



## (51) International Patent Classification:

G06F 17/30 (2006.01) G10L 25/18 (2013.01)  
G10L 25/48 (2013.01)

## (21) International Application Number:

PCT/JP2013/002581

## (22) International Filing Date:

16 April 2013 (16.04.2013)

## (25) Filing Language:

English

## (26) Publication Language:

English

## (30) Priority Data:

2012-094395 18 April 2012 (18.04.2012) JP

## (71) Applicant: SONY CORPORATION [JP/JP]; 1-7-1, Konan, Minato-Ku, Tokyo, 1080075 (JP).

## (72) Inventors: ABE, Mototsugu; c/o SONY CORPORATION, 1-7-1, Konan, Minato-ku, Tokyo, 1080075 (JP). NISHIGUCHI, Masayuki; c/o SONY CORPORATION, 1-7-1, Konan, Minato-ku, Tokyo, 1080075 (JP). KURATA, Yoshinori; c/o SONY CORPORATION, 1-7-1, Konan, Minato-ku, Tokyo, 1080075 (JP).

## (74) Agents: SASAKI, Eiji et al.; Daido Patent Attorneys Partnership Corporation KSK Bldg. West 8F, 3-25-9, Hatchobori, Chuo-Ku, Tokyo, 1040032 (JP).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

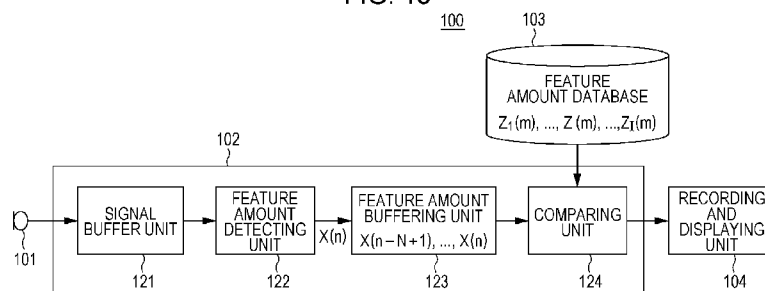
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

## Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

(54) Title: SOUND DETECTING APPARATUS, SOUND DETECTING METHOD, SOUND FEATURE VALUE DETECTING APPARATUS, SOUND FEATURE VALUE DETECTING METHOD, SOUND SECTION DETECTING APPARATUS, SOUND SECTION DETECTING METHOD, AND PROGRAM

FIG. 13



(57) Abstract: There is provided a sound detecting apparatus including: a feature value extracting unit which extracts a feature value per every predetermined time from an input time signal; a feature value maintaining unit which maintains a feature value sequence of a predetermined number of detection target sound items; and a comparison unit which respectively compares a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtains detection results of the predetermined number of detection target sound items every time the feature value extracting unit newly extracts a feature value, wherein the feature value extracting unit includes a time frequency transform unit and a likelihood distribution detecting unit, smooths the obtained likelihood distribution in frequency and time directions, and extracts a feature value per the predetermined time.

## Description

# Title of Invention: SOUND DETECTING APPARATUS, SOUND DETECTING METHOD, SOUND FEATURE VALUE DETECTING APPARATUS, SOUND FEATURE VALUE DETECTING METHOD, SOUND SECTION DETECTING APPARATUS, SOUND SECTION DETECTING METHOD, AND PROGRAM

### Technical Field

[0001] The present technology relates to a sound detecting apparatus, a sound detecting method, a sound feature value detecting apparatus, a sound feature value detecting method, a sound section detecting apparatus, a sound section detecting method and a program.

### Background Art

[0002] In recent years, home electrical appliances (electric devices for domestic use) generate various kinds of sound (hereinafter, referred to as "running state sound") such as control sounds, notification sounds, operating sounds, and alarm sounds in accordance with running state. If it is possible to observe such running state sounds by a microphone or the like installed at a certain place at home and detect when and which home electrical appliance performs what kind of operation, various application functions such as automatic collection of self action history, which is a so-called life log, visualization of notification sounds for people with hearing difficulties, and action monitoring for aged people who live alone can be realized.

[0003] The running state sound may be a simple buzzer sound, beep sound, music, voice sound, or the like, and a continuation time length is about 300 ms in a case of a short continuation time length and about several tens of seconds in a case of a long continuation time length. Such running state sound is reproduced by a reproduction device, sound from which is not sufficiently satisfactory, such as a piezoelectric buzzer or a thin speaker mounted on each home electrical appliance, and is made to propagate in the surroundings.

[0004] For example, PTL 1 discloses a technology in which partial fragmented data of a music composition is transformed into a time frequency distribution, a feature value is extracted and then compared with a feature value of a music composition, which has already been registered, and a name of the music composition is identified.

### Citation List

## Patent Literature

[0005] PTL 1: Japanese Patent No. 4788810

## Summary of Invention

### Technical Problem

[0006] It can be also considered that the same technology as that disclosed in PTL 1 is applied to detection of the aforementioned running state sound. In relation to the running state sound generated by home electrical appliances, however, the following facts that hinder such detection are present.

[0007] (1) It is necessary to recognize running state sound which is as short as several hundred milliseconds.

[0008] (2) Due to poor quality of a reproduction device, sound becomes distorted, or resonance occurs and in some cases frequency characteristics are extremely distorted.

[0009] (3) Amplitude and phase frequency characteristics are further distorted compared with sound generated by an actual domestic electrical appliance, due to propagation in the surroundings.

For example, Fig. 17A shows an example of a waveform of running state sound recorded at a position which is close to a domestic electrical appliance. On the other hand, Fig. 17B shows an example of a waveform of running state sound recorded at a position which is distant from the domestic electrical appliance, and the waveform is distorted.

[0010] (4) Relatively large noise and non-constant noise such as output sound from a television and conversation sound are superimposed in some cases due to propagation in the surroundings. For example, Fig. 17C shows an example of a waveform of running state sound recorded at a position which is close to a television as a noise source, and the running state sound is buried in noise.

[0011] (5) Since a level of sound from each domestic electrical appliance and distance to the microphone depends on each home electrical appliance, the volume of recorded sound varies.

[0012] It is desired to satisfactorily detect detection target sound such as running state sound generated from a home electrical appliance.

### Solution to Problem

[0013] An embodiment of the present technology relates to a sound detecting apparatus including: a feature value extracting unit which extracts a feature value per every predetermined time from an input time signal; a feature value maintaining unit which maintains a feature value sequence of a predetermined number of detection target sound items; and a comparison unit which respectively compares a feature value sequence extracted by the feature value extracting unit with a feature value sequence of

the maintained predetermined number of detecting target sound items and obtains detection results of the predetermined number of detection target sound items every time the feature value extracting unit newly extracts a feature value, wherein the feature value extracting unit includes a time frequency transform unit which performs time frequency transform on the input time signal for each time frame and obtains time frequency distribution, a likelihood distribution detecting unit which obtains tone likelihood distribution from the time frequency distribution, and a smoothing unit which smooths the likelihood distribution in a frequency direction and a time direction, and extracts the feature value per the predetermined time from the smoothed likelihood distribution.

- [0014] According to the present technology, the feature value extracting unit extracts the feature value per the predetermined time from the input time signal. In such a case, the feature value extracting unit performs time frequency transform on the input signal for each time frame, obtains the time frequency distribution, obtains tone likelihood distribution from the time frequency distribution, smooths the likelihood distribution in the frequency direction and the time direction, and extracts the feature value per the predetermined time from the smoothed likelihood distribution.
- [0015] For example, the likelihood distribution detecting unit may include a peak detecting unit which detects a peak in the frequency direction in each time frame of the time frequency distribution, a fitting unit which fits the tone model at each detected peak, and a scoring unit which obtains a score representing tone component likeliness at each detected peak based on the fitting result.
- [0016] The feature value maintaining unit maintains a feature value sequence of the predetermined number of detection target sound items. The detection target sound can include voice sound of a person or an animal or the like as well as running state sound generated from a domestic electrical appliance (control sounds, notification sounds, operating sounds, alarm sounds, and the like). Every time the feature value extracting unit newly extracts a feature value, the comparison unit respectively compares the feature value sequence extracted by the feature value extracting unit with the feature value sequence of the maintained predetermined number of detection target sound and obtains the detection results of the predetermined number of detection target sound items.
- [0017] For example, the comparison unit may obtain similarity based on correlation between corresponding feature values between the feature value sequence of the maintained detection target sound items and the feature value sequence extracted by the feature value extracting unit for each of the predetermined number of detection target sound items and obtain the detection results of the detection target sound items based on the obtained similarity.

- [0018] According to the present technology, the tone likelihood is obtained from the time frequency distribution of the input time signal, the feature value per every predetermined time is extracted and used from the likelihood distribution which has been smoothed in the frequency direction and the time direction, and it is possible to precisely detect detection target sound (running state sound and the like generated from a domestic electrical appliance) without depending on an installation position of the microphone.
- [0019] According to the present technology, for example, the feature value extracting unit may further include a thinning unit which thins the smoothed likelihood distribution in the frequency direction and/or the time direction. According to the present technology, for example, the feature value extracting unit may further include a quantizing unit which quantizes the smoothed likelihood distribution. In such a case, it is possible to reduce the data amount of the feature value sequence and to thereby reduce burden of the comparison computation.
- [0020] According to the present technology, for example, the apparatus may further include a recording control unit which records the detection results of the predetermined number of detection target sound items along with time information on a recording medium. In such a case, for example, it is possible to obtain a user action history at home such as an operation history of a domestic electrical appliance.
- [0021] Another concept of the present technology relates to a sound feature value extracting apparatus including: a time frequency transform unit which performs time frequency transform on an input time signal for each time frame and obtains time frequency distribution; a likelihood distribution detecting unit which obtains tone likelihood distribution from the time frequency distribution; and a feature value extracting unit which smooths the likelihood distribution in a frequency direction and a time direction and extracts a feature value per every predetermined time.
- [0022] According to the present technology, the time frequency transform unit performs time frequency transform on the input time signal for each time frame and obtains the time frequency distribution. The likelihood distribution detecting unit obtains the tone likelihood distribution from the time frequency distribution. For example, the likelihood distribution detecting unit may include a peak detecting unit which detects a peak in the frequency direction in each time frame of the time frequency distribution, a fitting unit which fits the tone model at each detected peak, and a scoring unit which obtains a score representing tone component likeliness at each detected peak based on the fitting result. In addition, the feature value extracting unit smooths the likelihood distribution in the frequency direction and the time direction and extracts the feature value per the predetermined time.
- [0023] As described above, according to the present technology, the tone likelihood dis-

tribution is obtained from the time frequency distribution of the input time signal, the feature value per every predetermined time is extracted from the likelihood distribution which has been smoothed in the frequency direction and the time direction, and it is possible to satisfactorily extract feature values of sound included in the input time signal.

[0024] According to the present technology, for example, the feature value extracting unit may further include a thinning unit which thins out the smoothed likelihood distribution in the frequency direction and/or the time direction. According to the present technology, for example, the feature value extracting unit may further include a quantizing unit which quantizes the smoothed likelihood distribution. In so doing, it is possible to reduce the data amount of the extracted feature values.

[0025] According to the present technology, for example, the apparatus may further include: a sound section detecting unit which detects a sound section based on the input time signal, and the likelihood distribution detecting unit may obtain tone likelihood distribution from the time frequency distribution within a range of the detected sound section. In so doing, it is possible to extract the feature values corresponding to the sound section.

[0026] In such a case, the sound section detecting unit may include a time frequency transform unit which performs time frequency transform on the input time signal for each time frame and obtains time frequency distribution, a feature value extracting unit which extracts feature value of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution, a scoring unit which obtains a score representing a sound section likeliness for each time frame based on the extracted feature values, a time smoothing unit which smooths the obtained score for each time frame in the time direction, and a threshold value determination unit which determinates a threshold value for the smoothed score for each time frame and obtains sound section information.

[0027] In addition, another embodiment of the present technology relates to a sound section detecting apparatus including: a time frequency transform unit which obtains time frequency distribution by performing time frequency transform on an input time signal for each time frame; a feature value extracting unit which extracts feature values of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution; and a scoring unit which obtains a score representing sound section likeliness for each time frame based on the extracted feature values.

[0028] According to the present technology, the time frequency transform unit performs time frequency transform on the input time signal for each time frame and obtains the time frequency distribution. The feature value extracting unit extracts the feature value

of the amplitude, the tone component intensity, and the spectrum approximate outline for each time frame based on the time frequency distribution. In addition, the scoring unit obtains the score representing the sound section likeliness for each time frame based on the extracted feature values. According to the present technology, for example, the apparatus may further include: a time smoothing unit which smooths the obtained score for each time frame in the time direction; and a threshold value determination unit which determines a threshold for the smoothed score for each time frame and obtains sound section information.

[0029] As described above, according to the present technology, the feature values of the amplitude, the tone component intensity, and the spectrum approximate outline for each time frame are extracted from the time frequency distribution of the input time signal, a score representing sound section likeliness for each time frame is obtained from the feature values, and it is possible to precisely obtain the sound section information.

### **Advantageous Effects of Invention**

[0030] According to the present technology, it is possible to satisfactorily detect detection target sound such as running state sound or the like generated by a domestic electrical appliance.

### **Brief Description of Drawings**

[0031] [fig.1]Fig. 1 is a block diagram showing a configuration example of a sound detecting apparatus according to an embodiment.

[fig.2]Fig. 2 is a block diagram showing a configuration example of a feature value registration apparatus.

[fig.3]Fig. 3 is a diagram showing an example of a sound section and noise sections which are present before and after the sound section.

[fig.4]Fig. 4 is a block diagram showing a configuration example of a sound section detecting unit which configures the feature value registration apparatus.

[fig.5A]Fig. 5A is a diagram illustrating a tone intensity feature value calculating unit.

[fig.5B]Fig. 5B is a diagram illustrating the tone intensity feature value calculating unit.

[fig.5C]Fig. 5C is a diagram illustrating the tone intensity feature value calculating unit.

[fig.5D]Fig. 5D is a diagram illustrating the tone intensity feature value calculating unit.

[fig.6]Fig. 6 is a block diagram showing a configuration example of a tone likelihood distribution detecting unit which is included in the tone intensity feature value calculating unit for obtaining distribution of scores  $S(n, k)$  of tone characteristic

likeliness.

[fig.7A]Fig. 7A is a diagram schematically illustrating a characteristic that a quadratic polynomial function fits well in the vicinity of a spectrum peak of a tone characteristic while the quadratic polynomial function does not fit well in the vicinity of a spectrum peak of a noise characteristic.

[fig.7B]Fig. 7B is a diagram schematically illustrating the characteristic that the quadratic polynomial function fits well in the vicinity of a spectrum peak of a tone characteristic while the quadratic polynomial function does not fit well in the vicinity of a spectrum peak of a noise characteristic.

[fig.8A]Fig. 8A is a diagram schematically showing a variation in a peak of the tone characteristic in a time direction.

[fig.8B]Fig. 8B is a diagram schematically showing fitting in a small region gamma on a spectrogram.

[fig.9]Fig. 9 is a flowchart showing an example of a processing procedure for detecting tone likelihood distribution by a tone likelihood distribution detecting unit.

[fig.10]Fig. 10 is a diagram showing an example of a tone component detecting result.

[fig.11]Fig. 11 is a diagram showing an example of a spectrogram of voice sound.

[fig.12]Fig. 12 is a block diagram showing a configuration example of a feature value extracting unit.

[fig.13]Fig. 13 is a block diagram showing a configuration example of a sound detecting unit.

[fig.14]Fig. 14 is a diagram illustrating operations of each part in the sound detecting unit.

[fig.15]Fig. 15 is a block diagram showing a configuration example of a compute apparatus which performs sound detection processing by software.

[fig.16]Fig. 16 is a flowchart showing an example of a procedure for detection target sound detecting processing by a CPU.

[fig.17A]Fig. 17A is a diagram illustrating the recorded state of sound generated by an actual domestic electrical appliance.

[fig.17B]Fig. 17B is a diagram illustrating the recorded state of sound generated by the actual domestic electrical appliance.

[fig.17C]Fig. 17C is a diagram illustrating a recorded state of sound generated by the actual domestic electrical appliance.

### **Description of Embodiment**

[0032] Hereinafter, a description will be given of an embodiment for implementing the present technology (hereinafter, referred to as an "embodiment"). In addition, the description will be given in the following order.

#### **1. Embodiment**



## 2. Modified Example

[0033] <1. Embodiment>

"Sound Detecting Apparatus"

[0034] Fig. 1 shows a configuration example of a sound detecting apparatus 100 according to an embodiment. The sound detecting unit 100 includes a microphone 101, a sound detecting unit 102, a feature value database 103, and a recording and displaying unit 104.

[0035] The sound detecting apparatus 100 executes a sound detecting process for detecting running state sound (control sounds, notification sounds, operating sounds, alarm sounds, and the like) generated by a home electrical appliance and records and displays the detection result. That is, in the sound detecting process, a feature value per every predetermined time is extracted from a time signal  $f(t)$  obtained by collecting sound by the microphone 101, and the feature value is compared with a feature value sequence of a predetermined number of detection target sound items registered in the feature value database. Then, if a comparison result that the feature value substantially coincides with the feature value sequence of the predetermined detection target sound is obtained in the sound detecting process, the time and a name of the predetermined detection target sound are recorded and displayed.

[0036] The microphone 101 collects sounds in a room and outputs the time signal  $f(t)$ . The sounds in the room also include running state sound (control sounds, notification sounds, operating sounds, alarm sounds, and the like) generated by the home electric appliances 1 to N. The sound detecting unit 102 obtains the time signal  $f(t)$ , which is output from the microphone 101, as an input and extracts a feature value per every predetermined time from the time signal. In this regard, the sound detecting unit 102 configures the feature value extracting unit.

[0037] In the feature value data base 103 which configures a feature value maintaining unit, a feature value sequence including a predetermined number of detection target sound items is registered and maintained in association with a detection target sound name. In this embodiment, the predetermined number of detection target sound items means all or a part of the running state sound generated by the home electrical appliances 1 to N, for example. The sound detecting unit 102 compares an extracted feature value sequence with a feature value sequence of the predetermined number of detection target sound items maintained in the feature value database 103 every time a new feature value is extracted and obtains a detection result of a predetermined number of detection target sound. In this regard, the sound detecting unit 102 configures a comparison unit.

[0038] The recording and displaying unit 104 records the detection target sound detecting result by the sound detecting unit 102 in a recording medium along with the time and

displays the detecting result on a display. For example, when the detection target sound detecting result by the sound detecting unit 102 indicate that notification sound A from the home electrical appliance 1 has been detected, the recording and displaying unit 104 records on the recording medium and displays on the display the fact that the notification sound A from the home electrical appliance 1 was produced and the time thereof.

[0039] Operations of the sound detecting apparatus 100 shown in Fig. 1 will be described. The microphone 101 collects sound in a room. The time signal output from the microphone 101 is supplied to the sound detecting unit 102. The sound detecting unit 102 extracts a feature value per every predetermined time from the time signal. Then, the sound detecting unit 102 compares the extracted feature value sequence with a feature value sequence of the predetermined number of detection target sound items maintained in the feature value database 103 every time a new feature value is extracted and obtains the detecting result of the predetermined number of detection target sound items. The detecting result is supplied to the recording and displaying unit 104. The recording and displaying unit 104 records on the recording medium and displays on the display the detecting result along with the time.

[0040] "Feature Value Registration Apparatus"

Fig. 2 shows a configuration example of a feature value registration apparatus 200 which registers a feature value sequence of detection target sound in the feature value database 103. The feature value registration apparatus 200 includes a microphone 201, a sound section detecting unit 202, a feature value extracting unit 203, and a feature value registration unit 204.

[0041] The feature value registration apparatus 200 executes a sound registration process (a sound section detecting process and a sound feature extracting process) and registers a feature value sequence of detection target sound (running state sound generated by a home electrical appliance) in a feature value database 103. Generally, noise sections are present before and after the detection target sound to be registered, which is recorded by the microphone 201. For this reason, a sound section including significant sound (detection target sound) to be actually registered is detected in the sound section detecting process. Fig. 3 shows an example of a sound section and noise sections which are present before and after the sound section. In the sound feature extracting process, a feature value which is useful for detecting the detection target sound is extracted from the time signal  $f(t)$  of the sound section which is obtained by the microphone 201 and registered in the feature value database 103 along with a detection target sound name.

[0042] The microphone 201 collects running state sound of a home electrical appliance, which is to be registered as detection target sound. The sound section detecting unit

202 obtains the time signal  $f(t)$ , which is output from the microphone 201, as an input and detects a sound section, namely a section of the running state sound generated by the home electrical appliance from the time signal  $f(t)$ . The feature value extracting unit 203 obtains the time signal  $f(t)$ , which is output from the microphone 201, as an input and extracts a feature value per every predetermined time from the time signal  $f(t)$ .

[0043] The feature value extracting unit 203 performs time frequency transform on the input time signal  $f(t)$  for every time frame, obtains time frequency distribution, obtains tone likelihood distribution from the time frequency distribution, smooths the likelihood distribution in a frequency direction and a time direction, and extracts a feature value per every predetermined time. In such a case, the feature value extracting unit 203 extracts the feature value in a range of a sound section based on sound section information supplied from the sound section detecting unit 202 and obtains a feature value sequence corresponding to a section of the operation condition sound generated by the home electrical appliance.

[0044] The feature value registration unit 204 associates and registers the feature value sequence corresponding to the running state sound generated by the home electrical appliance as a detection target sound, which has been obtained by the feature value extracting unit 203, with the detection target sound name (information on the running state sound) in the feature value database 103. In the example shown in the drawing, a state in which a feature value sequence including I detection target sound items  $Z1(m)$ ,  $Z2(m)$ , ...,  $Zi(m)$ , ...,  $ZI(m)$  are registered in the feature value database 103 is illustrated.

[0045] "Sound Section Detecting Unit"

Fig. 4 shows a configuration example of the sound section detecting unit 202. An input to the sound section detecting unit 202 is the time signal  $f(t)$  which is obtained by the microphone 201 recording the detection target sound to be registered (the running state sound generated by the home electrical appliance), and noise sections are also included before and after the detection target signal as shown in Fig. 3. In addition, an output from the sound detecting unit 202 is sound section information indicating a sound section including significant sound to be actually registered (detection target sound).

[0046] The sound section detecting unit 202 includes a time frequency transform unit 221, an amplitude feature value calculating unit 222, a tone intensity feature value calculating unit 223, a spectrum approximate outline feature value calculating unit 224, a score calculating unit 225, a time smoothing unit 226, and a threshold value determination unit 227.

[0047] The time frequency transform unit 221 performs time frequency transform on the

input time signal  $f(t)$  and obtains a time frequency signal  $F(n, k)$ . Here,  $t$  represents discrete time,  $n$  represents a number of a time frame, and  $k$  represents a discrete frequency. The time frequency transform unit 221 performs time frequency transform on the input time signal  $f(t)$  by short-time Fourier transform and obtains the time frequency signal  $F(n, k)$  as shown in the following Equation (1).

[Math.1]

$$F(n, k) = \log \left| \sum_{t=0}^{M-1} W(t) f(t - nR) e^{j2\pi kn} \right| \quad \dots (1)$$

Here,  $W(t)$  represents a window function,  $M$  represents a size of the window function, and  $R$  represents a frame time interval (= hop size). The time frequency signal  $F(n, k)$  represents a logarithmic amplitude value of a frequency component in a time frame  $n$  and at a frequency  $k$  and is a so-called spectrogram (time frequency distribution).

[0048] The amplitude feature value calculating unit 222 calculates an amplitude feature value  $x0(n)$  and  $x1(n)$  from the time frequency signal  $F(n, k)$ . Specifically, the amplitude feature value calculating unit 222 obtains an average amplitude  $A_{ave}(n)$  of a time section (with a length  $L$  before and after the target frame  $n$ ) in the vicinity of a target frame  $n$  for a predetermined frequency range (with a lower limit  $K_L$  and an upper limit  $K_H$ ), which is represented by the following Equation (2).

[Math.2]

$$A_{ave}(n) = \frac{1}{2L+1} \sum_{n=-L}^L \sum_{k=K_L}^{K_H} \exp(F(n, k)) \quad \dots (2)$$

[0049] In addition, the amplitude feature value calculating unit 222 obtains an absolute amplitude  $A_{abs}(n)$  in the target frame  $n$  for the predetermined frequency range (with a lower limit  $K_L$  and an upper limit  $K_H$ ), which is represented by the following Equation (3).

[Math.3]

$$A_{abs}(n) = \sum_{k=K_L}^{K_H} \exp(F(n, k)) \quad \dots (3)$$

[0050] Furthermore, the amplitude feature value calculating unit 222 obtains a relative amplitude  $A_{rel}(n)$  in the target frame  $n$  for the predetermined frequency range (with a lower limit  $K_L$  and an upper limit  $K_H$ ), which is represented by the following Equation (4).

[Math.4]

$$A_{\text{rel}}(n) = \frac{A_{\text{abs}}(n)}{A_{\text{ave}}(n)} \quad \dots (4)$$

[0051] In addition, the amplitude feature value calculating unit 222 regards the absolute amplitude  $A_{\text{abs}}(n)$  as an amplitude feature value  $x_0(n)$  and regards the relative amplitude  $A_{\text{rel}}(n)$  as an amplitude feature value  $x_1(n)$  as shown in the following Equation (5).

[Math.5]

$$x_0(n) = A_{\text{abs}}(n), \quad x_1(n) = A_{\text{rel}}(n) \quad \dots (5)$$

[0052] The tone intensity feature value calculating unit 223 calculates tone intensity feature value  $x_2(n)$  from the time frequency signal  $F(n, k)$ . The tone intensity feature value calculating unit 223 firstly transforms distribution of the time frequency signal  $F(n, k)$  (see Fig. 5A) into distribution of scores  $S(n, k)$  of tone characteristic likeliness (see Fig. 5B). Each score  $S(n, k)$  is a score from 0 to 1 which represents how much the time frequency component is "likely a tone component" in respective time  $n$  of  $F(n, k)$  at each frequency  $k$ . Specifically, the score  $S(n, k)$  is close to 1 at a position at which  $F(n, k)$  forms a peak of the tone characteristic in the frequency direction and is close to 0 at other positions.

[0053] Fig. 6 shows a configuration example of the tone likelihood distribution detecting unit 230 which is included in the tone intensity feature value calculating unit 223 for obtaining the distribution of the scores  $S(n, k)$  of the tone characteristic likeliness. The tone likelihood distribution detecting unit 230 includes a peak detecting unit 231, a fitting unit 232, a feature value extracting unit 233, and a scoring unit 234.

[0054] The peak detecting unit 231 detects a peak in the frequency direction in each time frame of the spectrogram (distribution of the time frequency signal  $F(n, k)$ ). That is, the peak detecting unit 231 detects whether or not a certain position corresponds to a peak (maximum value) in the frequency direction in all frames at all frequencies for the spectrogram.

[0055] The detection regarding whether or not the  $F(n, k)$  corresponds to a peak is performed by checking whether or not the following Equation (6) is satisfied, for example. Although a method using three points is exemplified as a peak detecting method, a method using five points is also applicable.

$$F(n, k - 1) < F(n, k) \text{ and } F(n, k) > F(n, k + 1) \dots (6)$$

[0056] The fitting unit 232 fits a tone model in a region in the vicinity of each peak, which has been detected by the peak detecting unit 231, as follows. First, the fitting unit 232

performs coordinate transform into coordinates including a target peak as an origin and sets a nearby time frequency region as shown by the following Equation (7). Here, delta N represents a nearby region (three points, for example) in the time direction, and delta k represents a nearby region (two points, for example) in the frequency direction. [Math.6]

$$\Gamma = [-\Delta_N \leq n \leq \Delta_N] \times [-\Delta_K \leq k \leq \Delta_K] \quad \dots (7)$$

[0057] Next, the fitting unit 232 fits a tone model of a quadratic polynomial function as shown by the following Equation (8), for example, to the time frequency signal in the nearby region. In such a case, the fitting unit 232 performs the fitting based on square error minimum criterion between the time frequency distribution in the vicinity of the peak and the tone model, for example.

[Math.7]

$$Y(k, n) = ak^2 + bk + ckn + dn^2 + en + g \quad \dots (8)$$

[0058] That is, the fitting unit 232 performs fitting by obtaining a coefficient which minimizes a square error, as shown in the following Equation (9), in the nearby region of the time frequency signal and the polynomial function as shown in the following Equation (10).

[Math.8]

$$J(a, b, c, d, e, g) = \sum_{\Gamma} (Y(k, n) - F(k, n))^2 \quad \dots (9)$$

$$(\hat{a}, \hat{b}, \hat{c}, \hat{d}, \hat{e}, \hat{g}) = \arg \min J(a, b, c, d, e, g) \quad \dots (10)$$

[0059] The quadratic polynomial function has a characteristic that the quadratic polynomial function fits well (the error is small) in the vicinity of the spectrum peak of the tone characteristic and does not fit well (the error is large) in the vicinity of a spectrum peak of the noise characteristic. Figs. 7A and 7B are diagrams schematically showing the state. Fig. 7A schematically shows a spectrum near a peak of the tone characteristic in n-th frame, which is obtained by the aforementioned Equation (1).

[0060] Fig. 7B shows a state in which a quadratic function  $f_0(k)$  shown by the following Equation (11) is applied to the spectrum in Fig. 7A. Here,  $a$  represents a peak curvature,  $k_0$  represents a frequency of an actual peak, and  $g_0$  represents a logarithmic amplitude value at a position of the actual peak. The quadratic function fits well around the spectrum peak of the tone characteristic component while the quadratic

function tends to greatly deviate around the peak of the noise characteristic.

[Math.9]

$$f_0(k) = a(k - k_0)^2 + g_0 \quad \cdots (11)$$

[0061] Fig. 8A schematically shows variation in the peak of the tone characteristic in the time direction. Amplitude and a frequency of the peak of the tone characteristic change in the previous and subsequent time frames while the approximate outline thereof is maintained. Although a spectrum which is actually obtained is a discrete point, the spectra are represented as a curve for convenience. One-dotted chain line shows a previous frame, a solid line shows a present frame, and a dotted line shows a next frame.

[0062] In many cases, the tone characteristic component is temporally continuous to some extent and can be represented as shift of quadratic functions with substantially the same shapes though a frequency and time slightly change. The variation  $Y(k, n)$  is represented by the following Equation (12). Since the spectrum is represented as logarithmic amplitude, a variation in the amplitude corresponds to displacement of the spectrum in the vertical direction. This is why an amplitude variation term  $f_1(n)$  is added. Here,  $\beta$  is a change rate of the frequency, and  $f_1(n)$  is a time function which represents a variation in the amplitude at a peak position.

[Math.10]

$$Y(k, n) = f_0(k - \beta n) + f_1(n) \quad \cdots (12)$$

[0063] The variation  $Y(k, n)$  can be represented by the following Equation (13) if  $f_1(n)$  is approximated by the quadratic function in the time direction. Since  $a$ ,  $k_0$ ,  $\beta$ ,  $d_1$ ,  $e_1$ , and  $g_0$  are constant, Equation (13) is equivalent to the aforementioned Equation (8) by appropriately transforming variables.

[Math.11]

$$\begin{aligned} Y(k, n) &= a(k - k_0 - \beta n)^2 + g_0 + d_1 n^2 + e_1 n \\ &= ak^2 - 2ak_0k - 2ak_0\beta n + a\beta^2 n^2 + d_1 n^2 + 2ak_0\beta n + e_1 n + ak_0^2 + g_0 \end{aligned} \quad \cdots (13)$$

[0064] Fig. 8B schematically shows fitting in the small region gamma on the spectrogram. Since similar shapes gradually change over time around the peak of the tone characteristic, Equation (8) tends to be well adapted. In relation to the vicinity of the peak of the noise characteristic, however, the shape and the frequency of the peak vary, and

therefore, Equation (8) is not well adapted, that is, a large error occurs even if Equation (8) is optimally made to fit.

[0065] The aforementioned Equation (10) shows calculation for fitting in relation to all coefficients a, b, c, d, e, and g. However, fitting may be performed after some coefficients are fixed to constants in advance. In addition, fitting may be performed with two or more dimensional polynomial function.

[0066] Returning to Fig. 6, the feature value extracting unit 233 extracts feature values (x0, x1, x2, x3, x4, and x5) as shown by the following Equation (14) based on the fitting result (see the aforementioned Equation (10)) at each peak obtained by the fitting unit 232. Each feature value is a feature value representing a characteristic of a frequency component at each peak, and the feature value itself can be used for analyzing voice sound or music sound.

[Math.12]

$$\begin{array}{ll}
 \text{(Curvature of Peak)} & x_0 = \hat{a} \\
 \text{(Frequency of Peak)} & x_1 = -\frac{b}{2\hat{a}} \\
 \text{(Logarithmic Amplitude Value of Peak)} & x_2 = \hat{g} \\
 \text{(Change Rate of Frequency)} & x_3 = -\frac{\hat{c}}{2\hat{a}} \\
 \text{(Change Rate of Amplitude)} & x_4 = \hat{e} \\
 \text{(Normalization Error in Fitting)} & x_5 = \frac{J(\hat{a}, \hat{b}, \hat{c}, \hat{d}, \hat{e}, \hat{g})}{\sum_{\Gamma} (F(n, k) - \hat{g})^2}
 \end{array} \quad \dots (14)$$

[0067] The scoring unit 234 obtains the score S(n, k) which represents the tone component likeliness of each peak by using the feature values extracted by the feature value extracting unit 233 for each peak, in order to quantize the tone component likeliness of each peak. The scoring unit 234 obtains the score S(n, k) as shown by the following Equation (15) by using one or a plurality of feature values from among the feature values (x0, x1, x2, x3, x4, and x5). In such a case, at least the normalization error x5 in fitting or the curvature of the peak in the frequency direction x0 is used.

[Math.13]

$$S(n, k) = \text{Sigm} \left( \sum_{i=0}^5 w_i H_i(x_i) + w_6 \right) \quad \dots (15)$$



[0068] Here,  $\text{Sigm}(x)$  is a sigmoid function,  $w_i$  is a predetermined load coefficient, and  $H_i(x_i)$  is a predetermined non-linear function for the  $i$ -th feature value  $x_i$ . It is possible to use a function as shown by the following Equation (16), for example, as the non-linear function  $H_i(x_i)$ . Here,  $u_i$  and  $v_i$  are predetermined load coefficients. Appropriate constant may be determined as  $w_i$ ,  $u_i$ , and  $v_i$  in advance, which can be automatically selected by steepest descent learning using multiple data items, for example.  
[Math.14]

$$H_i(x_i) = \text{Sigm}(u_i x_i + v_i) \quad \dots (16)$$

[0069] The scoring unit 234 obtains the score  $S(n, k)$  which represents the tone component likeliness for each peak by Equation (15) as described above. In addition, the scoring unit 234 sets the score  $S(n, k)$  at a position  $(n, k)$  other than the peak to 0. The scoring unit 234 obtains the score  $S(n, k)$  of the tone component likeliness, which is a value from 0 to 1, at each time and at each frequency of the time frequency signal  $f(n, k)$ .

[0070] The flowchart in Fig. 9 shows an example of a processing procedure for tone likelihood distribution detection by the tone likelihood distribution detecting unit 230. The tone likelihood distribution detecting unit 230 starts the processing in Step ST1 and then moves on to the processing in Step ST2. In Step ST2, the tone likelihood distribution detecting unit 230 sets a number  $n$  of a frame (time frame) to 0.

[0071] Next, the tone likelihood distribution detecting unit 230 determines whether or not  $n < N$  is satisfied in Step ST3. In addition, the frames of the spectrogram (time frequency distribution) are present from 0 to  $N - 1$ . If  $n < N$  is not satisfied, the tone likelihood distribution detecting unit 230 determines that the processing for all frames has been completed, and completes the processing in Step ST4.

[0072] If  $n < N$  is satisfied, the tone likelihood distribution detecting unit 230 sets a discrete frequency  $k$  to 0 in Step ST5. Then, the tone likelihood distribution detecting unit 230 determines whether or not  $k < K$  is satisfied in Step ST6. In addition, the discrete frequencies  $k$  of the spectrogram (time frequency distribution) are present from 0 to  $k - 1$ . If  $k < K$  is not satisfied, the tone likelihood distribution detecting unit 230 determines that the processing for all discrete frequencies has been completed, increments  $n$  in Step ST7, then returns to Step ST3, and moves on to the processing on the next frame.

[0073] If  $k < K$  is satisfied in Step ST6, the tone likelihood distribution detecting unit 230 determines whether or not  $F(n, k)$  corresponds to the peak in Step ST8. If  $F(n, k)$  does not correspond to the peak, the tone likelihood distribution detecting unit 230 sets the score  $S(n, k)$  to 0 in Step ST9, increments  $k$  in Step ST10, then returns to Step ST6, and moves on to the processing on the next discrete frequency.

- [0074] If  $F(n, k)$  corresponds to the peak in Step ST8, the tone likelihood distribution detecting unit 230 moves on to the processing in Step ST 11. In Step ST11, the tone likelihood distribution detecting unit 230 fits the tone model in a region in the vicinity of the peak. Then, the tone likelihood distribution detecting unit 230 extracts various feature values ( $x_0, x_1, x_2, x_3, x_4$ , and  $x_5$ ) based on the fitting result in Step ST12.
- [0075] Next, in Step ST13, the tone likelihood distribution detecting unit 230 obtains the score  $S(n, k)$ , which is a value from 0 to 1 representing the tone component likeliness of the peak, by using the feature values extracted in Step ST12. The tone likelihood distribution detecting unit 230 increments  $k$  in Step ST10 after the processing in Step ST14, then returns to Step ST6, and moves on to the processing on the next discrete frequency.
- [0076] Fig. 10 shows an example of distribution of the scores  $S(n, k)$  of the tone component likeliness obtained by the tone likelihood distribution detecting unit 230, which is shown in Fig. 6, from the time frequency distribution (spectrogram)  $F(n, k)$  as shown in Fig. 11. A larger value of the score  $S(n, k)$  is shown by a darker black color, and it can be observed that the peaks of the noise characteristic are not substantially detected while the peaks of the tone characteristic component (the component forming black thick horizontal lines in Fig. 11) are substantially detected.
- [0077] Returning to Fig. 4, the tone intensity feature value calculating unit 223 subsequently creates a tone component extracting filter  $H(n, k)$  (see Fig. 5C) which extracts only the component at a frequency position near a position at which the score  $S(n, k)$  is greater than a predetermined threshold value  $S_{thsd}$  (see Fig. 5B). The following Equation (17) represents the tone component extracting filter  $H(n, k)$ .
- [Math.15]

$$H(n, k) = \begin{cases} 1 & k_T - \Delta k < k < k_T + \Delta k \\ 0 & \text{otherwise} \end{cases} \quad \dots (17)$$

- [0078] However,  $k_T$  represents a frequency at which the tone component is detected, and  $\Delta k$  represents a predetermined frequency width. Here,  $\Delta k$  is preferably  $2/M$  when the size of the window function  $W(t)$  in the short-time Fourier transform (see Equation (1)) in order to obtain the time frequency signal  $F(n, k)$  as described above is  $M$ .
- [0079] The tone intensity feature value calculating unit 223 subsequently multiplies the original time frequency signal  $F(n, k)$  by the tone component extracting filter  $H(n, k)$  and obtains a spectrum (tone component spectrum)  $F_T(n, k)$  obtained by causing only the tone component to be left as shown in Fig. 5D. The following Equation (18)

represents the tone component spectrum  $F_T(n, k)$ .

[Math.16]

$$F_T(n, k) = H(n, k)F(n, k) \quad \cdots (18)$$

[0080] The tone intensity feature value calculating unit 223 finally sums up in a predetermined frequency region (with a lower limit  $K_L$  and an upper limit  $K_H$ ) and obtains tone component intensity  $A_{\text{tone}}(n)$  in the target frame  $n$ , which is represented by the following Equation (19).

[Math.17]

$$A_{\text{tone}}(n) = \sum_{k=K_L}^{K_H} \exp(F_T(n, k)) \quad \cdots (19)$$

[0081] Then, the tone intensity feature value calculating unit 223 regards the tone component intensity  $A_{\text{tone}}(n)$  as the tone intensity feature value  $x_2(n)$  as shown by the following Equation (20).

[Math.18]

$$x_2(n) = A_{\text{tone}}(n) \quad \cdots (20)$$

[0082] The spectrum approximate outline feature value calculating unit 224 obtains the spectrum approximate outline feature values  $x_3(n)$ ,  $x_4(n)$ ,  $x_5(n)$ , and  $x_6(n)$  as shown by the following Equation (21). Here,  $L$  represents a dimensional number of the feature value, and a case of  $L = 7$  is shown herein.

[Math.19]

$$\begin{aligned}
 x_3(n) &= \sum_{k=0}^{N/2-1} F(k, n) \cos(2\pi k / N) \\
 x_4(n) &= \sum_{k=0}^{N/2-1} F(k, n) \cos(4\pi k / N) \\
 x_5(n) &= \sum_{k=0}^{N/2-1} F(k, n) \cos(6\pi k / N) \\
 x_6(n) &= \sum_{k=0}^{N/2-1} F(k, n) \cos(8\pi k / N)
 \end{aligned}
 \quad \left. \vphantom{\begin{aligned} x_3(n) \\ x_4(n) \\ x_5(n) \\ x_6(n) \end{aligned}} \right\} \dots (21)$$

- [0083] The spectrum approximate outline feature value is a low-dimensional cepstrum obtained by developing a logarithm spectrum by discrete cosine transform. The above description was given of four or less dimensional coefficients, higher dimensional coefficients may be also used. Moreover, coefficients which are obtained by distorting a frequency axis and performing discrete cosine transform thereon, such as so-called MFCC (Mel-Frequency Cepstral Coefficients) may be also used.
- [0084] The aforementioned amplitude feature values  $x_0(n)$  and  $x_1(n)$ , the tone intensity feature value  $x_2(n)$ , and the spectrum approximate outline feature values  $x_3(n)$ ,  $x_4(n)$ ,  $x_5(n)$ , and  $x_6(n)$  configures L-dimensional (seven-dimensional in this case) feature value vector  $x(n)$  in the frame  $n$ . In addition, "volume of sound, a pitch of sound, and a tone of sound" are three sound factors, which are basic attributes indicating characteristics of the sound. Since the feature value vector  $x(n)$  is configured by amplitude (relating to volume of sound), tone component intensity (relating to a pitch of sound), and a spectrum approximate outline (relating to a tone of sound), the feature value vector  $x(n)$  configures a feature value relating to all the three sound factors.
- [0085] The score calculating unit 225 synthesizes the factors of the feature value vector  $x(n)$  and represents whether or not the frame  $n$  is a sound section including significant sound to be actually registered (detection target sound) by a score  $S(n)$  from 0 to 1. This can be obtained by the following Equation (22), for example. Here,  $\text{sigm}()$  is a sigmoid function,  $u_i$ ,  $v_i$ , and  $w_i$  ( $i = 0, \dots, L - 1$ ) are constants which are selected from sample data based on experiences.

[Math.20]

$$S(n) = \text{Sigm}\left(\sum_{i=0}^{L-1} w_i \xi_i(x_i(n)) + w_L\right)$$

$$\xi_i(x_i) = \text{Sigm}(u_i x_i(n) + v_i) \quad \dots (22)$$

[0086] The time smoothing unit 226 smooths the score  $S(n)$ , which has been obtained by the score calculating unit 225, in the time direction. In the smoothing processing, a moving average may be simply obtained, or a filter for obtaining a middle value such as a median filter may be used. The following Equation (23) shows an example in which the smoothed score  $S_a(n)$  is obtained by averaging processing. Here,  $\Delta n$  represents a size of the filter, which is a constant determined based on experiences.

[Math.21]

$$S_a(n) = \frac{1}{2\Delta n + 1} \sum_{\tau=n-\Delta n}^{n+\Delta n} S(n) \quad \dots (23)$$

[0087] The threshold value determination unit 227 compares the smoothed score  $S_a(n)$  in each frame  $n$ , which has been obtained by the time smoothing unit 226, with a threshold value, determines a frame section including a smoothed score  $S_a(n)$  which is equal to or greater than the threshold value as a sound section, and outputs sound section information indicating the frame section.

[0088] A description will be given of operations of the sound section detecting unit 202 shown in Fig. 4. The time signal  $f(t)$  which is obtained by collecting detection target sound to be registered (running state sound generated by a home electrical appliance) by a microphone 201 is supplied to the time frequency transform unit 221. The time frequency transform unit 221 performs time frequency transform on the input time signal  $f(t)$  and obtains the time frequency signal  $F(n, k)$ . The time frequency signal  $F(n, k)$  is supplied to the amplitude feature value calculating unit 222, the tone intensity feature value calculating unit 223, and the spectrum approximate outline feature value calculating unit 224.

[0089] The amplitude feature value calculating unit 222 calculates the amplitude feature value  $x_0(n)$  and  $x_1(n)$  from the time frequency signal  $F(n, k)$  (see Equation (5)). In addition, the tone intensity feature value calculating unit 223 calculates the tone intensity feature value  $x_2(n)$  from the time frequency signal  $F(n, k)$  (see Equation (20)). Furthermore, the spectrum approximate outline feature value calculating unit 224 calculates the spectrum approximate outline feature values  $x_3(n)$ ,  $x_4(n)$ ,  $x_5(n)$ , and

$x_6(n)$  (see Equation (21)).

- [0090] The amplitude feature values  $x_0(n)$  and  $x_1(n)$ , the tone intensity feature value  $x_2(n)$ , and the spectrum approximate outline feature value  $x_3(n)$ ,  $x_4(n)$ ,  $x_5(n)$ , and  $x_6(n)$  are supplied to the score calculating unit 225 as an L-dimensional (seven-dimensional in this case) feature value vector  $x(n)$  in the frame  $n$ . The score calculating unit 225 synthesizes the factors of the feature value vector  $x(n)$  and calculates a score  $S(n)$  from 0 to 1, which expresses whether or not the frame  $n$  is a sound section including significant sound to be actually registered (detection target sound) (see Equation (22)). The score  $S(n)$  is supplied to the time smoothing unit 226.
- [0091] The time smoothing unit 226 smooths the score  $S(n)$  in the time direction (see Equation (23)), and the smoothed score  $S_a(n)$  is supplied to the threshold value determination unit 227. The threshold value determination unit 227 compares the smoothed score  $S_a(n)$  in each frame  $n$  with the threshold value, determines a frame section including a smoothed score  $S_a$  which is equal to or greater than the threshold value as a sound section, and outputs sound section information indicating the frame section.
- [0092] The sound section detecting unit 202 shown in Fig. 4 extracts the feature values of amplitude, tone component intensity, and a spectrum approximate outline in each time frame from the time frequency distribution  $F(n, k)$  of the input time signal  $f(t)$  and obtains a score  $S(n)$  representing sound section likeliness of each time frame from the feature values. For this reason, it is possible to precisely obtain the sound section information which indicates the section of the detected sound even if the detected sound to be registered is recorded under a noise environment.
- [0093] "Feature Value Extracting Unit"
- Fig. 12 shows a configuration example of the feature value extracting unit 203. The feature value extracting unit 203 obtains as an input the time signal  $f(t)$  obtained by recording the detection target sound to be registered (the running state sound generated by the home electrical appliance) by a microphone 201 and, the time signal  $f(t)$  also includes noise sections before and after the detection target sound as shown in Fig. 3. In addition, the feature value extracting unit 203 outputs a feature value sequence extracted per every predetermined time in the section of the detection target sound to be registered.
- [0094] The feature value extracting unit 203 includes a time frequency transform unit 241, a tone likelihood distribution detecting unit 242, a time frequency smoothing unit 243, and a thinning and quantizing unit 244. The time frequency transform unit 241 performs time frequency transform on the input time signal  $f(t)$  and obtains the time frequency signal  $F(n, k)$  in the same manner as the aforementioned time frequency transform unit 221 of the sound section detecting unit 202. In addition, the feature value extracting unit 203 may use the time frequency signal  $F(n, k)$  obtained by the

time frequency transform unit 221 of the sound section detecting unit 202, and in such a case, it is not necessary to provide the time frequency transform unit 241.

[0095] The tone likelihood distribution detecting unit 242 detects tone likelihood distribution in the sound section based on the sound section information from the sound section detecting unit 202. That is, the tone likelihood distribution detecting unit 242 firstly transforms the distribution of the time frequency signals  $F(n, k)$  (see Fig. 5A) into distribution of scores  $S(n, k)$  of tone characteristic likeliness (see Fig. 5B) in the same manner as the aforementioned tone intensity feature value calculating unit 223 of the sound section detecting unit 202.

[0096] The tone likelihood distribution detecting unit 242 subsequently obtains tone likelihood distribution  $Y(n, k)$  in the sound section including significant sound to be registered (detection target sound) as shown by the following Equation (24) by using the sound section information.

[Math.22]

$$Y(n, k) = \begin{cases} S(n, k) & \text{When } n \text{ is inside sound section} \\ 0 & \text{When } n \text{ is outside sound section} \end{cases}$$

...(24)

[0097] The time frequency smoothing unit 243 smooths the tone likelihood distribution  $Y(n, k)$  in the sound section, which has been obtained by the tone likelihood distribution detecting unit 242, in the time direction and the frequency direction and obtains smoothed tone likelihood distribution  $Y_a(n, k)$  as shown by the following Equation (25).

[Math.23]

$$Y_a(n, k) = \sum_{\tau=-\Delta_n}^{\Delta_n} \sum_{\lambda=-\Delta_k}^{\Delta_k} Y(n-\tau, k-\lambda) H(\tau, \lambda) \quad \cdots (25)$$

[0098] Here, delta k represents a size of the smoothing filter on one side in the frequency direction, delta n represents a size thereof on one side in the time direction, and  $H(n, k)$  represents a quadratic impulse response of the smoothing filter. In addition, the above description was given of a case of a filter with no distortion in the frequency direction for simplification. However, the smoothing may be performed using a filter distorting a frequency axis, such as the Mel frequency.

[0099] The thinning and quantizing unit 344 thins out the smoothed tone likelihood distribution  $Y_a(n, k)$  obtained by the time frequency smoothing unit 243, further quantizes the tone likelihood distribution  $Y_a(n, k)$ , and create feature values  $Z(m, l)$  of the sig-

nificant sound to be registered (detection target sound) as shown by the following Equation (26).

[Math.24]

$$z(m, l) = \text{Quant} \left[ Y_a(mT, lK) \right] \quad (0 \leq m \leq M-1, 0 \leq l \leq L-1) \quad \dots (26)$$

[0100] Here, T represents a discretization step in the time direction, K represents a discretization step in the frequency direction, m represents thinned discrete time, and l represents a thinned discrete frequency. In addition, M represents a number of frames in the time direction (corresponding to time length of the significant sound to be registered (detection target sound)), L represents a number of dimensions in the frequency direction, and Quant[] represents a function of quantization.

[0101] The aforementioned feature values  $z(m, l)$  can be represented as  $Z(m)$  by collective vector notation in the frequency direction as shown by the following Equation (27).  
[Math.25]

$$Z(m) = [z(m, 0), \dots, z(m, L-1)] \quad (0 \leq m \leq M-1) \quad \dots (27)$$

[0102] In such a case, the aforementioned feature values  $Z(m, l)$  are configured by M vectors  $Z(0), \dots, Z(M-1), Z(M)$  which have been extracted per T in the time direction. Therefore, the thinning and quantizing unit 244 can obtain a sequence  $Z(m)$  of the feature values (vectors) extracted per every predetermined time in the section including the detecting target sound to be registered.

[0103] In addition, it can be also considered that the smoothed tone likelihood distribution  $Y_a(n, k)$  which has been obtained by the time frequency smoothing unit 243 is used as it is as an output from the feature value extracting unit 203, namely a feature value sequence. However, since the tone likelihood distribution  $Y_a(n, k)$  has been smoothed, it is not necessary to prepare all time and frequency data. It is possible to reduce an amount of information by thinning out in the time direction and the frequency direction. In addition, it is possible to transform data of 8 bits or 16 bits into data of 2 bits or 3 bits by quantization. Since thinning and quantization are performed as described above, it is possible to reduce the amount of information on the feature value (vector) sequence  $Z(m)$  and to thereby reduce processing burden for matching calculation by the sound detecting apparatus 100 which will be described later.

[0104] A description will be given of operations of the feature value extracting unit 203 shown in Fig. 12. The time signal  $f(t)$  obtained by collecting the detection target sound



(the running state sound generated by the home electrical appliance) to be registered by the microphone 201 is supplied to the time frequency transform unit 241. The time frequency transform unit 241 performs time frequency conversion on the input time signal  $f(t)$  and obtains the time frequency signal  $F(n, k)$ . The time frequency signal  $F(n, k)$  is supplied to the tone likelihood distribution detecting unit 242. In addition, the sound section information obtained by the sound section detecting unit 202 is also supplied to the tone likelihood distribution detecting unit 242.

- [0105] The tone likelihood distribution detecting unit 242 transforms distribution of the time frequency signals  $F(n, k)$  into distribution of scores  $S(n, k)$  of the tone characteristic likeliness, and further obtains the tone likelihood distribution  $Y(n, k)$  in the sound section including the significant sound to be registered (detection target sound) by using the sound section information (see Equation (24)). The tone likelihood distribution  $Y(n, k)$  is supplied to the time frequency smoothing unit 243.
- [0106] The time frequency smoothing unit 243 smooths the tone likelihood distribution  $Y(n, k)$  in the time direction and the frequency direction and obtains the smoothed tone likelihood distribution  $Y_a(n, k)$  (see Equation (25)). The tone likelihood distribution  $Y_a(n, k)$  is supplied to the thinning and quantizing unit 244. The thinning and quantizing unit 244 thins out the tone likelihood distribution  $Y_a(n, k)$ , further quantize the thinned tone likelihood distribution  $Y_a(n, k)$ , and obtains a feature values  $z(m, l)$  of the significant sound to be registered (detection target sound), namely the feature value sequence  $Z(m)$  (see Equations (26) and (27)).
- [0107] Returning to Fig. 2, the feature value registration unit 204 associates and registers the feature value sequence  $Z(m)$  of the detection target sound to be registered, which has been created by the feature value registration unit 204, with a detection target sound name (information on the operation condition sound) in the feature value database 103.
- [0108] A description will be given of operations of the feature value registration apparatus 200 shown in Fig. 2. The microphone 201 collects running state sound of a home electrical appliance to be registered as detection target sound. The time signal  $f(t)$  output from the microphone 201 is supplied to the sound section detecting unit 202 and the feature value extracting unit 203. The sound section detecting unit 202 detects the sound section, namely the section including the running state sound generated by the home electrical appliance, from the input time signal  $f(t)$  and outputs the sound section information. The sound section information is supplied to the feature value extracting unit 203.
- [0109] The feature value extracting unit 203 performs time frequency conversion on the input time signal  $f(t)$  for each time frame, obtains the distribution of the time frequency signals  $F(n, k)$ , and further obtains tone likelihood distribution, namely distribution of the scores  $S(n, k)$  from the time frequency distribution. Then, the feature value ex-

tracting unit 203 obtains the tone likelihood distribution  $Y(n, k)$  of the sound section from the distribution of the scores  $S(n, k)$  based on the sound section information, smooths the tone likelihood distribution  $Y(n, k)$  in the time direction and the frequency direction, and further performs thinning and quantizing processing thereon to create the feature value sequence  $Z(m)$ .

[0110] The feature value sequence  $Z(m)$  of the detection target sound to be registered (the running state sound of the home electrical appliance), which has been created by the feature value extracting unit 203, is supplied to the feature value registration unit 204. The feature value registration unit 204 associates and registers the feature value sequence  $Z(m)$  with the detection target sound name (information on the running state sound) in the feature value database 103. In the following description, it is assumed that  $I$  detection target sound items are registered, the feature value sequences thereof will be represented as  $Z1(m)$ ,  $Z2(m)$ , ...,  $Zi(m)$ , ...,  $ZI(m)$ , and the numbers of time frames in the feature value sequences (the number of vectors aligned in the time direction) will be represented as  $M1$ ,  $M2$ , ...,  $Mi$ , ...,  $MI$ .

[0111] "Sound Detecting Unit"

Fig. 13 shows a configuration example of the sound detecting unit 102. The sound detecting unit 102 includes a signal buffering unit 121, a feature value extracting unit 122, a feature value buffering unit 123, and a comparison unit 124. The signal buffering unit 121 buffers a predetermined number of signal samples of the time signal  $f(t)$  which is obtained by collecting sound by the microphone 101. The predetermined number means a number of samples with which the feature value extracting unit 122 can newly calculate a feature value sequence corresponding to one frame.

[0112] The feature value extracting unit 122 extracts feature values per every predetermined time based on the signal samples of the time signal  $f(t)$ , which has been buffered by the signal buffering unit 121. Although not described in detail, the feature value extracting unit 203 is configured in the same manner as the aforementioned feature value extracting unit 203 (see Fig. 12) of the feature value registration apparatus 200.

[0113] However, the tone likelihood detecting unit 242 in the feature value extracting unit 122 obtains the tone likelihood distribution  $Y(n, k)$  in all sections. That is, the tone likelihood distribution detecting unit 242 outputs the distribution of the scores  $S(n, k)$ , which has been obtained from the distribution of the time frequency signals  $F(n, k)$ , as it is. Then, the thinning and quantizing unit 244 outputs a newly extracted feature value (vector)  $X(n)$  per  $T$  (discretization step in the time direction) for all sections of the input time signal  $f(t)$ . Here,  $n$  represents a number of a frame of the feature value which is being currently extracted (corresponding to current discrete time).

[0114] The feature value buffering unit 123 saves the newest  $N$  feature values (vectors)  $X(n)$  output from the feature value extracting unit 122 as shown in Fig. 14. Here,  $N$  is at

least a number which is equal to or greater than a number of frames (the number of vectors aligned in the time direction) of the longest feature value sequences from among the feature value sequences  $Z1(m)$ ,  $Z2(m)$ , ...,  $Zi(m)$ , ...,  $ZI(m)$  registered (maintained) in the feature value database 103.

[0115] The comparison unit 124 sequentially compares the feature value sequences saved in the signal buffering unit 123 with feature value sequences of I detection target sound items registered in the feature value database 103 every time the feature value extracting unit 122 extracts the new feature value  $X(n)$ , and obtains detection results of the I detection target sound items. Here, if  $i$  represents the number of the detection target sound number, the length of each detection target sound item (frame number  $M_i$ ) differs from each other.

[0116] As shown in Fig. 14, the comparison unit 124 combines the latest frame  $n$  in the feature value buffering unit 123 with the last frame  $Zi(M_i - 1)$  of the feature value sequence of the detection target sound and calculates similarity by using a frame with a length of the feature value sequence of the detection target sound from among the  $N$  feature values saved in the feature value buffering unit 123. The similarity  $\text{Sim}(n, i)$  can be calculated by correlation between feature values as shown by the following Equation (28), for example. Here,  $\text{Sim}(n, i)$  means similarity with a feature value sequence of  $i$ -th detection target sound in the  $n$ -th frame. The comparison unit 124 determines that "the  $i$ -th detection target sound is generated at time  $n$ " and outputs the determination result when the similarity is greater than a predetermined threshold value. [Math.26]

$$\text{Sim}(n, i) = \frac{\sum_{m=0}^{M_i-1} X(n - M_i - 1 + m) Z_i(m)}{\sqrt{\sum_{m=0}^{M_i-1} X^2(n - M_i - 1 + m) \sum_{m=0}^{M_i-1} Z_i^2(m)}} \quad \dots (28)$$

[0117] A description will be given of operations of the sound detecting unit 102 shown in Fig. 13. The time signal  $f(t)$  obtained by collecting sound by the microphone 101 is supplied to the signal buffering unit 121, and the predetermined number of signal samples are buffered. The feature value extracting unit 122 extracts a feature value per very predetermined time based on the signal samples of the time signal  $f(t)$  buffered by the signal buffering unit 121. Then, the feature value extracting unit 122 sequentially outputs a newly extracted feature value (vector)  $X(n)$  per  $T$  (the discretization step in the time direction).

- [0118] The feature value  $X(n)$  which has been extracted by the feature value extracting unit 122 is supplied to the feature value buffering unit 123, and the latest  $N$  feature values  $X(n)$  are saved therein. The comparison unit 124 sequentially compares the feature value sequence saved in the signal buffering unit 123 with a feature value sequence of the  $I$  detection target sound items, which are registered in the feature value database 103, every time the new feature value  $X(n)$  is extracted by the feature value extracting unit 122, and obtains the detection result of the  $I$  detection target sound items.
- [0119] In such a case, the comparison unit 124 combines the latest frame  $n$  in the feature value buffering unit 123 with the last frame  $Z_i(M_i - 1)$  of the feature value sequence of the detection target sound and calculates similarity by using a frame with a length of the feature value sequence of the detection target sound (see Fig. 14). Then, the comparison unit 124 determines that "the  $i$ -th detection target sound is generated at time  $n$ " and outputs the determination result when the similarity is greater than the pre-determined threshold value.
- [0120] In addition, the sound detecting apparatus 100 shown in Fig. 1 can be configured as hardware or software. For example, it is possible to cause the computer apparatus 300 shown in Fig. 15 to include a part of or all the functions of the sound detecting apparatus 100 shown in Fig. 1 and performs the same processing of detecting detection target sound as that described above.
- [0121] The computer apparatus 300 includes a CPU (Central Processing Unit) 301, a ROM (Read Only Memory) 302, a RAM (Random Access Memory) 303, a data input and output unit (data I/O) 304, and an HDD (Hard Disk Drive) 305. The ROM 302 stores a processing program and the like of the CPU 301. The RAM 303 functions as a work area of the CPU 301. The CPU 301 reads the processing program stored on the ROM 302 as necessary, transfers to and develops in the RAM 303 the read processing program, reads the developed processing program, and executes tone component detecting processing.
- [0122] The input time signal  $f(t)$  is input to the computer apparatus 300 via the data I/O 304 and accumulated in the HDD 305. The CPU 301 performs the processing of detecting detection target sound on the input time signal  $f(t)$  accumulated in the HDD 305 as described above. Then, the detection result is output to the outside via the data I/O 304. In addition, a feature value sequence of  $I$  detection target sound items are registered and maintained in the HDD 305 in advance.
- [0123] The flowchart in Fig. 16 shows an example of a processing procedure for detecting the detection target sound by the CPU 301. In Step ST21, the CPU 301 starts the processing and then moves on to the processing in Step ST22. In Step ST22, the CPU 301 inputs the input time signal  $f(t)$  to the signal buffering unit configured in the HDD 305, for example. Then, the CPU 301 determines whether or not a number of samples

with which the feature value sequence corresponding to one frame can be calculated have been accumulated, in Step ST23.

- [0124] If the number of samples corresponding to one frame have been accumulated, the CPU 301 performs processing of extracting the feature value  $X(n)$  in Step ST24. The CPU 301 inputs the extracted feature value  $X(n)$  to the feature value buffering unit configured in the HDD 305, for example, in Step ST25. Then, the CPU 301 sets the number  $i$  of the detection target sound to zero in Step ST26.
- [0125] Next, the CPU 301 determines whether or not  $i < I$  is satisfied in Step ST27. If  $i < I$  is satisfied, the CPU 301 calculates similarity between the feature value sequence saved in the signal buffering unit and the feature value sequence  $Z_i(m)$  of the  $i$ -th detection target sound registered in the HDD 305 in Step ST28. Then, the CPU 301 determines whether or not the similarity  $>$  the threshold value is satisfied in Step ST29.
- [0126] If the similarity  $>$  the threshold value is satisfied, the CPU 301 outputs a result indicating coincidence in Step ST30. That is, a determination result that "the  $i$ -th detection target sound is generated at time  $n$ " is output as a detection output. Thereafter, the CPU 301 increments  $i$  in Step ST31 and returns to the processing in Step ST27. In addition, if the similarity  $>$  the threshold value is not satisfied in Step ST29, The CPU 301 immediately increments  $i$  in Step ST31 and returns to the processing in Step ST27. If  $i > I$  is not satisfied in Step ST27, the CPU 301 determines that the processing on the current frame has been completed, returns to the processing in Step ST22, and moves on to the processing on the next frame.
- [0127] Next, the CPU 301 sets the number  $n$  of the frame (time frame) to 0 in Step ST3. Then, the CPU 301 determines whether or not  $n < N$  is satisfied in Step ST4. In addition, it is assumed that the frames of the spectrogram (time frequency distribution) are present from 0 to  $N - 1$ . If  $n < N$  is not satisfied, the CPU 301 determines that the processing of all the frames has been completed and then completes the processing in Step ST5.
- [0128] If  $n < N$  is satisfied, the CPU 301 sets the discrete frequency  $k$  to 0 in Step ST6. Then, the CPU 301 determines whether or not  $k < K$  is satisfied in Step ST7. In addition, it is assumed that the discrete frequencies  $k$  of the spectrogram (time frequency distribution) are present from 0 to  $k-1$ . If  $k < K$  is not satisfied, the CPU 301 determines that the processing on all the discrete frequencies has been completed, increments  $n$  in Step ST8, then returns to Step ST4, and moves on to the processing on the next frame.
- [0129] If  $k < K$  is satisfied in Step ST7, the CPU 301 determines whether or not  $F(n, k)$  corresponds to a peak in Step ST9. If  $F(n, k)$  does not correspond to the peak, the CPU 301 sets the score  $S(n, k)$  to 0 in Step ST10, increments  $k$  in Step ST11, then returns to Step ST7, and moves on the processing on the next discrete frequency.

- [0130] If  $F(n, k)$  corresponds to the peak in Step ST9, the CPU 301 moves on to the processing in Step ST12. In Step ST12, the CPU 301 fits the tone model in the region in the vicinity of the peak. Then, the CPU 301 extracts various feature values ( $x_0, x_1, x_2, x_3, x_4$ , and  $x_5$ ) based on the fitting result in Step ST13.
- [0131] Next, in Step ST14, the CPU 301 obtains a score  $S(n, k)$ , which represents a tone component likelihood of the peak with a value from 0 to 1, by using the feature values extracted in Step ST13. The CPU 301 increments  $k$  in Step ST11 after the processing in Step ST14, then returns to Step ST7, and moves on to the processing on the next discrete frequency.
- [0132] As described above, the sound detecting apparatus 100 shown in Fig. 1 obtains the tone likelihood distribution from the time frequency distribution of the input time signal  $f(t)$  obtained by collecting sound by the microphone 101 and extracts and uses the feature value per every predetermined time from the likelihood distribution which has been smoothed in the frequency direction and the time direction. Accordingly, it is possible to precisely detect the detection target sound (running state sound and the like generated from a home electrical appliance) without depending on an installation position of the microphone 101.
- [0133] In addition, the sound detecting apparatus 100 shown in Fig. 1 records on a recording medium and displays on a display the detection result of the detection target sound, which has been obtained by the sound detecting unit 102, along with time. Accordingly, it is possible to automatically record running states of home electrical appliances and the like at home and obtains a self action history (so-called life log). In addition, it is possible to automatically visualize sound notification for people with hearing difficulties.
- [0134] <2. Modified Example>
- The above embodiment shows an example in which running state sound generated from a home electrical appliance (control sounds, notification sounds, operating sounds, alarm sounds, and the like) at home is detected. However, the present technology can be applied to use in automating detection relating to sound functions of a product fabricated in a production plant as well as domestic use. In addition, it is a matter of fact that the present technology can be applied not only to detection of running state sound but also to detection of voice sound of a specific person or a specific animal or other environmental sound.
- [0135] Although the above description was given of the embodiment in which the time frequency transform was performed based on the short-time Fourier transform, it can be also considered that the input time signal is subjected to the time frequency transform by using another transform method such as wavelet transform. In addition, although the above description was given of the embodiment in which the fitting was

performed based on the square error minimum criterion between the time frequency distribution in the vicinity of each detected peak and the tone model, it can be also considered that the fitting is performed based on a quadruplicate error minimum criterion, a minimum entropy criterion, or the like.

[0136] In addition, the present technology can be configured as follows.

(1) A sound detecting apparatus including: a feature value extracting unit which extracts a feature value per every predetermined time from an input time signal; a feature value maintaining unit which maintains a feature value sequence of a predetermined number of detection target sound items; and a comparison unit which respectively compares a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtains detection results of the predetermined number of detection target sound items every time the feature value extracting unit newly extracts a feature value, wherein the feature value extracting unit includes a time frequency transform unit which performs time frequency transform on the input time signal for each time frame and obtains time frequency distribution and a likelihood distribution detecting unit which obtains tone likelihood distribution from the time frequency distribution, smooths the obtained likelihood distribution in a frequency direction and a time direction, and extracts the feature value per the predetermined time.

(2) The apparatus according to (1), wherein the likelihood distribution detecting unit includes a peak detecting unit which detects a peak in the frequency direction in each time frame of the time frequency distribution, a fitting unit which fits the tone model at each detected peak, and a scoring unit which obtains a score representing tone component likeliness at each detected peak based on the fitting result.

(3) The apparatus according to (1) or (2), wherein the feature value extracting unit further includes a thinning unit which thins out the smoothed likelihood distribution in the frequency direction and/or the time direction.

(4) The apparatus according to (1) or (2), wherein the feature value extracting unit further includes a quantizing unit which quantizes the smoothed likelihood distribution.

(5) The apparatus according to any one of (1) to (4), wherein the comparison unit obtains similarity based on correlation between corresponding feature values between the feature value sequence of the maintained detection target sound items and the feature value sequence extracted by the feature value extracting unit for each of the predetermined number of detection target sound items and obtains the detection results of the detection target sound items based on the obtained similarity.

(6) The apparatus according to any one of (1) to (5), further including:

a recording control unit which records the detection results of the predetermined number of detection target sound items along with time information on a recording medium.

(7) A sound detecting method including: extracting a feature value per every predetermined time from an input time signal; and respectively comparing a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtaining detection results of the predetermined number of detection target sound items every time the feature value is newly extracted in the extracting of the feature value, wherein in the extracting of the feature value, time frequency transform is performed on the input time signal for each time frame, time frequency distribution is obtained, tone likelihood distribution is obtained from the time frequency distribution, the likelihood distribution is smoothed in a frequency direction and a time direction, and the feature value per the predetermined time is extracted.

(8) A program which causes a computer to perform: extracting a feature value per every predetermined time from an input time signal; and respectively comparing a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtaining detection results of the predetermined number of detection target sound items every time the feature value is newly extracted in the extracting of the feature value, wherein in the extracting of the feature value, time frequency transform is performed on the input time signal for each time frame, time frequency distribution is obtained, tone likelihood distribution is obtained from the time frequency distribution, the likelihood distribution is smoothed in a frequency direction and a time direction, and the feature value per the predetermined time is extracted.

(9) A sound feature value extracting apparatus including: a time frequency transform unit which performs time frequency transform on an input time signal for each time frame and obtains time frequency distribution; a likelihood distribution detecting unit which obtains tone likelihood distribution from the time frequency distribution; and a feature value extracting unit which smooths the likelihood distribution in a frequency direction and a time direction and extracts a feature value per every predetermined time.

(10) The apparatus according to (9), wherein the likelihood distribution detecting unit includes a peak detecting unit which detects a peak in the frequency direction in each time frame of the time frequency distribution, a fitting unit which fits the tone model at each detected peak, and a scoring unit which obtains a score representing tone component likeliness at each detected peak based on the fitting result.

(11) The apparatus according to (9) or (10), further including: a thinning unit which



thins out the smoothed likelihood distribution in the frequency direction and/or the time direction.

(12) The apparatus according to (9) or (10), further including: a quantizing unit which quantizes the smoothed likelihood distribution.

(13) The apparatus according to any one of (9) to (12), further including: a sound section detecting unit which detects a sound section based on the input time signal, wherein the likelihood distribution detecting unit obtains tone likelihood distribution from the time frequency distribution within a range of the detected sound section.

(14) The apparatus according to (13), wherein the sound section detecting unit includes a time frequency transform unit which performs time frequency transform on the input time signal for each time frame and obtains time frequency distribution, a feature value extracting unit which extracts feature value of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution, a scoring unit which obtains a score representing a sound section likeliness for each time frame based on the extracted feature values, a time smoothing unit which smooths the obtained score for each time frame in the time direction, and a threshold value determination unit which determinates a threshold value for the smoothed score for each time frame and obtains sound section information.

(15) A sound feature value extracting method including: obtaining time frequency distribution by performing time frequency transform on an input time signal for each time frame;

obtaining tone likelihood distribution from the time frequency distribution; and smoothing the likelihood distribution in a frequency direction and a time direction.

(16) A sound section detecting apparatus including: a time frequency transform unit which obtains time frequency distribution by performing time frequency transform on an input time signal for each time frame; a feature value extracting unit which extracts feature values of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution; and a scoring unit which obtains a score representing sound section likeliness for each time frame based on the extracted feature values.

(17) The apparatus according to (16), further including: a time smoothing unit which smooths the obtained score for each time frame in the time direction; and a threshold value determination unit which determines a threshold for the smoothed score for each time frame and obtains sound section information.

(18) A sound section detecting method including: obtaining time frequency distribution by performing time frequency transform on an input time signal for each time frame; extracting feature values of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution; and

obtaining a score representing sound section likeliness for each time frame based on the extracted feature values.

[0137] The present disclosure contains subject matter related to that disclosed in Japanese Priority Patent Application JP 2012-094395 filed in the Japan Patent Office on April 18, 2012, the entire contents of which are hereby incorporated by reference.

[0138] It should be understood by those skilled in the art that various modifications, combinations, sub-combinations and alterations may occur depending on design requirements and other factors insofar as they are within the scope of the appended claims or the equivalents thereof.

### **Reference Signs List**

[0139] 100: sound detecting apparatus  
101: microphone  
102: sound detecting unit  
103: feature value database  
104: recording and displaying unit  
121: signal buffering unit  
122: feature value extracting unit  
123: feature value buffering unit  
124: comparison unit  
200: feature value registration apparatus  
201: microphone  
202: sound section detecting unit  
203: feature value extracting unit  
204: feature value registration unit  
221: time frequency transform unit  
222: amplitude feature value calculating unit  
223: tone intensity feature value calculating unit  
224: spectrum approximate outline feature value calculating unit  
225: score calculating unit  
226: time smoothing unit  
227: threshold value determination unit  
230: tone likelihood distribution detecting unit  
231: peak detecting unit  
232: fitting unit  
233: feature value extracting unit  
234: scoring unit  
241: time frequency transform unit

242: tone likelihood distribution detecting unit

243: time frequency transform unit

244: thinning and quantizing unit

## Claims

- [Claim 1] A sound detecting apparatus comprising:  
a feature value extracting unit which extracts a feature value per every predetermined time from an input time signal;  
a feature value maintaining unit which maintains a feature value sequence of a predetermined number of detection target sound items;  
and  
a comparison unit which respectively compares a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtains detection results of the predetermined number of detection target sound items every time the feature value extracting unit newly extracts a feature value,  
wherein the feature value extracting unit includes a time frequency transform unit which performs time frequency transform on the input time signal for each time frame and obtains time frequency distribution and a likelihood distribution detecting unit which obtains tone likelihood distribution from the time frequency distribution, smooths the obtained likelihood distribution in a frequency direction and a time direction, and extracts the feature value per the predetermined time.
- [Claim 2] The apparatus according to Claim 1, wherein the likelihood distribution detecting unit includes a peak detecting unit which detects a peak in the frequency direction in each time frame of the time frequency distribution, a fitting unit which fits the tone model at each detected peak, and a scoring unit which obtains a score representing tone component likeliness at each detected peak based on the fitting result.
- [Claim 3] The apparatus according to Claim 1, wherein the feature value extracting unit further includes a thinning unit which thins out the smoothed likelihood distribution in the frequency direction and/or the time direction.
- [Claim 4] The apparatus according to Claim 1, wherein the feature value extracting unit further includes a quantizing unit which quantizes the smoothed likelihood distribution.
- [Claim 5] The apparatus according to Claim 1, wherein the comparison unit obtains similarity based on correlation between corresponding feature values between the feature value sequence of the maintained detection target sound items and the feature value sequence extracted by the

feature value extracting unit for each of the predetermined number of detection target sound items and obtains the detection results of the detection target sound items based on the obtained similarity.

[Claim 6]

The apparatus according to Claim 1, further comprising:  
a recording control unit which records the detection results of the predetermined number of detection target sound items along with time information on a recording medium.

[Claim 7]

A sound detecting method comprising:  
extracting a feature value per every predetermined time from an input time signal; and  
respectively comparing a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtaining detection results of the predetermined number of detection target sound items every time the feature value is newly extracted in the extracting of the feature value,  
wherein in the extracting of the feature value, time frequency transform is performed on the input time signal for each time frame, time frequency distribution is obtained, tone likelihood distribution is obtained from the time frequency distribution, the likelihood distribution is smoothed in a frequency direction and a time direction, and the feature value per the predetermined time is extracted.

[Claim 8]

A program which causes a computer to perform:  
extracting a feature value per every predetermined time from an input time signal; and  
respectively comparing a feature value sequence extracted by the feature value extracting unit with a feature value sequence of the maintained predetermined number of detecting target sound items and obtaining detection results of the predetermined number of detection target sound items every time the feature value is newly extracted in the extracting of the feature value,  
wherein in the extracting of the feature value, time frequency transform is performed on the input time signal for each time frame, time frequency distribution is obtained, tone likelihood distribution is obtained from the time frequency distribution, the likelihood distribution is smoothed in a frequency direction and a time direction, and the feature value per the predetermined time is extracted.

[Claim 9]

A sound feature value extracting apparatus comprising:

a time frequency transform unit which performs time frequency transform on an input time signal for each time frame and obtains time frequency distribution;

a likelihood distribution detecting unit which obtains tone likelihood distribution from the time frequency distribution; and

a feature value extracting unit which smooths the likelihood distribution in a frequency direction and a time direction and extracts a feature value per every predetermined time.

[Claim 10] The apparatus according to Claim 9, wherein the likelihood distribution detecting unit includes a peak detecting unit which detects a peak in the frequency direction in each time frame of the time frequency distribution, a fitting unit which fits the tone model at each detected peak, and a scoring unit which obtains a score representing tone component likeliness at each detected peak based on the fitting result.

[Claim 11] The apparatus according to Claim 9, further comprising:  
a thinning unit which thins out the smoothed likelihood distribution in the frequency direction and/or the time direction.

[Claim 12] The apparatus according to Claim 9, further comprising:  
a quantizing unit which quantizes the smoothed likelihood distribution.

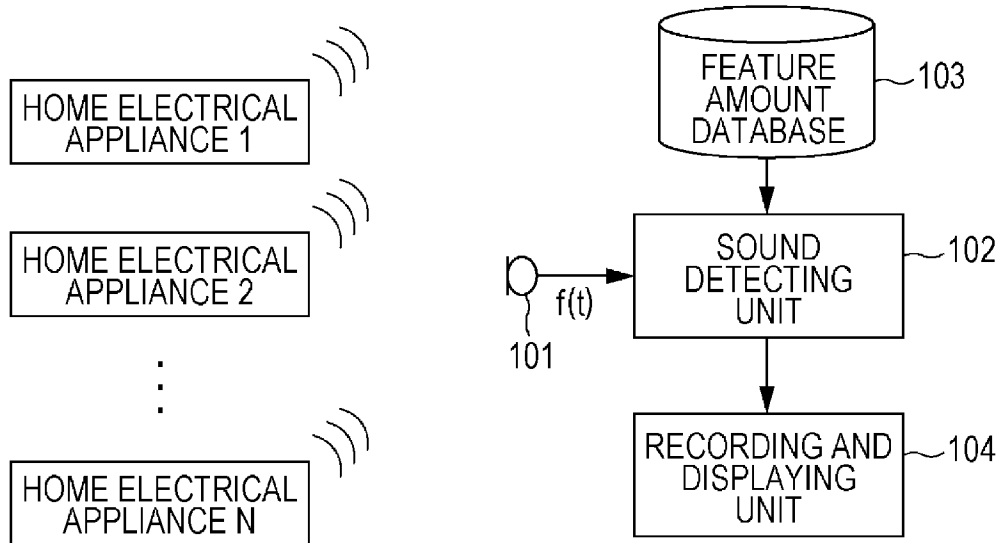
[Claim 13] The apparatus according to Claim 9, further comprising:  
a sound section detecting unit which detects a sound section based on the input time signal,  
wherein the likelihood distribution detecting unit obtains tone likelihood distribution from the time frequency distribution within a range of the detected sound section.

[Claim 14] The apparatus according to Claim 13, wherein the sound section detecting unit includes a time frequency transform unit which performs time frequency transform on the input time signal for each time frame and obtains time frequency distribution, a feature value extracting unit which extracts feature value of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution, a scoring unit which obtains a score representing a sound section likeliness for each time frame based on the extracted feature values, a time smoothing unit which smooths the obtained score for each time frame in the time direction, and a threshold value determination unit which determinates a threshold value for the smoothed score for each time frame and obtains sound section information.

- [Claim 15] A sound feature value extracting method comprising:  
obtaining time frequency distribution by performing time frequency transform on an input time signal for each time frame;  
obtaining tone likelihood distribution from the time frequency distribution; and  
smoothing the likelihood distribution in a frequency direction and a time direction.
- [Claim 16] A sound section detecting apparatus comprising:  
a time frequency transform unit which obtains time frequency distribution by performing time frequency transform on an input time signal for each time frame;  
a feature value extracting unit which extracts feature values of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution;  
and  
a scoring unit which obtains a score representing sound section likeliness for each time frame based on the extracted feature values.
- [Claim 17] The apparatus according to Claim 16, further comprising:  
a time smoothing unit which smooths the obtained score for each time frame in the time direction; and  
a threshold value determination unit which determines a threshold for the smoothed score for each time frame and obtains sound section information.
- [Claim 18] A sound section detecting method comprising:  
obtaining time frequency distribution by performing time frequency transform on an input time signal for each time frame;  
extracting feature values of amplitude, tone component intensity, and a spectrum approximate outline for each time frame based on the time frequency distribution; and  
obtaining a score representing sound section likeliness for each time frame based on the extracted feature values.

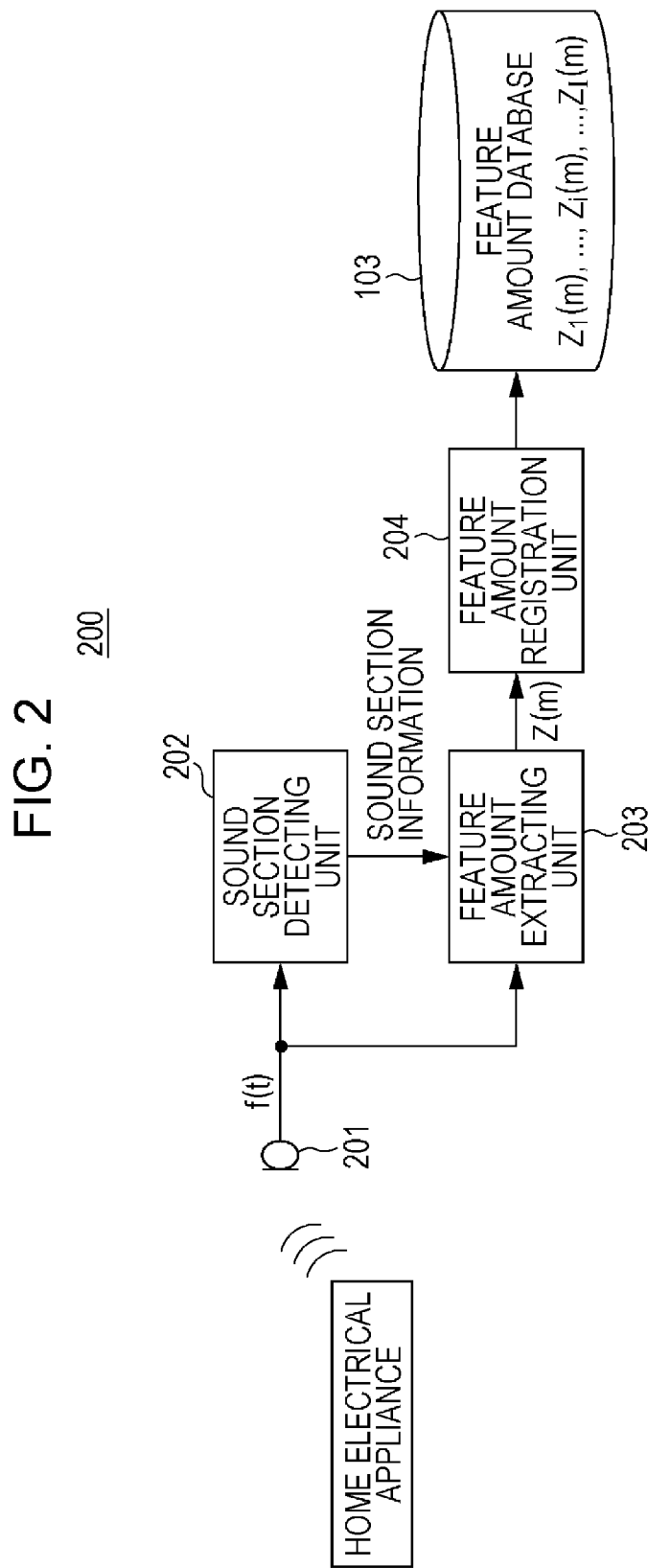
[Fig. 1]

FIG. 1

100

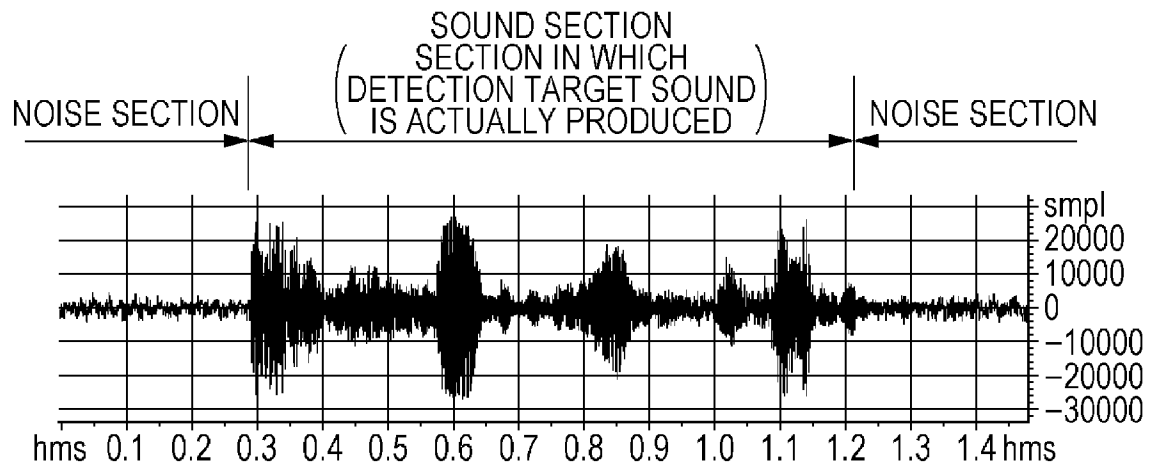


[Fig. 2]

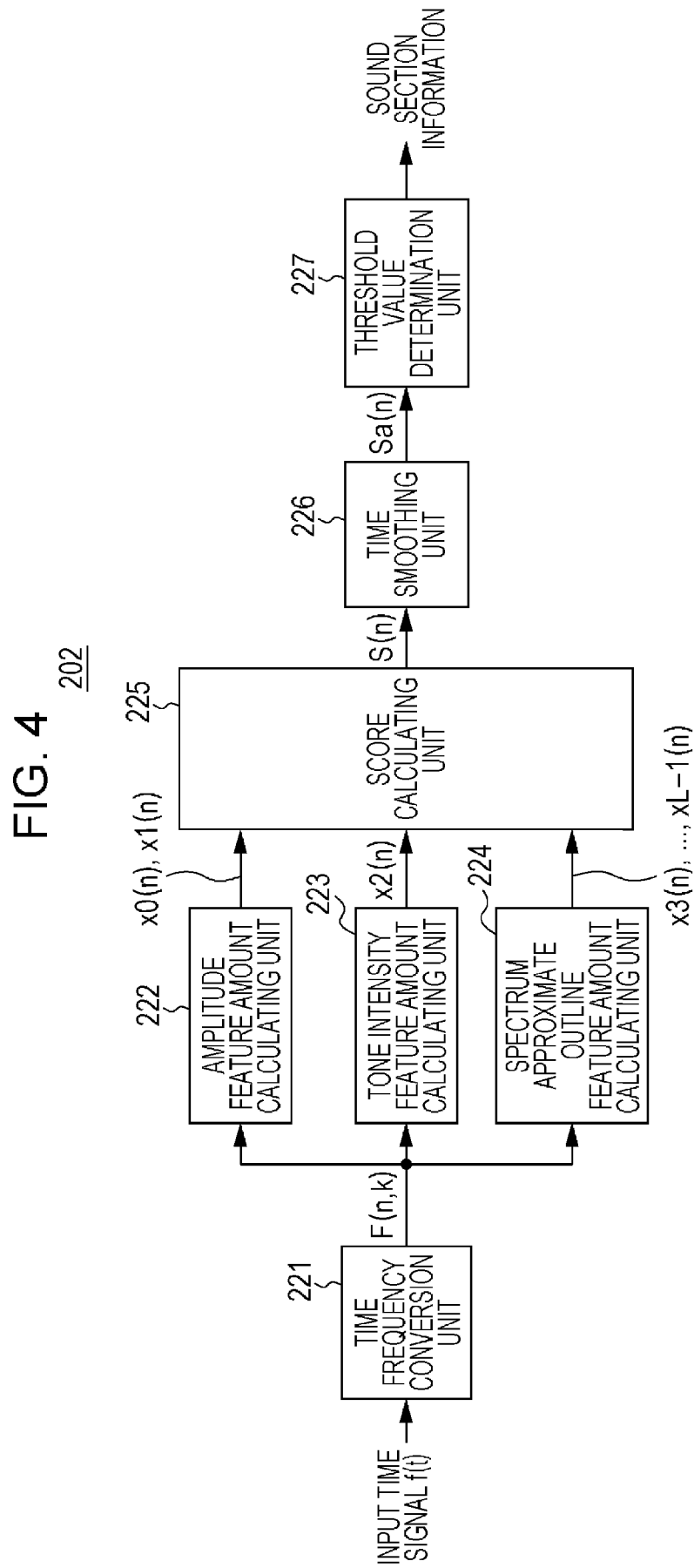


[Fig. 3]

FIG. 3

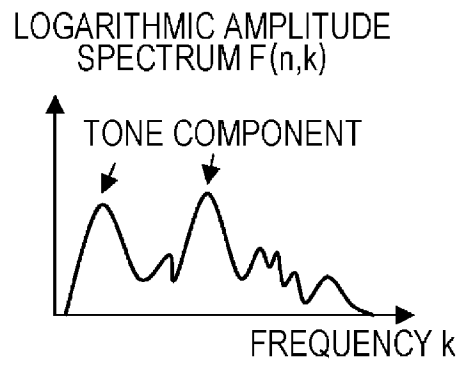


[Fig. 4]



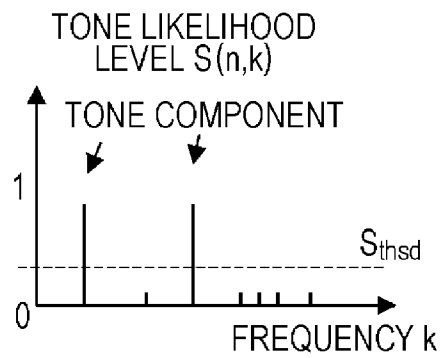
[Fig. 5A]

FIG. 5A



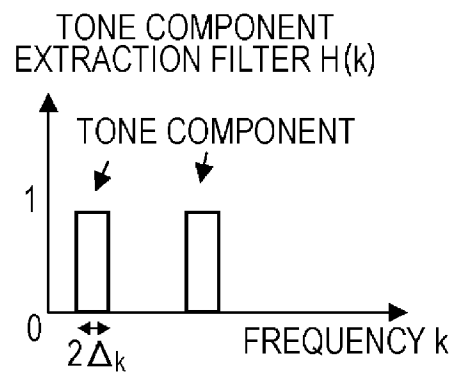
[Fig. 5B]

FIG. 5B



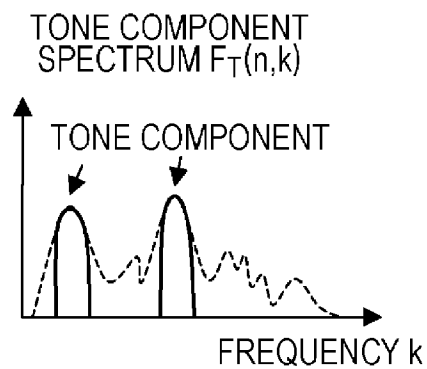
[Fig. 5C]

FIG. 5C



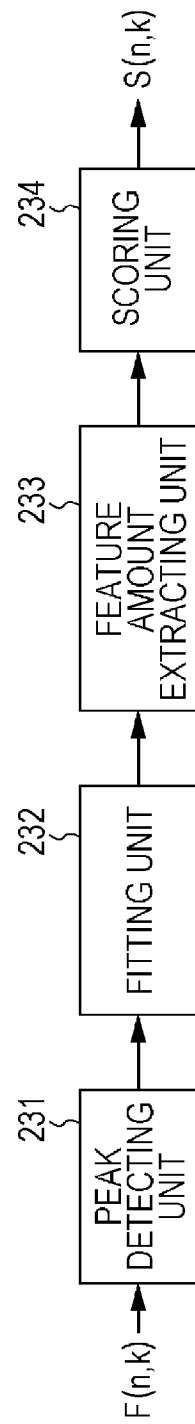
[Fig. 5D]

FIG. 5D



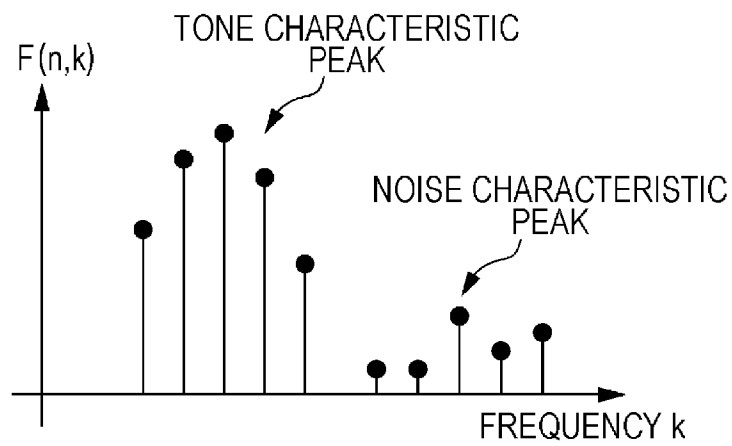
[Fig. 6]

FIG. 6

230

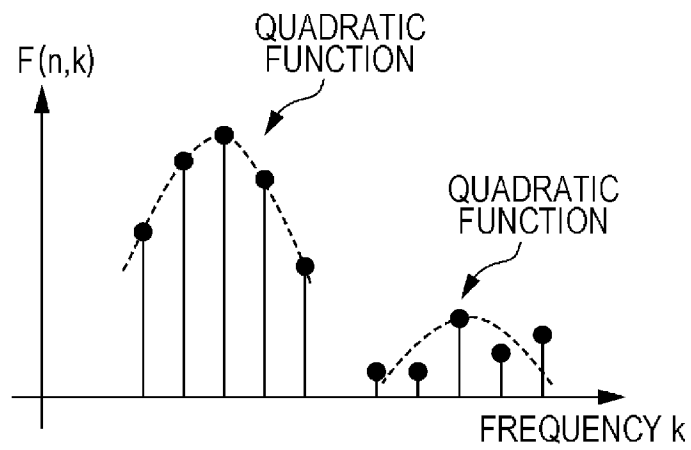
[Fig. 7A]

FIG. 7A



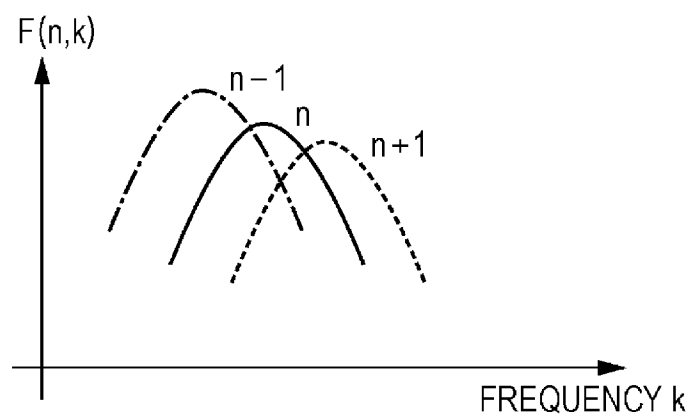
[Fig. 7B]

FIG. 7B



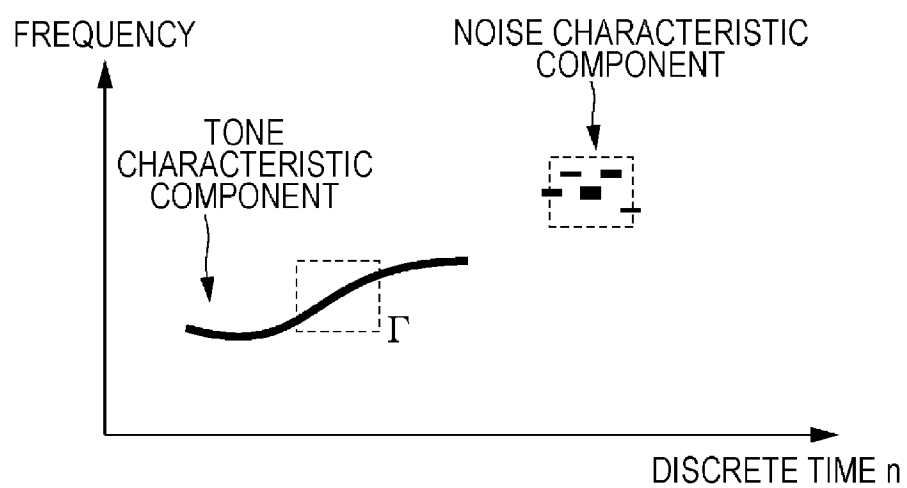
[Fig. 8A]

FIG. 8A



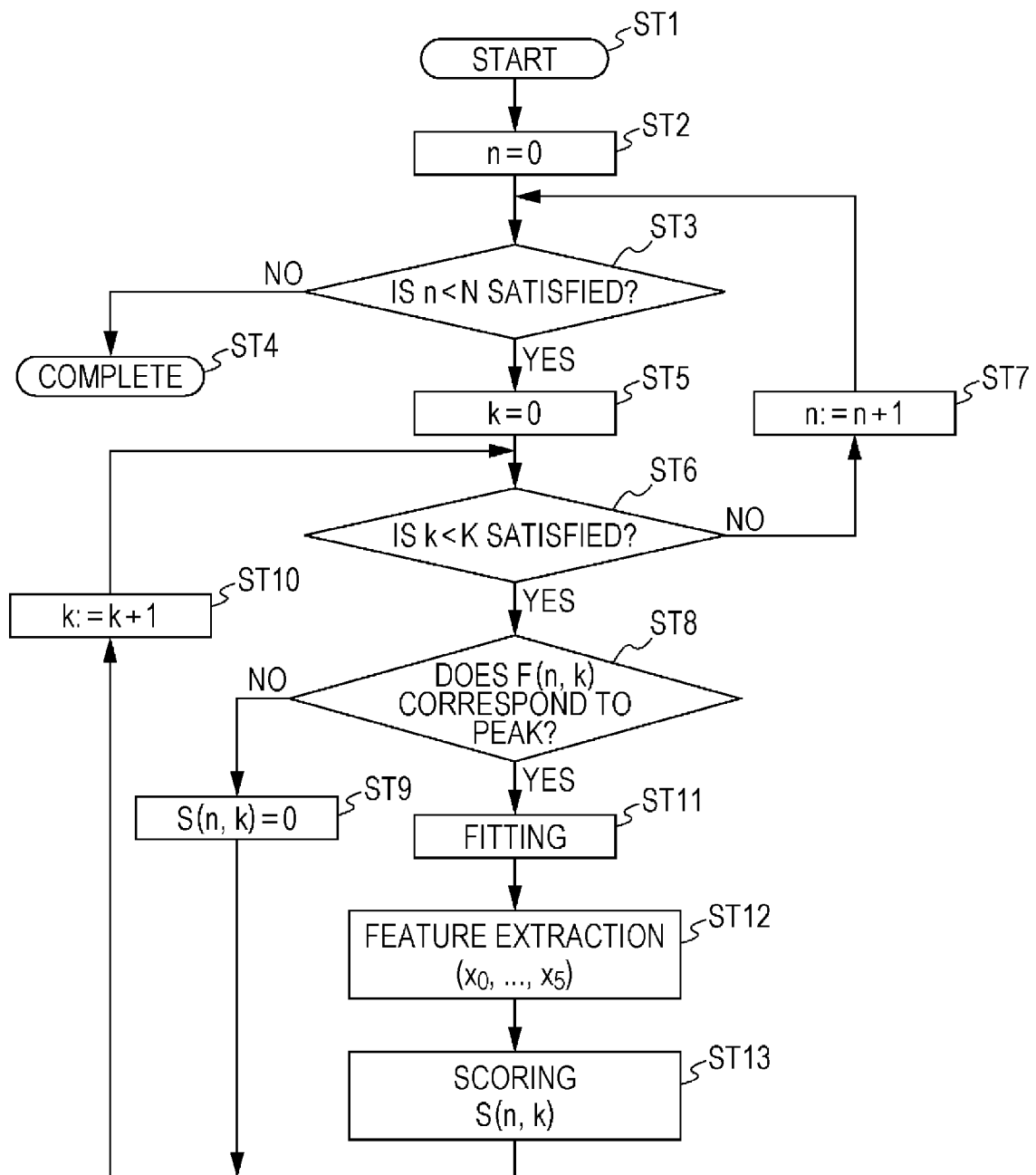
[Fig. 8B]

FIG. 8B



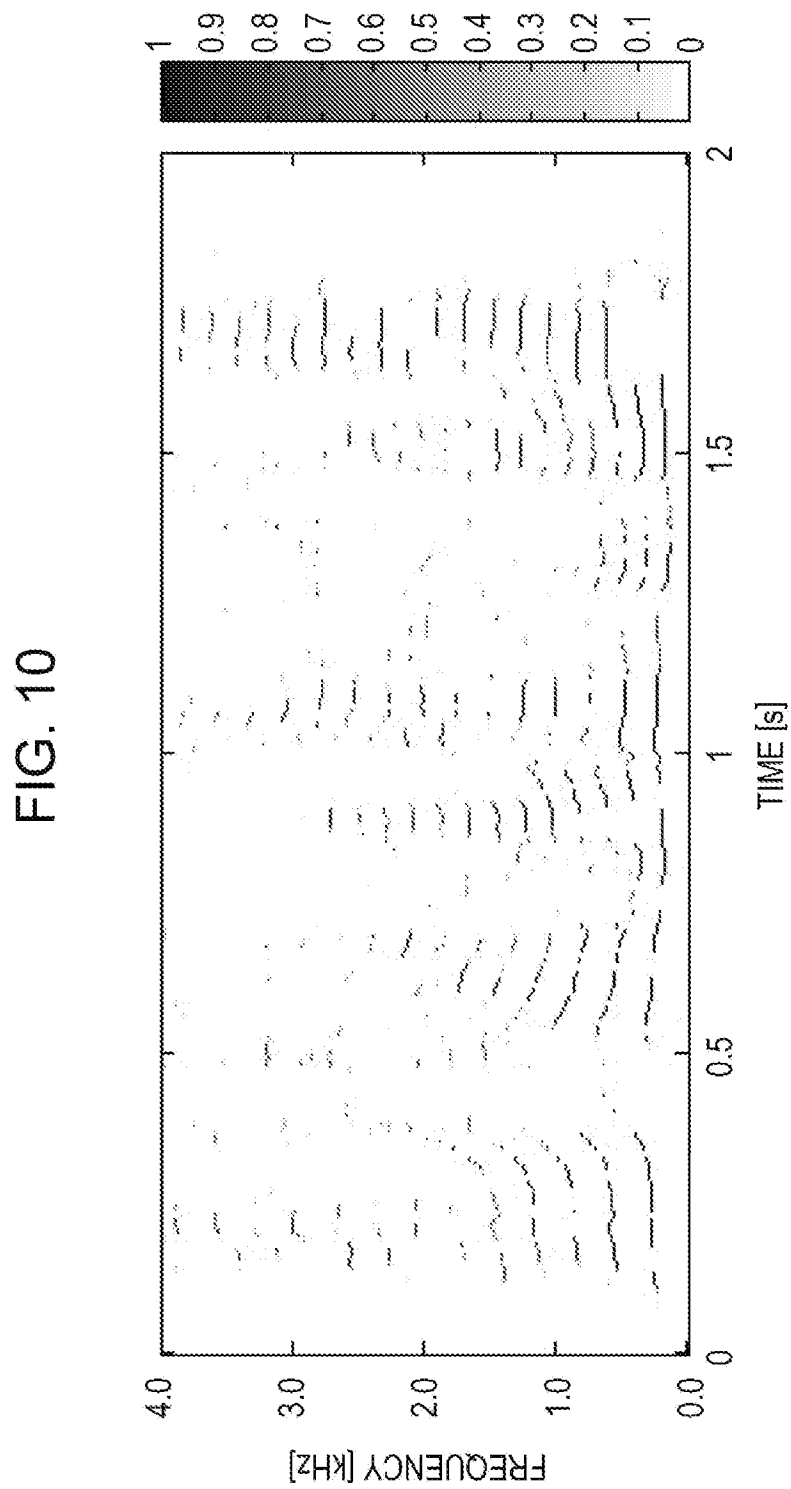
[Fig. 9]

FIG. 9

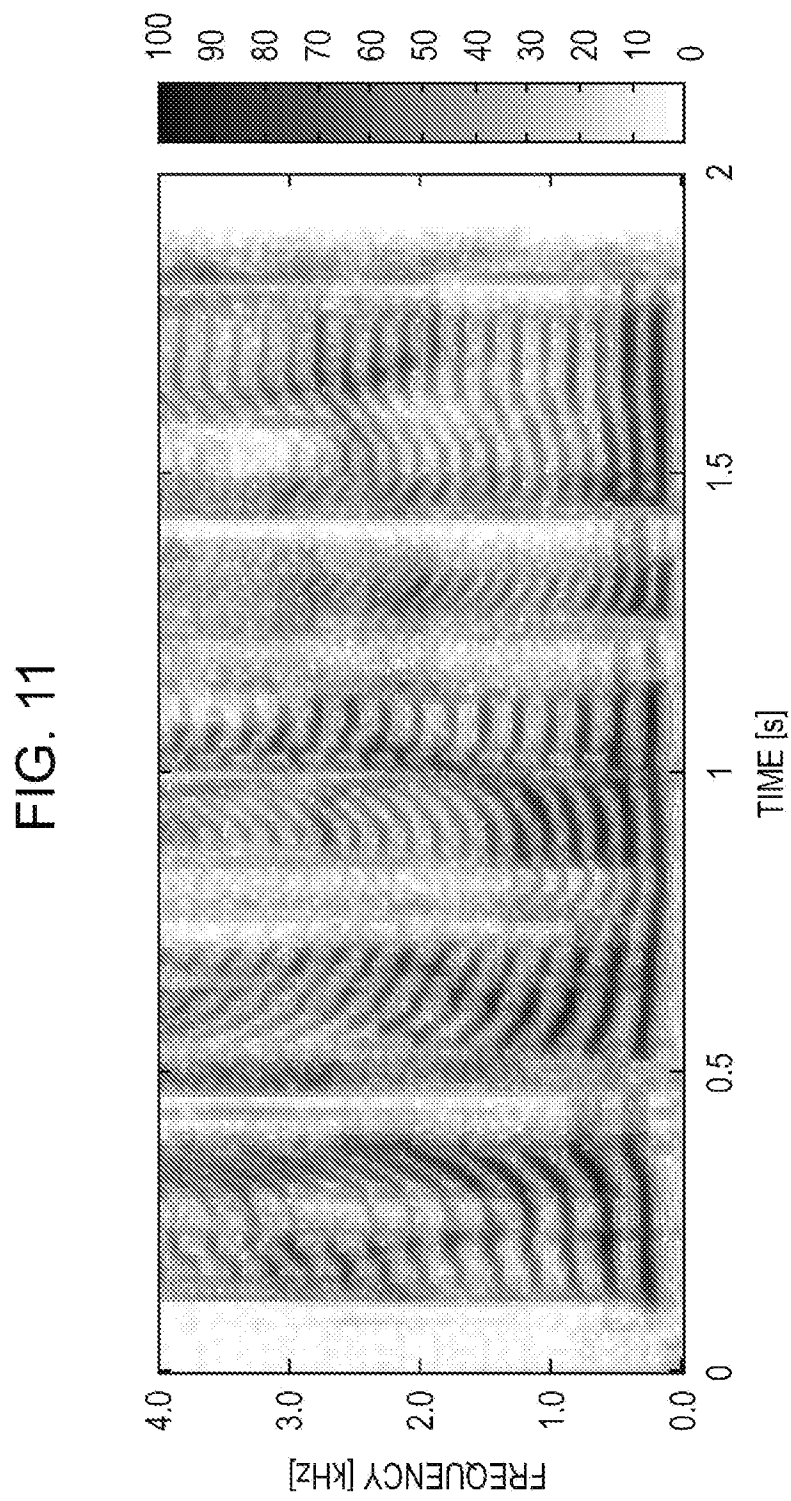




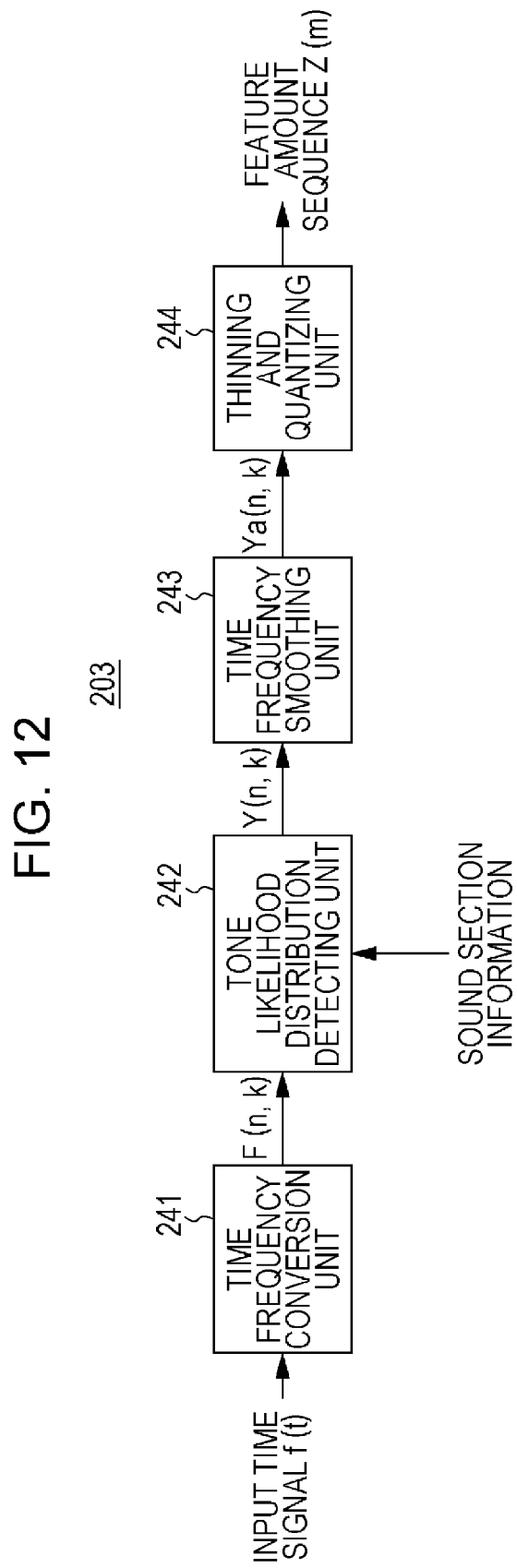
[Fig. 10]



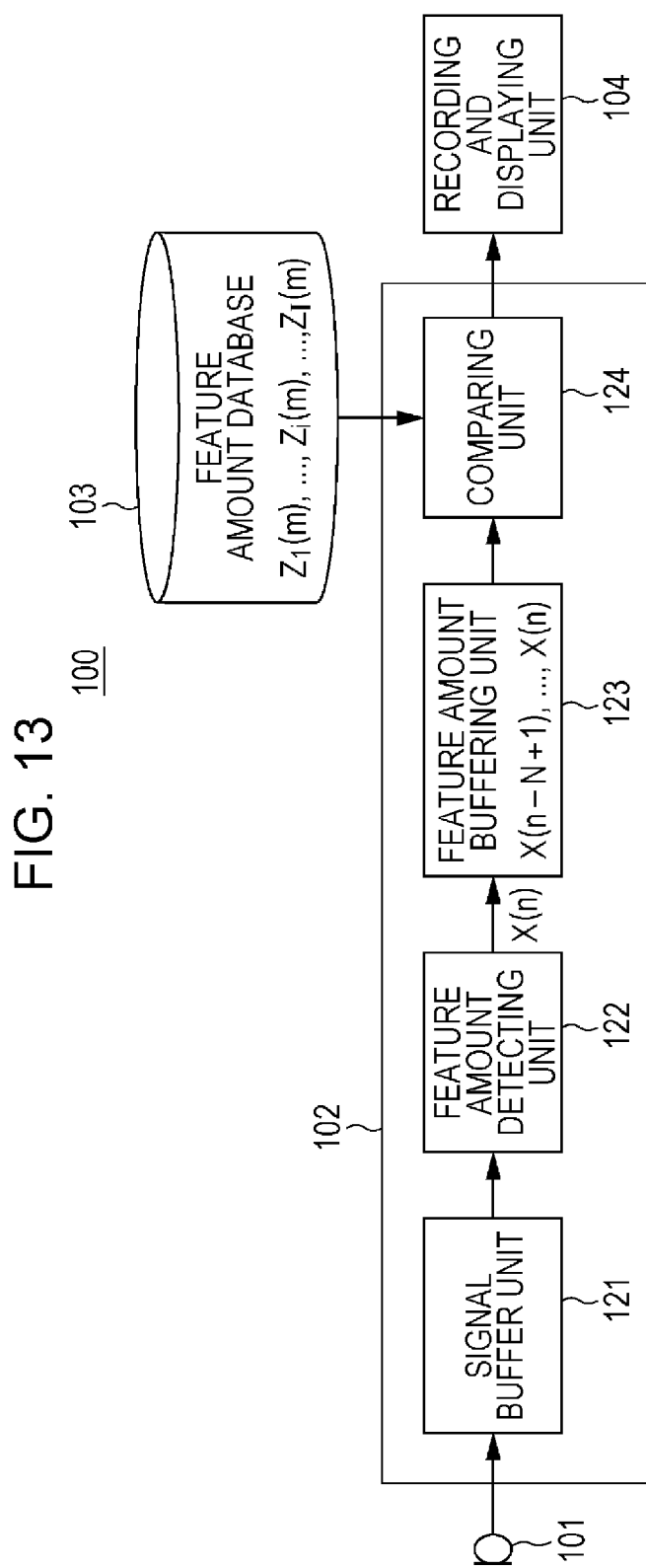
[Fig. 11]



[Fig. 12]

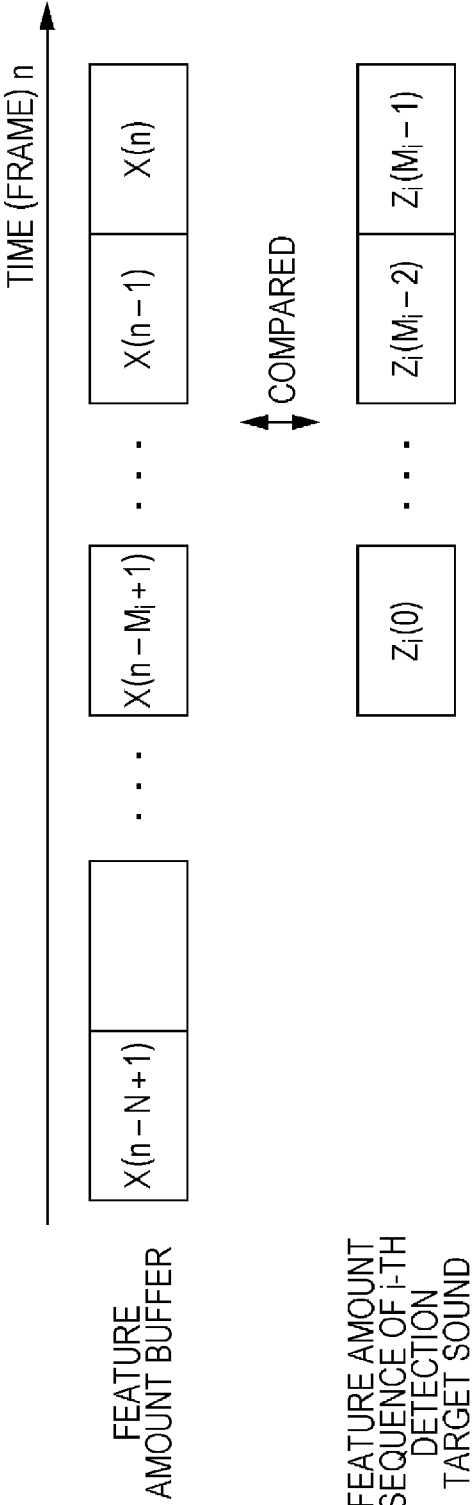


[Fig. 13]

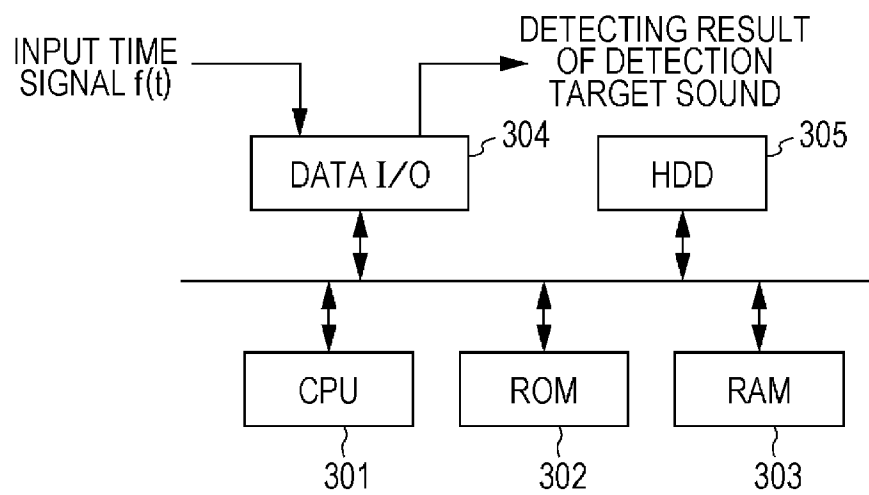


[Fig. 14]

FIG. 14

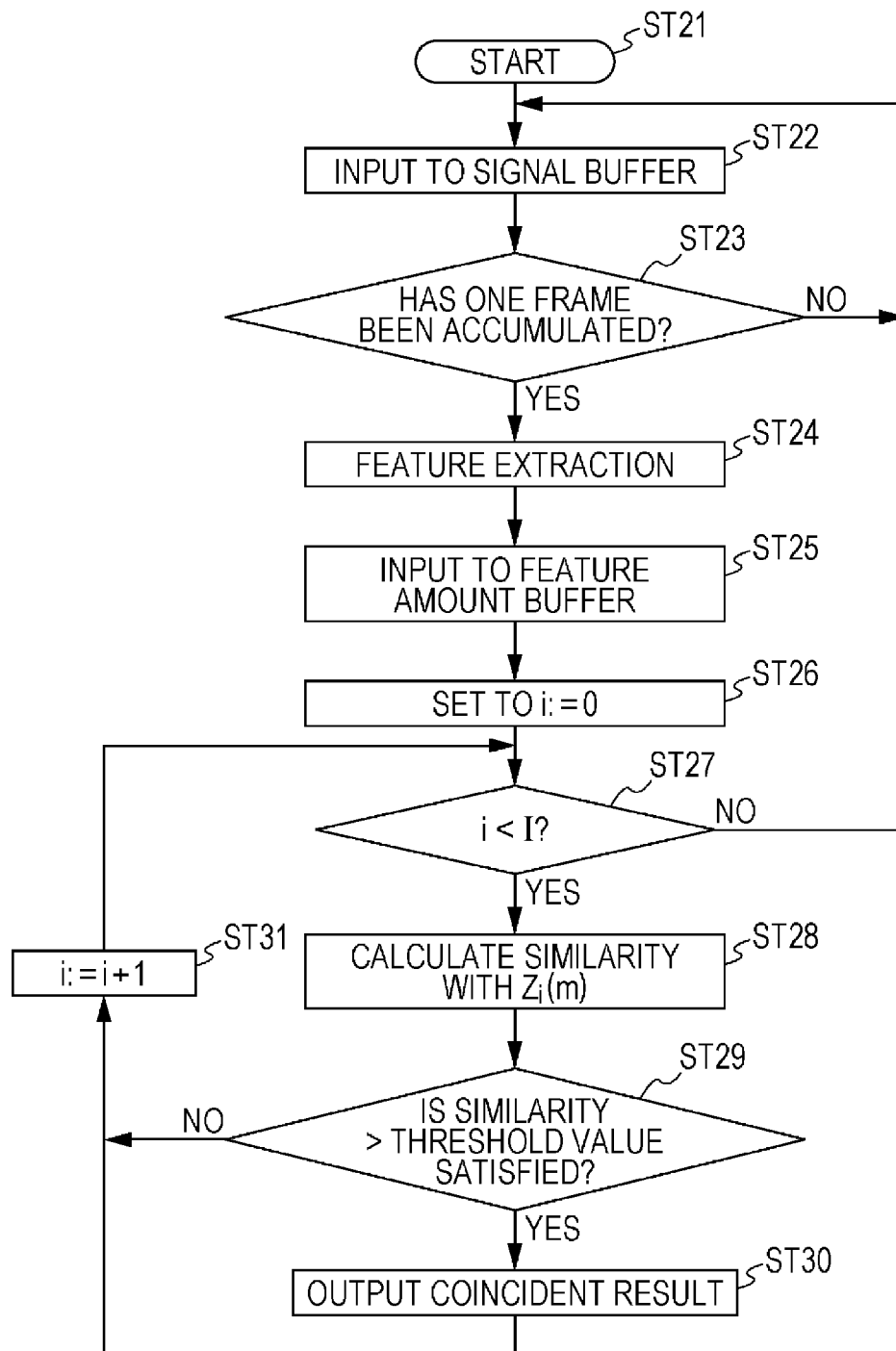


[Fig. 15]



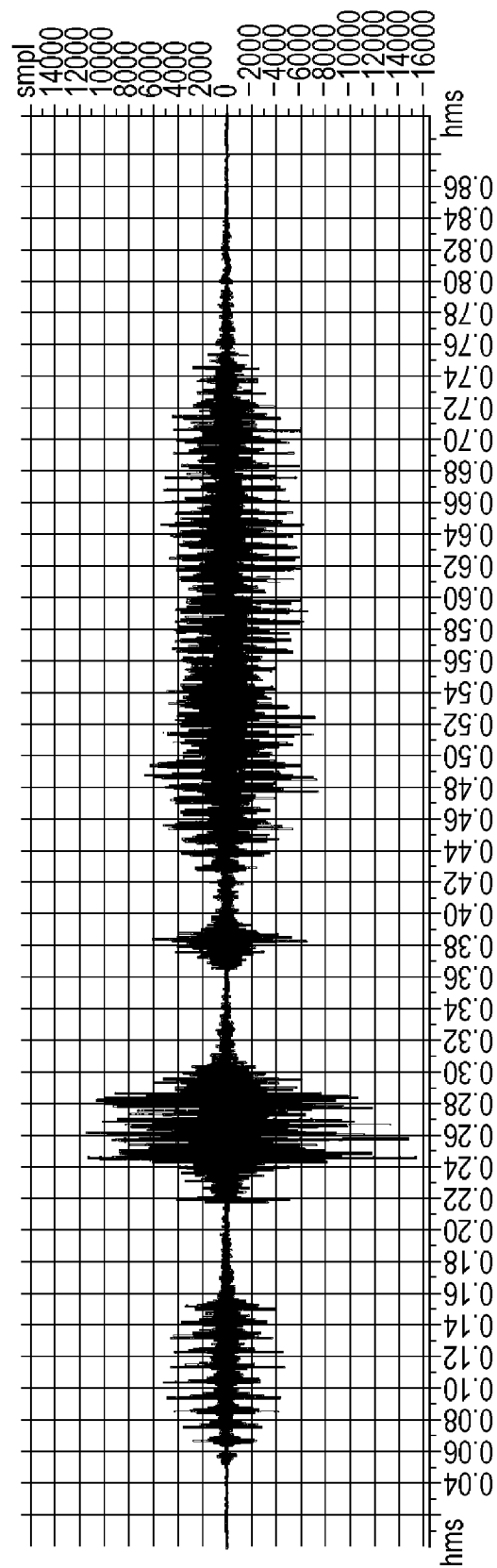
[Fig. 16]

FIG. 16



[Fig. 17A]

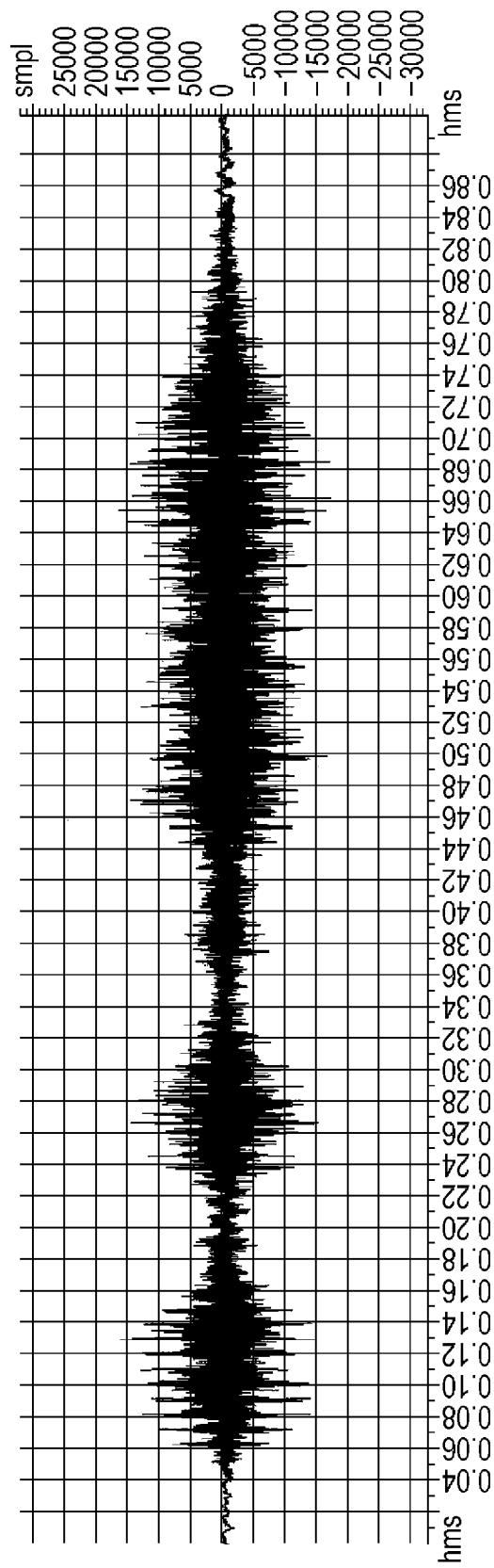
FIG. 17A





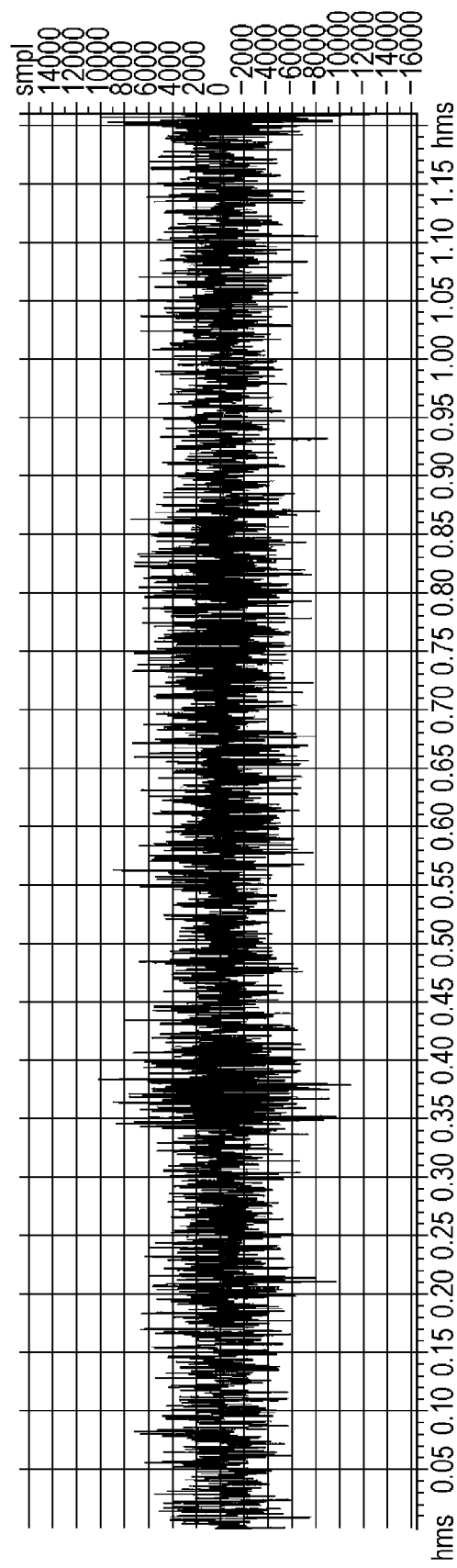
[Fig. 17B]

FIG. 17B



[Fig. 17C]

FIG. 17C



## INTERNATIONAL SEARCH REPORT

International application No

PCT/JP2013/002581

## A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F17/30 G10L25/48  
 ADD. G10L25/18

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	US 2010/332222 A1 (BAI MINGSIAN R [TW] ET AL) 30 December 2010 (2010-12-30) figure 1 paragraphs [0047] - [0048], [0053] - [0057] -----	1,3-9, 11-18 2,10
A	US 2009/002490 A1 (SAITO MUTSUMI [JP]) 1 January 2009 (2009-01-01) the whole document -----	1-18
A	US 2008/300702 A1 (GOMEZ EMILIA [ES] ET AL) 4 December 2008 (2008-12-04) figure 2 paragraphs [0059] - [0076] ----- -/-	1-18



Further documents are listed in the continuation of Box C.



See patent family annex.

## \* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&amp;" document member of the same patent family

Date of the actual completion of the international search

26 August 2013

Date of mailing of the international search report

05/09/2013

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2  
 NL - 2280 HV Rijswijk  
 Tel. (+31-70) 340-2040,  
 Fax: (+31-70) 340-3016

Authorized officer

Chétry, Nicolas

## INTERNATIONAL SEARCH REPORT

International application No

PCT/JP2013/002581

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>COWLING M ET AL: "Comparison of techniques for environmental sound recognition", PATTERN RECOGNITION LETTERS, ELSEVIER, AMSTERDAM, NL, vol. 24, no. 15, 1 November 2003 (2003-11-01), pages 2895-2907, XP004443655, ISSN: 0167-8655, DOI: 10.1016/S0167-8655(03)00147-8 the whole document -----</p>	1-18

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/JP2013/002581

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010332222	A1	30-12-2010	NONE
-----			
US 2009002490	A1	01-01-2009	JP 2009008823 A 15-01-2009
		US 2009002490 A1	01-01-2009
-----			
US 2008300702	A1	04-12-2008	NONE
-----			