

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
8 March 2001 (08.03.2001)

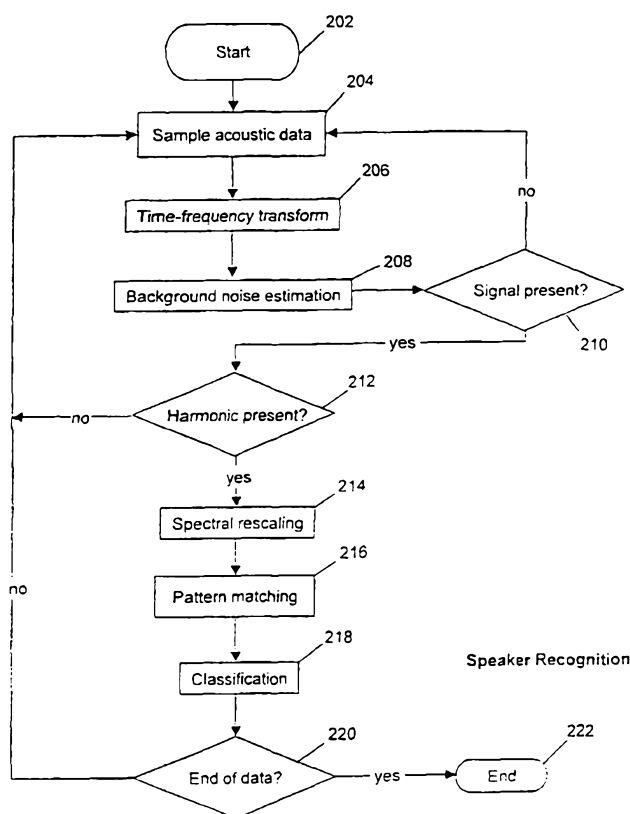
PCT

(10) International Publication Number  
**WO 01/16937 A1**

- (51) International Patent Classification<sup>7</sup>: **G10L 17/00**
- (21) International Application Number: **PCT/US00/23754**
- (22) International Filing Date: **29 August 2000 (29.08.2000)**
- (25) Filing Language: **English**
- (26) Publication Language: **English**
- (30) Priority Data:  
09/385,975      30 August 1999 (30.08.1999)      US
- (63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:  
US      09/385,975 (CON)  
Filed on      30 August 1999 (30.08.1999)
- (71) Applicant (for all designated States except US): **WAVE-MAKERS RESEARCH, INC.** [CA/CA]; 3328 West 2nd Avenue, Vancouver, British Columbia V6R 1J1 (CA).
- (72) Inventor; and  
(75) Inventor/Applicant (for US only): **ZAKARAUSKAS, Pierre** [CA/US]; 1723 Kennington Road, Encinitas, CA 92024 (US).
- (74) Agent: **LEE, Samuel, S.**; Fish & Richardson P.C., Suite 500, 4350 La Jolla Village Drive, San Diego, CA 92122 (US).
- (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.
- (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,

[Continued on next page]

(54) Title: **SYSTEM AND METHOD FOR CLASSIFICATION OF SOUND SOURCES**



(57) Abstract: A system and method to identify a sound source among a group of sound sources. The invention matches the acoustic input to a number of signal models, one per source class, and produces a goodness-of-match number for each signal model. The sound source is declared to be of the same class as that of the signal model with the best goodness-of-match if that score is sufficiently high. The data are recorded with a microphone, digitized and transformed into the frequency domain. A signal detector is applied to the transient. A harmonic detection method can be used to determine if the sound source has harmonic characteristics. If at least some part of a transient contains signal of interest, the spectrum of the signal after rescaling is compared to a set of signal models, and the input signal's parameters are fitted to the data. The average distortion is calculated to compare patterns with those of sources that used in training the signal models. Before classification can occur, a source model is trained with signal data. Each signal model is built by creating templates from input signal spectrograms when they are significantly different from existing templates. If an existing template is found that resembles the input pattern, the template is averaged with the pattern in such a way that the resulting template is the average of all the spectra that matched that template in the past.



IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

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

**Published:**

- *With international search report.*
- *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

## **SYSTEM AND METHOD FOR CLASSIFICATION OF SOUND SOURCES**

### **TECHNICAL FIELD**

This invention relates to systems and methods for automatic  
5 classification of acoustic (sound) sources, including text-independent speaker  
identification.

### **BACKGROUND**

There are several fields of research studying acoustic signal  
classification. Each field of research has adopted its own approaches to  
10 acoustic signal classification, with some overlap between them. At present, the  
main applications for automatic sound source classification are: speaker  
verification; speaker identification; passive sonar classification; and machine  
noise monitoring or diagnostics.

Speaker verification aims at verifying that a given speaker is indeed who  
15 he or she claims to be. In most speaker verification systems, a speaker  
cooperates in saying a keyword, and the system matches the way that keyword  
was said by the putative speaker with training samples of the same keywords. If  
the match is poor, the speaker is rejected or denied service (e.g., computer or  
premise access). A disadvantage of such methods is that the same keyword  
20 must be used at testing time as at training time, thus limiting application of  
such methods to access control. This method could not be used to label the  
speakers in a back-and-forth conversation for example.

Speaker identification aims at determining which among a set of voices  
best matches a given test utterance. Text-independent speaker identification  
25 tries to make such a determination without the use of particular keywords.

Passive sonar classification involves identifying a vessel according to  
the sound it radiates underwater. Machine noise monitoring and diagnostics

involves determining the state of a piece of machinery through the sound it makes.

In all of the above applications, a model of each sound source is first obtained by training a system with a set of example sounds from each source.

5 A test sample is then compared to the stored models to determine a sound source category for the test sample. Known methods require relatively long training times and testing samples that make such methods inappropriate in many cases. Further, such methods tend to require a large amount of memory storage and computational resources. Finally, these methods often are not  
10 robust to the presence of noise in the test signal, which prevents their use in many tasks. ("Signal" means a signal of interest; background and distracting sounds are referred to as "noise").

The inventor has determined that it would be desirable to be able to classify an acoustic signal even when some portions of the spectra are masked  
15 by noise, and require a minimum amount of training and testing. The present invention provides a system and method for acoustic signal classification that avoids the limitations of prior techniques.

### SUMMARY

The invention includes a method, apparatus, and computer program to  
20 classify a sound source. The invention matches the acoustic input to a number of signal models, one per source class, and produces a score for each signal model. The sound source is declared to be of the same class as that of the model with the best score if that score is sufficiently high. In the preferred embodiment, classification is accomplished by the use of a signal model  
25 augmented by learning. The input signal may represent human speech, in which case the goal would be to identify the speaker in a text-independent manner. However, it should be recognized that the invention may be used to classify any type of live or recorded acoustic data, such as musical instruments, birds, engine or machine noise, or human singing.

The preferred embodiment of the invention classifies input signals as follows. An input signal is digitized into binary data, which is transformed to a time-frequency representation (spectrogram). Background noise is estimated and a signal detector isolates periods containing signal. Periods without signal  
5 content are included in the noise estimate. The spectrogram of the input signal is rescaled and compared to spectrograms for a number of templates defining a signal model, where each signal model represents a source class. The average distortion between the measured spectrograms and the spectrograms of each signal model is calculated. The signal model with the lowest distortion is  
10 selected. If the average distortion of the selected signal model is sufficiently low, the source is declared to belong to the corresponding class. If not, the source is declared to be of unknown type.

The set of signal models is trained with signal data by creating templates from the spectrograms of the input signals when such spectrograms are  
15 significantly different from the spectrograms of existing templates. If an existing template is found that resembles the input signal spectrogram, that template is averaged with the input signal spectrogram in such a way that the resulting template is the average of all the spectra that matched that template in the past.

20 The invention has the following advantages: It is able to classify an acoustic signal source: independently of the sound the source happens to be emitting at the time of sampling; independently of sound levels; and even when some portions of the spectra of the acoustic signal are masked by noise. The invention also requires relatively few training, testing data, and computational  
25 resources.

The details of one or more embodiments of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

### DESCRIPTION OF DRAWINGS

FIG. 1 is block diagram of a prior art programmable computer system suitable for implementing the signal enhancement technique of the invention.

5        FIG. 2 is a flow diagram showing the basic method of the preferred embodiment of the invention.

FIG. 3 is a flow diagram showing a preferred process for estimating background noise parameters and detecting the presence of signal.

10       FIG. 4 is a flow diagram showing the preferred method to detect the presence of harmonically related peaks in a signal spectrum.

FIG. 5 is a flow diagram showing a preferred method for generating and using signal model templates.

Like reference numbers and designations in the various drawings indicate like elements.

15

## DETAILED DESCRIPTION

Throughout this description, the preferred embodiment and examples shown should be considered as exemplars rather than as limitations of the invention.

### 5 *Overview of Operating Environment*

FIG. 1 shows a block diagram of a typical prior art programmable processing system which may be used for implementing the acoustic signal classification system of the invention. An acoustic signal is received at a transducer microphone 10, which generates a corresponding electrical signal  
10 representation of the acoustic signal. The signal from the transducer microphone 10 is then preferably amplified by an amplifier 12 before being digitized by an analog-to-digital converter 14. The output of the analog-to-digital converter 14 is applied to a processing system, which applies the classification techniques of the invention. The processing system preferably  
15 includes a CPU 16, RAM 20, ROM 18 (which may be writable, such as a flash ROM), and an optional storage device 22, such as a magnetic disk, coupled by a CPU bus as shown. The output of the classification process can be displayed for the benefit of a human user by means of a video display controller 24 which drives a video display 26, or used by the system to customize its response to the  
20 identity of the sound source, or used to actuate external equipment (e.g., lock mechanisms in an access control application).

### *Functional Overview of System*

The following describes the functional components of an acoustic signal classification system. A first functional component of the invention is a pre-processor that transforms input data to a time-frequency representation. The  
25 patterns of the relative power in different frequency bands and how such

patterns change in the short term are used by the present system to classify an input signal.

The second and third functional components of the invention are a dynamic background estimator and a signal detector respectively, working in tandem. A signal detector is useful to discriminate against continuous background noise. It is important to ensure that classification is based on signal only, and is not influenced by background noise. The dynamic background noise estimation function is capable of separating transient sounds from background noise, and estimating the background noise alone. In one embodiment, a power detector acts in each of multiple frequency bands. Noise-only portions of the data are used to generate mean and standard deviation of the noise in decibels (dB). When the power exceeds the mean by more than a specified number of standard deviations in a frequency band, the corresponding time period is flagged as containing signal and is not used to estimate the noise-only spectrum.

The fourth functional component of the invention is a harmonic detector. In the case of harmonic sounds, the harmonic detector is also used to provide an estimate for the fundamental frequency of the signal that can be useful for classification. A harmonic detector is a useful filter to apply to the data since in many cases of interest (e.g., human voice, music, bird singing, engine and machinery), the signal has a harmonic structure. A preferred embodiment of a harmonic detector is described below. The harmonic detector counts the number of harmonically related peaks in the spectrum.

The fifth functional component is a spectral rescaler. The input signal can be weak or strong, close or far. Before measured spectra are matched against templates in a model, the measured spectra are rescaled so that the inter-pattern distance does not depend on the overall loudness of the signal. In the preferred embodiment, a weighting proportional to the signal-to-noise ratio (SNR) in decibels (dB) is applied to the frequency bands during rescaling. The

weights are bounded below and above by a minimum and a maximum value, respectively. The spectra are rescaled so that the weighted distance to each stored template is minimized.

The sixth functional component is a pattern matcher. The pattern  
5 matcher compares the spectrogram of the input signal to a set of signal models, each defining a class. Each signal model consists of a set of prototypical spectrograms of short duration (“templates”) obtained from signals of known identity. Signal model training is accomplished by collecting spectrograms that are significantly different from prototype spectrograms previously collected. In  
10 the preferred embodiment, the first prototype spectrogram is the first input signal spectrogram containing signal significantly above the noise level. For subsequent time epochs, if the input signal spectrogram is closer to any existing prototype spectrogram than a selected distance threshold, then that input signal spectrogram is averaged with the closest prototype spectrogram. If the input  
15 signal spectrogram is farther away from any prototype spectrogram than the selected threshold, then the input signal spectrogram is declared to be a new prototype spectrogram.

The distance between templates and the measured spectrogram of the input signal can be one of several appropriate metrics, such as the Euclidean  
20 distance or a weighted Euclidean distance. For each signal model class, the template with the smallest distance to the measured input signal spectrogram is selected as the best fitting prototype spectrogram for that class.

The seventh functional component is a classifier. A score for each class is accumulated for each input signal sample. When sufficient data has been  
25 collected from a suitable number of input signal samples, a final classification decision is made. Alternatively, a decision can be forced at any desired time or event (for example, if a period of speech is followed by a significant period of silence), and the best fitting class returned along with the score at that point.

### Overview of Basic Method

FIG. 2 is a flow diagram of the preferred method embodiment of the invention. The method shown in FIG. 2 is used for enhancing an incoming acoustic signal, which consists of a plurality of data samples generated as output from the analog-to-digital converter 14 shown in FIG. 1. The method begins at a Start state (Step 202). The incoming data stream (e.g., a previously generated acoustic data file or a digitized live acoustic signal) is read into a computer memory as a set of samples (Step 204). In the preferred embodiment, the invention normally would be applied to classify from a “moving window” of data representing portions of a continuous acoustic data stream, such that the entire data stream is processed. Generally, an acoustic data stream to be classified is represented as a series of data “buffers” of fixed length, regardless of the duration of the original acoustic data stream.

The samples of a current window are subjected to a time-frequency transformation, which may include appropriate conditioning operations, such as pre-filtering, shading, etc. (Step 206). Any of several time-frequency transforms can be used, such as the short-time Fourier transform, banks of filter analysis, discrete wavelet transform, *etc.*

The result of the time-frequency transformation is that the initial time series input signal  $x(t)$  is transformed into a time-frequency representation  $X(f, i)$ , where  $t$  is the sampling index to the time series  $x$ , and  $f$  and  $i$  are discrete variables respectively indexing the frequency and time dimensions of spectrogram  $X$ . In the preferred embodiment, the logarithm of the magnitude of  $X$  is used instead of  $X$  in subsequent steps unless specified otherwise, i.e.:

$$P(f, i) = 20 \log_{10}(|X(f, i)|).$$

The power level  $P(f, i)$  as a function of time and frequency will be referred to as a “spectrogram” from now on.

The power levels in individual frequency bands  $f$  are then subjected to background noise estimation (Step 208). A signal detector detects the presence

of signal buried in stationary background noise (Step 210), and passes only spectrograms that include signal. The background noise estimation updates the estimate of the background noise parameters when no signal is present.

A preferred embodiment for performing background noise estimation  
 5 comprises a power detector that averages the acoustic power in a sliding window for each frequency band  $f$ . When the power within a predetermined number of frequency bands exceeds a threshold, determined as a certain number of standard deviations above the background noise, the power detector declares the presence of signal, i.e., when:

10

$$P(f, i) > B(f) + \sigma(f),$$

where  $B(f)$  is the mean background noise power in band  $f$ ,  $\sigma(f)$  is the standard deviation of the noise in that same band, and  $c$  is a constant. In an alternative  
 15 embodiment, noise estimation need not be dynamic, but could be measured once (for example, during boot-up of a computer running software implementing the invention).

The spectrograms that are passed through the signal detector are then applied to a harmonic detector function (Step 212). This step allows the system  
 20 to discriminate against signals that are not of the same harmonic class as the input signal and therefore for which no further comparison is necessary. For example, the human voice is characterized by the presence of a set of harmonics between 0.1 and about 3 kHz, with a fundamental frequency (pitch) of between 90 Hz for adult males to 300 Hz for children.

25 The spectrograms  $P$  from Step 206 are then preferably rescaled so that they can be compared to stored templates (Step 214). One method of performing this step is to shift each element of the spectrogram  $P(f, i)$  up by a constant  $k(i, m)$  so that the root-mean-squared difference between  $P(f, i) + k(i,$

$m$ ) and the  $m^{\text{th}}$  template  $T(f, m)$  is minimized. This is accomplished by taking the following, where  $N$  is the number of frequency bands:

$$k(i, m) = \frac{1}{N} \sum_{f=1}^N [P(f, i) - T(f, m)]$$

In another embodiment, weighting is used to rescale the templates prior to comparison. The weights  $w(i)$  are proportional to the SNR  $r(f, i)$  in band  $f$  at time  $i$ , calculated as a difference of levels, *i.e.*  $r(f, i) = P(f, i) - B(f)$  for each frequency band. In this embodiment, each element of the rescaling factor is weighted by a weight defined as follows, where  $w_{\min}$  and  $w_{\max}$  are preset thresholds:

$$\begin{aligned} w(f, i) &= w_{\min} \text{ if } r(f, i) < w_{\min}; \\ w(f, i) &= w_{\max} \text{ if } r(f, i) > w_{\max}; \\ w(f, i) &= r(f, i) \text{ otherwise.} \end{aligned}$$

In the preferred embodiment, the weights are normalized by the sum of the weights at each time frame, *i.e.*:

$$\begin{aligned} w'(f, i) &= w(f, i) / \sum_f (w(f, i)), \\ w'_{\min} &= w_{\min} / \sum_f (w(f, i)), \\ w'_{\max} &= w_{\max} / \sum_f (w(f, i)). \end{aligned}$$

$$k(i, m) = \frac{1}{N} \sum_{f=1}^N [P(f, i) - T(f, m)] w'(f, i)$$

In that case, the rescaling constant is given by:

The effect of such rescaling is to preferentially align the frequency bands of the templates having a higher SNR. However, rescaling is optional and need not be used in all embodiments.

In another embodiment, the SNR of the templates is used as well as the SNR of the measured spectra for rescaling the templates. The SNR of template  $T(f, m)$  is defined as  $r_N(f, m) = T(f, m) - B_N(f)$ , where  $B_N(f)$  is the background noise in frequency band  $f$  at the time of training. In one embodiment of a

5 weighting scheme using both  $r$  and  $r_N$ , the weights  $w_N$  are defined as the square-root of the product of the weights for the templates and the spectrogram:

Other combinations of  $r_N$  and  $r$  are admissible. In the preferred embodiment, the weights are normalized by the sum of the weights at each time

$$\begin{aligned} w_2(f, i, m) &= w_{\min} \text{ if } \sqrt{r_N(f, m)r(f, i)} < w_{\min}; \\ w_2(f, i, m) &= w_{\max} \text{ if } \sqrt{r_N(f, m)r(f, i)} > w_{\max}; \\ w_2(f, i, m) &= \sqrt{r_N(f, m)r(f, i)} > w_{\max} \text{ otherwise.} \end{aligned}$$

frame, *i.e.*:

$$\begin{aligned} 10 \quad w'_2(f, i) &= w_2(f, i) / \sum_f(w_2(f, i)), \\ w'_{\min} &= w_{\min} / \sum_f(w_2(f, i)), \\ w'_{\max} &= w_{\max} / \sum_f(w_2(f, i)). \end{aligned}$$

After spectral rescaling, the preferred embodiment performs pattern matching to find a template  $T^*$  in a signal model that best matches the current

15 spectrogram  $P(f, i)$  (Step 216). There exists some latitude in the definition of the term “best match”, as well as in the method used to find that best match. In one embodiment, the template with the smallest r.m.s. (root mean square) difference  $d^*$  between  $P + k$  and  $T^*$  is found. In the preferred embodiment, the weighted r.m.s. distance is used, where:

$$20 \quad d(i, m) = \frac{1}{N} \sum_{f=1}^N [P(f, i) + k(i, m) - T(f, m)]^2 w'_2(f, i, m)$$

In this embodiment, the frequency bands with the least SNR contribute less to the distance calculation than those bands with more SNR. The best

matching template  $T^*(i)$  at time  $i$  is selected by finding  $m$  such that  $d^*(i) = \min_m (d(i, m))$ .

The last component is a classifier. A score for each class is accumulated, and when sufficient data has been collected, a decision is made. For example, a score can be the average of the distances  $d(i, m)$  over time  $i$ . In a typical embodiment, 8-20 scores are accumulated, each corresponding to a buffer of voiced speech (as opposed to unvoiced speech – consonants – since the buffers without voiced speech do not contain as much information as to the identity of the speaker. The classification decision may simply comprise comparing a score to a threshold, resulting in a binary determination, or one can use a “soft” classifier, such as a neural network. Alternatively, a decision can be forced at any desired time or event, and the best fitting class is returned, along with the score at that point. The score can include a component that relates the contribution of the fundamental frequency to the total score. The preferred embodiment of that component is of the form  $K(f_0 - f_{\text{source}})^2$ , where  $f_0$  is the measured fundamental frequency,  $f_{\text{source}}$  is the fundamental frequency of the source model, and  $K$  is a proportionality constant.

More particularly, in the preferred embodiment, the score is the average of the distance over time, plus a fundamental frequency term, *i.e.*,

$$s = \frac{1}{N} \sum_{i=I}^{I+N} d^*(i) + K(f_0 - f_{\text{source}})^2$$

where the average is taken over  $N$  points starting at time  $i = I$ . In this case, the score  $s$  needs to be minimized. If  $s$  does not meet a selected threshold value  $T_{\text{unknown}}$  for all models, then the source is declared to be of “unknown” type. Otherwise, the source is declared to belong to the class with the lowest score.

Single or multiple signal models, each comprising one or more templates, may be applied in various applications to classify an input acoustic signal. In the case of a single signal model, the classification is binary.

### Background Noise Estimation and Signal Detection

FIG. 3 is a flow diagram providing a more detailed description of the process of background noise estimation and signal detection which were briefly described as Steps 208 and 210 respectively in FIG. 2. The background noise estimation updates the estimates of the background noise parameters when no signal is present. A signal detector is useful to discriminate against continuous background noise. It is important to ensure that classification is based on signal only, and is not influenced by the background noise.

The process begins at a Start Process state (Step 302). The process needs a sufficient number (*e.g.*, 1 second) of samples of background noise before it can use the mean and standard deviation of the noise to detect signal. Accordingly, the routine determines if a sufficient number of samples of background noise have been obtained (Step 304). If not, the present sample is used to update the noise estimate (Step 306) and the process is terminated (Step 310). In one embodiment of the background noise update process, the spectrogram elements  $P(f, i)$  are kept in a ring buffer and used to update the mean  $B(f)$  and the standard deviation  $\sigma(f)$  of the noise in each frequency band  $f$ . The background noise estimate is considered ready when the index  $i$  is greater than a preset threshold.

If the background samples are ready (Step 304), then a determination is made as to whether the signal level  $P(f, i)$  of a current input signal sample is significantly above the background in some of the frequency bands (Step 308). In a preferred embodiment, when the power within a predetermined number of frequency bands is greater than a threshold, determined as a certain number of standard deviations above the background noise mean level, the determination step indicates that the power threshold has been exceeded, i.e., when

$$P(f, i) > B(f) + c \sigma(f),$$

where  $c$  is a constant predetermined empirically (Step 312). The process then ends (Step 310). If a sufficiently powerful signal is not detected in Step 308,

then the background noise statistics are updated in Step 306 and the process then ends (Step 310).

### *Harmonic Detector*

FIG.4 is a flow diagram providing a more detailed description of the process of harmonic detection which was briefly described as Step 212 in FIG. 2. The harmonic detector detects the presence of peaks in the spectrum of an input signal sample that have a harmonic relation between them. This step is often useful, since a large proportion of sources of interest have spectra that are characterized as having a harmonic relationship between their frequency components.

The process begins at a Start Process state (Step 402). The transformed spectrum of an input signal sample is scanned for local peaks for frequencies up to a maximum frequency of  $f_{\max}$  in order to “pick” a peak (Step 404). A local peak is declared at  $P(f)$  if  $P(f-1) < P(f) < P(f+1)$ . The peaks that stand above the neighboring spectrum values by more than a threshold  $\varepsilon$ , i.e., those  $f$  for which  $P(f-1) + \varepsilon < P(f) < P(f+1) + \varepsilon$ , are extracted (Step 406). Each of those peaks represents one “vote” for each of the fundamental frequencies  $f_0$  (Step 408). The preferred embodiment estimate of  $V_0(f_0)$  is  $\text{floor}(f_{\max}/f_0)$ . Since lower values of  $f_0$  have fewer harmonics for a given  $f_{\max}$  than higher  $f_0$ , the votes are normalized by the expected number of harmonics in the frequency range considered  $V_0(f_0)$  (Step 410). If the ratio  $V(f_0)/V_0(f_0)$  is greater than a threshold (Step 412), a harmonic relationship is declared to exist.

### *Pattern Matching*

FIG. 5 is a flow diagram providing a more detailed description of the process of pattern matching which was briefly described as Step 216 of FIG. 2. The process begins at a Start Process state (Step 502). The pattern matching process finds a template  $T^*$  in the signal model that best matches a current spectrogram  $P(f, i)$  (Step 504). The pattern matching process is also responsible

for the learning process of the signal model. There exists some latitude in the definition of the term “best match”, as well as in the method used to find that best match. In one embodiment, the template with the smallest r.m.s. difference  $d^*$  between  $P + k$  and  $T^*$  is found. In the preferred embodiment, the weighted  
 5 r.m.s. distance is used to measure the degree of match. In one embodiment, the r.m.s. distance is calculated by:

$$d(i, m) = \frac{1}{N} \sum_{f=1}^N [P(f, i) + k(i, m) - T(f, m)]^2 w_2(f, i, m)$$

In this embodiment, the frequency bands with the least SNR contribute less to the distance calculation than those bands with more SNR. The best matching template  $T^*(f, i)$  that is the output of Step 504 at time  $i$  is selected by  
 10 finding  $m$  such that  $d^*(i) = \min_m[d(i, m)]$ . If the system is not in learning mode (Step 506), then  $T^*(f, i)$  is also the output of the process as being the closest template (Step 508). The process then ends (Step 510)

If the system is in learning mode (Step 506), the template  $T^*(f, i)$  most similar to  $P(f, i)$  is used to adjust the signal model. The manner in which  $T^*(f, i)$  is incorporated in the model depends on the value of  $d^*(i)$  (Step 512). If  $d^*(i) < d_{max}$ , where  $d_{max}$  is a predetermined threshold, then  $T^*(f, i)$  is adjusted (Step  
 15 516), and the process ends (Step 510). The preferred embodiment of Step 516 is implemented such that  $T^*(f, i)$  is the average of all spectra  $P(f, i)$  that are used to compose  $T^*(f, i)$ . In the preferred embodiment, the number  $n_m$  of  
 20 spectra associated with  $T(f, m)$  is kept in memory, and when a new spectrum  $P(f, i)$  is used to adjust  $T(f, m)$ , the adjusted template is:

$$T(f, m) = [n_m T(f, m) + P(f, i)] / (n_m + 1),$$

and the number of patterns corresponding to template  $m$  is adjusted as well:

$$n_m = n_m + 1.$$

25 Referring back to Step 512, if  $d^*(i) > d_{max}$  then a new template is created,  $T^*(f, i) = P(f, i)$ , with a weight  $n_m = 1$  (Step 514), and the process ends (Step 510).

*Computer Implementation*

The invention may be implemented in hardware or software, or a combination of both (e.g., programmable logic arrays). Unless otherwise specified, the algorithms included as part of the invention are not inherently  
5 related to any particular computer or other apparatus. In particular, various general-purpose machines may be used with programs written in accordance with the teachings herein, or it may be more convenient to construct more specialized apparatus to perform the required method steps. However, preferably, the invention is implemented in one or more computer programs  
10 executing on programmable systems each comprising at least one processor, at least one data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. Each such programmable system component constitutes a means for performing a function. The program code is executed on the processors to  
15 perform the functions described herein.

Each such program may be implemented in any desired computer language (including machine, assembly, high level procedural, or object oriented programming languages) to communicate with a computer system. In any case, the language may be a compiled or interpreted language.

20 Each such computer program is preferably stored on a storage media or device (e.g., ROM, CD-ROM, or magnetic or optical media) readable by a general or special purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system  
25 may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

A number of embodiments of the invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. For example, some of the steps of various of the algorithms may be order independent, and  
5 thus may be executed in an order other than as described above. Accordingly, other embodiments are within the scope of the following claims.

## WHAT IS CLAIMED IS:

1. A method for classifying acoustic signal within a digitized acoustic input signal, including:
  - (a) transforming the digitized acoustic input signal to a time-frequency representation;  
5
  - (b) estimating a background noise level in the time-frequency representation;
  - (c) for each interval of the time-frequency representation containing significant signal levels, comparing the time-frequency representation of such interval with at least one signal model and determining at least  
10 one template in one such signal model that best matches the time-frequency representation of such interval, based in part on signal-to-noise ratio, and determining a score for each such comparison; and
  - (d) assigning the digitized acoustic input signal to the signal model with  
15 the best score.
2. The method of claim 1, wherein the step of assigning further includes the step of rejecting any signal model having a score that does not meet a selected threshold value.

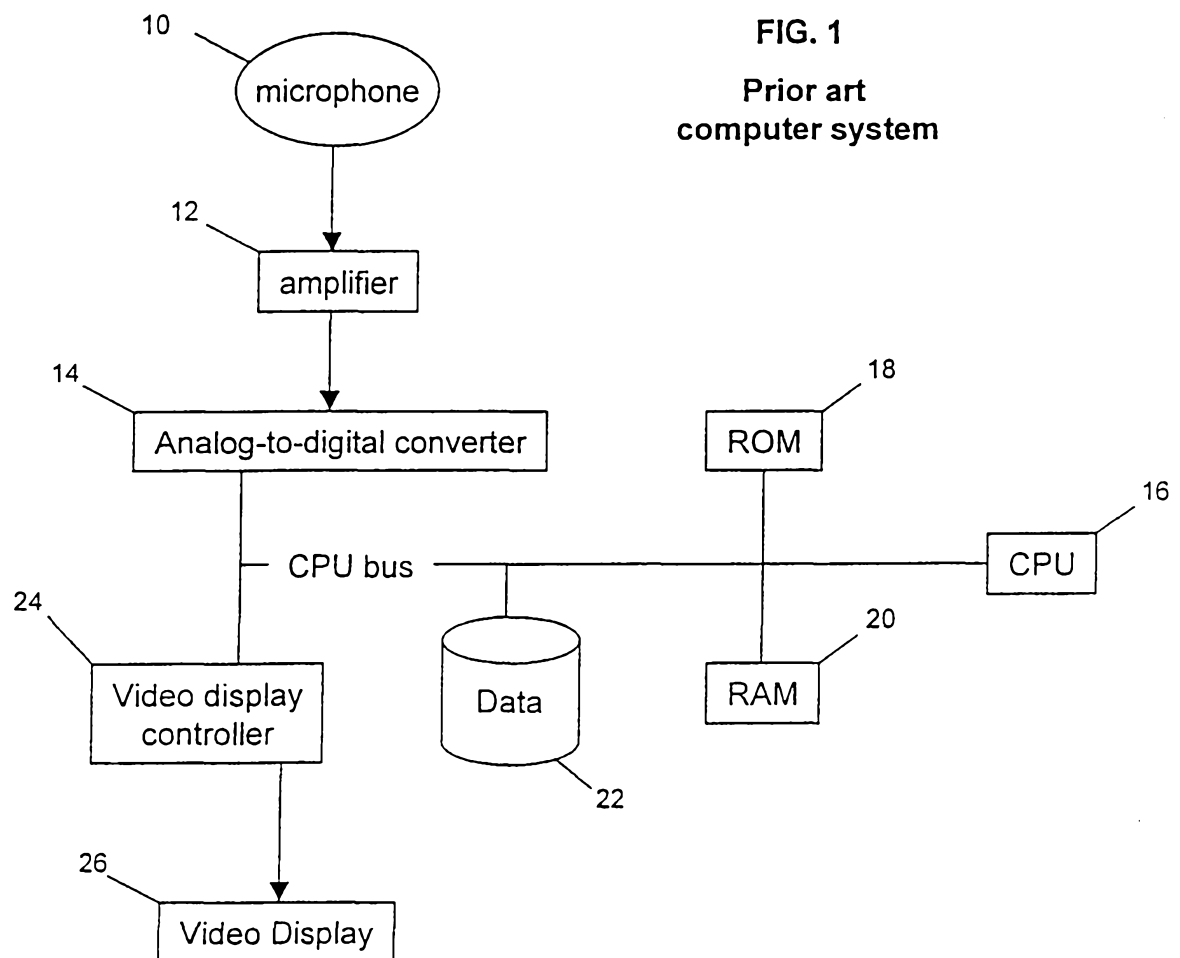
3. A method for classifying acoustic signal within a digitized acoustic input signal, including:
- (a) transforming the digitized acoustic input signal to a time-frequency representation;
  - 5 (b) isolating transient sounds within the time-frequency representation;
  - (c) estimating background noise and including long transients without signal content and background noise between transients in such estimating;
  - (d) detecting the presence of harmonics in the time-frequency  
10 representation;
  - (e) rescaling the time-frequency representation of the estimated background noise;
  - (f) comparing the rescaled time-frequency representation of each transient sound containing any signal of interest with at least one signal model  
15 and determining at least one template in one such signal model that best matches such representation; and
  - (g) assigning the digitized acoustic input signal to the signal model with the best score.
4. The method of claim 3, wherein the step of assigning further includes the  
20 step of rejecting any signal model having a score that does not meet a selected threshold value.

5. A system for classifying acoustic signal within a digitized acoustic input signal, including:
- a. computational means for transforming the digitized acoustic input signal to a time-frequency representation;
  - 5 b. computational means for estimating a background noise level in the time-frequency representation;
  - c. computational means for comparing the time-frequency representation of such interval with at least one signal model for each interval of the time-frequency representation containing significant signal levels, and  
10 determining at least one template in one such signal model that best matches the time-frequency representation of such interval, based in part on signal-to-noise ratio, and determining a score for each such comparison; and
  - d. computational means for assigning the digitized acoustic input signal to  
15 the signal model with the best score.
6. The system of claim 5, further including computational means for rejecting any signal model having a score that does not meet a selected threshold value.

7. A system for classifying acoustic signal within a digitized acoustic input signal, including:
- a. computational means for transforming the digitized acoustic input signal to a time-frequency representation;
  - 5 b. computational means for isolating transient sounds within the time-frequency representation;
  - c. computational means for estimating background noise and including long transients without signal content and background noise between transients in such estimating;
  - 10 d. computational means for detecting the presence of harmonics in the time-frequency representation;
  - e. computational means for rescaling the time-frequency representation of the estimated background noise;
  - f. computational means for comparing the rescaled time-frequency  
15 representation of each transient sound containing any signal of interest with at least one signal model and determining at least one template in one such signal model that best matches such representation; and
  - g. computational means for assigning the digitized acoustic input signal to the signal model with the best score.
- 20 8. The system of claim 7, further including computational means for rejecting any signal model having a score that does not meet a selected threshold value.

9. A computer program, residing on a computer-readable medium, for classifying acoustic signal within a digitized acoustic input signal, the computer program comprising instructions for causing a computer to:
- (a) transform the digitized acoustic input signal to a time-frequency representation;
  - (b) estimate a background noise level in the time-frequency representation;
  - (c) for each interval of the time-frequency representation containing significant signal levels, compare the time-frequency representation of such interval with at least one signal model and determine at least one template in one such signal model that best matches the time-frequency representation of such interval, based in part on signal-to-noise ratio, and determine a score for each such comparison; and
  - (d) assign the digitized acoustic input signal to the signal model with the best score.
10. The computer program method of claim 9, further including instructions for causing a computer to reject any signal model having a score that does not meet a selected threshold value.

11. A computer program, residing on a computer-readable medium, for classifying acoustic signal within a digitized acoustic input signal, the computer program comprising instructions for causing a computer to:
- (a) transform the digitized acoustic input signal to a time-frequency representation;
  - (b) isolate transient sounds within the time-frequency representation;
  - (c) estimate background noise and including long transients without signal content and background noise between transients in such estimating;
  - (d) detect the presence of harmonics in the time-frequency representation;
  - (e) rescale the time-frequency representation of the estimated background noise;
  - (f) compare the rescaled time-frequency representation of each transient sound containing any signal of interest with at least one signal model and determine at least one template in one such signal model that best matches such representation; and
  - (g) assign the digitized acoustic input signal to the signal model with the best score.
12. The method of claim 11, further including instructions for causing a computer to reject any signal model having a score that does not meet a selected threshold value.



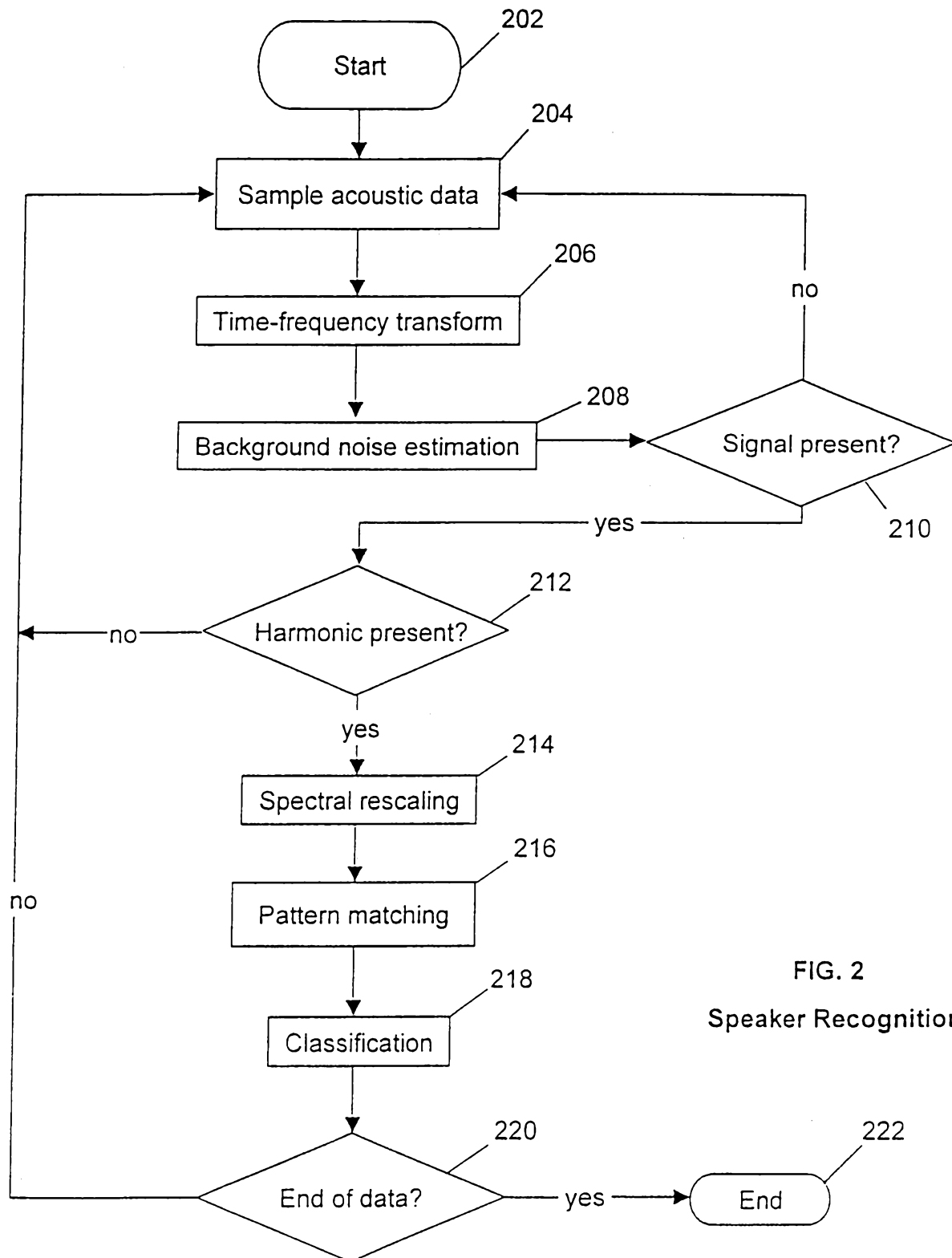


FIG. 2  
Speaker Recognition

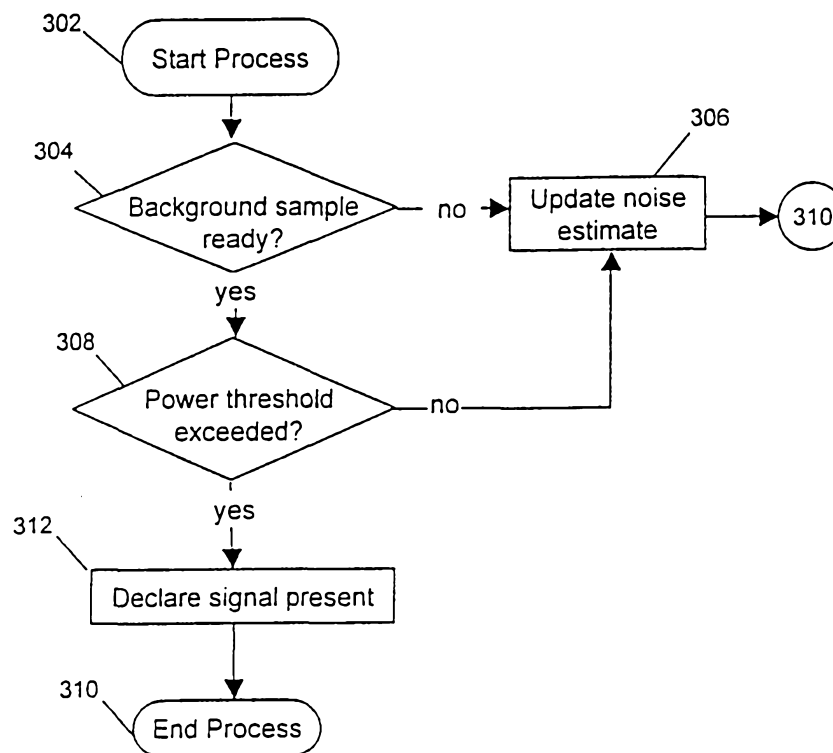


FIG. 3

Background noise estimation  
and signal detection

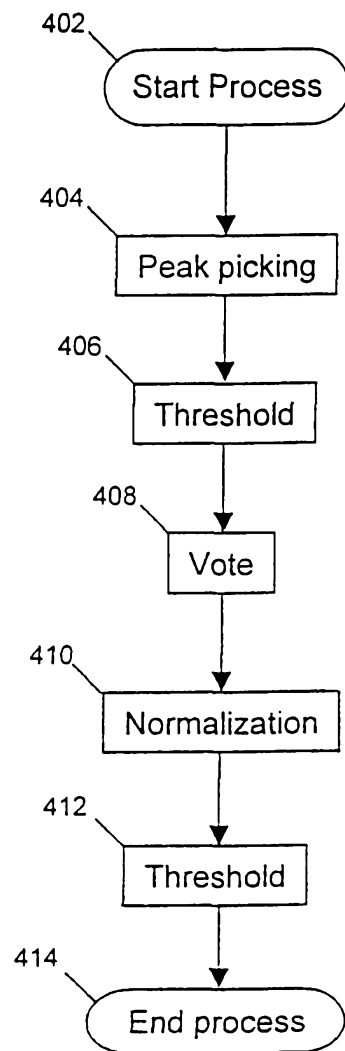


FIG. 4  
Harmonic detector

**FIG. 5**  
**Pattern matching routine**

