Recent Advances in Model-based Transform Audio Coding

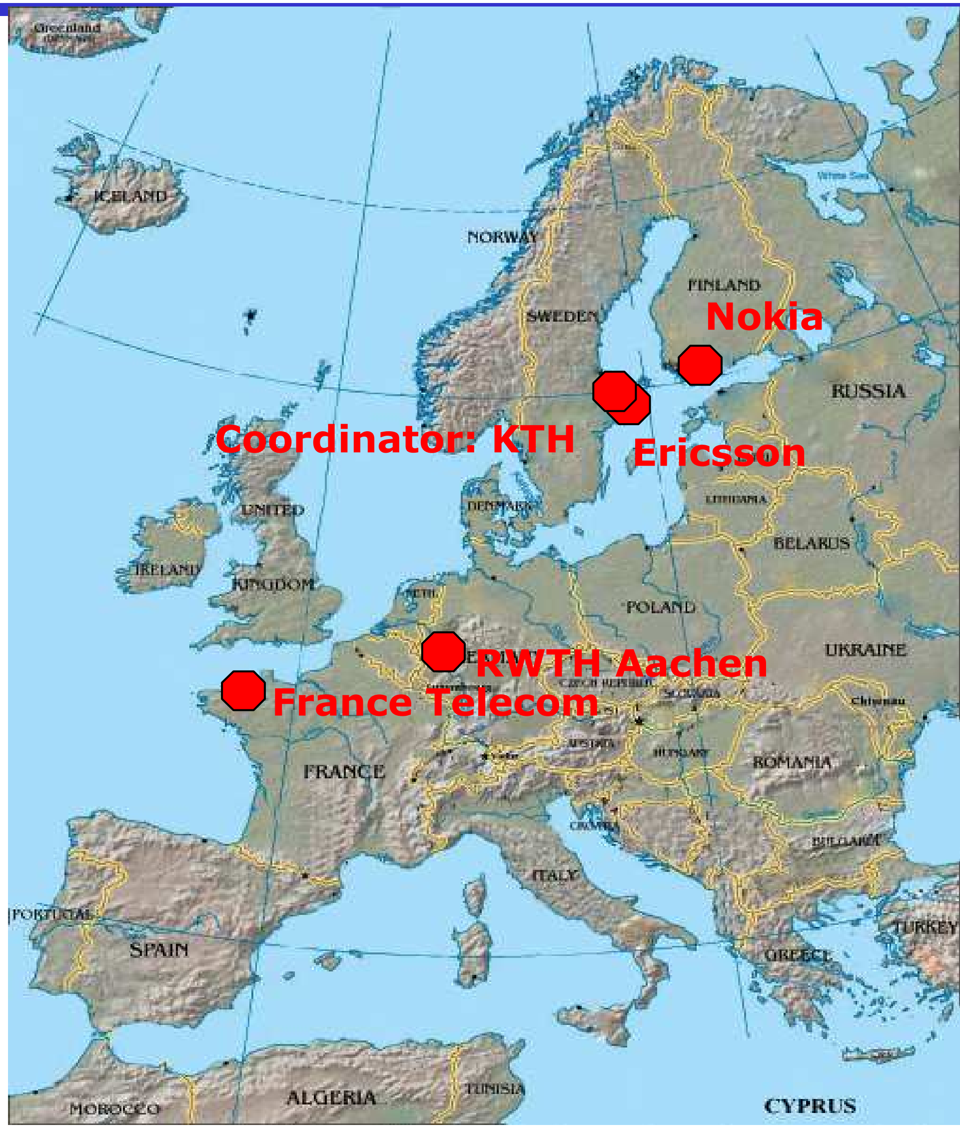
Marie Oger, Stéphane Ragot

Flexcode Seminar Stockholm

October 17, 2007

Overview of the Flexcode project

- Flexible Coding for Heterogeneous Networks
- Objectives: develop **flexible source-channel coding algorithms**
  - More flexible than current, application-specific coders
  - Flexibility through online design, **generic source, channel and distortion models**
  - **Focus on audio**
- **http://www.flexcode.eu**



Flexible Source Coding

- Approach
  - Start with existing models: transform and linear-predictive coding
  - "Flexcodize": analytic solutions $\rightarrow$ adaptive coding
- Tools for flexible coding include
  - High-rate quantization theory
  - Probability models (GMM, …) for quantizer design
    - Quantizer specification by equations
    - Estimate statistics for source
  - Distortion measures using sensitivity matrix

FlexCode

Motivations and objectives: existing model-based techniques

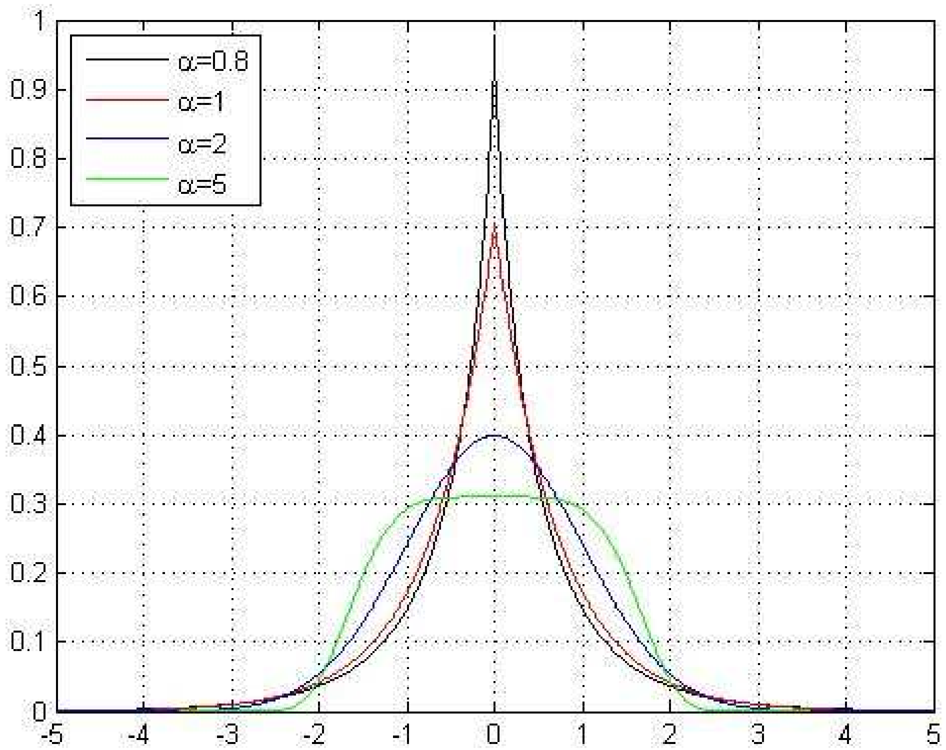
orange™

- GMM (Gaussian Mixture Model) based LPC quantization [Subramaniam 01] [Samuelsson 01]
  - LPC coefficients or prediction error are modeled by a GMM
  - A mean-removed Karhunen-Loeve transform (KLT) and normalization by standard deviations is applied to LPC coefficients
- Autoregressive GMM for speech coding [Samuelsson 04]
  - Companded GMM for vector quantizers (CGMM-VQ)
- Generalized Gaussian model for image coding [Parisot 03]
  - Wavelet coding for image (EBWIC Coder)
  - Wavelet coefficients are modeled by a generalized Gaussian model

Outline of presentation

- Generalized Gaussian model
  - Definition
  - Example
- Proposed stack-run coding with model-based deadzone
  - Principle of stack-run coding
  - Rate control based on asymptotic bit allocation
  - Model-based optimization of deadzone
  - Objective & Subjective results
  - Delay & Complexity
  - Audio samples
- Latest developments: model-based bit plane coding
  - Principle
  - Preliminary results
- Conclusion & perspectives

Generalized Gaussian pdf: definition



- The probability density function (pdf) of a zero-mean **generalized Gaussian variable** z of standard deviation σ is given by :

$$p_{\alpha,\sigma}(z) = \frac{A(\alpha)}{\sigma} e^{-|B(\alpha)z/\sigma|^{\alpha}}$$

where

$$A(\alpha) = \frac{\alpha B(\alpha)}{2\Gamma(1/\alpha)} \quad \text{and} \quad B(\alpha) = \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}$$

with $\Gamma(.)$ the Gamma function defined as

$$\Gamma(\alpha) = \int_0^{\infty} e^{-t}\, t^{\alpha+1}\, dt$$

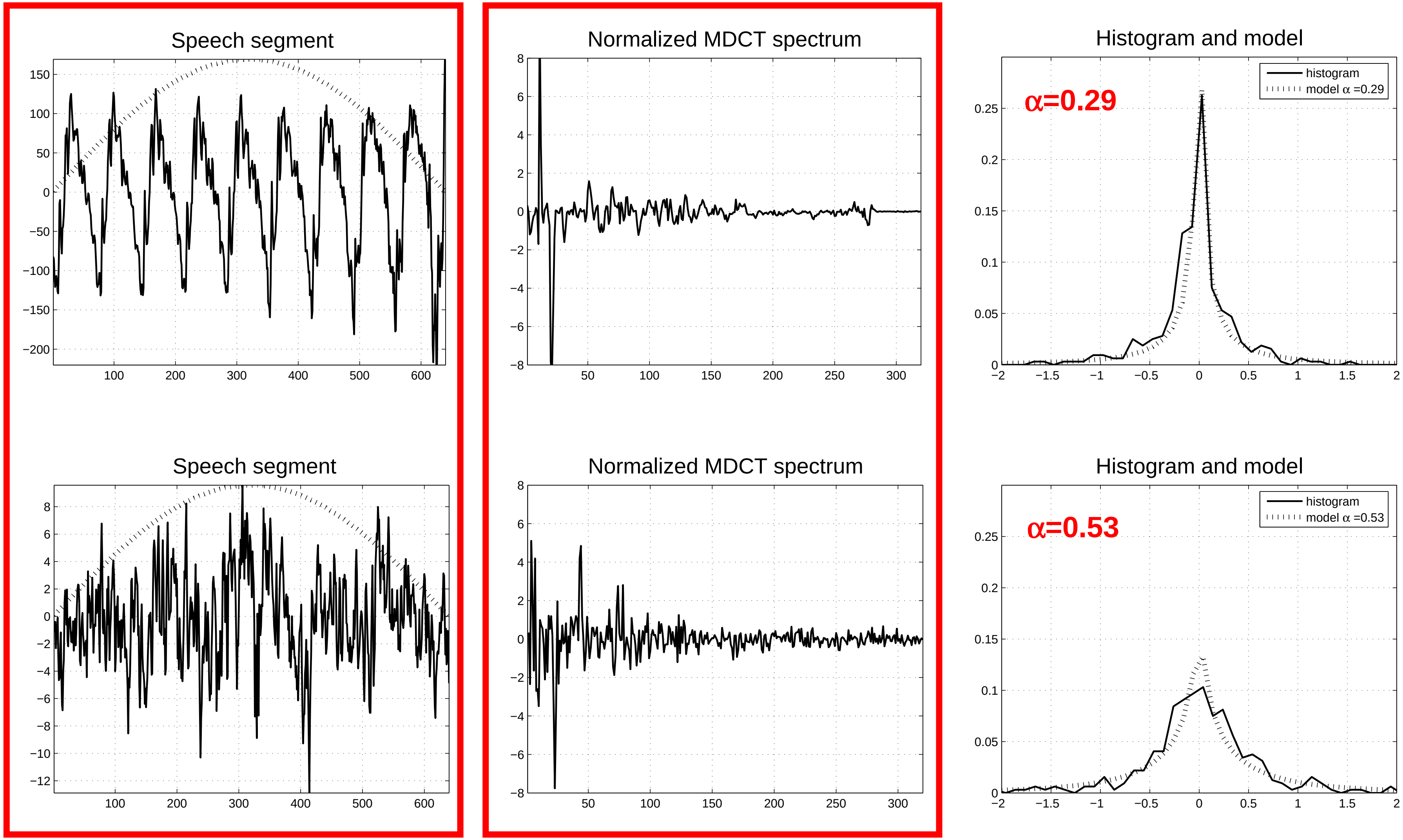
- The **method used to estimate α is proposed by Mallat** [Mallat 89]

$$F(\alpha) = \frac{E(|z|)}{\sqrt{E(z^2)}} = \frac{\Gamma(2/\alpha)}{\sqrt{\Gamma(1/\alpha)\Gamma(3/\alpha)}}$$

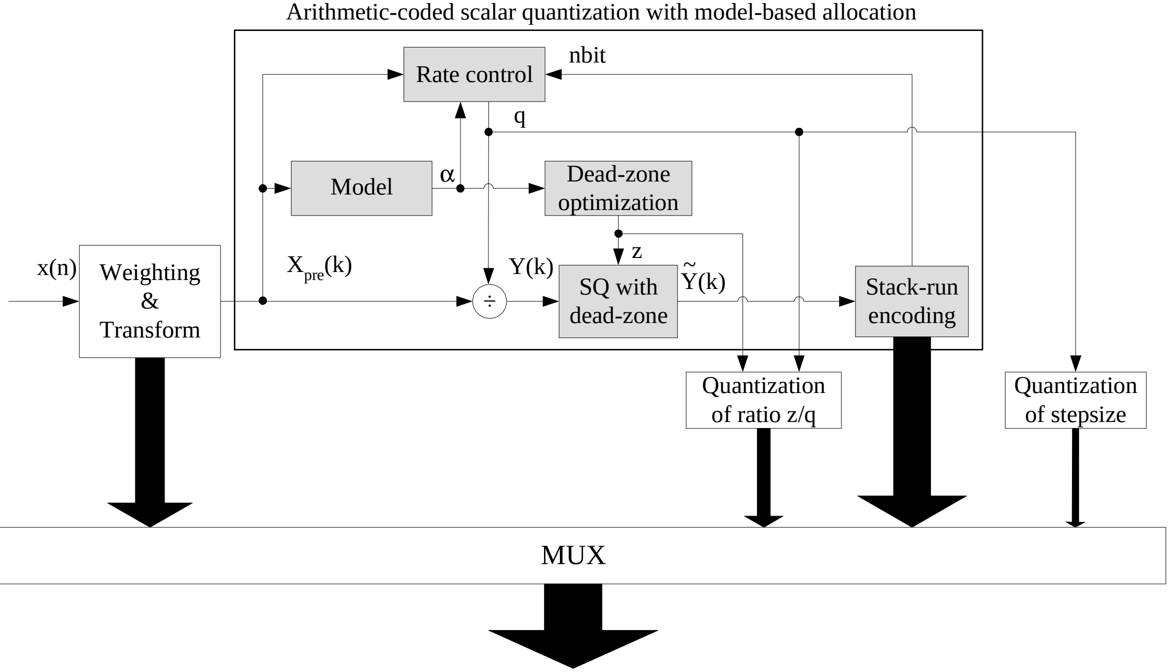
So $$\hat{\alpha} = F^{-1}\left(\frac{\hat{m}_1}{\sqrt{\hat{m}_2}}\right) = F^{-1}\left(\frac{\sum_{i=1}^{n} z_i^2}{\sqrt{\sum_{i=1}^{n} |z_i|}}\right)$$

FlexCode Estimation example: voiced & unvoiced speech

signal segment (time) spectrum (frequency)



FlexCode Stack-run coding with deadzone quantization



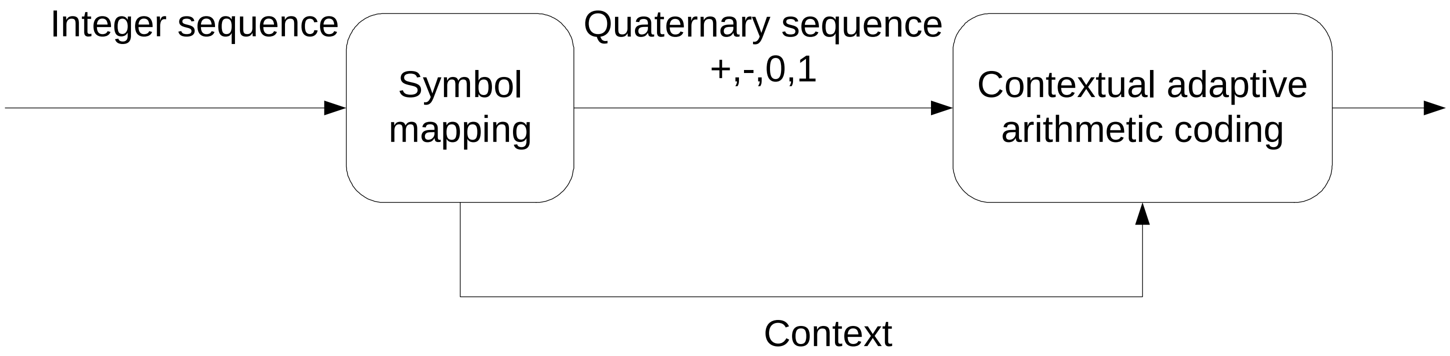
- Input/output signals sampled at 16 kHz
- Frame length of 20 ms with a lookahead of 25 ms (5 ms for LPC analysis and 20 ms for MDCT)
- Effective bandwith: 50-7000 Hz
- The perceptual weighting filter is defined as:

$$W(z) = \frac{A(z/\gamma)}{1-\beta z^{-1}} \quad \text{with} \quad \beta = 0.75 \quad \text{and} \quad \gamma = 0.92$$

- LPC coefficients quantized with a method based on GMM [Subramaniam 03] [Oger 06]
- MDCT implemented using the fast algorithm of [Duhamel 91]

FlexCode

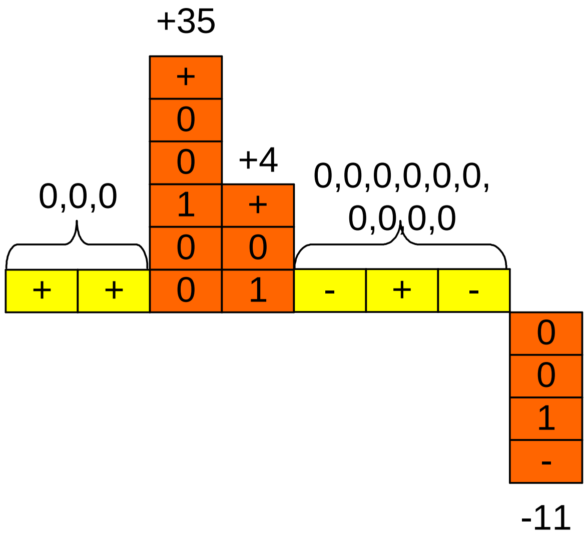
Stack-run coding [Tsai&Villasenor96]



- Stack-run coding is a **lossless coding method** representing integer sequences
  - Developed for wavelet image coding
- Adaptive arithmetic coding [Witten 87] using a **quaternary alphabet (0, 1, -, +)** and two contexts (one for "**runs**" and another for "**stacks**")
  - A run is a sequence of zeros
  - A stack is a non-zero signed integer

Stack-run coding

- Mapping rules for stack
  - The binary representation is unsigned
  - MSB is replaced by "+" if the coefficient is positive and "-" if it is negative.
  - The **absolute value is incremented by one**
  - The binary representation of "+4" is "+01" instead of "+00"
- The meanings of the symbol alphabet
  - "0" is used to signify a bit value of 0 in encoding of stack
  - "1" is used for bit value of 1 in stack, but it is not used for the MSB
  - "+" is used to represent the positive MSB of stack and for a bit value of **1 in representing run lengths**
  - "-" is used to represent the negative MSB of stack and for a bit value of **0 in representing run lengths**
- Mapping example for the sequence [0 0 0 +35 +4 0 0 0 0 0 0 0 0 0 0 -11]



Run	+ +			- + -		
Stack		00100+	10+			001-

High-rate distortion approximation

- Encoding of **N zero-mean independent variables x_i of variances σ_i^2**
- In case of **high-resolution** the mean square error D [Gersho & Gray 93] is given by

$$D \approx \sum_{i=1}^{N} h_i \sigma_i^2 2^{-2b_i}$$

where **h_i is a function of the pdf of the variable x_i** and b_i is the number of bits per sample used to code x_i

- For generalized Gaussian variables x_i the factor h_i is given by [Parisot 03] :

$$h_i = \frac{\Gamma\left(1/\alpha_i\right)^3}{3\alpha_i^2 \Gamma\left(3/\alpha_i\right)} e^{2/\alpha_i}$$

Model-based bit allocation : principles

- Encoding of **N zero-mean variables x_i of variances σ_i^2**
- The distortion D can be minimized by **Lagrangian techniques**:

$$J(b_i,\lambda) = D - \lambda\left(\sum_{i=1}^{N} b_i - B\right) \Rightarrow \lambda_{opt} = 2\ln(2)\sum_{i=1}^{N} h_i \sigma_i^2 2^{-2b_i}$$

where B is the target bit rate

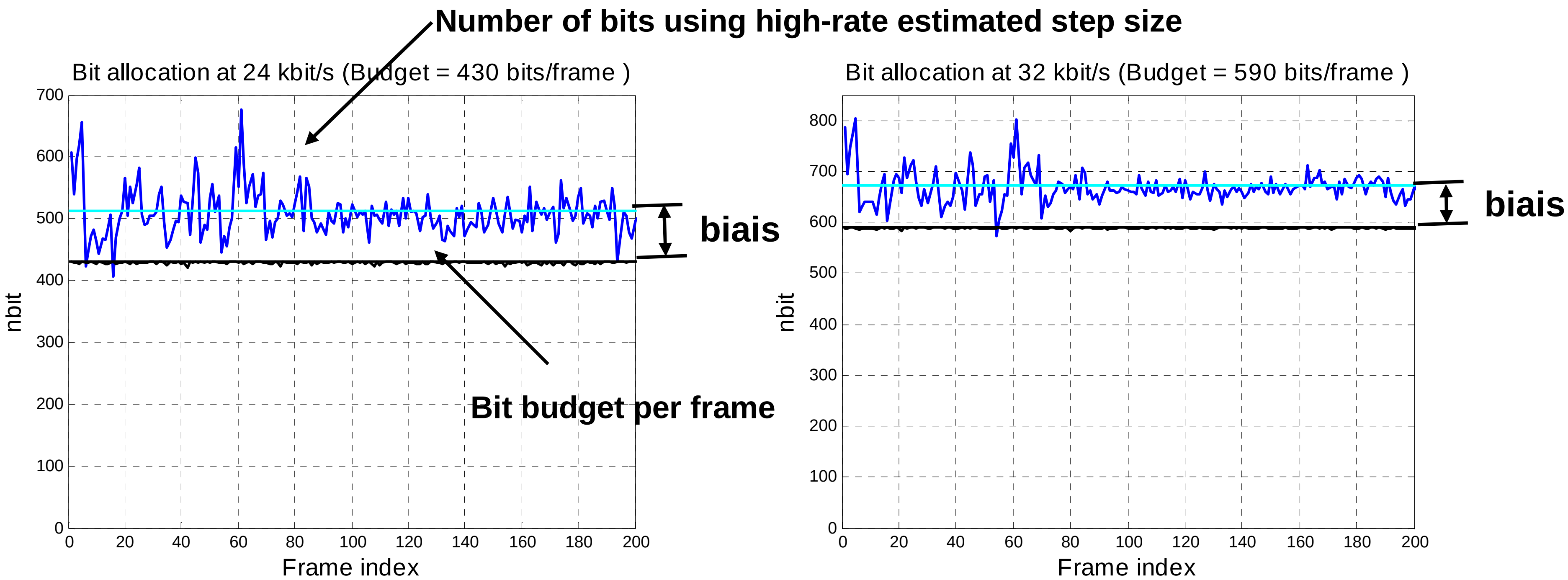
- Hence :

$$D_{opt} = \frac{\lambda_{opt}}{2\ln 2}$$

- In **case of high-resolution scalar uniform quantization** with step size q

$$D_{opt} = \frac{q_{opt}^2}{12} \Rightarrow \boxed{q_{opt} = \sqrt{\frac{6\lambda_{opt}}{\ln 2}}}$$

Model-based bit allocation: examples at 24 and 32 kbit/s



- Biais due to mismatch high-rate assumption and use of context-based lossless coding instead of zero-entropy coding

➡ A bisection search is used in order to be within the bit budget constraint

Deadzone optimization : principles [Parisot 03]

- Encoding of **N zero-mean generalized Gaussian variables x_i of variances σ_i^2**
- The distortion D is given by:

$$D(\alpha,z,q)=\frac{1}{\sigma^2}\int_{-z/2}^{z/2}x^2p_{\sigma,\alpha}(x)dx+\frac{2}{\sigma^2}\sum_{m=1}^{+\infty}\int_{-z/2+(m-1)q}^{z/2+mq}(x-\hat{x}_m)^2p_{\sigma,\alpha}(x)dx$$

- If the reconstruction level is set to centroid the distortion D is:

$$D(\alpha,z,q)=1-\sum_{m=1}^{+\infty}\frac{f_{1,m}\left(\alpha,\frac{z}{\sigma},\frac{q}{\sigma}\right)^2}{f_{0,m}\left(\alpha,\frac{z}{\sigma},\frac{q}{\sigma}\right)} \qquad \text{where} \quad f_{n,m}\left(\alpha,\frac{z}{\sigma},\frac{q}{\sigma}\right)=\int_{z/2\sigma+(m-1)q/\sigma}^{z/2\sigma+mq/\sigma}x^np_{1,\alpha}(x)dx$$

- If the reconstruction level is set to mid-value the distortion D is:

$$D(\alpha,z,q)=1+2\sum_{m=1}^{+\infty}\left(\frac{1}{2}\frac{z}{\sigma}+\left(m-\frac{1}{2}\right)\frac{q}{\sigma}\right)^2f_{0,m}\left(\alpha,\frac{z}{\sigma},\frac{q}{\sigma}\right)-4\sum_{m=1}^{+\infty}\left(\frac{1}{2}\frac{z}{\sigma}+\left(m-\frac{1}{2}\right)\frac{q}{\sigma}\right)f_{1,m}\left(\alpha,\frac{z}{\sigma},\frac{q}{\sigma}\right)$$

Deadzone optimization : principles

- The bit rate R is given by:

$$R = -P(0)\log_2 P(0) - 2\sum_{m=1}^{+\infty} P(m)\log_2 P(m)$$

- With $P(m) = \int_{z/2+(m-1)q}^{z/2+mq} x^n p_\alpha(x)\,dx = f_{0,m}\left(\alpha, \frac{z}{\sigma}, \frac{q}{\sigma}\right)$

- So the bit rate R is given by:

$$R = -f_{0,0}\left(\alpha, \frac{z}{\sigma}\right)\log_2 f_{0,0}\left(\alpha, \frac{z}{\sigma}\right) - 2\sum_{m=1}^{+\infty} f_{0,m}\left(\alpha, \frac{z}{\sigma}, \frac{q}{\sigma}\right)\log_2 f_{0,m}\left(\alpha, \frac{z}{\sigma}, \frac{q}{\sigma}\right)$$

Deadzone optimization : principles

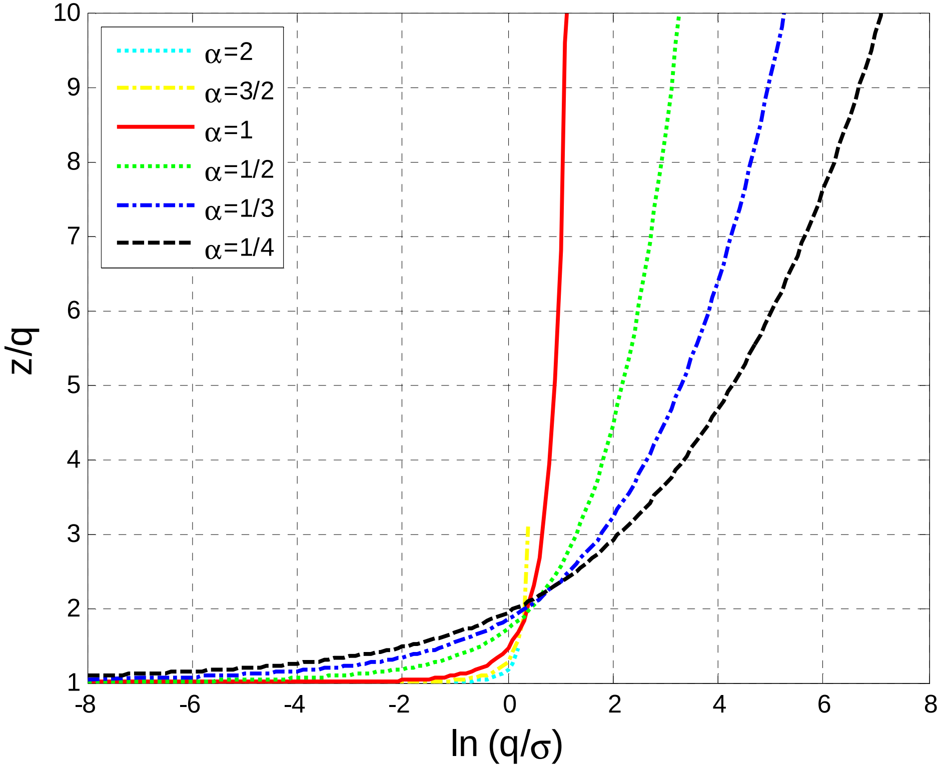
- Encoding of **N zero-mean generalized Gaussian variables x_i of variances σ_i^2**
- The distortion D can be minimized by **Lagrangian techniques**:

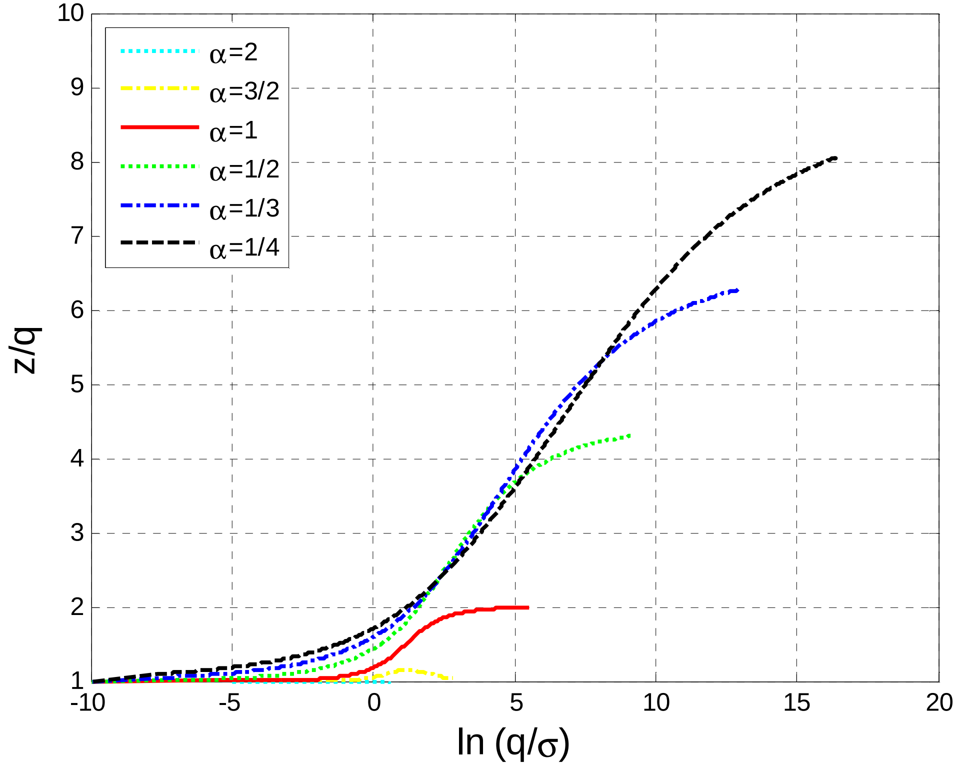
$$J(z_i, q_i, \lambda) = \sum_{i=1}^{N} \sigma_i^2 D\left(\alpha_i, \frac{z_i}{\sigma_i}, \frac{q_i}{\sigma_i}\right) + \lambda\left(\sum_{i=1}^{N} a_i R\left(\alpha_i, \frac{z_i}{\sigma_i}, \frac{q_i}{\sigma_i}\right) - R_{target}\right)$$

$$\Rightarrow \begin{cases} \dfrac{\frac{\partial D}{\partial \tilde{z}}(\alpha_i, \tilde{z}_i, \tilde{q}_i)}{\frac{\partial R}{\partial \tilde{z}}(\alpha_i, \tilde{z}_i, \tilde{q}_i)} = \dfrac{\frac{\partial D}{\partial \tilde{q}}(\alpha_i, \tilde{z}_i, \tilde{q}_i)}{\frac{\partial R}{\partial \tilde{q}}(\alpha_i, \tilde{z}_i, \tilde{q}_i)} \Rightarrow \text{Optimization of the deadzone z} \\[2ex] \dfrac{\frac{\partial D}{\partial \tilde{q}}(\alpha_i, \tilde{z}_i, \tilde{q}_i)}{\frac{\partial R}{\partial \tilde{q}}(\alpha_i, \tilde{z}_i, \tilde{q}_i)} = -\dfrac{\lambda a_i}{\sigma_i^2} \qquad \sum_{i=1}^{N} a_i R(\alpha_i, \tilde{z}_i, \tilde{q}_i) = R_{target} \end{cases}$$

where α_i, z_i, and q_i are respectively the shape parameter, the deadzone and the stepsize

Model-based deadzone optimization

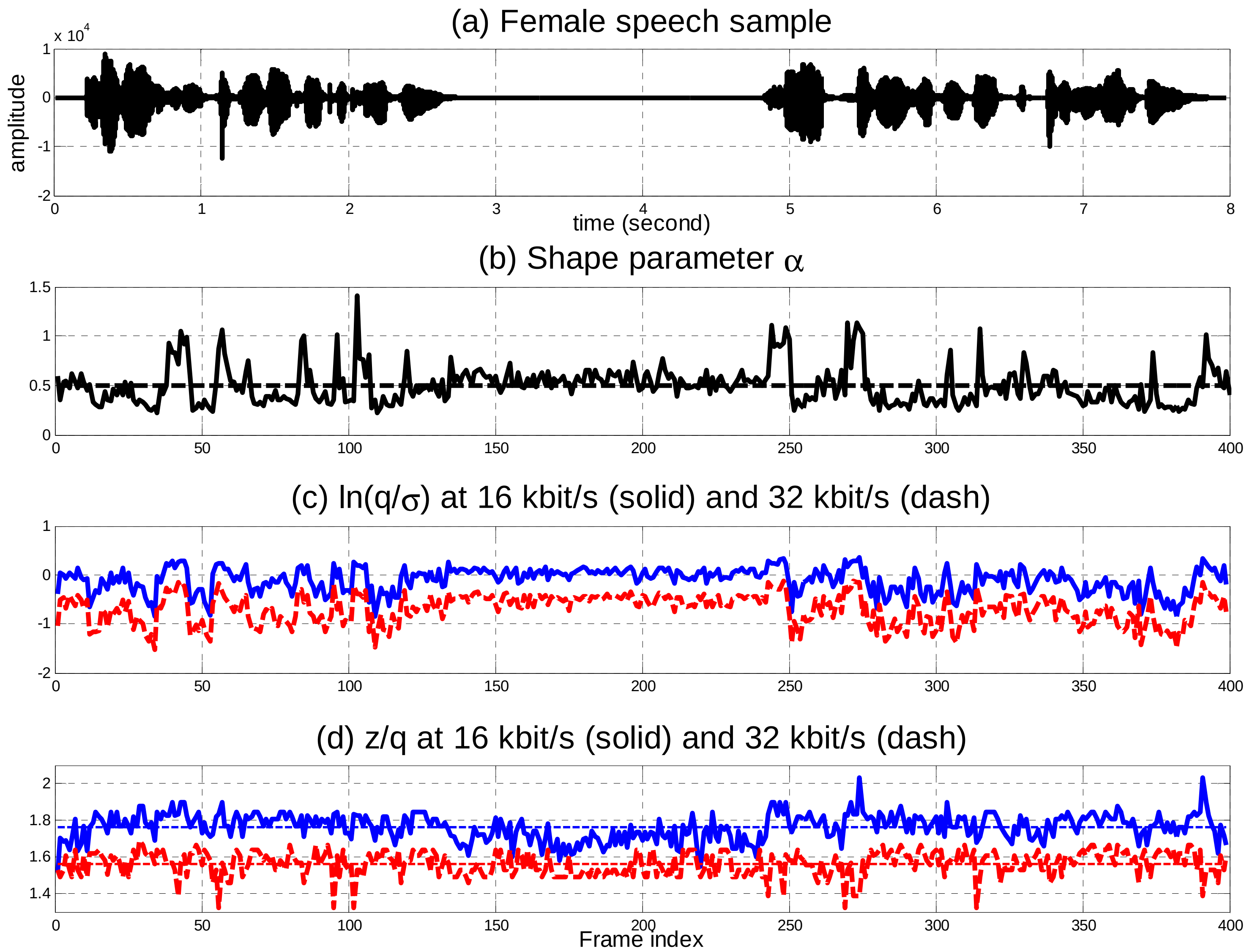




Scalar quantizer with reconstruction levels set to mid-value

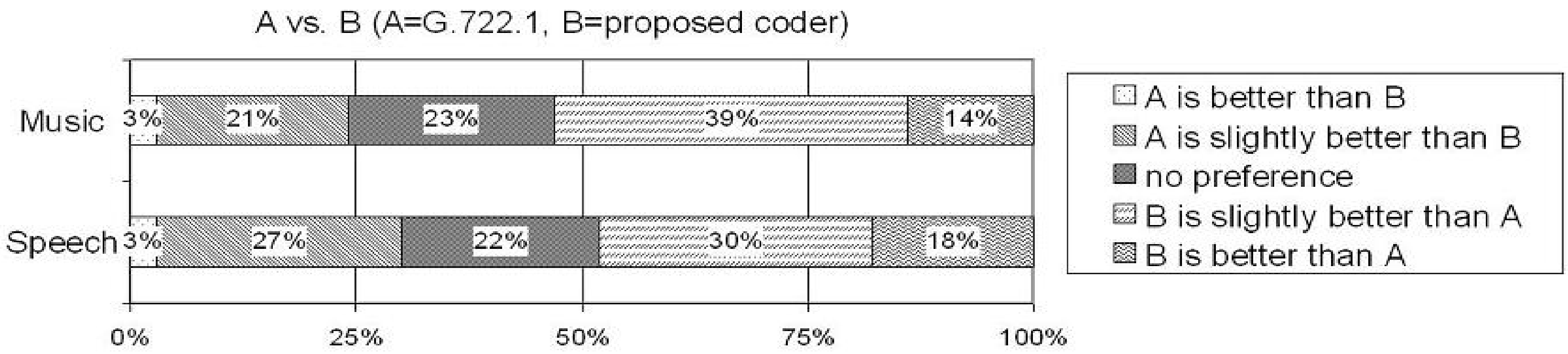
Scalar quantizer with reconstruction levels set to optimal centroid (Lloyd-Max)

Model-based deadzone optimization



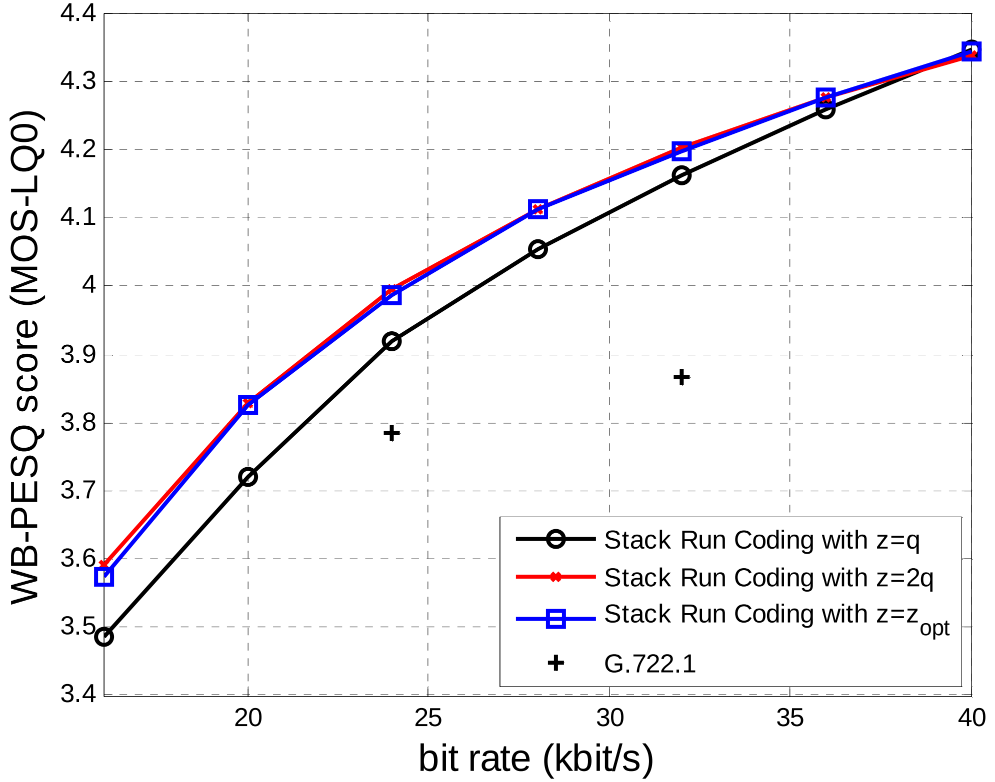
FlexCode Subjective quality results (no deadzone)

- 16 clean music samples (4 types × 4 sentence-pairs) of 8 seconds
- 24 clean speech samples in French language (6 male and female speakers × 4 sentence-pairs) of 8 seconds
- Two AB test at 24 kbit/s : one for speech, another for music



  - 8 expert listeners
  - The stack-run coding (z=q) was preferred:
    - In 53% cases for music
    - In 48% cases for speech
- Informal listening tests at 32 kbit/s
- **Stack-run coding (z=q) is better than ITU-T G.722.1 at 24 kbit/s and equivalent at 32 kbit/s**

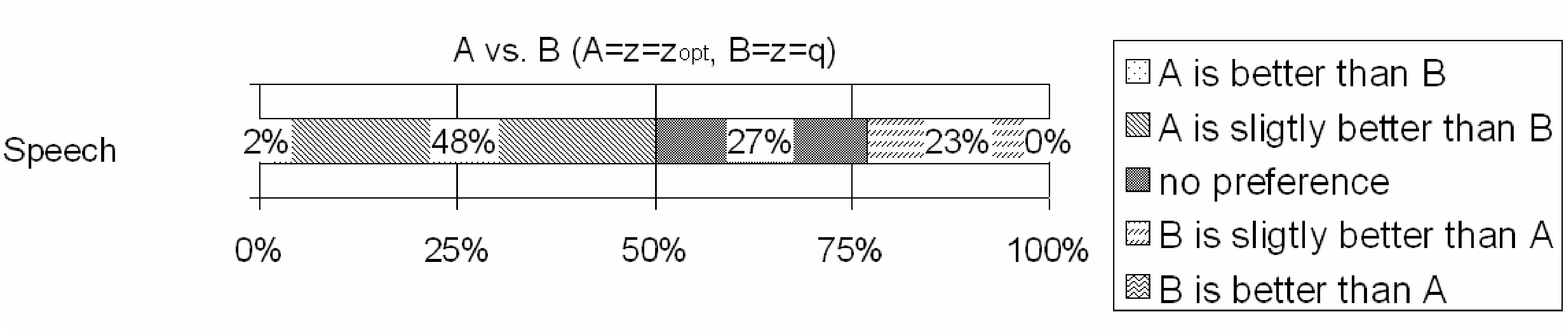
Objective results (with/without deadzone)



- 24 clean speech samples in French language (6 male and female speakers × 4 sentence-pairs) of 8 s
- Proposed coder: predictive MDCT coder with stack-run coding
- Results are presented with noise injection (injection similar to 3GPP AMR-WB+)
- These **objective results suggest that the inclusion of a dead-zone improves the performance**

FlexCode Subjective quality results (with deadzone)

- 20 clean speech samples in French language (5 male and female speakers $\times$ 4 sentence-pairs) of 8 seconds



- One AB test at 24 kbit/s for speech:
  - 9 expert listeners
  - Stack-run coding with $z=z_{opt}$ was preferred in 50% cases for speech
- Informal listening tests at 32 kbit/s
- **Stack-run coding with $z=z_{opt}$ is better than stack-run coding with z=q at low bitrate and is equivalent at high bitrate.**

Delay & complexity

- Algorithmic delay:
  - 45 ms (20 ms for the frame, 20ms for the MDCT and 5 ms for the lookahead) for the stack-run coder
  - 40 ms for the ITU-T G.722.1
- The computational complexity of ITU-T G.722.1 is very low
- The computational complexity of the stack-run coder is higher due to the use of bisection search for bit rate matching
  - Stack-run coding is performed several times per frame
- Storage requirements for the stack-run coder are low
  - Parameters for GMM based LPC quantization
  - MDCT tables (can be computed on lines)

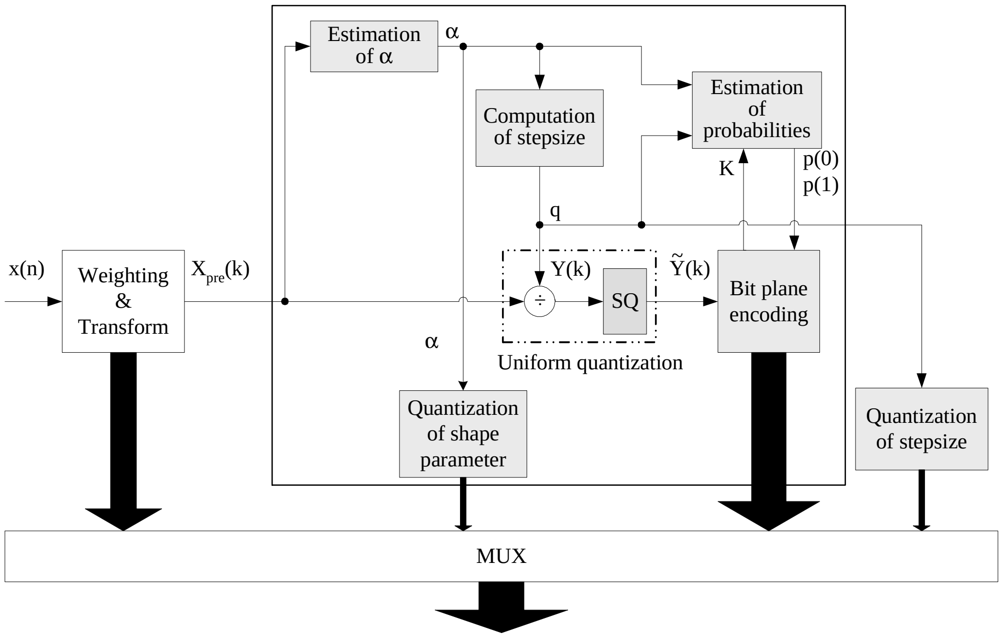
Audio samples

Comparison between stack-run coding and ITU-T G.722.1

	Stack-run coding with z=q	Stack-run coding with $z=z_{opt}$	ITU-T G.722.1
Music at 24 kbit/s			
Music at 32 kbit/s			
Speech at 24 kbit/s			
Speech at 32 kbit/s			

FlexCode Latest developments: Model-based bit plane coding

Bit-plane-coded scalar quantization with model-based allocation and probabilities



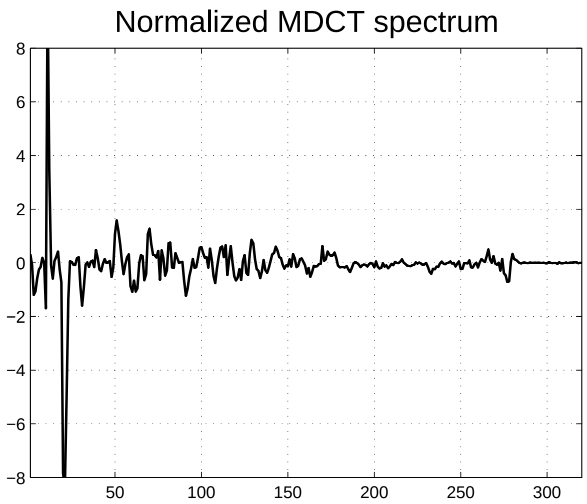
- Principle: replace stack-run coding by bit plane coding.
- Implicit rate control

➡ Computational complexity is much more lower than stack-run coding

FlexCode

Model-based estimation of symbol probabilities

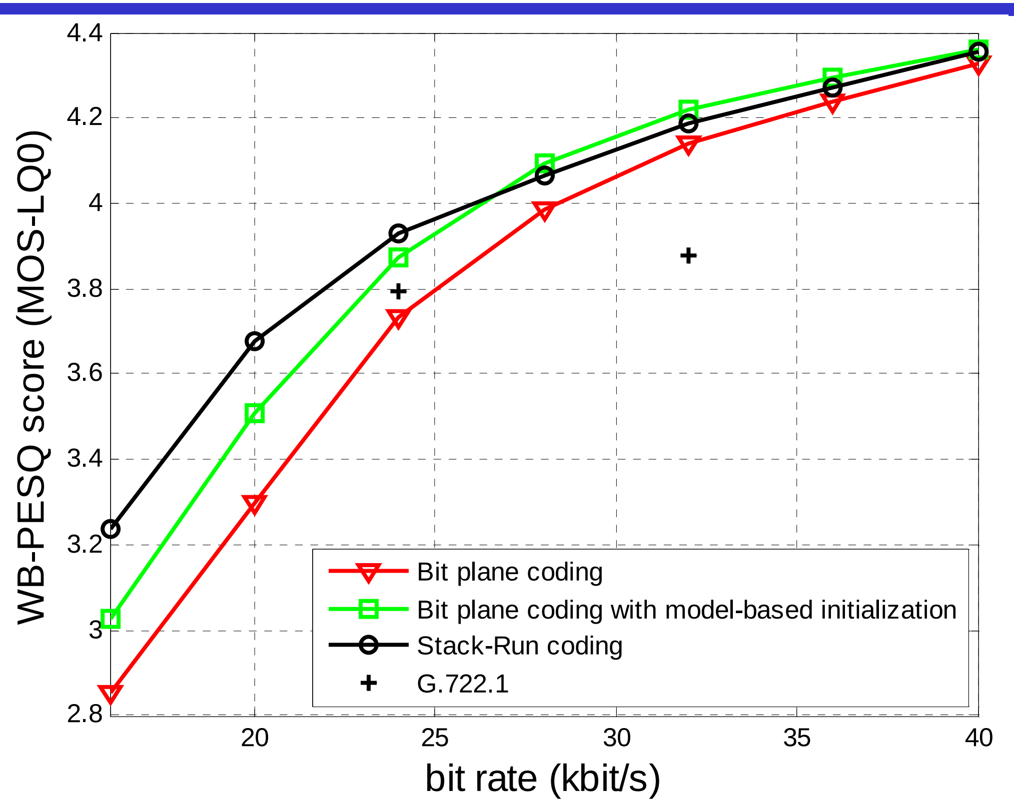
orange™



Y=		-2	+5	-3	……...	+1	0	+1	
		1	0	1	……...	0	0	0	→ Bit plane of signs
	2^2	0	1	0	0……...0	0	0	0	→ P_2=MSB
K=3	2^1	1	0	1	0……..0	0	0	0	
	2^0	0	1	1	0……..1	1	0	1	→ P_0=LSB

- The normalized MDCT spectrum $X_{pre}(k)$ is scalar quantized and we get an integer sequence $\tilde{Y}(k)$.
- This integer sequence is decomposed in binary format.
- The symbol probabilities in bit planes are estimated on the model of the pdf of $X_{pre}(k)$

Objective quality results



- 24 clean speech samples in French language (6 male and female speakers $\times$ 4 sentence-pairs) of 8 s
- Proposed coder: predictive MDCT coder with bit-plane coding
- Results are presented without noise injection
- These **objective results suggest that the speech quality of the proposed coder with model-based initialization of symbol probabilities is equivalent to reference coders at high bitrate**

FlexCode

Conclusion & perspectives

orange™

- We proposed a predictive MDCT coder with generalized Gaussian modeling for wideband speech and audio signals
- Generalized Gaussian modeling is used to:
  - Estimate the optimal step size
  - Optimize the deadzone
  - Estimate symbol probabitilies in bit planes
- Next step: Include sensitivity matrix into model-based coder
  - Linear-predictive filter $\rightarrow$ signal-adaptive transform

The End

Thank you!