Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2448182 A1 2002/12/05

(21) 2 448 182

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2002/02/26
(87) Date publication PCT/PCT Publication Date: 2002/12/05
(85) Entrée phase nationale/National Entry: 2003/11/24
(86) N° demande PCT/PCT Application No.: US 2002/005999
(87) N° publication PCT/PCT Publication No.: 2002/097792
(30) Priorités/Priorities: 2001/05/25 (60/293,825) US; 2002/01/11 (10/045,644) US; 2002/01/23 (60/351,498) US; 2002/02/12 (PCT/US02/04317) US

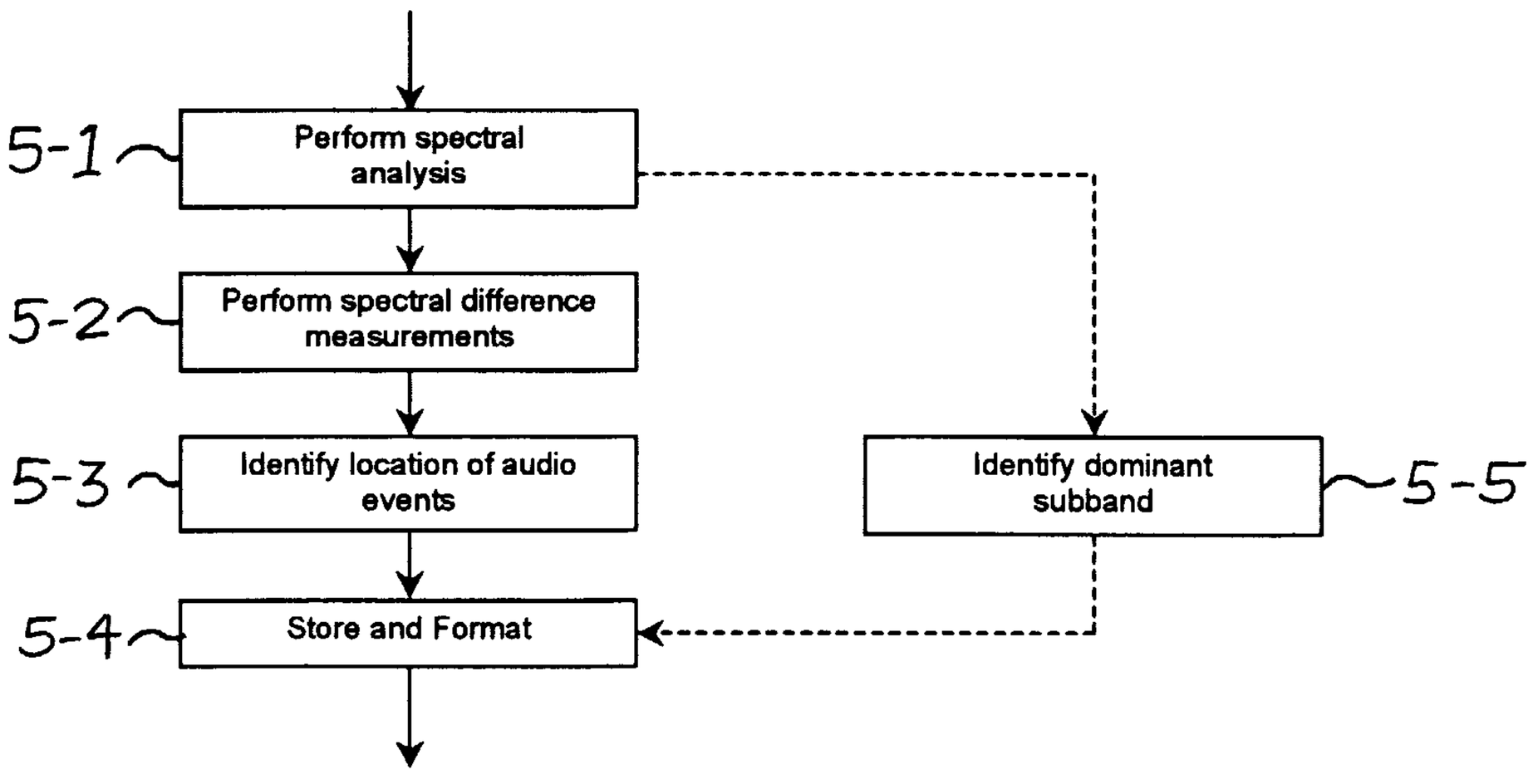
(51) Cl.Int.^{7}/Int.Cl.^{7} G10L 11/00

(71) Demandeur/Applicant:
DOLBY LABORATORIES LICENSING CORPORATION, US

(72) Inventeur/Inventor:
CROCKETT, BRETT G., US

(74) Agent: SMART & BIGGAR

(54) Titre : SEGMENTATION DE SIGNAUX AUDIO EN EVENEMENTS AUDITIFS
(54) Title: SEGMENTING AUDIO SIGNALS INTO AUDITORY EVENTS



(57) Abrégé/Abstract:
In one aspect, the invention divides an audio signal into auditory events, each of which tends to be perceived as separate and distinct, by calculating the spectral content of successive time blocks of the audio signal (5-1), calculating the difference in spectral content between successive time blocks of the audio signal (5-2), and identifying an auditory event boundary as the boundary between successive time blocks when the difference in the spectral content between such successive time blocks exceeds a threshold (5-3). In another aspect, the invention generates a reduced-information representation of an audio signal by dividing an audio signal into auditory events, each of which tends to be perceived as separate and distinct, and formatting and storing information relating to the auditory events (5-4). Optionally, the invention may also assign a characteristic to one or more of the auditory events (5-5).

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 December 2002 (05.12.2002)

PCT

(10) International Publication Number
WO 02/097792 A1

(51) International Patent Classification[7]: G10L 11/00

(21) International Application Number: PCT/US02/05999

(22) International Filing Date: 26 February 2002 (26.02.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/293,825	25 May 2001 (25.05.2001)	US
10/045,644	11 January 2002 (11.01.2002)	US
60/351,498	23 January 2002 (23.01.2002)	US
PCT/US02/04317	12 February 2002 (12.02.2002)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	60/293,825 (CIP)
Filed on	25 May 2001 (25.05.2001)
US	60/351,498 (CIP)
Filed on	23 January 2002 (23.01.2002)
US	10/045,644 (CIP)
Filed on	11 January 2002 (11.01.2002)
US	PCT/US02/04317 (CIP)
Filed on	12 February 2002 (12.02.2002)

(71) Applicant *(for all designated States except US)*: **DOLBY LABORATORIES LICENSING CORPORATION** [US/US]; 100 Potrero Avenue, San Francisco, CA 94103 (US).

(72) Inventor; and

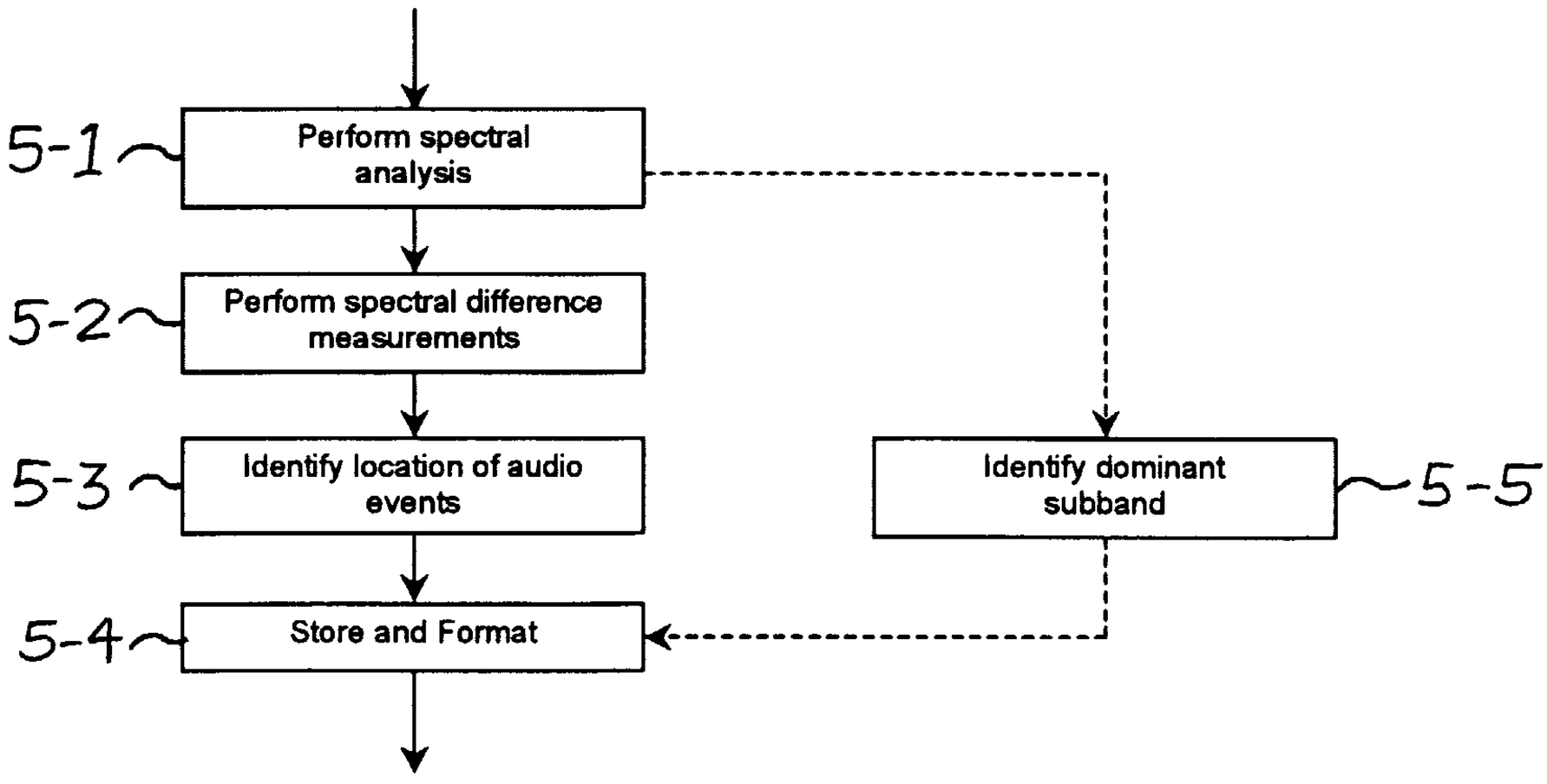
(75) Inventor/Applicant *(for US only)*: **CROCKETT, Brett, G.** [US/US]; 100 Potrero Avenue, San Francisco, CA 94103 (US).

(74) Agents: **GALLAGHER, Thomas, A.** et al.; Gallagher & Lathrop, 601 California Street, Suite 1111, San Francisco, CA 94108-2805 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT (utility model), AT, AU (petty patent), AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EC, EE (utility model), EE, ES, FI (utility model), FI,

[Continued on next page]

(54) Title: SEGMENTING AUDIO SIGNALS INTO AUDITORY EVENTS



(57) Abstract: In one aspect, the invention divides an audio signal into auditory events, each of which tends to be perceived as separate and distinct, by calculating the spectral content of successive time blocks of the audio signal (5-1), calculating the difference in spectral content between successive time blocks of the audio signal (5-2), and identifying an auditory event boundary as the boundary between successive time blocks when the difference in the spectral content between such successive time blocks exceeds a threshold (5-3). In another aspect, the invention generates a reduced-information representation of an audio signal by dividing an audio signal into auditory events, each of which tends to be perceived as separate and distinct, and formatting and storing information relating to the auditory events (5-4). Optionally, the invention may also assign a characteristic to one or more of the auditory events (5-5).

WO 02/097792 A1

WO 02/097792 A1

GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

DESCRIPTION

Segmenting Audio Signals into Auditory Events

TECHNICAL FIELD

The present invention pertains to the field of psychoacoustic processing of audio signals. In particular, the invention relates to aspects of dividing or segmenting audio signals into "auditory events," each of which tends to be perceived as separate and distinct, and to aspects of generating reduced-information representations of audio signals based on auditory events and, optionally, also based on the characteristics or features of audio signals within such auditory events. Auditory events may be useful as defining the MPEG-7 "Audio Segments" as proposed by the "ISO/IEC JTC 1/SC 29/WG 11."

BACKGROUND ART

The division of sounds into units or segments perceived as separate and distinct is sometimes referred to as "auditory event analysis" or "auditory scene analysis" ("ASA"). An extensive discussion of auditory scene analysis is set forth by Albert S. Bregman in his book *Auditory Scene Analysis - The Perceptual Organization of Sound*, Massachusetts Institute of Technology, 1991, Fourth printing, 2001, Second MIT Press paperback edition.) In addition, United States Patent 6,002,776 to Bhadkamkar, et al, December 14, 1999 cites publications dating back to 1976 as "prior art work related to sound separation by auditory scene analysis." However, the Bhadkamkar, et al patent discourages the practical use of auditory scene analysis, concluding that "[t]echniques involving auditory scene analysis, although interesting from a scientific point of view as models of human auditory processing, are currently far too computationally demanding and specialized to be considered practical techniques for sound separation until fundamental progress is made."

There are many different methods for extracting characteristics or features from audio. Provided the features or characteristics are suitably defined, their extraction can be performed using automated processes. For example "ISO/IEC JTC 1/SC 29/WG 11" (MPEG) is currently standardizing a variety of audio descriptors as part of the MPEG-7 standard. A common shortcoming of such methods is that they ignore auditory scene analysis. Such methods seek to measure, periodically, certain "classical" signal processing parameters such as pitch, amplitude, power, harmonic structure and spectral flatness. Such parameters, while providing useful information, do not analyze and characterize audio signals into elements perceived as separate and distinct according to human cognition. However, MPEG-7 descriptors may be useful in characterizing an Auditory Event identified in accordance with aspects of the present invention.

DISCLOSURE OF THE INVENTION

In accordance with aspects of the present invention, a computationally efficient process for dividing audio into temporal segments or "auditory events" that tend to be perceived as separate and distinct is provided. The locations of the boundaries of these auditory events (where they begin and end with respect to time) provide valuable information that can be used to describe an audio signal. The locations of auditory event boundaries can be assembled to generate a reduced-information representation, "signature, or "fingerprint" of an audio signal that can be stored for use, for example, in comparative analysis with other similarly generated signatures (as, for example, in a database of known works).

Bregman notes that "[w]e hear discrete units when the sound changes abruptly in timbre, pitch, loudness, or (to a lesser extent) location in space." (*Auditory Scene Analysis - The Perceptual Organization of Sound*, supra at page 469). Bregman also discusses the perception of multiple simultaneous sound streams when, for example, they are separated in frequency.

In order to detect changes in timbre and pitch and certain changes in amplitude, the audio event detection process according to an aspect of the present

invention detects changes in spectral composition with respect to time. When applied to a multichannel sound arrangement in which the channels represent directions in space, the process according to an aspect of the present invention also detects auditory events that result from changes in spatial location with respect to time. Optionally, according to a further aspect of the present invention, the process may also detect changes in amplitude with respect to time that would not be detected by detecting changes in spectral composition with respect to time.

In its least computationally demanding implementation, the process divides audio into time segments by analyzing the entire frequency band (full bandwidth audio) or substantially the entire frequency band (in practical implementations, band limiting filtering at the ends of the spectrum is often employed) and giving the greatest weight to the loudest audio signal components. This approach takes advantage of a psychoacoustic phenomenon in which at smaller time scales (20 milliseconds (ms) and less) the ear may tend to focus on a single auditory event at a given time. This implies that while multiple events may be occurring at the same time, one component tends to be perceptually most prominent and may be processed individually as though it were the only event taking place. Taking advantage of this effect also allows the auditory event detection to scale with the complexity of the audio being processed. For example, if the input audio signal being processed is a solo instrument, the audio events that are identified will likely be the individual notes being played. Similarly for an input voice signal, the individual components of speech, the vowels and consonants for example, will likely be identified as individual audio elements. As the complexity of the audio increases, such as music with a drumbeat or multiple instruments and voice, the auditory event detection identifies the "most prominent" (*i.e.*, the loudest) audio element at any given moment. Alternatively, the most prominent audio element may be determined by taking hearing threshold and frequency response into consideration.

While the locations of the auditory event boundaries computed from full-bandwidth audio provide useful information related to the content of an audio signal, it might be desired to provide additional information further describing the content of

an auditory event for use in audio signal analysis. For example, an audio signal could be analyzed across two or more frequency subbands and the location of frequency subband auditory events determined and used to convey more detailed information about the nature of the content of an auditory event. Such detailed information could provide additional information unavailable from wideband analysis.

Thus, optionally, according to further aspects of the present invention, at the expense of greater computational complexity, the process may also take into consideration changes in spectral composition with respect to time in discrete frequency subbands (fixed or dynamically determined or both fixed and dynamically determined subbands) rather than the full bandwidth. This alternative approach would take into account more than one audio stream in different frequency subbands rather than assuming that only a single stream is perceptible at a particular time.

Even a simple and computationally efficient process according to aspects of the present invention has been found usefully to identify auditory events.

An auditory event detecting process according to the present invention may be implemented by dividing a time domain audio waveform into time intervals or blocks and then converting the data in each block to the frequency domain, using either a filter bank or a time-frequency transformation, such as the FFT. The amplitude of the spectral content of each block may be normalized in order to eliminate or reduce the effect of amplitude changes. Each resulting frequency domain representation provides an indication of the spectral content (amplitude as a function of frequency) of the audio in the particular block. The spectral content of successive blocks is compared and changes greater than a threshold may be taken to indicate the temporal start or temporal end of an auditory event. FIG. 1 shows an idealized waveform of a single channel of orchestral music illustrating auditory events. The spectral changes that occur as a new note is played trigger the new auditory events 2 and 3 at samples 2048 and 2560, respectively.

As mentioned above, in order to minimize the computational complexity, only a single band of frequencies of the time domain audio waveform may be processed,

preferably either the entire frequency band of the spectrum (which may be about 50 Hz to 15 kHz in the case of an average quality music system) or substantially the entire frequency band (for example, a band defining filter may exclude the high and low frequency extremes).

Preferably, the frequency domain data is normalized, as is described below. The degree to which the frequency domain data needs to be normalized gives an indication of amplitude. Hence, if a change in this degree exceeds a predetermined threshold, that too may be taken to indicate an event boundary. Event start and end points resulting from spectral changes and from amplitude changes may be ORed together so that event boundaries resulting from either type of change are identified.

In the case of multiple audio channels, each representing a direction in space, each channel may be treated independently and the resulting event boundaries for all channels may then be ORed together. Thus, for example, an auditory event that abruptly switches directions will likely result in an "end of event" boundary in one channel and a "start of event" boundary in another channel. When ORed together, two events will be identified. Thus, the auditory event detection process of the present invention is capable of detecting auditory events based on spectral (timbre and pitch), amplitude and directional changes.

As mentioned above, as a further option, but at the expense of greater computational complexity, instead of processing the spectral content of the time domain waveform in a single band of frequencies, the spectrum of the time domain waveform prior to frequency domain conversion may be divided into two or more frequency bands. Each of the frequency bands may then be converted to the frequency domain and processed as though it were an independent channel in the manner described above. The resulting event boundaries may then be ORed together to define the event boundaries for that channel. The multiple frequency bands may be fixed, adaptive, or a combination of fixed and adaptive. Tracking filter techniques employed in audio noise reduction and other arts, for example, may be employed to define adaptive frequency bands (*e.g.*, dominant simultaneous sine waves at 800 Hz and 2 kHz could result in two adaptively-determined bands centered on those two

frequencies). Although filtering the data before conversion to the frequency domain is workable, more optimally the full bandwidth audio is converted to the frequency domain and then only those frequency subband components of interest are processed. In the case of converting the full bandwidth audio using the FFT, only sub-bins corresponding to frequency subbands of interest would be processed together.

Alternatively, in the case of multiple subbands or multiple channels, instead of ORing together auditory event boundaries, which results in some loss of information, the event boundary information may be preserved.

As shown in FIG. 2, the frequency domain magnitude of a digital audio signal contains useful frequency information out to a frequency of Fs/2 where Fs is the sampling frequency of the digital audio signal. By dividing the frequency spectrum of the audio signal into two or more subbands (not necessarily of the same bandwidth and not necessarily up to a frequency of Fs/2 Hz), the frequency subbands may be analyzed over time in a manner similar to a full bandwidth auditory event detection method.

The subband auditory event information provides additional information about an audio signal that more accurately describes the signal and differentiates it from other audio signals. This enhanced differentiating capability may be useful if the audio signature information is to be used to identify matching audio signals from a large number of audio signatures. For example, as shown in FIG. 2, a frequency subband auditory event analysis (with a auditory event boundary resolution of 512 samples) has found multiple subband auditory events starting, variously, at samples 1024 and 1536 and ending, variously, at samples 2560, 3072 and 3584 . It is unlikely that this level of signal detail would be available from a single, wideband auditory scene analysis.

The subband auditory event information may be used to derive an auditory event signature for each subband. While this would increase the size of the audio signal's signature and possibly increase the computation time required to compare multiple signatures it could also greatly reduce the probability of falsely classifying two signatures as being the same. A tradeoff between signature size, computational

complexity and signal accuracy could be done depending upon the application. Alternatively, rather than providing a signature for each subband, the auditory events may be ORed together to provide a single set of "combined" auditory event boundaries (at samples 1024, 1536, 2560, 3072 and 3584. Although this would result in some loss of information, it provides a single set of event boundaries, representing combined auditory events, that provides more information than the information of a single subband or a wideband analysis.

While the frequency subband auditory event information on its own provides useful signal information, the relationship between the locations of subband auditory events may be analyzed and used to provide more insight into the nature of an audio signal. For example, the location and strength of the subband auditory events may be used as an indication of timbre (frequency content) of the audio signal. Auditory events that appear in subbands that are harmonically related to one another would also provide useful insight regarding the harmonic nature of the audio. The presence of auditory events in a single subband may also provide information as to the tone-like nature of an audio signal. Analyzing the relationship of frequency subband auditory events across multiple channels can also provide spatial content information.

In the case of analyzing multiple audio channels, each channel is analyzed independently and the auditory event boundary information of each may either be retained separately or be combined to provide combined auditory event information. This is somewhat analogous to the case of multiple subbands. Combined auditory events may be better understood by reference to FIG. 3 that shows the auditory scene analysis results for a two channel audio signal. FIG. 3 shows time concurrent segments of audio data in two channels. ASA processing of the audio in a first channel, the top waveform of FIG. 3, identifies auditory event boundaries at samples that are multiples of the 512 sample spectral-profile block size, 1024 and 1536 samples in this example. The lower waveform of FIG. 3 is a second channel and ASA processing results in event boundaries at samples that are also multiples of the spectral-profile block size, at samples 1024, 2048 and 3072 in this example. A combined auditory event analysis for both channels results in combined auditory

combined auditory event analysis for both channels results in combined auditory event segments with boundaries at samples 1024, 1536, 2048 and 3072 (the auditory event boundaries of the channels are "ORed" together). It will be appreciated that in practice the accuracy of auditory event boundaries depends on the size of the spectral-profile block size (N is 512 samples in this example) because event boundaries can occur only at block boundaries. Nevertheless, a block size of 512 samples has been found to determine auditory event boundaries with sufficient accuracy as to provide satisfactory results.

FIG. 3A shows three auditory events. These events include the (1) quiet portion of audio before the transient, (2) the transient event, and (3) the echo / sustain portion of the audio transient. A speech signal is represented in FIG. 3B having a predominantly high-frequency sibilance event, and events as the sibilance evolves or "morphs" into the vowel, the first half of the vowel, and the second half of the vowel.

FIG. 3 also shows the combined event boundaries when the auditory event data is shared across the time concurrent data blocks of two channels. Such event segmentation provides five combined auditory event regions (the event boundaries are ORed together).

FIG. 4 shows an example of a four channel input signal. Channels 1 and 4 each contain three auditory events and channels 2 and 3 each contain two auditory events. The combined auditory event boundaries for the concurrent data blocks across all four channels are located at sample numbers 512, 1024, 1536, 2560 and 3072 as indicated at the bottom of the FIG. 4.

In principle, the processed audio may be digital or analog and need not be divided into blocks. However, in practical applications, the input signals likely are one or more channels of digital audio represented by samples in which consecutive samples in each channel are divided into blocks of, for example 4096 samples (as in the examples of FIGS. 1, 3 and 4, above). In practical embodiments set forth herein, auditory events are determined by examining blocks of audio sample data preferably representing approximately 20 ms of audio or less, which is believed to be the shortest auditory event recognizable by the human ear. Thus, in practice, auditory

events are likely to be determined by examining blocks of, for example, 512 samples, which corresponds to about 11.6 ms of input audio at a sampling rate of 44.1 kHz, within larger blocks of audio sample data. However, throughout this document reference is made to "blocks" rather than "subblocks" when referring to the examination of segments of audio data for the purpose of detecting auditory event boundaries. Because the audio sample data is examined in blocks, in practice, the auditory event temporal start and stop point boundaries necessarily will each coincide with block boundaries. There is a trade off between real-time processing requirements (as larger blocks require less processing overhead) and resolution of event location (smaller blocks provide more detailed information on the location of auditory events).

Other aspects of the invention will be appreciated and understood as the detailed description of the invention is read and understood.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is an idealized waveform of a single channel of orchestral music illustrating auditory.

FIG. 2 is an idealized conceptual schematic diagram illustrating the concept of dividing full bandwidth audio into frequency subbands in order to identify subband auditory events. The horizontal scale is samples and the vertical scale is frequency.

FIG. 3 is a series of idealized waveforms in two audio channels, showing audio events in each channel and combined audio events across the two channels.

FIG. 4 is a series of idealized waveforms in four audio channels showing audio events in each channel and combined audio events across the four channels.

FIG. 5 is a flow chart showing the extraction of audio event locations and the optional extraction of dominant subbands from an audio signal in accordance with the present invention.

FIG. 6 is a conceptual schematic representation depicting spectral analysis in accordance with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

In accordance with an embodiment of one aspect of the present invention, auditory scene analysis is composed of three general processing steps as shown in a portion of FIG. 5. The first step 5-1 ("Perform Spectral Analysis") takes a time-domain audio signal, divides it into blocks and calculates a spectral profile or spectral content for each of the blocks. Spectral analysis transforms the audio signal into the short-term frequency domain. This can be performed using any filterbank, either based on transforms or banks of bandpass filters, and in either linear or warped frequency space (such as the Bark scale or critical band, which better approximate the characteristics of the human ear). With any filterbank there exists a tradeoff between time and frequency. Greater time resolution, and hence shorter time intervals, leads to lower frequency resolution. Greater frequency resolution, and hence narrower subbands, leads to longer time intervals.

The first step, illustrated conceptually in FIG. 6 calculates the spectral content of successive time segments of the audio signal. In a practical embodiment, the ASA block size is 512 samples of the input audio signal. In the second step 5-2, the differences in spectral content from block to block are determined ("Perform spectral profile difference measurements"). Thus, the second step calculates the difference in spectral content between successive time segments of the audio signal. As discussed above, a powerful indicator of the beginning or end of a perceived auditory event is believed to be a change in spectral content. In the third step 5-3 ("Identify location of auditory event boundaries"), when the spectral difference between one spectral-profile block and the next is greater than a threshold, the block boundary is taken to be an auditory event boundary. The audio segment between consecutive boundaries constitutes an auditory event. Thus, the third step sets an auditory event boundary between successive time segments when the difference in the spectral profile content between such successive time segments exceeds a threshold, thus defining auditory events. In this embodiment, auditory event boundaries define auditory events having a length that is an integral multiple of spectral profile blocks with a minimum length of one spectral profile block (512 samples in this example). In principle, event

boundaries need not be so limited. As an alternative to the practical embodiments discussed herein, the input block size may vary, for example, so as to be essentially the size of an auditory event.

The locations of event boundaries may be stored as a reduced-information characterization or "signature" and formatted as desired, as shown in step 5-4. An optional process step 5-5 ("Identify dominant subband") uses the spectral analysis of step 5-1 to identify a dominant frequency subband that may also be stored as part of the signature. The dominant subband information may be combined with the auditory event boundary information in order to define a feature of each auditory event.

Either overlapping or non-overlapping segments of the audio may be windowed and used to compute spectral profiles of the input audio. Overlap results in finer resolution as to the location of auditory events and, also, makes it less likely to miss an event, such as a transient. However, overlap also increases computational complexity. Thus, overlap may be omitted. FIG. 6 shows a conceptual representation of non-overlapping 512 sample blocks being windowed and transformed into the frequency domain by the Discrete Fourier Transform (DFT). Each block may be windowed and transformed into the frequency domain, such as by using the DFT, preferably implemented as a Fast Fourier Transform (FFT) for speed.

The following variables may be used to compute the spectral profile of the input block:

N = number of samples in the input signal

M = number of windowed samples in a block used to compute spectral profile

P = number of samples of spectral computation overlap

Q = number of spectral windows/regions computed

In general, any integer numbers may be used for the variables above. However, the implementation will be more efficient if M is set equal to a power of 2 so that standard FFTs may be used for the spectral profile calculations. In addition, if N, M, and P are chosen such that Q is an integer number, this will avoid under-

running or over-running audio at the end of the N samples. In a practical embodiment of the auditory scene analysis process, the parameters listed may be set to:

M = 512 samples (or 11.6 ms at 44.1 kHz)

P = 0 samples (no overlap)

The above-listed values were determined experimentally and were found generally to identify with sufficient accuracy the location and duration of auditory events. However, setting the value of P to 256 samples (50% overlap) rather than zero samples (no overlap) has been found to be useful in identifying some hard-to-find events. While many different types of windows may be used to minimize spectral artifacts due to windowing, the window used in the spectral profile calculations is an M-point Hanning, Kaiser-Bessel or other suitable, preferably non-rectangular, window. The above-indicated values and a Hanning window type were selected after extensive experimental analysis as they have shown to provide excellent results across a wide range of audio material. Non-rectangular windowing is preferred for the processing of audio signals with predominantly low frequency content. Rectangular windowing produces spectral artifacts that may cause incorrect detection of events. Unlike certain encoder/decoder (codec) applications where an overall overlap/add process must provide a constant level, such a constraint does not apply here and the window may be chosen for characteristics such as its time/frequency resolution and stop-band rejection.

In step 5-1 (FIG. 5), the spectrum of each M-sample block may be computed by windowing the data by an M-point Hanning, Kaiser-Bessel or other suitable window, converting to the frequency domain using an M-point Fast Fourier Transform, and calculating the magnitude of the complex FFT coefficients. The resultant data is normalized so that the largest magnitude is set to unity, and the normalized array of M numbers is converted to the log domain. The array need not be converted to the log domain, but the conversion simplifies the calculation of the difference measure in step 5-2. Furthermore, the log domain more closely matches the nature of the human auditory system. The resulting log domain values have a

range of minus infinity to zero. In a practical embodiment, a lower limit can be imposed on the range of values; the limit may be fixed, for example -60 dB, or be frequency-dependent to reflect the lower audibility of quiet sounds at low and very high frequencies. (Note that it would be possible to reduce the size of the array to M/2 in that the FFT represents negative as well as positive frequencies).

Step 5-2 calculates a measure of the difference between the spectra of adjacent blocks. For each block, each of the M (log) spectral coefficients from step 5-1 is subtracted from the corresponding coefficient for the preceding block, and the magnitude of the difference calculated (the sign is ignored). These M differences are then summed to one number. Hence, for a contiguous time segment of audio, containing Q blocks, the result is an array of Q positive numbers, one for each block. The greater the number, the more a block differs in spectrum from the preceding block. This difference measure may also be expressed as an average difference per spectral coefficient by dividing the difference measure by the number of spectral coefficients used in the sum (in this case M coefficients).

Step 5-3 identifies the locations of auditory event boundaries by applying a threshold to the array of difference measures from step 5-2 with a threshold value. When a difference measure exceeds a threshold, the change in spectrum is deemed sufficient to signal a new event and the block number of the change is recorded as an event boundary. For the values of M and P given above and for log domain values (in step 5-1) expressed in units of dB, the threshold may be set equal to 2500 if the whole magnitude FFT (including the mirrored part) is compared or 1250 if half the FFT is compared (as noted above, the FFT represents negative as well as positive frequencies — for the magnitude of the FFT, one is the mirror image of the other). This value was chosen experimentally and it provides good auditory event boundary detection. This parameter value may be changed to reduce (increase the threshold) or increase (decrease the threshold) the detection of events.

For an audio signal consisting of Q blocks (of size M samples), the output of step 5-3 of FIG. 5 may be stored and formatted in step 5-4 as an array $B(q)$ of information representing the location of auditory event boundaries where $q = 0, 1, . .$

. , Q-1. For a block size of M = 512 samples, overlap of P = 0 samples and a signal-sampling rate of 44.1kHz, the auditory scene analysis function 2 outputs approximately 86 values a second. The array $B(q)$ may stored as a signature, such that, in its basic form, without the optional dominant subband frequency information of step 5-5, the audio signal's signature is an array $B(q)$ representing a string of auditory event boundaries.

Identify dominant subband (optional)

For each block, an optional additional step in the processing of FIG. 5 is to extract information from the audio signal denoting the dominant frequency "subband" of the block (conversion of the data in each block to the frequency domain results in information divided into frequency subbands). This block-based information may be converted to auditory-event based information, so that the dominant frequency subband is identified for every auditory event. Such information for every auditory event provides information regarding the auditory event itself and may be useful in providing a more detailed and unique reduced-information representation of the audio signal. The employment of dominant subband information is more appropriate in the case of determining auditory events of full bandwidth audio rather than cases in which the audio is broken into subbands and auditory events are determined for each subband.

The dominant (largest amplitude) subband may be chosen from a plurality of subbands, three or four, for example, that are within the range or band of frequencies where the human ear is most sensitive. Alternatively, other criteria may be used to select the subbands. The spectrum may be divided, for example, into three subbands. Useful frequency ranges for the subbands are (these particular frequencies are not critical):

Subband 1	300 Hz to 550 Hz
Subband 2	550 Hz to 2000 Hz
Subband 3	2000 Hz to 10,000 Hz

To determine the dominant subband, the square of the magnitude spectrum (or the power magnitude spectrum) is summed for each subband. This resulting sum for

each subband is calculated and the largest is chosen. The subbands may also be weighted prior to selecting the largest. The weighting may take the form of dividing the sum for each subband by the number of spectral values in the subband, or alternatively may take the form of an addition or multiplication to emphasize the importance of a band over another. This can be useful where some subbands have more energy on average than other subbands but are less perceptually important.

Considering an audio signal consisting of Q blocks, the output of the dominant subband processing is an array $DS(q)$ of information representing the dominant subband in each block ($q = 0, 1, \ldots, Q\text{-}1$). Preferably, the array $DS(q)$ is formatted and stored in the signature along with the array $B(q)$. Thus, with the optional dominant subband information, the audio signal's signature is two arrays $B(q)$ and $DS(q)$, representing, respectively, a string of auditory event boundaries and a dominant frequency subband within each block, from which the dominant frequency subband for each auditory event may be determined if desired. Thus, in an idealized example, the two arrays could have the following values (for a case in which there are three possible dominant subbands).

1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	(Event Boundaries)
1	1	2	2	2	2	1	1	1	3	3	3	3	3	3	1	1	(Dominant Subbands)

In most cases, the dominant subband remains the same within each auditory event, as shown in this example, or has an average value if it is not uniform for all blocks within the event. Thus, a dominant subband may be determined for each auditory event and the array $DS(q)$ may be modified to provide that the same dominant subband is assigned to each block within an event.

The process of FIG. 5 may be represented more generally by the equivalent arrangements of FIGS. 7, 8 and 9. In FIG. 7, an audio signal is applied in parallel to an "Identify Auditory Events" function or step 7-1 that divides the audio signal into auditory events, each of which tends to be perceived as separate and distinct and to an optional "Identify Characteristics of Auditory Events" function or step 7-2. The process of FIG. 5 may be employed to divide the audio signal into auditory events or

some other suitable process may be employed. The auditory event information, which may be an identification of auditory event boundaries, determined by function or step 7-1 is stored and formatted, as desired, by a "Store and Format" function or step 7-3. The optional "Identify Characteristics" function or step 7-3 also receives the auditory event information. The "Identify Characteristics" function or step 7-3 may characterize some or all of the auditory events by one or more characteristics. Such characteristics may include an identification of the dominant subband of the auditory event, as described in connection with the process of FIG. 5. The characteristics may also include one or more of the MPEG-7 audio descriptors, including, for example, a measure of power of the auditory event, a measure of amplitude of the auditory event, a measure of the spectral flatness of the auditory event, and whether the auditory event is substantially silent. The characteristics may also include other characteristics such as whether the auditory event includes a transient. Characteristics for one or more auditory events are also received by the "Store and Format" function or step 7-3 and stored and formatted along with the auditory event information.

Alternatives to the arrangement of FIG. 7 are shown in FIGS. 8 and 9. In FIG. 8, the audio input signal is not applied directly to the "Identify Characteristics" function or step 8-3, but it does receive information from the "Identify Auditory Events" function or step 8-1. The arrangement of FIG. 5 is a specific example of such an arrangement. In FIG. 9, the functions or steps 9-1, 9-2 and 9-3 are arranged in series.

The details of this practical embodiment are not critical. Other ways to calculate the spectral content of successive time segments of the audio signal, calculate the differences between successive time segments, and set auditory event boundaries at the respective boundaries between successive time segments when the difference in the spectral profile content between such successive time segments exceeds a threshold may be employed.

It should be understood that implementation of other variations and modifications of the invention and its various aspects will be apparent to those skilled

in the art, and that the invention is not limited by these specific embodiments described. It is therefore contemplated to cover by the present invention any and all modifications, variations, or equivalents that fall within the true spirit and scope of the basic underlying principles disclosed and claimed herein.

The present invention and its various aspects may be implemented as software functions performed in digital signal processors, programmed general-purpose digital computers, and/or special purpose digital computers. Interfaces between analog and digital signal streams may be performed in appropriate hardware and/or as functions in software and/or firmware.

CLAIMS

1. A method for dividing each of multiple channels of digital audio signals into auditory events, each of which tends to be perceived as separate and distinct, comprising

detecting changes in signal characteristics with respect to time in the audio signal in each of the channels, and

identifying the auditory event boundaries in the audio signal of a channel, each boundary in response to a change in signal characteristics with respect to time in the channel exceeding a threshold such that a set of auditory event boundaries is obtained for each channel, wherein an audio segment in a channel between consecutive boundaries constitutes an auditory event.

2. A method according to claim 1 further comprising identifying a combined auditory event boundary for the channels in response to the identification of an auditory event boundary in any channel.

3. A method according to claim 2 wherein the audio in respective channels represent respective directions in space.

4. A method according to claim 2 wherein the audio in respective channels represent frequency bands of an audio signal.

5. A method according to claim 1 wherein the audio in respective channels represent respective directions in space.

6. A method according to claim 1 wherein the audio in respective channels represent bands of frequencies of an audio signal.

7. A method according to any one of claims 1-6 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels includes dividing the audio signal into time blocks and converting the data in each block to the frequency domain.

8. A method according to claim 7 wherein said signal characteristics are one of:

(a) spectral content, or

(b) spectral content and amplitude content.

9. A method according to claim 7 wherein said signal characteristics are spectral content.

10. A method according to claim 9 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels detects the changes in spectral content between successive time blocks of the audio signal in each of the channels.

11. The method claim 10 wherein the audio data in consecutive time blocks is represented by coefficients and said detecting changes in spectral content between successive time blocks of the audio signal in each of the channels includes

subtracting coefficients of a block from corresponding coefficients of an adjacent block.

12. The method of claim 11 wherein detecting changes in spectral content between successive time blocks of the audio signal in each of the channels further includes

summing the magnitudes of the differences resulting from subtracting coefficients of a block from corresponding coefficients of an adjacent block and comparing the summed magnitudes to a threshold.

13. The method of claim 12 wherein an auditory event boundary is identified when the summed magnitudes exceed said threshold.

14. The method of claim 7 wherein said method further comprises assigning a characteristic to one or more of the auditory events.

15. The method of claim 14 wherein characteristics assignable to one or more of the auditory events include one or more of: the dominant subband of the frequency spectrum of the auditory event, a measure of power of the auditory event, a measure of amplitude of the auditory event, a measure of the spectral flatness of the auditory event, whether the auditory event is substantially silent, and whether the auditory event includes a transient.

16. The method of claim 15 further comprising formatting and storing the auditory event boundaries and identification of characteristics assigned to auditory events.

17. The method of any one of claims 7 further comprising formatting and storing the auditory event boundaries.

18. The method of claim 7 further comprising normalizing the frequency domain data.

19. The method of claim 18 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels detects changes in spectral content between successive time blocks of the audio signal in each of the channels.

20. The method of claim 19 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels further detects changes in amplitude between successive time blocks of the audio signal in each of the channels.

21. The method of claim 20 wherein said changes in amplitude are detected by the degree to which the frequency domain data is normalized.

22. The method of claim 7 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels further comprises normalizing the frequency domain data and converting the normalized data to the log domain.

23. The method of claim 22 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels further comprises detecting changes in spectral content between successive time blocks of the audio signal in each of the channels.

24. The method of claim 23 wherein said detecting changes in signal characteristics with respect to time in the audio signal in each of the channels

further includes detecting changes in amplitude between successive time blocks of the audio signal in each of the channels.

25. The method of claim 24 wherein said changes in amplitude are detected by the degree to which the frequency domain data is normalized.

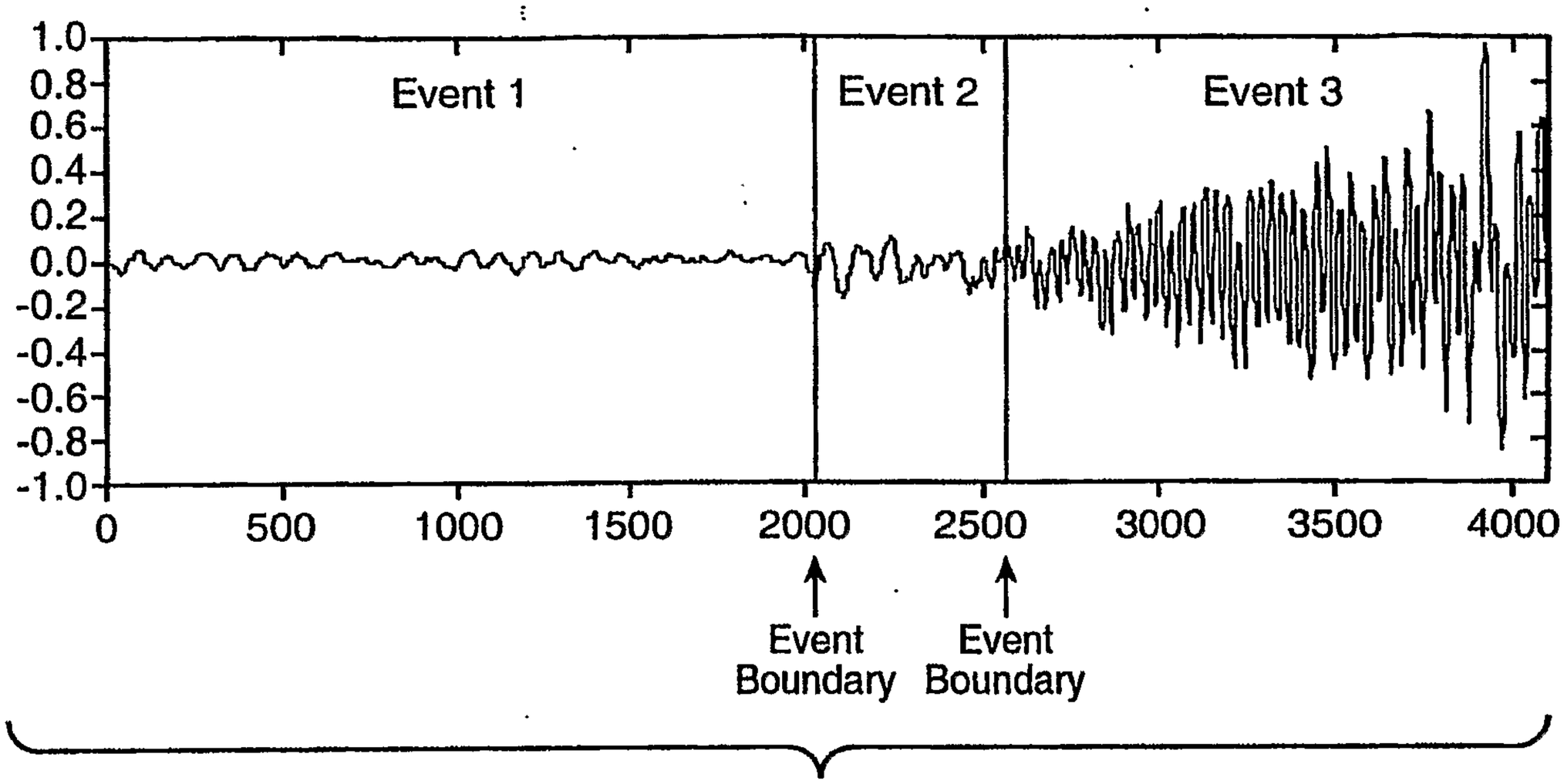

1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Event 1
Event 2
Event 3
0
500
1000
1500
2000
2500
3000
3500
4000
Event Boundary
Event Boundary

FIG._1

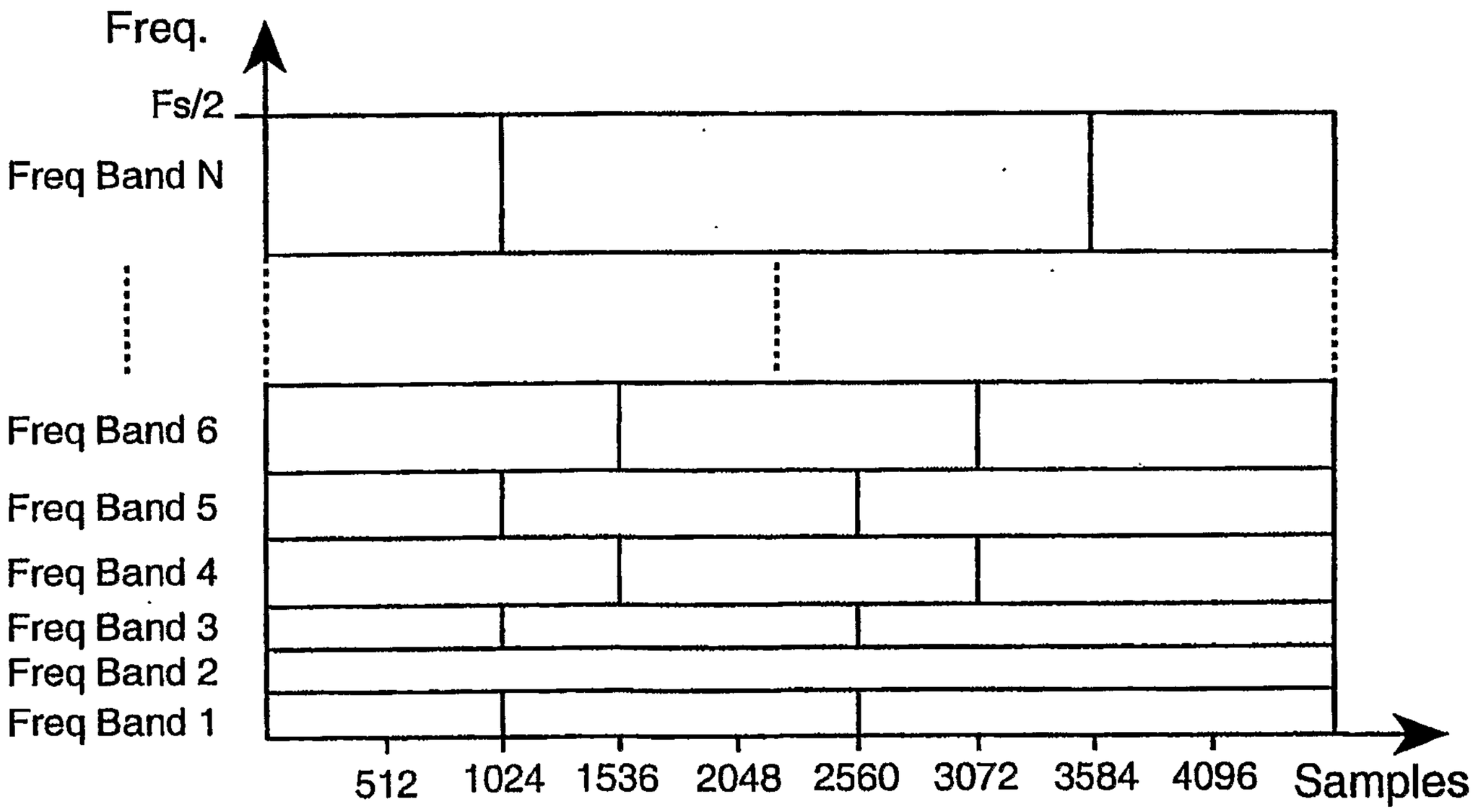

Freq.
Fs/2
Freq Band N
Freq Band 6
Freq Band 5
Freq Band 4
Freq Band 3
Freq Band 2
Freq Band 1
512
1024
1536
2048
2560
3072
3584
4096
Samples

FIG._2

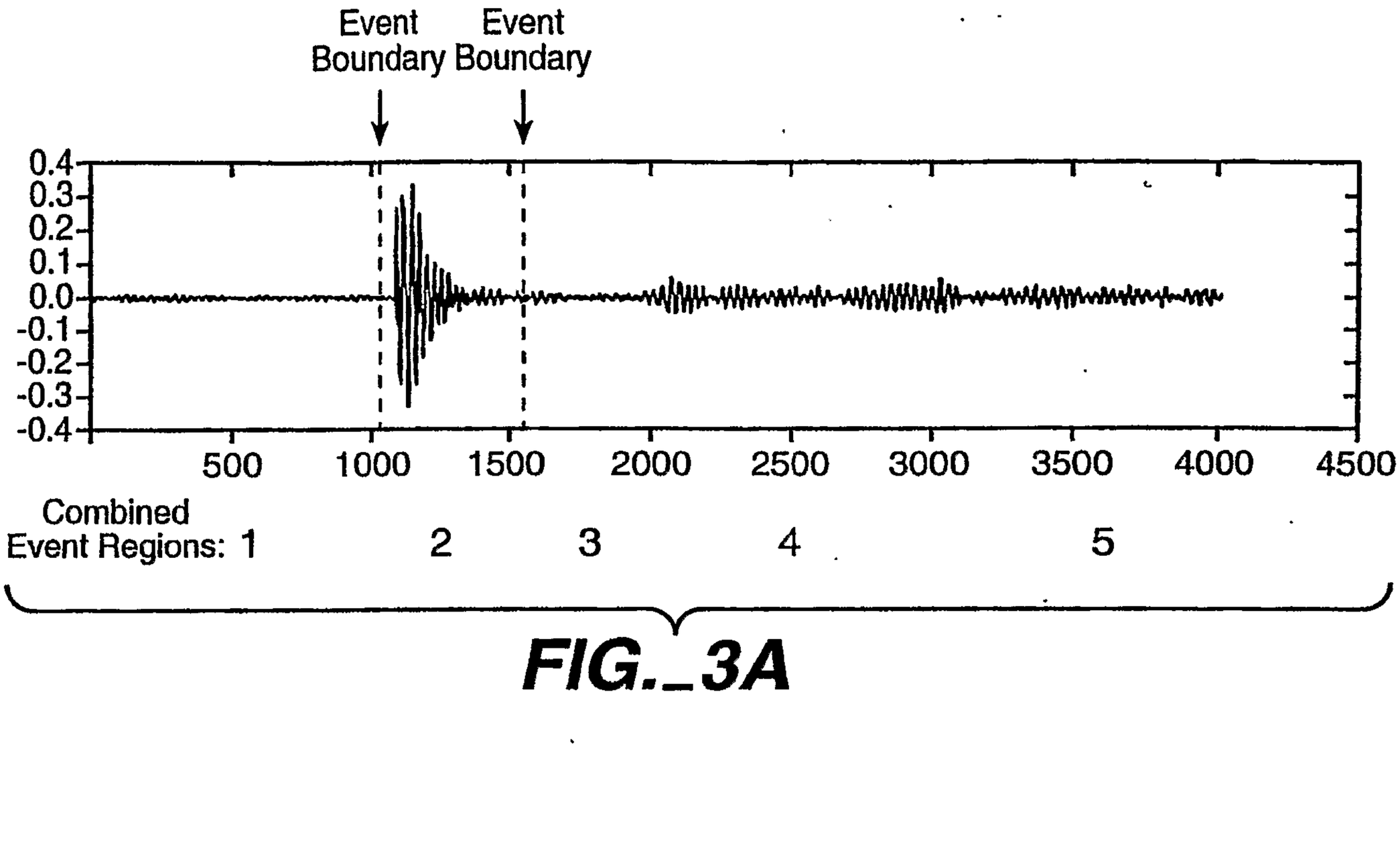

Event Boundary
Event Boundary
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
500
1000
1500
2000
2500
3000
3500
4000
4500
Combined Event Regions: 1
2
3
4
5

FIG._3A

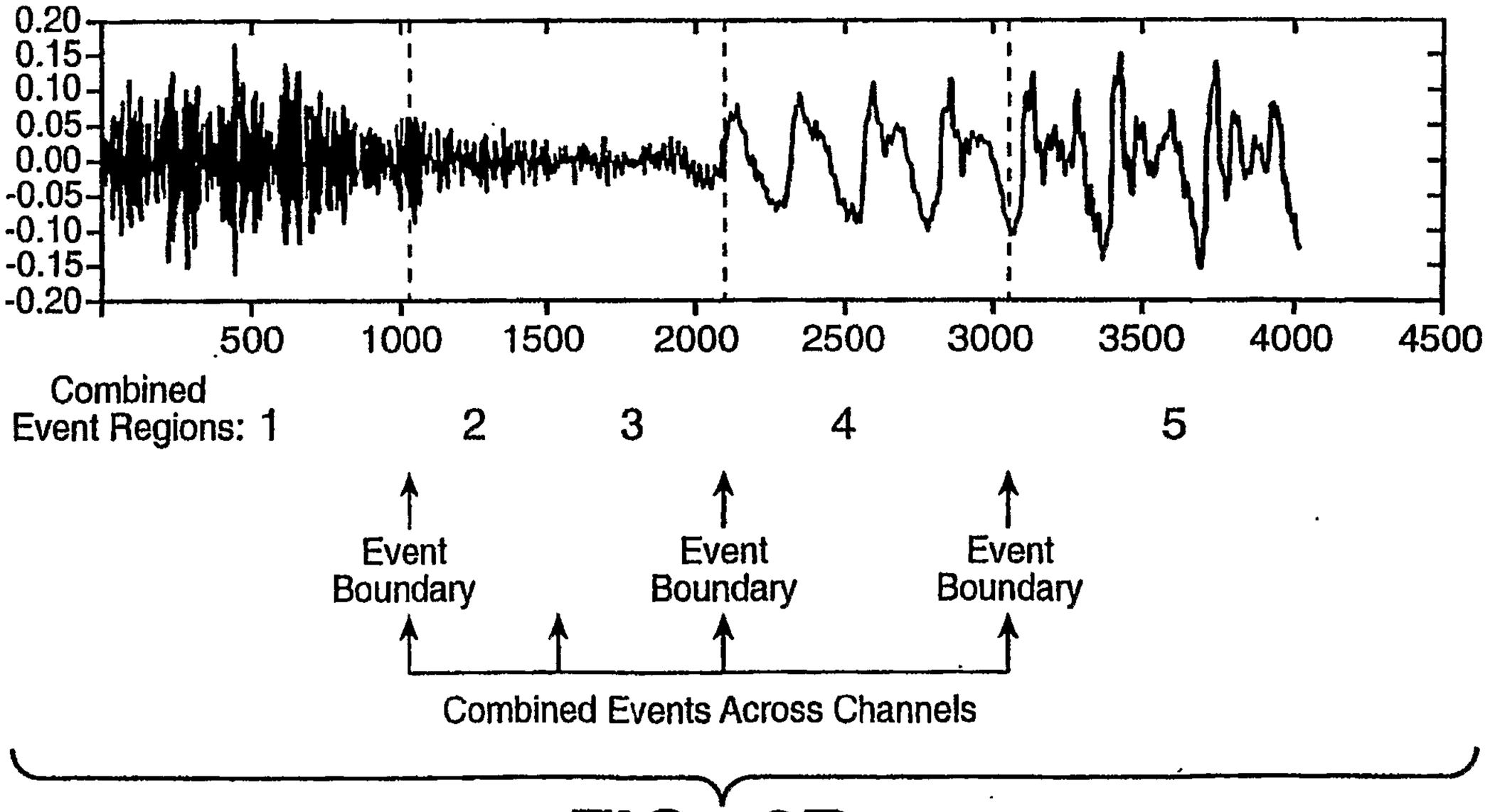

0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
500
1000
1500
2000
2500
3000
3500
4000
4500
Combined Event Regions: 1
2
3
4
5
Event Boundary
Event Boundary
Event Boundary
Combined Events Across Channels

FIG._3B

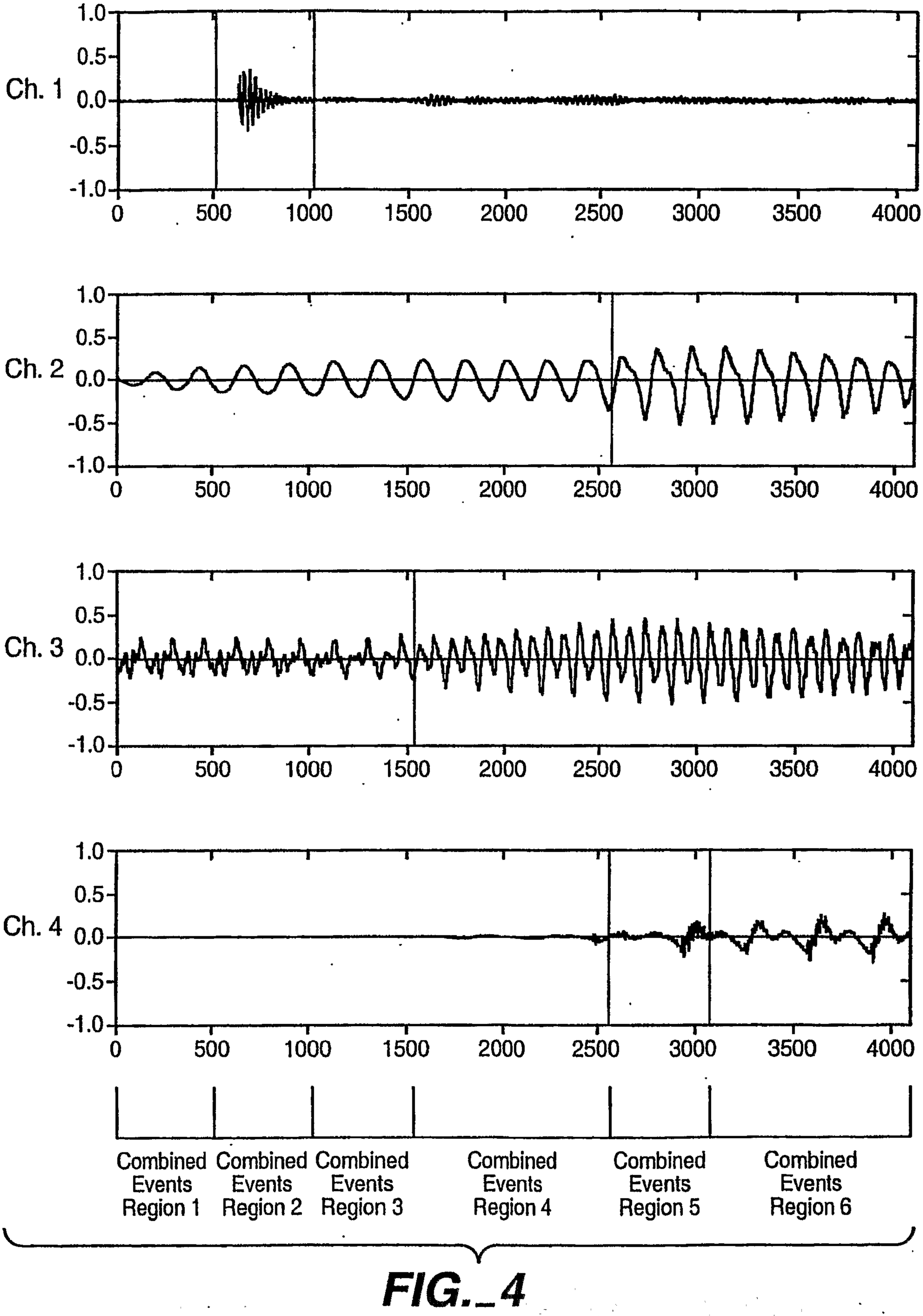

Ch. 1
Ch. 2
Ch. 3
Ch. 4
1.0
0.5
0.0
-0.5
-1.0
0
500
1000
1500
2000
2500
3000
3500
4000
Combined Events Region 1
Combined Events Region 2
Combined Events Region 3
Combined Events Region 4
Combined Events Region 5
Combined Events Region 6

FIG._4

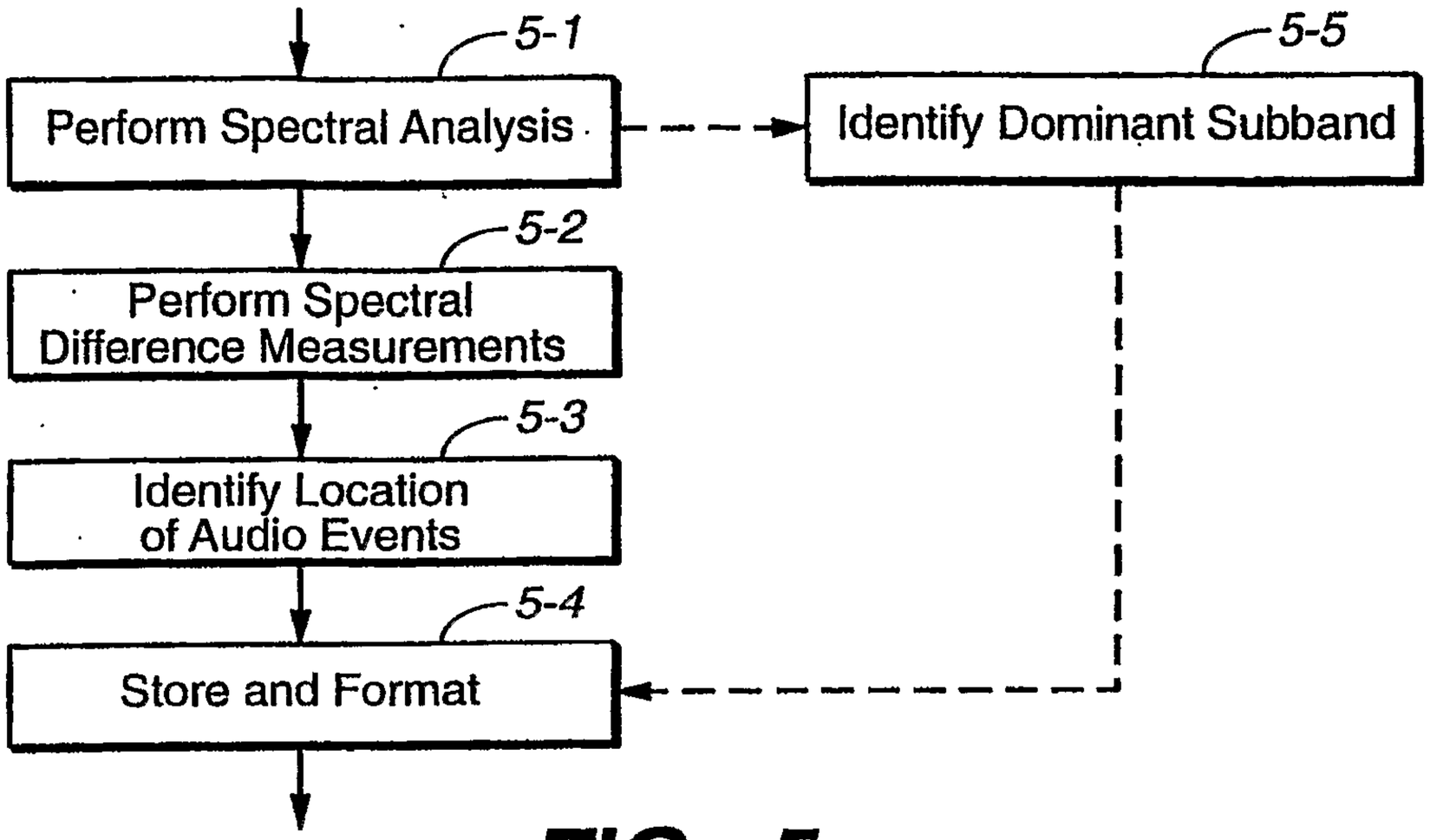

5-1
Perform Spectral Analysis
5-5
Identify Dominant Subband
5-2
Perform Spectral Difference Measurements
5-3
Identify Location of Audio Events
5-4
Store and Format

FIG._5

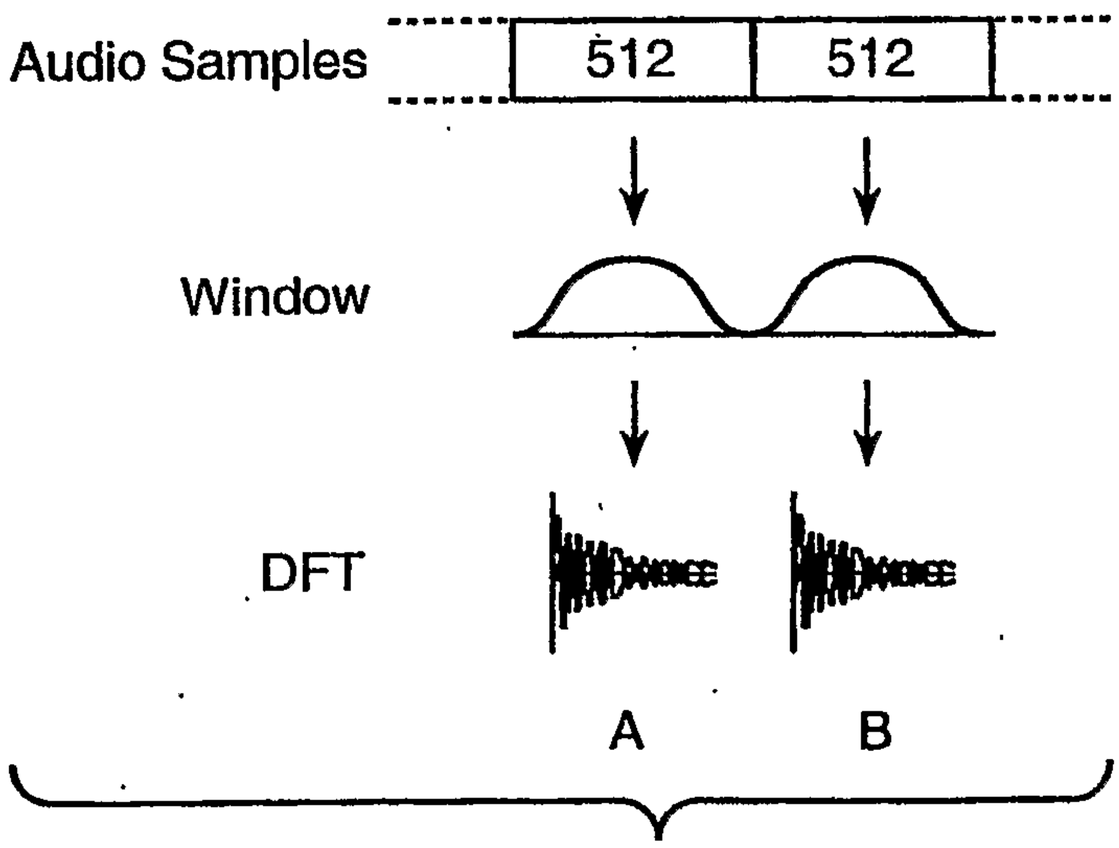

Audio Samples
512
512
Window
DFT
A
B

FIG._6

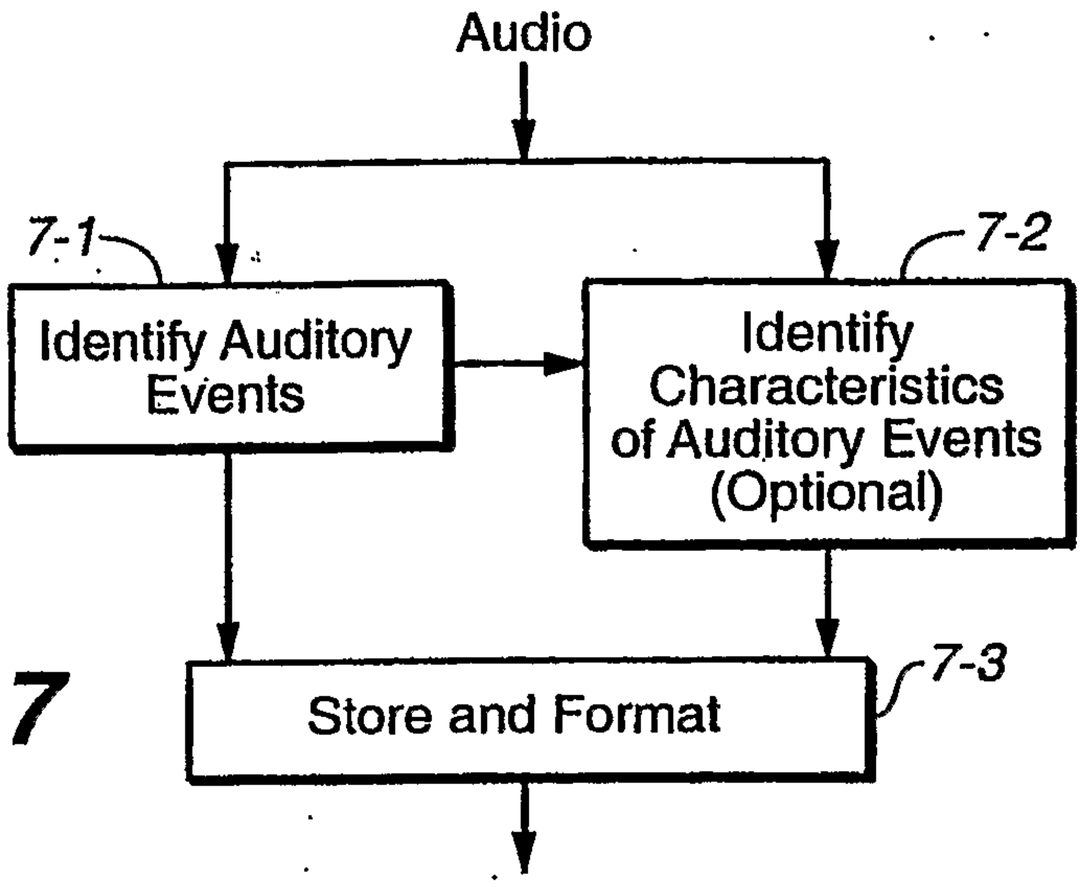


FIG._7

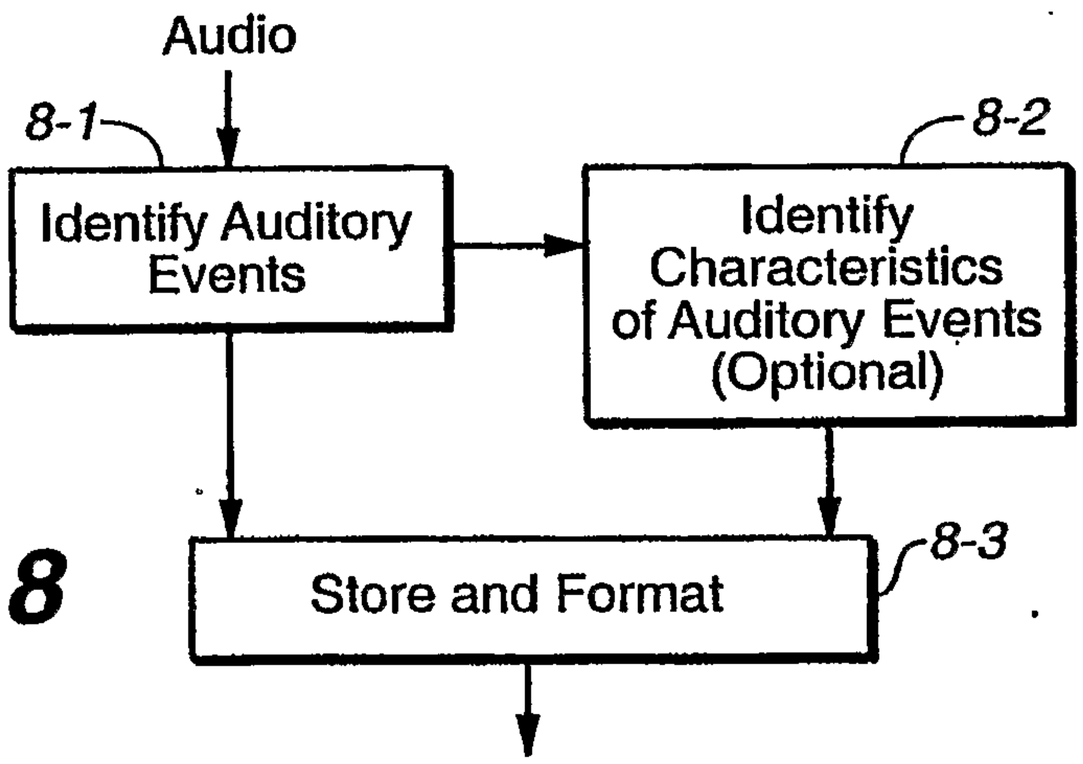


FIG._8

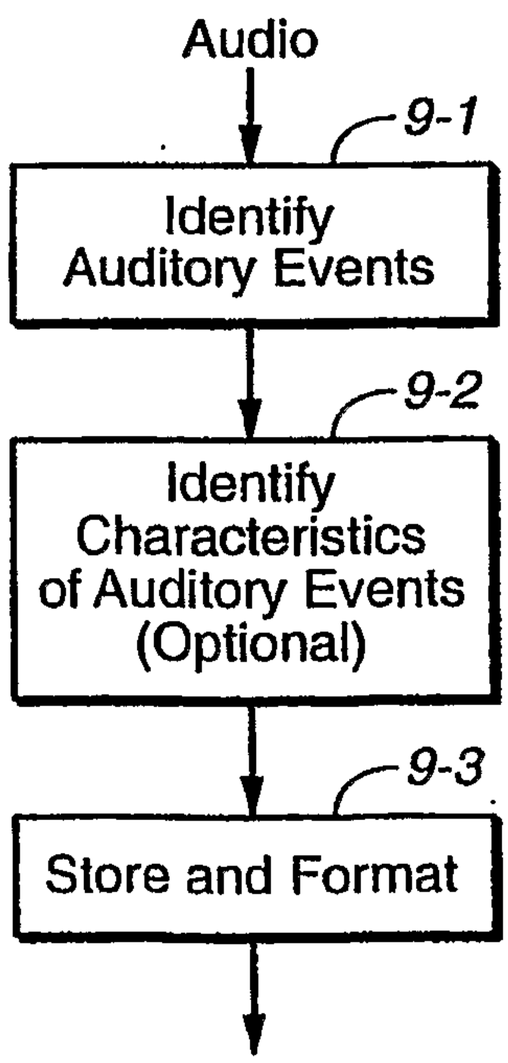
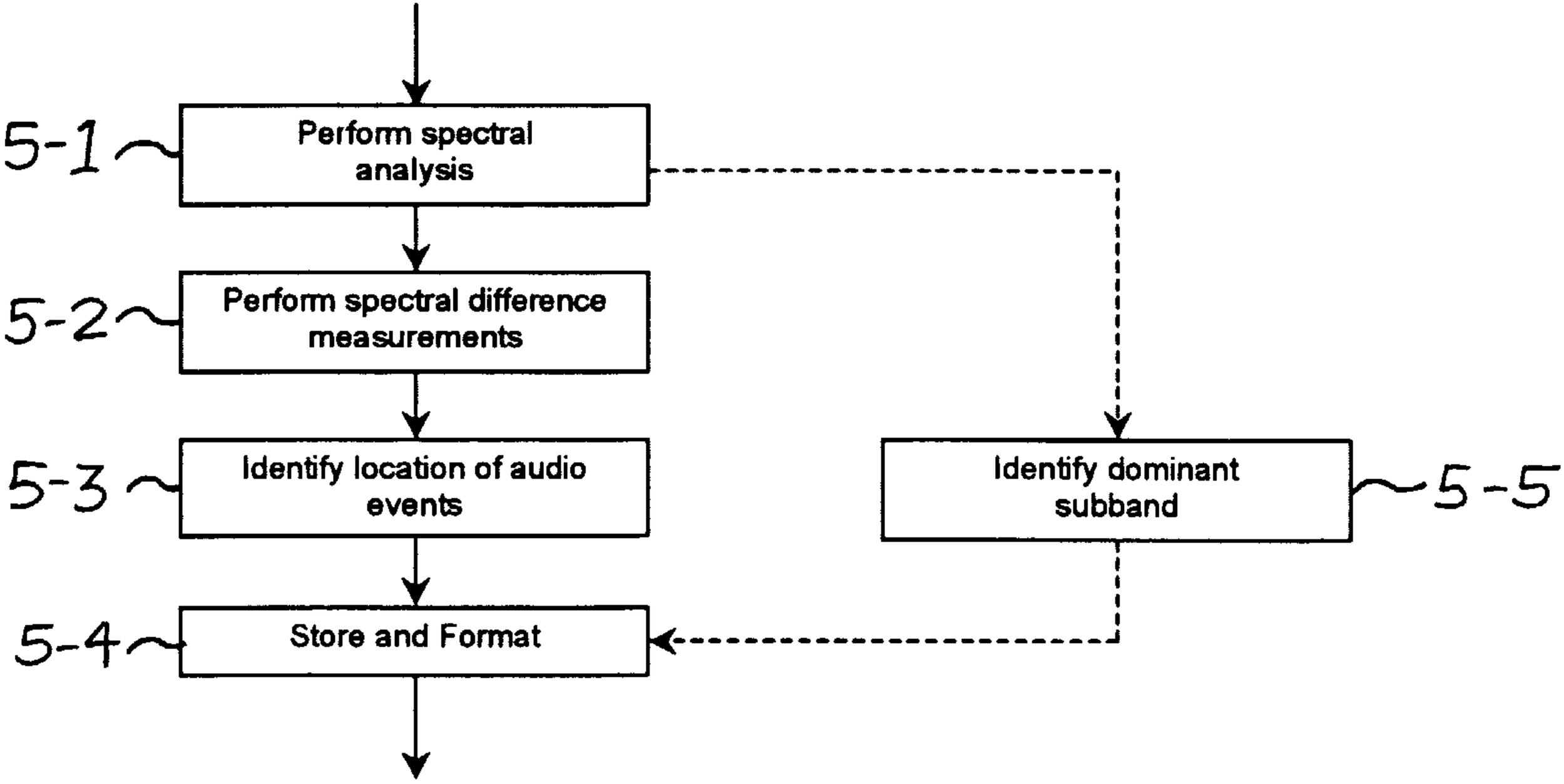


FIG._9


5-1
Perform spectral analysis
5-2
Perform spectral difference measurements
5-3
Identify location of audio events
5-4
Store and Format
Identify dominant subband
5-5