Instructions: This assignment is intended to familiarize you with looking at and thinking about speech signals. Parts of this assignment are expected to be quite challenging. Please ask for help if you feel stuck! Please submit answers to all underlined questions below. Please submit your solutions as a PDF file via the course Canvas site. The tex file used to create this PDF has also been provided so that you can enter your answers directly into this document.
Lateness policy: Late submissions submitted up to Monday 4/20/20 7:00 pm (the "late deadline") will receive a 20% reduction in credit. Submissions that are later than this will receive some non-zero credit, but we cannot guarantee how much. We cannot guarantee that we will carefully grade or give feedback on submissions after the "late deadline". In addition, you have four free "late days" that you can use throughout the term to extend homework (not project) deadlines without penalty. If you are using any of your "late days" for this assignment, mention that in the comment while submitting this assignment. You should state how many "late days" you are using and how many remaining "late days" you have. The number of late days used for any given homework must be an integer.
Collaboration: You are encouraged to discuss assignments and any aspect of the course material with others, but any material you submit (writeup, code, figures) should be produced on your own.

*** Before starting this homework you'll need to download and install Praat (a tool designed for detailed speech analysis). Once installed, you can listen to and view wav files using Praat. You can also look at spectrograms and play around with the various available tools/settings to estimate formant information, pitch information, and much more.

Once you've opened an audio file using Praat, you can click on 'View & Edit'. A window visualizing the waveform and spectrogram will appear. Within this new window, you can look under Spectrum > Spectrogram Settings to set the view range and the window length. To apply changes, click 'Apply' and 'Ok'. To return to the default settings click 'Standards'.

0. Please fill out this questionnaire. This is purely to help us calibrate assignment load and let us know what may need to be made clearer in class. Your responses will receive a small amount of credit, independent of the actual answers.

   (i) Did you collaborate on this assignment, and if so, with whom?
   Yes, my collaborator is Hanqi Zhang on our class.

   (ii) Approximately how many hours did this assignment take to complete?
   The assignment takes about 8-10 hours including reading related material.

   (iii) On a scale of 1 to 5 (where 1 = trivial and 5 = impossible), what was the difficulty of this assignment?
   4

   (iv) On a scale of 1 to 5 (where 1 = useless and 5 = essential), how useful has this assignment been to your understanding of the material?
   4

1. **Looking at Spectrograms using Praat**: The goal of this section is to solidify your understanding of certain concepts described in class, to get more comfortable with understanding speech signals visually, and to investigate properties of speech that may affect how we process it automatically.

   (i) Listen to and view the utterances of *beet, bet, bat, but, boot.*

   a. Investigate the spectrograms to measure formant frequencies.
   For this part, set view range to be from 0-8k Hz and window length to 0.005s.
   Estimate the formant frequencies for each vowel (measure at the center of the vowel).
   Feel free to look at Praat's automatic formant estimates, but take your own measurements visually from the spectrogram; automatic formant trackers can be quite unreliable!
   bat: F1: 750Hz F2: 2100Hz F3: 2800Hz
   boot: F1: 400Hz F2: 1400Hz F3: 3000Hz
   beet: F1:300Hz F22700Hz F33200Hz
   bet: F1: 500Hz F2: 2200Hz F3:2900Hz
   but: F1:750Hz F2:1500Hz F3:2900Hz

b. Measuring pitch.
Estimate the average pitch (fundamental frequency) of each utterance.
Again, feel free to look at Praat's automatic pitch estimates, but take your own measurements visually from the signal and/or spectrogram. There are multiple ways to estimate pitch, and you can use whichever approach you prefer.
bat: F0: 10/0.078 = 128Hz
boot: F0:10/0.071=141Hz
beet: F0:10/0.067 = 149Hz
bet: F0:10/0.068 = 147Hz
but: F0:10/0.081 = 123Hz

c. How did you estimate the pitch? What is one other approach you could have used?
I find the pitch by counting ten period of the vowel.And then use 10 divided by the time that 10 periods take. For example, in the bat case 10 period for a use 0.078 seconds and this give us the fundamental frequency of : 10/0.078 = 128Hz
One other approach I could have used is : searching the first peak of the spectral contour of vowels.

d. **EXTRA CREDIT:** Describe a **third** method you could have used to estimate pitch.
The third method is that we can use the differnce in Hz between Harmonics.

(ii) Understanding prosody: meaning beyond words.

a. Listen to the six recordings of the word "Amelia". The first three should sound different (in meaning) from the last three.
What is the main difference you notice in the perceived meanings of these two clusters of recordings?
The main difference is that the first three files, the name in an indicative way.While the second three files are in an interrogative way. Since these two cluster has different prosody, their accent are on the different letter.

b. Now investigate the corresponding spectrograms in Praat and try to relate your observations in the previous part with some of the speech features that you can view using Praat.
Which speech feature is the most relevant for explaining the difference between the two types of 'Amelia'? Describe a rule that relates this speech feature to the perceived meaning of the audio.
The most relevant feature is the pitch of the audio since you will have different pitch when saying the same word with different mood.
To related these two:
At the end of a sentence(or word), a falling trend of pitch is in declarative mood. A rising trend will likely to be a interrogative mood. And a rising and fall will be a incredulity mood.

(iii) Different pronunciations of the same word. For this part you will be phonetically transcribing words, using ARPAbet phonetic symbols. Please refer to this listing (https://en.wikipedia.org/wiki/ARPABET) of phonetic symbols in ARPAbet (we will use the two character version) and their corresponding International Phonetic Alphabet (IPA) symbols (for reference only – you will sometimes see IPA symbols used in our readings).

a. Listen to the two utterances "winter-1.wav" and "winter-2.wav" by the same speaker. You should notice a difference in the pronunciation of the word "winter". To help identify the difference, you can use Praat to listen to just the part of the signal where "winter" is uttered in each utterance.
Utterance transcripts:
winter-1: "spring and well i guess we're still in **winter** and uh"
winter-2: "we usually average oh anywhere from six to twelve inches during the **winter** and"
Write the phonetic transcription of "winter" for both utterances.
Winter-1:
WIHNTAXR
Winter-2:
WIHNNXAXR

b. Listen to the two utterances "and-1.wav" and "and-2.wav" by the same speaker. Again you should notice a difference in the pronunciation of the word "and" (marked in **bold**), and you can use Praat to listen to the portions corresponding to "and".
Utterance transcript:

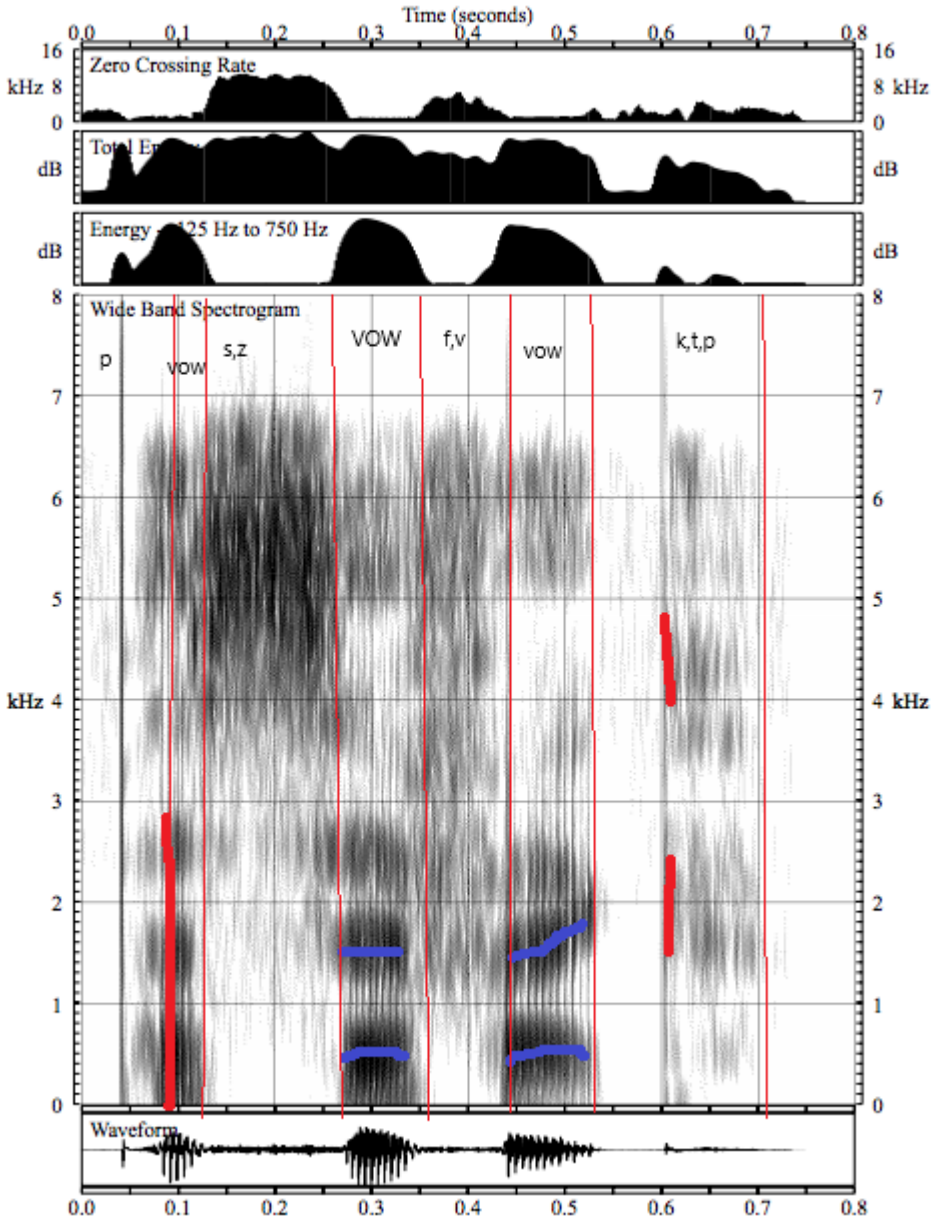
and-1: "summer **and** winter so"
and-2: "this was in the middle of the summer **and** we woke up one morning and it snowed for about fifteen minutes"
Write the phonetic transcription of "and" for both utterances.
Does the difference in pronunciation convey something about the meaning of the utterance? If so, what?
and-1:
AEND
and-2:
EN
The difference in pronunciation convey different meaning of "and". The first "and" is emphasizing on the word "winter". It has actual meaning of "parallel relationship" between winter and summer.
The second "and" has no actual meaning, it is just a sequence of move.
Therefore, the pronunciation will affect the meaning of the word or sentence.

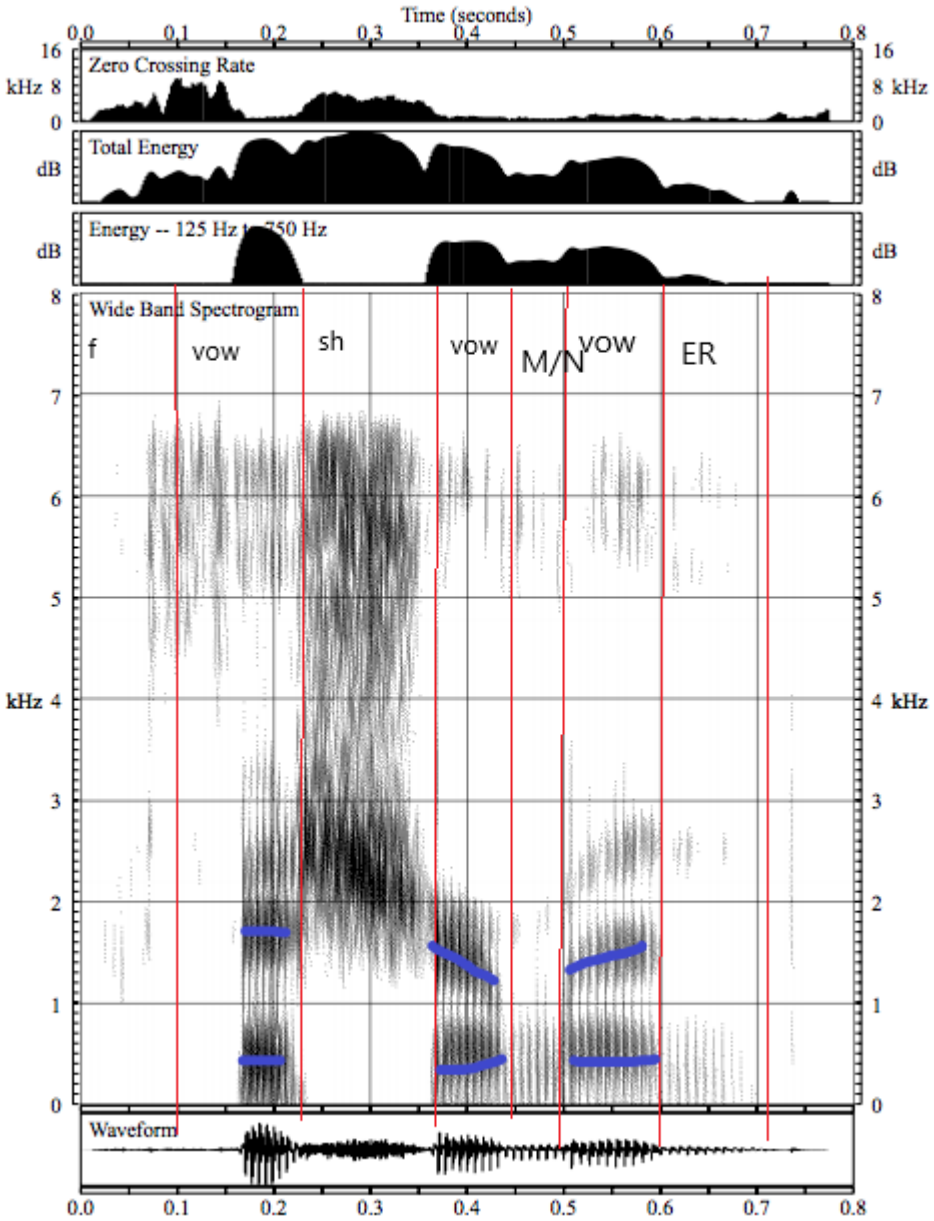
2. Work through slides 1-37 of **MIT_OCW_6.345_lec3-4.pdf** (from MIT's speech recognition course on OpenCourseWare), which is available on Canvas under Modules > Readings. Try to identify the spoken word that correpsonds to the spectrogram pictured on slide 24 and, optionally, also the one on slide 28 (Figures **??** & **??** below). This should be challenging, but give it a reasonable effort. The figures below contain several additional measurements (energy, zero crossing rate), but you should only need the spectrograms. A suggested workflow is as follows:

   (a) Mark the boundaries between phones with vertical lines. If you are unsure about some boundaries, mark all the ones you can.

   (b) Trace the first two or three formants wherever possible and label the stationary portions with their formant frequencies.

   (c) Below the spectrogram, mark each segment as corresponding to a vowel or consonant and with all phonetic labels that seem plausible for that segment. Once again please use ARPAbet symbols as listed here.


Time (seconds)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Zero Crossing Rate
16
kHz 8
0
dB
Energy
125 Hz to 750 Hz
Wide Band Spectrogram
p
vow
s,z
VOW
f,v
vow
k,t,p
8
7
6
5
kHz 4
3
2
1
0
Waveform



(d) Try to mentally "decode" a path through the resulting network of possible phonetic sequences. It may help to also think about the number of syllables (which should be the same as the number of vowels).
The syllables of two words.(also as an input to the python code)
P24: [stop,vowel,fricative,vowel,fricative,vowel,stop]
P24:[P-EY/AE/AX-S/Z-IH/IY/IX-F/V-IH/IY/IX-K/T/P]
P28: [fricative,vowel,fricative,vowel,nasal,vowel,nasal]
p28:[F-IH/AE/-SH/S-AXR/AX/-M/N-AE/IX-N/NG]

(e) Use the provided code in `hw1.py` (or write your own) to help automate the process of finding word hypotheses. The python output using the above input are:

```
Spectrogram p.24
['pacified'] ['P', 'AE', 'S', 'AX', 'F', 'AY', 'D']
['pussyfoot'] ['P', 'UH', 'S', 'IY', 'F', 'UH', 'T']
['pacific'] ['P', 'AX', 'S', 'IH', 'F', 'IX', 'K']
['disavowed'] ['D', 'IH', 'S', 'AX', 'V', 'AW', 'D']
['deficit'] ['D', 'EH', 'F', 'AX', 'S', 'AX', 'T']
['dervishich'] ['D', 'ER', 'V', 'IX', 'SH', 'IX', 'K']
['deciphered'] ['D', 'IX', 'S', 'AY', 'F', 'AXR', 'D']
Spectrogram p.28
['fisherman'] ['F', 'IH', 'SH', 'AXR', 'M', 'AE', 'N']
['fisherman'] ['F', 'IH', 'SH', 'AXR', 'M', 'AX', 'N']
['fishermen'] ['F', 'IH', 'SH', 'AXR', 'M', 'IX', 'N']
['fastening'] ['F', 'AE', 'S', 'AX', 'N', 'IX', 'NG']
['fashioning'] ['F', 'AE', 'SH', 'AX', 'N', 'IX', 'NG']
['seasoning'] ['S', 'IY', 'Z', 'AX', 'N', 'IX', 'NG']
['sassaman'] ['S', 'AE', 'S', 'AX', 'M', 'AX', 'N']
['softening'] ['S', 'AO', 'F', 'AX', 'N', 'IX', 'NG']
['sossamon'] ['S', 'OW', 'S', 'AA', 'M', 'AO', 'N']
['siphoning'] ['S', 'AY', 'F', 'AX', 'N', 'IX', 'NG']
[Finished in 14.4s]
```

What are the predicted phone sequence labelings and word hypotheses you come up with?

To summarize, I guess the word in the slides 24 to be pacific.

And the word on slides 28 to be fisherman.