EE-414 Speech Processing Lab

Lab-8 - 180020002

Aim:

- To understand motivation behind Cepstral Analysis of speech
- To understand basic Cepstral Analysis approach
- To perform vocal tract and source information separation by Cepstral Analysis
- To understand liftering concept in cepstral Analysis
- To develop a pitch determination method by Cepstral analysis.
- To develop a formant information determination method by Cepstral analysis.

Theory:

Speech is composed of excitation source and vocal tract system components. In order to analyze and model the excitation and system components of the speech independently and also use that in various speech processing applications, these two components have to be separated from the speech. The objective of *cepstral analysis* is to separate the speech into its source and system components without any previous knowledge about source and system.

According to the source filter theory of speech production, voiced sounds are produced by exciting the time varying system characteristics with periodic impulse sequence and unvoiced sounds are produced by exciting the time varying system with a random noise sequence. The resulting speech can be considered as the convolution of respective excitation sequence and vocal tract filter characteristics. If e(n) is the excitation sequence and h(n) is the vocal tract filter sequence, then the speech sequence s(n) can be expressed as follows:

$s(n) = e(n) * h(n)$

This can be represented in frequency domain as,

$S(\omega) = E(\omega). H(\omega)$

The speech sequence has to be deconvolved into the excitation and vocal tract components in the time domain. For this, multiplication of the two components in the frequency domain has to be converted to a linear combination of the two components. For this purpose cepstral analysis is used for transforming the multiplied source and system components in the frequency domain to linear combination of the two components in the cepstral domain.

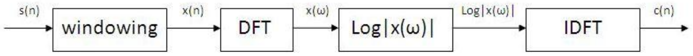
Basic principles of Cepstral Analysis

To linearly combine the E(ω) and H(ω) in the frequency domain, logarithmic representation is used. So the logarithmic representation of equation will be,

$$\log|S(\omega)| = \log|E(\omega)| + \log|H(\omega)|$$

The log operation transforms the magnitude speech spectrum where the excitation component and vocal tract component are multiplied, to a linear combination (summation) of these components. The separation can be done by taking the inverse discrete fourier transform (IDFT) of the linearly combined log spectra of excitation and vocal tract system components. IDFT of linear spectra transforms back to the time domain but the IDFT of log spectra transforms to *quefrency* domain or the cepstral domain which is similar to time domain. In the quefrency domain the vocal tract components are represented by the slowly varying components concentrated near the lower quefrency region and excitation components are represented by the fast varying components at the higher quefrency region.

$$c(n) = \text{ IDFT}(\log|S(\omega)|) = \text{IDFT}(\log|E(\omega)| + \log|H(\omega)|)$$



Methods have to be devised to extract to these vocal tract and excitation characteristics independently. For this purpose a ***liftering*** operation is performed in the quefrency domain.

Liftering

Liftering operation is similar to filtering operation in the frequency domain where a desired quefrency region for analysis is selected by multiplying the whole cepstrum by a rectangular window at the desired position. There are two types of liftering performed, low-time liftering and high-time liftering. Low-time liftering operation is performed to extract the vocal tract characteristics in the quefrency domain and high-time liftering is performed to get the excitation characteristics of the analysis speech frame.

Low-time liftering for Formant estimation

Low-time liftering is used for estimating slow varying vocal tract characteristics from the computed cepstrum of the given speech sequence. The low-time liftering window used for extracting vocal tract characteristics can be represented as follows,

$$w_e[n] = \begin{cases} 1\ ,0 \leq n \leq L_c \\ 0, L_c \leq n \leq \frac{N}{2} \end{cases}$$

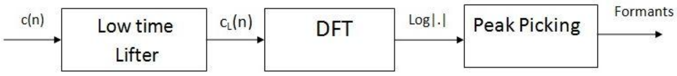
where Lc is the cut off length of the liftering window and N/2 is half the total length of the cepstrum. The vocal tract characteristics can be obtained by multiplying the cepstrum c(n) with the low-time liftering window

$$c_e(n) = w_e[n].\, c([n])$$

Applying DFT on the low-time liftered sequence takes to its log magnitude spectrum which is the vocal tract spectrum of the given short term speech

$$\mathrm{Log}[|H(w)|] = \mathrm{DFT}[c_e(n)]$$

The important vocal tract parameters like formant location and bandwidth can be computed from the vocal-tract spectrum. The formant locations can be estimated by picking the peaks from the smooth vocal tract spectrum.



High-time liftering for pitch estimation

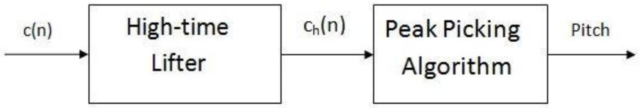
As the cepstrum computed from the analysis speech sequence is symmetric, half the length of the cepstrum is considered for the liftering. The excitation characteristic are obtained through a high time liftering operation using the following window,

$$w_h[n] = \begin{cases} 1\,, L_c \le n < \frac{N}{2} \\ 0, else \end{cases}$$

where ***Lc*** is the cut off length of the liftering window and ***N/2*** is the half the total length of the cepstrum. The excitation characteristics are obtained by multiplying high time liftering window with the cepstrum,

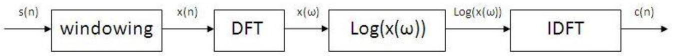
$$c_h(n) = w_h[n] * c([n])$$

Pitch can be estimated as the instant corresponds to the highest peak in the high-time liftered cepstrum. In the Figure 8, pitch period is the time instant corresponding to the largest peak in the high-time liftered cepstrum. The reciprocal of the pitch interval multiplied by the sampling frequency gives the pitch frequency of the analysis speech frame.



Complex Cepstrum

For the reconstruction of the sequence from the cepstrum, *complex cepstrum* is used. Instead of taking inverse fourier transform of the log magnitude spectrum for the real cepstrum, *the inverse fourier transform of the logarithm of complex spectrum* is used for computing complex cepstrum. As the logarithm of all the spectral values are used, the phase is preserved in the complex cepstral sequence which can be used for reconstructing back the sequence. The methods for computing pitch and formant parameters from the complex cepstrum remain same as that of the real cepstrum as these parameters are obtained from the magnitude of the complex cepstral coefficients.



The mathematical relation for computing complex cepstrum are as follow:

$$C_c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log(|s(\omega)|)e^{\int \omega n d\omega}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\log(|s(\omega)|e^{\int \angle s(\omega)})e^{\int \omega n d\omega}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\log(|s(\omega)|)e^{\int \omega n d\omega} + j\frac{1}{2\pi}\int_{-\pi}^{\pi}\angle s(\omega)e^{\int \omega n d\omega}$$

$$C_c(n) = C_r(n) + jC_i(n)$$

where $C_r(n)$ is the real cepstrum

$C_i(n)$ is the imaginary part of the complex cepstrum

Procedure:

Record (16kHz, 16bit) the word “speech signal”; truncate long silence regions.

A. Fundamentals of cepstral analysis of speech:

a. Select a frame (20 ms long) at the centre of a voiced segment. Plot the time waveform, the log-magnitude spectrum, and the cepstrum..

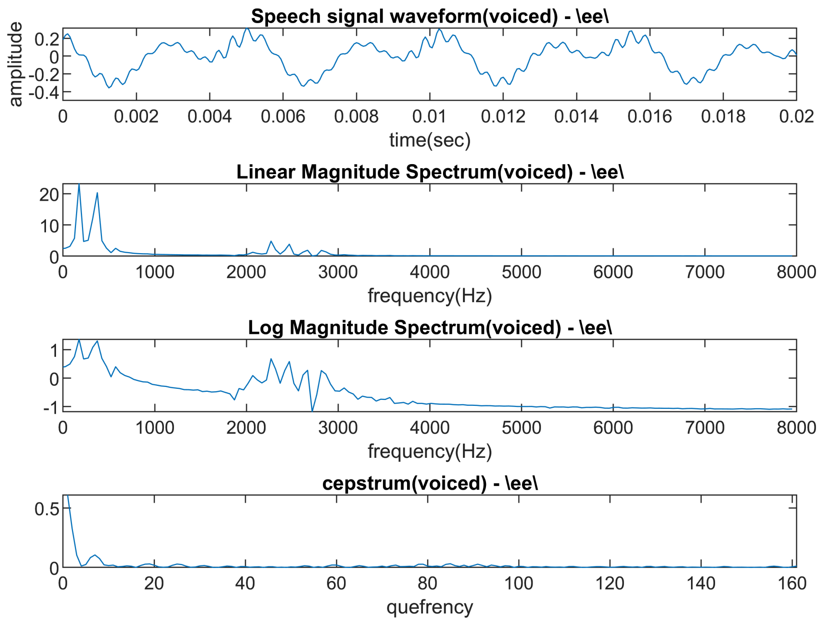
b. Repeat the above for an unvoiced segment.

c. Write the procedure to determine whether the segment is voiced/unvoiced by inspecting the cepstrum. Apply this procedure to the two segments (in a and b).

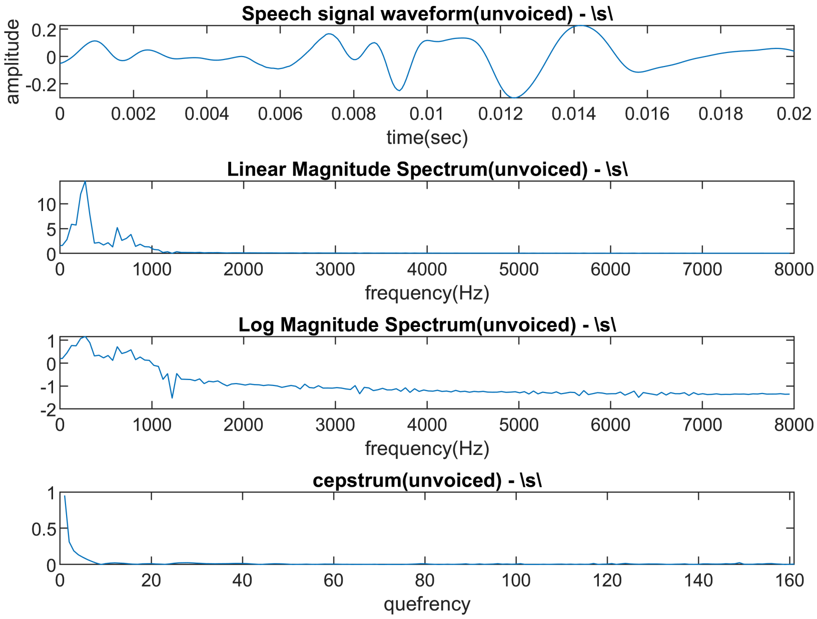
Code:

```
clc;clear all;close all;
warning('off');
%%Load the .wav file with 'speech signal'
[y,fs]=audioread('l7_speech_signal.wav');

%normalising the signal amplitudes to be in -1 to 1
y=y./(1.01*abs(max(y)));

%% Selecting 20ms frame at the centre

%/ee/- Voiced sound
y_v = y(ceil(0.226*fs) : floor(0.246*fs));
win= dsp.Window('Hamming'); % Applying hamming window
y_v = win(y_v);

%/s/ - Unvoiced sound
y_uv = y(ceil(0.056*fs) : floor(0.076*fs));
win = dsp.Window('Hamming'); % Applying hamming window
y_uv = win(y_uv);

%plotting time domain waveform of \ee\
t_v = 0 : 1 / fs : (length(y_v) - 1) / fs;
figure;
subplot(411);
plot(t_v, y_v);
xlabel('time(sec)');
```

```
ylabel('amplitude');
title('Speech signal waveform(voiced) - \ee\');

%plotting Linear Magnitude Spectrum of the speech signal
Y_v = fftshift(fft(y_v));
F_v = -fs/2 : fs/length(Y_v) : fs/2 - fs/length(Y_v);

subplot(412);
plot(F_v, abs(Y_v));
xlim([0, fs/2]);
xlabel('frequency(Hz)');
title('Linear Magnitude Spectrum(voiced) - \ee\');

%plotting Log Magnitude Spectrum of the speech signal
subplot(413);
plot(F_v, log10(abs(Y_v)));
xlim([0, fs/2]);
xlabel('frequency(Hz)');
title('Log Magnitude Spectrum(voiced) - \ee\');

%plotting IDFT of the speech signal
c_v = ifft(log10(abs(Y_v))); % ifft
% PLOT
subplot(414);
plot(abs(c_v));
xlim([0,ceil(length(c_v)/2)]);
xlabel('quefrency');
title('cepstrum(voiced) - \ee\');
```



```
%plotting time domain waveform of \p\
t_uv = 0 : 1 / fs : (length(y_uv) - 1) / fs;
figure;
subplot(411);
plot(t_uv, y_uv);
xlabel('time(sec)');
ylabel('amplitude');
title('Speech signal waveform(unvoiced) - \s\');

%plotting Linear Magnitude Spectrum of the speech signal
Y_uv = fftshift(fft(y_uv));
F_uv = -fs/2 : fs/length(Y_uv) : fs/2 - fs/length(Y_uv);

subplot(412);
plot(F_uv, abs(Y_uv));
xlim([0, fs/2]);
xlabel('frequency(Hz)');
title('Linear Magnitude Spectrum(unvoiced) - \s\');

%plotting Log Magnitude Spectrum of the speech signal
subplot(413);
plot(F_uv, log10(abs(Y_uv)));
xlim([0, fs/2]);
xlabel('frequency(Hz)');
title('Log Magnitude Spectrum(unvoiced) - \s\');

%plotting IDFT of the speech signal
c_uv = ifft(log10(abs(Y_uv))); % ifft
% PLOT
subplot(414);
plot(abs(c_uv));
xlim([0,ceil(length(c_uv)/2)]);
xlabel('quefrency');
title('cepstrum(unvoiced) - \s\');
```



Observations/procedure:

The variations in the lower quefrency region (near 0 axis) is due to vocal tract characteristics and the fast varying nature of the cepstrum towards the upper quefrency region represents the excitation characteristics of the short term speech segment.

For voiced, we are getting distinct peaks at some places whereas in unvoiced sounds we are getting random peaks. So, we can use this procedure to identify the voiced and unvoiced from the cepstrum.

B. Liftering:

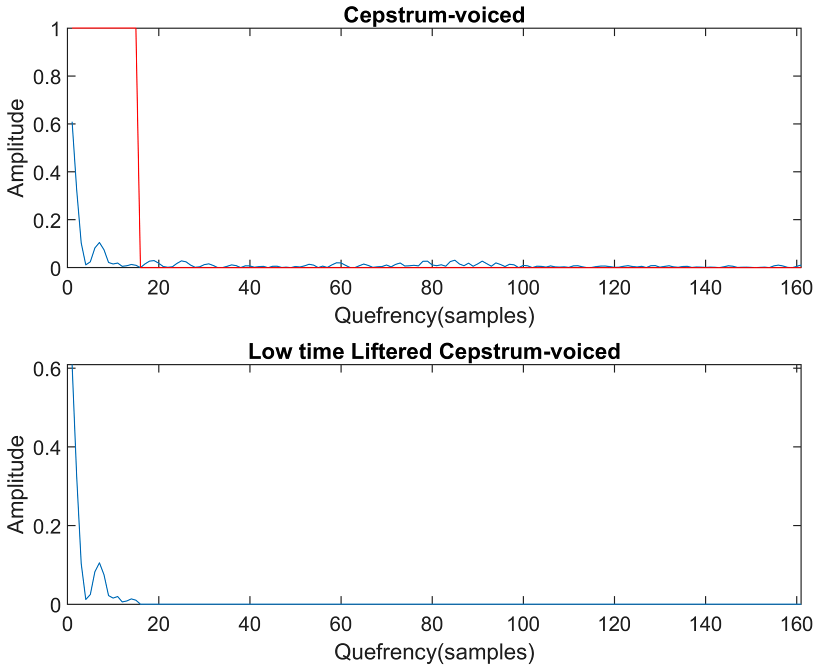
a. Extract the deconvolved vocal tract component and excitation component from the cepstrum by liftering.

b. Write about how you used Low-Time Liftering and High-Time Liftering for extracting the above components.

Code:

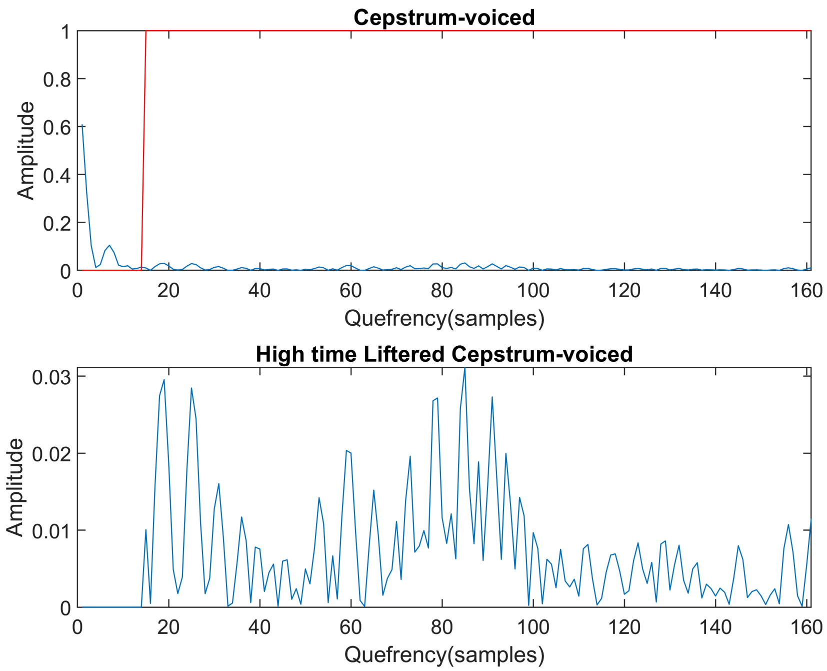
```
% voiced segment

%Low time liftering
c_v_1= c_v(1:ceil(length(c_v)/2)); % As the cepstrum is symmetric, half the cepstral coefficier
L=zeros(1,length(c_v_1)); %For defining liftering window
L(1:15)=1; %Liftering window
c_v_lt=c_v_1.*L; %Low time liftered cepstrum

figure;
subplot(211);
plot(abs(c_v_1));
hold on
plot(L,'r');
```

```
xlim([0,length(c_v_1)]);
title('Cepstrum-voiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
hold off
subplot(212);
plot(abs(c_v_lt));
xlim([0,length(c_v_1)]);
title('Low time Liftered Cepstrum-voiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
```

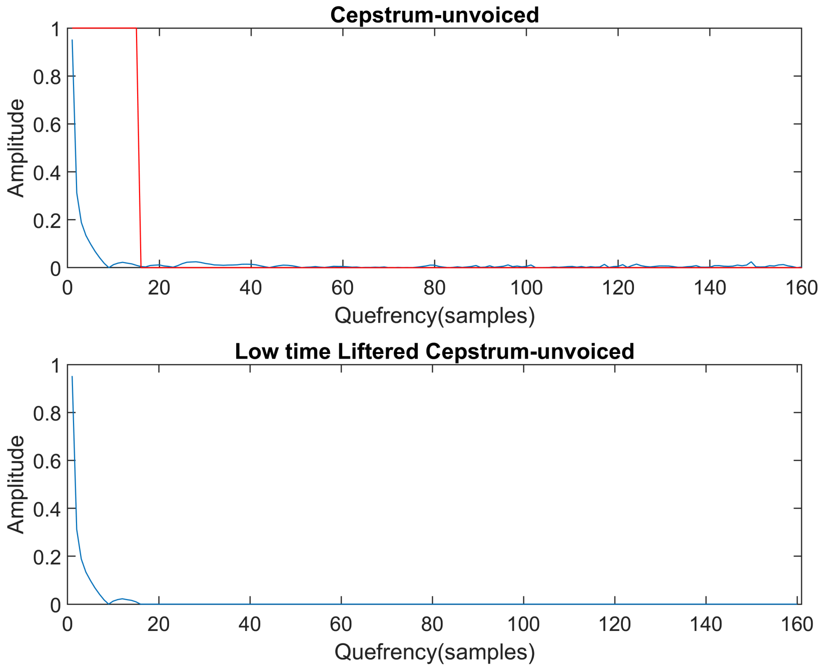


```

%High time liftering
L=zeros(1,length(c_v_1)); %For defining liftering window
L(15:length(L))=1; %Liftering window
c_v_ht=c_v_1.*L; %Low time liftered cepstrum

figure;
subplot(211);
plot(abs(c_v_1));
hold on
plot(L,'r');
xlim([0,length(c_v_1)]);
title('Cepstrum-voiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
hold off
subplot(212);
plot(abs(c_v_ht));
xlim([0,length(c_v_1)]);
title('High time Liftered Cepstrum-voiced');
```

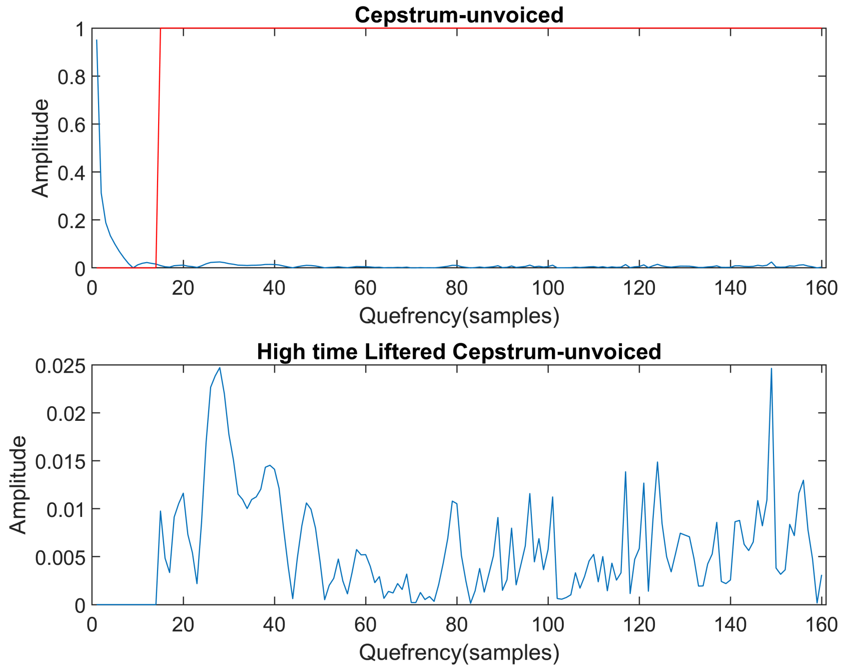
```
xlabel('Quefrency(samples)');
ylabel('Amplitude');
```



```
% unvoiced segment

%Low time liftering
c_uv_1= c_uv(1:length(c_uv)/2); % As the cepstrum is symmetric, half the cepstral coefficients
L=zeros(1,length(c_uv_1)); %For defining liftering window
L(1:15)=1; %Liftering window
c_uv_lt=c_uv_1.*L; %Low time liftered cepstrum

figure;
subplot(211);
plot(abs(c_uv_1));
hold on
plot(L,'r');
xlim([0,length(c_uv_1)]);
title('Cepstrum-unvoiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
hold off
subplot(212);
plot(abs(c_uv_lt));
xlim([0,length(c_v_1)]);
title('Low time Liftered Cepstrum-unvoiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
```



```
%High time liftering
L=zeros(1,length(c_uv_1)); %For defining liftering window
L(15:length(L))=1; %Liftering window
c_uv_ht=c_uv_1.*L; %Low time liftered cepstrum

figure;
subplot(211);
plot(abs(c_uv_1));
hold on
plot(L,'r');
xlim([0,length(c_v_1)]);
title('Cepstrum-unvoiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
hold off
subplot(212);
plot(abs(c_uv_ht));
xlim([0,length(c_v_1)]);
title('High time Liftered Cepstrum-unvoiced');
xlabel('Quefrency(samples)');
ylabel('Amplitude');
```



Observations/procedure:

We can use lowtime liftering to seperate out the vocal tract information and hightime liftering to seperate out the excitation information.

C. Pitch estimation by cepstral analysis:

a. In the case of the voiced segment, estimate the pitch of the voiced speech segment using the cepstral analysis. Explain your procedure.

Pitch can be estimated as the quefrency location of the highest peak in the high-time liftered cepstrum

Code:

```
[y_val,y_loc]=max(c_v_ht); % Finding the peak in the high time liftered cepstrum
% Location of the peak gives pitch period in quefrency samples
pitch_period=y_loc
```

```
pitch_period = 85
```

```
% converting pitch period in samples into frequency
pitch_frequency=(1/pitch_period)*fs
```

```
pitch_frequency = 188.2353
```

Observations/procedure:

Pitch can be estimated as the quefrency location of the highest peak in the high-time liftered cepstrum.

pitch period=85 samples

pitch frequency= 188.2353Hz

D. Formant estimation by cepstral analysis:

a. Using liftered cepstrum, estimate the frequencies of the first three resonances of the vocal tract of the voiced speech frame.. Explain your procedure. Plot the log magnitude spectrum that shows the formant information (while not having the excitation information).

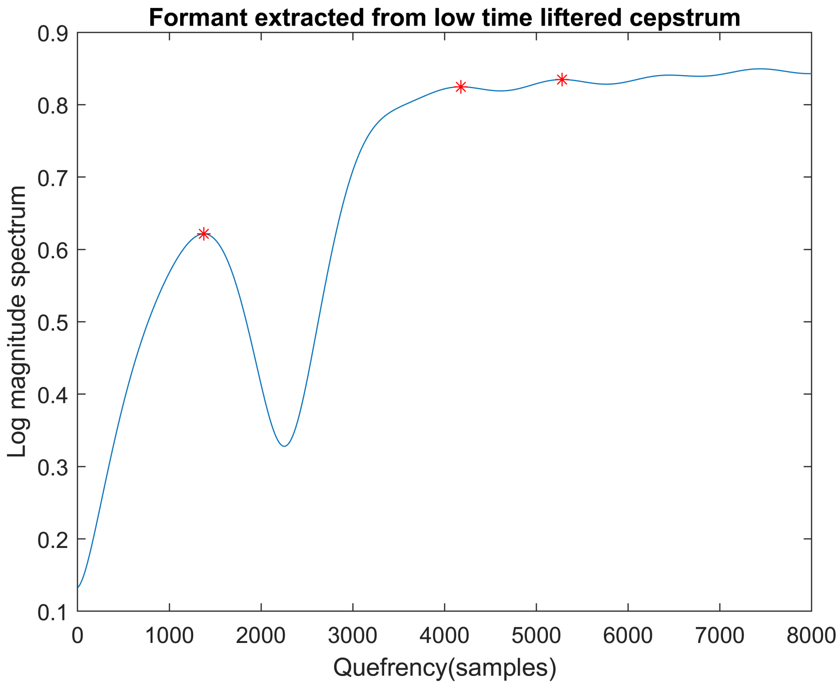
b. Repeat the above for the unvoiced speech frame.

Code:

```
%Low time liftering
c_v_lt_1=c_v_lt(1:15); % Taking the non-zero cepstral coefficients from the low time lifter cep
% 16000 point FFT is taken to find the log magnitude spectrum of the low
% time lifted cepstral coefficients
C_v_lt_1 = fftshift(fft(c_v_lt_1,16000));
C_v_lt_1=C_v_lt_1(1:8000);
C_v_lt_1=abs(C_v_lt_1);

%peak picking algorithm
k=1;
for i=2:length(C_v_lt_1)-1
    if (C_v_lt_1(i-1)<C_v_lt_1(i)) && (C_v_lt_1(i+1)<C_v_lt_1(i))
        formant_mag_v(k)=C_v_lt_1(i);
        formant_v(k)=i;
        k=k+1;
    else
        continue;
    end
end
formant_v(1:3)
```

```
ans = 1×3
        1376        4176        5280
```

```
figure;
plot(C_v_lt_1);
hold on;
plot(formant_v(1:3),formant_mag_v(1:3),'r*');
hold off;
title('Formant extracted from low time liftered cepstrum');
xlabel('Quefrency(samples)');
ylabel('Log magnitude spectrum');
```

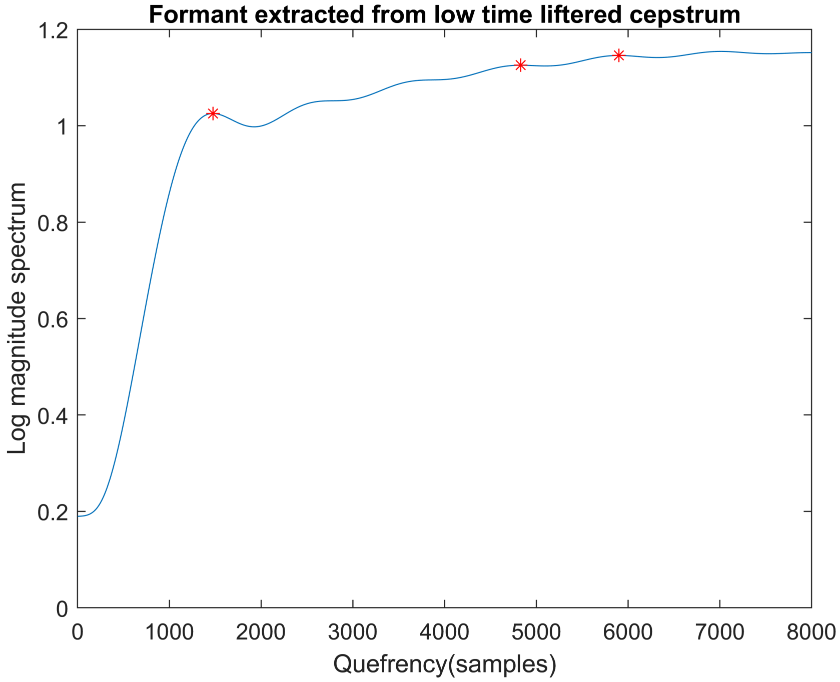


```
%Low time liftering
c_uv_lt_1=c_uv_lt(1:15); % Taking the non-zero cepstral coefficients from the low time lifter c
% 16000 point FFT is taken to find the log magnitude spectrum of the low
% time lifted cepstral coefficients
C_uv_lt_1 = fftshift(fft(c_uv_lt_1,16000));
C_uv_lt_1=C_uv_lt_1(1:8000);
C_uv_lt_1=abs(C_uv_lt_1);
F_uv = -fs/2 : fs/length(Y_uv) : fs/2 - fs/length(Y_uv);

%peak picking algorithm
k=1;
for i=2:length(C_uv_lt_1)-1
    if (C_uv_lt_1(i-1)<C_uv_lt_1(i)) && (C_uv_lt_1(i+1)<C_uv_lt_1(i))
        formant_mag_uv(k)=C_uv_lt_1(i);
        formant_uv(k)=i;
        k=k+1;
    else
        continue;
    end
end

formant_uv(1:3)
```

```
ans = 1×3
        1476        4828        5898
```

```
figure;
plot(C_uv_lt_1);
hold on;
plot(formant_uv(1:3),formant_mag_uv(1:3),'r*');
```

```
hold off;
title('Formant extracted from low time liftered cepstrum');
xlabel('Quefrency(samples)');
ylabel('Log magnitude spectrum');
```



Observations/procedure:

(voiced)Frequencies of the first three resonances: 1376Hz, 4176Hz, 5280Hz

(unvoiced)Frequencies of the first three resonances: 1476Hz, 4828Hz, 5898Hz

Formant locations can be estimated from vocaltract spectral characteristics. This can be computed from spectral representation of the low-time liftered cepstral coefficients. As the DFT of the low-time liftered cepstral coeffcients gives the corresponding smooth log magnitude spectrum, the formants can be located using a simple peak picking algorithm.

Both have a similar characterstic of increasing nature.