Voice coding basics

Andrea Bianco
Telecommunication Network Group
firstname.lastname@polito.it
http://www.telematica.polito.it/

Analog to digital conversion

- Human voice is an analog signal, i.e., a continuous-time signal assuming infinite possible values within a bounded interval
- To make transmission, elaboration and storage easier, it is normally transformed in a digital signal
  - Sampling and quantization

Sampling

- Sampling means observing the value assumed by the signal in given time instant
- A signal with bandwidth B can be reconstructed correctly if sampled at regular intervals with sampling frequency $f_s \geq 2B$ (Shannon theorem)
- Signal bandwidth may be limited by a low-pass filter
- The sampled signal is a discrete-time signal, but it may assume infinite possible values within a bounded interval
  - Often named PAM signal

Pulse Amplitude Modulation (PAM)

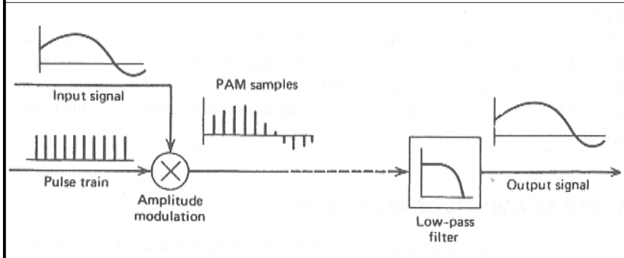


Figure 3.1 Pulse amplitude modulation

Quantization

- To correctly represent the sampled values, infinite precision numbers would be needed
- To avoid this, a quantization process is applied
  - All values assumed by the sampled signal within an *interval* of amplitude Δ are represented by a single value, normally the intermediate value of the interval, named *level*
- If all intervals have the same amplitude, the quantization is said to be uniform
- A quantizer operating on 2^N levels, represents each sampled value with an N bit string

Quantization noise

- The quantization process introduces an error, named quantization noise or error

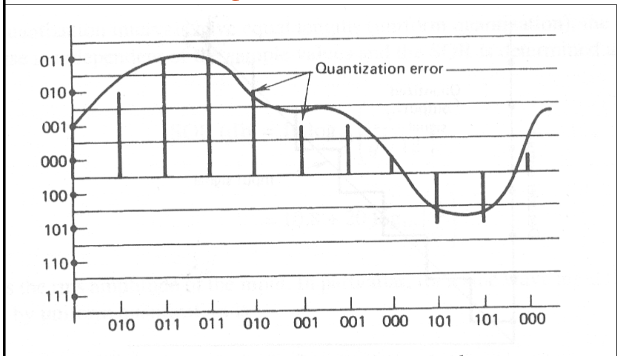
$$\hat{x} = x + e$$

- The SNR expressed in dB is proportional to the number of used bits:

$$SNR_{dB} = 10\log_{10}\frac{\sigma_x^2}{\sigma_e^2} \propto 6N$$

- For each additional bit, the SNR improves by 6 dB

Quantization



Pulse Code Modulation (PCM)

- The analog to digital conversion obtained through a sampling process with fixed frequency f_s, with a uniform quantization over N levels is named Pulse Code Modulation (PCM)
- The output of the PCM is a data flow of (N f_s) bit/s which can be further coded, compressed, etc
- Examples:
  - audio signal with CD quality
    - Bandwidth ~20 KHz ➔ f_s=44100 Hz
    - Quantization over 65536 levels ➔ N=16
    - ~700 Kb/s

Pulse Code Modulation (PCM)

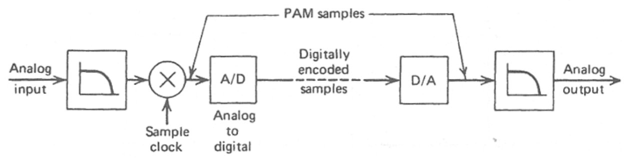
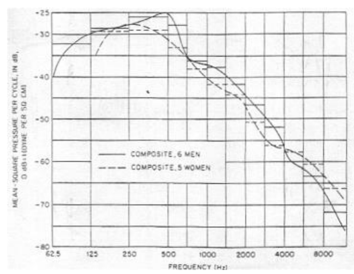


Figure 3.7 Pulse code modulation

Human voice signal characteristics

- Bandwidth ~ 4KHz, Dynamic ~ 60 dB



Coding quality

- The coded voice signal should be understandable and if possible “natural”
- Some objective measures to evaluate the quality of a coding system can be defined
  - However they not necessarily directly relate to the listener perceived quality
  - Often subjective metrics are used
- The more popular metric is the subjective MOS (Mean Opinion Score)
  - A group of listeners listens to a short coded spoken sentence and grades the subjective (perceived) quality on a scale ranging from 1 (Bad) to 5 (Excellent).

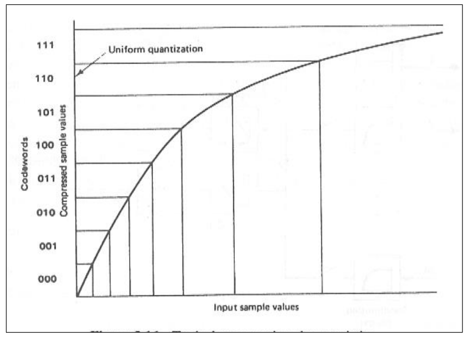
Voice coding through PCM

- The human voice can be coded with a good subjective quality using a PCM scheme with f_s=8kHz and N=12
- The resulting rate of 96 kbit/s is considered too high
- To reduce the rate, several variations of the basic PCM coding scheme were introduced
  - They try to exploit some peculiar characteristics of the human voice

Logarithmic (or companded) PCM

- The power of the voice signal is not uniformly distributed over the signal dynamics
  - Values around zero are more likely
- A logarithmic quantization is more efficient than a linear quantization
- To obtain the same SNR less levels can be used, or with the same number of levels a higher SNR can be obtained
- Logarithmic PCM (ITU G.711)
  - f_s=8 kHz, N=8 bit ➔ rate = 64 kbit/s

Companding



Differential PCM (DPCM)

- Consecutive samples of the voice signal are strongly correlated
- The dynamics of the *difference* among two consecutive samples is largely smaller than the dynamics of the signal itself
- By coding the differences, a smaller number of levels can be used without loosing in quality:

$$d[n] = x[n] - \alpha x[n-1] \qquad \alpha=0.9 \text{ (optimal)}$$

α is related to the correlation coefficient

Differential PCM

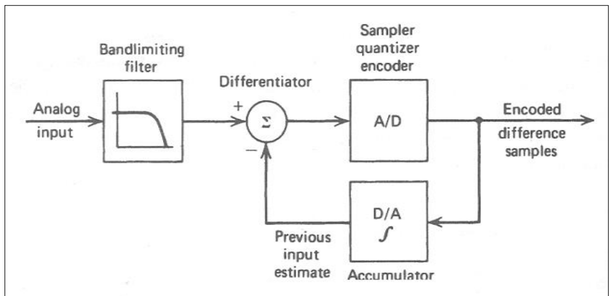


Figure 3.27 Functional block diagram of differential PCM.

DPCM

- More complex codecs consider a larger number of previous samples and code the sample through a weighted difference of previous samples:

$$d[n] = x[n] - \sum_{i=1}^{N} \alpha_i x[n-i]$$

- Weights α_i are fixed and computed so as to minimize σ_d^2, the difference dynamic

Adaptive PCM (APCM)

- The voice signal is non-stationary
- The quantizer can exploit this characteristic by adapting the energy levels of the signal
- Signal portions with lower energy levels are quantized with a finer granularity
- By fixing the number of used bits, the coding quality is improved
- More information (overhead) is needed at the receiver to correctly reconstruct the original signal

APCM Feed-Forward

- The quantizer estimates the local energy of the signal and it computes the current min-max values of the dynamic
- A traditional PCM coding is then used
- The min-max values are transferred to the receiver using reserved bits
- The energy estimate is recomputed typically every 20 ms (160 samples)

APCM Feed-Back

- The local energy estimate is based on a window of already coded samples
- The receiver has already received all the information needed to compute the local min-max values of the signal dynamic
- No transmission overhead is required
- The estimate can be recomputed more frequently, without increasing the coding rate

Adaptive Differential PCM (ADPCM)

- It combines the APCM and DPCM techniques,
  - A weighted difference between the current sample and previous samples is used, but weights α_i are time-variable
- Weights are independently computed by the transmitter and by the receiver so as to locally minimize the difference dynamic
- Standardized as ITU G.726
  - 4 bit per samples
  - Rate of 32 kbit/s

VOCODER

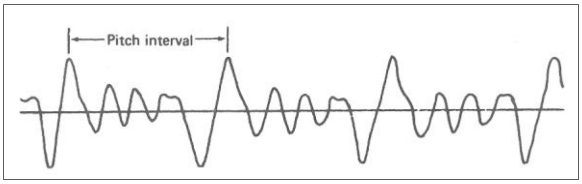
- With codecs using modifications of the basic PCM scheme, it is almost impossible to obtain rates smaller than 32 Kb/s
- For same applications (e.g., the GSM standard) this bit rate is too high
- An alternative approach is based on the idea of developing a production and perception model of the human voice signal
- At the receiver, the voice signal is reproduced synthetically on the basis of the values assumed by the model parameters, computed at the source and transmitted to the receiver in the voice stream

Model parameters and elements

- Lung:
  - generate an unstructured signal, a pressure signal, similar to a white noise. The air pressure is characterized by a **gain**.
- Vocal chords:
  - can be opened or closed
    - when opened (unvoiced signals) the pressure signal is not modulated
    - when closed (voiced signals) they create a periodic oscillation with a fundamental frequency (**pitch**)
  - a **flag** is used to represent this information

Waveform of Voiced/Unvoiced Sound

Figure 3.24 Time waveform of unvoiced sound.



Model parameters and elements

- Vocal tract (cavity):
  - Can be modeled as a time variable filter
  - The filter transfer function depends on the pronounced phonema
    - Each phonema is characterized by a spectrum with a set of peaks at different frequencies
    - These frequencies are named formants
    - Being difficult to extract the formant frequencies, normally the spectrum envelope is directly computed as a set of coefficients α_i (turn out to be the same coefficients of the ADPCM scheme).

LPC (Linear Prediction Code) VOCODER

- The human voice signal is analyzed over time windows lasting 20ms
- Over each time window, the extracted parameters are coded according to the following scheme:
  - GAIN ➔ 5 bit
  - FLAG (vocal chord open/close) ➔ 1 bit
  - PITCH ➔ 6 bit
  - 10 coefficients α_i ➔ 10*4 bit
  - Total: 52 bit
- At the receiver, the filter (representing the phonema) is reconstructed through the 10 α_i coefficients and excited either with white noise or with a periodic signal with frequency equal to the pitch frequency
- Resulting rate: 2.6 Kbit/s

Speech generation in LPC

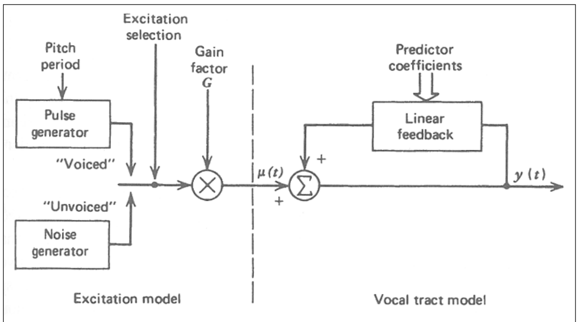
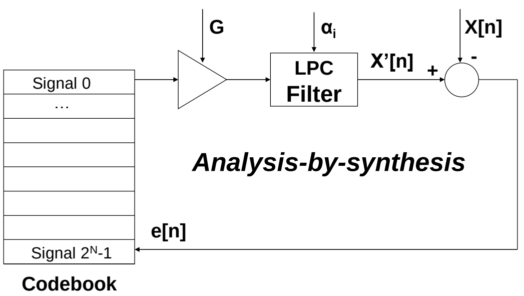


Figure 3.41 Speech generation model of linear predictive coding.

Code Excited LP (CELP)

- The human voice coded according to the LPC vocoder is understandable but its perceptive quality is unsatisfactory
- Main limitation is that in 20ms the voice signal is not either voiced or unvoiced
- To improve the quality, the CELP codec uses an improved exciting signal
- The exciting signal is chosen from a pre-defined library, named codebook, containing many different signals
- The transmitter uses a brute-force approach to select the exciting signal that minimizes the difference between the input signal and the synthetized signal
- A typical codebook size is 1024 signals

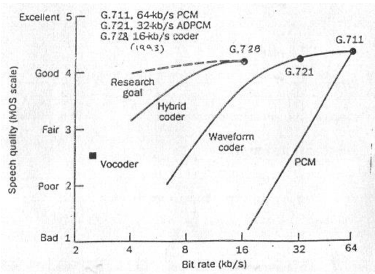
Scheme of a CELP codec



Comparison between PCM and VOCODER

- PCM:
  - Flexible
    - can be used for different types of signals
  - Robust
    - very insensitive to noise
  - High quality
  - Easy to implement
  - Low delays
  - High rates (at least 32Kb/s)
- VOCODER
  - Very low rates (as low as 1.6 Kb/s)
  - Low-medium quality
  - Can be used only for human voice coding
  - Sensitive to noise

Coding quality



Andrea Bianco – TNG group - Politecnico di Torino

Computer Network Design and Management- 31