



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 (GSAD)

CAUTION !

PREPUBLISHED RECOMMENDATION

This prepublication is an unedited version of a recently approved Recommendation. It will be replaced by the published version after editing. Therefore, there will be differences between this prepublication and the published version.

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/had not] 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 2010

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

Recommendation ITU-T G.720.1 (ex G.GSAD)

Generic sound activity detector (GSAD)

CONTENTS

1	Scope.....	3
2	References.....	4
3	Definitions.....	4
3.1	Terms defined elsewhere:.....	4
3.2	Terms defined in this Recommendation.....	4
4	Abbreviations and acronyms.....	4
5	Conventions	7
6	General description of GSAD algorithm	8
6.1	Input sampling rate	8
6.2	Operating frame size.....	8
6.3	Delay.....	8
6.4	Configurations	8
6.5	Complexity and memory cost.....	9
7	Detailed description of the GSAD algorithm.....	9
7.1	Detailed Description of the VAD Module.....	10
7.1.1	Pre-processing and power spectrum calculation	10
7.1.2	Differential Zero Crossing Rate (DZCR) calculation	10
7.1.3	Modified Segmental SNR (MSSNR) calculation	11
7.1.4	Long term SNR estimation.....	12
7.1.5	Background fluctuation estimation	13
7.1.6	Initial VAD decision	13
7.1.7	VAD hangover	13
7.1.8	Calculation of tone stability	14
7.1.9	Calculation of spectral peak fluctuation.....	15
7.1.10	Calculation of spectral peakiness	15
7.1.11	Calculation of frequency stability	16
7.1.12	Background music detection	16
7.1.13	Silence detection	17
7.1.14	Background estimate update	17
7.2	Detailed description of the Speech/Music Discrimination module.....	20
7.2.1	Calculation of the flux and the variance of the flux.....	20
7.2.2	Calculation of two spectral-peaks peakiness measures.....	21

7.2.3	Speech/Music Discrimination decision.....	22
8	Organization of the reference C code	24

Recommendation ITU-T G.720.1 (ex G.GSAD)

Generic sound activity detector (GSAD)

Summary

This Recommendation 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 narrowband or wideband audio input using 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.

An external control signal indicates to the GSAD algorithm which one of 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-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 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 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. 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://itu.int/net/ITU-T/sigdb/speaudio/Gseries.htm#G.720.1>.

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.

3.2.3 Foreground signal: The dominating information signal (speech or music) in the speech channel.

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

3.2.5 Quality-preferred operating point: 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 Bandwidth-saving operating point: 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.

3.2.7 Balanced operating point: 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.

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
DZCR	Differential zero crossing rate
FFT	Fast Fourier transform
GSAD	Generic sound activity detection/detector
GSAD_NB	Generic sound activity detecting mode for narrowband signals

Acronym	Description
GSAD_WB	Generic sound activity detecting mode for wideband signals
MSSNR	Modified segmental SNR
NB	Narrowband
RMS	Root mean square
SDF	Spectral density function
SiD	Silence detector
SMD	Speech/music discriminator
VAD	Voice activity detection/detector
VAD_NB	Voice activity detecting mode for narrowband signals
VAD_WB	Voice activity detecting mode for wideband signals
WB	Wideband
WMOPS	Weighted million operations per second
ZCR	Zero crossing rate

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

Name	Description
$hang_b_mus$	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
$peak_flux_cnt$	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