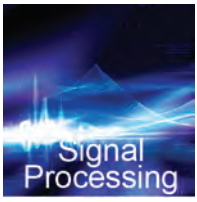
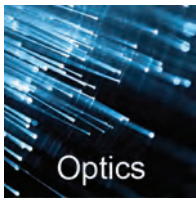




National Security Agency

2014 TECHNOLOGY CATALOG

TECHNOLOGY TRANSFER PROGRAM

ACOUSTICS

The NSA is a leader in acoustic research. These advanced acoustic technologies are a cornerstone of the NSA's mission to produce foreign signals intelligence. With the enormous increase in voice and acoustic data, the demand for faster, more accurate voice and acoustic signal analysis and filtering has never been greater. To meet this growing demand for voice signal intelligence, the NSA continually conducts research in acoustics signal analysis, phonetics, audio signal identification, and audio transcription.

Within the acoustics technology area, the NSA has several technologies available for license. These technologies include: methods for identification, extraction, and analysis of voice and voice signals; foreign language voice recognition; duplicate voice identification; and methods of measuring voice enhancement.

Real-Time Simultaneous Identification of Multiple Voices

Patent: 8,442,825

The invention provides multiple speaker identification by identifying voices in a manner that uniquely mimics the essence of ear-to-brain interconnection combined with observed human voice identification learning and recognition processes. The object is real-time or faster voice identification needing only relatively simple computing resources. Specifically, this invention looks for prosody matches (spectral patterns over time periods) that were trained by software in an Artificial Neural Network (ANN) description.

VALUE

Excludes non-speech sounds within audio
Real-time processing

Measuring Degree of Enhancement to a Voice Signal

Patent: 7,818,168

This technology is a method of measuring the degree of enhancement made to a voice signal. Typically, voice signals are statistically non-stationary and the more noise, or other corruption, introduced into a signal, the more stationary its distribution of values become. In this invention, the degree of reduction in stationarity is indicative of the degree of enhancement made to the signal. The method of determining the degree of enhancement begins with receiving and digitizing the signal. A user then identifies a number of formant regions and computes the stationarity for each formant. The voice signal is enhanced and formant regions in the enhanced signal are identified. The stationarity for the formants in the enhanced signal is found. Finally, a comparison is made between the stationarities of the original and enhanced signals.

VALUE

Quantifies voice enhancement
Reduces human arbitration and various listening tests

Comparing Voice Signals

Patent: 7,650,281

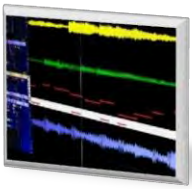
This technology tests the robustness of a given voice-matching algorithm by providing the algorithm with variants of a given digital file and testing the original against these variants including time-reversal, segmented re-arrangement, or a mixture of both time-reversal and segmented re-arrangement. In effect, this increases the corpus of ground truth, thus allowing realistic testing under controlled conditions.

VALUE

Reduces the Equal Error Rate (EER)
Larger corpus from fewer files

Identifying Digital Audio Signal Format

Patent: 7,620,469



This technology identifies the format of a digital audio signal including signals that are either self-defining or headerless. In this method, a digital audio file is received and converted from an assumed format and bit ordering to a user-definable format. The file is divided into blocks and the frequencies of occurrence are determined.

A first set of frequencies of occurrence less than and equal to most frequently occurring integer is created. Next, a second set of frequencies of occurrence greater than the most frequently occurring integer is created. Third and fourth sets of differences are created and replaced with polarity indicators. These indicators are summed and percentages calculated to determine maximum pairings. The statistics are then assigned to the converted file. This process is repeated with another format and bit ordering to identify the format with the maximum statistics.

VALUE

Identifies digital audio formats, including self-defining signals
Reduces need for human listening

Automated Detection of Duplicate Audio & Voice Recordings

Patent: 7,571,093

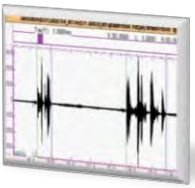
This invention detects duplicate audio and voice recordings in a collection. A recording is selected, divided into segments, and a pitch value is extracted for each segment. The total time a voice appears in the recording is estimated and pitch values that are less than or equal to a user-defined value are removed. Unique pitch values are identified and the frequency of occurrence is determined and normalized. The distribution percentiles are then calculated. This method is repeated for each recording where it is compared for total voice time, average pitch value, and distribution percentile.

VALUE

No manual transcription required
Language and content independent

Detecting Voice Activity

Patents: 7,127,392 and 6,556,967



This technology eliminates the need to manually search audio files for speech content by automatically locating speech intervals that may contain other signals such as music, noise, or empty space. It outputs the start and stop times of these intervals relative to the beginning of the file, ignoring non-speech portions of the file. This technology classifies signal segments as speech or non-speech and employs algorithms which consider characteristics derived from an audio signal's AM envelope.

VALUE

Reduces bandwidth and traffic
Improves performances of speaker recognition systems

United States Patent **8,442,825**
Sinutko **May 14, 2013**

Biomimetic voice identifier

Abstract

A device for voice identification including a receiver, a segmenter, a resolver, two advancers, a buffer, and a plurality of IIR resonator digital filters where each IIR filter comprises a set of memory locations or functional equivalent to hold filter specifications, a memory location or functional equivalent to hold the arithmetic reciprocal of the filter's gain, a five cell controller array, several multipliers, an adder, a subtractor, and a logical non-shift register. Each cell of the five cell controller array has five logical states, each acting as a five-position single-pole rotating switch that operates in unison with the four others. Additionally, the device also includes an artificial neural network and a display means.

Inventors: **Sinutko; Michael** (Glen Burnie, MD)

Applicant:

Name	City	State	Country	Type
Sinutko; Michael	Glen Burnie	MD	US	

Assignee: **The United States of America as Represented by the Director, National Security Agency** (Washington, DC)
N/A (N/A)

Family ID: **48225543**

Appl. No.: 13/200,034
Filed: August 16, 2011

Current U.S. Class:	**704/247**; 379/52; 379/88.01; 379/88.02; 379/88.07; 704/200; 704/231; 704/246; 704/249; 704/250; 704/270; 704/273; 704/275; 708/300; 84/602
Current CPC Class:	G06F 15/00 (20130101); G10H 7/00 (20130101); H04M 11/00 (20130101); G10L 15/02 (20130101); G10L 15/16 (20130101)
Current International Class:	G10L 17/00 (20060101); H04M 1/64 (20060101); G10L 15/00 (20060101); G10L 21/00 (20060101); G06F 17/10 (20060101); G06F 15/00 (20060101); G10H 7/00 (20060101); H04M 11/00 (20060101)
Field of Search:	;708/300

References Cited [Referenced By]

U.S. Patent Documents

4852170	July 1989	Bordeaux
5758023	May 1998	Bordeaux
6067517	May 2000	Bahl et al.
6240192	May 2001	Brennan et al.
6738486	May 2004	Kaulberg
7319959	January 2008	Watts
8117248	February 2012	Dobbek et al.
2011/0116666	May 2011	Dittberner et al.

Other References

K Chong, B. Gwee, and J. Chang, ! 16-Channel Low-Power Nonuniform Spaced Filter Bank Core for Digital Hearing Aids, Sep. 2006, IEEE Transactions on Circuits and Systems, vol. 53, No. 09 pp. 853-857. cited by examiner .

T. Irino and R. Patterson, A compressive gammachirp auditory filter for both physiological and psychophysical data, May 2001, J. Acoust. Soc. Am., vol. 109, No. 5, pp. 2008-2022. cited by examiner .
M.P. Leong, C.T. Jin, P.H.W. Leong, "An FPGA-Based Electronic Cochlea", EURASIP Journal on Applied Signal Processing 2003:7, pp. 629-638. cited by examiner .
N. Sawhney, Situational Awareness from Environmental Sounds, Jun. 13, 1997, http://www.mendeley.com/catalog/situational-awareness-environmental- -sounds/. cited by examiner .
H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing, vol. 2, No. 4, Oct. 1994, pp. 578-589. cited by examiner.

Primary Examiner: Shah; Paras D
Assistant Examiner: Thomas-Homescu; Anne
Attorney, Agent or Firm: Kurian; Roshni

Claims

What is claimed is:

1. A device for voice identification, comprising: a) a receiver, having an input, and having an output; b) a segmenter, having a first input, having a second input, having a third input, and having one output; c) a plurality of infinite impulse response (IIR) resonator digital filters, where each of said plurality of IIR resonator digital filters has an input connected to said output of said segmenter, an output, and further comprises: i. a first register, having a first input, having a second input, having a third input, having a fourth input, having a fifth input, having a first output, having a second output, having a third output, having a fourth output, and having a fifth output; ii. a multiplexer/demultiplexer, having a first input connected to said first output of said first register, having a second input connected to said second output of said first register, having a third input connected to said third output of said first register, having a fourth input connected to said

fourth output of said first register, having a fifth input connected to said fifth output of said first register, having a sixth input, having a seventh input, having a control input, having a first output connected to said first input of said first register, having a second output connected to said second input of said first register, having a third output connected to said third input of said first register, having a fourth output connected to said fourth input of said first register, having a fifth output connected to said fifth input of said first register, having a sixth output, having a seventh output, and having an eighth output; iii. a first multiplier, having a first input connected to said sixth output of said multiplexer/demultiplexer, having a second input, and having an output; iv. a second multiplier, having a first input connected to said seventh output of said multiplexer/demultiplexer, having a second input, and having an output; v. a third multiplier, having a first input, having a second input, and having an output connected to said sixth input of said multiplexer/demultiplexer; vi. an adder, having a first input connected to said output of said first multiplier, having a second input connected to said output of said second multiplier, having a third input connected to said output of said third multiplier, and having an output; vii. a subtractor, having a first input connected to said output of said adder, having a second input connected to said eighth output of said multiplexer/demultiplexer, having an output connected to said seventh input of said multiplexer/demultiplexer; and viii. a second register, having a first input, having a first output connected to said first input of said third multiplier, having a second output connected to said second input of said second multiplier, having a third output connected to said second input of said first multiplier, and having a fourth output; d) a resolver, having a plurality of outputs equal to said plurality of IIR resonator digital filters plus 3 additional outputs, and having a plurality of inputs equal to said plurality of IIR resonator digital filters plus one additional input, where each of said plurality of inputs is connected to each of said outputs of each of said plurality of IIR resonator digital filters; e) a first advancer, having an input connected to one of said plurality of outputs of said resolver and an output connected to said second input of said segmenter; f) a second advancer, having an input connected to one of said plurality of outputs of said resolver and an output connected to said one additional input of said resolver; g) a buffer, having a plurality of inputs, three less than said outputs of said resolver that are connected to said plurality of outputs of said resolver, and a plurality of outputs equal to a value of P; h) a display

means, where said display has a plurality of inputs such that each of said plurality of inputs of said display means is connected to one of said plurality of outputs of said buffer and said fourth output of said second register; and i) an artificial neural network having a plurality of outputs equal to a value of C and having a plurality of inputs equal to said value of P, where each of said plurality of inputs is connected to each of said plurality of outputs of said buffer.

2. The device of claim 1, wherein said multiplexer/demultiplexer includes five programming states selected from a group of programming states consisting of: a) a first logical state where said first input is connected to said sixth output, said second input is connected to said seventh output, said fifth input is connected to said eighth output, said sixth input is connected to said third output, said seventh input is connected to said fifth output; b) a second logical state where said fifth input is connected to said sixth output, said first input is connected to said seventh output, said fourth input is connected to said eighth output, said sixth input is connected to said second output, said seventh input is connected to said fourth output; c) a third logical state where said fourth input is connected to said sixth output, said fifth input is connected to said seventh output, said third input is connected to said eighth output, said sixth input is connected to said first output, said seventh input is connected to said third output; d) a fourth logical state where said third input is connected to said sixth output, said fourth input is connected to said seventh output, said second input is connected to said eighth output, said sixth input is connected to said fifth output, said seventh input is connected to said second output; and e) a fifth logical state where said second input is connected to said sixth output, said third input is connected to said seventh output, said first input is connected to said eighth output, said sixth input is connected to said fourth output, said seventh input is connected to said first output.

Description

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application is related to U.S. patent application Ser. No.

12/806,256, entitled "INFINITE IMPULSE RESPONSE RESONATOR DIGITAL FILTER" filed on Aug. 5, 2010.

FIELD OF THE INVENTION

The present invention pertains to signal processing, and in particular to a voice identification system where "voice" includes speech and non-speech sounds.

BACKGROUND OF THE INVENTION

Current voice identification systems are limited in their ability to efficiently process audio from environments that include high noise content and multiple speakers. Consider the following prior art.

U.S. Pat. No. 4,852,170, entitled "Real Time Computer Speech Recognition System," discloses a system that determines the frequency content of successive segments of speech. While U.S. Pat. No. 4,852,170 disclosed a system that is capable of analyzing speech digitally, the present invention offers a system that allows speech to be analyzed digitally in a way that is distinguishable from the device taught in U.S. Pat. No. 4,852,170. U.S. Pat. No. 4,852,170 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,758,023, entitled "Multi-Language Speech Recognition System," discloses a system that considers the frequency spectrum of speech. While U.S. Pat. No. 4,852,170 disclosed a system that is capable of analyzing speech digitally, the present invention offers a system that allows speech to be analyzed digitally in a way that is distinguishable from the device taught in U.S. Pat. No. 5,758,023. U.S. Pat. No. 5,758,023 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,067,517, entitled "Transcription of Speech Data with Segments from Acoustically Dissimilar Environments," discloses a technique for improving recognition accuracy when transcribing speech data that contains data from a variety of environments. U.S. Pat. No. 6,067,517 is hereby incorporated by reference into the specification of the

present invention.

U.S. Pat. No. 7,319,959, entitled "Multi-Source Phoneme Classification for Noise-Robust Automatic Speech Recognition," discloses a system and method for processing an audio signal by segmenting the signal into streams and analyzing each stream to determine phoneme-level classification. U.S. Pat. No. 7,319,959 is hereby incorporated by reference into the specification of the present invention.

While some of the prior art teaches methods of identifying multiple speakers, these methods often employ large, costly software programs that require substantial unique computing resources. Additionally, the known prior art fails to disclose an efficient and cost-effective way to identify multiple voices, in parallel, and identify the time periods that each such multiplicity of voices is present within the digitized audio that is being processed.

SUMMARY OF THE INVENTION

An object of the present invention is a device for identification of multiple arbitrary voices in parallel.

Another object of the present invention is a device for identifying the time locations of multiple arbitrary voices.

Yet another object of the present invention is a device that utilizes an infinite impulse response (IIR) resonator digital filter that is more efficiently implemented than existing IIR resonator digital filters.

The device for voice identification includes a receiver with an input and an output, a segmenter to store the output of the receiver, a plurality of IIR resonator digital filters, where a sequence of results from each IIR filter is sent to a resolver and where each IIR filter comprises a set of memory locations or functional equivalent to hold filter specifications, a memory location or functional equivalent to hold the arithmetic reciprocal of the filter's gain, a multiplexer/demultiplexer composed of a five cell controller array, a first multiplier, a second multiplier, a third multiplier, an adder, a subtractor, and a logical non-shift register to hold the intermediate states of

the present invention's filtering computations. Each cell of the five cell controller array has five logical states, each acting as a five-position single-pole rotating switch that operates in unison with the four others.

Additionally, the device also includes an artificial neural network having a fixed sequence of inputs connected to both a fixed sequence of a resolver's buffer outputs and a fixed sequence of inputs of a display means that enables presenting the resolver's audio spectrum output as presented to the neural network.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 depicts the prosody period structure.

FIG. 2 is a schematic depicting a system for voice identification;

FIG. 3 is a schematic depicting the infinite impulse response resonator digital filter;

FIG. 4A is a schematic of the first state of the controller array;

FIG. 4B is a schematic of the second state of the controller array;

FIG. 4C is a schematic of the third state of the controller array;

FIG. 4D is a schematic of the fourth state of the controller array; and

FIG. 4E is a schematic of the fifth state of the controller array.

DETAILED DESCRIPTION

The present invention is a device for identification of multiple voices in parallel. Additionally, the device is capable of identifying the time locations of these voices within the audio file by mimicking the essence of the human cochlea-to-brain interaction and implementing a voice identification technique.

In order to understand how the device mimics the essence of the human

cochlea-to-brain interaction, it is first necessary to understand how prosody and the prosody period structure are defined.

Prosody as used and defined in the present invention is biologically more accurate than prosody as commonly used in the technical field of voice identification, and is evidently unique among known published definitions of prosody. Specifically, prosody in the present invention is spectral pattern matching over an amount of time defined as the Prosody Period. The Prosody Period is held constant per provided audio input to the present invention. The Prosody Period Structure is depicted in FIG. 1 and described as follows: Each prosody period P.sub.t is composed of N prosody sub periods A. Within each prosody sub period process B, a succession of p samples are filtered by each filter f.sub.i.epsilon.{f.sub.1, f.sub.2, f.sub.3, . . . , f.sub.F}, where f.sub.1<f.sub.2<f.sub.3 . . . <f.sub.F.

Here, the vertical axis C represents the number of IIR filters and the horizontal axis D represents time. In one embodiment, the prosody period is composed of 40 prosody sub periods (N=40) in which 50 successive samples (p=50) per sub period are filtered per filter. Note the following definitions:

P=NF, where N=rP.sub.t/P;

P.sub.t=pN/r=N(t.sub.j+1-t.sub.j), j.epsilon.{0, 1, 2, . . . , N-1}, t.sub.j.epsilon.{t.sub.0, t.sub.1, t.sub.2, . . . , t.sub.N-1}.

p.sub.t=p/r=prosody sub period

Positive real number r is defined as the input audio's sampling rate in samples per second.

Positive integer F is defined as the number of IIR filters used.

Positive integer P is the number of filtering consensus scalars resulting from the Prosody Period Structure per prosody period.

Positive real number P.sub.t is the length of the user-specified prosody period in seconds.

Positive integer p is defined as the number of inner hair cell filament-interconnected stereocilia to be mimicked by the resolver as described below.

FIG. 2 is a schematic representing the device 1 for voice identification.

The device 1 includes a receiver 2 that has an input 3 for receiving a digital audio file for processing and an output 4. The device 1 also includes a segmenter 5 that has three inputs and one output 6. The segmenter 5, whose first input is connected to the output 4 of the receiver 2, segments the file received from the receiver 2 into "prosody period" cycles such that each cycle is P.sub.t seconds long. Each prosody period cycle is composed of N prosody sub periods where each prosody sub period processes p samples. Note that the initial starting time, t.sub.0 of P.sub.t, are specified by the present invention's end user.

The device 1 also incorporates F infinite impulse response (IIR) resonator digital filters 7A, 7B, 7X, where each filter is used to approximate the function of one inner hair cell of the human ear. The IIR resonator digital filters 7A, 7B, 7X are connected to the segmenter 5 via its output 6 which presents the amplitude of the digital audio sample currently being processed. The plurality of IIR resonator digital filters 7A, 7B, 7X are used to filter each sample from output 6 of the set of p samples in a prosody sub period. In one embodiment, the device 1 utilizes one hundred IIR resonator digital filters 7A, 7B, 7X (i.e. F=100) to filter each sample of a set of p samples, where the set of p samples is composed of 50 samples. Note that positive integer F represents the number of IIR filters 7A, 7B, 7X that filters each sample of a succession of p samples.

Each of the F IIR resonator digital filters 7A,7B, 7X has its own output 8A,8B, 8X, respectively, where each of the outputs 8A,8B, 8X is connected to an input of the resolver 9. The outputs 8A,8B, 8X corresponding to the resolver's 9 F inputs occurs for each of the p digital audio samples per prosody sub period. Per filter 7A, 7B, 7X, over p digital audio samples per prosody sub period, resolver 9, in aggregate functioning, mimics the functioning of p interconnected stereocilia atop one human inner hair cell.

In this embodiment, the resolver 9 mimics fifty such stereocilia (p=50) per prosody sub period per filter (hair cell) equating to 100 hair cell outputs 40 different times (40 successive sub prosody periods) over one prosody period. The resolver 9 also has an additional input 17 that that is connected to the output of a second advancer 16 (discussed below).

The resolver 9 also has a plurality of outputs 10A, 10B, 10X, which present F filtering consensus results following the processing of all p digital audio samples in a prosody sub period, that are connected initially to the first F inputs of a buffer 18, and additional outputs 11, 12, 13, and 23 that are connected to the buffer 18, the segmenter 5, a first advancer 14, and a second advancer 16 respectively. In total, the resolver 9 has F+1 inputs and F+3 outputs. Additionally, the resolver 9 also includes two counters (not shown). The first counter tracks the count of prosody sub periods per prosody period (i.e. from 1 to N) while the second counter tracks the number of samples per prosody sub period (i.e. from 1 to p).

For each prosody sub period, the resolver 9 computes, after each set of F filterings applied to each of p successive audio samples delivered to the filters via output 6 of segmenter 5, a consensus amplitude per frequency, thereby mimicking the stereocilia result of a frequency's corresponding inner hair cell to store in buffer 18 for subsequent delivery to an artificial neural network 21 (described below), thereby mimicking amplitudes to deliver to the brain's auditory cortex. In the preferred embodiment, each consensus amplitude is the maximum absolute value from among each filter's p output amplitudes per prosody sub period.

Whenever the resolver's 9 first counter (not shown) is equal to N, which occurs upon each Nth delivery of resolver's 9 F outputs 10A,10B, 10X to buffer 18, a control signal is sent via the resolver's 9 output 12 to the segmenter 5 causing the prosody period to restart resolution milliseconds later in the digital audio delivered via the receiver's 2 output 4 unless a new prosody period no longer fits in the remaining input audio, the latter condition stopping the device 1 for the current audio via input 3. In this embodiment, resolution=5. The resolver 9 also sends a control signal via its output 13 to the first advancer 14 upon sensing the completion of each set of F filterings per audio sample amplitude that is presented at output 6. The first advancer's 14 output 15 causes the segmenter 5 via its output 6

to advance one successive audio sample, within its digital audio input, to the IIR filters 7A,7B, 7X. Note that the initial sample is specified by the present invention's user.

Additionally, whenever the resolver's 9 second counter (not shown) is equal to p, the resolver 9 sends a control signal via the resolver's output 23 to a second advancer 16. The second advancer 16 causes the resolver 9 to release its current F outputs to the next available F inputs of a buffer 18, and causes the resolver 9 to begin accepting the next p groups of F outputs from filters 7A,7B, 7X via its corresponding inputs 8A,8B,8X.

The device 1 also includes an artificial neural network 21 that has a plurality of inputs equal to the value of P, where each of the plurality of inputs is connected to each of the corresponding outputs 19A,19B,19X of the buffer 18. The artificial neural network 21 also has a plurality of outputs 22A,22B,22X equal to the value of C, where C is defined as the number of classes trained into the artificial neural network 21.

The buffer 18 accumulates N prosody sub periods of data, then adjusts said data, then simultaneously releases the adjusted data at one time to the artificial neural network 21 in response to a signal via output 11 of resolver 19. Simultaneously, the buffer's 18 P output scalars 19A,19B,19X are sent to a spectral display means 20 which has been programmed to meaningfully interpret the given signals.

At the point when N=(rP.sub.t/p) iterations are complete, said adjustments to the accumulated data in buffer 18 include biasing the data according to a curve, amplifying the biased data, and normalizing the amplified results over the biased and amplified P scalars in buffer 18.

In the preferred embodiment, buffer 18 first applies a human hearing response biasing curve and then applies a logarithmic amplification to each scalar using natural log (aka, log base e), specifically, log(scalar+1). The result is then normalized to 1.0. The normalizing of the buffer's 18 biased and amplified contents after N resolver 9 to buffer 18 iterations are complete, i.e., when a prosody period is completed, mimics a human's concentrating on a sound within a prosody period while trying to identify it. After the N iterations are completed, resolver 9 signals authority to the

buffer 18, via output 11, to release its contents to the artificial neural network 21 and, via output 12, to advance the segmenter 5 resolution positive real milliseconds later where the segmenter 5 begins a new cycle if at least P.sub.t seconds of input audio are available at its input 4. If at least P.sub.t seconds of input audio are not available at its input 4, the segmenter 5 halts. Note that the number resolution is specified by the present invention's user. The successive segmenter 5 cycles, which apply the Prosody Period Structure described above, every succeeding resolution milliseconds, mimics the continual process of human listening. The likelihood that some subset of sounds trained into the artificial neural net 21 matches the input audio during the current prosody period is represented as a scalar at each of the artificial neural network's outputs 22A,22B,22X.

In this embodiment, this scalar is in the range 0 through 1.0 where 0 represents that there is no likelihood that some subset of sounds trained into the artificial neural net 21 matches the input audio within the current prosody period. In contrast, a scalar having the value of 1.0 represents the greatest likelihood that some subset of sounds trained into the artificial neural net 21 matches the input audio within the current prosody period. Additionally, logarithmically amplifying the buffer's 18 contents before normalization helps in overcoming a neural network's typical inability to effectively recognize widely dynamic scalars within its present range of comparison (the prosody period) by amplifying low amplitude scalars notably more than high amplitude scalars.

FIG. 3 is a schematic of each IIR resonator digital filter 30 that is utilized in the present invention.

Each IIR resonator digital filter 30 includes a first register 31 that has a first input 32, a second input 33, a third input 34, a fourth input 35, a fifth input 36, a first output 37, a second output 38, a third output 39, a fourth output 40, and a fifth output 41.

Each IIR resonator digital filter 30 also includes a multiplexer/demultiplexer 42. The multiplexer/demultiplexer 42 has a first input connected to the first output 37 of the first register 31, a second input connected to the second output 38 of the first register 31, a third input connected to the third output

39 of the first register 31, a fourth input connected to the fourth output 40 of the first register 31, a fifth input connected to the fifth output 41 of the first register 31, a sixth input 43, a seventh input 44, and a control input 45. The multiplexer/demultiplexer 42 also has several outputs including a first output connected to the first input 32 of the first register 31, a second output connected to the second input 33 of the first register 31, a third output connected to the third input 34 of the first register 31, a fourth output connected to the fourth input 35 of the first register 31, a fifth output connected to the fifth input 36 of the first register 31, a sixth output 46, a seventh output 47, and an eighth output 48.

Each IIR resonator digital filter 30 also includes a first multiplier 49, a second multiplier 52, and a third multiplier 55. The first multiplier 49 has a first input connected to the sixth output 46 of the multiplexer/demultiplexer 42, a second input 50, and an output 51.

The second multiplier 52 has a first input connected to the seventh output 47 of the multiplexer/demultiplexer 42, a second input 53, and an output 54.

The third multiplier 55 has a first input 56, a second input 57 which is supplied with successive audio samples via output 6 of segmenter 5 (shown in FIG. 1), an output 58 connected to the sixth input 43 of the multiplexer/demultiplexer 42.

Each IIR resonator digital filter 30 also includes an adder 59 that has a first input connected to the output 51 of the first multiplier 49, a second input connected to the output 54 of the second multiplier 52, a third input 60 connected to the output 58 of the third multiplier 55, and an output 61.

Additionally, each IIR resonator digital filter 30 includes a subtractor 62. The subtractor 62 has a first input connected to the output 61 of the adder 59, a second input connected to the eighth output 48 of the multiplexer/demultiplexer 42, and an output 63 that is connected to the seventh input 44 of the multiplexer/demultiplexer 42. Each IIR filter's 30 output appears at the output 63 of the subtractor 62 where output 63 supplies in succession, internal to resolver 9 (shown in FIG. 1), each of the p samples shown in the figure above (in paragraph 0023) from its

corresponding filter, per prosody sub period.

Finally, each IIR resonator digital filter 30 includes a second register 64. This second register 64 has an input 65 and a first output 56, a second output 53, a third output 50 and a fourth output 66. The second register 64 receives and stores four values via its input 65 prior to the activation of device 1 (shown in FIG. 1). These values represent an inverse of the user-definable gain (G.sup.-1), a first user-definable coefficient (C.sub.1), a second user-definable coefficient (C.sub.2), and a user-definable center frequency (CF) respectively for each IIR filters 30. G.sup.-1, C.sub.1, C.sub.2, and CF are programmed into the second register 64, via its input 65, and the IIR resonator digital filter 30, as implemented here, is a single-frequency pass filter. Additionally, the CF 66 value is eventually sent via an input 19Y to the spectral display means 20 described above (shown in FIG. 1).

FIGS. 4A-4E depict each of the programming states of the multiplexer/demultiplexer 42 in each IIR resonator digital filter's 30 (shown in FIG. 2) in more detail. With each additional sample, the programming state of each multiplexer/demultiplexer 42 in each IIR resonator digital filter's 30 rotates to the following state in response to control input 45 which activates upon sensing each new audio input sample delivered via output 6 (shown in FIG. 1).

FIG. 4A is a view of the first programming state of the multiplexer/demultiplexer 42 that has a control input 45. In this first programming state of the multiplexer/demultiplexer 42, the first input 37 is connected to the sixth output 46, the second input 38 is connected to the seventh output 47, the fifth input 41 is connected to the eighth output 48, the sixth input 43 is connected to the third output 34, and the seventh input 44 is connected to the fifth output 36.

FIG. 4B is a view of the second programming state of the multiplexer/demultiplexer 42 that has a control input 45. In this second programming state of the multiplexer/demultiplexer 42, the fifth input 41 is connected to the sixth output 46, the first input 37 is connected to the seventh output 47, the fourth input 40 is connected to the eighth output 48, the sixth input 43 is connected to the second output 33, and the seventh

input 44 is connected to the fourth output 35.

FIG. 4C is a view of the third programming state of the multiplexer/demultiplexer 42 that has a control input 45. In this third programming state of the multiplexer/demultiplexer 42, the fourth input 40 is connected to the sixth output 46, the fifth input 41 is connected to the seventh output 47, the third input 39 is connected to the eighth output 48, the sixth input 43 is connected to the first output 32, and the seventh input 44 is connected to the third output 34.

FIG. 4D is a view of the fourth programming state of the multiplexer/demultiplexer 42 that has a control input 45. In this fourth programming state of the multiplexer/demultiplexer 42, the third input 39 is connected to the sixth output 46, the fourth input 40 is connected to the seventh output 47, the second input 38 is connected to the eighth output 48, the sixth input 43 is connected to the fifth output 36, and the seventh 44 input is connected to the second output 33.

FIG. 4E is a view of the fifth programming state of the multiplexer/demultiplexer 42 that has a control input 45. In this fifth programming state of the multiplexer/demultiplexer 42, the second input 38 is connected to the sixth output 46, the third input 39 is connected to the seventh output 47, the first input 37 is connected to the eighth output 48, the sixth input 43 is connected to the fourth output 35, and the seventh input 44 is connected to the first output 32.

While the preferred embodiment has been disclosed and illustrated, a variety of substitutions and modifications can be made to the present invention without departing from the scope of the invention.

* * * * *

United States Patent **7,818,168**
Cusmariu **October 19, 2010**

Method of measuring degree of enhancement to voice signal

Abstract

A method of measuring the degree of enhancement made to a voice signal by receiving the voice signal, identifying formant regions in the voice signal, computing stationarity for each identified formant region, enhancing the voice signal, identifying formant regions in the enhanced voice signal that correspond to those identified in the received voice signal, computing stationarity for each formant region identified in the enhanced voice signal, comparing corresponding stationarity results for the received and enhanced voice signals, and calculating at least one user-definable statistic of the comparison results as the degree of enhancement made to the received voice signal.

Inventors: **Cusmariu; Adolf** (Eldersburg, MD)
Assignee: **The United States of America as represented by the Director, National Security Agency** (Washington, DC)
N/A (N/A)
Family ID: **42941270**
Appl. No.: **11/645,264**
Filed: **December 1, 2006**

Current U.S. Class:	**704/209**; 381/71.14; 704/208; 704/220; 704/225; 704/E21.002
Current CPC Class:	G10L 21/0364 (20130101)
Current International Class:	G10L 19/06 (20060101)
Field of Search:	;704/208,209,E21.002,220,225 ;381/71.14

References Cited [Referenced By]

U.S. Patent Documents

4827516	May 1989	Tsukahara et al.
5251263	October 1993	Andrea et al.
5742927	April 1998	Crozier et al.
5745384	April 1998	Lanzerotti et al.
5963907	October 1999	Matsumoto
6510408	January 2003	Hermansen
6618699	September 2003	Lee et al.
6704711	March 2004	Gustafsson et al.
7102072	September 2006	Kitayama
2001/0014855	August 2001	Hardy
2002/0167937	November 2002	Goodman
2004/0059572	March 2004	Ivanic et al.
2004/0167774	August 2004	Shrivastav
2004/0186716	September 2004	Morfitt, III et al.
2007/0047742	March 2007	Taenzer et al.
2009/0018825	January 2009	Bruhn et al.
2009/0063158	March 2009	Norden et al.

Other References

Purcell et al. "Compensation following real-time manipulation of formants in isolated vowels" Apr. 2006. cited by examiner .
Rohdenburg et al. "Objective Perceptual Quality Measures for the

Evaluation of Noise Reduction Schemes" 2005. cited by examiner .
Yan et al. "Formant-Tracking Linear Prediction Models for Speech Processing in Noisy Enviroments" 2005. cited by examiner .
Cohen et al. "Speech enhancement for non-stationarynoise environments" 2001. cited by examiner .
Gray et al. "A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis" 1974. cited by examiner .
Narendranath et al. "Transformation of formants for voice conversion using artificial neural networks" 1995. cited by examiner .
Martin et al. "A Noise Reduction Preprocessor for Mobile Voice Communication" 2004. cited by examiner .
Yan et al. "A Formant Tracking LP Model for Speech Processing in Car/Train Noise" 2004. cited by examiner .
Baer et al. "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times" 1993. cited by examiner .
Lee et al. "Formant Tracking Using Segmental Phonemic Information" 1999. cited by examiner.

Primary Examiner: Smits; Talivaldis Ivars
Assistant Examiner: Borsetti; Greg A
Attorney, Agent or Firm: Morelli; Robert D.

Claims

What is claimed is:

1. A method of measuring the degree of enhancement made to a voice signal, comprising the steps of: a) receiving, on a digital signal processor, the voice signal; b) identifying, on the digital signal processor, a user-definable number of formant regions in the voice signal; c)

computing, on the digital signal processor, stationarity for each formant region identified in the voice signal; d) enhancing, on the digital signal processor, the voice signal; e) identifying, on the digital signal processor, formant regions in the enhanced voice signal that correspond to those identified in step (b); f) computing, on the digital signal processor, stationarity for each formant region identified in the enhanced voice signal; g) comparing, on the digital signal processor, corresponding results of step (c) and step (f); and h) calculating, on the digital signal processor, at least one user-definable statistic of the results of step (g) as the degree of enhancement made to the voice signal.

2. The method of claim 1, further including the step of digitizing the received voice signal if the signal is received in analog format.

3. The method of claim 1, further including the step of segmenting the received voice signal into a user-definable number of segments.

4. The method of claim 1, wherein each step of identifying formant regions is comprised of the step of identifying formant regions using an estimate of a Cepstrum.

5. The method of claim 4, wherein the step of estimating a Cepstrum is comprised of selecting from the group of Cepstrum estimations consisting of a real Cepstrum and an absolute value of a complex Cepstrum.

6. The method of claim 1, wherein each step of computing stationarity for each formant region is comprised of the steps of: i) calculating an arithmetic average of the formant region; ii) calculating a geometric average of the formant region; iii) calculating a harmonic average of the formant region; and iv) comparing any user-definable combination of two results of step (i), step (ii), and step (iii).

7. The method of claim 6, wherein the step of comparing any user-definable combination of two results of step (i), step (ii), and step (iii) is comprised of the step of comparing any user-definable combination of two results of step (i), step (ii), and step (iii) using a comparison method selected from the group of comparison methods consisting of difference, difference divided by sum, and difference divided by one plus the

difference.

8. The method of claim 1, wherein each step of enhancing the voice signal is comprised of enhancing the voice signal using a voice enhancement method selected from the group of voice enhancement methods consisting of, echo cancellation, delay-time minimization, and volume control.

9. The method of claim 1, wherein the step of comparing corresponding results of step (c) and step (f) is comprised of comparing corresponding results of step (c) and step (f) using a comparison method selected from the group of comparison methods consisting of a ratio of corresponding results of step (c) and step (f) minus one and a difference of corresponding results of step (c) and step (f) divided by a sum of corresponding results of step (c) and step (f).

10. The method of claim 1, wherein the step of calculating at least one user-definable statistic of the results of step (g) is comprised of calculating at least one user-definable statistic of the results of step (g) using a statistical method selected from the group of statistical methods consisting of arithmetic average, median, and maximum value.

11. The method of claim 2, further including the step of segmenting the received voice signal into a user-definable number of segments.

12. The method of claim 11, wherein each step of identifying formant regions is comprised of the step of identifying formant regions using an estimate of a Cepstrum.

13. The method of claim 12, wherein the step of estimating a Cepstrum is comprised of selecting from the group of Cepstrum estimations consisting of a real Cepstrum and an absolute value of a complex Cepstrum.

14. The method of claim 13, wherein each step of computing stationarity for each formant region is comprised of the steps of: i) calculating an arithmetic average of the formant region; ii) calculating a geometric average of the formant region; iii) calculating a harmonic average of the formant region; and iv) comparing any user-definable combination of two results of step (i), step (ii), and step (iii).

15. The method of claim 14, wherein the step of comparing any user-definable combination of two results of step (i), step (ii), and step (iii) is comprised of the step of comparing any user-definable combination of two results of step (i), step (ii), and step (iii) using a comparison method selected from the group of comparison methods consisting of difference, ratio, difference divided by stun, and difference divided by one plus the difference.

16. The method of claim 15, wherein each step of enhancing the voice signal is comprised of enhancing the voice signal using a voice enhancement method selected from the group of voice enhancement methods consisting of echo cancellation, delay-time minimization, and volume control.

17. The method of claim 16, wherein the step of comparing corresponding results of step (c) and step (f) is comprised of comparing corresponding results of step (c) and step (f) using a comparison method selected from the group of comparison methods consisting of a ratio of corresponding results of step (c) and step (f) minus one and a difference of corresponding results of step (c) and step (f) divided by a sum of corresponding results of step (c) and step (f).

18. The method of claim 17, wherein the step of calculating at least one user-definable statistic of the results of step (g) is comprised of calculating at least one user-definable statistic of the results of step (g) using a statistical method selected from the group of statistical methods consisting of arithmetic average, median, and maximum value.

Description

FIELD OF INVENTION

The present invention relates, in general, to data processing and, in particular, to speech signal processing.

BACKGROUND OF THE INVENTION

Methods of voice enhancement strive to either reduce listener fatigue by minimizing the effects of noise or increasing the intelligibility of the recorded voice signal. However, quantification of voice enhancement has been a difficult and often subjective task. The final arbiter has been human, and various listening tests have been devised to capture the relative merits of enhanced voice signals. Therefore, there is a need for a method of quantifying an enhancement made to a voice signal. The present invention is such a method.

U.S. Pat. Appl. No. 20010014855, entitled "METHOD AND SYSTEM FOR MEASUREMENT OF SPEECH DISTORTION FROM SAMPLES OF TELEPHONIC VOICE SIGNALS," discloses a device for and method of measuring speech distortion in a telephone voice signal by calculating and analyzing first and second discrete derivatives in the voice waveform that would not have been made by human articulation, looking at the distribution of the signals and the number of times the signals crossed a predetermined threshold, and determining the number of times the first derivative data is less than a predetermined value. The present invention does not measure speech distortion as does U.S. Pat. Appl. No. 20010014855. U.S. Pat. Appl. No. 20010014855 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 20020167937, entitled "EMBEDDING SAMPLE VOICE FILES IN VOICE OVER IP (VoIP) GATEWAYS FOR VOICE QUALITY MEASUREMENTS," discloses a method of measuring voice quality by using the Perceptual Analysis Measurement System (PAMS) and the Perceptual Speech Quality Measurement (PSQM). The present invention does not use PAMS or PSQM as does U.S. Pat. Appl. No. 20020167937. U.S. Pat. Appl. No. 20020167937 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 20040059572, entitled "APPARATUS AND METHOD FOR QUANTITATIVE MEASUREMENT OF VOICE QUALITY IN PACKET NETWORK ENVIRONMENTS," discloses a device for and method of measuring voice quality by introducing noise into the voice signal, performing speech recognition on the signal containing noise. More noise

is added to the signal until the signal is no longer recognized. The point at which the signal is no longer recognized is a measure of the suitability of the transmission channel. The present invention does not introduce noise into a voice signal as does U.S. Pat. Appl. No. 20040059572. U.S. Pat. Appl. No. 20040059572 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 20040167774, entitled "AUDIO-BASED METHOD SYSTEM, AND APPARATUS FOR MEASUREMENT OF VOICE QUALITY," discloses a device for and method of measuring voice quality by processing a voice signal using an auditory model to calculate voice characteristics such as roughness, hoarseness, strain, changes in pitch, and changes in loudness. The present invention does not measure voice quality as does U.S. Pat. Appl. No. 20040167774. U.S. Pat. Appl. No. 20040167774 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 20040186716, entitled "MAPPING OBJECTIVE VOICE QUALITY METRICS TO A MOS DOMAIN FOR FIELD MEASUREMENTS," discloses a device for and method of measuring voice quality by using the Perceptual Evaluation of Speech Quality (PESQ) method. The present invention does not use the PESQ method as does U.S. Pat. Appl. No. 20040186716. U.S. Pat. Appl. No. 20040186716 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 20060093094, entitled "AUTOMATIC MEASUREMENT AND ANNOUNCEMENT VOICE QUALITY TESTING SYSTEM," discloses a device for and method of measuring voice quality by using the PESQ method, the Mean Opinion Score (MOS-LQO) method, and the R-Factor method described in International Telecommunications Union (ITU) Recommendation G.107. The present invention does not use the PESQ method, the MOS-LQO method, or the R-factor method as does U.S. Pat. Appl. No. 20060093094. U.S. Pat. Appl. No. 20060093094 is hereby incorporated by reference into the specification of the present invention.

SUMMARY OF THE INVENTION

It is an object of the present invention to measure the degree of enhancement made to a voice signal.

The present invention is a method of measuring the degree of enhancement made to a voice signal.

The first step of the method is receiving the voice signal.

The second step of the method is identifying formant regions in the voice signal.

The third step of the method is computing stationarity for each formant region identified in the voice signal.

The fourth step of the method is enhancing the voice signal.

The fifth step of the method is identifying the same formant regions in the enhanced voice signal as was identified in the second step.

The sixth step of the method is computing stationarity for each formant region identified in the enhanced voice signal.

The seventh step of the method is comparing corresponding results of the third and sixth steps.

The eighth step of the method is calculating at least one user-definable statistic of the results of the seventh step as the degree of enhancement made to the voice signal.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flowchart of the present invention.

DETAILED DESCRIPTION

The present invention is a method of measuring the degree of enhancement made to a voice signal. Voice signals are statistically

non-stationary. That is, the distribution of values in a signal changes with time. The more noise, or other corruption, that is introduced into a signal the more stationary its distribution of values becomes. In the present invention, the degree of reduction in stationarity in a signal as a result of a modification to the signal is indicative of the degree of enhancement made to the signal.

FIG. 1 is a flowchart of the present invention.

The first step 1 of the method is receiving a voice signal. If the voice signal is received in analog format, it is digitized in order to realize the advantages of digital signal processing (e.g., higher performance). In an alternate embodiment, the voice signal is segmented into a user-definable number of segments.

The second step 2 of the method is identifying a user-definable number of formant regions in the voice signal. A formant is any of several frequency regions of relatively great intensity and variation in the speech spectrum, which together determine the linguistic content and characteristic quality of the speaker's voice. A formant is an odd multiple of the fundamental frequency of the vocal tract of the speaker. For the average adult, the fundamental frequency is 500 Hz. The first formant region centers around the fundamental frequency. The second format centers around 1500 Hz. The third formant region centers around 2500 Hz. Additional formants exist at higher frequencies. Any number of formant regions derived by any sufficient method may be used in the present invention. In the preferred embodiment, the Cepstrum (pronounced kept-strum) is used to identify formant regions. Cepstrum is a jumble of the word "spectrum." It was arrived at by reversing the first four letters of the word "spectrum." A Cepstrum may be real or complex. A real Cepstrum of a signal is determined by computing a Fourier Transform of the signal, determining the absolute value of the Fourier Transform, determining the logarithm of the absolute value, and computing the Inverse Fourier Transform of the logarithm. A complex Cepstrum of a signal is determined by computing a Fourier Transform of the signal, determining the complex logarithm of the Fourier Transform, and computing the Inverse Fourier Transform of the logarithm. Either a real Cepstrum or an absolute value of a complex Cepstrum may be used in the present invention.

The third step 3 of the method is computing stationarity for each formant region identified in the voice signal. Stationarity refers to the temporal change in the distribution of values in a signal. A signal is deemed stationary if its distribution of values does not change within a user-definable period of time. In the preferred embodiment, stationarity is determined using at least one user-definable average of values in the user-definable formant regions (e.g., arithmetic average, geometric average, and harmonic average, etc.). The arithmetic average of a set of values is the sum of all values divided by the total number of values. The geometric average of a set of n values is found by calculating the product of the n values, and then calculating the nth-root of the product. The harmonic average of a set of values is found by determining the reciprocals of the values, determining the arithmetic average of the reciprocals, and then determining the reciprocal of the arithmetic average. The arithmetic average of a set of positive values is larger than the geometric average of the same values, and the geometric average of a set of positive values is larger than the harmonic average of the same values. The closer, or less different, these averages are to each other the more stationary is the corresponding voice signal. Any combination of these averages may be used in the present invention to gauge stationarity of a voice signal (i.e., arithmetic-geometric, arithmetic-harmonic, and geometric-harmonic). Any suitable difference calculation may be used in the present invention. In the preferred embodiment, difference calculations include difference, ratio, difference divided by sum, and difference divided by one plus the difference.

The fourth step 4 of the method is enhancing the voice signal received in the second step 2. In an alternate embodiment, a digitized voice signal and/or segmented voice signal is enhanced. Any suitable enhancement method may be used in the present invention (e.g., noise reduction, echo cancellation, delay-time minimization, volume control, etc.).

The fifth step 5 of the method is identifying formant regions in the enhanced voice signal that correspond to those identified in the second step 2.

The sixth step 6 of the method is computing stationarity for each formant

region identified in the enhanced voice signal.

The seventh step 7 of the method is comparing corresponding results of the third step 3 and the sixth step 6. Any suitable comparison method may be used in the present invention. In the preferred embodiment, the comparison method is chosen from the group of comparison methods that include ratio minus one and difference divided by sum.

The eighth step 8 of the method is calculating at least one user-definable statistic of the results of the seventh step 7 as the degree of enhancement made to the voice signal. Any suitable statistical method may be used in the present invention. In the preferred embodiment, the statistical method is chosen from the group of statistical methods including arithmetic average, median, and maximum value.

* * * * *

United States Patent **7,650,281**
Cusmariu **January 19, 2010**

Method of comparing voice signals that reduces false alarms

Abstract

A method of comparing voice samples to reduce false alarms by receiving a first voice sample, generating a model of the first voice sample, reordering the first voice sample, generating a model of the reordered first voice sample; receiving a second voice sample; generating a model of the second voice sample; reordering the second voice sample, generating a model of the reordered second voice sample, comparing at least one pairings of the models, and determining if the first voice sample matches the second voice sample if the model comparisons are within a user-definable threshold.

Inventors: **Cusmariu; Adolf** (Eldersburg, MD)
Assignee: **The U.S. Goverment as Represented By The Director, National Security Agency** (Washington, DC)
N/A (N/A)
Family ID: **41509952**
Appl. No.: **11/545,922**
Filed: **October 11, 2006**

Current U.S. Class:	**704/243**; 340/426.25; 455/411; 704/233; 704/246; 704/273; 704/274; 704/275
Current CPC Class:	G10L 17/08 (20130101)
Current International Class:	G10L 15/06 (20060101)
Field of Search:	;704/274,273,243,246,275,251,233 ;455/411 ;340/825.36,825.49,426.25

References Cited [Referenced By]

U.S. Patent Documents

4040013	August 1977	Carlson
4241329	December 1980	Bahler et al.
5629687	May 1997	Sutton et al.
5649055	July 1997	Gupta et al.
5736927	April 1998	Stebbins et al.
5842161	November 1998	Cohrs et al.
6054990	April 2000	Tran
6076055	June 2000	Bossemeyer, Jr. et al.
6233556	May 2001	Teunen et al.
6314401	November 2001	Abbe et al.
6765931	July 2004	Rabenko et al.
6931375	August 2005	Bossemeyer, Jr. et al.
7278028	October 2007	Hingoranee
2004/0236573	November 2004	Sapeluk
2005/0031097	February 2005	Rabenko et al.
2005/0107070	May 2005	Geupel
2005/0143996	June 2005	Bossemeyer, Jr. et al.

Primary Examiner: Chawan; Vijay B
Attorney, Agent or Firm: Morelli; Robert D.

Claims

What is claimed is:

1. A method of comparing voice samples to reduce false alarms: a) receiving a first voice sample; b) generating a first model of the first voice sample; c) reordering the first voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof; d) generating a second model of the result of step (c); e) receiving a second voice sample; f) generating a third model of the second voice sample; g) reordering the second voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof; h) generating a fourth model of the result of step (g); i) comparing at least one pair of models selected from the group of model pairs consisting of the first model and the third model, the first model and the fourth model, the second model and the third model, and the second model and the fourth model; and j) determining that the first voice sample matches the second voice sample if the results of step (i) are within a user-definable threshold.

2. The method of claim 1, wherein the step of receiving a first voice sample is comprised of the step of receiving a first voice sample in an analog format.

3. The method of claim 2, further including the step of digitizing the first voice sample.

4. The method of claim 2, further including the step of digitizing the first voice sample.

5. The method of claim 4, wherein the step of receiving a second voice sample is comprised of the step of receiving a first voice sample in an analog format.

6. The method of claim 5, further including the step of digitizing the first voice sample.

7. The method of claim 1, wherein the step of receiving a second voice sample is comprised of the step of receiving a first voice sample in an analog format.

8. A method of comparing voice samples to reduce false alarms: a) receiving a first voice sample; b) reordering the first voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof; c) combining the results of step (a) and step (b); d) generating a first model of the result of step (c); e) receiving a second voice sample; f) reordering the second voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof; g) combining the results of step (e) and step (f); h) generating a second model of the result of step (g); i) comparing the first model and the second model; and j) determining that the first voice sample matches the second voice sample if the result of step (i) is within a user-definable threshold.

9. The method of claim 8, wherein the step of receiving a first voice sample is comprised of the step of receiving a first voice sample in an analog format.

10. The method of claim 9, further including the step of digitizing the first voice sample.

11. The method of claim 9, further including the step of digitizing the first voice sample.

12. The method of claim 10, wherein the step of receiving a second voice sample is comprised of the step of receiving a first voice sample in an analog format.

13. The method of claim 8, wherein the step of receiving a second voice sample is comprised of the step of receiving a first voice sample in an analog format.

14. The method of claim 8, wherein the step of combining the results of step (e) and step comprised of the step of appending the result of step (f)

to the result of step (e).

15. The method of claim 14, further including the step of digitizing the first voice sample.

16. The method of claim 15, wherein the step of combining the results of step (e) and step (f) is comprised of the step of appending the result of step (f) to the result of step (e).

Description

FIELD OF INVENTION

The present invention relates, in general, to data processing for a specific application and, in particular, to digital audio data processing.

BACKGROUND OF THE INVENTION

Voice comparisons are used in speaker identification, recognition, authentication, and verification. False alarms occur in these methods when two voice samples that are different are compared and determined to be identical. False alarms reduce the performance of automated voice comparison methods. There is a need for a voice comparison method that reduces false alarms. The present invention is such a method.

U.S. Pat. No. 4,241,329, entitled "CONTINUOUS SPEECH RECOGNITION METHOD FOR IMPROVING FALSE ALARM RATES," discloses a method of recognizing keywords in continuous speech with improved false alarm rates by characterizing each keyword to be recognized by a template, selecting a sequence of patterns from a signal, comparing the pattern to the templates, determining which template matches the pattern, and applying a prosodic test to determine if the determination is a false alarm or not. The present invention does not characterize keywords with templates, compare patterns to templates, or apply a prosodic test to determine false alarms as does U.S. Pat. No. 4,241,329. U.S. Pat. No. 4,241,329 is hereby incorporated by reference

into the specification of the present invention.

U.S. Pat. Nos. 6,076,055 and 6,931,375 and U.S. Pat. Appl. No. 20050143996, each entitled "SPEAKER VERIFICATION METHOD," each disclose a method of verifying that a speaker is who the speaker claims to be by generating a code book, acquiring a number of training utterances from each of a number of speakers, receiving a number of test utterances, comparing the test utterances to the training utterances to form a number of decisions, weighting each decision, and combining the weighted decisions to form a verification decision. The present invention does not generate a code book, acquire a number of training utterances from each of a number of speakers, weight decisions, or combine weighted decisions to form a verification decision as does U.S. Pat. Nos. 6,076,055 and 6,931,375 and U.S. Pat. Appl. No. 20050143996. U.S. Pat. Nos. 6,076,055 and 6,931,375 and U.S. Pat. Appl. No. 20050143996 are hereby incorporated by reference into the specification of the present invention.

SUMMARY OF THE INVENTION

It is an object of the present invention to reduce false alarms in voice comparisons.

It is another object of the present invention to reduce false alarms in voice comparisons by reordering the voice samples being compared.

It is another object of the present invention to reduce false alarms in voice comparisons by reordering the voice samples being compared and conducting at least one comparison selected from the group of comparisons consisting of comparing a first signal as received to a second signal as received, comparing a first signal as received to a second signal which is reordered, and comparing a first signal which is reordered to a second signal as received.

The present invention is a method of comparing voice samples to reduce false alarms.

The first step of the method is receiving a first voice sample.

The second step of the method is generating a first model of the first voice sample.

The third step of the method is reordering the first voice sample.

The fourth step of the method is generating a second model of the reordered first voice sample.

The fifth step of the method is receiving a second voice sample.

The sixth step of the method is generating a third model of the second voice sample.

The seventh step of the method is reordering the second voice sample.

The eighth step of the method is generating a fourth model of the reordered second voice sample.

The ninth step of the method is comparing at least one pairing of the models.

The tenth step of the method is determining if the first voice sample matches the second voice sample if the model comparisons are within a user-definable threshold.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flowchart of the present invention; and

FIG. 2 is a flowchart of an alternate embodiment of the present invention.

DETAILED DESCRIPTION

The present invention is a method of comparing voice samples to reduce false alarms.

FIG. 1 is a flowchart of the present invention.

The first step 1 of the method is receiving a first voice sample. In the preferred embodiment, the first voice sample is received in an analog format. In an alternative embodiment, the first voice sample is received in a digital format. If the first voice sample is received in an analog format then the method of the present invention includes the step of digitizing the analog first voice sample.

The second step 2 of the method is generating a model of the first voice sample. Example of models of voice include spectral models, cepstral models, linear predictive coding (LPC), sinusoidal models, Hidden Markov models, and so on.

The third step 3 of the method is reordering the first voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof. Prior art method rely only on forwardly digitized recordings for model building and comparison, whereas the present invention relies on both forwardly digitized recordings and non-forwardly digitized recordings. A time-reversal of a digitally sampled voice recording is the recording reordered backwards. A segmented re-arrangement of the digitally sampled voice recording is first a determination of where the segments are in the recording, where the segments include phonemes, words, utterances, sentences, or any other acoustical or semantic grouping, and second a user-definable re-arrangement of the segments. Time-reversal and segmented re-arrangement may be combined to produce additional candidate sequences for voice model generation and comparison.

The fourth step 4 of the method is generating a model of the reordered first voice sample. A voice models generated from acoustically reordered digital voice recordings enhance the robustness of voice authentication by forcing an additional test derived from the same recording, yet containing novel features for validation.

The fifth step 5 of the method is receiving a second voice sample. In the preferred embodiment, the second voice sample is received in an analog format. In an alternative embodiment, the second voice sample is received in a digital format. If the second voice sample is received in an analog

format then the method of the present invention includes the step of digitizing the analog second voice sample.

The sixth step 6 of the method is generating a model of the second voice sample.

The seventh step 7 of the method is reordering the second voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof.

The eighth step 8 of the method is generating a model of the reordered second voice sample.

The ninth step 9 of the method is comparing at least one pairing of the models. In the preferred embodiment, the pairings include the model of the first voice sample and the model of the second voice sample, the model of the first voice sample and the model of the reordered second voice sample, the model of the reordered first voice sample and the model of the second voice sample, and the model of the first reordered voice sample and the model of the reordered second voice sample.

The tenth step 10 of the method is determining that the first voice sample matches the second voice sample if the model comparisons are within a user-definable threshold.

FIG. 2 is a flowchart of an alternate embodiment of the present invention.

The first step 21 of the alternative method is receiving a first voice sample. In the preferred embodiment, the first voice sample is received in an analog format. In an alternative embodiment, the first voice sample is received in a digital format. If the first voice sample is received in an analog format then the method of the present invention includes the step of digitizing the analog first voice sample.

The second step 22 of the alternative method is reordering the first voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a

combination thereof.

The third step 23 of the alternative method is combining the results of the first step 21 and the second step 22. In the preferred embodiment, the combination method is to append the result of the second step 22 to the result of the first step 21. However, any other suitable combination method may be employed.

The fourth step 24 of the alternative method is generating a model of the result of the third step 23.

The fifth step 25 of the alternative method is receiving a second voice sample. In the preferred embodiment, the second voice sample is received in an analog format. In an alternative embodiment, the second voice sample is received in a digital format. If the second voice sample is received in an analog format then the method of the present invention includes the step of digitizing the analog second voice sample.

The sixth step 26 of the alternative method is reordering the second voice sample using a reordering method selected from the group of reordering methods consisting of time-reversal, segmental rearrangement, and a combination thereof.

The seventh step 27 of the alternative method is combining the results of the fifth step 25 and the sixth step 26. In the preferred embodiment, the combination method is appending the result of the sixth step 26 to the result of the fifth step 25. However, any other suitable combination method may be employed.

The eighth step 28 of the alternative method is generating a model of the result of the seventh step 27.

The ninth step 29 of the alternative method is comparing the results of the fourth step 24 and the eighth step 28.

The tenth step 30 of the alternative method is determining that the first voice sample matches the second voice sample if the result of the ninth step 29 is within a user-definable threshold.

* * * * *

United States Patent **7,620,469**
Cusmariu **November 17, 2009**

Method of identifying digital audio signal format

Abstract

A method of identifying file format, converting file from assumed format and bit ordering to user-definable format, dividing file into blocks, determining frequencies of occurrence in blocks, creating first set of frequencies of occurrence less than and equal to most frequently occurring integer, creating second set of frequencies of occurrence greater than the most frequently occurring integer, creating third set of differences in first sets, creating fourth set of differences in second sets, replacing third and fourth sets with polarity indicators, summing polarity indicators, determining sum percentages, pairing percentages, determining pairing maximum number, determining statistics, determining maximum of statistics, assigning result to converted file, selecting another format and bit ordering and returning to third step, identifying converted file with maximum statistic, and determining format and bit ordering of file to be that of assumed format associated with converted file identified in last step.

Inventors: **Cusmariu; Adolf** (Eldersburg, MD)

Assignee: **The United States of America as represented by the Director of the National Security Agency** (Washington, DC)
N/A (N/A)

Family ID: 41279714
Appl. No.: 11/489,804
Filed: July 17, 2006

Current U.S. Class:	**700/94**; 381/58
Current CPC Class:	G10L 19/00 (20130101)
Current International Class:	G06F 17/00 (20060101)
Field of Search:	;700/94 ;381/58

References Cited [Referenced By]

U.S. Patent Documents

5374916	December 1994	Chu
5784544	July 1998	Stevens
6038400	March 2000	Bell et al.
6205223	March 2001	Rao et al.
6285637	September 2001	Manter et al.
6483988	November 2002	Noguchi et al.
6918554	July 2005	Stamm et al.
6999827	February 2006	Yong

Other References

JHOVE--JSTOR/Harvard Object Validation Environment, Available at http://hul.harvard.edu/jhove/, Dec. 12, 2006. cited by other.

Primary Examiner: Kuntz; Curtis
Assistant Examiner: Saunders, Jr.; Joseph
Attorney, Agent or Firm: Morelli; Robert D.

Claims

What is claimed is:

1. A method of identifying a format of a digital audio file, comprising the steps of: a) receiving the digital audio file; b) converting the digital audio file from a user-assumed digital audio format and bit ordering to a user-definable digital audio format and same bit ordering; c) dividing the converted digital audio file into user-definable blocks; d) determining, for each user-definable block, a list of unique integers therein and their frequencies of occurrence; e) creating, for each result of step (d), a first set that includes the frequencies of occurrence of the unique integers less than and equal to the most frequently occurring integer; f) creating, for each result of step (d), a second set that includes the frequencies of occurrence of the unique integers greater than the most frequently occurring integer; g) creating, for each first set, a third set that includes differences between adjacent frequencies of occurrence in the corresponding first set; h) creating, for each second set, a fourth set that includes differences between adjacent frequencies of occurrence in the second set; i) replacing each element in each third set and fourth set with a user-definable integer that indicates the polarity of the element; j) summing, for each third set, the polarity integers in the third set; k) summing, for each fourth set, the polarity integers in the fourth set; l) dividing each result of step (j) by a quantity of polarity integers in the corresponding third set and multiplying by 100; m) dividing each result of step (k) by a quantity of polarity integers in the corresponding fourth set and multiplying by 100; n) pairing each result of step (l) with the result of step (m) that corresponds to the same user-definable block; o) determining, for each result of step (n), the maximum integer in the pairing; p) determining, for each result of step (o), a user-definable set of statistics; q) determining the maximum of zero and the results of step (p); r) assigning the result of step (q) to the converted digital audio file; s) if additional digital audio formats and bit orderings are to be tested then selecting another digital audio format and bit ordering and returning to step (c), otherwise proceeding to the next step; t) identifying the converted digital audio file having the maximum assigned integer; and u) determining the format of the received digital audio file to be the assumed format and bit ordering associated with the converted digital audio file identified in step (t).

2. The method of claim 1, wherein the step of converting the digital audio file from a user-assumed digital audio format and bit ordering to a user-definable digital audio format and same bit ordering is comprised of the step of converting the digital audio file from a user-assumed digital audio format and bit ordering, where the bit ordering is selected from the group of bit orderings consisting of Most Significant Bit First and Least Significant Bit First.

3. The method of claim 1, wherein the step of converting the digital audio file from a user-assumed digital audio format and bit ordering to a user-definable digital audio format and same bit ordering is comprised of the step of converting the digital audio file to an 8-bit linear format sampled at 8 KHz and the same bit ordering.

4. The method of claim 1, wherein the step of dividing the converted digital audio file into user-definable blocks is comprised of the step of dividing the converted digital audio file into blocks containing 4 seconds of data sampled at 8 KHz.

5. The method of claim 1, wherein the step of determining, for each user-definable block, a list of unique integers therein and their frequencies of occurrence is comprised of the step of determining, for each user-definable block, a list of unique integers therein and their frequencies of occurrence, wherein the integers are listed in order from lowest integer to highest integer.

6. The method of claim 1, wherein the step of replacing each element in each third set and fourth set with a user-definable integer that indicates the polarity of the element is comprised of the step of replacing each element in each third set and fourth set with a 1 for each positive element and a -1 for each negative element.

7. The method of claim 1, wherein the step of determining, for each result of step (o), a user-definable integer of statistics is comprised of the step of determining, for each result of step (o), a mean and a median.

8. The method of claim 1, further including the step of removing from the

result of step (b) runs of integers that differ by no more than a user-definable integer.

9. The method of claim 8, wherein the step of removing from the result of step (b) runs of integers that differ by no more than a user-definable integer is comprised of the step of removing from the result of step (b) runs of integers that differ by no more than a integer selected from the group of integers consisting of 0, 1, and 2.

10. The method of claim 1, further including the step of removing from the result of step (b) integers outside of a user-definable range.

11. The method of claim 10, wherein the step of removing from the result of step (b) integers outside of a user-definable range is comprised of the step of removing from the result of step (b) integers outside of a range of -15 to 15.

12. The method of claim 11, wherein the step of converting the digital audio file from a user-assumed digital audio format and bit ordering to a user-definable digital audio format and same bit ordering is comprised of the step of converting the digital audio file to an 8-bit linear format sampled at 8 KHz and the same bit ordering.

13. The method of claim 12, wherein the step of dividing the converted digital audio file into user-definable blocks is comprised of the step of dividing the converted digital audio file into blocks containing 4 seconds of data sampled at 8 KHz.

14. The method of claim 13, wherein the step of determining, for each user-definable block, a list of unique integers therein and their frequencies of occurrence is comprised of the step of determining, for each user-definable block, a list of unique integers therein and their frequencies of occurrence, wherein the integers are listed in order from lowest integer to highest integer.

15. The method of claim 14, wherein the step of replacing each element in each third set and fourth set with a user-definable integer that indicates the polarity of the element is comprised of the step of replacing each element

in each third set and fourth set with a 1 for each positive element and a -1 for each negative element.

16. The method of claim 15, wherein the step of determining, for each result of step (o), a user-definable set of statistics is comprised of the step of determining, for each result of step (o), a mean and a median.

17. The method of claim 16, further including the step of removing from the result of step (b) runs of integers that differ by no more than a user-definable integer.

18. The method of claim 17, wherein the step of removing from the result of step (b) runs of integers that differ by no more than a user-definable integer is comprised of the step of removing from the result of step (b) runs of integers that differ by no more than a integer selected from the group of integers consisting of 0, 1, and 2.

19. The method of claim 18, further including the step of removing from the result of step (b) integers outside of a user-definable range.

20. The method of claim 19, wherein the step of removing from the result of step (b) integers outside of a user-definable range is comprised of the step of removing from the result of step (b) integers outside of a range of -15 to 15.

Description

FIELD OF INVENTION

The present invention relates, in general, to data processing for a specific application and, in particular, to digital audio data processing.

BACKGROUND OF THE INVENTION

Audio signals were initially recorded as analog signals. An analog representation of an audio signal has a continuous nature (e.g., a smooth

curving line), as opposed to a digital representation of an audio signal, which has a discrete nature. Each sample in a digital representation is a integer in base two, or binary, format, where each binary digit, or bit, in the integer is either a one or a zero.

It is difficult, if not impossible, to copy or transmit an analog representation of a signal perfectly, whereas it is easy to do the same for a digital representation of a signal. Any deviation in an analog representation of an audio signal as compared to the original signal represents loss of audio quality. Since digital representations of audio signals can be copied or transmitted perfectly, it is the preferred representation for audio signals.

There are many different formats for digitally representing an audio signal. The essential characteristics of a digital representation is its encoding scheme (e.g., .mu.-law (pronounced mu-law), a-law), the integer of bits that represent each sample in the signal (e.g., 8-bit, 16-bit, 32-bit), and the sampling rate per second used to digitize the signal (e.g., 8 KHz, 16 KHz, 32 KHz). The integer of bits that represent a integer is commonly referred to as the word, byte, or block length.

With audio signals increasingly being included in computer communication, different file formats have arisen. Some file formats are self-describing. That is, they include header information that says what digital representation was used to encode the audio signal. However, header information is not always accurate. Other file formats, referred to as headerless formats, do not say what digital representation was used to encode an audio signal. Such formats can be difficult to decipher, and may require one to listen to the audio file.

Computer files include extensions. For example, a file named filename.ext, has ".ext" as its file extension. The most common file extension on the INTERNET include .snd, .au, .aiff .wav, and .mov. The .snd extension is ambiguous because it could indicate the self-describing format of a Next Computer or the headerless format of an Apple Macintosh computer. The .au format is used in SUN Microsystems computers to indicate .mu.-law encoding. The .aiff format is used in Apple Macintosh computers. The .wav format is used on computers running the Microsoft Windows operating system. The .mov format is used in QuickTime movies. The extension is

supposed to indicate the format used to encode the file. However, just as headers in self-describing files do not always describe the file format used, neither do file extensions.

U.S. Pat. No. 6,285,637, entitled "METHOD AND APPARATUS FOR AUTOMATIC SECTOR FORMAT IDENTIFICATION IN AN OPTICAL STORAGE DEVICE," discloses a method of distinguishing between the formats for Compact Disc-Read Only Memory (CD-ROM) and Compact Disc-Digital Audio (CD-DA) on an optical storage device by examining a Q-channel data-type indicator bit. The value of the bit indicates whether the format of the optical storage device is CD-ROM or CD-DA. The present invention does not examine a Q-channel data-type bit to determine format as does U.S. Pat. No. 6,285,637. In addition, U.S. Pat. No. 6,285,637 does not disclose a method of distinguishing between digital audio formats as does the present invention. U.S. Pat. No. 6,285,637 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,483,988, entitled "AUDIO AND VIDEO SIGNALS RECORDING APPARATUS HAVING SIGNAL FORMAT DETECTION FUNCTION," discloses a method of determining if received audio is in AC-3 format (i.e., Digital Dolby) or in a format supported by MPEG by extracting bit stream and header information. The present invention does not use header information to determine digital audio format. U.S. Pat. No. 6,483,988 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,918,554, entitled "TAPE CARTRIDGE FORMAT IDENTIFICATION IN A SINGLE REEL TAPE HANDLING DEVICE," discloses a method of identifying the format of a tape by including information on a tape cartridge leader that indicates the format of the tape. The present invention does not use information of a leader of tape to determine format as does U.S. Pat. No. 6,918,554. U.S. Pat. No. 6,918,554 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,999,827, entitled "AUTO-DETECTION OF AUDIO INPUT FORMATS," discloses a device for distinguishing between two different digital audio formats, 12S and SPDIF, by detecting edge transmissions

and using a time counter to determine the time slot of the received signal. A time slot for 12S is in the range from 81.38 nanoseconds to 488.28 nanoseconds. A time slot for SPDIF is in the range from 5.2 microseconds to 250 microseconds. The format for whichever range encompasses the time slot determined by U.S. Pat. No. 6,999,827 is determined to be the format of the received signal The present invention does not use edge detection and time slot estimation to determine format as does U.S. Pat. No. 6,999,827. U.S. Pat. No. 6,999,827 is hereby incorporated by reference into the specification of the present invention.

JSTOR and Harvard University Library collaborated to develop a framework for format validation of various digital objects. JSTOR is a not-for-profit organization that maintains an archive of important scholarly journals. The framework that was developed is called JHOVE (pronounced "jove"), which stands for the JSTOR/Harvard Object Validation Environment. JHOVE identifies the format of various self-defining digital formats by determining whether or not the signal is formed according to the requirements of a particular digital format (e.g., does the signal contain a required integer at required byte offsets, does the signal contain all of the required components, does the signal include any components that it should not, etc.). The present invention does not determine format by determining whether or not the signal is formed according to the requirements of a particular digital format as does JHOVE. In addition, JHOVE cannot identify a headerless digital format as does the present invention.

There is a need for a method of identifying digital audio formats, whether self-defining or headerless. The present invention is such a method.

SUMMARY OF THE INVENTION

It is an object of the present invention to identify the format of a digital audio signal.

It is another object of the present invention to identify the format of a digital audio signal that is either self-defining or headerless.

The present invention is a method of identifying a format of a digital audio

file.

The first step of the method is receiving the digital audio file.

The second step of the method is converting the digital audio file from a user-assumed digital integer audio format and bit ordering to a user-definable digital integer audio format and same bit ordering.

The third step of the method is dividing the converted digital audio file into user-definable blocks.

The fourth step of the method is determining, for each block, a list of unique integers therein and their frequencies of occurrence.

The fifth step of the method is creating, for each result of the fourth step, a first set that includes the frequencies of occurrence of the unique integers less than and equal to the most frequently occurring integer, also known as the mode.

The sixth step of the method is creating, for each result of the fourth step, a second set that includes the frequencies of occurrence of the unique integers greater than the mode.

The seventh step of the method is creating, for each first set, a third set that includes differences between adjacent frequencies of occurrence in the corresponding first set.

The eighth step of the method is creating, for each second set, a fourth set that includes differences between adjacent frequencies of occurrence in the second set.

The ninth step of the method is replacing each element in each third set and fourth set with a user-definable integer that indicates the polarity (or sign) of the element, that is, positive or negative.

The tenth step of the method is summing, for each third set, the polarity integers in the third set.

The eleventh step of the method is summing, for each fourth set, the polarity integers in the fourth set.

The twelfth step of the method is dividing each result of the tenth step by the quantity of integers in the corresponding third set and multiplying by 100.

The thirteenth step of the method is dividing each result of the eleventh step by a quantity of integers in the corresponding fourth set and multiplying by 100.

The fourteenth step of the method is pairing each result of the twelfth step with the result of the thirteenth step that corresponds to the same user-definable block.

The fifteenth step of the method is determining, for each result of the fourteenth step, the maximum number in the pairing.

The sixteenth step of the method is determining, for each result of the fifteenth step, a user-definable number of statistical parameters; means and medians are typical, though not exclusive, examples.

The seventeenth step of the method is determining the maximum of zero and the results of the sixteenth step.

The eighteenth step of the method is assigning the result of the seventeenth step to the converted digital audio file.

The nineteenth step of the method is selecting another digital audio format and bit ordering and returning to the third step if additional digital audio formats and bit orderings are to be tested. Otherwise, proceeding to the next step.

The twentieth step of the method is identifying the converted digital audio file having the maximum assigned integer.

The twenty-first step of the method is determining the format and bit ordering of the received digital audio file to be that of the assumed format

associated with the converted digital audio file identified in the twentieth step.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flowchart of the steps of the present invention.

DETAILED DESCRIPTION

The present invention is a method of identifying a format of a digital audio file.

FIG. 1 is a flowchart of the present invention.

The first step 1 of the method is receiving the digital audio file, where the file includes binary integers that represent the components of the audio signal contained in the file. The received file may be in any digital audio format. Examples of some digital audio formats are listed above.

The second step 2 of the method is converting the digital audio file from a user-assumed digital audio format and bit ordering to a user-definable digital audio format and same bit ordering. In the present invention, the user assumes that the received file is in any integer of candidate formats and bit orderings. The received file will then be converted from the assumed format and analyzed. The converted file that is analyzed most favorably as described by the following steps will be identified as the correct format of the received file. In the first step 1, the user selects the first assumed format to be analyzed. In a subsequent step, another format and bit ordering will be selected and analyzed. This process will continue until the user has analyzed each format and bit ordering that he desires. Examples of bit ordering include Most Significant Bit First (MSBF) and Least Significant Bit First (LSBF). For example, if an audio sample is represented by the integer 23 then it may be represented in binary as either 10111 in MSBF or as 11101 in LSBF. Since each format assumed by the user, there are two possible bit orderings. Therefore, 2N analyses must be performed for N formats assumed. The format of the analysis that results in the highest figure-of-merit is determined to be the format of the received file. In the preferred embodiment, the received file is converted

from its assumed format and bit ordering to an 8-bit linear format sampled at 8 KHz, in the same bit ordering.

Converted digital audio files often include long runs of the same, or nearly the same, integer. Such runs take up processing time and do not add proportionately to the accuracy of the result. So, they may be eliminated. In an alternate embodiment, runs of the same, or nearly the same, integer are removed. In the preferred embodiment, a run includes nearly the same integer if no integer in the run differs from any other integer in the run by at most 2.

Digital audio files may employ a large range of integers for better fidelity (e.g., -128 to 128). Using the full range of values takes up processing time and does not add proportionately to the accuracy of result. Therefore, the range of integers in the converted file may be limited. In a second alternate embodiment, integers in the converted file that are outside of a user-definable range are removed. In the preferred embodiment, the user-definable integer range is -15 to 15.

The third step 3 of the method is dividing the converted digital audio file into user-definable blocks. In the preferred embodiment, the converted digital audio file is divided into blocks containing samples comprising 4 seconds in duration at a sampling rate of 8 KHz.

The fourth step 4 of the method is determining, for each user-definable block, a list of unique integers therein and their frequencies of occurrence. In the preferred embodiment, the integers are sorted in order from lowest integer to highest integer. For example, a block may include the following subset of integers: [-4 3 3 3 20 -4 -15 32 3 20 3 32 3 -15 -15 32 3 -28 -28 -4 -28 -15 -4 20 32 29 -4 3 29 20]. The unique integers in this block, from lowest to highest, are [-28 -15 -4 3 20 29 32]. The frequencies of occurrence, or density, for these unique integers are [3 4 5 8 4 2 4].

The fifth step 5 of the method is creating, for each result of the fourth step 4, a first set that includes the frequencies of occurrence of the unique integers less than and equal to the most frequently occurring integer. The most frequently occurring integer in a block of digital audio is commonly referred to as its mode. In the example above, the mode is 3. For the

example above, the first set is [3 4 5 8]. A first set will be created for each block in the converted file. The first set represents increasing density.

The sixth step 6 of the method is creating, for each result of the fourth step 4, a second set that includes the frequencies of occurrence of the unique integers greater than the most frequently occurring integer or mode. For the example above, the second set is [4 2 4]. A second set will be created for each block in the converted file. The second set represents decreasing density.

The seventh step 7 of the method is creating, for each first set, a third set that includes differences between adjacent frequencies of occurrence in the corresponding first set, in a next-minus-previous order. In the example above, the first set of [3 4 5 8] results in a third set of [1 1 3] (i.e., the differences between 4 and 3, 5 and 4, and 8 and 5), where the integer on the left is subtracted from the integer on the right. A third set is created for each block. The differences will be used to produce a measure of how often the sign, or polarity, of a segment increases or decreases relative to its length.

The eighth step 8 of the method is creating, for each second set, a fourth set that includes differences between adjacent frequencies of occurrence in the second set. In the example above, the second set of [4 2 4] results in a fourth set of [-2 2] (i.e., the differences between 2 and 4, and 4 and 2), where the integer on the left is subtracted from the integer on the right. A fourth set is created for each block. The differences will be used to produce a measure of how often the sign, or polarity, of a segment increases or decreases relative to its length.

The ninth step 9 of the method is replacing each element in each third set and fourth set with a user-definable integer that indicates the polarity of the element. In the preferred embodiment, a 1 is used to indicate a positive element and a -1 is used to indicate a negative element. In the example above, the third set of [1 1 3] is replaced with [1 1 1], and the fourth set of [-2 2] is replaced with [-1 1]. Similar replacements are made for each third and fourth set.

The tenth step 10 of the method is summing, for each third set, the polarity

integers in the third set. In the example above, the third block of [1 1 1] sums to 3. Similar sums are determined for each third set.

The eleventh step 11 of the method is summing, for each fourth set, the polarity integers in the fourth set. In the example above, the fourth block of [-1 1] sums to 0. Similar sums are determined for each fourth set.

The twelfth step 12 of the method is dividing each result of the tenth step 10 by a quantity of polarity integers in the corresponding third set and multiplying by 100. In the example above, the sum of the third set (i.e., 3) is divided by the integer of polarity integers in the third set (i.e., 3) to produce 1. The result (i.e., 1) is then multiplied by 100 to get 100, which is the percentage of the polarity of the elements with respect to its length. Similar percentages are created for each third set.

The thirteenth step 13 of the method is dividing each result of the eleventh step 11 by a quantity of polarity integers in the corresponding fourth set and multiplying by 100. In the example above, the sum of the fourth set (i.e., 0) is divided by the integer of polarity integers in the fourth set (i.e., 2) to produce 0. The result (i.e., 0) is then multiplied by 100 to get 0, which is the percentage of the polarity of the elements with respect to its length. Similar percentages are created for each fourth set.

The fourteenth step 14 of the method is pairing each result of the twelfth step 12 with the result of the thirteenth step 13 that corresponds to the same user-definable block. In the example, the pair for the associated third and fourth sets is [100, 0]. Similar pairs are created for each associated third and fourth sets. These measures represent the local monotonic nature of the density of each block, increasing or decreasing.

The fifteenth step 15 of the method is determining, for each result of the fourteenth step 14, the maximum integer in the pairing. In the example, the maximum element in the pair [100, 0] is 100. Similar maximums will be identified for each pairing.

The sixteenth step 16 of the method is determining, for each result of the fifteenth step 15, a user-definable set of statistics. In the preferred embodiment, the statistics are mean and median. However, other statistics

are possible. If in the example above included not only the pairing maximum pairing of 100 but also pairing maximums of 90, 85, 70, and 65 then the mean would be 82 and the median would be 85.

The seventeenth step 17 of the method is determining the maximum of zero and the results of the sixteenth step 16. In the example above, the maximum of 0, 82, and 85 is 85. The result of the seventeenth step is the numerical result of the analysis of the converted file of the received file, where the received file was assumed to be in a user-definable format and bit ordering. This integer will be compared to similarly generated integers for converted files of the received file, where different formats and bit orders are assumed.

The eighteenth step 18 of the method is assigning the result of the seventeenth step 17 to the converted digital audio file.

The nineteenth step 19 of the method is selecting another digital audio format and bit ordering and returning to the third step 3 if additional digital audio formats and bit orderings are to be tested. Otherwise, proceeding to the next step.

The twentieth step 20 of the method is identifying the converted digital audio file having the maximum assigned integer.

The twenty-first, and last, step 21 of the method is determining the format and bit ordering of the received digital audio file to be that of the assumed format associated with the converted digital audio file identified in the twentieth step 20.

In the present invention, the user converted the received file a user-definable number of times, assuming a different combination of format and bit ordering of the received file per conversion. Each converted file was then analyzed to generate a number which represents an estimation of the maximal polarity monotonicity percentage for each file. Then, the converted file that generated the highest such estimate was identified. Finally, the assumed format and bit ordering associated with the converted file with the highest integer was determined to be the format and bit ordering of the received file.

* * * * *

United States Patent **7,571,093**
Cusmariu **August 4, 2009**

Method of identifying duplicate voice recording

Abstract

A method of identifying duplicate voice recording by receiving digital voice recordings, selecting one of the recordings; segmenting the selected recording, extracting a pitch value per segment, estimating a total time that voice appears in the recording, removing pitch values that are less than and equal to a user-definable value, identifying unique pitch values, determining the frequency of occurrence of the unique pitch values, normalizing the frequencies of occurrence, determining an average pitch value, determining the distribution percentiles of the frequencies of occurrence, returning to the second step if additional recordings are to be processed, otherwise comparing the total voice time, average pitch value, and distribution percentiles for each recording processed, and declaring the recordings duplicates that compared to within a user-definable threshold for total voice time, average pitch value, and distribution percentiles.

Inventors: **Cusmariu; Adolf** (Eldersburg, MD)

Assignee: **The United States of America as represented by the Director, National Security Agency** (Washington, DC)
N/A (N/A)

Family ID: 40910223
Appl. No.: 11/506,090
Filed: August 17, 2006

Current U.S. Class: **704/207**; 704/201; 704/239; 704/270; 707/999.101
Current CPC Class: G10L 25/48 (20130101); G10L 25/90 (20130101)
Current International Class: G10L 11/04 (20060101); G06F 17/00 (20060101); G10L 15/00 (20060101); G10L 21/00 (20060101)

References Cited [Referenced By]

U.S. Patent Documents

6067444	May 2000	Cannon et al.
6766523	July 2004	Herley
7035867	April 2006	Thompson et al.
7120581	October 2006	Kahn et al.
7421305	September 2008	Burges et al.
2005/0182629	August 2005	Coorman et al.

Primary Examiner: Sked; Matthew J
Attorney, Agent or Firm: Morelli; Robert D.

Claims

What is claimed is:

1. A method of identifying duplicate voice recording, comprising the steps of: a) receiving a plurality of digital voice recordings; b) selecting one of said plurality of digital voice recordings; c) segmenting the selected digital voice recording; d) extracting a pitch value from each segment; e) estimating a total time that voice appears in the selected digital voice recording; f) removing pitch values that are less than and equal to a user-definable value; g) identifying unique pitch values in the result of step (f); h) determining the frequency of occurrence of the unique pitch values;

i) normalizing the result of step (h) so that the frequencies of occurrence are greater than zero and less than one; j) determining an average pitch value from the pitch values remaining after step (f); k) determining the distribution percentiles of the result of step (h); l) if additional digital voice recordings are to be processed then returning to step (b), otherwise proceeding to the next step; m) comparing the results of steps (e), (j), and (k) for each digital voice recording processed; and n) declaring the digital voice recordings duplicates that compared to within a user-definable threshold for each of the results of steps (e), (j), and (k).

2. The method of claim 1, wherein the step of receiving a plurality of digital voice recordings is comprised of the step of receiving a plurality of digital voice recordings in any digital format.

3. The method of claim 2, wherein the step of segmenting the selected digital voice recording is comprised of the step of segmenting the selected digital voice recording into 16 millisecond segments sampled at 8000 samples per second.

4. The method of claim 3, wherein the step of extracting a pitch value from each segment is comprised of the step of extracting a pitch value from each segment using any pitch extraction method.

5. The method of claim 4, wherein the step of estimating a total time that voice appears in the selected digital voice recording is comprised of the step of estimating a total time that voice appears in the selected digital voice recording using the pitch values.

6. The method of claim 5, wherein the step of removing pitch values that are less than and equal to a user-definable value is comprised of the step of removing pitch values that are less than and equal to zero.

7. The method of claim 6, further including the step of removing pitch values that vary from one pitch value to the next pitch value by less than or equal to a user-definable value.

8. The method of claim 7, wherein the step of normalizing the result of step (h) so that the frequencies of occurrence are greater than zero and less

than one is comprised of the step of dividing the result of step (h) by the number of pitch values remaining after step (f).

9. The method of claim 8, wherein the step of determining an average pitch value from the pitch values remaining after step (f) is comprised of the step of determining an average pitch value from the pitch values remaining after step (f) and rounding to the nearest integer.

10. The method of claim 1, wherein the step of segmenting the selected digital voice recording is comprised of the step of segmenting the selected digital voice recording into 16 millisecond segments sampled at 8000 samples per second.

11. The method of claim 1, wherein the step of extracting a pitch value from each segment is comprised of the step of extracting a pitch value from each segment using any pitch extraction method.

12. The method of claim 1, wherein the step of extracting a pitch value from each segment is comprised of the step of extracting a pitch value from each segment using a cepstral pitch extraction method.

13. The method of claim 1, wherein the step of estimating a total time that voice appears in the selected digital voice recording is comprised of the step of estimating a total time that voice appears in the selected digital voice recording using the pitch values.

14. The method of claim 1, wherein the step of removing pitch values that are less than and equal to a user-definable value is comprised of the step of removing pitch values that are less than and equal to zero.

15. The method of claim 1, further including the step of removing pitch values that vary from one pitch value to the next pitch value by less than or equal to a user-definable value.

16. The method of claim 1, wherein the step of normalizing the result of step (h) so that the frequencies of occurrence are greater than zero and less than one is comprised of the step of dividing the result of step (h) by the number of pitch values remaining after step (f).

17. The method of claim 1, wherein the step of determining an average pitch value from the pitch values remaining after step (f) is comprised of the step of determining an average pitch value from the pitch values remaining after step (f) and rounding to the nearest integer.

Description

FIELD OF INVENTION

The present invention relates, in general, to data processing for a specific application and, in particular, to digital audio data processing.

BACKGROUND OF THE INVENTION

Voice storage systems may contain duplicate voice recordings. Duplicate recordings reduce the amount of storage available for storing unique recordings.

Prior art methods of identifying duplicate voice recordings include manually listening to records and translating voice into text and comparing the resulting text. Listening to voice recordings is time consuming, and the performance of speech-to-text conversion is highly dependent on language, dialect, and content.

Identifying duplicate voice records is further complicated by the fact that two recordings of different lengths may be duplicates, and two recordings of the same length may not be duplicates. Therefore, there is a need for a method of identifying duplicate voice records that do not have the shortcomings of the prior art methods. The present invention is just such a method.

U.S. Pat. No. 6,067,444, entitled "METHOD AND APPARATUS FOR DUPLICATE MESSAGE PROCESSING IN A SELECTIVE CALL DEVICE," discloses a device for and method of receiving a first message that includes a message sequence number. A subsequent message is

received. If the subsequent message has the same message sequence number, address, vector type, length, data, and character total then the subsequent message is determined to be a duplicate. The present invention does not employ message sequence number, address, vector type, and character total as does U.S. Pat. No. 6,067,444. U.S. Pat. No. 6,067,444 is hereby incorporated by reference into the specification of the present invention.

SUMMARY OF THE INVENTION

It is an object of the present invention to identify duplicate voice recording.

It is another object of the present invention to identify duplicate voice recording without listening to the recording.

It is another object of the present invention to identify duplicate voice recording without converting the voice to text.

The present invention is a method of identifying duplicate voice recording.

The first step of the method is receiving digital voice recordings.

The second step of the method is selecting one of the recordings.

The third step of the method is segmenting the selected recording.

The fourth step of the method is extracting a pitch value per segment.

The fifth step of the method is estimating a total time that voice appears in the recording.

The sixth step of the method is removing pitch values that are less than and equal to a user-definable value.

The seventh step of the method is identifying unique pitch values.

The eighth step of the method is determining the frequency of occurrence of the unique pitch values.

The ninth step of the method is normalizing the frequencies of occurrence.

The tenth step of the method is determining an average pitch value.

The eleventh step of the method is determining the distribution percentiles of the frequencies of occurrence.

The twelfth step of the method is returning to the second step if additional recordings are to be processed. Otherwise, proceeding to the next step.

The thirteenth step of the method is comparing the total voice time, average pitch value, and distribution percentiles for each recording processed.

The fourteenth step of the method is declaring the recordings duplicates that compared to within a user-definable threshold for total voice time, average pitch value, and distribution percentiles.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flowchart of the steps of the present invention.

DETAILED DESCRIPTION

The present invention is a method of identifying duplicate voice recording.

FIG. 1 is a flowchart of the present invention.

The first step 1 of the method is receiving a plurality of digital voice recordings. Digital voice recordings may be received in any digital format.

The second step 2 of the method is selecting one of the digital voice recordings.

The third step 3 of the method is segmenting the selected digital voice recording. In the preferred embodiment, the selected digital voice recording is segmented into 16 millisecond segments sampled at 8000

samples per second.

The fourth step 4 of the method is extracting a pitch value from each segment. The pitch value may be extracted using any pitch extraction method. In the preferred embodiment, a cepstral method is used to extract pitch values.

The fifth step 5 of the method is estimating a total time that voice appears in the selected digital voice recording. In the preferred embodiment, the extracted pitch values are used to estimate the total time that voice appears in the selected digital voice recording.

The sixth step 6 of the method is removing pitch values that are less than and equal to a user-definable value. In the preferred embodiment, the user-definable value is zero. In an alternate embodiment, then method further includes a step of removing pitch values that vary from one pitch value to the next pitch value by less than or equal to a user-definable value.

The seventh step 7 of the method is identifying unique pitch values in the result of the sixth step 6.

The eighth step 8 of the method is determining the frequency of occurrence of the unique pitch values.

The ninth step 9 of the method is normalizing the result of the eighth step 8 so that the frequencies of occurrence are greater than zero and less than one. In the preferred embodiment, the results of the eighth step 8 are normalized by dividing the result of the eighth step 8 step by the number of pitch values remaining after the sixth step 6.

The tenth step 10 of the method is determining an average pitch value from the pitch values remaining after the sixth step 6. In the preferred embodiment, the average pitch value is rounded to the nearest integer.

The eleventh step 11 of the method is determining the distribution percentiles of the result of the eighth step 8.

The twelfth step 12 of the method is returning to the second step 2 if additional digital voice recordings are to be processed. Otherwise, proceeding to the next step.

The thirteenth step 13 of the method is comparing the results of the fifth step 5, the tenth step 10, and eleventh step 11 for each digital voice recording processed.

The fourteenth step 14 of the method is declaring the digital voice recordings duplicates that compared to within a user-definable threshold for each of the results of the fifth step 5, the tenth step 10, and the eleventh step 11.

* * * * *

United States Patent **7,127,392**
Smith **October 24, 2006**

Device for and method of detecting voice activity

Abstract

The present invention is a device for and method of detecting voice activity. First, the AM envelope of a segment of a signal of interest is determined. Next, the number of times the AM envelope crosses a user-definable threshold is determined. If there are no crossings, the segment is identified as non-speech. next, the number of points on the AM envelope within a user-definable range is determined. If there are less than a user-definable number of points within the range, the segment is identified as non-speech. Next, the mean, variance, and power ratio of the normalized spectral content of the AM envelope is found and compared to the same for known speech and non-speech. The segment is identified as being of the same type as the known speech or non-speech to which it most closely compares. These steps are repreated for each signal segment of interest.

Inventors: **Smith; David C.** (Columbia, MD)
Assignee: **The United States of America as represented by the National Security Agency** (Washington, DC)
N/A (N/A)
Family ID: **37110665**

Appl. No.: 10/370,309
Filed: February 12, 2003

Current U.S. Class: **704/233**; 704/208; 704/214; 704/215; 704/E11.003
Current CPC Class: G10L 25/78 (20130101)
Current International Class: G10L 15/20 (20060101)
Field of Search: ;704/214,204,210,213,215,216,219,217,500,233

References Cited [Referenced By]

U.S. Patent Documents

5619565	April 1997	Cesaro et al.
6023674	February 2000	Mekuria
6182035	January 2001	Mekuria
6249757	June 2001	Cason
2002/0103636	August 2002	Tucker et al.
2002/0147580	October 2002	Mekuria et al.

Other References

D Smith et al., "A Multivariate Speech Activity Detector Based on the Syllable Rate", Proceedings of SPIE, vol. 3461, pp. 68-78, 1998. cited by other .
D. Smith et al., "A Multivariate Speech Activity Detector Based on the Syllable Rate", Proceedings of ICASSP, vol. 1, pp. 73-76, 1999. cited by other.

Primary Examiner: Dorvil; Richemond
Assistant Examiner: Vo; Huyen X.
Attorney, Agent or Firm: Morelli; Robert D.

Claims

What is claimed is:

1. A voice activity detector, comprising: a) an absolute value squarer, having an input for receiving a signal, and having an output; b) a low-pass filter, having an input connected to the output of said absolute value squarer, and having an output; c) a first function block for finding a mean value, having an input connected to the output of the low pass-filter, and having an output; d) a second function block for finding a maximum value, having an input connected to the output of the low-pass filter, and having an output; e) a threshold-crossing detector, including a first user-definable threshold, having an input connected to the output of the low pass filter, and having an output; f) a third function block for finding a number of points between a user-definable range, having a first input connected to the output of the low-pass filter, having a second input connected to the output of the first function block, having a third input connected to the output of the second function block, and having an output; g) a comparator, having an input connected to the output of the third function block, and including a second user-definable threshold to which to compare; h) a subtractor, having a first input connected to the output of the low pass filter, having a second input connected to the output of the second function block, and having an output; i) a padder, having an input connected to the output of the subtractor, and having an output; j) a Digital Fast Fourier Transformer, having an input connected to the output of the padder, and having an output; k) a normalizer, having an input connected to the output of the Digital Fast Fourier Transformer, and having an output; l) a classifier, having an input connected to the output of the normalizer, and having an output; and m) a decision-logic block, having a first input connected to the output of the threshold-crossing detector, having a second input connected to the output of the comparator, having a third input connected to the output of the classifier, and having an output.

2. The voice activity detector of claim 1, wherein the threshold-crossing detector includes a first user-definable threshold that is 0.25 times the mean value of the output of the low-pass filter.

3. The voice activity detector of claim 1, wherein the third function block

includes a user-definable range from 0.25 times the mean value of the output of the low-pass filter to the maximum value of the low-pass filter minus 0.25 times the mean value of the low-pass filter.

4. The voice activity detector of claim 1, wherein the comparator includes 10 as the second user-definable threshold.

5. A method of detecting voice activity detector, comprising the steps of: a) receiving a signal; b) extracting a segment from the signal; c) computing an absolute value of the signal segment; d) squaring the result of the last step; e) finding an Amplitude Modulation (AM) envelope of the result of the last step; f) computing the mean of the last step; g) finding a first number of times the AM envelope crosses a first user-definable threshold; h) if the result of the last step is zero, identifying the signal segment as non-speech and returning to step (b) if there are more signal segments to process, otherwise stopping; i) finding the maximum value of the AM envelope; j) finding a second number points on the AM envelope that are within a user-definable range; k) if the result of the last step is less than a second user-definable threshold then identifying the signal segment as non-speech and returning to step (b) if there are more signal segments to process, otherwise stopping; l) subtracting the mean value of the AM envelope from the AM envelope; m) if the result of the last step is not a power of two then padding the result of the last step to form the next highest power of two; n) finding the spectral content of the AM envelope; o) finding a normalized vector of the result of the last step; p) computing a mean, variance, and power ratio of the result of the last step; and q) comparing the results of the last step to means, variances, and power ratios of known speech and non-speech, identifying the signal segment as a type to which they most closely compare, and returning to step (b) is there are more signal segments to process.

6. The method of claim 5, wherein the step of extracting a signal segment is comprised of the step of extracting a 0.5 second segment from the signal, where the signal segment overlaps a most resent previous signal segment by 0.4 seconds.

7. The method of claim 6, further including the steps of: a) retaining a number of consecutive 0.5 second frames; and b) using the number of

consecutive 0.5 second frames as votes to determine whether the 0.1 second interval common to the number of consecutive 0.5 second frames is speech or non-speech.

8. The method of claim 7, wherein said step of retaining a number of consecutive 0.5 second frames is comprised of the step of retaining five consecutive 0.5 second frames.

9. The method of claim 5, wherein said step of finding a first number of times the AM envelope crosses a first user-definable threshold is comprised of finding a first number of times the AM envelope crosses 0.25 times the mean of the AM envelope.

10. The method of claim 5, wherein the step of finding a second number points on the AM envelope that are within a user-definable range is comprised of the step of finding a second number points on the AM envelope that are within 0.25 times the mean value and the maximum value minus 0.25 times the mean value.

11. The method of claim 5, wherein the step of identifying the signal segment as non-speech if the result of the last step is less than a second user-definable threshold is comprised of identifying the signal segment as non-speech if the result of the last step is less than 10.

12. The method of claim 5, wherein the step of padding the result of the last step to form the next highest power of two is comprised of the step of padding the result of the last step with zeros to form the next highest power of two.

13. The method of claim 5, wherein the step of finding the spectral content of the AM envelope is comprised of the step of performing a Digital Fast Fourier Transform.

14. The method of claim 5, wherein the step of comparing the results of the last step to means, variances, and power ratios of known speech and non-speech is comprised of the step of performing a Quadratic Discriminant Analysis.

Description

FIELD OF THE INVENTION

The present invention relates, in general, to data processing and, in particular, to speech signal processing for identifying voice activity.

BACKGROUND OF THE INVENTION

A voice activity detector is useful for discriminating between speech and non-speech (e.g., fax, modem, music, static, dial tones). Such discrimination is useful for detecting speech in a noisy environment, compressing a signal by discarding non-speech, controlling communication devices that only allow one person at a time to speak (i.e., half-duplex mode), and so on.

A voice activity detector may be optimized for accuracy, speed, or some compromise between the two. Accuracy often means maximizing the rate at which speech is identified as speech and minimizing the rate at which non-speech is identified as speech. Speed is how much time it takes a voice activity detector to determine if a signal is speech or non-speech. Accuracy and speed work against each other. The most accurate voice activity detectors are often the slowest because they analyze a large number of features of the signal using computationally complex methods. The fastest voice activity detectors are often the least accurate because they analyze a small number of features of the signal using computationally simple methods. The primary goal of the present invention is accuracy.

Many prior art voice activity detectors only do a good job of distinguishing speech from one type of non-speech using one type of discriminator and do not do as well if a different type of non-speech is present. For example, the variance of the delta spectrum magnitude is an excellent discriminator of speech vs. music but it not a very good discriminator of speech vs. modem signals or speech vs. tones. Blind combination of specific discriminators does not lead to a general solution of speech vs.

non-speech. A dimension reduction technique such as principal components reduction may be used when a large number of discriminators are analyzed in an attempt to compress the data according to signal variance. Unfortunately, maximizing variance may not provide good discrimination.

Over the past few years, several voice activity detectors have been in use. The first of these is a simple energy detection method, which detects increases in signal energy in voice grade channels. When the energy exceeds a threshold, a signal is declared to be present. By requiring that the variance of the energy distribution also exceed a threshold, the method may be used to distinguish speech from several types of non-speech.

In two articles, both entitled "A multivariate speech activity detector based on the syllable rate," Proceeding of SPIE, Vol. 3461, pp. 68 78, 1998, and Proceeding of ICASSP, Vol. 1, pp. 73 76, 1999, Dr. David Smith et al. disclose a method of detecting voice by squaring the absolute value of a signal segment, finding the AM envelope of the signal segment, determining whether or not the AM envelope crosses a user-definable threshold, subtracting a mean of the AM envelope from the AM envelope, padding the result with zeros to make the result a power of two if necessary, finding the spectral components of the AM envelope, finding a normalized vector of the spectral components, and comparing the result to empirical models of speech and non-speech. The present invention is an improvement upon the method disclosed in these articles.

U.S. Pat. No. 5,619,565, entitled "VOICE ACTIVITY DETECTION METHOD AND APPARATUS USING THE SAME," discloses a device for and method of detecting voice, a single tone, and a dual tone by squaring a maximum value of a received signal, dividing the result by a measure of energy and comparing the ration to three threshold that represent voice, a single tone, and a dual tone, respectively. The present invention does not employ either the device or the method of U.S. Pat. No. 5,619,565. U.S. Pat. No. 5,619,565 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,023,674, entitled "NON-PARAMETRIC VOICE ACTIVITY DETECTION," discloses a device for and method of detecting voice

activity by extracting pitch period and signal energy information from an audio signal. The present invention does not employ either the device or the method of U.S. Pat. No. 6,023,674. U.S. Pat. No. 6,023,674 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,182,035, entitled "METHOD AND APPARATUS FOR DETECTING VOICE ACTIVITY," discloses a device for and method of detecting voice activity using wavelet transformation. The present invention does not use wavelet transformation to detect voice activity. U.S. Pat. No. 6,182,035 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 6,249,757, entitled "SYSTEM FOR DETECTING VOICE ACTIVITY," discloses a device for and method of detecting voice activity using two nonlinear filters, where one of the filter has a low time constant, and where the other filter has a high time constant. The present invention does not use two filters with differing time constants to detect voice activity. U.S. Pat. No. 6,249,757 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 2002/0103636, entitled "FREQUENCY-DOMAIN POST-FILTERING VOICE-ACTIVITY DETECTOR," discloses a device for and method of detecting voice activity by taking a currently received set of audio samples and a previously received set of audio samples in the time domain, converts the time-domain samples to the frequency domain, weights the energies of frequency ranges of the remaining frequencies proportionately to their frequencies, computes the total power of the ranges, and compares the power peaks to a threshold. The present invention does not weight the energies of frequency ranges to detect voice activity. U.S. Pat. Appl. No. 2002/0103636 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Appl. No. 2002/0147580, entitled "REDUCED COMPLEXITY VOICE ACTIVITY DETECTOR," discloses a device for and method of detecting voice activity by processing an audio signal to produce a train of signal samples, identifying signal peaks, computing values for quasi-pitch periods associated with the signal sample train, and selectively comparing the quasi-pitch periods with one another to determine the presence or

absence of a speech component. The present invention does not produce and compare quasi-pitch periods to detect voice activity. U.S. Pat. Appl. No. 2002/0147580 is hereby incorporated by reference into the specification of the present invention.

SUMMARY OF THE INVENTION

It is an object of the present invention to detect voice activity in a signal.

It is another object of the present invention to detect voice activity by in a manner than includes determining if the number of points on an AM envelope of a signal segment is within a user-definable range based on a mean value and maximum value of the AM envelope are above a user-definable threshold.

The present invention is a device for and method of detecting voice activity.

The device of the present invention implements the following method.

The first step of the method is receiving a signal.

The second step of the method is extracting a user-definable segment from the signal.

The third step of the method is finding the absolute value of the signal segment.

The fourth step of the method is squaring the absolute value.

The fifth step of the method is finding the Amplitude Modulation (AM) envelope of the signal segment.

The sixth step of the method is finding the mean value of the AM envelope.

The seventh step of the method is finding the number of times the AM envelope crosses a first user-definable threshold.

If the AM envelope doesn't cross the first user-definable threshold then the eighth step of the method is declaring the signal segment to be non-speech, returning to the second step if additional segments of the signal are to be processed, and stopping if there are no other signal segments to be processed. Otherwise, proceeding to the next step.

The ninth step of the method is finding the maximum value of the AM envelope.

The tenth step of the method is finding the number of points on the AM envelope within a user-definable range based on the mean and the maximum values of the AM envelope.

If N is less than a second user-definable threshold then the eleventh step of the method is declaring the signal segment to be non-speech, returning to the second step if there are additional signal segments to be processed, and stopping if there are no other signal segments to be processed. Otherwise, proceeding to the next step.

The twelfth step of the method is subtracting the mean value of the AM envelope from the AM envelope.

If the result of the last step is not a power of two then the thirteenth step of the method is padding the result of the last step so that it is a power of two. Otherwise, proceeding to the next step.

The fourteenth step of the method is finding the spectral content of the AM envelope.

The fifteenth step of the method is computing a normalized vector of the magnitude of the spectral content of the AM envelope.

The sixteenth step of the method is computing a mean, a variance, and a power ratio of the normalized vector.

The seventeenth, and last, step of the method is comparing the result of the last step to empirically-determined models of mean, variance, and power ratio of known speech and non-speech segments and declaring the

signal segment to be of the type of the empirically-determined model to which it most closely compares.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic of the present invention; and

FIG. 2 is a list of steps of the present invention.

DETAILED DESCRIPTION

The present invention is a device for and method of detecting voice activity. It is an improvement over the device and method disclosed in the two papers of Smith et al. disclosed above.

FIG. 1 is a schematic of the best mode and preferred embodiment of the present invention. The voice activity detector 1 receives a segment of a signal, computes feature vectors from the segment, and determines whether or not the segment is speech or non-speech. In the preferred embodiment, the segment is 0.5 seconds of a signal. In the preferred embodiment, the next segment analyzed is a 0.1 second increment of the previous segment. That is, the next segment includes the last 0.4 seconds of the first segment with an additional 0.1 seconds of the signal. Other segment sizes and increment schemes are possible and are intended to be included in the present invention. However, a segment length of 0.5 seconds was empirically determined to give the best balance between result accuracy and time window needed to resolve the syllable rate of speech.

The voice activity detector 1 receives the segment at an absolute value squarer 2. The absolute value squarer 2 finds the absolute value of the segment and then squares it. An arithmetic logic unit, a digital signal processor, or a microprocessor may be used to realize the function of the absolute value squarer 2.

The absolute value squarer 2 is connected to a low pass filter (LPF) 3. The low pass filter 3 blocks high frequency components of the output of the absolute value squarer 2 and passes low frequency components of the

output of the absolute value squarer 2. For speech purposes, low frequency is considered to be less than or equal to 60 Hz since the syllable rate of speech is within this range and, more particularly, within the range of 0 Hz to 10 Hz. The low pass filter 3 removes unnecessary high frequency components and simplifies subsequent computations. In the preferred embodiment, the low pass filter 3 is realized using a Hanning window. The output of the low pass filter 3 is often referred to as an Amplitude Modulated (AM) envelope of the original signal. This is because the high frequency, or rapidly oscillating, components have been removed, leaving only an AM envelope of the original segment.

The low pass filter 3 is connected to a first function block 4 for determining the maximum value of the AM envelope (MAX), a second function block 5 for determining the mean value of the AM envelope (MEAN), and a threshold-crossing detector 6. An arithmetic logic unit, a digital signal processor, or a microprocessor may be used to realize either of the first and second function blocks 4,5.

The output of second function block 5 is connected to the threshold-crossing detector 6. The threshold-crossing detector 6 counts the number of times the AM envelope dips below a first user-definable threshold. In the preferred embodiment, the first user-definable threshold is 0.25 times the mean of the AM envelope. If the segment presented to the threshold-crossing detector 6 does not cross the first user-definable threshold then the segment is identified as non-speech. However, just because the segment crosses the first user-definable threshold does not mean that the segment is speech. Therefore, processing of the segment continues if it crosses the first user-definable threshold. The threshold-crossing detector 6 has an output for indicating whether or the segment is non-speech. If the segment is non-speech then the output of the threshold-crossing detector 6 is a logic zero. Otherwise, the output of the threshold-crossing detector 6 is a logic one. A logic one output does not necessarily indicate that the segment is speech. Additional processing is required to make such a determination.

The outputs of the low-pass filter 3, the first function block 4, and the second function block 5 are connected to a third function block 7 for determining the number of points N on the AM envelope that lie within a

user-definable range. In the preferred embodiment, the user-definable range is from 0.25 times the mean of the AM envelope to MAX minus 0.25 times the mean of the AM envelope. An arithmetic logic unit, a digital signal processor, or a microprocessor may be used to realize the third function block 7.

The output of the third function block 7 is connected to a comparator 8 for determining whether or not N is greater than or equal to a second user-definable threshold. In the preferred embodiment the second user-definable threshold is 10. The comparator 8 has an output for indicating whether the segment is non-speech. If the number of points on the AM envelope within the user-definable range is less than the second user-definable threshold then the output of the comparators indicates that the signal segment is non-speech (e.g., a logic zero). Otherwise, the output of the comparator 8 is a logic one. A logic one output does not necessarily indicate that the segment is speech. Additional processing is required to make such a determination.

The first function block 4, the second function block 5, the third function block 7, and the comparator 8 represents the improvement over the device and method described by Smith et al. in the two articles described above. The improvement results in a speech activity detector that is more accurate than the one disclose by Smith et al. above.

The outputs of the low-pass filter 3 and the second function block 5 are connected to a subtractor 9. The subtractor 9 receives the AM envelope of the segment and the mean of the AM envelope and subtracts the mean of the AM envelope from the AM envelope. Mean subtraction improves the ability of the voice activity detector 1 to discriminate between speech and certain modem signals and tones. The subtractor 9 may be realized by an arithmetic logic unit, a digital signal processor, or a microprocessor.

The subtractor 9 is connected to a padder 10. If the output of the subtractor 9 is not a power of two, the padder 10 pads the output of the subtractor 9 with zeros so that the result is a power of two. In the preferred embodiment, eight bit values are used as a compromise between accuracy of resolving frequencies and the desire to minimize computation complexity. The padder 10 may be realized with a storage register and a counter.

The padder 10 is connected to a Digital Fast Fourier Transformer (DFFT) 11. The DFFT 11 performs a Digital Fast Fourier Transform on the output of the padder 10 to obtain the spectral, or frequency, content of the AM envelope. It is expected that there will be a peak in the magnitude of the speech signal spectral components in the 0 10 Hz range, while the magnitude of the non-speech signal spectral components in the same range will be small. The present invention establishes a spectral difference between speech signal and non-speech signal spectral components in the syllable rate range.

The DFFT 11 is connected to a normalizer 12. The normalizer 12 computes the normalized vector of the magnitude of the DFFT of the AM envelope, computes the mean of the normalized vector, computes the variance of the normalized vector, and computes the power ratio of the normalized vector. A normalized vector of a magnitude spectrum consists of the magnitude spectrum divided by the sum of all of the components of the magnitude spectrum. The normalized vector is a vector whose components are non-negative and sum to one. Therefore, the normalized vector may be viewed as a probability density. The power ratio of the normalized vector is found by first determining the average of the components in the normalized vector and then dividing the largest component in the normalized vector by this average. The result of the division is the power ratio of the normalized vector. The mean, variance, and power ratio of the normalized vector constitutes the feature vector of the segment received by the voice activity detector 1. The normalizer 12 may be realized by an arithmetic logic unit, a microprocessor, or a digital signal processor.

The normalizer 12 is connected to a classifier 13. The classifier 13 receives the mean, variance, and power ratio of the segment computed by the normalizer 12 and compares it to precomputed models which represent the mean, variance, and power ratio of known speech and non-speech segments. The classifier 13 declares the feature vector of the segment to be of the type (i.e., speech or non-speech) of the precomputed model to which it matches most closely. Various classification methods are known by those skilled in the art. In the preferred embodiment, the classifier 13 performs the classification method of Quadratic Discriminant

Analysis. The classifier 13 may determine whether the received segment is speech or non-speech based on the segment received or the classifier 13 may retain a number of, preferably five, consecutive 0.5 second segments and use them as votes to determine whether the 0.1 second interval common to these segments is speech or non-speech. Voting permits a decision every 0.1 seconds after the first number of frames are processed and improves decision accuracy. Therefore, voting is used in the preferred embodiment. The classifier 13 may be realized with an arithmetic logic unit, a microprocessor, or a digital signal processor.

The outputs of the classifier 13, the threshold-crossing detector 6, and the comparator 8 are connected to decision logic block 14 for determining whether the segment is speech or non-speech. In the preferred embodiment, the decision logic block 14 is an AND gate. That is, the threshold-detector 6, the comparator 8, and the classifier 13 each put out a logic one value to indicate speech and a logic zero value to indicate non-speech. So, a logic one value from each of the threshold-crossing detector 6, the comparator 8, and the classifier 13 is required to indicate that the segment is speech. However, a logic zero value from either the threshold-crossing detector 6, the comparator 8, or the classifier 13 would indicate that the segment is non-speech.

FIG. 2 is a list of steps of the method of the present invention.

The first step 21 of the method is receiving a signal.

The second step 22 of the method is extracting a user-definable segment from the signal. In the preferred embodiment, the segment is 0.5 seconds in length. A subsequent segment overlaps the most recent previous segment. In the preferred embodiment, a subsequent segment overlaps the most recent previous segment by 0.4 seconds so that the new part of the segment is only 0.1 seconds in length. In an alternate embodiment, the signal segments processed are retained as consecutive frames The frames (e.g., 5 frames) are then used as votes to determine whether the 0.1 second interval common to the number of consecutive 0.5 second frames is speech or non-speech.

The third step 23 of the method is finding the absolute value of the signal

segment.

The fourth step 24 of the method is squaring the absolute value.

The fifth step 25 of the method is finding the Amplitude Modulation (AM) envelope of the signal segment. In the preferred embodiment, the AM envelope is found by low-pass filtering the segment.

The sixth step 26 of the method is finding the mean value of the/AM envelope.

The seventh step 27 of the method is finding the number of times the AM envelope crosses a first user-definable threshold. In the preferred embodiment, the first user-definable threshold is 0.25 times the mean of the AM envelope.

If the AM envelope doesn't cross the first user-definable threshold then the eighth step 28 of the method is declaring the signal segment to be non-speech, returning to the second step 22 if additional segments of the signal are to be processed, and stopping if there are no other signal segments to be processed. Otherwise, proceeding to the next step.

The ninth step 29 of the method is finding the maximum value (MAX) of the AM envelope.

The tenth step 30 of the method is finding the number of points N on the AM envelope within a user-definable range based on the mean and maximum values of the AM envelope. In the preferred embodiment, the user-definable range is from 0.25 times the mean value to MAX minus 0.25 times the mean value.

If N is less than a second user-definable threshold then the eleventh step 31 of the method is declaring the signal segment to be non-speech, returning to the second step 22 if there are additional signal segments to be processed, and stopping if there are no other signal segments to be processed. Otherwise, proceeding to the next step. In the preferred embodiment, the second user-definable threshold is 10.

The twelfth step 32 of the method is subtracting the mean value of the AM envelope from the AM envelope.

If the result of the last step is not a power of two then the thirteenth step 33 of the method is padding the result of the last step so that it is a power of two. Otherwise, proceeding to the next step. In the preferred embodiment, the result of the last step is padded with zeros if necessary.

The fourteenth step 34 of the method is finding the spectral content of the AM envelope. In the preferred embodiment, spectral content is found by performing a Digital Fast Fourier Transform (DFFT).

The fifteenth step 35 of the method is computing a normalized vector of the magnitude of the spectral content of the AM envelope.

The sixteenth step 36 of the method is computing a mean, a variance, and a power ratio of the normalized vector.

The seventeenth, and last, step 37 of the method is comparing the result of the last step to empirically-determined models of mean, variance, and power ratio of known speech and non-speech segments and declaring the signal segment to be of the type of the empirically-determined model to which it most closely compares. In the preferred embodiment, the seventeenth step 37 of the method is conducted by performing a Quadratic Discriminant Analysis

* * * * *

United States Patent **6,556,967**
Nelson , et al. **April 29, 2003**

Voice activity detector

Abstract

The present invention is a device for and method of detecting voice activity by receiving a signal; computing the absolute value of the signal; squaring the absolute value; low pass filtering the squared result; computing the mean of the filtered signal; subtracting the mean from the filtered result; padding the mean subtracted result with zeros to form a value that is a power of two if the result is not already a power of two; computing a DFFT of the power of two result; normalizing the DFFT result of the last step; computing a mean of the normalization; computing a variance of the normalization; computing a power ratio of the normalization; classifying the mean, variance and power ratio as speech or non-speech based on how this feature vector compares to similarly constructed feature vectors of known speech and non-speech. The voice activity detector includes an absolute value squarer; a low pass filter; a mean subtractor; a zero padder; a DFFT; a normalizer; and a classifier.

Inventors: **Nelson; Douglas J.** (Columbia, MD)**, Smith; David C.** (Columbia, MD)**, Townsend; Jeffrey L.** (Columbia, MD)

Assignee: **The United States of America as represented by the National Security Agency** (Washington, DC)

Family ID: 23016092
Appl. No.: 09/266,811
Filed: March 12, 1999

Current U.S. Class:	**704/233**; 704/208; 704/214; 704/226; 704/227; 704/228; 704/E11.003
Current CPC Class:	G10L 25/78 (20130101)
Current International Class:	G10L 11/00 (20060101); G10L 11/02 (20060101); G10L 015/20 ()
Field of Search:	;704/233,226,227,228,214,208

References Cited [Referenced By]

U.S. Patent Documents

4351983	September 1982	Crouse et al.
4672669	June 1987	DesBlache et al.
5012519	April 1991	Adlersberg et al.
5255340	October 1993	Arnaud et al.
5276765	January 1994	Freeman et al.
5323337	June 1994	Wilson et al.
5459814	October 1995	Gupta et al.
5533118	July 1996	Cesaro et al.
5586180	December 1996	Degenhardt et al.
5598466	January 1997	Graumann
5611019	March 1997	Nakatoh et al.
5619565	April 1997	Cesaro et al.
5619566	April 1997	Fogel
5649055	July 1997	Gupta et al.
5657422	August 1997	Janiszewski et al.
5706394	January 1998	Wynn
5732141	March 1998	Chaoui et al.
5735716	April 1998	Bergstrom et al.

5737407	April 1998	Graumann
5749067	May 1998	Barrett
5809459	September 1998	Bergstrom et al.
5826230	October 1998	Reaves
5867574	February 1999	Erylimaz
5907824	May 1999	Tzirkel-Hancock
5963901	October 1999	Vahatalo et al.
5991718	November 1999	Malah
6061647	May 2000	Barrett
6182035	January 2001	Mekuria

Primary Examiner: Chawan; Vijay B
Attorney, Agent or Firm: Morelli; Robert D.

Claims

What is claimed is:

1. A voice activity detector, comprising: a) an absolute value squarer, having an input for receiving a signal, and having an output; b) a low pass filter, having an input connected to the output of said absolute value squarer, and having an output; c) a mean subtractor, having an input connected to the output of said low pass filter, and having an output; d) a zero padder, having an input connected to the output of said mean subtractor, and having an output; e) a Digital Fast Fourier Transformer, having an input connected to the output of said zero padder, and having an output; f) a normalizer, having an input connected to the output of said Digital fast Fourier Transformer, and having an output; and g) a classifier, having an input connected to the output of said normalizer, and having an output.

2. A voice activity detector, comprising: a) an absolute value squarer, having an input for receiving a signal, and having an output; b) a low pass filter, having an input connected to the output of said absolute value squarer, and having an output; c) a threshold-crossing detector, having a

user-definable threshold, having an input connected to the output of said low pass filter, having a first output, and having a second output; d) a mean subtractor, having an input connected to the first output of said zero crossing detector, and having an output; e) a zero padder, having an input connected to the output of said mean subtractor, and having an output; f) a Digital Fast Fourier Transformer, having an input connected to the output of said zero padder, and having an output; g) a normalizer, having an input connected to the output of said Digital Fast Fourier Transformer, and having an output; h) a classifier, having an input connected to the output of said normalizer, and having an output; and i) decision logic, having a first input connected to the second output of said zero crossing detector, having a second input connected to the output of said classifier, and having an output.

3. A method of detecting voice activity, comprising the steps of: a) receiving a signal; b) computing the absolute value of the signal; c) squaring the result of the last step; d) filtering the result of the last step to only pass low frequency components in the range of from 0-60 Hz; e) computing the mean of the last step; f) subtracting the mean computed in the last step from the result of step (d); g) padding the result of the last step with zeros to form the next highest power of two of the result of the last step if the result of the last step is not already a power of two; e) computing a Digital Fast Fourier Transform of the result of the last step; f) normalizing the result of the last step; g) computing a mean of the result of the last step; h) computing a variance of the result of step (f); i) computing a power ratio of the result of step (f); j) classifying the results of step (g), step (h), and step (i) as a type of known speech and known non-speech to which the results of step (g), step (h), and step (i) most closely compares, where the known speech and the known non-speech are each identified by a mean, a variance and a power ratio.

4. The method of claim 3, wherein said step of receiving a signal is comprised of the step of receiving a 0.5 second segment of a signal, where said segment was incremented by 0.1 seconds from a next previous segment.

5. The method of claim 4, further including the steps of: a) retaining a number of consecutive 0.5 second frames; and b) using the number of

consecutive 0.5 second frames as votes to determine whether the 0.1 second interval common to the number of consecutive 0.5 second frames is speech or non-speech.

6. The method of claim 5, wherein said step of retaining a number of consecutive 0.5 second frames is comprised of the step of retaining five consecutive 0.5 second frames.

7. The method of claim 6, wherein said step of classifying the results of step (g), step (h), and step (i) is comprised of performing a Quadratic Discriminant Analysis.

8. The method of claim 7, further including counting the number of times the result of filtering crosses a user-definable threshold.

9. The method of claim 8, wherein said step of counting the number of zero threshold crossings is comprised of the step of counting the number of times the result of filtering crosses a user-definable threshold, where the threshold is defined as 0.25 times the mean of an AM envelope of the signal.

10. The method of claim 3, wherein said step of classifying the results of step (g), step (h), and step (i) is comprised of performing a Quadratic Discriminant Analysis.

11. The method of claim 3, further including counting the number of times the result of filtering crosses a user-definable threshold.

12. The method of claim 11, wherein said step of counting the number of zero threshold crossings is comprised of the step of counting the number of times the result of filtering crosses a user-definable threshold, where the threshold is defined as 0.25 times the mean of an AM envelope of the signal.

Description

FIELD OF THE INVENTION

The present invention relates, in general, to data processing and, in particular, to speech signal processing for identifying voice activity.

BACKGROUND OF THE INVENTION

A voice activity detector is useful for discriminating between speech and non-speech (e.g., fax, modem, music, static, dial tones). Such discrimination is useful for detecting speech in a noisy environment, compressing a signal by discarding non-speech, controlling communication devices that only allow one person at a time to speak (i.e., half-duplex mode), and so on.

A voice activity detector may be optimized for accuracy, speed, or some compromise between the two. Accuracy often means maximizing the rate at which speech is identified as speech and minimizing the rate at which non-speech is identified as speech. Speed is how much time it takes a voice activity detector to determine if a signal is speech or non-speech. Accuracy and speed work against each other. The most accurate voice activity detectors are often the slowest because they analyze a large number of features of the signal using computationally complex methods. The fastest voice activity detectors are often the least accurate because they analyze a small number of features of the signal using computationally simple methods. The primary goal of the present invention is accuracy.

Many prior art voice activity detectors only do a good job of distinguishing speech from one type of non-speech using one type of discriminator and do not do as well if a different type of non-speech is present. For example, the variance of the delta spectrum magnitude is an excellent discriminator of speech vs. music but it not a very good discriminator of speech vs. modem signals or speech vs. tones. Blind combination of specific discriminators does not lead to a general solution of speech vs. non-speech. A dimension reduction technique such as principal components reduction may be used when a large number of discriminators are analyzed in an attempt to compress the data according to signal variance. Unfortunately, maximizing variance may not provide good

discrimination.

Over the past few years, several voice activity detectors have been in use. The first of these is a simple energy detection method, which detects increases in signal energy in voice grade channels. When the energy exceeds a threshold, a signal is declared to be present. By requiring that the variance of the energy distribution also exceed a threshold, the method may be used to distinguish speech from several types of non-speech.

FIG. 1 is an illustration of a voice activity detection method called the readability method 1. It is a variation of the energy method. A signal is filtered 2 by a pre-whitening filter. An autocorrelation 3 is performed on the pre-whitened signal. The peak in the autocorrelated signal is then detected 4. The peak is then determined to be within the expected pitch range 5 (i.e., speech) or not 6 (i.e., non-speech). Speech is declared to be present if a bulge occurs in the correlation function within the expected periodicity range for the pitch excitation function of speech. The readability method is similar to the energy method since detection is based on energy exceeding a threshold. The readability method 1 performs better that the energy method because the readability method 1 exploits the periodicity of speech. However, the readability method does not perform well if there are changes in the gain, or dynamic range, of the signal. Also, the readability method identifies non-speech as speech when non-speech exhibits periodicity in the expected pitch range (i.e., 75 to 400 Hz.). The pre-whitening filter removes un-modulated tones (i.e., non-speech) to prevent such tones from being identified as speech. However, such a filter does not remove other non-speech signals (e.g., modulated tones and FM signals) which may be present in a channel carrying speech. Such non-speech signals and may be falsely identified as speech.

FIG. 2 is an illustration of the NP method 20 which detects voice activity by estimating the signal to noise ratio (SNR) for each frame of the signal. A Fast Fourier Transform (FFT) is performed on the signal and the absolute value of the result is squared 21. The result of the last step is then filtered to remove un-modulated tones using a pre-whitening filter 22. The variance in the result of the last step is then determined 23. The result of the last step is then limited to a band of frequencies in which speech may occur 24. The power spectrum of each frame is computed and sorted 25

into either high energy components or low energy components. High energy components are assumed to be signal (speech which may include non-speech) or interference (non-speech) while low energy components are assumed to be noise (all non-speech). The highest energy components are discarded. The signal power is then estimated from the remaining high energy components 26. The noise power is estimated by averaging the low-energy components 27. The signal power is then divided by the noise power 28 to produce the SNR. The SNR is then compared to a user-definable threshold to determine whether or not the frame of the signal is speech or non-speech. Signal detection in the NP method is based on a power ratio measurement and is, therefore, not sensitive to the gain of the receiver. The fundamental assumption in the NP method is that spectral components of speech are sparse.

FIG. 3 illustrates a voice activity detector method named TALKATIVE 30 which detects speech by estimating the correlation properties of cepstral vectors. The assumption is that non-stationarity (a good discriminator of speech) is reflected in cepstral coefficients. Vectors of cepstral coefficients are computed in a frame of the signal 31. Squared Euclidean distances between cepstral vectors are computed 32. The squared Euclidean distances are time averaged 33 within the frame in order to estimate the stationarity of the signal. A large time averaged value indicates speech while a small time averaged value indicates a stationary signal (i.e., non-speech). The time averaged value is compared to a user-definable threshold 34 to determine whether or not the signal is speech or non-speech. The TALKATIVE method performs well for most signals, but does not perform well for music or impulsive signals. Also, considerable temporal smoothing occurs in the TALKATIVE method.

U.S. Pat. No. 4,351,983, entitled "SPEECH DETECTOR WITH VARIABLE THRESHOLD," discloses a device for and method of detecting speech by adjusting the threshold for determining speech on a frame by frame basis. U.S. Pat. No. 4,351,983 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 4,672,669, entitled "VOICE ACTIVITY DETECTION PROCESS AND MEANS FOR IMPLEMENTING SAID PROCESS," discloses a device for and method of detecting voice activity by comparing

the energy of a signal to a threshold. The signal is determined to be voice if its power is above the threshold. If its power is below the threshold then the rate of change of the spectral parameters is tested. U.S. Pat. No. 4,672,669 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,255,340, entitled "METHOD FOR DETECTING VOICE PRESENCE ON A COMMUNICATION LINE," discloses a method of detecting voice activity by determining the stationary or non-stationary state of a block of the signal and comparing the result to the results of the last M blocks. U.S. Pat. No. 5,255,340 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,276,765, entitled "VOICE ACTIVITY DETECTION," discloses a device for and a method of detecting voice activity by performing an autocorrelation on weighted and combined coefficients of the input signal to provide a measure that depends on the power of the signal. The measure is then compared against a variable threshold to determine voice activity. U.S. Pat. No. 5,276,765 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Nos. 5,459,814 and 5,649,055, both entitled "VOICE ACTIVITY DETECTOR FOR SPEECH SIGNALS IN VARIABLE BACKGROUND NOISE," discloses a device for and method of detecting voice activity by measuring short term time domain characteristics of the input signal, including the average signal level and the absolute value of any change in average signal level. U.S. Pat. Nos. 5,459,814 and 5,649,055 are hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Nos. 5,533,118 and 5,619,565, both entitled "VOICE ACTIVITY DETECTION METHOD AND APPARATUS USING THE SAME," discloses a device for and method of detecting voice activity by dividing the square of the maximum value of the received signal by its energy and comparing this ratio to three different thresholds. U.S. Pat. Nos. 5,533,118 and 5,619,565 are hereby incorporated by reference into the specification of the present invention.

U.S. Pat. Nos. 5,598,466 and 5,737,407, both entitled "VOICE ACTIVITY

DETECTOR FOR HALF-DUPLEX AUDIO COMMUNICATION SYSTEM," discloses a device for and method of detecting voice activity by determining an average peak value, a standard deviation, updating a power density function, and detecting voice activity if the average peak value exceeds the power density function. U.S. Pat. Nos. 5,598,466 and 5,737,407 are hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,619,566, entitled "VOICE ACTIVITY DETECTOR FOR AN ECHO SUPPRESSOR AND AN ECHO SUPPRESSOR," discloses a device for detecting voice activity that includes a whitening filter, a means for measuring energy, and using the energy level to determine the presence of voice activity. U.S. Pat. No. 5,619,566 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,732,141, entitled "DETECTING VOICE ACTIVITY," discloses a device for and method of detecting voice activity by computing the autocorrelation coefficients of a signal, identifying a first autocorrelation vector, identifying a second autocorrelation vector, subtracting the first autocorrelation vector from the second autocorrelation vector, and computing a norm of the differentiation vector which indicates whether or not voice activity is present. U.S. Pat. No. 5,732,141 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,749,067, entitled "VOICE ACTIVITY DETECTOR," discloses a device for and method of detecting voice activity by comparing the spectrum of the a signal to a noise estimate, updating the noise estimate, computing a linear predictive coding prediction gain, and suppressing updating the noise estimate if the gain exceeds a threshold. U.S. Pat. No. 5,749,067 is hereby incorporated by reference into the specification of the present invention.

U.S. Pat. No. 5,867,574, entitled "VOICE ACTIVITY DETECTION SYSTEM AND METHOD," discloses a device for and method of detecting voice activity by computing an energy term based on an integral of the absolute value of a derivative of a speech signal, computing a ration of the energy to a noise level, and comparing the ratio to a voice activity threshold. U.S. Pat. No. 5,867,574 is hereby incorporated by reference into

the specification of the present invention.

SUMMARY OF THE INVENTION

It is an object of the present invention to detect voice activity in a signal.

It is another object of the present invention to detect voice activity in a signal by squaring the absolute value of a signal, finding the low frequency components of the signal known as an AM envelope, subtracting the mean of the AM envelope from the AM envelope, padding the result with zeros if the result is not a power of two, transform the result using a Discreet Fast Fourier Transform, normalizing the result, computing a feature vector, and determining the presence of voice activity using Quadratic Discriminant Analysis.

It is another object of the present invention to remove music signals by observing threshold crossings of the AM envelope of the signal.

The present invention is a device for and method of detecting voice activity. A segment of a signal is received at an absolute value squarer, which computes the absolute value of the segment and then squares it.

The absolute value squarer is connected to a low pass filter, which blocks high frequency components of the output of the absolute value squarer and passes low frequency components of the output of the absolute value squarer.

The low pass filter is connected to a mean subtractor, which receives the AM envelope of the segment, computes the mean of the AM envelop and subtracts the mean of the AM envelope from the AM envelope.

The mean subtractor is connected to a zero padder, which pads the result of the mean subtractor with zeros to form a value that is a power of two.

The zero padder is connected to a Digital Fast Fourier Transformer (DFFT), which performs a Digital Fast Fourier Transform on the output of the zero padder.

The DFFT is connected to a normalizer, which computes a normalized magnitude vector of the DFFT of the AM envelope, computes the mean of the normalized magnitude vector, computes the variance of the normalized magnitude vector, and computes the power ratio of the normalized magnitude vector.

The normalizer is connected to a classifier, which receives the mean, variance, and power ratio of the normalizer magnitude vector and compares these features to models of similar features precomputed for known speech and known non-speech to determine whether the unknown segment received is speech or non-speech.

Alternate embodiments of the present invention may be realized by adding a threshold-crossing detector between the low pass filter and the mean subtractor to identify music as non-speech.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is an illustration of the prior art readability method;

FIG. 2 is an illustration of the prior art NP method;

FIG. 3 is an illustration of the prior art TALKATIVE method;

FIG. 4 is a schematic of the present invention;

FIG. 5 is a graph comparing the present invention to TALKATIVE; and

FIG. 6 is a schematic of an alternate embodiment of the present invention.

DETAILED DESCRIPTION

The present invention is a device for and method of detecting voice activity. FIG. 4 is a schematic of the best mode and preferred embodiment of the present invention. The voice activity detector 40 receives a segment of a signal, computes feature vectors from the segment, and determines whether or not the segment is speech or non-speech. In the preferred embodiment, the segment is 0.5 seconds of a signal. In the preferred

embodiment, the next segment analyzed is a 0.1 second increment of the previous segment. That is, the next segment includes the last 0.4 seconds of the first segment with an additional 0.1. seconds of the signal. Other segment sizes and increment schemes are possible and are intended to be included in the present invention. However, a segment length of 0.5 seconds was empirically determined to give the best balance between result accuracy and time window needed to resolve the syllable rate of speech.

The voice activity detector 40 receives the segment at an absolute value squarer 41. The absolute value squarer 41 finds the absolute value of the segment and then squares it. An arithmetic logic unit, a digital signal processor, or a microprocessor may be used to realize the function of the absolute value squarer 41.

The absolute value squarer 41 is connected to a low pass filter 42. The low pass filter 42 blocks high frequency components of the output of the absolute value squarer 41 and passes low frequency components of the output of the absolute value squarer 41. For speech purposes, low frequency is considered to be less than or equal to 60 Hz since the syllable rate of speech is within this range and, more particularly, within the range of 0 Hz to 10 Hz. The low pass filter 42 removes unnecessary high frequency components and simplifies subsequent computations. In the preferred embodiment, the low pass filter 42 is realized using a Hanning window. The output of the low pass filter 42 is often referred to as an Amplitude Modulated (AM) envelope of the original signal. This is because the high frequency, or rapidly oscillating, components have been removed, leaving only an AM envelope of the original segment.

The low pass filter 42 is connected to a mean subtractor 43. The mean subtractor 43 receives the AM envelope of the segment, computes the mean of the AM envelope, and subtracts the mean of the AM envelope from the AM envelope. Mean subtraction improves the ability of the voice activity detector 40 to discriminate between speech and certain modem signals and tones. The mean subtractor 43 may be realized by an arithmetic logic unit, a digital signal processor, or a microprocessor.

The mean subtractor 43 is connected to a zero padder 44. The zero

padder 44 pads the output of the mean subtractor 43 with zeros out to a power of two if the output of the mean subtractor 43 is not a power of two. In the preferred embodiment, nine bit values are used as a compromise between accuracy of resolving frequencies and the desire to minimize computation complexity. The zero padder 44 may be realized with a storage register and a counter.

The zero padder 44 is connected to a Digital Fast Fourier Transformer (DFFF) 45. The DFFT 45 performs a Digital Fast Fourier Transform on the output of the zero padder 44 to obtain the spectral, or frequency, content of the AM envelop. It is expected that there will be a peak in the magnitude of the speech signal spectral components in the 0-10 Hz range, while the magnitude of the non-speech signal spectral components in the same range will be small. Establishing a spectral difference between speech signal and non-speech signal spectral components in the syllable rate range is a key goal of the present invention.

The DFFT 45 is connected to a normalizer 46. The normalizer 46 computes the normalized vector of the magnitude of the DFFT of the AM envelope, computes the mean of the normalized vector, computes the variance of the normalized vector, and computes the power ratio of the normalized vector. A normalized vector of a magnitude spectrum consists of the magnitude spectrum divided by the sum of all of the components of the magnitude spectrum. The normalized vector is a vector whose components are non-negative and sum to one. Therefore, the normalized vector may be viewed as a probability density. The normalized vector may be viewed as a probability density. The power ratio of the normalized vector is found by first determining the average of the components in the normalized vector and then dividing the largest component in the normalized vector by this average. The result of the division is the power ratio of the normalized vector. The mean, variance, and power ratio of the normalized vector constitutes the feature vector of the segment received by the voice activity detector 40. The normalizer 46 may be realized by an arithmetic logic unit, a microprocessor, or a digital signal processor.

The normalizer 46 is connected to a classifier 47. The classifier 47 receives the mean, variance, and power ratio of the segment computed by the normalizer 46 and compares it to precomputed models which

represent the mean, variance, and power ratio of known speech and non-speech segments. The classifier 47 declares the feature vector of the segment to be of the type (i.e., speech or non-speech) of the precomputed model to which it matches most closely. Various classification methods are know by those skilled in the art. In the preferred embodiment, the classifier 47 performs the classification method of Quadratic Discriminant Analysis. The classifier 47 may determine whether the received segment is speech or non-speech based on the segment received or the classifier 47 may retain a number of, preferably five, consecutive 0.5 second segments and use them as votes to determine whether the 0.1 second interval common to these segments is speech or non-speech. Voting permits a decision every 0.1 seconds after the first number of frames are processed and improves decision accuracy. Therefore, voting is used in the preferred embodiment. The classifier 47 may be realized with an arithmetic logic unit, a microprocessor, or a digital signal processor.

The performance of the voice activity detector 40 was compared against the TALKATIVE voice activity detector. FIG. 5 is a graph of the comparison which plots, on the y-axis, the rate at which voice activity was falsely detected versus the rate at which voice activity was correctly detected, on the x-axis. As can be seen from FIG. 5, the present invention significantly outperformed the TALKATIVE method.

FIG. 6 is a schematic of an alternate embodiment of the present invention. The voice activity detector 60 of FIG. 6 is better able to identify music and quickly identify it as non-speech. The voice activity detector 60 does this by using the same circuit as the voice activity detector 40 of FIG. 4 and inserting therein a threshold-crossing detector 63. Each function of FIG. 6 performs the same function as its like-named counterpart of FIG. 4 and will not be re-described here. So, the segment is received by an absolute value squarer 61. The absolute value squarer 61 is connected to a low pass filter 62.

The low pass filter 62 is connected to the threshold-crossing detector 63. The threshold-crossing detector 63 counts the number of times the AM envelope dips below a user-definable threshold. In the preferred embodiment, the threshold is 0.25 times the mean of the AM envelope. If the segment presented to the threshold-crossing detector 63 does not

cross the threshold then the segment is identified as non-speech and the segment need not be processed further. However, just because the segment crosses the threshold does not mean that the segment is speech. Therefore, processing of the segment continues if it crosses the threshold. The threshold-crossing detector 63 may have two outputs, one for indicating that the segment is non-speech and another for transmitting the segment received to a mean subtractor 64.

The output of the threshold-crossing detector 63 that transmits the segment received is connected to the mean subtractor 64. The mean subtractor 64 is connected to a zero padder 65. The zero padder 65 is connected to a DFFT 66. The DFFT 66 is connected to a normalizer 67. The normalizer 67 is connected to a classifier 68. The classifier 68 and the non-speech indicating output of the threshold-crossing detector 63 are connected to decision logic 69 for determining whether the segment is speech or non-speech. The decision logic 69 may be as simple as an AND gate. That is, the threshold-detector 63 and the classifier 68 may each use a logic value of 1 to indicate speech and a logic value of 0 to indicate non-speech. So, a logic value of 1 from both the threshold-crossing detector 63 and the classifier 68 is required to indicate that the segment is speech. However, logic levels of 0 from either the threshold-crossing detector 63 or the classifier 68 would indicate that the segment is non-speech. The same options that exist for the voice activity detector 40 of FIG. 4 are available to the voice activity detector 60 of FIG. 6.

* * * * *