti:/, are also acceptable, although the reason for their confusability is not the frequency response of the telephone line.

9. For the following phoneme classes, explain briefly how they are produced: vowels, diphthongs, nasals, stops, and fricatives.

Vowels are made by keeping the vocal tract open and fixed at a certain shape, while the vocal cord vibrates periodically.

Diphthongs are made by shift the shape of the vocal tract from that of one vowel to that of another, while the vocal cord vibrates periodically.

Nasals are made by lowering the velum and blocking the air flow in the oral cavity, and allowing air to pass through the nasal cavity.

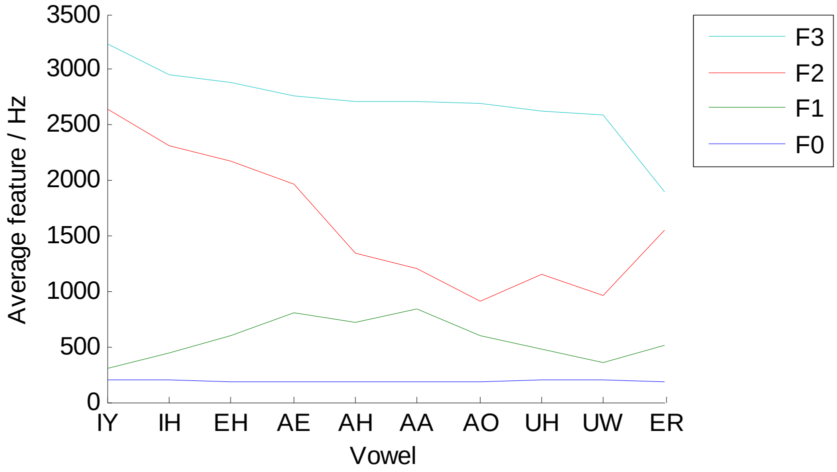
Stops are made by blocking the vocal tract at a certain point, and then suddenly releasing it, letting air flow pass through.

Fricatives are made by restricting the vocal tract at a certain point, but not blocking it completely, so a turbulent air flow forms at the restricted point.

10. Now let’s study the “Peterson-Barney” vowel database (see the attached file `pb.tgz`). In this database, there are 10 vowels from 76 speakers (33 men, 28 women and 15 children). Each vowel is represented by four features: the fundamental frequency (F_0) and the first three formant frequencies (F_1, F_2, F_3). The 10 vowels are listed as below:

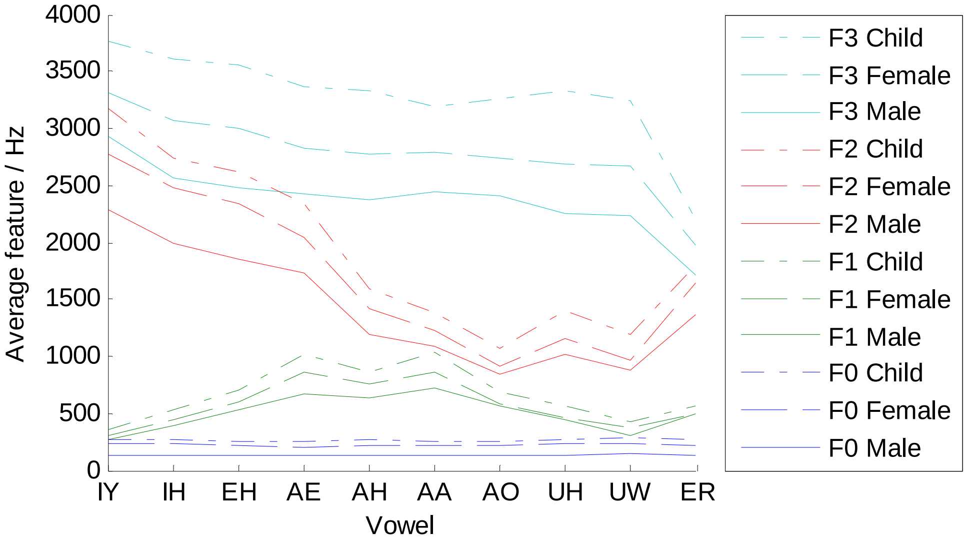
Arpabet	IY	IH	EH	AE	AH	AA	AO	UH	UW	ER
IPA	[i]	[ ]	[e]	[æ]	[ ]	[a]	[o]	[ ]	[u]	[ ]

a. Compute the average fundamental frequency and formant frequencies for each of the vowels, and present the results in a figure with the vowels on the horizontal axis and frequency as the vertical axis. Your figure should contain four lines, one line for each feature.



b. Compute the average fundamental frequency and formant frequencies for each vowel and each group (male, female, children – read the header file for their respective speaker labels), and present the results in a figure like in question (a). Your figure should contain 12 lines, one for each combination of vowel and group. Use colors, line types and/or a legend to indicate clearly

which line corresponds to which combination.
What differences do you observe in the figure between the different groups? Explain why.

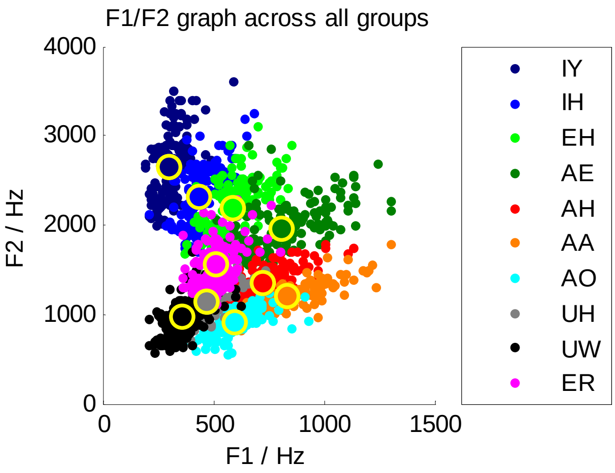


Both the fundamental frequency and the formants are highest for children, then for women, and lowest for men. This is because the size of children’s vocal organs is smaller than that of women, and the size of women’s vocal organs is smaller than that of men.

c. Look at the figure drawn in question (a). If we are building a vowel recognizer, which two features of the four do you think will be the most informative? Why?

F_1 and F_2 are the most informative features, because they vary a lot across different vowels.

d. Using the two features you selected in question (c) as the two axes, plot the average of each vowel as a point in the coordinate system. These are the positions of the vowels in a 2-D feature space. Name three pairs of vowels that are most confusable.



This figure not only plots the average of each vowel (big dots), but also the features for each instance in the database (small dots). The most confusable pairs of vowels are IH-EH, AH-AA, and UH-UW.

Problem 2. Classification (6 + 3 + 5 = 14 points)

Suppose we have 3 classes of data points on the *x*-axis:

Class Alpha: {-2, -1, 0, 1, 2}
Class Beta: {-9, -8, -7, 4, 6, 8}
Class Gamma: {-12, -11, -9, -5, 3, 11}

1. (KNN classifier) Classify -5 and 3 using the KNN classifier with K = 1, 3, 7 respectively.

K **Point**	**1**	**3**	**7**
-5	Gamma	Alpha, or Beta, or Gamma	Beta
3	Gamma	Alpha, or Beta, or Gamma	Alpha

2. (Gaussian classifier) What are the mean and variance of each of these three classes?

Class	**Mean**	**Variance (*n*)**	**Variance (*n*-1)**
Alpha	0.00	2.00	2.50
Beta	-1.00	50.67	60.80
Gamma	-3.83	68.81	82.57

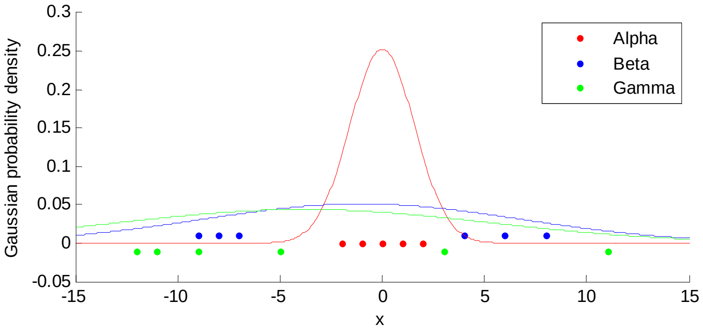
Note: There are two formulas for calculating the variance, differing only in the denominator: $\operatorname{var} X = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\operatorname{var} X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Both formulas are accepted.

3. Now, classify -5 and 3 again according to the Gaussian distributions: classify a point to the class that yields the largest probability density value at that point. Do you get the same result as the KNN classifier? Why or why not?

Point	**Gaussian density values**			**Classification Result**
	Alpha	**Beta**	**Gamma**	
-5	0.0017	0.0449	0.0435	Beta
3	0.0417	0.0449	0.0331	Beta

Note: The density values in the table are calculated with the variance whose denominator is *n*-1. If you used *n*, the density values would differ, but luckily the classification results stay the same.

The classification results of the Gaussian classifier and the KNN classifier are quite different. Even KNN classifiers with different values of K give totally different results. One reason for this is that the distributions of the Beta and Gamma



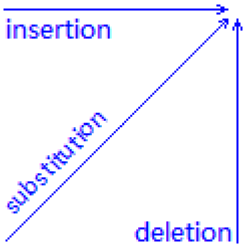
classes are split and widespread. This makes the KNN classifier very sensitive to the value of K, and the single-modal Gaussian distribution a poor model for their distributions.

Problem 3. Dynamic Time Warping (DTW) (3 * 8 = 24 points)

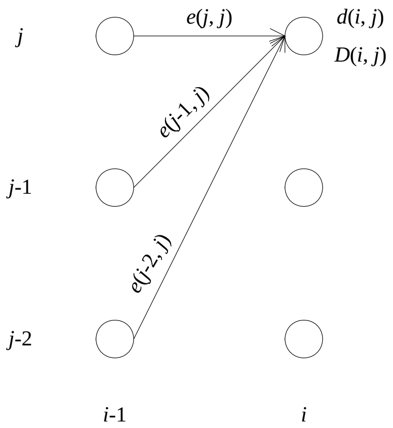
1. What kind of errors can a speech recognizer make? Can word error rate (WER) be higher than 100%? Explain.

   The errors a speech recognizer can make are insertion, deletion and substitution.
   WER can be higher than 100% when a lot of redundant words are inserted into the hypothesis.

2. Assuming the decoded sentence is on the *x*-axis and the reference sentence is on the *y*-axis, draw the path constraint diagram for WER computation. Indicate which path corresponds to which type of error.



3. Let's look at an actual path constraint diagram used for matching an utterance against a template (both represented by a sequence of feature vectors). The utterance is placed on the *x*-axis, and the template on the *y*-axis. Under these path constraints, each frame of the utterance must be match to one and only one frame of the template. Matching frame *i* of the utterance with frame *j* of the template incurs a matching cost $d(i, j)$, and matching two successive frames of the utterance to the frames j_1, j_2 of the template incurs a transition cost of $e(j_1, j_2)$. The accumulated cost up to the position (i, j) is denoted by $D(i, j)$. Write down the dynamic programming recursion formula for $D(i, j)$.



$$D(i,j) = \min \begin{Bmatrix} D(i-1, j) + e(j, j) \\ D(i-1, j-1) + e(j-1, j) \\ D(i-1, j-2) + e(j-2, j) \end{Bmatrix} + d(i, j)$$

4. Write down the pseudo-code to perform DTW based on the above path constraints, and the extra constraint that the first and last frames of the utterance must be matched to the first and last frames of the template. The utterance has L frames, and the template has M frames. Your code should return the cost of the optimal path, and the path f_1 ... f_L itself, where f_i is the number of the frame in the template that is matched with frame *i* of the utterance.

```
// Initialization
D(1,1) = d(1,1); D(1,2:M) = +inf;
// Recursion
for i = 2 to L
   for j = 1 to M
      D(i,j) = $\min_{j-2<=k<=j,k>=1}$ {D(i-1,k) + e(k,j)} + d(i,j);
      prev(i,j) = arg $\min_{j-2<=k<=j,k>=1}$ {D(i-1,k) + e(k,j)} + d(i,j);
   end
end
// Traceback
pos = M; f(L) = M;
for i = L downto 2
   pos = prev(i, pos);
   f(i-1) = pos;
end
return optimal cost D(L,M) and optimal path f(1:L);
```

5. Assume you are doing isolated word recognition, and you have N templates. Given an input utterance, how do you decide which word it is?

Match the input utterance against all the templates using the DTW algorithm, and choose the template that yields the smallest matching cost.

6. What is the time complexity of DTW?

The time complexity is $O(LM)$ for one template, and $O(LMN)$ for all the templates.
(Answering either one is accepted)

7. Pruning is a common strategy to speed up the DTW procedure. Describe the implementation of beam search in DTW.

After calculating each column of the D array, find the minimum accumulated cost in that column, say $D(i, j)$. If any accumulated cost in that column, say $D(i, k)$, is larger than $c * D(i, j)$ (c is a constant), then set $D(i, k)$ to +inf, which mean the node (i, k) is pruned.

Note: The key of beam search is to decide which nodes to keep in each column based on the **best** node in that column. It is okay to use $c + D(i, j)$ instead of $c * D(i, j)$. It is also okay to keep a fixed number of nodes in each column. It is **not** okay, however, to keep only the nodes close enough to the diagonal, because this is not based on the best node in each column.

8. Describe how DTW can be extended to continuous speech recognition.

We need to put the templates of all the words in the vocabulary on the y-axis, and allow transitions from the last frame of any word to the first frame of any word. As for the boundary conditions, we require that the first frame of the utterance be matched with the first frame of any template, and that the last frame of the utterance be matched with the last frame of any template.

Problem 4. Digital Signal Processing with Matlab (3*2 + 4.5*2 = 15 points)

Matlab is a popular tool for providing visual representations of mathematical functions. This problem will hopefully show that there is nothing magic about the front-end DSP in a speech recognition system: you can easily reproduce the effects by using a program like Matlab. By the way, this problem reproduces the “FFT for Spectral Analysis” demo in Matlab; you can find the demo in the help.

We're going to create a 0.25s-long signal sampled at 1000 Hz, so we create a sequence of time points from 0 to 0.25 with a step of 0.001.

```
    >> t = 0:.001:.25;
```

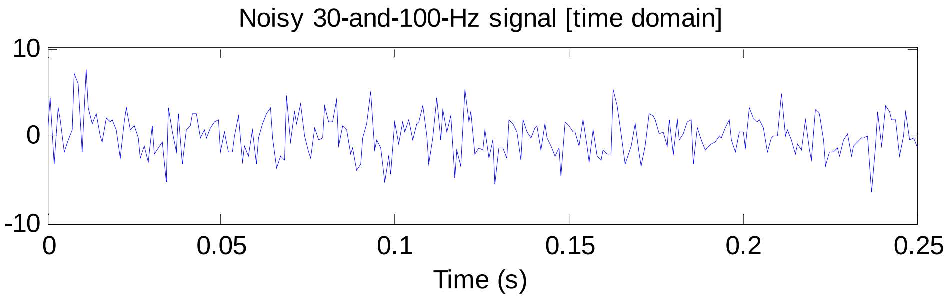
Next, we can form a signal consisting of two frequency components: 30 Hz and 100 Hz. Let's do it step by step.

```
    >> x30 = sin(2 * pi * 30 * t);
    >> plot(t,x30), title('Pure 30-Hz signal [time domain]')
    >> x100 = sin(2 * pi * 100 * t);
    >> plot(t,x100), title('Pure 100-Hz signal [time domain]')
    >> x = x30 + x100;
    >> plot(t,x), title('Pure 30-and-100-Hz signal [time domain]')
```

Now we add some random noise with a zero mean and standard deviation of 2 to produce a noisy signal *y*:

```
    >> y = x + 2 * randn(size(t));
    >> plot(t,y), title('Noisy 30-and-100-Hz signal [time domain]')
```

1. Include the plot of the noisy 30-and-100-Hz signal in your homework.



Clearly, it is difficult to identify the frequency components by looking at the original signal; that's why spectral analysis is so popular. It is easy to find the spectrum of the noisy signal *y* using the discrete Fourier transform. Let's take the 256-point fast Fourier transform (FFT):

```
    >> Y = fft(y,256);
```

The power spectral density, a measurement of the energy at various frequencies, is found using the complex conjugate function:

```
>> PY = Y .* conj(Y) / 256;
```

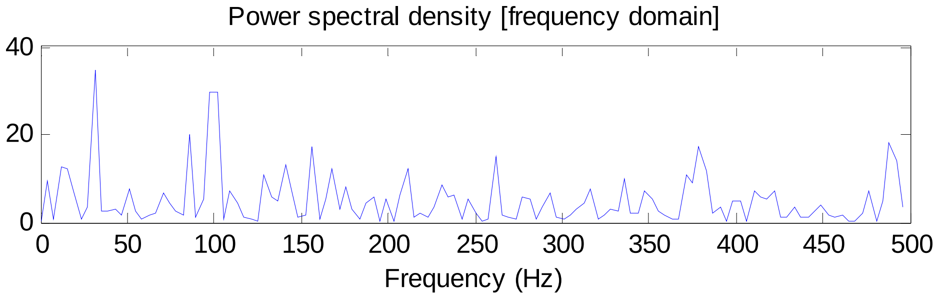
To plot the power spectral density, we must first form a frequency axis:

```
>> f = 1000/256*(0:128);
```

We do this for the first 129 points (the remaining points are symmetric to the 2~128th points). We can now plot the power spectral density:

```
>> plot(f,PY(1:128)), title('Power spectral density [frequency domain]')
>> xlabel('Frequency (Hz)')
```

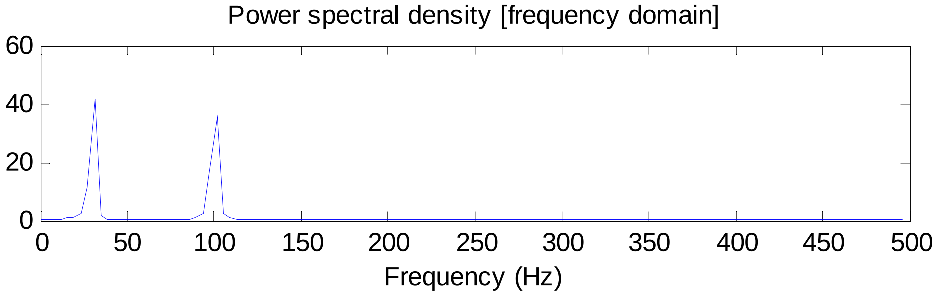
2. Include the plot of the power spectral density in your homework.



3. The spectrum, *Y*, is complex. That is, it has a real part and an imaginary part, or a magnitude and a phase. Which of these have a physical meaning, and how would you interpret them?

The magnitude and phase of the spectrum at a certain frequency point correspond to the amplitude and phase of that frequency component in the signal. The real and imaginary parts of the spectrum do not have a direct physical meaning.

4. Try plotting the power spectral density of the clean 30-and-100-Hz signal without noise. How does the time-domain noise affect the power spectral density?



The noise in the time domain imposes a noise across all frequencies in the power spectral density.