# Flexible Audio Coding Scheme Based on the Autoregressive Model

*Alexey Ozerov and W. Bastiaan Kleijn*

Royal Institute of Technology, Stockholm, Sweden

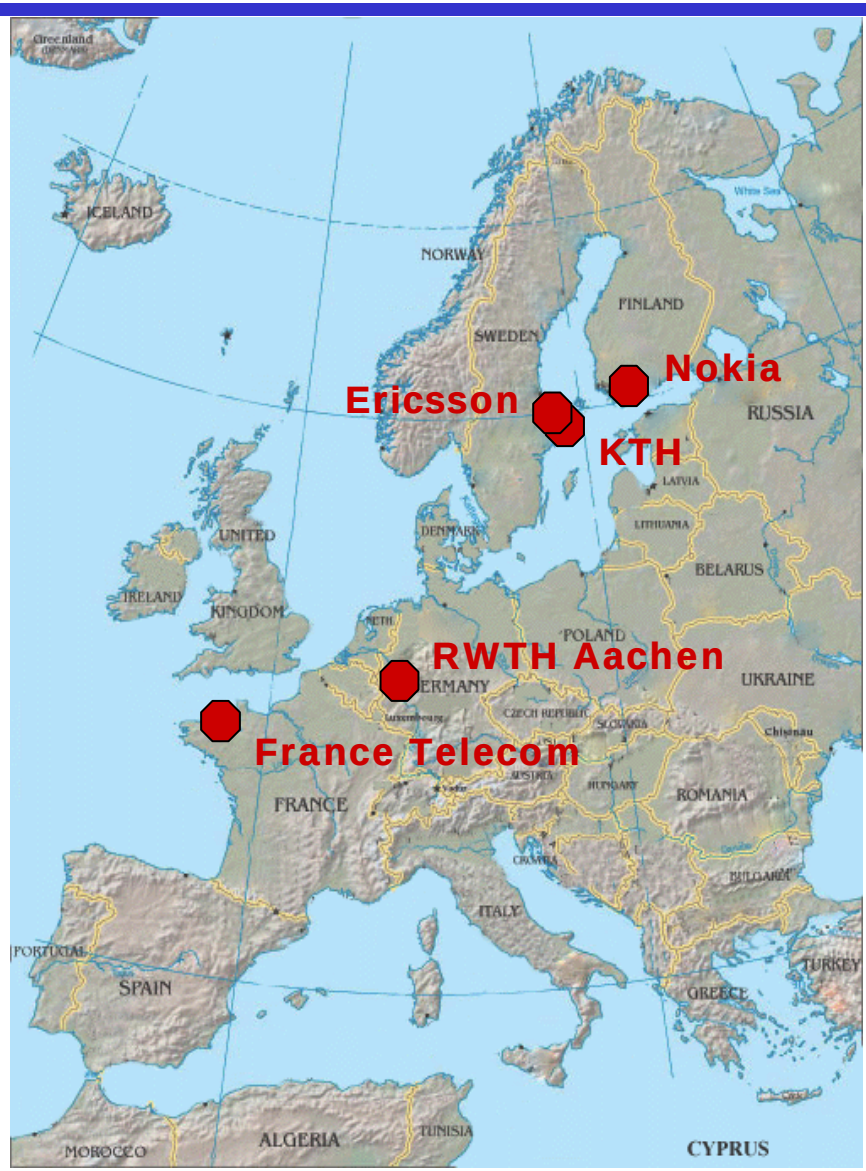
Seminar at IRISA
September 24, 2007

# Outline

- Flexcode in a Nutshell
- Basics of Adaptive Quantization under High-Rate Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

FlexCode

# Who?

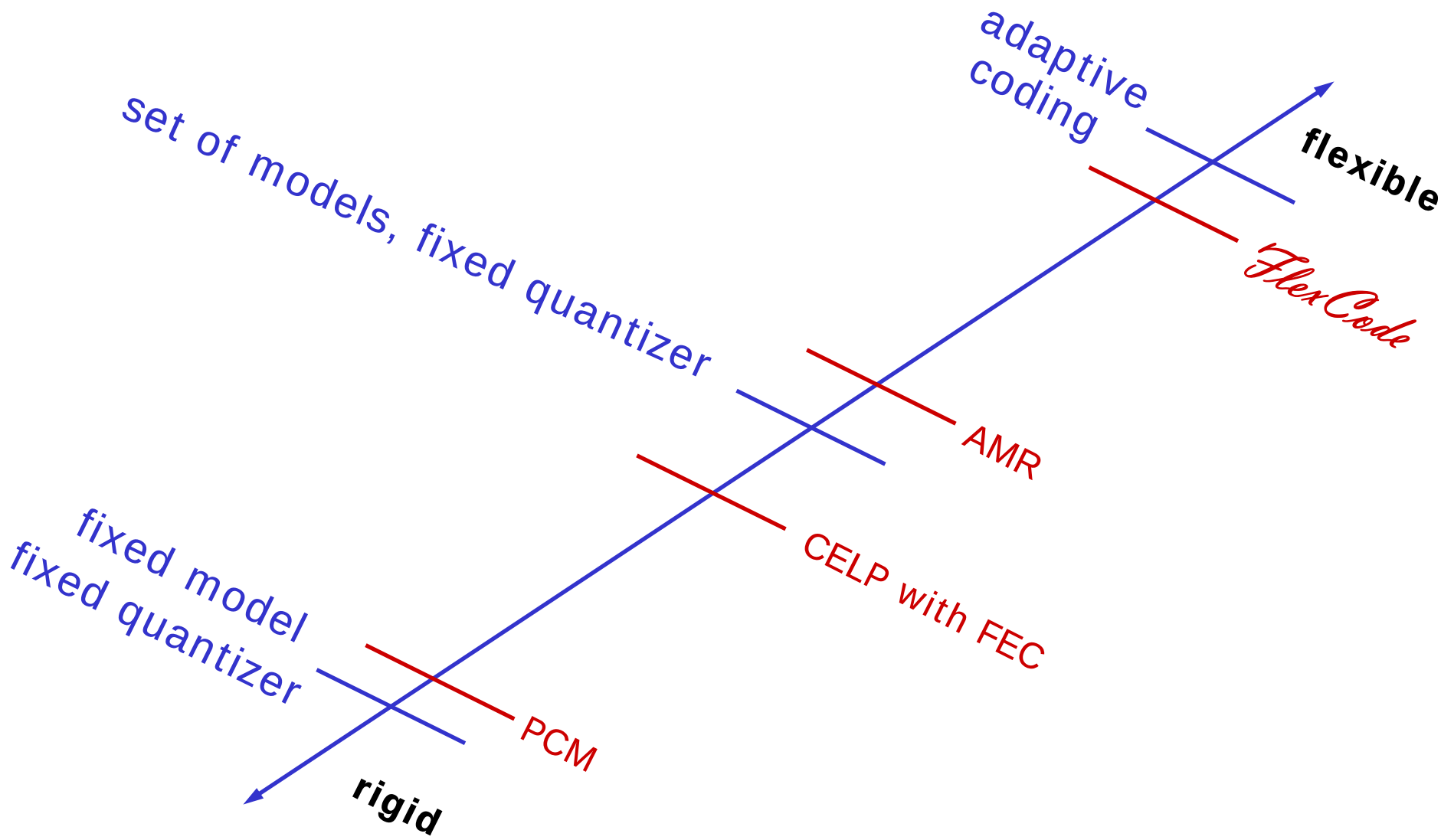


# The Problem

- Heterogeneity of networks increasing
- Networks inherently variable (mobile users)

- But:
  - Coders not designed for specific environment
  - Coders inflexible (codebooks and FEC)
  - Feedback channel underutilized

# Adaptation and Coding



# The Tools

- Tools include
  - Models of source, channel, receiver
  - High-rate quantization theory
  - Multiple description coding (MDC)
  - Iterative source-channel decoding
  - Distortion measures using the sensitivity matrix

# Outline

- Flexcode in a Nutshell
- Introduction
- Basics of Adaptive Quantization under High-Rate Theory Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

# How Can we Code an Audio Signal?

- Irrelevance
  - Parts of the signal that we cannot perceive
  - *FlexCode* : use advanced auditory model expressed in terms of sensitivity matrix

- Redundancy
  - Statistical dependencies that allows the information to be expressed with fewer bits
  - *FlexCode* : use a probabilistic model of the signal

# Motivation

- Redundancy reduction
- Conventional codebook-based approaches
  - Train a codebook for a particular rate
  - Or train a set of codebooks for a set of the rates
- In *FlexCode* we want
  - Coder that is able to run for any rate from the continuum of the rates
  - Computational complexity to be independent on the rate
- Thus, we cannot train codebooks, we need to compute them on the fly
- Probabilistic source modeling together with high-rate theory approximation allows that

# Outline

- Flexcode in a Nutshell
- Introduction
- Basics of Adaptive Quantization under High-Rate Theory Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

# High-Rate Theory

$x^k \in \mathbb{R}^k$ source vector

$f_{X^k}(x^k)$ its pdf

Distortion in a quantization cell :

$$D_i = \frac{1}{k}\frac{\int_{\mathbf{V_i}} f_{X^k}(x^k)\|x^k - Q(x^k)\|^2 dx^k}{\int_{\mathbf{V_i}} f_{X^k}(x^k) dx^k}$$

HR approx. $$\approx \frac{1}{k}\frac{f_{X^k}(c_i^k)\int_{\mathbf{V_i}} \|x^k - c_i^k\|^2 dx^k}{f_{X^k}(c_i^k)\int_{\mathbf{V_i}} dx^k} = \frac{1}{kV_i}\int_{\mathbf{V_i}} \|x^k - c_i^k\|^2 dx^k$$

- optimal for high-rate
- still work well for low-rate

# Two Types of Quantization

- Constrained Resolution (CR) quantization
  - Fixed number of bits per vector
  - R bits per vector = $2^R$ codewords in the codebook
- Constrained Entropy (CE) quantization
  - Any number of bits per vector (variable rate)
  - The average rate (or the entropy of the codeword indices) is constrained

$$H(I) = -\sum_{i \in \mathcal{I}} p_I(i) \log(p_I(i)) = R$$

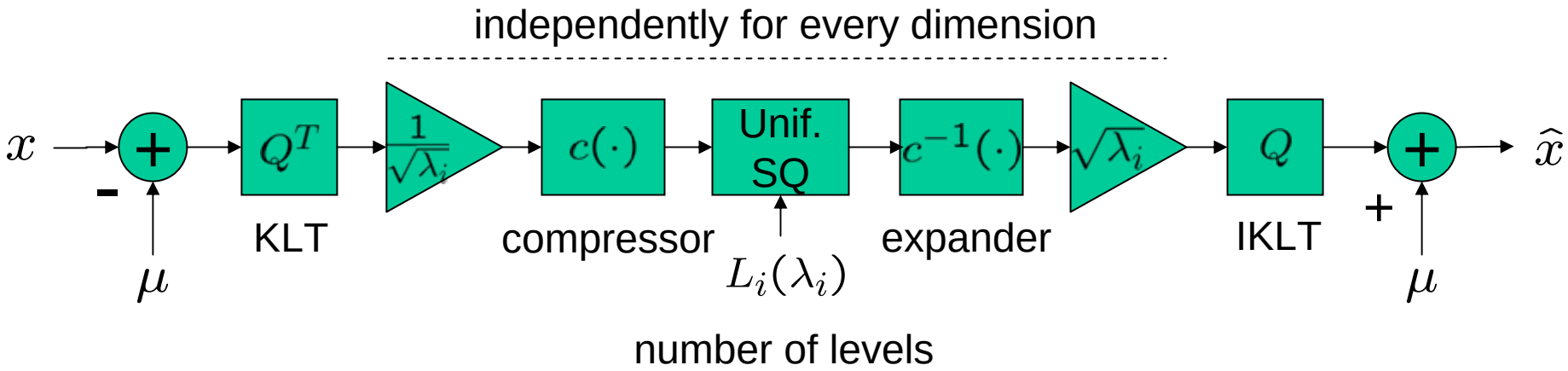
- CE performs better than CR, but needs a more flexible transmission channel

# Single Gaussian case

- CR quantization (with companded scalar quantizers)

$$X \in \mathbb{R}^k, \qquad X \sim \mathcal{N}(\mu, \Sigma)$$

EVD $\qquad \Sigma = Q \Lambda Q^T \qquad Q^T Q = I \qquad \Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$

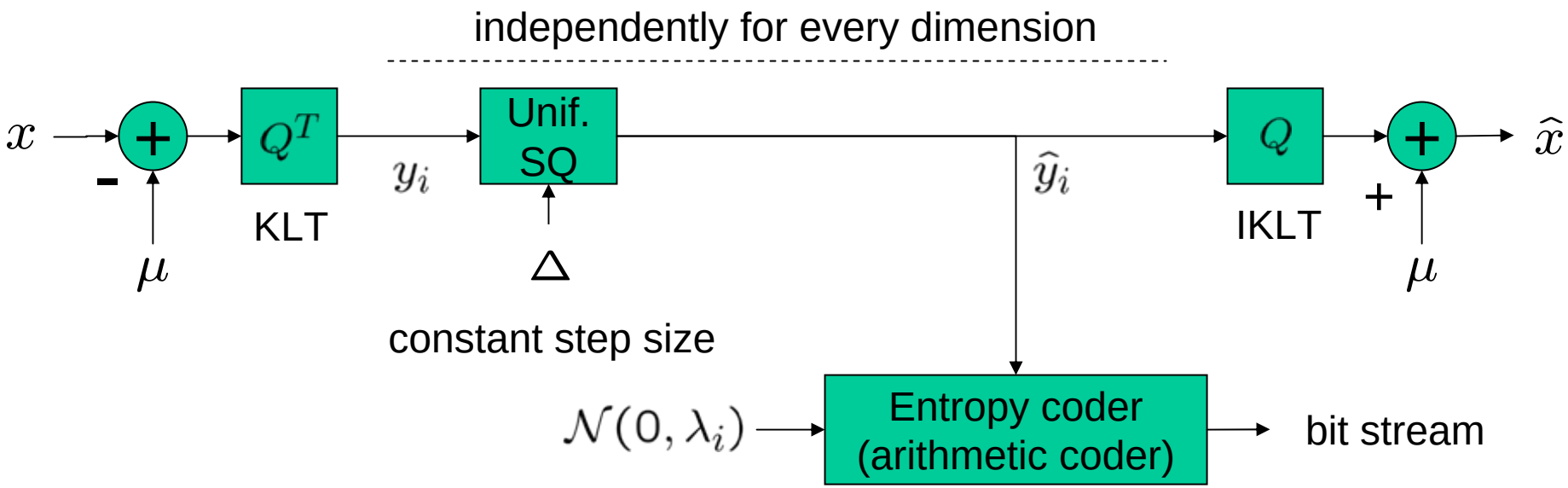


# Single Gaussian case

- CE quantization (with scalar quantizers)

$$X \in \mathbb{R}^k, \qquad X \sim \mathcal{N}(\mu, \Sigma)$$

EVD $\quad \Sigma = Q \Lambda Q^T \qquad Q^T Q = I \qquad \Lambda_i = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$



# Single Gaussian case

- We see that with this approach
  - we can quantize with any rate
  - the computational complexity is independent on the particular rate

- We can do better using vector lattice quantizers instead of scalar quantizers
  - we can gain up to 0.25 bits per sample in rate
  - which is equivalent to 1.5 dB in distortion

# Gaussian Mixture Model (GMM) case

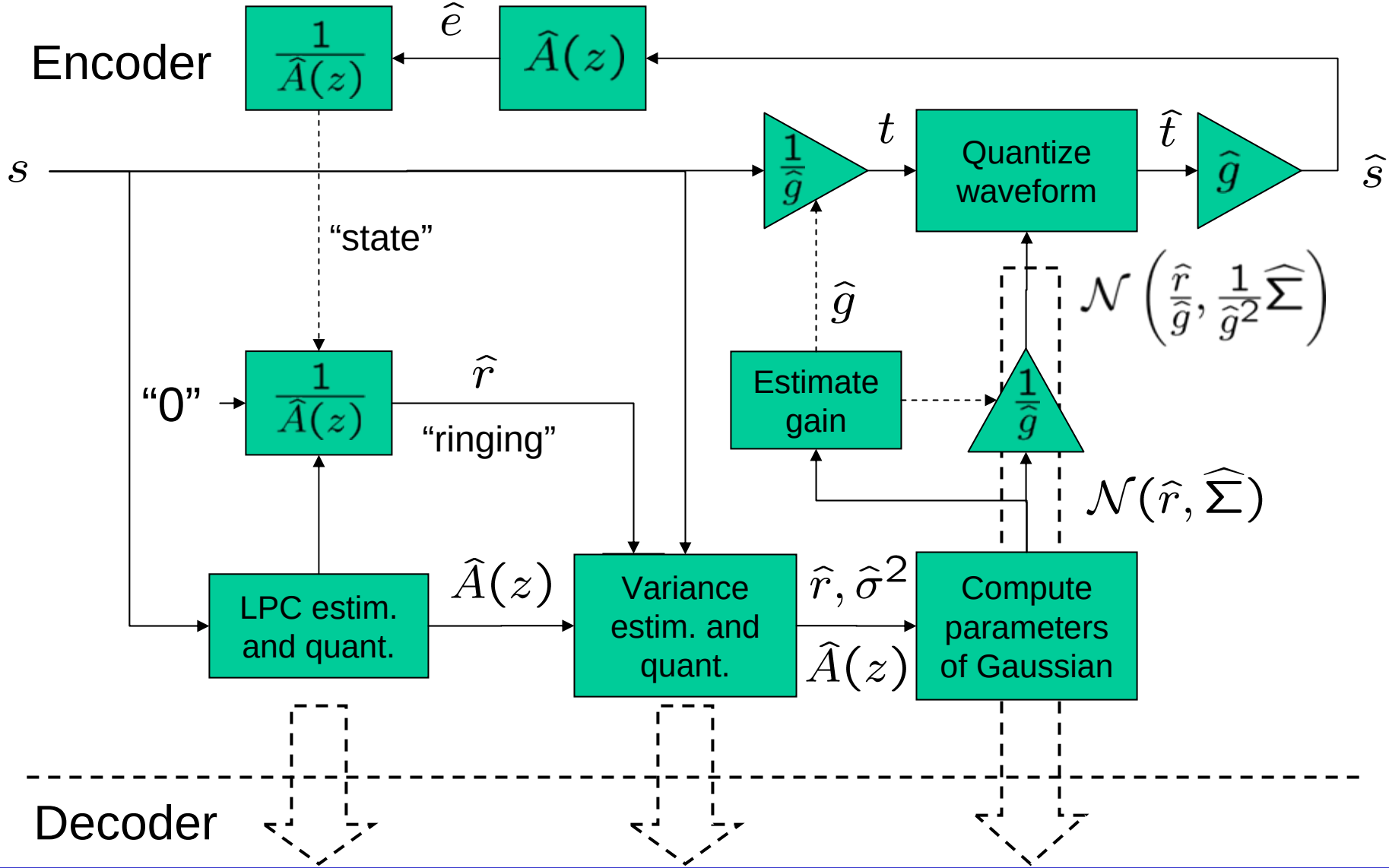
- With GMM the quantization consists in the following steps:
  - For each input vector *x*, choose the component (state) maximizing the *a posteriori* probability *p(i|x)*
  - Quantize using selected Gaussian component (in CR or CE case), as described before

- With this approach we loose in optimality, when the Gaussian components are not well separated

# Outline

- Flexcode in a Nutshell
- Introduction
- Basics of Adaptive Quantization under High-Rate Theory Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

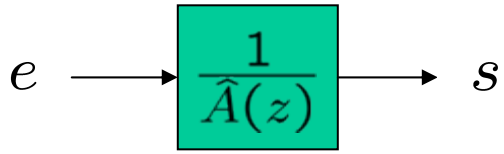
# Encoder structure



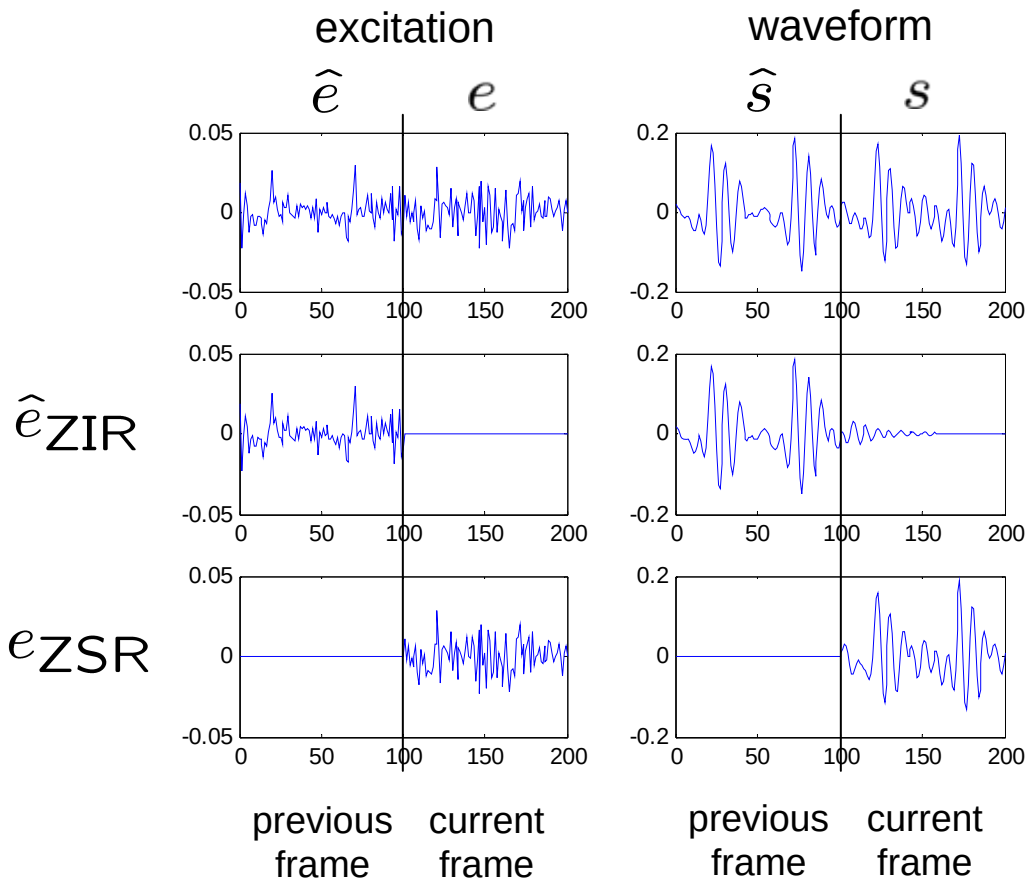
# LPC estimation and quantization

- LPCs are estimated as in AMR-WB coder
  - Estimate LPC for every 5 ms frame
- LPCs are quantized in LSP domain using a GMM
- Quantized LPC are interpolated in LSF domain for every 1.25 ms subframe (as in AMR-WB coder)

# "Ringing" (or ZIR) computation




$$s = \hat{s}_{\text{ZIR}} + s_{\text{ZSR}}$$




"Ringing" or

Zero Input Response (ZIR)

$$\hat{r} = \hat{s}_{\text{ZIR}}$$

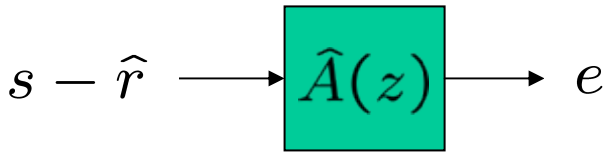
Zero State Response (ZSR)

$$s_{\text{ZSR}}$$

$$t = s_{\text{ZSR}} \rightarrow \hat{t}$$

# Variance estimation and quantization

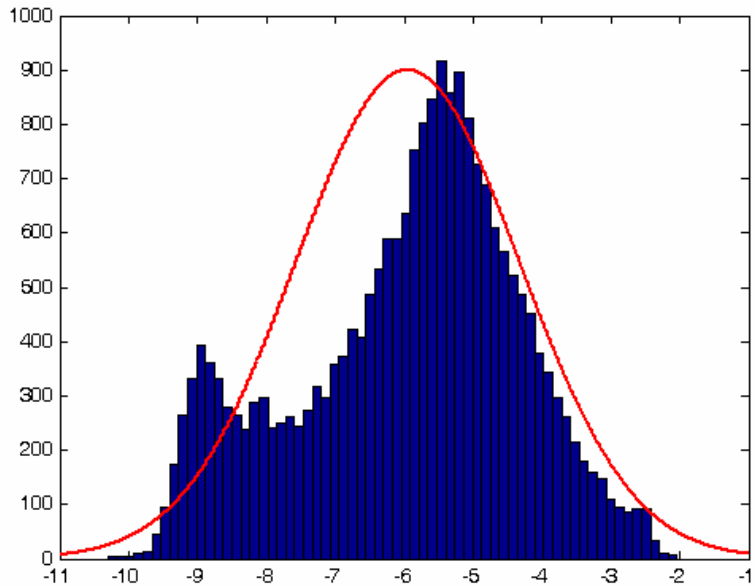
- Variance estimation
  - in ML sense




compute variance

$$\sigma^2 = \text{var}(e)$$

- Variance quantization
  - Modeled by a single Gaussian in log-domain
  - Quantized using this distribution



# Compute parameters of Gaussian

"ringing" $\hat{r}$

LPC and variance

$$\frac{\hat{\sigma}_e}{\hat{A}(z)} = \frac{\hat{\sigma}_e}{1+\hat{a}_1 z^{-1}+\ldots+\hat{a}_p z^{-p}}$$

then $\boxed{s \sim \mathcal{N}(\hat{r}, \widehat{\Sigma})}$

It is conceptually important to consider "ringing" as a part of the model

where $\widehat{\Sigma} = \hat{A}^{-1}\hat{A}^{-T}$

$\hat{A}$ is a lower triangular Toeplitz (k x k) matrix with as first column

$$\hat{\sigma}^{-1}[1, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p, 0, \ldots, 0]^T$$

# Gain estimation

- This gain should be considered as a part of the model of perception
- Have a sense for CE case only

$$s \sim \mathcal{N}(\hat{r}, \widehat{\Sigma})$$

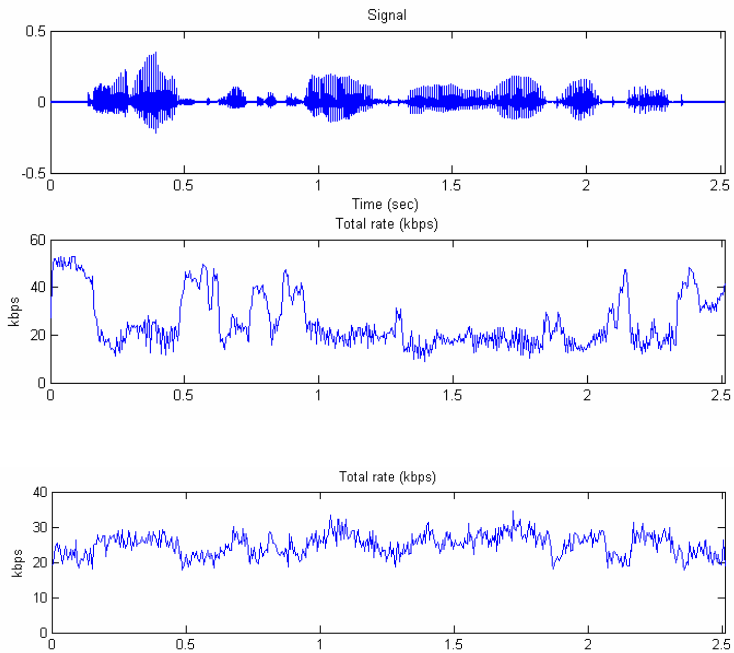
$$\hat{g} = \sqrt{\frac{1}{k} E[s^T s]} = \sqrt{\frac{1}{k}(\hat{r}^2 + \mathrm{tr}(\widehat{\Sigma}))}$$

seems to be good for SSNR

alternatively

$$\hat{g} = \hat{\sigma}$$

what is the optimal gain for ear?



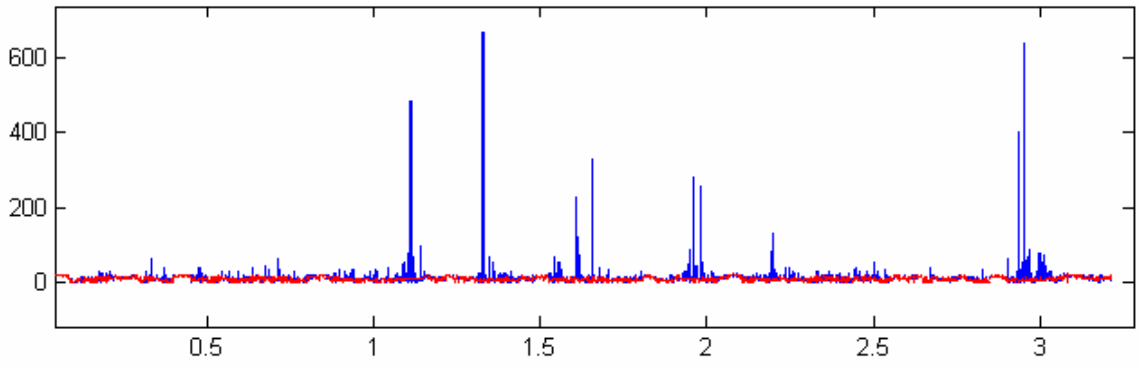
# Instability problem

- For low rates and short frames this system can become unstable
- The reason : closed loop between “ringing” computation and variance estimation
- More deep reason : $s \sim \mathcal{N}(\hat{r}, \widehat{\Sigma})$
  - there is no constraint for “ringing”

- That is why it is conceptually important to consider “ringing” as a part of the model

# Instability problem

- Solution 1 : estimate variance using “ringing” computed from non-quantized signal (so the closed loop is broken)
  - Works for CR case
  - Does not work for CE case, since it gives sometimes a very small likelihood => outliers => catastrophic bursts in the rate



# Instability problem

- Solution 2 : constraint “ringing” during signal quantization
  - Helps in CR case (in addition to solution 1)
  - Works in CE case
- Idea :

$$d_1(\hat{s}, s) = \|\hat{s} - s\|^2$$

$$d_2(\hat{s}, s) = \|\hat{s} - s\|^2 + \alpha \|\hat{r}_{\text{nxt}} - r_{\text{nxt}}\|^2$$

$$\begin{aligned} d_2(\hat{s}, s) &= \|\hat{s} - s\|^2 + \alpha \|B\hat{s} - Bs\|^2 \\ &= (\hat{s} - s)^T (I + \alpha^2 B^T B)(\hat{s} - s) \end{aligned}$$

# Comparison of AR model based quantization scheme with some existing works



- Kim and Klejn 2004
  - **Common point**: adaptive codebook in the speech domain
  - **Differences**:
    - this scheme needs the codebooks to be trained when rate changes;
    - it is based on full CB search (as in CELP), thus its computational complexity is dependent of the rate
- Samuelsson 2004
  - **Common points**:
    - using of model (he uses GMM);
    - computational complexity is independent of the speech rate
  - **Differences**:
    - this scheme needs a lot of data to train a representative GMM;
    - GMM with many states cannot be used (because of unmanageable computational complexity and scarcity of training data);
    - computational complexity is dependent of the model size (somehow can be viewed as model rate in our case);
    - we use optimal distribution of rate between signal and model

# Comparison with a CELP scheme

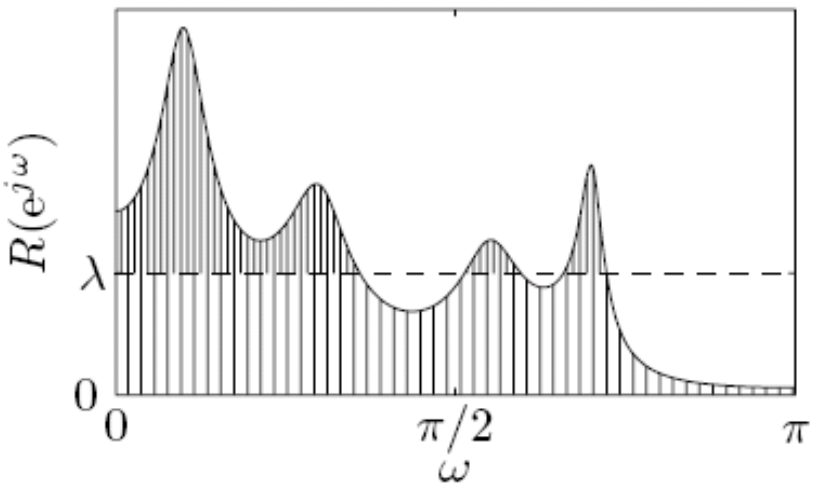
- We compare with a CELP (analysis by synthesis scheme) with a CB trained minimizing MSE in speech domain
  - 8 kHz speech, frame length = 5 samples,
  - Rate = 19.2 kbps (12 bits per frame)

| | AR coder (CR case) | AR coder (CE case) | CELP |
|---|---|---|---|
| Gain rate | 3 | 2.7 | 5 |
| Speech rate | 9 | 9.3 | 7 |
| Av. SSNR | 15.8 | 17.85 | 17.54 |

  - This is with scalar quantizers, and for quite rate

# Reverse water-filling

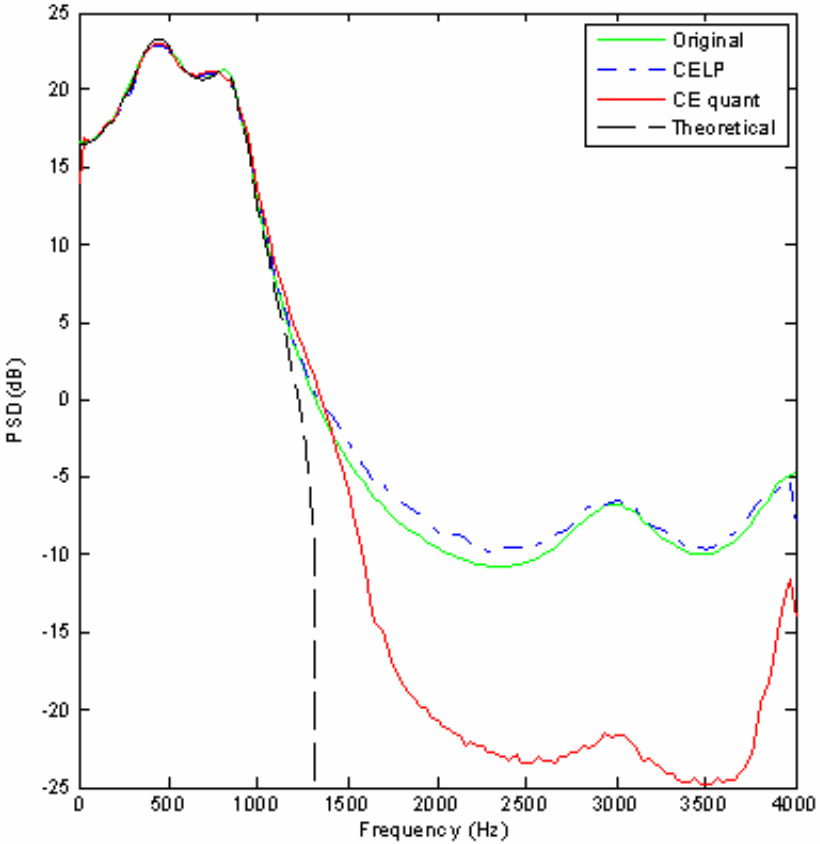
original power spectrum →

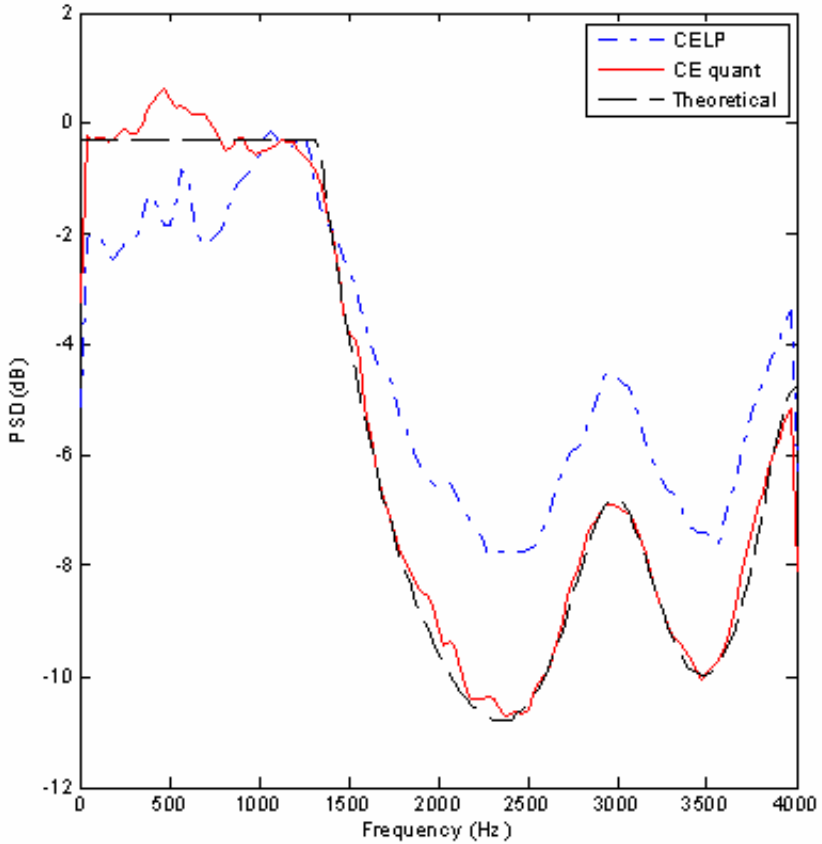



← reconstruction power spectrum

← distortion

# Reverse water-filling



# CE vs. CR quantization

- Disadvantage
  - Need a more flexible channel, since rate varies
- Advantages
  - Constant distortion is very good for perception
  - Constant distortion allows to avoid many instability problems (coder design becomes much easier than in CR case)

# Outline

- Flexcode in a Nutshell
- Introduction
- Basics of Adaptive Quantization under High-Rate Theory Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

# Optimal Bit Allocation between Signal and Model

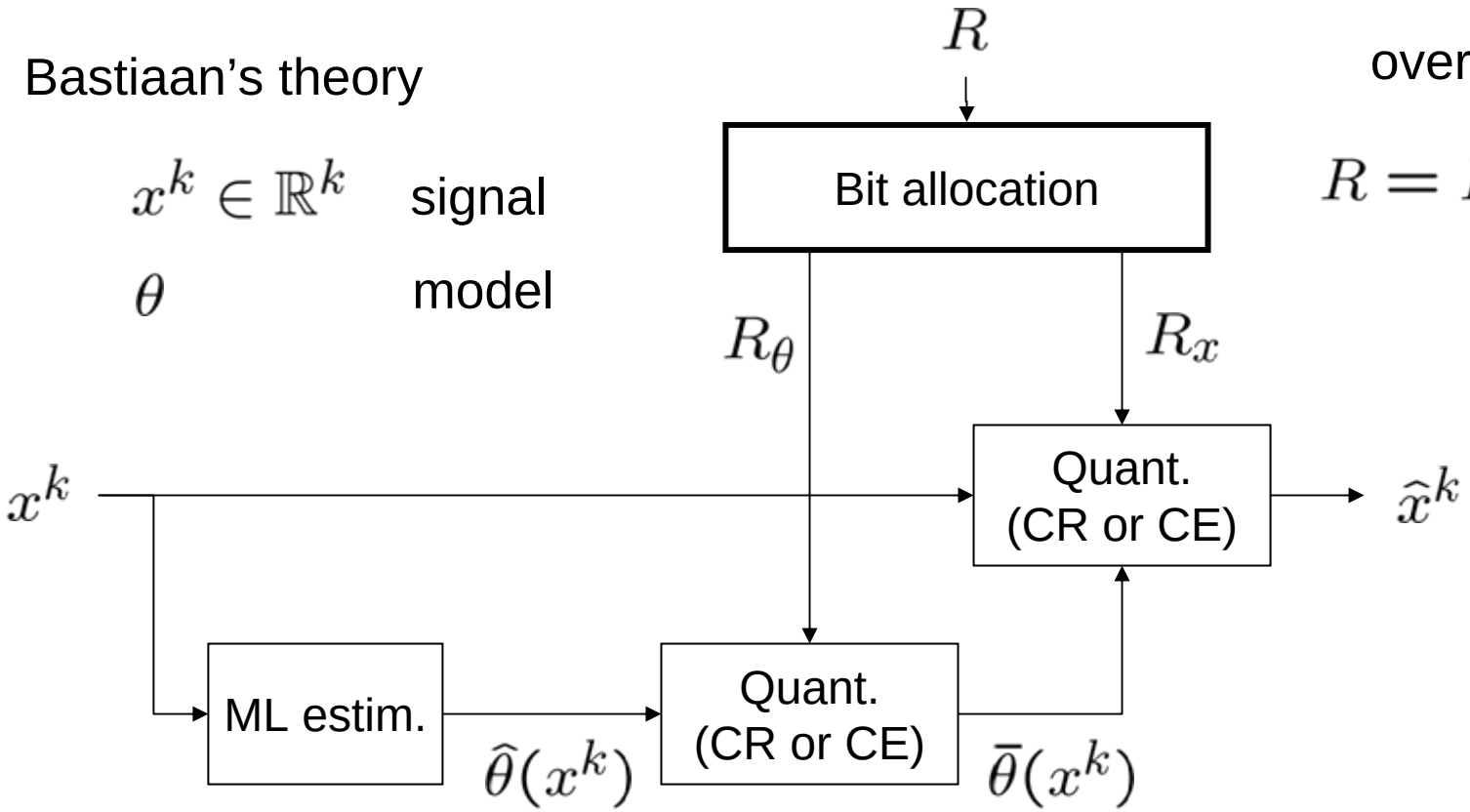
Bastiaan's theory

$x^k \in \mathbb{R}^k$ signal

$\theta$ model

overall rate

$R = R_x + R_\theta$




Problem: given the overall rate R, what is the optimal bit allocation between signal and model?

# Optimal Bit Allocation between Signal and Model

## Solution in CE case (in CR case it is analogous)

Codeword length

$$
\begin{aligned}
L(x^k) &= L_{\bar{\Theta}}(\bar{\theta}(x^k)) + L_{X^k|\bar{\Theta}}(x^k|\bar{\theta}(x^k)) \\
&= -\log\left(p_{\bar{\Theta}}(\bar{\theta}(x^k))\right) - \log\left(p_{X^k|\bar{\Theta}}(x^k|\bar{\theta}(x^k))\left(\frac{D}{C}\right)^{k/2}\right) \\
&= \psi\left(\bar{\theta}(x^k), \hat{\theta}(x^k), x^k\right) - \log\left(p_{X^k|\hat{\Theta}}(x^k|\hat{\theta}(x^k))\left(\frac{D}{C}\right)^{k/2}\right)
\end{aligned}
$$

$$
\psi\left(\bar{\theta}(x^k), \hat{\theta}(x^k), x^k\right) = -\log\left(p_{\bar{\Theta}}(\bar{\theta})\right) - \log\left(\frac{p_{X^k|\bar{\Theta}}(x^k|\bar{\theta})}{p_{X^k|\hat{\Theta}}(x^k|\hat{\theta})}\right)
$$

*index of resolvability*

# Optimal Bit Allocation between Signal and Model

Solution in CE case (in CR case it is analogous)

Overall rate

$$R = \mathbb{E}\left[L(X^k)\right] = \overbrace{-\mathbb{E}\left[\psi\left(\bar{\theta}(X^k), \hat{\theta}(X^k), X^k\right)\right]}^{\text{dependent on } R_\theta}$$
$$\underbrace{-\mathbb{E}\left[\log\left(p_{X^k|\hat{\Theta}}(X^k|\hat{\theta}(X^k))\right)\right] - \frac{k}{2}\log\left(\frac{D}{C}\right)}_{\textbf{const} \text{ indep. on rates}}$$

$$D = F(R, R_\theta) = C \exp\left\{\frac{k}{2}\left(R + \text{const} + \mathbb{E}\left[\psi\left(\bar{\theta}(X^k), \hat{\theta}(X^k), X^k\right)\right]\right)\right\}$$

$$R_\theta^{\text{opt}} = \arg\min_{R_\theta} F(R, R_\theta) = \arg\min_{R_\theta} \mathbb{E}\left[\psi\left(\bar{\theta}(X^k), \hat{\theta}(X^k), X^k\right)\right]$$

=> Under HR assumptions, the optimal rate for the model is independent on the overall rate

# Optimal Bit Allocation between Signal and Model

Then these general results are applied for speech coding using AR model

We also show using some approximations that in this particular case

$$\mathbb{E}\left[\psi\left(\bar{\theta}(X^k),\hat{\theta}(X^k),X^k\right)\right] \approx R_\theta + \frac{k}{4}\mathbb{E}\left[\mathsf{SLSD}\left(\bar{\theta}(X^k),\hat{\theta}(X^k)\right)\right]$$

SLSD = Square Log Spectral Distortion

# Optimal Bit Allocation between Signal and Model

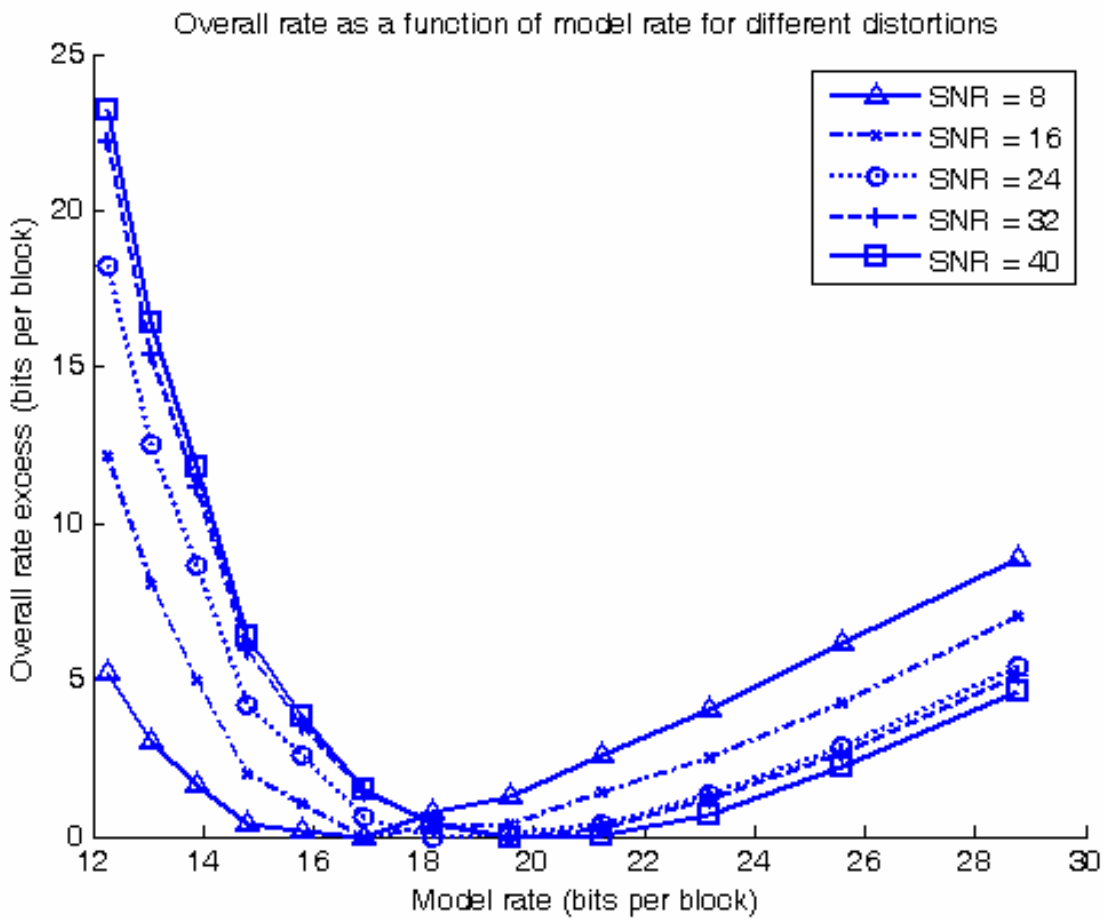
## Confirmation

Table 1: Bit rates of the AMR-WB coder [7].

| Rate | 6.6 | 8.85 | 12.65 | 14.25 | 15.85 | 18.25 | 19.85 | 23.05 |
|---|---|---|---|---|---|---|---|---|
| AR model parameters | 36 | 46 | 46 | 46 | 46 | 46 | 46 | 46 |
| pitch-model parameter | 23 | 26 | 30 | 30 | 30 | 30 | 30 | 30 |
| excitation | 48 | 80 | 144 | 176 | 208 | 256 | 288 | 352 |

# Optimal Bit Allocation between Signal and Model

## Experimental verification




Theoretically predicted rate =

19.0 bits per block

# Optimal Bit Allocation between Signal and Model

Number of sentences = 10,

Fs = 8000 Hz, Frame length = 2.5 ms (20 samples),

Total bits per frame

| Bits for gain | 20 bits | 40 bits | 60 bits | 80 bits | 100 bits |
|---|---|---|---|---|---|
| 2 bits | 7.21 | 13.17 | 18.30 | 23.66 | 28.90 |
| 3 bits | 8.97 | 15.45 | 21.02 | 26.85 | 32.76 |
| 4 bits | **9.37** | **16.06** | **21.63** | **27.38** | **33.46** |
| 5 bits | 9.19 | 15.97 | 21.53 | 27.31 | 33.38 |
| 6 bits | 8.83 | 15.70 | 21.29 | 26.98 | 33.10 |
| 7 bits | 8.41 | 15.43 | 21.00 | 26.71 | 32.80 |

# Optimal bit allocation between signal and model

Summary

- General result
  - When the signal is quantized based on some already quantized model and the HR assumptions are verified, **the optimal rate for the model is independent on the overall rate**
  - This result is true for any model and in both CR and CE cases
- More-over, for AR model quantization it is shown that minimization of mean Square Log Spectral Distortion it is near optimal (this result is obtained theoretically without using any perceptual knowledge)

# Outline

- Flexcode in a Nutshell
- Introduction
- Basics of Adaptive Quantization under High-Rate Theory Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

# Open Issues

- Integration of pitch model (long term prediction)
- Model of perception
  - Which model to use?
  - Transmit the information about the model of perception or not?
- Computational complexity
  - SVD has a comp. complex. of order $O(N^3)$ ($N = 80$)
  - Can we replace KLT by some fixed transform (DFT, DCT, MDCT)?

# Outline

- Flexcode in a Nutshell
- Introduction
- Basics of Adaptive Quantization under High-Rate Theory Assumptions
- Flexible Coding Scheme
- Optimal Bit Allocation between Signal and Model
- Open Issues
- Conclusion

# Conclusion

- Rate
  - This scheme can run for any rate from the continuum of the rates
  - Computational complexity is independent on the rate
- CE quantization has a lot of advanteges compared to CR quantization
- Clarity and simplicity of the scheme
  - Source and perception models are well separatred
  - No tweaking (at least at the current stage of development)