(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 June 2012 (28.06.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/083554 A1

(51) International Patent Classification:
G10L 11/02 (2006.01)

(21) International Application Number:
PCT/CN2010/080222

(22) International Filing Date:
24 December 2010 (24.12.2010)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **HUAWEI TECHNOLOGIES CO., LTD.** [CN/CN]; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **WANG, Zhe** [CN/CN]; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN).

(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, 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, JP, 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, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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))*

(54) **Title**: A METHOD AND AN APPARATUS FOR PERFORMING A VOICE ACTIVITY DETECTION

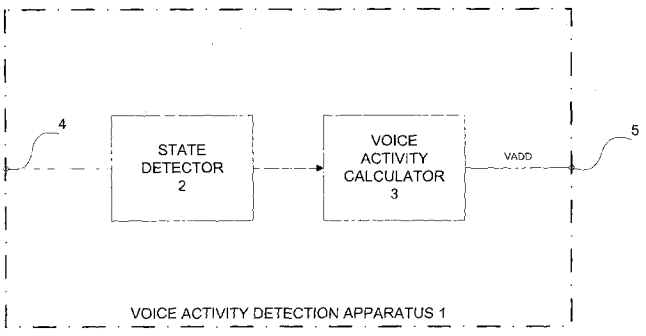

Fig.1

(57) **Abstract**: A voice activity detection apparatus (1) for determining a voice activity detection decision (VADD) for an input audio signal, wherein the voice activity detection apparatus (1) comprises a state detector (2) adapted to determine a current working state (WS) of at least two different working states of the voice activity detection apparatus (1) dependent on the input audio signal wherein each of the at least two different working states (WS) is associated with a corresponding working state parameter decision set (WSPDS) including at least one voice activity decision parameter (VADP) and a voice activity calculator (3) adapted to calculate a voice activity detection parameter value for the at least one VADP of the working state parameter decision set (WSPDS) associated with the current working state (WS) and to determine the voice activity detection decision (VADD) by comparing the calculated voice activity detection parameter value of the respective voice activity decision parameter (VADP) with a threshold.

WO 2012/083554 A1

TITLE

A method and an apparatus for performing a voice activity detection

TECHNICAL BACKGROUND

The invention relates to a method and an apparatus for performing a voice activity detection and in particular to a voice activity detection apparatus having at least two different working states using non-linearly processed sub-band segmental signal to noise ratio parameters.

Voice activity detection (VAD) is generally a technique which is provided to detect a voice activity in a signal. Voice activity detection is also known as a speech activity detection or simply speech detection. The function of VAD is to detect in communication channels the presence of absence of active signals such as speech or music. Networks thus can decide to compress a transmission bandwidth in periods where active signals are absent or perform other processing according to whether there is an active signal or not. In VAD a feature parameter or a set of feature parameters extracted from the input audio signal can be compared to corresponding threshold values to determine whether the input audio signal is an active signal or not based on the comparison result. There have been many parameters proposed for VAD. In general, energy based parameters are known to provide good performance. Thus, in recent years sub-band SNR based parameters as a kind of energy based parameters have been widely used for VAD. No matter what feature parameter or feature parameters are used by a voice activity detector these parameters exhibit a weak speech characteristic at the offsets of speech bursts, thus increasing the possibility of mis-detecting speech offsets. Usually, in order to ensure a correct detection of speech offsets a conventional voice activity detector performs some special processing at speech offsets. A conventional way to do this special processing is to apply a “hard” hangover to the VAD decision at speech offsets wherein the first group of frames detected as inactive by the voice activity detector at speech offsets is forced to active. Another possibility is to apply a “soft” hangover to the voice activity detection decision at speech offsets. In applying a soft hangover the VAD decision threshold at speech offsets is adjusted to favour speech detection for the first several offset frames of the audio signal. Accordingly, in this conventional voice activity detector when the input signal is a non speech offset signal the VAD decision is made in a normal way while in an offset state the VAD decision is made in a way favouring speech detection.

Although the application of a hard hangover process in order to ensure a correct detection of speech offsets can successfully help to diminish the possibility of a mis-detection at speech offsets the hard hangover scheme lacks efficiency. Many real inactive frames will be unnec-

essarily forced to active thus decreasing the VAD overall performance. On the other hand, although a soft hangover processing scheme as used for instance by the G.718 ITU-T standardized voice activity detector improves the hangover efficiency to a higher level the VAD performance can be still improved.

Accordingly, it is a goal of the present invention to provide a method and an apparatus for VAD which provide a higher VAD performance than conventional VAD apparatuses and methods.

SUMMARY OF THE INVENTION

According to a first aspect of the present invention a voice activity detection (VAD) apparatus for determining a VAD decision (VADD) for an input audio signal is provided
wherein the VAD apparatus comprises
a state detector adapted to determine a current working state (WS) of at least two different working states of the VAD apparatus dependent on the input audio signal,
wherein each of the at least two different working states (WS) is associated with a corresponding working state parameter decision set (WSPDS) including at least one VAD parameter (VADP); and
a voice activity calculator adapted to calculate a VAD parameter value for the VAD parameter (VADP) of the working state parameter decision set (WSPDS) associated with the current working state (WS) and to determine the VAD decision (VADD) by comparing the calculated VAD parameter value with a threshold.

Accordingly, the VAD apparatus according to the first aspect of the present invention comprises more than one working state (WS). The VAD apparatus according to the first aspect of the present invention uses at least two different parameters or two different sets of parameters for making VAD decisions for different working states.

In a possible implementation the VAD parameters can have the same general form but can comprise different factors. In a possible implementation the different VAD parameters can comprise modified sub-band segmental signal to noise ratio (SNR) based parameters which are non-linearly processed in a different manner.

The number of working states used by the VAD apparatus according to the first aspect of the present invention can vary. In a possible implementation of the VAD apparatus the apparatus comprises two different working states, i.e. a normal working state (NWS) and an offset working state (OWS).

In a possible implementation of the VAD apparatus according to the first aspect of the present invention for each working state (WS) of the VAD apparatus a corresponding working state parameter decision set (WSPDS) is provided each comprising at least one VAD parameter (VADP). The number and type of VAD parameters (VADPs) can vary for the different working state parameter decision sets (WSPDS) of the different working states (WS) of the VAD apparatus according to the first aspect of the present invention.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD decision (VADD) determined by said voice activity calculator is determined or calculated by using sub-band segmental signal to noise ratio (SNR) based VAD parameters (VADPs).

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD decision (VADD) for said input audio signal is determined by said voice activity calculator on the basis of the at least one VAD parameter (VADP) of the working parameter decision set (WSPDS) provided for the current working state (WS) of said VAD apparatus using a predetermined VAD processing algorithm provided for the current working state (WS) of said VAD apparatus. The used VAD processing algorithm can be reconfigured or configurable via an interface thus providing more flexibility for the VAD apparatus according to the first aspect of the present invention.

In a possible implementation of the VAD apparatus according to the present invention the VAD processing algorithm used for determining the VAD decision (VADD) can be adapted.

In a further possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD apparatus is switchable between different working states (WS) according to configurable working state transition conditions. This switching can be performed in a possible implementation under the control of the state detector.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD apparatus comprises a normal working state (NWS) and an offset working state (OWS) and can be switched between these two different working states according to configurable working state transition conditions.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD apparatus detects a change from voice activity being present to a voice activity being absent and/or switches from a normal working state (NWS) to an offset working state (OWS) in said input audio signal if in the normal working state (NWS) of said VAD

apparatus the VAD decision (VADD) determined on the basis of the at least one VAD parameter (VADP) of the normal working state parameter decision set (NWSPDS) of said normal working state (NWS) indicates a voice activity being present for a previous frame and a voice activity being absent in a current frame of said input audio signal. In a possible implementation of the VAD apparatus according to the first aspect of the present invention the VADD said VAD apparatus detects in its normal working state (NWS) forms an intermediate VADD ($VADD_{int}$), which may form the VADD or final VADD output by the VAD apparatus in case this intermediate VAD indicates that voice activity is present in the current frame. As described above, in case this intermediate VADD indicates that no voice activity is present in the current frame, this intermediate VADD may be used to detect a transition or change from a normal working state to an offset working state and to switch to the offset working state where the voice activity detector calculates for the current frame a voice activity voice detection parameter of the offset working state parameter decision set to determine the VADD or final VADD output by the VAD apparatus.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention if said VAD apparatus detects in its normal working state (NWS) that a voice activity is present in a current frame of said input audio signal this intermediate VAD decision ($VADD_{int}$) is output as a final VAD decision ($VADD_{fin}$).

In a further possible implementation of the VAD apparatus according to the first aspect of the present invention, wherein if said VAD apparatus detects in its normal working state (NWS) that a voice activity is present in the previous frame and that a voice activity is absent in a current frame of said input signal it is switched from its normal working state (NWS) to an offset working state (OWS) wherein the VAD decision (VADD) is determined on the basis of the at least one VAD parameter of the offset working state parameter decision set (OWSPDS).

In a still further possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD decision (VADD) determined in the offset working state (OWS) of said VAD apparatus forms the final VADD or VAD decision (VADD) output by the VAD apparatus if the VAD decision (VADD) determined on the basis of the at least one VAD parameter (VADP) of the offset working state parameter decision set (OWSPDS) indicates that a voice activity is present in the current frame of the input audio signal.

In a still further possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD decision (VADD) determined in the offset working state (OWS) of said VAD apparatus forms an intermediate VAD decision (VAD_{int}) if the VAD decision (VADD) determined on the basis of the at least one VAD parameter (VADP) of the offset

working state parameter decision set (OWSPDS) indicates that a voice activity is absent in the current frame of the input audio signal.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the intermediate VAD decision ($VADD_{int}$) undergoes a hard hangover processing to provide a final VAD decision ($VADD_{fin}$).

In a further possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD apparatus is switched from the normal working state (NWS) to the offset working state (OWS) if the VAD decision (VADD) determined by the voice activity calculator of said VAD apparatus in the normal working state (NWS) using a VAD processing algorithm and the working state parameter decision set (NWSPDS) provided for said normal working state (NWS) indicates an absence of voice in the input audio signal and a soft hangover counter (SHC) exceeds a predetermined threshold counter value.

In a further possible implementation of the VAD apparatus according to the first aspect of the present invention said VAD apparatus is switched from the offset working state (OWS) to the normal working state (NWS) if the soft hangover counter (SHC) does not exceed a predetermined threshold counter value.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the input audio signal consists of a sequence of audio signal frames and the soft hangover counter (SHC) is decremented in the offset working state (OWS) of said VAD apparatus for each received audio signal frame until the predetermined threshold counter value is reached.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention if a predetermined number of consecutive active audio signal frames of the input audio signal is detected the soft hangover counter (SHC) is reset to a counter value depending on a long term signal to noise ratio (lSNR) of the input audio signal.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention an active audio signal frame is detected if a calculated voice metric of the audio signal exceeds a predetermined voice metric threshold value and a pitch stability of said audio signal frame is below a predetermined stability threshold value.

In a possible implementation of the VAD apparatus according to the first aspect of the present invention the VAD parameters of a working state parameter decision set (WSPDS) of a

working state of said activity detection apparatus comprises energy based decision parameters and/or spectral envelope based parameters and/or entropy based decision parameters and/or statistic based decision parameters.

In a further possible implementation of the VAD apparatus according to the first aspect of the present invention an intermediate VAD decision ($VADD_{int}$) determined by said voice activity calculator of said VAD apparatus is applied to a hard hangover processing unit performing a hard hangover of said applied intermediate VAD decision ($VADD_{int}$).

According to a second aspect of the present invention an audio signal processing device is provided comprising a VAD apparatus according to the first aspect of the present invention and comprising an audio signal processing unit controlled by a VAD decision (VADD) generated by said VAD apparatus.

According to a third aspect of the present invention a method for performing a VAD is provided wherein a VAD decision (VADD) is calculated by a VAD apparatus for an input audio signal using at least one VAD parameter (VADP) of a working state parameter decision set (WSPDS) of a current working state detected by a state detector of said VAD apparatus.

BRIEF DESCRIPTION OF THE FIGURES

In the following possible implementations of different aspects of the present invention are described with reference to the enclosed figures.

Fig. 1 shows a block diagram of a VAD apparatus according to a possible implementation of the VAD apparatus according to the first aspect of the present invention.

Fig. 2 shows a block diagram of a possible implementation of an audio signal processing apparatus according to a second aspect of the present invention.

DETAILED DESCRIPTION OF EMBODIMENTS

Fig. 1 shows a block diagram of a possible implementation of a VAD apparatus 1 according to a first aspect of the present invention. As can be seen in fig. 1 the VAD apparatus 1 according to the first aspect of the present invention comprises in the exemplary implementation a state detector 2 and a voice activity calculator 3. The VAD apparatus 1 is provided for determining a VAD decision VADD for a received input audio signal applied to an input 4 of the VAD apparatus 1. The determined VAD decision VADD is output at an output 5 of the VAD apparatus

1. The state detector 2 is adapted to determine a current working state WS of the VAD apparatus 1 dependent on the input audio signal applied to the input 4. The VAD apparatus 1 according to the first aspect of the present invention comprises at least two different working states WS. In a possible implementation the VAD apparatus 1 comprises for example two working states WS. Each of the at least two different working states WS is associated with a corresponding working state parameter decision set WSPDS which includes at least one VAD parameter VADP.

The VAD apparatus 1 comprises in the shown implementation of fig. 1 further a voice activity calculator 3 which is adapted to calculate a VAD parameter value for the at least one VAD parameter VADP of the working state parameter decision set WSPDS associated with the current working state WS of the VAD apparatus 1. This calculation is performed to determine a VAD decision VADD by comparing the calculated VAD parameter value of the at least one VAD parameter with a corresponding threshold.

The state detector 2 as well as the voice activity calculator 3 of the VAD apparatus 1 can be hardware or software implemented. The VAD apparatus 1 according to the first aspect of the present invention has more than one working state. At least two different VAD parameters or two different sets of VAD parameters are used by the VAD apparatus 1 for generating the VAD decision VADD for different working states WS.

The VAD decision VADD determined for said input audio signal by said voice activity calculator 3 is determined in a possible implementation on the basis of at least one VAD parameter VADP of the working state parameter decision set WSPDS provided for the current working state WS of the VAD apparatus 1 using a predetermined VAD processing algorithm provided for the current working state WS of the VAD apparatus 1. The state detector 2 detects the current working state WS of the VAD apparatus 1. The determination of the current working state WS is performed by the state detector 2 dependent on the received input audio signal. In a possible implementation the VAD apparatus 1 is switchable between different working states WS according to configurable working state transition conditions. In a possible implementation the VAD apparatus 1 comprises two working states, i.e. a normal working state NWS and an offset working state OWS.

In a possible implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD apparatus 1 detects a change from a voice activity being present to a voice activity being absent in the input audio signal if a corresponding condition is met. If in the normal working state NWS of said VAD apparatus 1 the VAD decision VADD determined by the voice activity calculator 3 of said VAD apparatus 1 on the basis of the at least one VAD

parameter VADP of the normal working state parameter decision set NWSPDS of said normal working state NWS indicates a voice activity being present for a previous frame and a voice activity being absent in a current frame of said input audio signal the VAD apparatus 1 detects a change from voice activity being present in the input audio signal to a voice activity being absent in the input audio signal.

In a possible implementation of the VAD apparatus 1 according to the first aspect if the VAD apparatus 1 detects in its normal working state NWS that a voice activity is present in a current frame of the input audio signal this intermediate VAD decision $VADD_{int}$ can be output as a final VAD decision $VADD_{fin}$ at the output 5 of the VAD apparatus 1 for further processing.

In a further possible implementation of the VAD apparatus 1 according to the first aspect of the present invention if said VAD apparatus 1 detects in its normal working state NWS that a voice activity is present in the previous frame of the input audio signal and that a voice activity is absent in a current frame of the input audio signal it is switched automatically from its normal working state NWS to an offset working state OWS. In the offset working state OWS the VAD decision VADD is determined by the voice activity calculator 3 on the basis of the at least one VAD parameter VADP of the offset working state parameter decision set OWSPDS. The VAD parameters VADPs of the different working state parameter decision sets WSPDS can be stored in a possible implementation in a configuration memory of the VAD apparatus 1.

In a possible implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD decision VADD determined by the voice activity calculator 3 in the offset working state OWS forms an intermediate VAD decision $VADD_{int}$ if the VAD decision VADD determined on the basis of the at least one VAD parameter VADP of the offset working state parameter decision set OWSPDS indicates that a voice activity is absent in the current frame of the input audio signal. In a possible implementation this generated intermediate VAD decision undergoes a hard hangover processing before it is output as a final VAD decision $VADD_{fin}$ at the output 5 of the VAD apparatus 1.

In a possible implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD apparatus 1 is switched automatically from the normal working state NWS to the offset working state OWS if the VAD decision VADD determined by the voice activity calculator 3 of the VAD apparatus 1 in the normal working state NWS using a VAD processing algorithm and the working state parameter decision set WSPDS provided for this normal working state NWS indicates an absence of voice in the input audio signal and if a soft hang-over counter SHC exceeds at the same time a predetermined threshold counter value.

In a further possible implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD apparatus 1 is switched from the offset working state OWS to the normal working state NWS if a soft hangover counter SHC does not exceed at the same time a predetermined threshold counter value.

The input audio signal applied to the input 4 of the VAD apparatus 1 consists in a possible implementation of a sequence of audio signal frames wherein the soft hangover counter SHC employed by the VAD apparatus 1 is decremented in the offset working state OWS of said VAD apparatus 1 for each received audio signal frame until the predetermined threshold counter value is reached. In a possible implementation if a predetermined number of consecutive active audio signal frames of the input audio signal is detected the soft hangover counter SHC is reset to a counter value depending on a long term signal to noise ratio (lSNR) of the received input audio signal. This long term signal to noise ratio (lSNR) can be calculated by a long term signal to noise ratio estimation unit of the VAD apparatus 1. In a possible implementation of the VAD apparatus 1 according to the first aspect of the present invention an active audio signal frame is detected if a calculated voice metric of the audio signal frame exceeds a predetermined voice metric threshold value and a pitch stability of the audio signal frame is below a predetermined stability threshold value.

In a possible implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD parameters VADPs of a working state parameter decision set WSPDS of a working state WS of the VAD apparatus 1 can comprise energy based decision parameters and/or spectral envelope based decision parameters and/or entropy based decision parameters and/or statistic based decision parameters. In a specific implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD decision VADD determined by the voice activity calculator 3 uses sub-band segmental signal to noise ratio (SNR) based VAD parameters VADPs.

In a further possible implementation of the VAD apparatus 1 an intermediate VAD decision VADD determined by the voice activity calculator 3 of the VAD apparatus 1 can be applied to a further hard hangover processing unit performing a hard hangover of the applied intermediate VAD decision VADD.

The VAD apparatus 1 according to the first aspect of the present invention can comprise in a possible implementation two operation states wherein the VAD apparatus 1 operates either in a normal working state NWS or in a offset working state OWS. A speech offset is a short period at the end of the speech burst within the received audio signal. Thus, a speech offset contains relatively low speech energy. A speech burst is a speech period of the input audio signal be-

tween two adjacent speech pauses. The length of a speech offset typically extends over several continuous signal frames and can be sample dependent. The VAD apparatus 1 according to the first aspect of the present invention continuously identifies the starts of speech offsets in the input audio signal and switches from the normal working state NWS to the offset working state OWS when a speech offset is detected and switches back to the normal working state NWS when the speech offset state ends. The VAD apparatus 1 selects one VAD parameter or a set of parameters for the normal working state NWS and another VAD parameter or set of parameters for the offset working state OWS. Accordingly, with a VAD apparatus 1 according to the first aspect of the present invention different VAD operations are performed for different parts of the received audio signal and specific VAD operations are performed for each working state WS. The VAD apparatus 1 according to the first aspect of the present invention performs a speech burst and offset detection in the received audio input signal wherein the offset detection can be performed in different ways according to different implementations of the VAD apparatus 1.

In a possible implementation of the VAD apparatus 1 the input audio signal is segmented into signal frames and inputted to the VAD apparatus 1 at input 4. The input audio signal can for example comprise signal frames of 20ms length. In a possible specific implementation for each input signal frame an open loop pitch analysis can be performed twice each for a sub-frame having 10ms. The pitch lags searched for the two sub-frames of each input frame are denoted as T(0), T(1) respectively and the corresponding correlations are denoted respectively as voicing (0) and voicing(1). The voicing metric(V) of the audio signal frame V(0) is calculated by:

$$V(0) = (voicing(-1) + voicing(0) + voicing(1))/3 + corr\_shift$$

where voicing(-1) represents the corresponding correlation as a pitch lag of the second sub-frame of the previous input signal frame and wherein corr_shift is a compensation value depending on the background noise level.

The pitch stability (S) of said audio signal frame can be calculated by:

$$S_r(0) = [abs(T(-1) - T(-2)) + abs(T(0) - T(-1)) + abs(T(1) - T(0))]/3$$

wherein T(-1), T(-2) are the first and second pitch lags of the previous input signal frame and abs() means the absolute value. In a possible specific implementation the input frame is considered as a voice frame or active frame when the following condition is met:

$$V(0) > 0.65 \&\& S_T(0) < 14$$

In a possible implementation if three consecutive active frames are detected a voiced burst of the input audio signal is detected and a soft hangover counter SHC is reset to non-zero value determined depending on the signal long term SNR lSNR. When the VAD apparatus 1 according to the first aspect of the present invention is working in a normal working state NWS and the determined intermediate VAD decision VADD falls after previous frames have been classified or determined as active to inactive for a current signal frame and if the soft hangover counter SHC is greater than 0 the input audio signal is assumed to enter a speech offset and the VAD apparatus 1 switches from the normal working state NWS into the offset working state OWS. The length of the soft hangover counter SHC defines the length of the VAD offset working state OWS. In a possible implementation the soft hangover counter SHC is decremented or elapsed by one at each signal frame within the VAD speech offset working state OWS. The speech offset working state OWS of the VAD apparatus 1 ends when the software hangover counter SHC decrements to a predetermined threshold value such as 0 and the VAD apparatus 1 switches back to its normal working state NWS at the same time.

In a possible specific implementation three parameters are used by the VAD apparatus 1 for making an intermediate VAD decision $VADD_{int}$. One parameter is the voicing metric (V-1) of the preceding frame and the two other parameters are given by:

$$mssnr_{nor} = \begin{cases} \sum_{i}^{N}(snr(i)+\alpha)^{4} & snr(i)+\alpha \geq 1, lsnr > 18 \\ \sum_{i}^{N}(snr(i)+\alpha)^{10} & snr(i)+\alpha \geq 1, 8 < lsnr \leq 18 \\ \sum_{i}^{N}(snr(i)+\alpha)^{15} & snr(i)+\alpha \geq 1, lsnr \leq 8 \\ \sum_{i}^{N}(snr(i)+\alpha)^{9} & otherwise \end{cases}$$

$$mssnr_{off} = \begin{cases} \sum_{i}^{N}(snr(i)+\alpha+\beta)^{4} & snr(i)+\alpha \geq 1, lsnr > 18 \\ \sum_{i}^{N}(snr(i)+\alpha+\beta)^{10} & snr(i)+\alpha \geq 1, 8 < lsnr \leq 18 \\ \sum_{i}^{N}(snr(i)+\alpha+\beta)^{15} & snr(i)+\alpha \geq 1, lsnr \leq 8 \\ \sum_{i}^{N}(snr(i)+\alpha+\beta)^{9} & otherwise \end{cases}$$

wherein snr(i) is the modified log SNR of the i^{th} spectral sub-band of the input signal frame,

N is the number of sub-bands per frame,
lsnr is the long term SNR estimate and
α, β are two configurable coefficients.
The first coefficient α can be determined in a possible implementation by:

$$\alpha = f(i, lsnr) = a(i) \cdot lsnr + b(i)$$

where a(i) and b(i) are two real or floating numbers determined by the sub-band index i. The second coefficient β can be determined by the voicing metric V(-1) wherein if V(-1)>0.65 β = 0.2 and if V(-1) ≤ 0.65 β = 0.1. In a possible implementation the calculation of the SNR of each sub-band snr(i) is given by:

$$snr(i) = \log_{10}\left(\frac{E(i)}{E_n(i)}\right)$$

wherein E(i) is the energy of the i^{th} sub-band of the input frame,
$E_n(i)$ is the energy of the i^{th} sub-band of the background noise estimate.

In a possible implementation the energy of each sub-band of the background noise estimate can be estimated by moving averaging the energies of each sub-band among background noise frames detected as follows:

$$E_n(i) = \lambda \cdot E_n(i) + (1-\lambda) \cdot E(i)$$

wherein E(i) is the energy of the i^{th} sub-band of the frame detected as background noise,
λ is a forgetting factor usually in a range between 0.9 – 0.99. The power spectrum related in the above calculation can in a possible implementation be obtained by a fast Fourier transformation FFT.

In the normal working state NWS the VAD apparatus 1 according to the first aspect of the present invention the apparatus uses the modified segmental SNR $mssnr_{nor}$ to make an intermediate VAD decision $VADD_{int}$. This intermediate VAD decision $VADD_{int}$ can be made by comparing the calculated modified segmental SNR $mssnr_{nor}$ to a threshold thr which can be determined by:

$$thr = \begin{cases} 135 & lsnr > 18 \\ 35 & 8 < lsnr \leq 18 \\ 10 & lsnr \leq 8 \end{cases}$$

The intermediate VAD decision $VADD_{int}$ is active if the modified SNR $msnr_{nor}$>thr, otherwise the intermediate VAD decision $VADD_{int}$ is inactive.

In the speech offset state the VAD apparatus 1 uses in a possible implementation both the modified SNR $msnr_{off}$ and the voice metric V(-1) for making an intermediate VAD decision $VADD_{int}$. The intermediate VAD decision $VADD_{int}$ is made as active if the modified segmental SNR $mssnr_{off}$>thr or the voice metric V(-1) > a configurable threshold value of e.g. 0.7, otherwise the intermediate VAD decision $VADD_{int}$ is made as inactive.

In a possible implementation a hard hangover can be optionally applied to the intermediate VAD decision $VADD_{int}$. In this specific implementation if a hard hangover counter HHC is greater than a predetermined threshold such as 0 and if the intermediate VAD decision $VADD_{int}$ is inactive the final VAD decision $VADD_{fin}$ is forced to active and the hard hangover counter HHC is decremented by 1. In a possible implementation the hard hangover counter HHC is reset to its maximum value according to the same rule applied to the soft hangover counter SHC resetting.

In a still further possible implementation of the VAD apparatus 1 according to the first aspect of the present invention the VAD apparatus 1 selects in this specific implementation only two VAD parameters for its intermediate VAD decision, i.e. $mssnr_{nor}$ and $mssnr_{off}$.

$$mssnr_{nor} = \begin{cases} \sum_{i}^{N} (snr(i)+\alpha)^{4} & snr(i)+\alpha \geq 1, lsnr > 18 \\ \sum_{i}^{N} (snr(i)+\alpha)^{9} & snr(i)+\alpha \geq 1, 8 < lsnr \leq 18 \\ \sum_{i}^{N} (snr(i)+\alpha)^{13} & snr(i)+\alpha \geq 1, lsnr \leq 8 \end{cases}$$

$$mssnr_{off} = \begin{cases} \sum_{i}^{N} (snr(i)+\alpha+\beta)^{5} & lsnr > 18 \\ \sum_{i}^{N} (snr(i)+\alpha+\beta)^{11} & 8 < lsnr \leq 18 \\ \sum_{i}^{N} (snr(i)+\alpha+\beta)^{15} & lsnr \leq 8 \end{cases}$$

wherein the modified segmental SNR $mssnr_{nor}$ is used in the normal working state NWS and the modified segmental SNR $mssnr_{off}$ is used in the offset working state OWS. The coefficient β is determined in this implementation not only by the metric V(-1) but also by the sub-band index

i wherein for the sub-band index i greater than an integer value of m, if V(-1)>0.65 the coefficient β is set to 0.2 otherwise the coefficient β is set to 0.1. Further, for the sub-band index i being not greater than m if V(-1) > 0.65 the second coefficient β is set to β = 0.2 / + 1.5 otherwise the second coefficient β is set to 0.1 · 1,5. In this specific embodiment another set of thresholds the are defined for the offset working state OWS to be different from the set of thresholds the for the normal working state NWS.

The invention further provides as a second aspect an audio signal processing apparatus as shown in fig. 2 comprising a VAD apparatus 1 supplying a final VAD decision VADD to an audio signal processing unit 7 of the audio signal processing apparatus 6. Accordingly, the audio signal processing unit 7 is controlled by a VAD decision VADD generated by the VAD apparatus 1. The audio signal processing unit 7 can perform different kinds of audio signal processing on the applied audio signal such as speech encoding depending on the VAD decision.

According to a third aspect the present invention provides a method for performing a VAD wherein the VAD decision VADD is calculated by a VAD apparatus for an input audio signal using at least one VAD parameter VADP of a working state parameter decision set WSPDS of a current working state WS detected by a state detector of said VAD apparatus. According to a possible implementation of the method an input frame of the applied input audio signal is received. Then, a signal type of the input signal can be identified from a set of predefined signal types. In a further step a working state WS of the VAD apparatus is selected or chosen among several possible working states WS according to the identified input signal type. In a further step the VAD parameters are selected corresponding to the selected working state WS of the VAD apparatus among a larger set of predefined VAD decision parameters. Finally, a VAD decision VADD is made based on the chosen or selected VAD parameters.

A possible implementation of the method according to a third aspect of the present invention the set of predefined signal types can consist of a speech offset type and a non-speech offset type. Several possible working states WS can include a state for speech offset defined as a short period of the applied audio signal at the end of the speech bursts. The speech offset can be identified typically by a few frames immediately after the intermediate decision of the VAD apparatus working in the non-speech offset working state falls to inactive from active in a speech burst. A speech burst can be detected e. g. when a more than 60ms long active speech signal is detected. In a possible implementation of the method according to the third aspect of the present invention the set of predefined VAD parameters can include sub-band segmental SNR based parameters with different forms. In a possible implementation the sub-band seg-

mental SNR based parameters with different forms are sub-band segmental SNR parameters processed by different non-linear functions.

CLAIMS

WHAT IS CLAIMED IS:

1. A voice activity detection apparatus (1) for determining a voice activity detection decision (VADD) for an input audio signal, wherein the voice activity detection apparatus (1) comprises:
a state detector (2) adapted to determine a current working state (WS) of at least two different working states of the voice activity detection apparatus (1) dependent on the input audio signal wherein each of the at least two different working states (WS) is associated with a corresponding working state parameter decision set (WSPDS) including at least one voice activity decision parameter (VADP); and
a voice activity calculator (3) adapted to calculate a voice activity detection parameter value for the at least one VADP of the working state parameter decision set (WSPDS) associated with the current working state (WS) and to determine the voice activity detection decision (VADD) by comparing the calculated voice activity detection parameter value of the respective voice activity decision parameter (VADP) with a threshold.

2. The voice activity detection apparatus according to claim 1,
wherein said voice activity detection decision (VADD) is determined by said voice activity calculator (3) by using sub-band segmental signal to noise ratio (SNR) based voice activity decision parameters (VADPs).

3. The voice activity detection apparatus according to one of the preceding claims 1 to 2,
wherein said voice activity detection decision (VADD) for said input audio signal is determined on the basis of the at least one voice activity decision parameter (VADP) of the working state parameter decision set (WSPDS) provided for the current working state (WS) of said voice activity detection apparatus (1) using a predetermined voice activity detection processing algorithm provided for the current working state (WS) of said voice activity detection apparatus (1).

4. The voice activity detection apparatus according to one of the preceding claims 1 to 3,
wherein said voice activity detection apparatus (1) is switchable between different working states (WS) according to configurable working state transition conditions.

5. The voice activity detection apparatus according to one of the preceding claims 1 to 4,
wherein said voice activity detection apparatus (1) comprises a normal working state (NWS) and an offset working state (OWS).

6. The voice activity detection apparatus according to claim 5, wherein said voice activity detecting apparatus (1) detects a change from voice activity being present to voice activity being absent in said input audio signal if in the normal working state (NWS) of said input audio signal the voice activity detection decision (VADD) determined on the basis of the at least one voice activity detection parameter (VADP) of the normal working state parameter decision set (NWSPDS) of said normal working state (NWS) indicates a voice activity being present for a previous frame and a voice activity being absent in a current frame of said input audio signal.

7. The voice activity detection apparatus according to one of claims 5 to 6, wherein if said voice activity detection apparatus (1) detects in its normal working state (NWS) that a voice activity is present in the previous frame and that a voice activity is absent in a current frame of said input audio signal it is switched from its normal working state (NWS) to an offset working state (OWS) in which the voice activity detection decision (VADD) is determined on the basis of the at least one voice activity detection parameter (VADP) of the offset working state parameter decision set (OWSPDS).

8. The voice activity detection apparatus according to claim 5 to 7, wherein the voice activity detection decision (VADD) determined in the offset working state (OWS) forms an intermediate voice activity detection decision ($VADD_{int}$) if the voice activity detection decision (VADD) determined on the basis of the at least one voice activity detection parameter (VADD) of the offset working state parameter decision set (OWSPDS) indicates that a voice activity is absent in the current frame of the input audio signal.

9. The voice activity detection apparatus according to claim 8, wherein the intermediate voice activity detection decision (VADD) undergoes a hard hangover processing to provide a final voice activity detection decision ($VADD_{fin}$).

10. The voice activity detection apparatus according to claim 5, wherein said voice activity detection apparatus (1) is switched from the normal working state (NWS) to the offset working state (OWS) if the voice activity detection decision (VADD) determined by the voice activity calculator (3) of said voice activity detection apparatus (1) in the normal working state (NWS) using a voice activity detection processing algorithm and the working state parameter decision set (NWSPDS) provided for said normal working state (NWS) indicates an absence of voice in the input audio signal and a soft hangover counter (SHC) exceeds a predetermined threshold counter value.

11. The voice activity detection apparatus according to claim 5, wherein said voice activity detection apparatus (1) is switched from the offset working state (OWS) to the normal working state (NWS) if the soft hangover counter (SHC) does not exceed a predetermined threshold counter value.

12. The voice activity detection apparatus according to claim 10 or 11, wherein said input audio signal consists of a sequence of audio signal frames and said software hangover counter (SHC) is decremented in the offset working state (OWS) of said voice activity detection apparatus (1) for each received audio signal frame until the predetermined threshold counter value is reached.

13. The voice activity detection apparatus according to one of the preceding claims 10 to 12, wherein if a predetermined number of consecutive active audio signal frames of the input audio signal is detected said software hangover counter (SHC) is reset to a counter value depending on a long term signal to noise ratio (ISNR) of the input audio signal.

14. The voice activity detection apparatus according to one of the preceding claims 10 to 13, wherein an active audio signal frame is detected if a calculated voice metric (V) of the audio signal frame exceeds a predetermined voice metric threshold value and a pitch stability (S) of said audio signal frame is below a predetermined stability threshold value.

15. The voice activity detection apparatus according to on of the preceding claims 1 to 14, wherein said voice activity decision parameters (VADPs) of a working state parameter decision set (WSPDS) of a working state (WS) of said voice activity detection apparatus comprises
energy based decision parameters,
spectral envelope based decision parameters,
and/or statistic based decision parameters.

16. The voice activity detection apparatus according to one of the preceding claims 1 to 15, wherein an intermediate voice activity detection decision ($VADD_{int}$) determined by said voice activity calculator (3) is applied to a hard hangover processing unit performing a hard hangover of said applied intermediate voice activity detection decision ($VADD_{int}$).

17. An audio signal processing device (6) comprising a voice activity detection apparatus (1) according to one of the preceding claims 1 to 16 and an audio signal processing unit (7)

controlled by a voice activity detecting decision (VADD) generated by said voice activity detection apparatus (1).

18. A method for performing a voice activity detection
wherein a voice activity detection decision (VADD) is calculated by a voice activity detection apparatus (1) for an input audio signal using at least one voice activity detection parameter (VADP) of a working state parameter decision set (WSPDS) of a current working state (WS) detected by a state detector (2) of said voice activity detection apparatus.

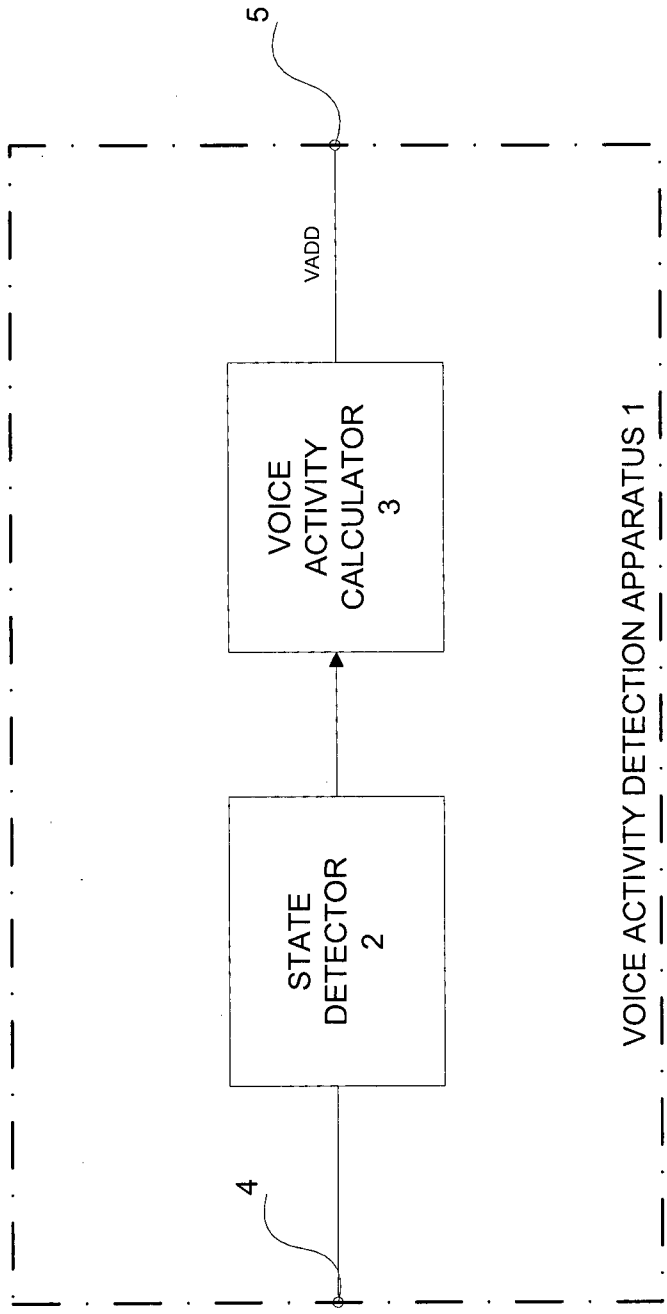
5
VADD
VOICE ACTIVITY CALCULATOR 3
STATE DETECTOR 2
4
VOICE ACTIVITY DETECTION APPARATUS 1

Fig. 1

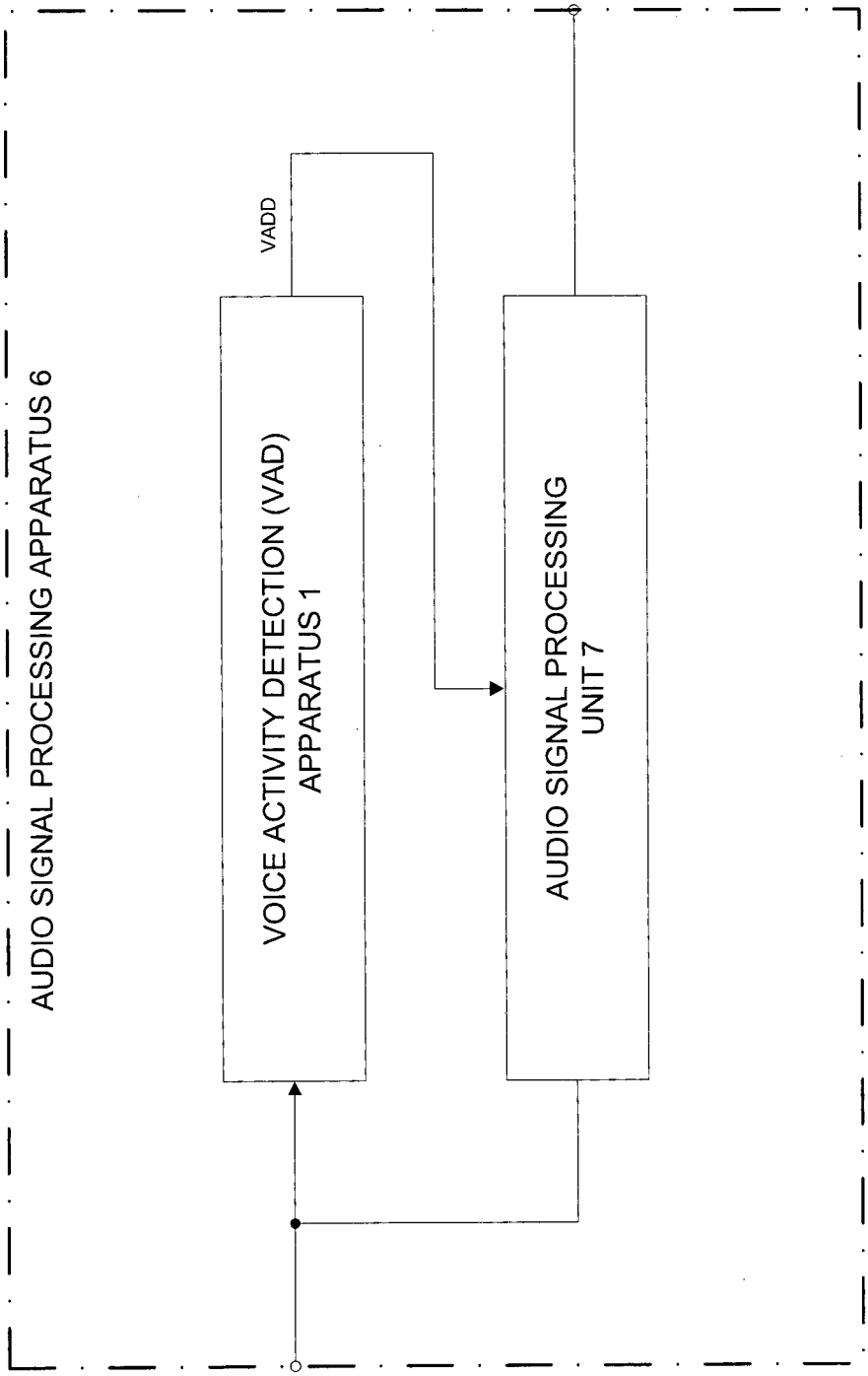
AUDIO SIGNAL PROCESSING APPARATUS 6
VADD
VOICE ACTIVITY DETECTION (VAD) APPARATUS 1
AUDIO SIGNAL PROCESSING UNIT 7

Fig. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2010/080222

A. CLASSIFICATION OF SUBJECT MATTER

G10L 11/02 (2006.01) i

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)

IPC: G10L, H04L, H04W, H04Q, H04M

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)

CNPAT,CNKI,WPI,EPODOC,IEEE: VAD, VADP, WSPDS, voice, audio, activity, detect???, decision, parameter?, working, state, set, signal, threshold, noise

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO00/17856A1 (CONEXANT SYSTEMS, INC.) 30 Mar. 2000 (30.03.2000) the whole document	A
A	CN1166723A (SAMSUNG ELECTRONICS CO., LTD.) 03 Dec. 1997 (03.12.1997) the whole document	A
A	CN101236742A (ZTE CORP.) 06 Aug. 2008 (06.08.2008) the whole document	A
A	CN101790752A (QUALCOMM INC.) 28 Jul. 2010 (28.07.2010) the whole document	A
A	EP0790599A1 (NOKIA MOBILE PHONES LTD.) 20 Aug. 1997 (20.08.1997) the whole document	A

☐ 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

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 May 2011 (30.05.2011)	**30 Jun. 2011 (30.06.2011)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer WANG, Lunjie Telephone No. (86-10)62413491

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2010/080222

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
WO00/17856A1	30.03.2000	TW442774A	23.06.2001
		US6188981B1	13.02.2001
CN1166723A	03.12.1997	KR970073003A	07.11.1997
CN101236742A	06.08.2008	NONE	
CN101790752A	28.07.2010	WO2009042948A1	02.04.2009
		CA2695231A1	02.04.2009
		US2009089053A1	02.04.2009
		JP2010541010T	24.12.2010
		EP2201563A1	30.06.2010
		INMUMNP201000244E	09.07.2010
		TW200926151A	16.06.2009
EP0790599A1	20.08.1997	EP0784311B1	05.09.2001
		JP9212195A	15.08.1997
		JP9204196A	05.08.1997
		JP2007179073A	12.07.2007
		JP2008293038A	04.12.2008
		FI100840B1	27.02.1998
		DE69630580E	11.12.2003
		DE69614989E	11.10.2001
		WO9722116A2	19.06.1997
		US5963901A	05.10.1999
		US5839101A	17.11.1998
		AU1067797A	03.07.1997