

SPEECH COMPRESSION

Linear Predictive Coding

- Back in 1972, state of the art in speech coding was straight PCM at 64Kb/sec
- The required bandwidth for the transmission of one voice channel was 64KHz
- Keep in mind that one voice channel in telephony is only 4KHz wide

SPEECH VIA SYNTHESIS

- In order to compress speech below 8bits/sample, a completely different approach has been adopted
- This approach is based on modeling speech and transmitting the parameters of the model rather than the waveform itself

AN EXAMPLE

- We are all familiar with the Fourier series

$$s(t) = \sum_n c_n e^{jn\omega_c t}$$

- In theory we should be able to just keep the coefficients and then use them to reconstruct the speech

IT IS NOT THAT SIMPLE...

- Speech is modeled by a process the parameters of which are first extracted from the speech itself
- At the receiving end, the same parameters are used to synthesize speech

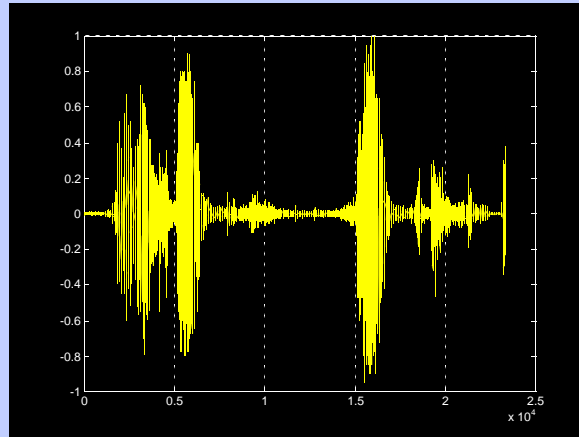
COMPRESSION

- Let's say we need to send 16 parameters, each represented by an 8 bit code, every 32 msec.
- What is the average bit rate?
 - $R = (16 \times 8) / 32 \times 10^{-3} = 4 \text{ kb/s}$
- This is a 16:1 reduction compared to 64 Kb/s PCM

SPEECH CHARACTERISTICS

- Speech is a series of “sounds” of duration ~64 msec.
- Human vocal tract produces sounds that fall under one of the following
 - voiced(dominant periodic component)
 - unvoiced(dominant noise component)
- Speech is analyzed in blocks of less than 64 msec. (20-->30 msec.)

SAMPLE SPEECH



©1997 BG Mobasseri

7

ECE 8700

MODELING SPEECH

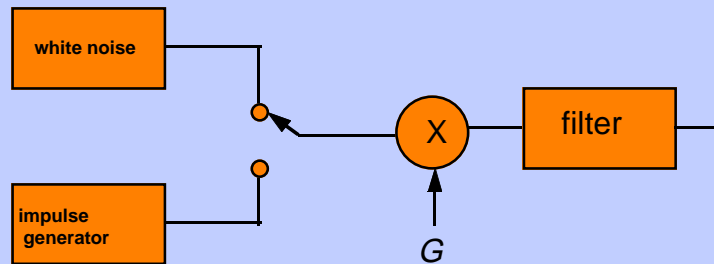
- For voiced speech, excitation can be modeled as an impulse train at frequency f_0 which is fed to a linear filter
- f_0 is called *pitch period*
- Unvoiced speech can be modeled by a white noise process

©1997 BG Mobasseri

8

ECE 8700

SYSTEM VIEW



VOCAL TRACT FILTER

- Synthesis filter can be modeled by

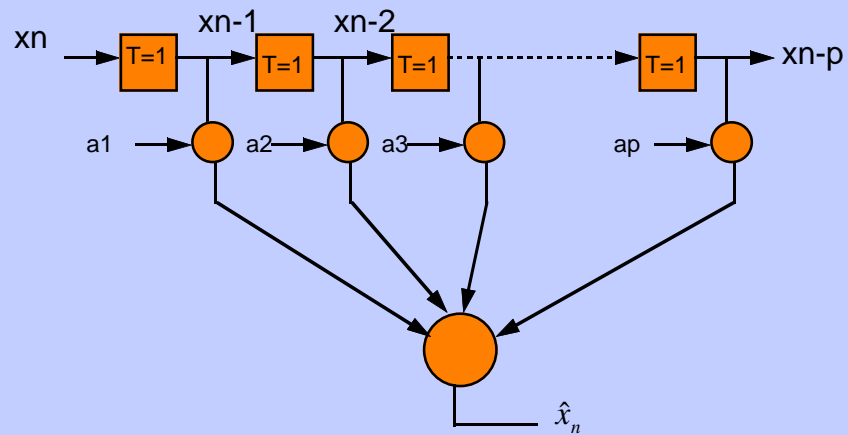
$$X_n = \sum_{i=0}^p a_n X_{n-i} + Gw_n$$

w_n : *input sequence*(white noise or impulses)

a_n : filter coeff.

p : filter order

LPC AS A TRANSVERSAL FILTER



SHORT TERM STATIONARY

- Speech signals are short time stationary. Therefore, the excitation is either white noise or periodic.
- During each 20-30 msec. time, filter coefficients remain approximately fixed.

DETAILED STEPS

- Speech is filtered to 3KHz and sampled at 8000 samples/sec.
- Samples are divided into 20 msec. intervals; about 160 samples each
- From this data, the encoder computes the parameters to be transmitted to the receiver

PREDICTION FILTER

- Filter performs a linear prediction as follows

$$\hat{x}_n = \sum_{k=1}^p a_k x_{n-k} \quad 1 < n < N$$

- Since we have all the samples, we can compute the prediction error

PREDICTION ERROR

- The difference between the actual speech sample and its predicted value:

$$\begin{aligned}e_n &= x_n - \hat{x}_n \\ &= x_n - \sum_{k=1}^p a_k x_{n-k}\end{aligned}$$

FINDING FILTER COEFF.

- Minimize the prediction error wrt $\{a_n\}$

$$E_p = \frac{1}{N} \sum_{n=1}^N e_n^2 = \frac{1}{N} \sum_{n=1}^N \left(x_n - \sum_{k=1}^p a_k x_{n-k} \right)^2$$

SOLUTION

- The vector of optimum coefficients, is the solution to the following linear eq.

$$\mathbf{a} = \hat{\mathbf{R}}^{-1} \mathbf{r}$$

$\hat{\mathbf{R}}$: $p \times p$ matrix, (i, j) element = \hat{R}_{i-j}

\mathbf{r} : vector with components \hat{R}_i

$$\hat{R}_i = \frac{1}{N} \sum_{n=-\infty}^{\infty} x_n x_{n-i}$$

WHAT IS THE MIN. ERROR?

- Plugging optimal filter coeff. into the prediction error:

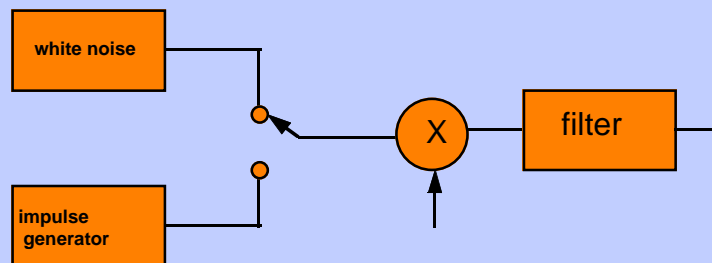
$$E_p^{\min} = G^2 \left[\frac{1}{N} \sum_{n=1}^N w_2^n \right]$$

WHAT IS TRANSMITTED?

- Voiced-unvoiced information
- Pitch period
- Prediction filter coeff.
- gain

HOW IS SPEECH RECONSTRUCTED?

- Each block of sampled speech is reconstructed based on the parameters on the previous slide



PERFORMANCE RESULTS

- Speech coding via LPC can reduce bit rate to as low as 4800bit/sec, or using vector quantization of LPC parameters, to 2400 bits/sec.

Code Excited Linear Prediction(CELP)

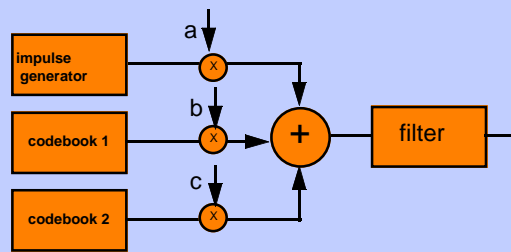
An 8 kb/s LPC algorithm

- Coder and decoder have predetermined code books of random excitation
- For each speech signal, the codebook is searched for the stochastic excitation to the linear filter producing the best sound

Vector-Sum Excited Linear Prediction VSELP

An 8 kb/s LPC algorithm

- In VSELP a mix of excitations are used



VSELP DATA

- Sample rate: 8KHz
- Frame size: 20msec (160 samples)
- Encoder parameters: 159 bits divided as
 - filter coeff.....38
 - frame energy....5
 - pitch.....28
 - codebook 1.....28
 - codebook 2.....28
 - gains.....32
- Data rate: 7950 bps

WHERE DO WE STAND TODAY?

- The most recent ITU recommendation is G.729 at 8Kb/s(1995)
- There are also cellular standards
 - GSM.....13 kb/s(1987)
 - IS-54.....7.95 kb/s(1990)
 - IS-96.....8.5 kb/s(1993)

ADDITIONAL READING ...

- *IEEE Communication Magazine*, Sept. 1997
- *IEEE Signal Processing Magazine*, Sept. 1996
- *IEEE Signal Processing Magazine*, Sept. 1997.