

DESCRIPTION

Method and Apparatus for Processing Audio
Information Using Time-Aligned Blocks of Encoded
Audio in Video/Audio Applications to Facilitate Audio
Switching

TECHNICAL FIELD

The present invention relates generally to audio signal processing in video/audio applications. More particularly, the present invention relates to block-coding methods for encoding sequences of blocks of audio information separated by gaps or guard bands so that normal variations in signal processing delays will not disrupt the alignment of video and audio information.

BACKGROUND ART

Several international standards have been developed which define various aspects of embedding digital audio information into frames of video information. For example, standard SMPTE 259M published by the Society of Motion Picture and Television Engineers (SMPTE) defines a Serial Digital Interface (SDI) in which up to four channels of digital audio information may be embedded into component and composite serial digital video signals. Standard SMPTE 272M provides a full definition of how digital audio information is to be embedded in ancillary data spaces within frames of the video information.

The serial transmission of digital audio information itself is the subject of various international standards. For example, standard AES3 (ANSI S4.40) published by the Audio Engineering Society (AES), defines serial transmission of two-channel digital audio represented in a linear pulse code modulation (PCM) form. According to this standard, PCM samples for two channels are interleaved and conveyed in pairs.

A common activity in nearly all recording and broadcasting applications is editing or cutting embedded video/audio information streams and splicing the cut information streams to form a new single stream. Similar activities generate an information stream by merging multiple information streams or by switching between multiple streams. The video information is normally the primary synchronizing reference so that an edit or cut point is normally aligned with a video frame.



G00002



166166

Standards such as AES11 define recommended practices for synchronizing digital audio equipment in studio operations. AES11 is directed toward controlling timing uncertainties caused by jitter or processing delays and provides for aligning video frame information with the two-sample frames of AES3 digital audio information streams. Equipment and methods that adhere to this standard can ensure that synchronized signals have the same number of frames over a given period of time and contain samples that have a common timing. Unfortunately, no standards or practices currently exist which define an alignment between video information and larger intervals of audio information. As a result, equipment from different manufacturers and even from the same manufacturer have variations in timing and in processing delays that introduce a significant amount of uncertainty in the relative alignment of audio and video information.

This uncertainty in alignment is of little consequence in applications that use linear representations of audio information such as that defined in the AES3 standard. Because edit points are constrained to occur between the two-sample frames of audio information, any uncertainty in video/audio alignment will not result in the loss of audio information. It will only affect the relative timing of sound and picture as presented to a person, which is unlikely to be discernable.

There is, however, a growing number of applications that use bit-rate-reduction encoding techniques to embed greater numbers of audio channels into a video/audio data stream. These encoding techniques are often applied to sample blocks of 128 or more audio samples to generate blocks of encoded information. These sample blocks typically represent audio information that spans an interval of 3 to 12 ms. Each block of encoded information generated by these encoding processes represents the smallest unit of information from which a reasonably accurate replica of a segment of the original audio information can be recovered. Split-band coding techniques reduce bit rates by applying psychoacoustic-based coding to frequency-subband representations of an audio signal. The frequency-subband representations may be generated by application of a plurality of bandpass filters or one or more transforms. For ease of discussion, these split-band coding techniques are described here in terms of applying a filterbank to generate subband signals.

The uncertainty in alignment mentioned above is significant in these block-coding applications because an edit point falling within the boundaries of an encoded

block will result in part of that block being cut from the remaining signal. The partial loss of an encoded block will be manifested by a loss in the recovered signal for a duration typically of 3 ms or more. It is likely that such a loss would be discernable to the human auditory system.

5 This problem may be avoided by using a post-processing process in which a PCM representation of the original audio signals is recovered by applying a decoding process to the encoded audio, editing the recovered PCM representation as required, and generating a new encoded representation by applying an encoding process to the edited PCM audio information. This solution is unattractive because of the additional
10 costs and degradation in audio quality resulting from the decoding/re-encoding processes. In addition, for reasons that will be better understood after reading the discussion set forth below, post-processing is unattractive because the decoding/re-encoding processes introduce additional delays in the audio information stream.

15 **DISCLOSURE OF INVENTION**

It is an object of the present invention to provide for the processing of embedded video/audio information streams that permits activities like editing and switching while avoiding the aforementioned problems.

20 According to the teachings of one aspect of present invention, a method receives an input audio signal conveying audio information, receives video frame references indicating time references for a sequence of video frames, generates blocks of encoded audio information from the audio information in a bit-rate reduced form by applying a block-encoding process to the input audio signal and time compresses the blocks of encoded audio information, and assembles the time-compressed blocks
25 into an encoded audio stream comprising a plurality of sequences of the time-compressed blocks such that a beginning block in a respective sequence is separated from an ending block in a preceding sequence by a gap that is time aligned with a respective video frame reference.

30 The various features of the present invention and its preferred embodiments may be better understood by referring to the following discussion and the accompanying drawings in which like reference numerals refer to like elements in the several figures. The contents of the following discussion and the drawings are set

embodiment is mentioned here to show that no particular signal format is essential to practice the present invention.

Video tape recorder (VTR) 16 receives video information from path 1 and audio information from path 2 and records this video/audio information on tape.

5 Subsequently, VTR 16 reads the video/audio information recorded on tape and generates along path 21 a playback signal conveying the video information with embedded audio information. In a similar manner, VTR 17 records video and audio information received from paths 3 and 4, respectively, and subsequently generates along path 22 a playback signal conveying the video information with embedded
10 audio information.

VTR 16, VTR 17 and VTR 18 include circuitry such as a Serial Digital Interface (SDI) embedder for embedding audio information into the video information during playback.

SDI embedder 11 receives video and audio information from paths 5 and 6,
15 respectively, and generates along path 14 a signal conveying digital video information with embedded digital audio information. VTR 18, which includes circuitry such as a SDI disembedder, extracts audio information from the video/audio data signal and records the separated video and audio information on tape. Subsequently, VTR 18 recovers the video and the audio information from tape and uses circuitry such as an
20 SDI embedder to generate along path 23 a playback signal conveying the video and embedded audio information. If a digital data recorder is substituted for VTR 18, however, neither embedder nor disembedder circuitry is required in the recorder because the video/audio data stream itself can be recorded and played back.

SDI embedder 12 receives video and audio information from paths 7 and 8,
25 respectively, and generates along path 24 a signal conveying digital video information with embedded digital audio information.

SDI router 31 receives video/audio signals from paths 21, 22, 23 and 24 and selectively routes or switches these signals along path 34 to playback/record device 41. The number of signals received by SDI router 31 is not significant.

30 Playback/record device 41 represents an arbitrary device that uses the signal passed along path 34. For example, it can be a recording device such as VTR or a playback device such as a television set. Furthermore, playback/record 41 may be located

remotely with respect to SDI embedder 31, in which case path 34 represents a communication or broadcast channel.

Shifts in Video/Audio Alignment

Circuit delays in VTR 16, 17, 18 and in SDI embedder 11 and 12 may alter the relative alignment of video information and audio information. As a result, the alignment of video/audio information in playback signal 21, for example, may be shifted with respect to the alignment between video information and audio information as received from paths 1 and 2, respectively. The amount of the change in alignment varies among equipment from different manufacturers, can vary among different pieces of equipment from the same manufacturer, and can even vary within a given piece of equipment as a function of the initialization state of buffers, for example.

Referring to Fig. 2A, signal 111 represents audio information having a particular alignment with video frame references 101 and 102. Each of these video frame references indicates a particular reference point in a respective video frame. A common reference point for NTSC video information, for example, coincides with the video information for line 10 in each frame. A common reference point for PAL video information coincides with line 1 in each frame. No particular alignment is critical to the practice of the present invention.

In Fig. 2B, signal 121 represents the same audio information as that conveyed by signal 111 but is delayed with respect to signal 111. As a result, the alignment between signal 121 and the video frame references is shifted with respect to the alignment for signal 111. In Fig. 2C, signal 131 represents the same audio information as that conveyed by signal 111 but is advanced with respect to signal 111; hence, the alignment between signal 131 and the video frame references is shifted in a manner opposite to the shift in alignment of signal 121.

Referring to Fig. 1, suppose the audio information and alignment represented in Fig. 2A is conveyed by paths 1/2, 3/4, 5/6 and 7/8. Different shifts in alignment such as those illustrated in Figs. 2A through 2C will likely exist in the signals generated along paths 21 through 24. Suppose further that the alignments illustrated in Figs. 2A through 2C exist in the signals generated along paths 21 through 23, respectively. When SDI router 31 switches between signals received from these three paths, a small discontinuity will occur in the audio information embedded in the

signal passed along path 34. If the audio information is represented in a linear form such as PCM, this discontinuity probably cannot be perceived by a human listener because the discontinuity is only for a few samples. It would be especially difficult to discern a discontinuity between two signals having different audio content.

5

Effects of Encoding

As explained above, however, there is a growing interest to embed greater numbers of audio channels into a video/audio data stream. When the capacity of the information in these greater number of audio channels exceeds the capacity of the space available for audio information, some form of bandwidth or bit-rate
10 compression is employed. One example of such compression is audio coding based on psychoacoustic principles.

These encoding techniques are often applied to blocks of audio samples to generate blocks of encoded information. These sample blocks typically represent audio information that spans an interval of 3 to 12 ms. Each block of encoded
15 information generated by these encoding processes represents the smallest unit of information from which a reasonably accurate replica of a segment of the original audio information can be recovered.

A sequence of encoded information blocks 112 is represented as a train of pulses in Fig. 2A. The information conveyed by these blocks is an encoded
20 representation of the audio information in signal 111. The shape and size of the pulses is not significant. The pulse train is intended only to suggest a sequence of blocks that convey encoded information corresponding to blocks of audio samples that may abut one another or, preferably, overlap one another. In the example shown in Fig. 2A, the audio information spanning the interval between adjacent video frame references is
25 represented by six blocks of encoded information. Various considerations for improving the quality of audio coding in video/audio applications is disclosed in WO-A 99/21187.

When block encoding techniques are used in the system of Fig. 1, the signals that SDI router 31 receives from paths 21 through 24 contain audio information
30 encoded in blocks. As discussed above, varying shifts in alignment may occur between the encoded information blocks and the video frame references. This is illustrated by the different alignments between video frame reference 101, for example, and blocks 112, 122 and 132 shown in Figs. 2A, 2B and 2C, respectively.

Suppose, as discussed above, that the alignments illustrated in Figs. 2A through 2C exist in the signals generated along paths 21 through 23, respectively. When SDI router 31 switches at video frame reference 101 from the signal received via path 22, illustrated in Fig. 2B, to the signal received via path 23, illustrated in Fig. 2C, a significant amount of audio information at the switch point cannot be recovered from the signal routed along path 23. The audio information conveyed in block 123 prior to the switch point cannot be recovered because, on the one hand, the entire block is needed to recover the audio information but, on the other hand, the portion of that block after the switch point is missing. Similarly, the audio information conveyed in block 133 after the switch point cannot be recovered because the portion of block 133 prior to the switch point is missing.

This problem is not unique to the type of system illustrated in Fig. 1. For example, the problem also occurs with tape edits or audio dubbing for a single VTR.

As will be explained more fully below, the present invention overcomes this problem by forming guard bands or gaps in the encoded audio stream so that considerable variation in video/audio alignment can be tolerated without loss of audio information.

Encoding Signal Processor

Fig. 3 illustrates a video/audio signal processor that can be incorporated in a variety of ways into a system such as that illustrated in Fig. 1. In the embodiment shown, a plurality of signals conveying video information with embedded audio information are received from input signal paths 61-1, 61-2 and 61-3. Three input signal paths are shown in the figure; however, embodiments of the present invention may have signal paths for essentially any number of input signals. Signal distributor 62 represents a wide range of signal distribution processes including switching, merging, editing, splicing and storage/retrieval. For simplicity, the illustration and discussion herein assume signal distributor 62 receives a plurality of video/audio signals and processes and/or distributes those signals in some manner to generate along path 63 a single signal conveying video information with embedded audio information. Deformatter 64 receives the video/audio information from path 63, extracts the embedded audio information and passes it along path 65. The video information may be passed along path 69. Audio signal processor 66 receives audio information from path 65 and applies a block-encoding process to the audio

information to generate along path 67 blocks of encoded information. Formatter 68 receives the blocks of encoded information from path 67 and generates along path 70 an output signal comprising a plurality of sequences of blocks of encoded information with gaps or guard bands between a beginning block in one sequence and an ending block in a preceding sequence. Using a reference signal such as a master clock signal, the gap or guard band is time aligned with the video information.

As mentioned above, the figures do not illustrate signal paths for carrying master clock signals used to synchronize the equipment. In a preferred embodiment, audio signal processor 66 forms audio sample blocks that are aligned with the master clock signal. This alignment is illustrated in Fig. 2A where boundaries between adjacent sample blocks coincide with video frame references 101 and 102; however, other alignments may be used.

Referring to Fig. 5A, sequence of blocks 112-2 conveys encoded information representing signal segment 111-2, which is a hypothetical time-compressed representation of the portion of signal 111 between video frame references 101 and 102. Similarly, sequence of blocks 112-1 conveys encoded information representing signal segment 111-1 and sequence of blocks 112-3 conveys encoded information representing signal segment 111-3. Audio signal processor 66 and formatter 68 generate sequences of blocks conveying an encoded representation of the audio information in which, for example, a guard band or gap has been formed between the ending block in sequence 112-1 and the beginning block in sequence 112-2.

The shifts in alignment illustrated in Figs. 2A through 2C are also illustrated in Figs. 5A through 5C. In these figures, encoded information in sequences 122-1, 122-2, 122-3, 132-1, 132-2 and 132-3 convey encoded information representing signal segments 121-1, 121-2, 121-3, 131-1, 131-2 and 131-3, respectively. As may be seen from Figs. 5B and 5C, no loss in audio information occurs as a result of shifts in alignment because potential switch points at video frame references 101 and 102 occur within a guard band.

The signal processor illustrated in Fig. 3 can be incorporated into an SDI router, for example, to process video signals containing embedded AES3 or PCM audio information. An embodiment omitting signal distributor 62 can be incorporated into a VTR or SDI embedder. Another embodiment that also omits deformatter 64 can be incorporated into a VTR or into the input circuits of an SDI embedder.

Fig. 4 illustrates one embodiment of an encoding audio signal processor that is suitable for incorporation into the embodiment illustrated in Fig. 3 and also has separate utility as will be explained below. According to this embodiment, audio signal processor 66 comprises a plurality of filterbanks 71, 72 and 73. Filterbank 71 generates a plurality of frequency subband signals along paths 75-1 through 75-3 in response to the signal received from path 65-1. Filterbank 72 generates a plurality of frequency subband signals along paths 76-1 through 76-3 in response to the signal received from path 65-2. Filterbank 73 generates a plurality of frequency subband signals along paths 77-1 through 77-3 in response to the signal received from path 65-3. Filterbanks 71, 72 and 73 may be implemented in a number of ways including a bank of bandpass filters, a cascading set of bandsplitting filters, and one or more time-domain-to-frequency-domain transforms. Only three filterbanks are shown and only three subband signals are shown for each filterbank, however, an embodiment may include many more filterbanks that each generate twenty-four or more subband signals, each representing frequency subbands having bandwidths commensurate with or less than the critical bandwidths of the human auditory system. Encoder 79 applies a block encoding process to the subband signals and generate along path 67 a sequence of blocks representing in an encoded form the audio information received via paths 65-1, 65-2 and 65-3.

Split-band coding is not essential to practice the present invention. Other forms of coding such as block-companded PCM or delta-modulation may be used.

In one practical embodiment, an encoding audio signal processor receives eight channels of audio information in linear PCM form or, alternatively, four AES3 data streams, and uses eight filterbanks and an encoder that applies a block encoding process to generate blocks of encoded information with guard bands that can be conveyed in a space or bandwidth no larger than that required to convey two channels of audio information in linear PCM form or, alternatively, one AES3 data stream.

Overlapping Blocks and Window Functions

The train of pulses used in the figures to represent blocks of information suggest that adjacent blocks abut but do not overlap one another. Although no particular arrangement of blocks is critical to the practice of the present invention, preferred embodiments process blocks that overlap one another. Generally, overlapped blocks of audio information are weighted or modulated by a window

function so that the sum of overlapped samples in adjacent blocks is substantially equal to a constant.

Fig 6 illustrates a sequence of blocks. Beginning block 141 in the sequence overlaps adjacent block 142. All blocks in the sequence are represented by an envelope that has the shape of a window function used to weight corresponding audio information in the time domain. Ending block 146 in the sequence overlaps the preceding block and a subsequent block not shown in the figure. The amount of overlap and the choice of window function can have significant effects on coding performance but no particular window function or amount of overlap is critical to the present invention. In preferred embodiments, the amount of overlap is equal to one-half the block length and the window function is derived from the Kaiser-Bessel function.

As explained above, audio signal processor 86 generates audio information that is aligned with video frame references. In embodiments that generate sequences of blocks of audio information, the alignment may be such that a video frame reference is coincident with essentially any point in any block of the sequence. In the example shown in Fig. 6, the start of beginning block 141 is coincident with video frame reference 100.

In some applications, the exact point of coincidence may vary from video frame to video frame. For example, in applications that combine digital audio information with NTSC video information, successive video frames may have varying numbers of audio samples because the audio sampling rate is not an integer multiple of the video frame rate.

Various considerations for block length, window function, and video/audio alignment are discussed in WO-A 99/21187, referred to above.

CLAIMS

1. A method for processing audio information comprising:

receiving an input audio signal conveying said audio information,

receiving video frame references indicating time references for a sequence of video frames,

generating blocks of encoded audio information from said audio information in a bit-rate reduced form by applying a block-encoding process to said input audio signal and time compressing said blocks of encoded audio information, and

assembling said time-compressed blocks into an encoded audio stream comprising a plurality of sequences of said time-compressed blocks such that a beginning block in a respective sequence is separated from an ending block in a preceding sequence by a gap that is time aligned with a respective video frame reference.

2. A method according to claim 1 wherein said block-encoding process encodes said audio information into a form that contains less redundancy and/or less perceptual irrelevance.

3. A method according to claim 2 wherein said block-encoding process comprises:

applying either a bank of bandpass filters or one or more transforms to said input audio signal to generate a plurality of frequency subband representations of said input audio signal, and

generating said blocks of encoded audio information by adaptively allocating bits to said frequency subband representations according to psychoacoustic principles.

4. A method according to any one of claims 1 through 3 wherein said video frames comprise video information arranged in a plurality of video lines and each of said gaps have a duration greater than that of three video lines.

5. A method according to any one of claims 1 through 3 wherein each of said video frames has a duration equal to a frame interval and the duration of said gaps is within a range from about one percent of said frame interval to about twenty percent of said frame interval.

6. A method according to any one of claims 1 through 5 that applies said block-encoding process to overlapping blocks of said audio information.

5 7. An apparatus for processing audio information comprising:
 means for receiving an input audio signal conveying said audio information,
 means for receiving video frame references indicating time references for a
sequence of video frames,
 means for generating blocks of encoded audio information from said audio
10 information in a bit-rate reduced form by applying a block-encoding process to said
input audio signal and time compressing said blocks of encoded audio information,
and
 means for assembling said time-compressed blocks into an encoded audio
stream comprising a plurality of sequences of said time-compressed blocks such that a
15 beginning block in a respective sequence is separated from an ending block in a
preceding sequence by a gap that is time aligned with a respective video frame
reference.

20 8. An apparatus according to claim 7 wherein said block-encoding process encodes
said audio information into a form that contains less redundancy and/or less perceptual
irrelevance.

 9. An apparatus according to claim 8 wherein said block-encoding process
comprises:
25 means for applying either a bank of bandpass filters or one or more transforms
to said input audio signal to generate a plurality of frequency subband representations
of said input audio signal, and
 means for generating said blocks of encoded audio information by adaptively
allocating bits to said frequency subband representations according to psychoacoustic
30 principles.

10. An apparatus according to any one of claims 7 through 9 wherein said video frames comprise video information arranged in a plurality of video lines and each of said gaps have a duration greater than that of three video lines.

5 11. An apparatus according to any one of claims 7 through 9 wherein each of said video frames has a duration equal to a frame interval and the duration of said gaps is within a range from about one percent of said frame interval to about twenty percent of said frame interval.

10 12. An apparatus according to any one of claims 7 through 11 wherein the means for generating said blocks of encoded audio information applies said block-encoding process to overlapping blocks of said audio information.