Scalable Speech and Audio Coding for Heterogeneous Networks

W. Bastiaan Kleijn

School of Electrical Engineering

KTH - Royal Institute of Technology

Stockholm, Sweden

Work with: Alexey Ozerov, Janusz Klejsa

Guoqiang Zhang, Moo Young Kim

Outline

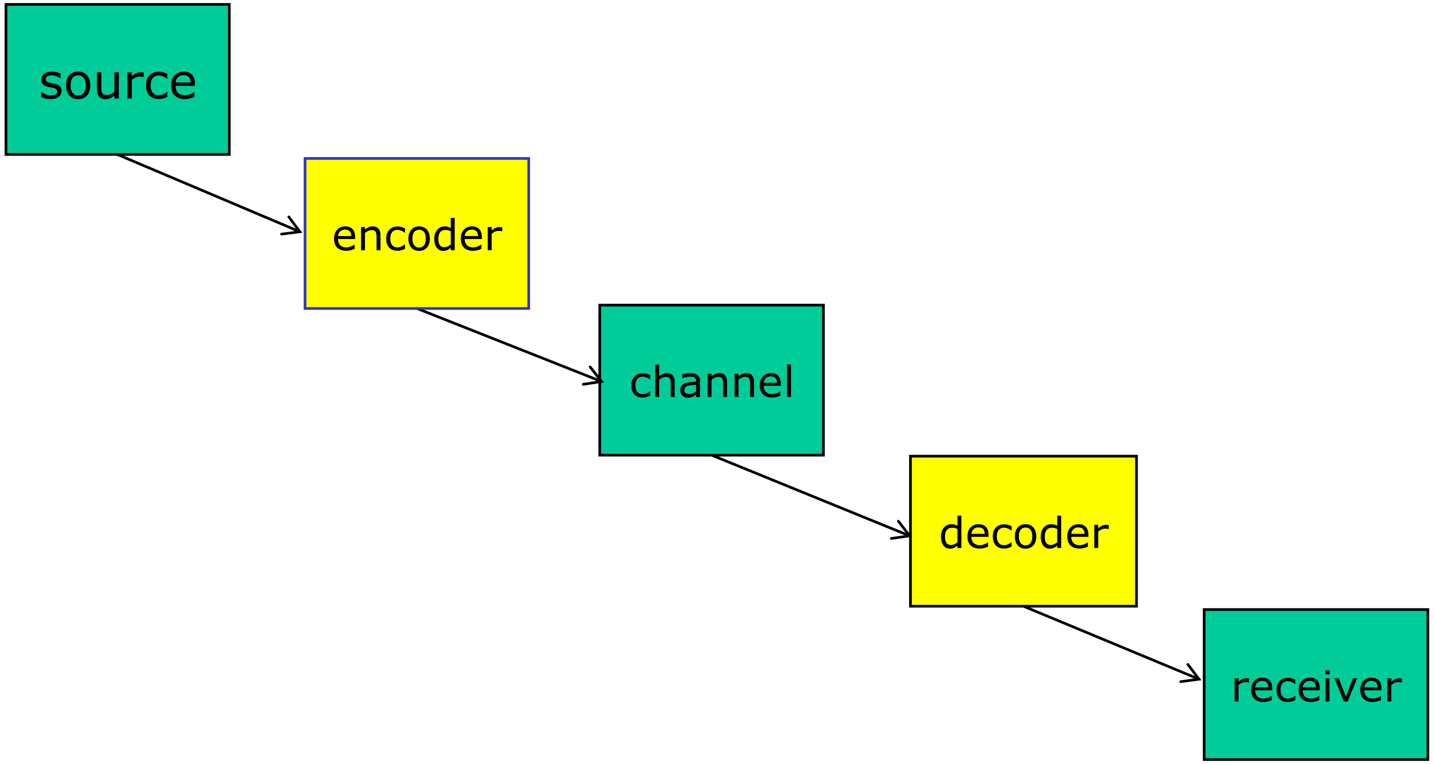
- Introduction
- Techniques:
  - Rate constraint used in coder design
  - Scalable model-based coding
  - Scalable multiple-description coding (MDC)
- Model-based coding architectures

Objective

- Audio coder with following attributes:
  - Good rate-distortion performance
  - Scalable in rate
  - Scalable in robustness to packet loss

Approach: *Model Everything*

- Statistical models of
  - Source
  - Channel
  - Receiver
  - Encoder
  - Decoder
  - Estimate / optimize in real time



Selection En/Decoder Model

- Rate-distortion theory (Shannon, 1959)
  - Needs densities; bounds for simple densities only
  - Variable-rate only
  - No direct relation to practical systems
- Lloyd algorithm (Lloyd, 1958)
  - Not a model; leads directly to quantizer
  - Iterative / results in codebooks / not scalable
  - Locally optimal / no need density function
- High-rate theory (Bennett, 1948)
  - Assumes signal density constant in quantization cell
  - Asymptotically optimal
  - Fixed and variable rate
  - Analytic solutions / scalable
  - Provides centroid density / requires additional step

Summary of Introduction

- To design (near-)optimal coders we need models of source, encoder, channel, decoder, receiver

- High-rate theory provides
  - Relation distortion and rate for coder
  - Analytic solution reconstruction point density

Outline

- Introduction
- Techniques:
  - Rate constraint used in coder design
  - Scalable model-based coding
  - Scalable multiple-description coding (MDC)
- Model-based coding architectures

Traditional Rate Constraints

- Fixed rate = *constrained resolution*
  - Good for circuit-switched networks
  - Variable distortion
- Variable rate = *constrained entropy*
  - Good for packet-switched networks?
  - Fixed cell density: fixed mean distortion per cell

Constrained-Resolution Coding

- Mean distortion of cell:

$$D(x^k) = C(k,G)\ v(x^k)^{\frac{2}{k}}$$

- Mean distortion:

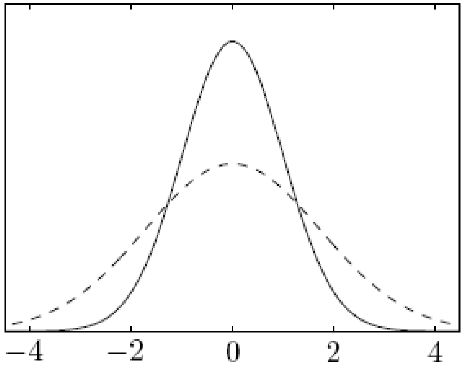
$$D = C \int f_{X^k}(x^k) g(x^k)^{-\frac{2}{k}} dx^k$$

- Constraint:

$$N = \int g(x^k)\, dx^k$$

- Solution:

$$\boxed{g(x^k) = N\,\beta\, f_{X^k}(x^k)^{\frac{k}{k+2}}}$$



Constrained-Entropy Coding

- Mean distortion of cell:

$$D(x^k) = C(k,G)\; v(x^k)^{\frac{2}{k}}$$

- Mean distortion:

$$D = C \int f_{X^k}(x^k) g(x^k)^{-\frac{2}{k}} dx^k$$

- Constraint:

$$\boxed{H(I) = \text{constant} \quad \Rightarrow \quad \int f_{X^k}(x^k) \log(g(x^k)) dx^k = \text{constant}}$$

- Solution:

$$\boxed{g(x^k) = \text{constant} = \exp(H(I) - h(X^k))}$$

Do Existing Solutions Suffice?

- (Solutions are same for high dimensionality)
  - Data density uniform in region of support
- Constrained-resolution coding:
  - Distortion outliers generally dominate perceived quality
- Constrained-entropy coding:
  - Rate outliers can be severe
- More outliers if data density incorrect
  - Mismatch due to assumptions, inaccurate misestimation, etc.
  - Backward adaptation (low delay): large mismatch at transitions
- Iterative source-channel decoding
  - Exploit redundancy in quantizer; leads to mismatch of criterion

Alternative Approach

- Constrained-entropy constrains index entropy

$$H(I) = -\sum p_I(i) \underbrace{\log(p_I(i))}_{\text{minus codeword length}}$$

- Alternative: constrain exp-waited codeword length

$$J_\gamma(I) = \sum p_I(i) \underbrace{p_I(i)^{-\frac{\gamma}{k}}}_{\substack{\text{exponentially weighted} \\ \text{codeword length}}}$$

Variable-Constraint Coding

- Mean distortion of cell:

$$D(x^k) = C(k,G)\; v(x^k)^{\frac{2}{k}}$$

- Mean distortion:

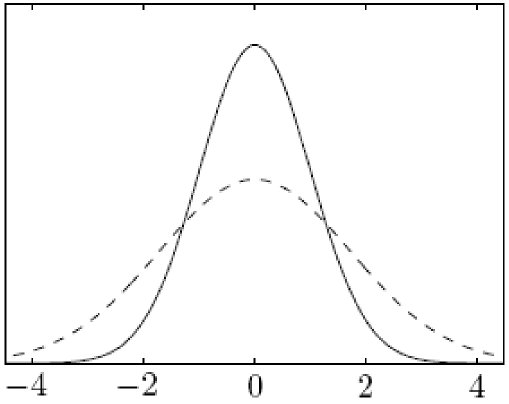
$$D = C\int f_{X^k}(x^k) g(x^k)^{-\frac{2}{k}} dx^k$$

- Constraint:

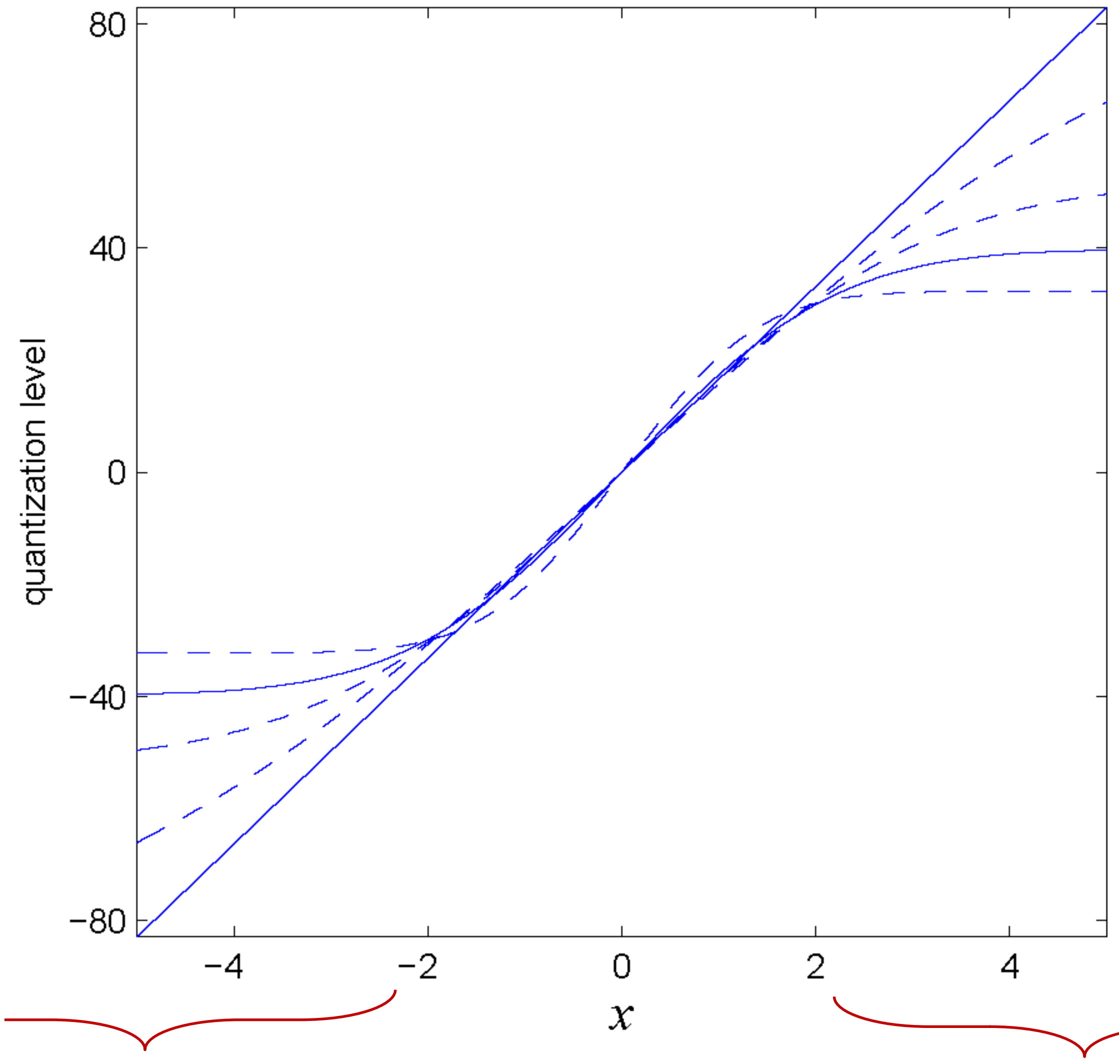
$$\boxed{J_\gamma(I) = \sum p_I(i) p_I(i)^{-\frac{\gamma}{k}} \quad \Rightarrow \quad \int f_{X^k}(x^k)^{1-\frac{\gamma}{k}} g(x^k)^{\frac{\gamma}{k}} dx^k = \text{constant}}$$

- Solution:

$$\boxed{g(x^k) = N\,\beta\, f_{X^k}(x^k)^{\frac{\gamma}{\gamma+2}}}$$



Variable-Constraint Companders



high distortion and/or rate

Summary of Rate Constraints

- Standard constraints are resolution constraint entropy constraint
  - Distortion or rate outliers
- In many applications a compromise is better
  - Satisfy both network and perception views
  - Iterative source-channel decoding
  - Model mismatch often important
- Variable-constraint theory facilitates compromises

Outline

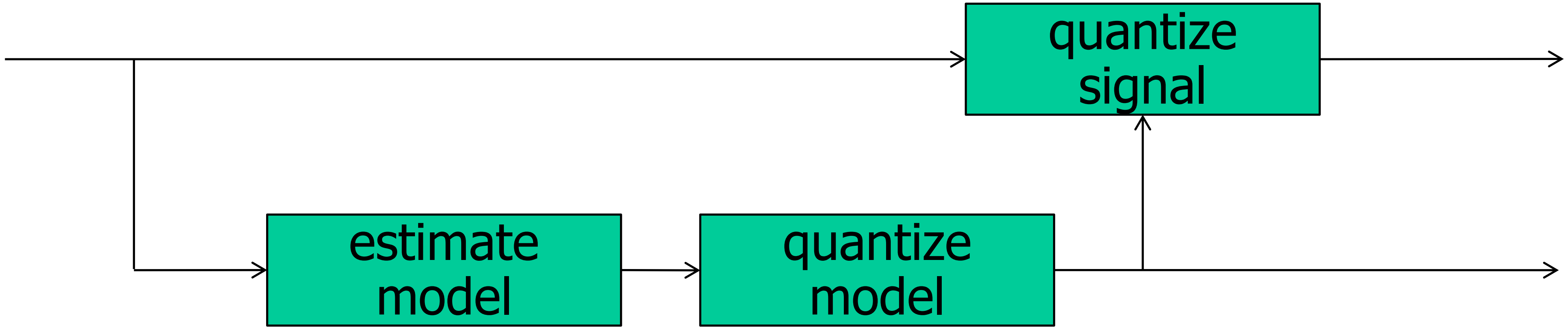
- Introduction
- Techniques:
  - Rate constraint used in coder design
  - Scalable model-based coding
  - Scalable multiple-description coding (MDC)
- Model-based coding architectures

Rate Distribution

- High-rate theory leads to computable=scalable quantizers

- For scalable model-based coder:

  How many bits for model versus how many bits for signal-given-model?



Rate Distribution: Quantizer Design

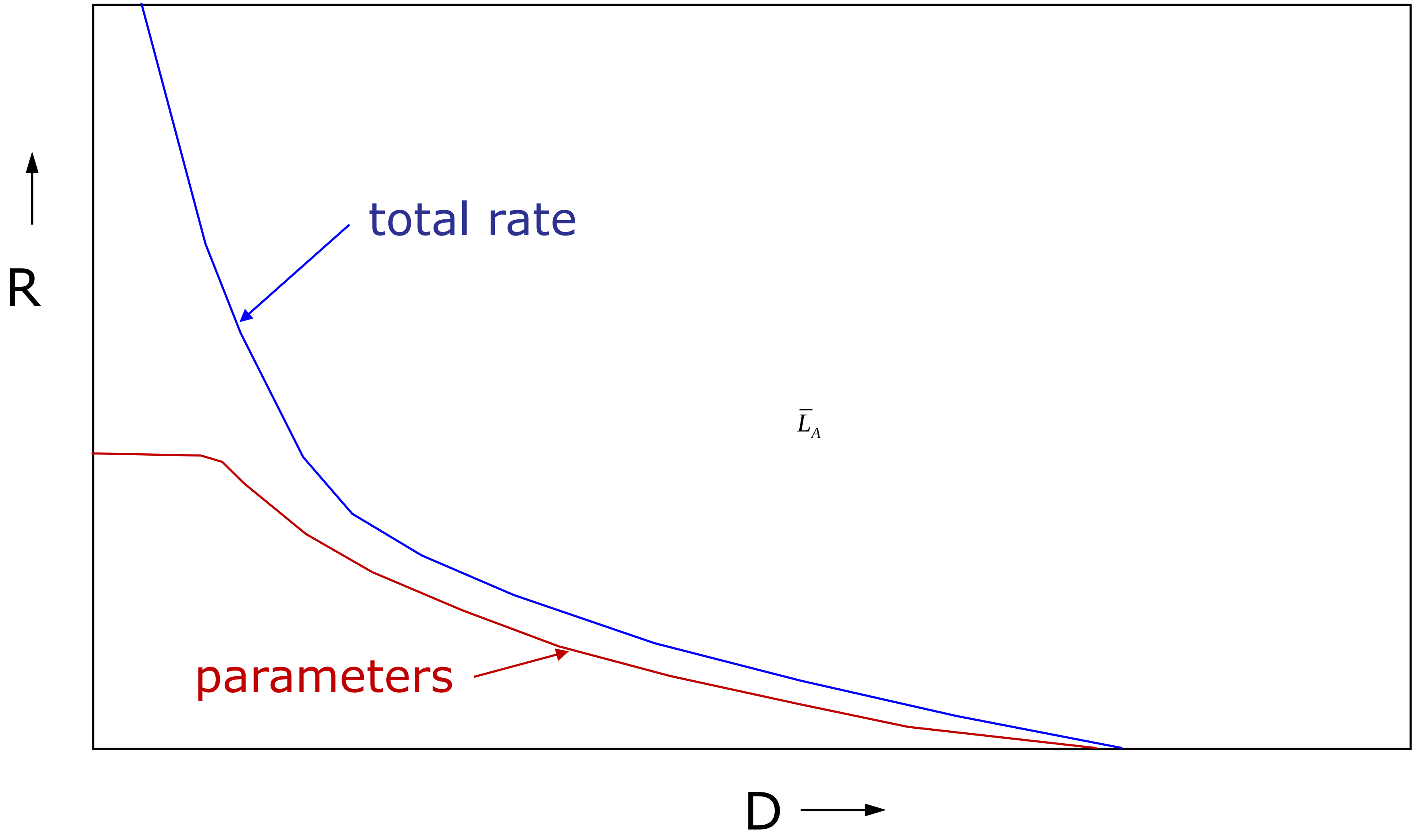
- Select quantizer that minimizes mean codeword length
- Constrained-entropy case

$$\mathrm{E}\left[L_A\right] = -\mathrm{E}\Big[\overbrace{\log(p_\Theta(\tilde{\theta}(x^k)))}^{\substack{\text{mean quantized}\\ \text{model code length}}}\Big] - \mathrm{E}\Big[\overbrace{\log(v\, f_{X^k|\Theta}(x^k \,|\, \tilde{\theta}(x^k)))}^{\text{mean signal code length}}\Big]$$

$$= \underbrace{-\mathrm{E}\Big[\log(p_\Theta(\tilde{\theta}(x^k)))\Big] + \mathrm{E}\left[\log\left(\frac{f_{X^k|\Theta}(x^k \,|\, \hat{\theta}(x^k))}{f_{X^k|\Theta}(x^k \,|\, \tilde{\theta}(x^k))}\right)\right]}_{\substack{\text{mean index of resolvability}\\ \textit{independent of distortion}}} - \mathrm{E}\Big[\log(v\, f_{X^k|\Theta}(x^k \,|\, \hat{\theta}(x^k)))\Big]$$

only term relating to distortion

Rate Distribution



From low distortion to high distortion is

from hybrid coding to parametric coding

Constrained Resolution

- Coding a sequence x^k with fixed-rate allocation for sequence and for model:

$$L = L_m + L(x^k)$$

$$= L_m + \log(N)$$

$$= L_m - \mathrm{E}\left[\log\left(p_{X^k|\Theta}(X^k|\tilde{\theta})^{\boxed{\frac{k}{k+2}}}\right)\right] - \frac{k}{2}\log\left(\frac{D_{CR}}{C}\right)$$

$$= L_m \underbrace{+ \underbrace{\boxed{\frac{k}{k+2}}\mathrm{E}\left[\log\left(\frac{p_{X^k|\Theta}(X^k|\hat{\theta})}{p_{X^k|\Theta}(X^k|\tilde{\theta})}\right)\right]}_{\text{regret=excess code length}}}_{\text{mean index of resolvability}} - \underbrace{\boxed{\frac{k}{k+2}}\mathrm{E}\left[\log\left(f_{X^k|\Theta}(X^k|\hat{\theta})\right)\right] - \frac{k}{2}\log\left(\frac{D_{CR}}{C}\right)}_{\text{mean signal code length for optimal model}}$$

A Practical Coder: AMR-WB*

	Rate, kb/s	6.6	8.85	12.65	14.25	15.85	18.25	19.85	23.05
Model Parameters	AR model	36	46	46	46	46	46	46	46
	pitch	23	26	30	30	30	30	30	30
	gains	24	24	28	28	28	28	28	28
	LTP flag	0	0	4	4	4	4	4	4
	VAD flag	1	1	1	1	1	1	1	1
Coefficients	excitation	48	80	144	176	208	256	288	352

$*$ AMR−WB coder uses 20 ms blocks

Coding with Autoregressive Models

- Autoregressive models used in essentially all mobile telephones

- Interesting application of the theory
  - What does the index of resolvability correspond to?

- Our model assumption is that the signal is Gaussian
  - Multivariate Gaussian:

$$p_{X^k|\Theta}(x^k) = \frac{1}{\sqrt{2\pi \det(R_\Theta)}} \exp\left(-\tfrac{1}{2}\, x^{kH} R_\Theta^{-1} x^k\right)$$

  - For large k:

$$\log\left(p_{X^k|\Theta}(x^k \mid \theta)\right) \approx -\frac{1}{2}\log(2\pi) - \frac{k}{4\pi}\int_0^{2\pi} \log(R_\theta(\mathrm{e}^{j\omega})d\omega - \frac{k}{4\pi}\int_0^{2\pi} \frac{R_X(\mathrm{e}^{j\omega})}{R_\theta(\mathrm{e}^{j\omega})}\, d\omega$$

I of Resolvability Autoregressive Models

- Mean index of resolvability:

constrained-resolution case

$$\psi = L_m + \mathrm{E}\left[\log\left(\frac{p_{X^k|\Theta}(X^k\,|\,\hat{\theta})^{\frac{k}{k+2}}}{p_{X^k|\Theta}(X^k\,|\,\tilde{\theta})^{\frac{k}{k+2}}}\right)\right]$$

effect of quantization

effect of modeling averages to 1
if optimal gain is used

$$\approx L_m + \frac{k}{4\pi}\underbrace{\int_0^{2\pi} -\log\left(\frac{R_{\hat{\theta}}(\mathrm{e}^{j\omega})}{R_{\theta}(\mathrm{e}^{j\omega})}\right) + \left(\frac{R_{\hat{\theta}}(\mathrm{e}^{j\omega})}{R_{\theta}(\mathrm{e}^{j\omega})} - 1\right)\frac{R_X(\mathrm{e}^{j\omega})}{R_{\hat{\theta}}(\mathrm{e}^{j\omega})}\,d\omega}_{\text{Itakura-Saito criterion if } R_W=1}$$

$R_W=1$ and small spectral error

$$\approx L_m + \frac{k}{8\pi}\underbrace{\int_0^{2\pi}\Big(\log\big(R_{\hat{\theta}}(\mathrm{e}^{j\omega})\big) - \log\big(R_{\theta}(\mathrm{e}^{j\omega})\big)\Big)^2 d\omega}_{\substack{\text{mean square log spectral error =}\\ \text{signal rate penalty because model is not right}}} \quad = \quad L_m + D(\tilde{\theta},\hat{\theta})$$

Threshold for Constr Resolution

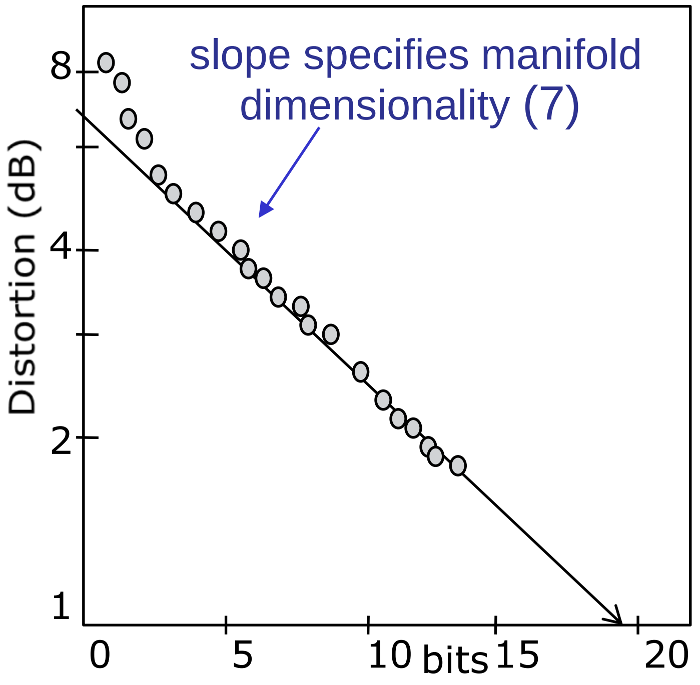
- Mean index of resolvability:

$$\psi \approx L_m + \frac{k}{8\pi}\int_0^{2\pi}\left(\log\left(R_{\hat{\theta}}(e^{j\omega})\right)-\log\left(R_{\theta}(e^{j\omega})\right)\right)^2 d\omega$$

- Second term depends on parameter distribution
  - is known in literature (Paliwal-Kleijn 1995)

- Minimize rate:

  Threshold 1.25 dB = 20 bits
  - Common usage is 1 dB!
    - Based on "perception"



Rate Distribution

- Index of resolvability:

$$\psi \approx -\log\left(P(\tilde{\theta})\right) + D(R_{\theta}, R_{\tilde{\theta}})$$

- High-rate relation to differential entropy

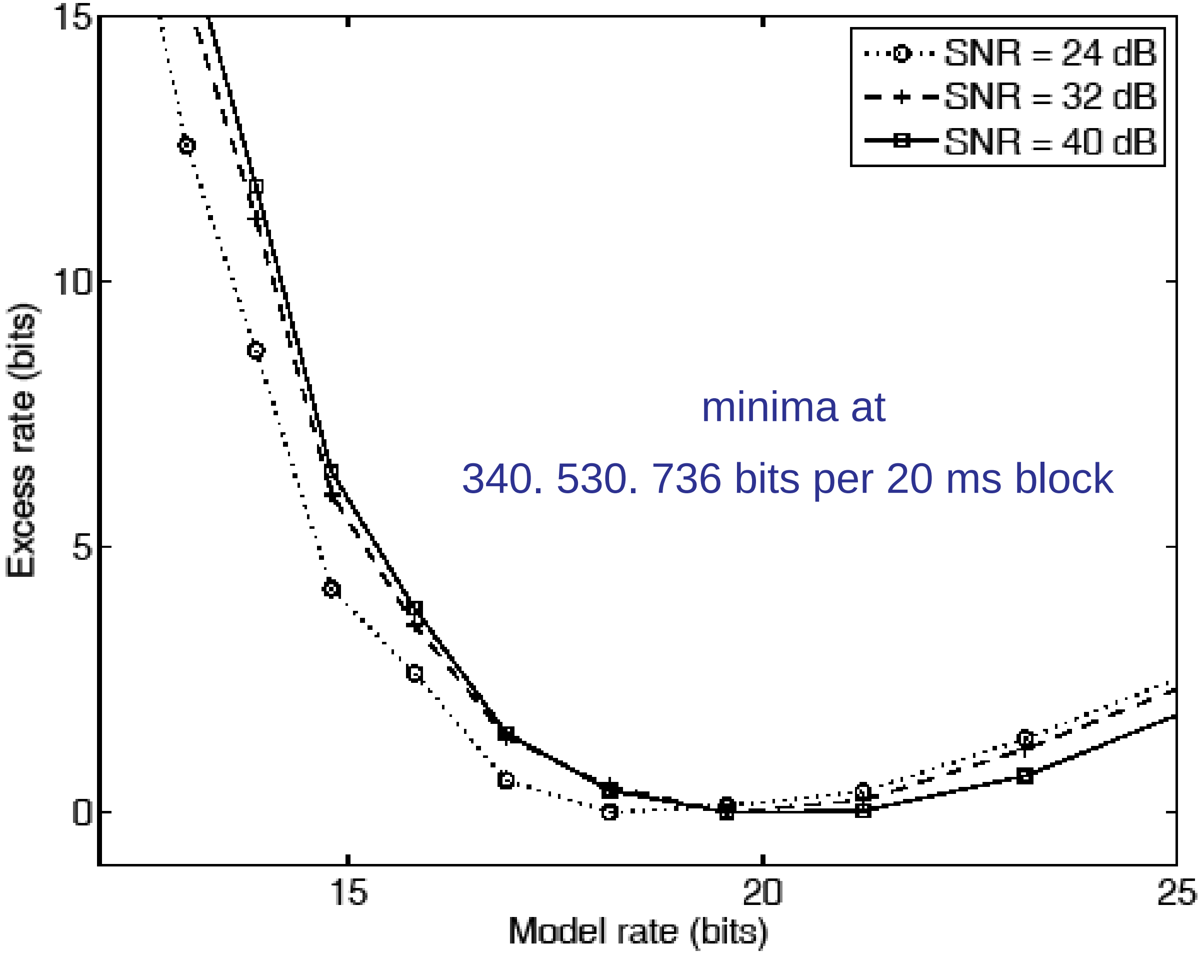
$$D(R_{\tilde{\theta}}, R_{\theta}) = dCe^{-\frac{2}{d}[R(\tilde{\theta}) - h(\theta)]}$$

- Set derivative to zero:

$$\boxed{R_{\tilde{\theta}} = h(\theta) + \frac{d}{2}\log\left(\frac{k}{2}C\right)}$$

- Threshold for 8 kHz sampled speech, AR model k=160, d=8, $C = 1/12$ or $C = 1/2\pi e$ 19 and 17.2 bits, corresponds to 1.29 dB
  - Again disproves common belief that 1 dB threshold motivated by perception; it simply leads to lowest mean squared error

Rate Distribution: Confirmation



minima at

340. 530. 736 bits per 20 ms block

W. Bastiaan Kleijn, “Principles of Speech Coding”, in *Speech Processing*, Eds. Benesty, Huang, Sondhi, Springer, pp. 283-306, 2007

W. Bastiaan Kleijn and Alexey Ozerov. ”Rate distribution between model and signal”. Proc. IEEE WStockhoshop App Sign Process Audio Acoust, WASPAA, pp. 243-246, 2007

Summary of Rate Distribution

- Cannot use trial-and-error for rate distribution between model and signal in adaptive coding
- New theory provides optimal distribution
  - Fixed rate for model
- Theory predicts existing heuristic results
  - Contrast to common belief:

    Rate distribution is *not* governed by perceptual effects

Outline

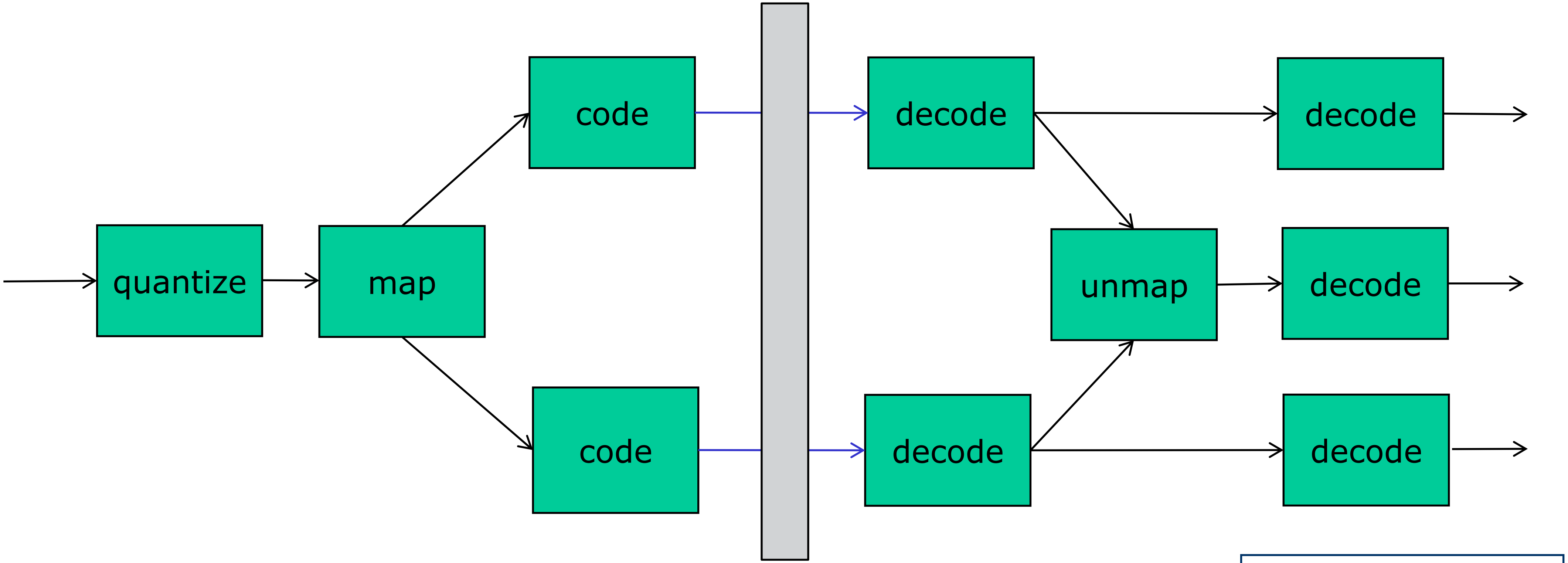
- Introduction
- Techniques:
  - Rate constraint used in coder design
  - Scalable model-based coding
  - Scalable multiple-description coding (MDC)
- Model-based coding architectures

Objective of Scalable MDC

- Optimal coding performance under packet-loss
- Optimal redundancy under all circumstances
  - Never more redundancy than needed
  - No redundancy if channel is perfect
- Should work with model-based coders

MDC Principle

- MDC = Multiple Description Coding
  - Each description facilitates signal reconstruction
  - Quality improves with number of received descriptions
  - Trade-off between max quality and “incomplete” quality



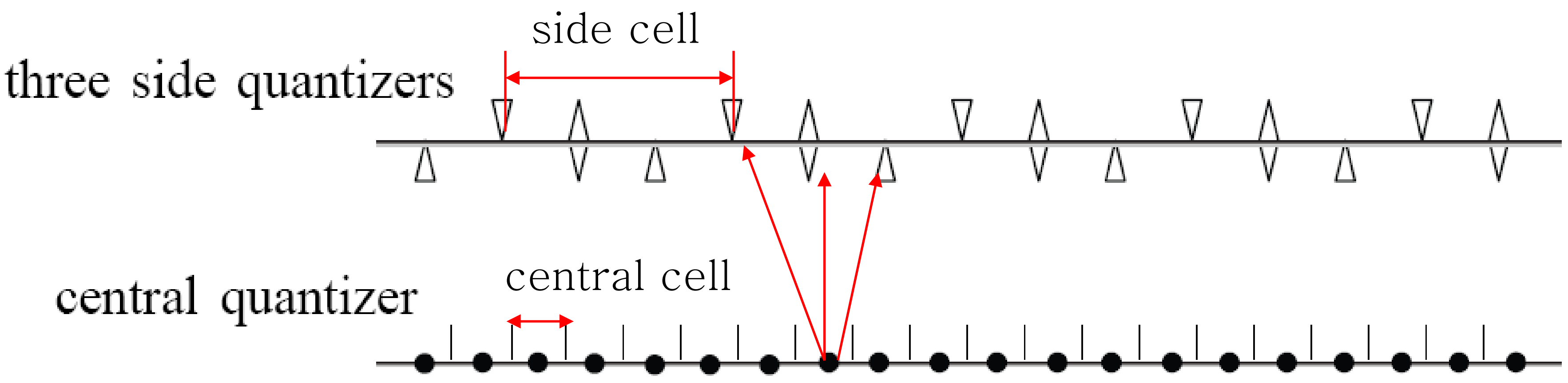
MDC

- A form of *joint source-channel coding*
  - Integral part of source coder design
  - Can provide optimal performance
- Alternative is forward error correction (FEC)
  - MDC has “soft” failure, FEC has “hard” failure
  - FEC facilitates modular design
  - MDC generally *inflexible*
- Usage in context of model-based coding not clear

Principle General Scalar MDC

- Design Principle:
  1. Define central and coarse side quantizers
  2. Mapping from central points to K side quantizer points (K-tuples)

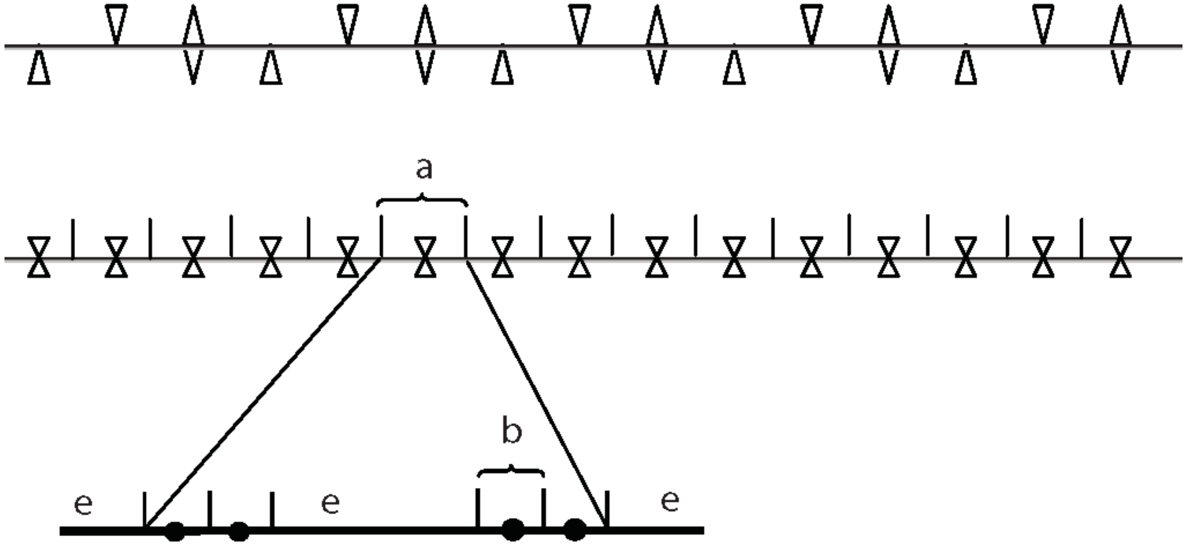
$$\alpha : \mathcal{A}_c \rightarrow \mathcal{A}_0 \times \mathcal{A}_1 \times \ldots \times \mathcal{A}_{K-1}$$



Redundancy factor: N= $\frac{\text{side cell volume}}{\text{central cell volume}}$

Scalable K-Description Scalar MDC

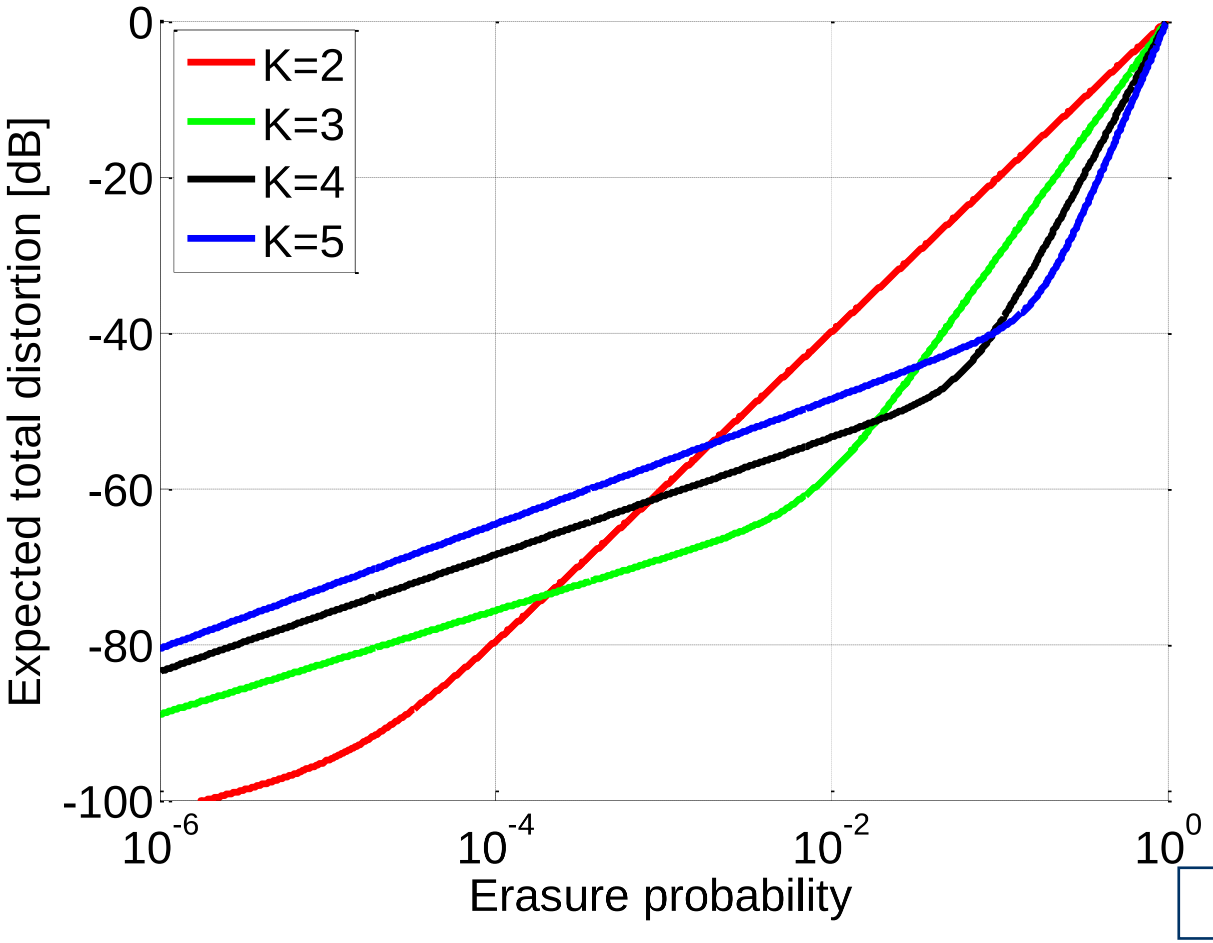
- Exploits reference quantizer = union of side quantizers
  - Reference centroid is *mean of associated K-tuples*
  - Find K-tuples that minimize spread of side cells
    - No need for search; optimal & elegant solution
- Example: three descriptions, $K=3$
  - Redundancy: $N=$ central quantizer cells per side description



G. Zhang, J. Klejsa, and W. B. Kleijn, "Optimal Index Assignment for Multiple Description Scalar Quantization", in preparation.

Behavior K-Description MDC

- More descriptions for higher loss rate
- Relations are rate dependent



Model-Based MDC

- Example: source coding with AR model

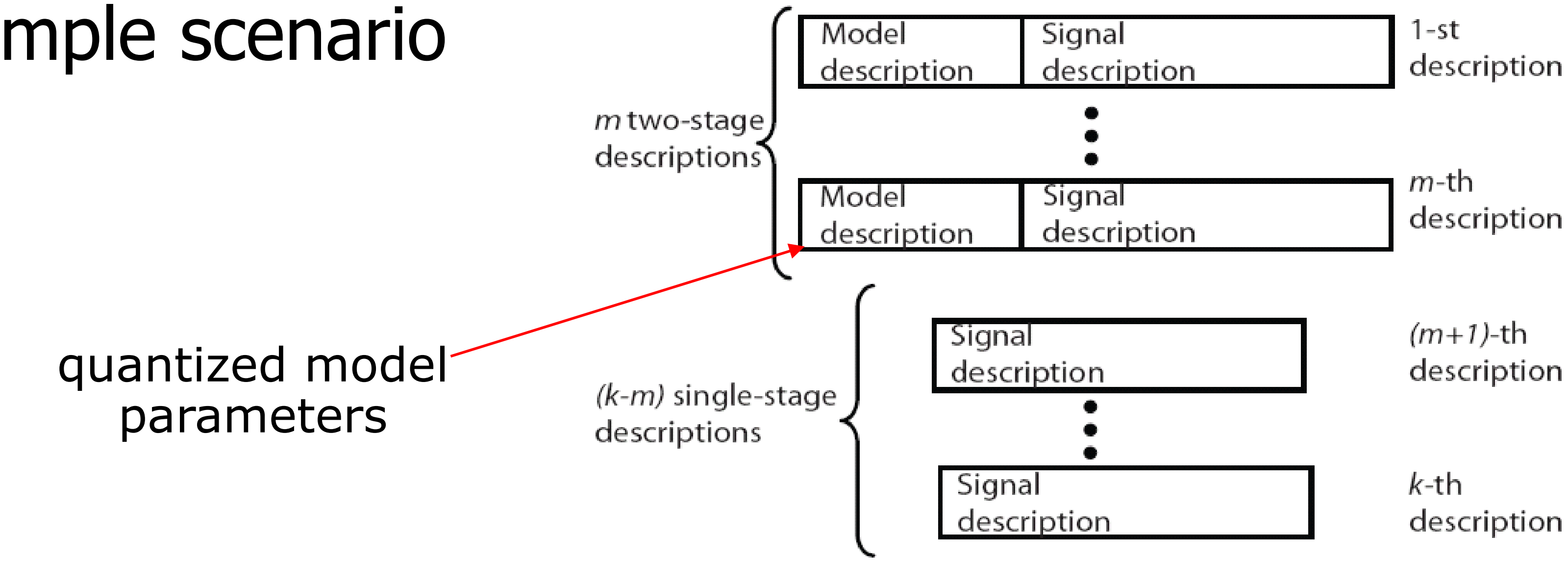
- Can we perform MDC on the model?

  NO!

- A description is a signal description
  - With or without model
  - How many signal descriptions carry a model description?

Model-Based MDC

- Example scenario



- Find optimal rate distribution model and signal

$$R_T = -\underbrace{mE_X\{\log(f_{\bar{\Theta}}(\bar{\Theta}(X))\}}_{\text{model rate}} - k\underbrace{E_X\{\log(f_{X|\bar{\Theta}}(X|\theta(X))V)\}}_{\text{signal rate}}$$

Summary of Scalable MDC

- MDC is a form of joint source-channel coding
  - High performance
- Problem: inflexible in design
  - Not commonly used; FEC more flexible
- Our methods lead to flexible MDC
  - Optimal redundancy at all times
- Model-based MDC
  - Generally optimal to include model with each description

Outline

- Introduction
- Techniques:
  - Rate constraint used in coder design
  - Scalable model-based coding
  - Scalable multiple-description coding (MDC)
- Model-based coding architectures

Coding Architecture Goal

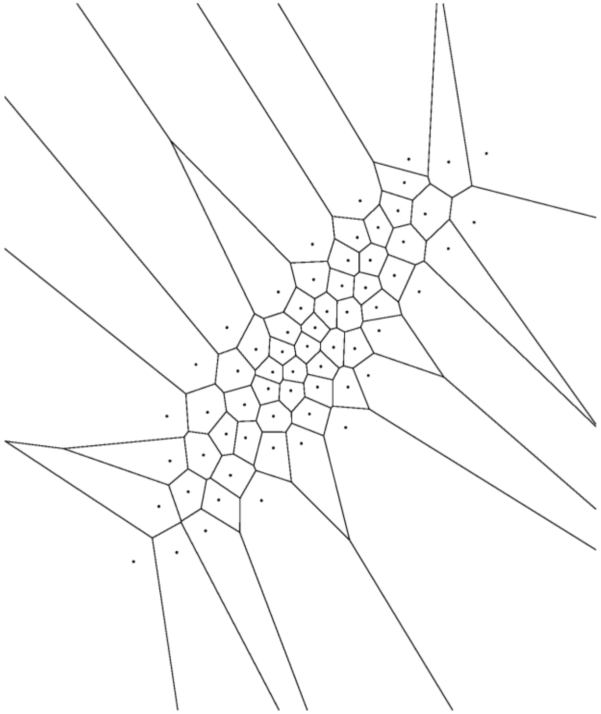
- Vector quantization is optimal
  - Search computationally complex (CR)
  - Indexing complicated (CE)

- Goal:

  to make scalar quantization effective

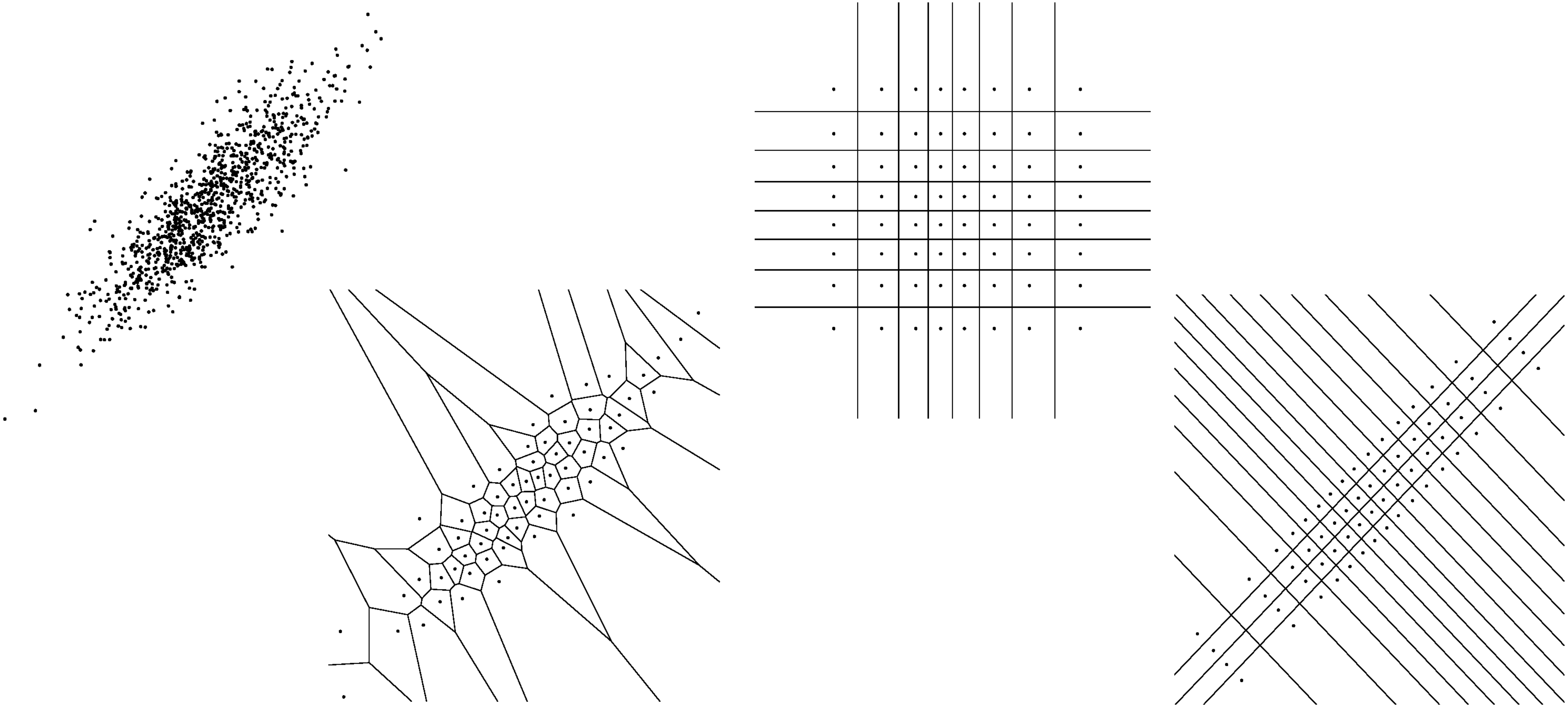
  – (Or low-dimensional VQ)

  – Remove advantages of VQ
    - Memory advantage
    - Space-filling advantage
    - (Shape advantage, CR only)



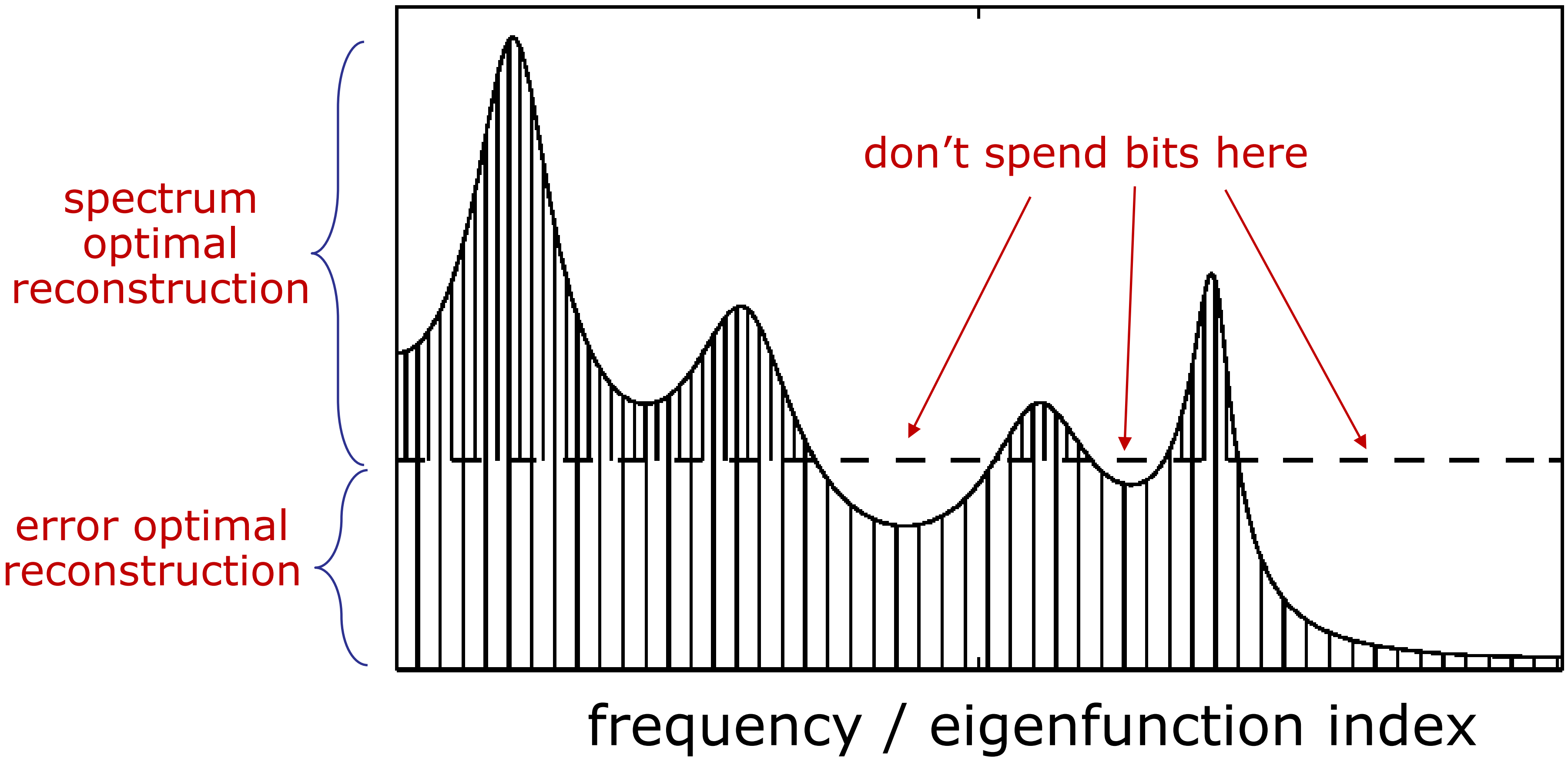
Architecture Goal Illustration

- 3-bit/dimension constrained-resolution quantizer



Reverse Waterfilling

- Code only where needed



spectrum optimal reconstruction

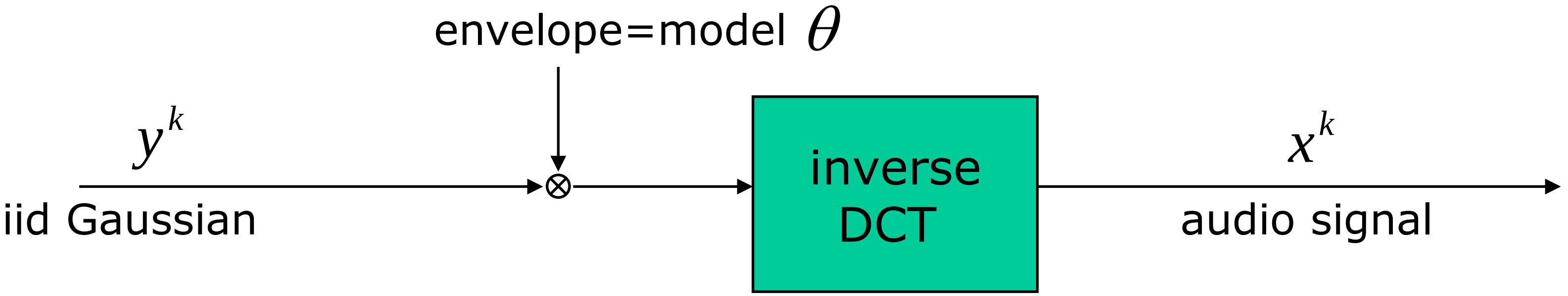
error optimal reconstruction

don’t spend bits here

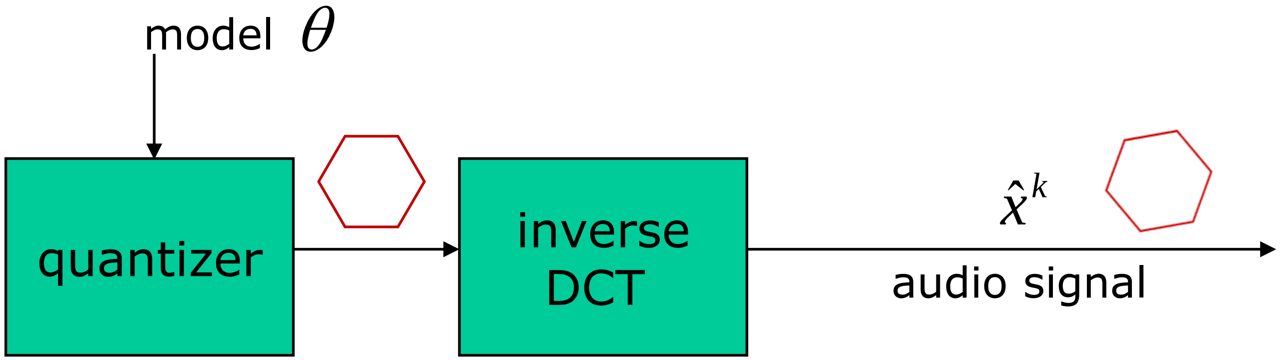
frequency / eigenfunction index

Architecture: Transform

- Model



- Model-based transform coder (CR case)

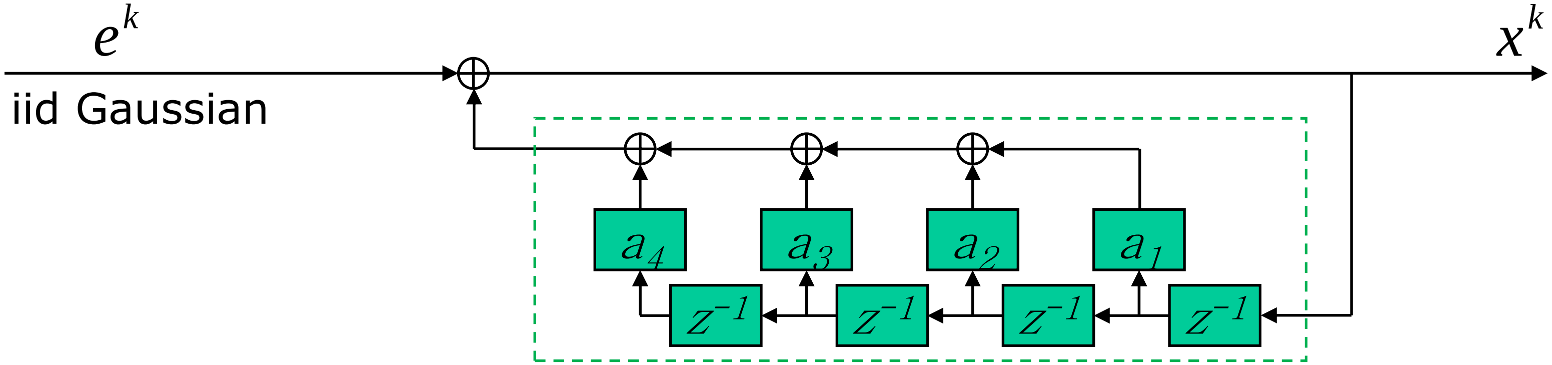


Kleijn 0811

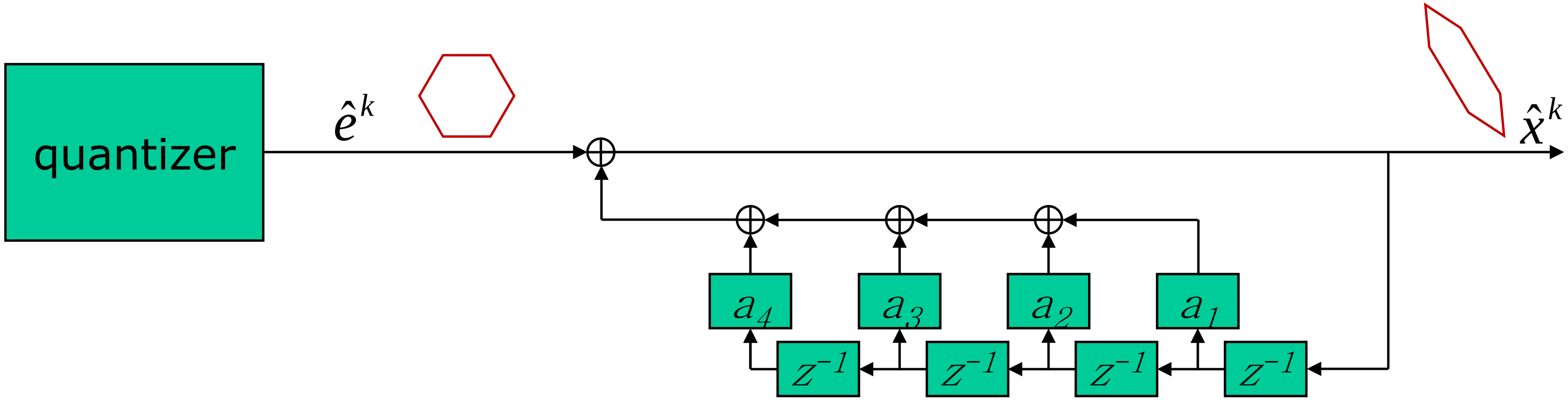
Architecture: AR Model

- Model

$$f_{X^k|\Theta}(x^k \mid \theta) = \frac{1}{\sqrt{2\pi \det(R_{X^k})}} \exp\left(-\tfrac{1}{2} x^k R_{X^k}^{-1} x^k\right)$$



- CELP coding

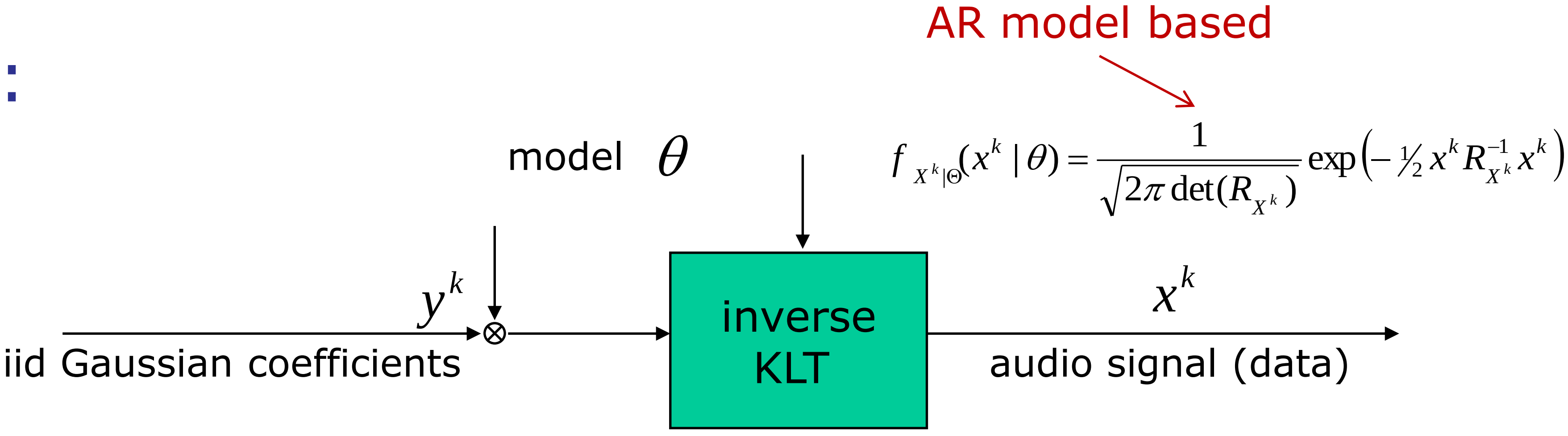


Architecture Comparison

- Unitary transform
  - Does not affect space-filling
  - Reverse water filling
  - Imperfect decorrelation for fixed transform
  - Model not specified
- CELP (analysis-by-synthesis AR coder)
  - AR model functions well
  - Inefficient space filling
  - No inherent reverse water filling (requires *postfilter*)
  - Nightmare for adaptive coding (no theory)

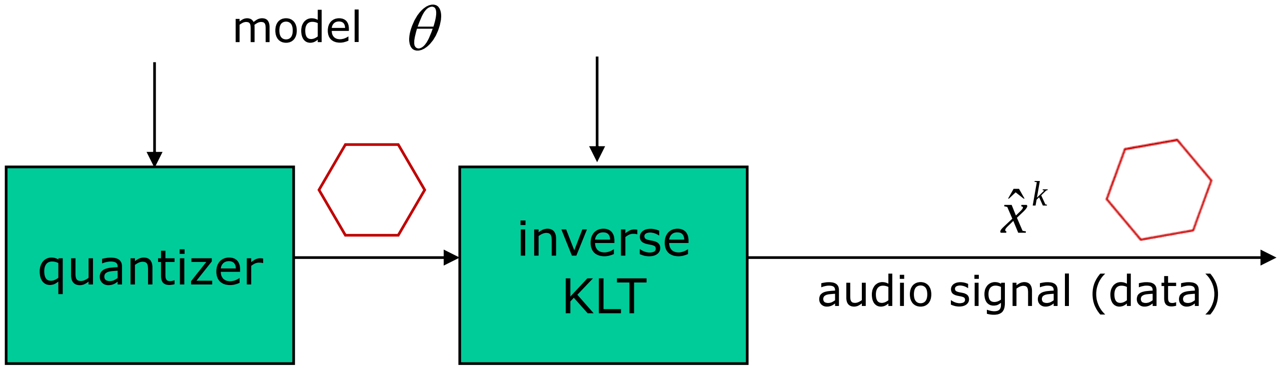
FlexCode Architecture

- Model:

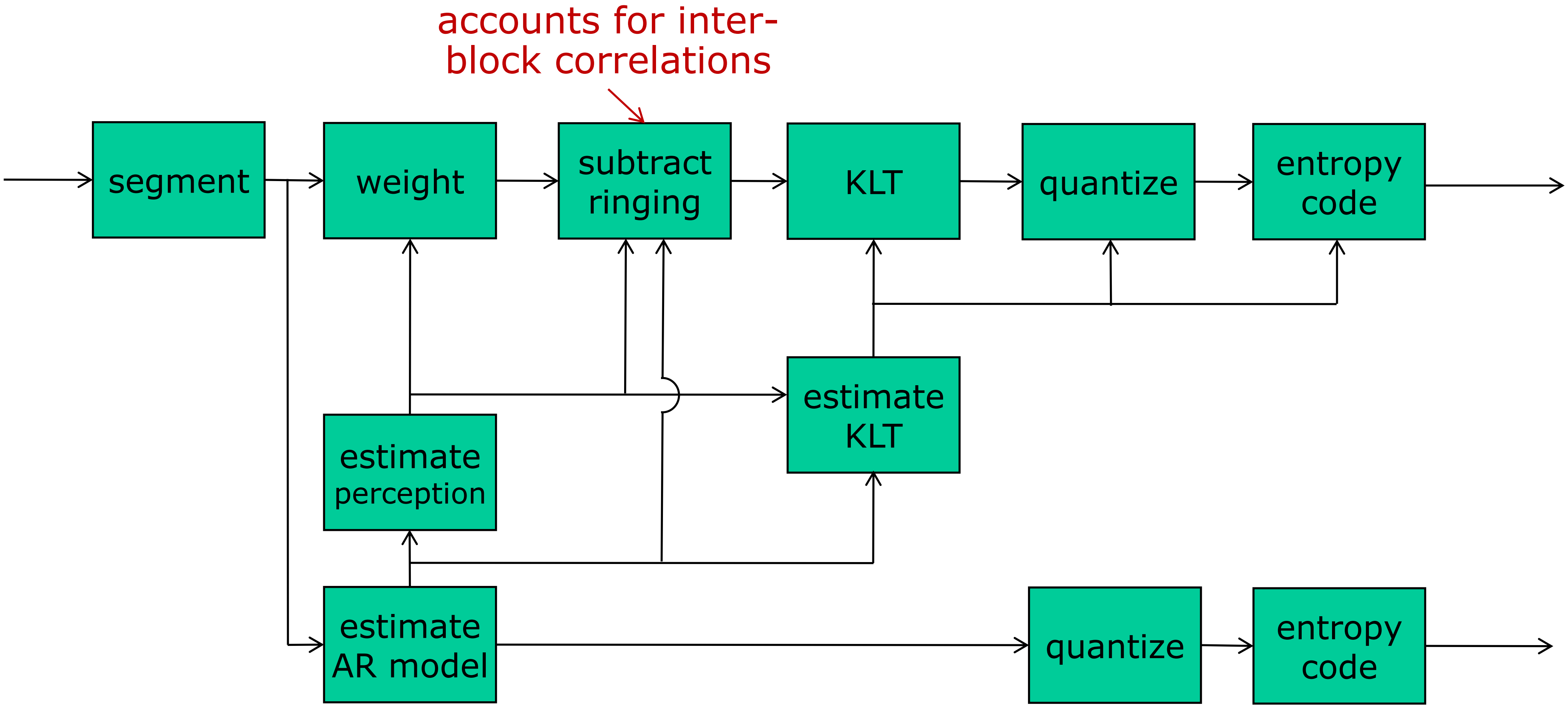


$$f_{X^k|\Theta}(x^k \mid \theta) = \frac{1}{\sqrt{2\pi \det(R_{X^k})}} \exp\left(-\tfrac{1}{2} x^k R_{X^k}^{-1} x^k\right)$$

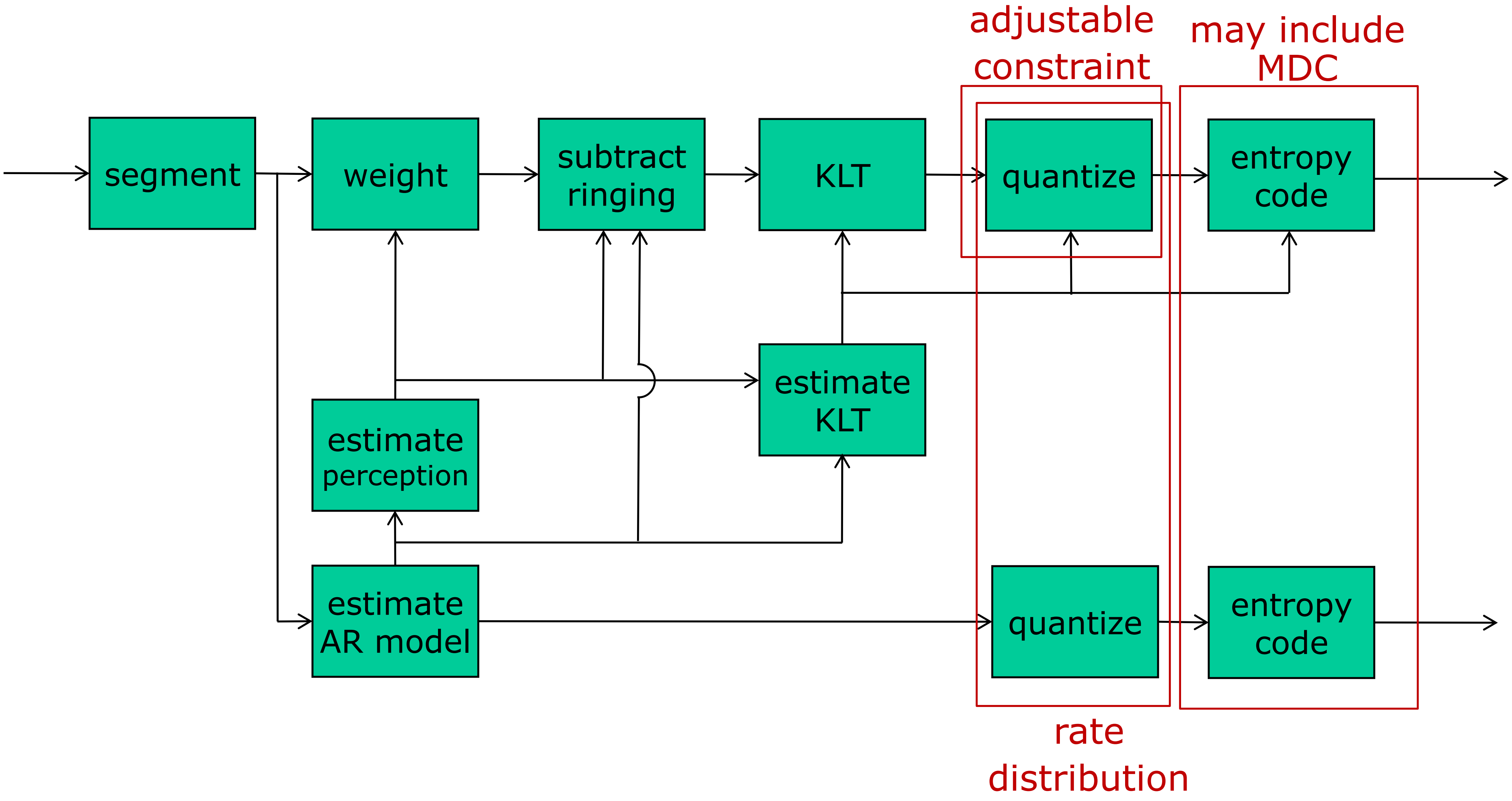
- Model-based transform coder
- KLT based on estimated AR model



FlexCode Architecture



FlexCode Architecture



Performance FlexCode Architecture

- Extensive MUSHRA testing
- Comparison to:
  - 3GPP AMR wide-band/G.722.2
  - G.729.1
  - G.722.1
- Performance FlexCode scalable architecture
  - Worse at 14 kb/s
  - Equivalent at 24 kb/s
  - Better at 32 kb/s
  - KLT performs better than DCT

Summary of Architecture

- Goal is to make scalar quantization effective
- CELP not optimal
  - Poor cell shape
  - No reverse waterfilling
  - Requires postfilters
- Proposed architecture: adaptive transform
  - Best of transforms
  - Best of modeling

Conclusions

- Network heterogeneity ⇒ adaptive coders
- *Modeling everything* facilitates adaptive coding
- Specific techniques
  - Variable-constraint coding reduces effect outliers
  - Rate-dist theory enables adaptive model-based coding
    - *Predicts existing results without invoking perception*
  - Scalable MDC extends scalable coding to environments with packet loss
- Architecture sets effectiveness low-D coding
  - CELP not naturally scalable
  - KLT-based architecture performs best