A Differential Perceptual Audio Coding Method with Reduced Bitrate Requirements

M. Paraskevas and J. Mourjopoulos, *Member, IEEE*

***Abstract* - A new audio transform coding technique is proposed that reduces the bitrate requirements of the Perceptual Transform Audio Coders, by utilizing the stationarity characteristics of the audio signals. The method detects the frames which have significant audible content and codes them in a way similar to conventional Perceptual Transform Coders. However, when successive data frames are found to be similar to those sections, then their audible differences are only coded. An error analysis for the proposed method is presented and results from tests on different types of audio material are listed, indicating that an average of 30% in compression gain (over the conventional Perceptual Audio Coders bitrate) can be achieved, with a small deterioration in the audio quality of the coded signal. The proposed method has the advantage of easy adaptation within the Perceptual Transform Coders architecture and add only small computational overhead to these systems.**

I. INTRODUCTION

In recent years the introduction of digital audio as a method for storing, processing and transmitting high-fidelity acoustic signals has helped in the evolution of numerous applications in the field of consumer electronics and professional audio. It is also envisaged that, in the near future, significant new techniques will become commercially available and novel applications will emerge based on the manipulation of audio data within multimedia or audiovisual technologies. However, the feasibility of such future applications, as well as some current ones, greatly depends on the use of data compression techniques which reduce the data transmission rate and memory storage requirements. Given the existence of such techniques, terrestrial or satellite transmission channels can be economically employed for single or multi-channel audio data transmission, and also data storage media can be efficiently utilized for storing lengthy segments of acoustic signals.

The storage and transmission of such high-quality audio data (here it will be considered as reference the Compact Disc format, based on a 44.1 kHz sampling rate and 16-bit resolution) results in the relatively high bit-rate of 706 kBits/s, per data channel. This data rate can be technically or economically prohibitive for many applications, and this necessitates the introduction of data compression, preferably by using low-complexity methods (so that real-time implementations are not impeded), and without the insertion of perceptually detectable distortions. Applications which have emerged or are expected to appear with strong dependence on such signal compression technology [1], are in the area of high-fidelity audio for radio broadcasting (especially for the Digital Audio Broadcasting - DAB format [2]), in the area of multichannel audio for HDTV, in storing and processing of audio signals for domestic (e.g. multimedia or home studio) and professional applications (multichannel music recording), in transmitting audio data through computer or communication networks (e.g. ISDN), etc.

Coding and data compression methods for acoustic signals have been known for at least 4 decades, but until recently they were mainly concerned with speech signals [3], [4]. More recently, Transform Coding and data compression methods have been extensively used for wideband audio applications [4],[5] and during the last few years, the approach most suited for achieving the required data compression for high quality audio applications, has been shown to be based on utilising the masking properties of the human auditory system [6]. In this way, data irrelevancy can be removed from acoustic signals, without any noticeable effect to the listener. According to the findings in psychoacoustics [7], the masking mechanism occurs in the inner ear and dictates that noise spectral components can be inaudible provided that they coexist with other components of stronger amplitude. Perceptual Audio Coders (PAC) [8-22] utilize this phenomenon and shape quantization noise components below the masking threshold of the signal. The lower bound for the compression gain achieved by such Perceptual Audio Coders was determined by Johnston [23], who has shown that the information capacity of the human auditory system is in the region of 2 bits/sample, a bitrate which is sufficient for efficient transparent coding. This lower bound for PACs is achieved by Perceptual Entropy (PE) coding methods, which generate CD-quality audio at 128 kb/s per channel; slightly lower audio quality ("CD-like") requires 64 kb/s per channel, for a 20 kHz bandwidth signal [1],[16],[17].

For the implementation of such PACs, it is important to adopt an optimal strategy for time-frequency analysis, so that the required resolution in both domains is achieved and computational efficiency is maintained. Historically, several different approaches towards this objective have been attempted, belonging to the general class of the Lapped Orthogonal Transforms [24]. One approach, the Transform Coding technique [8-17], employs block FFT or DCT [25], usually on overlapping data sections. Such an approach, due to sub-critical sampling will generate a higher number of spectral coefficients compared to the time-domain data and may result in lower data compression efficiency. To avoid these problems, a Time Domain Aliasing Cancelling (TDAC) technique is adopted, often in conjunction with a Modified DCT (MDCT) method [26]. Such a technique was followed by the ASPEC audio coding system [10].

The second family of PAC methods, those based on Subband Coding [17-20], rely on parallel filter banks in order to achieve the frequency domain analysis and final reconstruction. Usually such filters are realized by QMF tree filter structures, which achieve the required non-constant with frequency spectral resolution and perfect signal reconstruction at the expense of computational efficiency [27]. Equally-spaced filter banks may be also used, often employing Polyphase structures [28], which allow sufficient spectral and time resolution. Such an approach, using 32 banks, was adopted by the MUSICAM audio coding system [20], as used by the Layers I and II of the recent ISO/MPEG audio coding Standard [29], and by the PASC system used in the Digital Compact Cassette (DCC) recording format. Finally, hybrid systems utilising both Polyphase filter banks and Modified DCT have been proposed in [21] and are employed by the Layer III of the ISO/MPEG audio coding Standard [29] and the MiniDisc recording format [14].

Although the performance of such Perceptual Audio Coders is quite satisfactory, there is a continuous need for a further reduction in data rate. As a contribution to this objective, a novel coding method is proposed here, which attempts to remove redundant time domain information from the signal by utilising its long and medium-term stationarity, whilst at the same time removing the perceptually redundant spectral information. As is known, by coding successive differences of a waveform, coding rate gains can be achieved. For

Manuscript received July 20, 1993; revised March 2, 1995. The associate editor coordinating the review of this paper and approving it for publication was Prof. James M. Kates.

The authors are with the Wire Communications Laboratory, Electrical Engineering Department, University of Patras, Patras, Greece.

IEEE Log Number 9414950

example, such principles are employed in ADPCM for time domain coding [4]. However this technique is mostly appropriate for speech coding and suffers from poor audio performance so that it is not suitable for high-quality audio applications. Here, a comparable technique is employed but for frequency domain coding according to which, the past history of the signal is manipulated in order to code current data by utilising subjectively significant information. It is also known that music sounds exhibit some stationarity of varying degree ranging from less than 100 ms, for extremely rapid pieces of music, to more than 1 second for sustained notes [12]. Since these intervals are longer than ordinary transform blocklength (20 msec) employed by perceptual coders, the correlation between spectra in successive frames is generally high, especially for coefficients corresponding to harmonics of the signal. Most audio coding techniques ignore the medium or long-term stationarity of the music signal and do not remove this type of redundancy in the audio data. In [11],[17],[29] a window switching mechanism was used, mainly in order to suppress pre-echo effects, and also to reduce coding bitrate. According to this approach, long analysis windows are used for stationary portions of the musical material and short analysis windows are used when a transient is detected. In order to choose between long and short window type, a special start or stop window type is employed. Another comparable approach was presented in [12]. According to this, a linear prediction technique was applied for predictive coding of the harmonics. In this case, the data blocklength was fixed.

Since the proposed method relies on coding the inter-frame differences of the perceptually significant spectral information, it is termed Differential Perceptual Audio Coding (DPAC) method. The main advantage of the DPAC approach over the existing PAC methods is that it achieves significant coding gains (up to 30%), without significant deterioration in coding signal quality. However, it has slightly higher coding complexity than these methods, but nevertheless this increased complexity can be easily accommodated within existing implementations of PAC methods.

The paper is organized as follows: in the section II, an initial description of the Transform Coding principles is given with specific reference in the perceptual models employed by such systems. In the section III, the principles and the theory of the proposed differential DPAC method are presented, followed by an analysis of the coding error generated by this technique. Based on this analysis, performance criteria are introduced in section IV, which are employed for the evaluation of the errors and the coding gains achieved by the method, when compared to established Perceptual Audio Coders. Finally, in the section V, such comparative results are presented, obtained from the analysis of a variety of musical segments and other audio signals.

II. OVERVIEW OF TRANSFORM CODING TECHNIQUES

A. Basic Concepts

Transform Coding methods are traditionally employed for high-quality audio data compression applications mainly because they operate in the frequency domain, which in turns results in performance gains obtained from the uncorrelated nature of the spectral components and the ability to minimize perceived distortions by optimal shaping of quantization noise in this domain [3]. As will be discussed in more detail later, this advantage can be further exploited by using the masking phenomenon of the human auditory system. Fig. 1 gives the block diagram of a Transform Coder and Decoder.

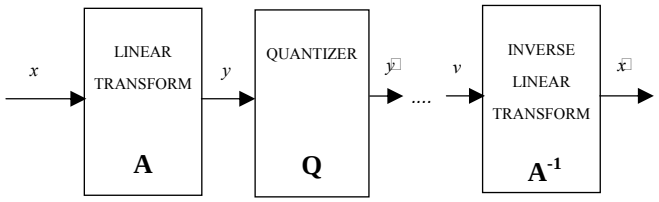

Fig. 1. Block diagram of an audio Transform Coder - Decoder (TC).

Let us assume the samples of a source signal x_n, $n=0,1,...N-1$. Using matrix notation, this signal sequence can be described as:

$$x = (x_0, x_1, ..., x_{N-1})^T \quad (1)$$

An N-dimensional linear transform applied on this vector, can be defined as the vector's multiplication, according to:

$$y = (y_0, y_1, ..., y_{N-1})^T = A\, x \quad (2)$$

where the $\boldsymbol{A}$-array is the forward transformation kernel. The components of vector y are called the transform coefficients (or frequency components or spectral coefficients), which are subsequently quantized by a set of memoryless quantizers $Q(.)$, to produce the quantized version:

$$\hat{y} = Q(y) = (\hat{y}_0, \hat{y}_1, ..., \hat{y}_{N-1})^T \quad (3)$$

The vector $\hat{y}$ is encoded and transmitted to the receiver. Assuming an error free data channel (i.e. $v = \hat{y}$), the decoder obtains an approximation $\hat{x}$ of the original signal vector x, by an inverse transformation on the received version $\hat{y}$, described by $\hat{x} = A^{-1} \hat{y} = B \hat{y}$, where $\boldsymbol{B}$ is the inverse transformation kernel. For an orthogonal transform the inverse is just the transpose of $B = A^{-1} = A^T$. Assuming a stationary input sequence with zero mean (μ_x=0) and variance σ_x^2, the average variance of the total reconstruction error between input x and output $\hat{x}$ is given [4], by

$$\sigma_r^2 = \frac{1}{N} E[(x-\hat{x})^T (x-\hat{x})] = \frac{1}{N} E[(y-\hat{y})^T (y-\hat{y})] = \frac{1}{N} E[q^T q] = \sigma_q^2 \quad (4)$$

where $q = y - \hat{y}$ is the vector of quantization error of the transform coefficients y. Hence, the reconstruction error variance equals the quantization error of the transform coefficients. The general problem addressed by any coding method is the optimum allocation of bits R_k, so that the average error variance is minimized, subject to a global constraint on the average bitrate R, which imposes a predefined constant bitrate value. It has been shown in [4] that the optimum bit allocation is given by the expression:

$$R_k = \frac{R_0}{N} + \frac{1}{2} \left[log_2 \sigma_k^2 - \frac{1}{N} \sum_{j=0}^{N-1} log_2 \sigma_j^2 \right] \quad (5)$$

where σ^2_k is the variance of k-indexed transform coefficient y_k, and R_0 is the total available number of bits per data block. In practice, when a quantizer is used for coefficients quantization, then the number of bits per coefficient R_k, must have a non-negative integer value.

B. Perceptual Audio Transform Coding

Perceptual Audio Coders are based on Transform Coding methods which have been developed during the last decade and are mainly based on an optimized time/frequency mapping and a dynamic noise shaping according to perceptual criteria. The operations of a PAC are described in Fig. 2. As can be observed, the PAC method employs a time/frequency transformation stage, a Psychoacoustic Model for estimation of Noise Masking Threshold (NMT) [29], and bit allocation routines [9], [11-13], [29] which ensure that quantization noise is kept below the NMT. The allocation of bits and the scale factor (maxima in spectral regions) are transmitted as side information, necessary for signal reconstruction by the decoder.

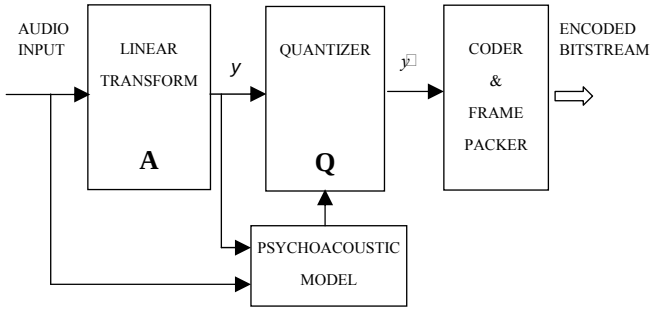

Fig. 2. Basic block diagram of a Perceptual Audio Coder (PAC).

C. The Masking phenomenon and its modelling

The lengthy research in psychoacoustics [7] has helped in the evolution of efficient algorithms or hardware realizations of auditory processes [30],[31]. With respect to masking phenomenon, digital signal processing methods have resulted in the evolution of many applications in the area of data compression [8] and in the evaluation of the quality of audio signals [32],[33]. In accordance with these models [8],[32],[33], the time/frequency mapping which occurs in the inner ear, can be represented by the Short-Time Fourier Transform (STFT) $X_f(k)$, $k=0,1,...N-1$, of the incoming signal $x(n)$, $n=0,1,...N-1$, derived from successive overlapping time windows of the audio data. For the practical implementation of this method, a STFT is used with a Hanning window of about 20 ms (1024 points for sampling frequency 44.1 kHz). From the FFT spectra the power spectrum $X_p(k)$, is calculated, corresponding to the rectification process occurring in the inner ear. At the next stage, the physical variable of frequency f, has to be mapped to the physiological variable of Bark scale, z, which describes the unequal incrementation of the critical band centre frequencies, and is given [7] by the expression

$$z = \frac{26.81\ f}{(\ 1960\ +\ f\ )} - 0.53 \qquad (6)$$

From linear spectral energy density $X_p(k)$ and using the above described frequency scale warping f (Hz) $\rightarrow$ z (Bark), the Bark spectrum energy density $X_b(i)$, can be computed by the expression

$$X_b(i) = \sum_{k=f_{li}}^{f_{hi}} X_p(k) \qquad (7)$$

where i is the discrete value of the continuous Bark scale z, and corresponds to the critical band number. Also f_{li} and f_{hi} are the lower and upper boundaries of critical band i, respectively. The Bark spectrum energy density $X_b(i)$, within each of these bands must be converted to basilar membrane excitation, evaluated by the convolution of $X_b(i)$ with the characteristic basilar membrane spreading function $\Lambda(z)$ [6], given by

$$10\log_{10}\Lambda(z) = 15.8114 + 7.5(z+0.474) - 17.5(1+(z+0.474)^2)^{1/2} \qquad (8)$$

Hence the discrete time critical band spectrum $E(i)$, is given by [8]

$$E(i) = X_b(i) * \Lambda(i) = \sum_j X_b(j)\ \Lambda(i-j) \qquad (9)$$

where i is the index of critical band of the masked signal, and j is the index of critical band of the masking signal.

The output of this convolution is the critical band spread spectrum $E(i)$, from which the NMT is then evaluated. It is generally accepted that when a tone is masking noise then this threshold is at 14.5+i dB below $E(i)$ [8] and when noise is masking a tone, then the threshold is at 5.5 dB below this function, [34]. Since each particular frame of the signal's spectral content may have properties between these two extreme cases, the spectral tonality measure $\alpha(i)$ in each critical band, is used. This measure is corresponding to predictability of the signal and is estimated in the spectral magnitude/phase domain [21]. This index $\alpha(i)$, is then employed for the estimation of the offset masking energy $O(i)$ (dB), which is given by

$$O(i) = \alpha(i)\ (\ 14.5 + i\ ) + (\ 1 - \alpha(i)\ )\ 5.5 \qquad (10)$$

The offset masking energy gives the value which must be subtracted from the critical band spread spectrum $E(i)$, to produce the "raw" masking threshold spectrum, $T(i)$, according to

$$T(i) = 10^{\log_{10} E(i) - \frac{O(i)}{10}} \qquad (11)$$

Then, this function is renormalized, (taking into account the non-normalized nature of the spreading function), and is converted to the frequency domain function $T(k)$, using the inverse frequency scale warping z (Bark) $\rightarrow$ f (Hz). Finally, the absolute threshold of hearing [7] is compared to the previously estimated function in order to define regions where the NMT falls below the lower thresholds of audibility. In such cases, the value of the absolute threshold is substituted in the estimated NMT data.

III. THEORY OF DIFFERENTIAL PERCEPTUAL AUDIO CODING

A. Overview of the Technique

In the proposed technique, the elimination of data redundancy, is obtained by replacing spectral coefficients with zeroes, when no audible changes corresponding to these coefficients are detected in successive audio data frames [35]. Such zeroed spectral components can be reconstructed in the receiver, using the respective values from the previously decoded data frame. At this point, two definitions must be given:

a) **Reference** Frames are defined as the frames which include dominant sound transients and are characterized by significant audible differences compared to previous frames. These frames are coded using the established PAC schemes and at a bitrate comparable to that achieved by the existing perceptual coders.

b) **Simple** Frames, are defined as the frames which have small audible differences compared to previous frames. These frames are coded using only the spectral components which differ from a previously defined Reference Frame whilst the rest of their spectral values are coded as zeroes. In this way, the bitrate required for coding such frames is significantly reduced with no loss in their perceived quality.

In the current work, the indices R and S are used to denote a Reference or a Simple Frame respectively. Also, the prime is used to signify parameters which have been evaluated through the proposed Spectral Coefficient Selection (SCS) procedure, as will be discussed later.

Fig. 3 shows the basic structure of the proposed audio coder. The proposed DPAC scheme can be briefly described by the following steps:

1) For the input vector $\boldsymbol{x}$, the spectral domain coefficients $\boldsymbol{y}$, are obtained using the Modified DCT [26].
2) A Frame Selection (FS) module, defines the status of each frame as was defined in the previous paragraph. The decision is based on qualitive and quantitive criteria which will be discussed in detail, in Section IV-C.
3) A Spectral Coefficient Selection (SCS) module, is enabled from the FS module only when Simple Frames are detected and replaces some of the spectral components of the current frame by the respective components of the Reference Frame. The detailed operation of this module will be given in Section III-B.
4) The psychoacoustic masking function is estimated according to the discussion in section II C, in order to define the allocation of bits and the noise shaping. In this work, the Psychoacoustic Model

2 of the ISO/MPEG Standard (as used in the Layer III of ISO Audio Coder), was employed [29].

5) A uniform quantizer (zero memory Lloyd-Max quantizer [4]) is used to map the transform coefficients to the corresponding quantized data.
6) The quantized values are coded via a Huffman Coder and the data are multiplexed and transmitted to the receiver.

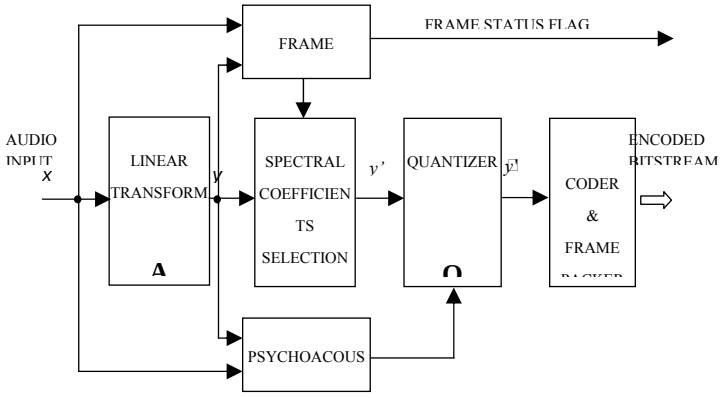

Fig. 3. Block diagram of the Differential Perceptual Audio Coder (DPAC).

B. Error Analysis for the DPAC

The Spectral Coefficient Selection (SCS) module replaces the spectral components of a Simple Frame using the respective components of the Reference Frame, when the replacement criterion is satisfied. Furthermore, the operations of this module differ depending on its application to Simple or Reference Frames. Clearly, the SCS results in a substitution which will generate errors which differ from those given by equation (4). Here, a study of this error will be provided which will be complemented by experimental results given in section V.

1) Reference Frame Quantization Error: If the current frame has been identified (by the FS module) as a Reference Frame, then the vector of the transform coefficients is given by $y_R = A\,x_R$. In this case, the output of the SCS module will be:

$$y_{R'} = S(y_R) = y_R \qquad (12)$$

Therefore, the spectral substitution error is $t_R = y_R - y_{R'} = 0$, and its variance will be obviously $\sigma_{tR}^2 = 0$, since no spectral substitution has taken place. The quantized version of the vector $y_{R'}$, is given by the expression:

$$\hat{y}_{R'} = Q(y_{R'}) = Q(y_R) = \hat{y}_R \qquad (13)$$

and the quantization error will be $q_R = y_R - \hat{y}_R$. Therefore, the total reconstruction error (in the time domain) will be $r_R = x_R - \hat{x}_R$, and it can be easily proven (from (4)) that its variance equals to the variance of the TC quantization error, i.e. $\sigma_{rR}^2 = \sigma_{qR}^2$. Hence, no additional noise is added to these frames due to the proposed differential coding approach.

2) Simple Frame Quantization Error: For a Simple Frame, the vector of the transform coefficients is given by $y_S = A\,x_S$. Then, the vector y_D of the absolute difference between the compared spectra y_R and y_S, is computed as:

$$y_D = \{y_{Dk}\} = \{|y_{Rk} - y_{Sk}|\},\ k = 1,\ldots,N \qquad (14)$$

and its average value $\bar{y}_D$ can be also estimated. Then, the vector y_D^* is composed from those components of the vector y_D, which exceed a predetermined threshold which is defined as $\xi \bullet \bar{y}_D$, i.e.

$$y_D^* = \{y_{Dk}^*\} = \begin{Bmatrix} 0 & if\ y_{Dk} \le \xi \bullet \bar{y}_D\ (M\ values) \\ y_{Dk} & if\ y_{Dk} > \xi \bullet \bar{y}_D\ (L\ values) \end{Bmatrix},\quad k = 1,\ldots,N \qquad (15)$$

where $M+L=N$.

The factor ξ is estimated from experimental results and represents the amount of the allowed inter-frame spectral component variation having typical values between 0.1 and 0.3. Also, the vector $\bar{y}_D^*$ is the complementary to y_D^* and is defined by:

$$\bar{y}_D^* = \{\bar{y}_{Dk}^*\} = \begin{Bmatrix} y_{Dk} & if\ y_{Dk} \le \xi \bullet \bar{y}_D\ (M\ values) \\ 0 & if\ y_{Dk} > \xi \bullet \bar{y}_D\ (L\ values) \end{Bmatrix}$$
$$k = 1,\ldots,N \qquad (16)$$

Therefore, the substitution operation *S(.)* performed by the SCS module (for a Simple Frame), can be described by the expression:

$$y_{S'} = \{y_{S'k}\} = S(y_{Sk}) = \begin{Bmatrix} y_{Sk} & if\ y_{Dk} = y_{Dk}^*\ (L\ values) \\ y_{Rk} & if\ y_{Dk} = \bar{y}_{Dk}^*\ (M\ values) \end{Bmatrix}$$
$$k = 1,\ldots,N \qquad (17)$$

The $y_{Dk} = y_{Dk}^*$ indicates the L spectral components which are differ significantly compared to the Reference Frame, whereas the $y_{Dk} = \bar{y}_{Dk}^*$ indicates the M spectral components which will not be changed significantly compared to the Reference Frame. Hence, the spectral substitution error vector is defined by:

$$t_S = y_S - y_{S'} = \{y_{Sk} - y_{S'k}\} = \begin{Bmatrix} 0 & if\ y_{Dk} = y_{Dk}^*\ (L\ values) \\ y_{Sk} - y_{Rk} & if\ y_{Dk} = \bar{y}_{Dk}^*\ (M\ values) \end{Bmatrix},$$
$$k = 1,\ldots,N \qquad (18)$$

and its variance, (assuming that the vectors $\boldsymbol{y_R}$ and $\boldsymbol{y_S}$ are uncorrelated), can be described by:

$$\sigma_{tS}^2 = \frac{1}{N} E\left[t_S^T\, t_S\right] = \frac{1}{N} \sum_{k=0}^{N-1} E\left[t_{Sk}^2\right]$$

$$= \frac{1}{N}\left[\sum_M E\left[y_{Sk}^2\right] - \sum_M E\left[y_{Rk}^2\right]\right] \qquad (19)$$

$$= \frac{1}{N}\left[\sum_M \sigma_{Sk}^2 - \sum_M \sigma_{Rk}^2\right]$$

The above equation shows that the variance of the total spectral substitution error for the complete N-point block signal, i.e. $N.\sigma_{tS}^2$, equals to the difference of the variances σ_S^2 and σ_R^2, only with respect to the substituted coefficients. The quantized version of the vector $y_{S'}$, is given by:

$$\hat{y}_{S'} = \{\hat{y}_{S'k}\} = Q(y_{S'k}) = \begin{Bmatrix} \hat{y}_{Sk} & if\ y_{Dk} = y_{Dk}^*\ (L\ values) \\ \hat{y}_{Rk} & if\ y_{Dk} = \bar{y}_{Dk}^*\ (M\ values) \end{Bmatrix},\ k = 1,\ldots,N$$
$$k = 1,\ldots,N \qquad (20)$$

whereas from equations (20) and (17), the vector of quantization noise is:

$$q_S = \{q_{Sk}\} = y_{S'k} - \hat{y}_{S'k} = \begin{Bmatrix} y_{Sk} - \hat{y}_{Sk} & if\ y_{Dk} = y_{Dk}^*\ (L\ values) \\ y_{Rk} - \hat{y}_{Rk} & if\ y_{Dk} = \bar{y}_{Dk}^*\ (M\ values) \end{Bmatrix}$$
$$k = 1,\ldots,N \qquad (21)$$

Finally, assuming an error free communication channel, then the

receiver will obtain an approximation of the input signal using the inverse transform on the vector $\hat{y}_{S'}$, according to $\hat{x}_S = A^{-1}\hat{y}_{S'}$. Therefore, the total DPAC reconstruction error in the time domain is described by:

$$r_S = x_S - \hat{x}_S = A^{-1}(y_S - \hat{y}_{S'}) = \left\{ \begin{array}{l} A^{-1}(y_{Sk} - \hat{y}_{Sk}) \text{ if } y_{Dk} = y^*_{Dk}\ (I) \\ A^{-1}(y_{Sk} - \hat{y}_{Rk}) \text{ if } y_{Dk} = \bar{y}^*_{Dk}\ (II) \end{array} \right\}$$
$$k = 1,\ldots,N \qquad (22)$$

An expression which defines the variance of the total DPAC reconstruction error for the Simple Frames will be now given. According to the previous discussion, the total distortion introduced in a Simple Frame is caused by the spectral substitution and the quantization processes. The branch *(I)* corresponds to the quantization noise of the L spectral components, which are not affected by the SCS operation, i.e.:

$$(I): A^{-1}(y_S - \hat{y}_S) = A^{-1}(q_S) \qquad (23)$$

The branch *(II)* corresponds both to the spectral substitution operation (17) and to the quantization of the reference components (13) and is therefore expressed as:

$$(II): A^{-1}(y_S - \hat{y}_R) = A^{-1}(y_S - y_R + y_R - \hat{y}_R) = A^{-1}(t_S + q_R) \qquad (24)$$

Therefore, the variance of the total reconstruction error is given by:

$$\sigma^2_{rS} = \frac{1}{N} E\left[r_S{}^T r_S \right] = \frac{1}{N} E\left[(y_S - \hat{y}_{S'})^T (y_S - \hat{y}_{S'}) \right]$$

$$= \frac{1}{N} E\left[\sum_{k=0}^{N-1} [y_{Sk} - \hat{y}_{S'k}]^2 \right]$$

$$= \frac{1}{N} E\left[\sum_{L} [y_{Sk} - \hat{y}_{Sk}]^2 + \sum_{M} [y_{Sk} - \hat{y}_{Rk}]^2 \right]$$

$$= \frac{1}{N} E\left[\sum_{L} [q_{Sk}]^2 + \sum_{M} [t_{Sk} + q_{Rk}]^2 \right]$$

Assuming that the vectors of the quantization and the spectral substitution are uncorrelated, then it is:

$$\sigma^2_{rS} = \frac{1}{N}\left[\sum_{L} E[q^2_{Sk}] + \sum_{M} E[t^2_{Sk}] + \sum_{M} E[q^2_{Rk}] \right]$$

$$= \frac{1}{N}\left[\sum_{j=1}^{L} \sigma^2_{qSj} + \sum_{j=1}^{M} \sigma^2_{tSj} + \sum_{j=1}^{M} \sigma^2_{qRj} \right]$$

The above equation shows that the variance of the total DPAC error, which is introduced by the spectral substitution and the quantization procedures in a Simple Frame, is the summation of the variance of the quantization noise for these components which are not affected by the SCS module, the variance of the quantization noise for the spectral components of the corresponding Reference Frame which are modified by the SCS module, and the variance of the spectral substitution error as is described by equation (26). This indicates that the additional error of the proposed DPAC method, is mainly due to the difference in variance between the frames' original and substituted spectral coefficients. The above conclusion shows that the DPAC error performance mainly depends on the statistical differences observed between such spectral components and is, in principle, dependent on the type of audio material (and the SCS module decision threshold value). For this reason, performance results are be presented in section V, giving the DPAC error as was obtained from the analysis of different audio signals.

IV. IMPLEMENTATION OF THE DPAC

A. Encoding Process

In more detail, the operations of the proposed DPAC scheme are described by the following steps:

1) A Modified Discrete Cosine Transformation (MDCT), using the Time Domain Aliasing Cancellation method (TDAC) is used. Several fast algorithms are known for its implementation [24],[36], which can even realized in real time [37].

Let us assume that $x_t(n)$, $n=0,\ldots,N-1$, is the t-th data block of the input signal and $y_t(k)$, $k=0,\ldots,(N/2)-1$, are the generated spectral coefficients. The forward MDCT is defined by [26]:

$$y_t(k) = \sum_{m=0}^{N-1} f(m) x_t(m) \cos\left(\frac{\pi}{2N} (2m + 1 + \frac{N}{2})(2k+1) \right)$$
$$k = 0,\ldots,\frac{N}{2} - 1 \qquad (27)$$

where $f(.)$ is the analysis-synthesis window, which is a symmetric window such that its added-overlapped effect is producing a unity gain in the signal.

The inverse MDCT is given by [26]:

$$u_t(p) = f(p) \sum_{k=0}^{N/2-1} y_t(k) \cos\left(\frac{\pi}{2N} (2p + 1 + \frac{N}{2})(2k+1) \right)$$
$$p = 0,\ldots,N-1 \qquad (28)$$

hence, the reconstructed signal is given by an overlap-add operation, i.e.:

$$\tilde{x}_t(n) = u_{t-1}\left(n + \frac{N}{2} \right) + u_t(n) \qquad n = 0,\ldots,\frac{N}{2} - 1 \qquad (29)$$

2) Initially, the first frame is assumed to be a Reference Frame, and then the following functions are calculated:

1) The MDCT coefficients vector $\boldsymbol{y}_{\mathbf{R}}$.

2) The respective Noise Masking Threshold, (NMT) Thr_R.

This vector is quantized by a uniform quantizer and entropy-coded by a Huffman coder [4].

3) The next data frame is initially assumed to be a Simple Frame, but its exact status must be evaluated according to the FS procedure. The following functions are then calculated:

a) The MDCT coefficients vector $\boldsymbol{y}_{\mathbf{S}}$.

b) The respective NMT, Thr_S.

4) In this step, the status of the current frame is evaluated. This task is performed by the Frame Selection (FS) module, using a variety of selection criteria. These criteria are described in detail in the next section. Hence, if the current frame is found to be a Reference one, then the original spectrum $\boldsymbol{y}_{\mathbf{S}}$ is quantized and coded using the standard PAC approach (section II B). Otherwise, (i.e. if the current frame is found to be a Simple one), then the components y_{Sk}, for $y_{Dk} = y^*_{Dk}$ are quantized and coded. The substituted values from the Reference Frame y_{Rk}, for $y_{Dk} = \bar{y}^*_{Dk}$, are not quantized, coded or transmitted. In this way a reduction in the overall data rate transmission is obtained. This approach yields always improved coding performance, in contrast to an alternative approach which could be based on coding differences between successive frame spectral coefficients. In such a case, coding requirements could increase since due to phase variation, such difference could have greater absolute value than the coefficient itself.

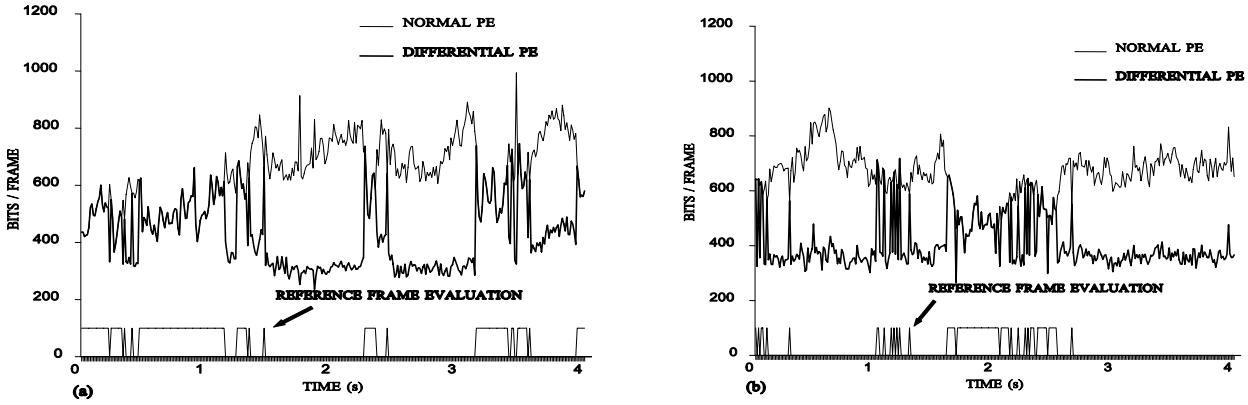

Fig. 4. PE bitrate requirements, expressed as function of time, for the conventional PAC method and the proposed DPAC method: (a) results for big band jazz segment (JZZ). (b) results for orchestral music segment (MZT).

It is clear from the above, that the adoption of the proposed coding scheme (i.e. separation into Reference and Simple Frames), causes a non-uniform variation (in time domain) of the required bitrate, as is also shown in Fig. 4. Therefore, since it is usually necessary to keep the bitrate R constant, a time buffer must be employed by the system. This buffer must have the capacity for storing data for approx. 2-4 s as was found from the experimental results discussed in section V, and also in [23]. Moreover, the more sophisticated mechanism of "bit reservoir" [17],[19], can be also used in order to preserve the accuracy of time buffering. According to this, unused bits in one frame can be used by other frames, i.e. when the instantaneous bitrate demand is higher.

B. Decoding Process

The decoder has a simpler structure compared to the encoder. One bit flag which is transmitted as side information, allows the discrimination of the type of the processed frame. If a Reference Frame is processed, then the decoding operation is the same as with the established perceptual decoders. If a Simple Frame is decoded, then the zeroed quantized values are substituted by the corresponding values of the Reference Frame. It is obvious that the differential decoder has a more complex structure compared to the established perceptual decoder schemes and also requires some amount of memory (i.e. $N/2$ memory location), in order to store the decoded values of the Reference Frame. However, this additional complexity is by no means prohibitive for its real time implementation of the DPAC method. The block diagram of the differential decoder is shown in the Fig. 5.

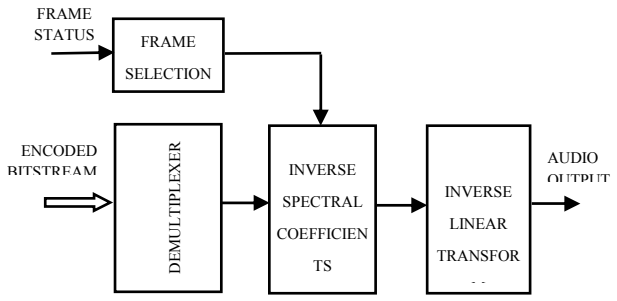

Fig. 5. Block diagram of the Differential Perceptual Audio Decoder.

C. Frame Selection Criteria

The performance of the proposed audio coding technique largely depends on the accuracy of frame status selection and especially on the exact determination of the similarity of the spectral components of the Reference and Simple Frames. In principle, different selection criteria can be tested, since they may perform differently when applied to various types of audio material. In this work, three criteria were tested based on mathematical and information theory concepts.

1) The first criterion, is based on the Time Domain Euclidean Distance (TDED) (interblock distance), between the initial or earlier detected Reference Frame and frame under test, normalized according to the Reference Frame energy. This criterion is defined as

$$\lambda = \frac{\left[\frac{1}{N} \sum_{n=0}^{N-1} \{ x_R(n) - x_S(n) \}^2 \right]^{1/2}}{\left[\frac{1}{N} \sum_{n=0}^{N-1} x_R^2(n) \right]^{1/2}} \qquad (30)$$

where $x_R(n)$ is the n-th time-domain sample in the Reference Frame of block size N, and $x_S(n)$ is the n-th time-domain sample in the frame under test. The decision level λ was obtained from experimental results and was also found to depend on the type of the processed audio material. The above criterion is entirely objective, since no psychoacoustic weighting is employed, and represents the difference between the two waveforms. According to Parceval's theorem this criterion has an equivalent frequency domain interpretation. A similar criterion was proposed in [11], in order to select the optimal block-size in an ATC technique with adaptive block-size MDCT.

2) The second criterion is based on the well known information concept of Perceptual Entropy (PE). As is known [23], the PE represents the minimum number of bits per spectral coefficient, required by an ideal Transform Coder to achieve transparent audio coding. In the case of perceptual coders the quantization noise which is injected at each frequency line, is shaped just below the NMT and hence no audible distortions are generated. In our case, the Perceptual Entropy $PE_{S'}$, of the modified spectrum $y_{S'}$ was calculated, as well as the Perceptual Entropy PE_S of the original spectrum $\boldsymbol{y}_S$. If the ratio λ of these two entropies, was not found to exceed some predefined threshold, then the frame under test was assumed to be a Simple Frame, i.e.

$$\lambda = PE_{S'} / PE_S \qquad (31)$$

The value of this decision threshold was obtained from experimental results and was not found to depend on the type of the processed audio material.

3) The third criterion is based on the ratio of the spectral shape for the two spectra under test, as measured by the Spectral Flatness Measure (SFM) [38]. The Spectral Flatness Measure (SFM) γ_x^2, of an zero-mean process $x(n)$ with discrete power spectral density $S_x(k)$, is expressed by [4]

$$\gamma_x^2 = \frac{\left[\prod_{k=1}^{N} S_x(k) \right]^{1/N}}{\frac{1}{N} \sum_{k=1}^{N} S_x(k)} \qquad (32)$$

i.e. the SFM can be computed as the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum. The inverse of the SFM, γ_x^{-2}, is a measure of signal predictability and typically has values between 3 and 16 for long-term power spectral densities of speech signals [4]. Viewed in a different way, the SFM represents the global tonality of the short time spectrum of the signal.

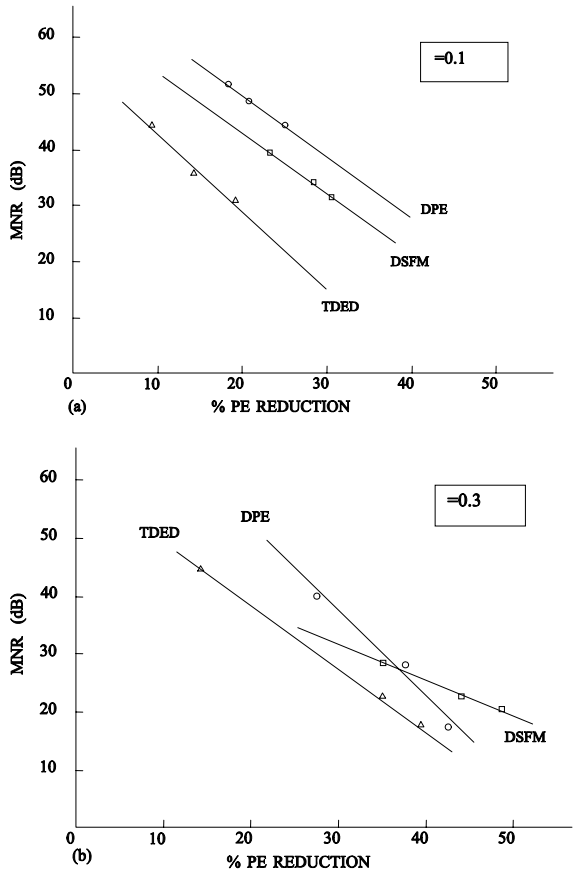

Fig. 6. MNR and corresponding PE reduction, for the different values of parameter ξ: (a) ξ=0.1 and (b) ξ=0.3.

Finally, as a frame selection criterion, the measure of difference (λ) between Reference Frame SFM measurement and the frame under test was employed, i.e.

$$\lambda = Abs\left(10\ log_{10}\ \frac{SFM_S}{SFM_R} \right), \quad (dB) \qquad (33)$$

If this value was found not to exceed a predefined (from experimental results) value, then this indicate that no significant audible changes existed between the compared spectra and in this case the frame under test was assumed to be a Simple Frame.

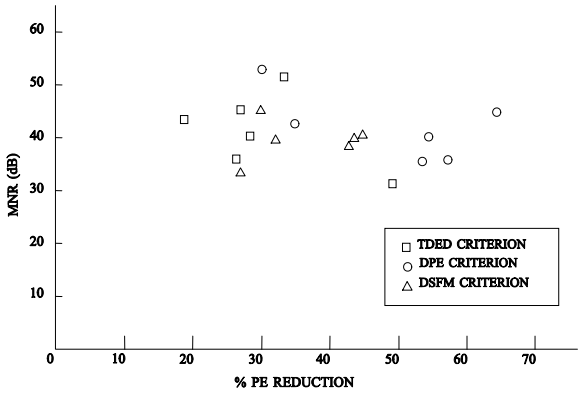

Fig. 7. PE reduction and corresponding audio quality for the different FS criteria, and for the complete audio database.

V. TESTS AND RESULTS

In this section, the experimental results are presented, which were obtained by the use of the DPAC technique on a variety of audio data. This presentation is divided into two subsections. In the first subsection, the PE reduction achieved by the DPAC technique is tested with respect to the resulting sound quality. The subjective audio quality of the coded signal was tested by the known objective Noise-to-Mask Ratio (NMR) criterion [39], expressed here for clarity as the Mask-to-Noise Ratio (MNR). As was found out in [33], this objective criterion has a very high correlation with subjective results obtained by a panel of listeners. Therefore, this criterion was the most appropriate from a perceptual point of view and was also found by the authors to give a good impression of the audio quality of the coder material. Also, the calibration procedure for parameters λ and ξ is described and the appropriate FS criterion is selected. In the second subsection, the overall performance of the proposed DPAC system (i.e. including the quantization and coding processes) is measured as compared to existing PAC systems. For these tests, a database of audio signals was created, containing segments of different pieces of music and other audio material, contained in the author's database and the ITU-R Recommendation [40], listed in Table I. All these data were sampled at 44.1 kHz and normalized according to the dynamic range of a 16-bit A/D converter.

TABLE I
DESCRIPTION OF AUDIO MATERIAL USED FOR THE EVALUATION TESTS

ITEM NO	FILE CODE	MATERIAL CATEGORY	MATERIAL DESCRIPTION	DATABASE
1	GSP	SPEECH	GERMAN MALE SPEECH	ITU-R
2	DST	POP MUSIC	DIRE STRAITS	-"-
3	SVG	POP MUSIC (VOCALS)	SUZAN VEGA	-"-
4	CLM	CONTEMPORARY JAZZ	ORNETTE COLEMAN	-"-
5	RVL	CLASSICAL	RAVEL	-"-
6	BSG	SOLO INSTRUMENT	BASS GUITAR	-"-
7	HRP	SOLO INSTRUMENT	HARPSICHORD	-"-
8	TRN	SOLO INSTRUMENT	TRIANGLES	-"-
9	JZZ	BIG BAND JAZZ	HARRY JAMES	AUTHOR'S
10	BTH	CLASSICAL	BEETHOVEN ORCHESTRAL	-"-
11	MZT	CLASSICAL	MOZART INSTRUMENTAL	-"-
12	MSG	CLASSICAL	MUSSORGSKY ORCHESTRAL	-"-
13	SPC	SPEECH	ENGLISH FEMALE SPEECH	-"-
14	POP	POP MUSIC	TRACY CHAPMAN	-"-

A. Audio Signal Perceptual Entropy reduction

All the following results were obtained by measurements made at the output of the SCS stage of the DPAC system, i.e. without the use of quantization and coding. This evaluation stage has been included in order to accurately assess the effects of the proposed Spectral Coefficient Substitution technique without the subsequent contribution of the quantization stage. Furthermore, this evaluation stage will investigate the following topics:

i) The effect of the parameters λ (Frame Selection Decision Threshold) and ξ (Spectral Substitution Decision Threshold), on the resulting compression gain and audio quality, as well as the experimental evaluation of the optimal values for the above parameters.

ii) The experimental determination of the optimum Frame Selection (FS) criterion.

iii) The evaluation of the spectral substitution error properties and its relation to the MNR audio quality measure.

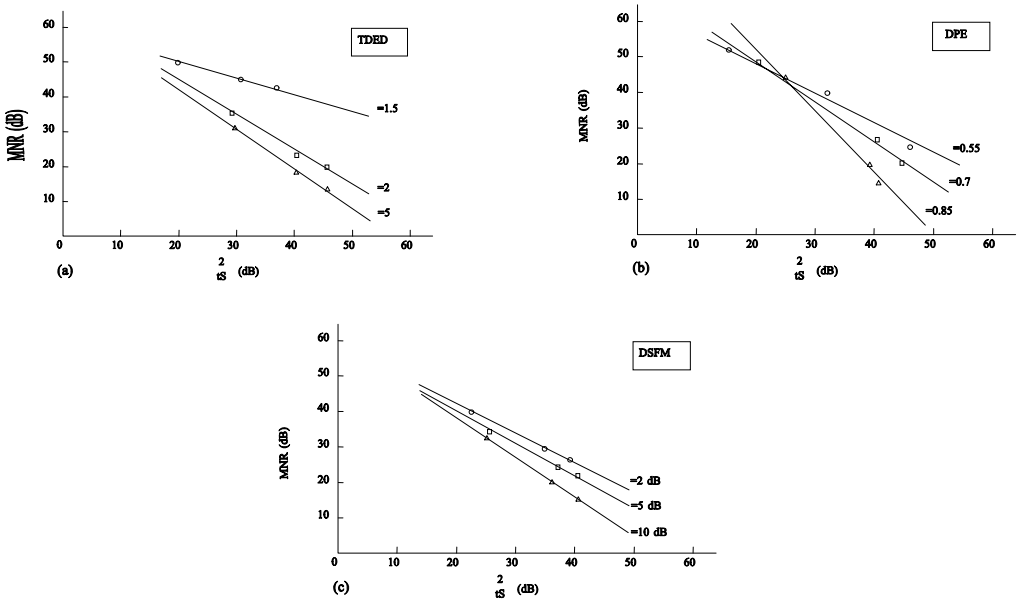

Fig. 8. Variance of the Spectral Substitution error in relation to MNR for the different FS criteria (JZZ music piece).

For the above tests, firstly the Perceptual Entropy reduction (compared to the PAC methods) and the effect on the audio quality (measured via Segmental SNR and MNR measures) was assessed, for different values of the parameters λ and ξ, for all types of the music and for all FS criteria. From these experimental results, and as is shown in Fig. 6, it was found that the PE reduction and audio quality measures are linearly related, independently to the FS criterion employed. Also, as is shown from comparison of the Fig.6a and 6b, an increase of parameter ξ causes a linear increase of the PE reduction and a linear decrease of the sound quality (MNR measure), for all FS criteria. Finally from these tests, a first indication of the superior FS criterion has appeared, since the DPE criterion has the best behaviour, allowing optimum PE reduction and good audio quality. From these tests, the optimum value for the parameter ξ was found to lie between 0.1 and 0.3.

A clearer indication of the performance of the different FS criteria is shown in Fig.7, where the PE reduction and quality results are shown for the complete database, and for the optimal values of ξ and λ. This figure confirms the superiority of the DPE criterion, with second best, the DSFM criterion. Further tests have confirmed that the optimum value for the parameter λ, when the DPE criterion has been selected, was $0.6 < \lambda < 0.7$.

It is also significant that for all criteria, the amount of variance due to Spectral Substitution (equation (19)) was found to be linearly related to the audio quality, as is shown in Fig.8.

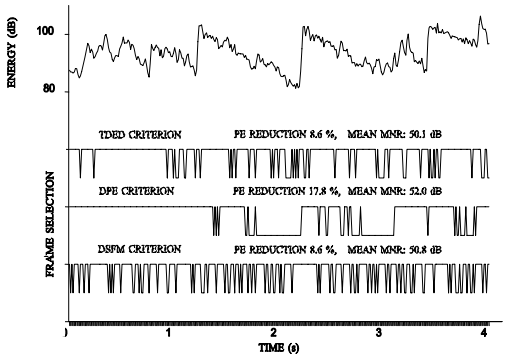

Fig. 9. Reference Frame allocation along a music piece (JZZ), for the different FS criteria.

Nevertheless, the resulting audio quality has also significant variations across each audio track, and this is also related to the specific FS criterion. A good illustration of this is given by Fig. 9, where the allocation between reference and simple frames is shown, along the time evolution of a music piece (JZZ), and for each of the 3 FS criteria. In this figure, the number of reference frames allocated by each criterion was kept the same and the corresponding PE reduction and audio quality is also shown.

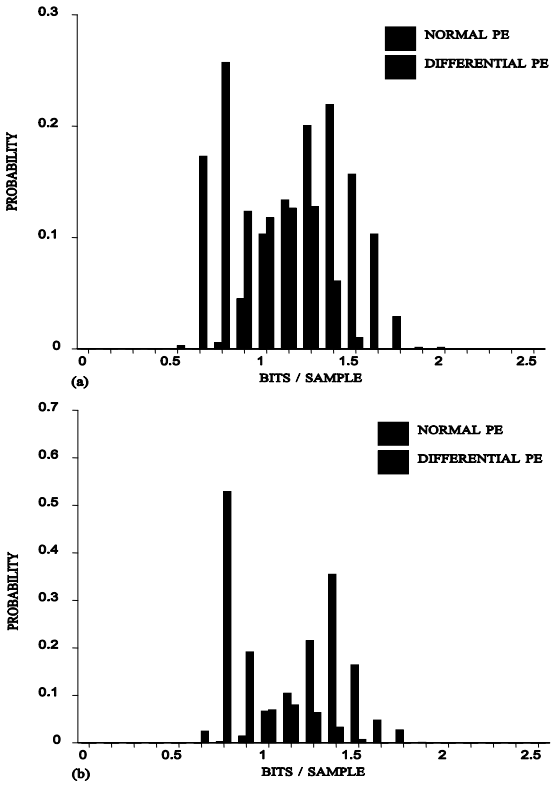

Fig. 10. Perceptual Entropy distribution for the conventional PAC and the proposed DPAC technique, (a) results for big band jazz music (JZZ) and, (b) results for orchestral music (MZT).

Significant data compression gains can be obtained by the application of the DPAC method. Fig. 10 shows a comparison between PE for the same music segments (JZZ and MZT respectively), for the conventional PAC (Normal PE), and the proposed DPAC (Differential

PE) technique. It is clear that the distribution of PE has been shifted towards lower bit/sample values, indicating an average data reduction of about 30% and 35% respectively, compared to the existing PAC method. Similar results were also obtained for the other types of audio material.

Fig. 11 shows the bitrate gain of the DPAC method, expressed as average reduction in the value (in bits) of the audio samples. This diagram is obtained from the difference of DPAC and PAC values, shown in Fig. 10.

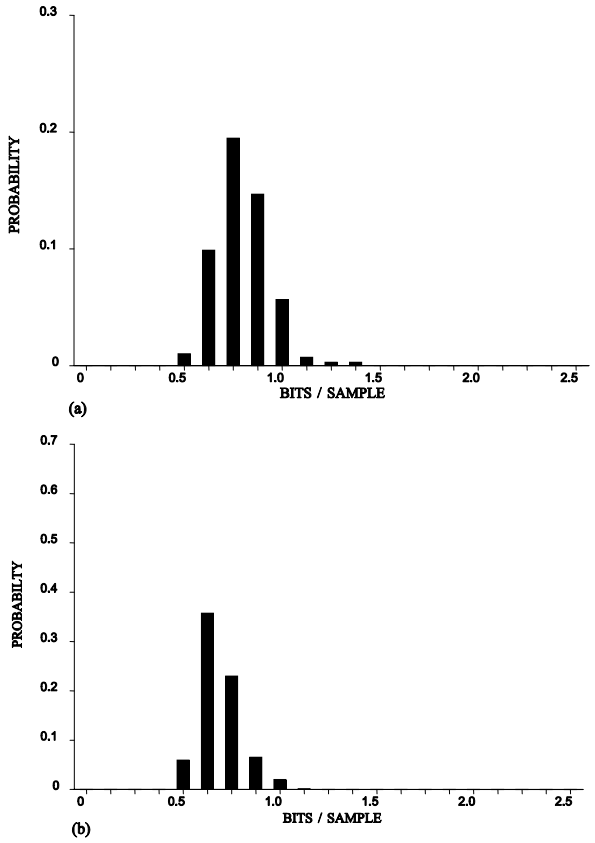

Fig. 11. Probability of reduction in Perceptual Entropy requirements over those achieved by PAC and DPAC, and for the same music segments as those used in Fig. 10.

B. DPAC performance evaluation

In this section, results are presented concerning the audio quality and compression gain performance for the complete DPAC method (i.e. including the quantization and coding stages). Furthermore, these results are compared to the performance of a reference Perceptual Transform Coder (PAC) [8]. It must be also noted here that the DPAC quantization stage is similar to the one used for this PAC system. For these two systems typical results were obtained for the MNR and segmental SNR criterion for different types of music and for fixed decision threshold parameter values, as is shown in Table 2(a). As can be observed bitrate reduction and up to 38% additional (to the PAC method) compression gain was achieved by the DPAC method, as is listed at the right-hand column of this Table. In all cases, full bandwidth (44.1 kHz sampling rate) audio material was employed, which corresponds to a 128 kb/s bitrate for the PAC system. The mean MNR value for the DPAC method was found to depend on both the SCS module and the quantization and coding stages (as was predicted by equation (26)), as is shown in Table 3. Typically, the MNR of the DPAC method was found to be lower than that achieved by the conventional PAC, following a variation over time in a manner depending on the music data, as is shown in Fig. 12, for 128 kb/s bitrate.

Additional comparative tests were carried-out in order to assess the audio quality of DPAC method in respect to PAC method, both set at the same bitrate. These tests were conducted for a range of rates, ranging from low compression (192 kBits/s) to high compression (64 kBits/s). The results, for all music material are given in Tables 2(b), (c) and in Fig. 13. These results indicate that for relative high bitrates, the DPAC method degrades audio quality, whereas at low bitrates, the DPAC method improves the quality in the cases of audio material which has some inherent regular pattern (e.g. for pop-music (POP), jazz (JZZ), harpsichord (HRP) or speech (SPC)). In other cases however, especially for classical music with variable character, e.g. the Mozart Symphony piece (MZT) and the Ravel orchestral piece (RVL), the technique performed worst than the PAC method, although the degree of such degradation was smaller for lower bitrates than for the higher ones.

Table II
Typical comparative performance results for the PAC and proposed DPAC method. The PAC system bitrate was fixed at 128 kb/s (for 44.1 kHz sampling rate audio), whereas the DPAC bitrate was varying with the audio data, resulting to the additional (to PAC's) compression gain shown.

FILE CODE	PAC		DPAC			
	MEAN OVERALL MNR (dB)	SEGM. SNR (dB)	MEAN OVERALL MNR (dB)	SEGM. SNR (dB)	BITRATE (kBits/s)	ADDITIONAL COMPRESSION GAIN (%)
JZZ	25.1	33.5	15.9	22.4	89	30.4
BTH	18.1	34.5	11.0	26.5	94	26.5
MZT	24.4	35.9	10.3	20.5	82	36.1
MSG	23.8	33.6	11.3	19.7	89	30.7
POP	15.9	32.4	6.3	23.2	83	34.8
SPC	16.2	29.2	7.4	20.5	79	38.0

(a)

FILE CODE	PAC		DPAC	
	MEAN OVERALL MNR (dB)	SEGM. SNR (dB)	MEAN OVERALL MNR (dB)	SEGM. SNR (dB)
GSP	19.0	28.2	16.6	23.4
DST	22.5	28.3	15.7	19.6
SVG	17.6	23.1	14.1	10.5
CLM	24.1	31.1	9.9	15.0
RVL	28.0	33.0	14.5	18.6
BSG	6.0	14.8	4.2	19.0
HRP	28.0	31.3	22.1	24.1
TRN	12.0	9.0	8.9	12.1

(b)

FILE CODE	PAC		DPAC	
	MEAN OVERALL MNR (dB)	SEGM. SNR (dB)	MEAN OVERALL MNR (dB)	SEGM. SNR (dB)
GSP	5.1	10.4	2.4	6.8
DST	7.1	8.1	0.5	2.4
SVG	6.2	6.4	0.2	3.6
CLM	7.1	9.1	3.1	7.3
RVL	17.0	24.1	7.1	13.2
BSG	0.2	9.4	2.0	8.0
HRP	11.0	11.3	14.0	17.3
TRN	8.0	8.2	9.0	9.1

(c)

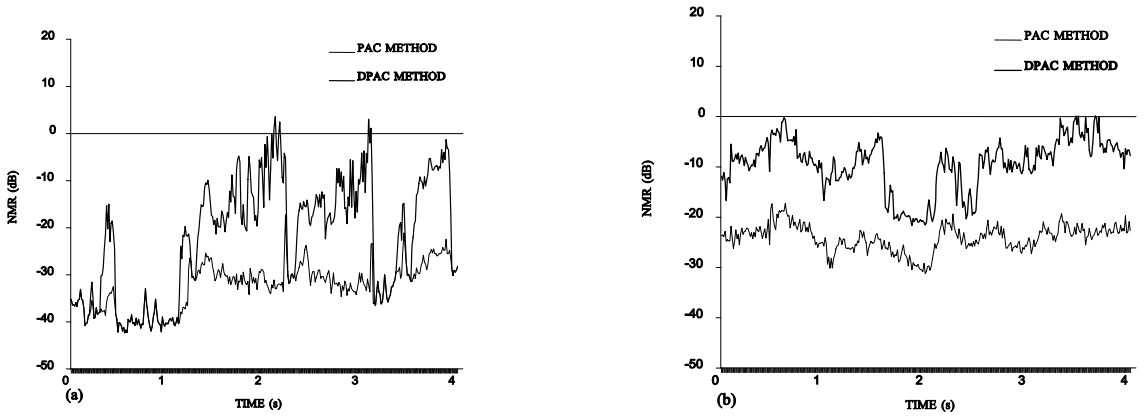

Fig. 12. Overall NMR measures for PAC and DPAC method along the evolution of the same music segment: (a) results for big band jazz music (JZZ) and, (b) results for orchestral music (MZT) (bitrate for both tests: 128 kBits/s).

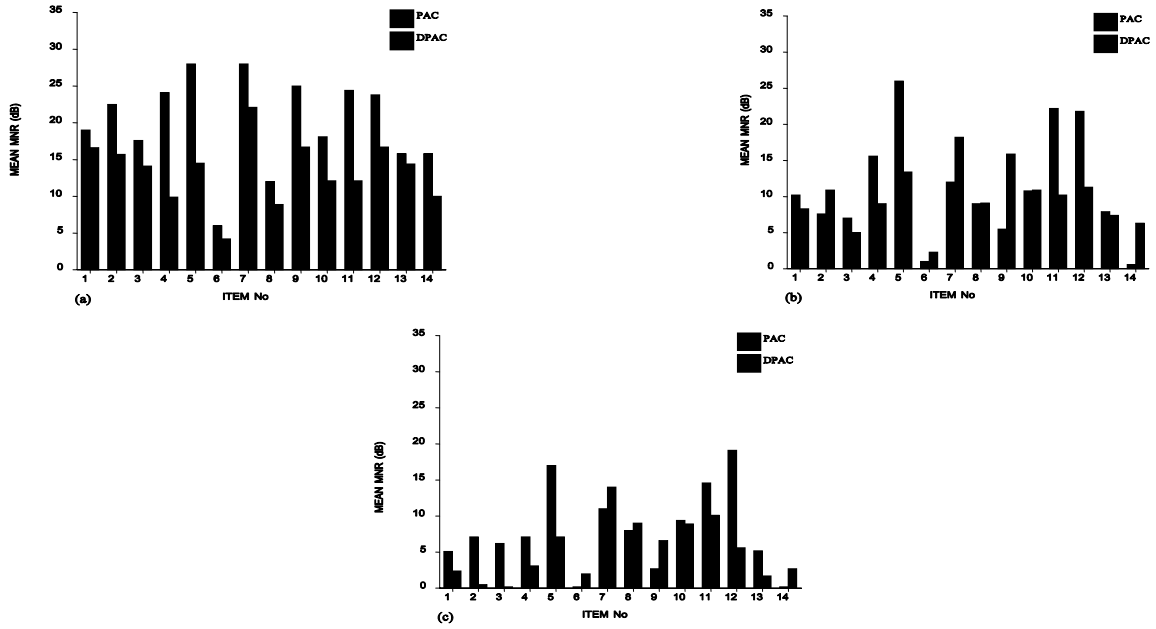

Fig. 13. Comparison in audio performance between PAC and DPAC methods, at similar bitrates: (a) 128 kBits/s, (b) 88 kBits/s, (c) 64 kBits/s.

This aspect of the DPAC compression method is clearly illustrated by Fig. 14, where the difference in audio performance (MNR difference) between DPAC and PAC is plotted for different bitrates and for a variety of audio material. These results indicate that the DPAC method may be better suited for low-bitrate, lower-quality audio applications, and mainly for popular music material, where it appears to have an advantage over the traditional PAC techniques.

Table III
DPAC method NMR Measurements for the Spectral Substitution process and the complete coding process for different types of music.

FILE CODE	JZZ	BTH	MZT	MSG	POP	SPC
MEAN MNR (dB) (SPECTRAL SUBSTITUTION ONLY)	53.4	50.1	44.3	45.2	43.9	49.8
MEAN OVERALL MNR (dB) (SPECTRAL SUBSTITUTION + QUANTIZATION)	15.9	11.0	10.3	11.3	6.3	7.4

II. CONCLUSION

The proposed audio signal coding method (DPAC) aims at lowering the bitrate requirements of Perceptual Audio Coders, by reducing the Perceptual Entropy of such data. This reduction is achieved by utilizing the stationarity of audio signals and by detecting and coding only the audible spectral differences between data frames. The decoding process of the DPAC method employs the spectral components from certain reference frames, kept in a buffer, in order to reconstruct the spectrum for frames of small perceptual significance. The structure of the system is based on modules which can be added to a conventional Perceptual Audio Coder, namely the Frame Selection Module and the Spectral Coefficient Substitution module. In all other respects, the DPAC method uses conventional PAC architecture.

It was shown, that the addition of the above modules can contribute some errors on top of the quantization effects produced by conventional coders. Such errors were found to arise from the Spectral Coefficient Substitution module and were shown to depend on the statistical similarity of audio signal spectral components between successive data frames, i.e. the stationarity of the signal. Given that these errors are strongly dependant on the type of audio material, experiments were conducted using a database of representative music

segments upon which tests were performed in order to assess the performance of the DPAC method.

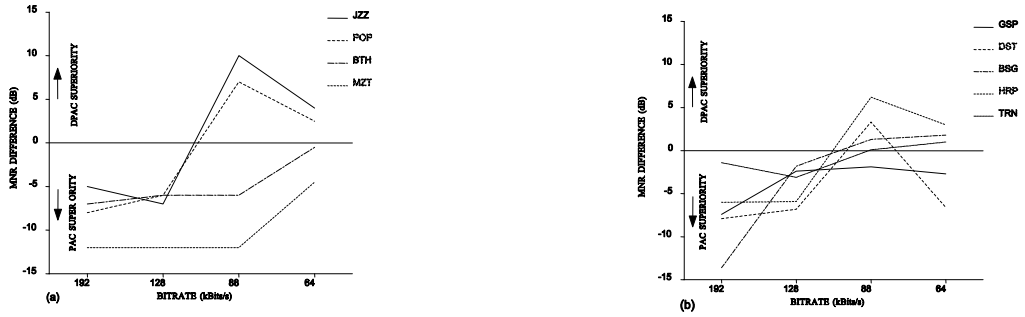

Fig. 14. Audio quality difference between DPAC and PAC method at different bitrates and for different types of audio material: (a) author's database and (b) ITU-R database.

According to these tests, it was found that the DPAC method achieved an average improvement of approximately 30% in audio data compression compared to existing PAC techniques, resulting in an overall bitrate requirement in the range between 60 and 100 kBits/s. In this respect, the initial aims of the method were met, but at the cost of some deterioration in signal quality as was manifested by Noise to Mask Ratio (MNR) measurement which correlate very strongly with the objective impressions of listeners.

A significant finding of the tests was that although the audio quality of the DPAC method was in average worst than the corresponding PAC method for relatively low compression ratios (e.g. at 128 kBits/s), this performance disadvantage was reduced for higher compression ratios, so that at 64 kb/s the DPAC method was superior to PAC, for some types of audio material. This indicates that the DPAC method may be preferable to PAC techniques for lower quality broadcasting and audio transmission applications. These results are encouraging, given that the proposed method introduces a novel approach in the utilization of the signal's stationarity and can be further improved in terms of both compression gain and signal quality. It is significant that it was shown that specific "anchor" frames of audio signal transience can be detected (which were best detected by using a Differential Perceptual Entropy (DPE) criterion) and can be optimally coded by the Perceptual Transform Coder approach. Furthermore, it is also useful that many intermediate audio data frames can be coded at a reduced rate, without prohibitive computational load or hardware complexity, compared to existing Transform Coding systems. Nevertheless, additional work must be undertaken to improve the present method, mainly with respect to the optimal bit allocation and the choice of Spectral Coefficient Substitution in the stationary segments of the audio signals. In this way, it may be possible to reduce the DPAC overall coding error rate (i.e. the combined effect at Spectral Substitution and quantization), using the well established perceptual criteria.

REFERENCES

[1] Jayant N., "Signal Compression: Technology Targets and Research Directions," IEEE Trans. on Selected Areas in Commun. Vol-10, No. 5, pp. 796-818, June 1992.

[2] Plenge G., "DAB - A new sound broadcasting system. Status of the development - Routes to its introduction," EBU Review - Technical No. 246, April 1991.

[3] Zelinski R., Noll P., "Adaptive Transform Coding of Speech Signals," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No.4, pp. 299-309, August 1977.

[4] Jayant N.S., Noll P. "Digital Coding of Waveforms, Principles and Applications to Speech and Video," Prentice-Hall, Englewood Cliffs, New Jersey 1984.

[5] Brandenburg K., Schramm H., "A 16 Bit Transform Coder for Real Time processing of sound signals," Signal Processing II: Theories and Applications", North-Holland 1983, pp. 359-362.

[6] Schroeder M.R., Atal B.S., Hall J.L., "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," J. Acoust. Soc. Amer. Vol. 66, No. 6, pp. 1647-1651, Dec. 1979.

[7] Zwicker E., Fastl H., "Psychoacoustics, Facts and Models," Spinger-Verlag 1990.

[8] Johnston J.D. "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Trans. on Selected Areas in Commun. Vol. 6, No. 2, February 1988.

[9] Brandenburg K., "OCF: A New Coding Algorithm for High Quality Sound Signals," in Proc. ICASSP '87, pp. 141-145, 1987.

[10] Brandenburg K., Herre J., Johnston J.D., Mahieux Y., Schroeder E., "ASPEC: Adaptive Spectral Entropy Coding of High Quality Music Signals," 90th Convention of Audio Eng. Soc., Paris, 1991.

[11] Iwadare M., Sugiyama A., Hazu F., Hirano A., Nishitani T., "A 128 kb/s Hi-Fi Audio CODEC Based on Adaptive Transform Coding with Adaptive Block Size MDCT," IEEE Trans. on Select. Areas in Commun., Vol. 10, No. 1, pp. 138-144, January 1992.

[12] Mahieux Y., Petit J.P., Charbonnier A., "Transform Coding of audio using correlation between successive transform blocks," in Proc. ICASSP '89, pp. 2021-2024, 1989.

[13] Paillard B., Mabilleau P., Morissete S., "Transparent Coding of a Monophonic Audio Signal at 100 Kb/s," 92th Convention of Audio Eng. Soc., Vienna 1992.

[14] Tsutsui K., Suzuki H., Shimoyoshi O., Sonohara M., Akagiri K., Heddle R., "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc," 93rd Convention of Audio Eng. Soc., San Francisco, 1992.

[15] Fesseler P., Thierer G., "A wideband audio codec with a bitrate of 32 kb/s for real time implementation on a single DSP," Signal Processing VI: Theories and Applications, pp. 311-314, 1992.

[16] Brandenburg K., Seitzer D., "Low bit rate coding of high quality digital audio: Algorithms and evaluation of quality," May 1989, AES 7th International Conference, Audio in Digital Times.

[17] Brandenburg K., Stoll G., "The ISO/MPEG Audio Codec: A Generic Standard for Cod-ing of High Quality Digital Audio," 92nd Convention of Audio Eng. Soc. Vienna, 1992.

[18] Theile G., Stoll G., Link M., "Low bit-rate coding of high quality audio signals. An introduction to the MASCAM system," EBU Review - Technical No. 230, pp. 158-181, August 1988.

[19] Wiese D., Stoll G., "Bitrate reduction of high quality audio signal by modelling the ears masking thresholds," 89th Convention of Audio Eng. Soc., Los Angeles 1990.

[20] Stoll G., Dehery Y.F., "MUSICAM: High Bit-Rate Reduction System Family for Different Applications" CCIR Final Study Group Meetings, Geneva, October 1989.

[21] Brandenburg K., Johnston J.D., "Second Generation Perceptual Audio Coding: The Hybrid Coder," 82nd Convention of the Audio Eng. Soc., London, 1987.

[22] Turgeon A., Soumagne J., Mabilleau P., Morissete S., Pailard B., "A Study of Strategies for the Perceptual Coding of Audio Signals," 90th Convention of Audio Eng. Soc., Paris 1991.

[23] Johnston J.D., "Estimation of Perceptual Entropy Using Noise Masking Criteria," in Proc. ICASSP '88, pp. 2524-2527, 1988.

[24] Malvar H.S., "Lapped Transforms for Efficient Transform/Subband Coding", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-38,

No.6, pp. 969-978, June 1990.

[25] Ahmed N., Natarajan T., Rao K.R. "Discrete Cosine Transform," IEEE Trans. on Computers, Vol. C-23, pp. 90-93, January 1974.

[26] Princen J.P, Bradley A.B, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-34 No.5, pp. 1153-1161, Oct.1986.

[27] Cox R.V., "The Design of Uniformly and Nonuniformly Spaced Pseudoquadrature Mirror Filters," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-34, No. 5, pp. 1090-1096, October 1986.

[28] Rothweiler J.H. "Polyphase Quadrature Filters - A New Subband Coding Technique," in Proc. ICASSP '83, pp. 1280-1284, 1983.

[29] ISO/IEC MPEG Audio Coder 11172-3: 1993(E), CH1211 Geneve 20, Switzerland.

[30] Kates J.M., "A time-domain digital cochlear model," IEEE Trans. Signal Processing, Vol. SP-39, No. 12, pp. 2573-2592, December 1991.

[31] Lyon R.F., Mead C.A., "An analog Electronic Cochlea," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-36, No.7, pp. 1119-1134, 1988.

[32] Paillard B., Mabilleau P., Morissete S., "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signal," Journal of AES, Vol. 40, No. 1/2, pp. 21-31, January/February 1992.

[33] Beerends J.G., Stemerdink J.A., "A perceptual audio quality measure based on a psychoacoustic sound representation," Journal of AES, Vol. 40, No. 12, pp. 963-978, December 1992.

[34] Hellman R.P. "Asymmetry of masking between noise and tone," Perception and Psychophysics, Vol. 11, pp. 241-246, 1972.

[35] Paraskevas M., Mourjopoulos J., Kokkinakis G., "Audio Coding based on subjective differences," 94th Conv. of Audio Eng. Soc., Berlin, 1993.

[36] Duhamel P., Mahieux Y., Petit J.P. "A Fast Algorithm for the Implementation of Filter Banks Based on "Time Domain Aliasing Cancellation," in Proc. ICASSP '91, pp. 2209-2212, 1991.

[37] Sporer T., "The use of Multirate Filter Banks for Coding of High Quality Digital Audio," EUSIPCO 1992, pp. 211-214, 1992.

[38] Herre J., Eberlein E., "Evaluation of Concealment Techniques for Compressed Digital Audio," 94th Convention of Audio Eng. Soc., Berlin, 1993.

[39] Herre J., Eberlein E., Schott H., Brandenburg K., "Advanced Audio Measurement System using Psychoacoustic Properties," 92th Convention of Audio Eng. Soc., Vienna, 1992.

[40] ITU-R Document, "Chairman report of the fourth meeting of the TG 10/2," Geneva, November 1993.

Michael Paraskevas was born in Greece in 1964. He received an electrical engineering degree from the University of Patras, Greece in 1989. The topic of his final year project was room equalization using all-pole filter methods. In 1995 he received the PhD degree in audio compression techniques using perceptual model of human hearing. He has also worked at a number of E.C. projects, mainly on speech synthesis and broadcasting applications.

His current research interests are digital signal processing techniques, audio data compression using perceptual encoders, audio data transmition over networks, designing and implementation of networks and active noise canceling from signals. He is a member of the Technical Chamber of Greece and the Audio Engineering Society.

John Mourjopoulos was born in Drama, Greece in 1954. In 1977 he received the B.Sc. degree in engineering from Lanchester Polytechnic (now Coventry University) in the United Kingdom and in 1979 the M.Sc. degree in acoustics from the Institute of Sound and Vibration Research (ISVR), University of Southampton. In 1984 he completed the Ph.D. degree at the same institute, working in the areas of digital signal processing and room acoustics.

He also worked at ISVR as a researcher fellow. Since 1986 he has been with the Wire Communications Laboratory, Electrical & Computer Engineering Department, University of Patras, where he is currently an assistant professor in electroacoustics and head of the Audio Group. His research interests are in the areas of audio and room acoustics equalization, audio digital signal processing, analysis, modeling and coding, as well as speech enhancement and recognition.

He has also been active as musician, participating in recordings and concerts and composing music for film and television. He is a member of the Audio Engineering Society and was chairman of the AES Greek Section for 1993-95. He is also a member of the IEEE and the Helenic Acoustical Society.