Innovation, Sciences et Développement économique Canada
Office de la Propriété Intellectuelle du Canada

Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

CA 2955652 C 2022/04/05

(11)(21) 2 955 652

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2014/10/24
(87) Date publication PCT/PCT Publication Date: 2015/08/13
(45) Date de délivrance/Issue Date: 2022/04/05
(85) Entrée phase nationale/National Entry: 2017/01/18
(86) N° demande PCT/PCT Application No.: CN 2014/089490
(87) N° publication PCT/PCT Publication No.: 2015/117410
(30) Priorité/Priority: 2014/07/18 (CN201410345942.3)

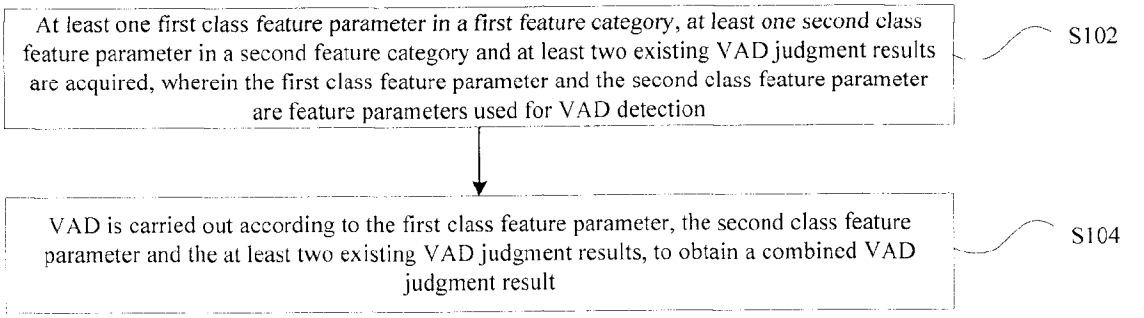
(51) Cl.Int./Int.Cl. *G10L 25/78* (2013.01)

(72) Inventeurs/Inventors:
ZHU, CHANGBAO, CN;
YUAN, HAO, CN

(73) Propriétaire/Owner:
ZTE CORPORATION, CN

(74) Agent: FASKEN MARTINEAU DUMOULIN LLP

(54) Titre : PROCEDE ET DISPOSITIF DE DETECTION D'ACTIVITE VOCALE
(54) Title: VOICE ACTIVITY DETECTION METHOD AND APPARATUS



(57) Abrégé/Abstract:
Provided are a Voice Activity Detection (VAD) method and apparatus. The method includes that: at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results are acquired, the first class feature and the second class feature are features used for VAD detection (S102); and VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results, to obtain a combined VAD judgment result (S104). By means of the technical solution, the technical problems of low detection accuracy of a VAD solution are solved, and the accuracy of VAD is improved, thereby improving the user experience.

Abstract

Provided are a Voice Activity Detection (VAD) method and apparatus. The method includes that: at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results are acquired, the first class feature and the second class feature are features used for VAD detection (S102); and VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results, to obtain a combined VAD judgment result (S104). By means of the technical solution, the technical problems of low detection accuracy of a VAD solution are solved, and the accuracy of VAD is improved, thereby improving the user experience.

Voice Activity Detection Method and Apparatus

Technical Field

The present disclosure relates to the field of communications, and in particular to a Voice Activity Detection (VAD) method and apparatus.

Background

In a normal voice call, a user is sometimes talking, and sometimes listening. Under such a scenario, an inactive speech stage occurs in the call process. The total inactive speech stage of a calling party and a called party under normal circumstances occupies more than 50% of the total voice coding duration. In an inactive speech stage, there is only some background noise which usually does not have any useful information. In consideration of this fact, an active speech and a non-active speech are detected by means of a VAD algorithm in a voice signal processing procedure, and are processed using different methods respectively. Many voice coding standards currently adopted, such as an Adaptive Multiple Rate (AMR) and an Adaptive Multiple Rate-WideBand (AMR-WB), support the VAD function. In terms of efficiency, VAD of these coders cannot achieve good performance under all typical background noises. Specifically, the VAD efficiency of these coders is relatively low under an unstable noise circumstance. VAD may be wrong sometimes for a music signal, which greatly reduces the performance of a corresponding processing algorithm. In addition, the current VAD technologies have the problem of inaccurate judgment. For instance, some VAD technologies have relatively low detection accuracy when detecting several frames before a voice segment, and some VAD technologies have relatively low detection accuracy when detecting several frames after a voice segment.

An effective solution for the above problems in the related art has not been proposed yet.

Summary

The embodiments of the present disclosure provide a VAD method and apparatus, which at least solve the technical problems of low detection accuracy of a conventional VAD solution in the related art.

According to one embodiment of the present disclosure, a VAD method is provided, which may include that: at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results are acquired, in the embodiment, the first class feature and the second class feature are features used for VAD detection; and VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results, to obtain a combined VAD judgment result.

In an exemplary embodiment, the first class feature in the first feature category may include at least one of: the number of continuous active frames, an average total signal-to-noise ratio (SNR) of all sub-bands and a tonality signal flag, in the embodiment, the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames. The second class feature in the second feature category may include at least one of: a flag of noise type, a smoothed average long-time frequency domain SNR, the number of continuous noise frames and a frequency domain SNR.

In an exemplary embodiment, the step that VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results may include that: a) one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD; b) if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, and otherwise, Step c) is executed, in the embodiment, the VAD flag is used for indicating that the VAD judgment result is an active frame or an inactive frame; c) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, Step d) is executed, and otherwise, the VAD judgment result selected in Step a) is selected as the combined VAD judgment result; d) when a preset condition is met, a logical operation OR is carried out on the at least two existing VAD judgment results and the result of the logical operation OR is used as the combined VAD judgment result, and otherwise, Step e) is executed; and e) if the flag of noise type indicates

that the noise type is silence, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, and otherwise, the VAD judgment result selected in Step a) is selected as the combined VAD judgment result.

In an exemplary embodiment, the step that VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results may include that: a) one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD; b) if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, and otherwise, Step c) is executed, in the embodiment, the VAD flag is used for indicating that the VAD judgment result is an active frame or an inactive frame; c) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, Step d) is executed, and otherwise, the VAD judgment result selected in Step a) is selected as the combined VAD judgment result; d) when a preset condition is met, a logical operation OR is carried out on the at least two existing VAD judgment results and the result of the logical operation OR is used as the combined VAD judgment result, and otherwise, Step e) is executed; and e) a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result.

In an exemplary embodiment, the step that VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results may include that: a) one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD; and b) if the flag of noise type indicates that the noise type is silence, the smoothed average long-time frequency domain SNR is greater than a threshold and the tonality signal flag indicates a non-tonal signal, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, in the embodiment, the VAD flag is used for indicating that the VAD judgment result is an active frame or an

inactive frame.

In an exemplary embodiment, the step that VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results may include that: a) one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD; and b) if the noise type is non-silence and a preset condition is met, a logical operation OR is carried out on the at least two existing VAD judgment results, and the result of the logical operation OR is used as the combined VAD judgment result.

In an exemplary embodiment, the preset condition may include at least one of: condition 1: the average total SNR of all sub-bands is greater than a first threshold; condition 2: the average total SNR of all sub-bands is greater than a second threshold, and the number of continuous active frames is greater than a preset threshold; and condition 3: the tonality signal flag indicates a tonal signal.

In an exemplary embodiment, the step that VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results may include that: if the number of continuous noise frames is greater than a first appointed threshold and the average total SNR of all sub-bands is smaller than a second appointed threshold, a logical operation AND is carried out on the at least two existing VAD judgment results, and the result of the logical operation AND is used as the combined VAD judgment result; and otherwise, one existing VAD judgment result is randomly selected from the at least two existing VAD judgment results as the combined VAD result.

In an exemplary embodiment, the smoothed average long-time frequency domain SNR and the flag of noise type may be determined by means of the following modes:

calculating average energy of active frames of a current frame and average energy of background noise of the current frame according to any one VAD judgment result in a combined VAD judgment result of a previous frame of the current frame or at least two existing VAD judgment results corresponding to the previous frame, average energy of active frames of the previous frame within a first preset time period and average energy of background noise of the previous frame;

calculating a long-time SNR of the current frame within a second time period according to the average energy of background noise and average energy of active frames of the current frame within the second preset time period;

calculating a smoothed average long-time frequency domain SNR of the current frame within a third preset time period according to any one VAD judgment result in the combined VAD judgment result of the current frame or at least two existing VAD judgment results corresponding to the previous frame and a frequency domain SNR of the previous frame; and

determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR.

In an exemplary embodiment, determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR may include:

setting the flag of noise type to non-silence, and setting, when the long-time SNR is greater than a first preset threshold and the average frequency domain SNR is greater than a second preset threshold, the flag of noise type to silence.

According to another embodiment of the present disclosure, a VAD apparatus is provided, which may include: an acquisition component, arranged to acquire at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results, in the embodiment, the first class feature and the second class feature are features used for VAD detection; and a detection component, arranged to carry out, according to the first class feature, the second class feature and the at least two existing VAD judgment results, VAD to obtain a combined VAD judgment result.

In an exemplary embodiment, the acquisition component may include: a first acquisition unit, arranged to acquire the first class feature in the first feature category which includes at least one of: the number of continuous active frames, an average total signal-to-noise ratio (SNR) of all sub-bands and a tonality signal flag, in the embodiment, the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames; and a second acquisition unit, arranged to acquire the second class feature in the second feature category which includes at least one of: a flag of noise type, a smoothed

average long-time frequency domain SNR, the number of continuous noise frames and a frequency domain SNR.

In the embodiments of the present disclosure, combined detection is carried out according to at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results. By virtue of the above technical means, the technical problems of low detection accuracy of a VAD solution in the related art are solved, and the accuracy of VAD is improved, thereby improving the user experience.

Brief Description of the Drawings

The drawings illustrated herein are used to provide further understanding of the embodiments of the present disclosure, and form a part of the present disclosure. The schematic embodiments and illustrations of the present disclosure are used to explain the present disclosure, and do not form improper limits to the present disclosure. In the drawings:

Fig. 1 is a flowchart of a VAD method according to an embodiment of the present disclosure;

Fig. 2 is a structural diagram of a VAD apparatus according to an embodiment of the present disclosure;

Fig. 3 is another structural diagram of a VAD apparatus according to an embodiment of the present disclosure; and

Fig. 4 is a flowchart of a VAD method according to an embodiment 1 of the present disclosure.

Detailed Description of the Embodiments

The present disclosure will be illustrated below with reference to the drawings and in conjunction with the embodiments in detail. It is important to note that the embodiments of the present disclosure and the features in the embodiments can be combined under the condition of no conflicts.

In order to solve the problem of low detection accuracy of VAD, the following embodiments provide corresponding solutions, which will be illustrated in detail.

Fig. 1 is a flowchart of a VAD method according to an embodiment of the present disclosure. As shown in Fig. 1, the method includes the steps S102 to

S104 as follows.

Step S102: At least one first class feature in a first feature category (also called as a feature category 1), at least one second class feature in a second feature category (also called as a feature category 2) and at least two existing VAD judgment results are acquired, the first class feature and the second class feature are features used for VAD detection.

Step S104: VAD is carried out according to the first class feature, the second class feature and the at least two existing VAD judgment results, to obtain a combined VAD judgment result.

By means of all the above processing steps, combined VAD can be carried out according to at least one feature in a first feature category, at least one feature in a second feature category and at least two existing VAD judgment results, thus improving the accuracy of VAD.

In the present embodiment, the first class feature in the first feature category may include at least one of: the number of continuous active frames, an average total SNR of all sub-bands and a tonality signal flag, where the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames.

In the present embodiment, the second class feature in the second feature category may include at least one of: a flag of noise type, a smoothed average long-time frequency domain SNR, the number of continuous noise frames and a frequency domain SNR, the smoothed average long-time frequency domain SNR can be interpreted as: a frequency domain SNR obtained by smoothing the average of a plurality of frequency domain SNRs within a predetermined time period (long time).

There are multiple implementations for Step S104. For instance, Step S104 may be implemented by means of the modes as follows.

Judgment ending in the following several implementations is only representative of process ending of a certain implementation, and does not mean that a combined VAD judgment result is no longer modified after this process is ended.

A first implementation is executed in accordance with the following steps:

a) one VAD judgment result is selected from the at least two existing VAD

judgment results as an initial value of combined VAD;

b) if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, and otherwise, Step c) is executed, the VAD flag is used for indicating that the VAD judgment result is an active frame or an inactive frame;

c) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, Step d) is executed, and otherwise, the VAD judgment result selected in Step a) is selected as the combined VAD judgment result;

d) when a preset condition is met, a logical operation OR is carried out on the at least two existing VAD judgment results and the result of the logical operation OR is used as the combined VAD judgment result, and otherwise, Step e) is executed; and

e) if the flag of noise type indicates that the noise type is silence, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result.

A second implementation is executed in accordance with the following steps:

a) one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD;

b) if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, and otherwise, Step c) is executed, the VAD flag is used for indicating that the VAD judgment result is an active frame or an inactive frame;

c) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, Step d) is executed, and otherwise, the VAD judgment result selected in Step a) is selected as the combined VAD judgment result;

d) when a preset condition is met, a logical operation OR is carried out on the at least two existing VAD judgment results and the result of the logical operation

OR is used as the combined VAD judgment result, and otherwise, Step e) is executed; and

e) a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result.

A third implementation is executed in accordance with the following steps:

one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD; and

if the flag of noise type indicates that the noise type is silence, the smoothed average long-time frequency domain SNR is greater than a threshold and the tonality signal flag indicates a non-tonal signal, a VAD flag, which is not selected as the initial value, in the at least two existing VAD judgment results is selected as the combined VAD judgment result, the VAD flag is used for indicating that the VAD judgment result is an active frame or an inactive frame.

A fourth implementation is executed in accordance with the following steps:

a) one VAD judgment result is selected from the at least two existing VAD judgment results as an initial value of combined VAD; and

b) if the noise type is non-silence and a preset condition is met, a logical operation OR is carried out on the at least two existing VAD judgment results, and the result of the logical operation OR is used as the combined VAD judgment result.

It is important to note that the preset condition involved in the first implementation, the second implementation and the third implementation may include at least one of:

condition 1: the average total SNR of all sub-bands is greater than a first threshold;

condition 2: the average total SNR of all sub-bands is greater than a second threshold, and the number of continuous active frames is greater than a preset threshold; and

condition 3: the tonality signal flag indicates a tonal signal.

It is important to note that the third implementation and the fourth implementation can be used in conjunction.

A fifth implementation is executed in accordance with the following steps:

if the number of continuous noise frames is greater than a first appointed

threshold and the average total SNR of all sub-bands is smaller than a second appointed threshold, a logical operation AND is carried out on the at least two existing VAD judgment results and the result of the logical operation AND is used as the combined VAD judgment result; and otherwise, one existing VAD judgment result is randomly selected from the at least two existing VAD judgment results as the combined VAD result.

It is important to note that the fifth implementation and the above four implementations can be used in conjunction.

In an exemplary embodiment of the present embodiment, the smoothed average long-time frequency domain SNR and the flag of noise type may be determined by means of the following modes:

calculating average energy of active frames of a current frame and average energy of background noise of the current frame according to any one VAD judgment result in a combined VAD judgment result of a previous frame of the current frame or at least two existing VAD judgment results corresponding to the previous frame, average energy of active frames of the previous frames within a first preset time period and average energy of background noise of the previous frames;

calculating a long-time SNR of the current frame within a second time period according to the average energy of background noise and average energy of active frames of the current frame within the second preset time period;

calculating a smoothed average long-time frequency domain SNR of the current frame within a third preset time period according to any one VAD judgment result in the combined VAD judgment result of the current frame or at least two existing VAD judgment results corresponding to the previous frame and a frequency domain SNR of the previous frame; and

determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR.

It is important to note that the smoothed average long-time frequency domain SNR is obtained by smoothing an average frequency domain SNR within a predetermined time period.

In an exemplary implementation, the flag of noise type may be determined based on the following manner, but is not limited to:

setting the flag of noise type to non-silence, and setting, when the long-time SNR is greater than a first preset threshold and the average frequency domain SNR is greater than a second preset threshold, the flag of noise type to silence.

In an exemplary implementation, the number of continuous active frames and the number of continuous noise frames are determined by means of the following modes:

when a current frame is a non-initialized frame, calculating the number of continuous active frames and number of continuous noise frames of the current frame according to a combined VAD judgment result of a previous frame of the current frame, or

when the current frame is a non-initialized frame, selecting one VAD judgment result from at least two existing VAD judgment results of the previous frame and the combined VAD judgment result of the previous frame, and calculating the number of continuous active frames and number of continuous noise frames of the current frame according to the currently selected VAD judgment result.

In an exemplary implementation process of the present embodiment, the number of continuous active frames and the number of continuous noise frames are determined by means of the following modes:

when a VAD flag for the combined VAD judgment result of the previous frame or for the currently selected VAD judgment result indicates an active frame, adding 1 to the number of continuous active frames, and otherwise, setting the number of continuous active frames to 0; and when a VAD flag for the combined VAD judgment result of the previous frame or for the currently selected VAD judgment result indicates a noise frame, adding 1 to the number of continuous noise frames, and otherwise, setting the number of continuous noise frames to 0.

In the present embodiment, a VAD apparatus is also provided. As shown in Fig. 2, the VAD apparatus includes:

an acquisition component 20, arranged to acquire at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results, the first class feature and the second class feature are features used for VAD detection; and

a detection component 22, coupled with the acquisition component 20, and

arranged to carry out, according to the first class feature, the second class feature and the at least two existing VAD judgment results, VAD to obtain a combined VAD judgment result.

In an exemplary embodiment, as shown in Fig. 3, the acquisition component 20 may also include the following processing units:

a first acquisition unit 200, arranged to acquire the first class feature in the first feature category which includes at least one of: the number of continuous active frames, an average total SNR of all sub-bands and a tonality signal flag, the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames; and

a second acquisition unit 202, arranged to acquire the second class feature in the second feature category which includes at least one of: a flag of noise type, a smoothed average long-time frequency domain SNR, the number of continuous noise frames and a frequency domain SNR.

It is important to note that all the components involved in the present embodiment can be implemented by means of software or hardware. In an exemplary implementation, the components may be implemented by means of hardware in the following modes: the acquisition component 20 is located in a first processor, and the detection component 22 is located in a second processor; or the two components are located in, but not limited to, the same processor.

In order to better understand the above embodiment, detailed illustrations will be made below in conjunction with exemplary embodiments.

An OR operation and an AND operation involved in the following embodiments are defined as follows.

If any one VAD output flag in two VADs is an active frame, the result of the logical operation OR of the two VADs is an active frame, and when the two VADs are both inactive frames, the result of the logical operation OR is an inactive frame.

If any one VAD output flag in two VADs is an inactive frame, the result of the logical operation AND of the two VADs is an inactive frame, and when the two VADs are both active frames, the result of the logical operation AND is an active frame.

Note: if it is not specified which VAD(s) the following embodiment is/are

referring to, it represents that the VAD(s) may be two existing VADs or a combined VAD or other VADs capable of achieving corresponding functions.

Judgment ending in the following embodiments is only representative of process ending of a certain implementation, and does not mean that a combined VAD judgment result is no longer modified after this process is ended.

Embodiment 1

The present embodiment provides a VAD method. As shown in Fig. 4, the method includes the steps as follows.

Step S402: Two existing VAD output results are obtained.

Step S404: A sub-band signal and spectrum amplitude of a current frame are obtained.

The embodiments of the present disclosure are specifically illustrated with an audio stream of which a frame length is 20ms and a sampling rate is 32kHz. Under the conditions of other frame lengths and sampling rates, a combined VAD method provided by the embodiments of the present disclosure is also applicable.

A time domain signal of a current frame is input into a filter bank, and sub-band filtering calculation is carried out to obtain a filter bank sub-band signal.

In the present embodiment, a 40-channel filter bank is adopted. The technical solutions provided by the embodiments of the present disclosure are also applicable to filter banks with other channel amounts.

A time domain signal of a current frame is input into the 40-channel filter bank, and sub-band filtering calculation is carried out to obtain filter bank sub-band signals $X[k,l]$ of 40 sub-bands on 16 time sampling points, $0 \le k < 40$, and $0 \le l < 16$, where k is an index of a sub-band of the filter bank, and its value represents a sub-band corresponding to a coefficient; and l is a time sampling point index of each sub-band. The implementation steps are as follows.

1: 640 latest audio signal samples are stored in a data cache.

2: Data in the data cache are shifted by 40 positions to shift 40 earliest samples out of the data cache, and 40 new samples are stored at positions 0 to 39.

Data x in the cache is multiplied by a window coefficient to obtain an array z, a calculation formula being as follows:

$$z[n] = x[n] \bullet W_{qmf}[n]; 0 \le n < 640;$$

where W_{qmf} is a window coefficient of the filter bank.

80-point data u is calculated using the following pseudo-code:

```
for ( n=0;  n<80;  n++)
{   u[n] = 0;
for ( j=0;  j<8;  j++)
{
u[n]+= z[n+ j •80];
        }
}
```

Arrays r and i are calculated using the following formula:

$$\begin{aligned} \mathrm{r}[n] &= u[n] - u[79-n] \\ \mathrm{i}[n] &= u[n] + u[79-n] \end{aligned}, 0 \le n < 40$$

40 sub-band complex samples on the first time sampling point are calculated using the following formula: $X[k,l] = R(k) + iI(k), 0 \le k < 40$, where $R(k)$ and $I(k)$ are real part and imaginary part of a coefficient of the filter bank sub-band signal X on the l^{th} time sampling point, respectively. The calculation formula is as follows.

$$\begin{aligned} R(k) &= \sum_{n=0}^{39} \mathrm{r}(n) \cos[\frac{\pi}{40}(k+\frac{1}{2})n] \\ I(k) &= \sum_{n=0}^{39} \mathrm{i}(n) \cos[\frac{\pi}{40}(k+\frac{1}{2})n] \end{aligned}, 0 \le k < 40$$

3: The calculation process in Step 2 is repeated until all data of the present frame are filtered by the filter bank, and the final output result is filter bank sub-band signal $X[k,l]$.

4: After the above calculation process is completed, the filter bank sub-band signal $X[k,l]$ of 40 sub-bands on 16 time sampling points are obtained, where $0 \le k < 40$, and $0 \le l < 16$.

Then, time-frequency transform is carried out on the filter bank sub-band signal, and spectrum amplitudes are calculated.

The embodiments of the present disclosure can be implemented by carrying out time-frequency transform on all or part of filter bank sub-bands and calculating spectrum amplitudes. A time-frequency transform method in the embodiments of the present disclosure may be a Discrete Fourier Transform (DFT) method, a Fast Fourier Transformation (FFT) method, a Discrete Cosine Transform (DCT) method or a Discrete Sine Transform (DST) method. In the embodiments of the present disclosure, a specific implementation method is illustrated taking the use of DFT as an example. A calculation process is as follows.

16-point DFT is carried out on data of 16 time sampling points of each filter bank sub-band indexed from 0 to 9 so as to further improve the spectrum resolution. The amplitude of each frequency point is calculated to obtain spectrum amplitude X_{DFT_AMP}.

The calculation formula for time-frequency transform is as follows.

$$X_{DFT}[k,j]=\sum_{l=0}^{15}X[k,l]e^{-\frac{2\pi}{16}jl};0\le k<10,0\le j<16.$$

The process of calculating the amplitude of each frequency point is as follows.

Firstly, energy of an array $X_{DFT}[k][j]$ on each frequency point is calculated, the calculation formula being as follows:

$X_{DFT\_POW}[k,j]=((\mathrm{Re}(X_{DFT}[k,j]))^2+(\mathrm{Im}(X_{DFT}[k,j]))^2);0\le k<10,0\le j<16$, where $\mathrm{Re}(X_{DFT}[k,j])$ and $\mathrm{Im}(X_{DFT}[k,j])$ represent the real part and the imaginary part of the spectrum coefficient $X_{DFT}[k,j]$, respectively.

If k is an even number, the spectrum amplitude on each frequency point is calculated using the following formula:

$X_{DFT\_AMP}[8\bullet k+j]=\sqrt{\mathrm{X_{DFT\_POW}}[k,\mathrm{j}]+\mathrm{X_{DFT\_POW}}[k,15\text{-}\mathrm{j}]};0\le k<10;0\le j<8;$; and

If k is an odd number, the spectrum amplitude on each frequency point is calculated using the following formula:

$X_{DFT\_AMP}[8\bullet k+7-j]=\sqrt{\mathrm{X_{DFT\_POW}}[k,\mathrm{j}]+\mathrm{X_{DFT\_POW}}[k,15\text{-}\mathrm{j}]};0\le k<10;0\le j<8;$,

where X_{DFT_AMP} is a spectrum amplitude subjected to time-frequency transform.

Step S406: A frame energy feature is a weighted accumulated value or directly accumulated value of all sub-band signal energies.

The frame energy feature of the current frame is calculated according to sub-band signals. Specifically,

sb_power[k]= $\sum_{l=0}^{15}((\mathrm{Re}(X[k,l]))^2+(\mathrm{Im}(X[k,l]))^2)$ 0<=k< band_num.

Frame energy 2 can be obtained by accumulating energy sb_power in certain sub-bands.

frame_energy2= $\sum_{n=e\_sb\_start}^{e\_sb\_end} sp\_power[n]$.

Frame energy 1 is frame_energy= frame_energy2 + fac*sb_power[0].

A plurality of SNR sub-bands can be obtained by sub-band division, and a SNR sub-band energy frame_sb_energy of the current frame can be obtained by accumulating energy in respective sub-band.

frame_sb_energy[i]= $\sum_{j=\mathrm{Nregion\_index[i]}}^{\mathrm{Nregion\_index[i+1]-1}} sp\_power[j]$.

Background noise energy, including sub-band background noise energy and background noise energy of all sub-bands, of the current frame is estimated according to a modification value of a flag of background noise, the frame energy feature of the current frame and the background noise energy of all sub-bands of previous frame. Table 1 gives a calculation method for a frame energy feature parameter. Calculation of a flag of background noise is shown in Step S430.

Step S408: The spectral centroid features are the ratio of the weighted sum to the non-weighted sum of energies of all sub-bands or partial sub-bands, or the value is obtained by applying a smooth filter to this ratio. The spectral centroid features can be obtained in the following steps.

A sub-band division for calculating the spectral centroid features is as follows.

Table 1 QMF sub-band division for spectral centroid features

Spectral centroid feature number k	Start sub-band index spc_start_band	End sub-band index spc_end_band

2	0	9
3	1	23

Two spectral centroid features, respectively the spectral centroid feature in the first interval and the spectral centroid feature in the second interval, are calculated using the subband division for calculating the spectral centroid features as shown in Table 1 and the following formula:

$$sp\_center[k] = \frac{\sum_{n=spc\_start\_band(k)}^{spc\_end\_band(k)} (n+1) * sp\_power[n] + Delta1}{\sum_{n=spc\_start\_band(k)}^{spc\_end\_band(k)} sp\_power[n] + Delta2}; 2 \le k < 4$$

Smooth the spectral centroid feature in the second interval sp_center[2], and obtain the smoothed spectral centroid feature in the second interval according to the following formula: sp_center[0]= fac*sp_center[0]+(1-fac)*sp_center[2].

Step S410: The time-domain stability features are the ratio of the variance of the sum of amplitudes to the expectation of the square of amplitudes, or this ratio multiplied by a factor. The time-domain stability features are computed with the energy features of the most recent N frame. Let the energy of the nth frame be frame_energy[n]. The amplitude of frame_energy[n] is computed by $Amp_{t1}[n] = \sqrt{\text{frame\_energy}[n]} + e\_offset; 0 \le n < N$, where e_offset is an offset value within a range of [0,0.1].

By adding together the energy amplitudes of two adjacent frames from the current frame to the N^{th} previous frame, N/2 sums of energy amplitudes are obtained as $Amp_{t2}(n) = Amp_{t1}(-2n) + Amp_{t1}(-2n-1); 0 \le n < 20$,

where when n=0, Amp_{t1} [n] represents the energy amplitude of a current frame, and when n<0, Amp_{t1} [n] represents the energy amplitude of the n^{th} previous frame with respect to the current frame.

Then the ratio of the variance to the average energy of the N/2 recent sums is computed to obtain the time-domain stability feature ltd_stable_rate. The calculation formula is as follows:

$$ltd\_stable\_rate = \sum_{n=0}^{N/2-1} (Amp_{t2}[n] - \frac{1}{N/2} \sum_{n=0}^{N/2-1} Amp_{t2}[n])^2 / (\sum_{n=0}^{N/2-1} Amp_{t2}[n]^2 + delta)$$

Note that the value of N is different when computing different time-domain

stability features.

Step S412: The tonality features are computed with the spectrum amplitudes. More specifically, they are obtained by computing the correlation coefficient of the amplitude difference of two adjacent frames, or with a further smoothing the correlation coefficient. The tonality features may be computed in the following steps.

a) Compute the amplitudes difference of two adjacent frames. If the difference is smaller than 0, set it to 0. In this way, a group of non-negative spectrum differential coefficients spec_low_dif[] is obtained.

b) Compute the correlation coefficient between the non-negative amplitude difference of the current frame obtained in Step a) and the non-negative amplitude difference of the previous frame to obtain the first tonality features. The calculation formula is as follows:

$$f\_tonality\_rate=\frac{\sum_{i=0}^{N} spec\_low\_dif[i]*pre\_spec\_low\_dif[i]}{\sqrt{\sum_{i=0}^{N} spec\_low\_dif[i]^2 * pre\_spec\_low\_dif[i]^2}},$$

where pre_spec_low_dif is the amplitude difference of the previous frame. Various tonality features can be calculated according to the following formula:

f_tonality_rate[0]=f_tonality_rate;

f_tonality_rate[1]=pre_f_tonality_rate[1]*0.96f+f_tonality_rate*0.04f;

f_tonality_rate[2]=pre_f_tonality_rate[2]*0.90f+f_tonality_rate*0.1f;

where pre_f_tonality_rate is the tonality features of the previous frame.

Step S414: Spectral Flatness Features are the ratio of the geometric mean to the arithmetic mean of certain spectrum amplitude, or this ratio multiplied by a factor. The spectrum amplitude spec_amp[] is smoothed to obtain a smoothed spectrum amplitude: smooth_spec_amp[i] = smooth_spec_amp[i]*fac + spec_amp[i]*(1-fac), 0<=i< SPEC_AMP_NUM. The smoothed spectrum amplitude is divided for three frequency regions, and the spectral flatness features are computed for these three frequency regions. Table 3 shows frequency region division for spectrum flatness.

Table 2 frequency region division of spectrum amplitude for spectral flatness

Spectral flatness number	Start sub-band index	End sub-band index

k	spc_amp_start[k]	spc_amp_end[k]
0	5	19
1	20	39
2	40	64

The spectral flatness features are the ratio of the geometric mean geo_mean[k] to the arithmetic mean ari_mean[k] of the spectrum amplitude or the smoothed spectrum amplitude. The number of the spectrum amplitudes used to compute the spectral flatness feature SFF[k] is N[k]=spec_amp_end[k]-spec_amp_start[k]+1.

$$geo\_mean[k] = (\prod_{n=spec\_amp\_start[k]}^{spe\_amp\_end[k]} smooth\_spec\_amp[n])^{1/N[k]}$$

$$ari\_mean[k] = (\sum_{n=spec\_amp\_start[k]}^{spec\_amp\_end[k]} smooth\_spec\_amp[n]) / N[k]$$

SFF[k]=geo_mean[k]/ari_mean[k]

The spectral flatness features of the current frame are further smoothed to obtain smoothed spectral flatness features sSFM[k]= fac*sSFM[k]+(1-fac) SFF [k].

Step S416: A SNR feature of the current frame is calculated according to the estimated background noise energy of the previous frame, the frame energy feature and the SNR sub-band energy of the current frame. Calculation steps for the frequency domain SNR are as follows.

When a flag of background noise of the previous frame is 1, sub-band background noise energy is updated, update pseudo-codes being as follows:

sb_bg_energy[i] = sb_bg_energy[i]*0.90f + frame_sb_energy[i]*0.1f.

A SNR of each sub-band is calculated according to the sub-band energy of the current frame and the estimated sub-band background noise energy of the previous frame, and the SNR of each sub-band smaller than a certain threshold is set to 0. Specifically,

snr_sub[i]=log2((frame_sb_energy[i]+0.0001f)/(sb_bg_energy[i]+0.0001f)), where snr_sub[i] smaller than -0.1 is set as zero.

An average value of SNRs of all sub-bands is a frequency domain SNR (snr). Specifically,

$$snr = \frac{1}{\text{SNR\_sb\_num}} \sum_{i=0}^{\text{SNR\_sb\_num}-1} snr\_sub[i]$$

Step S418: A flag of noise type is obtained according to a smooth long-time

frequency domain SNR and a long-time SNR lt_snr_org.

The long-time SNR is average energy of long-time active frames and average energy of long-time background noise. The average energy of long-time active frames and the average energy of long-time background noise are updated according to a VAD flag of a previous frame. When the VAD flag is an inactive frame, the average energy of background noise is updated, and when the VAD flag is an active frame, the average energy of long-time active frames is updated. Specifically,

the average energy of long-time active frames is lt_active_eng= fg_energy/fg_energy_count;

the average energy of background noise is lt_inactive_eng= bg_energy/ bg_energy_count,

where $fg\_energy = \sum_{j=0}^{fg\_energy\_count-1} frame\_energy[j]$, i is an active frame index value, $bg\_energy = \sum_{j=0}^{bg\_energy\_count-1} frame\_energy[j]$, and j is an inactive frame index value; and

the long-time SNR is lt_snr_org=log10(lt_active_eng /lt_inactive_eng).

An initial flag of noise type is set to non-silence, and when lf_snr_smooth is greater than a set threshold THR1 and lt_snr_org is greater than a set threshold THR2, the flag of noise type is set to silence.

A calculation process of lf_snr_smooth is shown in Step S420.

The VAD used in Step S418 may be, is not limited to, one VAD in two VADs, and may also be a combined VAD.

Step S420: A calculation method for the smoothed average long-time frequency domain SNR lf_snr_smooth is as follows:

lf_snr_smooth= lf_snr_smooth*fac + (1-fac)*l_snr,

where l_snr = l_speech_snr/l_speech_snr_count - l_silence_snr/l_silence_snr_count,

where l_speech_snr and l_speech_snr_count are respectively an accumulator of frequency domain SNR and a counter for the active frames, and l_silence_snr and l_silence_snr_count are respectively an accumulator of frequency domain SNR and a counter for the inactive frames. When the current frame is an initial frame, initialization is carried out as follows.

l_silence_snr=0.5f;

l_speech_snr=5.0f;

l_silence_snr_count=1; and

l_speech_snr_count=1.

When the current frame is not an initial frame, the above four parameters are updated according to a VAD flag. When the VAD flag indicates that the current frame is an inactive frame, the parameters are updated in accordance with the following formula:

l_silence_snr = l_silence_snr + snr;

l_silence_snr_count = l_silence_snr_count + 1.

When the VAD flag indicates that the current frame is an active frame,

l_speech_snr = l_speech_snr + snr;

l_speech_snr_count = l_speech_snr_count + 1.

The VAD in Step S420 may be, but is not limited to, one VAD in two VADs, and may also be a combined VAD.

Step S422: An initial value is set for the number of continuous noise frames during a first frame, the initial value being set to 0 in this embodiment. During a second frame and subsequent frames, when VAD judgment indicates an inactive frame, the number of continuous inactive frames is added with 1, and otherwise, the number of continuous noise frames is set to 0.

The VAD in Step S422 may be, but is not limited to, one VAD in two VADs, and may also be a combined VAD.

Step S424: A tonality signal flag of the current frame is calculated according to the frame energy feature, tonality feature f_tonality_rate, time-domain stability feature ltd_stable_rate, spectral flatness feature sSFM and spectral centroid feature sp_center of the current frame, and it is judged whether the current frame is a tonal signal. When the current frame is judged to be a tonal signal, the current frame is considered to be a music frame. The following operations are executed.

a) Suppose current frame signal is a non-tonal signal, and a tonality frame flag music_background_frame is used to indicate whether the current frame is a tonal frame. When the value of music_background_frame is 1, it represents that the current frame is a tonal frame, and when the value of music_background_frame is 0, it represents that the current frame is non-tonal.

b) If the tonality feature f_tonality_rate[0] or its smoothed value f_tonality_rate[1] is greater than their respectively preset thresholds, Step c) is executed, and otherwise, Step d) is executed.

c) If time-domain stability feature ltd_stable_rate[5] is smaller than a set threshold, a spectral centroid feature sp_center[0] is greater than a set threshold and one of three spectral flatness features is smaller than its threshold, it is determined that the current frame is a tonal frame, the value of the tonality frame flag music_background_frame is set to 1, and Step d) is further executed.

d) A tonal level feature music_background_rate is updated according to the tonality frame flag music_background_frame, an initial value of the tonal level feature music_background_rate is set when a VAD apparatus starts to work, in the region [0, 1].

If the current tonality frame flag indicates that the current frame is a tonal frame, the tonal level feature music_background_rate is updated using the following formula:

music_background_rate = music_background_rate*fac + （1-fac）.

If the current frame is not a tonal frame, the tonal level feature music_background_rate is updated using the following formula:

music_background_rate = music_background_rate*fac.

e) It is judged whether the current frame is a tonal signal according to the updated tonal level feature music_background_rate, and the value of the tonality signal flag music_background_f is set correspondingly.

If the tonal level feature music_background_rate is greater than a set threshold, it is determined that the current frame is a tonal signal, and otherwise, it is determined that the current frame is a non-tonal signal.

Step S426: The average total SNR of all sub-bands is an average of SNR over all sub-bands for a plurality of frames. A calculation method is as follows.

When a background update flag of the previous frame is 1, frame energy of the current frame is accumulated to a background noise energy accumulator of all sub-bands, and the value of a background noise energy counter of all sub-bands tbg_energy_count is added with 1.

Background noise energy of all sub-bands is calculated according to the following formula: t_bg_energy= t_bg_energy_sum/tbg_energy_count.

An SNR of all sub-bands for the current frame is calculated according to the frame energy of the current frame.

tsnr= log2(frame_energy+0.0001f)/(t_bg_energy+0.0001f).

SNRs of all sub-bands for a plurality of frames are averaged to obtain an average total SNR of all sub-bands.

$$snr\_flux = \frac{1}{N}\sum_{i=0}^{N-1} tsnr[i]$$

,

where N represents N latest frames, and tsnr[i] represents tsnr of the i^{th} frame.

Step S428: An initial value is set for the number of continuous active frames during a first frame. The initial value is set to 0 in this embodiment. When the current frame is the second frame and a speech frame behind the second frame, a current number of continuous active frames is calculated according to a VAD judgment result. Specifically,

When the VAD flag is 1, the number of continuous active frames is added with 1, and otherwise, the number of continuous active frames is set to 0.

The VAD in Step S428 may be, but is not limited to, one VAD in two VADs, and may also be a combined VAD.

Step S430: An initial flag of background noise of the current frame is calculated according to the frame energy feature, spectral centroid feature, time-domain stability feature, spectral flatness feature and tonality feature of the current frame, the initial flag of background noise is modified according to a VAD judgment result, tonality feature, SNR feature, tonality signal flag and time-domain stability feature of the current frame to obtain a final flag of background noise, and background noise detection is carried out according to the flag of background noise.

The flag of background noise is used for indicating whether to update background noise energy, and the value of the flag of background noise is set to 1 or 0. When the value of the flag of background noise is 1, the background noise energy is updated, and when the value of the flag of background noise is 0, the background noise energy is not updated.

Firstly, suppose the current frame is a background noise frame, and when any of the following conditions is satisfied, it can be determined that the current

frame is not a noise signal.

The time-domain stability feature ltd_stable_rate[5] is greater than a set threshold which ranges from 0.05 to 0.30.

The spectral centroid feature sp_center[0] and the time-domain stability feature ltd_stable_rate[5] are greater than corresponding thresholds, respectively, the threshold corresponding to sp_center[0] ranges from 2 to 6, and the threshold corresponding to ltd_stable_rate[5] ranges from 0.001 to 0.1.

The tonality feature f_tonality_rate[1] and the time-domain stability feature ltd_stable_rate[5] are greater than corresponding thresholds, respectively, the threshold corresponding to f_tonality_rate[1] ranges from 0.4 to 0.6, and the threshold corresponding to ltd_stable_rate[5] ranges from 0.05 to 0.15.

The spectral flatness features of each sub-band or the smoothed spectral flatness features of each sub-band are smaller than correspondingly set thresholds which range from 0.70 to 0.92.

The frame energy frame_energy of the current frame is greater than a set threshold, the threshold ranges from 50 to 500, or the threshold is dynamically set according to long-time average energy.

The tonality feature f_tonality_rate is greater than a corresponding threshold.

The initial flag of background noise can be obtained by Step a) to Step f), and then the initial flag of background noise is modified. When the SNR feature, the tonality feature and the time-domain stability feature are smaller than corresponding thresholds, and when vad_flag and music_background_f are set to 0, the background noise update flag is updated to 1.

The VAD in Step S430 may be, but is not limited to, one VAD in two VADs, and may also be a combined VAD.

Step S432: A final combined VAD judgment result is obtained according to at least one feature in the feature category 1, at least one feature in the feature category 2 and two existing VAD judgment results.

In the following exemplary embodiment, the two existing VADs are VAD_A and VAD_B, output flags are respectively vada_flag and vadb_flag, and an output flag of a combined VAD is vad_flag. When the VAD flag is 0, it is indicative of an inactive frame, and when the VAD flag is 1, it is indicative of an active frame. A specific judgment process is as follows.

a) vadb_flag is selected as an initial value of vad_flag.

b) If the flag of noise type indicates that the noise type is silence, a frequency domain SNR is greater than a set threshold such as 0.2 and the initial value of vad_flag of the combined VAD is 0, vada_flag is selected as the combined VAD, and the judgment ends; and otherwise, Step c) is executed.

c) If the smoothed average long-time frequency domain SNR is smaller than a set threshold such as 10.5, or the noise type is not silence, Step d) is executed, and otherwise, the initial value of vad_flag selected in Step a) is selected as the combined VAD judgment result.

d) If any one of the following conditions is satisfied, a result of logical operation OR of the two VADs is used as the combined VAD, and the judgment ends; and otherwise, Step e) is executed.

Condition 1: An average total SNR of all sub-bands is greater than a first threshold such as 2.2.

Condition 2: An average total SNR of all sub-bands is greater than a second threshold such as 1.5, and the number of continuous active frames is greater than a threshold such as 40.

Condition 3: A tonality signal flag is 1.

e) If the flag of noise type indicates that the noise type is silence, vada_flag is selected as the combined VAD, and the judgment ends.

Embodiment 2:

Step S432 in the embodiment 1 may also be implemented in accordance with the following modes.

A final combined VAD judgment result is obtained according to at least one feature in a feature category 1, at least one feature in a feature category 2 and two existing VAD judgment results.

In the present exemplary embodiment, the two existing VADs are VAD_A and VAD_B, output flags are respectively vada_flag and vadb_flag, and an output flag of a combined VAD is vad_flag. When the VAD flag is 0, it is indicative of an inactive frame, and when the VAD flag is 1, it is indicative of an active frame. A specific judgment process is as follows.

a) vadb_flag is selected as an initial value of vad_flag.

b) If a noise type is silence, a frequency domain SNR is greater than a set

threshold such as 0.2 and the initial value of vad_flag of the combined VAD is 0, vada_flag is selected as the combined VAD, and the judgment ends; and otherwise, Step c) is executed.

c) If a smoothed average long-time frequency domain SNR is smaller than a set threshold such as 10.5 or the noise type is not silence, Step d) is executed, and otherwise, the initial value of vad_flag selected in Step a) is selected as a combined VAD judgment result.

d) If any one of the following conditions is satisfied, a result of logical operation OR of the two VADs is used as the combined VAD, and the judgment ends; and otherwise, Step e) is executed.

Condition 1: An average total SNR of all sub-bands is greater than a first threshold such as 2.0.

Condition 2: An average total SNR of all sub-bands is greater than a second threshold such as 1.5, and the number of continuous active frames is greater than a threshold such as 30.

Condition 3: A tonality signal flag is 1.

e) vada_flag is selected as the combined VAD, and the judgment ends.

Embodiment 3:

Step S432 in the embodiment 1 may also be implemented in accordance with the following modes.

A final combined VAD judgment result is obtained according to at least one feature in a feature category 1, at least one feature in a feature category 2 and two existing VAD judgment results.

In the present exemplary embodiment, the two existing VADs are VAD_A and VAD_B, output flags are respectively vada_flag and vadb_flag, and an output flag of a combined VAD is vad_flag. When the VAD flag is 0, it is indicative of an inactive frame, and when the VAD flag is 1, it is indicative of an active frame. A specific judgment process is as follows.

a) vadb_flag is selected as an initial value of vad_flag.

b) If a noise type is silence, Step c) is executed, and otherwise, Step d) is executed.

c) If a smoothed average long-time frequency domain SNR is greater than 12.5 and music_background_f is 0, vad_flag is set as vada_flag, and otherwise,

the initial value of vad_flag selected in Step a) is selected as a combined VAD judgment result.

d) If an average total SNR of all sub-bands is greater than 2.0, or an average total SNR of all sub-bands is greater than 1.5 and the number of continuous active frames is greater than 30, or a tonality signal flag is 1, a result of logical operation OR of the two VADs, i.e., OR (vada_flag, vadb_flag) is used as the combined VAD, and otherwise, the initial value of vad_flag selected in Step a) is selected as a combined VAD judgment result.

Embodiment 4:

Step S432 in the embodiment 1 may also be implemented in accordance with the following modes.

A final combined VAD judgment result is obtained according to at least one feature in a feature category 1, at least one feature in a feature category 2 and two existing VAD judgment results.

In the following exemplary embodiment, the two existing VADs are VAD_A and VAD_B, output flags are respectively vada_flag and vadb_flag, and an output flag of a combined VAD is vad_flag. When the VAD flag is 0, it is indicative of an inactive frame, and when the VAD flag is 1, it is indicative of an active frame. A specific judgment process is as follows.

a) vadb_flag is selected as an initial value of vad_flag.

b) If a noise type is silence, Step c) is executed, and otherwise, Step d) is executed.

c) If a smoothed average long-time frequency domain SNR is greater than 12.5 and music_background_f is 0, vada_flag is set as vad_flag, and otherwise, Step e) is executed.

d) If an average total SNR of all sub-bands is greater than 1.5, or an average total SNR of all sub-bands is greater than 1.0 and the number of continuous active frames is greater than 30, or a tonality signal flag is 1, a result of logical operation OR of two VADs, i.e., OR (vada_flag, vadb_flag), is used as the combined VAD, and otherwise, Step e) is executed.

e) If the number of continuous noise frames is greater than 10 and the average total SNR of all sub-bands is smaller than 0.1, a result of AND operation on the two existing VAD output flags, i.e., AND (vada_flag, vadb_flag), is used as

the combined VAD, and otherwise, vadb_flag is selected as the combined VAD.

Embodiment 5:

Step S432 in the embodiment 1 may also be implemented in accordance with the following modes.

A final combined VAD judgment result is obtained according to at least one feature in a feature category 1, at least one feature in a feature category 2 and two existing VAD judgment results.

In the following exemplary embodiment, the two existing VADs are VAD_A and VAD_B, output flags are respectively vada_flag and vadb_flag, and an output flag of a combined VAD is vad_flag. When the VAD flag is 0, it is indicative of an inactive frame, and when the VAD flag is 1, it is indicative of an active frame. A specific judgment process is as follows.

a) vadb_flag is selected as an initial value of vad_flag.

b) If the noise type is silence, Step c) is executed, and otherwise, Step d) is executed.

c) If music_background_f is 0, the result of logical operation OR of the two VADs, i.e., OR (vada_flag, vadb_flag), is used as the combined VAD, and otherwise, vada_flag is selected as the combined VAD.

d) If an average total SNR of all sub-bands is greater than 2.0, or an average total SNR of all sub-bands is greater than 1.5 and the number of continuous active frames is greater than 30, or a tonality signal flag is 1, the result of logical operation OR of the two VADs, i.e., OR (vada_flag, vadb_flag), is used as the combined VAD, and otherwise, the initial value of vad_flag selected in Step a) is selected as a combined VAD judgment result.

In another embodiment, software is also provided, which is arranged to execute the technical solution described in the above embodiments and exemplary implementations.

In another embodiment, a storage medium is also provided. The software is stored in the storage medium. The storage medium includes, but is not limited to, an optical disk, a floppy disk, a hard disk, an erasable memory and the like.

Obviously, those skilled in the art shall understand that all components or all steps in the present disclosure may be implemented using a general calculation apparatus, may be centralized on a single calculation apparatus or may be

distributed on a network composed of a plurality of calculation apparatuses. Optionally, they may be implemented using executable program codes of the calculation apparatuses. Thus, they may be stored in a storage apparatus and executed by the calculation apparatuses, the shown or described steps may be executed in a sequence different from this sequence under certain conditions, or they are manufactured into each integrated circuit component respectively, or a plurality of components or steps therein is manufactured into a single integrated circuit component. Thus, the present disclosure is not limited to a combination of any specific hardware and software.

The above is only the exemplary embodiments of the present disclosure, and is not used to limit the present disclosure. There may be various modifications and variations in the present disclosure for those skilled in the art. Any modifications, equivalent replacements, improvements and the like within the principle of the present disclosure shall fall within the protection scope defined by the appended claims of the present disclosure.

Industrial Applicability

Based on the above technical solution provided by the embodiments of the present disclosure, combined detection can be carried out according to at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results. The technical problems of low detection accuracy of a VAD solution in the related art can be solved, and the accuracy of VAD can be improved, thereby improving the user experience.

CLAIMS:

1. A Voice Activity Detection (VAD) method, comprising:

acquiring (S102) at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results, wherein the first class feature and the second class feature are features used for VAD detection; and

carrying out (S104), according to the first class feature, the second class feature and the at least two existing VAD judgment results, VAD to obtain a combined VAD judgment result;

wherein the second class feature in the second feature category comprises: a flag of noise type, a smoothed average long-time frequency domain SNR, a frequency domain SNR;

carrying out (S104) VAD according to the first class feature, the second class feature and the at least two existing VAD judgment results comprises:

a) selecting one VAD judgment result from the at least two existing VAD judgment results as an initial value of combined VAD;

b) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, and otherwise, executing Step c), wherein the another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, is used for indicating that a VAD judgment result is an active frame or an inactive frame;

c) executing Step d) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, and otherwise, selecting the VAD judgment result selected in Step a) as the combined VAD judgment result;

d) carrying out a logical operation OR on the at least two existing VAD judgment results and using the result of the logical operation OR as the combined VAD judgment result when a preset condition is met, and otherwise, executing Step e); and

Date Reçue/Date Received 2021-04-20

e) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result if the flag of noise type indicates that the noise type is silence, and otherwise, selecting the VAD judgment result selected in Step a) as the combined VAD judgment result.

2. The method as claimed in claim 1, wherein the first class feature in the first feature category comprises at least one of: the number of continuous active frames, an average total signal-to-noise ratio (SNR) of all sub-bands or a tonality signal flag, wherein the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames.

3. The method as claimed in claim 1, wherein the preset condition comprises at least one of:

condition 1: an average total SNR of all sub-bands is greater than a first threshold;

condition 2: the average total SNR of all sub-bands is greater than a second threshold, and the number of continuous active frames is greater than a preset threshold; or

condition 3: a tonality signal flag indicates a tonal signal.

4. The method as claimed in claim 1, wherein the smoothed average long-time frequency domain SNR and the flag of noise type are determined by means of the following modes:

calculating average energy of long-time active frames of a current frame and average energy of long-time background noise of the current frame according to any one VAD judgment result in a combined VAD judgment result of the previous frame of the current frame or at least two existing VAD judgment results corresponding to the previous frame, average energy of long-time active frames of the previous frame within a first preset time period and average energy of long-time background noise of the previous frame;

calculating a long-time SNR of the current frame within a second time period according to the average energy of long-time background noise and average energy of long-time active frames of the current frame within the second preset time period;

Date Reçue/Date Received 2021-04-20

calculating a smoothed average long-time frequency domain SNR of the current frame within a third preset time period according to any one VAD judgment result in the combined VAD judgment result of the current frame or at least two existing VAD judgment results corresponding to the previous frame and average frequency domain SNR of the previous frame; and

determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR.

5. The method as claimed in claim 4, wherein determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR comprises:

setting the flag of noise type to non-silence, and setting, when the long-time SNR is greater than a first preset threshold and the smoothed average long-time frequency domain SNR is greater than a second preset threshold, the flag of noise type to silence.

6. A Voice Activity Detection (VAD) method, comprising:

acquiring (S102) at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results, wherein the first class feature and the second class feature are features used for VAD detection; and

carrying out (S104), according to the first class feature, the second class feature and the at least two existing VAD judgment results, VAD to obtain a combined VAD judgment result, wherein the second class feature in the second feature category comprises: a flag of noise type, a smoothed average long-time frequency domain SNR, a frequency domain SNR;

carrying out (S104) VAD according to the first class feature, the second class feature and the at least two existing VAD judgment results comprises:

a) selecting one VAD judgment result from the at least two existing VAD judgment results as an initial value of combined VAD;

b) selecting another VAD judgment result from the at least two existing VAD judgement results, which is not selected as the initial value, as the combined VAD judgment result if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value

Date Reçue/Date Received 2021-04-20

indicates an inactive frame, and otherwise, executing Step c), wherein the another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, is used for indicating that a VAD judgment result is an active frame or an inactive frame;

c) executing Step d) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, and otherwise, selecting the VAD judgment result selected in Step a) as the combined VAD judgment result;

d) carrying out a logical operation OR on the at least two existing VAD judgment results and using the result of the logical operation OR as the combined VAD judgment result when a preset condition is met, and otherwise, executing Step e); and

e) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result.

7. The method as claimed in claim 6, wherein the first class feature in the first feature category comprises at least one of: a number of continuous active frames, an average total signal-to-noise ratio (SNR) of all sub-bands or a tonality signal flag, wherein the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames.

8. The method as claimed in claim 6, wherein the preset condition comprises at least one of:

condition 1: an average total SNR of all sub-bands is greater than a first threshold;

condition 2: the average total SNR of all sub-bands is greater than a second threshold, and the number of continuous active frames is greater than a preset threshold; or

condition 3: a tonality signal flag indicates a tonal signal.

Date Reçue/Date Received 2021-04-20

9. The method as claimed in claim 6, wherein the smoothed average long-time frequency domain SNR and the flag of noise type are determined by means of the following modes:

calculating average energy of long-time active frames of a current frame and average energy of long-time background noise of the current frame according to any one VAD judgment result in a combined VAD judgment result of the previous frame of the current frame or at least two existing VAD judgment results corresponding to the previous frame, average energy of long-time active frames of the previous frame within a first preset time period and average energy of long-time background noise of the previous frame;

calculating a long-time SNR of the current frame within a second time period according to the average energy of long-time background noise and average energy of long-time active frames of the current frame within the second preset time period;

calculating a smoothed average long-time frequency domain SNR of the current frame within a third preset time period according to any one VAD judgment result in the combined VAD judgment result of the current frame or at least two existing VAD judgment results corresponding to the previous frame and average frequency domain SNR of the previous frame; and

determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR.

10. The method as claimed in claim 9, wherein determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR comprises: setting the flag of noise type to non-silence, and setting, when the long-time SNR is greater than a first preset threshold and the smoothed average long-time frequency domain SNR is greater than a second preset threshold, the flag of noise type to silence.

11. A Voice Activity Detection (VAD) apparatus, comprising:

an acquisition component (20), arranged to acquire at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results, wherein the first class feature and the second class feature are features used for VAD detection; and

Date Reçue/Date Received 2021-04-20

a detection component (22), arranged to carry out, according to the first class feature, the second class feature and the at least two existing VAD judgment results, VAD to obtain a combined VAD judgment result;

wherein the second class feature in the second feature category comprises: a flag of noise type, a smoothed average long-time frequency domain SNR, a frequency domain SNR, and the detection component (22) is arranged to carry out VAD as follows:

a) selecting one VAD judgment result from the at least two existing VAD judgment results as an initial value of combined VAD;

b) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, and otherwise, executing Step c), wherein the another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, is used for indicating that a VAD judgment result is an active frame or an inactive frame;

c) executing Step d) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, and otherwise, selecting the VAD judgment result selected in Step a) as the combined VAD judgment result;

d) carrying out a logical operation OR on the at least two existing VAD judgment results and using the result of the logical operation OR as the combined VAD judgment result when a preset condition is met, and otherwise, executing Step e); and

e) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result if the flag of noise type indicates that the noise type is silence, and otherwise, selecting the VAD judgment result selected in Step a) as the combined VAD judgment result.

Date Reçue/Date Received 2021-04-20

12. A Voice Activity Detection, VAD, apparatus, comprising:

an acquisition component (20), arranged to acquire at least one first class feature in a first feature category, at least one second class feature in a second feature category and at least two existing VAD judgment results, wherein the first class feature and the second class feature are features used for VAD detection; and

a detection component (22), arranged to carry out, according to the first class feature, the second class feature and the at least two existing VAD judgment results, VAD to obtain a combined VAD judgment result;

wherein the second class feature in the second feature category comprises: a flag of noise type, a smoothed average long-time frequency domain SNR, a frequency domain SNR, and the detection component (22) is arranged to carry out VAD as follows:

a) selecting one VAD judgment result from the at least two existing VAD judgment results as an initial value of combined VAD;

b) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result if the flag of noise type indicates that the noise type is silence, the frequency domain SNR is greater than a preset threshold and the initial value indicates an inactive frame, and otherwise, executing Step c), wherein the another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, is used for indicating that a VAD judgment result is an active frame or an inactive frame;

c) executing Step d) if the smoothed average long-time frequency domain SNR is smaller than a preset threshold or the noise type is not silence, and otherwise, selecting the VAD judgment result selected in Step a) as the combined VAD judgment result;

d) carrying out a logical operation OR on the at least two existing VAD judgment results and using the result of the logical operation OR as the combined VAD judgment result when a preset condition is met, and otherwise, executing Step e); and

e) selecting another VAD judgment result from the at least two existing VAD judgment results, which is not selected as the initial value, as the combined VAD judgment result.

Date Reçue/Date Received 2021-04-20

13. The apparatus as claimed in claim 11 or 12, wherein the acquisition component (20) comprises:

a first acquisition unit (200), arranged to acquire the first class feature in the first feature category which comprises at least one of: a number of continuous active frames, an average total signal-to-noise ratio (SNR) of all sub-bands or a tonality signal flag, wherein the average total SNR of all sub-bands is an average of SNR over all sub-bands for a predetermined number of frames; and

a second acquisition unit (202), arranged to acquire the second class feature in the second feature category which comprises at least one of: a flag of noise type, a smoothed average long-time frequency domain SNR, the number of continuous noise frames or a frequency domain SNR.

14. The apparatus as claimed in claim 11 or 12, wherein the preset condition comprises at least one of:

condition 1: an average total SNR of all sub-bands is greater than a first threshold;

condition 2: the average total SNR of all sub-bands is greater than a second threshold, and the number of continuous active frames is greater than a preset threshold; and

condition 3: a tonality signal flag indicates a tonal signal.

15. The apparatus as claimed in claim 13, wherein the smooth long-time average frequency domain signal-to-noise ratio and the noise type flag are determined by means of the following modes:

calculating average active audio frame energy of a current frame and average background noise energy of the current frame according to any one VAD judgment result in a combined VAD judgment result of a previous frame of the current frame or at least two existing VAD judgment results corresponding to the previous frame, average active audio frame energy of the previous frame within a first preset time period and average background noise energy of the previous frame;

calculating a long-time signal-to-noise ratio of the current frame within a second time period according to the average background noise energy and average active audio frame energy of the current frame within the second preset time period;

Date Reçue/Date Received 2021-04-20

calculating a smooth long-time average frequency domain signal-to-noise ratio of the current frame within a third preset time period according to any one VAD judgment result in the combined VAD judgment result of the current frame or at least two existing VAD judgment results corresponding to the previous frame and a frequency domain signal-to-noise ratio of the previous frame; and

determining the noise type flag according to the long-time signal-to-noise ratio and the smooth long-time average frequency domain signal-to-noise ratio;

wherein preferably, determining the flag of noise type according to the long-time SNR and the smoothed average long-time frequency domain SNR comprises:

setting the flag of noise type to non-silence, and setting, when the long-time SNR is greater than a first preset threshold and the smoothed average long-time frequency domain SNR is greater than a second preset threshold, the flag of noise type to silence.

Date Reçue/Date Received 2021-04-20

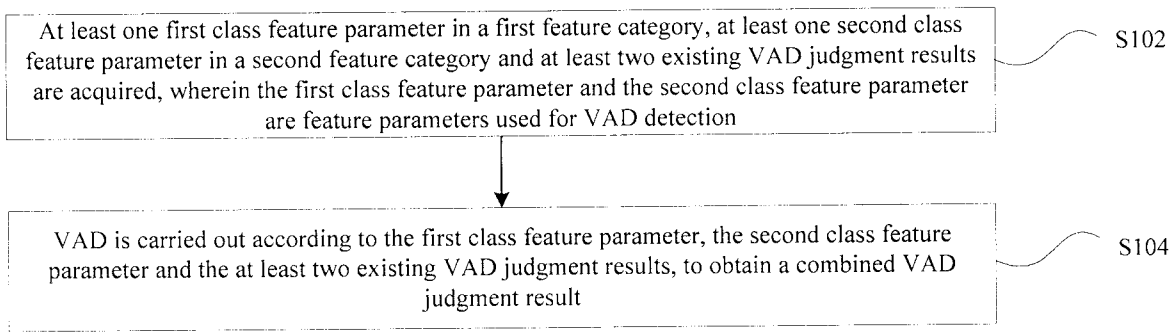


Fig. 1

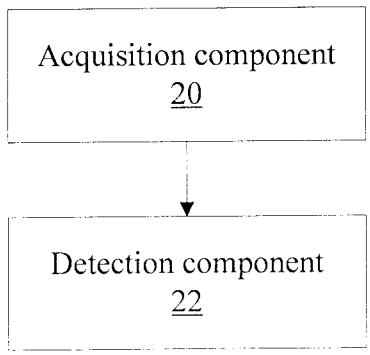


Fig. 2

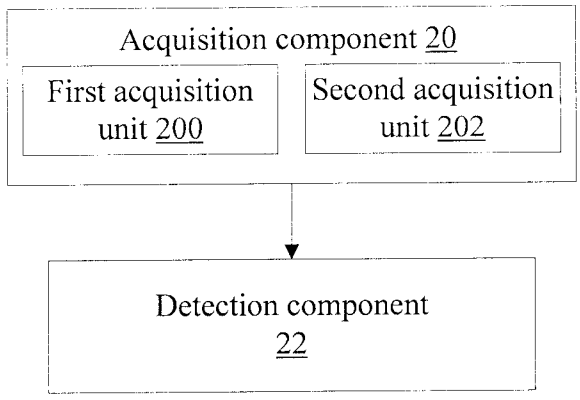


Fig. 3

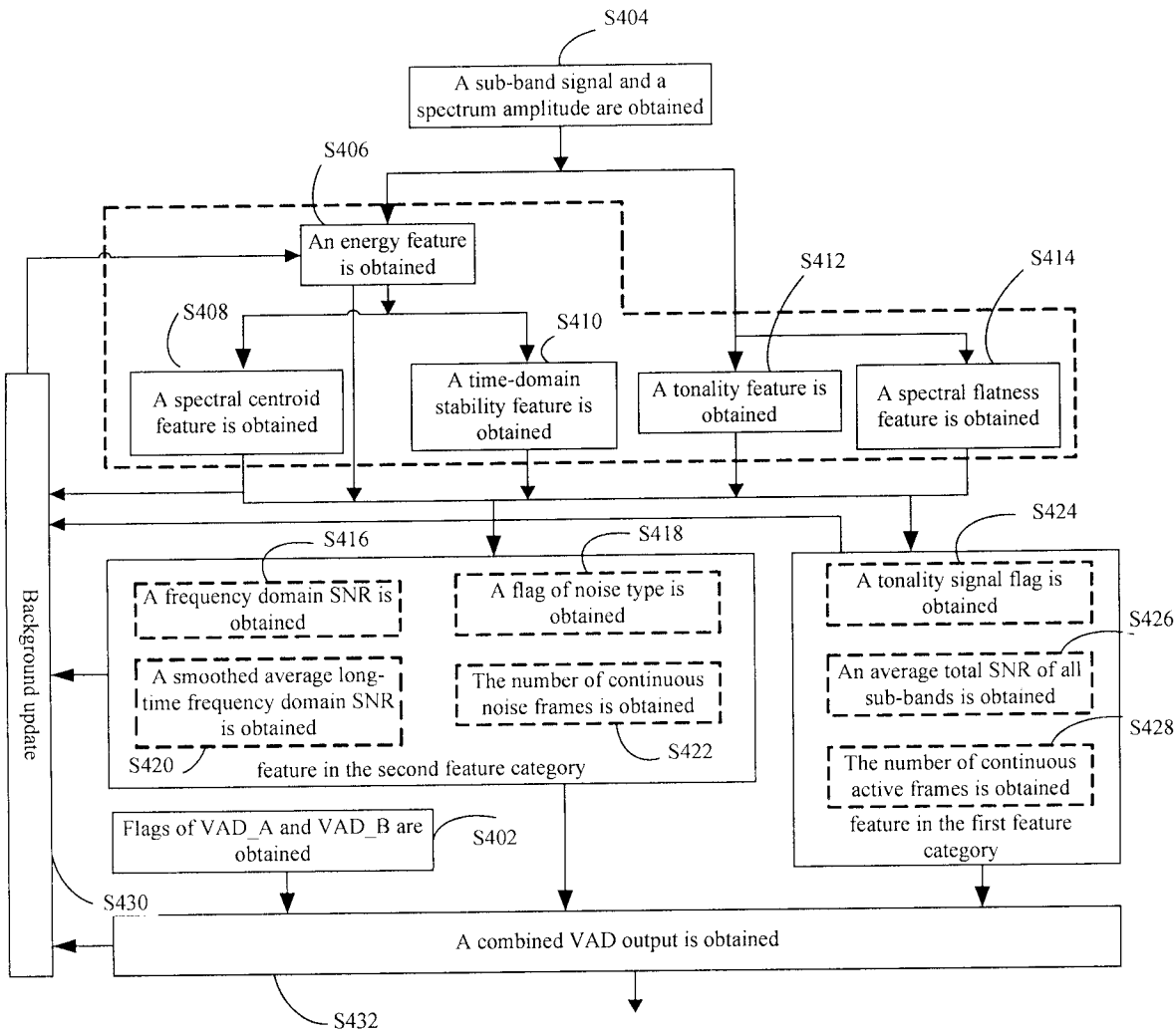
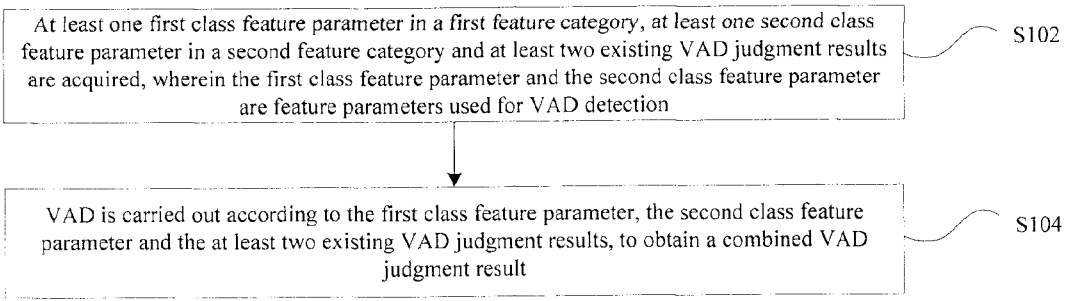

S404
A sub-band signal and a spectrum amplitude are obtained
S406
An energy feature is obtained
S408
A spectral centroid feature is obtained
S410
A time-domain stability feature is obtained
S412
A tonality feature is obtained
S414
A spectral flatness feature is obtained
S416
A frequency domain SNR is obtained
S418
A flag of noise type is obtained
A smoothed average long-time frequency domain SNR is obtained
S420
The number of continuous noise frames is obtained
S422
feature in the second feature category
S424
A tonality signal flag is obtained
S426
An average total SNR of all sub-bands is obtained
S428
The number of continuous active frames is obtained
feature in the first feature category
Flags of VAD_A and VAD_B are obtained
S402
Background update
S430
A combined VAD output is obtained
S432

Fig. 4


At least one first class feature parameter in a first feature category, at least one second class feature parameter in a second feature category and at least two existing VAD judgment results are acquired, wherein the first class feature parameter and the second class feature parameter are feature parameters used for VAD detection
S102
VAD is carried out according to the first class feature parameter, the second class feature parameter and the at least two existing VAD judgment results, to obtain a combined VAD judgment result
S104