# 11-751 Speech Recognition and Understanding
# Homework 1

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

## Problem 1. General Spoken Language Processing, Speech Production

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

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

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

4. Why are vowels important in speech recognition?

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.

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 two other factors.

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?

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)

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

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] | [I] | [e] | [ae] | [^] | [a] | [o] | [U] | [u] | [3] |

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.

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?

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.

## Problem 2. Classification

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. Use majority voting to determine class labels.

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

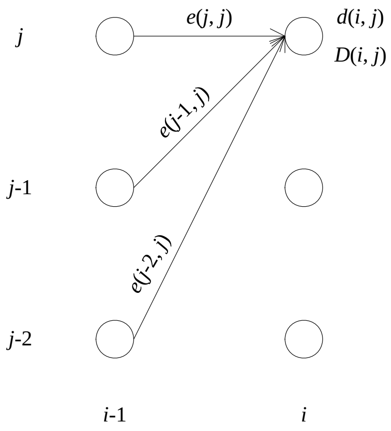
3. Now, classify -5 and 3 again according to the Gaussian distributions: classify a point to the class that yields the largest probability density value at that point. Do you get the same result as the KNN classifier? Why or why not?

# Problem 3. Dynamic Time Warping (DTW)

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

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 $D(i, j)$.



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 $I$ frames, and the template has $J$ frames. Your code should return the cost of the optimal path, and the path $f_1 \dots f_I$ itself, where $f_i$ is the number of the frame in the template that is matched with frame $i$ of the utterance.

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?

6. What is the time complexity of DTW?

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

8. Describe how DTW can be extended to continuous speech recognition. (Hint: how can you recognize an utterance with multiple words, rather than a single word?)

# Problem 4. Digital Signal Processing with Matlab

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.
If Matlab is not on your machine, please download the installation file from:
http://www.cmu.edu/computing/software/all/matlab/download.html

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?

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?