EE 519 Digital Speech Processing for Multimedia: Coding, Recognition and Synthesis
Fall 2002 University of Southern California

HOMEWORK # 3

Assigned on September 27, 2002 Friday
Due on October 4, 2002 Friday, 5pm

- Class web page: http://learn.usc.edu/
  Please check the web page regularly for the latest course news and notes.
- Speech Labs (SAIL) web page: http://sail.usc.edu

1. Suppose that the window sequence, \( w(n) \) used in short-time Fourier analysis is causal and has a z-transform of the form

\[
W(z) = \frac{\sum_{r=0}^{N_x} b_r z^{-r}}{1 - \sum_{k=1}^{N_k} a_k z^{-k}}
\]  

(1)

a) What properties should \( W(z) \) (or equivalently \( W(e^{j\omega}) \)) have in order that it is suitable for this application?

b) Obtain a recurrence formula for \( X_n(e^{j\omega}) \) in terms of the signal \( x(n) \) and previous values of \( X_n(e^{j\omega}) \).

\[
X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m}
\]  

(2)

c) Consider the case

\[
W(z) = \frac{1}{1 - az^{-1}}
\]  

(3)

How should \( a \) be chosen to obtain a frequency resolution of approximately 100Hz at a sampling rate of 10kHz?

d) The value of \( a \) required in (c) suggests that problems may arise in implementing very narrow-band time-dependent Fourier analysis recursively. Briefly discuss the nature of these problems.

2. Prove that

\[
\sum_{k=0}^{N-1} e^{j2\pi kn/N} = N \sum_{r=-\infty}^{\infty} \delta(n - rN) = \begin{cases} N, & n = rN, r = 0, \pm 1, \ldots \\ 0, & \text{otherwise} \end{cases}
\]  

(4)

3. In implementing time-dependent Fourier representations, we employ sampling in both the time and frequency dimensions. In this problem we investigate the effects of both types of sampling. Consider a sequence \( x(n) \) with conventional Fourier transform

\[
X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m}
\]  

(5)
a) If the periodic function \( X(e^{j\omega}) \) is sampled at frequencies \( w_k = 2\pi k/N, k = 0, 1, 2, \ldots, N - 1 \), we obtain
\[
\hat{X}(k) = \sum_{m=-\infty}^{\infty} x(m)e^{-j(2\pi km/N)}
\]
(6)

These samples can be thought as the discrete Fourier transform of the sequence \( \hat{x}(n) \) given by
\[
\hat{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}(k)e^{j(2\pi kn/N)}
\]
(7)

Show that
\[
\hat{x}(n) = \sum_{r=-\infty}^{\infty} x(n + rN)
\]
(8)

b) What are the conditions on \( x(n) \) so that no aliasing distortion occurs in the time domain when \( X(e^{j\omega}) \) is sampled?

c) Now consider "sampling" the sequence \( x(n) \), i.e., let us form the new sequence
\[
y(n) = x(nM)
\]
(9)

consisting of every \( M^{th} \) sample of \( x(n) \). Show that the Fourier transform of \( y(n) \) is
\[
Y(e^{j\omega}) = \frac{1}{M} \sum_{k=0}^{M-1} X(e^{j(\omega-2\pi k/M)})
\]
(10)

In providing this result you may wish to begin by considering the sequence
\[
v(n) = x(n)p(n)
\]
(11)

where
\[
p(n) = \sum_{r=-\infty}^{\infty} \delta(n + rM)
\]
(12)

then note that \( y(n) = v(nM) = x(nM) \).

4. The purpose of this problem is to show that, under certain conditions, only one frequency sample of the STFT magnitude is needed for unique representation of the sequence \( x(n) \) for a time decimation factor of unity. The STFT is defined by
\[
X(n, w) = \sum_{m=-\infty}^{\infty} x(m)w(n - m)e^{-jwm}
\]
(13)

where \( w(n) \) is the analysis window of length \( N \). Suppose that the following conditions on \( x(n) \) and \( w(n) \) hold,

A: \( x(n) \) is a non-negative (i.e. \( x(n) \geq 0 \)), real and right-sided sequence whose first non-zero value falls at \( n = n_0 \).

B: The analysis window \( w(n) \) is \( N \) point along, real and positive (i.e., \( w(n) > 0 \)) over the interval \( 0 \leq n \geq N - 1 \).

You are now asked to show that the sequence \( x(n) \) is specified by one appropriately chosen frequency sample of \( X(n, w) \).

a) Consider the smallest value of \( n \), namely \( n_0 \), such that \( x(n) \) is non-zero. Show that \( x(n_0) \) can be determined by the expression
\[
x(n_0) = \frac{X(n_0), 0}{w(0)}
\]
(14)
b) Suppose now that \(x(n)\) is known up to time \(n - 1\) and we wish to compute the sample \(x(n)\) from the previous values and \(X(n, w)\). Show that,

\[
X(n, w) = Y(n, w) + x(n)w(0)e^{-j\omega n}
\]

where

\[
Y(n, w) = \sum_{m=n-N+1}^{n-1} x(m)w(n - m)e^{-j\omega m}
\]

\(Y(n, w)\) is known, since it is a function of samples prior to time \(n\).

c) Argue that \(x(n)\) can be recovered recursively for \(n > n_0\) from only the DC values of the STFT and previous samples of \(x(n)\) and \(w(n)\). Show the recursive relation.

5. MATLAB PROBLEM

In this problem you use the speech waveform \(speech2_10k\) in the workspace \(ex7M1.mat\). This problem helps you to develop an understanding of the limitations of the STFT in achieving good time-frequency resolution. The speech was sampled at 10000 samples/s. 

a) Plot the speech signal \(speech2_10k\). Identify the unvoiced/voiced regions. Plot triangular windows of durations 30ms and 5ms, respectively, created using MATLAB function \(triang.m\).

b) Plot the first eight STFT log-magnitudes of \(speech2_10k\) by sliding the 30-ms triangular window in 15-ms intervals. Use a 1024-point FFT and display only the first 512 points of the STFT log-magnitude. Use the command \(subplot(221)\) so you can display two sets of four functions. Also, make a matrix of your eight STFT log-magnitudes (512 points each), and then use the \(mesh.m\) MATLAB function to plot the 2-D time-frequency function.

(c) Repeat part (b) with the 5-ms triangular analysis window.

d) Repeat parts (b) and (c), but display spectograms using the MATLAB function \(spectogram.m\) or \(spectrogram_ex7p21.m\), rather than a mesh plot.

e) Comment on the time-frequency resolution tradeoffs in using the long and short triangular windows.

6. OPTIONAL BONUS CREDIT PROBLEM

You will write a program to automatically segment a recorded sentence into three different components: silence (non-speech), voiced speech and unvoiced speech.

a) Describe (in words) your strategy in writing and improving your program. A successful labeling program should utilize at least a short-term energy and a short-term zero crossing measure. However, as usual, you may add whatever you like to further improve the performance of your program. Make sure to turn in all code you use as the appendix of this homework.

b) Show a plot that, you feel, best depicts the labeling of the test sentence.

c) Have matlab create a table of the starting location of each labelled segment. For instance, your output should look something like the following:

| Silence: Sample 1 - Sample 200 |
| Voiced: Sample 200 - Sample 500 |
| Unvoiced: Sample 500 - Sample 600 |

and so on...

d) Comment on the accuracy of your algorithm. Make sure to run your code on the sentences provided for Hw 2 to see how generally is can be applied. Show the labeling for one other sentence that you have recorded. In what cases does your algorithm tend to have problems?

Good luck...