International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

G.720.1

(01/2010)

SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

Digital terminal equipments – Coding of voice and audio signals

Generic sound activity detector

Recommendation ITU-T G.720.1

ITU-T

ITU International Telecommunication Union

ITU-T G-SERIES RECOMMENDATIONS

TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

INTERNATIONAL TELEPHONE CONNECTIONS AND CIRCUITS	G.100–G.199
GENERAL CHARACTERISTICS COMMON TO ALL ANALOGUE CARRIER-TRANSMISSION SYSTEMS	G.200–G.299
INDIVIDUAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON METALLIC LINES	G.300–G.399
GENERAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON RADIO-RELAY OR SATELLITE LINKS AND INTERCONNECTION WITH METALLIC LINES	G.400–G.449
COORDINATION OF RADIOTELEPHONY AND LINE TELEPHONY	G.450–G.499
TRANSMISSION MEDIA AND OPTICAL SYSTEMS CHARACTERISTICS	G.600–G.699
DIGITAL TERMINAL EQUIPMENTS	G.700–G.799
General	G.700–G.709
Coding of voice and audio signals	**G.710–G.729**
Principal characteristics of primary multiplex equipment	G.730–G.739
Principal characteristics of second order multiplex equipment	G.740–G.749
Principal characteristics of higher order multiplex equipment	G.750–G.759
Principal characteristics of transcoder and digital multiplication equipment	G.760–G.769
Operations, administration and maintenance features of transmission equipment	G.770–G.779
Principal characteristics of multiplexing equipment for the synchronous digital hierarchy	G.780–G.789
Other terminal equipment	G.790–G.799
DIGITAL NETWORKS	G.800–G.899
DIGITAL SECTIONS AND DIGITAL LINE SYSTEM	G.900–G.999
MULTIMEDIA QUALITY OF SERVICE AND PERFORMANCE – GENERIC AND USER-RELATED ASPECTS	G.1000–G.1999
TRANSMISSION MEDIA CHARACTERISTICS	G.6000–G.6999
DATA OVER TRANSPORT – GENERIC ASPECTS	G.7000–G.7999
PACKET OVER TRANSPORT ASPECTS	G.8000–G.8999
ACCESS NETWORKS	G.9000–G.9999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T G.720.1

Generic sound activity detector

Summary

Recommendation ITU-T G.720.1 describes an independent front-end processing module implementing a generic sound activity detector (GSAD) that can be applied prior to signal processing applications and can operate on narrow-band or wideband audio input using a 10-ms frame length (without lookahead), such as used by speech or audio codecs. The primary function of the GSAD is to indicate the input frame activity for performing voice activity detection (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. In order to apply GSAD in specific cases, an adaptation layer may be required.

An external control signal indicates to the GSAD algorithm which one of the three different operating points to use, namely: bandwidth-saving, balanced and quality-preferred operating points. 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-user's 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 speech and music classification for the active frames, which can be utilized for 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: signal to noise ratio (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.

The main body of 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.

Annex A describes a standalone generic voice activity detector (GVAD) that can be applied prior to signal processing applications and can operate on narrow-band or wideband audio input using 10 ms frame length (without lookahead), such as used by speech or audio codecs. Its function is to indicate the input frame activity. In order to apply GVAD in specific cases, an adaptation layer may be required.

The Recommendation also contains an electronic attachment with the ANSI C source code which forms an integral part of this Recommendation, and a set of test vectors. The set of test vectors is also available for download from the ITU-T Test Signal Database at: http://www.itu.int/net/ITU-T/sigdb/speaudio/Gseries.htm#G.720.1.

History

Edition	Recommendation	Approval	Study Group
1.0	ITU-T G.720.1	2010-01-13	16

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

© ITU 2011

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

Recommendation ITU-T G.720.1

Generic sound activity detector

1 Scope

The generic sound activity detector (GSAD) is an independent front-end processing module which can be applied prior to signal processing applications that operate on narrow-band 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. In order to apply GSAD in specific cases, an adaptation layer may be required.

This Recommendation is organized as follows.[1] 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 normative reference to any other standards.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation does not use terms defined elsewhere.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 background signal: The interfering signals (e.g., background noise) at the background of the dominating speech or music signals.

3.2.2 balanced operating point: Generic sound activity detector (GSAD) operating point that provides a middle point performance between the other two operating points in either voice activity detection or speech/music discrimination. See clause 6.4.

3.2.3 bandwidth-saving operating point: Generic sound activity detector (GSAD) operating point privileging accurate detection for background signals in voice activity detection, or privileging accurate detection for speech signals in speech/music discrimination. See clause 6.4.

3.2.4 foreground signal: The dominating information signal (speech or music) in the speech channel.

---

[1] This Recommendation contains an electronic attachment with the ANSI C source code which forms an integral part of this Recommendation, and a set of test vectors. The set of test vectors is also available for download from the ITU-T Test Signal Database at: http://www.itu.int/net/ITU-T/sigdb/speaudio/Gseries.htm#G.720.1.

3.2.5 quality-preferred operating point: Generic sound activity detector (GSAD) operating point privileging accurate detection for foreground signals in voice activity detection, or privileging accurate detection for music signals in speech/music discrimination. See clause 6.4.

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

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

4 Abbreviations and acronyms

This Recommendation uses the abbreviations and acronyms defined in clause 4.1 and the symbols used throughout this Recommendation are listed in clause 4.2.

4.1 Glossary of acronyms

DZCR	Differential Zero Crossing Rate
FFT	Fast Fourier Transform
GSAD	Generic Sound Activity Detection/Detector
GSAD_NB	Generic Sound Activity Detecting mode for Narrow-band signals
GSAD_WB	Generic Sound Activity Detecting mode for Wideband signals
GVAD	Generic Voice Activity Detection/Detector
GVAD_NB	GVAD mode for Narrow-band signals
GVAD_WB	GVAD mode for Wideband signals
MSSNR	Modified Segmental SNR
NB	Narrow-band
RMS	Root Mean Square
SDF	Spectral Density Function
SiD	Silence Detector
SMD	Speech/Music Discriminator
SNR	Signal to Noise Ratio
VAD	Voice Activity Detection/Detector
VAD_NB	Voice Activity Detecting mode for Narrow-band signals
VAD_WB	Voice Activity Detecting mode for Wideband signals
WB	Wideband
WMOPS	Weighted Million Operations Per Second
ZCR	Zero Crossing Rate

4.2 Glossary of symbols

Δ_{op}	An offset factor for $diff_{hist}^{[m]}$
$avrg\_P_1^{[m]}$	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

bgd_frm_cnt	The counter counting the number of background frames in the background music detection
con_frm_cnt	The counter counting the number of consecutive frames in the background estimate update procedure
C_{op}	A constant that depends on the operating point, used to calculate $T_{MSSNR}^{[m]}$
$D_{p2s}(i)$	The power distance between the *i*-th local spectral peak and its adjacent four spectral bins on its two sides
$D_{p2v}(j)$	The normalized peak to valley distance of the *j*-th local spectral peak
$diff_{hist}^{[m]}$	An initial difference measure between $high_{bin}^{[m]}$ and $low_{bin}^{[m]}$
$diff_{hist}^{avg}$	An average measure of $diff_{hist}^{[m]}$
$diff_{hist}^{final}$	A final difference measure between $high_{bin}^{[m]}$ and $low_{bin}^{[m]}$
dtmf_flg	The flag indicating the presence of a DTMF signal
DZCR	The differential zero crossing rate of the input frame
$E_{band}(i)$	The energy of the *i*-th sub-band
$E_{band\_buf\_min}(i)$	The minimum energy of the *i*-th sub-band of the past frames
$E_{band\_old}(i)$	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
$\overline{E_{band}(i)}$	The moving average of the power of the *i*-th sub-band over past frames
E_{band_mean}	The power mean of the 16 whitened sub-bands in the calculation of frequency stability
$E_{vh}(j)$	The higher frequency spectral valley of the *j*-th local spectral peak
$E_{vl}(j)$	The lower frequency spectral valley of the *j*-th local spectral peak
$flux_{bgd}$	The fluctuation of the background signal
$flux^{[m]}$	A spectral deviation measure for the *m*-th frame
FP_{dBov}	The signal power in dBov
hang_b_mus	The hangover counter for background music
hang_f_mus	The hangover counter for foreground music
hang_sp	The hangover counter for speech
$high_{bin}^{[m]}$	A counter on the *MSSNR* high values
high_fst_cnt	The counter counting the number of frames with high frequency-stability
$idx\_peak_{loc}(i)$	The position of the *i*-th local spectral peak in the spectrum
$idx\_peak_{max}(j)$	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
$idx\_peak_{gLb}^{[m]}$	The location of $peak_{gLb}^{[m]}$ in its spectrum
ivad	The initial VAD decision
$low_{bin}^{[m]}$	A counter on the *MSSNR* low values
low_SP_cnt	The counter counting the number of frames with low spectral peakiness
lsnr	The long term SNR of the input frame

$max_{MSSNR}^{[m]}$	A maximal value of *MSSNR* over past frames
$mov\_flux^{[m]}$	The moving average of $flux^{[m]}$ at the *m*-th frame
msnr(*i*)	The modified SNR of the *i*-th sub-band
MSSNR	The modified segmental SNR
P_1	A first peakiness measure
P_2	A second peakiness measure
$peak_{loc}(i)$	The power of the *i*-th local spectral peak
$peak_{gLb}^{[m]}$	The largest local spectral peak of the *m*-th input frame
peak_flux_cnt	The counter counting the spectral peak fluctuation
reset_flg	The flag indicating whether to reset counters used in the background estimate update procedure
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
s(*i*)	The *i*-th sample in the input frame
snr(*i*)	The SNR of the *i*-th sub-band of 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
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
sta_{fq}	The frequency stability of the input frame
$T_{var\_flux}^{[m]}$	Adaptive threshold on $var\_flux^{[m]}$
$T_{MSSNR}^{[m]}$	An adaptive threshold of the *MSSNR*
T_{op}^{down}	Lowest desired value for $T_{var\_flux}^{[m]}$
T_{op}^{up}	Highest desired value for $T_{var\_flux}^{[m]}$
thr_{bgd}	The threshold for background frame identification
thr_{fst}	The threshold for high frequency-stability identification
thr_{SP_low}	The threshold for low spectral peakiness identification
thr_{vad}	The threshold for the initial VAD decision
tone_sta_cnt	The counter counting the number of frames with high tone stability
update_cnt	A counter for counting consecutive background frames
update_flg	The flag indicating whether to update the background estimate
$var\_flux^{[m]}$	The variance of $flux^{[m]}$ at the *m*-th frame
X_T	A threshold on $diff_{hist}^{avg}$
ZCR	The zero crossing rate of the input frame

$\overline{ZCR_n}$ The moving average of the *ZCR* of the estimated background signal

5 Conventions

The following conventions apply to this Recommendation:

– $|x|$ denotes the absolute value of x; e.g., $|12| = 12$, $|-3| = 3$

– $\mathrm{sgn}[x]$ denotes the sign of x; e.g., $\mathrm{sgn}[3] = 1$, $\mathrm{sgn}[-5] = -1$

– $\mathrm{MAX}[x_0, x_1, x_2, \ldots x_{N-1}]$ denotes the maximum of $x_0, x_1, x_2, \ldots x_{N-1}$; e.g., $\mathrm{MAX}[0, -1, 3, 5, 2] = 5$

– $\mathrm{MIN}[x_0, x_1, x_2, \ldots x_{N-1}]$ denotes the minimum of $x_0, x_1, x_2, \ldots x_{N-1}$; e.g., $\mathrm{MIN}[0, -1, 3, 5, 2] = -1$

– $\log(x)$ denotes the logarithm in base 10 of x; e.g., $\log(100) = 2$

– $p^{[m]}$ denotes the value of parameter p at the m-th input frame. Current frame is assumed when [m] is omitted

– Σ denotes summation

– o.w. denotes "otherwise"

6 General description of the GSAD algorithm

6.1 Input sampling rate

The GSAD can accept both narrow-band (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., frames of 80 samples for NB audio input and of 160 samples for WB audio input. Consecutive GSAD indications can be combined to indicate the activity for frames with a multiple of 10 ms.

6.3 Delay

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

6.4 Configurations

GSAD operates on narrow-band (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 GSAD to operate in a generic sound activity detecting mode (GSAD_NB and GSAD_WB) or in a voice activity detecting mode (VAD_NB and VAD_WB). When 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 differentiated 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 1 describes the GSAD output bits.

Table 1 – 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 GSAD operates in a generic sound activity detecting mode, both bits are output, which can be interpreted as four possible values indicating a speech ("3"), music ("2"), audible noise ("1") or silence frame ("0"). 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 an active or inactive frame, respectively.

In addition, a third control signal selects the desired operating point from three possible ones. Setting the third control signal to "0" selects the balanced operating point, setting it to "1" selects the quality-preferred operating point and setting the third control signal to "2" selects the bandwidth-saving operating point. When GSAD operates in a voice activity detection mode, the three operating points provide the ability to select a preferred balance 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 correctly classify most of the active speech frames, and the quality-preferred operating point is tuned to correctly classify most of the active music frames. The balanced operating point provides a middle point performance between the other two operating points.

6.5 Complexity and memory cost

Table 2 shows the GSAD complexity in WMOPS for its different modes and signal sampling frequencies. The RAM used for GSAD is 3284 bytes and the table ROM is 1674 bytes.

Table 2 – Complexity of the GSAD

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

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 frames by the speech/music discriminator (SMD) and inactive frames will be further classified into silence or audible noise frames by the silence detector (SiD).

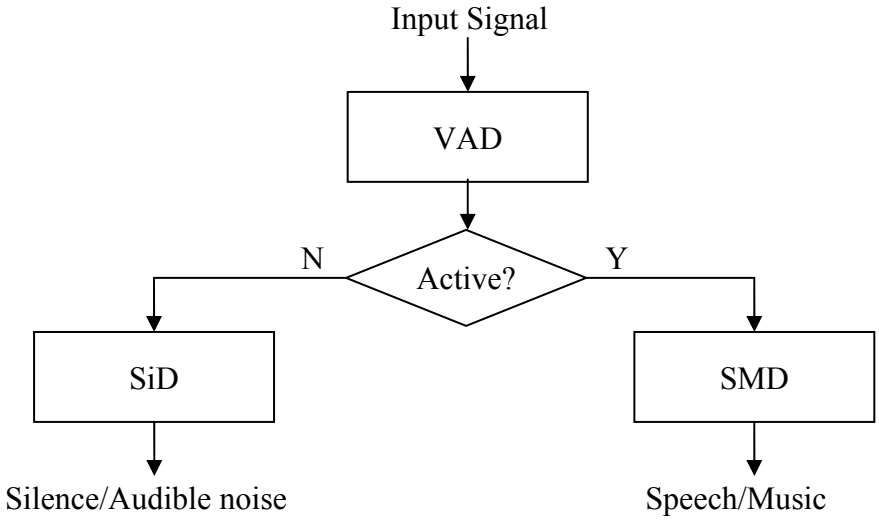


Figure 1 – Block diagram of the GSAD system

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 indicates the number of the relevant clauses in the text.

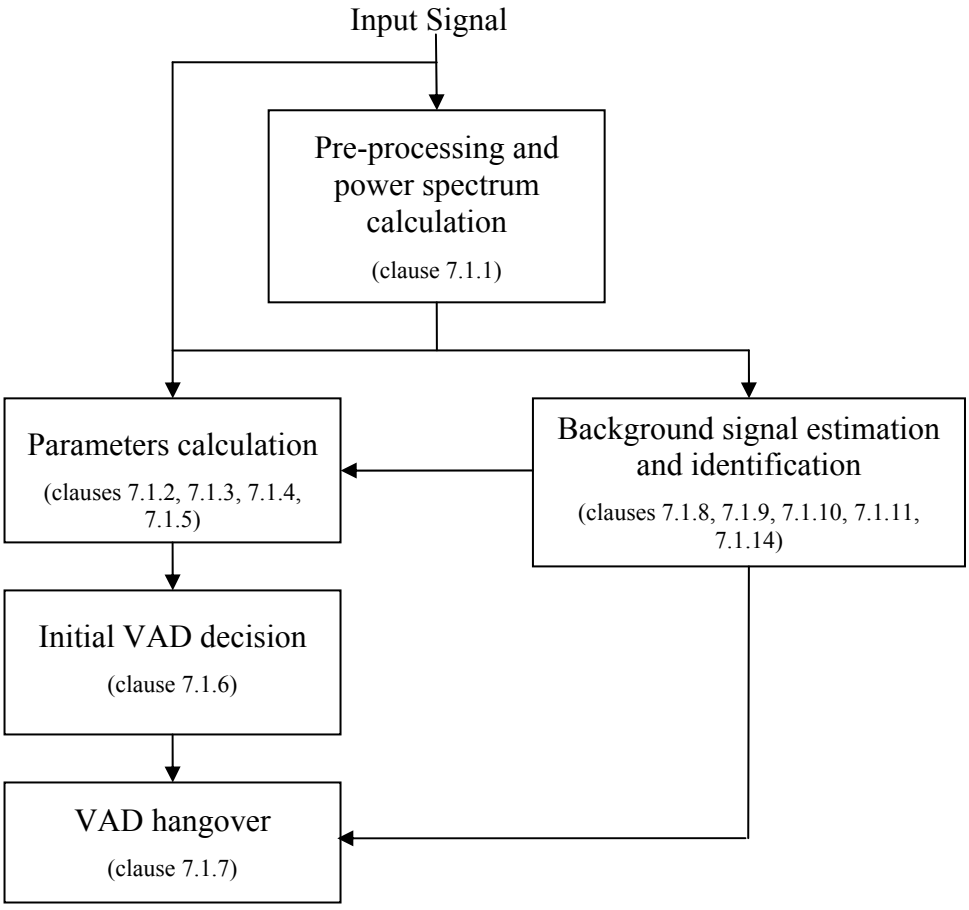


Figure 2 – General diagram of the VAD module

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

$$s_{emp}(i) = s(i-1) + 0.8 \cdot s(i) \tag{7.1-1}$$

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) at the end 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 the square root of the sum of squares of the real and the imaginary parts of the complex FFT coefficients, where *k* is the index for the power spectrum elements. In the following, $S^{[m]}(k)$ 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:

$$ZCR = \frac{1}{2}\sum_{i=0}^{N-2}\left|\mathrm{sgn}[s(i)] - \mathrm{sgn}[s(i+1)]\right| \tag{7.1-2}$$

where *N* is the number of samples per frame (*N* = 80 for NB input and *N* = 160 for WB input) and *s*(*i*) is the *i*-th original sample. The *DZCR* of the input frame is calculated by:

$$DZCR = ZCR - \overline{ZCR_n} \tag{7.1-3}$$

where $\overline{ZCR_n}$ is the moving average of the *ZCR* of the estimated background signal. For the calculation of $\overline{ZCR_n}$, refer 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:

$$E_{band}(i) = \frac{\alpha}{h(i) - l(i) + 1}\sum_{k=\mathrm{l(i)}}^{\mathrm{h(i)}} S(k) + (1-\alpha)E_{band\_old}(i) \tag{7.1-4}$$

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 3 for NB input and in Table 4 for WB input.

Table 3 – Boundaries of the sub-bands for NB input

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

Table 4 – Boundaries of the sub-bands for WB input

i	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
l(i)	2	4	6	8	10	12	14	17	20	23	27	37	49	63	79	97
h(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 (for the calculation of *lsnr*, refer to clause 7.1.4):

$$\alpha = \begin{cases} 0.75 & lsnr > 35 \\ 0.55 & lsnr \le 35 \end{cases} \tag{7.1-5}$$

For the first input frame, α is set to 1.

The sub-band SNR is calculated as:

$$snr(i) = 10\log(E_{band}(i) / \overline{E_{band\_n}(i)}) \tag{7.1-6}$$

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 $\overline{E_{band\_n}(i)}$, refer to clause 7.1.14).

The modified sub-band SNR is calculated by:

$$msnr(i) = \begin{cases} MAX\left[MIN\left[\dfrac{snr^3(i)}{64}, snr(i) \right], 0 \right] & 1 < i \le 12 \\ MAX\left[MIN\left[\dfrac{snr^3(i)}{25}, snr(i) \right], 0 \right] & 0 \le i \le 1 \; or \; i > 12 \end{cases} \tag{7.1-7}$$

Finally, the modified segmental SNR (*MSSNR*) is obtained by:

$$MSSNR = \sum_{i=0}^{15} msnr(i) \tag{7.1-8}$$

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:

$$rms = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s^2(i)} \tag{7.1-9}$$

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).

Following, a long term RMS of the background signal and a long term RMS of the foreground signal are estimated. 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:

$$rms_{bgd}^{[m]} = \beta_b \cdot rms_{bgd}^{[m-1]} + (1 - \beta_b) \cdot rms^{[m]} \tag{7.1-10}$$

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)-th frames respectively, and β_b is an adaptive factor controlling the update speed of the $rms_{bgd}^{[m]}$:

$$\beta_b = \begin{cases} 0.95 & rms^{[m]} > 1.4125 \cdot rms_{bgd}^{[m-1]} \text{ or } rms_{bgd}^{[m-1]} > 1.4125 \cdot rms^{[m]} \\ 0.99 & o.w. \end{cases} \tag{7.1-11}$$

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:

$$rms_{fgd}^{[m]} = \beta_f \cdot rms_{fgd}^{[m-1]} + (1-\beta_f) \cdot rms^{[m]} \tag{7.1-12}$$

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

$$\beta_f = \begin{cases} 0.99 & rms^{[m]} > 2 \cdot rms_{fgd}^{[m-1]} \\ 0.999 & 1.6 \cdot rms_{fgd}^{[m-1]} \le rms^{[m]} \le 2 \cdot rms_{fgd}^{[m-1]} \\ 0.99999 & rms^{[m]} < 1.6 \cdot rms_{fgd}^{[m-1]} \end{cases} \tag{7.1-13}$$

Finally, the long term SNR, *lsnr*, is obtained by:

$$lsnr = 0.85 \cdot [20 \cdot \log(rms_{fgd}/32767) - 20 \cdot \log(rms_{bgd}/32767)] \tag{7.1-14}$$

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 input frame is considered as background frame and the $flux_{bgd}$ is updated by:

$$flux_{bgd} = \chi \cdot flux_{bgd} + (1-\chi) \cdot MSSNR \tag{7.1-15}$$

where χ (the factor controlling the update speed of the $flux_{bgd}$) is determined by:

$$\chi = \begin{cases} 0.955 & MSSNR > flux_{bgd} \text{ ; during initialization period} \\ 0.995 & MSSNR \le flux_{bgd} \text{ ; during initialization period} \\ 0.997 & MSSNR > flux_{bgd} \text{ ; after initialization period} \\ 0.9997 & MSSNR \le flux_{bgd} \text{ ; after initialization period} \end{cases} \tag{7.1-16}$$

In equation 7.1-16, 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:

$$\begin{aligned} &\textit{if } \ MSSNR > thr_{\text{vad}} && ivad = 1 \\ &\textit{if } \ MSSNR - \lambda \cdot DZCR > thr_{\text{vad}} && ivad = 1 \\ &\textit{else} && ivad = 0 \end{aligned} \tag{7.1-17}$$

where λ is a factor determined by the long term SNR and thr_{vad} is a VAD decision threshold that is determined jointly by the long term SNR, background fluctuation and the operating point. The initial VAD decision, *ivad*, is either "1" (active) or "0" (inactive).

The determination of λ and thr_{vad} is described below. First, the long term SNR, *lsnr*, is classified into four categories: very high SNR, high SNR, medium SNR and low SNR, where each corresponds to the following values of λ:

$$\lambda = \begin{cases} 0 & lsnr > 35 \\ 2.7778 & 35 \geq lsnr > 25 \\ 2.2222 & 25 \geq lsnr > 15 \\ 1.667 & lsnr \leq 15 \end{cases} \tag{7.1-18}$$

Further, the background fluctuation is classified into three categories: high, medium and low fluctuation. Threshold thr_{vad} is chosen from a NB or WB table indexed by the long term SNR category, the background fluctuation category and the operating point. Basically, thr_{vad} is biased for different operating points for the same condition (i.e., the same long term SNR and the same background fluctuation), but for the very high SNR condition, the bias is ignored.

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 (see Figure 3).

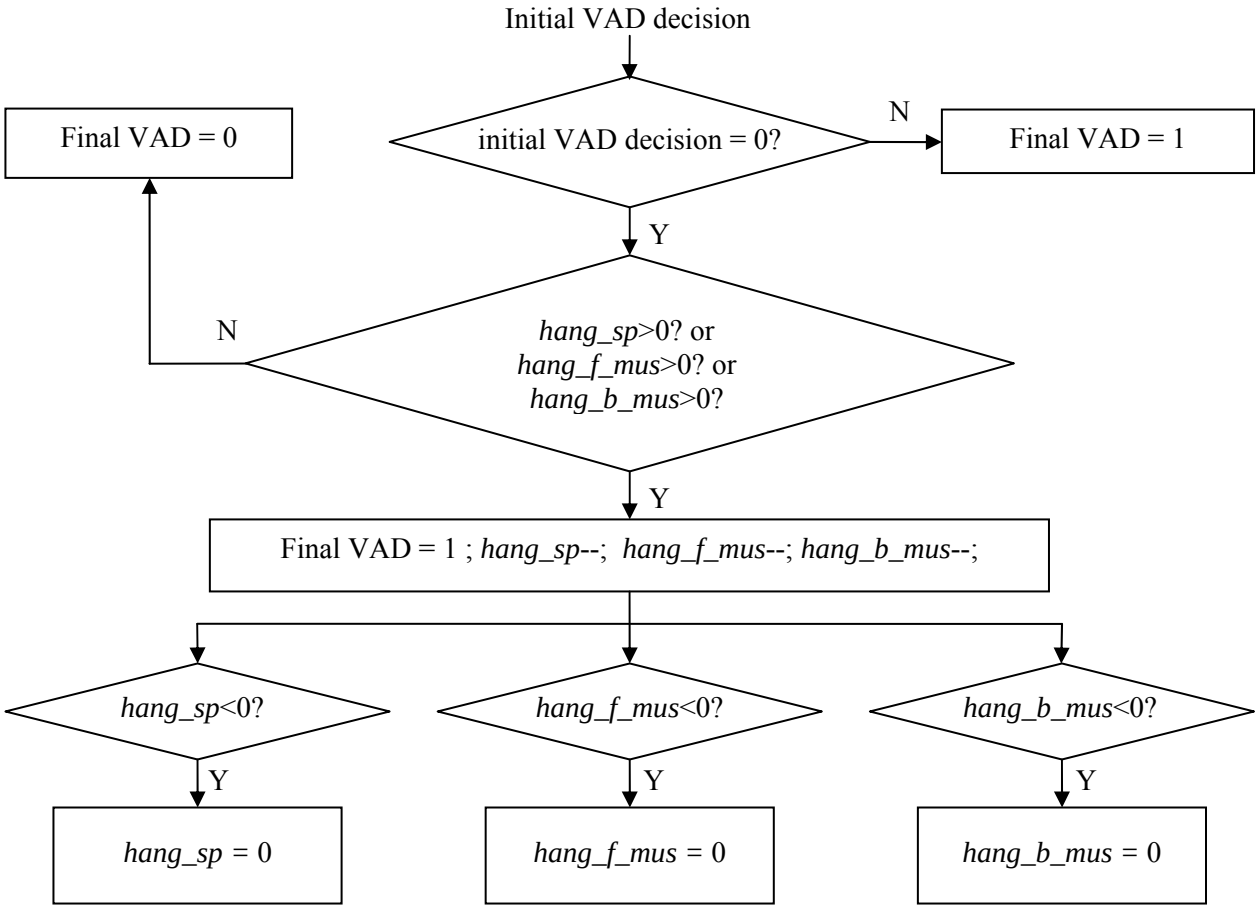


Figure 3 – Flow chart of the hangover procedure

In Figure 3, *hang_sp*, *hang_f_mus* and *hang_b_mus* are three independent hangover counters respectively 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 of 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 clauses 7.1.12 and 7.1.14, respectively.

7.1.8 Calculation of tone stability and DTMF detection

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 neighbouring 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:

$$D_{p2s}(i) = \left| 4 \cdot peak_{loc}(i) - S_{idx\_peak_{loc}(i-1)} - S_{idx\_peak_{loc}(i-2)} - S_{idx\_peak_{loc}(i+1)} - S_{idx\_peak_{loc}(i+2)} \right| \quad (7.1\text{-}19)$$

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, and $idx\_peak_{loc}(i)$ 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 largest three D_{p2s} values are found and denoted as $idx\_peak_{\max}(0)$, $idx\_peak_{\max}(1)$ and $idx\_peak_{\max}(2)$, where:

$$D_{p2s}(idx\_peak_{\max}(0)) \geq D_{p2s}(idx\_peak_{\max}(1)) \geq D_{p2s}(idx\_peak_{\max}(2)) \qquad (7.1\text{-}20)$$

For each of the three stored spectral positions, their distance to the position of the spectral peak that has the maximum D_{p2s} in the previous frame, $idx\_peak_{\max\_old}$, is obtained by:

$$\begin{aligned} D_{p2old}(0) &= idx\_peak_{\max}(0) - idx\_peak_{\max\_old} \\ D_{p2old}(1) &= idx\_peak_{\max}(1) - idx\_peak_{\max\_old} \\ D_{p2old}(2) &= idx\_peak_{\max}(2) - idx\_peak_{\max\_old} \end{aligned} \qquad (7.1\text{-}21)$$

If any of $D_{p2old}(0)$, $D_{p2old}(1)$, or $D_{p2old}(2)$ is less than 2, the counter *tone_sta_cnt* for tone stability is incremented by 1, and it is used for the background update decision (see clause 7.1.14). The value of $idx\_peak_{max}(0)$ is used to update $idx\_peak_{max\_old}$ for next frame's analysis. 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 third to the 64th spectral bins, except for 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, a 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 that frame. The position of the *m*-th 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 value of *peak_flux_cnt* is used in the background update decision (see clause 7.1.14).

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 peaks, 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 four 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 four FFT bins with frequency above the bin of the *j*-th local spectral peak, respectively. This can be mathematically expressed as:

$$\begin{aligned} E_{vl}(j) = MIN[&S(idx\_peak_{loc}(j)-1), S(idx\_peak_{loc}(j)-2), \\ &S(idx\_peak_{loc}(j)-3), S(idx\_peak_{loc}(j)-4)] \end{aligned} \qquad (7.1\text{-}22)$$

$$\begin{aligned} E_{vh}(j) = MIN[&S(idx\_peak_{loc}(j)+1), S(idx\_peak_{loc}(j)+2), \\ &S(idx\_peak_{loc}(j)+3), S(idx\_peak_{loc}(j)+4)] \end{aligned} \qquad (7.1\text{-}23)$$

where $S(k)$ is the k-th power spectral bin of the input frame, $idx\_peak_{loc}(j)$ is the location of the j-th local spectral peak in the spectrum. The normalized peak-to-valley distance is calculated by:

$$D_{p2v}(j) = \frac{62 \cdot (2 \cdot peak_{loc}(j) - E_{vl}(j) - E_{vh}(j))}{\sum_{i=2}^{63} S(i)} \quad (7.1\text{-}24)$$

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 frequency stability sta_{fq} is used in the background update decision. The spectrum of the input frame is whitened by a moving average of the spectra over past frames and sta_{fq} is obtained as the power deviation of the 16 whitened sub-bands:

$$sta_{fq} = \frac{1}{16} \sum_{i=0}^{15} [E_{band}(i) / \overline{E_{band}(i)} - E_{band\_mean}]^2 \quad (7.1\text{-}25)$$

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 energy of the i-th sub-band over past frames, E_{band_mean} is the power mean of the 16 whitened sub-bands calculated by:

$$E_{band\_mean} = \frac{1}{16} \sum_{i=0}^{15} E_{band}(i) / \overline{E_{band}(i)} \quad (7.1\text{-}26)$$

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

$$\overline{E_{band}(i)}^{[m]} = 0.9 \cdot \overline{E_{band}(i)}^{[m-1]} + (1 - 0.9) \cdot E_{band}^{[m]}(i) \quad (7.1\text{-}27)$$

7.1.12 Background music detection

Detection for 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, SP_{sum}, is above a music detection threshold. A hangover of 1000 frames is set once music background is detected, but it will be fast exited if SP_{sum} is sufficiently low. Moreover, the music detection threshold is further "biased" depending on whether the background music hangover exits. The procedure of the background music detection is depicted in Figure 4.

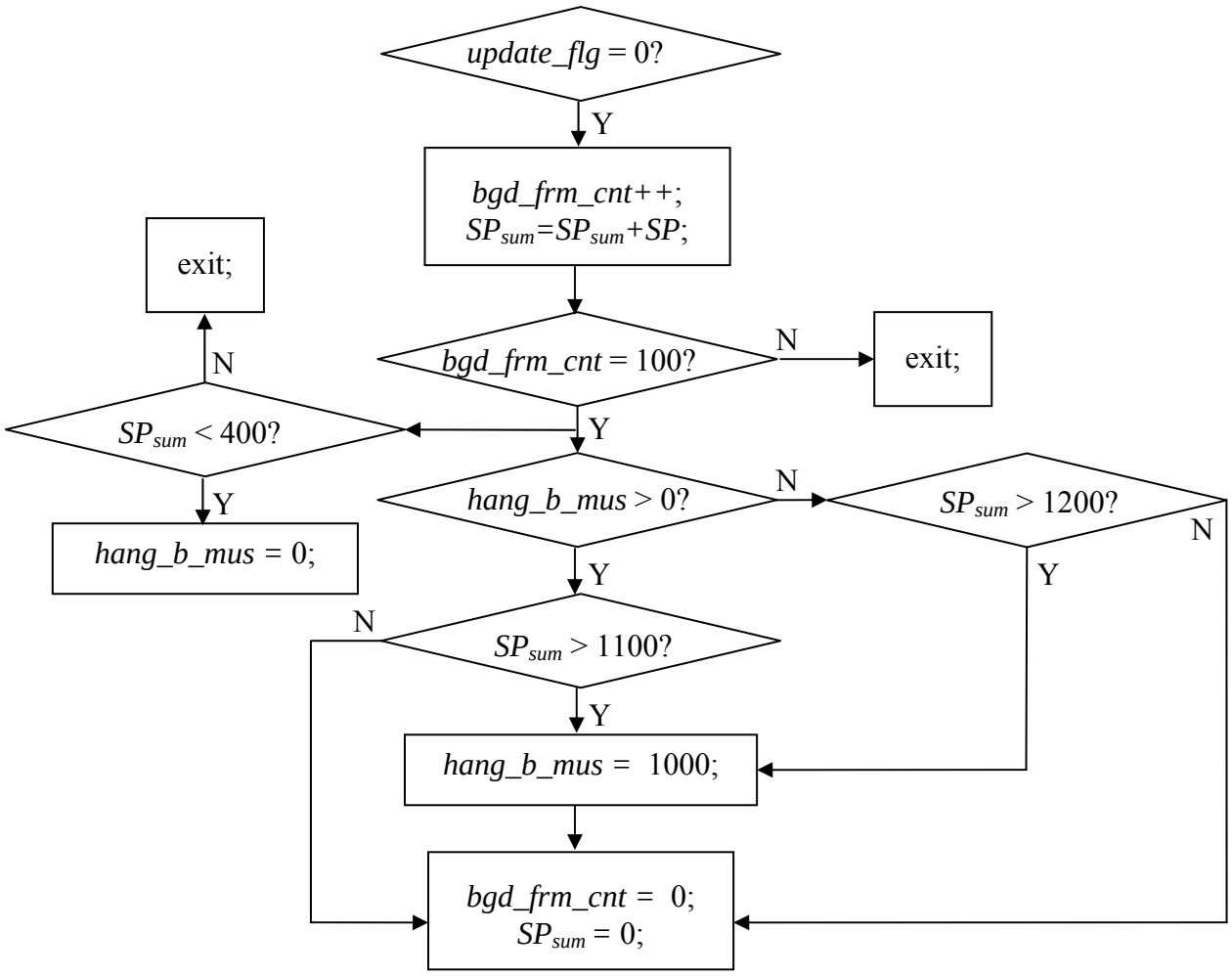


Figure 4 – Flow chart of the background music detection procedure

In Figure 4, 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:

$$FP_{dBov} = 20 \cdot \log_{10}(rms/32767) \tag{7.1-28}$$

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

$$thr_{bgd} = \begin{cases} 40(WB\ input) \text{ or } 34.375(NB\ input) & lsnr > 18 \text{ and } hang\_b\_mus > 0 \\ 15 & o.w. \end{cases} \tag{7.1-29}$$

The decision whether to update the background signal estimate consists of the following steps (see also the flowchart in Figure 5).

Step 1) If the *MSSNR* of the input frame is not greater than thr_{bgd}, a counter for counting consecutive background frames, *update_cnt*, is incremented by 1. If *update_cnt* is greater than *M* frames and no DTMF signal is detected and the foreground music hangover *hang_f_mus* is not greater than 0, the background update flag *update_flg*, and the reset flag, *reset_flg*, are set. The value of *M* is determined from *lsnr*:

$$M = \begin{cases} 2 & lsnr > 15 \\ 4 & lsnr \le 15 \end{cases} \tag{7.1-30}$$

If *update_flg* is set, the input frame will be used to update the background estimate. The purpose of *reset_flg* is described later in this clause. The counting for consecutive background frames tolerates occasional background frames whose *MSSNR*s are greater than thr_{bgd}. If $update\_cnt \ge M$ but the *MSSNR* of the input frame is higher than thr_{bgd}, *update_cnt* is set to be *M*. The *update_cnt* is reset to 0 if more than nine consecutive frames whose *MSSNR*s are greater than 50 are detected.

Step 2) While the *update_flg* is not set at Step 1, background noise is sought within time windows of 30 frames in sequence and used to update the background estimate if there is one. The possibility of containing background frames is estimated for each time window by analysing the tonality and the stability of the frames within the time window. If all or most of the frames within the time window are recognized as background frames and the few non-background frames are not at the end of the time window, the latest frame within the time window (i.e. the input frame) is used to update the background estimate by setting the background update flag *update_flg*. Otherwise, if the time window is believed (but with less confidence) to contain background frames, the minimum energies for each sub-band over frames within the time window are used to update the sub-band energies of the background estimate. Otherwise, the time window is considered not to have background frames and analysis for another time window starts. If the input frame (the latest frame within the current time window) is used for background estimate update, successive frames which have high frequency stability are also used for the background estimate update until a frame with low frequency stability is encountered. Details of step 2 are described below.

For each input frame, if the *update_flg* is not set at Step 1 and if no DTMF signal is detected and the foreground music hangover *hang_f_mus* is not greater than 0, the following operations a) to d) are performed, otherwise the *reset_flg* is set:

a) a counter *con_frm_cnt* for counting the number of consecutive frames is incremented by 1;

b) the spectral peakiness *SP* of the input frame is compared to a threshold thr_{SP_low} which is 15 for NB input and 12.7 for WB input. If $SP < thr_{SP\_low}$, counter *low_SP_cnt* (that counts the number of frames with low spectral peakiness) is incremented by 1;

c) the frequency stability sta_{fq} is compared to a 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;

d) the sub-band energies $E_{band}(i)$, i=0,1,...15, of the input frame are used to update the minimum sub-band energy buffer $E_{band\_buf\_min}(i)$, i=0,1,...15, for those bands where $E_{band}(i) < E_{band\_buf\_min}(i)$.

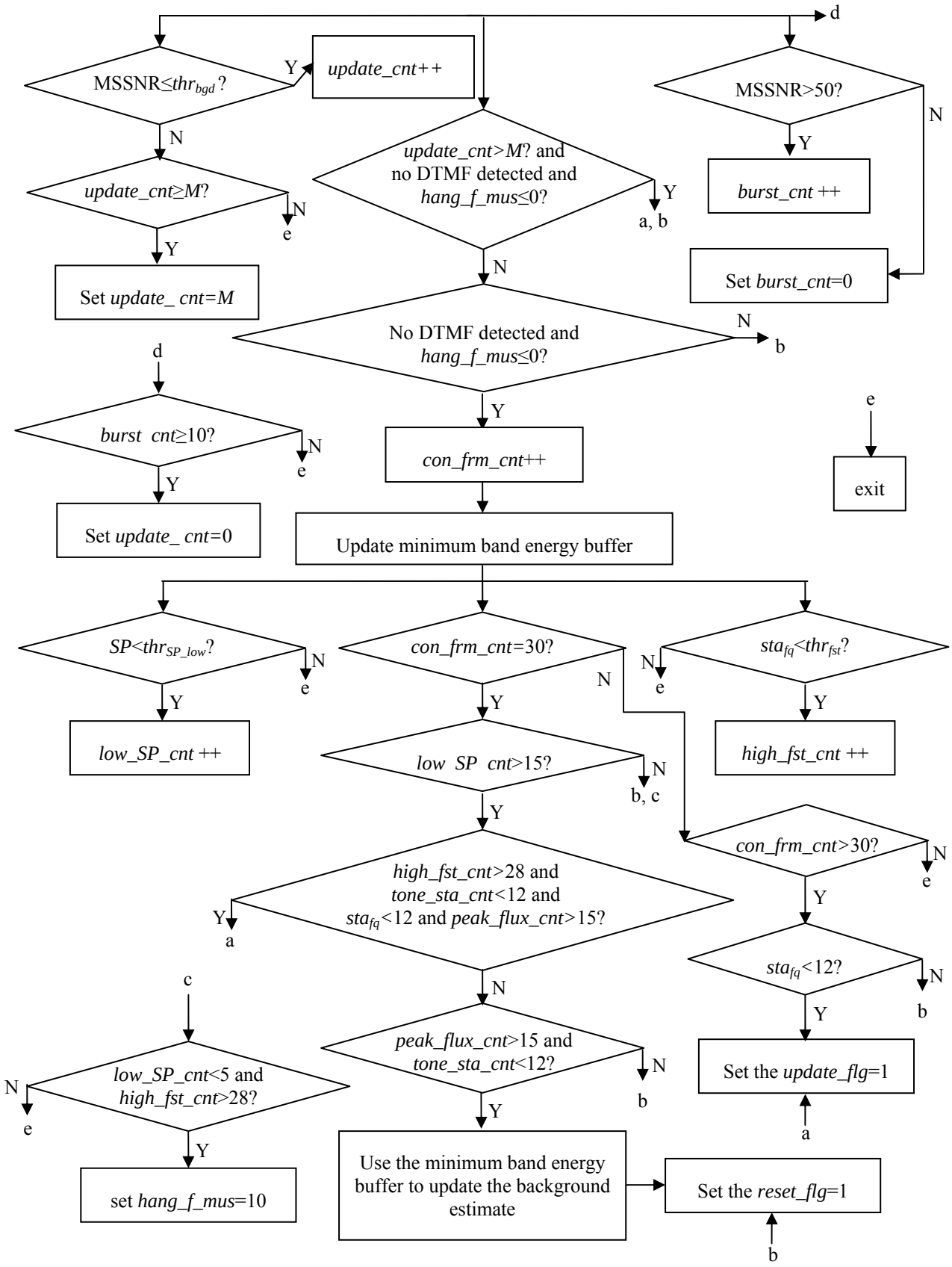


Figure 5 – Flowchart of the background estimate update procedure

Once *con_frm_cnt* reaches 30, the *low_SP_cnt* is compared to a threshold of 15. If *low_SP_cnt* is not greater than 15, the *reset_flg* is set. Otherwise, the following condition is examined:

$$\begin{aligned} &if\ (high\_fst\_cnt > 28\ \&\&\ tone\_sta\_cnt < 12 \\ &\&\&\ sta_{fq} < 12\ \&\&\ peak\_flux\_cnt > 15) \end{aligned} \tag{7.1-31}$$

where *tone_sta_cnt* and *peak_flux_cnt* are obtained in clauses 7.1.8 and 7.1.9, respectively. If the condition above is satisfied, the background update flag, *update_flg*, is set to 1. Otherwise, if the following condition is satisfied:

$$if\ (peak\_flux\_cnt > 15\ \&\&\ tone\_sta\_cnt < 12) \tag{7.1-32}$$

the sub-band energies $E_{band\_buf\_min}(i)$, i=0,1,...15, stored in the minimum sub-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 to 1.

If *reset_flg* is set to 1, the parameters *con_frm_cnt*, *low_SP_cnt*, *high_fst_cnt*, *tone_sta_cnt* and *peak_flux_cnt* are all reset to 0.

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

$$\overline{ZCR_n}^{[m]} = 0.95 \cdot \overline{ZCR_n}^{[m-1]} + (1-0.95) \cdot ZCR \tag{7.1-33}$$

$$\overline{E_{band\_n}(i)}^{[m]} = 0.9 \cdot \overline{E_{band\_n}(i)}^{[m-1]} + (1-0.9) \cdot E_{band}(i) \tag{7.1-34}$$

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,…,15. $\overline{ZCR_n}$ is initialized at the first four input frames as 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 as the energy of the *i*-th sub-band of the input frame, which is further limited by a minimum value. If a 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; the calculation 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 narrow-band (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:

$$flux^{[m]} = \frac{\sum_{n=1}^{3}\sum_{i=1}^{47}\left|S^{[m]}(i) - S^{[m-n]}(i)\right|}{\sum_{n=1}^{3}\sum_{i=1}^{47}\left(S^{[m]}(i) + S^{[m-n]}(i)\right)} \tag{7.2-1}$$

The variance of the flux is calculated for the last 20 frames that 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 a FIFO buffer, which is updated only for active frames. The *MSSNR* of the *m*-th frame is denoted in the sequel by $MSSNR^{[m]}$. It should be noted 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:

$$mov\_flux^{[m]} = 0.99 \cdot mov\_flux^{[m-1]} + 0.01 \cdot flux^{[m]} \tag{7.2-2}$$

The variance of the *flux* for the *m*-th frame, $\text{var}\_flux^{[m]}$, is calculated as:

$$var\_flux^{[m]} = \sum_{k=0}^{19}\left(flux^{[m-k]} - mov\_flux^{[m]}\right)^2 \tag{7.2-3}$$

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, denoted as $S^{[m]}(i),\ \ i=1,...,K,\ K \le 5$, are used (if less peaks are found, *K* might be less than 5).

The first peakiness measure, P_1, is calculated by:

$$P_1 = \frac{\sqrt{\frac{1}{K}\sum_{i=1}^{K} S^2(i)}}{\frac{1}{K}\sum_{i=1}^{K}\left|S(i)\right|} - 1 \tag{7.2-4}$$

The second peakiness measure, P_2, is calculated by:

$$P_2 = \frac{\max_{i=1}^{K} |S(i)|}{\frac{1}{K}\sum_{i=1}^{K} |S(i)|} - 1 \tag{7.2-5}$$

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

$$avrg\_P_1^{[m]} = 0.995 \cdot avrg\_P_1^{[m-1]} + 0.005 \cdot P_1 \tag{7.2-6}$$

and

$$avrg\_P_2^{[m]} = 0.995 \cdot avrg\_P_2^{[m-1]} + 0.005 \cdot P_2 . \tag{7.2-7}$$

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 decreased 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 that depends on the operating point:

$$T_{MSSNR}^{[m]} = C_{op} \cdot max_{MSSNR}^{[m]} \tag{7.2-8}$$

where C_{op} is 0.5 for the bandwidth-saving and the balanced operating points, and 0.45 for the quality-preferred operating point.

The information about $MSSNR^{[m]}$ 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, when 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 $high_{bin}^{[m]}$, the histogram difference measure $diff_{hist}^{[m]}$, which is between –1 and 1, is calculated according to:

$$diff_{hist}^{[m]} = \frac{high_{bin}^{[m]} - low_{bin}^{[m]}}{M} = 1 - \frac{2 \cdot high_{bin}^{[m]}}{M} \tag{7.2-9}$$

The final histogram difference measure is generated as a moving 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:

$$diff_{hist}^{avg} = 0.9 \cdot diff_{hist}^{avg} + 0.1 \cdot \left(diff_{hist}^{[m]} + \Delta_{op} \right) \tag{7.2-10}$$

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 $var\_flux^{[m]}$ is given by the linear equation:

$$T_{var\_flux}^{[m]} = A \cdot diff_{hist}^{final} + B \tag{7.2-11}$$

where *A* and *B* are calculated by:

$$A = \frac{T_{op}^{up} - T_{op}^{down}}{2 \cdot X_T} \tag{7.2-12}$$

$$B = \frac{T_{op}^{up} + T_{op}^{down}}{2} \tag{7.2-13}$$

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 (which 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 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, when it is fixed. For the first 75 active frames, the ratio is "biased" toward speech. This is done to help avoiding some speech-to-music misclassification at the beginning, where the decision is based on the yet-unstable adaptive threshold.

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 of time. As 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 in 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 last for 150 frames. Once a low *MSSNR* segment is detected, the SMD algorithm is being "reset". The following operation is to set $diff_{hist}^{final}$ to Δ_{op} for a period of 400 frames, thus neutralizing the previously calculated preference for music. The reset also sets *J* to 0, causing the previously accumulated values of *var_flux* in the buffer not to be used, as well as the application of a "bias" toward speech for the next 75 frames.

The values of the two moving averages of the spectral-peak 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 5 lists the files that comprise the ANSI-C source code attached to this Recommendation and which is an integral part of this Recommendation. For this edition of the Recommendation, the software is labelled as "Release 1.0a". Proper compilation of the code can be checked with the test vectors also provided as an electronic attachment to this Recommendation.

Table 5 – Organization of the GSAD source files

File/directory 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
rom_fx.c/h	ROM tables
constdef_fx.h	Macro defined for constants
wmops.c/h	Complexity and RAM measuring tool
stl2005_basop	Directory containing the source code for the ITU-T fixed-point basic operators adapted from [b-ITU-T G.191 (2005)] (basop32_fx.c, basop32_fx.h, control_fx.c, control_fx.h, count_fx.c, count_fx.h, enh1632_fx.c, enh1632_fx.h, enh40_fx.c, enh40_fx.h, move_fx.h, patch_fx.h, stl_fx.h, typedef_fx.h, and typedefs_fx.h)

Annex A

Generic voice activity detector

(This annex forms an integral part of this Recommendation)

A.1 Scope

The generic voice activity detector (GVAD) is an independent front-end processing module which can be applied prior to signal processing applications such as, but not limited to, speech or audio codecs. GVAD operates on narrow-band or wideband audio input at 10-ms frame length (without lookahead). Its function is to indicate the input frame activity. In order to apply GVAD in specific cases, an adaptation layer may be required.

The GVAD is bit-exact with the VAD module of the main body of this Recommendation, with the speech/music discriminator, the silence detector of the main Recommendation removed to save implementation cost for applications requiring VAD only.

This annex is organized as follows. References, definitions, abbreviations/acronyms and conventions are defined in clauses A.2, A.3, A.4 and A.5, respectively. Clause A.6 gives a general description of the GVAD 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 GVAD algorithm is described in clause A.7. Finally in clause A.8, use of the simulation software is described. References to the main Recommendation are made for those parts of the description that are the same as the main Recommendation.

A.2 References

See clause 2.

A.3 Definitions

A.3.1 Terms defined in this annex

This annex defines the following terms:

A.3.1.1 balanced operating point: Generic voice activity detector (GVAD) operating point that provides a middle point performance between the other two operating points. See clause A.6.4.

A.3.1.2 bandwidth-saving operating point: Generic voice activity detector (GVAD) operating point privileging accurate detection for background signals. See clause A.6.4.

A.3.1.3 quality-preferred operating point: Generic voice activity detector (GVAD) operating point privileging accurate detection for foreground signals. See clause A.6.4.

A.4 Abbreviations and acronyms

See clause 4.

A.5 Conventions

This annex does not define any conventions.

A.6 General description of the GVAD algorithm

A.6.1 Input sampling rate

The GVAD can accept both narrow-band (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 GVAD by an external signal.

A.6.2 Operating frame size

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

A.6.3 Delay

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

A.6.4 Configurations

GVAD includes two control parameters. GVAD operates on narrow-band (NB) and wideband (WB) audio input signals, where a control signal indicates the bandwidth of the input signal.

In addition, a second control signal selects the desired operating point from three possibilities. Setting this second control signal to "0" selects the balanced operating point, setting it to "1" selects the quality-preferred operating point and setting it to "2" selects the bandwidth-saving operating point. These three operating points provide the ability to select a preferred balance 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 ITU-T G.720.1 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.

GVAD outputs a binary value where "1" indicates an active frame and "0" indicates an inactive frame.

A.6.5 Complexity and memory cost

Table A.1 shows the GVAD complexity in WMOPS for NB and WB signal sampling frequencies respectively. The RAM used for GVAD (static and dynamic) is 2436 bytes and the table ROM is 1674 bytes.

Table A.1 – Complexity of the GVAD

Modes	Complexity (WMOPS)
GVAD_WB	2.397
GVAD_NB	1.475

A.7 Detailed description of the GVAD algorithm

The detailed description of the GVAD algorithm follows exactly that of the VAD module (see clause 7.1).

A.8 Use of the simulation software

The bit-exact, fixed point simulation software of GVAD is defined by compilation flag "LC" in the ANSI-C source code of the main body of this Recommendation. By defining the compilation flag "LC" in the ITU-T G.720.1 source code, the compiler will generate the executable for GVAD. The command line instruction for GVAD thus becomes:

gsad[2] rate_option op_option speech_file decision_file

Required arguments:

rate_option: 8k or 16k

==========

8k: 8 kHz sampling rate
16k: 16 kHz sampling rate

==========

op_option: 0, 1 or 2

==========

0: Balanced operating point
1: Quality-preferred operating point
2: Bandwidth-saving operating point

==========

speech_file: Input speech file name

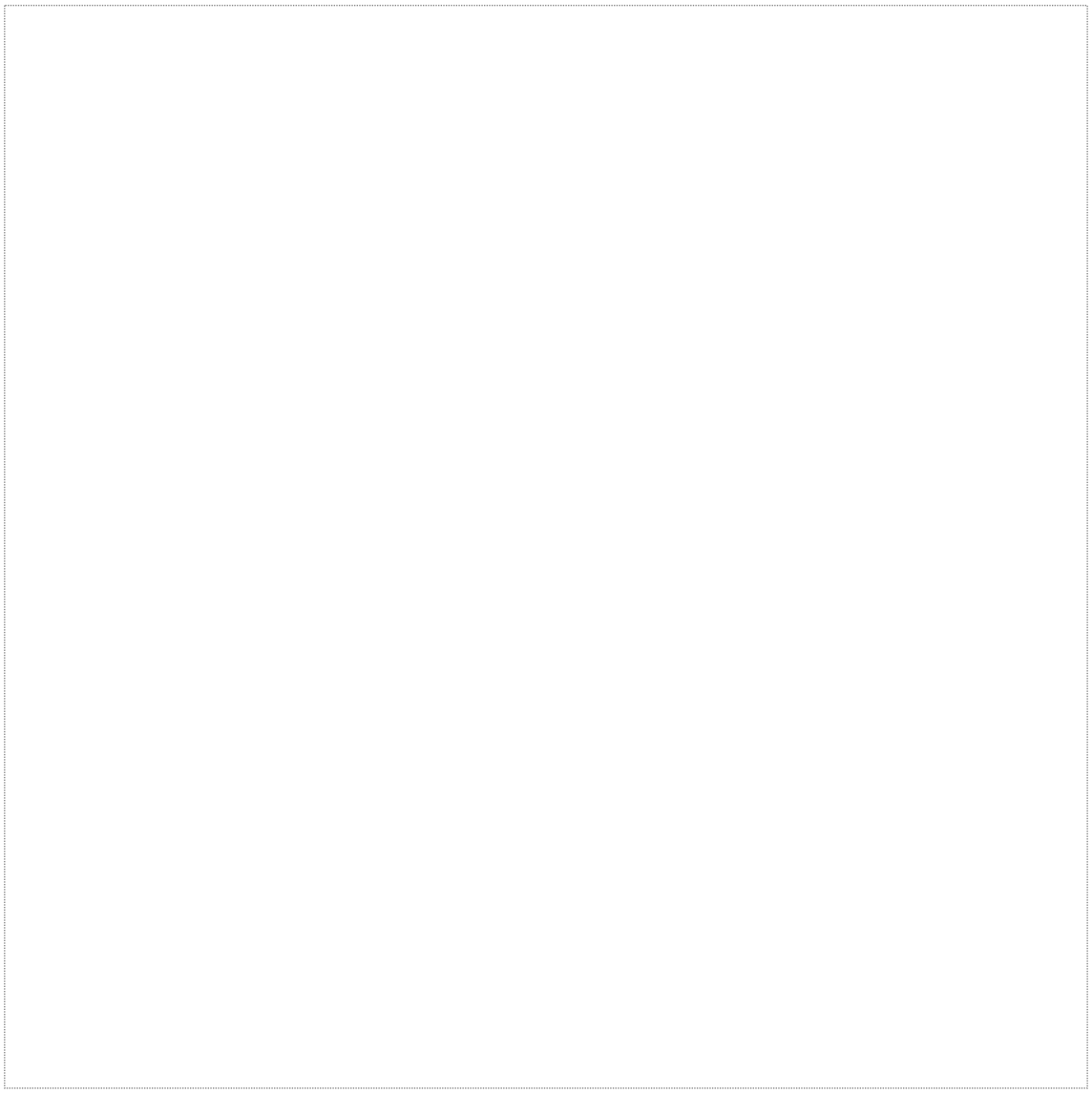
decision_file: Output signal class indicator file name

Proper compilation of the code can be checked with the test vectors provided for ITU-T G.720.1 bit-exactness check.

[2] Note that, because the C source code of the main body of this Recommendation is used to generate the GVAD executable, the default name of the generated GVAD executable is kept the same to that of the main body of ITU-T G.720.1 ("gsad").

Bibliography

[b-ITU-T G.191 (2005)] Recommendation ITU-T G.191 (2005), *Software tools for speech and audio coding standardization*.



SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	**Transmission systems and media, digital systems and networks**
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Terminals and subjective and objective assessment methods
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2011