EE6424 Digital Audio Signal Processing
Part 2
Lecture 4:
Digital Speech Coding

Outline of lecture

- Overview of speech coding
- Linear predictive coding (LPC)
- Differential PCM (DPCM)
- Code Excited Linear Prediction (CELP)

EE6424 Part 2: Lecture 4.1

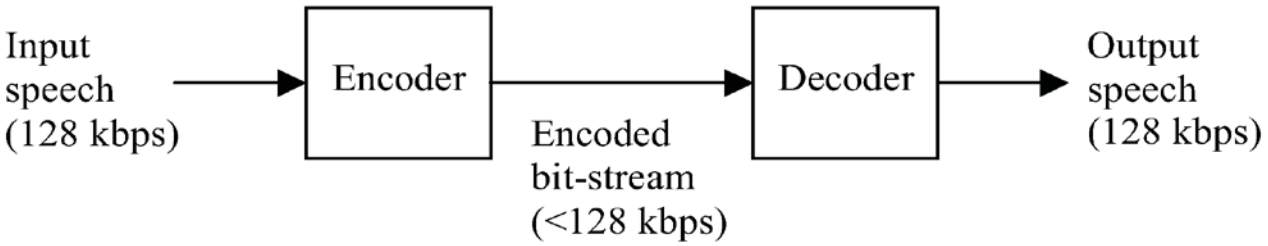
OVERVIEW OF SPEECH CODING

Introduction

- **Speech coding** is a procedure to represent a digitized speech signal using as few bits as possible, while maintaining a reasonable quality.
- The linear or logarithmic quantization can be considered as a form of **coding** as it results in a representation of the signal with a limited number of bits.
- A not so common name having the same meaning is speech **compression**.
- Compression means that a **restricted number of bits** (a design specification) are used to represent a given quantity of **information**, which would require a higher bit rate if no compression was taking place.
- Before compression, the signal contains redundancy and compression is achieved by **removing** this **redundancy** and representing the speech signal with **reduced precision**.

Structure of a speech coder

- Speech coders are designed to reduce the **reference bit-rate** of **128 kbps**, assuming a sampling frequency of 8 kHz and resolution of 16 bit.
- The input speech enters the **encoder** to produce the **compressed** speech with a **lower bit rate** than that of the input speech.
- The **decoder** produces the output speech, being an **approximation** to and **having the same rate** as the input speech.
- The encoder/decoder pair is called a **speech coder**.



Classification by bit rate

- Speech coders could be classified depending on their **bit rate**.
- The minimum bit-rate that speech coders can be achieved is limited by the information content of the speech signal. It is commonly believed that the minimum is around **60 bps**.
- Current coders could work at **2 kbps and above** with good quality. There is plenty room for future improvement.

Category	Bit-Rate Range
High bit-rate	>15 kbps
Medium bit-rate	5 to 15 kbps
Low bit-rate	2 to 5 kbps
Very low bit-rate	<2 kbps

Key principles of speech coding

- First coding principle:

  Any redundancy contained in the signal should not be transmitted (or stored). All bits are to carry as much information as possible given limited binary resources.

- Second coding principle:

  Only what we hear should be transmitted.

- Redundancy appears in the following forms:
  - **Statistical** redundancy due to amplitude distribution of the signal
  - **Temporal** redundancy due to correlation (predictability) of the signal
  - **Perceptual** redundancy (inaudible components)

- Perceptual redundancy referred to the portion of information in the input speech when removed leads to either **unperceivable** or **un-annoying** distortion (e.g., based on the frequency masking property of hearing).
- The process is **lossy** and **irreversible** as opposed to the **lossless** process when statistical or temporal redundancy is removed.
- Speech coders use a combination of **lossless** and **lossy coding** scheme, where **bit-rate is reduced** via the following two main components
  - Processing step (filtering, transformation, etc.) to **remove** inherent **redundancy** (statistical, temporal, and perceptual) in the signal.
  - Bit allocation – a procedure to **distribute** the binary resources over the speech parameters with allowable **precision** so as to minimize the distortions thus incurred. Reducing the precision reduces the bit rate.

Temporal redundancy

- A **redundant** signal is **predictable**.
- The predictability of a signal manifests itself as the presence of **correlations** between samples, i.e., correlation means redundancy.

  Example: a sine wave contains a lot of redundancy. The frequency, phase and amplitude are sufficient to describe such a signal.
- Suppressing the **temporal redundancy** consists of decreasing the signal **correlation**.
- In the frequency domain, signal correlation corresponds to a **non-flat spectrum**.
- A white noise has **flat spectrum**. It contains no redundancy which makes it useless trying to predict the value of a sample from its past and therefore is impossible to compress.

- **Spectral flatness** can be used to measure **temporal redundancy** quantitatively.
- Let $\Gamma\left(e^{j\omega}\right)$ denote the **power spectral density** of a signal $x(n)$. The spectral flatness is defined as

$$\xi = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \Gamma\left(e^{j\omega}\right) d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma(e^{j\omega}) d\omega}$$

- The power spectral density of a signal could be obtained by taking the **discrete-time Fourier transform** of the auto-correlation of the signal $x(n)$, assuming stationary.

- The **spectral flatness** has the following properties:
  - $\xi = 1$ corresponds to completely flat spectrum (zero redundancy as for white noise)
  - $\xi = 0$ if some part of the spectrum is equal to zero
  - The closer ξ is to 1, the flatter the spectrum
  - The closer ξ is to 0, the more contrastive the spectrum
- The **redundancy** (in number of bits per sample) caused by the temporal correlation is expressed as

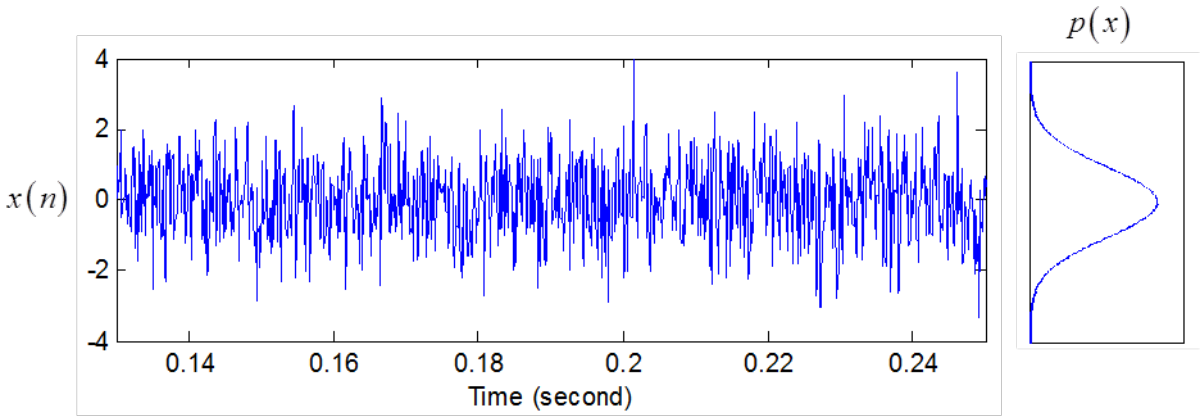
$$R = -\frac{1}{2}\log_2 \xi$$

- Temporal redundancy can be removed by **flattening the spectrum**, which can be obtained using a whitening filter.

- The main feature of speech spectrum is the presence of **formants** leading to **contrastive spectrum**. Therefore temporal redundancy is high and compression is possible.
- Temporal correlation is usually the **predominant** source of redundancy in speech coding.

Statistical redundancy

- Correlation or predictability of signal is a temporal phenomenon. A **white Gaussian noise** is not predictable based on past samples.



- The term **white** refers to its **temporal correlation** which is absent here due to its flat/white spectrum.
- The term **Gaussian** refers to the **amplitude distribution** of the samples which are normally distributed

- Statistical redundancy exists due to non-uniform amplitude distribution. There are more samples with smaller amplitude compared to those with higher amplitude.
- **Non-uniform** amplitude distribution leads to **statistical redundancy**

  A signal with a non-uniform amplitude distribution contains mores redundancy than that of a signal with a uniform distribution.

- In **logarithmic quantization**, step size is smaller (and therefore more steps) for value around zeros.

Perceptual redundancy

- **Perceptual** redundancy is due to psychoacoustic phenomenon of **auditory masking**, predominantly frequency masking.
- **Frequency masking** (simultaneous masking)

  A particular sound with a given level, which is perfectly audible on its own, can become inaudible when associated with another louder sound with a neighboring frequency. The masker and maskee appear at the same time but close to each other in terms of their frequencies.
- **Temporal masking** (non-simultaneous masking)

  A louder sound can mask a weaker sound that occurs after (post-masking) and before (pre-masking) it within a small interval.
- See EE6424 Part 01 for more details.

EE6424 Part 2: Lecture 4.2

LINEAR PREDICTIVE CODING

LP model of speech production

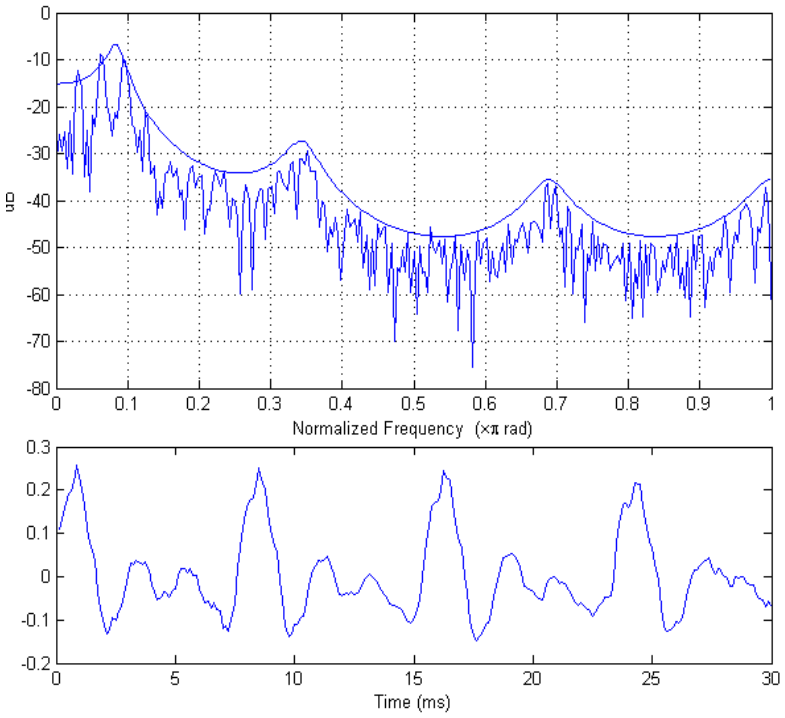
- The **Linear Predictive** (LP) model is based on the **source-filter** model with the assumption that the sound source and the vocal tract are fully uncoupled.
- Speech $s(n)$ is described as the output of an **all-pole filter** excited by a **white Gaussian noise** (or **periodic pulses**) $e(n)$ with flat spectral envelope. Here, p is the order of the filter and $a_0 = 1$ by default.

$$s(n) = e(n) - \sum_{i=1}^{p} a_i s(n-i)$$

- Let $S(z)$ and $E(z)$ be the z-transform of the speech and excitation signals, the LP model can be equivalently expressed as

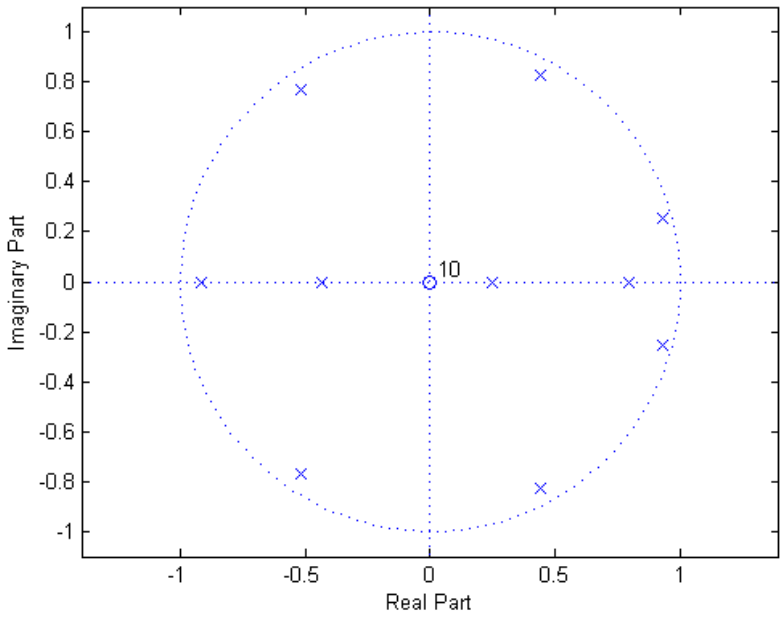
$$S(z) = E(z)\frac{1}{A(z)} = E(z)\left(\frac{1}{\sum_{i=0}^{p} a_i z^{-i}}\right)$$

- The figure below shows the spectrum of a 30 ms **speech frame** superimposed with the spectral envelop given by the frequency response of an all-pole filter $1/A(z)$.



- At the bottom panel is the speech frame of 30 ms of a voiced sound.
- Take note on the period and amplitude, which are slightly changing from one cycle to the next (jitter and shimmer).

- Figure below shows the poles and zeros of the all-pole filter, where the order $p = 10$.



- Take note on the locations of the **poles** (0 to π, or half circle) corresponding to the **formant locations** on the envelope of the power spectrum (normalized frequency from 0 to 1).

LP estimation algorithm

- For a given speech signal $s(n)$, imposing the coefficients $\{a_i\}$ in the linear predictive (LP) model results in the **prediction residual**

$$e(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i)$$

$$E(z) = S(z)A(z)$$

- The principle of LP estimation is to choose the set $\{a_1, a_2, \dots, a_p\}$ which **minimizes** the variance of the prediction residual

$$\{a_i\}^{opt} = \underset{a_1,\dots,a_p}{\operatorname{argmin}} E\{e^2(n)\}$$

- The LP coefficients could be obtained by solving the so-called **Yule-Walker** equation:

$$\left[\begin{array}{c|cccc} \gamma_{ss}(0) & \gamma_{ss}(1) & \gamma_{ss}(2) & \cdots & \gamma_{ss}(p) \\ \hline \gamma_{ss}(1) & \gamma_{ss}(0) & \gamma_{ss}(1) & \cdots & \gamma_{ss}(p-1) \\ \gamma_{ss}(2) & \gamma_{ss}(1) & \gamma_{ss}(0) & \cdots & \gamma_{ss}(p-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{ss}(p) & \gamma_{ss}(p-1) & \gamma_{ss}(p-2) & \cdots & \gamma_{ss}(0) \end{array}\right]\begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma_e^2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $\gamma_{ss}(l) = \sum_{n=0}^{N-1-l} s(n)s(n+l)$

- From the first row:

$$\sigma_e^2 = \sum_{l=1}^{p} \gamma_{ss}(l)a_l + \gamma_{ss}(0) = \gamma_{ss}(0)\left[\sum_{l=1}^{p} \frac{\gamma_{ss}(l)}{\gamma_{ss}(0)} a_l + 1\right] = \gamma_{ss}(0)\left[\sum_{l=1}^{p} r(l)a_l + 1\right]$$

Hence, $$\frac{\gamma_{ss}(0)}{\sigma_e^2} = \frac{\sigma_s^2}{\sigma_e^2} = \left[\sum_{l=1}^{p} r(l)a_l + 1\right]^{-1} = G$$

From the remaining rows:

$$\begin{bmatrix} \gamma_{ss}(1) & \gamma_{ss}(0) & \gamma_{ss}(1) & \cdots & \gamma_{ss}(p-1) \\ \gamma_{ss}(2) & \gamma_{ss}(0) & \gamma_{ss}(1) & \cdots & \gamma_{ss}(p-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{ss}(p) & \gamma_{ss}(p-1) & \gamma_{ss}(p-2) & \cdots & \gamma_{ss}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \gamma_{ss}(0) & \gamma_{ss}(1) & \cdots & \gamma_{ss}(p-1) \\ \gamma_{ss}(1) & \gamma_{ss}(0) & \cdots & \gamma_{ss}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{ss}(p-1) & \gamma_{ss}(p-2) & \cdots & \gamma_{ss}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = -\begin{bmatrix} \gamma_{ss}(1) \\ \gamma_{ss}(2) \\ \vdots \\ \gamma_{ss}(p) \end{bmatrix}$$

- The $p \times p$ **correlation matrix** has the form of **Toeplitz** matrix, which is symmetric, and has the same values along the lines parallel to the diagonal.
- The correlation matrix is positive definite which guarantees that an inverse matrix exists.
- We could solve the equation to obtain $\{a_i\}$ by taking the inverse of the correlation matrix.
- A more efficient method is by using the **Levinson-Durbin** algorithm which takes into account the Toeplitz structure of the correlation matrix.

Linear predictability of speech

- **Inverse filtering** using a MA filter (FIR)

  Passing a speech signal $s(n)$ through a linear predictive (LP) filter with a proper set of coefficients results in the prediction residual signal

$$e(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i) \quad \rightarrow \quad E(z) = A(z)S(z)$$

- **Synthesis** process using an AR filter (IIR)

$$\tilde{s}(n) = \tilde{e}(n) - \sum_{i=1}^{p} a_i \tilde{s}(n-i) \quad \rightarrow \quad \tilde{S}(z) = \frac{1}{A(z)}\tilde{E}(z)$$

- The filter $A(z)$ is call the **inverse filter** and $1/A(z)$ is called the **synthesis filter**.

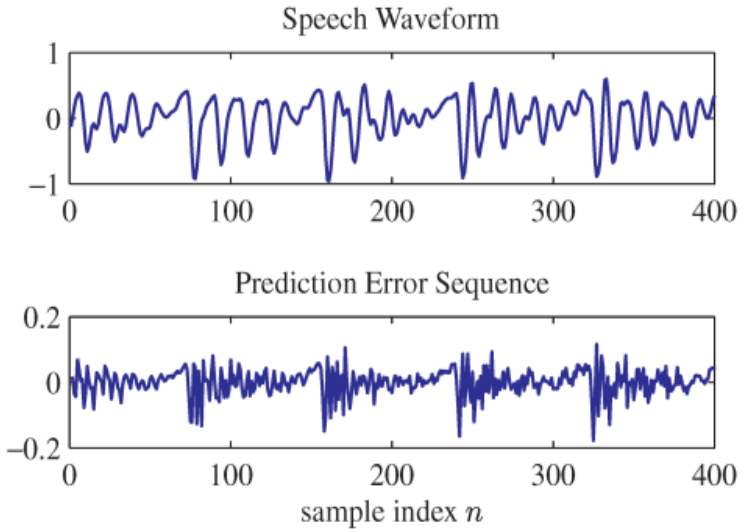
Prediction order

- The **prediction order** p is chosen such that the resulting synthesis filter $1/A\,(z)$ has enough **degrees of freedom** to replicate the spectral envelope of the input speech.
- There is approximately **one formant per kHz** of bandwidth of speech. Using **two poles for a single formant**, a signal bandwidth of F_{BW} kHz requires $2F_{\mathrm{BW}}$ of poles.
- Two more poles are usually added for modeling the glottal cycle waveform (and also empirically, because the resulting LPC speech sounds better).
- For telephone-based applications, working with a sampling frequency of 8 kHz, this leads to $p = 10$.

Linear predictive coding

- Inverse filtering **decorrelates** speech signal. The resulting prediction residual signal $e(n)$ has the following properties:
  - A **flatter** spectral envelop and therefore less redundancy
  - **Smaller** variance than that of the **original**
- The variance of the residual $e(n)$ is less than the variance of the input $s(n)$ if the linear prediction is effective.
- For efficient coding, the **prediction residual** is quantized instead of the original speech. This is termed as the **Linear Predictive Coding** (LPC).
- For a fixed number of bits, it will be possible to used a **smaller step size** for quantizing $e(n)$ and therefore to **reconstruct** $\tilde{s}(n)$ **with less error** than if $s(n)$ was quantized directly.

- The upper plot is a segment of speech signal normalized to maximum amplitude of 1.
- The lower plot is the prediction residual using a LP of order $p = 12$. The amplitude is a factor of five lower than that of the original speech signal.



Prediction gain

- The signal-to-quantization noise ratio (SQNR) is given by the ratio between the input variance $E\{s_n^2\}$ and the MSQE $E\{q_n^2\}$ on the prediction residual $e(n)$:

$$\text{SQNR} = \frac{E\{s_n^2\}}{E\{q_n^2\}} = \frac{E\{s_n^2\}}{E\{e_n^2\}} \times \frac{E\{e_n^2\}}{E\{q_n^2\}}$$

- The second factor $E\{e_n^2\}/E\{q_n^2\}$ is the SQNR value given by the **quantizer on the prediction residual** (assuming Gaussian and a loading factor of four):

$$\frac{E\{e_n^2\}}{E\{q_n^2\}} = 6.02\ B - 7.27\ dB$$

- The first factor is the **prediction gain**

$$G = \frac{E\{s_n^2\}}{E\{e_n^2\}}$$

- The SNR is increased by a factor of G

$$\text{SNR} = G \times (6.02\ B - 7.27\ dB)$$

- The **prediction gain** when the pth order linear prediction is used is represented as

$$G = \left(1 + \sum_{i=1}^{p} a_i r_i\right)^{-1}$$

- Here, $r_l = \gamma(l)/\gamma(0)$ are the normalized auto-correlation coefficients of the input ($r_0 = 1$).
- The smaller the prediction residual, the larger the prediction gain becomes, and a larger SQNR.
- The prediction gain is invariant to scaling (i.e., it does not depend on the amplitude of the input signal).

- When $p = 1$, the prediction gain becomes

$$G = (1 + a_1 r_1)^{-1} = (1 - r_1^2)^{-1}$$

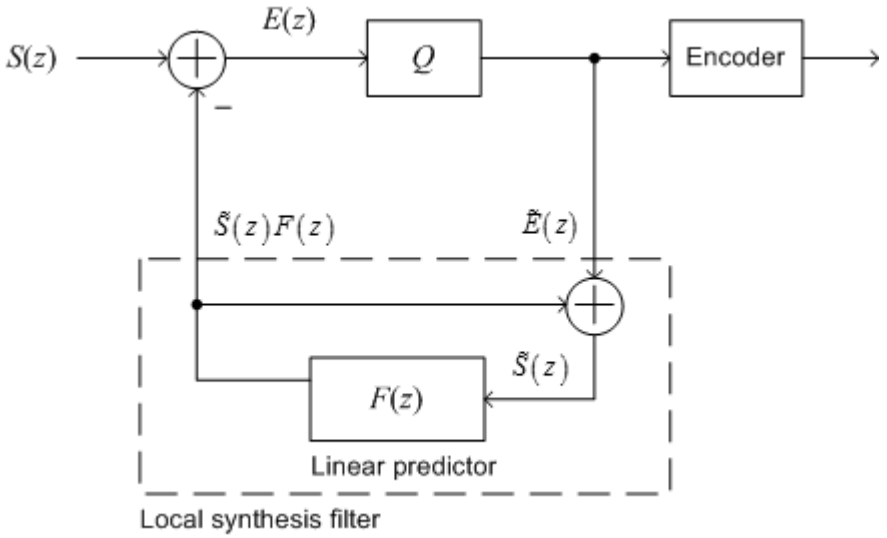
- This implies that $G > 1$ when $0 < |r_1| \leq 1$. The correlation coefficients could have positive or negative value.
  - Maximum normalized correlation is$|r_1| = 1$, whereby $G = \infty$ and adjacent samples are the same.
  - Zero temporal correlation when $|r_1| = 0$, where $G = 1$, where there is no redundancy to be removed.

EE6424 Part 2: Lecture 4.3

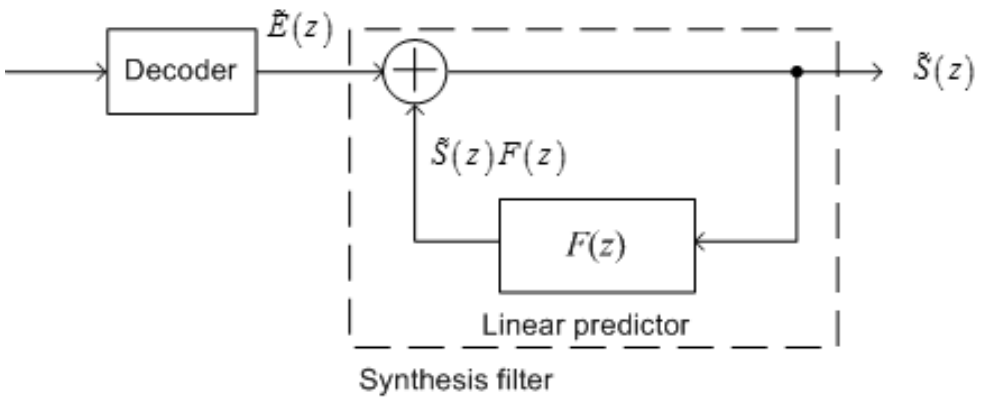
DIFFERENTIAL PULSE CODE MODULATION

Differential PCM

- DPCM is a special case of Linear Predictive Coding (LPC) with a first order linear predictor, $p = 1$ and $a_1 = -1$.
- The prediction residual is obtained by taking the difference between the current and previous samples $e(n) = s(n) - s(n-1)$, quantized, encoded and transmitted.



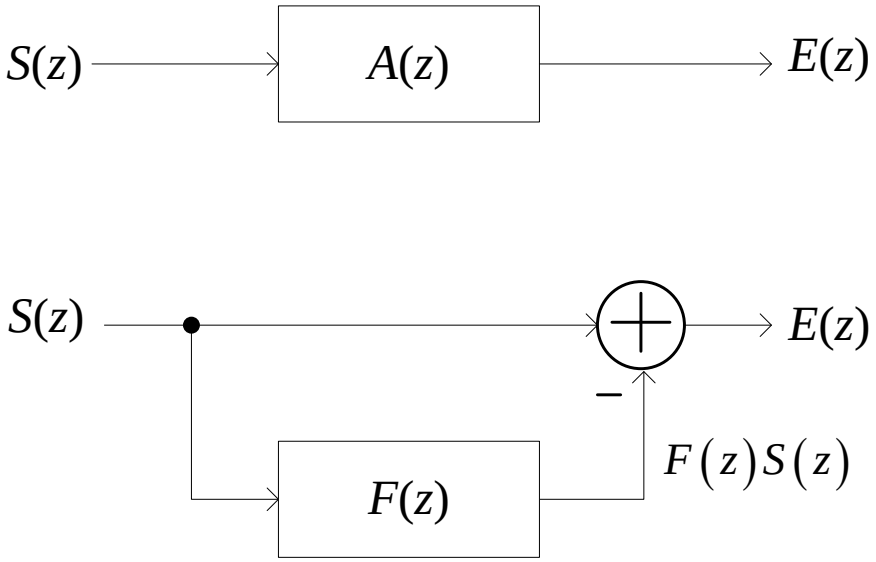
- A **local synthesis filter** similar to that in the receiver is located in the transmitter.
- It works in a similar fashion to obtain the quantized speech $\tilde{S}(z)$ from quantized residual $\tilde{E}(z)$.



- At the receiver, the quantized prediction residual $\tilde{E}(z)$ is combined with the linear predictor output to form the quantized speech $\tilde{S}(z)$.

- The quantizer can be **uniform** or **non-uniform** (i.e., logarithmic scale).
- In general LPC, the prediction coefficients could change with time so as to adapt to the input signal properties.
- The prediction coefficients have to be quantized and transmitted as well. Reflection coefficients or line spectrum pair (LSP) polynomials are used in practice for better quantization properties.

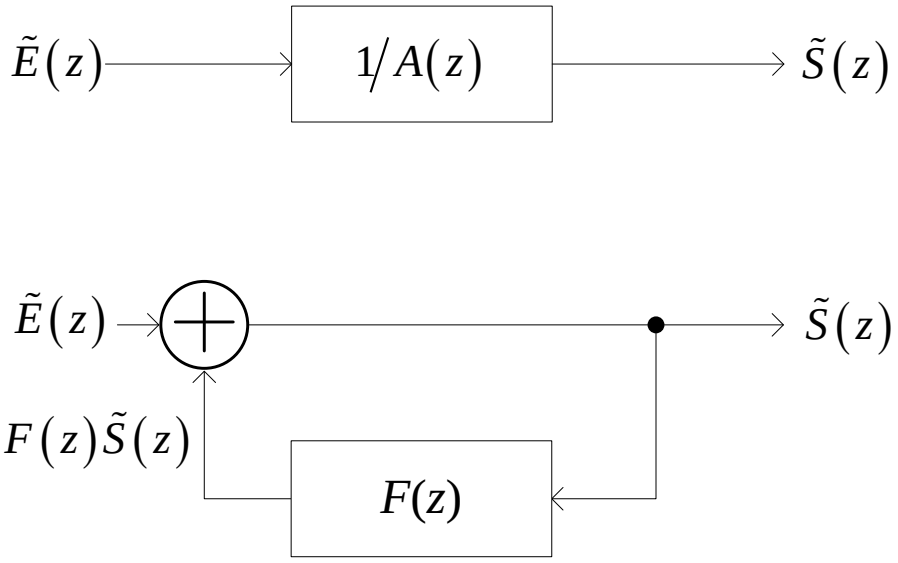
Inverse filtering (whitening)



Linear predictor

$$
\begin{aligned}
E(z) &= A(z)S(z) \\
&= \left[1+\sum_{i=1}^{p} a_i z^{-i}\right] S(z) \\
&= \left[1-\sum_{i=1}^{p} -a_i z^{-i}\right] S(z) \\
&= \left[1-F(z)\right] S(z) \\
&= S(z) - F(z)S(z)
\end{aligned}
$$

Local synthesis filter



Linear predictor

$$\tilde{S}(z) = \frac{1}{A(z)}\tilde{E}(z)$$

$$= \frac{1}{\left[1 - F(z)\right]}\tilde{E}(z)$$

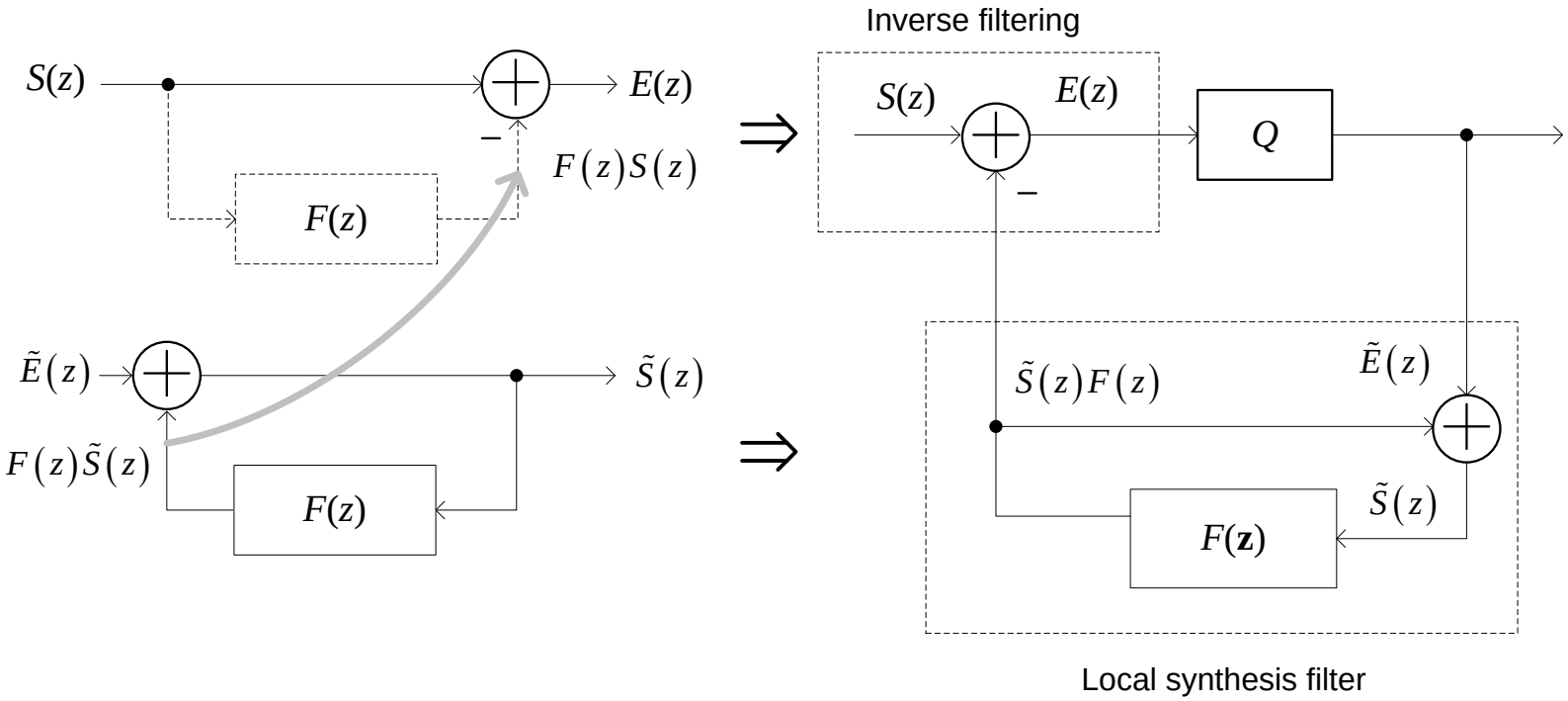
$$\tilde{S}(z)\left[1 - F(z)\right] = \tilde{E}(z)$$

$$\tilde{S}(z) - F(z)\tilde{S}(z) = \tilde{E}(z)$$

$$\tilde{S}(z) = \tilde{E}(z) + F(z)\tilde{S}(z)$$

Local synthesis filter

- A local synthesis filter is used as opposed to a direct inverse filter in the transmitter of a DPCM. The main reason is that the decoder has no access to the original input $S(z)$.
- This copes with the problem of accumulated encoder error by maintaining synchronization between encoder and decoder.
- This is done by replacing $S(z)F(z)$ with the estimate $\tilde{S}(z)F(z)$.


Inverse filtering
S(z)
E(z)
F(z)
F(z)S(z)
Q
Ẽ(z)
S̃(z)
F(z)S̃(z)
S̃(z)F(z)
Ẽ(z)
S̃(z)
F(z)
Local synthesis filter

Prediction gain of DPCM

- The **prediction gain** is given by

$$G = \frac{E\{s^2(n)\}}{E\{e^2(n)\}}$$

- The variance of the **prediction residual**

$$\begin{aligned} E\{e^2(n)\} &= E\{[s(n) - s(n-1)]^2\} \\ &= E\{s^2(n) - 2s(n)s(n-1) + s^2(n-1)\} \\ &= 2E\{s^2(n)\} - 2\gamma(l) \\ &= 2E\{s^2(n)\}\left(1 - \frac{\gamma(l)}{E\{s^2(n)\}}\right) \\ &= 2E\{s^2(n)\}(1 - r_1) \end{aligned}$$

- Using these results, the **prediction gain** of the DPCM is given by

$$G = \frac{E\{s^2(n)\}}{2E\{s^2(n)\}(1-r_1)} = \frac{1}{2(1-r_1)}$$

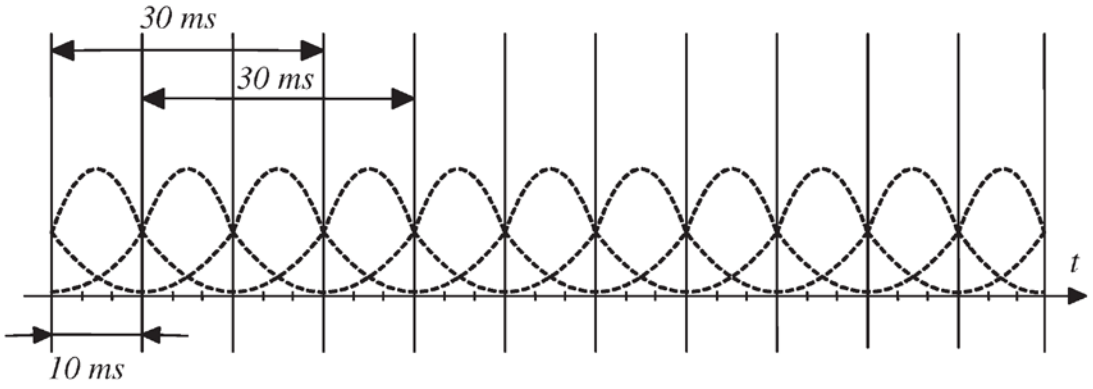
- The differential coding is effective ($G > 1$), when $r_1 > 0.5$.
- **Delta modulation** (DM or ΔM) is an extreme case of DPCM:
  - The sampling frequency is raised very high such that the difference between adjacent samples is extremely small (highly correlated between adjacent samples).
  - This leads to high prediction gain (small prediction residual), which allows 1-bit quantization to be used.

EE6424 Part 2: Lecture 4.4

CODE EXCITED LINEAR PREDICTION

Frame-based speech processing

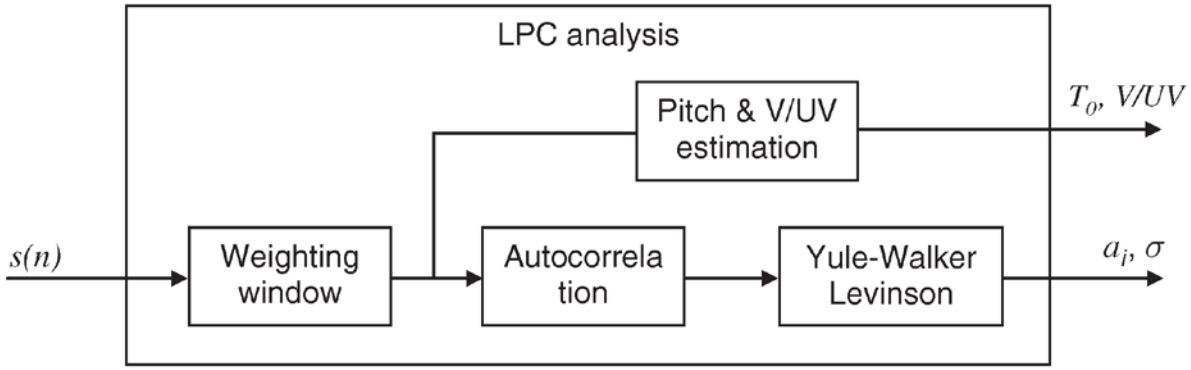
- Speech properties could be assumed **stationary** within a 10 to 30 ms frame given the inertia of the articulators.
- The LP model is applied on speech frames, typically 30 ms and with an **overlap** of 20 ms (or equivalently a **shift** of 10 ms).
- Speech frames are weighted using a **weighting window** (e.g., Hamming) to prevent the first few samples of the frame from having too much weight in the auto-correlation function and LP coefficients estimation.



- The linear prediction of a current sample is based on previous samples; the first few samples in a frame cannot be correctly predicted. This justifies the use of **weighting window**.
- The LP model does not replicate the exact speech waveform. It models the **spectral envelop** based on the idea that our ear is more sensitive to the amplitude spectrum than to the phase spectrum.
- The **prediction coefficients** $\{a_i\}$ are computed for every **frame** depending on the frame rate.

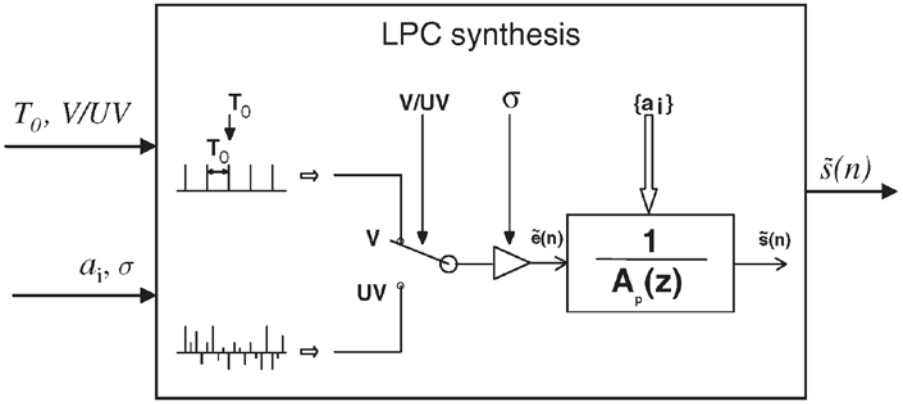
LPC10 (FS-1015) Linear Predictive Coder

- Developed by US Department of Defense and later by NATO.
- The **excitation** of the LP model is assumed to be one of the following:
  - Periodic pulses with period T_0 and amplitude σ for **voiced** speech
  - White Gaussian noise with variance σ^2 for **unvoiced** speech
- In both cases, the **spectral envelope** of the excitation is flat.



- Each frame of speech (22.5 ms) in coded with **54 bits**.
- The available bits are distributed over parameters so as to obtain an optimal quality for the coded speech:
  - Pitch period (T_0) and voiced/unvoiced decision (7 bits)
  - 10 prediction coefficients $\{a_i\}$ (42 bits)
  - Gain σ (5 bits).
- With no frame overlap, the bit rate is $54 \times \frac{1}{22.5\times10^{-3}} = 2400$ bps.
- Each frame of 180 samples, coded with 54 bits, amounts to an average of **0.3 bit per sample**.

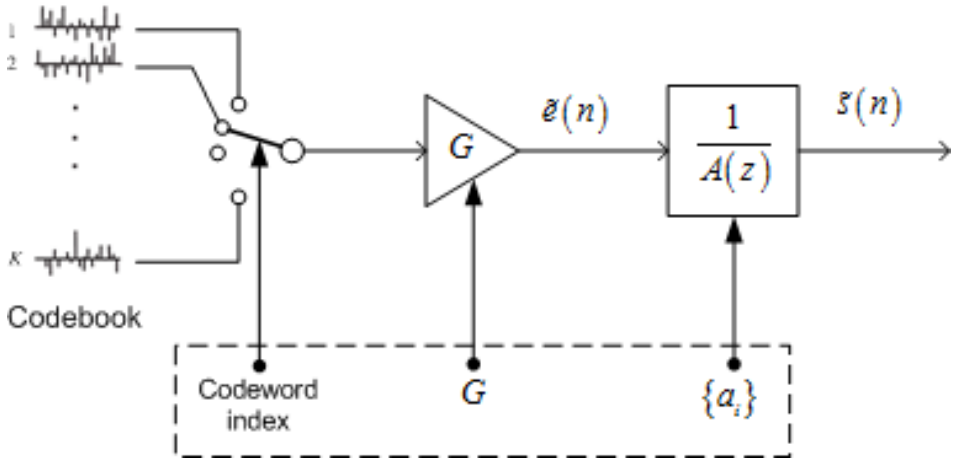
- The **excitation** $\tilde{e}(n)$ is regenerated at the receiver using the voiced/unvoiced decision and pitch period information.
- The **prediction residual** is never transmitted, which degrades the quality of the LPC10 coder.
- Transmitting the prediction residual, even after quantization, requires a huge number of bits for the quality to be acceptable.



- The LPC10 coder is very sensitive to the efficiency of its **voiced/unvoiced** decision and **pitch period** T_0 estimation, which is a difficult problem, for the following reasons:
  - Glottal cycle amplitude and duration varies from period to period (shimmer and jitter).
  - Pitch period T_0 has to be estimated from the filtered version of the glottal pulses (the voiced speech as we heard).
- One way to enhance the quality of the LPC10 coder is to reduce the constraint on the LP model excitation $\tilde{e}(n)$, as for the case of CELP.

Code excited linear prediction

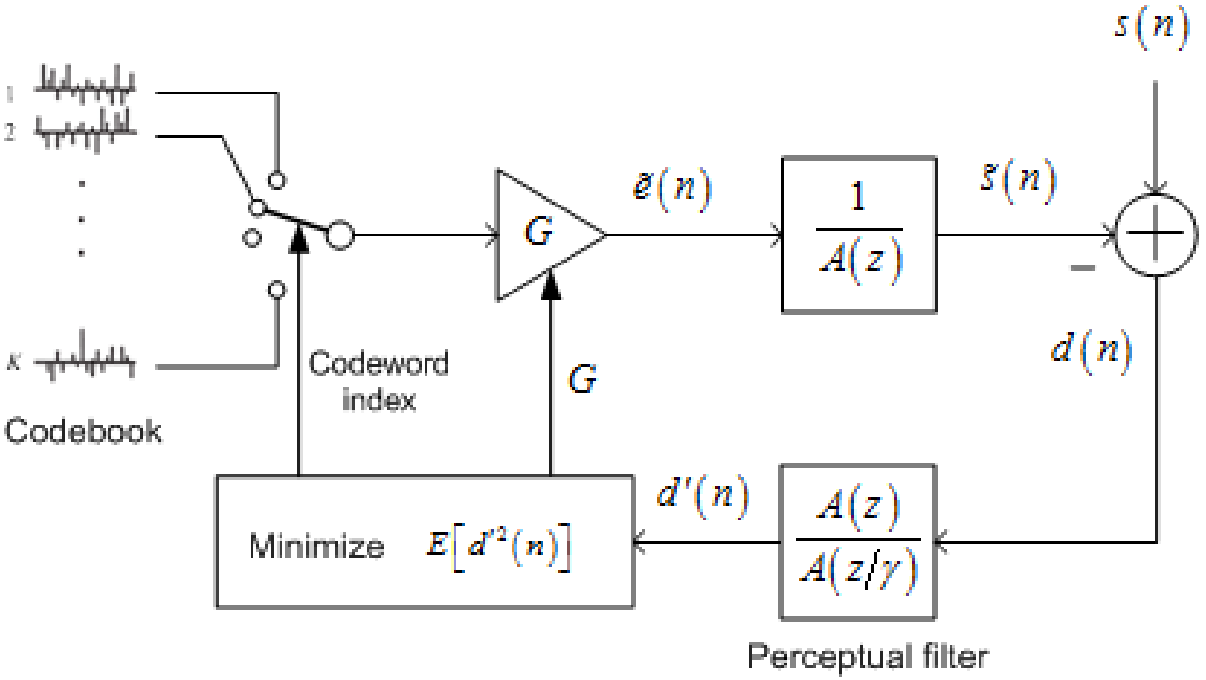
- The central idea of CELP is to perform vector quantization (VQ) on the LP residual (equivalently the LP excitation sequence).
- The encoder selects one excitation sequence from a predefined **stochastic codebook** of possible sequences.
- The index of the selected sequence is sent to the decoder together with the quantized LP coefficients and a gain factor G.



CELP decoder

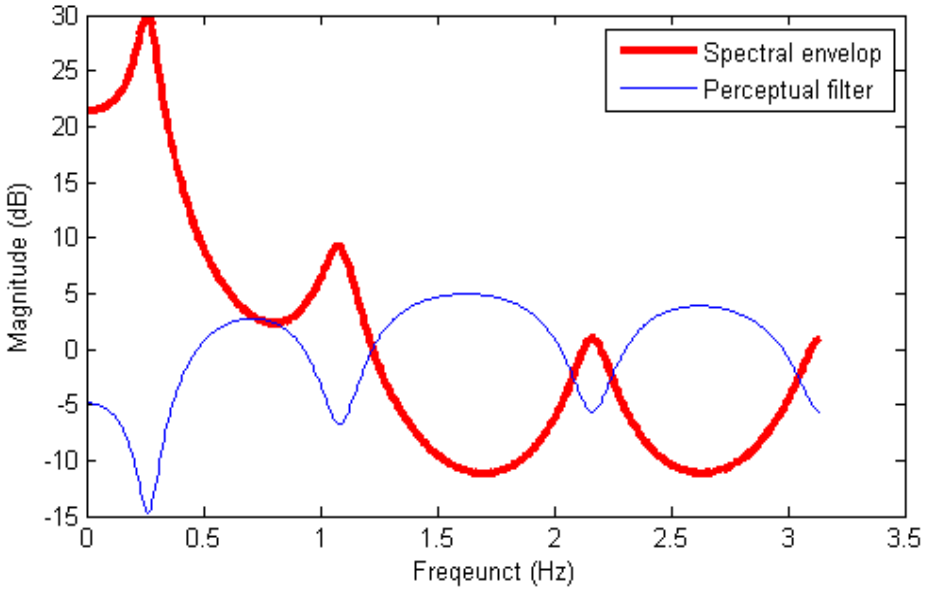
- The **best codeword** (**LP excitation**) is chosen iteratively so as to minimize the energy of modeling error $d(n)$ taken as the difference between the original speech $s(n)$ and the synthesized speech $\tilde{s}(n)$.
- Different from the standard VQ, the best match excitation is selected in a closed-loop through an **analysis-by-synthesis** process.
  - In the **standard VQ**, the selected codeword has the smallest distance to the LP residual.
  - In CELP, the selected codeword is not necessary the one that is most similar to the LP residual. This is due to the existence of synthesis and perceptual filters in the loop.
- Pitch estimation and voice and unvoiced decision are no-longer required.

- The modeling error $d(n)$ is filtered by a perceptual filter before its energy $E\left[d'^2(n)\right]$ is computed.



CELP encoder

- The **perceptual filter** is constructed based on the **frequency masking** property of human hear.
- It scales up the error at the frequencies with less energy for an input frame. Quantization noise in these frequencies will not be masked as compared to those at the peak of the spectral envelop.



EE6424 Part 2: Lecture 4.5

HISTORY AND FUTURE PROSPECTS

Telephone bandwidth speech coder

- Most speech coding systems operate over the **telephone bandwidth** with the frequency contents limited between 300 to 3400 Hz sampled at 8 kHz.
- The **reference bit rate** is taken as 128 kbps (16 bit, 8 kHz), and is referred to as **toll quality**.

Coder	Bit-rate	Bit rate
Reference bit-rate	128 kbps	16 bits × 8 kHz = 128 kbps
12-bit linear PCM	96 kbps	12 bits × 8 kHz = 96 kbps
8-bit μ-law PCM	64 kbps	8 bits × 8 kHz = 64 kbps
Adaptive DPCM (ADPCM)	32 kbps	4 bits × 8 kHz = 32 kbps

- **ADPCM** operates at 32 kbps and provides a speech quality comparable to PCM at 64 kbps. This corresponds to a bit rate reduction by a factor of 2.
- The 32 kbps ADPCM coder was standardized by CCITT in 1984 as the G.721 standard. Extensions at 40, 24, and 16 kbps were proposed in 1988 as the G.726 standard.
- ADPCM is still of widespread use worldwide.

Coding Standards for Cell Phone

- The **LP10** standard (NATO Standard STANAG-4198) was used for satellite transmission of speech communication formally from 1984 till 1996.
- In 1996, the LPC10 was replaced by **MELP** (Mixed excitation LP), a variant of the **CELP** coder, to be the US Federal Standard for coding at 2.4 kbps.
- In 1996, a variant of the CELP termed as the **algebraic CELP** (ACELP) has adapted as the **enhanced full-rate** (EFR) codec for the **GSM** (Global System for Mobile communication).
- The EFR codec operates at **11.2 kbps** and produces a much better speech quality. Also, it implements an efficient codebook searching algorithm thereby eliminating one main drawback of the original CELP coder.
- A variant of the ACELP coder has been standardized by ITU-T as G.729 for operation at a bit rate of **8 kbps**.

Coder	Bit-rate	Bit rate
Reference bit-rate	128 kbps	16 bits × 8 kHz = 128 kbps
12-bit linear PCM	96 kbps	12 bits × 8 kHz = 96 kbps
8-bit μ-law PCM	64 kbps	8 bits × 8 kHz = 64 kbps
Adaptive DPCM (ADPCM)	32 kbps	4 bits × 8 kHz = 32 kbps
LP10 and MELP	2.4 kbps	0.3 bits × 8 kHz = 2.4 kbps
GSM EFR	11.2 kbps	1.4 bits × 8 kHz = 11.2 kbps
ITU-T G.729	8 kbps	1 bit × 8 kHz = 8 kbps

Summary

- One might wonder if speech coding still relevant consider the following facts:
  - Memory capacity of semiconductor almost doubles for every 18 months
  - High bit rate digital transmissions are available through optical fibers
- Speech coding will continue to be in demand for the following reasons:
  - Wider bandwidth with better quality is required in more and more applications, which means higher bit-rate.
  - Optical fibers links has limited coverage. Radio communication remains the only viable option with inherent bit rate constraint.