A DEVICE AND PROCESS FOR ENCODING AUDIO DATA

FIELD OF THE INVENTION

The present invention relates to a device and process for encoding audio data, and in particular to a process for determining encoding parameters for use in MPEG audio encoding.

BACKGROUND

The MPEG-1 audio standard, as described in the International Standards Organisation (ISO) document ISO/IEC 11172-3: Information technology - *Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbps* ("the MPEG-1 standard"), defines processes for lossy compression of digital audio and video data. The MPEG-1 standard defines three alternative processes or "layers" for audio compression, providing progressively higher degrees of compression at the expense of increasing complexity. The third layer, referred to as MPEG-1-L3 or MP3, provides an audio compression format widely used in consumer audio applications. The format is based on a psychoacoustic or perceptual model that allows significant levels of data compression (*e.g.*, typically 12:1 for standard compact disk (CD) digital audio data using 16-bit samples sampled at 44.1 kHz), whilst maintaining high quality sound reproduction, as perceived by a human listener. Nevertheless, it remains desirable to provide even higher levels of data compression, yet such improvements in compression are usually attended by an undesirable degradation in perceived sound quality. Accordingly, it is desired to address the above or at least to provide a useful alternative.

SUMMARY OF THE INVENTION

In accordance with the present invention, there is provided a process for encoding audio data, including:

determining a first encoding parameter for encoding a block of audio data if a temporal masking transient is not detected in said block of audio data; and

determining a second encoding parameter for encoding said block of audio data if a temporal masking transient is detected in said block of audio data to enable greater compression of said audio data.

The present invention also provides a scalefactor generator for an audio encoder, said scalefactor generator adapted to generate scalefactors for use in quantizing respective portions of a block of audio data if a temporal masking transient is not detected in said block of audio data; and to select one of said scalefactors for use in quantizing each of said portions if a temporal masking transient is detected in said block of audio data to enable greater compression of said audio data.

The present invention also provides a scalefactor modifier for an audio encoder, said scalefactor modifier adapted to output scalefactors for use in quantizing respective portions of a block of audio data if a temporal masking transient is not detected in said block of audio data; and to select one of said scalefactors for use in quantizing each of said portions if a temporal masking transient is detected in said block of audio data to enable greater compression of said audio data.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the present invention are hereinafter described, by way of example only, with reference to the accompanying drawings, wherein:

Figure 1 is a block diagram of a preferred embodiment of an audio encoder;

Figure 2 is a flow diagram of a scalefactor generation process executed by the audio encoder;

Figure 3 is a barchart of the increase in compression of encoded audio data generated by the audio encoder over that generated by a prior art audio encoder; and

Figure 4 is a graph comparing the quality of encoded audio data generated by the audio encoder and a prior art audio encoder.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

As shown in Figure 1, an audio encoder includes an input pre-processing module 102, a fast Fourier transform (FFT) analysis module 104, a masking threshold generator 106, a windowing module 108, a filter bank and modified discrete cosine transform (MDCT) module 110, a joint stereo coding module 112, a scalefactor generator 114, a scalefactor modifier 115, a quantization module 116, a noiseless coding module 118, a rate distortion/control module 120, and a bit stream multiplexer 122. The audio encoder executes an audio encoding process that generates encoded audio data 124 from input audio data 126. The encoded audio data 124 constitutes a compressed representation of the input audio data 126.

In the described embodiment, the audio encoder is a standard digital signal processor (DSP) such as a TMS320 series DSP manufactured by Texas Instruments, and the modules 102 to 122 of the encoder are software modules stored in the firmware of the DSP-core. However, it will be apparent that at least part of the audio encoding modules 102 to 122 could alternatively be implemented as dedicated hardware components such as application-specific integrated circuits (ASICs).

The audio encoding process executed by the encoder performs encoding steps based on MPEG-1 layer 3 processes described in the MPEG-1 standard. The input audio data 126 is time-domain pulse code modulated (PCM) digital audio data, which may be of DVD quality, using a sample rate of 48,000 samples per second. As described in the MPEG-1 standard, the time-domain input audio data 126 is divided into 32 sub-bands and (modified) discrete cosine transformed by the filter bank and MDCT module 110, and the resulting frequency-domain (spectral) data undergoes stereo redundancy coding, as performed by the joint stereo coding module 112. The scalefactor generator 114 then generates scalefactors that determine the quantization resolution, as described below, and the audio data is then quantized by the quantization module 106 using quantization parameters determined by the rate distortion/control module 120. The bitstream multiplexer 122 then generates the encoded audio data or bitstream 124 from the quantized data.

The quantizer 106 performs bit allocation and quantization based upon masking data generated by the masking threshold generator 106. The masking data is generated from the input audio data 126 on the basis of a psychoacoustic model of human hearing and aural perception. The psychoacoustic modelling takes into account the frequency-dependent thresholds of human hearing, and a psychoacoustic phenomenon referred to as masking, whereby a strong frequency component close to one or more weaker frequency components tends to mask the weaker components, rendering them inaudible to a human listener. This makes it possible to omit the weaker frequency components when encoding audio data, and thereby achieve a higher degree of compression, without adversely affecting the perceived quality of the encoded audio data 124. The masking data comprises a signal-to-mask ratio value for each frequency sub-band. These signal-to-mask ratio values represent the amount of signal masked by the human ear in each frequency sub-band, and are therefore also referred to as masking thresholds. The quantizer 116 uses this information to decide how best to use the available number of data bits to represent the input audio data 116, as described in the MPEG-1 standard. Information describing how

the available bits are distributed over the audio spectrum is included as side information in the encoded audio bitstream 124.

The MPEG-1 standard specifies the layer 3 encoding of audio data in long blocks comprising three groups of twelve samples (*i.e.*, 36 samples) over the 32 sub-bands, making a total of 1152 samples. However, the encoding of long blocks gives rise to an undesirable artefact if the long block contains one or more sharp transients, for example, a period of silence followed by a percussive sound, such as from a castanet or a triangle. The encoding of a long block containing a transient can cause relatively large quantization errors which are spread across an entire frame when that frame is decoded. In particular, the encoding of a transient typically gives rise to a pre-echo, where the percussive sound becomes audible prior to the true transient. To alleviate this effect, the MPEG-1 standard specifies the layer 3 encoding of audio data using two block lengths: a long block of 1152 samples, as described above, and a short block of twelve samples for each sub-band, *i.e.*, 12*32 = 384 samples. The short block is used when a transient is detected to improve the time resolution of the encoding process when processing transients in the audio data, thereby reducing the effects of pre-echo.

A psychoacoustic effect referred to as temporal masking can disguise such effects. In particular, the human auditory system is insensitive to low level sounds in a period of approximately 20 milliseconds prior to the appearance of a much louder sound. Similarly, a post-masking effect renders low level sounds inaudible for a period of up to 200 milliseconds after a comparatively loud sound. Accordingly, the use of short coding blocks for encoding transients can mask pre-echoes if the time spread is of the order of a few milliseconds. Furthermore, MPEG-1 layer 3 encoding processes control pre-echo by reducing the threshold of hearing used by the masking threshold generator 106 when a transient is detected.

As shown in Figure 2, the audio encoder executes a scalefactor generation process that generates scalefactors for use by the quantization module 116 and the rate distortion/control module 120 to determine suitable quantization parameters for quantizing

spectral components of the audio data. When encoding blocks of spectral data which do not contain appreciable transients, the data is encoded in long blocks of 1152 samples, as described above. The process begins at step 202 by determining whether the block of spectral data from the joint stereo coding module 112 is a long block or a short block, indicating whether a transient was detected by the input pre-processing module 102. In this case, the block is a long block, and hence no transient was detected and standard processing is therefore performed at step 204. That is, scalefactors are generated by the scalefactor generator 114 in accordance with the MPEG-1 layer 3 standard. These scalefactors are then passed to the quantization module 116. Alternatively, if a short block has been passed to the scalefactor generator 114, then a test is performed to determine whether standard pre-echo control, as described above, is to be used. If so, then the process performs standard processing at step 204. This involves limiting the value of the scalefactors to reduce transient pre-echo, as described in the MPEG-1 standard. Alternatively, if pre-echo control is not invoked, then three scalefactors scf_m, $m = 1,2,3$ are generated by the scalefactor generator 114 for three respective groups of twelve spectral coefficients generated by the filter bank and MDCT module 110.

At step 210, the scalefactor modifier 115 selects the greatest of these three scalefactors as scf_{max}. Thus instead of normalizing the three groups of spectral coefficients by their respective scalefactors, as per the MPEG-1 layer 3 standard, all three groups of coefficients can be normalized by the maximum scalefactor scf_{max}. The use of the maximum scalefactor reduces the dynamic range of the encoded spectral coefficients. The Huffman coding performed by the noiseless coding module 118 ensures that input samples which occur more often are assigned fewer bits. Consequently, quantization and coding of these smaller values results in fewer bits in the encoded audio data 124; *i.e.*, greater compression.

However, normalizing by a greater scalefactor would also increase the quantization error. In particular, the signal-to-noise ratio for the mth spectral coefficient (SNR_m) is given by

$$SNR_m = 10.\log\frac{P_s}{\kappa_m^2.P_n} = 10.\log\frac{P_s}{P_n} - 10.\log\kappa_m^2,$$

where

$$\kappa_m = \frac{scf_{\max}}{scf_m}$$

where P_s is the signal power, and P_n is the quantization noise power, given by:

$$P_n = \int_{-\Delta/2}^{\Delta/2} e^2 . p(e) de$$

where, e represents the error, *i.e.*, the difference between a true spectral coefficient and its quantized value, $p(e)$ is the probability density function of the quantization error, and Δ is the quantizer step size. In the case of linear quantizers, the error is uniformly distributed over a range $-\Delta/2$ to $+\Delta/2$, where Δ is the quantizer step size. Δ varies for power law quantizers, which are used in MPEG 1 Layer 3 encoders.

Accordingly, the degree of degradation Err_m of the SNR_m resulting from using the maximum scalefactor value is given by:

$$Err_m = 20.\log \kappa_m$$

This degree of degradation Err_m is determined at step 214.

At step 216, the sound energy E_m, m = 1,2,3 in each group of 12 samples is determined from the MDCT coefficients $X_m(k)$, as follows:

$$E_m = \sum_{k=1}^{12} X_m(k)^2,$$

The energy in each group is used to determine the duration of the temporal pre-masking and post-masking effects of the transient signal under consideration, as described below.

In a short block, the scalefactors are generated from the MDCT spectrum, which depends on the 12 samples output from each subband filter of the filterbank and MDCT module 110. In standard MP3 encoders, 3 sets of 12 samples are grouped together.

Applying the principles of temporal masking to short blocks, if the signal energy E_2 in the second group is higher than the signal energy E_1 of the previous set of 12 samples, the effect of the first set of samples will be masked by the second set due to pre-masking. This is possible as 12 samples at a sampling rate of 48,000 samples per second corresponds to a period of 0.25 ms. Similarly, 24 samples correspond to 0.5 ms, which is much smaller than the 20 ms pre-masking period.

In the human auditory system, post-masking is more dominant than pre-masking. Consequently, quantization errors are more likely to be perceived when relying on pre-masking. The worst cases occur when the third set of 12 samples is relied on to pre-mask the previous 24 samples. Consequently, a test is performed at step 218 to detect this situation by determining whether the energies of each group of 12 samples are in ascending order, *i.e.*, whether $E_1 < E_2 < E_3$. If so, then a further test is performed at step 222 by comparing the degradation Err_m of the SNR that would result from using the maximum scalefactor to the SNR associated with quantization noise. If the noise Err_m introduced by increasing the scalefactors is greater than 30% of the SNR, the encoder performs standard processing at step 204; *i.e.*, the respective scalefactors scf_m are used, as per the MPEG-1 layer 3 standard.

The scalefactor modifier 115 is activated only after the scalefactors are generated at step 208. This ensures that higher numbers of bits are not allocated for the modified scalefactors and allows the effect of temporal masking to be taken into account.

The encoded audio stream 124 generated by the audio encoder is compatible with any standard MPEG-1 Layer 3 decoder. In order to quantify the improved compression of the encoder, it was used to encode 17 audio files in the waveform audio '.wav' format and

sizes of the resulting encoded files are compared with those for a standard MPEG Layer 3 encoder in Figure 3. To achieve a higher compression, both encoders were tested at variable bitrates and using the lowest quality factor. Figure 3 shows that, for the particular audio files tested, the improvement in compression produced by the audio encoder is at least 1%, and is nearly 10% in some cases. The amount of compression will, of course, depend on the number of transients present in the input audio data 126.

In order to assess the quality of the audio encoder, a quality-testing software program known as OPERA (Objective PERceptual Analyzer) was used, as described at http://www.opticom.de. This program objectively evaluates the quality of wide-band audio signals by simulating the human auditory system. OPERA is based on the most recent perceptual techniques, and is compliant with PEAQ (Perceptual Evaluation of Audio Quality), an ITU-R standard.

Using OPERA, the quality of the ISO MPEG-1 Layer 3 encoder was compared to that of the audio encoder. Figure 4 is a graph comparing objective difference grade (ODG) values generated for each of the '.wav' files represented in Figure 3 and the corresponding input audio data 126. The ODG values for the audio encoder are joined by a solid line 402 and those for a standard MP3 audio encoder are shown as a dashed line 404. ODG values can range from –4.0 to 0.4, with more positive ODG values indicating better quality. A zero or positive ODG value corresponds to an imperceptible impairment, and –4.0 corresponds to an impairment judged as annoying. The tradeoff in quality due to higher compression of the audio files is apparent by the marginally more negative ODG values 402 for the audio encoder compared to those 404 for the standard MP3 audio encoder. As can be observed, files with higher compression have a marginally lower ODG value, with a typical higher compression ratio of 4-5 % leading to a decrease in ODG value by only 0.16.

Although the audio encoding process described above has been described in terms of determining scalefactors for use in quantizing audio data to generate MPEG audio data, it will be apparent that alternative embodiments of the invention can be readily envisaged in which encoding errors produced by any lossy audio encoding process are allowed to increase in selected portions of audio data that are masked by temporal transients. Thus the

resulting degradation in quality, which would be apparent if the encoding errors were not masked, is instead hidden from a human listener by the psychoacoustic effects of temporal masking.

Many modifications will be apparent to those skilled in the art without departing from the scope of the present invention as herein described with reference to the accompanying drawings.

CLAIMS:

1. A process for encoding audio data, including:

determining a first encoding parameter for encoding a block of audio data if a temporal masking transient is not detected in said block of audio data; and

determining a second encoding parameter for encoding said block of audio data if a temporal masking transient is detected in said block of audio data to enable greater compression of said audio data.

2. A process as claimed in claim 1, wherein said step of determining a second encoding parameter includes:

generating an error value representing an encoding error for encoding using said second encoding parameter; and

selecting, on the basis of said error value, one of said first encoding parameter and said second encoding parameter for encoding said block of audio data.

3. A process as claimed in claim 1, wherein said first encoding parameter and said second encoding parameter are scalefactors for use in quantizing said block of audio data.

4. A process as claimed in claim 3, wherein said step of determining a first encoding parameter includes generating first scalefactors for use in quantizing respective portions of said block of audio data; and wherein said step of determining a second encoding parameter includes selecting one of said first scalefactors for use in quantizing each of said portions if a temporal masking transient is detected in said block of audio data.

5. A process as claimed in claim 4, wherein said portions correspond to groups of audio samples, and said selecting includes selecting the maximum of said first scalefactors.

6. A process as claimed in claim 4, including determining whether said temporal masking transient is in a last portion of said block, and, if so, then generating an error value

representing an encoding error for encoding using the selected scalefactor, and selecting the selected scalefactor for encoding said block of audio data if said error value satisfies an error criterion.

7. A process as claimed in claim 6, wherein the temporal masking transient is determined to be in a last portion of said block if respective energies of said portions are in ascending order.

8. A process as claimed in claim 6, wherein said error criterion is satisfied if said error value is less than a predetermined fraction of a corresponding quantization error value.

9. A process as claimed in claim 8, wherein said predetermined fraction is substantially equal to 0.3.

10. A process as claimed in claim 8, wherein said quantization error value represents a signal to noise ratio for quantization, and said error value represents the degradation of signal to noise ratio resulting from encoding using the selected scalefactor.

11. A process as claimed in any one of claims 1 to 10, wherein the process generates MPEG encoded audio data.

12. A process as claimed in any one of claims 1 to 11, wherein the process is an MPEG-1 layer 3 audio encoding process.

13. A computer readable storage medium having stored thereon program code for executing the steps of any one of claims 1 to 12.

14. An audio encoder having components for executing the steps of any one of claims 1 to 12.

15. A scalefactor generator having components for executing the steps of any one of claims 1 to 12.

16. A scalefactor generator for an audio encoder, said scalefactor generator adapted to generate scalefactors for use in quantizing respective portions of a block of audio data if a temporal masking transient is not detected in said block of audio data; and to select one of said scalefactors for use in quantizing each of said portions if a temporal masking transient is detected in said block of audio data to enable greater compression of said audio data.

17. A scalefactor generator as claimed in claim 16, wherein the maximum of said scalefactors is selected.

18. A scalefactor generator as claimed in claim 16, wherein said scalefactor generator is further adapted to determine whether said temporal masking transient is in a last portion of said block, and, if so, to generate an error value representing an encoding error for encoding using the selected scalefactor, and to select the selected scalefactor for encoding said block of audio data if said error value satisfies an error criterion.

19. A scalefactor modifier for an audio encoder, said scalefactor modifier adapted to output scalefactors for use in quantizing respective portions of a block of audio data if a temporal masking transient is not detected in said block of audio data; and to select one of said scalefactors for use in quantizing each of said portions if a temporal masking transient is detected in said block of audio data to enable greater compression of said audio data.