

Lecture 15: Stochastic and Adaptive Codebooks

Mark Hasegawa-Johnson

ECE 537: Speech Processing

- 1 Review: LPC and MPLPC
- 2 Long-Term Predictor
- 3 Stochastic Codebook
- 4 Adaptive Codebook
- 5 Summary

Outline

- 1 Review: LPC and MPLPC
- 2 Long-Term Predictor
- 3 Stochastic Codebook
- 4 Adaptive Codebook
- 5 Summary

Review: LPC

- LPC models the resonant structure of speech using a p^{th} -order linear predictor:

$$s[n] = v[n] + \sum_{k=1}^p a_k s[n-k]$$

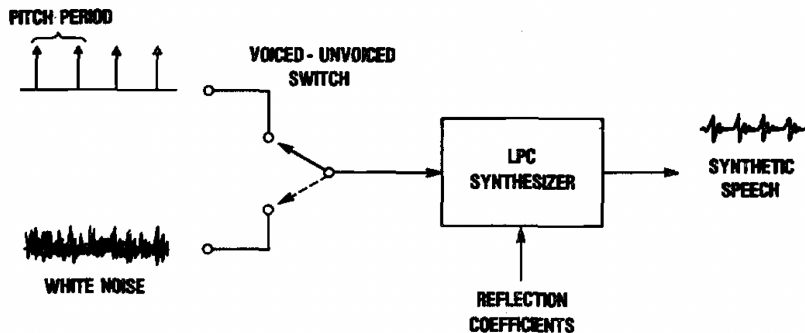
- Minimizing the mean-squared error results in p linear equations in p unknowns, which can be solved as

$$\mathbf{r} = \mathbf{R}\mathbf{a}$$

- In almost all cases, the formant frequencies (and their bandwidths) are among the poles of the LPC synthesis filter:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_k a_k z^{-k}} = \frac{Gz^p}{\prod_{i=1}^p (z - z_i)}$$

The Vocoder Model



Atal and Remde, 1982, Fig. 1

Review: Multipulse Excitation

- Multipulse excitation is $v_k[n] = \sum_{\ell=1}^k a_{\ell} \delta[n - n_{\ell}]$
- The resulting synthesized speech is $\hat{s}_k[n] = \sum_{\ell=1}^k a_{\ell} h[n - n_{\ell}]$ where $h[n]$ is the LPC impulse response
- The impulse times n_k and weights a_k are chosen to minimize MSE:

$$\mathcal{E}_k = \sum_n (s[n] - \hat{s}_k[n])^2$$

$$n_k = \operatorname{argmin} \mathcal{E}_k = \operatorname{argmax} R_{EH}[n_k]$$

$$a_k = \operatorname{argmin} \mathcal{E}_k = \frac{R_{EH}[n_k]}{R_{HH}[0]}$$

$$R_{EH}[n_k] = \sum_n (s[n] - \hat{s}_{k-1}[n]) h[n - n_k]$$

$$R_{HH}[0] = \sum_n h^2[n]$$

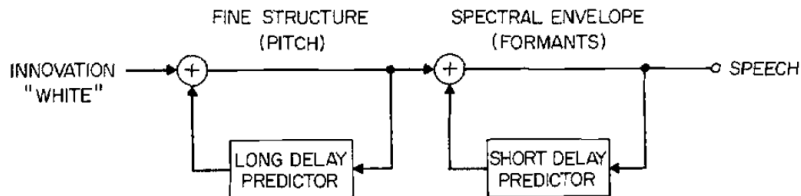
Outline

- 1 Review: LPC and MPLPC
- 2 Long-Term Predictor
- 3 Stochastic Codebook
- 4 Adaptive Codebook
- 5 Summary

Voiced vs. Unvoiced Speech

- **Unvoiced Speech:** the excitation is white noise, representing frication
- **Voiced Speech:** the excitation is periodic. Using an impulse train makes the speech intelligible, but not very natural. Can we do better?

Long-Term Predictor



- Atal (1982) proposed modeling voiced excitation by passing white noise through a long-term predictor, as shown in this figure from Schroeder and Atal (1985).
- The short-time predictor is LPC, representing formants.
- The long-term predictor represents pitch.

Integer versus Fractional Pitch Periods

If the pitch period T_0 is an integer, the long-term predictor is:

$$s[n] = v[n] + \beta s[n - T_0]$$

If the pitch period is not an integer, then the ideal long-term predictor is

$$\begin{aligned} s[n] &= v[n] + \sum_{m=-\infty}^{\infty} s[n-m] \text{sinc}(\pi(m - T_0)) \\ &\approx v[n] + \sum_{k=-1}^1 \beta_k s[n - \text{round}(T_0) + k] \end{aligned}$$

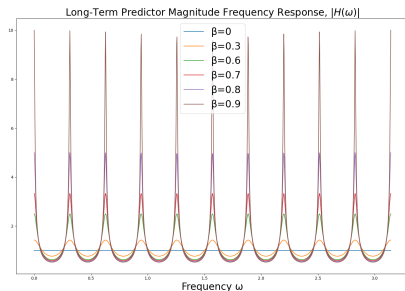
Frequency Response of the Long-Term Predictor

The filter

$s[n] = v[n] + \beta s[n - T_0]$ has the following frequency response, which resembles an impulse train as $\beta \rightarrow 1$:

$$H(\omega) = \frac{1}{1 - \beta e^{-j\omega T_0}}$$

$$|H(\omega)| = \frac{1}{\sqrt{(1 + \beta^2) - 2\beta \cos(\omega T_0)}}$$



Outline

- 1 Review: LPC and MPLPC
- 2 Long-Term Predictor
- 3 Stochastic Codebook**
- 4 Adaptive Codebook
- 5 Summary

Is the Prediction Residual Really White?

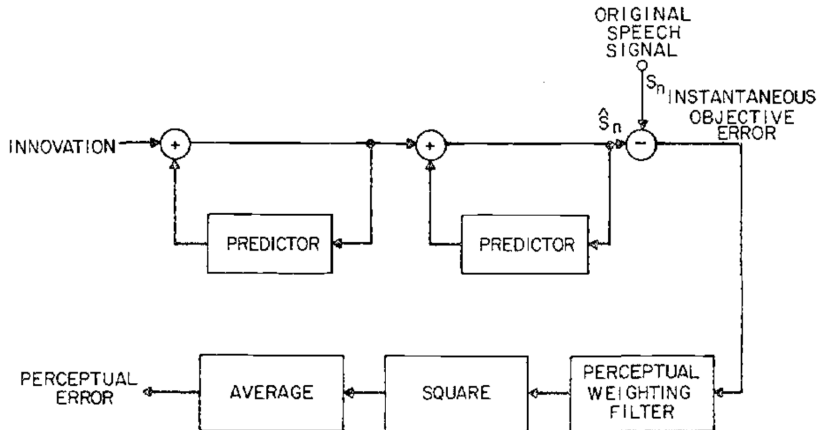
- Atal (1982) proposed that the residual, after both LTP and LPC, is approximately white. But does white noise really work?
- Not quite: It contains some additional detail that is hard to model using a linear filter.

Code-Excited LPC (CELP)

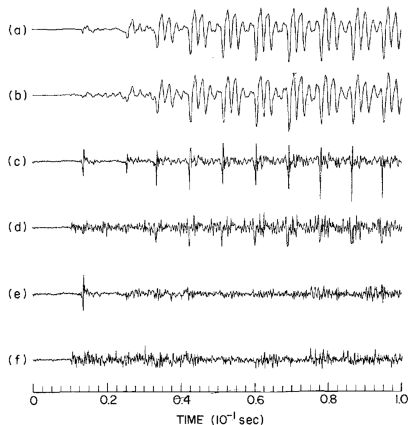
Schroeder and Atal (1995) proposed the following algorithm:

- Assume that you have a codebook containing ~ 1000 different white noise signals, $c_1[n]$ through $c_{1000}[n]$.
- For each $c_k[n]$, generate $\hat{s}_k[n]$.
- Choose the one that has the lowest error.

Code-Excited LPC (CELP)



Schroeder and Atal, 1985, Fig. 2



- a. Original speech
- b. Synthetic speech
- c. LPC residual
- d. Synthetic LPC residual
- e. LPC-LTP residual
- f. Codevector from 10-bit codebook

What should be in the stochastic codebook?

- CELP: unit-normal Gaussian random sequences
- Algebraic CELP (ACELP): sparse (mostly zero) codevectors chosen $\sim$ uniformly spaced on a 40-dimensional sphere
- Vector-Sum Excited LPC (VSELP): sums of codevectors
- Glottal-excited LPC: Parametric models of the derivative of the glottal volume velocity

Outline

- 1 Review: LPC and MPLPC
- 2 Long-Term Predictor
- 3 Stochastic Codebook
- 4 Adaptive Codebook**
- 5 Summary

Multiple Excitations

Rose and Barnwell proposed an excitation formed as:

$$v[n] = \sum_{i=1}^M \beta_i f_i[n],$$

... where each $f_i[n]$ can be:

- ... an impulse (MPLPC), or
- ... a Gaussian vector (CELP), or
- ... a delayed copy of $v[n]$

Multiple Excitations

Let $h[n]$ be the LPC impulse response. Rose and Barnwell's proposal can be written as:

$$v[n] = \beta_1 \delta[n - n_1] + \beta_2 c_k[n] + \beta_3 v[n - T_0],$$

$$\hat{s}[n] = \beta_1 h[n - n_1] + \beta_2 h[n] * c_k[n] + \beta_3 \hat{s}[n - T_0],$$

... where β_1 , n_1 , β_2 , k , β_3 , and T_0 are all chosen to minimize a perceptually weighted copy of

$$\mathcal{E} = \sum_n (s[n] - \hat{s}[n])^2$$

Sequential Optimization

- **Adaptive Codebook:** Choose T_0 and β_3 to minimize a perceptually weighted copy of

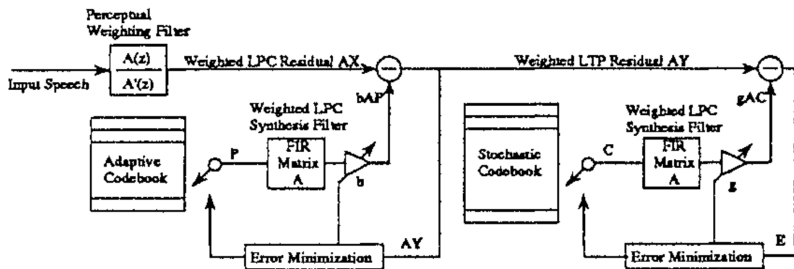
$$\mathcal{E}_{\text{Adaptive}} = \sum_n (s[n] - \beta_3 \hat{s}[n - T_0])^2$$

... where $\hat{s}[n - T_0]$ is synthetic speech from the previous frame (if T_0 is less than one frame length, use an integer multiple of T_0 instead)

- **Stochastic Codebook:** Then choose β_2 and k to minimize a perceptually weighted copy of

$$\mathcal{E}_{\text{Stochastic}} = \sum_n (s[n] - \beta_2 h[n] * c_k[n] - \beta_3 \hat{s}[n - T_0])^2$$

Sequential Optimization



Taniguchi et al., 1990, Fig. 1(a)

Outline

- 1 Review: LPC and MPLPC
- 2 Long-Term Predictor
- 3 Stochastic Codebook
- 4 Adaptive Codebook
- 5 Summary

Summary

- Long-Term Predictor: $s[n] = v[n] + \beta s[n - T_0]$
- Code Excited LPC: Choose $v[n]$ from a set of pseudo-random codevectors in order to minimize sum-squared error.
- Adaptive Codebook: Choose β and T_0 from an adaptive codebook made up of delayed copies of the previous frame's excitations