Title: Term Project

Members: Wenjun Zhang Jinzhao Feng

A53218995 A53213089

Class: ECE252 – Speech Compression

Date: 03.17.2017

Description

In this project, we record a sentence “I have a pen, I have an apple” with a sampling rate of 8kHz. Following the guidelines described in Question 9.4 in the text, we implement an LPC-10 encoder and design a decoder which can read stored data and synthesize the speech.

Approach

Encoder

A referenced block diagram of the LPC encoder is shown in Figure.1. The input speech is first segmented into non-overlapping frames. A pre-emphasis filter is used to adjust the spectrum of the input signal. But in our project, by adding the pre-emphasis filter the reconstructed sound actually becomes worse, so we will just skip this procedure.

The voicing detector classifies the current frame as voiced or unvoiced and outputs one bit indicating the voicing state.

The segmented signal is used for LPC analysis, where10 LPCs are derived for voiced segments and 4 LPCs for unvoiced segments.

Instead of from the prediction-error, in our project, pitch period is estimated from the segmented signal, since by doing so we actually obtain higher quality reconstruction.

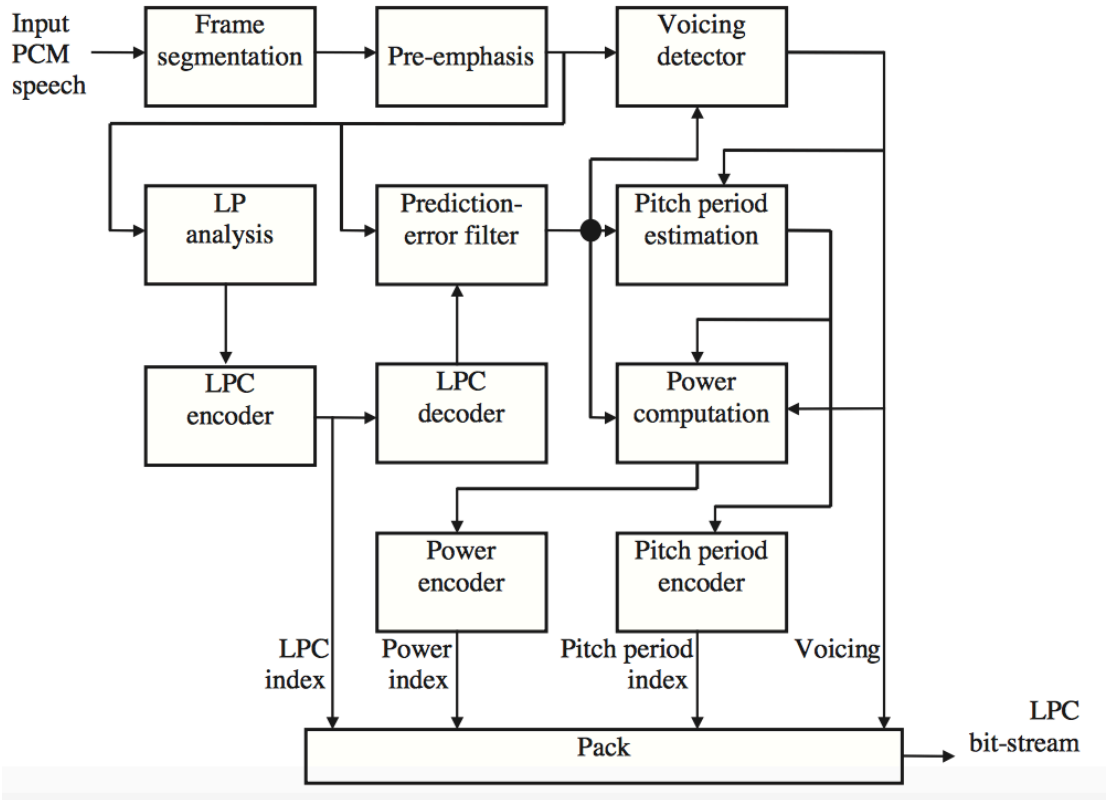


Figure.1 Block diagram of LPC encoder

Voiced / Unvoiced decision

We use 2 methods to decide whether a frame is voiced or unvoiced, which is low band energy and zero crossing rate. If a frame satisfies either of the following conditions, we will decide it to be voiced.

Low Band Energy

This is the most obvious and simple indicator of "voicedness". Typically, voiced sounds are several order of magnitude higher in energy than unvoiced signals. For the frame (of length N) ending at instant m, the energy is given by

$$E[m] = \sum_{n=m-N+1}^{m} s^2[n]$$

Since voiced speech has energy concentrated in the low-frequency region, due to the relatively low value of the pitch frequency, better discrimination can be obtained by low-pass filtering the speech signal prior to energy calculation. That is, only energy of low frequency components is taken into account. A bandwidth of 800Hz is applied for the purpose since the highest pitch frequency is around 500 Hz.

In the project, if the low band energy of a frame is higher than 5, we decide it to be voiced.

Zero Crossing Rate

The zero crossing rate of the frame ending at time instant m is defined by

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

with sgn() the sign function returning ± 1 depending on the sign of the operand. Note that a zero crossing is said to occur if successive samples have different signs.

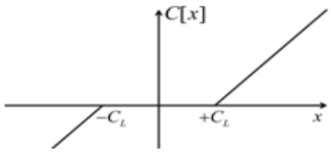
For voiced speech, the zero crossing rate is relatively low due to the presence of the pitch frequency component (of low-frequency nature), whereas for unvoiced speech, the zero crossing rate is high due to the noise-like appearance of the signal with a large portion of energy located in the high-frequency region.

In the project, if the zero crossing rate of a frame is non-zero and lower than 45, we decide it to be voiced.

Pitch Periods Estimation

Speech is stationary in short time, so segmented speech is stationary. A periodic function has a periodic autocorrelation, so we can estimate the pitch period of a frame by finding the correct peak.

We apply a center clipper before the autocorrelation as below and take $C_L=10\%X_{max}$. A clipper can help eliminate most of the extraneous peaks and a clear indication of periodicity is retained.



- C_L=% of A_{max} (e.g., 30%)
- Center Clipper definition:
  - if $x(n) > C_L$, $y(n)=x(n)-C_L$
  - if $x(n) \leq C_L$, $y(n)=0$

Computing the autocorrelation of the segmented signal, we first removed the first big samples, then find the peak index of the autocorrelation. And finally, the pitch period estimation will be the index plus 10.

We tried to use AMDF in this part, but the quality becomes worse. So we choose to only use autocorrelation method.

Gain

It is reasonable to expect the model gain g to be determined by matching the signal energy with the energy of the linearly predicted samples.

For unvoiced frame:

$$p = \frac{1}{N}\sum_{n=0}^{N-1} e^2[n] \qquad g = \sqrt{p}$$

For voiced frame:

$$p = \frac{1}{\lfloor N/T \rfloor T}\sum_{n=0}^{\lfloor N/T \rfloor T-1} e^2[n] \qquad g = \sqrt{Tp}$$

Waveform

The waveform of a voiced frame is shown in Figure.2. We can see that it's periodic and with high amplitudes.

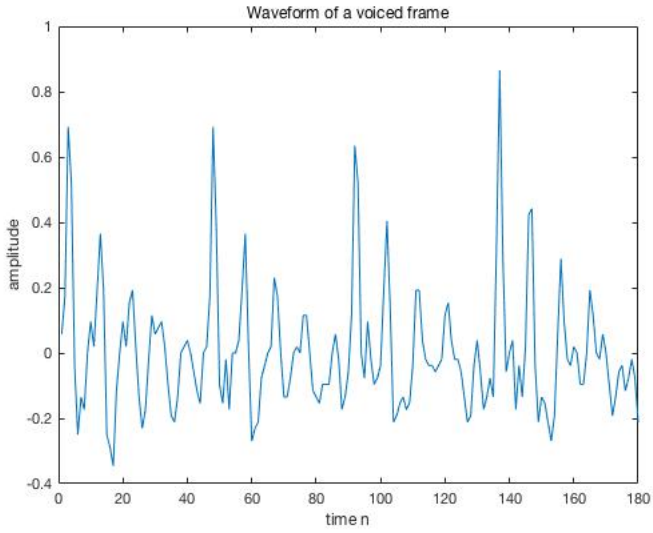


Figure.2 Waveform of a voiced frame

Decoder

A referenced block diagram of the LPC decoder is shown in Figure.3. It's essentially the LPC model of speech production with parameters controlled by the bit-stream. It is assumed that the output of the impulse train generator is comprised of a series of unit-amplitude impulses. While the white noise generator has unit-variance output.

Denoting the gain by g, gain computation is performed as follows.

$$g = \begin{cases} \sqrt{p}, & for\ unvoiced\ case \\ \sqrt{Tp}, & for\ voiced\ case \end{cases}$$

And similarly, we skip the de-emphasis part.

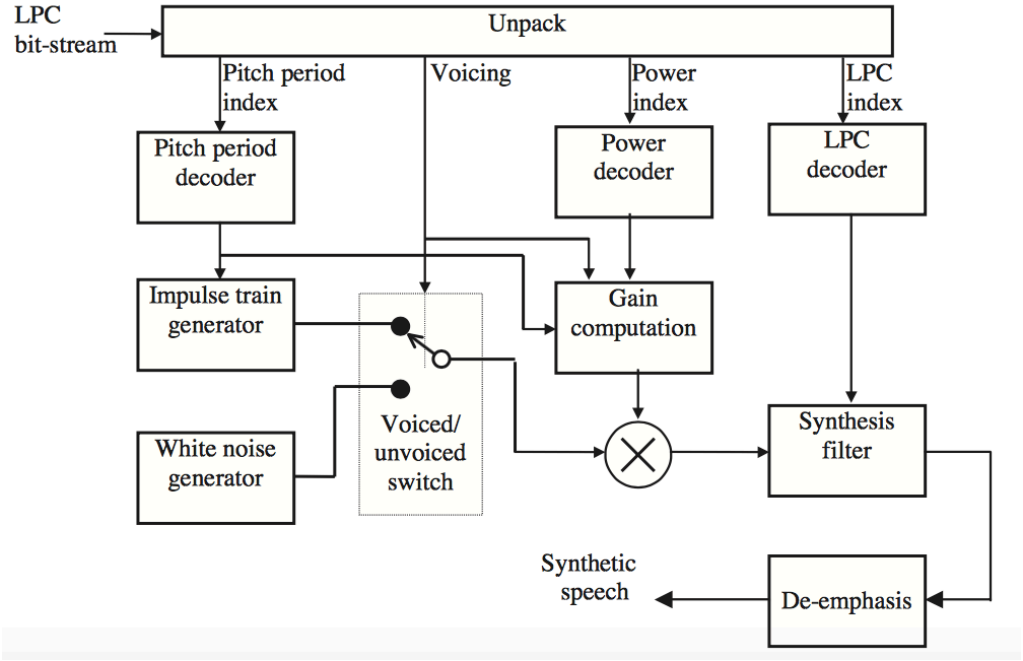


Figure.3 Block diagram of LPC decoder

Following the diagram above, we are able to synthesisze the speech in the decoder with the parameters stored in the ".mat" files (which served as LPC bit-stream).

Comment

The quality of the synthesized speech is not good but identifiable. This is because the perfect excitation is the error computed in the encoder, while the excitation we generate is actually quite different from it. And it's hard to generate excitation like the error using LPC-10 method.