Attorney Docket No. 83063

TIME-SERIES SEGMENTATION

TO ALL WHOM IT MAY CONCERN:

BE IT KNOWN THAT (1) PAUL M. BAGGENSTOSS, citizen of the United States of America, employee of the United States Government, resident of Newport, County of Newport, State of Rhode Island has invented certain new and useful improvements entitled as set forth above of which the following is a specification:

JAMES M. KASISCHKE, ESQ.
Reg. No. 36562
Naval Undersea Warfare Center
Division, Newport
Newport, RI 02841-1708
TEL: 401-832-4763
FAX: 401-832-1231

I hereby certify that this correspondence is being deposited with the U.S. Postal Service as U.S EXPRESS MAIL, Mailing Label No. EL578538533US in envelope addressed to: Assistant Commissioner for Patents, Washington, DC 20231 on 8/31/2001 (DATE OF DEPOSIT)

James M. Kasischke
APPLICANT'S ATTORNEY

8/31/2001
DATE OF SIGNATURE

DEPARTMENT OF THE NAVY
OFFICE OF COUNSEL
NAVAL UNDERSEA WARFARE CENTER DIVISION
1176 HOWELL STREET
NEWPORT RI 02841-1708

IN REPLY REFER TO:

Attorney Docket No. 83063
Date: 12 May 2003

The below identified patent application is available for licensing. Requests for information should be addressed to:

PATENT COUNSEL
NAVAL UNDERSEA WARFARE CENTER
1176 HOWELL ST.
CODE 00OC, BLDG. 112T
NEWPORT, RI 02841

Serial Number 09/949,409

Filing Date 8/31/01

Inventor Paul M. Baggenstoss

If you have any questions please contact James M. Kasischke, Acting Deputy Counsel, at 401-832-4736.

Attorney Docket No. 83063

TIME-SERIES SEGMENTATION

STATEMENT OF GOVERNMENT INTEREST

The invention described herein may be manufactured and used by or for the Government of the United States of America for governmental purposes without the payment of any royalties thereon or therefor.

CROSS REFERENCE TO OTHER PATENT APPLICATIONS

Not applicable.

BACKGROUND OF THE INVENTION

(1) Field of the Invention

This invention generally relates to a method and system for identifying data segments within a signal by using naturally occurring boundaries in the signal and updating sample-by-sample.

More particularly, the invention is directed to solving the problem of dividing an input signal, such as an acoustic data signal or a speech signal, consisting of multiple "events" into frames where the signal within each frame is statistically "consistent". Once the data has been segmented, detection and

classification of events is greatly facilitated. In speech signals, for example, the data becomes segmented into phonetically constant frames or frames in which there are an integer number of pitch periods. This makes determination of pitch more accurate and reliable.

(2) Description of the Prior Art

Prior to this invention, it has not been known how to divide a time-series (signal) into segments with a fine enough resolution corresponding to individual pitch interval boundaries. The current art for optimally segmenting a time-series consists of first segmenting the data into fixed-size segments, then performing a second stage of segmentation to group together numbers of the fixed-size segments into larger blocks. This approach has a resolution no finer than the size of the fixed-size segments.

Because speech signals contain features that are very short in duration, it would be preferable to segment the data to a finer resolution, such as to a resolution of one sample. The current art cannot be used to segment the data to a resolution of one sample because it requires first segmenting to fixed-size segments large enough to extract meaningful features. Furthermore, the existing dynamic-programming solution is computationally impractical because the data has to be processed at each delay and at each segment length.

Thus, a problem exists in the art whereby it is necessary to develop a computationally efficient and practical method of segmenting multiple events into frames to a resolution of one sample necessary to identify individual pitch intervals.

By way of example of the state of the art, reference is made to the following papers, which are incorporated herein by reference. References pertaining to the prior art are contained in the following references:

[1] Euler, S.A.; Juang, B.H.; Lee, C.H.; Soong, F.K., *Statistical Segmentation and Word Modeling Techniques in Isolated Word Recognition*, 1990 International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp. 745 -748.

[2] Svendsen, F. Soong, *On the Automatic Segmentation of Speech Signals*, 1987 International Conference on Acoustics, Speech, and Signal Processing, pp. 77-80, Dallas, 1986.

[3] R. Bellman, S. Dreyfus, *Applied Dynamic Programming*, Princeton Univ. Press, 1962

[4] R. Kenefic, *An Algorithm to Partition DFT Data into Sections of Constant Variance*, IEEE Trans AES, July 1998

Referring further to the current state of the art as developed in the field to date, it should be understood that detection and classification of short signals is a high priority for the Navy. Segmentation of a time series is a method that facilitates detection and classification.

In segmentation of short signals, the following is an illustration of the current state of the art. Let there be N samples $x=[x_1..x_N]$. One would like to divide these samples into a number of segments, for example:

$$x = [x_1..x_a][x_{a+1}..x_b][x_{b+1}..x_c][x_{c+1}..x_N],$$

such that the total score, Q, where:

$$Q = Q(x_1..x_a)+Q(x_{a+1}..x_b)+Q(x_{b+1}..x_c)+Q(x_{c+1}..x_N)$$

is as high as possible.

To do this, the score function, Q(n,t), must be known for a segment of length n ending at time t. Assuming it is known, the problem is to find the best number of segments and their start times {a, b, c, d . .}. The standard dynamic-programming approach disclosed in Bellman and also Soong, above, is to first compute the score for all possible segment lengths at all possible end-times. In other words, compute Q(n,t) for $t = n_{min}..N$ and $n = n_{min}$ to n_{max} where n_{min} and n_{max} are the range of allowed segment lengths. The problem is solved by starting at sample n_{min} because the best solution for segmenting the data up

to sample n_{min} is immediately known, it is just the value of the score function Q when $n=n_{min}$ and $t=n_{min}$, $Q(n_{min}, n_{min})$. Let this be called $Q_b(n_m)$. The best solutions for later samples are then easily found as follows:

$Q_b(t)=Q(n,t)+Q_b(t-n)$ maximized over n.

Since $Q_b(t-n)$ was already computed, all of the necessary information is available. The value of n for this solution is also saved and is called $n_m(t)$. This process proceeds until $Q_b(N)$, $n_b(N)$. The problem is then solved. The maximum total score is $Q_b(N)$ and the length of the last segment is $n_b(N)$. The other segment lengths are found by working backwards. For example, the length of the next-to-last segment is $n_b(N-n_b(n))$, which was previously stored. This is the standard approach taught in the prior art.

In many problems, it is needed to have the best segments and also to pick the best models for each segment. In speech, for example, it may be necessary to know if a segment is voiced or unvoiced speech or it might be necessary to choose the best model order. Let p be an index that ranges over all possible models. To find the best combination of segment lengths and model indexes, first the score function Q(p,n,t) must be known. A slight modification is then made to the above procedure by carrying out the maximizations at each time over both n and p jointly.

What has been described so far is the standard approach taught by the Bellman and Soong references. The problem with applying the method to speech processing and other fields is that computing the score function is time-consuming and the method is not practical to apply sample-by-sample as data is acquired. Instead, it is necessary to apply the method to a coarse resolution defined by the frame-processing interval taught by Soong. Features of the data finer than the frame processing interval are filtered out of the data.

As mentioned, sample-by-sample processing is normally impractical. If the score function is computed on samples $[x_{t-n+1}..x_t]$, and it is desired to move over one sample to $[x_{t-n+2}..x_{t+1}]$, it is necessary to re-compute the entire score function. This is because the state of the art in signal processing in speech and other fields uses the Fast Fourier Transform (FFT) and a "window" function such as a Hanning window. Window functions are necessary to smooth transitions in the data and eliminate edge effects. This is because the data is processed in "chunks" which are not always aligned with the naturally occurring event boundaries.

It should be understood that the present invention would in fact enhance the functionality of the above cited art by the combined effect of eliminating the window function previously used, and providing sample-by-sample updates.

SUMMARY OF THE INVENTION

Therefore it is an object of this invention to provide an improved method of time series segmentation.

Another object of this invention is to provide a method for dividing a signal into plural segments of data.

Still another object of this invention is to provide a method for dividing a signal into plural segments of data in the absence of a window function.

Yet another object of the invention is to provide a method for dividing a signal into plural segments of data and updating segment scores thereof one sample at a time.

In accordance with one aspect of this invention, there is provided a method for segmenting a signal into segments having similar spectral characteristics is provided. Initially the method generates a table of previous values from older signal values that contains a scoring value for the best segmentation of previous values and a segment length of the last previously identified segment. The method then receives a new sample of the signal and computes a new spectral characteristic function for the signal based on the received sample. A new scoring function is computed from the spectral characteristic function. Segments of the signal are recursively identified based on the newly computed scoring function and the table of previous

values. The spectral characteristic function can be a selected one of an autocorrelation function and a discrete Fourier transform. An example is provided for segmenting a speech signal.

BRIEF DESCRIPTION OF THE DRAWINGS

The appended claims particularly point out and distinctly claim the subject matter of this invention. The various objects, advantages and novel features of this invention will be more fully apparent from a reading of the following detailed description in conjunction with the accompanying drawings in which like reference numerals refer to like parts, and in which:

The FIG. is an example of segmentation of speech illustrating the result of the method and system of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

In general, the present invention is directed to solving the problem of dividing an input signal, such as acoustic data or a speech signal, consisting of multiple "events" into frames where the signal within each frame is statistically "consistent". Once the data has been segmented, detection and classification of events is greatly facilitated. In speech signals, for example, the data becomes segmented into

phonetically constant frames or frames in which there are an integer number of pitch periods. This makes determination of pitch more accurate and reliable.

This invention was disclosed by the inventor in the following presentation, which is incorporated by reference herein.

P. M. Baggenstoss et al., *A Theoretically Optimal Probabilistic Classifier Using Class-Specific Features*, 2000 International Conference on Pattern Recognition, Barcelona, Spain, September 2, 2000.

The invention automatically divides an arbitrary time-series signal into arbitrary-length frames or segments wherein the data in each frame is "consistent". This ability to determine a consistent frame of data facilitates detection and classification of each frame of the data as well as the data as a whole. Current detectors locate events only to an FFT frame. The proposed method can locate events to a resolution of one sample. The results of experiments show that the segmentation occurring in the present invention is as good as possible by a human operator.

As indicated above, the problem is to divide a time-series signal such as a digitized audio stream into segments corresponding to the naturally occurring events in the signal. The invention provides a non-windowed processing method (in

contrast to the state of the art which uses windowing) which allows recursive update of a spectral feature function such as one of a Discrete Fourier Transform (DFT) and a circular Auto Correlation Function (ACF). This method has the added benefit of causing the resulting segments to be perfectly aligned to event boundaries.

The method of the present invention allows sample by sample updating of spectral feature function which does not require a window function. The window function is not necessary because the segments will be exactly aligned to the "event" boundaries in the signal. Also, because no window function is used, it is possible to update the score function efficiently by accounting only for the added and dropped samples.

When a spectral feature function such as a discrete Fourier transform (DFT) is computed on samples $[x_{t-n+1}..x_t]$, denoted $X_t[k]$ where t is the sample and k is the transform variable, and it is desired to compute it on samples $[x_{t-n+2}..x_{t+1}]$, denoted $X_{t+1}[k]$, $X_t[k]$ is related to $X_{t+1}[k]$ by the following equation:

$$X_{t+1}[k] = e^{j2k/n} [X_{t+1}[k] - (x_{t-n+1} - x_{t+1})] \quad (1)$$

If the spectral feature function is a circular autocorrelation function (ACF) computed on samples $[x_{t-n+1}..x_t]$, denoted $r_t[\tau]$ where τ is the correlation variable, and it is desired to compute it on samples $[x_{t-n+2}..x_{t+1}]$, denoted $r_{t+1}[\tau]$ then:

$$r_{t+1}[\tau] = r_t[\tau] + (x_{t+1} - x_{t-n+1})(x_{t-n+1} - x_{t+1-\tau})/n \qquad (2)$$

Score functions that are computed from the spectral feature function can be computed efficiently at each sample. Other types of efficiently-computed score functions are also possible. Previous values of the score function for the best segmentation and the length of the last segment can be stored in a table. Upon segmentation, the score and length of the latest value can be utilized with the table values for efficiently obtaining the current best segmentation. Accordingly, at any time the invention using dynamic programming can segment a stream of time series data into segments having like characteristics. These segments can then be classified.

Applying the current invention for speech processing the Autocorrelation function (ACF) is used as the spectral feature function. Because the ACF is sensitive to spectral features in the data, the resulting segments are on boundaries where the spectrum changes.

In speech data, a "reward" is additionally added to the score function for segments matching the pitch interval exactly. To determine the score of a segment, the ACF is computed, and then the Levinson recursion is used to compute the linear prediction error variance for every model order up to a maximum (of about 16). The score for a given model order p on a segment of length n is:

$$Q(p,n) = (-n/2)(\log(\sigma^2[p,n])+1) - (p/2)*\log(n) + K \qquad (3)$$

where n is the segment length and $\sigma^2[p,n]$ is the prediction error variance for model order p, and K is a "reward" value for periodicity. The well-known Levinson-Durbin algorithm can be used to compute $\sigma^2[p,n]$ from the ACF efficiently.

The term (p/2)*log(n) is the well-known Minimum Description Length (MDL) penalty score. To "reward" the segment for matching the pitch interval, a positive number K is added to Q(p,n) if the ACF of the segment shows "periodicity". To determine periodicity in the speech application, every division factor d = 2 up to d = 6 is tested. The meaning of d is the number of pitch intervals in the segment. For each value of d, the smallest ACF lag in the set {r[0], r[n/d], r[2n/d], .. r[n/2]} is determined. d_{max} is determined as the division factor producing the largest minimum ACF value. If d_{max} is greater than a fraction of r[0], it can be labeled as periodic with a division factor d_{max} and thus the period is n/d_{max}. The fraction is established by trial and error based on the given application. For speech recognition .5 has been found to be an effective fraction. While this only happens rarely, it is bound to happen for some segment (and all segments are tested), thus the method works. The reward value used is a monotonically increasing function of d_{max}.

The FIG. is an illustration of a segmented speech signal 10. Identified segments 12 are indicated by dashed lines. The Autocorrelation Function is provided for three identified segments (A), (B) and (C). (D) is provided as the autocorrelation function of an arbitrary region of the speech signal that is not indicated by the scoring function as a segment. The segments (A), (B) and (C) enclose well-defined events or periodic (voiced) areas of exactly 3, 2, and 5 pitch intervals, respectively. Further, the illustrated non-windowed ACF functions for these segments begin and end at the same levels showing almost perfect periodicity. The non-windowed ACF for the arbitrary region (D) which is slightly smaller than segment (C) does not have this property. From this example, it can be seen how the segmentation works hand-in hand with the non-windowed ACF.

The key feature of this invention is the use of non-windowed processing which permits fast computation of a spectral feature function such as a DFT or ACF on a sample-by sample basis. Thus, variations of the method include any method that uses the DFT, ACF or other recursively computed spectral feature function, as described herein.

In view of the above detailed description, it is anticipated that the invention herein will have far reaching applications other than those specifically described.

This invention has been disclosed in terms of certain embodiments. It will be apparent that many modifications can be made to the disclosed apparatus without departing from the invention. Therefore, it is the intent of the appended claims to cover all such variations and modifications as come within the true spirit and scope of this invention.

Attorney Docket No. 83063

TIME-SERIES SEGMENTATION

ABSTRACT OF THE DISCLOSURE

A method for segmenting a signal into segments having similar spectral characteristics is provided. Initially the method generates a table of previous values from older signal values that contains a scoring value for the best segmentation of previous values and a segment length of the last previously identified segment. The method then receives a new sample of the signal and computes a new spectral characteristic function for the signal based on the received sample. A new scoring function is computed from the spectral characteristic function. Segments of the signal are recursively identified based on the newly computed scoring function and the table of previous values. The spectral characteristic function can be a selected one of an autocorrelation function and a discrete Fourier transform. An example is provided for segmenting a speech signal.

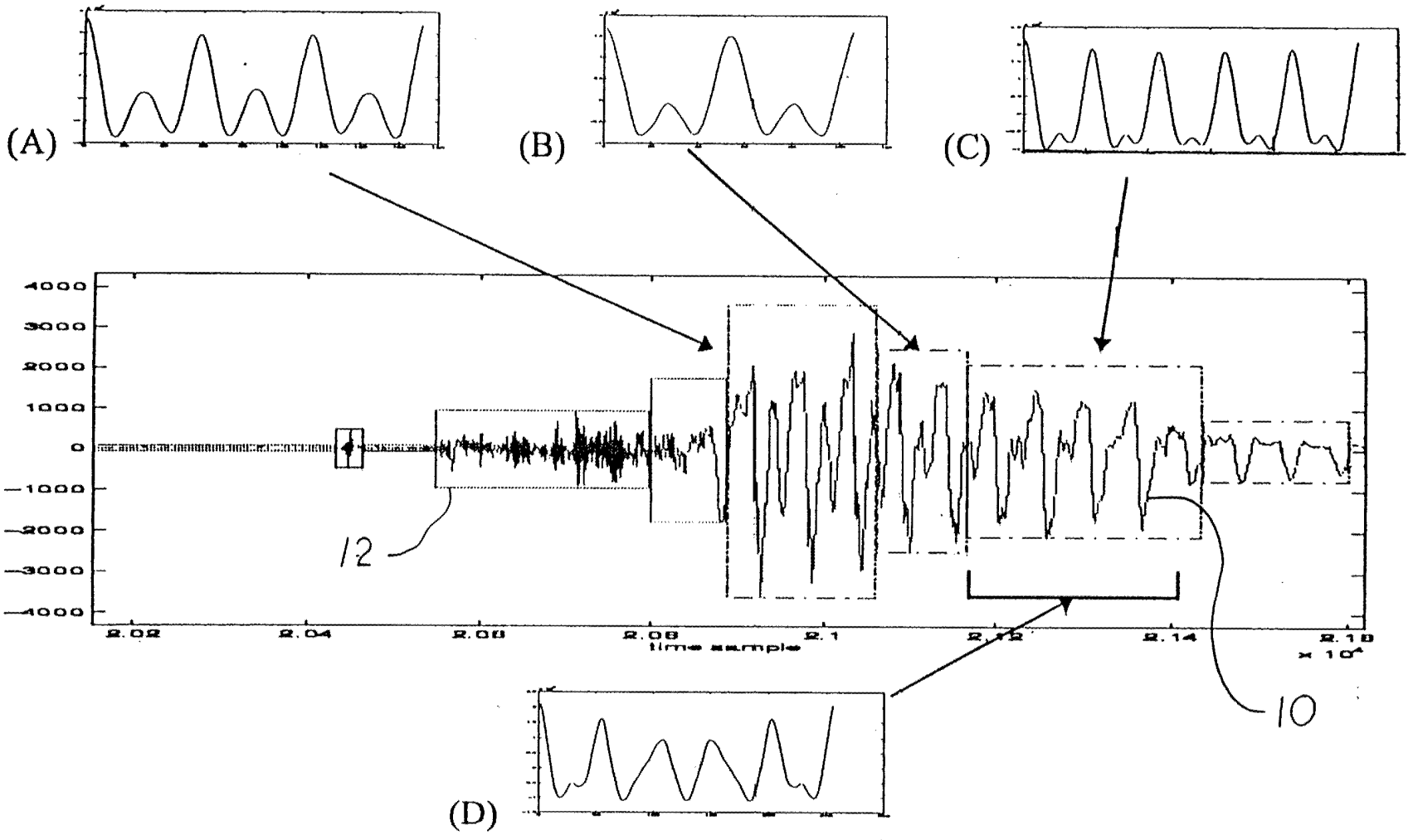
(A)
(B)
(C)
(D)
4000
3000
2000
1000
0
-1000
-2000
-3000
-4000
2.02
2.04
2.06
2.08
2.1
2.12
2.14
2.16
time sample
x 10^4
12
10

FIG.