Speech Coding: Fundamentals and Applications

Mark Hasegawa-Johnson

University of Illinois

October 12, 2018

Outline

1. Waveform Coding: PCM, DPCM
2. Sub-Band Coding: Audio Coding, e.g., MP3
3. LPC-Based Analysis by Synthesis: MPLPC, CELP, ACELP
4. LPC Vocoders: LPC-10e, MELP, MBE, PWI
5. Conclusions

Pulse code modulation (PCM)

Pulse code modulation (PCM) is the name given to memoryless coding algorithms which quantize each sample of $s(n)$ using the same reconstruction levels $\hat{s}_k$, $k = 0, \ldots, m, \ldots, K$, regardless of the values of previous samples. The reconstructed signal $\hat{s}(n)$ is given by

$$\hat{s}(n) = \hat{s}_m \quad \text{s.t.} \quad (s(n) - \hat{s}_m)^2 = \min_{k=0,\ldots,K} (s(n) - \hat{s}_k)^2$$

Uniform PCM

Uniform PCM is the name given to quantization algorithms in which the reconstruction levels are uniformly distributed between S_{max} and S_{min}. Suppose that a signal is quantized using B bits per sample. If zero is a reconstruction level, then the quantization step size Δ is

$$\Delta = \frac{S_{max} - S_{min}}{2^B - 1}$$

Assuming that quantization errors are uniformly distributed between $\Delta/2$ and $-\Delta/2$, the quantization error power is

$$10 \log_{10} E[e^2(n)] = 10 \log_{10} \frac{\Delta^2}{12}$$

$$\approx \text{Constant} + 20 \log_{10}(S_{max} - S_{min}) - 6B$$

Companded PCM

Companded PCM is the name given to coders in which the reconstruction levels $\hat{s}_k$ are not uniformly distributed. Such coders may be modeled using a compressive nonlinearity, followed by uniform PCM, followed by an expansive nonlinearity:

$$s(n) \rightarrow \boxed{\text{Compress}} \rightarrow t(n) \rightarrow \boxed{\text{Uniform PCM}} \rightarrow \hat{t}(n)$$

$$\hat{t}(n) \rightarrow \boxed{\text{Expand}} \rightarrow \hat{s}(n)$$

A typical example is the μ-law companding function (ITU standard G.711), which is given by

$$t(n) = S_{max} \frac{\log\left(1 + \mu |s(n)/S_{max}|\right)}{\log(1+\mu)} \text{sign}(s(n))$$

where μ is typically between 0 and 256 and determines the amount of non-linear compression applied.

The Mu-Law Compander

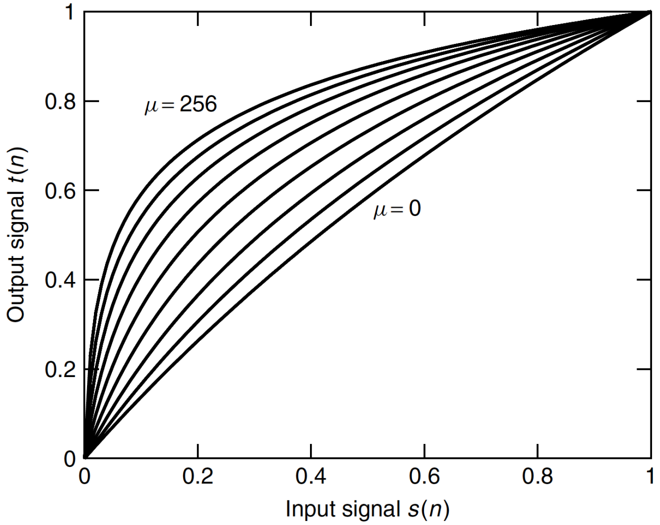


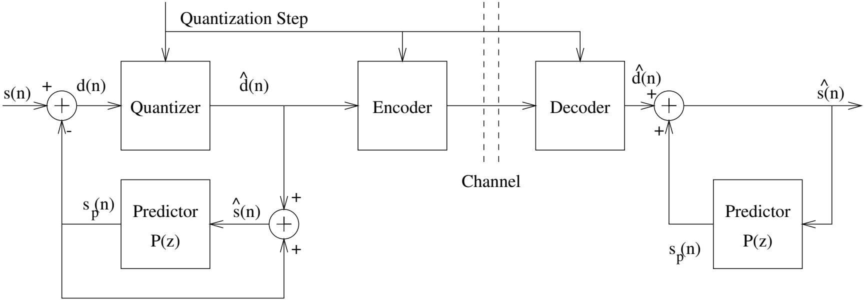
Figure 1. μ-law companding function, $\mu = 0, 1, 2, 4, 8, \ldots, 256$.

Differential PCM (DPCM)

In differential PCM, each sample $s(n)$ is compared to a prediction $s_p(n)$, and the difference is called the prediction residual $d(n)$. $d(n)$ has a smaller dynamic range than $s(n)$, so for a given error power, fewer bits are required to quantize $d(n)$. Common sub-types of DPCM include:

- Sigma-Delta coder: $s[n]$ is upsampled by a factor of 8 or 16, then $d[n]$ is quantized using only one bit per sample. Often used **inside an A/D**.
- Adaptive differential PCM (ADPCM): G.726 ADPCM is frequently used at 32 kbps in land-line telephony. The predictor in G.726 consists of an adaptive second-order IIR predictor in series with an adaptive sixth-order FIR predictor.

Differential PCM (DPCM)



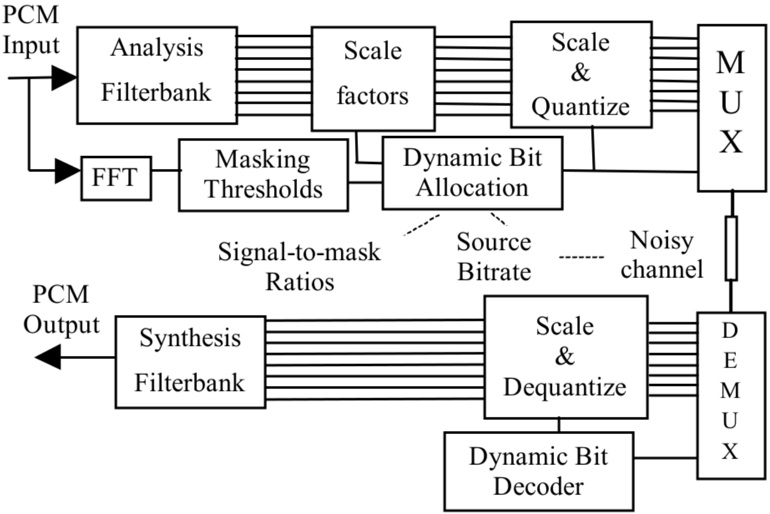
Outline

1. Waveform Coding: PCM, DPCM
2. Sub-Band Coding: Audio Coding, e.g., MP3
3. LPC-Based Analysis by Synthesis: MPLPC, CELP, ACELP
4. LPC Vocoders: LPC-10e, MELP, MBE, PWI
5. Conclusions

Sub-Band Coding

In subband coding, an analysis filterbank is first used to filter the signal into a number of frequency bands and then bits are allocated to each band by a certain criterion. Because of the difficulty in obtaining high-quality speech at low-bit rates using subband coding schemes, these techniques have been mostly used for wideband medium-to-high bit rate speech coders and for audio coding (e.g., MP3).

Structure of a perceptual subband speech coder



Credit: from Tang et al., 1997, by permission.

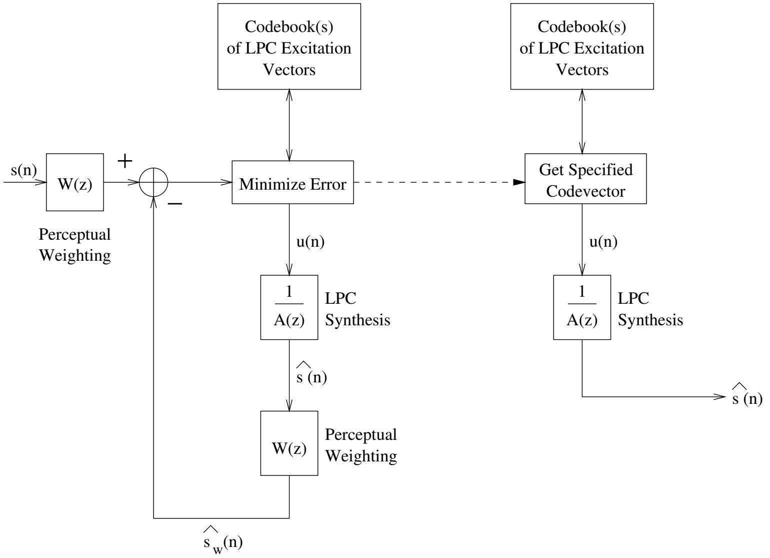
Outline

Analysis-by-Synthesis

An analysis-by-synthesis coder consists of the following components:

- A model of speech production which depends on certain parameters θ:
$$\hat{s}(n) = f(\theta)$$
- A list of K possible parameter sets for the model,
$$\theta_1, \ldots, \theta_k, \ldots, \theta_K$$
- An error metric $|E_k|^2$ which compares the original speech signal $s(n)$ and the coded speech signal $\hat{s}(n)$. In LPC-AS coders, $|E_k|^2$ is typically a perceptually-weighted mean-squared error measure.

LPC-Based Analysis by Synthesis



(a) LPC-AS Coder (b) LPC-AS Decoder

The Basic LPC Model

In LPC-based coders, the speech signal $S(z)$ is viewed as the output of a linear time invariant (LTI) system whose input is the excitation signal $U(z)$, and whose transfer function is represented by the following:

$$S(z) = \frac{U(z)}{A(z)} = \frac{U(z)}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

Line Spectral Frequencies (LSFs)

The LPC coefficients are quantized by converting them to *line-spectral frequencies* (LSFs). The LSFs are the roots of the polynomials $P(z)$ and $Q(z)$:

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1})$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1})$$

The operations above make the signal $p[n] \leftrightarrow P(z)$ symmetric in the time domain, while $q[n] \leftrightarrow Q(z)$ is anti-symmetric. For this reason, their zeros are **pure imaginary** numbers:

$$P(z) = \prod_{n=1}^{(p+1)/2} (1 - e^{jp_n}z^{-1})(1 - e^{-jp_n}z^{-1})$$

$$Q(z) = (1 - z^{-2}) \prod_{n=1}^{(p-1)/2} (1 - e^{jq_n}z^{-1})(1 - e^{-jq_n}z^{-1})$$

Line Spectral Frequencies (LSFs)

The LSFs have some interesting characteristics, very useful for quantization:

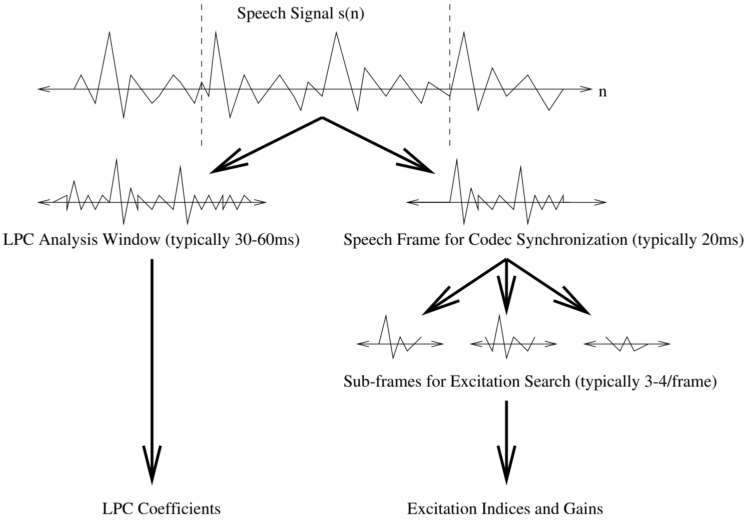
- the frequencies $\{p_n\}$ and $\{q_n\}$ are related to the formant frequencies
- the dynamic range of $\{p_n\}$ and $\{q_n\}$ is limited and the two alternate around the unit circle $(0 \leq p_1 \leq q_1 \leq p_2 \ldots \leq \pi)$
- $\{p_n\}$ and $\{q_n\}$ change slowly from one frame to another, hence, inter-frame prediction is possible.

The Frame Structure of LPC-AS

All real-world LPC-AS systems work at multiple time scales:

- The LPC filter coefficients, $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$, are chosen, converted to LSFs, quantized, and transmitted once per **frame**. A frame is typically about 30ms.
- The excitation signal $u[n]$ is chosen, quantized, and transmitted once per **subframe**. A subframe is typically about 7.5ms, thus there are typically about 4 subframes/frame.

The Frame Structure of LPC-AS

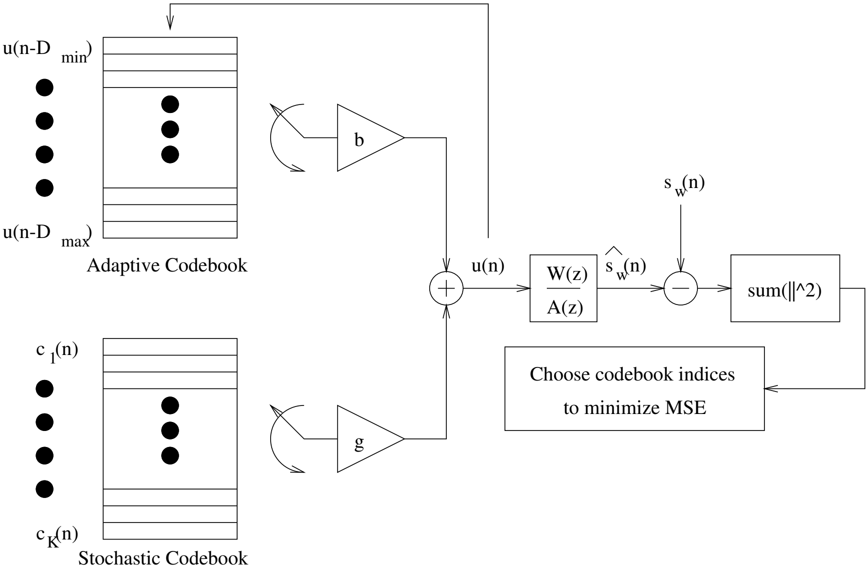


The Excitation of LPC-AS

The excitation signal is composed of two parts, as $u[n] = gc[n] + bu[n - D]$, where

- **Adaptive Codebook = Pitch Excitation**. The pitch excitation is $bu[n - D]$, i.e, it's just delayed samples of previously transmitted excitation. Here D is the pitch period, and b is the pitch prediction coefficient. D is chosen from a pre-determined range of possible pitch periods, $D_{min} \leq D \leq D_{max}$.
- **Stochastic Codebook = Noise Excitation**. Everything else (everything not explained by the adaptive codebook) has to be explained by choosing some kind of random noise signal, $c[n]$, and then scaling it by some gain term g.

The Excitation of LPC-AS



Pitch Prediction Filtering

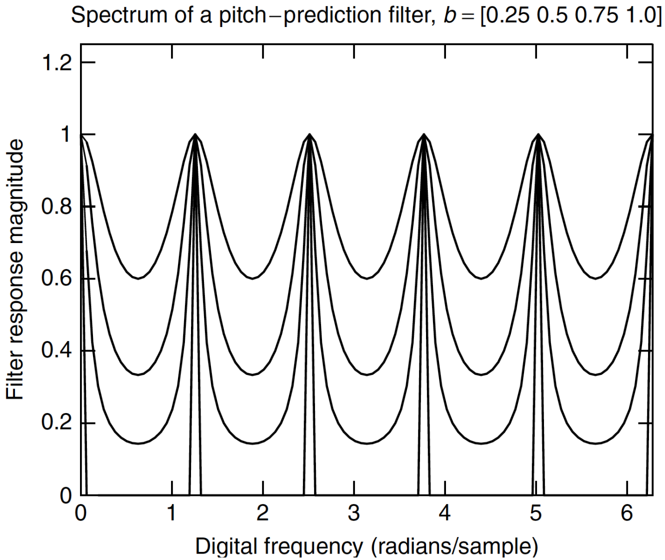
In an LPC-AS coder, the LPC excitation is allowed to vary smoothly between fully voiced conditions (as in a vowel) and fully unvoiced conditions (as in /s/). Intermediate levels of voicing are often useful to model partially voiced phonemes such as /z/. The partially voiced excitation in an LPC-AS coder is constructed by passing an uncorrelated noise signal $c(n)$ through a pitch prediction filter. A typical pitch prediction filter is

$$u[n] = gc[n] + bu[n-D]$$

where D is the pitch period. If $c[n]$ is unit-variance white noise, then according to Equation 23 the spectrum of $u[n]$ is

$$|U(e^{j\omega})|^2 = \frac{g^2}{1 + b^2 - 2b\cos\omega D}$$

Pitch Prediction Filtering



Stochastic Codebook

The noise part of the excitation, $c[n]$, is chosen from a codebook of noise-like signals.

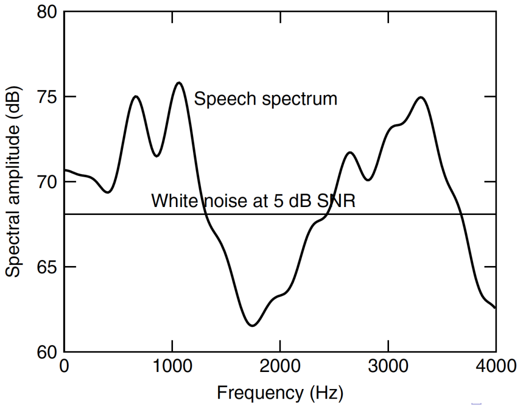
$$c[n] = c_k[n], \quad k = \arg\min \sum_{n=0}^{N-1} (s[n] - \hat{s}[n])^2$$

- CELP (code-excited LPC): $c_k[n] =$ samples generated in advance using a Gaussian random noise generator.
- MPLPC (multi-pulse LPC): $c_k[n] = g_1\delta[n - d_1] + g_2\delta[n - d_2]$, where the pulse gains, g_k, and delays, d_2, are chosen one at a time.
- ACELP (algebraic CELP): $c_k[n] =$ samples from an algebraic codeword, e.g., the vertex of an N-dimensional polyhedron.

Minimum-Error Codeword Selection

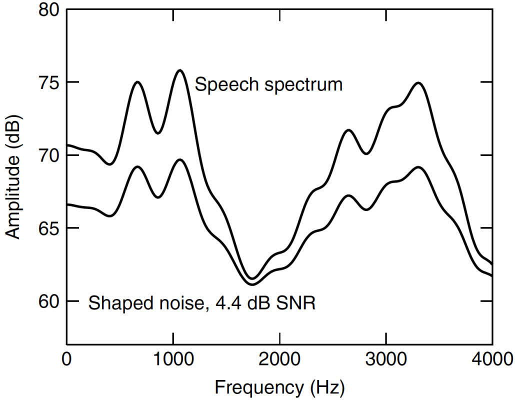
Early LPC-AS coders minimized the mean-squared error, which usually results in a noise signal that is nearly white.

$$\sum_n e^2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega$$



Perceptual Error Weighting

Not all noises are equally audible. Noise components near peaks of the speech spectrum are hidden by a "masking spectrum" $M(e^{j\omega})$, which is like a smoothed copy of the speech spectrum. Here's an example of optimally masked noise:



Noise-to-Masker Ratio

The audibility of noise may be estimated using a noise-to-masker ratio $|E_w|^2$:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|E(e^{j\omega})|^2}{|M(e^{j\omega})|^2}d\omega$$

The masking spectrum $M(e^{j\omega})$ has peaks and valleys at the same frequencies as the speech spectrum, but the difference in amplitude between peaks and valleys is somewhat smaller than that of the speech spectrum. One of the simplest model masking spectra which has the properties just described is based on the LPC spectrum, $1/A(z)$, as:

$$M(z) = \frac{|A(z/\gamma_2)|}{|A(z/\gamma_1)|}, \qquad 0 < \gamma_2 < \gamma_1 \leq 1$$

Perceptible Error

The noise-to-masker ratio may be efficiently computed by filtering the speech signal using a perceptual weighting filter $W(z) = 1/M(z)$. The perceptually weighted input speech signal is

$$S_w(z) = W(z)S(z)$$

Likewise, for any particular candidate excitation signal, the perceptually weighted output speech signal is

$$\hat{S}_w(z) = W(z)\hat{S}(z)$$

Given $s_w(n)$ and $\hat{s}_w(n)$, the noise-to-masker ratio may be computed as follows:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(e^{j\omega}) - \hat{S}_w(e^{j\omega})|^2 d\omega = \sum_n (s_w^2(n) - \hat{s}_w^2(n))$$

Zero-Input-Response and Zero-State-Response

Let the filter $H(z)$ be defined as the composition of the LPC synthesis filter and the perceptual weighting filter, thus $H(z) = W(z)/A(z)$. The computational complexity of the excitation parameter search may be greatly simplified if $\hat{S}_w$ is decomposed into the zero-input response (ZIR) and zero-state-response (ZSR) of $H(z)$, thus

$$\hat{S}_w = \hat{S}_{ZIR} + \hat{S}_{ZSR} = \hat{S}_{ZIR} + UH$$

where $\hat{S}_{ZIR}$ contains samples of the zero input response of $H(z)$, and the vector UH contains the zero state response of filter matrix H in response to codevector U. Given a candidate excitation vector U, the perceptually weighted error vector E may be defined as

$$E_w = S_w - \hat{S}_w = S_w - \hat{S}_{ZIR} - UH$$

Optimum Gain and Optimum Excitation

Recall that the excitation vector U is modeled as the weighted sum of a number of codevectors $C = [C_1, \ldots, C_M]$, $m = 1, \ldots, M$, scaled by gains $G = [g_1, \ldots, g_M]$, thus $U = GC$. The perceptually weighted error is therefore:

$$|E|^2 = |\tilde{S} - GCH|^2 = \tilde{S}\tilde{S}^T - 2GCH\tilde{S}^T + GCH(GCH)^T$$

Suppose we define the following additional bits of notation:

$$R = CH\tilde{S}^T, \quad \Sigma = CH(CH)^T$$

Then, for any given set of shape vectors X, G is chosen so that $|E|^2$ is minimized, which yields

$$G = R^T \Sigma^{-1}$$

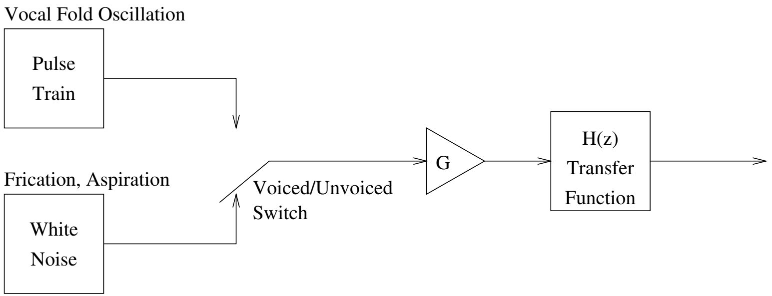
Outline

1. Waveform Coding: PCM, DPCM
2. Sub-Band Coding: Audio Coding, e.g., MP3
3. LPC-Based Analysis by Synthesis: MPLPC, CELP, ACELP
4. LPC Vocoders: LPC-10e, MELP, MBE, PWI
5. Conclusions

The LPC-10e Vocoder

The 2.4 kbps LPC-10e vocoder is one of the earliest and one of the longest-lasting standards for low-bit-rate digital speech coding. This standard was originally proposed in the 1970s, and was not officially replaced until the selection of the MELP 2.4 kbps coding standard in 1996. Speech coded using LPC-10e sounds metallic and synthetic, but it is intelligible. "Robot voice" effects are often produced using LPC-10e.

The LPC-10e Vocoder



Excitation of LPC-10e

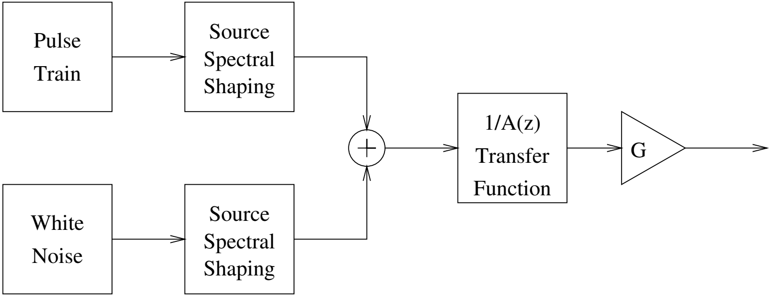
The residual signal $d(n)$ is modeled using either a periodic train of impulses (if the speech frame is voiced) or an uncorrelated Gaussian random noise signal (if the frame is unvoiced). The voiced/unvoiced decision is based on the average magnitude difference function (AMDF),

$$\Phi_d(m) = \frac{1}{N-|m|} \sum_{n=|m|}^{N-1} |d(n) - d(n-|m|)|$$

The frame is labeled as voiced if there is a trough in $\Phi_d(m)$ which is large enough to be caused by voiced excitation.

Mixed Excitation Linear Prediction (MELP)

MELP uses the same excitation signals $p[n]$ (pulse train) and $e[n]$ (white noise) as LPC-10e, but both are added together as $u[n] = ge[n] + bp[n]$.



Other Vocoders

- Multi-Band Excitation (MBE): V/UV decision is made separately in different frequency bands. Example: Inmarsat-M coding standard at 6.4kbps.
- Prototype Waveform Interpolative (PWI) Coding: The technique is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20-30 ms.

Outline

1. Waveform Coding: PCM, DPCM
2. Sub-Band Coding: Audio Coding, e.g., MP3
3. LPC-Based Analysis by Synthesis: MPLPC, CELP, ACELP
4. LPC Vocoders: LPC-10e, MELP, MBE, PWI
5. Conclusions

- Waveform coders quantize each sample. Very low complexity, in fact, they can be implemented using a single diode. Therefore often used inside an A/D circuit.
- Sub-band coders use attributes of human hearing, but don't use attributes of speech production, hence they are usually used for general audio coding.
- LPC analysis-by-synthesis is used for digital speech, e.g., cellular.
- LPC vocoders are used for very-low-bit-rate speech coders, or to generate robot voice.

Internet codecs, like MP4, use a variety of codecs. The packet header specifies which type of codec is used. This is selected depending on the type of audio contained in the packet.

Speech Coder	Rates (kbps)	Complexity	Applications
Waveform	16-64	Low	Land-line
Subband	12-256	Medium	Audio
LPC-AS	4.8-16	High	Cellular
LPC vocoder	2.0-4.8	High	Military