The FlexCode Speech and Audio Coding Approach[1]

Stefan Bruhn[†], *Volodya Grancharov*[†], *W. Bastiaan Kleijn*[*], *Janusz Klejsa*[*], *Minyue Li*[*], *Jan Plasberg*[*], *Harald Pobloth*[†], *Stephane Ragot*[◇], *Adriana Vasilache*[‡]

[*]ACCESS Linnaeus Center, Electrical Engineering, KTH - Royal Institute of Technology, 10044 Stockholm, Sweden
[‡]Nokia Research Center, Visiokatu 1, 33720 Tampere, Finland
[◇]France Télécom, R&D/TECH/SSTP, Av. Pierre Marzin, 22307 Lannion Cedex, France
[†]Multimedia Technologies, Ericsson Research, Sweden
Web: www.flexcode.eu

Abstract

The increasing heterogeneity of communication networks and the variability in user requirements form a challenge for source coding algorithms. To address this challenge, the aim of the FlexCode source coding approach is to create a speech and audio coder that can adapt instantaneously to network and user requirements. The approach is based on a statistical description of the source, a model of perception, and analytic relations between bit rate, distortion, packet-loss rate, and the reconstruction-point distribution density of the quantizers. The method facilitates analytic optimization of the source coder for given network and input signal conditions and user requirements at any time. Preliminary results confirm that the flexibility in rate and robustness can be obtained without a loss in performance compared to state-of-the-art coders optimized for particular conditions.

1 Introduction

The transmission of audio-visual signals plays a central role in modern society. Scarcity in bandwidth and the loss and damage of transmitted information have led to the introduction of source and channel coding algorithms. To maximize performance, specific algorithms have been defined for various applications. This has led to a proliferation of coding standards. While each algorithm generally performs well for the environment that it was designed for, the algorithms usually perform poorly under other circumstances. For example, many speech-coding algorithms use large codebooks that were obtained with off-line learning. Such codebooks do not perform well for music. Existing algorithms are not easily adapted to new applications with different constraints on quality, bit rate, delay, complexity, and packet robustness. If an environment varies as a function of time, the lack of flexibility is a severe handicap leading to low performance.

The main objective of FlexCode is to develop a coding paradigm that is inherently flexible. The resulting coding algorithm can adapt instantaneously to application constraints set by the users and the network environment. That is, the coder immediately optimizes its configuration for any particular combination of overall rate, packet loss rate, and bit error rate. The coder is layered to allow dropping of bits at less capable terminals or in network bottlenecks and has modest computational and memory requirements. To facilitate optimal performance, the coder makes use of sophisticated perceptual criteria.

A secondary objective of FlexCode is to validate the paradigm using a practical implementation that is applied to real-world scenarios. Thus, a real-time implementation of the coding algorithm is being developed for speech and audio signals. The implementation is to be compared with the performance of state-of-the-art systems at transmission rates of 10 to 30 kb/s.

In this paper, we first outline the architecture of the FlexCode source coder. We then discuss some specific advances that were developed to realize the paradigm, including methods that make the the coder robust against packet loss, and finally provide some initial experimental results.

2 FlexCode Source-Coding Approach

The FlexCode approach facilitates flexible and efficient source coding at relatively low rates. Fundamental to the approach are the usage of the source and perception models. Conventional speech coders generally rely both on modeling and on the training of quantizers using a data base and are aimed at a particular coding rate. Important knowledge about the signal is stored in codebooks. Audio coders generally rely less on signal models and perform less well at low rates. In contrast, in the FlexCode approach source models are specified based on data and (asymptotically) optimal quantizers are computed based on these source models and on models of perception. The inputs to the quantizer computation are the rate (or quality) required and the parameters of the probabilistic source model and the model of perception.

2.1 The Signal Model

In FlexCode the audio signal is specified by a sequence of random vectors S_n with realization s_n, where n is a time index. The vectors s_n are sequential segments of the signal, signal vectors after removal of a suitable zero-input response of a filter, or successive vectors of coefficients of a lapped transform. In the simplest case, the vectors s_n are modeled as independent. The distribution for the vectors is modeled by $f_{S|\Theta}(s|\theta)$ where Θ describes the parameters of the signal model and where we omit the time index for convenience. A convenient generic model for $f_{S|\Theta}(s|\Theta)$ is the ubiquitous mixture model:

$$f_{S|\theta}(s|\Theta) = \sum_{i\in\mathscr{I}} p_I(i) g_{S|\Theta}(s|\theta_i), \tag{1}$$

where $g_{S|\Theta}(s|\theta_i)$ is the distribution of component i, and $p_I(i)$ is the component probability. In FlexCode we use the Gaussian mixture model with zero-mean Gaussians. This implies that the components can be written as

$$g_{S|\Theta}(s|\theta) = \frac{1}{\sqrt{(2\pi)^k \det(R(\theta))}} \exp\left(-\frac{1}{2} s^T R(\theta)^{-1} s\right), \tag{2}$$

where $R(\theta)$ is the model covariance matrix for S. If we use the commonly used autoregressive (AR) signal model, then $R(\theta)$ is determined by the AR model parameters θ. Assuming that a good signal model is a model that minimizes the required rate to encode the signal, we showed in (cf. section 3.1) that for optimal AR modeling of the speech at 8 kHz sampling rate about one million components are required.

For speech it is natural to associate the components of the mixture model (1) with a particular configuration of the vocal tract. Only one component is active in each vector s. For audio signals this is less natural as composite sounds will occur. If the kernel densities $g_{S|\Theta}(s|\theta_i)$ are approximated as not overlapping, then each vector s corresponds to a particular mixture component. While the mixture model description of the signal is not common in speech and audio coding, the commonly used linear-prediction algorithm in speech and audio coding is consistent with a mixture model with nonoverlapping components.

[1] This work was supported in part by the European Union under Grant FP6-2002-IST-C 020023-2 FlexCode

2.2 Quantization and Lossless Compression

The FlexCode approach is to compute the quantizer for the vector S from the signal model. In the early version of the FlexCode coder the signal is quantized using scalar quantizers. To maximize the efficiency of scalar quantizers, dependencies between the vector components must be minimized. Let us assume an unweighted squared error distortion measure (which will be motivated in section 2.3) and a multi-variate Gaussian signal model. Dependencies between vector components can then be eliminated without affecting the distortion measure by performing the unitary transform on the signal vector S that diagonalizes its covariance matrix. The diagonalizing transform is the Karhunen-Loève transform. In the FlexCode approach, we diagonalize the covariance matrix of the *signal model* [1], which is known to both encoder and decoder.

The base-line FlexCode system facilitates both entropy-constrained and resolution-constrained quantization of the signal. The optimal entropy-constrained quantizer is uniform and the signal probability density forms an input to lossless coding. An arithmetic coder is employed for the entropy-constrained case. The constrained-resolution quantizer is implemented by means of companding.

As the covariance matrix is specified by the signal model, the Karhunen-Loève transform is known to both encoder and decoder. Assuming the signal model is accurate, the loss in coding efficiency of a scalar quantizer relative to a vector quantizer [2] (asymptotically with high rate) is reduced to the space-filling advantage (less than 1.5 dB) and the shape advantage (constrained-resolution only) of vector quantization.

The FlexCode approach has a theoretical performance advantage over the ubiquitous code-excited linear prediction (CELP) algorithm [3]. CELP is based on a signal codebook that is stored in a so-called "excitation" domain and then shaped with the AR filter corresponding to the segment. This means that the geometry of the CELP quantizer cells is generally not optimal. This advantage of the FlexCode approach was discussed in [4].

2.3 Distortion Measure

FlexCode supports the use of sophisticated distortion measures. The current approach facilitates any distortion measure representing human perception $d(s,y)$ that can be approximated as locally quadratic. Then we have that

$$d(s,y) \approx (s-y)^T M(s)(s-y), \tag{3}$$

where $M(s)$ is the so-called *sensitivity matrix* [5].

The FlexCode approach assumes that a distortion measure $d(s,y)$ can be represented by an invertible mapping (compressor) $F(\cdot)$ to a "perceptual domain" where the distortion measure is the squared error. The quantizer operates in the perceptual domain. The reconstructed signal Y then consistes of the inverse mapping of the quantized signal:

$$S \rightarrow F(\cdot) \rightarrow Q(\cdot) \rightarrow F^{-1}(\cdot) \rightarrow Y. \tag{4}$$

It has been shown that, under certain conditions and for high dimensions and low distortions, this quantization procedure can approach the rate-distortion limit arbitrarily closely for the case of constrained-entropy quantization [6]. It was also shown that the optimal compressor satisfies $F'(s)^T F'(s) = cM(x)$ where c is a scalar and where $F'(s)$ is the Jacobian of $F(\cdot)$.

Using the sensitivity matrix approach, two distortion measures have been implemented in FlexCode. The distortion measure described by [7] is simpler to implement but only considers frequency domain effects. The distortion measure described by [8] requires a larger computational effort but incorporates both time and frequency domain effects.

The assumption that human perception can be approximated as locally quadratic leads to problems at lower rates. The human auditory system is sensitive to the appearance and disappearance of spectral holes in successive signal segments. Such holes are a natural result of so-called reverse waterfilling, which is a consequence of usage of the squared-error distortion measure at low rates. FlexCode is working on a distortion measure that is a combination of a locally quadratic and a power-spectral error based criterion to resolve this problem in an elegant manner.

2.4 Audio Coding Architecture

Based on the principles described in sections 2.1, 2.2, and 2.3, the source coding architecture shown in Fig. 1 was developed. The coder facilitates implementation based on the Karhunen Loève transform and on the modulated lapped transform (MLT). In this description, we focus on the first implementation. The signal is first subjected to segmentation (the segments may overlap). The AR model is then estimated from the signal segment and quantized. For a particular input signal environment, a FlexCode result (cf. section 3.1) is that the optimal coding rate of the model is independent of the overall rate, and so it is possible to use a codebook for this quantizer. However, if the coder is to be used for different input signal environments, then codebooks are not desirable. Thus, in the current FlexCode implementation, scalar and lattice (cf. section 3.2) quantizers are used for quantization of the AR model.

Based on the sequence of estimated signal models, the Jacobian $F'(\cdot)$ of the perceptual model is computed and this Jacobian is used to weight the incoming signal segment.

For the spectral distortion measure [7], the Jacobian $F'(s)$ is Toeplitz and the weighting is implemented by a filtering operation, resulting in a pre- and post-architecture similar to that of [9] but with a different method for computing the weighting filter. The spectral weighting method is followed by a subtraction of the zero-input response of the signal model, which renders a sequence of largely independent signal vectors S_n. After the computation of a composite model that accounts for the weighting and for the subtraction of the zero-input response, the Karhunen-Loève transform is performed on these vectors, rendering a set of coefficients that form the input to scalar quantizers that are optimal for the signal model (under high-rate assumptions). The quantizers can be constrained-entropy or constrained-resolution quantizers. The quantizer indices are coded using arithmetic coding in the case of constrained-entropy quantization.

The audio-coding architecture includes a pitch model (single-tap AR model). For this model the delay (tap location) is adjusted closed-loop based on minimizing the residual remaining after subtracting the zero-input response.

At the decoder the quantized values are decoded and the invertible signal processing steps inverted.

The delay of system of Fig. 1 is mostly determined by the segment length and by the requirement. The segment length is determined by the duration over which the input signal is modeled well by a stationary AR model. Typical segment lengths are 5 to 10 ms. A related coder with a delay of only a few samples, facilitated by backward adaptation of the signal model, was described in [10].

3 Specific FlexCode Technologies

To realize the flexibility envisioned in the FlexCode paradigm and to maintain state-of-the-art performance for any particular coder configuration, new technologies and improvements to existing technologies were needed. In this section we highlight a number of these technologies.

3.1 Rate Distribution

In the coding system envisioned by FlexCode, the rate is an adjustable parameter. This means that one must find an optimal balance between the rate for the quantization of the signal model parameters and the rate for the quantizers that operate directly on the signal, with knowledge of the quantized signal model for each condition for which the coder is optimized. We have shown

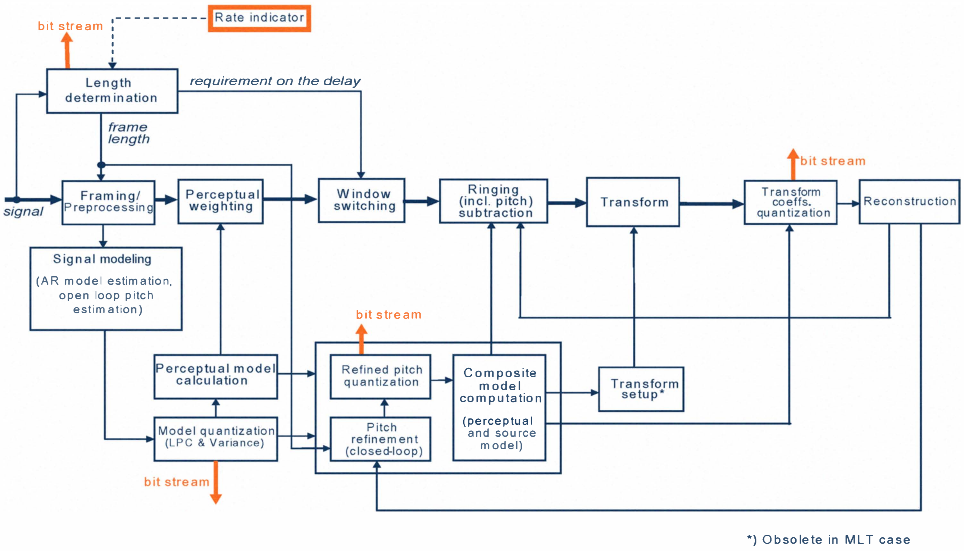
Rate indicator
bit stream
Length determination
requirement on the delay
frame length
bit stream
signal
Framing/ Preprocessing
Perceptual weighting
Window switching
Ringing (incl. pitch) subtraction
Transform
Transform coeffs. quantization
Reconstruction
Signal modeling (AR model estimation, open loop pitch estimation)
bit stream
Perceptual model calculation
Refined pitch quantization
Composite model computation (perceptual and source model)
Transform setup*
Model quantization (LPC & Variance)
Pitch refinement (closed-loop)
bit stream
*) Obsolete in MLT case

Figure 1: The source-coding architecture of FlexCode.

[11, 12] that the optimal rate distribution follows a simple asymptotic rule that is easily implemented in a practical coding system.

The optimal distribution of rate between the model and the signal can be obtained using a method that is closely related to the minimum-description length principle [13]. The total rate used for encoding the signal can be separated into three rate components: *i)* the rate for the signal using the ideal (unquantized) model *ii)* the rate for the quantized model, and *iii)* the penalty term: the increase in the rate for the signal because the model is not the ideal model. The optimal rate distribution is governed by the rate components *ii* and *iii*. We have solved the problem for the squared-error distortion measure and both the constrained-entropy and the constrained-resolution cases. For the constrained-entropy case, the signal rate for a given model is

$$R(s|\theta) = -\log\left(f_{S|\Theta}(s|\theta)(\frac{D}{C})^{\frac{k}{2}}\right), \qquad (5)$$

where k is the dimensionality of s, D is the distortion, and C is a constant, the coefficient of quantization.

It is seen from (5) that the difference between the rate for the signal given the ideal model and the rate for the signal given the quantized model depends only on the ratio of their likelihoods and not on the distortion (which cancels). This immmediately implies that the rate for the model is independent of the overall distortion D and, thus, the overall source-coder rate.

We have worked out the results for the case of the AR model. Based on (2) and (5) the penalty term *iii* can be related to log spectral distortion. The rate required for the model can be related to the differential entropy of the model parameters in the log spectral domain. We can then create an expression for the optimal rate for the model. For 8 kHz sampled speech and using 20 ms segments, this rate is 20 bits and 1.3 dB spectral distortion. This result justifies older empirical estimates of what is a "transparent" log spectral distortion level for coding (about 1 dB, [14]), and the corresponding rates a bit allocation of 18-25 bits/second (e.g., [14, 15]).

3.2 Lattice Quantization

For the squared-error criterion and constrained-entropy quantization, and under the high-rate assumption, optimal rate-distortion performance is reached by a uniform quantizer at any dimensionality. This suggests the use of lattice quantization, which can provide optimal quantizer cell shapes (e.g., [2]) and leads to low quantizer computational complexity. The computational advantage has meant that lattice quantization is also used for constrained-resolution quantization, using companding to obtain a uniform distribution. (In general, in this context companding is not rate-distortion optimal for a dimensionality higher than one.)

The use of lattice quantizers depends on the existence of an indexing algorithm for the lattice points. This generally involves the need for lattice truncation. Within the context of FlexCode we have developed new indexing algorithms that use generalized rectangular lattice truncations. The methods lead to effective coding for mixture models and companded lattice quantization.

At moderate bit-rates (below 2 bits per sample), for non-symmetric sources and singular data, lattice border effects are significant and the rate distortion curves depend on the orientation of the lattice truncation. Lattice rotation such that the denser direction corresponds to the denser direction in the data results in improved performance. Thus, significant improvement of performance has been observed for correlated data [16]. The method is expected to lead to a significant improvement in the coding efficiency of transform coefficients of the FlexCode coder.

3.3 Bit-Stream Scalable Coding

Bit-stream scalable coding, also known as embedded coding, encompasses all coding methods that facilitate the progressive dropping of bits from the coded bit stream without significant impairment of the coding efficiency at the various resulting (decoding) rates. It is well-known that for certain signals and distortion measures, including Gaussian signals with the squared-error distortion measure, rate-distortion optimal *successive refinement* (iterative improvement of the coded approximation) is possible [17] and this motivates us to study embedded coding in the context of the FlexCode source coder [18], [19].

The bit-stream scalable coding integrates in the constrained-entropy scalar quantization of the transform coefficients of the FlexCode coder. The pre-processed transform coefficients are modeled as independent identically distributed (iid) variables, and are uniformly quantized. The coefficient distribution model is an adaptive generalized Gaussian distribution that can describe a range of distributions including Gaussian and Laplacian. The sign and the absolute value of the coefficients are separated. The sign bit is transmitted only if the amplitude is nonzero. The bits of the quantization indices are arranged in bit planes and the planes are coded from most-significant bit to least-significant bit using arithmetic coding. To obtain efficient encoding, the arithmetic coder considers in the encoding of a bit of coefficient i in a particular bit plane j, the values of the already decoded bits for that coefficient in bit planes $k > j$, as well as the generalized Gaussian distribution that applies to the coefficient. Our experimental results [19] indicate that the new model-based bit-plane coding method matches the performance of conventional coding schemes that do not have bit-stream scalability.

3.4 Scalable Multiple Description Coding

The FlexCode source coding paradigm aims to provide efficient and robust transmission over a packet network with an arbitrary and possibly time-varying packet loss rate. Multiple-description coding (MDC) is a general method used to combat the effects of packet loss by introducing redundancy and exploiting diversity offered by the network [20]. Feedback information about the packet-loss rate is generally present and can, at least in principle, be utilized to determine the optimal redundancy for adaptive coding. However, state-of-the-art MDC schemes are generally not suited to applications where the channel varies, since severe design complexity or significant storage requirements prevent adaptive coding.

We have developed quantization-based MDC techniques that have low-design complexity and are scalable in both redundancy and rate. One method is an adaptive MDC for a two-description scalar setup that can achieve optimality for both the constrained entropy and constrained resolution case, for any packet-loss rate [21]. Systems that allow an arbitrary number of descriptions and that allow optimization of the number of descriptions are in an advanced stage of development.

A quantization-based MDC scheme consists of one central and number of side quantizers with an invertible mapping from the central codebook to the side codebooks. The main design complexity is associated with optimization of this mapping. The FlexCode two-description scalar MDC method is based on the use of predefined, parameterized mapping algorithms (so-called index-assignment schemes). The parametrization forms the basis for the analytic optimization of the MDC configuration. This scalable, two-description quantization-based MDC scheme will be used for the transform coefficients of the FlexCode coder.

4 Performance

While the FlexCode source coder is evolving rapidly, some preliminary performance results can be provided. The results only relate to source coding and not to robustness against packet loss. The coder was subjected to formal testing using the MUSHRA procedure on four music items, six speech items, three mixed speech and music items, and four noise speech items, at rates of 14, 24, and 32 kb/s. The items ranged from 4 to 25 seconds in duration. As a reference the AMR-WB [22] and ITU G.729.1 standards were used. The test was performed in professional laboratories at Ericsson (6 listeners), Nokia (12 listeners) and Orange/FT (10 listeners). The test results were consistent across the laboratories and across the items. At 24 kb/s the FlexCode coder performed similar to the reference coders. At 32 kb/s the FlexCode coder performed better than the reference coders. At 14 kb/s the performance is worse than that of the AMR-WB coder. The latter result is related to a problem with the implementation of the pitch predictor. Overall the results indicate that scalable coding does not imply a loss of performance.

5 Conclusion

From the current results of the FlexCode project we conclude that practical scalable audio coding without loss of performance compared to state-of-the-art audio coders designed for a particular coding rate is possible. Embedding based on newly developed bit plane coding techniques facilitates the stripping of bits in the coded bit stream, which is particularly useful for multicast scenarios. New methods for scalable multiple-description coding allow coders to adapt instantaneously to any packet loss rate.

Acknowledgement

The authors like to thank Catherine Colomes of Orange/France Telecom for designing and managing the quality tests.

References

[1] M. Y. Kim and W. B. Kleijn, "KLT-based adaptive classified vector quantization of the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 277–289, 2004.

[2] T. Lookabough and R. Gray, "High-resolution theory and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 5, pp. 1020–1033, 1989.

[3] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech at very low bit rates," in *Proc. Int. Conf. Comm.*, Amsterdam, 1984, pp. 1610–1613.

[4] A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," in *Proceedings Asilomar Conference on Signals, Systems & Computers*, Nov. 2007, pp. 535–539.

[5] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 367–381, 1995.

[6] T. Linder and R. Zamir, "High-resolution source coding for non-difference distortion measures: The rate-distortion function," *IEEE Trans. Information Theory*, vol. 45, no. 2, pp. 533–547, 1999.

[7] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 2002, pp. 1805–1808.

[8] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615 – 3622, 1996.

[9] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and postfiltering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Istanbul, 2000, pp. 881–884.

[10] M. Li and W. B. Kleijn, "A low-delay audio coder with constrained-entropy quantization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Nov. 2007, pp. 191–194.

[11] W. B. Kleijn, "Principles of speech coding," in *Springer Handbook of Speech Processing*, J. Benesty, A. Huang, and M. Sondhi, Eds. Springer, Nov. 2007, ch. 14, pp. 283–306.

[12] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Nov. 2007, pp. 243–246.

[13] J. Rissanen, "Modeling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[14] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 3–14, 1993.

[15] P. Hedelin, "Single stage spectral quantization at 20 bits," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 1994, pp. 525–528.

[16] A. Vasilache, "Indexing of lattice codevectors applied to error resilient audio coding," in *Proceedings of the AES* 30^{th} *International Conference*, Saariselkä, Finland, March, 15-17 2007.

[17] W. H. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, p. 269275, 1991.

[18] M. Oger, S. Ragot, and M. Antonini, "Transform audio coding with arithmetic-coded scalar quantization and model-based bit allocation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 4, 2007, pp. 545–548.

[19] T. M. N. Hoang, M. Oger, S. Ragot, and M. Antonini, "Embedded transform coding of audio signals by model-based bit plane coding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 4013–4016.

[20] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, pp. 74–93, 2001.

[21] J. Klejsa, M. Kuropatwinski, and W. B. Kleijn, "Adaptive resolution-constrained scalar multiple-description coding," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2008, pp. 2945–2948.

[22] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, and H. Mikkola, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech Audio*, vol. 6, no. 8, pp. 620–636, 2002.