Internet of Things:

Communication Technologies
Codecs

Harinath Garudadri and Ganz Chockalingam

Qualcomm Institute of Calit2

University of California, San Diego

Lesson 1

The big picture about codecs

Lesson 1 | The big picture about codecs

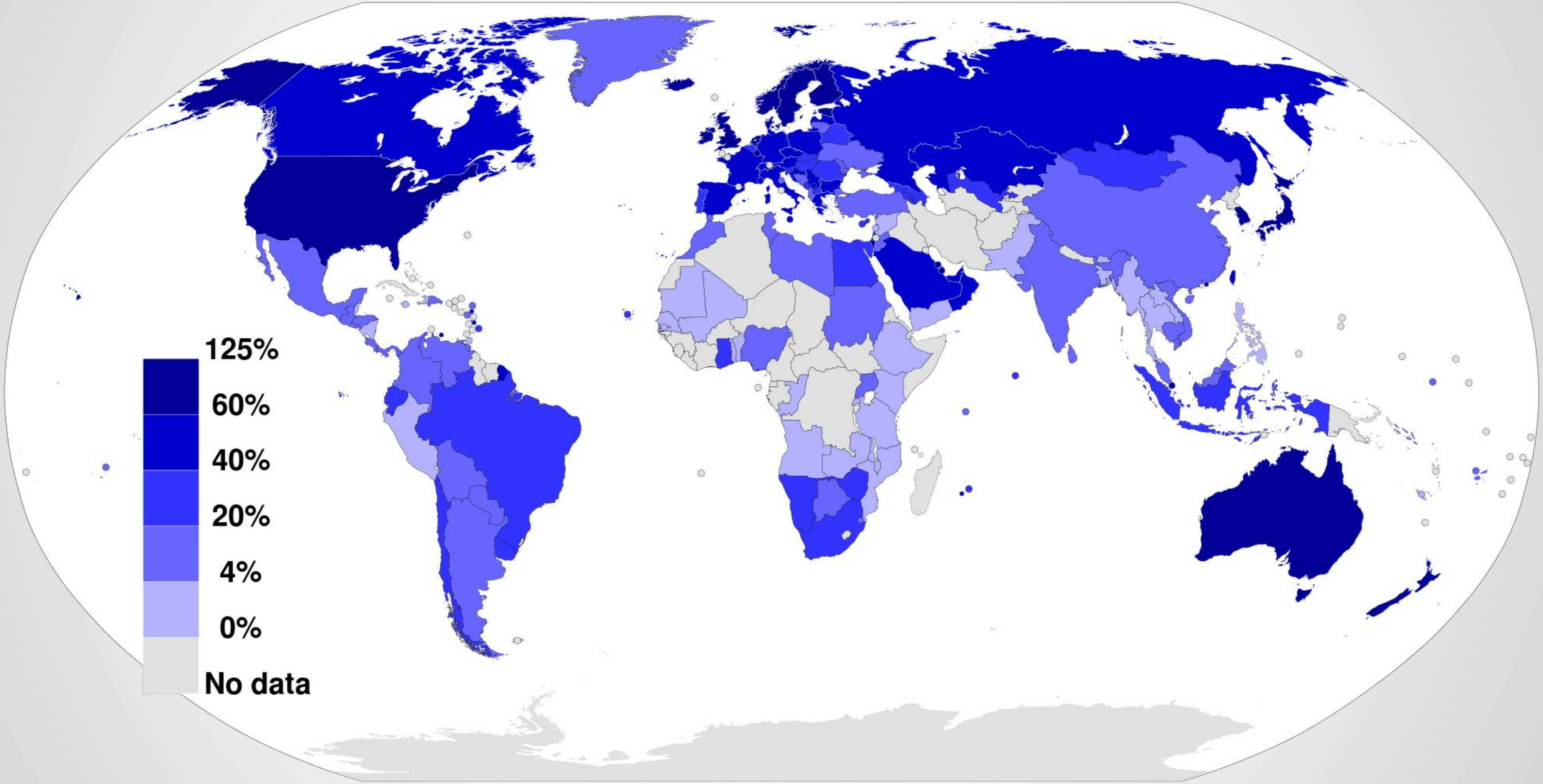
1 - Why compress data?

2 - Why does it work?

3 - What to look for?

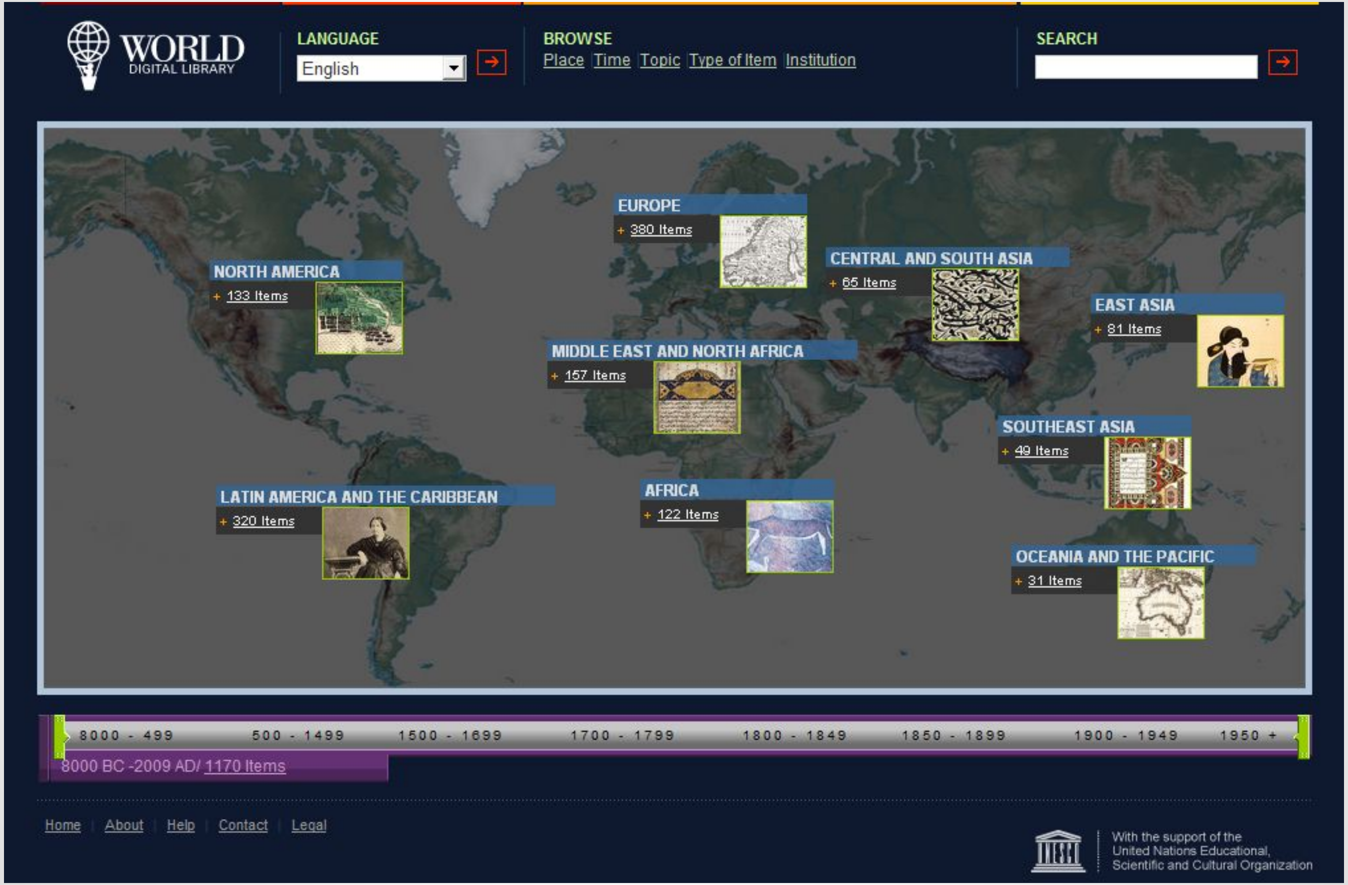
1 - Why compress data?

1 | Why compress data?



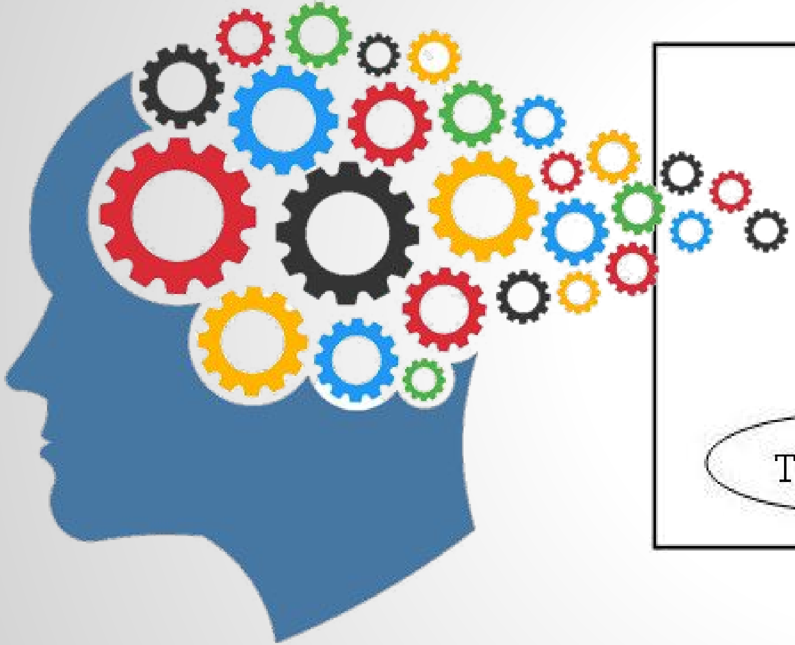
Mobile broadband subscriptions as a percentage of the population in 2012 (ITU-T)

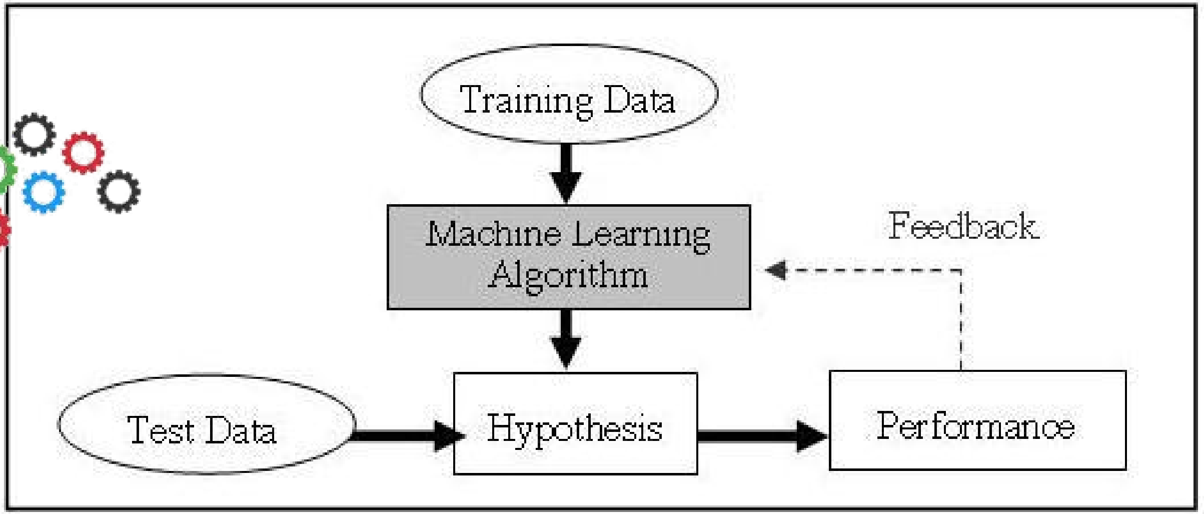
1 | Why compress data?



https://en.wikipedia.org/wiki/World_Digital_Library

1 | Why compress data?



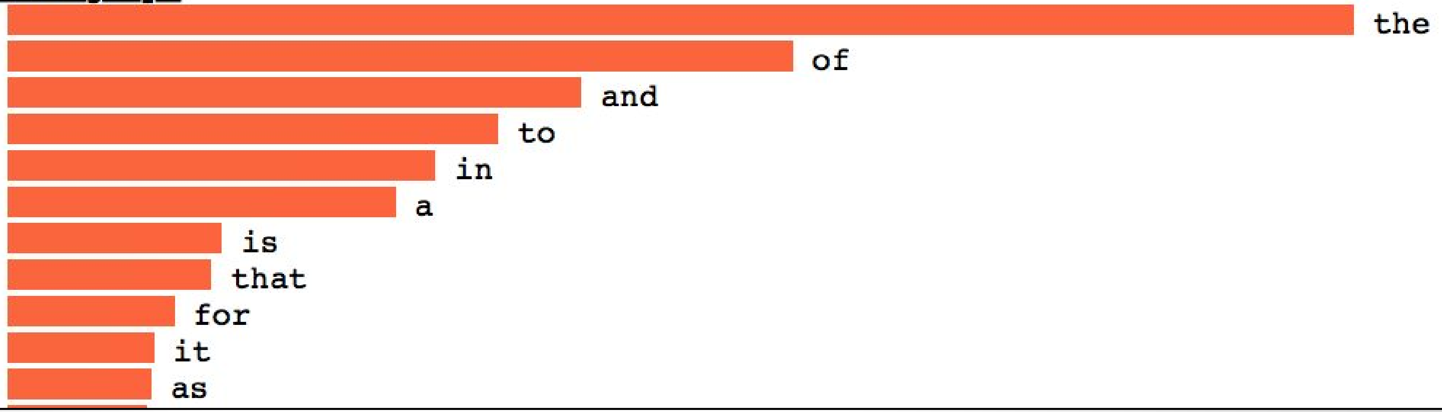


2 - Why does it work?

2 | Why does it work?

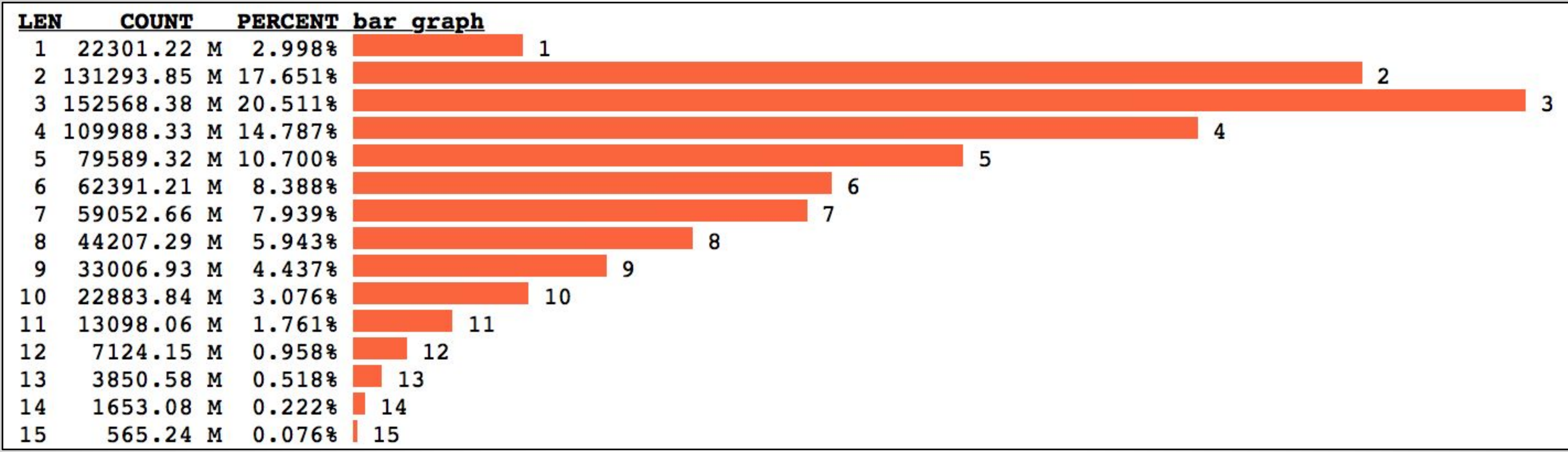
2 | Why does it work?

WORD	COUNT	PERCENT	bar graph
the	53.10 B	7.14%	
of	30.97 B	4.16%	
and	22.63 B	3.04%	
to	19.35 B	2.60%	
in	16.89 B	2.27%	
a	15.31 B	2.06%	
is	8.38 B	1.13%	
that	8.00 B	1.08%	
for	6.55 B	0.88%	
it	5.74 B	0.77%	
as	5.70 B	0.77%	



Redundancy in English Text: http://norvig.com/mayzner.html – ETAOIN SRHLDCU

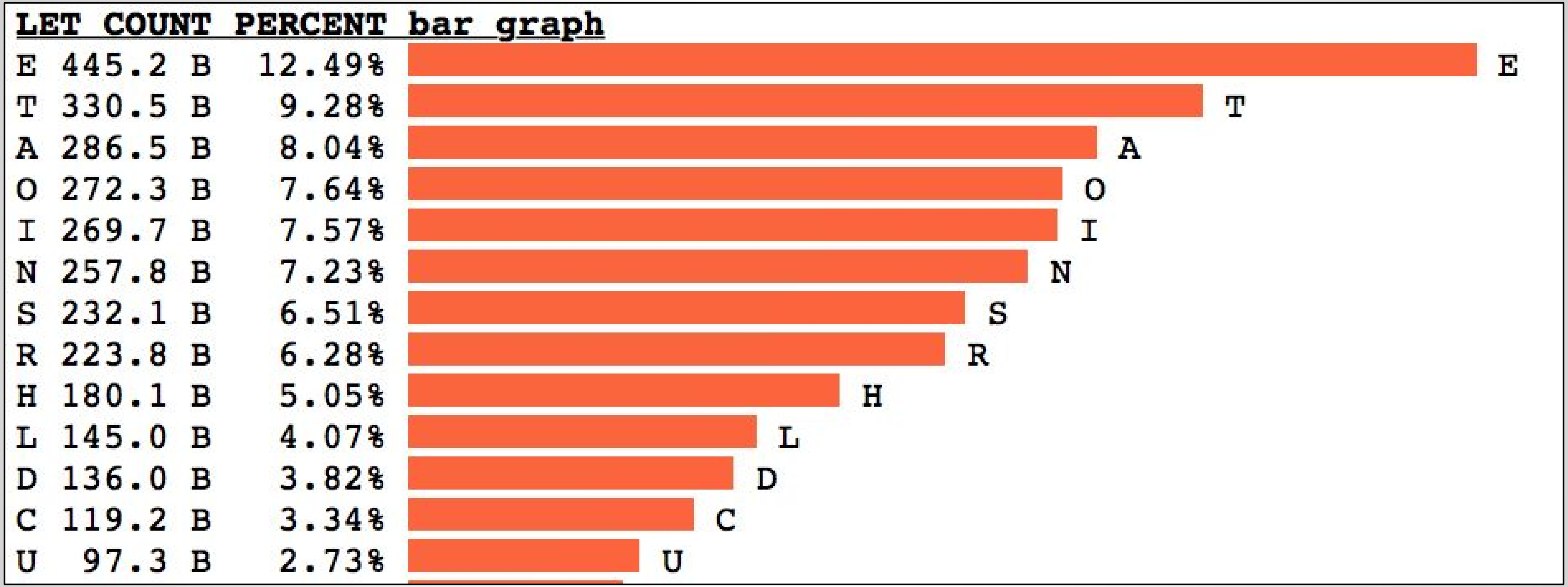
2 | Why does it work?



LEN	COUNT	PERCENT	bar graph
1	22301.22 M	2.998%	1
2	131293.85 M	17.651%	2
3	152568.38 M	20.511%	3
4	109988.33 M	14.787%	4
5	79589.32 M	10.700%	5
6	62391.21 M	8.388%	6
7	59052.66 M	7.939%	7
8	44207.29 M	5.943%	8
9	33006.93 M	4.437%	9
10	22883.84 M	3.076%	10
11	13098.06 M	1.761%	11
12	7124.15 M	0.958%	12
13	3850.58 M	0.518%	13
14	1653.08 M	0.222%	14
15	565.24 M	0.076%	15

Redundancy in English Text: http://norvig.com/mayzner.html – ETAOIN SRHLDCU

2 | Why does it work?



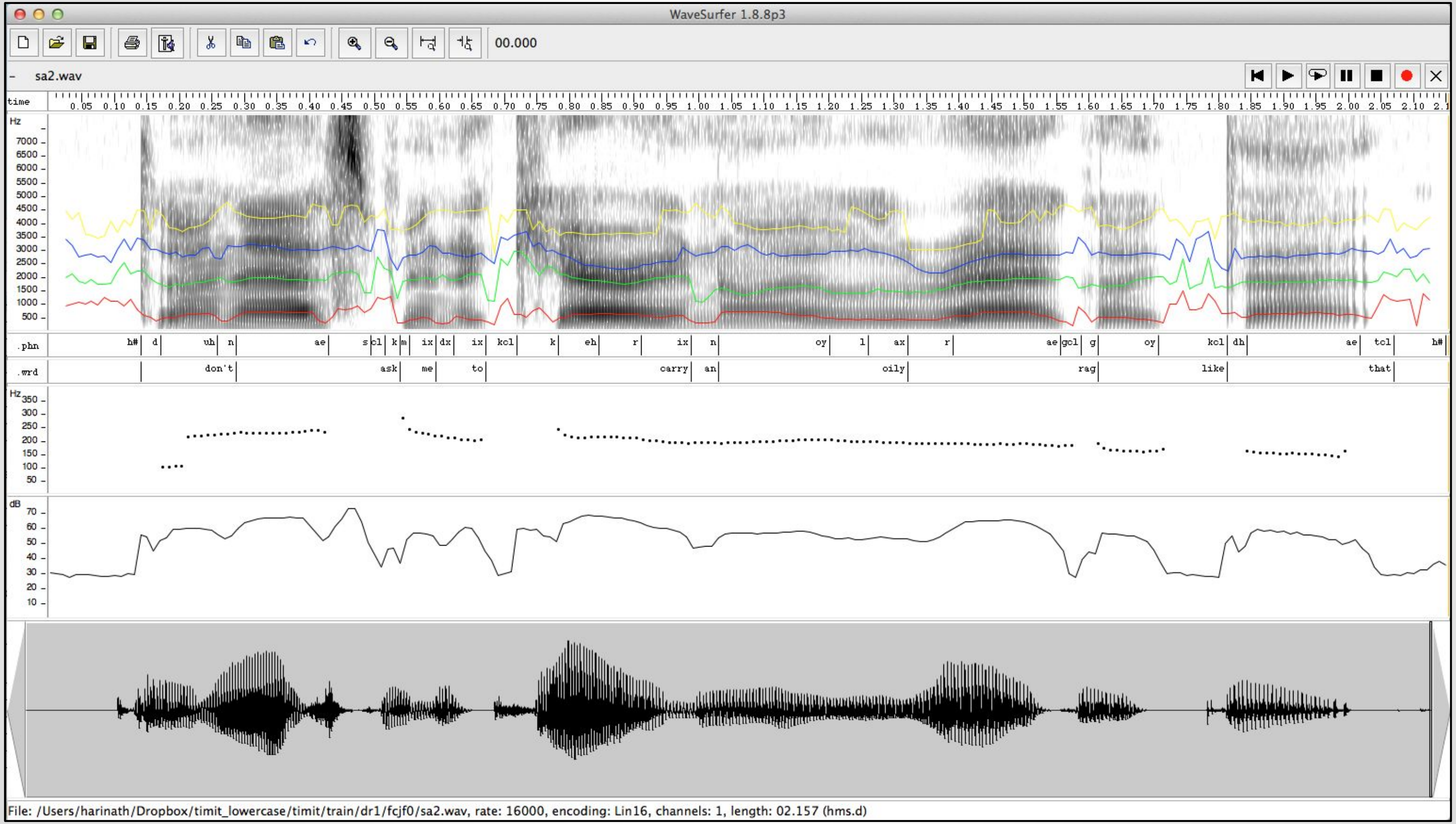
LET	COUNT	PERCENT	bar graph
E	445.2 B	12.49%	E
T	330.5 B	9.28%	T
A	286.5 B	8.04%	A
O	272.3 B	7.64%	O
I	269.7 B	7.57%	I
N	257.8 B	7.23%	N
S	232.1 B	6.51%	S
R	223.8 B	6.28%	R
H	180.1 B	5.05%	H
L	145.0 B	4.07%	L
D	136.0 B	3.82%	D
C	119.2 B	3.34%	C
U	97.3 B	2.73%	U

Redundancy in English Text: http://norvig.com/mayzner.html – ETAOIN SRHLDCU

2 | Why does it work?

ETAOIN SRHLDCU – QZ

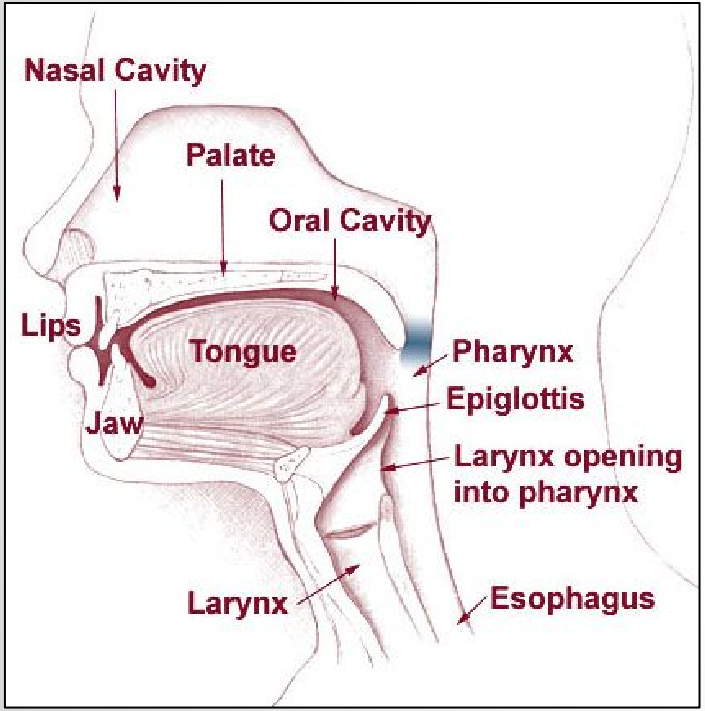
2 | Why does it work?



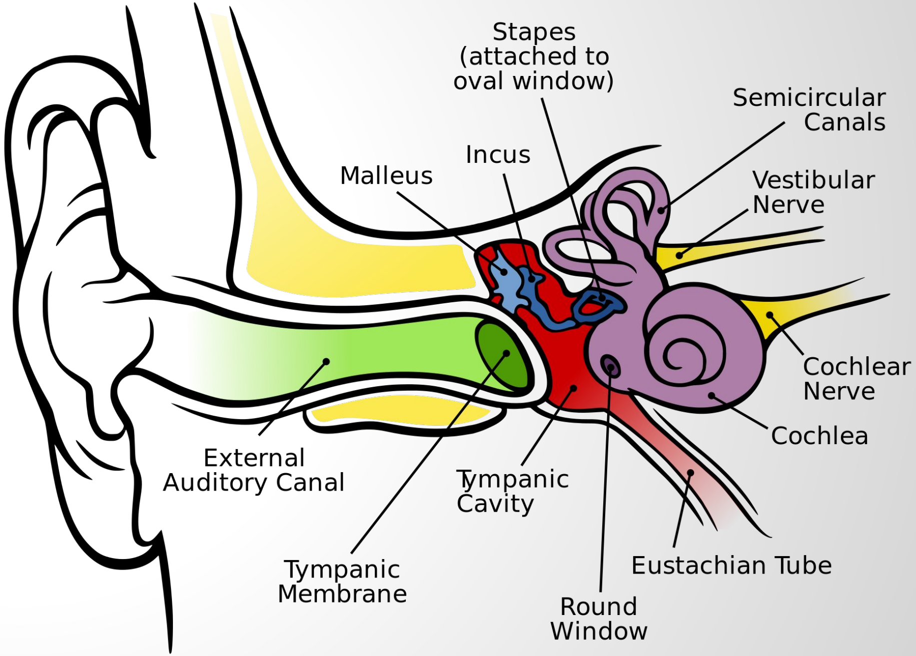
Redundancy in Amplitude, Time and Frequency

2 | Why does it work?

Source



Sink



Knowledge of Speech Production and Perception

3 - What to look for?

3 | What to look for?

- Compression Efficiency
- Reconstruction Quality
- Algorithm Latency
- Error Resiliency
- Computational Complexity

Lesson 2

Waveform Codecs

Lesson 2 | Waveform Codecs

1 - Amplitude - Log Companding (PCM)

2 - Time - Adaptive, Delta Quantization (ADPCM)

3 - Frequency - ADPCM in multiple subbands (Subband ADPCM)

1 - Amplitude - Log Companding (PCM)

1 | Amplitude - Log Companding (PCM)

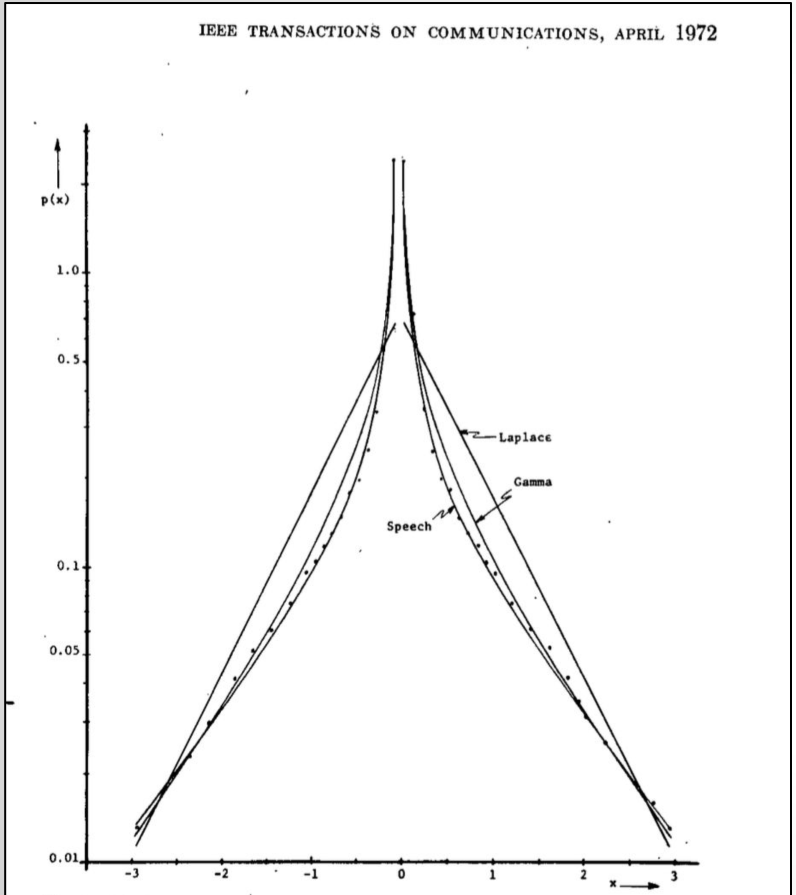


Fig. 3. Real speech and theoretical gamma and Laplace probability densities.

M.D. Paez and T.H. Glisson,

"Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems",

IEEE Transactions on Communications, April, 1972

1 | Amplitude - Log Companding (PCM)

G.711 μ-Law

$$y(n) = \text{sign}(x)\frac{ln(1 + \mu\,|x(n)|)}{ln(1+\mu)},$$

$$\text{sign}(x) = \begin{cases} +1, & x(n) > 0, \\ 0, & x(n) = 0, \\ -1, & x(n) < 0 \end{cases}$$

$$\mu = 255$$

1 | Amplitude - Log Companding (PCM)

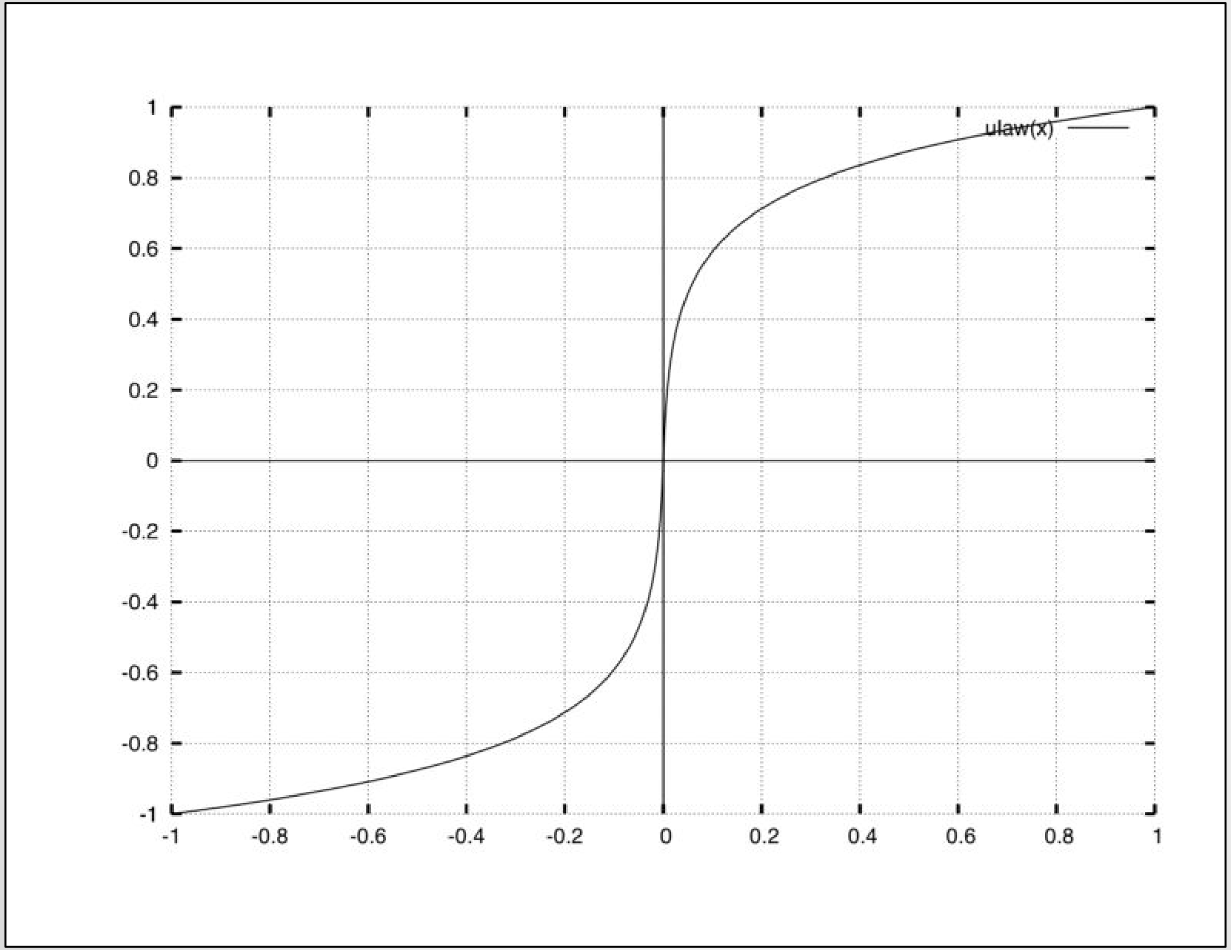
G.711 *A*-Law

$$y(n) = \begin{cases} \mathrm{sign}(x)\frac{A\,|x(n)|}{1+\ln(A)}, & \text{for } 0 \leq |x(n)| < \frac{1}{A}, \\ \mathrm{sign}(x)\frac{A\,|x(n)|}{1+\ln(A)}, & \text{for } \frac{1}{A} \leq |x(n)| < 1. \end{cases}$$

$$A = 87.6$$

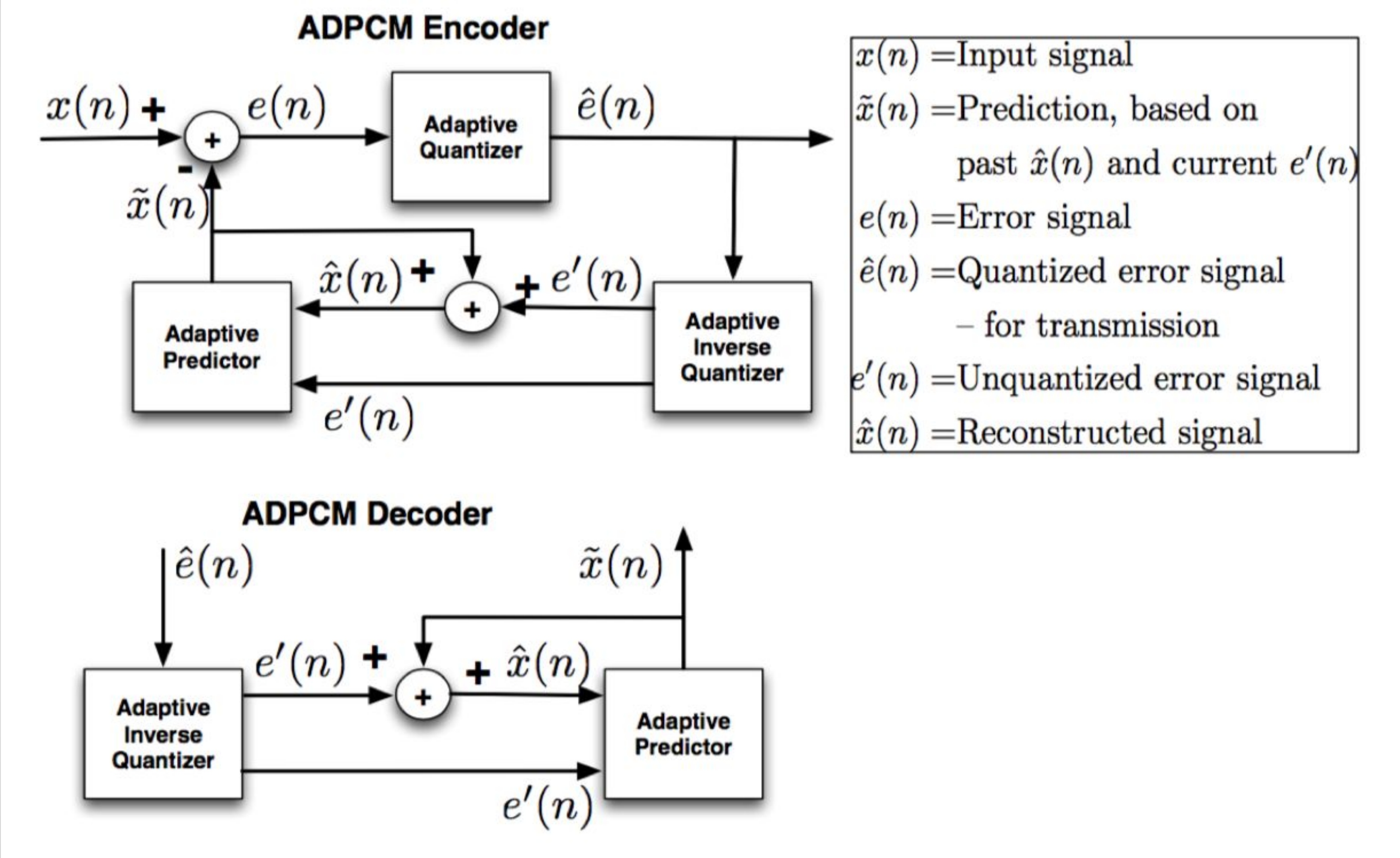
A-Law -- 13 bit signed sample → 8 bit quantized sample
u-Law -- 14 bit 2’s complement → 8 bit quantized sample

1 | Amplitude - Log Companding (PCM)



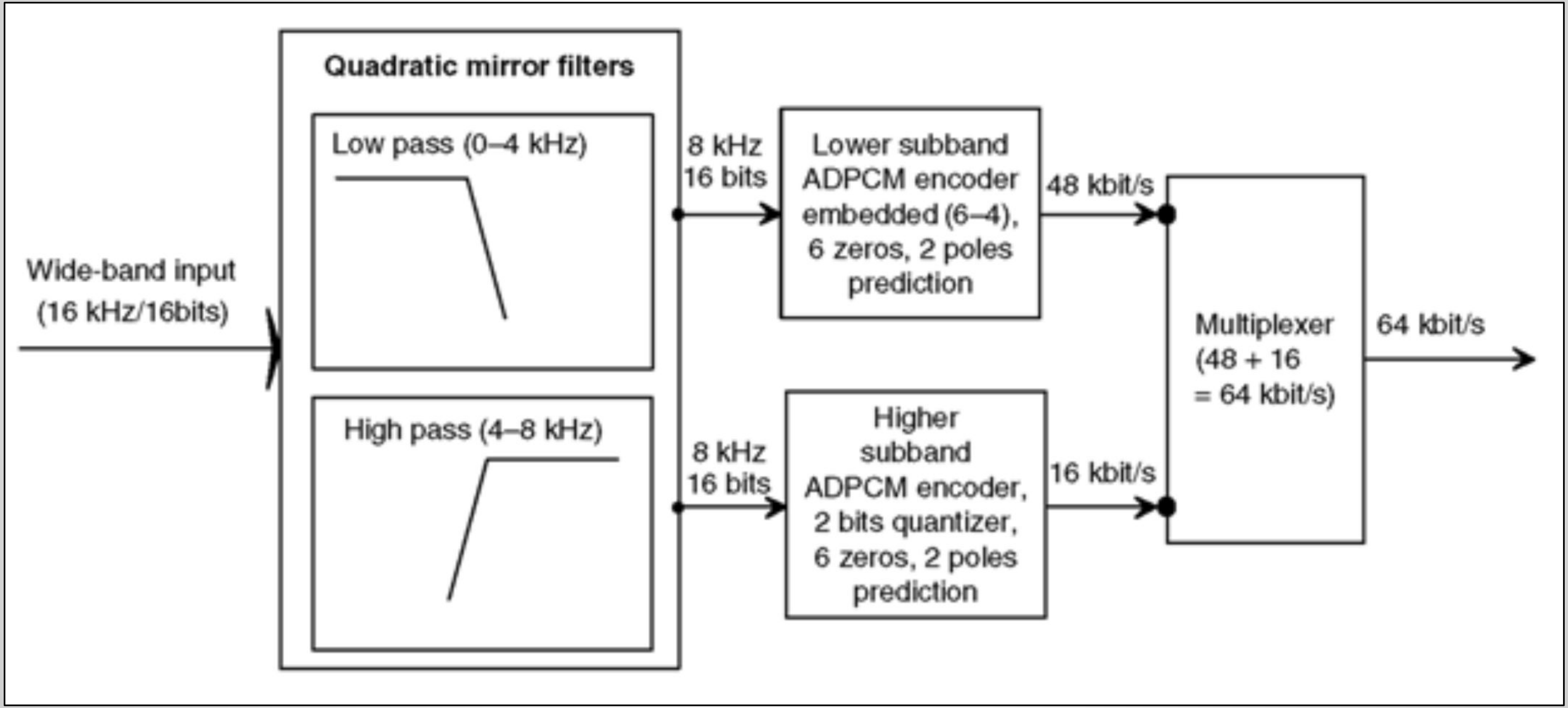
2 - Time - Adaptive, Delta Quantization (ADPCM)

2 | Time - Adaptive, Delta Quantization (ADPCM)

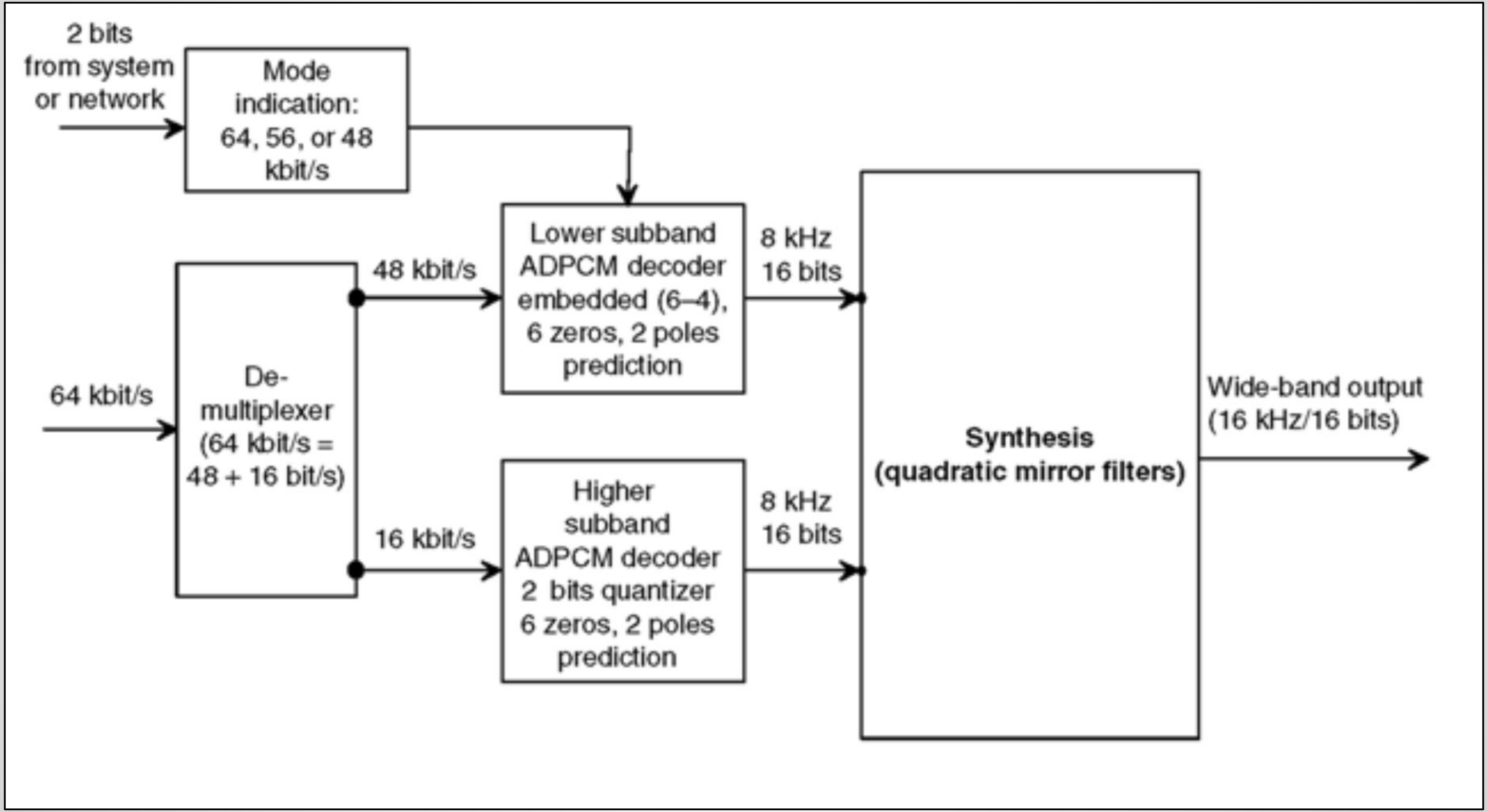


3 - Frequency - ADPCM in multiple subbands (Subband ADPCM)

3 | Frequency - ADPCM in multiple subbands



3 | Frequency - ADPCM in multiple subbands



Lesson 3

Parametric Codecs

Lesson 3 | Parametric Codecs

1 – Speech production model

2 – Linear Predictive Coding of Speech

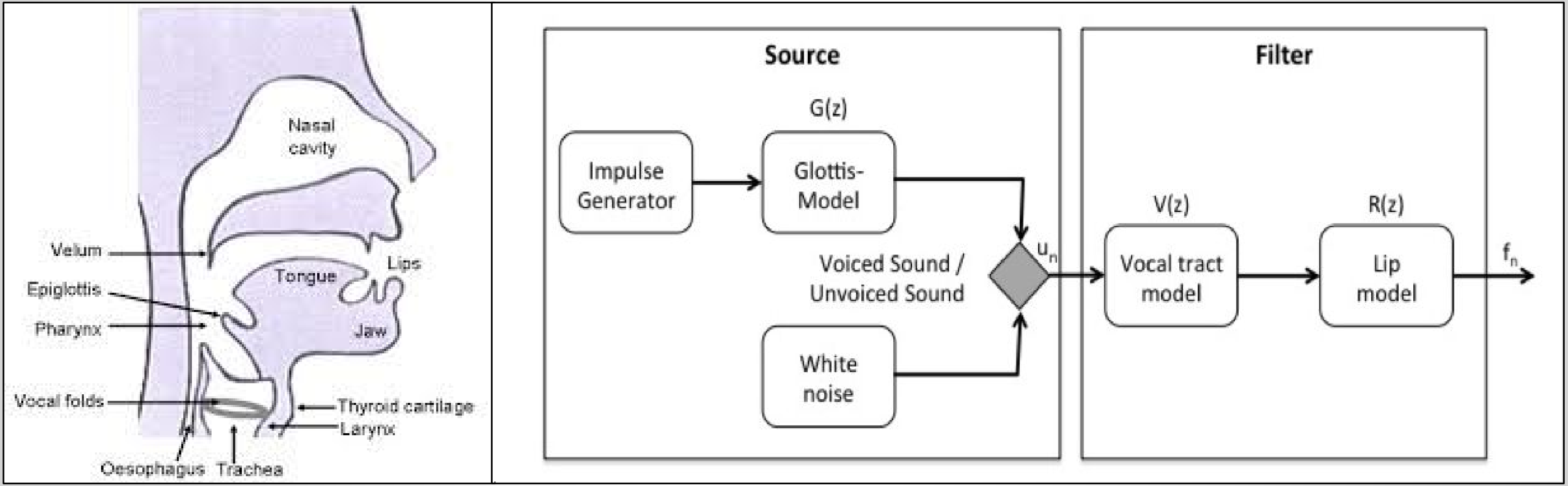
3 – 2G/3G Codecs

4 – Voice Quality

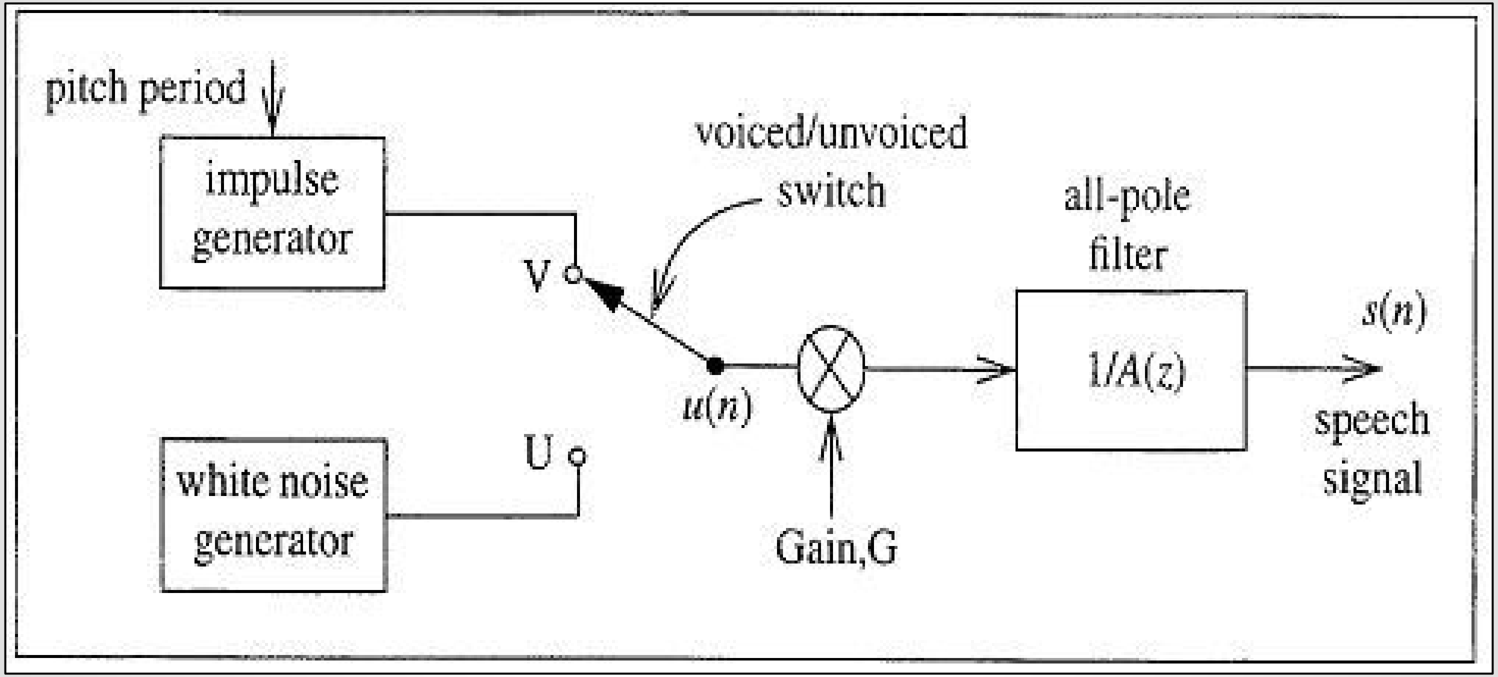
5 – VoIP over Wireless

1 - Speech production Model

1 | Speech production model



1 | Speech production model



0.5 to 1.5 bits / sample instead of 8 bits / sample

2 - Linear Predictive Coding of Speech

2 | Linear Predictive Coding of Speech

LPC Model of Speech Production

$$x(n) = \sum_{i=1}^{p} a(i)\, x(n-i) \quad + G\, u(n)$$

Given a speech segment $x(n), n = 0..N-1$, the task is to estimate

- the LPC filter $a(i), i = 1..p$ that approximates the vocal tract filter;
- input to the filter $u(n)$ that approximates the source; and
- the gain G for the source excitation signal.

2 | Linear Predictive Coding of Speech

LPC Formulation

$x(n)$ is predicted as linearly weighted sum of past samples

$$x(n) \approx \tilde{x}(n) = \sum_{i=1}^{p} a(i)\, x(n-i)$$

The error captures the ignorance missed by the model

$$\begin{aligned} e(n) &= x(n) - \tilde{x}(n) \\ &= x(n) - \sum_{i=1}^{p} a(i)\, x(n-i) \end{aligned}$$

$$E_n = \sum_{N} [e(n)]^2$$

Minimize E_n by differentiating E_n w.r.t each $a(i)$ and set to 0

2 | Linear Predictive Coding of Speech

Autocorrelation Method

The set of equations for LPC parameters $a(i)$ are

$$\sum_{k=1}^{p} a(i)R(|i-k|) = R(i) \quad 1 \leq i \leq p,$$

where, $R(i)$, the autocorrelation at lag i is given by

$$R(i) = \sum_{m=0}^{N-1-i} x(m)x(m+i)$$

The matrix $R(|i-k|), 1 \leq k \leq p, 1 \leq i \leq p$ is a Toeplitz matrix in which each descending diagonal from left to right is constant, i.e. $R_{i,j} = R_{i+1,j+1} = R(|i-k|)$. Levinson-Durbin recursive solution is an efficient method to solve for $a(i)$.

2 | Linear Predictive Coding of Speech

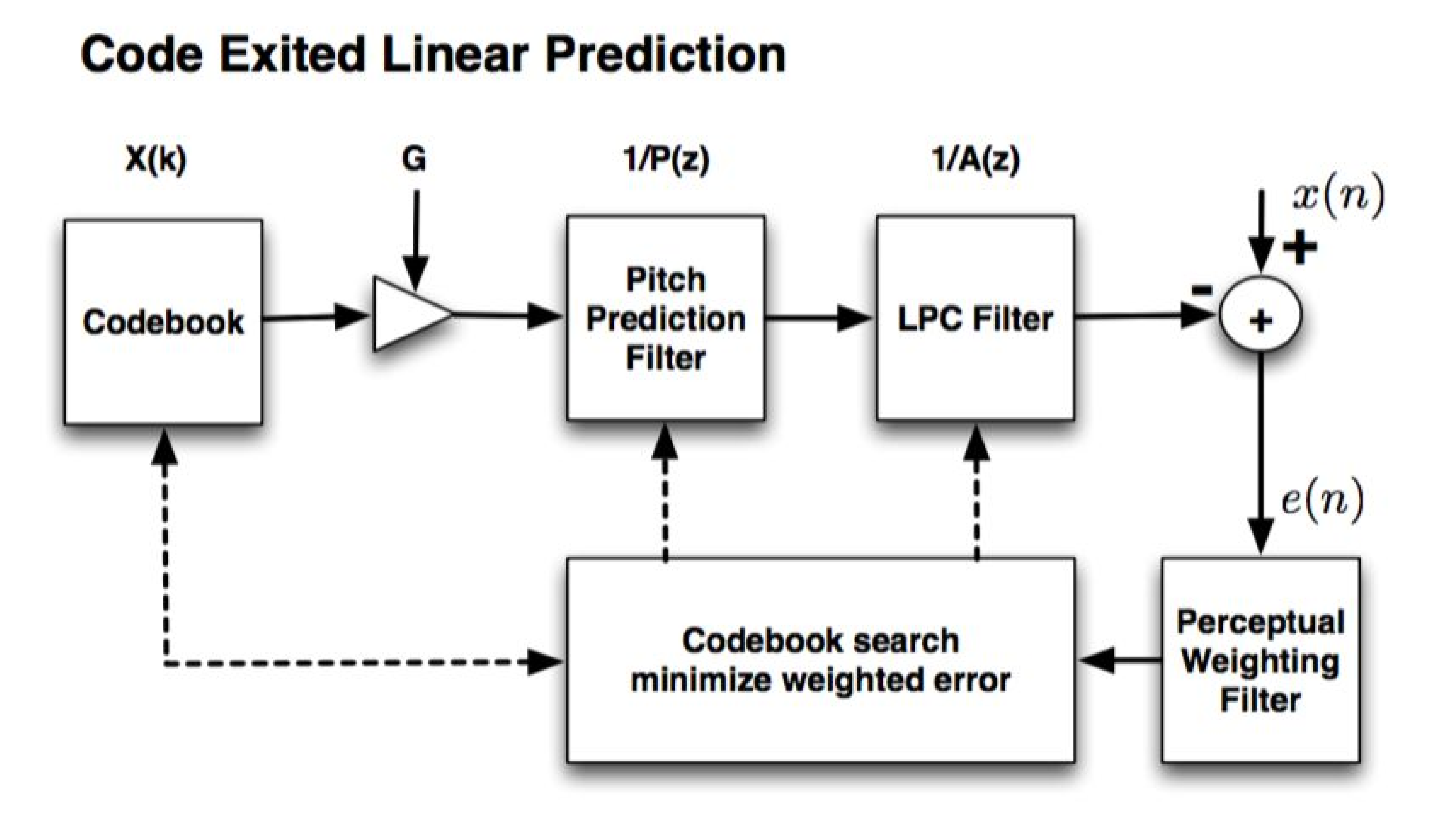
Source: Gain and $u(n)$

$$G^2 = R(0) - \sum_{k-1}^{p} a(k)R(k)$$

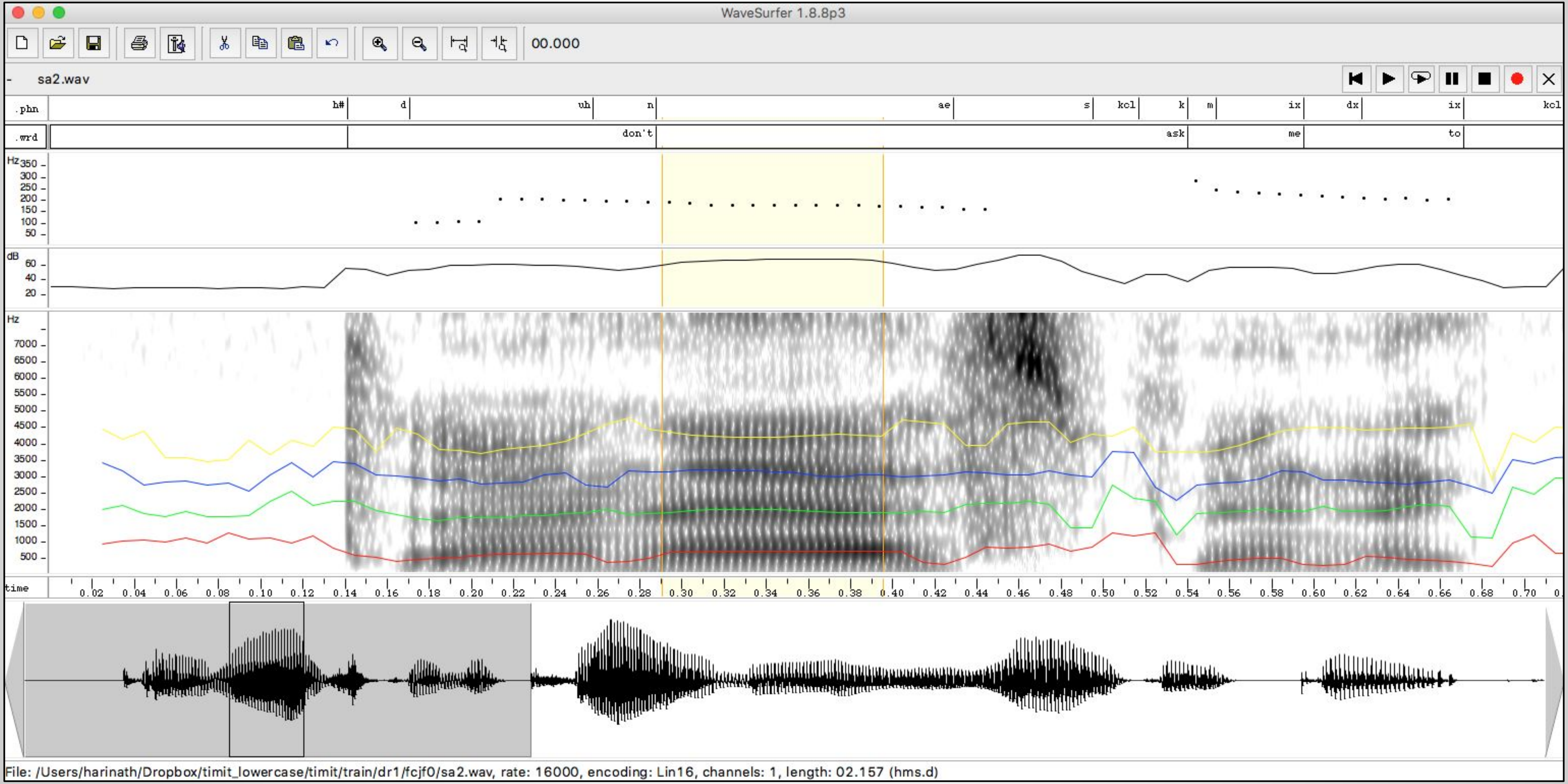
- $u(n)$ is bimodal – Voiced and Unvoiced. Voiced segments are periodic with the pitch period.
- Pitch period can be estimated using autocorrelation of low passed speech, center clipped speech, etc [1].
- When there is discernible peak in the autocorrelation function, it is labelled Unvoiced.
- $u(n)$ is a train of impulses at pitch period for voiced signals.
- $u(n)$ is random noise for unvoiced signals.

[1] Lawrence Rabiner, "On the Use of Autocorelation Analysis for Pitch Detection" IEEE Transactions On Acoustics, Spech, and signal Procesing,Vol. ASSP-25, No.1, February 1977

3 - 2G/3G Codecs

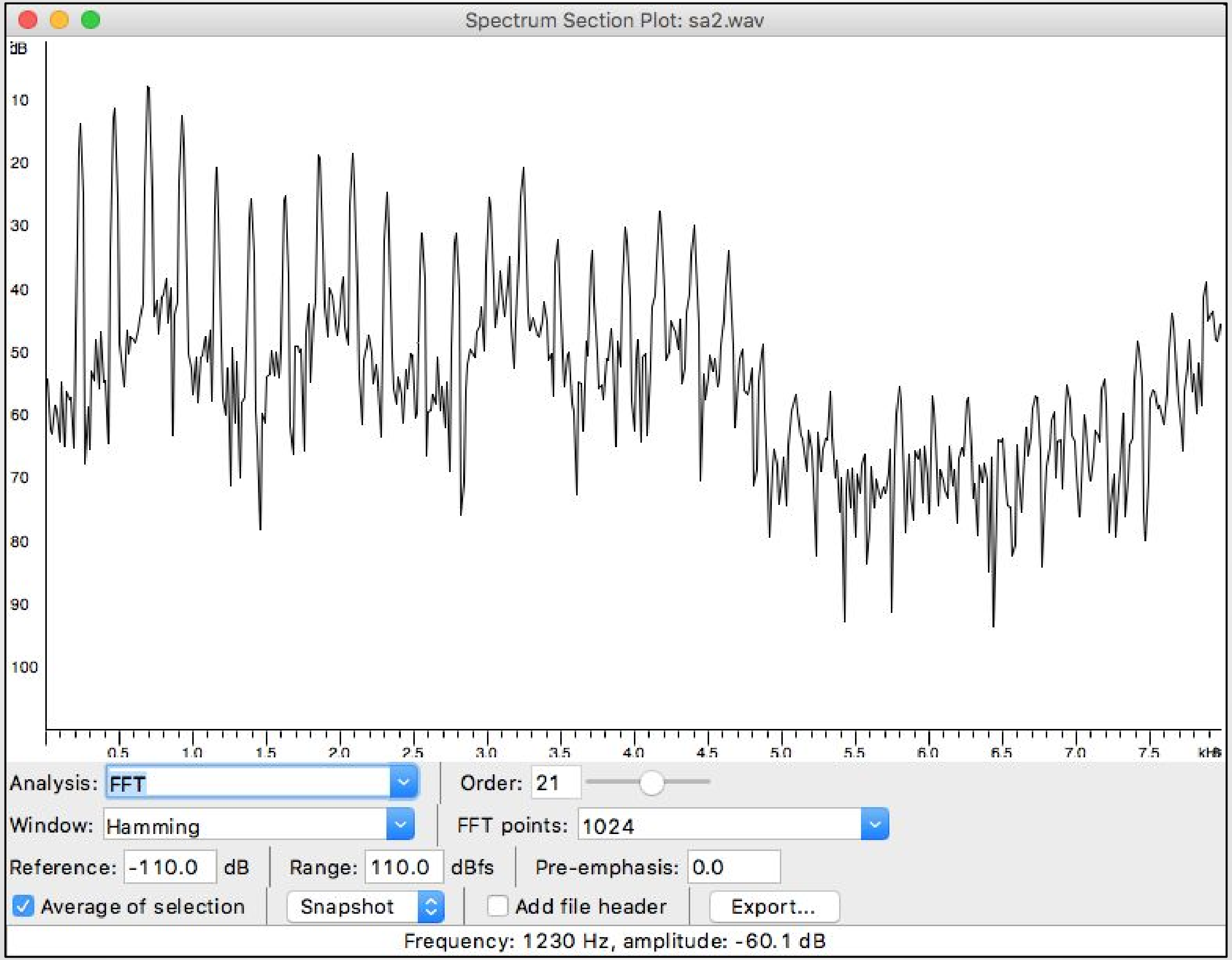

Code Exited Linear Prediction
X(k)
G
1/P(z)
1/A(z)
$x(n)$
+
Codebook
Pitch Prediction Filter
LPC Filter
-
+
$e(n)$
Codebook search minimize weighted error
Perceptual Weighting Filter

3 | 3G Codecs



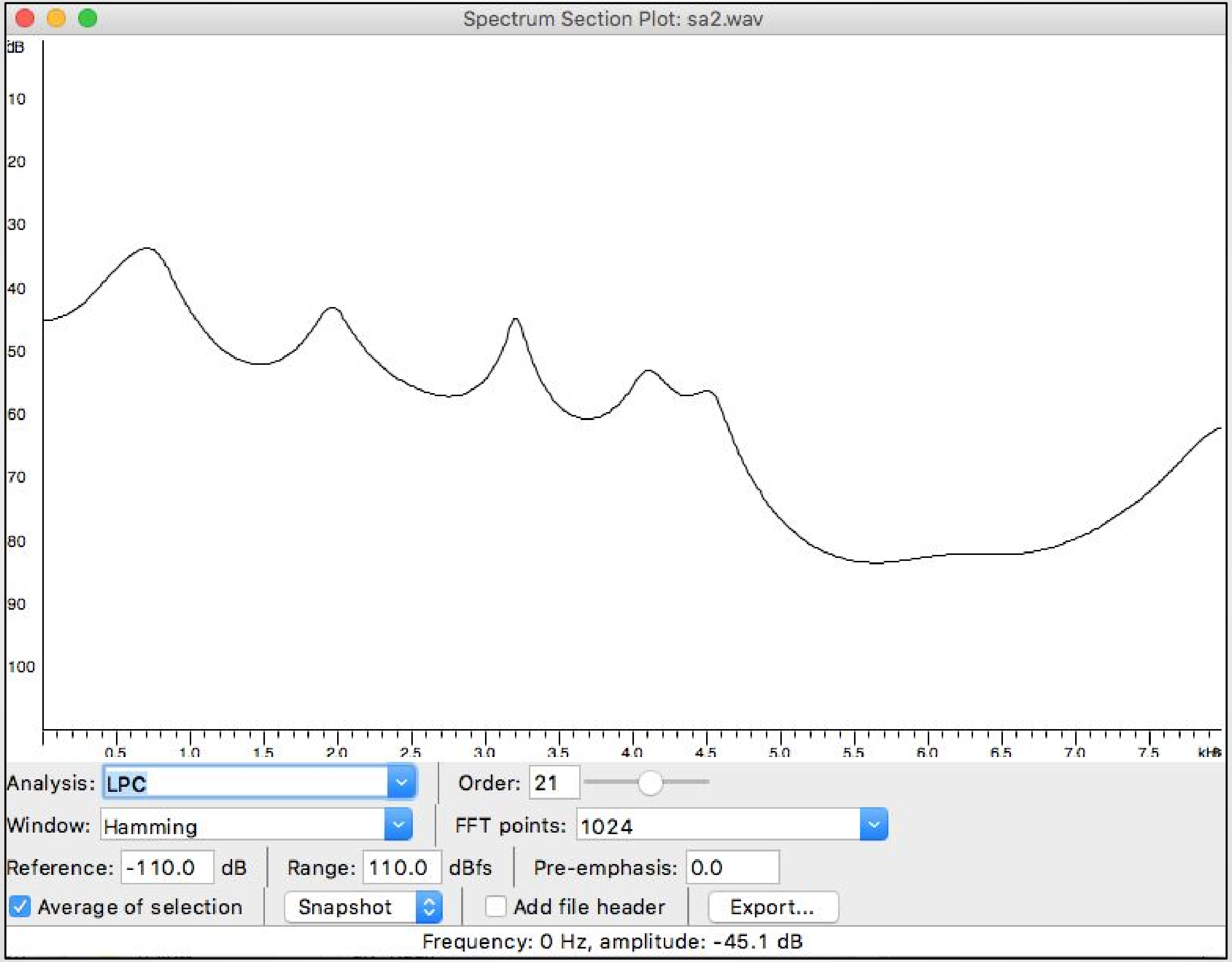
WaveSurfer 1.8.8p3
00.000
sa2.wav
.phn
h#
d
uh
n
ae
s
kcl
k
m
ix
dx
ix
kcl
.wrd
don't
ask
me
to
Hz
dB
time
File: /Users/harinath/Dropbox/timit_lowercase/timit/train/dr1/fcjf0/sa2.wav, rate: 16000, encoding: Lin16, channels: 1, length: 02.157 (hms.d)

3 | 3G Codecs



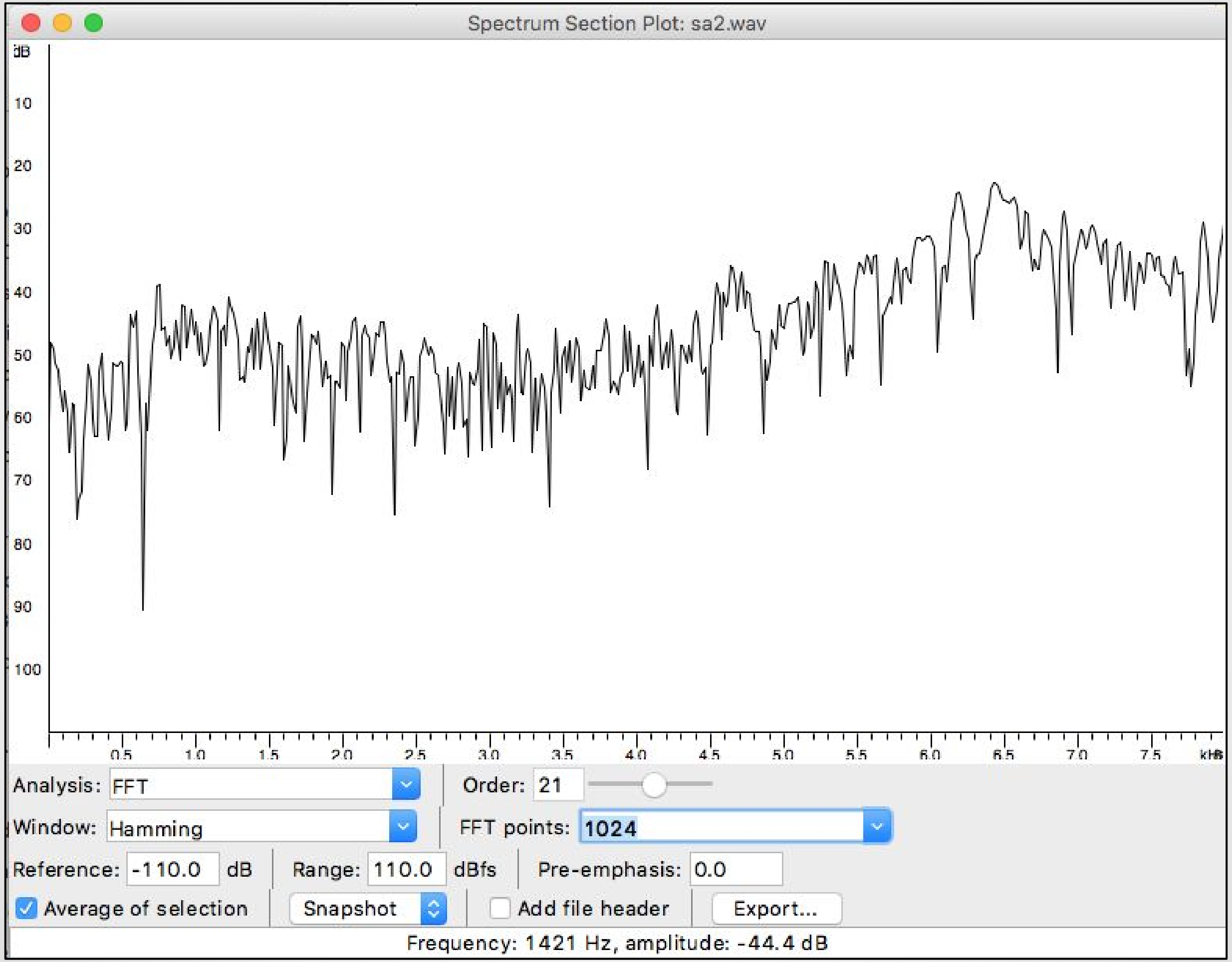
Spectrum Section Plot: sa2.wav
dB
10
20
30
40
50
60
70
80
90
100
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
kHz
Analysis: FFT
Order: 21
Window: Hamming
FFT points: 1024
Reference: -110.0 dB
Range: 110.0 dBfs
Pre-emphasis: 0.0
Average of selection
Snapshot
Add file header
Export...
Frequency: 1230 Hz, amplitude: -60.1 dB

3 | 3G Codecs



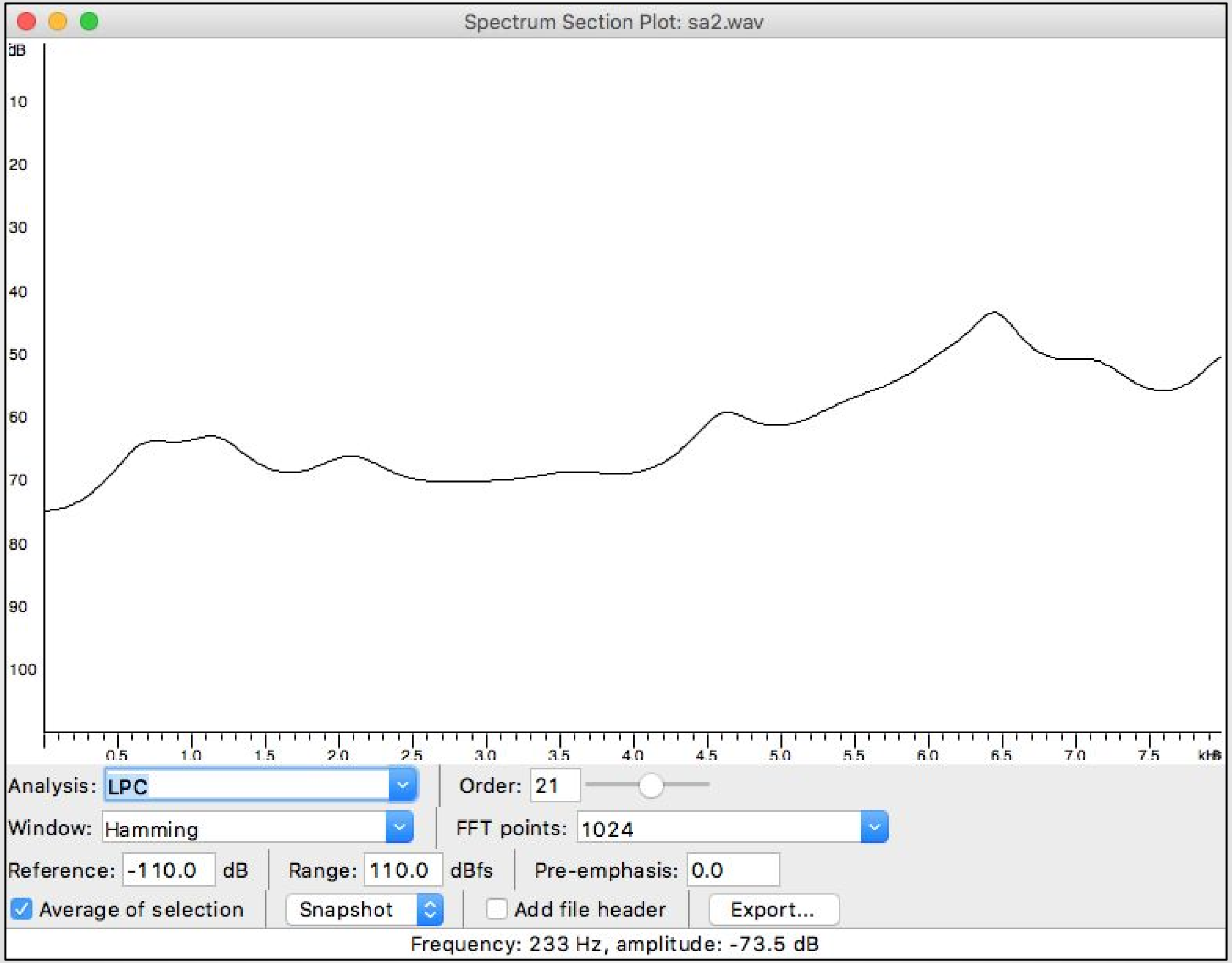
Spectrum Section Plot: sa2.wav
dB
10
20
30
40
50
60
70
80
90
100
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
kHz
Analysis: LPC
Order: 21
Window: Hamming
FFT points: 1024
Reference: -110.0 dB
Range: 110.0 dBfs
Pre-emphasis: 0.0
Average of selection
Snapshot
Add file header
Export...
Frequency: 0 Hz, amplitude: -45.1 dB

3 | 3G Codecs



Spectrum Section Plot: sa2.wav
dB
10
20
30
40
50
60
70
80
90
100
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
kHz
Analysis: FFT
Order: 21
Window: Hamming
FFT points: 1024
Reference: -110.0 dB
Range: 110.0 dBfs
Pre-emphasis: 0.0
Average of selection
Snapshot
Add file header
Export...
Frequency: 1421 Hz, amplitude: -44.4 dB

3 | 3G Codecs



Spectrum Section Plot: sa2.wav
dB
10
20
30
40
50
60
70
80
90
100
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
kHz
Analysis: LPC
Order: 21
Window: Hamming
FFT points: 1024
Reference: -110.0 dB
Range: 110.0 dBfs
Pre-emphasis: 0.0
Average of selection
Snapshot
Add file header
Export...
Frequency: 233 Hz, amplitude: -73.5 dB

4 - Voice Quality

4 | Voice Quality

Assessing Voice Quality

$$\text{SNR} = 10 \log_{10} \left\{ \frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} (x(n) - \tilde{x}(n))^2} \right\}$$

$$\text{SEGSNR} = \frac{10}{K} \sum_{k=0}^{K-1} \log_{10} \left\{ \frac{\sum_{n=0}^{N-1} x^2(kN+n)}{\sum_{n=0}^{N-1} (x(kN+n) - \tilde{x}(kN+n))^2} \right\}$$

4 | Voice Quality

Mean opinion score (MOS)

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

4 | Voice Quality

https://en.wikipedia.org/wiki/Mean_opinion_score

Codec	Data rate [kbit/s]	Mean opinion score (MOS)
G.711 (ISDN)	64	4.1
iLBC	15.2	4.14
AMR	12.2	4.14
G.729	8	3.92
G.723.1 r63	6.3	3.9
GSM EFR	12.2	3.8
G.726 ADPCM	32	3.85
G.729a	8	3.7
G.723.1 r53	5.3	3.65
G.728	16	3.61
GSM FR	12.2	3.5

4 | Voice Quality

AMR-WB MOS values for different rates

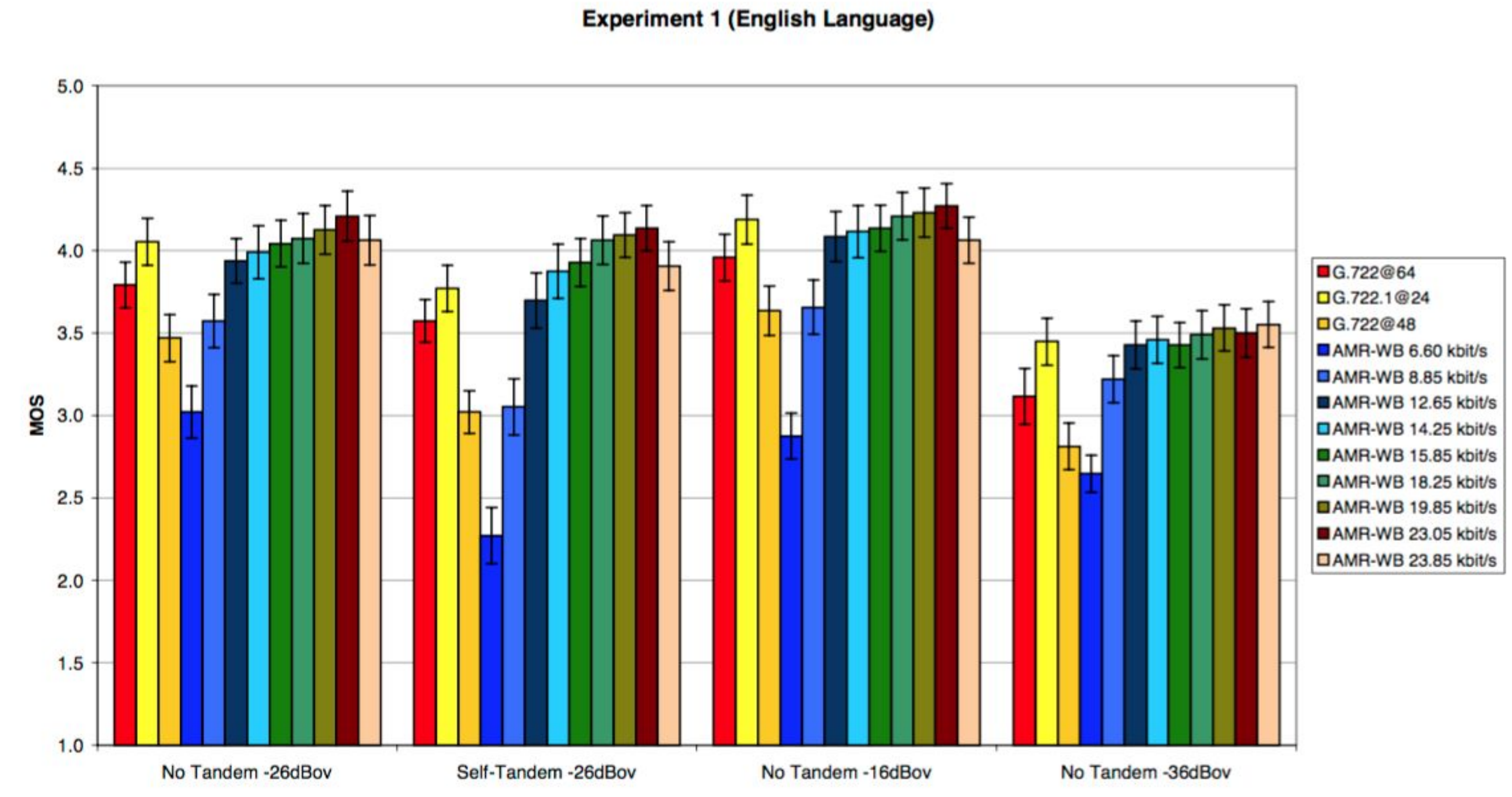
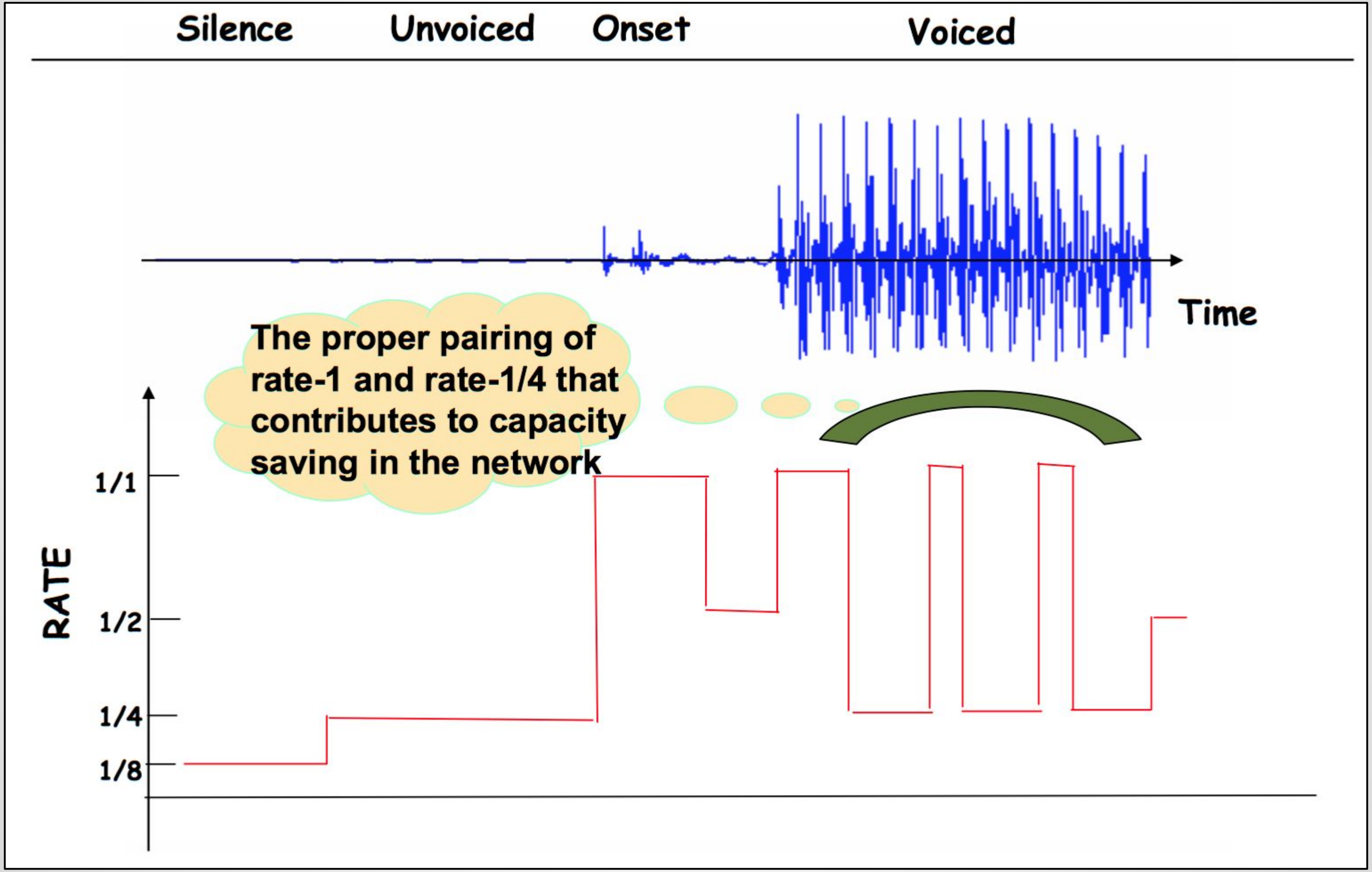


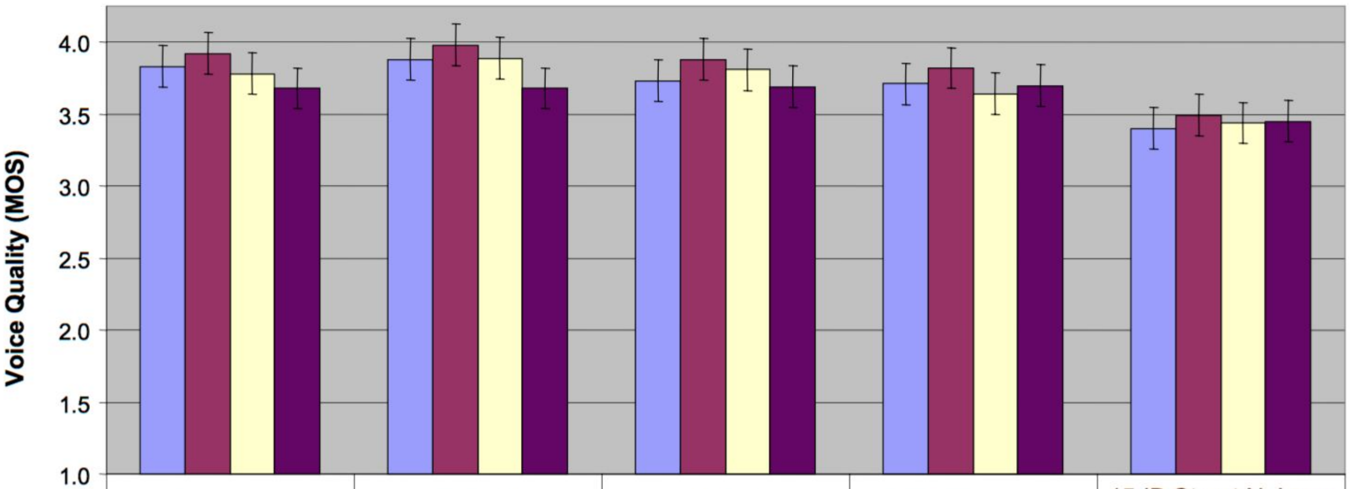
Figure 8.2: Experiment 1, testing tandeming and input levels with English language

4 | Voice Quality

EVRC Variable Rate Operation



4 | Voice Quality EVRC MOS values for different average data rates



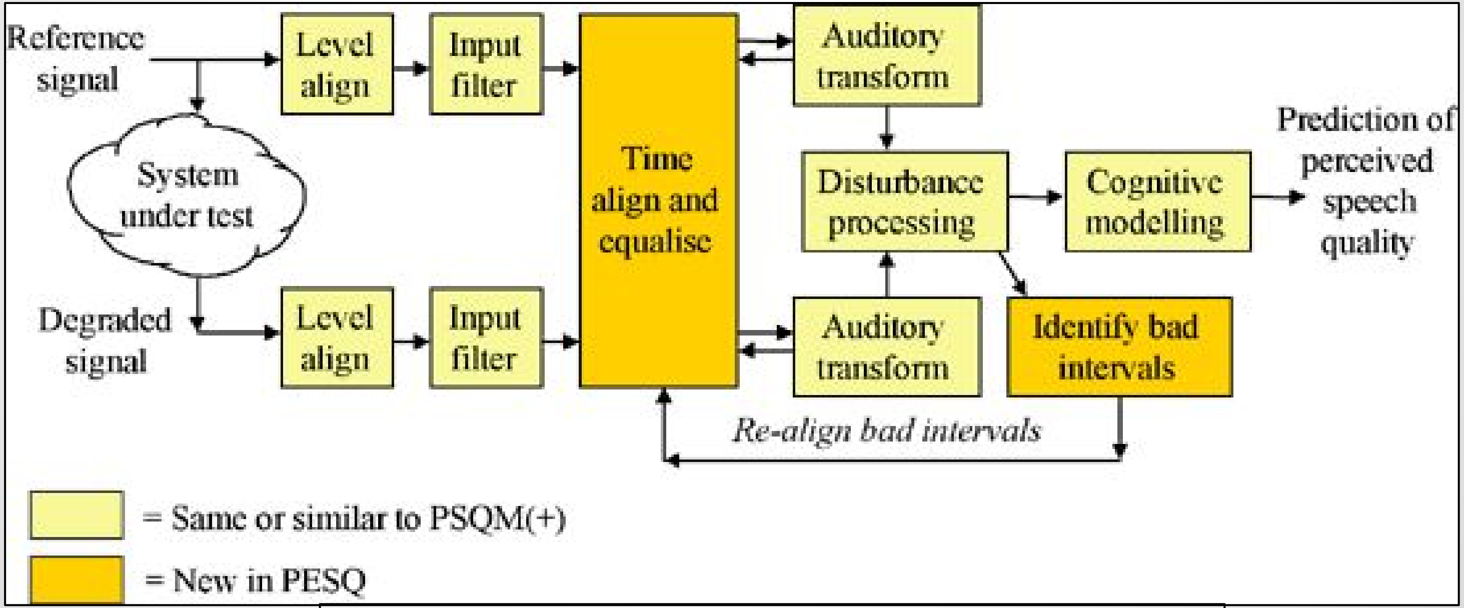
	High Input Level	0% FER	1% FER	EVRC Tandem	15dB Street Noise + 1% FER
EVRC	3.83	3.88	3.73	3.71	3.40
EVRC-B M0	3.92	3.98	3.88	3.82	3.49
EVRC-B M4	3.78	3.89	3.81	3.64	3.44
EVRC-B M6	3.68	3.68	3.69	3.70	3.45

Source data from 3GPP2 C11-20060327-015, Dynastat EVRC-B Characterization Test Report

M0 = 9.3 kbps; M4 = 6.6 kbps; M6 = 5.8 kbps

Other Objective Metrics

- **PESQ**, *Perceptual Evaluation of Speech Quality*, ITU-T recommendation P.862
- **POLQA,** *Perceptual Objective Listening Quality Assessment*, ITU-T recommendation P.863



http://www.opticom.de/technology/pesq.php

4 | Voice Quality

3GPP2-C11-20080114-016, “Report and analysis of PESQ under-prediction of EVRC family of speech codecs”, Hollywood, CA, Jan 14-18, 2008

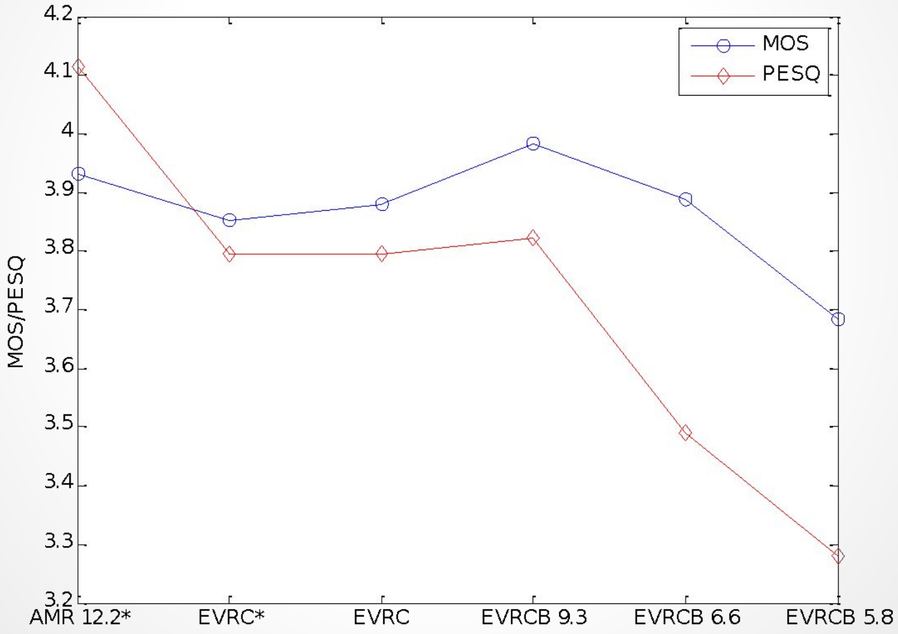
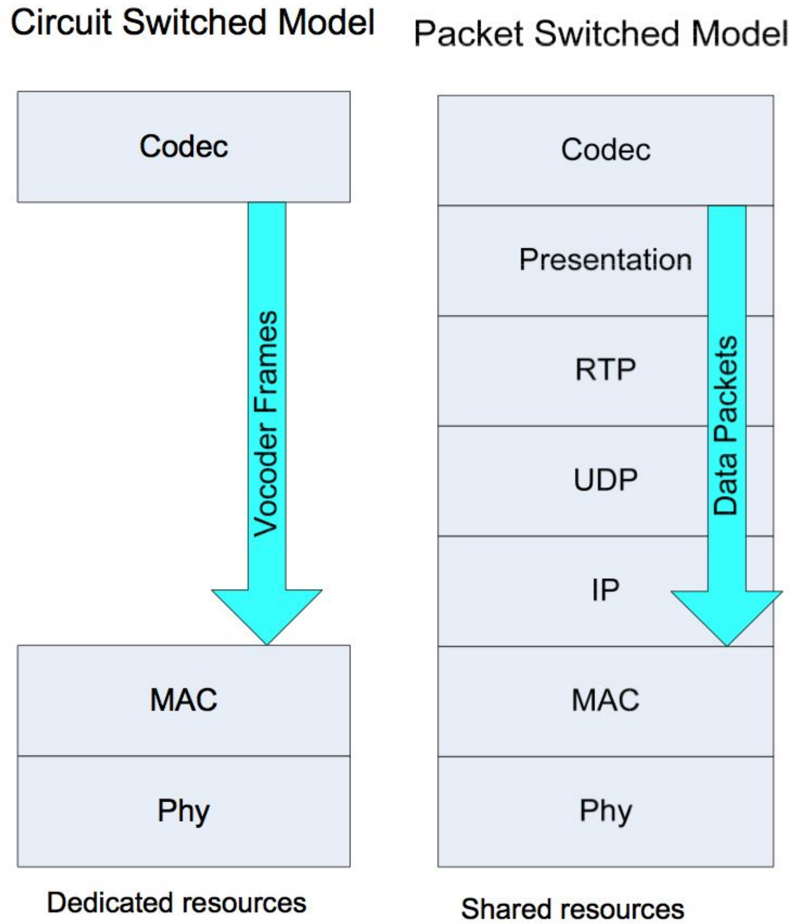


Figure 6 Comparison of PESQ and MOS for different codecs under 0% frame erasure

5 - VoIP over Wireless

5 | VoIP over Wireless



Implications

- Overheads at each layer
- Additional latency
- Jitter in packet arrival
- Fragmentation and loss of Goodput
-

5 | VoIP over Wireless

- **RFC 3550**: *RTP*: A Transport Protocol for Real-Time Applications
- **RFC 3551**: *RTP* Profile for Audio and Video Conferences with Minimal Control
- **RFC 4867**: *RTP* Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs
- **RFC 5188**: *RTP* Payload Format for the Enhanced Variable Rate Wideband Codec (EVRC-WB) and the Media Subtype Updates for EVRC-B Codec

5 | VoIP over Wireless

RFC 4995 and RFC 5225: RObust Header Compression (ROHC)

- At 4—12 kbps payload traffic, we have a total of 500—1500 bytes per second
- If we transmit data at 20 msec frames, we have a total of 50 frames / second
- So, we have 10—30 bytes in each VoIP packet
- In *IPv4*, the overhead per packet is 40 bytes (IP=20; UDP=8; RTP=12)
- In *IPv6*, the overhead per packet is 60 bytes (IP=40; UDP=8; RTP=12)

- *ROHC* compresses *VoIP* headers to 1—3 bytes
- Robust to lossy channels, long round trip delays, out of sequence arrivals
- Takes advantage of static fields and fields with known properties (e.g. Timestamp)
- Applied to only low bandwidth links (e.g. cellular)

5 | VoIP over Wireless

Further Study

1. **G.191 : Software tools for speech and audio coding standardization** https://www.itu.int/rec/T-REC-G.191-201003-I/en
2. **Digital Processing of Speech Signals,** Lawrence R. Rabiner, Ronald W. Schafer https://books.google.com/books/about/Digital_Processing_of_Speech_Signals.html?id=wltTAAAAMAAJ

Back up slides