E9 261 – Speech Information Processing

Homework # 3
Due date: May 2, 2021

Please upload (in the course webpage) your recordings and codes as a zipped folder FirstName_LastName_HW3.zip (or in multiple zip files with filenames have part1 part2 etc. each not exceeding 100Mb). In the zipped folder the program names should be self explanatory. Filename of each program should contain the question number it is associated to.

1. Record your own voice using Praat for at least 3 seconds. Use Praat to obtain and plot the followings (one below another in a time synchronous manner).

   (a) Speech waveform

   (b) Speech/silence decisions

   (c) Voiced/Unvoiced decisions

   (d) Pitch frequency over all voiced regions

   (e) Time-varying intensity contour

   (f) Time-varying formant (first three) frequencies over all voiced regions

   Choose a short-time window type and duration of your choice. Please do not take a screen shot of Praat windows and put them in your report. Obtain the values and plot them in Matlab, for example. Give proper axis labels with legible values on the axes.

2. Emphasis of the high frequency region of the spectrum is often accomplished using a first difference. In this problem, we examine the effect of such operations on the short-time Fourier transform.

   (a) Let $y[n] = x[n] - x[n-1]$. Show that

   $$Y_n(e^{\mathrm{j}\omega}) = X_n(e^{\mathrm{j}\omega}) - e^{-\mathrm{j}\omega} X_{n-1}(e^{\mathrm{j}\omega})$$

   (b) Under what conditions can we make the approximation

   $$Y_n(e^{\mathrm{j}\omega}) \approx (1 - e^{-\mathrm{j}\omega}) X_n(e^{\mathrm{j}\omega})$$

   In general, $x[n]$ may be linearly filtered as in

   $$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k]$$

(c) Show that $Y_n(e^{j\omega})$ is related to $X_n(e^{j\omega})$ by an expression of the form

$$Y_n(e^{j\omega}) = X_n(e^{j\omega}) * h_\omega[n]$$

Find $h_\omega[n]$ in terms of $h[n]$.

(d) Is it reasonable to expect that

$$Y_n(e^{j\omega}) = H(e^{j\omega})X_n(e^{j\omega})$$

3. The short-term energy is defined as

$$E_n = \sum_{m=-N}^{N} h[m]x^2[n-m]$$

Suppose we wish to compute E_n at each sample of the input

(a) Let $h[m]$ be

$$h[m] = a^{|m|}, \quad |m| \leq N$$
$$= 0 \qquad otherwise$$

Find a recurrence relation (i.e., a difference equation) for E_n.

(b) What is the savings in number of multiplications obtained by using the recurrence relation rather than directly computing E_n?

(c) Draw a digital network diagram of the recurrence formula for E_n. (As defined, $h[m]$ is noncausal. Therefore an appropriate delay must be inserted.)

4. The short-time average zero-crossing rate is defined as

$$Z_n = \frac{1}{2N} \sum_{m=n-N+1}^{n} |sgn(x[m]) - sgn(x[m-1])|$$

Show that Z_n can be expressed as

$$Z_n = Z_{n-1} + \frac{1}{2N}\{|sgn(x[n]) - sgn(x[n-1])| - |sgn(x[n-N]) - sgn(x[n-N-1])|\}$$

5. The short-time autocorrelation function is defined as

$$R_n[k] = \sum_{m=-\infty}^{\infty} x[m]w[n-m]x[m+k]w[n-k-m]$$

(a) Show that

$$R_n[k] = R_n[-k]$$

i.e., show that $R_n[k]$ is an even function of k.

(b) Show that $R_n[k]$ can be expressed as

$$R_n[k] = \sum_{m=-\infty}^{\infty} x[m]x[m-k]h_k[n-m]$$

where

$$h_k[n] = w[n]w[n+k]$$

(c) Suppose that

$$w[n] = a^n \quad n \geq 0$$
$$= 0 \quad n < 0$$

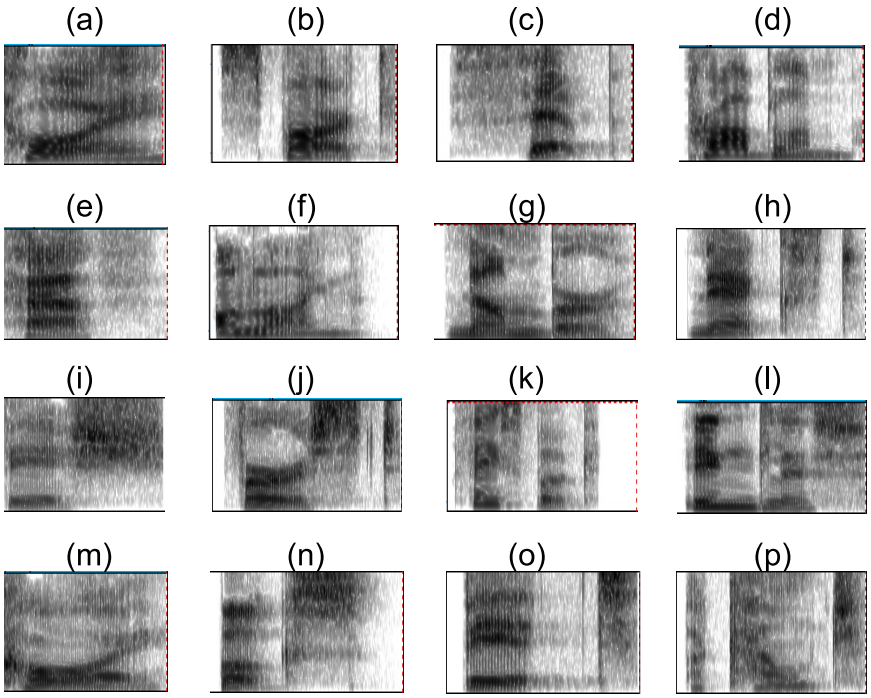
FInd the impulse response, $h_k[n]$, for computing the k^{th} lag.

(d) Find the Z-transform of $h_k[n]$ in (c) and from it obtain a recursive implementation for $R_n[k]$. Draw a digital network implementatyion for computing $R_n[k]$ as a function of n for the window of (c).

(e) Repeat parts (c) and (d) for

$$w[n] = na^n \quad n \geq 0$$
$$= 0 \quad n < 0$$

6. Consider the following 16 spectrograms (over 0-5kHz) for the following words in random order: coy, english, facebook, problem, spark, toy, sip, account, first, online, bit, books, fish, next, number, path. Match each word with its spectrogram.



7. Collect the recordings from all your classmates in SIP course related to the last question on LTAS in HW2. We will experiment with human ability to count number of speakers from a recording with multiple speakers speaking. For this identify four listeners (excluding yourself; listeners should not have any hearing impairment).

Consider any N speakers (among all classmates) and take one random sentence spoken by each speaker. Add these N recordings (you should truncate all N recording to their minimum length) denoted by $x_N[n]$. Present $x_N[n]$ for listening to every listener (can be

done through google form) and ask how many speakers he/she perceives to be present in $x_N[n]$. Repeat this for five times by choosing different sets of N speakers and report how many times (among five) each listener correctly predicts the number of speakers in $x_N[n]$ they heard. Repeat this experiment for by varying N from 2 to 8 with a step of 1. Present the accuracy in a table with one column for each listener and each row for every value of N.

8. Consider $\hat{x}[n] \leftrightarrow \hat{X}(z)$ is the cepstrum of $x[n] \leftrightarrow X(z)$. Show that $\frac{\mathrm{d}}{\mathrm{d}z}X(z) = X(z) \times \frac{\mathrm{d}}{\mathrm{d}z}\hat{X}(z)$. Use this to formulate a recursive way of computing $\hat{x}[n]$ from $x[n]$ (and vice-versa) when $x[n]$ is a causal and minimum phase sequence. Note that such a recursive computation of cepstrum is free of any aliasing or phase unwrapping error.

9. Implement cepstral based pitch estimation and compare with pitch from praat by computing the accuracy of the voiced/unvoiced classification and the mean absolute error in estimated pitch in the original voiced segments. Use at least 20 sentences' recordings from the last question on LTAS in HW2.

10. Record three different sentences in your voice each approximately 3 seconds long and save them as a 8kHz .wav files and do the following LPC analysis/synthesis for each of the files.

    (a) Consider 20msec non-overlapping segment and compute linear prediction (LP) coefficients and all-pole filter gain using Matlab command `lpc`. Compute prediction error signal using LP analysis equation (FIR filtering). Make sure you carry over the filter state from one segment to the next segment to avoid any discontinuity. Listen to the 3 seconds long prediction error signal. Do you recognize the sentence? Justify your observations

    (b) Implement the LP synthesis equation (IIR filtering) with the prediction error signal and the LP coefficients in all segments. Listen to the synthesized signal. Is there any difference between the synthesized signal and the original signal?

    (c) Implement the LP synthesis equation (IIR filtering) with white Gaussian noise in each segment in place of the prediction error signal (but with the same energy as that of the prediction error signal). Listen to the synthesized signal. Do you recognize the sentence?

11. Record five sustained vowels /a/ (had), /i/ (hid), /u/ (hood), /o/ (hod) and /aa/ (hard) in your own voice each for at least one second and save a 8kHz .wav file. Consider five randomly chosen 40ms segments in each of these sustained vowels and do the followings:

    (a) For each segment, compute the spectral envelope using cepstrum as discussed in the class. Use a lifter as $w[n] = 1, \quad 0 \leq n < L$ and $0, \quad n \geq L$. Choose L as half of the the pitch period. Plot the signal spectrum and plot the cepstrum based spectral envelope on top of that (using FFT order 1024).

    (b) Compute a spectral envelope using a p-order AR model. Vary p such that that the Itakura-Saito (IS) distance between the AR model spectrum and the cepstrum based spectral envelope (from previous question) is minimum. Provide a plot of IS distance vs p.

    (c) Does the best choice of p vary among different 40ms segments of a vowel? How about across different vowels?