EURASIP Journal on Applied Signal Processing 2003:10, 980–992

Progressive Syntax-Rich Coding of Multichannel Audio Sources

Dai Yang
Integrated Media Systems Center and Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089-2564, USA
Email: daiyang@alumni.usc.edu

Hongmei Ai
Integrated Media Systems Center and Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089-2564, USA
Email: aimee@ieee.org

Chris Kyriakakis
Integrated Media Systems Center and Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089-2564, USA
Email: ckyriak@imsc.usc.edu

C.-C. Jay Kuo
Integrated Media Systems Center and Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089-2564, USA
Email: cckuo@sipi.usc.edu

Received 6 May 2002 and in revised form 5 March 2003

Being able to transmit the audio bitstream progressively is a highly desirable property for network transmission. MPEG-4 version 2 audio supports fine grain bit rate scalability in the generic audio coder (GAC). It has a bit-sliced arithmetic coding (BSAC) tool, which provides scalability in the step of 1 Kbps per audio channel. There are also several other scalable audio coding methods, which have been proposed in recent years. However, these scalable audio tools are only available for mono and stereo audio material. Little work has been done on progressive coding of multichannel audio sources. MPEG advanced audio coding (AAC) is one of the most distinguished multichannel digital audio compression systems. Based on AAC, we develop in this work a progressive syntax-rich multichannel audio codec (PSMAC). It not only supports fine grain bit rate scalability for the multichannel audio bitstream but also provides several other desirable functionalities. A formal subjective listening test shows that the proposed algorithm achieves an excellent performance at several different bit rates when compared with MPEG AAC.

Keywords and phrases: multichannel audio, progressive coding, Karhunen-Loéve transform, successive quantization, PSMAC.

1. INTRODUCTION

Multichannel audio technologies have become much more mature these days, partially pushed by the need of the film industry and home entertainment systems. Starting from the monophonic technology, new systems, such as stereophonic, quadraphonic, 5.1 channels, and 10.2 channels, are penetrating into the market very quickly. Compared with the mono or stereo sound, multichannel audio provides end users a more compelling experience and becomes more appealing to music producers. As a result, an efficient coding scheme for multichannel audio storage and transmission is in great demand. Among several existing multichannel audio compression algorithms, Dolby AC-3 and MPEG advanced audio coding (AAC) [1, 2, 3, 4] are two most prevalent perceptual digital audio coding systems. Both of them can provide perceptually indistinguishable audio quality at the bit rate of 64 Kbps/ch.

In spite of their success, they can only provide bitstreams with a fixed bit rate, which is specified during the encoding phase. When this kind of bitstream is transmitted over variable bandwidth networks, the receiver can either successfully decode the full bitstream or ask the encoder to retransmit a bitstream with a lower bit rate. The best solution to this problem is to develop a scalable compression algorithm to transmit and decode the audio content in an embedded manner. To be more specific, a bitstream generated by a scalable coding scheme consists of several partial bitstreams, each of which can be decoded on their own in a meaningful way. Therefore, transmission and decoding of a subset of the total bitstream will result in a valid decodable signal at a lower bit rate and quality. This capability offers a significant advantage in transmitting contents over networks with variable channel capacity and heterogeneous access bandwidth.

MPEG-4 version 2 audio coding supports fine grain bit rate scalablility [5, 6, 7, 8, 9] in its generic audio coder (GAC). It has a bit-sliced arithmetic coding (BSAC) tool, which provides scalability in the step of 1 Kbps per audio channel for mono or stereo audio material. Several other scalable mono or stereo audio coding algorithms [10, 11, 12] were proposed in recent years. However, not much work has been done on progressive coding of multichannel audio sources. In this work, we propose a progressive syntax-rich multichannel audio codec (PSMAC) based on MPEG AAC. In PSMAC, the interchannel redundancy inherent in original physical channels is first removed in the preprocessing stage by using the Karhunen-Loéve transform (KLT). Then, most coding blocks in the AAC main profile encoder are employed to generate spectral coefficients. Finally, a progressive transmission strategy and a context-based QM-coder are adopted to obtain the fully quality-scalable multichannel audio bitstream. The PSMAC system not only supports fine-grain bit rate scalability for the multichannel audio bitstream, but also provides several other desirable functionalities, such as random access and channel enhancement, which have not been supported by other existing multichannel audio codecs (MAC).

Moreover, compared with the BSAC tool provided in MPEG-4 version 2 and most of the other scalable audio coding tools, a more sophisticated progressive transmission strategy is employed in PSMAC. PSMAC does not only encode spectral coefficients from MSB to LSB and from low to high frequency so that the decoder can reconstruct these coefficients more and more precisely with an increasing bandwidth as the receiver collects more and more bits from the bitstream, but also utilizes the psychoacoustic model to control the subband transmission sequence so that the most sensitive frequency area is more precisely reconstructed. In this way, bits used to encode coefficients in those nonsensitive frequency area can be saved and used to encode coefficients in the sensitive frequency area. As a result of this subband selection strategy, a perceptually more appealing audio can be reconstructed by PSMAC, especially at very low bit rates such as 16 Kbps/ch. The side information required to encode the subband transmission sequence is carefully handled in our implementation so that the overall overhead will not have significant impact on the audio quality even at very low bit rates. Note that Shen et al. [12] proposed a subband selection rule to achieve progressive coding. However, Shen's scheme demands a large amount of overhead in coding the selection order.

Experimental results show that, when compared with MPEG AAC, the decoded multichannel audio generated by the proposed PSMAC's mask-to-noise-ratio (MNR) progressive mode has comparable quality at high bit rates, such as 64 Kbps/ch or 48 Kbps/ch, and much better quality at low bit rates, such as 32 Kbps/ch or 16 Kbps/ch. We also demonstrate that our PSMAC can provide better quality of single-channel audio when compared with MPEG-4 version 2 GAC at several different bit rates.

The rest of the paper is organized as follows. Section 2 gives an overview of the proposed design. Section 3 briefly introduces how interchannel redundancy can be removed via the KLT. Sections 4 and 5 describe progressive quantization and subband selection blocks in our system, respectively. Section 6 presents the complete compression system. Experimental results are shown in Section 7. Finally, conclusion remarks are given in Section 8.

2. PROFILES OF PROPOSED PROGRESSIVE SYNTAX-RICH AUDIO CODEC

In the proposed progressive syntax-rich codec, the following three user-defined profiles are provided.

(1) The MNR progressive profile. If the flag of this profile is on, it should be possible to decode the first n bytes of the bitstream per second, where n is a user-specified value or a value that the current network parameters allowed.

(2) The random access profile. If the flag of this profile is present, the codec will be able to independently encode a short period of audio more precisely than other periods. It allows users to randomly access a certain part of audio that is more of interest to end users.

(3) The channel enhancement profile. If the flag of this profile is on, the codec will be able to independently encode an audio channel more precisely than other channels. Either these channels are of more interest to end users or the network situation does not allow the full multichannel audio bitstream to be received on time.

Figure 1 illustrates a simple example of three user-defined profiles. Among all profiles, the MNR progressive profile is the default one. In the other two profiles, that is, the random access and the channel enhancement, the MNR progressive feature is still provided as a basic functionality and the decoding of the bitstream can be stopped at any arbitrary point. With these three profiles, the proposed codec can provide a versatile set of functionalities desirable in variable bandwidth network conditions with different user access bandwidth.

3. INTERCHANNEL DECORRELATION

For a given time instance, removing interchannel redundancy would result in a significant bandwidth reduction. This can be done via an orthogonal transform $MV = U$, where V and U denote the vector whose n elements are samples in original channels and transformed channels, respectively. Among several commonly used transforms, including the discrete cosine transform (DCT), the Fourier transform

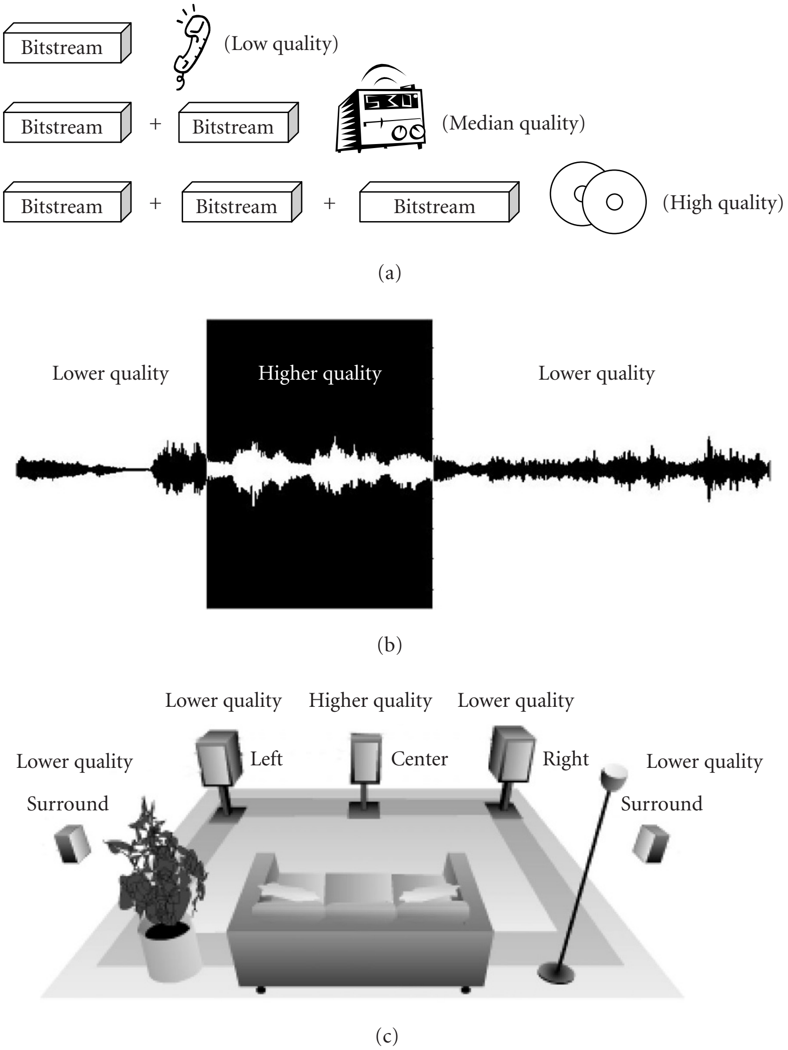


Figure 1: Illustration of three user-defined profiles: (a) the MNR progressive profile, (b) the random access profile, and (c) the channel enhancement with the enhanced center channel.

(FT), and the KLT, the signal-dependent KLT is adopted in the preprocessing stage because it is theoretically optimal in decorrelating signals across channels. If M is the KLT matrix, we call the corresponding transformed channels eigenchannels. Figure 2 illustrates how KLT is performed on multichannel audio signals, where the columns of the KLT matrix are composed of eigenvectors calculated from the covariance matrix C_V associated with original multichannel audio signals V.

Suppose that an input audio signal has n channels, then the covariance of KL transformed signals is

$$E[\bar{U}\bar{U}^T] = E[(M\bar{V})(M\bar{V})^T] = ME[\bar{V}\bar{V}^T]M^T$$
$$= MC_VM^T = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \qquad (1)$$

where $\bar{X}$ ($X = U, V$) represents the mean-removed signal of X, and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are eigenvalues of C_V. Thus, the transform produces statistically decorrelated channels in the sense of having a diagonal covariance matrix for transformed signals. Another property of KLT, which can be used in the reconstruction of audio of original channels, is that the inverse transform matrix of M is equal to its transpose. Since C_V is real and symmetric, the matrix formed by normalized eigenvectors is orthonormal. Therefore, we have $V = M^TU$ in reconstruction. From KL expansion theory [13], we know that selecting eigenvectors associated with the largest eigenvalues can minimize the error between original and reconstructed channels. This error will go to zero if all eigenvectors are used. KLT is thus optimum in the least square error sense.

The KLT preprocessing method was demonstrated to improve the multichannel audio coding efficiency in our previous work [14, 15, 16]. After the preprocessing stage, signals in these relatively independent channels called eigenchannels are further processed.

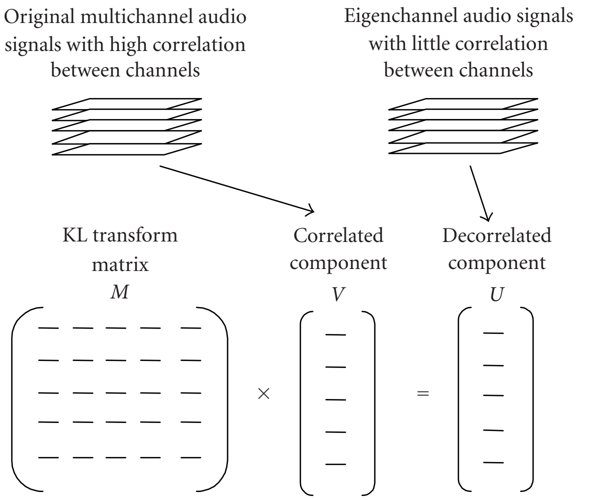

Original multichannel audio signals with high correlation between channels
Eigenchannel audio signals with little correlation between channels
KL transform matrix
M
Correlated component
V
Decorrelated component
U
×
=

Figure 2: Interchannel decorrelation via KLT.

4. SCALABLE QUANTIZATION AND ENTROPY CODING

The major difference between the proposed progressive audio codec and other existing nonprogressive audio codecs such as AAC lies in the quantization block and the entropy coding block. The dual iteration loop used in AAC to calculate the quantization step size for each frame and each channel coefficients is replaced by a progressive quantization block. The Huffman coding block used in the AAC to encode quantized data is replaced by a context-based QM-coder. This will be explained in detail below.

4.1. Successive approximation quantization (SAQ)

The most important component of the quantization block is called successive approximation quantization (SAQ). The SAQ scheme, which is adopted by most embedded wavelet coders for progressive image coding, is crucial to the design of embedded coders. The motivation for successive approximation is built upon the goal of developing an embedded code that is in analogy to find an approximation of binary representation of a real number [17]. Instead of coding every quantized coefficient as one symbol, SAQ processes the bit representation of coefficients via bit layers sliced in the order of their importance. Thus, SAQ provides a coarse-to-fine, multiprecision representation of the amplitude information. The bitstream is organized such that a decoder can immediately start reconstruction based on the partially received bitstream. As more and more bits are received, more accurate coefficients and higher quality multichannel audio can be reconstructed.

SAQ sequentially applies a sequence of thresholds $T_0, T_1, \ldots, T_{N+1}$ for refined quantization, where these thresholds are chosen such that $T_i = T_{i-1}/2$. The initial threshold T_0 is selected such that $|C(i)| < 2T_0$ for all transformed coefficients in one subband, where $C(i)$ represents the ith spectral coefficient in the subband. To implement SAQ, two separate lists, the dominant list and the subordinate list, are maintained both at the encoder and the decoder. At any point of the process, the dominant list contains the coordinates of those coefficients that have not yet been found to be significant, while the subordinate list contains magnitudes of those coefficients that have been found to be significant. The process that updates the dominate list is called the significant pass, and the process that updates the subordinate list is called the refinement pass.

In the proposed algorithm, SAQ is adopted as the quantization method for each spectral coefficient within each subband. This algorithm (for the encoder part) is listed below.

Successive approximation quantization (SAQ) algorithm

(1) Initialization: For each subband, find out the maximum absolute value $C_{\max}$ for all coefficients $C(i)$ in the subband, and set the initial quantization threshold to be $T_0 = C_{\max}/2 + \Delta$, where Δ is a small constant.

(2) Construction of the significant map (significance identification). For each $C(i)$ contained in the dominant list, if $|C(i)| \geq T_k$, where T_k is the threshold of the current layer (layer k), add i to the significant map, remove i from the dominant list, and encode it with "1s," where "s" is the sign bit. Moreover, modify the coefficient's value to

$$C(i) \longleftarrow \begin{cases} C(i) - 1.5T_k, & \forall\, C(i) > 0, \\ C(i) + 1.5T_k, & \text{otherwise.} \end{cases} \tag{2}$$

(3) Construction of the refinement map (refinement). For each $C(i)$ contained in the significant map, encode the bit at layer k with a refinement bit "D" and change the value of $C(i)$ to

$$C(i) \longleftarrow \begin{cases} C(i) - 0.25T_k, & \forall\, C(i) > 0, \\ C(i) + 0.25T_k, & \text{otherwise.} \end{cases} \tag{3}$$

(4) Iteration. Set $T_{k+1} = T_k/2$ and repeat steps (2)–(4) for $k = 0, 1, 2, \ldots$.

At the decoder side, the decoder performs similar steps to reconstruct coefficients' values. Figure 3 gives a simple example to show how the decoder reconstructs a single coefficient after one significant pass and one refinement pass. As illustrated in this figure, the magnitude of this coefficient is recovered to 1.5 times of the current threshold T_k after the significant pass, and then refined to $1.5T_k - 0.25T_k$ after the first refinement pass. As more refinement steps follow, the magnitude of this coefficient will approach its original value gradually.

4.2. Context-based QM-coder

The QM-coder is a binary arithmetic coding algorithm designed to encode data formed by a binary symbol set. It was the result of the effort by JPEG and JBIG committees, in which the best features of various arithmetic coders are integrated. The QM-coder is a lineal descendent of the Q-coder, but significantly enhanced by improvements in the two building blocks, that is, interval subdivision and

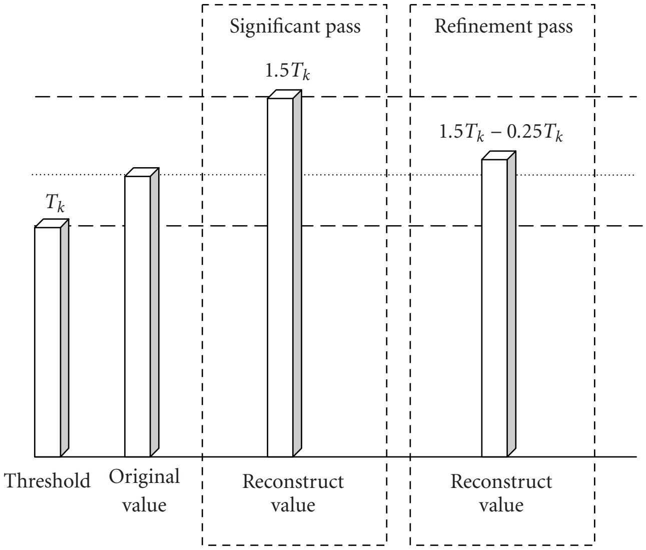


Figure 3: An example to show how the decoder reconstructs a single coefficient after one significant pass and one refinement pass.

probability estimation [18]. Based on the Bayesian estimation, a state-transition table, which consists of a set of rules to estimate the statistics of the bitstream depending on the next incoming symbols, can be derived. The efficiency of the QM-coder can be improved by introducing a set of context rules. The QM arithmetic coder achieves a very good compression result if the context is properly selected to summarize the correlation between coded data.

Six classes of contexts are used in the proposed embedded audio codec as shown in Figure 4. They are the general context, the constant context, the subband significance context, the coefficient significance context, the coefficient refinement context, and the coefficient sign context. The general context is used in the coding of the configuration information. The constant context is used to encode different channel header information. As their names suggest, the subband significance context, the coefficient significance context, the coefficient refinement context, and the coefficient sign context are used to encode the subband significance, coefficient significance, coefficient refinement, and coefficient sign bits, respectively. These contexts are adopted because different classes of bits may have different probability distributions. In principle, separating their contexts should increase the coding performance of the QM-coder.

5. CHANNEL AND SUBBAND TRANSMISSION STRATEGY

5.1. *Channel selection rule*

In the embedded MAC, we should put the most important bits (in the rate-distortion sense) to the cascaded bitstream first so that the decoder can reconstruct the optimal quality of multichannel audio given a fixed number of bits received. Thus, the importance of channels should be determined for an appropriate order of the bitstream.

The first instinct about the metric of channel importance would be the energy of the audio signal in each channel. However, this metric does not work well in general. For example, for some multichannel audio sources, especially for those that have been reproduced artificially in a music studio, the side channel which does not normally contain the main melody may even have a larger energy than the center channel. Based on our experience with multichannel audio, loss or significant distortion of the main melody in the center channel would be much more annoying than loss of melodies in side channels. In other words, the location of channels also plays an important role. Therefore, for a regular 5.1 channel configuration, the order of channel importance from the largest to the least should be

(1) center channel,
(2) left and right (L/R) channel pair,
(3) left surround and right surround (Ls/Rs) channel pair,
(4) low-frequency channel.

Between channel pairs, their importance can be determined by their energy values. This rule is adopted in our experiments, given in Section 7.

After KLT, eigenchannels are no longer the original physical channels, and sounds in different physical channels are mixed in every eigenchannel. Thus, spatial dependency of eigenchannels is less trivial. We observe from experiments that although it is true that one eigenchannel may contain sounds from more than one original physical channel, there still exists a close correspondence between eigenchannels and physical channels. To be more precise, audio of eigenchannel 1 would sound similar to that of the center channel, audio of eigenchannels 2 and 3 would sound similar to that of the L/R channel pair, and so forth. Therefore, if eigenchannel 1 is lost in transmission, we would end up with a very distorted center channel. Moreover, it happens that, sometimes, eigenchannel 1 may not the channel with a very large energy and could be easily discarded if the channel energy is adopted as the metric of channel importance. Thus, the channel importance of eigenchannels should be similar to that of physical channels, that is, eigenchannel 1 corresponding to the center channel, eigenchannels 2 and 3 corresponding to the L/R channel pair, and eigenchannels 4 and 5 corresponding to the Ls/Rs channel pair. Within each channel pair, the importance is still determined by their energy values.

5.2. *Subband selection rule*

In principle, any quality assessment of an audio channel can be either performed subjectively by employing a large number of expert listeners or done objectively by using an appropriate measuring technique. While the first choice tends to be an expensive and time-consuming task, the use of objective measures provides quick and reproducible results. An optimal measuring technique would be a method that produces the same results as subjective tests while avoiding all problems associated with the subjective assessment procedure. Nowadays, the most prevalent objective measurement is the MNR technique, which was first introduced by

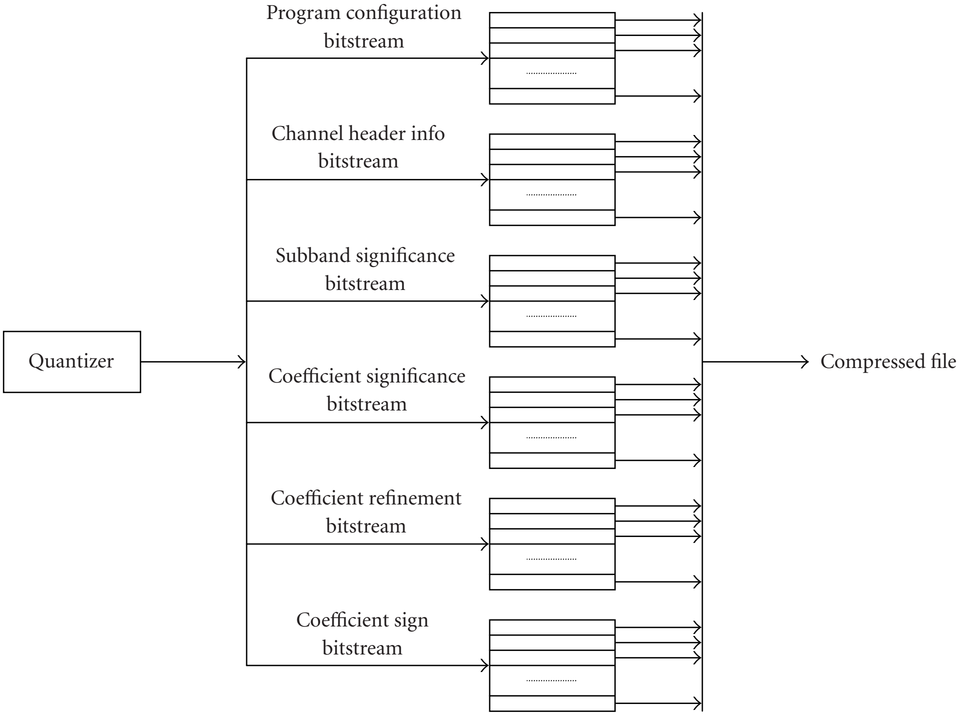

Program configuration bitstream
Channel header info bitstream
Subband significance bitstream
Coefficient significance bitstream
Coefficient refinement bitstream
Coefficient sign bitstream
Quantizer
Compressed file

FIGURE 4: The adopted context-based QM-coder with six classes of contexts.

Brandenburg [19] in 1987. It is the ratio of the masking threshold with respect to the error energy. In our implementation, the masking is calculated from the general psychoacoustic model of the AAC encoder. The psychoacoustic model calculates the maximum distortion energy which is masked by the signal energy, and outputs the signal to mask ratio (SMR).

A subband is masked if the quantization noise level is below the masking threshold, so the distortion introduced by the quantization process is not perceptible to human ears. As discussed earlier, SMR represents the human auditory response to the audio signal. If SNR of an input audio signal is high enough, the noise level will be suppressed below masking threshold, and the quantization distortion will not be perceived. Since SNR can be easily calculated by

$$\mathrm{SNR} = \frac{\sum_i \left| S_{\mathrm{original}}(i) \right|^2}{\sum_i \left| S_{\mathrm{original}}(i) - S_{\mathrm{reconstruct}}(i) \right|^2}, \tag{4}$$

where $S_{\mathrm{original}}(i)$ and $S_{\mathrm{reconstruct}}(i)$ represent the ith original and the ith reconstructed audio signal value, respectively, thus, MNR is just the difference between SNR and SMR (in dB) or

$$\mathrm{SNR} = \mathrm{MNR} + \mathrm{SMR}. \tag{5}$$

A side benefit of the SAQ technique is that an operational rate versus distortion plot (or, equivalently, an operational rate versus the current MNR value) for the coding algorithm can be computed online.

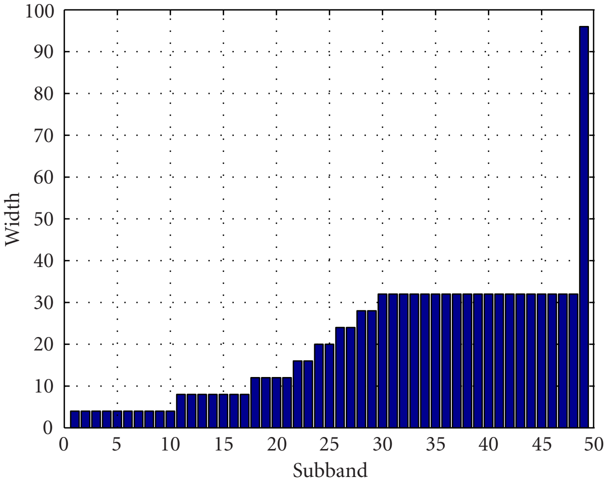


FIGURE 5: Subband width distribution.

The basic ideas behind choosing the subband selection rules are simple. They are presented as follows:

(1) the subband with a better rate deduction capability should be chosen earlier to enhance the performance;
(2) the subband with a smaller number of coefficients should be chosen earlier to reduce the computational complexity if the rate reduction performances of two subbands are close.

The first rule implies that we should allocate more bits to those subbands with larger SMR values (or smaller MNR

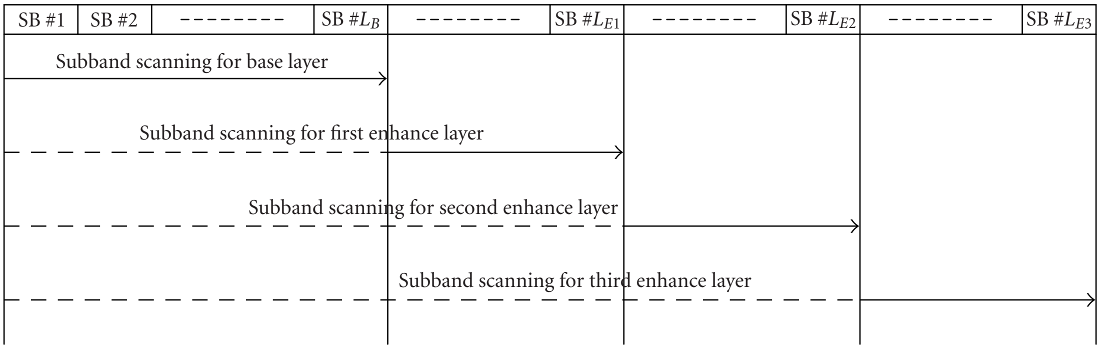

SB #1
SB #2
SB #L_B
SB #L_{E1}
SB #L_{E2}
SB #L_{E3}
Subband scanning for base layer
Subband scanning for first enhance layer
Subband scanning for second enhance layer
Subband scanning for third enhance layer

Figure 6: Illustration of the subband scanning rule, where the solid line with an arrow means that all subbands inside this area are scanned, and the dashed line means that only those nonsignificant subbands inside the area are scanned.

values). In other words, we should send out bits belonging to those subbands with larger SMR values (or smaller MNR values) first. The second rule tells us how to decide the subband scanning order. As we know about the subband[1] formation in MPEG AAC, the number of coefficients in each subband is nondecreasing with the increase of the subband number. Figure 5 shows the subband width distribution used in AAC for 44.1 kHz and 48 kHz sampling frequencies and long block frames. Thus, a sequential subband scanning order from the lowest number to the highest number is adopted in this work.

In order to save bits, especially at very low bit rates, only information corresponding to lower subbands will be sent into the bitstream at the first layer. When the number of layers increases, more and more subbands will be added. Figure 6 shows how subbands are scanned for the first several layers. At the base layer, the priority is given to lower-frequency signals so that only subbands numbered up to L_B will be scanned. As the information of enhancement layers is added to the bitstream, the subband scanning upper limit increases (as indicated by values of L_{E1}, L_{E2}, and L_{E3} as shown in Figure 6) until it reaches the effective psychoacoustic upper bound of all subbands N. In our implementation, we choose $L_{E3} = N$, which means that all subbands are scanned after the third enhance layer. Here, the subband scanning upper limits in different layers, that is, L_B, L_{E1}, and L_{E2}, are empirically determined values that provide a good coding performance.

A dual-threshold coding technique is proposed in this work. One of the thresholds is the MNR threshold, which is used in subband selection. The other is the magnitude threshold, which is used for coefficients quantization in each selected subband. A subband that has its MNR value smaller than the current MNR threshold is called the significant subband. Similar to the SAQ process for coefficient quantization, two lists, that is, the dominant subband list and the subordinate subband list, are maintained in the encoder and the decoder, respectively. The dominant subband list contains the indices of those subbands that have not become significant yet, and the subordinate subband list contains the indices of those subbands that have already become significant. The process that updates the subband dominant list is called the subband significant pass, and the process that updates the subband subordinate list is called the subband refinement pass.

Different coefficient magnitude thresholds are maintained in different subbands. Since we would like to deal with the most important subbands first and get the best result with only a little amount of information from the resource, and, since sounds in different subbands have different impacts on human ears according to the psychoacoustic model, it is worthwhile to consider each subband independently rather than all subbands in one frame simultaneously.

We summarize the subband selection rule below.

(1) MNR threshold calculation. Determine empirically the MNR threshold value $T_{i,k}^{\text{MNR}}$ for channel i at layer k. Subbands with smaller MNR value at the current layer are given higher priority.

(2) Subband dominant pass. For those subbands that are still in the dominant subband list, if subband j in channel i has the current MNR value $\text{MNR}_{i,j}^{k} < T_{i,k}^{\text{MNR}}$, add subband j of channel i into the significant map, remove it from the dominant subband list, and send 1 to the bitstream, indicating that this subband is selected. Then, apply SAQ to coefficients in this subband. For subbands that have $\text{MNR}_{i,j}^{k} \geq T_{i,k}^{\text{MNR}}$, send 0 to the bitstream, indicating that this subband is not selected in this layer.

(3) Subband refinement pass. For a subband already in the subordinate list, perform SAQ to coefficients in the subband.

(4) MNR values update. Recalculate and update MNR values for selected subbands.

(5) Repeat steps (1)–(4) until the bitstream meets the target rate.

Figure 7 gives a simple example of the subband selection rule. Suppose that, at layer k, channel i has the MNR threshold equal to $T_{i,k}^{\text{MNR}}$. In this example, among all scanned

[1] The term "subband" defined in this paper is equivalent to the "scale factor band" implemented in MPEG AAC.

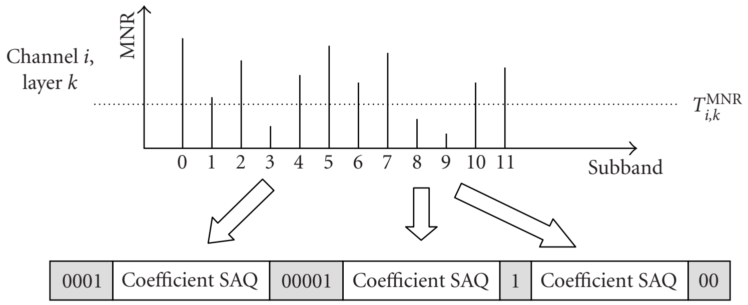


Figure 7: An example of the subband selection rule.

subbands, that is, subbands 0 to 11, only subbands 3, 8, and 9 have their current MNR values smaller than $T_{i,k}^{\mathrm{MNR}}$. Therefore, according to rule (2), three 0 bits and one 1 bit are first sent into the bitstream indicating nonsignificant subbands 0, 1, and 2 and significant subband 3. These subband selecting bits are represented in the left-most shaded area in Figure 7. Similarly, subband selecting bits for subbands 4 to 11 are illustrated in the rest of shaded areas. Coefficients SAQ bits of significant subbands are sent immediately after each significant subband bit as shown in this example.

6. COMPLETE DESCRIPTION OF PSMAC

The block diagram of a complete PSMAC encoder is shown in Figure 8. The perceptual model, the filter bank, the temporal noise shaping (TNS), and the intensity blocks in our progressive encoder are the same as those in the AAC main profile encoder. The interchannel redundancy removal block via KLT is implemented after the input audio signals are transformed into the modified discrete cosine transform (MDCT) domain. Then, a dynamic range control block follows to avoid any possible data overflow in later compression stages. Masking thresholds are then calculated in the perceptual model based on the KL transformed signals. The progressive quantization and lossless coding parts are finally used to construct the compressed bitstream. The information generated at the first several coding blocks will be sent into the bitstream as the overhead.

Figure 9 provides more details of the progressive quantization block. The channel and the subband selection rules are used to determine which subband in which channel should be encoded at this point, and then coefficients within this selected subband will be quantized via SAQ. The user-defined profile parameter is used for the syntax control of the channel selection and the subband selection. Finally, based on several different contexts, the layered information together with all overhead bits generated during previous coding blocks will be losslessly coded by using the context-based QM-coder.

The encoding process performed by using the proposed algorithm will stop when the bit budget is exhausted. It can cease at any time, and the resulting bitstream contains all lower rate coded bitstreams. This is called the full embedded property. The capability to terminate the decoding of an embedded bitstream at any specific point is extremely useful in a coding system that is either rate constrained or distortion constrained.

7. EXPERIMENTAL RESULTS

The proposed PSMAC system has been implemented and tested. The basic audio coding blocks [1] inside the MPEG AAC main profile encoder, including the psychoacoustic model, filter bank, TNS, and intensity/coupling, are still adopted. Furthermore, an interchannel removal block, a progressive quantization block, and a context-based QM-coder block are added to construct the PSMAC.

Two types of experimental results are shown in this section. One is measured by an objective metric, that is, the MNR, and the other is measured in terms of a subjective metric, that is, listening test score. It is worthwhile to mention that the coding blocks adopted from AAC have not been modified to improve the performance of the proposed PSMAC for fair comparison. Moreover, test audio that produces the worst performance by the MPEG reference code was not selected in the experiment.

7.1. Results using MNR measurement

Two multichannel audio materials are used in this experiment to compare the performance of the proposed PSMAC algorithm with MPEG AAC [1] main profile codec. One is a one-minute long ten-channel[2] audio material called “Messiah,” which is a piece of classical music recorded live in a real concert hall. Another one is an eight-second long five-channel[3] music called “Herre,” which is a piece of pop music and was used in the MPEG-2 AAC standard (ISO/IEC 13818-7) conformance work.

7.1.1. MNR progressive mode

The performance comparison of MPEG AAC and the proposed PSMAC for the normal MNR progressive mode are

[2] The ten channels include center (C), left (L), right (R), left wide (Lw), right wide (Rw), left high (Lh), right high (Rh), left surround (Ls), right surround (Rs), and back surround (Bs).

[3] The five channels include C, L, R, Ls, and Rs.

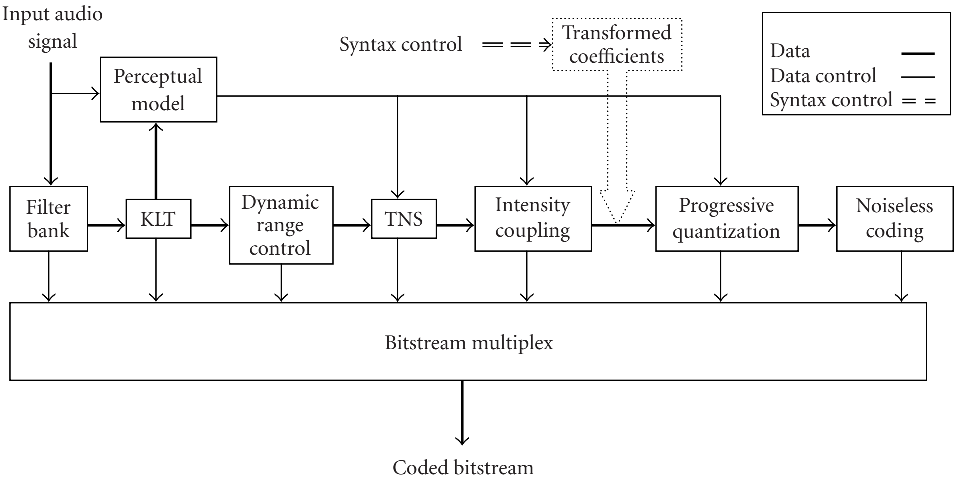


Figure 8: The block diagram of the proposed PSMAC encoder.

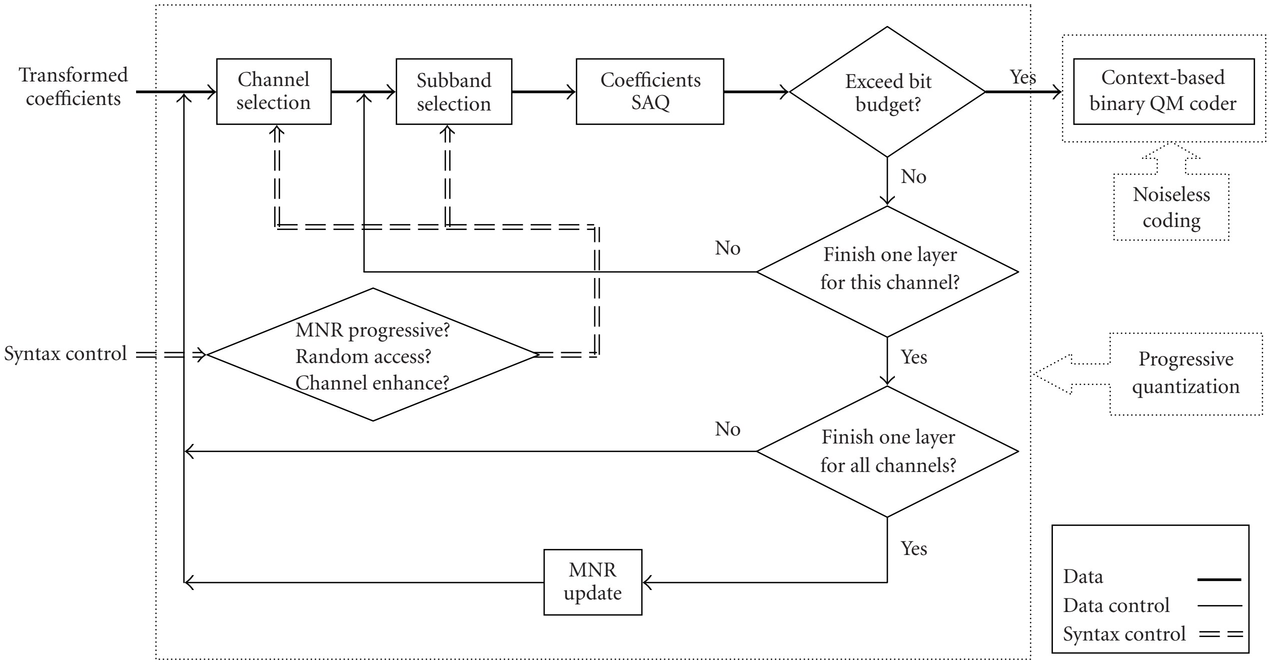


Figure 9: Illustration of the progressive quantization and lossless coding blocks.

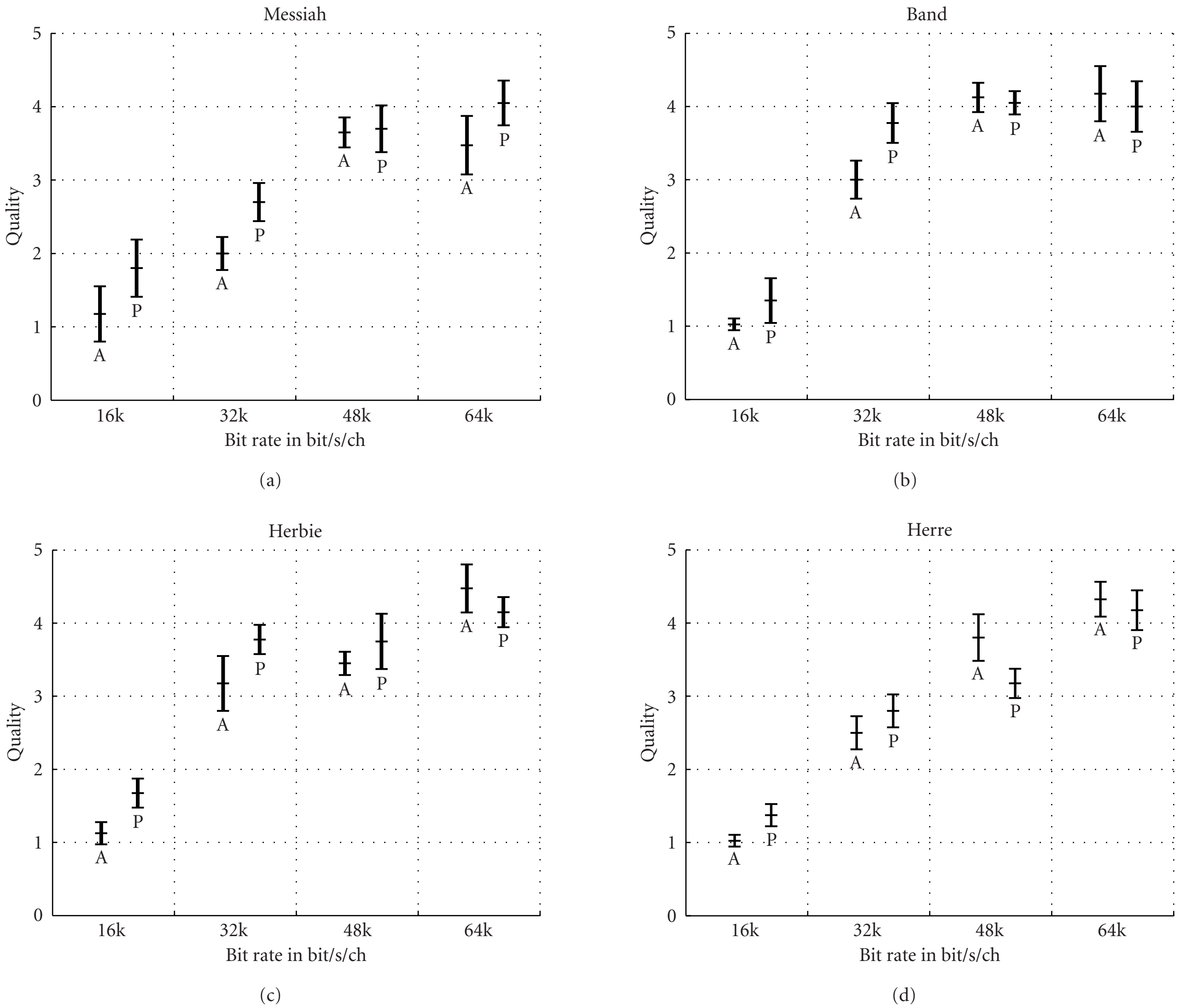
shown in Table 1. The average MNR shown in the table is calculated by

$$\text{mean MNR}_{\text{subband}} = \frac{\sum_{\text{channel}} \text{MNR}_{\text{channel, subband}}}{\text{number of channels}},$$
$$\text{average MNR} = \frac{\sum_{\text{subband}} \text{mean MNR}_{\text{subband}}}{\text{number of subband}}. \quad (6)$$

Table 1 shows the MNR values for the performance comparison of the nonprogressive AAC algorithm and the proposed PSMAC algorithm when working in the MNR progressive profile. Values in this table clearly show that our codec outperforms AAC for both testing materials at lower bit rates and it only has a small performance degradation at higher bit rates. In addition, the bitstream generated by MPEG AAC only achieves an approximate bit rate and is normally a little

Table 1: MNR comparison for MNR progressive profiles.

Bit rate (bit/s/ch)	Average MNR values (dB/subband/ch)			
	Herre		Messiah	
	AAC	PSMAC	AAC	PSMAC
16k	−0.90	6.00	14.37	21.82
32k	5.81	14.63	32.40	34.57
48k	17.92	22.32	45.13	42.81
64k	28.64	28.42	54.67	47.84



(a)

(b)

(c)

(d)

Figure 10: Listening test results for multichannel audio sources where A = MPEG AAC and P = PSMAC.

bit higher than the desired one while our algorithm achieves a much more accurate bit rate in all experiments carried out.

7.1.2. *Random access*

The MNR result after the base-layer reconstruction for the random access mode by using the test material "Herre" is shown in Table 2. When listening to the reconstructed music, we can clearly hear the quality difference between the enhance period and the rest of the other period. The MNR value given in Table 2 verifies the above claim by showing that the mean MNR value for the enhanced period is much better (more than 10 dB per subband) than the rest of other periods. It is common that we may prefer a certain part of a music to others. With the random access profile, the user can individually access a period of music with better quality than others when the network condition does not allow a full high quality transmission.

7.1.3. *Channel enhancement*

The performance result using the test material "Herre" for the channel enhancement mode is also shown in Table 2. Here, the center channel has been enhanced with enhancement parameter 1. Note that the total bit rate is kept the same for both codecs, that is, each has an average bit rate of 16 Kbps/ch. Since we have to separate the quantization and the coding control of the enhanced physical channel as well as to simplify the implementation, KLT is disabled in the channel enhancement mode. Compared with the normal MNR progressive mode, we find that the enhanced center channel has an average of more than 10 dB per subband MNR improvement, while the quality of other channels is only degraded by about 3 dB per subband.

When an expert subjectively listens to the reconstructed audio, the one with the enhanced center channel has a much better performance and is more appealing, compared with

Table 2: MNR comparison for random access and channel enhancement profiles.

Average MNR values (dB/subband/ch)					
Random access		Channel enhancement			
		Enhanced channel		Other channels	
Other area	Enhanced area	w/o enhance	w/ enhance	w/o enhance	w/ enhance
3.99	13.94	8.42	19.23	1.09	−2.19

the one without channel enhancement. This is because the center channel of "Herre" contains more musical information than other channels, and a better reconstructed center channel will give listeners a better overall quality, which is basically true for most multichannel audio materials. Therefore, this experiment suggests that, with a narrower bandwidth, audio generated by the channel enhancement mode of the PSMAC algorithm can provide the user a more compelling experience with either a better reconstructed center channel or a channel which is more interesting to a particular user.

7.2. Subjective listening test

In order to further confirm the advantage of the proposed PSMAC algorithm, a formal subjective listening test according to ITU recommendations [20, 21, 22] was conducted in an audio lab to compare the coding performance of PSMAC and the MPEG AAC main profile. At the bit rate of 64 Kbps/ch, the reconstructed sound clips are supposed to have a perceptual quality similar to that of the original ones, which means that the difference between PSMAC and AAC would be so small that nonprofessionals can hardly hear it. According to our experience, nonprofessional listeners tend to give random scores if they cannot tell the difference between two sound clips, which makes their scores nonrepresentative. Therefore, instead of inviting a large number of nonexpert listeners, four well-trained professionals, who have no knowledge of any algorithms, participated in the listening test [22]. For each test sound clip, subjects listened to three versions of the same sound clip, that is, the original one followed by two processed ones (one by PSMAC and one by AAC in a random order), subjects were allowed to listen to these files as many times as possible until they were comfortable to give scores to the two processed sound files for each test material.

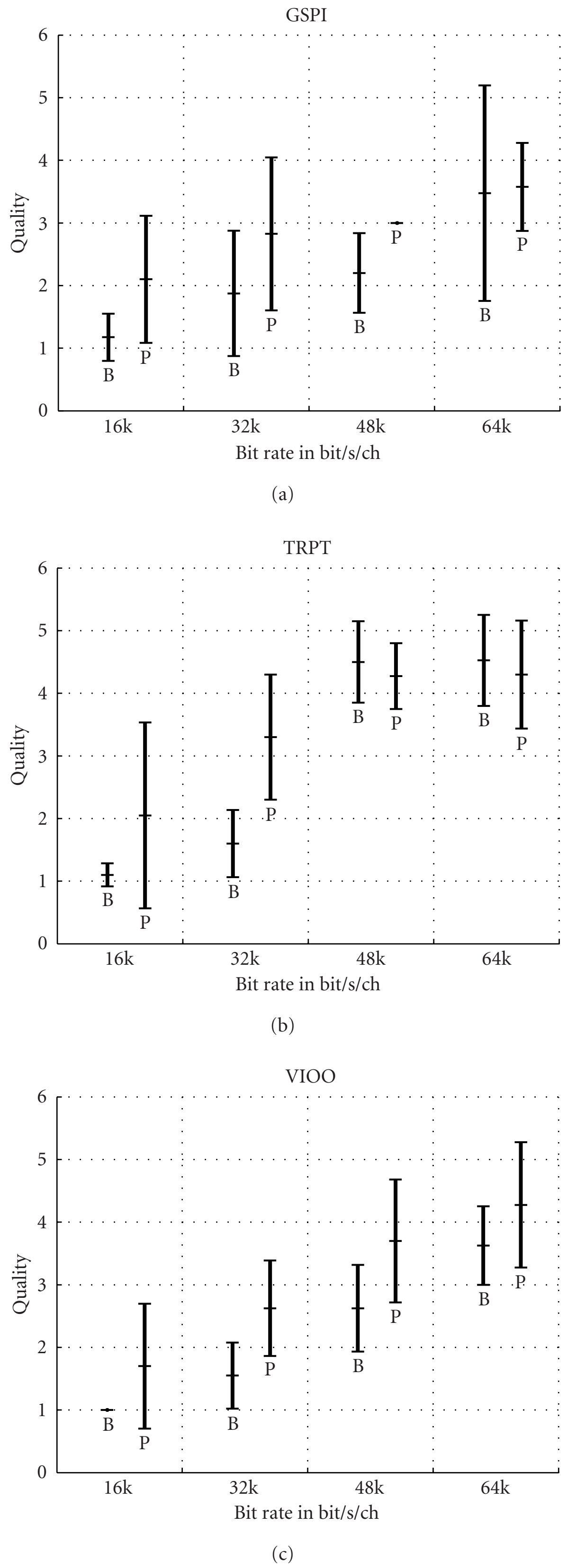
The five-grade impairment scale given in Recommendation ITU-R BS. 1284 [21] was adopted in the grading procedure and utilized for final data analysis. Besides "Messiah" and "Herre," another two ten-channel audio materials called "Band" and "Herbie" were included in this subjective listening test, where "Band" is a rock band music lively recorded in a football field, and "Herbie" is a piece of music played by an orchestra. According to ITU-R BS. 1116-1 [20], audio files selected for listening test only contained short durations, that is, 10 to 20 seconds long.

Figure 10 shows the score given to each test material coded at four different bit rates during the listening test for multichannel audio materials. The solid vertical line represents the 95% confidence interval, where the middle line shows the mean value and the other two lines at the boundary of the vertical line represent the upper and lower confidence limits [23]. It is clear from Figure 10 that, at lower bit rates, such as 16 Kbps/ch and 32 Kbps/ch, the proposed PSMAC algorithm outperforms MPEG AAC in all four test materials. To be more precise, at these two bit rates for all four test materials, the proposed PSMAC algorithm achieves statistically significantly better results.[4] At higher bit rates, such as 48 Kbps/ch and 64 Kbps/ch, PSMAC achieves either comparable or slightly degraded subjective quality when compared with MPEG AAC.

To demonstrate that the PSMAC algorithm achieves an excellent coding performance even for single-channel audio files, another listening test for the mono sound was also carried out. Three single-channel single-instrument test audio materials, which are downloaded and processed from MPEG sound quality assessment material, known as "GSPI" (http://www.tnt.uni-hannover.de/project/mpeg/audio/sqam/), "TRPT" (http://www.tnt.uni-hannover.de/project/mpeg/audio/sqam/), and "VIOO" (http://www.tnt.uni-hannover.de/project/mpeg/audio/sqam/), were used in this experiment, and the performance between the standard fine-grain scalable audio coder provided by MPEG-4 BSAC [6, 8] and the proposed PSMAC was compared.

Figure 11 shows the listening test results for the three single-channel audio materials. For cases where no confidence intervals are shown, it means that all four listeners happened to give the same score to the given sound clip. From this figure, we can clearly see that at lower bit rates, for example, 16 Kbps/ch and 32 Kbps/ch, our algorithm generates better sound quality for all test sequences. In all cases, except "GSPI" coded at 32 Kbps/ch, PSMAC achieves statistically significantly better performance than that of MPEG-4 BSAC. At higher bit rates, for example, 48 Kbps/ch and 64 Kbps/ch, our algorithm outperforms MPEG-4 BSAC for two out of three test materials and is only slightly worse for the "TRPT" case.

[4] We call algorithm A statistically significantly better than algorithm B if the mean value given to the sound clip processed by algorithm A is above the upper 95% confidence limit given to sound clip processed by algorithm B.



(a)

(b)

(c)

Figure 11: Listening test results for single-channel audio sources where B = BSAC and P = PSMAC.

8. CONCLUSION

A PSMAC algorithm was presented in this research. This algorithm utilized KLT as a preprocessing block to remove interchannel redundancy inherent in the original multichannel audio source. Then, rules for channel selection and subband selection were developed and the SAQ process was used to determine the importance of coefficients and their layered information. At the last stage, all information was losslessly compressed by using the context-based QM-coder to generate the final multichannel audio bitstream.

The distinct advantages of the proposed algorithm over most existing MACs not only lie in its progressive transmission property which can achieve a precise rate control but also in its rich-syntax design. Compared with the new MPEG-4 BSAC tool, PSMAC provides a more delicate subband selection strategy such that the information, which is more sensitive to the human ear, is reconstructed earlier and more precisely at the decoder side. It was shown by experimental results that PSMAC has a comparable performance as nonprogressive MPEG AAC at several different bit rates when using the multichannel test material while PSMAC achieves better reconstructed audio quality than MPEG-4 BSAC tools when using single-channel test materials. Moreover, the advantage of the proposed algorithm over the other existing audio codec is more obvious at lower bit rates.

ACKNOWLEDGMENTS

This is research has been funded by the Integrated Media Systems Center and National Science Foundation Engineering Research Center, Cooperative Agreement no. EEC-9529152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

REFERENCES

[1] ISO/IEC 13818-5, *Information technology—Generic coding of moving pictures and associated audio information—Part 5: Software simulation*, 1997.

[2] ISO/IEC 13818-7, *Information technology—Generic coding of moving pictures and associated audio information—Part 7: Advanced audio coding*, 1997.

[3] K. Brandenburg and M. Bosi, "ISO/IEC MPEG-2 advanced audio coding: overview and applications," in *Proc. 103rd Convention of Audio Engineering Society (AES)*, New York, NY, USA, September 1997.

[4] M. Bosi, K. Brandenburg, S. Quackenbush, et al., "ISO/IEC MPEG-2 advanced audio coding," in *Proc. 101st Convention of Audio Engineering Society (AES)*, Los Angeles, Calif, USA, November 1996.

[5] S.-H. Park, Y.-B. Kim, S.-W. Kim, and Y.-S. Seo, "Multi-layer bit-sliced bit-rate scalable audio coding," in *Proc. 103rd Convention of Audio Engineering Society (AES)*, New York, NY, USA, September 1997.

[6] ISO/IEC JTC1/SC29/WG11 N2205, *Final Text of ISO/IEC FCD 14496-5 Reference Software*.

[7] ISO/IEC JTC1/SC29/WG11 N2803, *Text ISO/IEC 14496-3 Amd 1/FPDAM*.

[8] ISO/IEC JTC1/SC29/WG11 N4025, *Text of ISO/IEC 14496-5:2001*.

[9] J. Herre, E. Allamanche, K. Brandenburg, et al., "The integrated filterbank based scalable MPEG-4 audio coder," in *Proc. 105th Convention of Audio Engineering Society (AES)*, San Francisco, Calif, USA, September 1998.

[10] J. Zhou and J. Li, "Scalable audio streaming over the internet with network-aware rate-distortion optimization," in *Proc. IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 2001.

[11] M. S. Vinton and E. Atlas, "A scalable and progressive audio codec," in *Proc. IEEE International Conference on Aoustics Speech and Signal Processing*, Salt Lake City, Utah, USA, May 2001.

[12] Y. Shen, H. Ai, and C.-C. J. Kuo, "A progressive algorithm for perceptual coding of digital audio signals," in *Proc. 33rd Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, Calif, USA, Octorber 1999.

[13] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[14] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "An interchannel redundancy removal approach for high-quality multichannel audio compression," in *Proc. 109th Convention of Audio Engineering Society (AES)*, Los Angeles, Calif, USA, September 2000.

[15] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "An exploration of Karhunen-Loéve transform for multichannel audio coding," in *Proc. SPIE on Digital Cinema and Microdisplays*, vol. 4207 of *SPIE Proceedings*, pp. 89–100, Boston, Mass, USA, November 2000.

[16] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "High fidelity multichannel audio coding with Karhunen-Loéve transform," *IEEE Trans. Speech, and Audio Processing*, vol. 11, no. 4, 2003.

[17] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3445–3462, 1993.

[18] W. Pennebaker and J. Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, NY, USA, 1993.

[19] K. Brandenburg, "Evaluation of quality for audio encoding at low bit rates," in *Proc. 82nd Convention of Audio Engineering Society (AES)*, London, UK, 1987.

[20] ITU-R Recommendation BS.1116-1, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*.

[21] ITU-R Recommendation BS.1284, *Methods for the subjective assessment of sound quality – general requirements*.

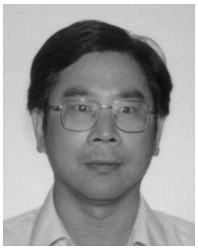
[22] ITU-R Recommendation BS.1285, *Pre-selection methods for the subjective assessment of small impairments in audio systems*.

[23] R. A. Damon Jr. and W. R. Harvey, *Experimental Design, ANOVA, and Regression*, Harper & Row Publishers, New York, NY, USA, 1987.

Dai Yang received the B.S. degree in electronics from Peking University, Beijing, China in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, Calif in 1999 and 2002, respectively. She is currently a Postdoctoral Researcher in NTT Cyber Space Laboratories in Tokyo, Japan. Her research interests are in the areas of digital signal and image processing, audio, speech, video, graphics coding, and their network/wireless applications.

Hongmei Ai received the B.S. degree in 1991, and the M.S. and Ph.D. degrees in 1996 all in electronic engineering from Tsinghua University. She was an Assistant Professor (1996–1998) and Associate Professor (1998–1999) in the Department of Electronic Engineering at Tsinghua University, Beijing, China. She was a Visiting Scholar in the Department of Electrical Engineering at the University of Southern California, Los Angeles, Calif from 1999 to 2002. Now she is a Principal Software Engineer at Pharos Science & Applications, Inc., Torrance, Calif. Her research interests focus on signal and information processing and communications, including data compression, video, and audio processing, and wireless communications.

Chris Kyriakakis received the B.S. degree from California Institute of Technology in 1985, and the M.S. and Ph.D. degrees from the University of Southern California in 1987 and 1993, respectively, all in electrical engineering. He is currently an Associate Professor in the Department of Electrical Engineering, University of Southern California. He heads the Immersive Audio Laboratory. His research focuses on multichannel audio acquisition, synthesis, rendering, room equalization, streaming, and compression. He is also the Research Area Director for sensory interfaces in the Integrated Media Systems Center which is the National Science Foundation's Exclusive Engineering Research Center for multimedia and Internet research at the University of Southern California.



C.-C. Jay Kuo received the B.S. degree from the National Taiwan University, Taipei, in 1980, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively, all in electrical engineering. Dr. Kuo was a computational and applied mathematics (CAM) Research Assistant Professor in the Department of Mathematics at the University of California, Los Angeles, from October 1987 to December 1988. Since January 1989, he has been with the Department of Electrical Engineering-Systems and the Signal and Image Processing Institute at the University of Southern California, where he currently has a joint appointment as a Professor of electrical engineering and mathematics. His research interests are in the areas of digital signal and image processing, audio and video coding, wavelet theory and applications, and multimedia technologies and large-scale scientific computing. He has authored around 500 technical publications in international conferences and journals, and graduated more than 50 Ph.D. students. Dr. Kuo is a member of SIAM and ACM, and a Fellow of IEEE and SPIE. He is the Editor-in-Chief of the Journal of Visual Communication and Image Representation. Dr. Kuo received the National Science Foundation Young Investigator Award (NYI) and Presidential Faculty Fellow (PFF) Award in 1992 and 1993, respectively.