EE599 Speech Processing Spring 2019
Project 1 Specification
Due date February 14, 2019

Any common programming language may be used for this project, but Matlab is recommended. Turn in all code and specified plots for the problems given. Label all axes appropriately throughout, and title all figures. Answer the analysis questions on a separate page.

This project will focus on basic signal analysis of a simple speech signal, the word "sun" spoken by an adult male speaker. The data file, 'sun.wav', is 12-bit quantized with 10kHz sampling rate.

1. **Preprocessing**: Zero-mean the signal, and choose an appropriate frame size (a power of two) with a 50% overlap. Make an initial plot of the signal that has the time series signal at the top and a spectrogram at the bottom. Customize the spectrogram parameters to get a good visual balance between time and frequency resolution. Explain your reasoning for choice of frame size and spectrogram parameters.
2. **Energy** and **Zero-crossing rate**: Plot the signal and its energy and zero-crossing rate vs. time on a single plot (this can be either a single axis or a 3x1 plot as you prefer). Define energy as the mean squared signal amplitude over each frame. Zero-crossing rate is a measure correlated with high-frequency content, often useful for simple signal processing tasks, defined as:

$$zc = \frac{1}{2N}\sum_{n=1}^{N-1}\left|\text{sgn}(s(n)) - \text{sgn}(s(n-1))\right|, \text{ where } \text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

   What does the energy and zero-crossing rate information tell you about these phonemes? Specifically, how do you think they relate to phoneme category and voiced/unvoiced status?
3. **Phoneme selection:** Select three frames to be representatives of each of the three phonemes in the word, and plot them as a 3x1 figure against a time axis.
4. **Log magnitude FFT:** Apply a Hamming window to each selected frame, and compute the log magnitude (dB units), using whatever NFFT you think appropriate. Plot these (3x1 figure).
5. **Cepstrum:** Compute the real cepstrum of each hamming-windowed frame (computing directly, not using the 'rceps' or other built-in command). Plot these (3x1 figure). For the voiced phonemes, can you see a clear peak representing the pitch? If so, what is the corresponding f_0?
6. **Liftered spectrum:** Select a single appropriate cepstrum cutoff to separate vocal tract and excitation information. Lifter the cepstra, and then take the inverse FFT and plot the log magnitude spectral envelope corresponding to the liftered cepstrum of each selected frame (3x1 figure). Explain your choice of the threshold for liftering.

7. **LPC spectra:** Compute the LPC of the hamming-windowed frames, for p=4, 14, and 40. Plot the log magnitude spectral envelope corresponding to the equivalent all-pole spectral model (full page 3x3 figure). Discuss which order you think gives the most informative representation of the vocal tract spectrum.

8. **LPC Residual:** Use the linear prediction error filter with p=14 to calculate the LPC residual, for the entire waveform. Make sure to calculate the residual from the original (not hamming-windowed) signal, and also be careful to make sure that the residual is contiguous at frame boundaries without any edge effects (the easiest way to do this is to append some points from the previous frame before running the filter to get an accurate residual for the entire frame). Make two plots, one of the LPC residual of the entire signal, and one of the residual for each selected frame.

9. **Pitch tracking:** Use Praat to calculate the fundamental frequency contour of the signal, and estimate the formant tracks. Plot the spectrogram with the pitch and formant contours.

10. **Formant analysis:**

    Review the various spectra from the DFT, liftered spectrum, and all three orders of LPC analysis. Make a table for each phoneme, estimating the formants F1, F2, and F3 for each of the five spectral plots without referring to the other plots or using knowledge of what you think the values should be. Add an additional column for the Praat formant estimates, taken from the time location of your selected frames. For the vowel phoneme, add one additional column with the formant values obtained from a chart (cite your specific source for reference). Compare the different estimates, and identify which one you think is the most accurate. Do you think your estimates are accurate? Overall, which of the analysis methods were most useful to you in clearly identifying formants? Specifically discuss the effect of varying the order p of the LPC analysis.