Open lab –Speech/Audio signal Processing using MATLAB

Lab Sheet 5

Cepstral Analysis of Speech Signals

Aim

- To understand basic Cepstral Analysis approach, applied to speech
- 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.

De-convolution

A signal coming out from a system is due to the input excitation and also the response of the system. From the signal processing point of view, the output of a system can be treated as the convolution of the input excitation with the system response. At times, we need each of the components separately for study and/or processing. The process of separating the two components is termed as de-convolution.

In the first case, if we knew the input excitation, then the system component can be separated/ constructed by exciting the system with the inputs and collecting its responses. This is what is done in same channel estimation problems. In the second case, if we knew the system response, then the input excitation can be recovered using the inverse filter theory concept. For instance, Linear Prediction (LP) analysis of speech to recover excitation. There is yet another type of deconvolution, where the assumption is both input excitations as well as system responses are unknown. The present study of cepstral analysis of speech comes under this category.

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 a priori knowledge about source and / or 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:

$$s(n) = e(n) * h(n) \qquad (1)$$

This can be represented in frequency domain as,

$$S(\omega) = E(\omega).H(\omega) \qquad (2)$$

The Eqn. (2) indicates that the multiplication of excitation and system components in the frequency domain for the convolved sequence of the same in the time domain. 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

From the Eqn. (2) the magnitude spectrum of given speech sequence can be represented as,

$$|S(\omega)| = |E(\omega)|.|H(\omega)| \quad (3)$$

To linearly combine the E(ω) and H(ω) in the frequency domain, logarithmic representation is used. So the logarithmic representation of Eqn. (3) will be,

$$log|S(\omega)| = log|E(\omega)| + log|H(\omega)| \quad (4)$$

As indicated in Eqn. (4), 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 i.e. log operation converted the "*" operation into "+" operation in the frequency domain. 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. It should be noted that 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. This is mathematically explained in Eqn. (5). 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.

$$c(n) = IDFT(log|S(\omega)|) = IDFT(log|E(\omega)| + log|H(\omega)|) \quad (5)$$

Figure 1 details the various steps involved in converting the given short term speech signal to its cepstral domain representation. The obtained cepstrum contains vocal tract components which are linearly combined according Eqn.(5). As the cepstrum is derived from the log magnitude of the linear spectrum, it is also symmetrical in the quefrency domain. Here also only one symmetric part of the cepstrum is used for plotting.

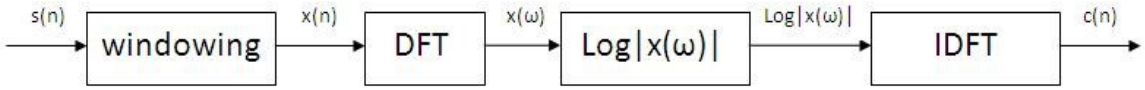

Figure 1: Block diagram representing computation of cepstrum

Experiment1

1. Convert the speech into short-term segments of size 15-30 ms. Here the frame size is fixed to 30 ms. Then each frame is multiplied by a hamming window. Then cepstral representation of short-term speech is computed by finding the IDFT of the log magnitude spectrum.
2. Compare the cepstrum of voiced and unvoiced speech segments

It can be observed that 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. 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. Following section

describes about the liftering operation performed to extract the vocal tract and excitation features independently from 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,

$$w_e[n] = \begin{cases} 1, 0 \leq n \leq L_c \\ 0, L_c \leq n \leq \frac{N}{2} \end{cases} \quad (6)$$

where Lc is the cut off length of the liftering window and N/2 is half the total length of the cepstrum. Usually Lc is used as 15 or 20. The vocal tract characteristics can be obtained by multiplying the cepstrum c(n) with the low-time liftering window as indicated in Eqn. (7).

$$c_e(n) = w_e[n] . c(n) \quad (7)$$

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 as given in Eqn. (8).

$$\text{Log}[|H(w)|] = \text{DFT}[c_e(n)] \quad (8)$$

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. The block diagram given in Figure 2 shows the process of formant estimation using low-time liftering. The formants locations obtained from the peaks in the vocal tract spectrum.

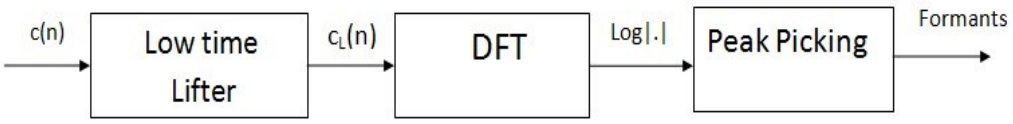

Figure 2: Block diagram representing low-time liftering

Experiment2

3. Find the Formants from cepstrum using low-time liftering

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,

$$w_h[n] = \begin{cases} 1, L_c \leq n < \frac{N}{2} \\ 0, else \end{cases} \quad (9)$$

where *Lc* is the cut off length of the liftering window and *N/2* is the half the total length of the cepstrum. Usually *Lc* is used as 15 or 20. The excitation characteristics are obtained by multiplying high time liftering window with the cepstrum obtained as given in Eqn. (7).

$$c_h(n) = w_h(n) * c(n) \quad (10)$$

The block diagram given in Figure 3 indicates the high-time liftering process for pitch estimation. The computation of high-time liftered cepstrum from the cepstrum using high-time liftering window .Pitch can be estimated as the instant corresponds to the highest peak in the high-time liftered cepstrum. 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.

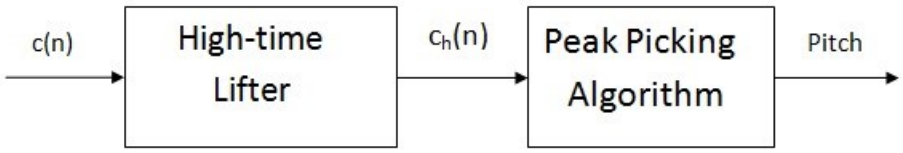

Figure 3: Block low-time liftering diagram representing high-time liftering

Experiment3

4. Find the Pitch from cepstrum using high-time liftering