LABORATARY EXERCISE 1
Speech Modeling, Analysis, Synthesis and Compression

Mengbai Tao, Guillermo Rodríguez Ferrández and Yimeng Hou

March 31, 2014

Task 1: Stationary (Single Tone) Speech Signals: Modeling,Analysis and Synthesis

1.1 Generate a stationary speech signal

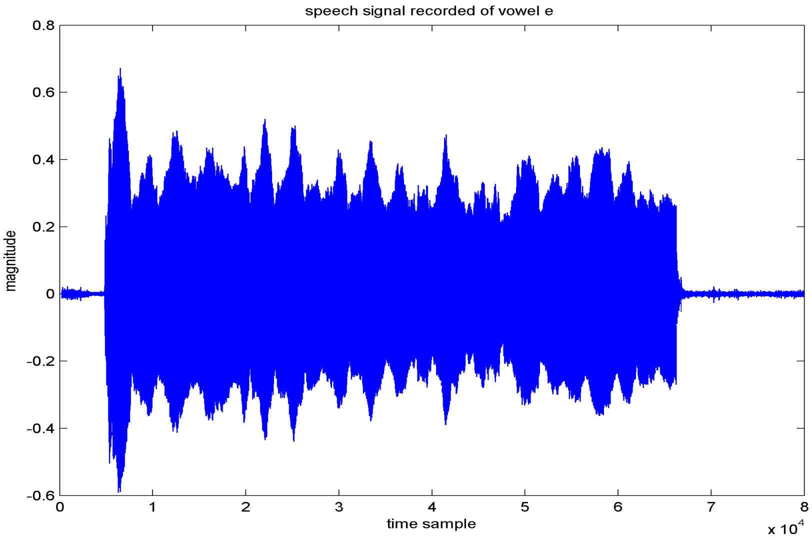

Figure 1.1: A vowel signal of ’e’.

1.2 Estimate the LPC model parameters

The number of samples contained in the frame (300ms,fs =8kHz): $L = 0.3 * 8000 = 2400$

The estimated H(z) LPC parameters are: $a_j, j = 1, 2...p,$ =
1 -0.7333 -0.7882 0.0821 0.6158 0.2426 -0.3082 -0.0730 0.2831 -0.3821 0.2490

The variance of prediction errors $\approx 3.1839e^{-4}$

1.3 Calculate the residual sequence e(n)

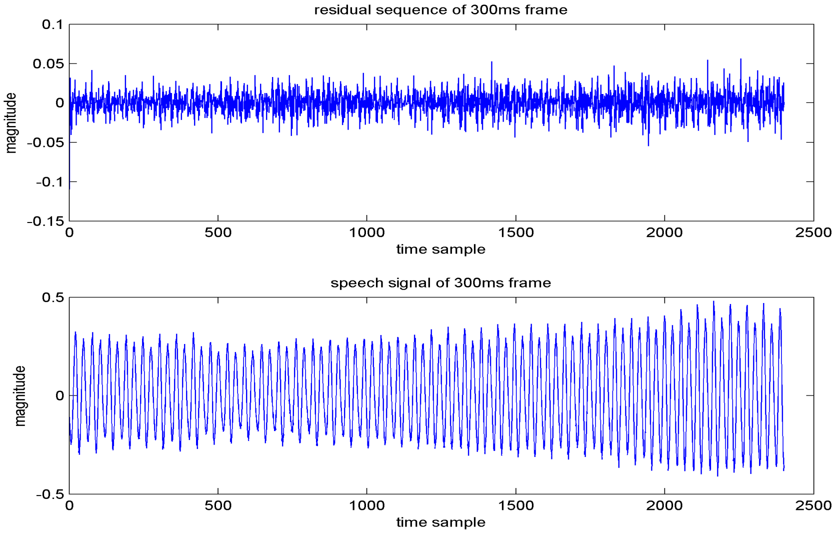

Figure 1.2: Comparsion between original signal and corresponding residual sequence.

1.4 Re-synthesize the speech using the estimated parameters

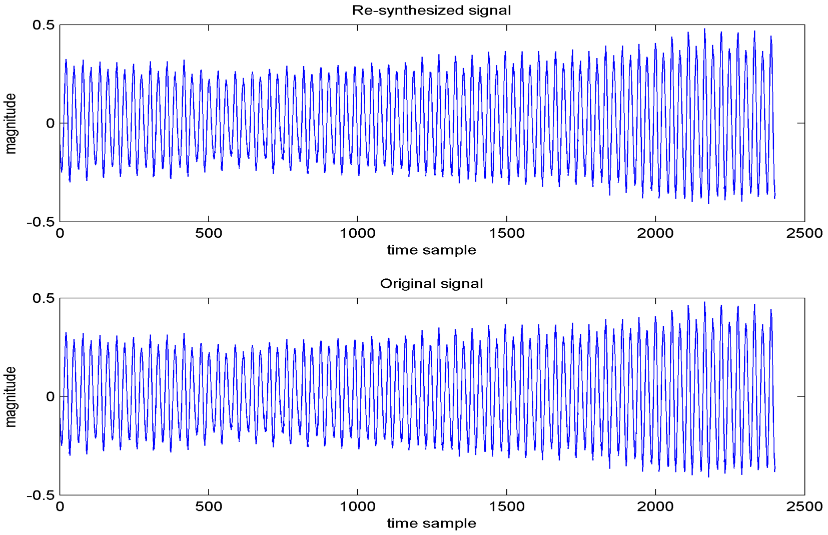

Figure 1.3: Comparsion between original speech and re-synthesized speech

Summary

In this part, linear prediction coding (LPC) is used to predict the the future sample based on the previous p samples. By doing this, the 'redundant' parts of transmitted signal is removed and can be predict by receiver and thus, save space,time and power. The quality difference between these two speech is almost unpreceivable.

Task-2: Nonstationary Speech Signals: Modeling, Analysis and Synthesis

2.1 Recording a "natural" speech signal containing a sentence

The original speech .wav file is included in the .zip file.

2.2 Block-based speech analysis

The total number of samples in the speech signal = 120000 The total number of blocks: TotalBlocks = 750 The sampling rate f_s= 8000 Hz

2.3 Block-based estimation of residual sequence $\hat{e}(n)$

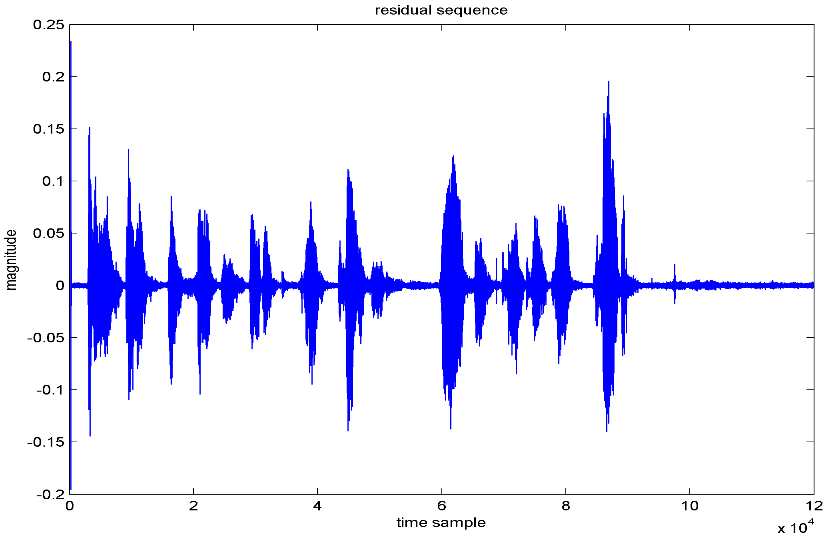

Figure 2.1: The entire residual sequence of the speech sentence.

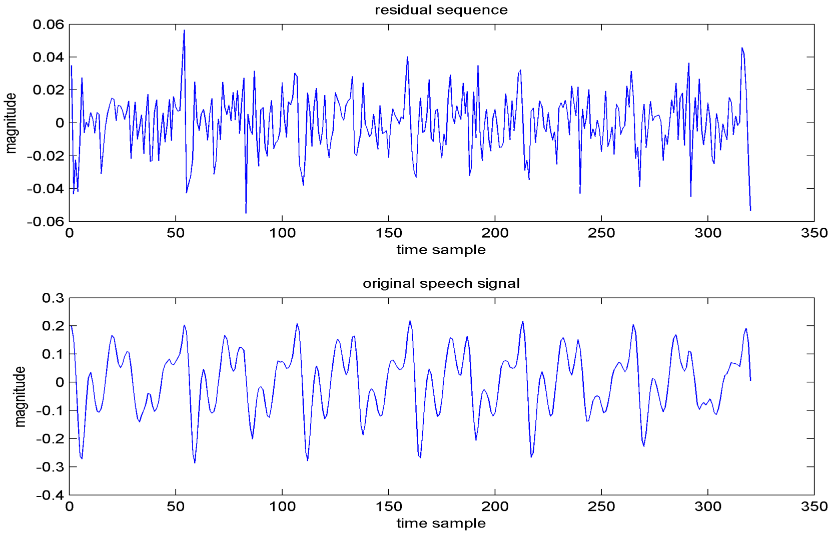

Figure 2.2: Comparsion between original sequence and residual sequence.

Compare the plots in 2.3.a and 2.3.b, the second one is easier to observe the pitch period

Count the approximate number of samples in one pitch period (manually counted) = 55

The speech frequency is approximately $1/(55/8000) \approx 145Hz$

2.4 Block-based speech re-synthesis

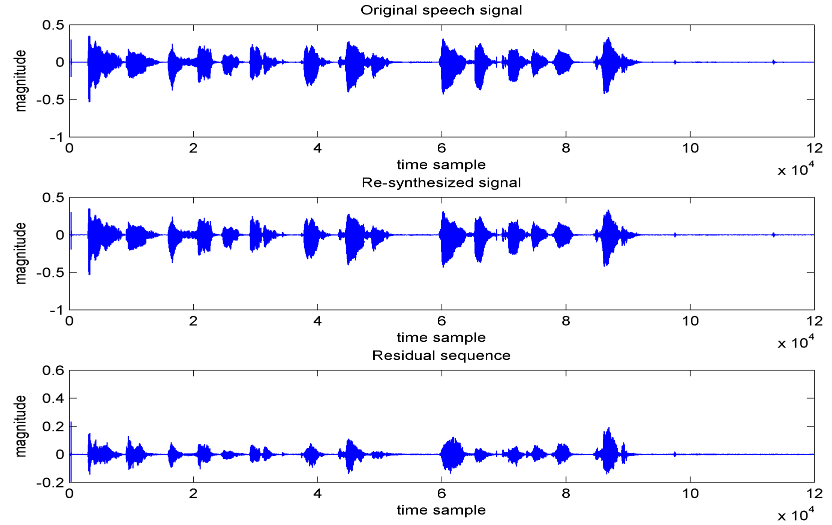

Figure 2.3: The original signal, the re-synthesized signal and the residual sequence.

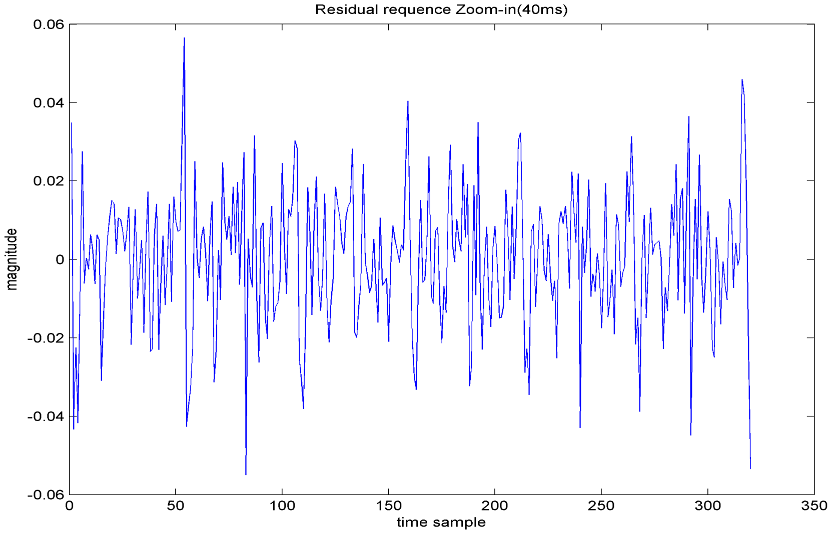

Figure 2.4: Zoom-in of the residual sequence.

From the figures and sounds, it is concluded that the synthetic speech is distinguishable and mostly clear but a little distorted regarding to the original speech. Especially when another consonant is started after vowel sounds.This distortion may due to the unperfect coefficients estimate from the LPC model, i.e. $\sigma^2 \neq 0$

Summary

For nonstationary speech signals, it is decomposed and characterized by the residual sequence and LPC parameters. To re-synthesize the signal, it can be simply done by inversely fliltering the residual sequences with a all-zero filter. The quality of the re-synthesized signal is a little distorted regarding to the original one, especially when the filter order p is large.

Task-3: Re-synthesize Speech by Using K Most Significant Residuals/Block as the Excitations

The re-synthesized speech .wav file is also included in the .zip file. From this figure, we could see the difference between task 2.

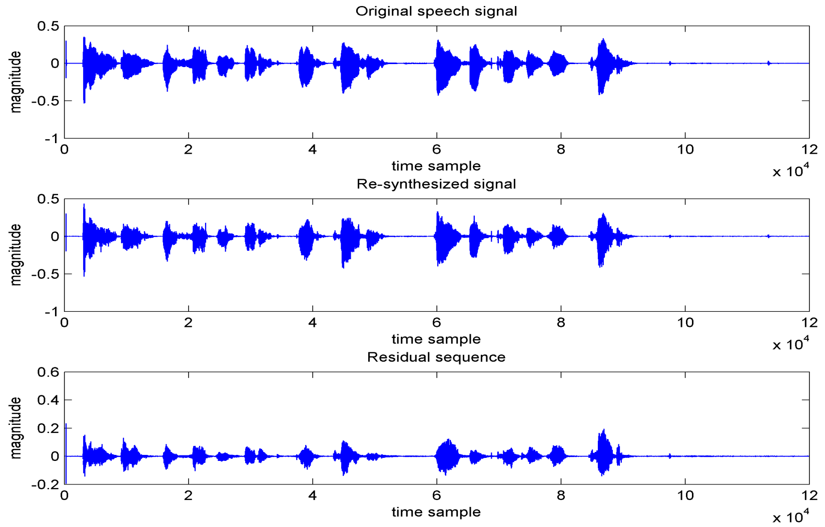

Figure 3.1: The original speech, the re-synthesized speech and the residual sequence.

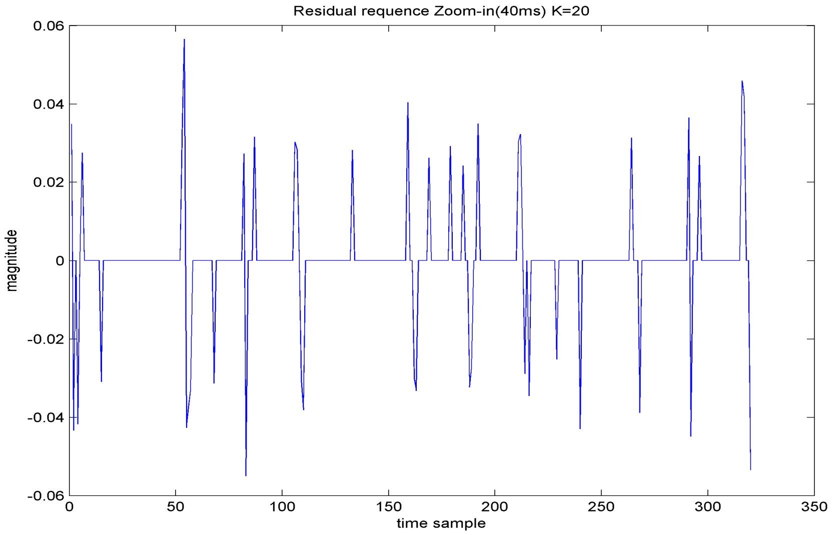

Figure 3.2: Zoom-in of the residual sequence.

Summary

By using the K most significant residuals in each block as excitations, it is concluded that the speech is still clearly enough to understand. However the distortion is aggravated comparing to task 2. In this case, greater compression rate is achieved at the cost of quality degradation.

Task-4: Using Cepstrum to Estimate Pitch Periods in Voiced Speech

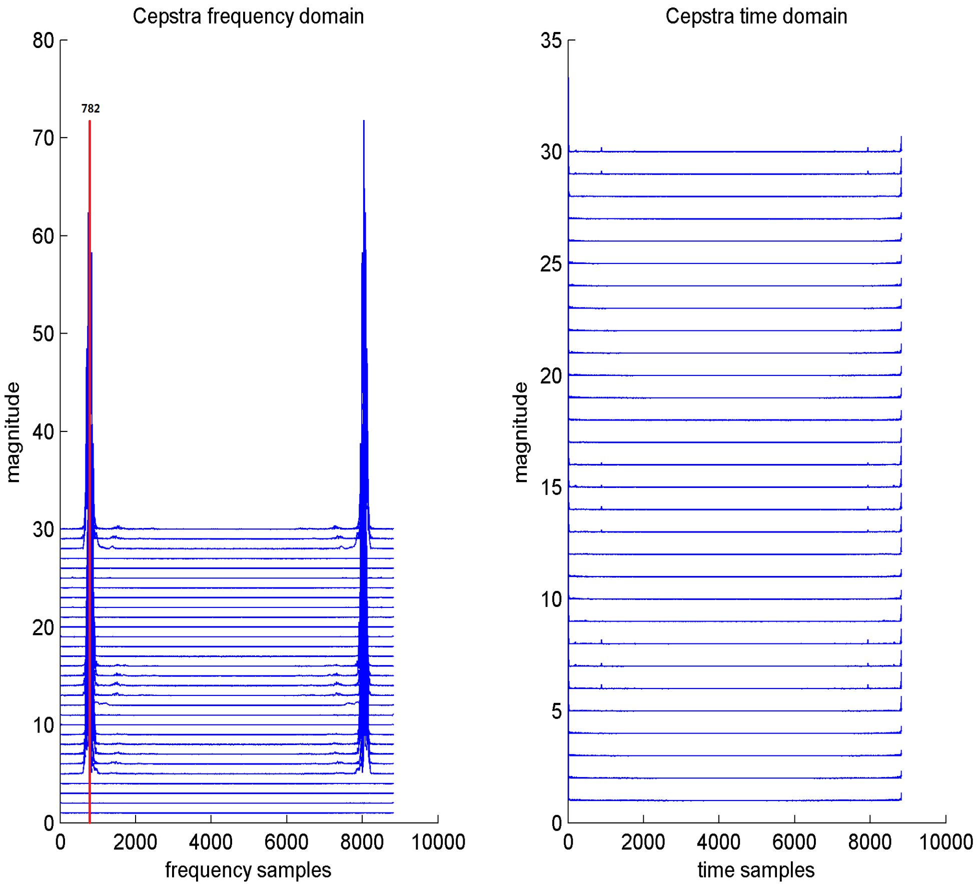

Figure 4.1: The cepstrum of voiced speech

Summary

The pitch should be equal to 782/8820*44100 ≈ 3910 Hz from the frequency-domain. Or equal to 782/882*(44100/10) ≈ 3910 Hz from the time-domain.

The pitch frequency may be easily covered by other frequency components. By using logarithm operation and switching the nonlinear (multiply) relation between those components into linear (addition) relation, the frequency corresponding to the pitch component will be easier to be obvious. On the other hand, using cepstrum method will increase some related frequency component and hence, the SNR is decreased.

Task 5: Objective Measures for Synthetic Speech Quality

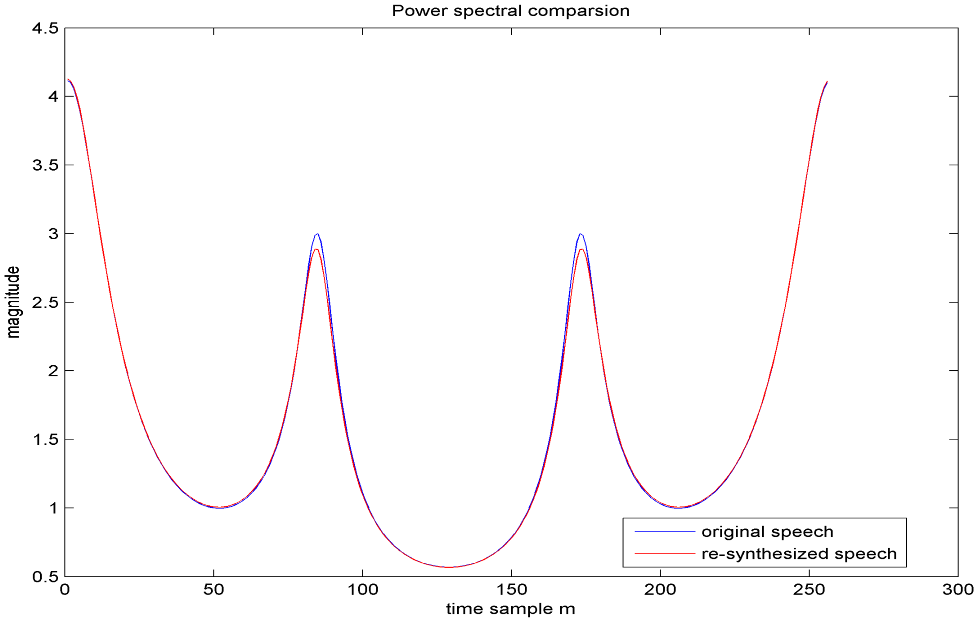

Figure 5.1: Power spectral Comparsion.

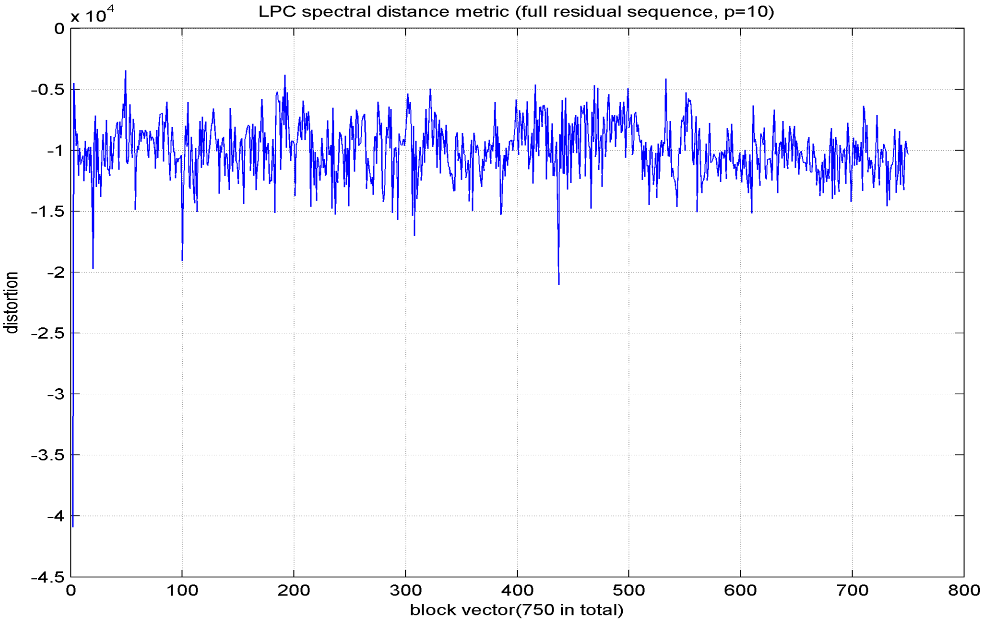

Figure 5.2: Distortion in different block when filter order p = 10, complete residual sequence case.

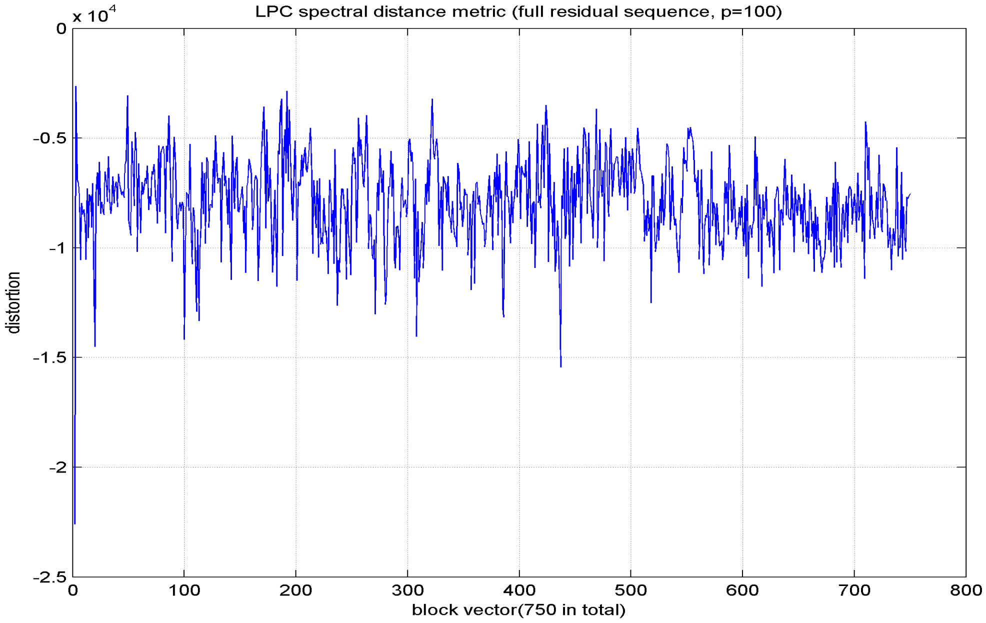

Figure 5.3: Distortion in different block when filter order p = 100, complete residual sequence case.

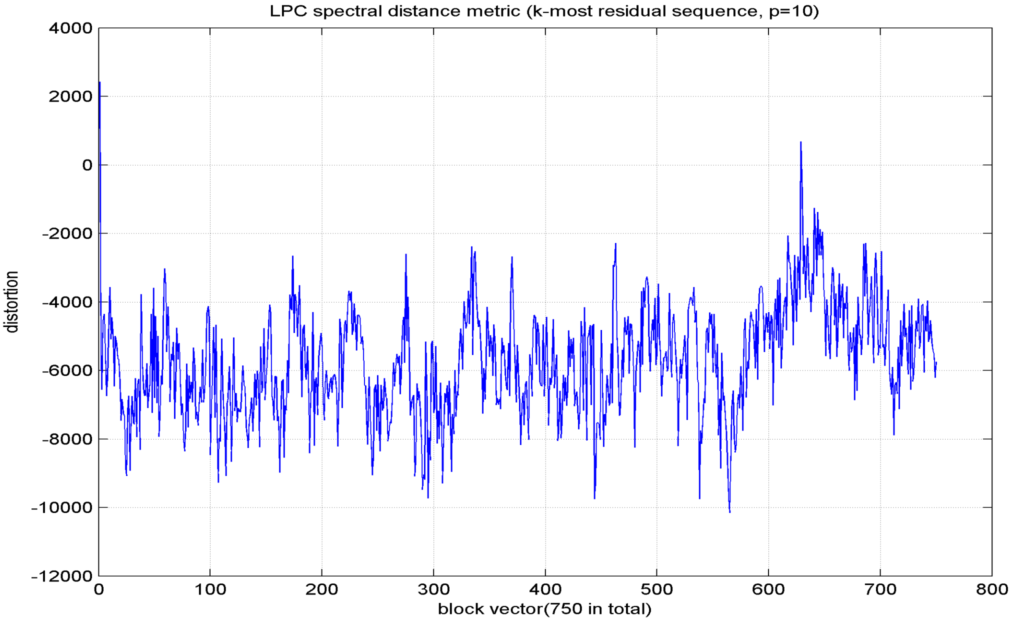

Figure 5.4: Distortion in different block when filter order p = 10, k-most residual sequence case.

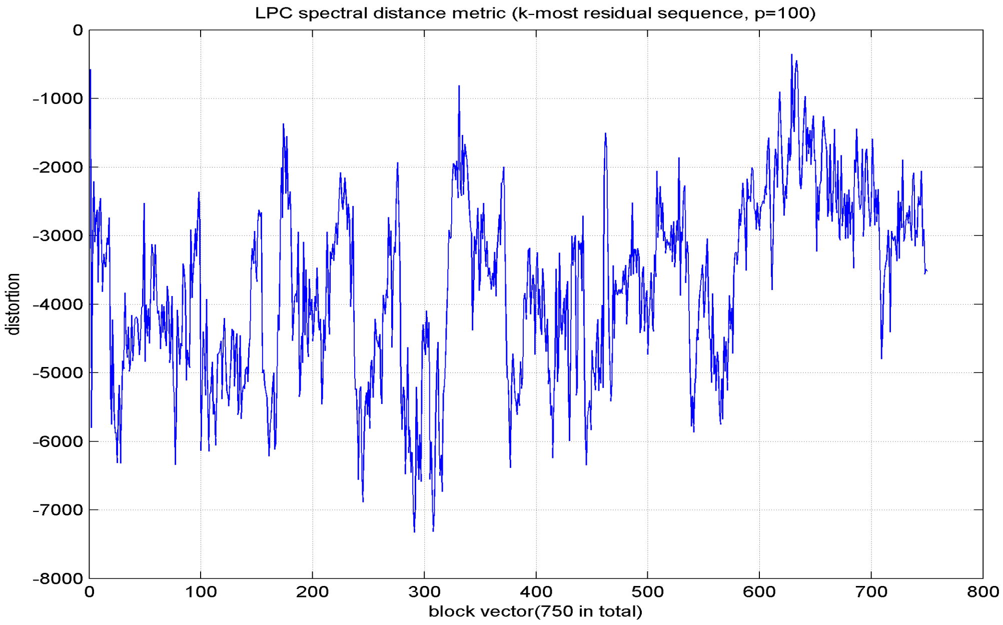

Figure 5.5: Distortion in different block when filter order p = 100, k-most residual sequence case.

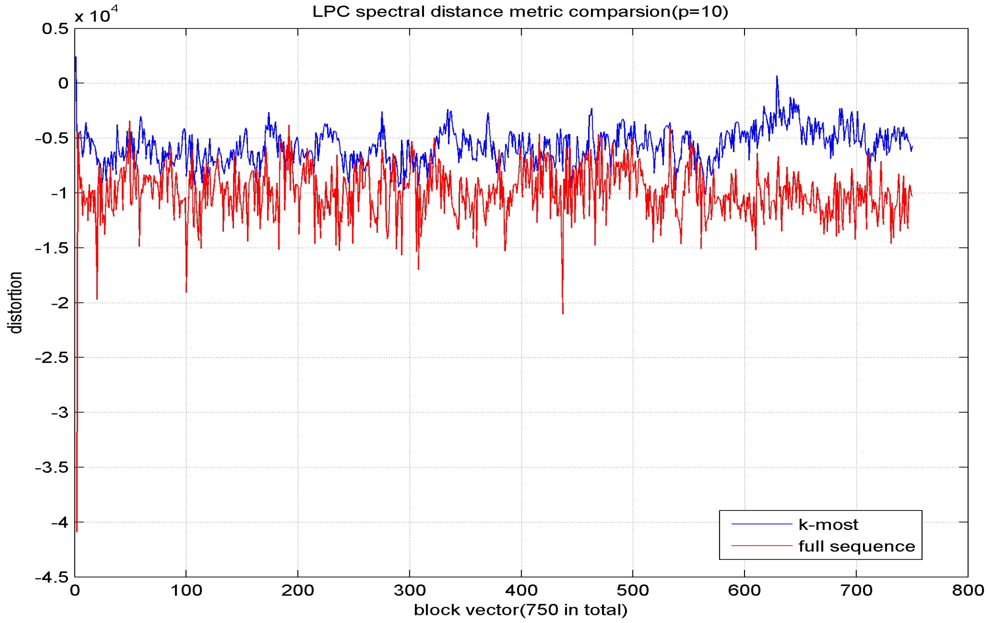

Figure 5.6: Distortion in different block when filter order p = 10, task 2 and task 3 comparsion.

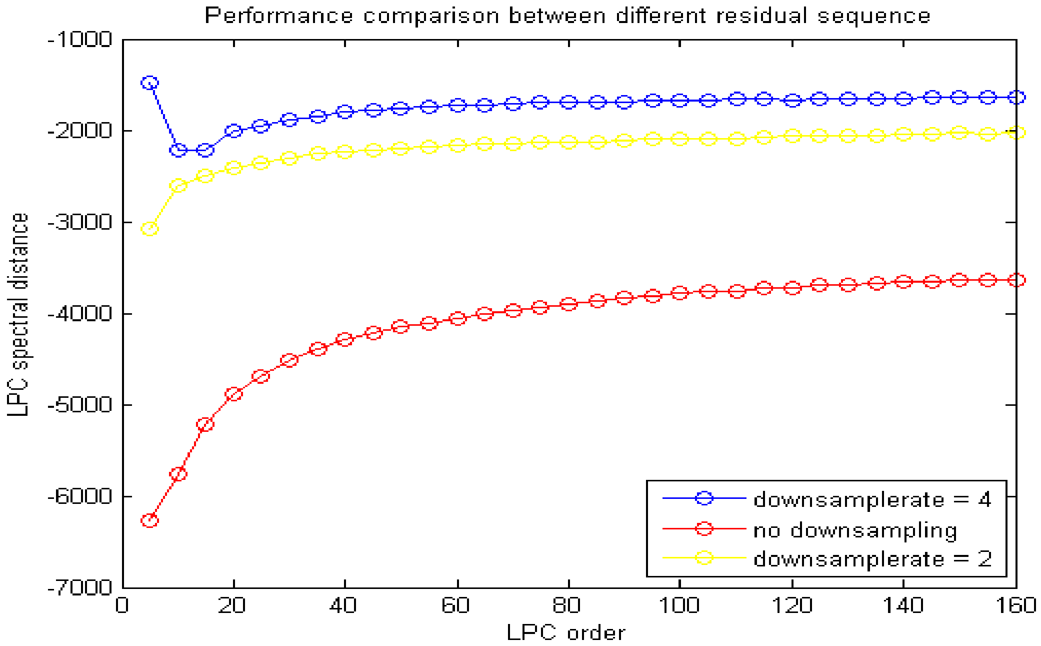

Figure 5.7: Distortion in different LPC orders and different residual sequences (using k-most method), every single circle in the figure means the averaged distortion value of all the blocks.

Summary

The distortion in different cases are shown in above figures. For figure 5.1, since for the metric, minus infinity means absolutely accurate estimate and 0 means totally wrong, the compression quality is acceptable. However, changing the LPC model order p or the number of residuals has different effects on the performance. Figure 5.7 simply shows the distortion with different filter order and different number of residual sequences. Specifically, the residual sequence will be down-sampled at a rate of 2 or 4 . Then after the re-synthesizion, the samples will be up-sampled to the original length. The purpose of multirate sampling is to change the number of residual sequence in a decent manner.Then the distortion vector in each block will be averaged and presented only by a scalar. Generally, it is concluded that a large filter order will lead to worse speech quality. It is because the filter updating speed is too low to follow and predict the sudden changes,i.e. each beginning of consonants. In addition, insufficient filter coefficients or lack of residual samples are both possible factors that degrade the speech performance.

An interesting phenomenon observing from Figure 5.2 - Figure 5.5 is that distortion variance increases with the reduction of residual sequences. This may simply because less residual sequence leads to large differences between each samples, i.e obvious bias in the zero-value samples in the residual sequences. Another conclusion derived from Figure 5.6 is that the quality from task 2 is obvious better than task 3. By using the k-most significant samples from residual sequence, some information is lost. So the conclusion is consistent with theory.

Task 6: Writing the Report

1. How speech signal compression is achieved in this Lab. work? (brief, in less than 2 lines of text!)

It is achieved by presenting speech signal in terms of compact and sufficient paramaters from LPC models plus residual sequence to save space.

2. Suppose each speech signal sample consists of 8 bits, and the sample rate isfs=8 kHz. How many bits are required to encode one block (20ms) of uncompressed speech signals?

$N = 8000 * 0.02 * 8 = 1280$

3. Using the (LPC) speech model in Fig.2 where the all-pole model is specified by Eq.(1), list the minimum number of parameters that are required to encode one block (20ms) of speech.

12 parameters in are needed to encode each frame.
(10 paramaters from LPC coefficients ,one from error variance and the last one from pitch)

4. Assume each parameter in the above step 3 is encoded (in average) by 8 bits. How many bits do you require to encode one block of compressed speech under such a model?

$N = 12 * 8 = 96$

5. What is the compression ratio between the uncompressed speech (step 2) and the LPC compressed speech (steps 3 and 4) ?

In this implementation, the audio size of original speech is $120000 * 8 = 960000$, assuming $f_s = 8000$
After picking the k-most significant samples, compressed audio size is (k+12)* N_{block} * 8 bits. When k=20, this led to a compression ratio of 20.

6. There are many possible ways to compress audio/speech signals ? Write downone name of those speech compression methods that are based on non-parametric methods.

Sub-band coding