Digital Transmission - Part 2
by
Lajos (Layosh) Hanzo

School of ECS, UoS, SO17 1BJ, UK.
http://www-mobile.ecs.soton.ac.uk

References


a For detailed contents please refer to http://www-mobile.ecs.soton.ac.uk
b For detailed contents please refer to http://www-mobile.ecs.soton.ac.uk

Speech Signals and Introduction to Speech Coding

1 Basic Characterisation of Speech Signals

- In contrast to the so-called deterministic signals, random signals, such as speech, music, video, etc information signals cannot be described by the help of analytical formulae.
- They are typically characterised by the help of statistical characteristics, such as the PSD, ACF, CDF and PDF.
- In simple terms human speech is generated by emitting sound pressure waves, radiated primarily from the lips, although significant energy emanates in case of some sounds also from the nostrils, throat, etc.
The air compressed by the lungs excites the vocal cords in two typical modes.

When generating voiced sounds, the vocal cords vibrate and generate a high-energy quasi-periodic speech waveform.

For unvoiced sounds the vocal cords do not participate in the voice production and the source behaves similarly to a noise generator.

The excitation signal denoted by $E(z)$ is then filtered through the vocal apparatus, which behaves like a spectral shaping filter with a transfer function of $H(z) = 1/A(z)$ that is constituted by the spectral shaping action of the glottis, which is defined as the opening between the vocal folds, that of the vocal tract, lip radiation characteristics, etc.

The excitation signal $E(z)$ is then filtered through the vocal apparatus, which behaves like a spectral shaping filter with a transfer function of $H(z) = 1/A(z)$ that is constituted by the spectral shaping action of the glottis, which is defined as the opening between the vocal folds, that of the vocal tract, lip radiation characteristics, etc.

Figure 1: Linearly separable speech source model

Figure 2: Typical voiced speech segment and its PSD for a male speaker

Figure 3: Typical unvoiced speech segment and its PSD for a male speaker
2 Classification of Speech Codecs

![Diagram: Speech Quality versus Bitrate Classification of Speech Codecs]

- EXCELLENT
- GOOD
- FAIR
- POOR

- Vocoder
- Complex Delay
- Waveform Codecs
- Hybrid Codecs

![Diagram: Vocoder schematic]

Figure 4: Typical voiced speech segment and its ACF for a male speaker

Figure 5: Typical unvoiced speech segment and its ACF for a male speaker

Figure 6: Speech Quality versus Bitrate Classification of Speech Codecs

Figure 7: Vocoder schematic
3 Waveform Coding

3.1 Digitisation of Speech

- Anti-aliasing low-pass filtering to 4 kHz neglects about 1% of speech energy, hence 8 kHz wideband channels are used in commentary channels.
- The bandlimited speech is sampled according to the Nyquist theorem.
- Lastly, amplitude discretisation or quantisation must be invoked.

\[ \text{Speech} \xrightarrow{\text{LPF}} \text{Sampling} \xrightarrow{\text{Quantisation}} \text{Parallel To Serial Converter} \xrightarrow{\text{Binary Speech Bits}} \]

![Figure 8: Digitisation of analogue speech signals](image)

\[ \text{Figure 8: Digitisation of analogue speech signals} \]

3.2 Quantisation Characteristics

- The original speech signal is contaminated by quantisation noise, which is a function of the signal’s distribution, the quantiser’s resolution and its transfer characteristic.
- Linear quantisers exhibit a linear transfer function within the dynamic range and saturation above that.
- In Figure 10 according to \( R = 3 \) there are \( 2^3 = 8 \) reconstruction levels and a so-called mid-tread quantiser is featured.
- The mid-riser quantiser’s transfer function exhibits a level change at the abscissa value of zero.
- When the quantiser characteristic saturates at its maximum output level, the quantisation error increases without limit.

![Figure 9: Sampled and quantised analogue speech signal](image)

\[ \text{Figure 9: Sampled and quantised analogue speech signal} \]

![Figure 10: Linear quantisers and their quantisation errors: a/ Midtread, b/ Non-uniform](image)

\[ \text{Figure 10: Linear quantisers and their quantisation errors: a/ Midtread, b/ Non-uniform} \]
3.3 Quantisation Noise and Rate-Distortion Theory

- Observe in Figure 10 that the instantaneous quantisation error $e(x)$ is dependent on the instantaneous input signal level.
- In other words, $e(x)$ is non-uniform across the quantiser’s dynamic range and some amplitudes are represented without quantisation error, while others are associated with larger errors.
- If the input signal’s dynamic range exceeds the quantiser’s linear range, the quantiser’s output voltage saturates at its maximum level and the quantisation error may become arbitrarily high.
- Hence the knowledge of the input signal’s statistical distribution is important for minimising the overall granular and overload distortion.

$$\hat{x}(t) = x(t) + e(t)$$  \hspace{1cm} (1)

- Explicitly, for a rv $x$ with variance of $\sigma_x^2$ and quantised value $\hat{x}$ the distortion is defined as the mean squared error (mse) expression given by:

$$D = E[(x - \hat{x})^2] = E[e(t)^2],$$  \hspace{1cm} (3)

where $E$ represents the expected value.
- If $R_D = 0$ bits are used to quantise the quantity $x$, then the distortion is given by the signal’s variance $D = \sigma_x^2$. If, however, more than zero bits are used, i.e. $R_D > 0$, then intuitively one additional bit is needed every time we want to halve the rms value of ‘$D’$, or quadruple the signal-to-noise ratio $\text{SNR} = \sigma_x^2 / D$, which suggests a logarithmic relation between $R_D$ and $D$. After Shannon and Gallager we can write:

$$R_D = \frac{1}{2} \log_2 \frac{\sigma_x^2}{D} \text{ if } D \leq \sigma_x^2,$$  \hspace{1cm} (4)

If no amplitude discretisation is used, a sampled analogue source has formally an infinite entropy, requiring an infinite transmission rate.
- If the analogue speech samples are quantised to $R$-bit accuracy, there are $q = 2^R$ legitimate samples, each of which has a probability of occurrence $p_i, i = 1, 2 \ldots q$.
- The above $R$ bit/symbol channel capacity requirement can be further reduced to the value of the source’s entropy given by:

$$H(x) = -\sum_{i=1}^{q} p_i \cdot \log_2 p_i$$  \hspace{1cm} (2)

using entropy coding without inflicting any further coding impairment, if an infinite delay coding scheme is acceptable, although this is not practical.
- The so-called rate-distortion theorem, which quantifies the minimum required average bit rate $R_D$ in terms of [bit/sample] in order to represent a rv with less than $D$ distortion.

$$R_D = \begin{cases} \frac{1}{2} \log_2 \frac{\sigma_x^2}{D} & D \leq \sigma_x^2 \\ 0 & D > \sigma_x^2. \end{cases}$$  \hspace{1cm} (5)

The relationship of $D$ versus $R_D$ is shown in Figure 11.

Figure 11: Stylised $D$ versus $R_D$ curve
This simple result is useful for quick SNR estimates and it is also intuitively plausible, since every new bit used halves the quantisation error and hence doubles the SNR.

In practice the speech PDF is highly nonuniform and the quantiser’s dynamic range cannot be fully exploited, in order to minimise the quantiser characteristic overload error.

Hence Equation 10 over-estimates the expected SNR.

### 3.4 Non-uniform Quantisation: Companding

- If the input signal’s PDF is known and can be considered stationary, higher SNR can be achieved by appropriately matched non-uniform quantisation (NUQ) than in case of uniform quantisers.

The input signal’s dynamic range is partitioned in to non-uniformly spaced segments as we have seen in Figure 10.

- The quantisation intervals are more dense near the origin in order to quantise the typically high-probability low-magnitude samples more accurately.

- In contrast, the lower probability signal PDF tails are less accurately quantised.

- In contrast to uniform quantisation, where the maximum error was constant across the quantiser’s dynamic range, for non-uniform quantisers the SNR becomes more or less constant across the signal’s dynamic range.

- It is intuitively advantageous to render the length of the quantisation intervals or quantiles inversely proportional to the signal PDF, since a larger quantisation error is affordable in case of infrequent signal samples and vice versa.

- One of the possible system models is shown in Figure 12, where the input signal is first compressed using a so-called non-linear compander characteristic and then uniformly quantised.
The qualitative effect of non-linear compression on the signal’s PDF is portrayed in Figure 13, where it becomes explicit, why the compressed PDF can be quantised by a linear quantiser.

Observe that the compander has a more gentle slope, where larger quantisation intervals are expected and vice versa, implying that the compander’s slope is proportional to the quantisation interval density and inversely proportional to the stepsize in any given input signal interval.

Following Bennett’s approach Jayant and Noll have shown that if the signal’s PDF \( p(x) \) is a smooth, known function and sufficiently fine quantisation is used, then the quantisation error variance can be expressed as:

\[
\sigma_q^2 \approx \frac{q^2}{12} \int_{-x_{\text{max}}}^{x_{\text{max}}} \frac{p(x)}{|\dot{C}(x)|^2} dx,
\]

(11)

where \( \dot{C}(x) = dC(x)/dx \) represents the slope of the compander’s characteristic.

Jayant and Noll have also shown that the minimum quantisation error variance is achieved by the compander characteristic given by:

\[
C(x) = x_{\text{max}} \frac{\int_{0}^{x} \sqrt{p(x)} dx}{\int_{0}^{x_{\text{max}}} \sqrt{p(x)} dx},
\]

(12)

where the denominator constitutes a normalising factor.

Hence a simple practical compander design algorithm can be devised by evaluating the signal’s histogram in order to estimate the PDF \( p(x) \) and then graphically integrating \( \sqrt{p(x)} \) according to Equation 12 up to the abscissa value \( x \).

Although this technique minimises the quantisation error variance or maximises the SNR, if the input signal’s PDF or variance is time-variant, the compander’s performance degrades.

In many practical scenarios this is the case and hence often it is advantageous to optimise the compander’s characteristic to maximise the SNR independently of the shape of the PDF.
3.5 Logarithmic Compression

- The input signal’s variance is computed in case of an arbitrary PDF $p(x)$ as follows:
  \[ \sigma_x^2 = \int_{-\infty}^{\infty} x^2 p(x) dx. \]  
  (13)
- Assuming no saturation distortion, the SNR can be expressed from Equations 11 and 13 as follows:
  \[ \text{SNR} = \frac{\sigma_x^2}{\sigma_e^2} = \frac{\int_{-x_{\text{max}}}^{x_{\text{max}}} x^2 p(x) dx}{\frac{q_T}{12} \int_{-x_{\text{max}}}^{x_{\text{max}}} \left( p(x)/|\hat{C}(x)|^2 \right) dx}. \]  
  (14)
- In order to maintain an SNR value that is independent of the PDF $p(x)$ the numerator must be a constant times the denominator, which is equivalent to requiring that:
  \[ |\hat{C}(x)|^2 = \left| \frac{K}{x} \right|^2 \]  
  or alternatively that:
  \[ \hat{C}(x) = K/x \]  
  (16)
and hence:
  \[ C(x) = \int_{x}^{x_{\text{max}}} \frac{K}{z} dz = K \cdot \ln x + A \]  
  (17)
- This compander characteristic is shown at the left hand side of Figure 14 and ensures a constant SNR across the signal’s dynamic range, irrespective of the shape of the signal’s PDF.
- Intuitively, large signals can have a large error while small signal must maintain a low distortion, which gives a constant SNR for different input signal levels.

- The logarithmic characteristic is implemented as a 16-segment piece-wise linear approximation, as seen in Figure 15 and below:

\[
\begin{align*}
\text{sign} & & b_7 & b_6 & b_5 & b_4 & b_3 & b_2 & b_1 & b_0 \\
\text{segments} & & & & & \text{uniform quant.} & \text{in each segment}
\end{align*}
\]
- The transmission rate is $8 \times 8 = 64$ kbit/s (kbps), giving unimpaired speech quality, which would require about 12 bits in case of linear quantisation.
- There are two practical logarithmically companded standard speech codecs, namely the European A-law compander and the American $\mu$-law compander, both of which operate at 64 kbps.
Predictive Coding

1 Forward Predictive Coding

- **Motivation:** One could argue that if the input signal is correlated, the previous sample can be used to predict the present one.
- The prediction error (PE) constituted by the difference of the current sample and the previous one is significantly smaller on the average than the input signal, reducing the region of uncertainty in which the signal to be quantised can reside, and whence allows us to use either a reduced number of quantisation bits or a better resolution in coding.
- Recall that redundancy exhibits itself both in terms of the PSD and the ACF, as it was demonstrated by Figures 2 and 4 in case of voiced speech signals. The more flat the ACF and non-flat the PSD, the the better the predictions.

\[ e_q(n) = e(n) = s(n) - \hat{s}(n), \] (18)

we can generate the decoded speech \( s_q(n) = s(n) \) by the help of the predictor at the decoder’s end of the speech link by simply adding the quantised predicted value \( \hat{s}_q(n) \) to \( e_q(n) = e(n) \) as follows:

\[ s_q(n) = \hat{s}_q(n) + e_q(n). \] (19)

2 DPCM Codec Schematic

- In our forward-predictive codec we assumed that no transmission errors occurred. In the presence of Tx errors or if there is quantisation distortion, which is typically the case, if bit rate economy is an important factor - Fwd-predictive schemes perform purely.
- These problems can be circumvented by the backward predictive scheme of Figure 40.

\[ s_q(n) = \hat{s}(n) + e_q(n). \] (20)

- Observe in Figure 40 that in contrast to the forward predictive scheme of Figure 16, the input signal \( s(n) \) is predicted not from the previous values of \( s(n-k), \ k = 1 \ldots p \), but from:

\[ s_q(n) = \hat{s}(n) + e_q(n). \] (20)

- Since the locally re-constructed signal \( s_q(n) \) is contaminated by the quantisation noise of \( q(n) = e(n) - e_q(n) \) inherent in \( e_q(n) \), one could argue that this prediction will be probably a less confident one than that based on \( s(n-k), \ k = 1 \ldots p \), which might affect the coding efficiency of the scheme.

- Observe however that the signal \( s_q(n) \) is also available at the decoder, irrespective of the accuracy of the quantiser’s resolution. Although in case of transmission errors this is not the case, due to the codec's stabilising predictive feed-back loop the effect of transmission errors decays, while in case of the forward predictive scheme of Figure 16 the transmission errors persist.
Note in Figure 40 that the encoder’s backward oriented section is identical to the decoder’s schematic, which is referred to as the local decoder, outputting the locally re-constructed signal.

\[ s(n) = q(n) + e(n) \]

\[ s(n) = q(n) + e_q(n) \]

\[ s(n) = \tilde{s}(n) + e_q(n) \]

\[ \tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \]

\[ e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \]

\[ e_q(n) = Q[e(n)] \]

\[ s_q(n) = \tilde{s}(n) + e_q(n) \]

3 Predictor Design

3.1 Problem Formulation

- Due to the redundancy inherent in speech any present sample can be predicted as a linear combination of \( p \) past speech samples as follows:

\[ s(n) = \sum_{k=1}^{p} a_k s(n-k), \] (22)

where \( p \) is the predictor order, \( a_k \) represents the linear predictive coding (LPC) coefficients and \( \tilde{s}(n) \) the predicted speech samples.

- The prediction error, \( e(n) \), is then given by

\[ e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \]

\[ = \sum_{k=0}^{p} a_k s(n-k) \text{ where } a_0 = 1. \] (23)

- However, when \( e(n) \) is quantised to \( e_q(n) \) for bit rate economy, this is not true. It is probably not the best, but an attractive approach to determine \( a_k \) by minimising the mean-squared PE of Equation 23.

- Generating the synthesised speech can also be portrayed as exciting the all-pole synthesis filter \( H(z) = 1/A(z) \) with the excitation signal \( E(z) \).

- If the predictor removes the redundancy from the speech signal by minimising the prediction residual, \( e(n) \) becomes unpredictable, with an essentially flat spectrum, while \( H(z) = 1/A(z) \) models the spectral envelope of the speech.

- Due to the relationship \( A(z) = H^{-1}(z) \) \( A(z) \) is often referred to as the LPC inverse filter.
The expected value \( E \) of the mean-squared prediction error of Equation 23 can be written as:

\[
E[e^2(n)] = E[s(n) - \sum_{k=1}^{p} a_k s(n-k)]^2.
\] (27)

- To derive the optimum LPC coefficients we compute the partial derivative of Equation 27 with respect to all LPC coefficients and set \( \partial E / \partial a_i = 0 \) for \( i = 1 \ldots p \), which yields a set of \( p \) unknown LPC coefficients \( a_i \) as follows:

\[
\frac{\partial E[e^2(n)]}{\partial a_i} = -2 \cdot E \left\{ s(n) - \sum_{k=1}^{p} a_k s(n-k) \right\} s(n-i) = 0,
\] (28)

yielding:

\[
E\{s(n)s(n-i)\} = E \left\{ \sum_{k=1}^{p} a_k s(n-k)s(n-i) \right\}.
\] (29)

\[ C(i,k) \] ought to be determined from Equation 31 over an infinite interval, but this is impractical and 'rigid'.

- In low-complexity codecs or for stationary inputs \( C(i,k) \) can be determined using a long 'off-line' training.

- Then the set of \( p \) Equations 32 can be solved for example by Gauss-Jordan elimination, or more efficiently by the Berlekamp-Massey algorithm, the Levinson-Durbin algorithm to be described later, etc.

- In complex, LBR codecs the LPC coefficients are determined adaptively for quasi-stationary intervals to improve the efficiency of the predictor, to reduce the prediction error’s variance and hence to improve the coding efficiency.

- These time-variant LPC coefficients must be quantised and transmitted to the decoder, hence this technique is often referred to as forward adaptive prediction (FAP).

- Upon exchanging the order of the summation and expected value computation at the RHS of Equation 29 we arrive at:

\[
E\{s(n)s(n-i)\} = \sum_{k=1}^{p} a_k E\{s(n-k)s(n-i)\}, \quad i = 1, \ldots, p. \] (30)

where

\[
C(i,k) = E\{s(n-i)s(n-k)\}, \] (31)

represents the input signal’s covariance coefficients, leading to:

\[
\sum_{k=1}^{p} a_k C(i,k) = C(i,0), \quad i = 1, \ldots, p. \] (32)

### 3.2 Covariance Coefficient Computation

- Apart from LPC theory these Equations accrue in optimising other adaptive filters, such as channel equalisers or in the auto-regressive filter representation of FEC block codes.

- In backward adaptive prediction (BAP) the LPC coefficients are not transmitted to the decoder, they are recovered from the previously decoded signal.

- Again, to ensure the identical operation of the local and distant decoders, the encoder also uses previous decoded signal segments, rather than unquantised input signal segments for LPC coefficient computation.

- For efficient prediction the delay associated with BAP must be as low as possible, while the decoded signal quality has to be as high as possible, suggesting that this technique is applicable to high-quality, high-rate coding. It is not used in LBR applications, where the higher delay and coding distortion would reduce the predictor’s efficiency.

- Later we will return to these codec classes and augment their main features by referring to practical standardised coding arrangements belonging to both families.
In spectrally efficient high quality FAP codecs $C(i, k)$ of Equation 31 is typically computed for intervals, during which the signal’s statistics can be considered quasi-stationary, say 20–30 ms.

This has a bitrate-impact on FAP coding, which is not so in BAP coding, although too frequent LPC-updates are limited by the complexity of the frequent solution of the high order set of Equations 32, since the low-delay spectral estimation requirement does not tolerate the too infrequent updating of the LPC coefficients.

### 3.3 Predictor Coefficient Computation

In the autocorrelation method the prediction error term of Equation 27 is minimised over the finite interval of $0 \leq n \leq L_a - 1$, rather than $-\infty < n < \infty$.

Hence the set of $p$ Equations 32 can be reformulated:

$$\sum_{k=1}^{p} a_k R(|i-k|) = R(i), \quad i = 1, \ldots, p. \quad (37)$$

Equation 37 can be re-written in a matrix form as:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \ldots & R(p-1) \\ R(1) & R(0) & R(1) & \ldots & R(p-2) \\ R(2) & R(1) & R(0) & \ldots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \ldots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \quad (38)$$

The $p \times p$ autocorrelation matrix above has a Toeplitz structure, where all the elements along a diagonal are identical.

Equation 38 can be solved using matrix inversion by Gauss-Jordan elimination having a complexity of $\approx p^3$, while the often-used Levinson-Durbin (LD) algorithm has a complexity proportional to $6 \cdot p^2$, where the latter is lower for $p \leq 6$.

Hence $C(i, k)$ is now computed from the short-term expected value:

$$C(i, k) = \sum_{n=0}^{L_a-1} s(n-i)s(n-k), \quad i = 1, \ldots, p, \quad (33)$$

Upon setting $m = n - i$, Equation 33 can be expressed as

$$C(i, k) = \sum_{m=0}^{L_a-1-(i-k)} s(m)s(m + i - k), \quad (34)$$

which suggests that $C(i, k)$ is the short-time autocorrelation of the input signal $s(m)$ evaluated at a displacement of $(i-k)$, giving:

$$C(i, k) = R(i-k), \quad (35)$$

where

$$R(j) = \sum_{n=0}^{L_a-1-j} s(n)s(n+j) = \sum_{n=j}^{L_a-1} s(n)s(n-j), \quad (36)$$

represents the speech autocorrelation coefficients.

Let us define the prediction gain, as the ratio of the expected value of the input signal’s energy, namely $R_e(0)$, and that of the prediction error energy $R_e(0)$ expressed in terms of the corresponding autocorrelation coefficients as follows:

$$G = \frac{R_e(0)}{R_e(0)} \quad (39)$$

It is easily seen from Equations 32 and 38 that the so-called optimum one-tap predictor’s coefficient minimising the expectation of the prediction error is given by:

$$a_1 = \frac{R_e(1)}{R_e(0)} \quad (40)$$

which is the normalised one-step correlation between adjacent input samples.

**Example:** Given $R_e(1)/R_e(0) = 0.9$, determine $G$, when using the optimum one-tap predictor and non-optimum prediction using the previous sample as predicted value.
4 DPCM Performance

- In case of DPCM the quantiser operates on \( e(n) \) having a variance of \( \sigma_e^2 = E[|e(n)|^2] \), while the quantisation error variance is \( \sigma_q^2 \).
  
  Then applying the rate-distortion formula of Equation 37, the DPCM coding rate is given by:
  
  \[
  R_{\text{DPCM}} = \frac{1}{2} \log_2 \frac{\sigma_e^2}{\sigma_q^2} \text{[bits/pixel].} \quad (44)
  \]

- The coding rate reduction due to using DPCM is:
  
  \[
  \Delta R = R_{\text{PCM}} - R_{\text{DPCM}} = \frac{1}{2} \log_2 \frac{\sigma_e^2}{\sigma_q^2} - \frac{1}{2} \log_2 \frac{\sigma_e^2}{\sigma_q^2} = \frac{1}{2} \log_2 \frac{\sigma_e^2}{\sigma_q^2} \quad (45)
  \]

  giving a coding rate gain of:
  
  \[
  \Delta R \approx 1.66 \cdot \log_{10} \left( \frac{\sigma_e^2}{\sigma_q^2} \right)^2 \text{[bits/pixel].} \quad (46)
  \]

- For example, if \( \sigma_s = 10 \cdot \sigma_e \), then we have \( \Delta R = 3.332 \), giving more than 3 bits/sample bit rate saving.

- This leads to:
  
  \[
  \text{SNR}_{\text{DPCM}} \leq 10 \log_{10} \frac{\sigma_e^2}{\sigma_q^2 \cdot f(R)} \quad (49)
  \]

  where \( f(R) \) is the quantiser mean square distortion function for \( R \) number of quantisation bits in case of a unit variance input signal.

  For an equal number of quantisation bits the SNR improvement of DPCM over PCM is then given by:
  
  \[
  \Delta \text{SNR} = \text{SNR}_{\text{DPCM}} - \text{SNR}_{\text{PCM}} \quad (50)
  \]

  \[
  = 10 \log_{10} \frac{\sigma_s^2}{\sigma_q^2} - 10 \log_{10} \frac{\sigma_s^2}{\sigma_q^2} = 10 \log_{10} \frac{\sigma_s^2}{\sigma_q^2} \leq 10 \log_{10} \frac{\sigma_s^2}{\sigma_q^2}
  \]

- Again, assuming for example that \( \sigma_s = 10 \cdot \sigma_e \), we have a 20 dB SNR improvement over PCM, while maintaining the same coding rate.
In general the gains achievable will depend on the signal's statistics, as well as on the predictor (P) and quantiser (Q) designs. Usually Max-Lloyd quantisation (MLQ) is used, which is designed to match the prediction error’s PDF.

If prediction error’s PDF matches a Gaussian, Laplacian or Gamma distribution, the analytic quantiser designs tabulated in the literature can be invoked, otherwise specially trained ML quantisers must be employed.

As a result of accurate prediction the prediction residual typically becomes unpredicable or pseudo-random, which we described by the help of waveform coding techniques using an adaptive quantiser.

Although the high-quality encoding of the prediction residual is a sufficient criterion for perceptually high speech quality, it is not a necessary condition. Hence we will endeavour to improve the bit rate economy, while maintaining perceptually high speech quality.

Analysis-by-synthesis Principles

1 Motivation

- In the linearly separable speech production (LSSP) model the excitation signal is filtered through a slowly varying spectral shaping system to generate the speech signal.
- In an inverse approach one could view speech production as filtering the excitation $E(z)$ through the spectral shaping system $H(z) = 1/A(z)$ in order to generate $E(z)$.
- In predictive codecs a slowly varying adaptive predictor is used to estimate the incoming signal's spectrum.
- It was also shown, how this so-called short-term predictor can be made block-wise adaptive in order to accommodate changes in the signal's statistics and the LPC coefficients are transmitted to the predictive decoder.

2 Analysis-by-synthesis codec structure

- In the Analysis-by-synthesis (ABS) codec structure shown in Figure 19 vector quantisation techniques are invoked, where the synthesis filter is excited by an excitation vector of typically 5 ms or 40 samples length.
- A closed-loop structure is used, where the prediction error between the original input signal and the synthesized speech signal is evaluated for each candidate excitation vector and the specific excitation vector minimising the weighted error, rather than the conventional $\text{mse}$ is deemed to produce the best synthetic speech.
- As seen in Figure 19 the slowly varying synthesis filter(s) are excited by the so-called innovation sequences $u(n)$ of the excitation generator in order to produce the synthetic speech $\tilde{s}(n)$, which is compared with the input speech $s(n)$ about to be encoded.
The prediction error residual $e(n) = s(n) - \hat{s}(n)$ is formed and weighted by the error weighting filter, which will be described during our further discourse, in order to produce the perceptually weighted error $e_w(n)$.

ABS codecs - instead of minimising the usual mse term to provide best wave-form reproduction - they minimise the perceptually weighted error $e_w(n)$. Thereby they actually degrade the wave-form representation in favour of better subjective speech quality.

This is achieved at the cost of high complexity, since the synthetic speech is computed for all legitimate innovation sequences, sometimes several thousand times. The error weighting filter is designed to de-emphasize the weighted error in the vicinity of spectral formant regions, where the speech signal’s spectral prominences mask a higher reconstruction error, rendering the SNR near-constant over the speech signal’s frequency range.

The short-term synthesis filter is modelling the spectral envelope of the speech waveform. Its coefficients are computed by minimising the error of predicting a speech sample from a few, typically 8-10, previous samples, where minimisation is carried out over a quasi-stationary period of some 20 ms or 160 samples at $f_s = 8$ kHz.

The optional long-term synthesis filter models the ‘fine structure’ of the speech spectrum, which predicts the long-term periodicity of speech persisting after short term prediction, reflecting the pitch periodicity of the residual.

The decoder uses an identical structure to that of the encoder for generating the synthetic speech. However, its complexity is considerably lower, since the innovation sequence that minimised the perceptual error is transmitted to the decoder and it is the only sequence to which the synthesis filter’s response is computed.

Following the predictor optimization loop, the ‘remainder’ of the speech information is carried by the prediction residual, which is not modelled directly.
Instead, the best excitation for this short-term synthesis filter is determined by minimizing the weighted error between the input- and the synthetic speech over say 5 ms, yielding 4 optimisation intervals/20ms LPC-frame.

The quantised filter parameters and the vector quantised excitation are transmitted to the decoder, where the synthesised speech is generated by filtering the decoded excitation signal through the synthesis filter(s). Let us now consider the effects of choosing different parameters for the short-term synthesis filter.

The pred. gain and the SEGSNR of a CELP codec with the order $p$ of its synthesis filter are shown in Figure 20 using 20 ms frames and the LPC coeff. were unquantized. The excitation parameters for the codec were determined identically to our 7.1 kbits/s codec as described later, except no error weighting was used. We preferred an order of $p=10$ as a compromise between high prediction gain and a low bit-rate.

The bit-rate also depends on the frame length $L$. Hence we fixed $p = 10$ and the LPC coeffs. were calculated using Hamming windowed frames of $L$ samples. The results are shown in Figure 21.

The LPC coefficients are not suitable for quantization, because the frequency response of the synthesis filter is very sensitive to quantisation errors, which may lead to instability.

Some schemes use $k_i$, which are related to the lossless tube model of the vocal tract and are calculated as a by-product of using the LD algorithm to solve Equation ??? Using these coefficients the stability of the synthesis filter can be easily ensured by limiting the magnitude of all the coefficients to be less than one.

## 4 Long-Term Prediction

### 4.1 Open-loop Optimisation of LTP parameters
Most ABS codecs incorporate a long-term predictor (LTP) to improve speech quality and bit rate economy by further reducing the PE, which is achieved by predicting and removing the long-term redundancy of the speech.

While the STP removes the adjacent sample correlation and models the spectral envelope, i.e., the formant structure, it still leaves some long-term peaks in the STP residual, since at the on-set of quasi-periodic waveform segments of voiced sounds it fails to predict the signal adequately, as seen in Figure 22 for an 800-sample or 100 ms long speech segment.

The pitch-related, quasi-periodic PE peaks can be reduced by the LTP.

The operation of the LTP can be explained in a first approximation as subtracting a 'pitch-synchronously' positioned or delayed segment of the previous LPC residual from the current segment.

If the pitch periodicity is quasi-stationary, then the properly positioned previous segment will have co-located pitch pulses with the current segment. Hence after subtracting the previous LPC segment the pitch-synchronous prediction residual pulses of Figure 22b can be eliminated, as evidenced by Figure 22c portraying the LTP residual.

The performance of the above 'simplistic' LTP can be improved if, before subtracting the previous 'history', we scale the previous segment by a gain factor $G$, which can be optimised to minimise the energy of the LTP residual.

Both the LTP delay and gain will have to be transmitted to the decoder to reconstruct the LPC residual. Furthermore, since at the decoder only the previous reconstructed residual is available, which is based on the transmitted innovation sequence and LTP parameters, such as the delay and gain, the encoder also uses the previously reconstructed LPC residual segments, rather than the original ones.

Since periodic signals exhibit a line-spectrum, the quasi-periodic prediction residual’s spectrum seen in Figure?? has a periodic fine structure showing peaks and valleys, which is the manifestation of the time-domain long-term periodicity.

Hence the LTP models the spectral fine-structure of the speech that is similar to the line spectrum of a periodic signal. This pitch-related periodicity is strongly speaker- and gender-dependent: 100-300 Hz or about 3-10 ms.

The LTP residual is near-random, which is often modelled by a zero-mean, unit-variance random Gaussian codebook, yielding an extremely efficient vector quantiser. This concept leads to Code Excited Linear Predictive Coding (CELP), constituting the most prominent member of the family of ABS codecs.

The best innovation does not necessarily resemble the LTP residual, nor does it guarantee the best waveform match between the original speech and synthetic speech. It rather endeavours to produce the perceptually best speech quality.
• The LTP residual $e(n)$ is computed as:

$$e_L(n) = r(n) - Gr(n - \alpha)$$  \hspace{1cm} (51)

where $r(n)$ is the STP residual from which its delayed version $r(n - \alpha)$ is subtracted after scaling by an optimum gain factor $G$, computed by minimising the LTP residual error. The z-transform of Equation 51 is given by:

$$E_L(z) = R(z)[1 - Gz^{-\alpha}],$$

which can be re-arranged in the following form:

$$R(z) = \frac{E(z)}{[1 - Gz^{-\alpha}]} = \frac{E(z)}{P(z)},$$

where $P(z) = [1 - Gz^{-\alpha}]$ is the z-domain transfer function of the LTP.

• The total mean-squared LTP residual error $E$ computed over a segment of $N$ samples can be formulated as follows:

$$E = \sum_{n=0}^{N-1} e_L^2(n) = \sum_{n=0}^{N-1} [r(n) - Gr(n - \alpha)]^2$$

$$= \sum_{n=0}^{N-1} r^2(n) - 2Gr(n) - 2Gr^2(n - \alpha) + \sum_{n=0}^{N-1} G^2r^2(n - \alpha).$$

• Setting $\partial E / \partial G = 0$ gives

$$\sum_{n=0}^{N-1} -2r(n)r(n - \alpha) + \sum_{n=0}^{N-1} 2Gr^2(n - \alpha) = 0,$$  \hspace{1cm} (54)

yielding the optimum gain factor $G$ in the following form:

$$G = \frac{\sum_{n=0}^{N-1} r(n)r(n - \alpha)}{\sum_{n=0}^{N-1} [r(n - \alpha)]^2}.$$  \hspace{1cm} (55)

• Observe that the gain factor computed can be interpreted as the normalised cross correlation of $r(n)$, where the normalisation factor in the denominator represents the energy of the STP residual segment.
The effect of both the STP and LTP is shown in Figure refxwd1-lpc-ltp-pdf.

- When using a LTP, our ABS speech codec schematic seen in Figure 19 can be re-drawn as portrayed in Figure 25. The choice of the appropriate error weighting filter is crucial as regards the codec’s performance and its transfer function is based on findings derived from the theory of auditory masking.
- When generating for example a sinusoidal signal, this tone is capable of masking a highlevel, but spectrally more spread noise signal residing within the same frequency band. This is due to the inability of the human ear to resolve the two signals.
- Due to the speech signal’s spectral prominancies in the frequency regions of the formants this property can be exploited by allowing more quantisation noise to be concentrated around them. Clearly, an adaptive quantisation noise spectrum shaping filter is required, which de-weighs the quantisation noise in the formant regions, thereby allowing more noise in these frequency bands than in the spectral valleys.

Figure 24: Logarithmic PDF of a typical speech signal, LPC residual and LTP residual

Figure 25: Analysis-by-synthesis codec schematic with LTP

Figure 26: Modified analysis-by-synthesis codec schematic with LTP
The filter’s transfer function has to depend on the momentary signal spectrum, which is evaluated in the codec in terms of the filter coefficients $a_i$, describing $A(z)$. A convenient choice is to define it as:

$$\frac{W’(z)}{A(z/\gamma)} = \frac{1 - \sum_{k=1}^{p} a_k z^{-k}}{1 - \sum_{k=1}^{p} a_k \gamma^k z^{-k}}, \quad (57)$$

where the constant $\gamma$ determines, to which extent the error spectrum is de-emphasised in the formant regions.

- Typical values of $\gamma$ are in the range of $0.6 \ldots 0.85$. The schematic diagram of Figure 25 can also be re-arranged as in Figure 26.

- Recently other forms of error weighting have been suggested for speech codecs. For example in the ITU 16 kbits/s G.728 codec a filter

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \quad (58)$$

is used where $\gamma_1 = 0.9$ and $\gamma_2 = 0.6$.

### 4.2 Closed-loop Optimisation of LTP parameters

- Previously the LTP parameters were computed from the LPC residual using a simple correlation technique, as suggested by Equation 56, in a sub-optimum two-stage approach often referred to as open-loop optimisation.

- However, it was suggested by Singhal and Atal that a substantially improved speech quality can be attained at the cost of a higher complexity, if the LTP parameters are computed inside the ABS loop, which leads to the so-called adaptive codebook approach featured in the schematic of Figure 28.

- This is, because the adaptive codebook is replenished regularly using the previous composite excitation patterns $u(n)$ after a delay of one subsegment duration.

- The composite excitation signal $u(n)$ in Figure 28 is given by:

$$u(n) = v(n) + G u(n - \alpha) = v(n) + G u(n - \alpha), \quad (59)$$