7. Speech Coding
What is speech coding?

• Speech coding (or digital speech coding) is the process by which a speech signal can be temporally compressed into less bits/second and then decompressed, while preserving its most important content.

• The main objectives of digital speech coding are to **lower the bit rate** used to represent a speech signal while maintaining an adequate level of **perceptual fidelity**. In addition, for some applications we need to consider the **complexity** (computation required to encode/decode)

• Many other secondary objectives can also be defined (see slide)

• The most important uses of speech coding are for transmission (e.g. telephone line or cell network) and for storage (e.g. MP3)
Speech Coding objectives (additional objectives)

- High perceived quality (how well a human perceives the audio is)
- High measured intelligibility (The message is understood)
- Low bit rate (bits per second of speech)
- Low computational requirement (MIPS)
- Robustness to successive encode/decode cycles
- Robustness to transmission errors (e.g. intermittent cuts in the channel)

Objectives for real-time only

- Low coding/decoding delay (ms)
- Work with non-speech signals (e.g. touch tone)
Digital speech coding example: Digital telephone communication system

Diagram:
- **Transmitting Telephone**
  - A/D
  - Analysis
  - Quantize
  - Encode
    - Bit Pattern: 10110
      - Encrypt
        - 01011
          - Modulate (modem)
            - Analog Waveform
              - Telephone Channel
                - Analog Waveform

- **Receiving Telephone**
  - Demodulate
    - 01011
      - Unscramble
        - Decode
          - 10110
            - Quantized Parameters
              - Synthesis
                - D/A
                  - Analog Waveform

Diagram includes symbols for Analog Waveform, Transmitting Telephone, Receiving Telephone, A/D, Analysis, Quantize, Encode, Encrypt, Modulate, Telephone Channel, Demodulate, Unscramble, and Decode.
Minimum Bit Rate for speech

The bit rate corresponds to the information (number of bits) transmitted/compressed per unit of time.

In order to compute the minimum bitrate necessary to transmit speech we consider that the speech information transmitted is more or less equivalent to the sequence of phonemes uttered

- Considering we speak 10 phonemes / sec
- Consider we have from 30 to 50 phonemes for a language $32 = 2^5$ (can encode them in 5 bits)
- Minimum bit rate is $5 \times 10 = 50 \text{ bps}$! (lots of redundancy in the actual speech)
- To be compared with the 33.6 Kbps allowed on analog telephone line (and 64 Kbps on a digital line)
Signal-to-Noise Ratio (SNR)

- The SNR is one of the most common objective measures for evaluating the performance of a compression algorithm:

\[
SNR = 10 \log_{10} \left( \frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1} (s(n) - \hat{s}(n))^2} \right)
\]

where \( s(n) \) is the original speech data while \( \hat{s}(n) \) is the coded speech data. It therefore corresponds to the log-ratio between short-term power of signal and noise.
Mean opinion score (MOS)

- It is used to measure the perceived speech quality.
- The way to compute it is defined through ITU-T Recommendations.
- Has the following grades:
  - Excellent – 5
  - Good – 4
  - Fair – 3
  - Poor – 2
  - Bad – 1
- A minimum of 30 people has to grade speech by listening to voice samples or in conversations.
Classification of coding algorithms (I)

There are several possible classifications:

1. Based on how the speech production and perception models are incorporated (Used in the recommended article):
   - Open-loop codecs (also called vocoders): Extract an excitation signal + a vocal tract system to later resynthesize the speech, without any attempt to preserve the waveform shape. These more or less correspond to the Model-based codecs.
   - Closed-loop codecs: Apply the source/system model of speech within a feedback loop to iteratively reduce the difference between the original and quantized signals.

Closed-loop example: analysis-by-synthesis codec
There are several possible classifications:

2. Based on how the waveform of the speech signal is preserved with the process:

   • **Waveform codecs**: aim at reproducing the speech waveform as faithfully as possible. No previous knowledge is applied to the mechanisms that might have created the audio (for this reason it is not limited to speech). Has higher bit rates but high quality and intelligibility.

   • **Model-based codecs** (also called source-codecs or vocoders): preserve only the spectral properties of speech in the encoded signal. They transmit parameters obtained by modeling speech with a sound production model. Achieves lower bit rates but very synthetic.

   • **Hybrid codecs**: bring the best of both (i.e. bring certain waveform matching while using the sound production model). These are also called Analysis by synthesis codecs.
Reduced Bit-rate Speech Coding Schemes

Phonetics
Vocoders

LPC Homomorphic
Channel Formant Phase

Hybrid Coders

APC RELP MPLPC CELP SELP
SBC ATC Sinusoidal Harmonic MBE

PCM DM APCM DPCM ADPCM

Subjective Quality (MOS)

Hybrid Coders
Waveform Coders
Vocoders

Bit Rate (kb/s)
3. Based on the nature of the encoded parameters:
   - **Time-domain coding**: speech samples are directly encoded
   - **Parametric coding**: Acoustic parameters are first derived from the speech samples. These are usually estimated through linear prediction.
   - **Transform coding**: They also send parameters extracted from the signal. In this case they exploit the redundancy of the signal in the transform domain (Fourier, sinusoid, etc...)
Index for the remainder of the session

- Quantization
  - Scalar quantization
    - Uniform
    - Non-uniform
  - Vector quantization

- Time-domain coding
  - DPCM
  - ADPCM

- Parametric coding
  - LPC
  - Hybrid coding
  - MP-LPC, CELP

- Transform coding
  - Sub-band coders
  - Sinusoidal coding
  - MPEG-1 layer 3
Quantization

- Quantization algorithms are at the core of any bigger coding algorithms. They are a clear example of waveform coding techniques, although used in isolation are not very good.
- Scalar quantizers classify each speech sample into one of several levels, represented by B bits.
- Vector quantizers classify several speech samples together into one of several clusters/codebook vectors, each one also represented by a B bits code.
Scalar Quantization: Uniform Quantization

The decision and reconstruction levels are uniformly spaced.

Uniform Pulse-Code Modulation (PCM): Quantizes amplitudes by rounding off each samples to one of a set of discrete values.

It is the simplest waveform coding algorithm we will see.

It is only optimum for signals with a uniform PDF along the peak-to-peak range $2X_m$. 

![Diagram of 3-bit quantizer](image)
Sources of error in quantization

We can derive 2 sources of error:
• Errors within the dynamic range (see next slide)
• Errors above the dynamic range
  • They are correlated with the signal and need to be avoided by defining well the peak-to-peak range
Sources of error in quantization (II)

For samples within the range of the quantizer, the error/noise is defined as

\[ e[n] = \hat{x}[n] - x[n] \]

Which has to satisfy

\[ -\frac{\Delta}{2} < e[n] < \frac{\Delta}{2} \]

Where, if \( 2X_m \) is the peak-to-peak range of the signal (no clipping occurs) and \( B \) the number of bits of the quantizer:

\[ \Delta = \frac{2X_m}{2^B} \]

We can approximate the SNR of the quantizer as

\[ SNR_Q = 6.02B + 4.78 - 20\log_{10}\left(\frac{X_m}{\sigma_x}\right) \]

Where \( \sigma_x \) is the RMS power of the input signal. We therefore gain 6dB for each extra bit we use.
Scalar Quantization: Nonuniform Quantization
Reconstruction and decision levels do not have equal spacing.

Low level signals are more common than high level ones, thus we make quantization steps closer together at the commoner signal levels.

The log quantizers are called m-law (US) and A-law (Europe). A 7-bit log quantizer for speech achieves the performance of a 12-bit uniform quantizer.

They are usually called log-PCM coding algorithms and have been used for many years by telephone companies.
Non-uniform quantization

**µ - law**

\[
y[n] = X_{\max} \frac{\log \left[ 1 + \mu \frac{|x[n]|}{X_{\max}} \right]}{\log[1 + \mu]} \text{sign}\{x[n]\}
\]

**A - law**

\[
y[n] = X_{\max} \frac{1 + \log \left( \frac{A|x[n]|}{X_{\max}} \right)}{1 + \log A} \text{sign}\{x[n]\}
\]

- Usually \(A=87.7\), and \(\mu=255\) in the standards, and implemented as filters + uniform quantizer.
- The transformation has the effect of distributing samples linearly in the lower levels and logarithmically in the higher levels.
Vector Quantization (VQ)

- A generalization of scalar quantization in which a block of scalars (e.g., input speech samples) are coded together, rather than individually.

- The M-dimensional regions delimiting each cluster/codebook vector need to be set a priori or trained using, for example, k-means alg.

- Vector quantization is essential in achieving low bit rates in model-based coders.
K-Means Clustering

1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2) Assign each object to the group that has the closest centroid.

3) When all objects have been assigned, recalculate the positions of the K centroids.

4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.
K-means clustering
Lineal predictive coding

Using the LPC theory we saw earlier, we can estimate the current speech value from the previous “p” values

\[ \hat{x}[n] = \sum_{k=1}^{p} \alpha_k \hat{x}[n - k], \]

We can then take the difference between the real value (known) and the estimate:

\[ d[n] = x[n] - \hat{x}[n] \]

If the prediction is good \((G_p > 1)\) by quantizing \(d[n]\) we can reduce the quantization error

\[ SNR = G_p \cdot SNR_q \]

This leads us to a family of methods that use differential coding that are usually called differential PCM (DPCM)
In the coder we eliminate the predictable information from the signal, which is later reincorporated in the decoder. Given that the speech signal changes its characteristics over time, the predictor parameters will need to be recomputed and therefore sent together with the residual.
Time-Domain Coding: Delta Modulation (DM)

- It is the simplest of the DPCM codecs
- It normally uses only 1-bit quantizer and a first-order predictor
- It is VERY simple to implement and VERY computationally efficient
- For it to achieve good SNR we need a high bit rate (over the nyquist frequency)
- Only the quantized output needs to be sent

![Diagram of Delta Modulation](image-url)
Time-Domain Coding: Delta Modulation (DM)

- We can get lots of error when:
  - The signal moves fast and our $\Delta$ is too small
  - The signal moves slowly and our $\Delta$ is too big

When the coding freq. is lot high enough we might not be able to follow the signal.
Time-domain Coding: Adaptive differential PCM

• We can improve DM by automatically adjusting the step size according to the signal level:
  • When there are big changes on the signal we decrease the step size
  • When changes are small we increase the step size
• In general, this corresponds to the systems called Adaptive Differential PCM (ADPCM), which fall within the waveform codecs type
• ADPCM usually operates on short blocks of input samples, therefore introducing some delay.
• We can classify the ADPCM systems between feed-forward and feed-backwards (most used)
• Both the prediction residual and the predictor parameters need to be quantized before sending them
  • VQ or non-uniform quantization are the most common techniques
• The best ADPCM codecs can go as low as 10kbps with acceptable quality
ADPCM systems

- Feed-forward ADPCM

- Feed-backwards ADPCM
Model-based/Parametric codecs

- These codecs usually take into account the source-filter model of speech production to encode speech with acceptable quality at lower bit rates than ADPCM.
- They are usually referred to as VOCODERS. They are open-loop codecs, as they do not have a feedback loop.
- When they use a speech production model, these algorithms are very good for speech, but create distortions and lack of quality when encoding other signals (e.g. music).
- By sending only the **source/excitation type** and the **filter parameters**, vocoders can achieve very low bitrates.
- With respect to the estimation of the vocal tract parameters, we can classify the vocoders into:
  - Homomorphic: Cepstral parameters are extracted and encoded.
  - LPC: Lineal predictive analysis is used to obtain vocal tract parameters.
Parametric Coding: VQ LPC Coder

Transmitter

$s[n]$ → LPC Analysis → $k$ → Codebook $\hat{k}$ → VQ → Encoder → Bit Pattern

Pitch Voicing

Receiver

Bit Pattern → Decoder → Codebook $\hat{k}$ → Lattice → $\hat{s}[n]$

Pitch/Voicing
Hybrid/analysis by synthesis systems

- These systems incorporate some speech-production model (like the source-filter model) to obtain optimum parameters, that are then sent together with a synthetic excitation signal.
- Such excitation needs to be efficient to code and produce high quality output speech.
- In these the signal is also processed in blocks, creating some delay.
- They are called analysis-by-synthesis because a synthetic vocal tract filter is built iteratively to match the input speech for every given block. Also the excitation is generated from a fixed collection of input components

\[
\hat{x}[n] = x[n] - d[n]
\]

\[
\hat{d}[n] = \text{Excitation Generator}
\]

\[
\text{Vocal Tract Filter } H(z)
\]

\[
\text{Perceptual Weighting } W(z)
\]

Fig. 7.8 Structure of analysis-by-synthesis speech coders.

Note the similarity of Figure 7.8 to the core ADPCM diagram of Figure 7.5. The perceptual weighting and excitation generator inside the dotted box play the role played by the quantizer in ADPCM, where an adaptive quantization algorithm operates on \(d[n]\) to produce a quantized difference signal \(\hat{d}[n]\), which is the input to the vocal tract system. In ADPCM, the vocal tract model is in the same position in the closed-loop system, but instead of the synthetic output \(\hat{x}[n]\), a signal \(\tilde{x}[n]\) predicted from \(\hat{x}[n]\) is subtracted from the input to form the difference signal. This is a key difference. In ADPCM, the synthetic output differs from the input \(x[n]\) by the quantization error. In analysis-by-synthesis, \(\hat{x}[n] = x[n] - d[n]\), i.e., the reconstruction error is \(-d[n]\), and a perceptually weighted version of that error is minimized in the mean-squared sense by the selection of the excitation \(\hat{d}[n]\).
**Perceptual weighting filter**

- The quantization error does not affect all frequencies in the same way.
- In analysis-by-synthesis systems a perceptual weighting filter changes the frequency shape of the filter residual to modify such error.
- It is equivalent to the preemphasis filtering in ADPCM.
- The perceptual weighting filter \( W() \) is designed by inverting the vocal tract filter \( H() \).

\[
W(z) = \frac{A(z/\alpha_1)}{A(z/\alpha_2)} = \frac{H(z/\alpha_2)}{H(z/\alpha_1)}.
\]  

(7.14)

The poles of \( W(z) \) lie at the same angle but at \( \alpha_2 \) times the radii of the poles of \( H(z) \), and the zeros of \( W(z) \) are at the same angles but at \( \alpha_1 \) times the radii of the poles of \( H(z) \). If \( \alpha_1 > \alpha_2 \) the frequency response is like a "controlled" inverse filter for \( H(z) \), which is the shape desired. Figure 7.9 shows the frequency response of such a filter, where typical values of \( \alpha_1 = 0.9 \) and \( \alpha_2 = 0.4 \) are used in (7.14). Clearly, this filter tends to emphasize the high frequencies (where the vocal tract filter gain is low) and it deemphasizes the low frequencies in the error signal. Thus, it follows that the error will be distributed in frequency so that relatively more error occurs at low frequencies, where, in this case, such errors would be masked by the high amplitude low frequencies. By varying \( \alpha_1 \) and \( \alpha_2 \) in (7.14) the relative distribution of error can be adjusted.
Excitation signal

- A finite set of signals $f_{\gamma k}$ (known at source and destination) can be used to generate the signal as

$$\hat{d}[n] = \sum_{k=1}^{N} \beta_k f_{\gamma k}[n].$$

Where $N$ is the number of signals used to generate the impulse

- The selection of the optimum signals and $\beta_k$ is found to optimize the error

$$E = \sum_{n=0}^{L-1} ((x[n] - h[n] * \hat{d}[n]) * w[n])^2,$$

Where $L$ is the number of samples per block

- The excitation signals are defined differently in each algorithm
Hybrid Coding: Multi-Pulse Linear Prediction (MPLP)

It was the first codec using analysis by synthesis methods. The excitation waveform is represented by more than one impulse, as

$$\hat{d}[n] = \sum_{K=0}^{M-1} A_k \delta[n - n_k]$$

Usually N=4-5 and L=5ms (40 samples at 8KHz)
Usually at 10-16kbps it achieves acceptable quality
Hybrid Coding: Multi-Pulse Linear Prediction

Original Speech

Excitation

Synthetic Speech

Error

Time

Number of pulses used

1 2 3 4 5

TDP: Speech Coding
Code-Excited Linear Prediction (CELP)

- Represents the excitation (residual) from the linear prediction on each frame by codewords from a VQ-generated codebook, rather than multipulse.
- Usually 256-1024 sized codebooks (8-10 bits) are used
- On each frame a codeword is chosen from a codebook of residuals such as to minimize the mean-squared error between the synthesized and original speech waveforms.
- A codebook can be formed by applying a k-means clustering algorithm to a large set of residuals training vectors.
- CELP can achieve acceptable quality at 4,8kbps
- As the search for the optimum sequence is exhaustive it is very computationally intensive.
Quality versus data rate of coders
Transform-based coding

• These codecs apply a transformation of the signal to the frequency domain and encode the different bands independently.

• It takes advantage of the fact that different frequencies have different perception characteristics by human in order to reduce bitrate
Transform Coding: Sub-Band Coders

Sub-band and transform coders exploit the redundancy of the signal in the transform domain.

AT&T sub-band coder and decoder
Transform Coding: Sinusoidal Transform Coders (STC)
Transform Coding: MPEG-1

Also known as MPEG2 audio layer 3, or MP3
A transparently lossy audio compression system based on the weaknesses of the human ear.
Can provide compression by a factor of 6 and retain sound quality.
Quantification and coding is done every 8ms
MPEG-1 Audio Features

- PCM sampling rate of 32, 44.1, or 48 kHz
- Four modes:
  - Monophonic and Dual-monophonic
  - Stereo and Joint-stereo
- Three modes (layers in MPEG-I speak):
  - Layer I: Computationally cheapest, bit rates 384Kbps in stereo (used in DCC)
  - Layer II: Bit rate ~ 192 kbps in stereo (used in VCD)
  - Layer III: Most complicated encoding/decoding, bit rates ~ 128kbps in stereo, originally intended for streaming audio (mp3)
Filter-bank processing

- The FFT is computed every 512/1024 samples (delay of 5.33/10.66ms @48KHz)
- 32 bands are then obtained uniformly spaced every 750Hz
- A polyphase filterbank is used: efficient and low delay
Data input

- In MPEG-1 layer I blocks of 512 samples are used in the analysis, shifting them 32 samples per step.
- The encoded output packets cover 12 consecutive shifts/blocks
MPEG 1: Psychoacoustic Model

- We use a psychoacoustic model to determine which frequencies I need to allocate more bits to.
- We first use a Hanning weighting window on the signal block, and then a DFT to convert to frequency domain.
- The block size is either 512 (layer I) or 1024 (layer II and III) samples.
- Example:

  lowpass noise + 11.250 kHz Sinewave
  sampling rate = 48 kHz

[Graph showing frequency analysis with peaks at 11.250 kHz and lowpass noise]
MPEG 1: Psychoacoustic Model

- In MPEG1 there are 2 possible psychoacoustic models applied:
  - Model 1: Less computationally intensive. It applies some compromises in what it assumes a user won’t hear. Determines the audible threshold using a masking function. Uses 512 samples (layer I) or 1024 samples (Layers II and III).
  - Model 2: Has more features, and more computational requirements. Uses a spreading function to determine the masking threshold. Always 1024 samples per window.
In the FFT, among other processing, a critical band masking threshold is computed for each critical band frequency. The total masking due to the critical bands is the sum of all contributions.

The total masking due to the critical bands is the sum of all contributions.
MPEG 1: Psychoacoustic Model

In addition, the minimum auditory threshold is considered to obtain a total masking threshold. Anything below that threshold is not heard, therefore, does not harm our perceived quality. We calculate a masking threshold for each subband in the polyphase filter bank.
MPEG 1: Psychoacoustic Model

In order to use the total masking threshold we compute the signal-to-mask ratio (SMR) per each of the 32 subbands

\[ \text{SMR} = \frac{\text{signal energy}}{\text{masking threshold}} \]

Those bands with negative SMR can be encoded with less bits as the generated noise will not affect our perceived quality.
MPEG 1: Psychoacoustic Model

What we actually send:
MPEG 1: Layer I Specific Coding

The stream of samples is processed in groups: We group 12 samples from each subband and encode them in each frame (= 384 samples in Layer I). In layer II and III the same is done for 3 groups of 12 samples (3x12x32=1152 samples).

Each group is encoded with 0-15 bits/sample.

Each group has an additional 6-bit scale factor.

Note: Each subband filter produces 1 sample out for every 32 samples in.