# Multiple Description Perceptual Audio Coding with Correlating Transforms

Ramon Arean, Jelena Kovačević, *Senior Member, IEEE*, and Vivek K Goyal, *Member, IEEE*

***Abstract*—In audio communication over a lossy packet network, concealment techniques are used to mitigate the effects of lost packets. This concealment is markedly improved if the compressed representation retains redundancy to aid in the estimation of lost information. A perceptual audio coder employing multiple description correlating transforms demonstrates this phenomenon.**

***Index Terms*—Audio coding, multiple descriptions, packetized audio, robust communication.**

## I. Introduction

MOST state-of-the-art audio coders combine source coding principles with perceptual modeling. These coders, called *perceptual audio coders*, use human hearing models to determine perceptual relevance and then eliminate redundancy with the minimal degradation of relevant information [1], [2]. Perceptual audio coders are naturally frame- or packet-based because perceptual masking thresholds are computed for finite input blocks. Each packet contains certain control information followed by entropy-coded quantized subband samples. In a communication environment, packets may be lost due to network congestion or uncorrected bit errors on a radio link. The decoder must then conceal the loss as much as possible.

A naive approach is to use a stationary model for the signal and replace the lost frame with the conditional expectation of that frame given the surrounding received frames. This results in a reconstructed frame which is nearly zero and has an audible dropout. A perceptually superior technique is to interpolate by extending the sinusoidal components from neighboring frames, but this too will leave an audible impairment unless the duration of the lost segment is very short.

This paper addresses the design of a system that is robust to packet losses. We will concentrate on Internet applications, where we may assume that packets either arrive correctly or are lost completely and that packets are identified by headers. Pairs of packets are made statistically predictable from each other. When one of a pair is lost, a reasonable estimate can be computed, but there is a price to be paid: correlation implies a reduction of source coding efficiency. The tool for this is the multiple description correlating transform (MDCT) introduced by Wang *et al*. [3], [4] and developed further by Goyal and Kovačević [5], [6]. The MDCT allows the correlation and predictability to be continuously adjustable.

The introduction of MDCT's in the Bell Labs PAC coder [2] yields a multiple description PAC (MDPAC) coder. The modification of the existing coder is simple because the perceptual masking is undisturbed. Nevertheless, MDPAC achieves considerable perceptual improvement with only a small increase in bit rate when the packet loss probability is moderate. This demonstrates another application domain for MDCT, which thus far has been limited to image coding [3], [7].

Manuscript received August 10, 1998; revised July 14, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Dr.-Ing. Juergen H. Herre.

R. Arean was with Bell Labs, Lucent Technologies, Murray Hill, NJ 07974 USA. He is now with Orange Communications, SA, Lausanne, Switzerland (e-mail: ramon.arean@orange.ch).

J. Kovačević and V. K Goyal are with the Mathematics of Communications Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: jelena@bell-labs.com; v.goyal@ieee.org).

Publisher Item Identifier S 1063-6676(00)01722-3.

## II. Perceptual Audio Coding

Human perception plays a key role in compression of audio material. As a result, recent audio standards work has concentrated on a class of audio coders known as *perceptual coders*. Rather than trying to model the source, perceptual coders model the listener and attempt to remove *irrelevant* information contained in the input signal. For a given bit rate, a perceptual coder will typically have a lower SNR than a lossy source coder design to maximize SNR, but will provide superior perceived quality to the listener. The combination of an appropriate *signal representation*, by means of a transform or a filter bank, and the *psychoacoustic model* of the destination provide the means to achieve efficient compression.

### A. Bell Labs PAC Coder

A block diagram of the PAC coder is as shown in Fig. 1. PAC divides the input signal into 1024-sample blocks of data—*frames*—used throughout the encoding process. It consists of five basic parts:

- The *analysis filter bank* converts the time-domain data to frequency domain. First, the 1024-sample block is analyzed and, depending on its characteristics, such as stationarity and time resolvability, a modified discrete cosine transform or a discrete wavelet transform is applied [2]. The total given bit rate and the sampling rate also play a role in the design of the transform. The analysis filtering produces either 1024- or 128-sample blocks of frequency-domain coefficients. In either case, the base unit for further processing is a block of 1024 samples.
- The *perceptual model* is used in computing the frequency-domain threshold of masking both from the time-domain signal and from the output of the analysis filter bank. A threshold of masking is the maximum noise one can add

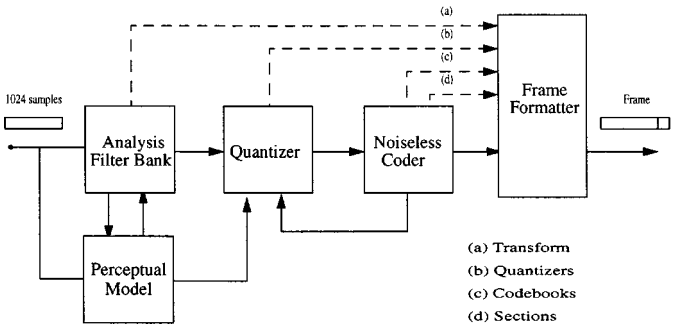



Fig. 1. PAC encoder block diagram.

to the audio signal at a given frequency without perceptibly altering it. Depending on the transform that was used previously, each 1024-sample block is split into a predefined number of groups of bands—*scale factor bands*. Within each scale factor band, a perceptual threshold value is computed.

- *Quantization:* Within each scale factor band the quantization step sizes are adjusted according to the computed perceptual threshold values in order to meet the noise level requirements. The quantization step sizes may also be adjusted to comply with a target bit rate, hence the feedback from the noiseless coder to the quantizer.
- *Noiseless coding:* Huffman coding is used to provide an efficient representation of the quantized coefficients. A set of optimized codebooks is used; each codebook codes sets of two or four coefficients. For efficiency, consecutive scale factor bands with the same quantization step size are grouped into sections, and the same codebook is used within each section. Failure to meet the target bit rate may trigger a recomputation of quantization step sizes.
- The *frame formatter* forms the bit stream, adding to the coded quantized coefficients the side information needed at the decoder to reconstruct the 1024-sample block. This block is defined as the *frame* and contains, along with one 1024-sample or eight 128-sample blocks, the following side information for each of them: the transform used in the analysis filter bank, section boundaries, codebooks, and quantization step sizes for sections. Side information accounts for 15% to 20% of the total bit rate of the coded signal.

At the decoder, the entropy coding, quantization, and transform blocks are inverted and an error mitigation block is added between the inverse quantization and the synthesis filter bank. In this block, lost frames are interpolated based on the preceding and following frames.

## III. Multiple Description Correlating Transforms

Traditionally, source and channel coding are separate; in essence, the source coding is designed with the assumption of a single lossless channel. When there are delay or complexity constraints, the channel coding will sometimes fail, resulting in poor performance.[1] This problem is particularly acute for channels with unpredictable variation, such as the Internet. In multiple description (MD) coding, the source encoder produces distinct descriptions for each of $M$ channels. A reconstruction may be formed from any subset of the channels and the problem is to make these reconstructions simultaneously "good." Of course, for a given total rate over the channels, there is a tradeoff between the qualities of the various reconstructions: At the one extreme, all the channels carry the same information and the reconstruction from any one channel is good (as a function of the received bit rate), but the reconstruction quality does not improve when more channels are received. At the other extreme, the bits from a source code are arbitrarily allocated to the channels. In this case, the reconstruction is good only when all the channels are received.

In this work, we equate channels with packets and limit our attention to $M = 2$ channels; extensions to $M > 2$ will be clear. Audio segments will be encoded in a long sequence of pairs of packets with the multiple description character. The first practical MD coding technique was quantizer-centric [8], but here we apply a transform-centric technique developed in [3]–[6].[2]

Suppose we are given a zero-mean jointly Gaussian two-dimensional source vector $x$. Without loss of generality—applying a Karhunen-Loève transform if necessary—we may assume $E[xx^T] = \mathrm{diag}(\sigma_1^2, \sigma_2^2), \sigma_1^2 \geq \sigma_2^2$. Suppose $x$ is encoded using a standard transform coder consisting of a linear transform followed by a scalar quantizer and a scalar entropy coder. An optimal transform, giving maximal coding gain, is the identity transform. Suppose each transform coefficient is a "description" in an MD scheme. Now what happens if one component is erased? When the low-variance component is lost the distortion is low, but when the high-variance component is lost the distortion is high. To improve upon this, each transform coefficient must capture some of the first component; in other words, the basis vectors should be skewed toward the first principal axis, or toward the component with larger variance. For a given quantizer, this reduces the average distortion when a component is lost. However, there is a price to be paid: Since the transform coefficients are correlated, the rate needed to transmit them is increased.

The specific steps to implement an MDCT of the source vector $x$ are as follows.

1) Uniform quantization: $x_{q_i} = [x_i]_\Delta$, where $[\cdot]_\Delta$ denotes rounding to the nearest multiple of $\Delta$.
2) $y = T(x_q)$ where $T : \Delta\mathbb{Z}^2 \rightarrow \Delta\mathbb{Z}^2$ is invertible.
3) The descriptions $y_1$ and $y_2$ are separately entropy coded.

The use of a discrete transform is to ensure cubic partition cells, as suggested in [4].

When both $y_1$ and $y_2$ are received, $x_q$ is recovered exactly. Otherwise, when one is lost the reconstruction of $x$ is the conditional expectation given the received data. Since $x$ is Gaussian, the conditional expectation has a simple form when the quantization error is small [5]; we will use this form although the data is not actually Gaussian.

[1]Here retransmission is considered part of channel coding.

[2]See [6], [8], and [9] for comprehensive introductions to MD coding.

It is shown in [5] that if $y_1$ and $y_2$ are equally likely to be lost, an optimal transform over a certain set of quasilinear transforms is[3]

$$T_\alpha(x_q) = \left[\begin{bmatrix} 1 & 0 \\ 1-2\alpha & 1 \end{bmatrix}\left[\begin{bmatrix} 1 & (2\alpha)^{-1} \\ 0 & 1 \end{bmatrix} \times \left[\begin{bmatrix} 1 & 0 \\ 2\alpha(\alpha-1) & 1 \end{bmatrix} x_q\right]_\Delta\right]_\Delta\right]_\Delta.$$

The invertibility of this transform is easy to verify. The relevant conditional expectations give reconstructions

$$\hat{x} = \frac{2\alpha}{4\alpha^4\sigma_1^2 + \sigma_2^2}\begin{bmatrix} 2\alpha^2\sigma_1^2 \\ \sigma_2^2 \end{bmatrix} y_1 \tag{1}$$

and

$$\hat{x} = \frac{2\alpha}{4\alpha^4\sigma_1^2 + \sigma_2^2}\begin{bmatrix} -2\alpha^2\sigma_1^2 \\ \sigma_2^2 \end{bmatrix} y_2. \tag{2}$$

The parameter $\alpha$ controls the trade-off between the redundancy and the average distortion when one component is lost. When distortion is measured in MSE per component and redundancy $\rho$ in bits per component and the quantization error is neglected, the trade-off follows [5]

$$D_1 = \frac{1}{2}\sigma_2^2 + \frac{\sigma_1^2 - \sigma_2^2}{4 \cdot 2^{2\rho}\left(2^{2\rho} + \sqrt{2^{4\rho} - 1}\right)}. \tag{3}$$

To encode more than two variables, one can form pairs of components and apply an MDCT to each pair. The allocation of redundancy to pairs is a simple convex optimization. The optimal coupling of variables pairs the largest variance component with the smallest variance component, the second largest with the second smallest, etc. [6], [9].

## IV. Multiple Description Perceptual Audio Coder

Fig. 2 depicts the MD version of the PAC coder. The only change to the PAC coder is the addition of an MDCT block with off-line design of transform parameters.

- An *MDCT* block is inserted between the quantizer and the noiseless coder. Within each 1024-sample unit, MDCT is applied to pairs of quantized coefficients, producing pairs of *MD-domain coefficients*. Within each pair, one MD-domain coefficient is assigned to each of Channel 1 and Channel 2. The pairings and the $\alpha$ parameters of the transforms are side information.

  In the decoder we add an inverse MDCT block that uses the side information (see Fig. 3).
- *Inverse MD transform*—This block performs the estimation and recovery of lost MD-domain coefficients when necessary. When both channels are received, this block simply inverts the MDCT's. When one channel is received, estimation follows (1)–(2). When both channels are lost, the built-in loss mitigation feature of PAC is used.

[3]Optimization criteria for other numbers of components and probabilities of loss are also given in [5].

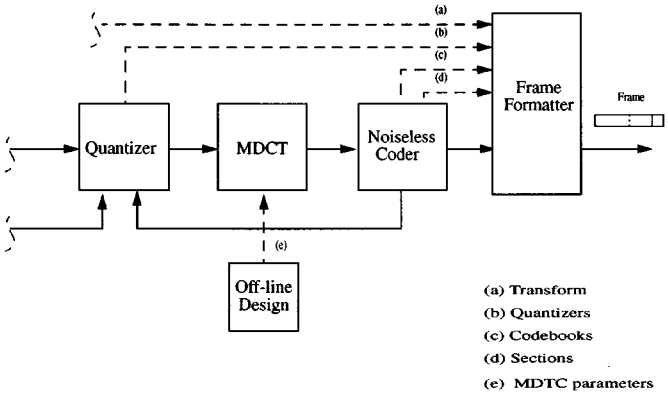



Fig. 2. MDPAC encoder block diagram.

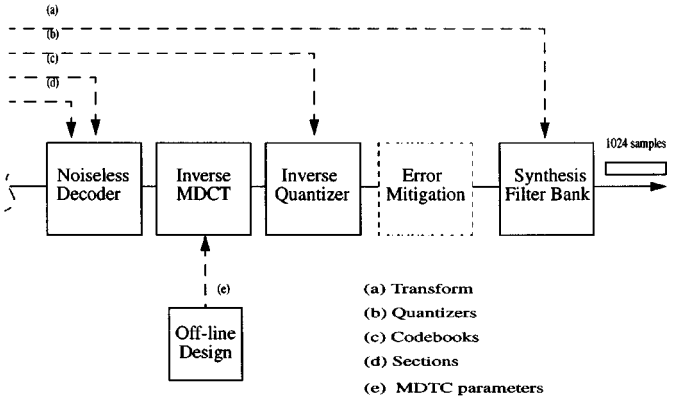



Fig. 3. MDPAC decoder block diagram.

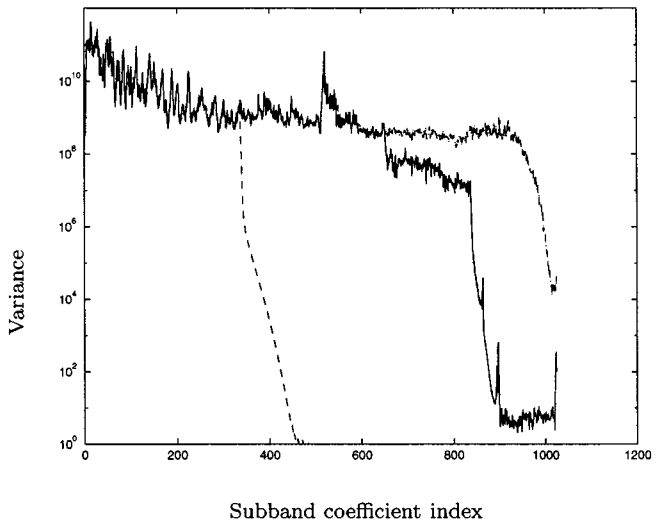



Fig. 4. Frequency-domain coefficient variances at bit rates (from left to right) 20 kbps, 30 kbps, and 48 kbps for File 9.

### A. Audio File Statistics

The second-order statistics of the source are needed for designing the optimal pairing and transform and for the estimation of lost coefficients. In the PAC structure, the bit rate affects the choice of analysis transform and thus the coefficient variances. This can be seen in Fig. 4, which gives the frequency-domain coefficient variances for an audio segment at three different bit rates. A bit rate of 20 kbps, suitable for Internet applications,

TABLE I
DESCRIPTIONS OF THE ANALYZED AUDIO FILES

| File | Sampling rate | Description | Duration [s] |
|---|---|---|---|
| 1 | 44.1 kHz | castagnette | 7.66 |
| 2 | 48.0 kHz | pipe | 14.42 |
| 3 | 44.1 kHz | violin (Ravel) [10, Track 59] | 29.00 |
| 4 | 44.1 kHz | piano (Schubert) [10, Track 60] | 92.00 |
| 5 | 44.1 kHz | symphonic (Strauss) [10, Track 65] | 112.00 |
| 6 | 44.1 kHz | symphonic (Mozart) [10, Track 67] | 82.00 |
| 7 | 44.1 kHz | symphonic (Baird) [10, Track 68] | 164.00 |
| 8 | 22.05 kHz | pop/rock (*Come Dancing* by The Kinks) | 42.71 |
| 9 | 22.05 kHz | pop/rock (*Underneath The Stars* by Mariah Carey) | 48.40 |

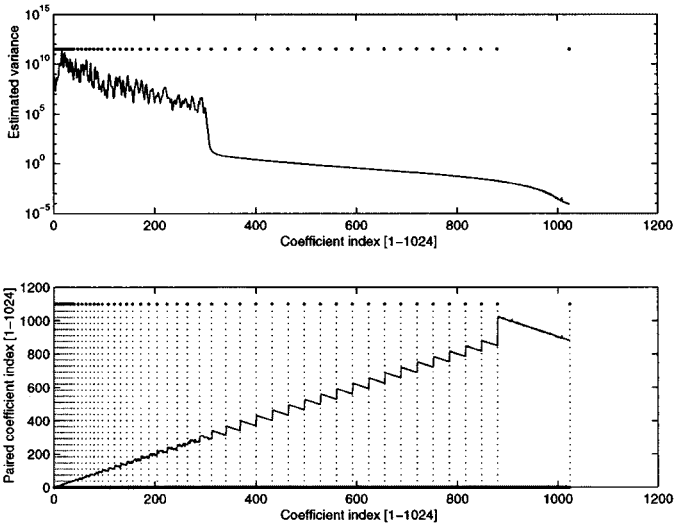



Fig. 5. Pairing design for audio file 6 coded at 20 kbps. Dotted vertical lines indicate scale factor band boundaries.

is used in subsequent analyses. Five files recommended by the European Broadcast Union [10] and four other files were analyzed. Table I gives brief descriptions of the files used in this project.

### *B. Pairing and Transform Design*

In the theoretical development of MDCT in [3]–[6], each transform coefficient has equal weight in the distortion metric and is quantized with an identical quantizer. In perceptual audio coding, the perceptual relevance and quantizer scaling of a transform coefficient are determined by its scale factor band; hence, to apply the MDCT without modification we should pair coefficients only within scale factor bands. Within each band, the optimal pairing described in Section III can be used. Fig. 5 shows this pairing for audio file 6 compressed to 20 kbps (mono).

The allocation of redundancies between pairs follows (3). Fig. 6 depicts, for the same audio file, the optimal redundancy allocation between pairs and the optimal transform parameter $\alpha$ for each pair. For comparison, mean redundancies of 0.1 and 0.5 bits per variable are shown; subsequent experiments use $\rho = 0.1$. (In the actual implementation, the very small values of $\alpha$ in the scale factor band where the variance drops sharply are adjusted upward.)

### *C. Entropy Coding and Side Information*

The entropy coding of PAC is unchanged, aside from insuring that the channels are coded independently.[4] The basic PAC side information must be duplicated so that it appears in both Channel 1 and Channel 2. In the monophonic case, this leads to an increase in the total bit rate of up to 20%. The MDCT parameters also constitute side information. We assume this is transmitted once, reliably, at the beginning of the transmission. With excessively fine coding (512 $\alpha$'s at 32 bits each and 512 10-bit integers to describe the pairing), the side information is less than 3 kB. This is small when amortized across the whole audio segment.

## V. EXPERIMENTAL RESULTS

To sensibly judge the performance of the MDPAC system one must *hear* the results. To enable this, sample audio files have been provided on-line in *aiff*, *next*, and *wave* formats at http://cm.bell-labs.com/who/jelena/Interests/MD/AudioDemo/DemoList.html.

The reader is invited to listen to the samples as the experiments are described.

Experiments were performed with the following two coders and bit rates:

- Coder 1: Original (single description) PAC at 20 kbps.
- Coder 2: MDPAC at 20 kbps or 26 kbps.

The rate of 26 kbps was chosen so that *with no packet losses* there is no perceptual difference between Coder 1 and Coder 2. Of the extra 6 kbps, 4 kbps are due to duplication of side information for independent transmission over each channel and 2 kbps are due to redundancy introduced by MDCT. The coders were simulated with various packet loss probabilities. In all cases packet losses were assumed to be independent. For the MD coder, the two channels were assigned separate packet loss probabilities to allow greater generality in the simulations.

[4]We did not redesign the Huffman codes, so the performance could probably be improved slightly.

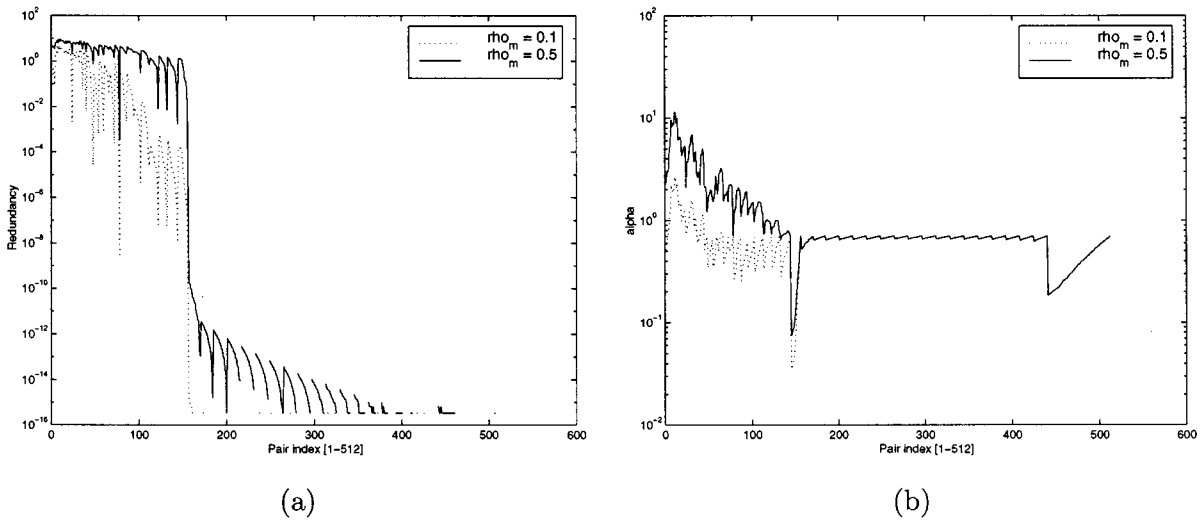



Fig. 6. Transform design for audio file 6 coded at 20 kbps: (a) optimally allocated redundancies; and (b) $\alpha$'s for each of the 512 pairs.

First let us note the performance with no packet losses. The performance of Coder 2 at 20 kbps is worse than that of Coder 1; the difference is small but certainly noticeable. As mentioned above, the performance of Coder 2 at 26 kbps is virtually identical to that of Coder 1. For the remainder of the section we compare Coder 2 at 26 kbps to Coder 1 with various packet loss probabilities in order to ascertain the robustness gained with the extra 6 kbps.

First suppose Coder 2 is operated such that one channel experiences no packet losses and the other has $2P$% losses. In comparison to Coder 1 with $P$% losses, Coder 2 performs comparably for $P = 5$ or $P = 20$; it sounds dramatically better for $P = 50$.

In many cases, the two virtual channels will have identical packet loss probabilities. Coders 1 and 2 were simulated for various packet loss probabilities. At packet loss probabilities of 5 to 20%, the performances of the coders are comparable because the simple frame interpolation in PAC works well. At higher packet loss probabilities, such as 50%, both coders are significantly degraded, but Coder 2 sounds much better. The MD coder might be improved with attention to smoothing the transitions between correctly received frames, frames estimated from one channel, and frames that are lost completely.

In conclusion, we find that inclusion of multiple description correlating transforms is a very easy way to improve the robustness of a perceptual audio coder. The MDCT does not affect the perceptual masking threshold calculation, but performance might be improved by accounting for cross-frequency masking. Further gains are expected from reducing side information.

## ACKNOWLEDGMENT

The authors would like to thank D. Sinha, G. Schuller, and P. Kroon of Bell Labs for their help and support. They also thank B. Yu of Bell Labs for fruitful discussions.

## REFERENCES

[1] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[2] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The Digital Signal Processing Handbook*, New York: IEEE Press, 1998, pp. 42.1–42.18.

[3] Y. Wang, M. T. Orchard, and A. R. Reibman, "Multiple description image coding for noisy channels by pairing transform coefficients," in *Proc. IEEE Workshop Multimedia Signal Processing*, Princeton, NJ, June 1997, pp. 419–424.

[4] M. T. Orchard, Y. Wang, V. Vaishampayan, and A. R. Reibman, "Redundancy rate-distortion analysis of multiple description coding using pairwise correlating transforms," in *Proc. IEEE Int. Conf. Image Processing*, vol. I, Santa Barbara, CA, Oct. 1997, pp. 608–611.

[5] V. K. Goyal and J. Kovačević, "Optimal multiple description transform coding of Gaussian vectors," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 1998, pp. 388–397.

[6] ——, "Generalized multiple description coding with correlating transforms," *IEEE Trans. Inform. Theory*, to be published.

[7] V. K. Goyal, J. Kovačević, R. Arean, and M. Vetterli, "Multiple description transform coding of images," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Chicago, IL, Oct. 1998, pp. 674–678.

[8] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821–834, May 1993.

[9] V. K. Goyal, "Beyond traditional transform coding," Ph.D. dissertation, Univ. Calif., Berkeley, 1998. (Univ. Calif., Electron. Res. Lab. Memo. UCB/ERL M99/2, Jan. 1999).

[10] PolyGram, Eur. Broadcast Union, CD-Sound Quality Assessment Material: Recording for Subjective Tests, Germany, 1997.

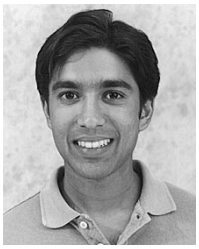
**Ramon Arean** was born in Lausanne, Switzerland, in 1974. He received the Certificate in mobile communications (with highest distinction) after a one-year program at the Institut Eurecom, Sophia Antipolis, France, in 1997, and the Dipl. Communications Systems degree from École Polytechnique Fédérale de Lausanne in 1998.

To complete his degree, he was with the Mathematics of Communications Department, Bell Labs, Lucent Technologies, Murray Hill, NJ, for six months during 1998. He is currently an Advanced Systems Engineer with Orange Communications, SA, Lausanne, working on value-added services such as PrePay and WAP.

**Jelena Kovačević** (S'88–M'91–SM'96) received the Dipl.Electr.Eng. degree from the University of Belgrade, Belgrade, Yugoslavia, in 1986, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, in 1988 and 1991, respectively.

In November 1991, she joined AT&T Bell Laboratories (now Lucent Technologies), Murray Hill, NJ, as a Member of Technical Staff. In the Fall of 1986, she was a Teaching Assistant at the University of Belgrade. From 1987 to 1991, she was a Graduate Research Assistant at Columbia University. In the summer of 1985, she worked for Gaz de France, Paris, France, during the summer of 1987 for INTELSAT, Washington, DC, and in the summer of 1988 for Pacific Bell, San Ramon, CA. Her research interests include wavelets, multirate signal processing, data compression, and signal processing for communications. She is a co-author of the book (with M. Vetterli) *Wavelets and Subband Coding* (Englewood Cliffs, NJ: Prentice-Hall, 1995). She is on the Editorial Boards of the *Journal of Applied and Computational Harmonic Analysis*, *Journal of Fourier Analysis and Applications*, and *Signal Processing Magazine*.

Dr. Kovačveić received the Belgrade October Prize, highest Belgrade Prize for student scientific achievements awarded for the Engineering Diploma Thesis in October 1986, and the E. I. Jury Award at Columbia University for outstanding achievement as a graduate student in the areas of systems, communication or signal processing. She served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and as a Guest Co-Editor (with I. Daubechies) of the Special Issue on Wavelets of the PROCEEDINGS OF THE IEEE. She is on the IMDSP Technical Committee of the Signal Processing Society of the IEEE and was a General Co-Chair (with Jan Allebach) of the Ninth Workshop on Image and Multidimensional Signal Processing.



**Vivek K Goyal** (S'92–M'98) was born in Waterloo, IA, in 1971. He received the B.S. degree in mathematics and the B.S.E. in electrical engineering (both with highest distinction), in 1993, from the University of Iowa, Iowa City. He received the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1995 and 1998, respectively.

He was a Research Assistant in the Laboratoire de Communications Audiovisuelles at École Polytechnique Fédérale de Lausanne, Switzerland, in 1996. He was with the Mathematics of Communications Department at Lucent Technologies' Bell Laboratories in 1997, where since 1998 he has been a Member of Technical Staff. His research interests include source coding theory, quantization theory, practical compression, and computational complexity.

Dr. Goyal is a member of Phi Beta Kappa, SIAM, Tau Beta Pi, and Eta Kappa Nu. In 1998, he received the Eli Jury Award of the University of California, Berkeley, awarded to a graduate student or recent alumnus for outstanding achievement in systems, communications, control, or signal processing.