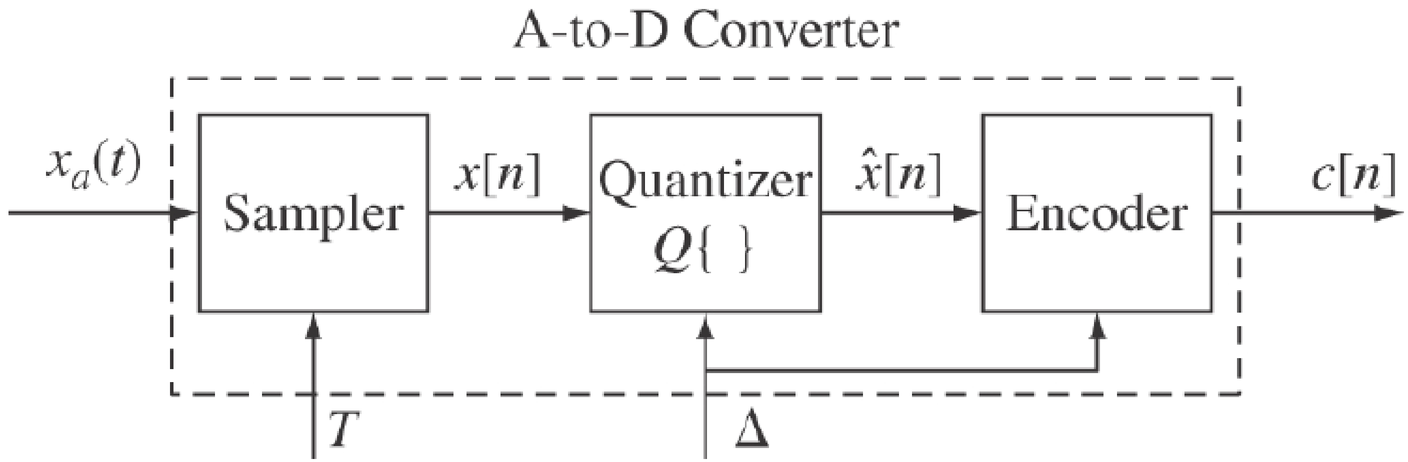
Lecture 14
Chapter 11. Digital Coding of Speech Signals

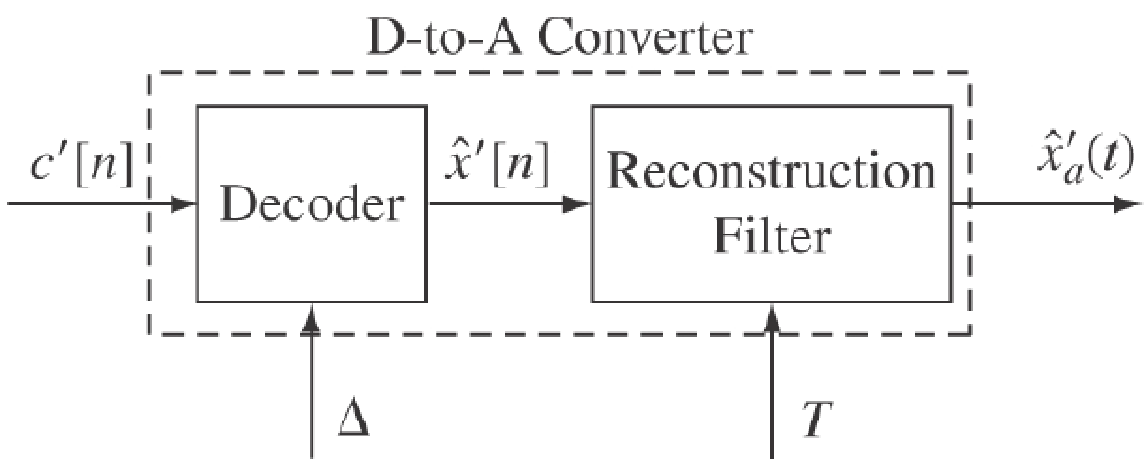
DEEE725 Speech Signal Processing Lab

Gil-Jin Jang

Original slides from Lawrence Rabiner

Analog-to-Digital Conversion: CODEC (enCOder and DECoder)





History of Speech Coding

- 1926 – pulse code modulation (PCM); first conceived in 1937.
- 1952 – delta modulation proposed, differential PCM (DPCM) invented.
- 1957 – A-law and μ-law encoding proposed; standardized for telephone network in 1972 (G.711)
- 1974 – ADPCM developed
- 1984 – CELP (code-excited linear prediction) vocoder proposed; majority of coding standards for speech signal today use a variation on CELP

Type of Speech Codecs

- Waveform codecs
  - Encode waveform directly
  - High-quality and not complex
  - Large amount of bandwidth

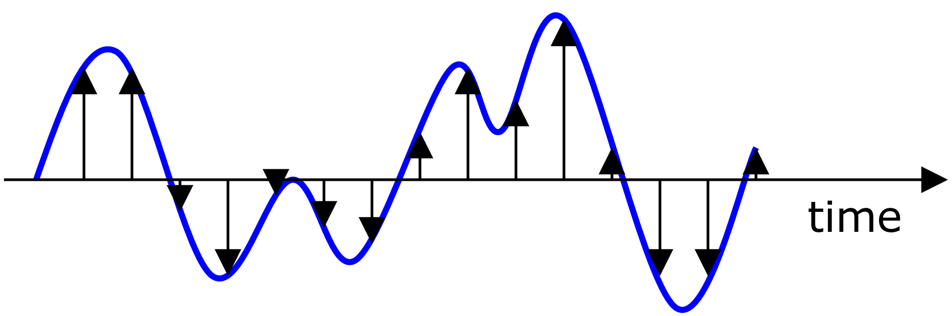
- Source codecs (voice coders – **<u>vocoders</u>**)
  - Match incoming signal to a math/physiological model
  - Linear-predictive filter model of the vocal tract
  - A voiced/unvoiced flag for the excitation
  - The information is sent rather than the signal
  - Low bit rates, but sounds synthetic
  - Higher bit rates do not improve much

Waveform Codecs

- PCM (pulse code modulation)
  - Sample input waveform directly
  - Uniform quantization requires > 10 bits/sample.
    - typically 16 bits/sample
    - 16 bits / sample x 8000 samples / second = **128 kbit/s**.
- DPCM (differential PCM)
  - Encode difference between consecutive samples
- Adaptive DPCM
  - Adapt step size for quantization based on speech statistics
  - Example: G.726 (1974); based on six previous differences.
  - Gain-adaptive 15 quantization levels results in **32 kbit/s**.

Voice Sampling

- A-to-D
  - discretize the analog waveform by some number of bits
  - A signal can be reconstructed if it is sampled at a minimum of twice the maximum frequency (Nyquist Theorem)
- Human speech
  - Typical bandwidth in 300-3800 Hz
  - 8000 samples per second (8 kHz sampling rate)



Each sample is encoded into an 16-bit PCM code word (e.g. 0011011001100101)
➔ 8000 x 16 bit/s = 128kbps

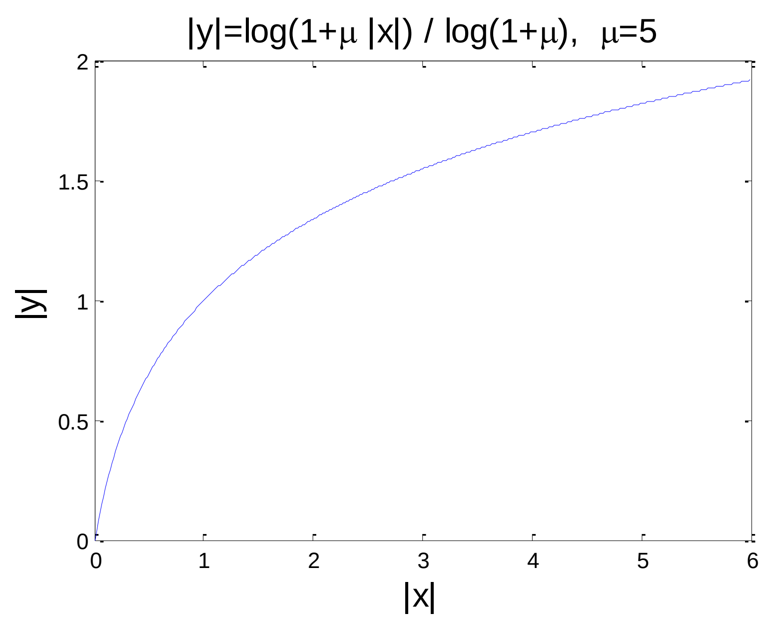
Quantization

- How many bits is used to represent
- Quantization noise
  - The difference between the actual level of the input analog signal
- Uniform quantization levels
  - Louder talkers sound better

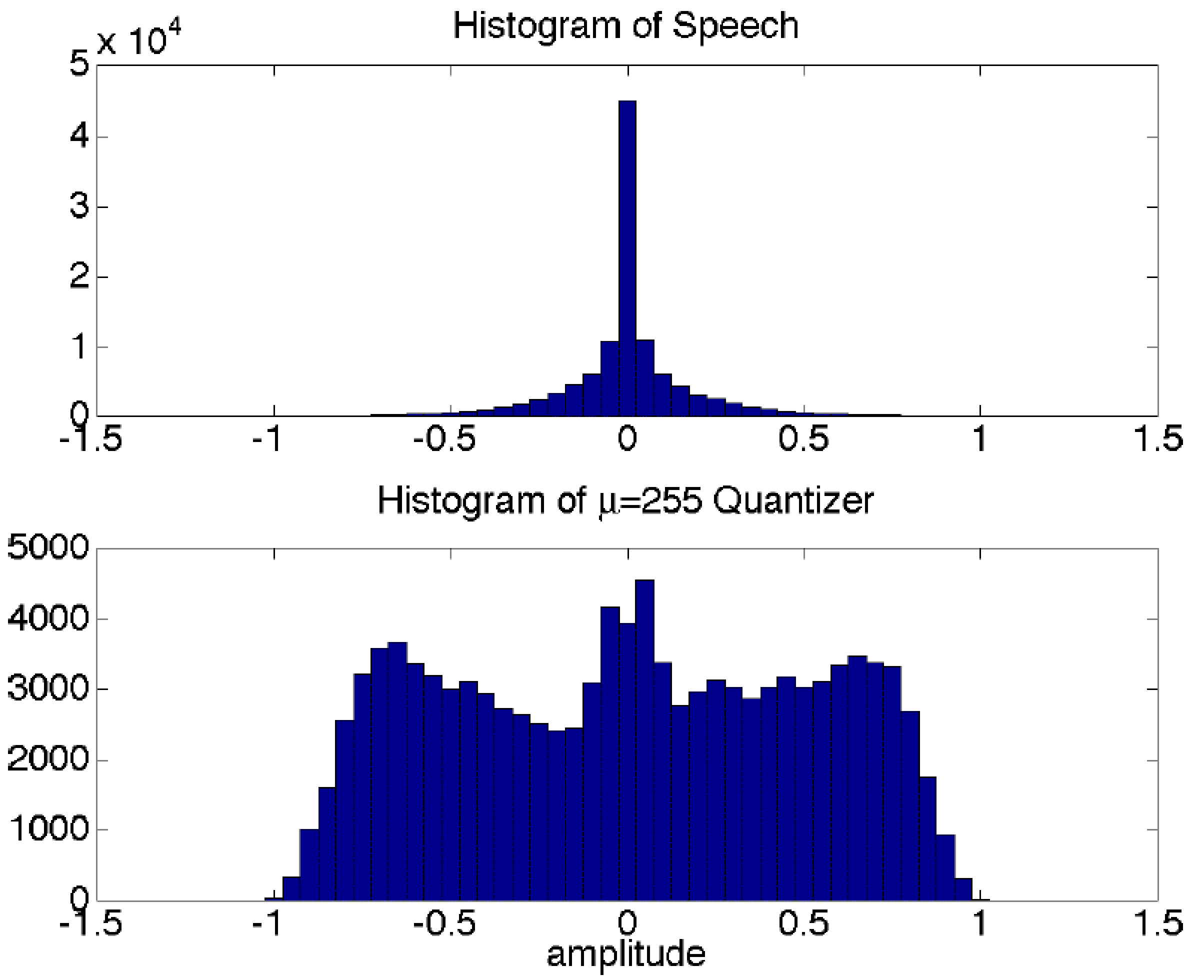
Non-uniform Quantization

- Smaller quantization steps at smaller signal levels to spread signal-to-noise ratio more evenly
- **Logarithmic scaling** (A-law in Europe and μ-law in US)

Non-uniform quantization
(G.711, μ-law & a-law, 1972)
Quantizing y to 8 bits yields
64 kbit/s at 8kHz



Histogram for mu-Law



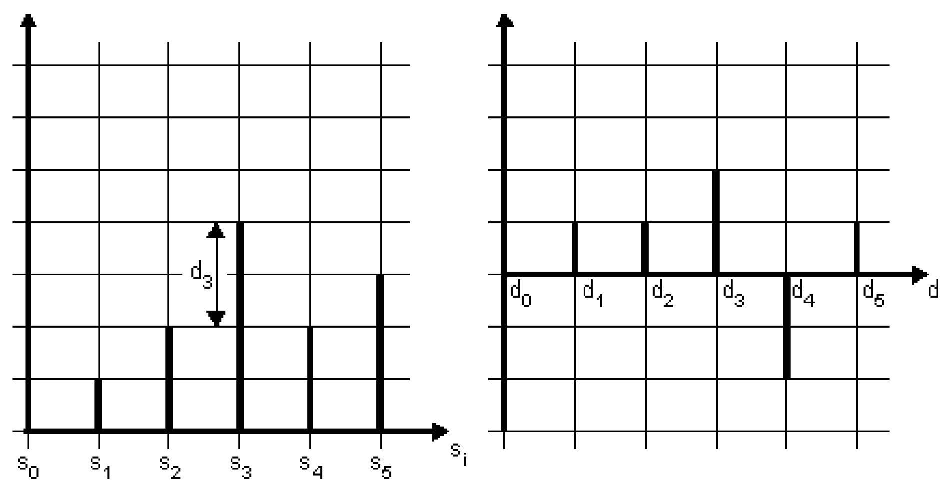
Speech waveform

Output of μ-Law compander

Becomes closer to uniform distribution

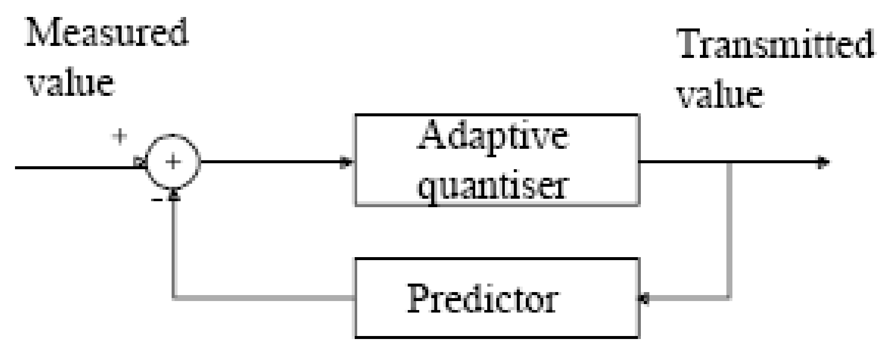
DPCM

- DPCM, Differential PCM
  – Only transmit the difference between the predicated value and the actual value
  – The receiver perform the same prediction
- No algorithmic delay



ADPCM (Adaptive DPCM)

- Predicts sample values based on
  - Past samples
  - Factoring in some knowledge of how speech varies over time
- The error is quantized and transmitted
  - Fewer bits are required



- G.721
  - 32 kbps
- G.726
  - A-law/mu-law PCM -> 16, 24, 32, 40 kbps
  - An MOS of about 4.0 at 32 kbps

Codec quality assessment

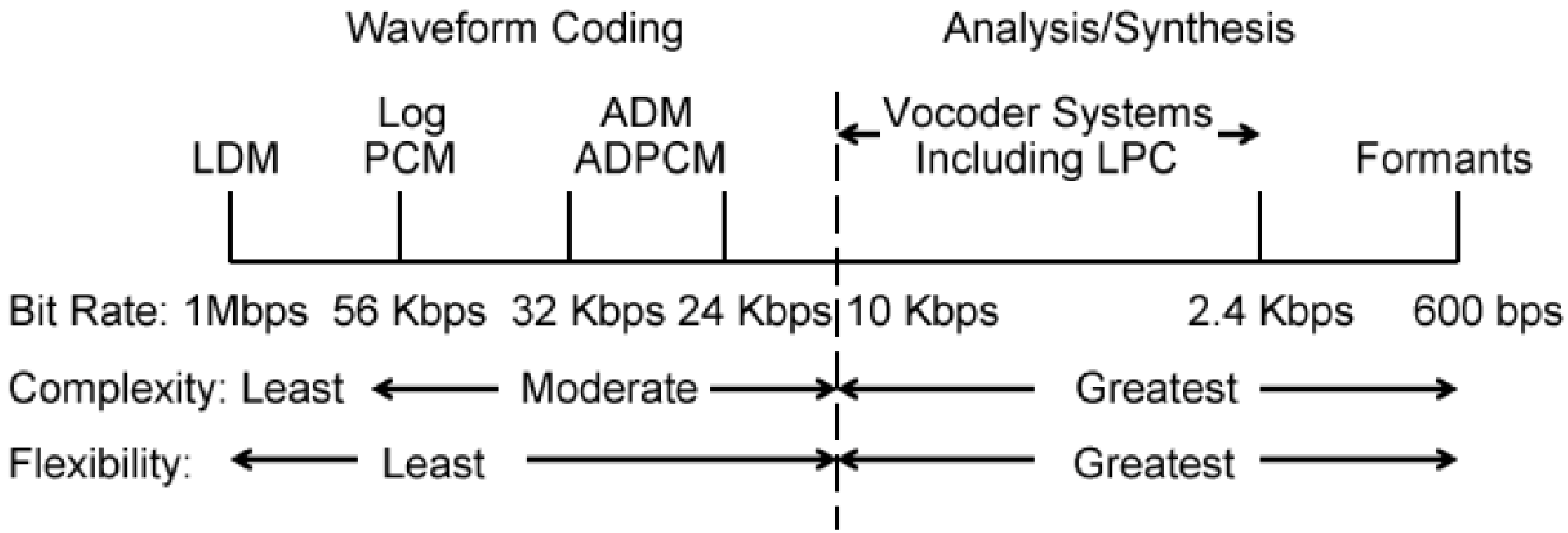
- Subjective evaluation (human listening test)
  - Preference test – “Which one is the best?”, usually binary
  - MOS (mean opinion score, 1-5)
    - “Rate the sound on a scale of 1 to 5.”
  - MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor)
    - Add a dummy to prevent random rating
- Objective measures (numerical calculation)
  - SNR (signal-to-noise ratio, dB)
    - $10 \log_{10} \frac{\sum_n (s[n] - \hat{s}[n])^2}{\sum_n s[n]^2}$
  - PESQ (perceptual evaluation of speech quality, 1-5)
    - Approximation of MOS values
    - ITU-T recommendation P.862

VOCODERS

Speech Information Rates

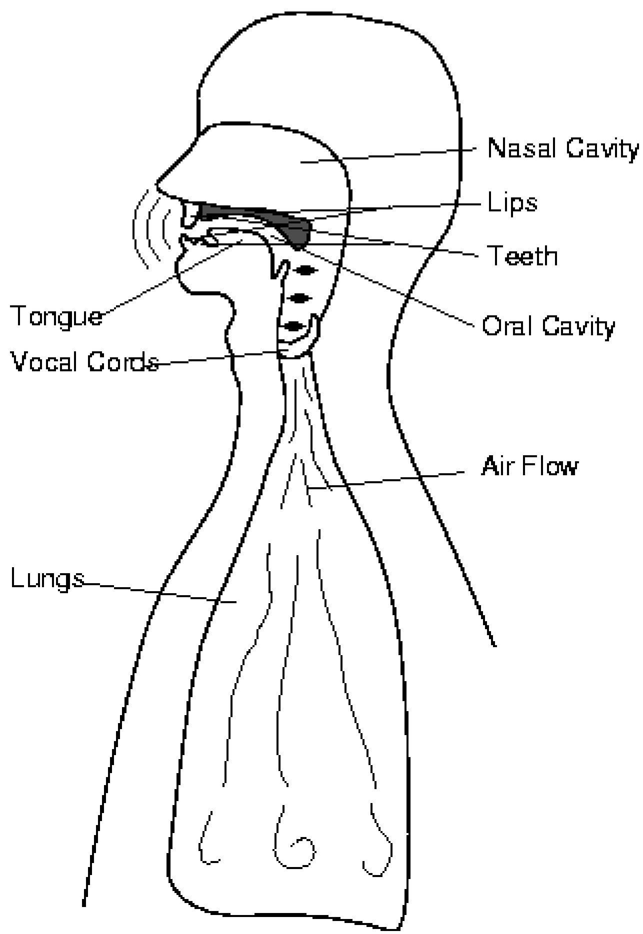
- Production level:
  - 10-15 phonemes/second for continuous speech
  - 32-64 phonemes per language ➔ 6 bits/phoneme
  - Information Rate = **60 – 90 bps** at the source
- Waveform level
  - speech bandwidth is 4 – 10 kHz
    - sampling rate is 8 – 20 kHz
  - need 12-16 bit quantization for high quality digital coding
  - Information Rate = **96-320 Kbps**
- More than 3 orders of magnitude (> x1000) difference in Information Rates between the production and waveform levels

Speech Coder Comparisons



- waveform coders characterized by:
  - high bit rates (24 Kbps – 1 Mbps)
  - low complexity / low flexibility
- analysis/synthesis systems characterized by:
  - low bit rates (600 bps – 10 Kbps)
    - 8 Kbps for 2G voice communication standard
  - high complexity
  - great flexibility (e.g., time expansion/compression)

Human Speech Production System



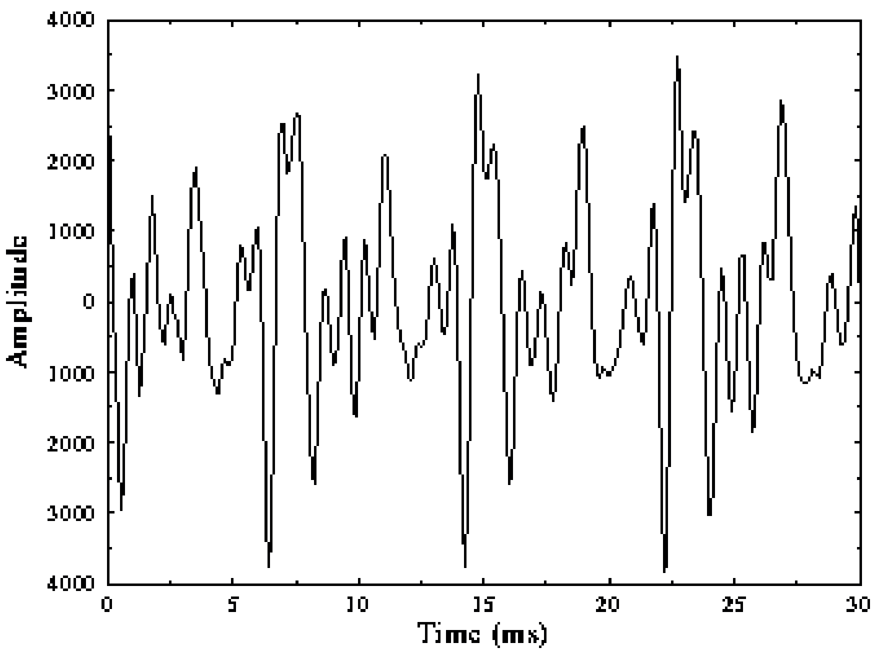
- Air flow forced from lungs to vocal tract
  - The basic vibrations – vocal cords
  - Filter with resonances (called formants)
- Model the vocal tract as a filter
  - The shape changes relatively slowly
- The vibrations at the vocal cords
  - The excitation signal
- Speech sound classes
  - **Voiced sounds**
    - Voice cord vibration
    - Long-term periodicity
  - **Unvoiced sounds**
    - Constriction in the vocal tract
    - No long-term periodicity
  - **Plosive sounds**
    - Release of air pressure behind mouth

Voiced Speech

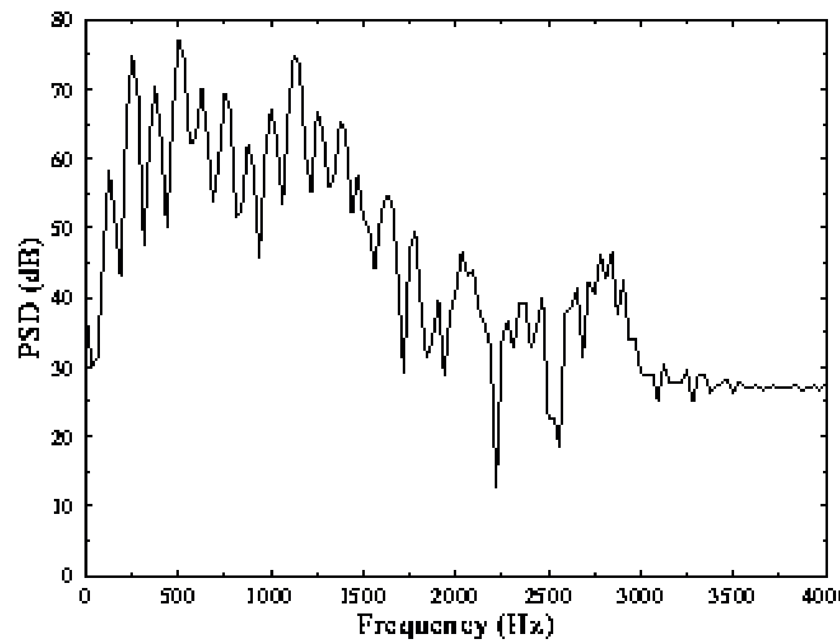
- The vocal cords vibrate open and close
  – Interrupt the air flow
  – Quasi-periodic pluses of air
  – **<u>Pitch</u>** – the rate of the opening and closing
- A high degree of periodicity at the pitch period
  – 2-20 ms

Voiced Speech

Voiced speech



Power spectral density

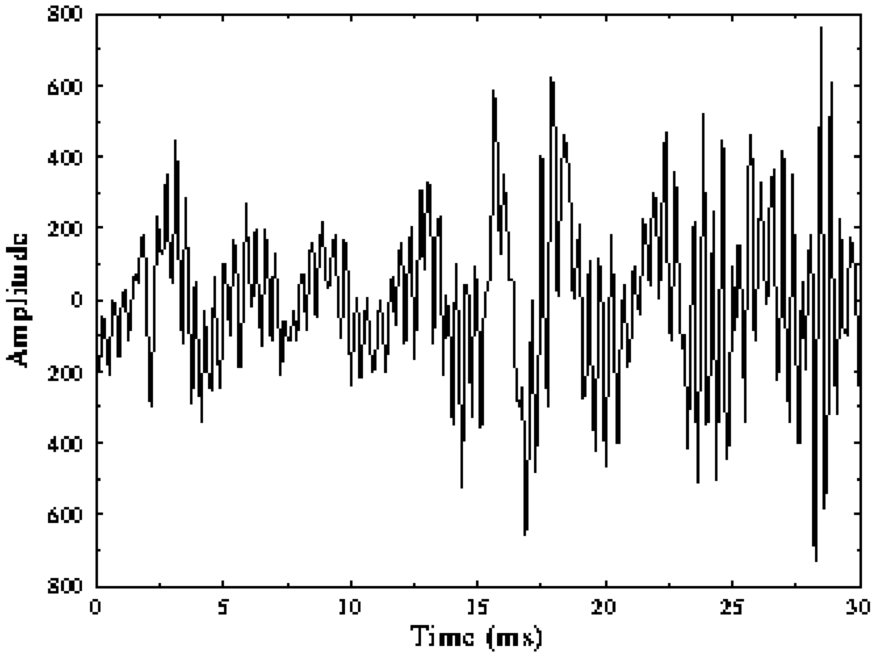


Unvoiced Speech

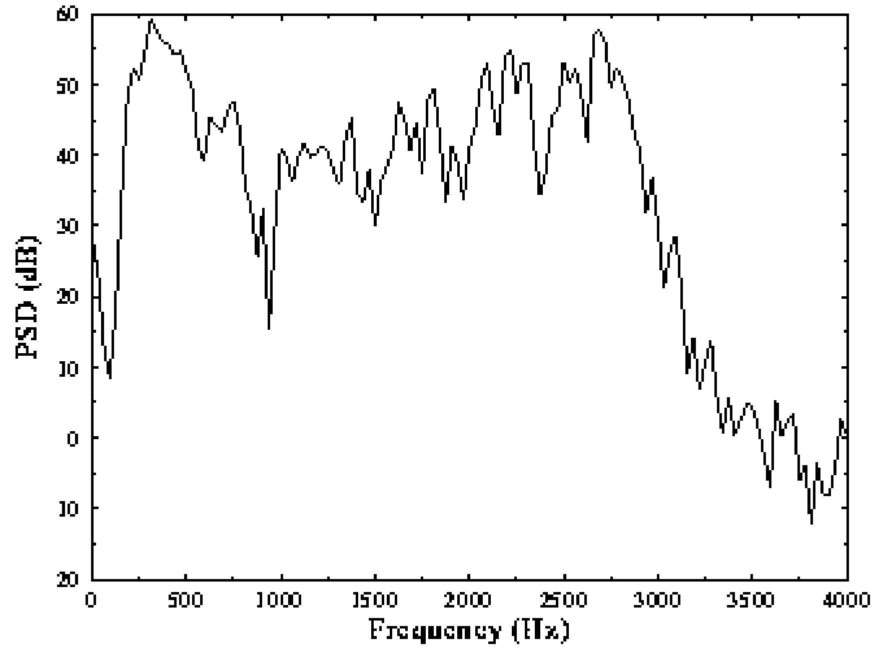
- Forcing air at high velocities through a constriction
- The glottis is held open
- Noise-like turbulence
- Show little long-term periodicity
- Short-term correlations still present

Unvoiced Speech

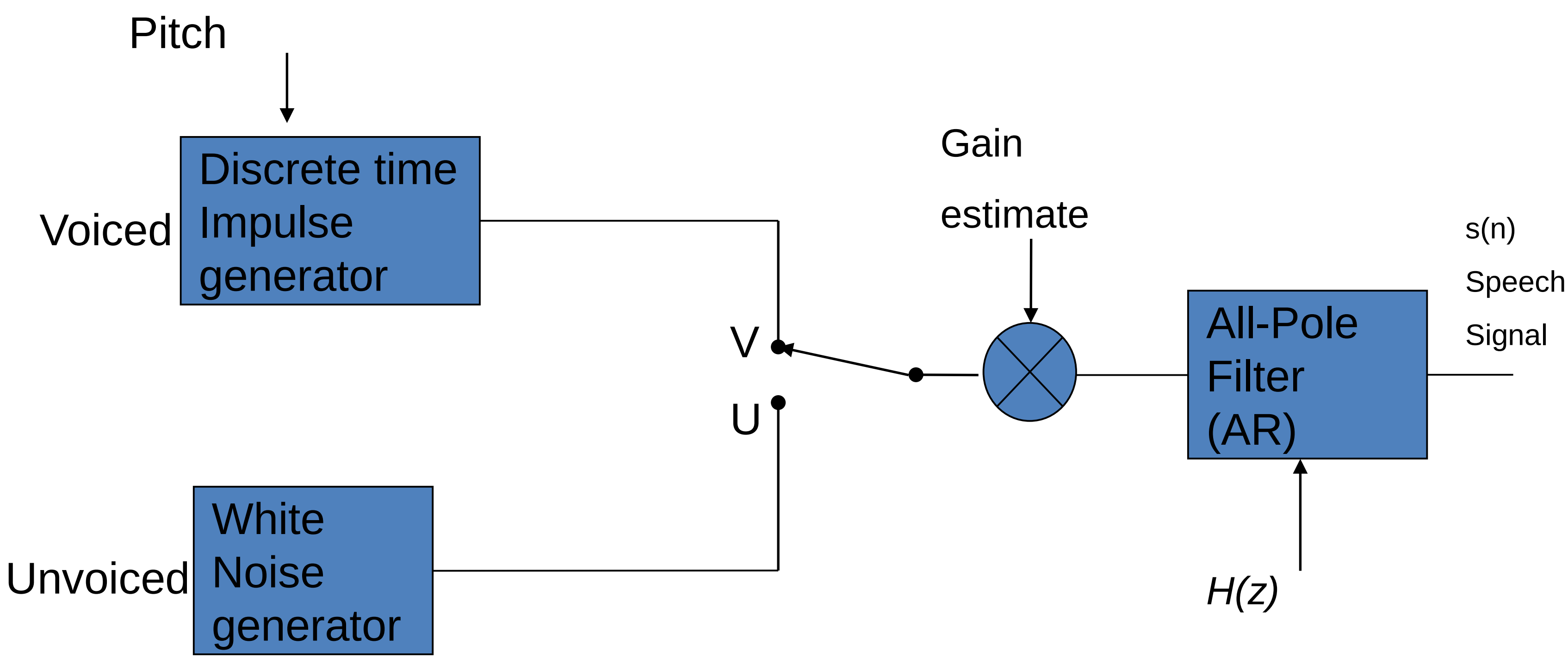
unvoiced speech



Power spectral density



Vocoders: Using LP Analysis



necessary components:
V/UV decision, pitch estimation, white noise generation, LPC, LPC-to-LSF transformation, scalar quantization, vector quantization, resynthesis

LPC Sounds

Number of LP Coefficients (LPC order)

- 2
- 6
- 8
- 10
- 12

Reference

LPC vocoder requirements

Things to Encode

- Excitation signal
  - For voiced: pitch period – 1 real number
  - For unvoiced: ??
  - U / UV flag: 1 bit
- Excitation gain
  - 1 real number
- Vocal tract filter H(z)
  - 10 real numbers for 8 kHz (LPC order of 10)

Data Rates

- for every 10 ms, we have at least 1+1+10 reals + 1 bit
- using 4 byte float for a real number, we need 12*4*8 + 1 = 385 bits
- the bandwidth to transmit this information is then 38500 bits / sec = **38.5 kbps**.

Comparison to Other Coders

- Audio codec (44.1 kHz)
  - CD quality: 44.1 kHz * 16 bits * stereo = 1411.2 kbps
  - MP3: 128 kbps / 192 kbps / 320 kbps
  - WMA: 64 kbps
- Speech codec (8kHz)
  - PCM: 8 kHz * 16 bits = 128 kbps
  - μ-law: 8 kHz * 8 bits = 64 kbps
  - ADPCM: 8 kHz * 4 bits = 32 kbps
  - <u>LPC-10, using floats: **38.5 kbps** – is it useful</u>?

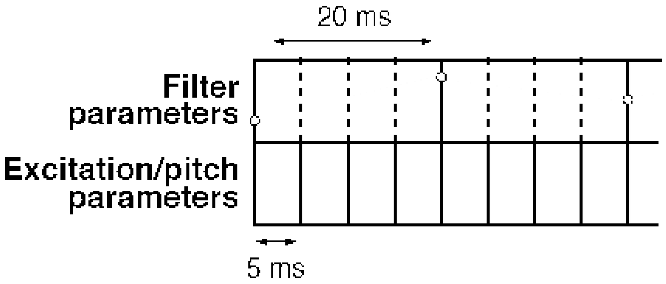
EFFICIENT VOCODER DESIGN USING LINE SPECTRAL FREQUENCIES

LPC Coder / Decoder Modules

- Excitation coding
  - **V/UV decision**
  - Voiced excitation – **Pitch estimation**; glottal pulse generation by pulse train / sinusoids / interpolation
  - Unvoiced excitation – random sound / **vector quantization**
- Vocal tract modeling
  - LPC – conversion to **LSP** (line spectral pair), **scalar / vector quantization**
- **Resynthesis**
  - combine regenerated, segmented frames back to time series

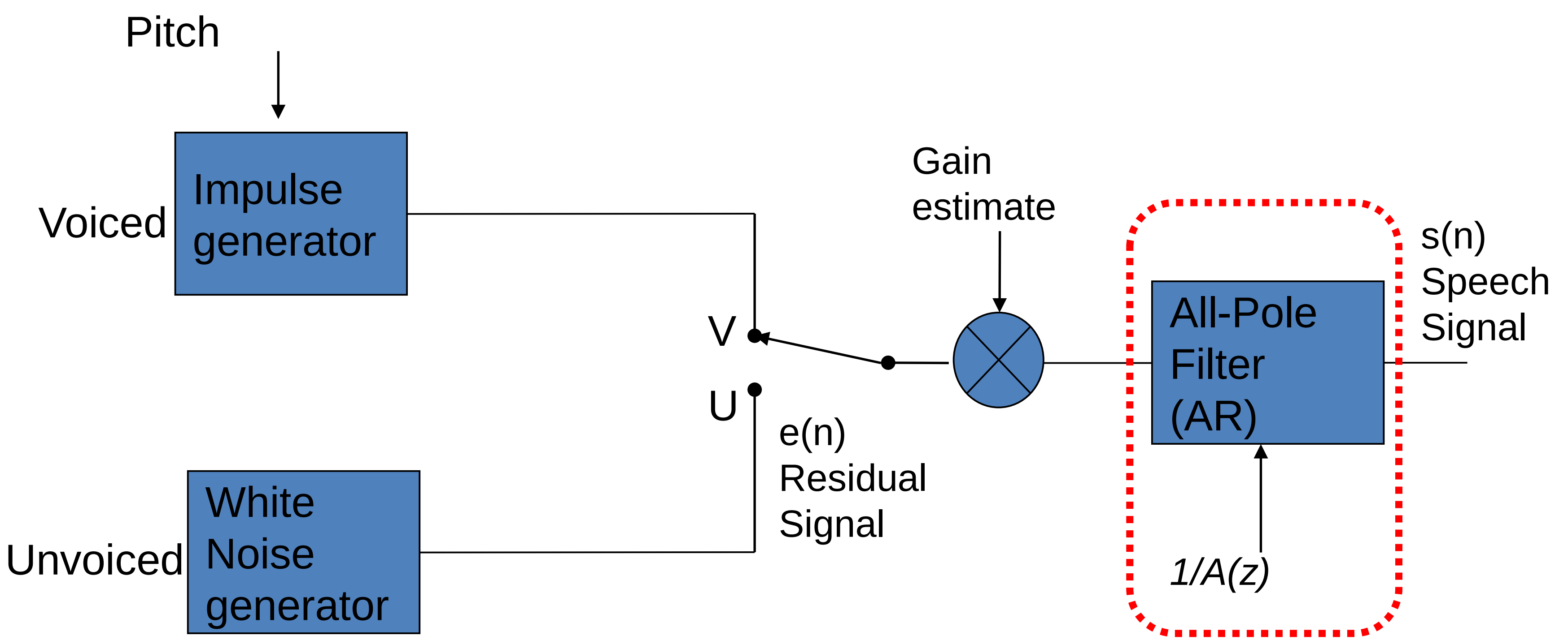
LPC-based Vocoder Design Procedure

1. Extract a_k parameters properly (LPC analysis)
   - a_k parameters change relatively slowly (**20ms**)
2. Transform a_k to LSFs and **quantize (code)** them properly so that there is little quantization error
   - The sensitivity of a_k to noise is not consistent, so use LSFs
   - Relatively small number of bits go into coding the a_k coefficients



3. Represent $e(n)$ via:
   - Change relatively fast (**5ms/10ms subframe**)
   - Pitch pulses and white noise – LPC coding
   - Codebook vectors – CELP (codebook excited linear prediction)
     - Almost all of the coding bits go into coding of $e(n)$

LPC-based Vocoder Design

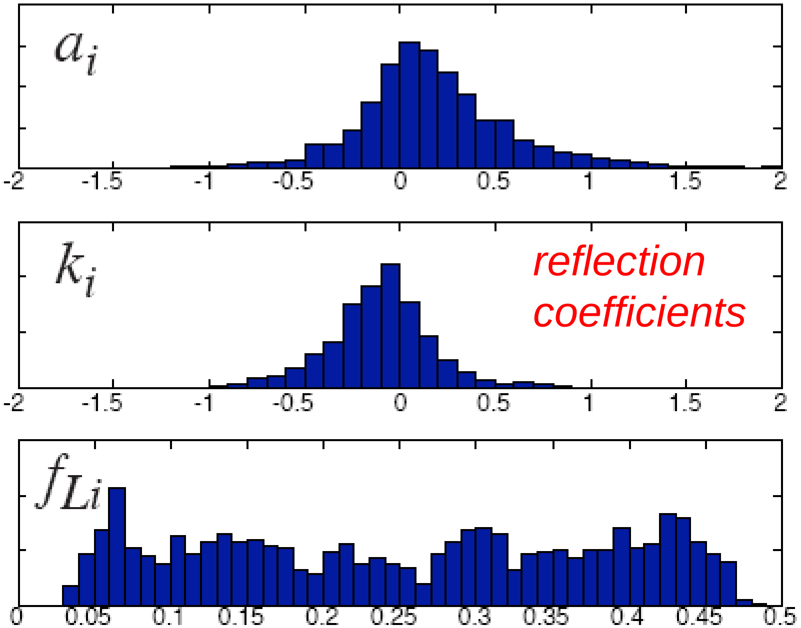


necessary components:
V/UV decision, pitch estimation, residual modeling by Impulse / noise
LPC, LPC-to-LSF transformation
LSF quantization

LPC Encoding

- For communications quality:
  - 8 kHz sampling (4 kHz bandwidth)
  - 10th order LPC (up to 5 pole pairs)
  - update every 20-30 ms ➔ 300-500 parameters/s
- LSF transformation
  - In MATLAB, use function **`poly2lsf(A)`**
  - In addition, LSF to LPC is by **`lsf2poly(L)`**
- Bit allocation:
  - FS1016 (4.8 kbps): 10 LSPs x 3-4 bits / 30 ms = 1.1 kbps

poor distribution, can't interpolate



LSFs – almost uniform sensitivities; errors are linear

Uniform Scalar Quantization

- If the distribution is uniform on a closed interval, uniform scalar quantization is the most efficient coding scheme
- Code for the input can be obtained very simply

```
c = max(1,min(max_code,fix((x-bias)/step)));
```

- However, if the input distribution is not uniform, non-uniform scalar quantization is necessary
- Moreover, if the multivariate input dimension is correlated, vector quantization is necessary

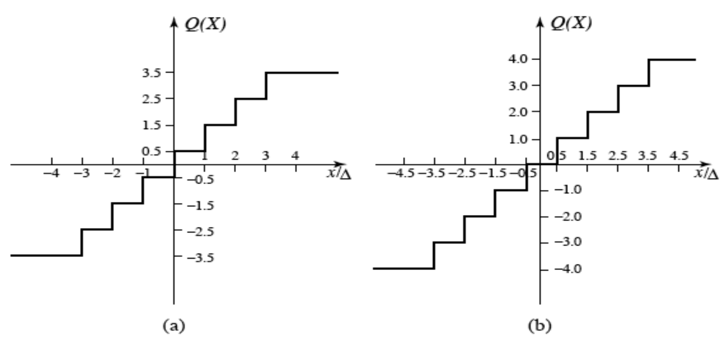


Fig. 8.2: Uniform Scalar Quantizers: (a) Midrise, (b) Midtread.

Optimal Coding

- Shannon information:
  An unlikely occurrence is more 'informative'

$p(A) = 0.5 \quad p(B) = 0.5$

ABBBBAAABBABBABBABB

A, **B** equiprobable

$p(A) = 0.9 \quad p(B) = 0.1$

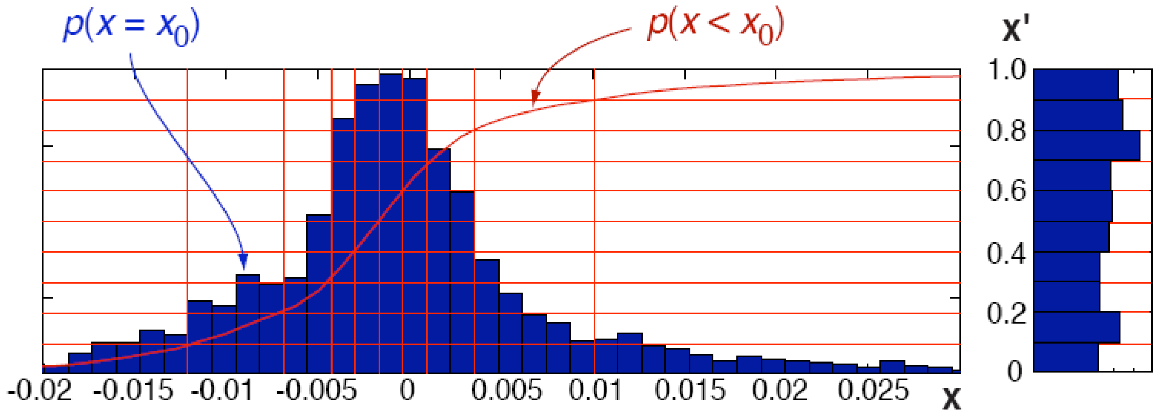
AAAAABBAAAAAABAAAAB

A is expected;
B is 'big news'

- Information in bits $I = -\log_2(probability)$
  - clearly works when all possibilities equiprobable
- Optimal bitrate $\rightarrow$ av.token length $=$ entropy $H = E[I]$
  - .. equal-length tokens are equally likely
- How to achieve this?
  - transform signal to have uniform pdf
  - nonuniform quantization for equiprobable tokens
  - variable-length tokens $\rightarrow$ Huffman coding

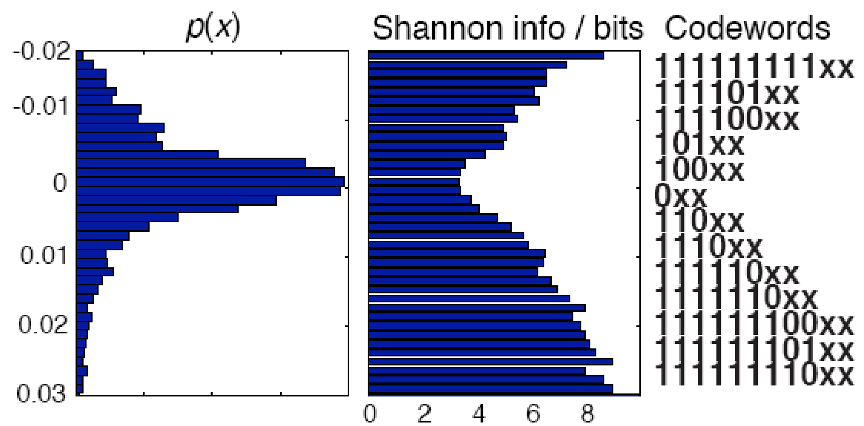
Scalar Quantization for Optimum Bitrate

- Quantization should reflect pdf of signal:



  - cumulative pdf $p(x < x_0)$ maps to uniform x'

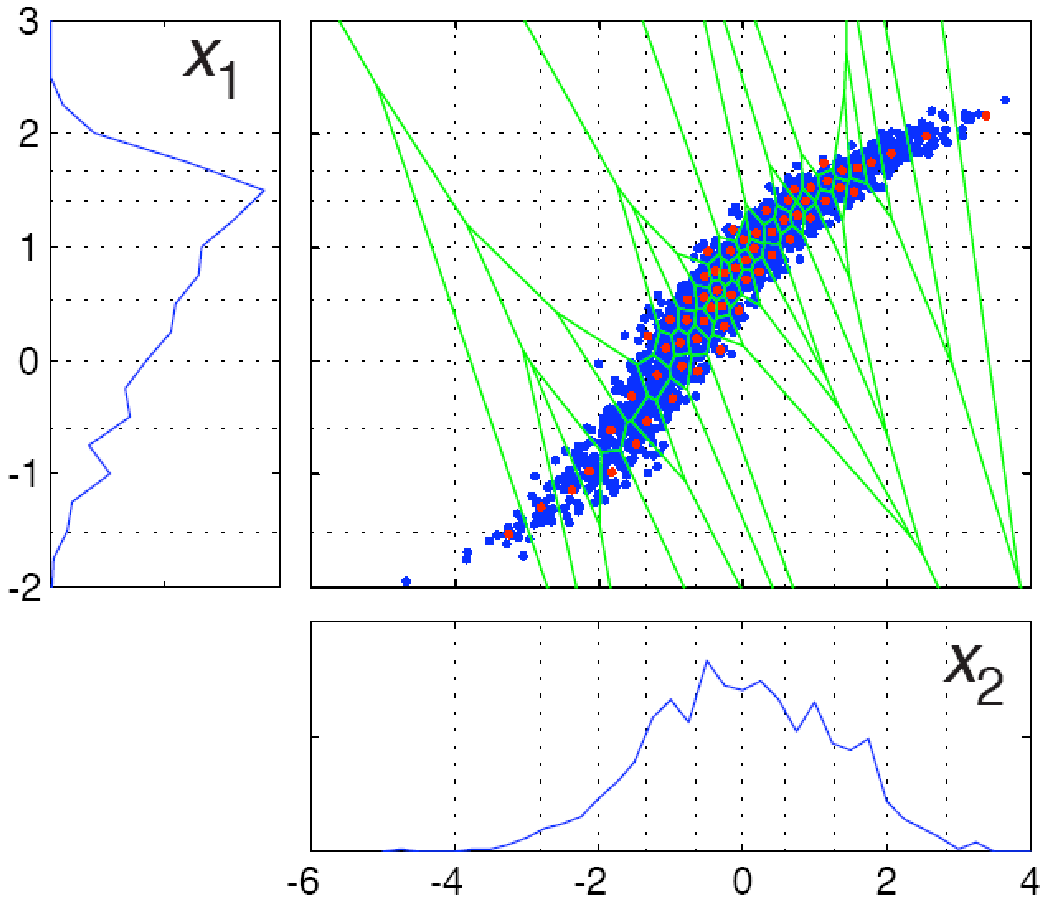
- Or, codeword length per Shannon $\log_2(p(x))$:



  - Huffman coding: tree-structured decoder

Vector Quantization

- Quantize mutually dependent values in joint space:

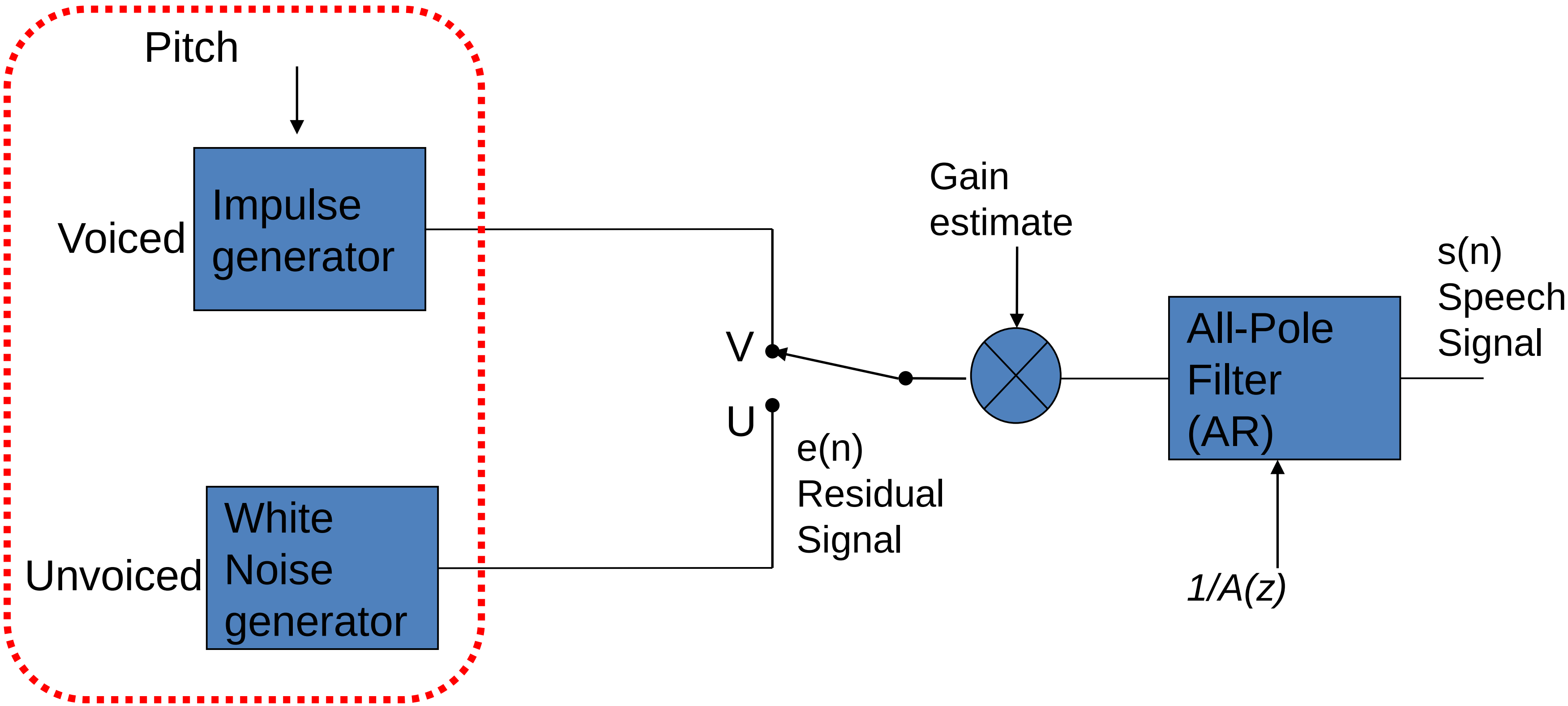


- May help even if values are largely independent
  - larger space x1,x2 is easier for Huffman

Some LSP Coding Strategies

- Uniform scalar quantization: NA
- Non-uniform scalar quantization
  - 1 x 3 + 4 x 4 + 5 x 3 = 34 bits / 20 ms = 1.7 kbps
    - Although the standard adopted 30 ms for frame size, we are using 20 ms
- Split vector quantization of 3/3/4, 256 codewords
  - 3 x 8 = 24 bits / 20 ms = 1.2 kbps
- Full vector quantization, codebook size 512
  - 16 bits / 20 ms = 0.8 kbps

LPC-based Vocoder Design



necessary components:
V/UV decision
pitch estimation, residual modeling by Impulse / noise
LPC, LPC-to-LSF transformation
LSF quantization

Excitation Encoding

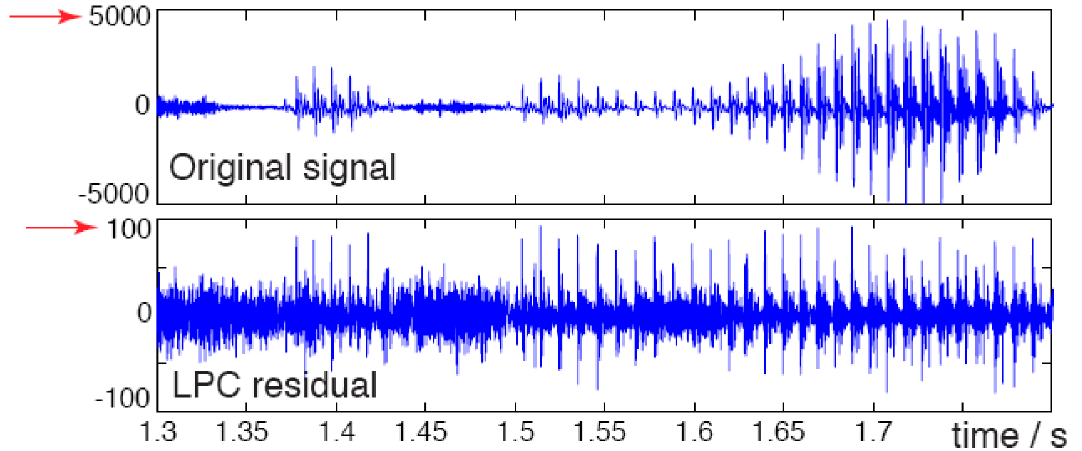
- Done on subframes (usually 5 ms-10ms)
- V/UV decision
- Compute gain and divide the residual signals
- For voiced segments, find pitch period and delay
- For unvoiced segments, in CELP, find the closest codeword from the codebook
- Information to be transferred
  - Gain / VUV flag / (voiced) pitch period / (unvoiced) excitation code (1 real + 1 bit + 1 real + 1 int)

Excitation Decoding

- Voiced segments
  - generate pulse train of the given period
  - adjust phase (shift) so that the first pulse be one pitch period away from the last pulse of previous frame, to prevent audible discontinuities
- Unvoiced segments
  - generate random signal (LPC10)
  - or use the closest codeword from the codebook (CELP)
- (Common) Gain modification
  - normalize the excitation so that it has unit variance
  - multiply gain

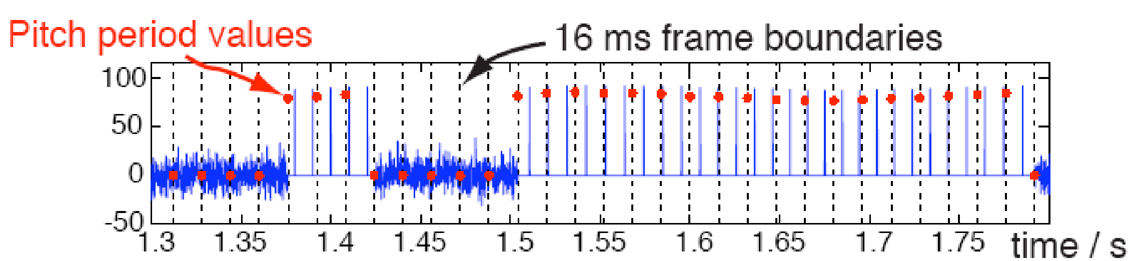
Excitation Encoding Illustration

- Excitation already better than raw signal:



  - save several bits/sample, but still > 32 Kbps

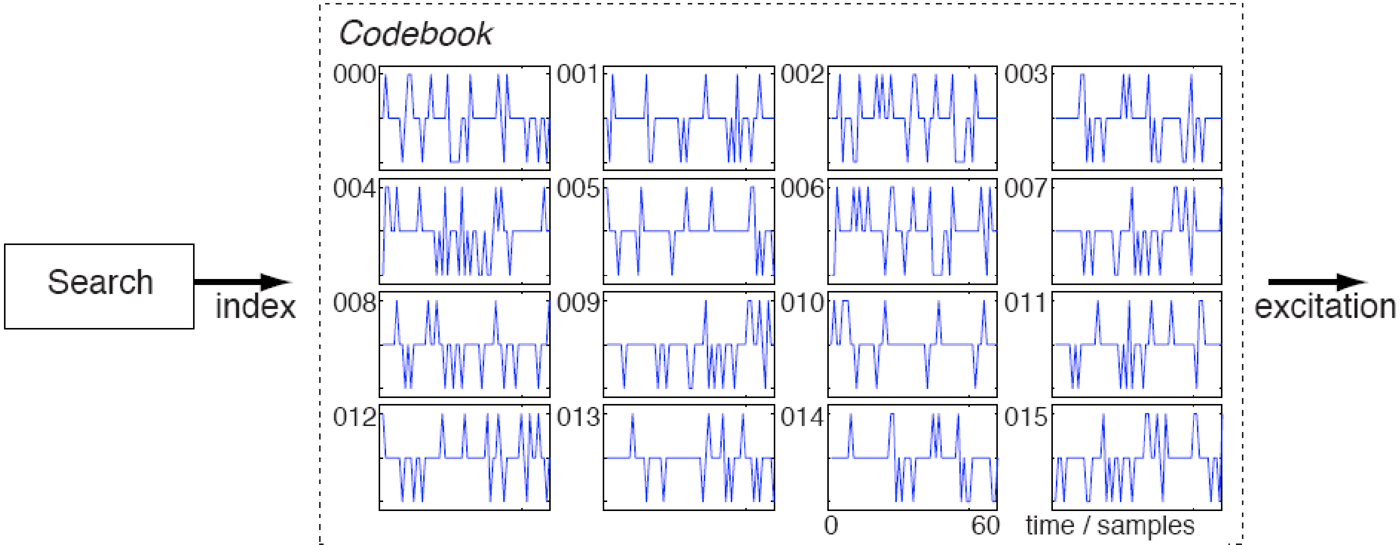
- Crude model: U/V flag + pitch period
  - $\sim$ 7 bits / 5 ms = 1.4 Kbps $\rightarrow$ LPC10 @ 2.4 Kbps



Case Study: CELP

- Represent excitation with codebook

  e.g. 512 sparse excitation vectors



  - linear search for minimum weighted error?

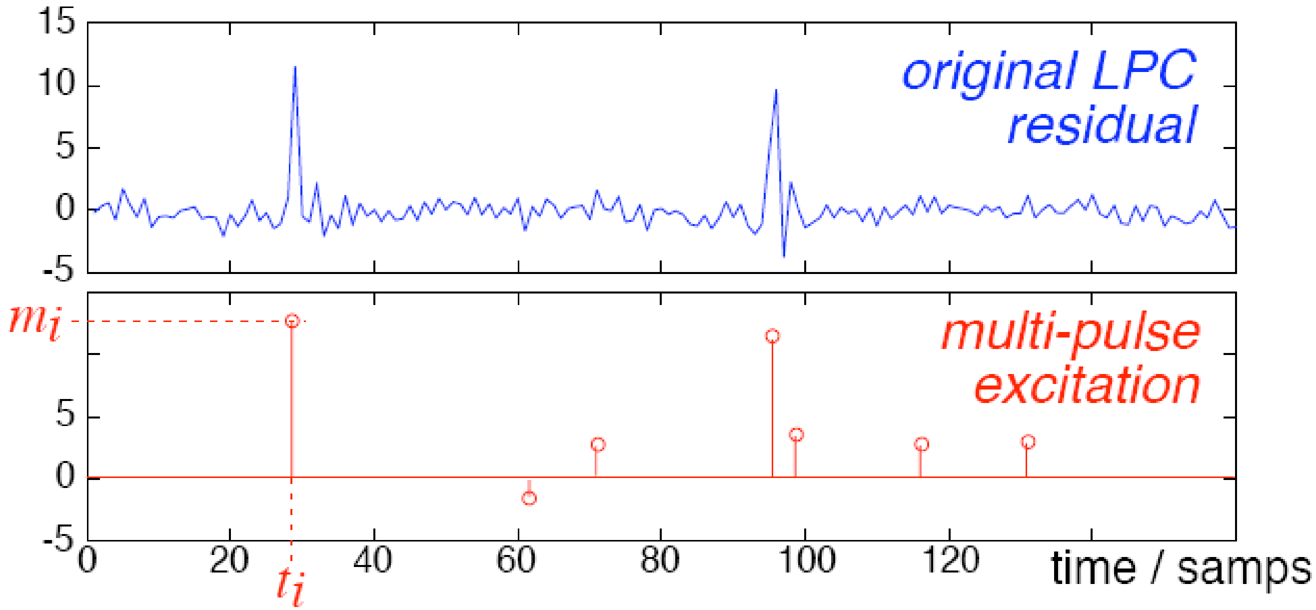
- FS1016 4.8 Kbps CELP (30ms frame = 144 bits):

10 LSPs	4x4 + 6x3 bits =	34 bits
Pitch delay	4 x 7 bits =	28 bits
Pitch gain	4 x 5 bits =	20 bits
Codebk index	4 x 9 bits =	36 bits
Codebk gain	4 x 5 bits =	20 bits

▸ 138 bits

Case Study: Multi-Pulse Excitation (MPE-LPC)

- Stylize excitation as N discrete pulses



  - encode as $N \times (t_i, m_i)$ pairs

Generic Mixed Signal Coders Comparison

- **Reference**
- Uniform PCM: 64 kbps, 8bits
- ADPCM: 8bits per sample
- LPC
- CELP

Bit Rate Comparison

"A lathe is a big tool. Grab every dish of sugar"

- MU-law PCM: **64** kbps, 8bits
- ADPCM: **32** kbps, 4bits
- CELP: **4.8** kbps, 0.6 bits
- LPC-10: **2.4** kbps, 0.3 bits

Data from Data-Compression.com

Sentence has a good mix of sounds; unvoiced sounds: "th" (lathe) and "sh" (dish, sugar); difficult consonants "b" (big), "t" (tool) and "G" (grab), which are part voiced, part unvoiced, and the "s" in "is" which is pronounced like a "z" simultaneously voiced and unvoiced.
Reference:
www.cs.ucl.ac.uk/teaching/GZ05/samples/index.html#speech

Appendix: Coding Standards

- Only enough for intelligibility and speaker identification
- Coding is only useful if recipient knows the code
- Standardization efforts are important
- (American) federal standards of low bit-rate secure voice:
  - FS1015e: LPC-10 2.4 Kbps
  - FS1016: 4.8 Kbps CELP
- ITU G.x series (also H.x for video)
  - G.726 ADPCM
  - G.729 Low delay CELP
- MPEG (audio)
  - MPEG 1 Audio layers 1,2,3 (mp3)
  - MPEG 2 Advanced Audio Codec (AAC)
  - MPEG 4 Synthetic-Natural Hybrid Codec