Lecture 13: Predictive Coding of Speech at Low Bit Rates, part 3

Mark Hasegawa-Johnson

ECE 537: Speech Processing, Fall 2022

Outline

What's the Error Spectrum?

$$q[n] = s[n] - \sum_{k=1}^{M+p+1} \alpha_k \hat{s}[n-k]$$

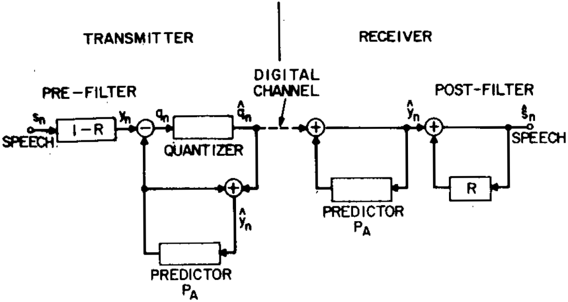
$$\hat{q}[n] = q[n] + \epsilon[n]$$

$$\begin{aligned}\hat{s}[n] &= \hat{q}[n] + \sum_{k=1}^{M+p+1} \alpha_k \hat{s}[n-k] \\ &= s[n] + \epsilon[n],\end{aligned}$$

where

- $\epsilon[n]$ is a random error, uniformly distributed between $-\frac{\Delta}{2}$ and $\frac{\Delta}{2}$, where Δ is the quantizer step size.
- If the quantizer step size is small enough, then $\epsilon[n]$ is uncorrelated with $\epsilon[n-m]$.
- In other words, $\epsilon[n]$ is white noise!

The Noise-Shaping Filter



The structure above shapes the noise by $\frac{1}{|1-R(e^{j\omega})|^2}$:

$$Y(z) = (1 - R(z))S(z)$$

$$\hat{y}[n] = y[n] + \epsilon[n]$$

$$\hat{S}(z) - S(z) = \frac{1}{1 - R(z)}\epsilon(z)$$

$$E\left[\left|\hat{S}(e^{j\omega}) - S(e^{j\omega})\right|^2\right] = \left|\frac{1}{1 - R(e^{j\omega})}\right|^2$$

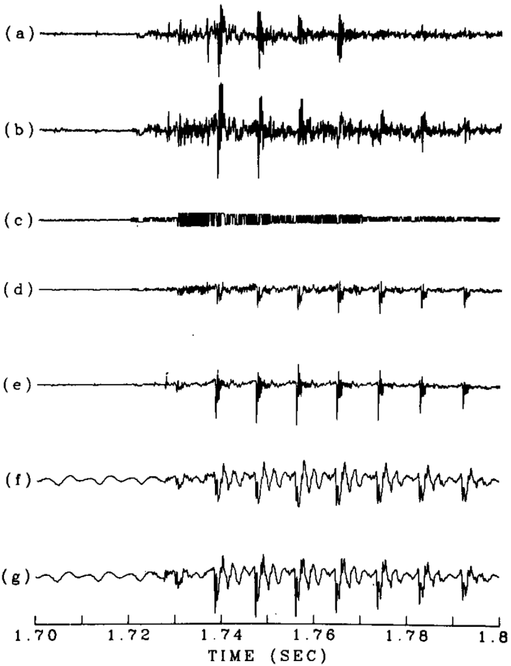
Strategies for Quantizing the Prediction Residual

- One bit per sample
- Adaptive center clipping
- Tree-based lookahead
- Multi-pulse LPC
- CELP (Code excited LPC)

Outline

Atal (1982), Figure 14:

(a) Prediction residual, $v[n]$, w/frame-wise center-clipping threshold
(b) Quantizer input, $q[n]$, w/sample-wise center-clipping threshold
(c) Quantized residual, $\hat{q}[n]$
(d) Reconstructed $\hat{d}[n]$
(e) Original $d[n]$
(f) Reconstructed $\hat{s}[n]$
(g) Original $s[n]$



Bit Rate

$F_s = 8000$Hz, so this coder uses 8000 bits/second for the residual, plus information about the predictor coefficients.

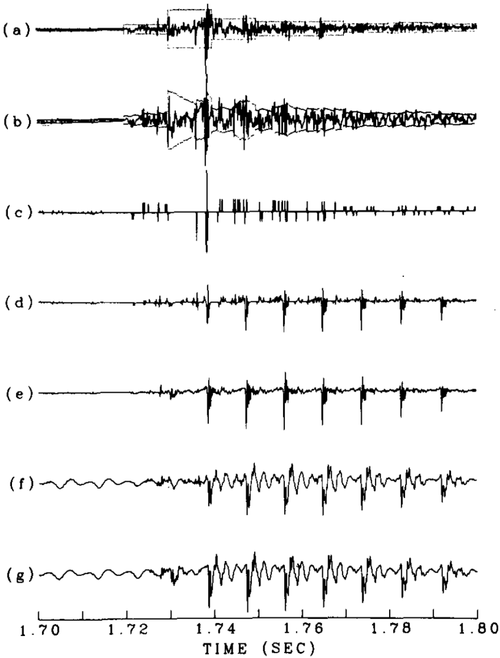
Outline

Adaptive Center Clipping

- Observation: high-amplitude samples of $q[n]$ have a much bigger perceptual impact than low-amplitude samples.
- Strategy:
  - Samples smaller than a threshold are set to zero
  - Samples larger than the threshold are quantized with ~ 8 different quantization levels
- Each 10-bit code-word specifies the number of zero-valued samples (0-127: 7 bits), and the amplitude of the next non-zero sample (3 bits)

Atal (1982), Figure 18:

(a) Prediction residual, $v[n]$, w/frame-wise center-clipping threshold
(b) Quantizer input, $q[n]$, w/sample-wise center-clipping threshold
(c) Quantized residual, $\hat{q}[n]$
(d) Reconstructed $\hat{d}[n]$
(e) Original $d[n]$
(f) Reconstructed $\hat{s}[n]$
(g) Original $s[n]$



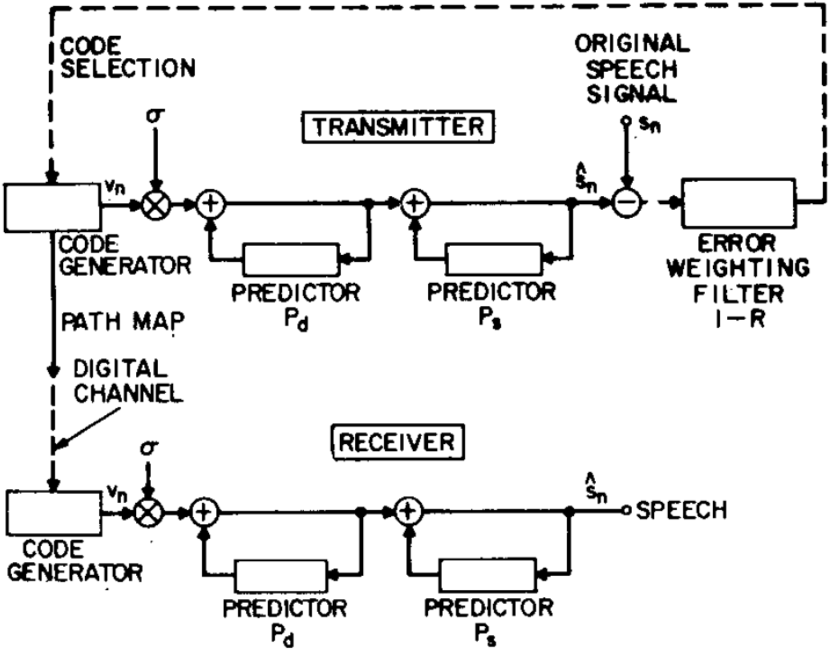
Bit Rate

Example in the article uses 5.6 kbps for the residual, to code an 8000 samples/second signal.

Outline

Why should $\hat{q}[n]$ be related to $q[n]$?

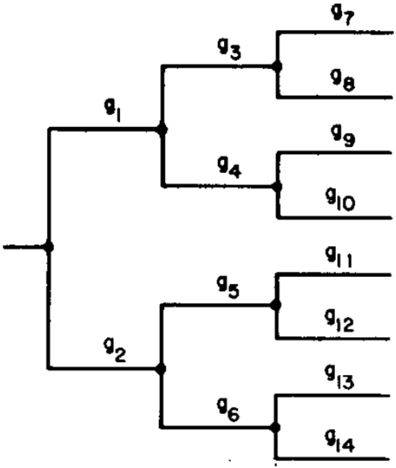
- The encoder calculates the best possible LPC excitation sequence $\hat{q}[n]$, and sends it to the decoder.
- Why should $\hat{q}[n]$ be related to the LPC analysis residual?
- Why not just find the excitation sequence that minimizes $\hat{y}[n] - y[n]$?


CODE
SELECTION
ORIGINAL
SPEECH
SIGNAL
σ
TRANSMITTER
s_n
$\hat{s}_n$
v_n
CODE
GENERATOR
PREDICTOR
P_d
PREDICTOR
P_s
ERROR
WEIGHTING
FILTER
1 − R
PATH MAP
DIGITAL
CHANNEL
RECEIVER
σ
v_n
$\hat{s}_n$
SPEECH
CODE
GENERATOR
PREDICTOR
P_d
PREDICTOR
P_s

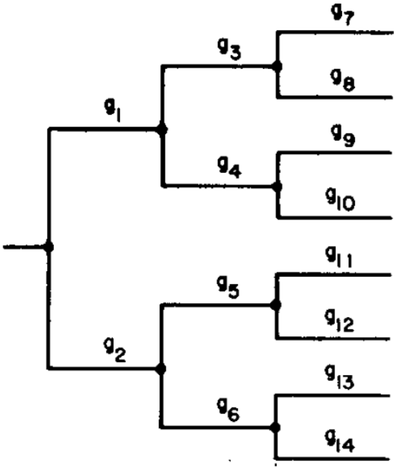
Latency

- The synthesis filter, $\left(\frac{1}{1-P_d(z)}\right)\left(\frac{1}{1-P_s(s)}\right)$, is IIR.
- $v[n]$ therefore has a strong effect on samples $\hat{s}[n+L]$ for pretty long L, at least dozens of samples.
- It's necessary to use some kind of lookahead.

- Fill a tree with pseudo-random numbers, in a sequence that is known to both encoder and decoder.
- Assume that the best M paths are known up to level $L-1$.



- From each level-$(L-1)$ path, test 2 paths to level-L, thus there are a total of $2M$ paths.
- Set $v[n], \ldots, v[n+L-1]$ equal to numbers on a path.
- $E = \sum_{m=n}^{n+L-1} (\hat{y}[n] - y[n])^2$.
- Choose the path with minimum E.
- Transmit its first bit.
- Repeat.



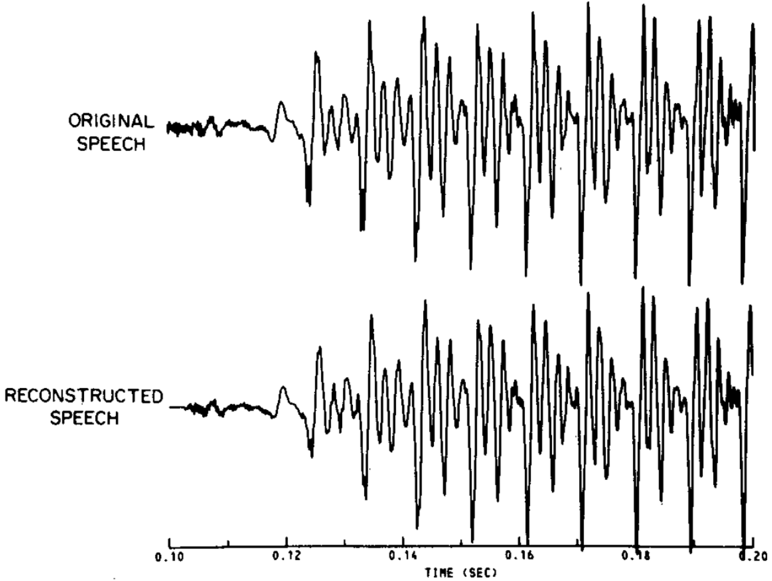


Fig. 25. Example of the waveforms of original and coded speech signals using a binary tree (1 bit/sample) with $M = 64$ and $L = 60$.

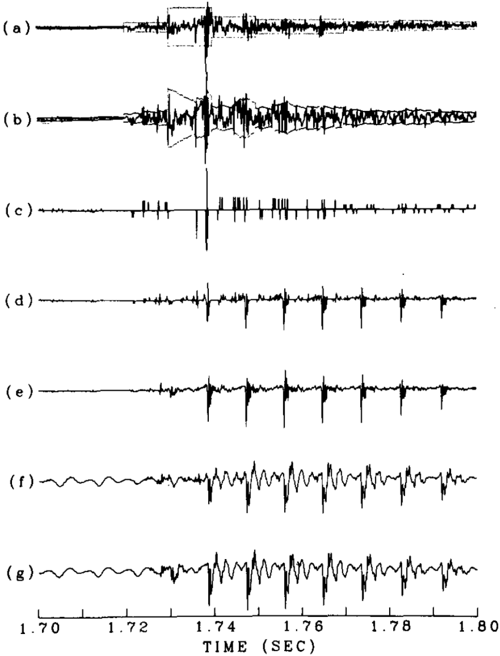
Bit Rate

One bit per sample, thus 8 kbps plus the bits required for predictor coefficients.

Outline

Atal (1982), Figure 18:

(a) Prediction residual, $v[n]$, w/frame-wise center-clipping threshold
(b) Quantizer input, $q[n]$, w/sample-wise center-clipping threshold
(c) Quantized residual, $\hat{q}[n]$
(d) Reconstructed $\hat{d}[n]$
(e) Original $d[n]$
(f) Reconstructed $\hat{s}[n]$
(g) Original $s[n]$



Multi-Pulse LPC

In every L-sample frame,

$$v[n] = \sum_{m=1}^{M} g_m \delta[n - d_m],$$

where $M \ll L$; (d_m, g_m) are the position and scale of the m^{th} pulse.

Multi-Pulse LPC

If you feed the signal $\delta[n-d]$ to the predictor filters $H(z) = \left(\frac{1}{1-P_d(z)}\right)\left(\frac{1}{1-P_s(z)}\right)$, the result is the delayed impuse response:

$$\delta[n-d] \xrightarrow{\mathcal{H}} h[n-d]$$

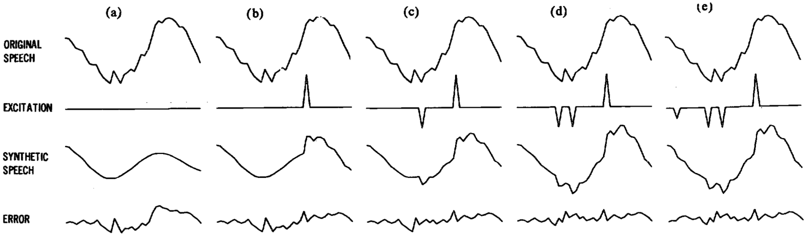
Multi-Pulse LPC

- For $1 \leq m \leq M$:
  - For $1 \leq d \leq L$:

$$\gamma_d = \frac{\sum_n y[n]h[n-d]}{\sum_n h^2[n-d]}$$
$$\epsilon_d = \sum_n \left(y[n] - \gamma_d h[n-d]\right)^2$$

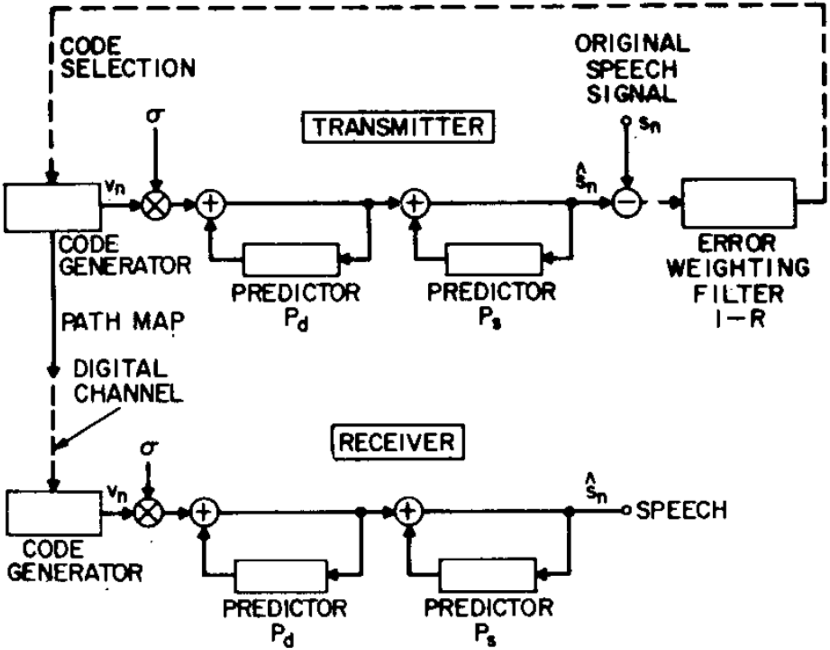
  - Set

$$d_m = \underset{d}{\operatorname{argmin}}\, \epsilon_d$$
$$g_m = \gamma_{d_m}$$



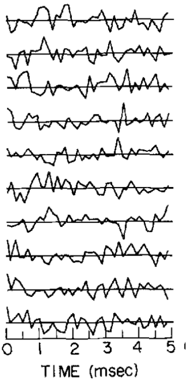
Copyright IEEE, permission granted for academic use: Atal and Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," 1982, Fig. 6

Outline


CODE
SELECTION
ORIGINAL
SPEECH
SIGNAL
σ
TRANSMITTER
s_n
v_n
$\hat{s}_n$
CODE
GENERATOR
PREDICTOR
P_d
PREDICTOR
P_s
ERROR
WEIGHTING
FILTER
1−R
PATH MAP
DIGITAL
CHANNEL
RECEIVER
σ
v_n
$\hat{s}_n$
SPEECH
CODE
GENERATOR
PREDICTOR
P_d
PREDICTOR
P_s

Code-Excited LPC

- Generate a "codebook" containing 1024 different pseudo-random 5ms sequences, $v[n]$.
- Choose the one that minimizes the error.

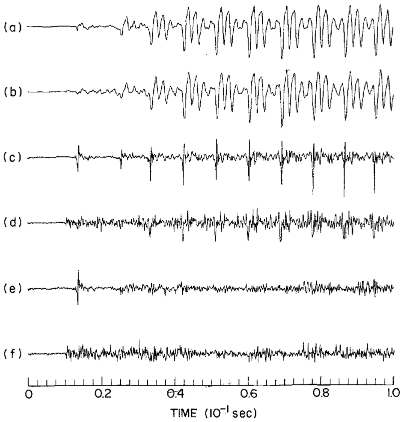


Copyright IEEE, permission granted for academic use:

Schroeder & Atal, 1985, Fig. 6(a)

Schroeder & Atal (1985), Figure 18:

(a) Original $s[n]$
(b) Synthetic $\hat{s}[n]$
(c) Original $d[n]$
(d) Synthetic $\hat{d}[n]$
(e) Original $v[n]$
(f) Synthetic $\hat{v}[n]$



Schroeder & Atal, 1985, Fig. 4

Bit Rate

10 bits per 5ms, thus 2 kbps plus predictor coefficients.

Outline

Conclusions

- Adaptive center-clipping: use bits to code the high-amplitude samples.
- Multi-pulse LPC: build up $v[n]$ one impulse at a time.
- Tree-coding and CELP: Just find the excitation that gives the best speech, who cares whether or not it's related to the true LPC residual.