Low Delay, Low Complexity Voice Coding and Bounding the Rate Distortion Performance of Voice Codecs

Jerry D. Gibson

Department of Electrical and Computer Engineering
University of California, Santa Barbara
gibson@ece.ucsb.edu

This research has been supported by NSF Grant Nos. CCF-0728646 and CCF-0917230

Voice Coding Problem

- Develop a Low Average bit-rate speech coder with:
  - Low Delay—Less than 10 msec
  - Low Complexity—Less Complex than CELP
  - Good Speech Quality and Intelligibility—PESQ MOS Near 4.0 for Narrowband Speech
  - Good Tandem performance with standard codecs such as G.729 and AMR—Small drop in PESQ MOS in either direction

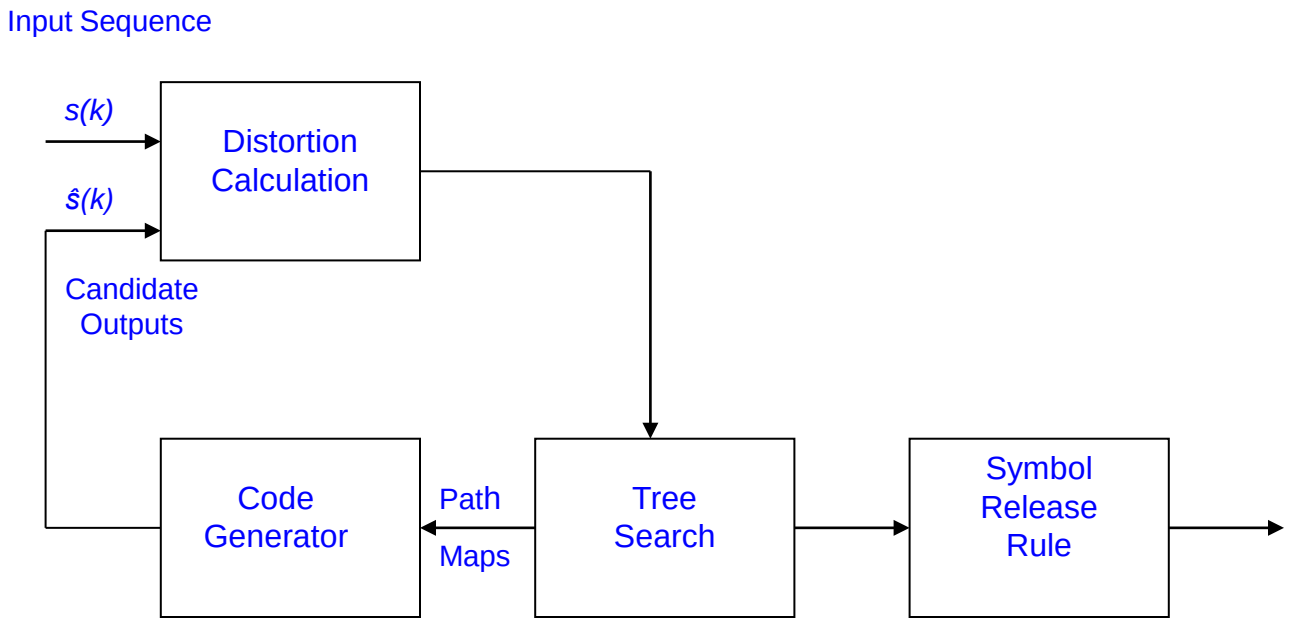
UCSB

Approach – Tree Coding

- Tree coding method – Search all paths to depth *L* and find the best fit
- Advantages of Tree Coding
  - Asymptotically optimal with increasing search depth
  - Less delay than CELP
  - Perceptually weighted error measure
  - Nearer waveform following than CELP

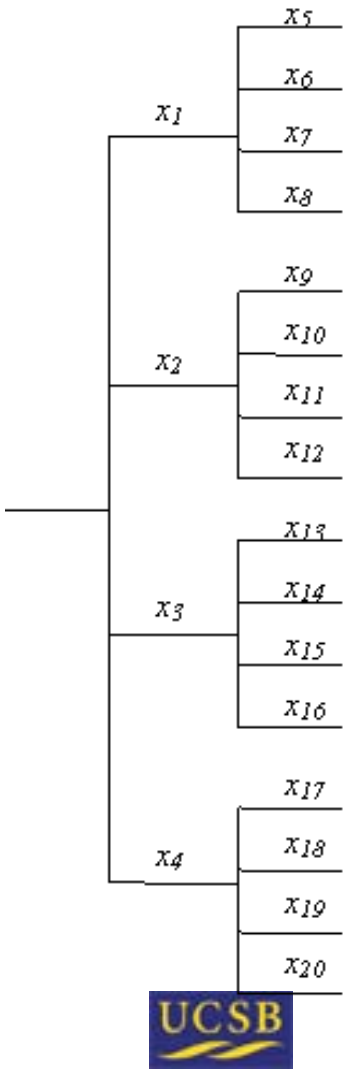
UCSB

Functional diagram of Tree Coder



- Functional blocks
  - Code generator
  - Tree search
  - Distortion calculation
  - Symbol Release Rule

ADPCM Code Tree, Four Level Quantizer

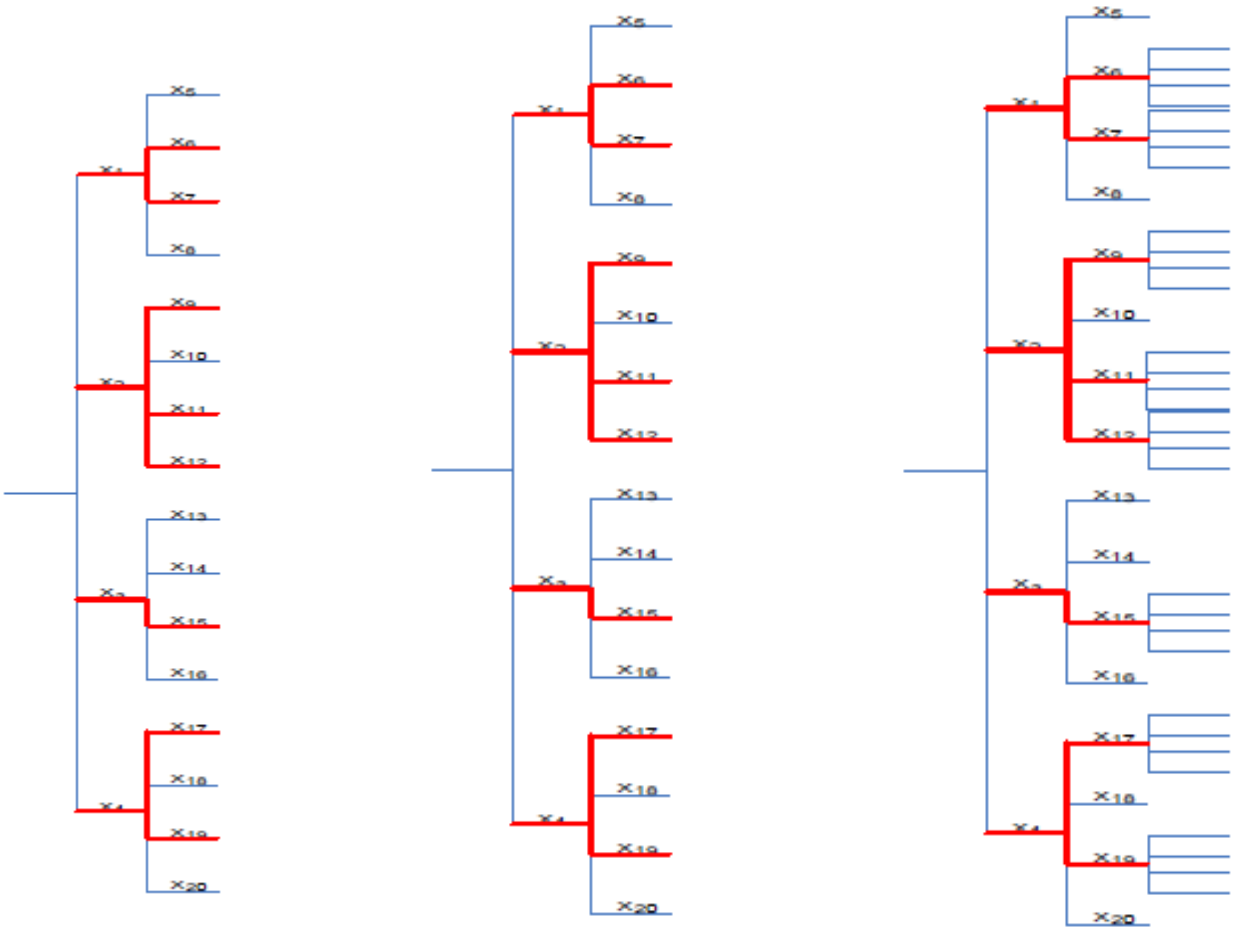


$R = {}^{1}/_{1} \log_2 4 = 2$ bits / symbol

M-L search in Tree coder

- Tree Search uses *M-L*-algorithm with *M* = 4 and *L* = 10
- *M-L* algorithm is used for tree coding to reduce the complexity of searching all possible paths
- Steps in *M-L* algorithm:
  - *M* most likely paths which yield minimum error for encoding previous sample are stored
  - Those paths are extended out to depth *L* for encoding current sample
  - The first symbol of the path giving minimum distortion is encoded

Tree Extension and Pruning



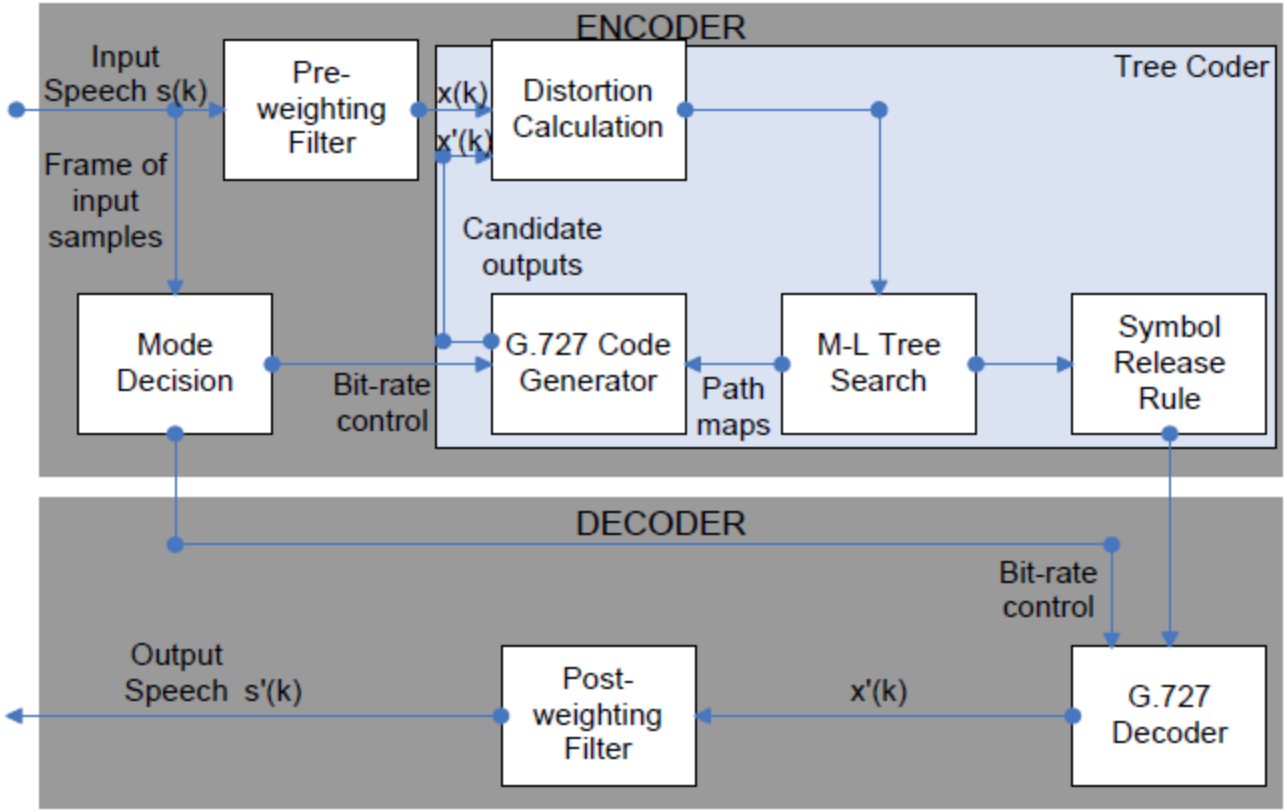
Distortion Calculation in Tree Coder

- Perceptually weighted error is used as the distortion measure
  - Weights error along each chosen path
  - Perceptual error weighting filter:

$$W(z) = \frac{1 - \sum_{i=1}^{N} a_i z^{-i}}{1 - \sum_{i=1}^{N} \mu^i a_i z^{-i}}$$ where, $\mu = 0.86$ and a_i 's are the predictor coefficients

- Symbol Release Rule--First Symbol corresponding to minimum error path is sent

G.727 Code Generator



Multimode Speech Coding

- Classify speech into four different modes:
  - Silence
  - Unvoiced
  - Voiced
  - Onset
- Code Each Mode Differently
- Advantage—Lower Average Bit Rate
- Disadvantages
  - Increased Delay
  - Increased Complexity

Results – Clean speech

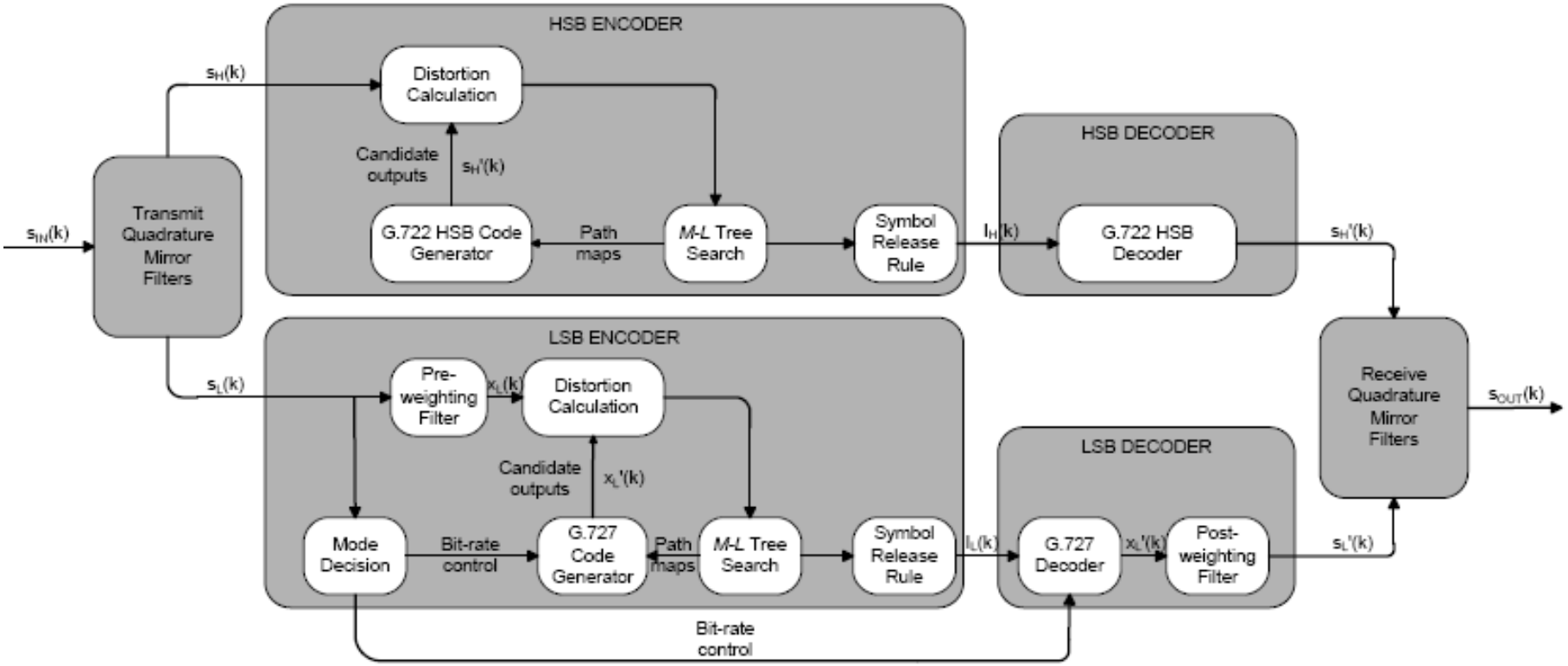
Sequence (PESQ)	MM Tree Coder	AMR-NB 12.2 kbps	G.727 32 kbps
T04	3.844	3.772	3.771
T05	3.958	4.029	3.901
T06	3.820	3.875	4.002
T07	3.912	3.732	4.007
T08	4.001	4.064	4.087
T12	3.848	3.898	3.819
T13	3.924	4.104	3.852
average	3.901	3.925	3.920

UCSB

Rate, Delay, Complexity Comparisons

Speech Coder Attributes	MM Tree Coder	AMR-NB 12.2 kbps	G.727 32 kbps
Avg Bit-rate (kbps)	11.96	<7	32
Delay (ms)	6.125	20	0.125
Complexity (wMOPS)	3.161	16.75 [10]	1.25 [11]

Wideband Tree Coding Structure



Wideband Codec Performance

Sequence	Multimode Tree Coder		AMR-WB 23.05 kbps		G.722 64 kbps	
	WPESQ	bit-rate (kbps)	WPESQ	bit-rate (kbps)	WPESQ	bit-rate (kbps)
F1	3.735	23.21	3.669	11.24	3.832	64
F2	3.980	33.43	3.477	15.33	3.955	64
F3	3.617	35.78	4.055	17.56	4.184	64
M1	4.268	47.77	4.281	22.75	4.397	64
M2	3.569	38.09	3.946	16.26	3.983	64
M3	3.260	36.04	3.742	17.21	3.333	64
Average	3.738	35.72	3.862	16.73	3.947	64

UCSB

Wideband Performance Comparisons

	Multimode Tree Coder	AMR-WB 23.05 kbps	G.722 64 kbps
WPESQ	3.2-4.3	3.4-4.3	3.3-4.4
Avg. Bit-rate (kbps)	23.21-47.77	11.24-22.75	64
Delay (ms)	12.375	25	1.625
Complexity (WMOPS)	12.74	39.0	<10

UCSB

Bit Rate Scalable Wideband Tree Coding

Sequence	3 enhancement bits		2 enhancement bits		1 enhancement bit	
	WPESQ	bit-rate (kbps)	WPESQ	bit-rate (kbps)	WPESQ	bit-rate (kbps)
F1	3.796	26.77	3.592	23.21	3.131	19.65
F2	3.598	38.70	3.459	33.43	3.099	28.15
F3	3.649	41.51	3.601	35.78	3.416	30.05
M1	4.114	55.55	4.163	47.77	3.986	40.00
M2	3.647	44.01	3.533	38.09	3.370	32.16
M3	3.235	41.64	3.134	36.04	3.013	30.44
Average	3.673	41.36	3.580	35.72	3.336	30.08

G.722 Performance

Sequence	G.722 64 kbps	G.722 56 kbps	G.722 48 kbps
F1	3.832	3.705	3.306
F2	3.955	3.925	3.658
F3	4.184	4.176	4.054
M1	4.397	4.393	4.380
M2	3.983	3.901	3.710
M3	3.333	3.318	3.197
Average	3.947	3.903	3.718

UCSB

The Perromance Bounding Problem

- Voice coding is Ubiquitous Today
- No Rate Distortion Performance Bounds for State-of-the-Art Voice Codecs Have Ever Been Obtained
- Rate Distortion Performance Bounds Can
  - Show how much performance can be gained
  - Guide the design of future codecs

Context of the Problem

Gallager [1968] notes that information theory has been more useful for channel coding than for source coding and that the reason, “ . . . appears to lie in the difficulty of obtaining reasonable probabilistic models and meaningful distortion measures for sources of practical interest." He goes on to say, “ . . . it is not clear at all whether the theoretical approach here will ever be a useful tool in problems such as speech digitization “

Prior Work on Rate Distortion Bounds for Speech

- Brehm and Trottler [1986] Spherically Invariant Random Process Models for Speech and MSE Distortion Measure
- Kalveram and Meissner [1988] *R(D)* Bounds for Composite Speech Models, MSE Distortion Measure, and Small *D*
- Kalveram and Meissner [1989] Estimate the Subsource Switch Sequence and Bounds Based on MSE
- De and Kabal [1992] *R(D)* Bounds for Speech Based on the Histogram of Cochlear Firing Values

Rate Distortion Theory for Time Discrete Gaussian Processes

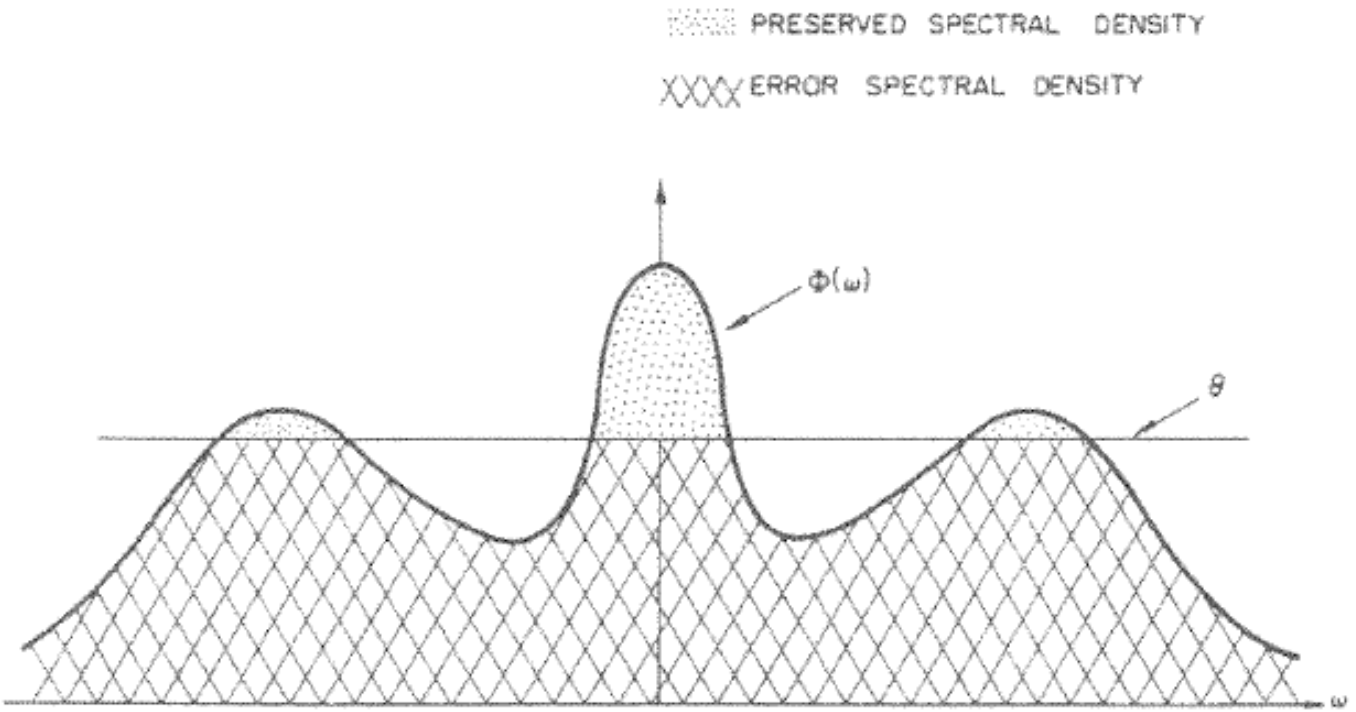
- For a time discrete Gaussian source with given power spectral density and MSE fidelity criterion,

$$R(D_\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \max\left[0, \log \frac{S(\omega)}{\theta}\right] d\omega$$

- where

$$D_\theta = \int_{-\pi}^{\pi} \min\left[\theta, S(\omega)\right] d\omega$$

Reproduced and Error Spectral Densities



Frequency Weighted Error Criteria

- For a frequency weighted squared error fidelity criterion,

$$R(D) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log\left[S(\omega) / \min\left\{ S(\omega), \theta / W(\omega) \right\} \right] d\omega$$

where

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) \min\left[S(\omega), \theta / W(\omega) \right] d\omega$$

Reverse Water-Filling

Theorem 1: Rate distortion Function for a Parallel Gaussian Source

Let $X_i \sim n(0,\sigma_i^2), i=1,2,\ldots,N$ be independent Gaussian random variables and let the distortion measure be

$$D(x^N,\hat{x}^N)=\sum_{i=1}^{N} W_i(x_i-\hat{x}_i)^2$$

Then the rate distortion function is

$$R(D)=\sum_{i=1}^{N}\frac{1}{2}\log\frac{W_i\sigma_i^2}{D_i}$$

where $D_i=\begin{cases}\lambda & if\ \lambda<W_i\sigma_i^2\\ W_i\sigma_i^2 & if\ \lambda\geq W_i\sigma_i^2\end{cases}$

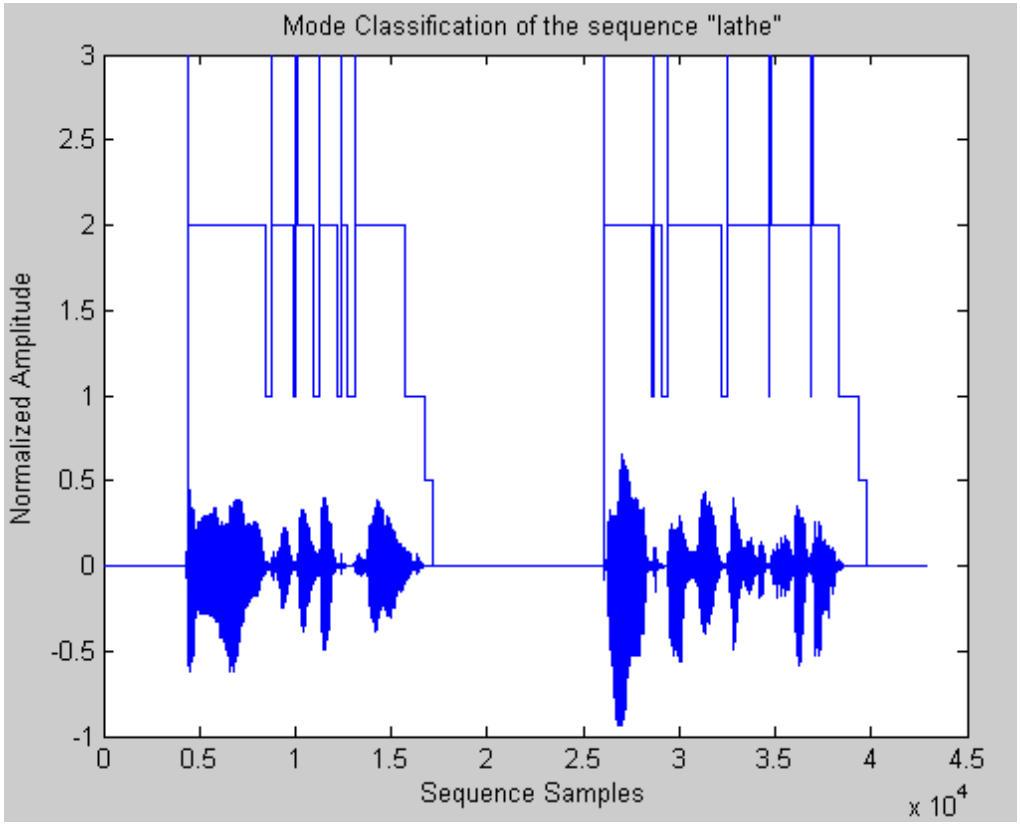
Approach for Speech

- Develop a Composite Source Model
- Use Known *R(D)* Results for MSE
- Map MSE Distortion Measure into PESQ-MOS

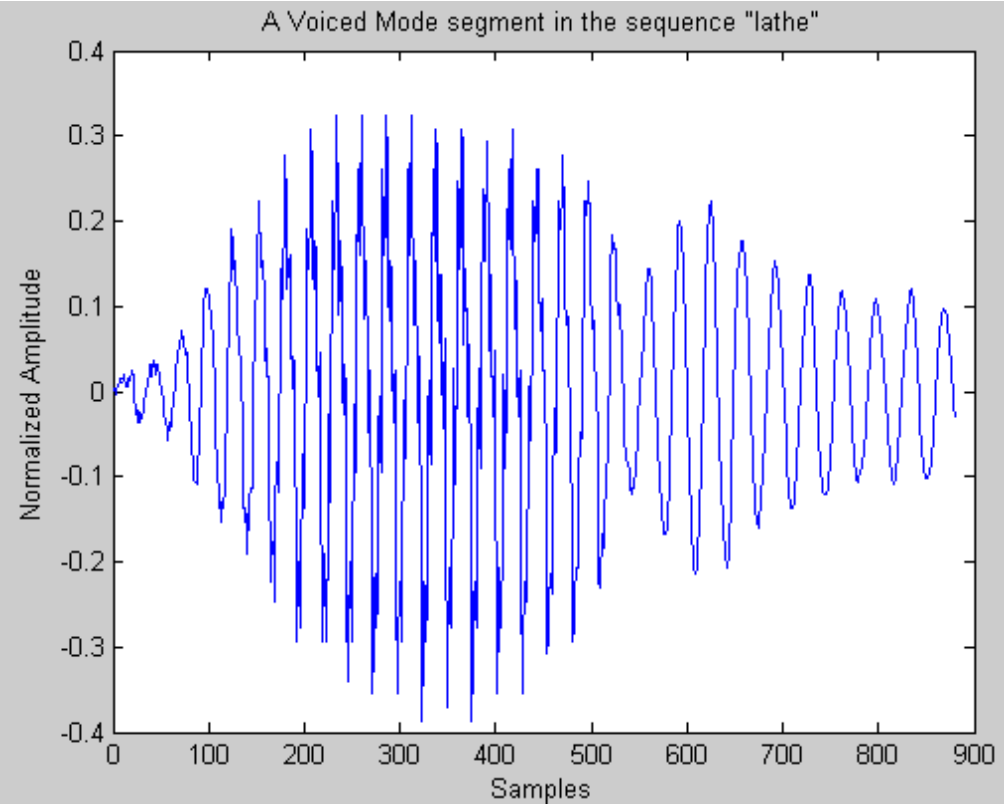
Composite Source Model

- Four Modes plus Silence
  - Voiced—10th Order Autoregressive
  - Onset—4th Order Autoregressive
  - Hangover—4th Order Autoregressive
  - Unvoiced—Memoryless Gaussian
  - Silence coded separately

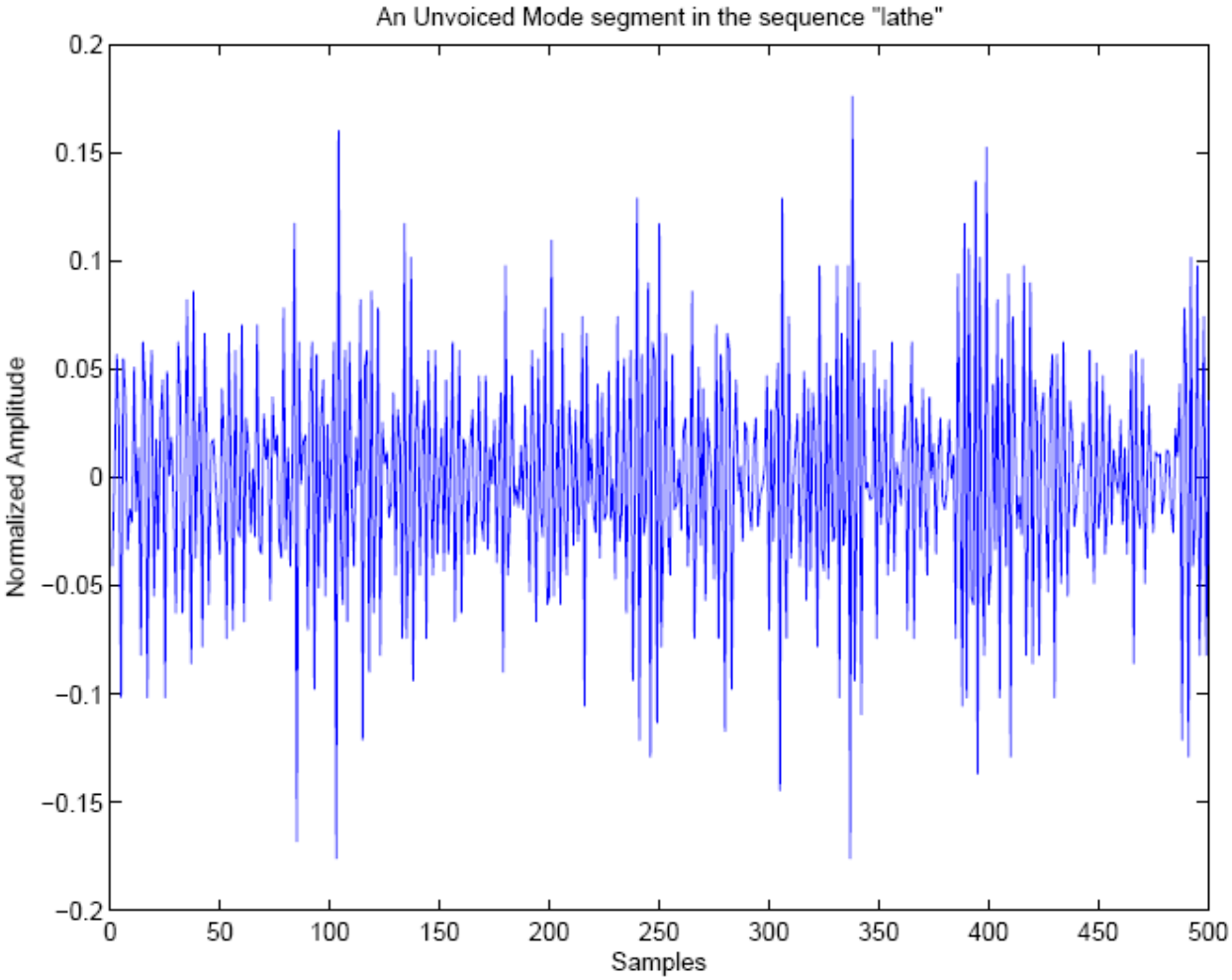
Segments enclosed with values 3, 2, 1, 0.5, 0 are onsets, voiced, unvoiced, hangover, and silence



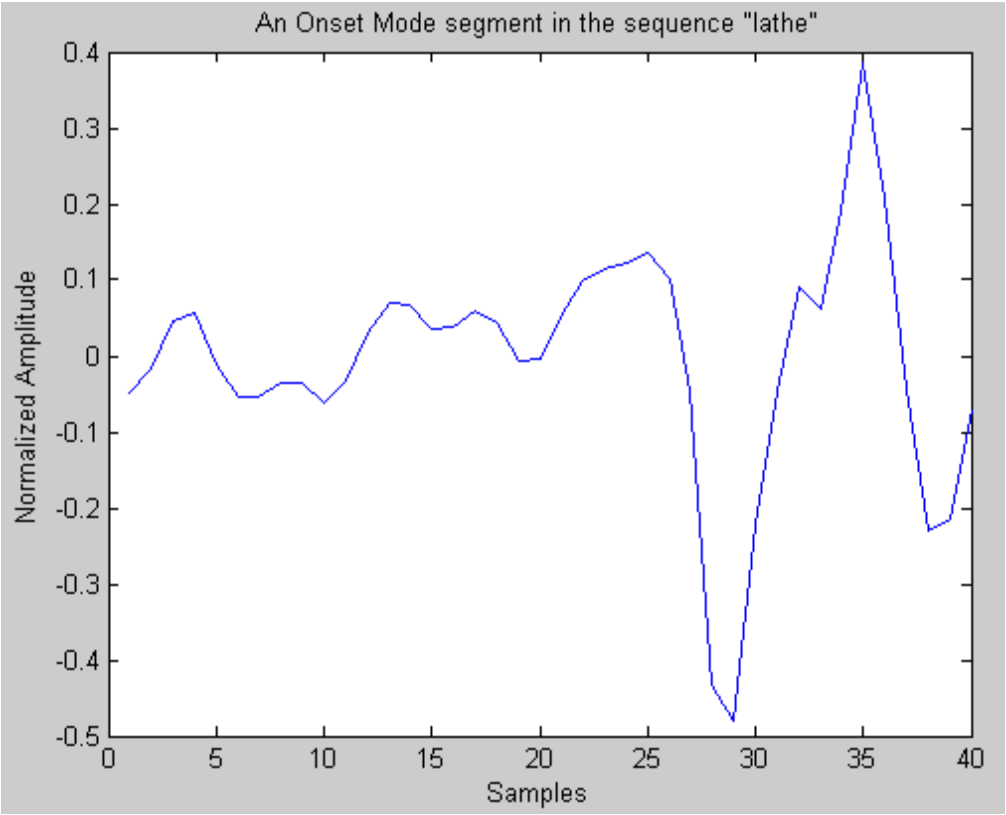
Voiced Mode


A Voiced Mode segment in the sequence "lathe"
Normalized Amplitude
Samples
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
100
200
300
400
500
600
700
800
900

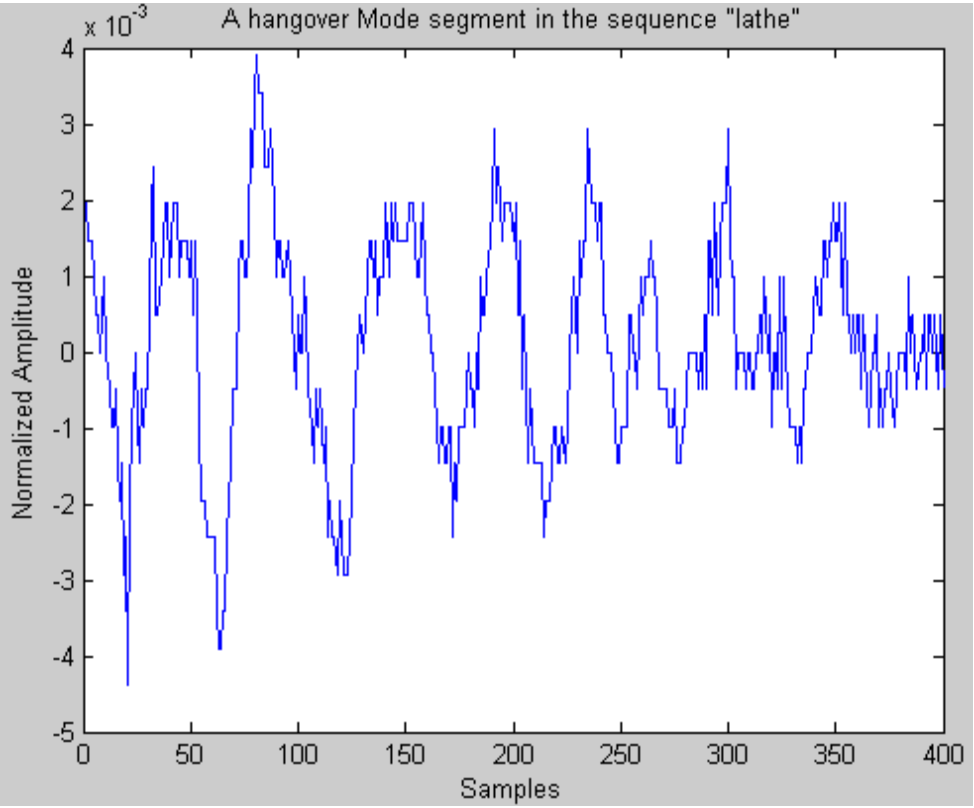
Unvoiced Mode



Onset Mode


An Onset Mode segment in the sequence "lathe"
Normalized Amplitude
Samples
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
0
5
10
15
20
25
30
35
40

Hangover Mode



Composite Speech Models

Sequence	Mode	Autocorrelation coefficients for V, ON, H Average frame energy for UV	Mean Square Prediction Error	Probability
"lathe"	V	[1 0.8217 0.5592 0.3435 0.1498 0.0200 −0.0517 −0.0732 −0.0912 −0.1471 −0.2340]	0.0656	0.5265
	ON	[1 0.8495 0.5962 0.3979 0.2518]	0.0432	0.0093
	H	[1 0.2709 0.2808 0.1576 0.1182]	0.7714	0.0186
	UV	0.1439	0.1439	0.0771
	S			0.3685

Sequence	Mode	Autocorrelation coefficients for V, ON, H Average frame energy for UV	Mean Square Prediction Error	Probability
"we were away"	V	[1 0.8014 0.5176 0.2647 0.0432 −0.1313 −0.2203 −0.3193 −0.3934 −0.4026 −0.3628]	0.0780	0.9842
	ON	[1 0.8591 0.7215 0.6128 0.5183]	0.0680	0.0053
	H			0
	UV			0
	S			0.0105

Conditional Rate Distortion Functions

- The conditional rate distortion function of a source ***X*** with side information ***Y*** is defined as

$$R_{X|Y}(D) = \min I(X; \hat{X} \mid Y)$$

where the minimum is taken over $p(\hat{x} \mid x, y)$: $D(X, \hat{X} \mid Y) \le D$ with $I(X; \hat{X} \mid Y)$ and $D(X, \hat{X} \mid Y)$ given by the usual expressions.

Conditional Rate Distortion Functions

- Gray shows that this conditional rate distortion function can be expressed as [1]

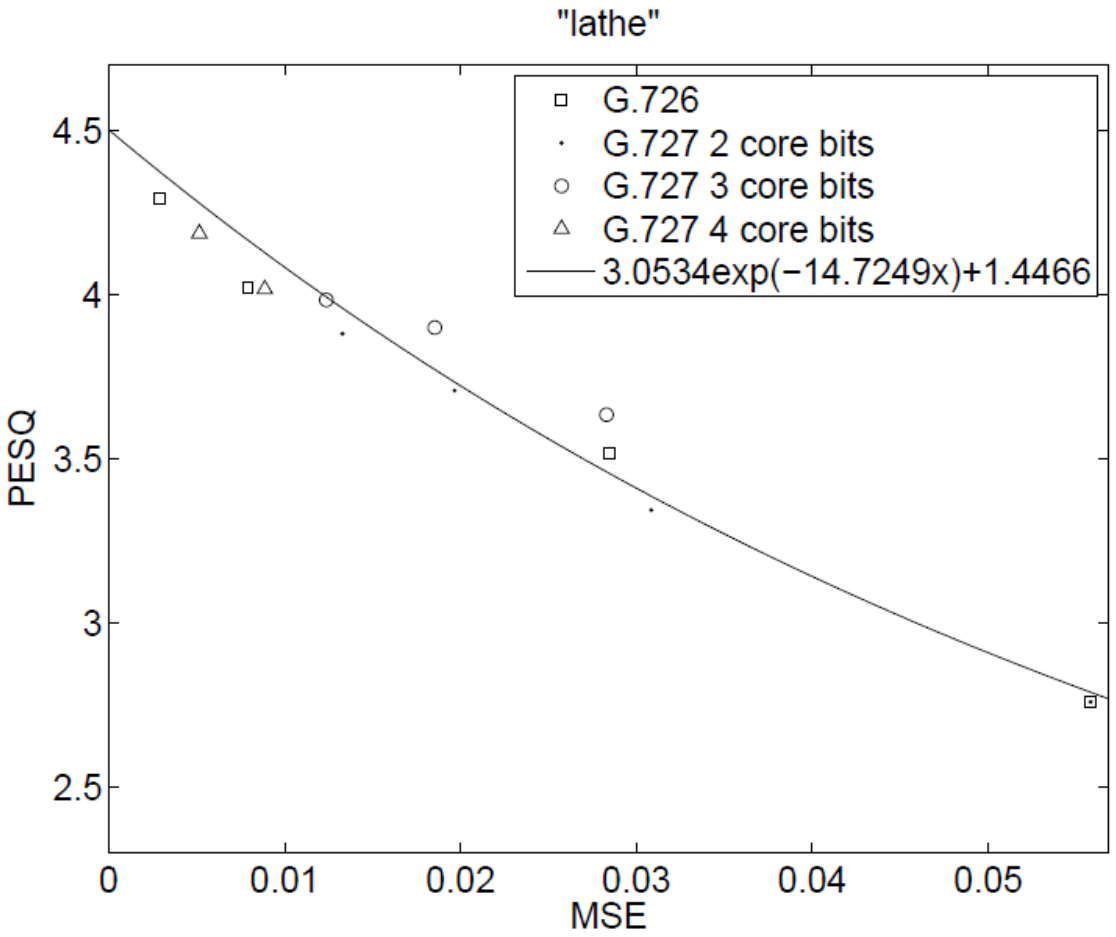
$$R_{X|Y}(D) = \min \sum_{y} R_{X|y}(D_y)\, p(y)$$

with the minimum taken over D_y's: $D(X, \hat{X} \mid Y) = \sum_{y} D_y\, p(y) \le D$

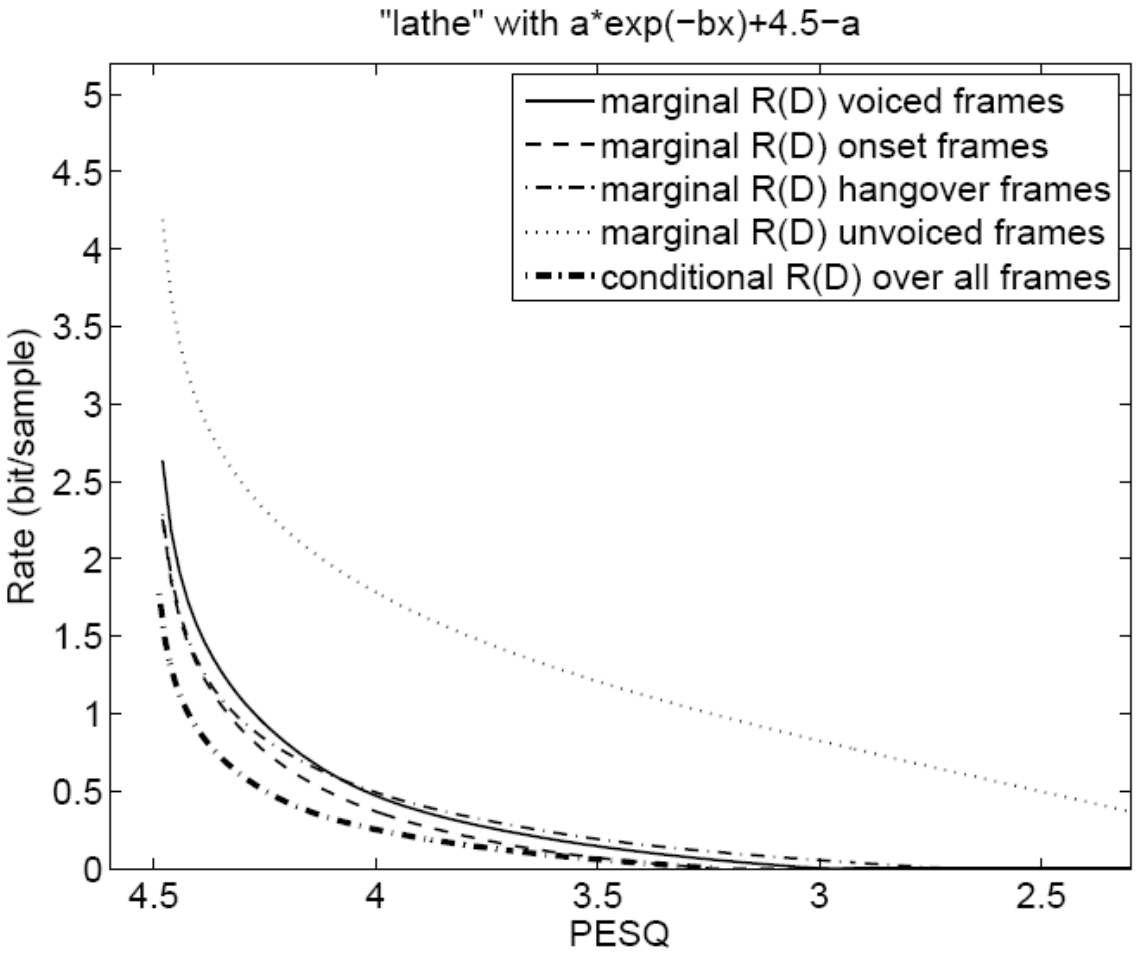
The minimum is achieved at D_y's where the slopes $\partial R_{X|Y=y}(D_y) / \partial D_y$ are equal for all y and $\sum_{y} D_y P[Y = y] = D$

[1] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Information Theory*, pp. 480-489, 1973.

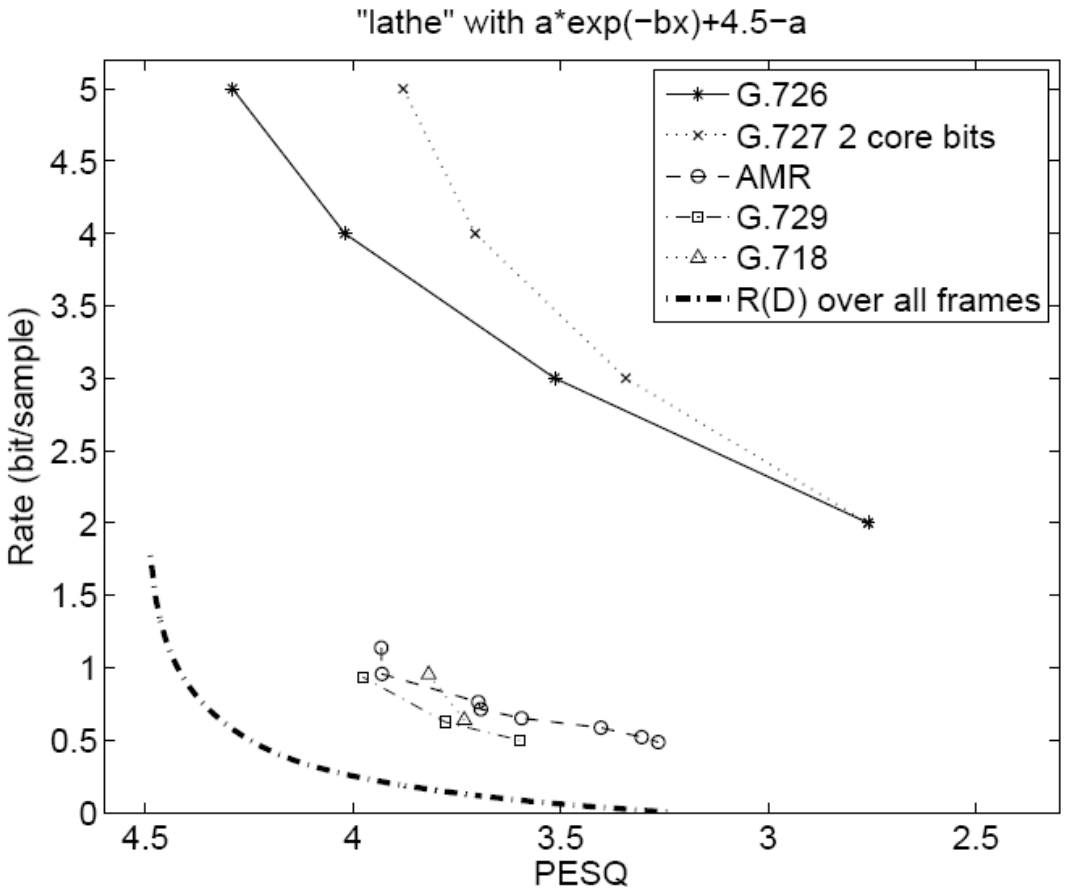
MSE to PESQ-MOS Mapping



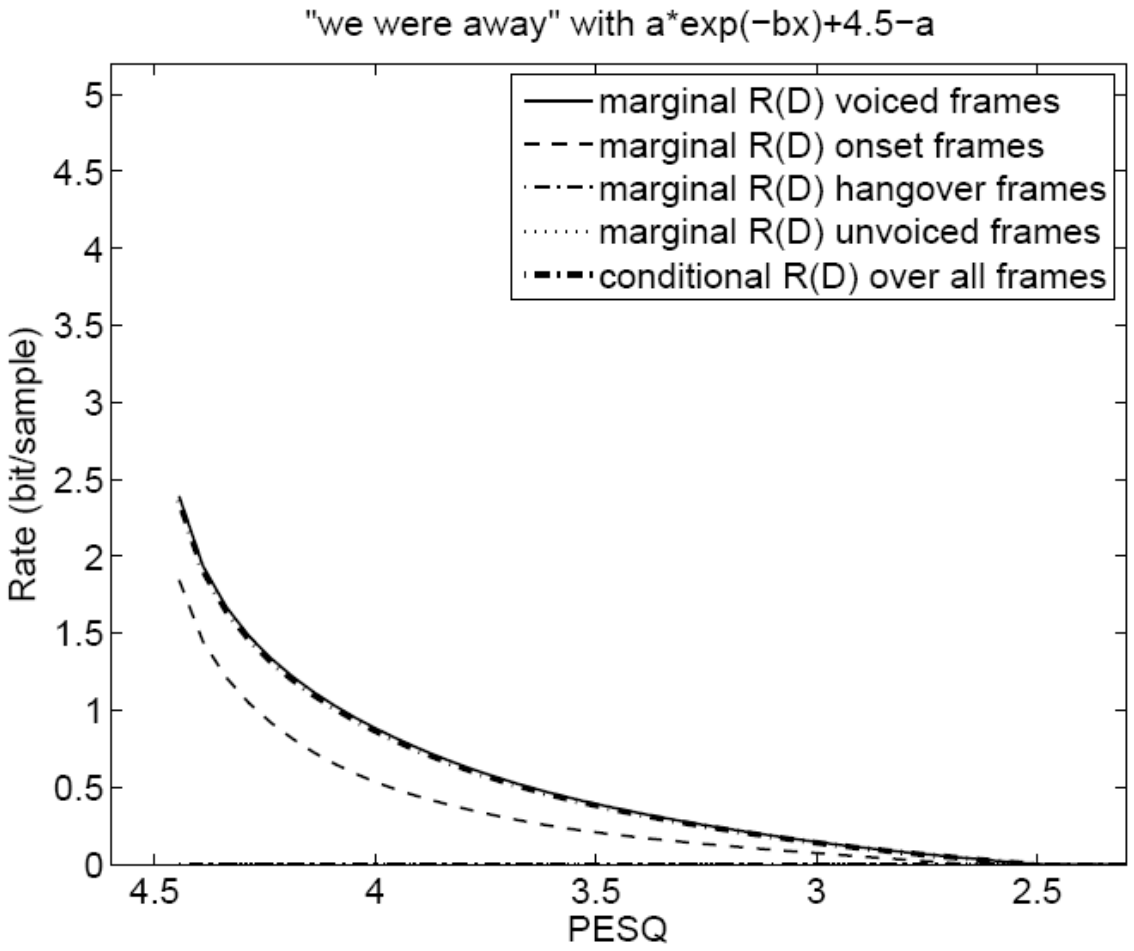
Marginal and Conditional Rate Distortion Bounds with Mapped MSE/MOS Distortion Measure



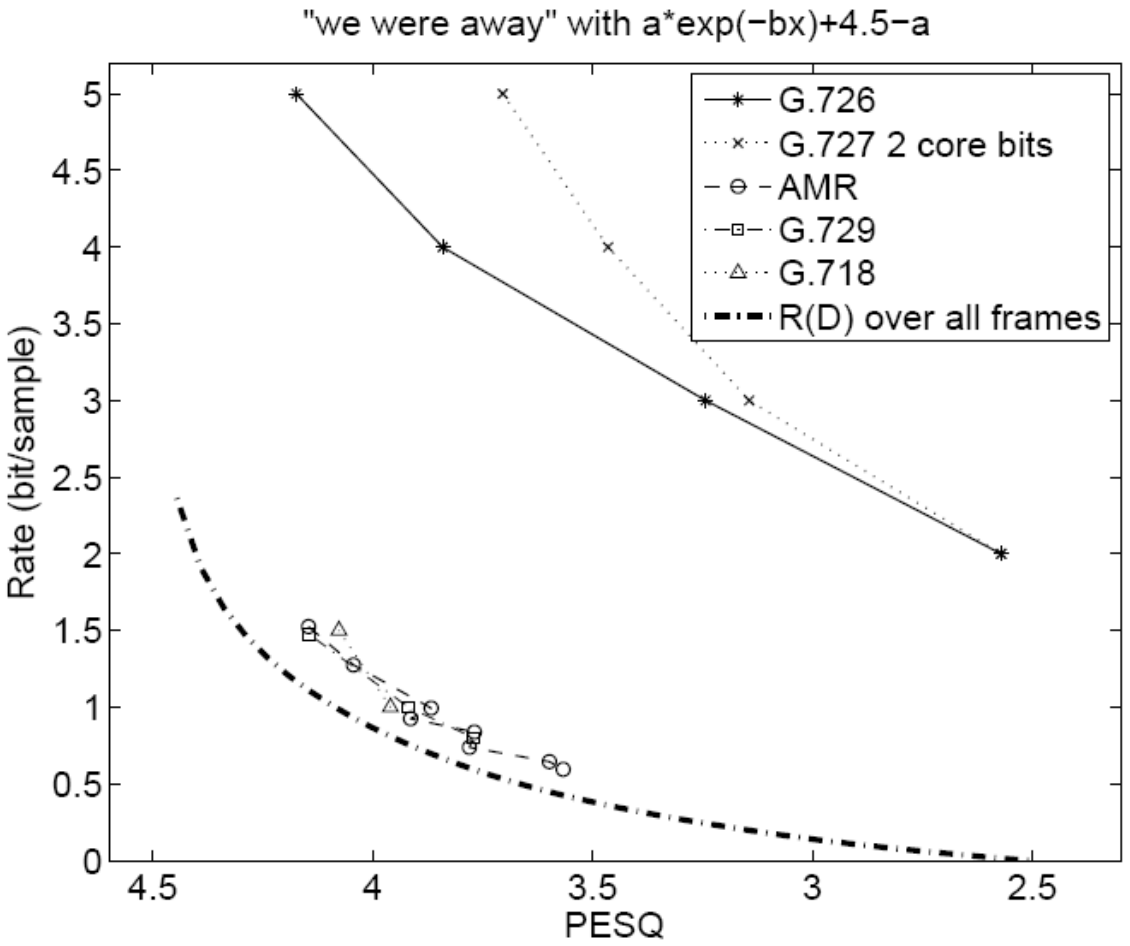
Performance of Best Known Voice Codecs



Marginal and Conditional Rate Distortion Bounds with Mapped MSE/MOS Distortion Measure



Performance of Best Known Voice Codecs

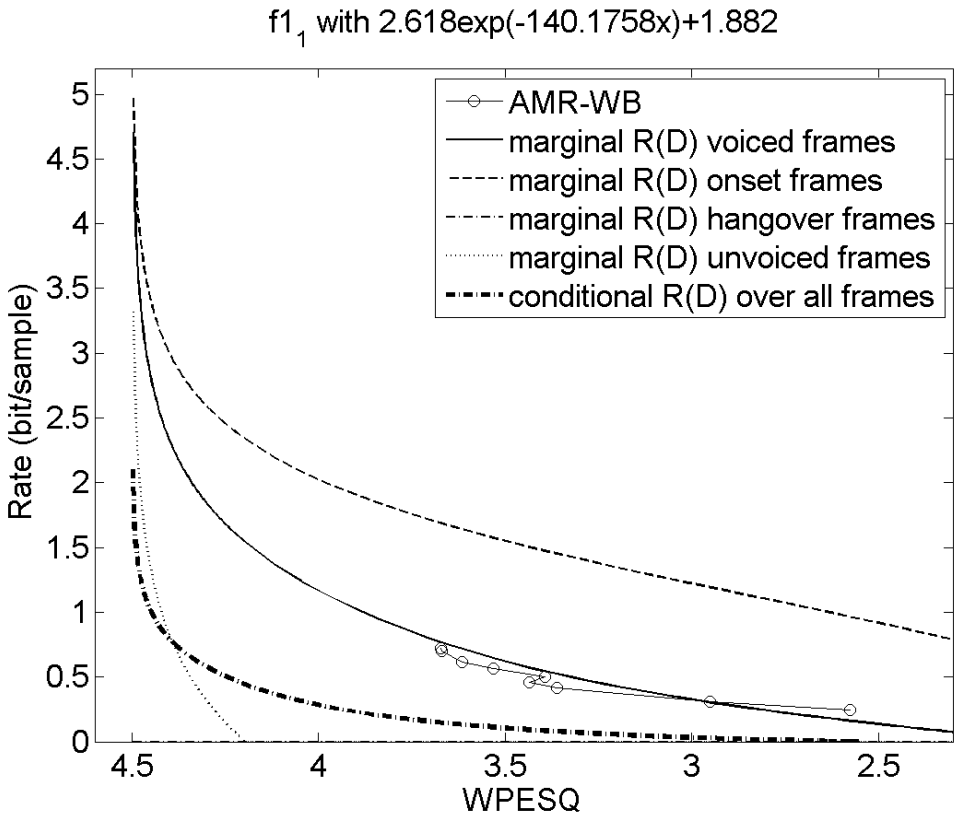


UCSB

Wideband Composite Sources

Sequence	Mode	Autocorrelation coefficients	MSE	Prob.
F1	V	[1 0.8448 0.5891 0.4132 0.3156 0.2670 0.2122 0.1462 0.0599 -0.0987 -0.3028 -0.4109 -0.3816 -0.3084 -0.2673 -0.2879 -0.3293 -0.3403 -0.3214 -0.2646 -0.1669]	0.0250	0.4406
	ON	[1 0.1226 -0.2917 0.2239 -0.0034]	0.5241	0.0043
	H		0	0
	UV	0.0009	0.0009	0.0028
	S			0.5523

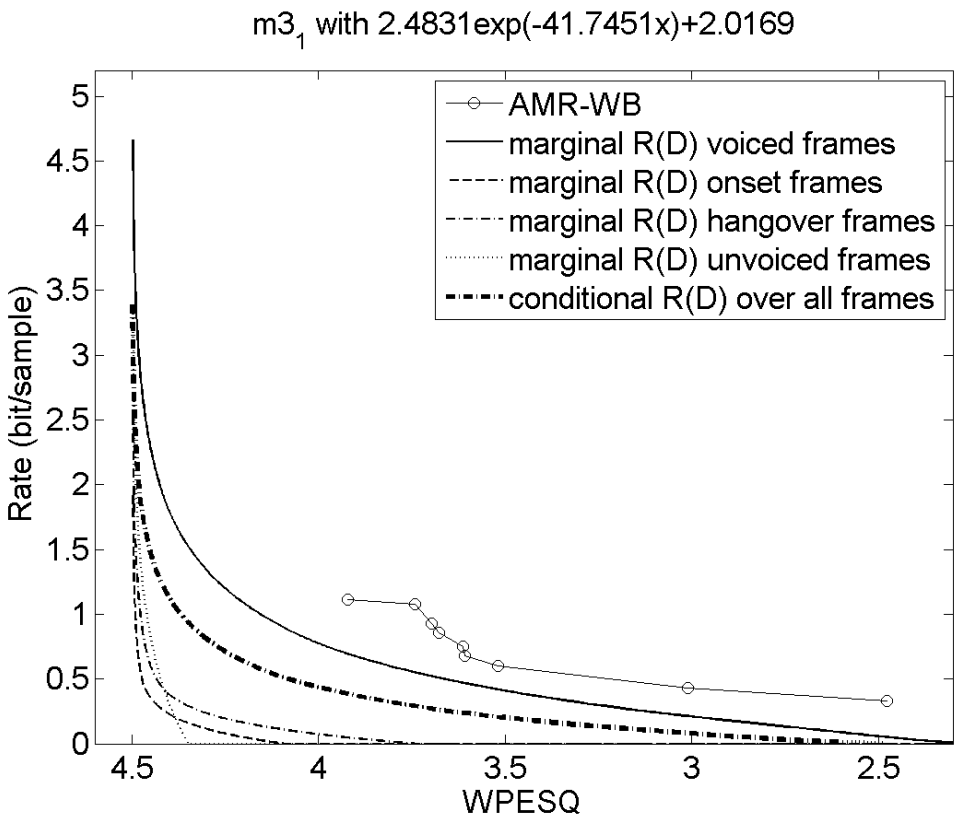
Wideband Bounds and Codec Comparisons



Wideband Composite Source Model

Sequence	Mode	Autocorrelation coefficients	MSE	Prob.
M3	V	[1 0.7954 0.6612 0.4775 0.2864 0.2398 0.2004 0.2169 0.2214 0.2248 0.2022 0.1613 0.1333 0.1075 0.1334 0.1759 0.1662 0.1343 0.0771 -0.0283 -0.0855]	0.0815	0.6939
	ON	[1 0.9564 0.9334 0.9104 0.8862]	0.0066	0.0069
	H	[1 0.9387 0.9028 0.8696 0.8257]	0.0129	0.0461
	UV	0.0015	0.0015	0.0064
	S			0.2467

Wideband Performance Bounds and Codec Comparisons



Voice Conclusions

- Introduced New Composite Source Models for Speech
- Calculated New Rate Distortion Bounds Based on These Composite Source Models
- Compared to Speech Codecs for Mapped MSE/MOS Distortion Measure
- Lower Bounds Best Known Existing Voice Codecs