
Question(s): 8/16

Geneva, 26 October - 6 November 2009

TEMPORARY DOCUMENT**Source:** Editor G.GSAD**Title:** Draft new ITU-T Recommendation G.720.1 (ex G.GSAD) "Generic sound activity detector" (for Consent)

This document contains the AAP summary and text for draft new ITU-T Recommendation G.GSAD "Generic Sound Activity Detector", which is proposed by Q8/16 for Consent at this WP 3/16 meeting.

AAP Summary

Voice Activity Detection (VAD) is a well-known technique which has been used for many years in telecommunication systems. It is used to detect the presence or absence of speech in a communications channel, and plays an important role as a pre-processing stage for enabling several functions. For example, VAD can reduce bandwidth usage by enabling silence compression algorithms in order to use the communications channel only if speech is detected.

However, current VAD algorithms are designed to handle voice signals and do not perform well in the presence of other types of audio signals which are increasingly appearing on telecommunications networks, particularly music. At the same time, there are many VAD algorithms available, which are typically associated with a specific codec only. Unfortunately, they are based on different design approaches, and some perform better than others. This leads to difficulty in improving the overall performance of VAD algorithms, and adds additional time and cost when developing a new codec since the existing VAD algorithms can not easily be re-used.

ITU-T Recommendation G.GSAD "Generic Sound Activity Detector" provides a solution to these problems. The GSAD is an independent front-end processing module which can be applied prior to signal processing applications that operate on narrowband input or wideband input at 10 ms frame length (without lookahead), such as speech or audio codecs. Its primary function is to indicate the input frame activity. For active frame it further indicates if the input frame is speech or music, and for inactive frame it indicates whether the frame is a silence frame or an audible noise frame. The GSAD can also operate with only the primary function of indicating the input frame activity.

The GSAD can further operate on three different operating points. For the activity detection functionality the operating points provide selectable balancing between bandwidth saving and audio quality, which can be utilized for high-performance silence compression schemes that can balance between end-users speech and audio subjective quality needs and the system and network traffic requirements. For example, for the speech database used in the GSAD selection phase with added

Contact: Wang Zhe
Huawei Technologies
P.R. ChinaTel: +86 10 82836050
Fax: +86 10 82836920
Email: wang.zhe@huawei.com

Attention: This is not a publication made available to the public, but an internal ITU-T Document intended only for use by the Member States of ITU, by ITU-T Sector Members and Associates, and their respective staff and collaborators in their ITU related work. It shall not be made available to, and used by, any other persons or entities without the prior written consent of ITU-T.
--

car, office and babble background noises, the Bandwidth Saving operating point saves approximately 30% of the bandwidth compared to the Balanced operating point and approximately 50% of the bandwidth compared to the Quality Preferred operating point.

The three different operating points also control the GSAD emphasis and balance between on speech and music classification for the active frames, which can be utilized for the fine-tuning of source-controlled audio compression algorithms.

Draft G.720.1 being put forward for Consent contains the following electronic attachments:

- ANSI-C source code
http://itu.int/md/dologin_md.asp?id=T09-SG16-091026-TD-PLEN-0186!A1!ZIP-E&type=mitems
- Test vectors: [http://ifa.itu.int/t/2009/sg16/exchange/wp3/q08/g720.1/tv-091106/G.720.1\(ex-G.GSAD\)-test-vectors-fixed-point.zip](http://ifa.itu.int/t/2009/sg16/exchange/wp3/q08/g720.1/tv-091106/G.720.1(ex-G.GSAD)-test-vectors-fixed-point.zip) (71.6MB) – Available from the FTP site due to its large size

(Full text available only in the electronic version)

CONTENTS

1	Scope.....	4
2	References.....	5
3	Definitions.....	5
	3.1 Terms defined elsewhere:.....	5
	3.2 Terms defined in this Recommendation.....	5
4	Abbreviations and acronyms.....	5
5	Conventions	8
6	General description of GSAD algorithm	8
	6.1 Input sampling rate	8
	6.2 Operating frame size.....	8
	6.3 Delay.....	9
	6.4 Configurations	9
	6.5 Complexity and memory cost.....	10
7	Detailed description of the GSAD algorithm.....	10
	7.1 Detailed Description of the VAD Module.....	10
	7.1.1 Pre-processing and power spectrum calculation.....	11
	7.1.2 Differential Zero Crossing Rate (DZCR) calculation.....	12
	7.1.3 Modified Segmental SNR (MSSNR) calculation	12
	7.1.4 Long term SNR estimation	13
	7.1.5 Background fluctuation estimation	14
	7.1.6 Initial VAD decision	14
	7.1.7 VAD hangover	15
	7.1.8 Calculation of tone stability	16
	7.1.9 Calculation of spectral peak fluctuation.....	16
	7.1.10 Calculation of spectral peakiness	17
	7.1.11 Calculation of frequency stability	17
	7.1.12 Background music detection.....	17
	7.1.13 Silence detection	18
	7.1.14 Background estimate update	19
	7.2 Detailed description of the Speech/Music Discrimination module.....	22
	7.2.1 Calculation of the flux and the variance of the flux.....	22
	7.2.2 Calculation of two spectral-peaks peakiness measures.....	23
	7.2.3 Speech/Music Discrimination decision.....	23
8	Organization of the reference C code	25

Recommendation ITU-T G.GSAD

Generic sound activity detector (GSAD)

Summary

This Recommendation describes an independent front-end processing module (Generic Sound Activity Detector – GSAD) which can be applied prior to signal processing applications that operate on narrowband or wideband audio input at 10 ms frame length (without lookahead), such as speech or audio codecs. Its primary function is to indicate the input frame activity (VAD). For an active frame it further indicates if the input frame is speech or music (speech/music discrimination), and for an inactive frame it indicates whether the frame is a silence frame or an audible noise frame (silence detection). The GSAD can also operate when only the primary function, of indicating the input frame activity, is used.

An external control signal indicates to the GSAD algorithm which one of the three different operating points to use. For the activity detection functionality, these operating points provide selectable balancing between bandwidth saving and audio quality, which can be utilized for high-performance silence compression schemes that can balance between the end-users speech and audio subjective quality needs and the system and network traffic requirements.

The three different operating points also control the GSAD emphasis and balance between on speech and music classification for the active frames, which can be utilized for the fine-tuning of source-controlled audio compression systems.

The VAD module uses a dual-parameters classification scheme, where one parameter is a differential zero crossing rate measure and the other parameter is a modified segmental SNR measure. An initial VAD decision is made with a pair of inequalities, with factors that are adaptive to the long term SNR of the input signal. A final VAD decision is obtained by an adaptive hangover scheme. The Speech/Music Discrimination module calculates the variance of a spectral deviation measure and applies an adaptive threshold to make an initial decision between speech and music. Two spectral peakiness measures further modify that initial decision and a one-frame hangover is used to obtain the final speech/music discrimination decision. The Silence Detection module uses an energy threshold to discriminate between a silence frame and an audible noise frame.

This Recommendation provides a detailed description of the overall GSAD configuration, including the operating points; the VAD module; the speech/music discrimination module and the silence detection module. This Recommendation further contains an electronic attachment with the ANSI C source code which forms an integral part of this Recommendation and a set of test vectors.

1 Scope

The GSAD is an independent front-end processing module which can be applied prior to signal processing applications that operate on narrowband or wideband audio input at 10 ms frame length (without lookahead), such as speech or audio codecs. Its primary function is to indicate the input frame activity. For an active frame it further indicates if the input frame is speech or music, and for an inactive frame it indicates whether the frame is a silence frame or an audible noise frame.

This Recommendation is organized as follows. References, definitions, abbreviations/acronyms and conventions are defined in clauses 2, 3, 4 and 5 respectively. Clause 6 gives a general description of the GSAD algorithm including the input sampling rate, the operating frame length, the algorithmic delay, the configurations and the complexity and memory cost. The detailed description of the

GSAD algorithm is described in clause 7 where clause 7.1 describes the VAD module and the speech/music discrimination module is described in clause 7.2. Finally in clause 8 the organization of the ANSI C code is described.

2 References

This Recommendation does not make reference to any other Standards.

3 Definitions

3.1 Terms defined elsewhere:

This Recommendation does not use definitions defined elsewhere.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 Speech/Music discriminator: A function or device which classifies audio input into either speech or music.

3.2.2 Silence detector: A function or device which identifies frames whose levels are below a silence threshold.

4 Abbreviations and acronyms

This Recommendation uses the abbreviations and acronyms defined in Table 1 and Symbols used throughout this Recommendation are listed in Table 2.

Table 1 – Glossary of acronyms

Acronym	Description
GSAD	Generic Sound Activity Detection/Detector
VAD	Voice Activity Detector
SMD	Speech/Music Discriminator
SiD	Silence Detector
NB	Narrowband
WB	Wideband
FFT	Fast Fourier Transform
SDF	Spectral Density Function
ZCR	Zero Crossing Rate
DZCR	Differential Zero Crossing Rate
MSSNR	Modified Segmental SNR
RMS	Root Mean Square
WMOPS	Weighted Million Operations Per Second

Table 2 – Glossary of symbols

Name	Description
$s(i)$	The i^{th} sample in the input frame
$S_{emp}(i)$	The i^{th} pre-emphasized sample
$S^{[m]}(k)$	The k^{th} spectral bin of the m^{th} frame
\overline{ZCR}_n	The moving average of the ZCR of the estimated background signal
ZCR	The zero crossing rate of the input frame
DZCR	The differential zero crossing rate of the input frame
$E_{band}(i)$	The energy of the i^{th} sub-band
$E_{band_old}(i)$	The energy of the i^{th} sub-band of the previous frame
$\overline{E}_{band_n}(i)$	The moving average of the energy of the i^{th} sub-band of estimated background signal
$snr(i)$	The SNR of the i^{th} sub-band of the input frame
$msnr(i)$	The modified SNR of the i^{th} sub-band
MSSNR	The modified segmental SNR
rms	The RMS of the input frame
$rms_{bgd}^{[m]}$	The long term RMS of the background signal of the m^{th} frame
$rms_{fgd}^{[m]}$	The long term RMS of the foreground signal of the m^{th} frame
$lsnr$	The long term SNR of the input frame
$flux_{bgd}$	The fluctuation of the background signal
thr_{vad}	The threshold for the initial VAD decision
$ivad$	The initial VAD decision
$hang_sp$	The hangover counter for speech
$hang_f_mus$	The hangover counter for foreground music
$hang_b_mus$	The hangover counter for background music
$peak_{loc}(i)$	The power of the i^{th} local spectral peak
$idx_peak_{loc}(i)$	The position of the i^{th} local spectral peak in the spectrum
$D_{p2s}(i)$	The power distance between the i^{th} local spectral peak and its adjacent four spectral bins on its two sides
$idx_peak_{max}(j)$	The location of the local spectral peak which has the j^{th} largest D_{p2s} in the spectrum
$idx_peak_{max_old}$	The location of the local spectral peak which has the maximum D_{p2s} in the previous frame
$dtmf_flg$	The flag indicating the presence of DTMF signal
$peak_{gLb}^{[m]}$	The largest local spectral peak of the m^{th} input frame
$idx_peak_{gLb}^{[m]}$	The location of $peak_{gLb}^{[m]}$ in its spectrum

Name	Description
$peak_flux_cnt$	The counter counting the spectral peak fluctuation
$E_{vl}(j)$	The lower frequency spectral valley of the j^{th} local spectral peak
$E_{vh}(j)$	The higher frequency spectral valley of the j^{th} local spectral peak
$D_{p2v}(j)$	The normalized peak to valley distance of the j^{th} local spectral peak
$\overline{E_{band}(i)}$	The moving average of the power of the i^{th} sub-band over past frames
sta_{fq}	The frequency stability of the input frame
E_{band_mean}	The power mean of the 16 whitened sub-bands in the calculation of frequency stability
bgd_frm_cnt	The counter counting the number of background frames in the background music detection
SP	The spectral peakiness of the input frame
SP_{sum}	The spectral peakiness accumulator storing the sum of the frame peakinesses in the background music detection
FP_{dBov}	The signal power in dBov
thr_{bgd}	The threshold for background frame identification
$update_flg$	The flag indicating whether to update the background estimate
$reset_flg$	The flag indicating whether to reset counters used in the background estimate update procedure
con_frm_cnt	The counter counting the number of consecutive frames in the background estimate update procedure
thr_{SP_low}	The threshold for low spectral peakiness identification
low_SP_cnt	The counter counting the number of frames with low spectral peakiness
thr_{fst}	The threshold for high frequency-stability identification
$high_fst_cnt$	The counter counting the number of frames with high frequency-stability
$E_{band_buf_min}(i)$	The minimum energy of the i^{th} sub-band of the past frames
$tone_sta_cnt$	The counter counting the number of frames with high tone stability
$flux^{[m]}$	A spectral deviation measure for the m^{th} frame
$mov_flux^{[m]}$	The moving average of $flux^{[m]}$ at the m^{th} frame
$var_flux^{[m]}$	The variance of $flux^{[m]}$ at the m^{th} frame
P_1	A first peakiness measure
P_2	A second peakiness measure
$avrg_P_1^{[m]}$	The moving average of P_1 at the m^{th} frame
$avrg_P_2^{[m]}$	The moving average of P_2 at the m^{th} frame
$T_{var_flux}^{[m]}$	Adaptive threshold on $var_flux^{[m]}$
$max_{MSSNR}^{[m]}$	A maximal value of $MSSNR$ over past frames

Name	Description
$T_{MSSNR}^{[m]}$	An adaptive threshold of the <i>MSSNR</i>
$high_{bin}^{[m]}$	A counter on the <i>MSSNR</i> high values
$low_{bin}^{[m]}$	A counter on the <i>MSSNR</i> high values
$diff_{hist}^{[m]}$	An initial difference measure between $high_{bin}^{[m]}$ and $low_{bin}^{[m]}$
$diff_{hist}^{avg}$	An average measure of $diff_{hist}^{[m]}$
Δ_{op}	An offset factor for $diff_{hist}^{[m]}$
X_T	A threshold on $diff_{hist}^{avg}$
$diff_{hist}^{final}$	A final difference measure between $high_{bin}^{[m]}$ and $low_{bin}^{[m]}$
T_{op}^{up}	Highest desired value for $T_{var_flux}^{[m]}$
T_{op}^{down}	Lowest desired value for $T_{var_flux}^{[m]}$

5 Conventions

The following conventions apply to this Recommendation:

- $|x|$ denotes the absolute value of x : $|12| = 12$, $|-3| = 3$;
- $sgn[x]$ denotes the sign of x : $sgn[3] = 1$, $sgn[-5] = -1$;
- $MAX[x_0, x_1, x_2, \dots, x_{N-1}]$ denotes the maximum of $x_0, x_1, x_2, \dots, x_{N-1}$: $MAX[0, -1, 3, 5, 2] = 5$;
- $MIN[x_0, x_1, x_2, \dots, x_{N-1}]$ denotes the minimum of $x_0, x_1, x_2, \dots, x_{N-1}$: $MIN[0, -1, 3, 5, 2] = -1$;
- $\log(x)$ denotes the logarithm in base 10 of x : $\log(100) = 2$;
- $p^{[m]}$ denotes the value of parameter p at the m^{th} input frame. Current frame is assumed when $[m]$ is omitted;
- \sum denotes summation;
- o.w. denotes “otherwise”.

6 General description of GSAD algorithm

6.1 Input sampling rate

The GSAD can accept both narrowband (NB) and wideband (WB) audio input signals sampled at 8 and 16 kHz, respectively. The sampling rate of the input signal is provided to the GSAD by an external signal.

6.2 Operating frame size

GSAD operates on a 10 ms basis, i.e. frame of 80 samples for NB input and 160 samples for WB audio input. Consecutive GSAD indications can be combined to indicate the activity for frames with multiple of 10 ms.

6.3 Delay

The GSAD does not introduce lookahead, therefore the added delay of the GSAD is 0 ms (algorithmic delay of 10 ms).

6.4 Configurations

The GSAD operates on narrowband (NB) and wideband (WB) audio input signals, where a control signal indicates the bandwidth of the input signal. For both bandwidths, a second control signal sets the GSAD to operate in a generic sound activity detecting mode or in a voice activity detecting mode. When the GSAD operates in a generic sound activity detecting mode, the input signal first passes through the voice activity detector. Frames that are detected as active signal frames are further discriminated as either speech frames or music frames. Frames that are detected as inactive signals are further classified as either audible noise frames or silence frames.

Table 3 describes the GSAD output bits.

Table 3 – Description of the GSAD output bits

VAD decision	GSAD decision	Main Output	
		First bit	Second bit
Active	Speech	1	1
	Music	1	0
Inactive	Noise	0	1
	Silence	0	0

When the GSAD operates in a generic sound activity detecting mode both bits are outputted, which can be interpreted as four possible values of “3”, “2”, “1” or “0”, indicating speech, music, audible noise or silence frame, respectively. When the GSAD operates in a voice activity detection mode, only the first bit is outputted, which can be interpreted as “1” or “0” indicating active or inactive frame, respectively.

In addition, a third control signal selects the desired operating point from three possible operating points. Setting the third control signal to “0” selects a Balanced operating point, setting it to “1” selects a Quality Preferred operating point and setting the third control signal to “2” selects a Bandwidth Saving operating point. When the GSAD operates in a voice activity detection mode, the three operating points provide the ability to select a preferred balancing between the bandwidth saving and the audio quality. The Bandwidth Saving operating point provides maximal bandwidth saving while maintaining acceptable audio quality, while the Quality Preferred operating point provides higher audio quality with less bandwidth saving. The Balanced operating point provides an optimum performance which is between the Bandwidth Saving operating point and the Quality Preferred operating point. For example, for the speech database used in the GSAD selection phase with added car, office and babble background noises, the Bandwidth Saving operating point saves approximately 30% of the bandwidth compared to the Balanced operating point and approximately 50% of the bandwidth compared to the Quality Preferred operating point. When the GSAD operates in a generic sound activity detecting mode, the operating points also control the operation of the speech/music discrimination module. The Bandwidth Saving operating point is tuned to classify most active speech frames correctly and the Quality Preferred operating point is tuned to classify most active music frames correctly, while the Balanced operating point provides a middle point performance between the other two operating points.

6.5 Complexity and memory cost

The complexity of the GSAD for its different modes and signal sampling frequencies are provided in Table 4.

Table 4 – Complexity of the GSAD

Modes	Complexity (WMOPS)
GSAD_WB	2.935
GSAD_NB	1.897
VAD_WB	2.397
VAD_NB	1.475

The RAM used for the GSAD is 3284 bytes and the table ROM is 1674 bytes.

7 Detailed description of the GSAD algorithm

Figure 1 is the block diagram of the GSAD system. The output of the VAD is a binary flag indicating the activity of the input frame. Active frames will be further classified into speech or music by the SMD and inactive frames will be further classified into silence or audible noise frame by the SiD.

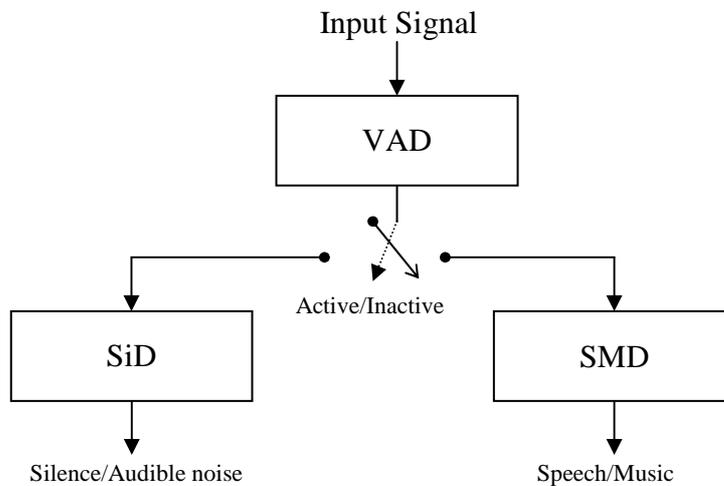


Figure 1 – Block diagram of the GSAD system

7.1 Detailed Description of the VAD Module

Figure 2 is the general diagram depicting the high level operation of the VAD module, where each box contains the number of the relevant clauses in the text.

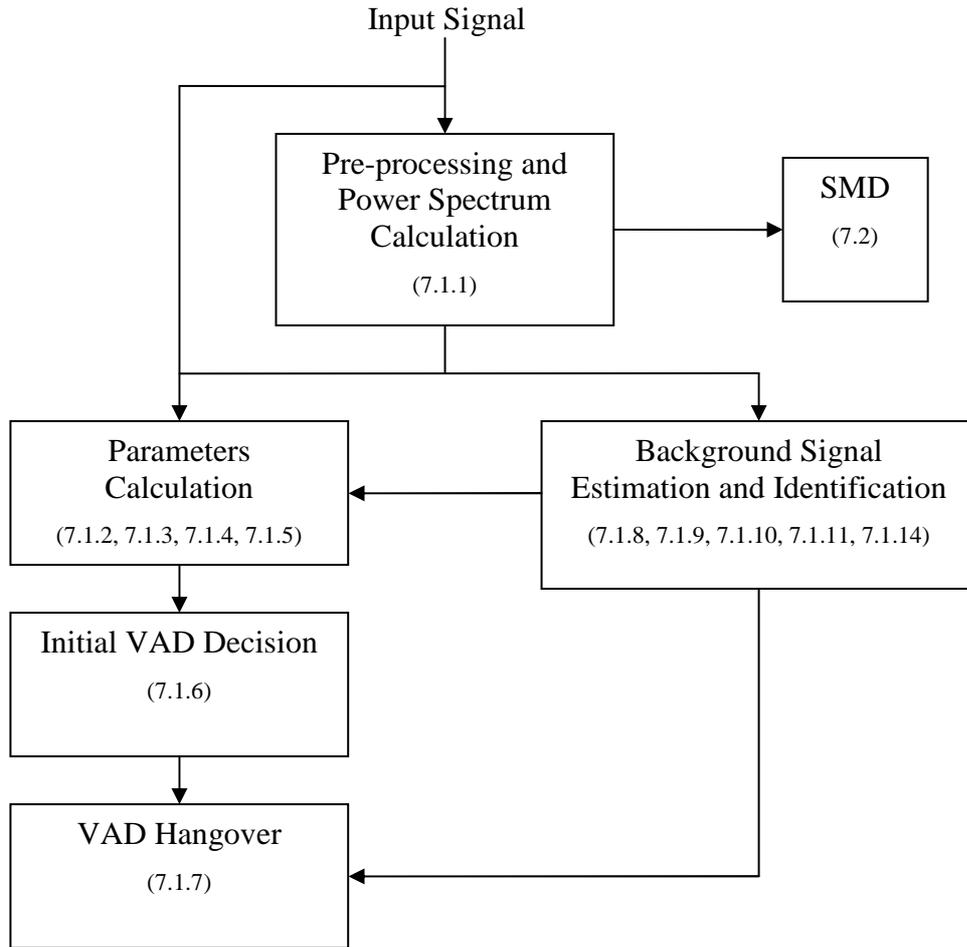


Figure 2 – General diagram of the VAD module

7.1.1 Pre-processing and power spectrum calculation

The input signal to the GSAD is a sampled audio signal at 10 ms frames, which means that each frame consists of 80 samples for NB input and of 160 samples for WB input. The input samples are first pre-emphasized then windowed by an asymmetric window. The pre-emphasis is done by

$$s_{emp}(i) = s(i-1) + 0.8 \cdot s(i) \quad (7.1.1)$$

where $s_{emp}(i)$ is the i^{th} pre-emphasized sample, $s(i)$ is the i^{th} original sample. The asymmetric window (128 points for NB and 256 points for WB) covers all samples of the current frame and also samples (48 samples for NB input and 96 samples for WB input) of the past frame. A fast Fourier transform (FFT) of 128 points for NB input and of 256 points for WB input is performed on the windowed samples. For the m^{th} frame, the power spectrum elements $S^{[m]}(k)$ are obtained by square root of the sum of square of the real and the imaginary parts of the complex FFT coefficients, where k is the index for the power spectrum elements. In the sequel, $S^{[m]}(k)$ is denoted by $S(k)$ when describing the processing of the current frame, for brevity.

7.1.2 Differential Zero Crossing Rate (DZCR) calculation

The zero crossing rate (*ZCR*) of the input frame is extracted from the original non-preprocessed time domain samples of the input frame.

$$ZCR = \frac{1}{2} \sum_{i=0}^{N-2} |\text{sgn}[s(i)] - \text{sgn}[s(i+1)]| \quad (7.1.2)$$

where *N* is the number of samples per frame (*N* = 80 for NB input and *N* = 160 for WB input). The *DZCR* of the input frame is calculated by

$$DZCR = ZCR - \overline{ZCR}_n \quad (7.1.3)$$

where \overline{ZCR}_n is the moving average of the *ZCR* of the estimated background signal. For the calculation of \overline{ZCR}_n , refers to Clause 7.1.14.

7.1.3 Modified Segmental SNR (MSSNR) calculation

The spectrum of the input frame is divided into 16 non-equal sub-bands in the frequency domain. The energy of each sub-band is calculated by

$$E_{band}(i) = \frac{\alpha}{h(i) - l(i) + 1} \sum_{k=l(i)}^{h(i)} S(k) + (1 - \alpha) E_{band_old}(i) \quad (7.1.4)$$

where *i* is the index of sub-band, *l*(*i*) is the lower bound of the *i*th sub-band, *h*(*i*) is the higher bound of the *i*th sub-band, *S*(*k*) is the power spectrum of the *k*th spectral bin, $E_{band_old}(i)$ is the energy of the *i*th sub-band of the previous frame and α is a weighting factor. The exact lower and higher bounds of each sub-band are given in Table 5 for NB input and table 6 for WB input, respectively.

Table 5 – Boundaries of the sub-bands for NB input

<i>i</i>	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>l</i> (<i>i</i>)	2	4	6	8	10	12	14	17	20	23	27	31	36	42	49	56
<i>h</i> (<i>i</i>)	3	5	7	9	11	13	16	19	22	26	30	35	41	48	55	63

Table 6 – Boundaries of the sub-bands for WB input

<i>i</i>	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>l</i> (<i>i</i>)	2	4	6	8	10	12	14	17	20	23	27	37	49	63	79	97
<i>h</i> (<i>i</i>)	3	5	7	9	11	13	16	19	22	26	36	48	62	78	96	127

The weighting factor α is determined by the long term SNR, *lsnr*, as:

$$\alpha = \begin{cases} 0.75 & \text{lsnr} > 35 \\ 0.55 & \text{lsnr} \leq 35 \end{cases} \quad (7.1.5)$$

(For the calculation of *lsnr* refer to Clause 7.1.4.) For the first input frame, α is set to 1.

The sub-band SNR is calculated as

$$\text{snr}(i) = 10 \log(E_{band}(i) / \overline{E_{band_n}}(i)), \quad (7.1.6)$$

where $\overline{E_{band_n}(i)}$ is the moving average of the energy of the i^{th} sub-band of the estimated background signal (for the calculation of $\overline{E_{band_n}(i)}$ refers to Clause 7.1.14). The modified sub-band SNR is calculated by

$$msnr(i) = \begin{cases} \text{MAX} \left[\text{MIN} \left[\frac{snr^3(i)}{64}, 1 \right], 0 \right] & 1 < i \leq 12 \\ \text{MAX} \left[\text{MIN} \left[\frac{snr^3(i)}{25}, 1 \right], 0 \right] & 0 \leq i \leq 1 \text{ or } i > 12 \end{cases} \quad (7.1.7)$$

The *MSSNR* is obtained by

$$MSSNR = \sum_{i=0}^{15} msnr(i). \quad (7.1.8)$$

7.1.4 Long term SNR estimation

The root mean square (RMS) of the input frame is calculated based on the original non-processed samples of the input frame:

$$rms = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s^2(i)}, \quad (7.1.9)$$

where *rms* is the RMS of the input frame, $s(i)$ is the i^{th} sample of the non-preprocessed input frame, N is the number of samples per frame ($N = 80$ for NB input and $N = 160$ for WB input). A long term RMS of the background signal and a long term RMS of the foreground signal are estimated respectively, where foreground signal denotes the dominating information signal (speech or music) and background signal denotes the interfering signals (e.g., background noise) at the background of the dominating speech or music signals. For the m^{th} frame, when the background update flag *update_flg* is set (for the setting of *update_flg*, refer to Clause 7.1.14), the long term RMS of the background signal is updated by

$$rms_{bgd}^{[m]} = \beta_b \cdot rms_{bgd}^{[m-1]} + (1 - \beta_b) \cdot rms^{[m]} \quad (7.1.10)$$

where $rms_{bgd}^{[m]}$ and $rms_{bgd}^{[m-1]}$ are the long term RMSs of the background signal of the m^{th} and the $(m-1)^{\text{th}}$ frames respectively, β_b is an adaptive factor controlling the update speed of the $rms_{bgd}^{[m]}$.

$$\beta_b = \begin{cases} 0.95 & rms^{[m]} > 1.4125 \cdot rms_{bgd}^{[m-1]} \text{ or } rms_{bgd}^{[m-1]} > 1.4125 \cdot rms^{[m]} \\ 0.99 & o.w. \end{cases} \quad (7.1.11)$$

Similarly, for the m^{th} frame, when its *MSSNR* is greater than a threshold, the long term RMS of the foreground signal is updated by

$$rms_{fgd}^{[m]} = \beta_f \cdot rms_{fgd}^{[m-1]} + (1 - \beta_f) \cdot rms^{[m]} \quad (7.1.12)$$

where $rms_{fgd}^{[m]}$ and $rms_{fgd}^{[m-1]}$ are the long term RMSs of the foreground signal of the m^{th} and the $(m-1)^{\text{th}}$ frames respectively, β_f is an adaptive factor controlling the update speed of the $rms_{fgd}^{[m]}$.

$$\beta_f = \begin{cases} 0.99 & rms^{[m]} > 2 \cdot rms_{fgd}^{[m-1]} \\ 0.999 & 1.6 \cdot rms_{fgd}^{[m-1]} < rms^{[m]} < 2 \cdot rms_{fgd}^{[m-1]} \\ 0.99999 & rms^{[m]} < 1.6 \cdot rms_{fgd}^{[m-1]} \end{cases} \quad (7.1.13)$$

The long term SNR $lsnr$ is obtained by

$$lsnr = 0.85 \cdot [20 \cdot \log(rms_{fgd} / 32767) - 20 \cdot \log(rms_{bgd} / 32767)] \quad (7.1.14)$$

7.1.5 Background fluctuation estimation

The fluctuation of the background signal $flux_{bgd}$ is estimated over background frames. When the $MSSNR$ of the input frame is below a threshold of 15, the $flux_{bgd}$ is updated by

$$flux_{bgd} = \chi \cdot flux_{bgd} + (1 - \chi) \cdot MSSNR \quad (7.1.15)$$

where χ is a factor controlling the update speed of the $flux_{bgd}$. χ is determined by

$$\chi = \begin{cases} 0.955 & MSSNR > flux_{bgd} ; \text{during initialization period} \\ 0.995 & MSSNR \leq flux_{bgd} ; \text{during initialization period} \\ 0.997 & MSSNR > flux_{bgd} ; \text{after initialization period} \\ 0.9997 & MSSNR \leq flux_{bgd} ; \text{after initialization period} \end{cases} \quad (7.1.16)$$

where the initialization period consists of the first 100 frames whose $MSSNR$ is below the threshold of 15.

7.1.6 Initial VAD decision

The initial VAD decision is made based on a set of inequalities involving the calculated $DZCR$ and $MSSNR$, where the inequalities' factors are adapted according to the long term SNR. The initial VAD decision is made by comparing a pair of linear combinations of $DZCR$ and $MSSNR$ to a threshold that adapts to the long term SNR, the background fluctuation and the operating point. The decision rule is:

$$\begin{aligned} \text{if } MSSNR > thr_{vad} & \quad ivad = 1 \\ \text{if } MSSNR - \lambda \cdot DZCR > thr_{vad} & \quad ivad = 1, \\ \text{else} & \quad ivad = 0 \end{aligned} \quad (7.1.17)$$

where λ is a factor determined by the long term SNR and thr_{vad} is a VAD decision threshold which is determined jointly by the long term SNR, background fluctuation and the operating point. The $ivad$ is the initial VAD decision, where the value of "1" means active and the value of "0" means inactive.

The determination of λ and thr_{vad} are described below. First, the long term SNR is classified into four categories; very high SNR, high SNR, medium SNR and low SNR, each corresponds to a value of λ as:

$$\lambda = \begin{cases} 0 & lsnr > 35 \\ 2.7778 & 35 \geq lsnr > 25 \\ 2.2222 & 25 \geq lsnr > 15 \\ 1.667 & lsnr \leq 15 \end{cases} \quad (7.1.18)$$

Further, the background fluctuation is classified into three categories: high fluctuation, medium fluctuation and low fluctuation. Each combination of long term SNR category, background fluctuation category and operating point corresponds to a value for thr_{vad} . The thr_{vad} is chosen from a threshold table indexed by the long term SNR category, the background fluctuation category and the operating point. The threshold tables are different for NB and WB input.

7.1.7 VAD hangover

The final VAD decision is made by passing the initial VAD decision through a hangover procedure. The hangover procedure consists of three independent hangover mechanisms operating in parallel. This is depicted in Figure 3 below:

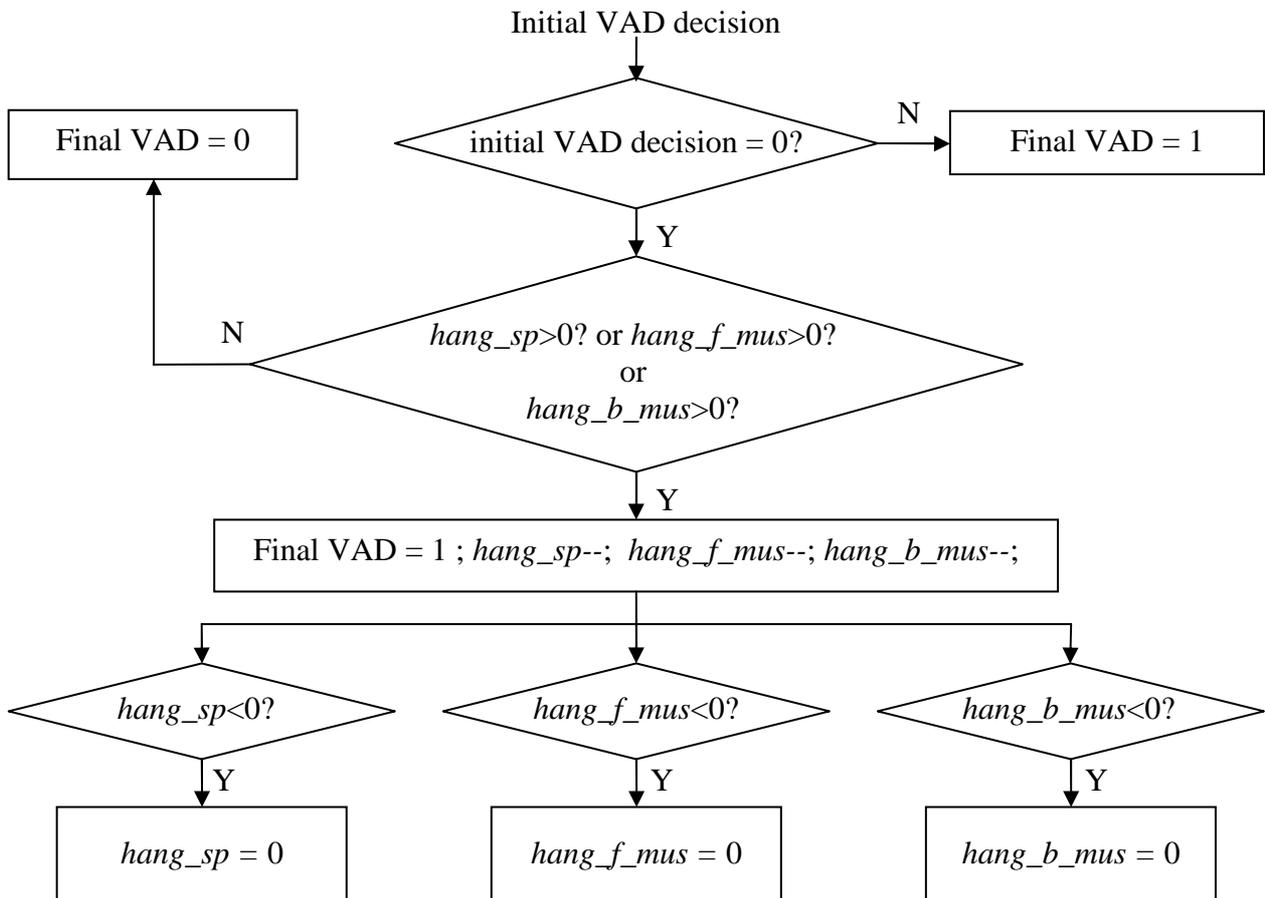


Figure 3 – Flow chart of the hangover procedure

where $hang_sp$, $hang_f_mus$ and $hang_b_mus$ are three independent hangover counters for speech, foreground music and background music. The hangover for speech addresses low energy frames where misdetection can occur at the offset of speech bursts. The hangover for foreground music

helps to resolve occasional misdetections at foreground music signals and the hangover for background music helps to maintain the quality of the music background. In addition, the first 200 inactive final VAD indications are forced to “1” to protect the beginning of the input where misdetections can occur due to false initialization of the background estimate. The $hang_sp$ is reset to a maximum value $hang_s$ when five successive foreground frames are detected. The value of $hang_s$ is determined jointly by the long term SNR $lsnr$ and the operating point. For the reset of $hang_b_mus$ and $hang_f_mus$, refer to Clause 7.1.12 and 7.1.14 respectively.

7.1.8 Calculation of tone stability

A set of local spectral peaks of the power spectrum of the input frame is identified within the spectral band between 250 Hz and 3300 Hz. A local spectral peak is identified as being higher than its immediate left and right neighboring spectral bins.

For each i^{th} local spectral peak, $D_{p2s}(i)$, which is the power distance between it and its adjacent four spectral bins on its two sides, is calculated:

$$D_{p2s}(i) = \left| 4 \cdot peak_{loc}(i) - S_{idx_peak_{loc}(i-1)} - S_{idx_peak_{loc}(i-2)} - S_{idx_peak_{loc}(i+1)} - S_{idx_peak_{loc}(i+2)} \right| \quad (7.1.19)$$

where S_j is the power of the j^{th} spectral bin, $peak_{loc}(i)$ is the power of the i^{th} local spectral peak, $idx_peak_{loc}(i)$ is the position of the i^{th} local spectral peak in the spectrum (the frame index $[m]$ is omitted for brevity). The positions of the three local spectral peaks with the maximal values of D_{p2s} are denoted by $idx_peak_{max}(0)$, $idx_peak_{max}(1)$ and $idx_peak_{max}(2)$, where:

$$D_{p2s}(idx_peak_{max}(0)) \geq D_{p2s}(idx_peak_{max}(1)) \geq D_{p2s}(idx_peak_{max}(2)) \quad (7.1.20)$$

For each of the three stored spectral positions, its distance to the position of the spectral peak which has the maximum D_{p2s} in the previous frame, $idx_peak_{max_old}$, is obtained by:

$$\begin{aligned} D_{p2old}(0) &= idx_peak_{max}(0) - idx_peak_{max_old} \\ D_{p2old}(1) &= idx_peak_{max}(1) - idx_peak_{max_old} \\ D_{p2old}(2) &= idx_peak_{max}(2) - idx_peak_{max_old} \end{aligned} \quad (7.1.21)$$

If any of the $D_{p2old}(0)$, $D_{p2old}(1)$, $D_{p2old}(2)$ is less than 2, the counter $tone_sta_cnt$ for tone stability is incremented by 1. The $idx_peak_{max}(0)$ is used to update $idx_peak_{max_old}$ for the next frame. Further, the three maxima of the local spectral peaks are identified and normalized by the average of the spectral power of the input frame from the 3rd to the 64th spectral bins, without the three spectral peak maxima. If the sum of the three normalized maxima is greater than one threshold or the sum of the two largest of the three normalized maxima is greater than a second threshold or the largest among the three normalized maxima is greater than a third threshold, DTMF signal is detected and indicated by setting the DTMF flag $dtmf_flg$.

7.1.9 Calculation of spectral peak fluctuation

The global spectral peak of the m^{th} input frame $peak_{gLB}^{[m]}$ is identified as the largest local spectral peak of the input frame. The position of the input frame’s global spectral peak in its spectrum $idx_peak_{gLB}^{[m]}$ is compared to that of the previous frame $idx_peak_{gLB}^{[m-1]}$. If $idx_peak_{gLB}^{[m]}$ is different from $idx_peak_{gLB}^{[m-1]}$, the counter for spectral peak fluctuation $peak_flux_cnt$ is incremented by 1. The $peak_flux_cnt$ is used in the background signal estimation.

7.1.10 Calculation of spectral peakiness

The local spectral peaks identified in Clause 7.1.8 are used to calculate the spectral peakiness of the input frame. For each of the j^{th} local spectral peak, a lower frequency spectral valley $E_{vl}(j)$ and a higher frequency spectral valley $E_{vh}(j)$ are defined by the lowest value of the power spectrum in the 4 FFT bins with frequency below the bin of the j^{th} local spectral peak and by the lowest value of the power spectrum in the 4 FFT bins with frequency above the bin of the j^{th} local spectral peak, respectively. This can be mathematically expressed as:

$$E_{vl}(j) = \text{MIN}[S(\text{idx_peak}_{loc}(j)-1), S(\text{idx_peak}_{loc}(j)-2), \\ S(\text{idx_peak}_{loc}(j)-3), S(\text{idx_peak}_{loc}(j)-4)] \quad (7.1.22)$$

$$E_{vh}(j) = \text{MIN}[S(\text{idx_peak}_{loc}(j)+1), S(\text{idx_peak}_{loc}(j)+2), \\ S(\text{idx_peak}_{loc}(j)+3), S(\text{idx_peak}_{loc}(j)+4)] \quad (7.1.23)$$

where, $S(k)$ is the k^{th} spectral bin of the input frame, $\text{idx_peak}_{loc}(j)$ is the location of the j^{th} local spectral peak in the spectrum. The normalized peak to valley distance is calculated by

$$D_{p2v}(j) = \frac{62 \cdot (2 \cdot \text{peak}_{loc}(j) - E_{vl}(j) - E_{vh}(j))}{\sum_{i=2}^{63} S(i)} \quad (7.1.24)$$

The spectral peakiness SP is obtained by summing the three maxima of $D_{p2v}(j)$ over all values of j .

7.1.11 Calculation of frequency stability

The parameter of frequency stability is used in the background update decision. The spectrum of the input frame is whitened by a moving average of the spectrums over past frames. The frequency stability is obtained as the power deviation of the 16 whitened sub-bands.

$$\text{sta}_{fq} = \frac{1}{16} \sum_{i=0}^{15} [E_{band}(i) / \overline{E_{band}(i)} - E_{band_mean}]^2, \quad (7.1.25)$$

where $E_{band}(i)$ is the energy of the i^{th} sub-band of the input frame (see Clause 7.1.3), $\overline{E_{band}(i)}$ is the moving average of the power of the i^{th} sub-band over past frames, E_{band_mean} is the power mean of the 16 whitened sub-bands. E_{band_mean} is calculated by

$$E_{band_mean} = \frac{1}{16} \sum_{i=0}^{15} E_{band}(i) / \overline{E_{band}(i)} \quad (7.1.26)$$

and $\overline{E_{band}(i)}$ is updated every time after the sta_{fq} calculation as

$$\overline{E_{band}(i)} = 0.9 \cdot \overline{E_{band}(i)} + (1-0.9) \cdot E_{band}(i) \quad (7.1.27)$$

7.1.12 Background music detection

Detection of background music is done for every 100 past background frames. Background music is detected if the sum of the spectral peakinesses over the 100 past background frames is sufficiently high. A hangover of 1000 frames is set once music background is detected, but will be fast exited if the sum of the spectral peakinesses over the 100 past background frames is sufficiently low. Moreover, the decision threshold for whether the sum of the spectral peakiness is high enough is

biased according to whether the background music hangover exits. The procedure of the background music detection is depicted in flow chart below:

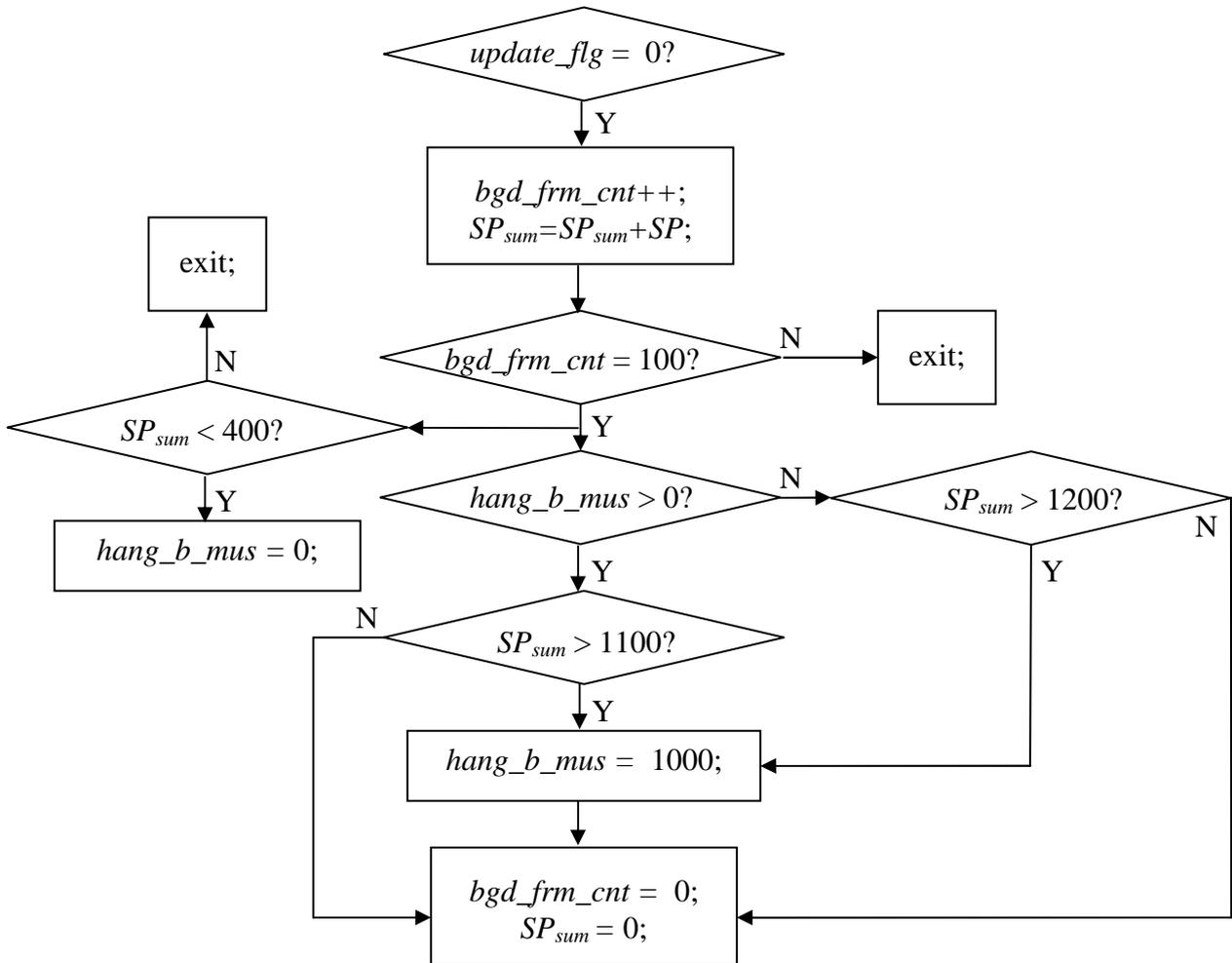


Figure 4 – Flow chart of the background music detection procedure

where the background frame is indicated by the background update flag *update_flg* (“0” for background frame), *bgd_frm_cnt* is a counter counting the number of background frames, *hang_b_mus* is the hangover counter for background music, *SP_{sum}* is the spectral peakiness accumulator storing the sum of the frame peakinesses.

7.1.13 Silence detection

The signal power in dBov of the input frame is calculated by:

$$FP_{dBov} = 20 \cdot \log_{10}(rms/32767), \quad (7.1.28)$$

where, *rms* is the RMS of the input frame (see Clause 7.1.4). If *FP_{dBov}* is below a silence threshold of -56 dBov by default, the input frame is indicated as silence frame.

7.1.14 Background estimate update

A background identification threshold thr_{bgd} is first determined jointly from the long term SNR, $lsnr$, and the background music hangover $hang_b_mus$.

$$thr_{bgd} = \begin{cases} 40 \text{ (WB input) or } 34.375 \text{ (NB input)} & lsnr > 18 \text{ and } hang_b_mus > 0 \\ 15 & o.w. \end{cases} \quad (7.1.29)$$

The decision whether to update the background signal estimate consists of following steps:

1) If M consecutive frames (hereafter called consecutive M background frames) whose $MSSNRs$ are not greater than thr_{bgd} are detected and no DTMF signal is detected and the foreground music hangover $hang_f_mus$ is not greater than 0, the input frame is used to update the background estimate by setting the background update flag $update_flg$, and a reset flag $reset_flg$ is also set (the purpose of the $reset_flg$ is described later in this clause). The number of consecutive background frames M is determined from $lsnr$:

$$M = \begin{cases} 2 & lsnr > 15 \\ 4 & lsnr \leq 15 \end{cases} \quad (7.1.30)$$

The counting for the consecutive M background frames tolerates occasional frames whose $MSSNRs$ are greater than thr_{bgd} . If M or more than M consecutive background frames have just been detected and the $MSSNR$ of the input frame is higher than thr_{bgd} , the number of counted consecutive background frames is set to be $M - 1$. In any case, if 10 or more than 10 consecutive frames whose $MSSNRs$ are greater than 50 are detected, the counting for consecutive background frames is reset to be 0. This step applies when the background estimate converges to the real background signal.

2) While the background estimate is not converged to the real background signal, background characteristic is sought within time windows of 30 frames one by one. The possibility of containing background frames for the time window is determined based on the analysis of the tonality and the stability of the frames within the time window. If all the frames contained within the time window or most of the frames but a few not at the end of the time window contained within the time window are recognized as background frames, the last frame (the input frame) within the time window is used to update the background estimate by setting the background update flag $update_flg$. Otherwise, if the time window is believed to contain background frames but with less confidence, the minimum characteristic within the time window is used to update the background estimate. Details of the aforementioned method are described as: If the condition in step 1 is not satisfied, and if no DTMF signal is detected and the foreground music hangover $hang_f_mus$ is 0, following operations are performed, otherwise the $reset_flg$ is set. A counter con_frm_cnt for counting the number of consecutive frames is incremented by 1; The spectral peakiness SP of the input frame is compared to the threshold thr_{SP_low} which is 15 (for NB input) and 12.7 (for WB input). If $SP < thr_{SP_low}$ a counter low_SP_cnt for counting the number of frames with low spectral peakiness is incremented by 1; The frequency stability sta_{fq} is compared to the threshold thr_{fst} which is 12 (for NB input) and 10 (for WB input). If $sta_{fq} < thr_{fst}$, a counter $high_fst_cnt$ for counting the number of frames with high frequency-stability is incremented by 1; The energy of each sub-band $E_{band}(i)$ of the input frame is compared to the energy of the corresponding band $E_{band_buf_min}(i)$ (stored in a minimum band energy buffer). The $E_{band_buf_min}(i)$ in the buffer is updated by replacing the old $E_{band_buf_min}(i)$ by the $E_{band}(i)$ if $E_{band}(i) < E_{band_buf_min}(i)$. Once con_frm_cnt reaches 30, the low_SP_cnt is compared to a threshold of 15. If low_SP_cnt is no greater than 15, the $reset_flg$ is set. Otherwise, the following condition is examined

$$\begin{aligned} & \text{if } (high_fst_cnt > 28 \ \&\& \ tone_sta_cnt < 12 \\ & \ \&\& \ sta_{fq} < 12 \ \&\& \ peak_flux_cnt > 15) \end{aligned} \quad (7.1.31)$$

If the condition above is satisfied, the background update flag *update_flg* is set to 1. Otherwise if following condition is satisfied,

$$\text{if } (\text{peak_flux_cnt} > 15 \ \&\& \ \text{tone_sta_cnt} < 12) \quad (7.1.32)$$

the sub-band energies $E_{band_buf_min}(i)$ stored in the minimum band energy buffer are used to update the background estimate. Otherwise the *reset_flg* is set. Moreover, if *low_SP_cnt* < 5 and *high_fst_cnt* > 28, foreground music is considered to be detected and the foreground music hangover *hang_f_mus* is set to 10. If *con_frm_cnt* exceeds 30, only the frequency stability *sta_{fq}* is examined. If *sta_{fq}* < 12, the background update flag *update_flg* is set. Otherwise, the *reset_flg* is set. Step 1 and 2 are depicted by flow chart below

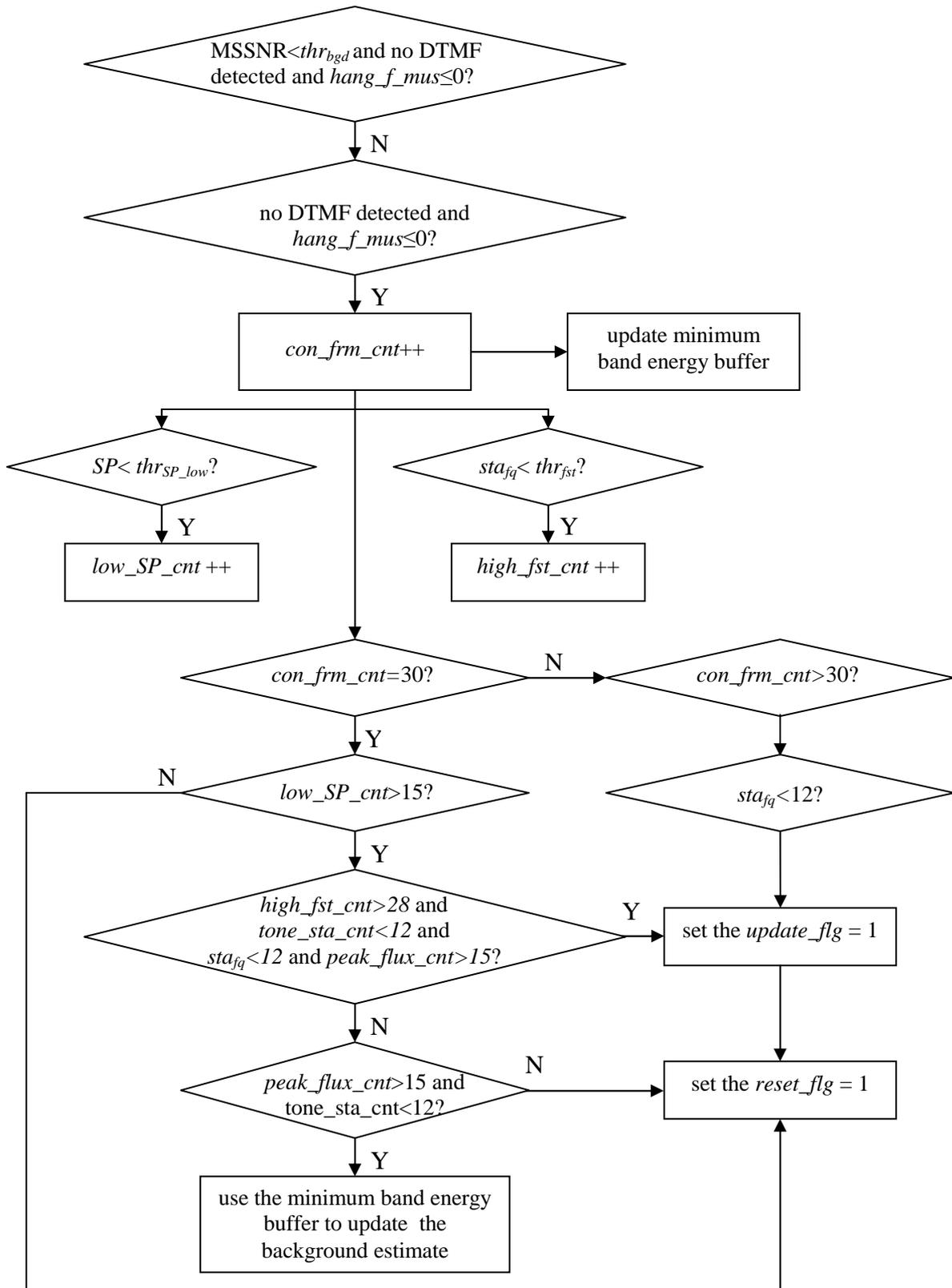


Figure 5 – Flow chart of the background estimate update procedure

If *update_flg* is set to 1, the background estimate for ZCR and sub-band energies are updated as

$$\overline{ZCR}_n^{[m]} = 0.95 \cdot \overline{ZCR}_n^{[m-1]} + (1 - 0.95) \cdot ZCR \quad (7.1.33)$$

$$\overline{E_{band_n}(i)}^{[m]} = 0.9 \cdot \overline{E_{band_n}(i)}^{[m-1]} + (1-0.9) \cdot E_{band}(i) \quad (7.1.34)$$

where $\overline{ZCR}_n^{[m]}$ and $\overline{ZCR}_n^{[m-1]}$ are the moving average of the ZCR of the background signal at the current and the previous frames respectively, $\overline{E_{band_n}(i)}^{[m]}$ and $\overline{E_{band_n}(i)}^{[m-1]}$ are the moving averages of the energy of the i^{th} sub-band of the background signal at the current and the previous frames, $i = 0,1,2,\dots,15$. \overline{ZCR}_n is initialized at the first four input frames being the average of the current and the past frames' ZCR. $\overline{E_{band_n}(i)}$ for all i is initialized at the first four input frames being the energy of the i^{th} sub-band of the input frame which is further limited by a minimum value. If successive high level signal is detected during the first four input frames, \overline{ZCR}_n and $\overline{E_{band_n}(i)}$ for all i are reset.

7.2 Detailed description of the Speech/Music Discrimination module

This clause describes the Speech/Music Discrimination module. The calculation of the main discrimination parameter, the variance of the flux, is described in Clause 7.2.1 and the calculations of two peakiness measures is described in Clause 7.2.2. Clause 7.2.3 provides the details of the Speech/Music Discrimination decision algorithm.

7.2.1 Calculation of the flux and the variance of the flux

The *flux* parameter measures the normalized distance between the power spectrum of the current frame and the power spectrum of the previous frames. The difference is measured for the spectral band between 62 Hz to 2914 Hz for both narrowband (NB) and wideband (WB) signals. Let $S^{[m]}(i)$ denote the i^{th} power spectrum component of the m^{th} frame. The *flux* for the m^{th} frame is calculated by:

$$flux^{[m]} = \frac{\sum_{n=1}^3 \sum_{i=1}^{47} |S^{[m]}(i) - S^{[m-n]}(i)|}{\sum_{n=1}^3 \sum_{i=1}^{47} (S^{[m]}(i) + S^{[m-n]}(i))} \quad (7.2.1)$$

The variance of the flux is calculated for the last 20 frames which are detected as “active” frames by a simple threshold on the *MSSNR* value (see Clause 7.1.3 for the description of the *MSSNR* calculations). The values of the flux for each active frame are stored in FIFO buffer, which is updated only for active frames. The *MSSNR* of the m^{th} frame is denoted in the sequel by $MSSNR^{[m]}$. (Note that the active frames for the calculation of the variance of the flux, which is based on *MSSNR*, are different from the active frames indicated by the VAD algorithm part of the GSAD.)

For the first 20 active frames the variance is set to 0 and at the 20th active frame an average of the flux, *mov_flux*, is initialized to the arithmetic average of the flux value of the first 20 active frames. For each frame after the first 20 active frames, the average is updated for each active frame by:

$$mov_flux^{[m]} = 0.99 \cdot mov_flux^{[m-1]} + 0.01 \cdot flux^{[m]} \quad (7.2.2)$$

The variance of the *flux* for the m^{th} frame, $var_flux^{[m]}$, is calculated as:

$$var_flux^{[m]} = \sum_{k=0}^{19} \left(flux^{[m-k]} - mov_flux^{[m]} \right)^2 \quad (7.2.3)$$

where the index m is incremented only for the active frames. The values of $var_flux^{[m]}$ for the first 20 active frames are multiplied by a linearly ramping window. A buffer of the last 120 frames of the active $var_flux^{[m]}$ is used by the SMD algorithm.

7.2.2 Calculation of two spectral-peaks peakiness measures

Two measures of the peakiness of the large peaks of the power spectrum are calculated. The peaks of the power spectrum are found by searching the power spectrum for samples that are higher than their immediate neighbouring samples. The K highest-value spectral peaks, up to $K=5$, which we denote by $S^{[m]}(i)$, $i = 1, \dots, K$, are used (if less peaks are found, K might be lower than 5.).

The first peakiness measure is calculated by:

$$P_1 = \frac{\sqrt{\frac{1}{K} \sum_{i=1}^K S^2(i)}}{\frac{1}{K} \sum_{i=1}^K |S(i)|} - 1 \quad (7.2.4)$$

The second peakiness measure is calculated by:

$$P_2 = \frac{\max_{i=1}^K |S(i)|}{\frac{1}{K} \sum_{i=1}^K |S(i)|} - 1 \quad (7.2.5)$$

The running averages of the two peakiness measures are calculated by:

$$avg_P_1^{[m]} = 0.995 \cdot avg_P_1^{[m-1]} + 0.005 \cdot P_1 \quad (7.2.6)$$

and

$$avg_P_2^{[m]} = 0.995 \cdot avg_P_2^{[m-1]} + 0.005 \cdot P_2. \quad (7.2.7)$$

7.2.3 Speech/Music Discrimination decision

The SMD algorithm uses a buffer of $L=120$ past $var_flux^{[m]}$ parameters and a binary-value buffer that contains information on $MSSNR^{[m]}$ for the past 512 frames.

An adaptive threshold $T_{var_flux}^{[m]}$ is calculated using $MSSNR^{[m]}$ parameter for the past 512 frames. At the first step, a parameter which follows the maximal value of $MSSNR^{[m]}$ for each frame m is calculated. The maximal value parameter, $max_{MSSNR}^{[m]}$, is set to $MSSNR^{[m]}$ if $MSSNR^{[m]}$ is larger than $max_{MSSNR}^{[m]}$ and is otherwise decremented by a multiplicative factor of 0.9999. This reduction allows a slow adaptation of the maximum value when $MSSNR^{[m]}$ is decreased over a long period. An adaptive threshold is set by multiplying $max_{MSSNR}^{[m]}$ by a constant which depends on the operating point:

$$T_{MSSNR}^{[m]} = C_{op} \cdot max_{MSSNR}^{[m]} \quad (7.2.8)$$

where C_{op} is 0.5 for the Bandwidth Saving and the Balanced operating points and is 0.45 for the Quality Preferred operating point.

The information about the past $MSSNR^{[m]}$ parameters is saved in a binary buffer which indicates, for each entry m , if $MSSNR^{[m]}$ is higher than $T_{MSSNR}^{[m]}$. Using this binary buffer, for each frame after the first 300 frames, a 2-bin histogram of the last K frames is calculated, where $high_{bin}^{[m]}$ counts the number of frames in the buffer where $MSSNR^{[m]}$ is higher than $T_{MSSNR}^{[m]}$ and $low_{bin}^{[m]}$ counts the number of frames in the buffer where $MSSNR^{[m]}$ is not higher than $T_{MSSNR}^{[m]}$. Obviously, only $high_{bin}^{[m]}$ needs to be calculated, since $low_{bin}^{[m]} = L - high_{bin}^{[m]}$. The value of K is incremented by 1 for each frame after the first 300 frames until it reaches 512, where it is fixed. The calculation of $high_{bin}^{[m]}$ is done simply by considering the current frame m (increment $high_{bin}^{[m]}$ if $MSSNR^{[m]}$ is higher than $T_{MSSNR}^{[m]}$) and the frame $m-512$ (decrement $high_{bin}^{[m]}$ if the last element in the buffer is 1, which means that $MSSNR^{[m-512]} > T_{MSSNR}^{[m-512]}$). From $high_{bin}^{[m]}$, a histogram difference measure, which is between -1 and 1, is calculated according to:

$$diff_{hist}^{[m]} = \frac{high_{bin}^{[m]} - low_{bin}^{[m]}}{M} = 1 - \frac{2 \cdot high_{bin}^{[m]}}{M}. \quad (7.2.9)$$

The final histogram difference measure is generated as a running average of the instantaneous histogram difference and is further biased with an offset factor, Δ_{op} , which depends on the operating point and is a small negative or positive number:

$$diff_{hist}^{avg} = 0.9 \cdot diff_{hist}^{avg} + 0.1 \cdot (diff_{hist}^n + \Delta_{op}) \quad (7.2.10)$$

where for the first 300 frames $diff_{hist}^{avg}$ is set to Δ_{op} . The average difference measure is then limited to be between $-X_T$ and X_T , where $X_T = 0.6$, and is denoted by $diff_{hist}^{final}$. The adaptive threshold for var_flux_n is given by the linear equation:

$$T_{var_flux}^{[m]} = A \cdot diff_{hist}^{final} + B \quad (7.2.11)$$

where A and B are calculated by:

$$A = \frac{T_{op}^{up} - T_{op}^{down}}{2 \cdot X_T} \quad (7.2.12)$$

$$B = \frac{T_{op}^{up} + T_{op}^{down}}{2} \quad (7.2.13)$$

T_{op}^{up} is the desired highest value for $T_{var_flux}^{[m]}$ and T_{op}^{down} is the desired lowest value for $T_{var_flux}^{[m]}$. For music signals (that typically have a lower var_flux) the threshold is set to a higher value, which helps in preferring the detection of music signals. Similarly, for speech signals (that typically have a higher var_flux) the threshold is set to a lower value, which helps in preferring the detection of speech signals.

Using the adaptive threshold, $T_{var_flux}^{[m]}$, the ratio of the times $var_flux^{[m]}$ was above that adaptive threshold for the past J frames is calculated for the frames which are declared active based on the threshold on $MSSNR$, or the first frame where the LC-VAD is active and the threshold of $MSSNR$ was not achieved yet. This definition holds also for the reset mechanism described below. If the

ratio is above 0.5, the raw SMD decision is set to speech and if the ratio is below 0.5 the raw SMD decision is set to music. The value of J is incremented for each frame declared active by the criterion above until it reaches 120, where it is fixed.. For the first 75 active frames, the ratio is ‘biased’ toward speech, to help avoiding some speech-to-music misclassification at the beginning, where the decision based on the yet-unstable adaptive threshold might be misguided.

A special mechanism is used to detect the end of long music segments and to reset the SMD at such conditions. This mechanism uses the fact that typical music segments are characterized by high $MSSNR$ over a long period. At a first step, a segment of at least 500 frames where $MSSNR$ is continuously above the fixed value of 8 is detected. At the end of such segment, which is called “falling edge”, the algorithm checks if the next segment of 75 frames if for at least 40 of these frames $MSSNR$ is below 15. If such low $MSSNR$ segment is detected, the algorithm declares an end of music segment. The “falling edge” conditions lasts for 150 frames. Once a low $MSSNR$ segment is detected, the SMD algorithm is being “reset”. The reset operation sets $diff_{hist}^{final}$ to Δ_{op} for a period of 400 frames, which means that the previously calculated preference for music is neutralized. The reset also sets J to 0, which means that the previously accumulated values of var_flux in the buffer are not used, as well as applying the bias toward speech for the next 75 frames.

The values of the two running averages of the spectral-peaks peakiness measures are used to correct the raw SMD decision toward music. For NB signals, if $avrg_P_1^{[m]} > 0.43$ or $avrg_P_2^{[m]} > 1.76$, the raw SMD decision is set to music. For WB signals, if $avrg_P_1^{[m]} > 0.52$ or $avrg_P_2^{[m]} > 2.05$, the raw SMD decision is set to music.

The final SMD decision is obtained by a one frame hangover of the raw SMD decision.

8 Organization of the reference C code

Table 7 – Organization of the GSAD source files

File name	Description
commandpars.c/h	Reading the command line
ereal_fft_fx.c/h	Modified real FFT
Gsad.c	Main function
gsad_math_adv_fx.c/h	Arithmetic functions
Parameters_fx.c/h	Parameter calculation functions
Preproc_fx.c/h	Pre-processing functions
vad_fx.c/h	Main VAD function
smd_fx.c/h	Main SMD function
typedef.h	Macro defines for data type
Rom_fx.c/h	Tables for rom
constdef_fx.h	Macro defines for constants