11-751 Speech Recognition and Understanding

Nikolas Wolfe
Homework 1
September 30th, 2013

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?

   Overall motivation for SR: Vocalized speech is the most efficient and effective natural communication tool possessed by human beings. It is therefore of great interest to find a way to leverage the power of modern computers to take advantage of the human ability to produce speech information through recognition and, if possible, understanding.

   Examples of useful SR applications:

   1.) Speech-To-Text (STT) Applications: Dictation, STT chatting on Google/Facebook, Siri

   Motivation: Allows people to be hands-free while having their words and thoughts seamlessly converted to text, which is a far more liquid form of information in terms of space and the ease of search.

   2.) Language-Learning Applications and Pronunciation Scoring:

   Motivation: It is in general far more difficult to learn to speak a language than to write one, and while computers are pretty good at allowing humans to learn language on their own through written text, one thing that humans still do far better than machines is pronunciation training. Building SR with the ability to accurately and reliably score pronunciation would thus be a great leap forward in the effort to make entirely computationally based language training software.

   3.) Voice Translation Applications, Google Translate, Jibbigo et al

   Motivation: Machine translation of textual information is an obviously desirable technology. It allows books and websites and other materials to be translated into foreign languages, thereby becoming accessible to a wider audience. But being able to do the analogous task for voice information is vastly more difficult, although equally desirable given the proliferation of multimedia information in the world today, specifically voice and auditory information. And the fundamental first obstacle to the automatic translation of auditory information, of course, is Speech Recognition.

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

   1.) **Noise**: Recorded speech signals are often accompanied by a great deal of background noise, often in the same frequency range as human voice. Separating human voice from background noise is difficult and often results in the loss or corruption of speech information.

   2.) **Spontaneous/Continuous Speech:** When people speak naturally they do not take into consideration the way they blend words together (e.g. "whaddya doin?") or make phonological alterations in the way they say words (e.g. Pittsburgese: "Sammiches, Stillers, Arns"), which often makes it difficult to determine the words being said, or the word segmentation (e.g. distinguishing when someone says, "Wreck a nice beach" vs. "Recognize speech."). And this doesn't even touch the issue of code-switching between different languages in spontaneous speech, (e.g. "Saying, 'How are you?' in Spanish is 'Como estas?'")

   3.) **Computational Intensity of SR:** There are a lot of words out there, a lot of people saying them, a lot of phonemes, and a lot of phonological variations. Getting enough information about a given language to develop accurate (and fast) SR is both time consuming and difficult – as well as computationally intense, at least for one computer. Currently, projects like IARPA Babel are trying to cut down the amount of speech

data required to train recognizers (which are built on extremely powerful computer clusters, no less) but the current standard for "how little" voice data is required to build and train a speech recognizer is about 10 hours of recorded speech in a given language. One of the ways companies like Google are building robust SR is by dividing up the task of computing the most likely sequence of words among many millions of distributed machines, and simultaneously using the many thousands of user-contributed hours of voice data to train their recognizers. But this is obviously hugely expensive and a monumental amount of work to accomplish something that a single human brain seems to do so effortlessly.

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 a speech sound, an individual segment of sound produced or uttered. A **phoneme** is the smallest individual unit of meaning in a given language's phonology, which is either composed of or maps to a set of individual **phones**. As Gimson (2008) writes, "[phonemes are] the smallest contrastive linguistic unit which may bring about a change of meaning." Phonemes are the abstractions of phonological units, whereas phones are what is actually uttered or spoken. Sometimes phones which are distinct can map to the same phoneme, such as the aspirated and voiced versions of the English phoneme /p/, e.g. the way the letter *p* is pronounced in the word *pin* and in the word *spin*. In the word *pin* the /p/ is aspirated, and in the word *spin* it is not. When two phones map to the same phoneme, they are called **allophones**.

   Phonemes can also be composed of several phones, such as the digraphs *gb* and *kp* which appear in many West African languages such as Ewe, Dangme, Konkomba, and Dagbani. In these digraphs, the speaker is actually pronouncing (to some degree) two individual phones simultaneously, or indistinguishably closely together. In the digraph *kp*, for instance, the speaker will shape their mouth in such a way as to produce the phone [k] but will voice the phone [p], such as the word *Kpandai* (which sounds like 'pandai' to English speakers), the name of a particular city in Ghana.

4. Why are vowels important in speech recongnition?

   Vowels are the anchors in speech around which words can be distinguished from noise. Because of the way vowels are produced, (where the air-flow through the vocal tract is uninhibited) they tend to be louder than consonants and are therefore easier to distinguish. Furthermore, vowels can be easily identified by their regular formant frequency signatures – a feature lacking in most consonant frequency signatures, which are distinguished in many cases by bursts of high frequency noise. Finally, syllables tend to form around vowels. When voice recognizers are built, it is generally impossible to model every single word in a language. If voice recognizers operated at the word level, they would be incapable of recognizing out-of-vocabulary words. What recognizers actually do is look for phonemes and/or syllables in a given language, which in many cases are distinguished by their accompanying vowel sounds. So, vowels provide the markers from which the syllabic and phonological construction of words in a language can be derived. Without vowels, it would be exceedingly difficult to develop acoustic models of language.

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.

Consonants	**Examples**
Plosive	[p], [b]
Fricative	[f], [s], [z]
Affricate	English: *ch,* [j]
Nasal	[n], [m]
Lateral	[l]
Retroflex	Swedish: [rd], as in *nord*
Glide	[w], [j]

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.

    1.) Environmental factors such as visual cues, body language, and other external information have a significant effect on limiting the things we expect to hear when we interpret speech. For example, supposing were to find ourselves being chased down by a tornado and a friend were trying to yell something at us while flailing his arms amidst the roaring winds, we are much more likely to interpret the word "run!" instead of, say, "fun."

    2.) Contextual cues are often used to pick up the meaning of speech. When we try to understand people speaking in a different language than we are used to, we will very often not understand every word that is being said, but will rather try to piece together the semantics from the context, that is, the words we actually *do* understand. For example, if someone is speaking to us about an event for which we have no present environmental reference cues, but we detect words that sound like "marry," "daughter", and "reception," we can make the reasonable guess that the person is talking about a wedding.

7. Give a short answer to the following questions:

    1.) What is the fundamental frequency (F0)?
    The fundamental frequency is the lowest frequency present in a spoken voice, between 85 and 180 Hz for an adult male and 165 to 255 Hz for an adult female

    2.) What are formants?
    Gunnar Fant (1960): "The spectral peaks of the sound spectrum $|P(f)|$ are called *formants*"
    Formants are the meaningful component frequencies that help distinguish vowels.

    3.) Is there any relationship between the fundamental frequency and the formants? Why or why not?
    The fundamental frequency is the lowest frequency in any particular voice. Formants are the frequencies which are above it, either as integer multiples (harmonics) or not, and the combination of the formants and the fundamental frequency are called **partials**.

    4.) What kind of voice characteristics does F0 reflect?
    F0 is the lowest component frequency of the **vowels** in human 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 typical frequency response of a telephone line is in the range of 300 Hz – 3000 Hz. Letters which are either consonant-sounding when they are pronounced or involve very short consonant sounds followed by loud vowels are likely to be the most difficult to distinguish over a telephone line, because certain distinguishing phonological features of consonants are at or above 3000 Hz. For instance, nasal consonants like *M* and *N* are likely to be confused because they contain formants near and above 3000 Hz. In a paper by Ronald Cole and Mark Fanty on "Spoken Letter Recognition," it was determined that the letters B, M, N, T, and P were the most likely letters to be confused with something else. Common mistakes were combinations like M vs. N, B vs. D, B vs. P, and C vs. Z. In cases like C vs. Z, B vs. P, and C vs. Z, the consonant sounds are both short and have similar spectra. This makes them likely to be confused because the frequency response of a telephone line does not always capture them correctly.

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

    Vowels: Produced by opening the vocal tract, allowing air to flow freely past the glottis

Diphthongs: A combination of consecutive vowel sounds produced by altering the position of the tongue gradually while vocalizing, e.g. in the word "sign," there is no distinctive [g] sound, it's rather pronounced like "siyn," and the two vowel sounds [i] and [e] are heard in succession.

Nasals: Formed by passing air through the nasal cavity and blocking the passage of air through the lips, using the oral cavity as a resonance chamber.

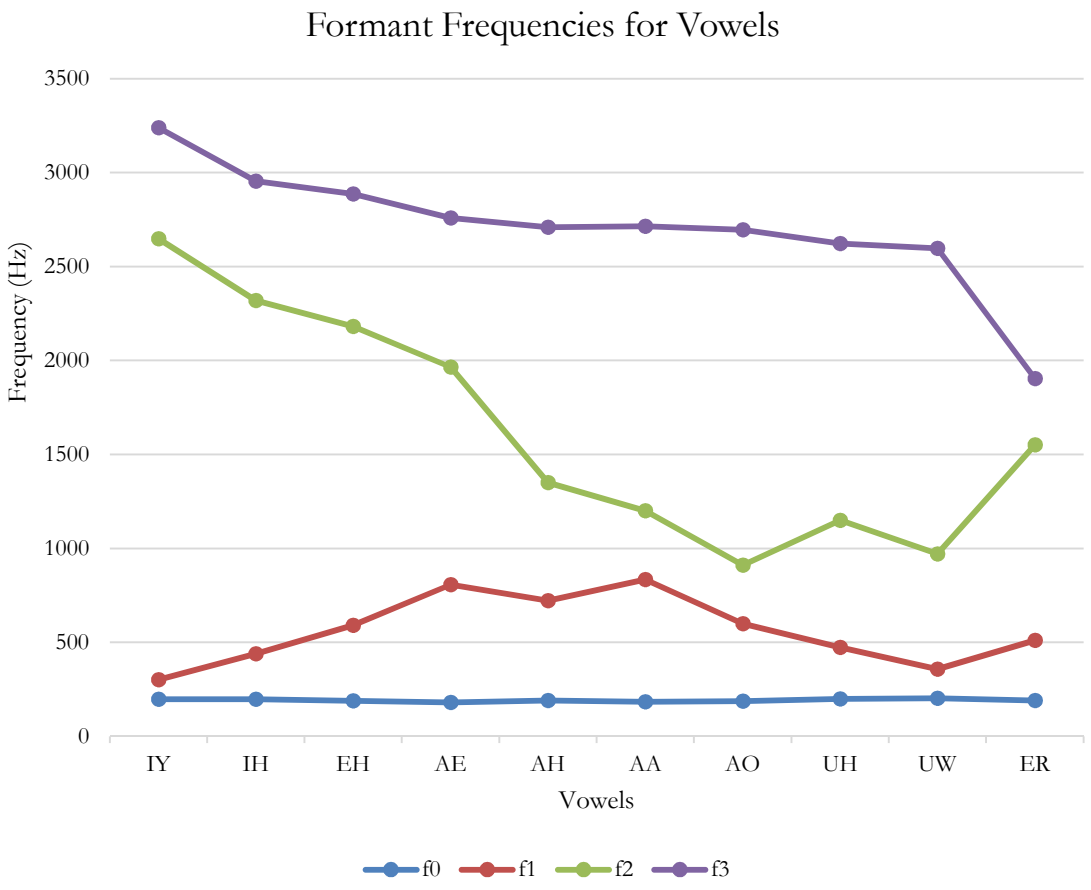
Stops: A plosive – a burst of air through the lips – which is immediately stopped by the glottis or tongue.

Fricatives: Formed by creating turbulence in mouth/oral cavity by forcing air through a narrow channel, usually created by positioning the tongue against the teeth or some other part of the velum/muscular palate.

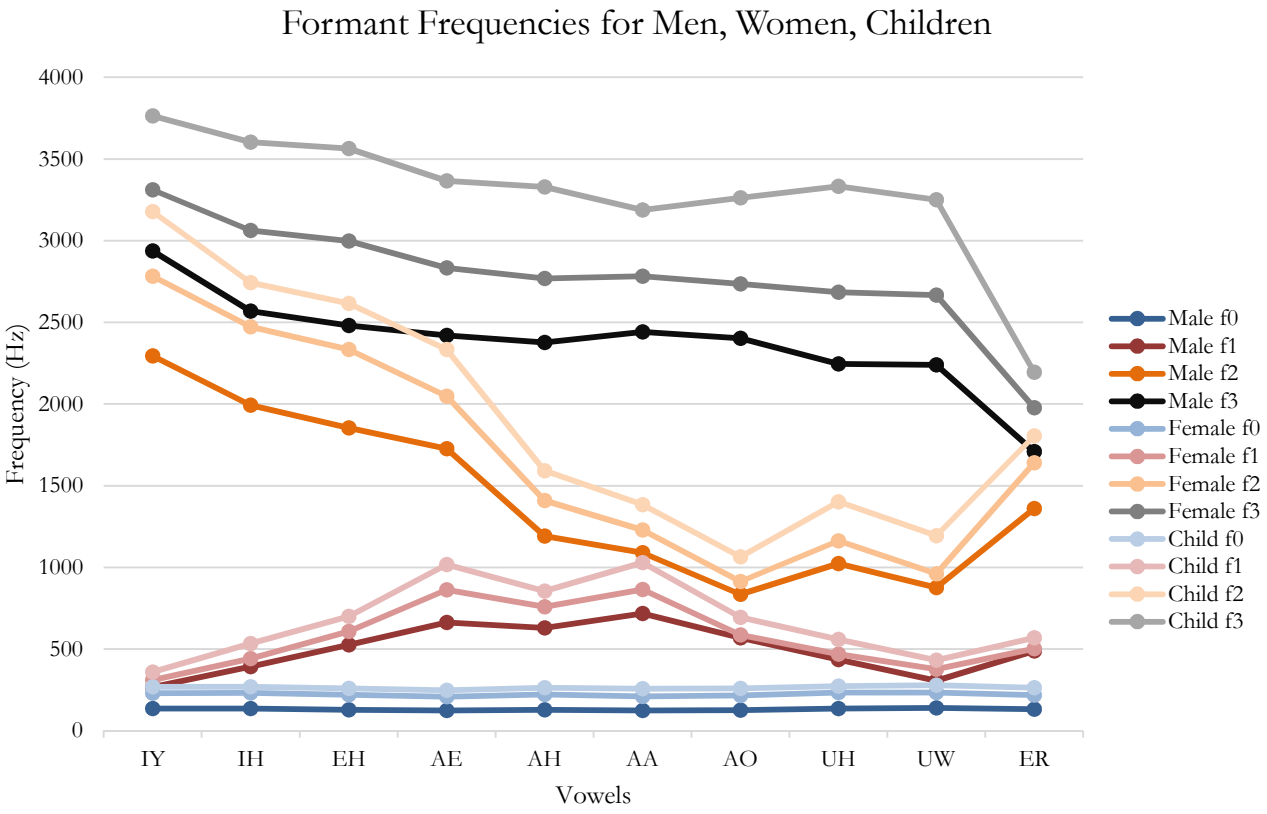
10. Now let's study the "Peterson-Barney" vowel database. In this database, there are 10 vowels from 76 speakers (33 men, 28 women and 15 children). Each vowel is represented by four features: the fundamental frequency (F0) and the first three formant frequencies (F1, F2, F3). The 10 vowels are listed 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.

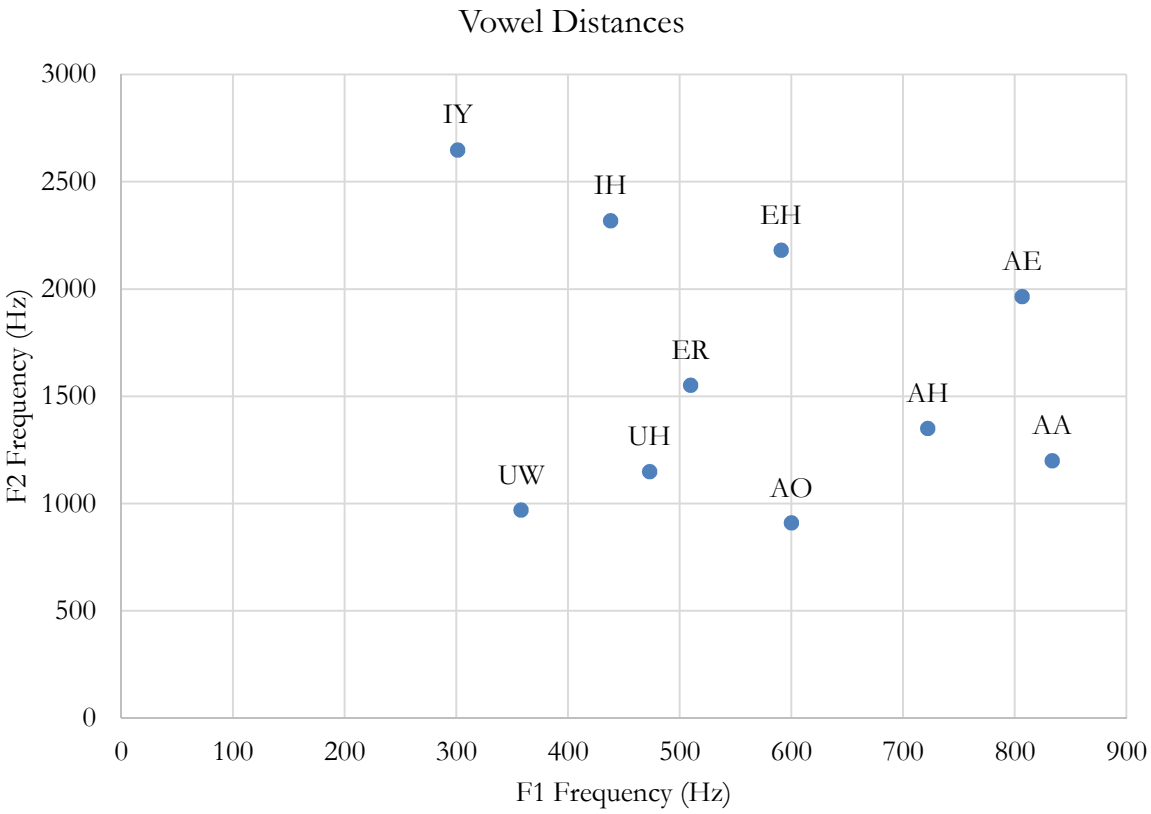


As can be clearly seen from the diagram, the formants for men, women, and children all follow the same general pitch contours, which suggests that vowels have similar formant frequency ratios. What we also notice is a trend whereby the formants for men are always lower than or equal to the formants for women, which are always lower than or equal to the formants for children, in that order. Furthermore, the distance between the third formants for the male, female, and child speakers has the highest differential, whereas the fundamental frequencies are almost all the same. This makes sense, because it is known that men generally have lower-pitched voices than women and children. It is interesting, however, to note that the fundamental frequency (which is sometimes only perceived and not actually uttered) is almost the same for all three groups.

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?

Formants f1 and f2 will be the most useful because they have the greatest variability with respect to each other in the frequency range below 3 KHz, thus forming distinct intervals which can be detected and distinguished. In fact, in practice the formants f1 and f2 are the standard minimum pieces of information required to distinguish vowels apart. The fundamental frequency is always the same and thus gives us no information as to whether we are looking at one vowel or another. Additionally, f3 gives us very little information because it does not vary much until we look at the vowel ER. Ergo f1 and f2 are the most informative features to distinguish 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.



The closest vowels to each other, according to this plot, are ER vs. UH, AH vs. AA, and UH vs. UW. Because of this closeness, we can posit that they are the most likely vowels to be confused with one another.

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.

   [**-12**, **-11**, -10, **-9**, **-8**, **-7**, -6, **-5**, -4, -3, **-2**, **-1**, **0**, **1**, **2**, **3**, **4**, 5, **6**, 7, **8**, 9, 10, **11**, 12]

K-value	Class for -5	Reasoning	Class for 3	Reasoning
1	Gamma	-5 is a member of Class Gamma	Gamma	3 is a member of Class Gamma
3	Gamma	The 3 nearest neighbors are -5, -7, and -2. All three belong to different classes so -5 has the smallest Euclidean distance from -5	Gamma	The 3 nearest neighbors are 2, 3, and 4. All three belong to different classes so 3 has the smallest Euclidean distance from 3
7	Beta	The 7 nearest neighbors are equally divided between Class Beta and Class Alpha, but the sum of the distances to the items in Class Beta is smaller than Class Alpha	Alpha	The 7 nearest neighbors involve 4 from Class Alpha, 1 from Class Gamma, and 2 from Class Beta, so by majority vote, we choose Alpha

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

   Class Alpha
   Mean: 0
   Variance: 2

   Class Beta
   Mean: -1
   Variance: 50.66

   Class Gamma
   Mean: -3.83
   Variance: 68.81

3. Now, classify -5 and 3 again according to the Gaussian distributions: classify a point in 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?

   I don't understand this question. Sorry.

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.

Speech recognizers can make a variety of errors. They can misinterpret a word as something else (an incorrect substitution), they can fail to recognize a word (an incorrect deletion), or they can add words which are not present in the speech signal at all (an incorrect insertion.) These three measures contribute to the Word Error Rate (WER), which is the general criteria to measure for the performance of a speech recognizer. The WER is defined as the number of errors divided by the number of spoken words. The WER can in fact be higher than 100% if a recognizer is so bad that it inserts incorrect words (which are by definition wrong because they don't even exist in the spoken input) in addition to misrecognizing spoken words. Then the number of errors could potentially exceed the number of spoken words, causing the word error rate to be over 100%.

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.

   I don't understand this question. Sorry.

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 $j1, j2$ of the template incurs a transition cost of $e(j1, j2)$. The accumulated cost up to the position (i, j) is denoted by $D(i, j)$. Write down the dynamic programming recursion formula for $D(i, j)$.

   I don't understand this question. Sorry.

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 $f1 \ldots fI$ itself, where *fi* is the number of the frame in the template that is matched with frame *i* of the utterance.

   I don't understand this question. Sorry.

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?

   Using only DTW, you can make a reasonable guess as to which input utterance corresponds to which template. The template with the shortest distance to the input utterance is the most likely match.

6. What is the time complexity of DTW?

   Assuming the comparison operation between two frames takes constant time, having to compare *n* frames for a template with *m* frames for the decoded signal (without limiting the window or using beam search) is quadratic time complexity, or $O(nm)$. Limiting the window will speed this time up considerably.

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

   I didn't have time to do this. Sorry.

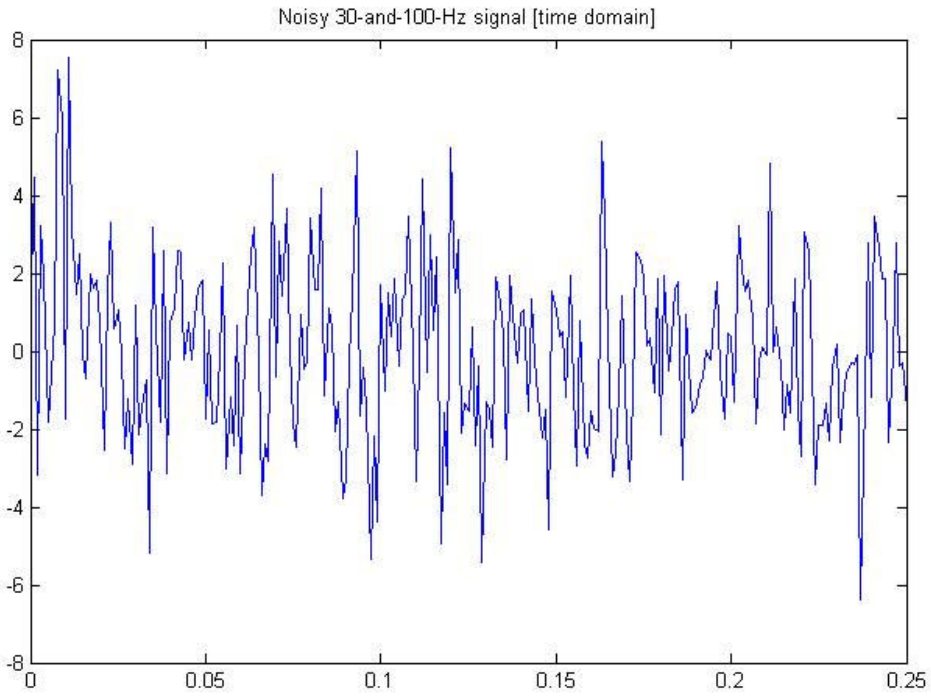
8. Describe how DTW can be extended to continuous speech recognition. (Hint: how can you recognize an utterance with multiple words, rather than a single word?)

   I didn't have time to do this. Sorry.

Problem 4: Digital Signal Processing with MATLAB

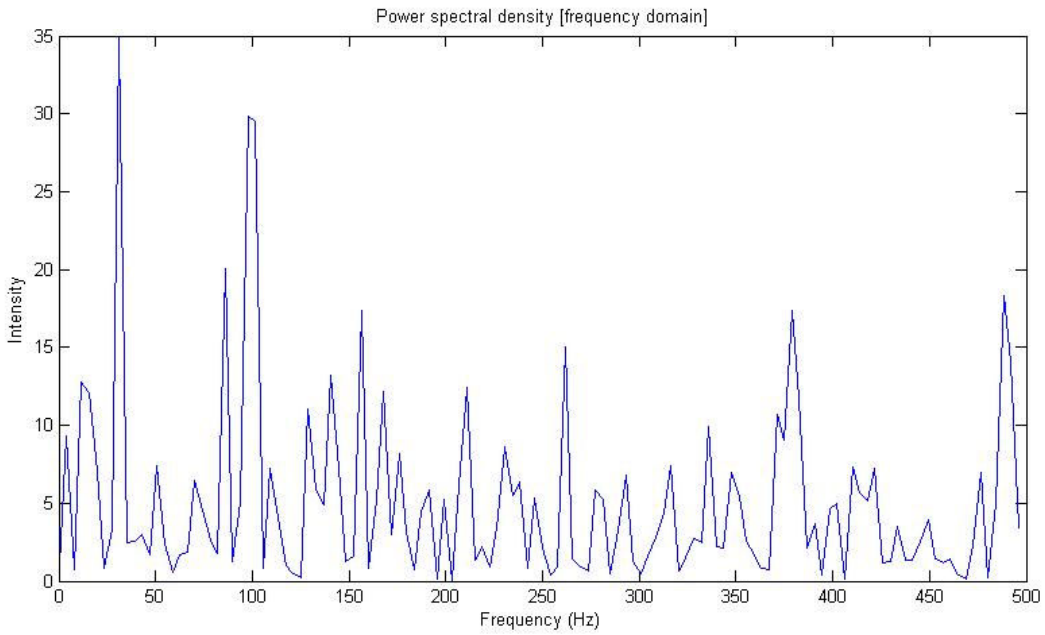
1. Include the plot of the noisy 30-and-100-Hz signal in your homework.

```
t = 0:.001:.25;
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]');
y = x + 2 * randn(size(t));
plot(t,y), title('Noisy 30-and-100-Hz signal [time domain]');
```



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

```
Y = fft(y,256);
PY = Y .* conj(Y) / 256;
f = 1000/256*(0:127);
plot(f,PY(1:128)), title('Power spectral density [frequency domain]');
xlabel('Frequency (Hz)');
ylabel('Intensity');
```



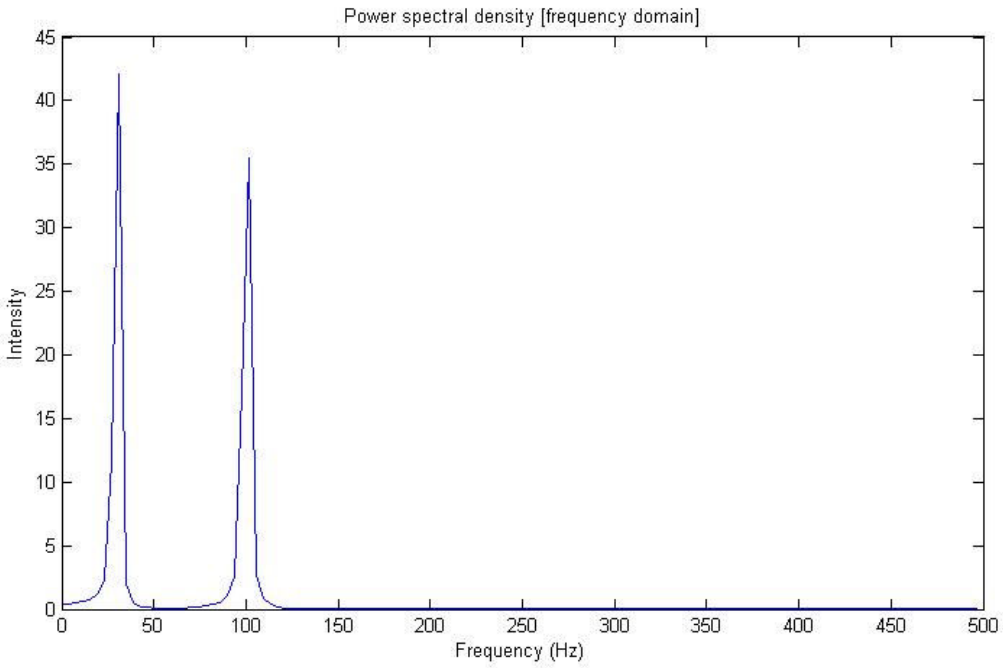
3. The spectrum, *Y*, is complex. That is, it has a real part and an imaginary part, or a magnitude and a phase. Which of these have a physical meaning, and how would you interpret them?

   The magnitude (real) part of the spectrum is what has physical meaning. It corresponds to the decibel value (the energy) of the signal at various frequencies, which is an indication of what "component" frequencies are present in the spectrum. This allows us to distinguish noise from information, provided the background noise is not too loud so as to completely mask the frequencies we are concerned with. In the frequency spectrum above, we see two definite spikes at 30 Hz and 100 Hz, which are the component sine waves we constructed in Part 1. The rest of the spikes, though loud, are the loudest component frequencies of the noise which we added to the signal. But it is clear that if we are interested in determining the presence of 30 Hz and 100 Hz waves in the signal, they are very obviously present when we examine the power spectrum.

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?

```
t = 0:.001:.25;
x30 = sin(2 * pi * 30 * t);
x100 = sin(2 * pi * 100 * t);
x = x30 + x100;
X = fft(x, 256);
PX = X .* conj(X) / 256;
f = 1000/256*(0:127);
plot(f,PX(1:128)), title('Power spectral density [frequency domain]')
xlabel('Frequency (Hz)')
ylabel('Intensity')
```

   The time domain noise affects the signal by causing a range of other frequencies to register intensities in the power spectrum. When you simply plot the spectral density of the pure 30Hz/100Hz sine wave, the plot appears clear of any frequency intensities with the exception of two (expected) spikes at 30 Hz and 100 Hz. An ideal signal devoid of any other noise breaks down beautifully into its component frequencies. The plot is shown on the next page.


Power spectral density [frequency domain]
Intensity
Frequency (Hz)
0
5
10
15
20
25
30
35
40
45
0
50
100
150
200
250
300
350
400
450
500