|  |  |
| --- | --- |
| 3GPP TS 26.194 V18.0.0 (2024-03) | |
| Technical Specification | |
| 3rd Generation Partnership Project;  Technical Specification Group Services and System Aspects;  Speech codec speech processing functions;  Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Voice Activity Detector (VAD)  (Release 18) | |
|  | |
|  |  |
|  | |
| The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP. The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented. This Specification is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification. Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices. | |

|  |
| --- |
|  |
| ***3GPP***  Postal address  3GPP support office address  650 Route des Lucioles - Sophia Antipolis  Valbonne - FRANCE  Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16  Internet  http://www.3gpp.org |
| ***Copyright Notification***  No part may be reproduced except as authorized by written permission. The copyright and the foregoing restriction extend to reproduction in all media.  © 2024, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).  All rights reserved.  UMTS™ is a Trade Mark of ETSI registered for the benefit of its members  3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners  GSM® and the GSM logo are registered and owned by the GSM Association |

Contents

Foreword 4

1 Scope 5

2 Normative References 5

3 Technical Description 5

3.1 Definitions, symbols and abbreviations 5

3.1.1 Definitions 5

3.1.2 Symbols 5

3.1.2.1 Variables 5

3.1.2.2 Constants 6

3.1.2.3 Functions 7

3.1.3 Abbreviations 8

3.2 General 8

3.3 Functional description 8

3.3.1 Filter bank and computation of sub-band levels 9

3.3.2 Tone detection 10

3.3.3 VAD decision 11

3.3.3.1 Hangover addition 12

3.3.3.2 Background noise estimation 13

3.3.3.3 Speech level estimation 14

4 Computational details 14

Annex A (informative) : Change history 15

# Foreword

This Technical Specification has been produced by the 3GPP.

This document specifies the Voice Activity Detector (VAD) to be used in the Discontinuous Transmission (DTX) as described in [3].

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of this TS, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 Indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the specification;

# 1 Scope

This document specifies the Voice Activity Detector (VAD) to be used in the Discontinuous Transmission (DTX) as described in [3].

The requirements are mandatory on any VAD to be used either in User Equipment (UE) or Base Station Systems (BSS)s that utilize the AMR wideband speech codec.

# 2 Normative References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TS 26.173: "ANSI-C code for the Adaptive Multi-Rate Wideband speech codec" .

[2] 3GPP TS 26.190: "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions" .

[3] 3GPP TS 26.193: "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Source controlled rate operation".

[4] ITU, The International Telecommunications Union, Blue Book, Vol. III, Telephone Transmission Quality, IXth Plenary Assembly, Melbourne, 14-25 November, 1988, Recommendation G.711, Pulse code modulation (PCM) of voice frequencies.

[5] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

# 3 Technical Description

## 3.1 Definitions, symbols and abbreviations

### 3.1.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [5] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [5].

**frame:** Time interval of 20 ms corresponding to the time segmentation of the speech   
transcoder.

### 3.1.2 Symbols

For the purposes of this TS, the following symbols apply.

#### 3.1.2.1 Variables

**bckr\_est[n]** background noise estimate at the frequency band "n"

**burst\_count** counts length of a speech burst, used by VAD hangover addition

**hang\_count** hangover counter, used by VAD hangover addition

**level[n]** signal level at the frequency band "n"

**new\_speech** pointer of the speech encoder, points a buffer containing last received samples of a speech frame [2]

**noise\_level** estimated noise level

**pow\_sum** input power

**s(i)** samples of the input frame

**snr\_sum** measure between input frame and noise estimate

**speech\_level** estimated speech level

**stat\_count** stationary counter

**stat\_rat** measure indicating stationary of the input frame

**tone\_flag** flag indicating the presence of a tone

**vad\_thr** VAD threshold

**VAD\_flag** Boolean VAD flag

**vadreg** intermediate VAD decision

#### 3.1.2.2 Constants

**ALPHA\_UP1** constant for updating noise estimate (see subclause 3.3.5.2)

**ALPHA\_DOWN1** constant for updating noise estimate (see subclause 3.3.5.2)

**ALPHA\_UP2** constant for updating noise estimate (see subclause 3.3.5.2)

**ALPHA\_DOWN2** constant for updating noise estimate (see subclause 3.3.5.2)

**ALPHA3** constant for updating noise estimate (see subclause 3.3.5.2)

**ALPHA4** constant for updating average signal level (see subclause 3.3.5.2)

**ALPHA5** constant for updating average signal level (see subclause 3.3.5.2)

**BURST\_HIGH** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**BURST\_P1** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**BURST\_SLOPE** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**COEFF3** coefficient for the filter bank (see subclause 3.3.1)

**COEFF5\_1** coefficient for the filter bank (see subclause 3.3.1)

**COEFF5\_2** coefficient for the filter bank (see subclause 3.3.1)

**HANG\_HIGH** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**HANG\_LOW** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**HANG\_P1** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**HANG\_SLOPE** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

**FRAME\_LEN** size of a speech frame, 256 samples (20 ms)

**MIN\_SPEECH\_LEVEL1** constant for speech estimation (see subclause 3.3.5.3)

**MIN\_SPEECH\_LEVEL2** constant for speech estimation (see subclause 3.3.5.3)

**MIN\_SPEECH\_SNR** constant for VAD threshold adaptation (see subclause 3.3.5)

**NO\_P1** constant for VAD threshold adaptation (see subclause 3.3.5)

**NO\_SLOPE** constant for VAD threshold adaptation (see subclause 3.3.5)

**NOISE\_MAX** maximum value for noise estimate (see subclause 3.3.5.2)

**NOISE\_MIN** minimum value for noise estimate (see subclause 3.3.5.2)

**POW\_TONE\_THR** threshold for tone detection (see subclause 3.3.5)

**SP\_ACTIVITY\_COUNT** constant for speech estimation (see subclause 3.3.5.3)

**SP\_ALPHA\_DOWN** constant for speech estimation (see subclause 3.3.5.3)

**SP\_ALPHA\_UP** constant for speech estimation (see subclause 3.3.5.3)

**SP\_CH\_MAX** constant for VAD threshold adaptation (see subclause 3.3.5)

**SP\_CH\_MIN** constant for VAD threshold adaptation (see subclause 3.3.5)

**SP\_EST\_COUNT** constant for speech estimation (see subclause 3.3.5.3)

**SP\_P1** constant for VAD threshold adaptation (see subclause 3.3.5)

**SP\_SLOPE** constant for VAD threshold adaptation (see subclause 3.3.5)

**STAT\_COUNT** threshold for stationary detection (see subclause 3.3.5.2)

**STAT\_THR** threshold for stationary detection (see subclause 3.3.5.2)

**STAT\_THR\_LEVEL** threshold for stationary detection (see subclause 3.3.5.2)

**THR\_HIGH** constant for VAD threshold adaptation (see subclause 3.3.5)

**TONE\_THR** threshold for tone detection (see subclause 3.3.3)

**VAD\_POW\_LOW** constant for controlling VAD hangover addition (see subclause 3.3.5.1)

#### 3.1.2.3 Functions

**+** Addition

**-** Subtraction

**\*** Multiplication

**/** Division

**| x |** absolute value of x

**AND** Boolean AND

**OR** Boolean OR

**** 

**MIN(x,y)** =

**MAX(x,y)** =

### 3.1.3 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [5] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [5].

ANSI American National Standards Institute

DTX Discontinuous Transmission

VAD Voice Activity Detector

CNGComfort Noise Generation

## 3.2 General

The function of the VAD algorithm is to indicate whether each 20 ms frame contains signals that should be transmitted, e.g. speech, music or information tones. The output of the VAD algorithm is a Boolean flag (VAD\_flag) indicating presence of such signals.

## 3.3 Functional description

The block diagram of the VAD algorithm is depicted in Figure 1. The VAD algorithm uses parameters of the speech encoder to compute the Boolean VAD flag (VAD\_flag). This input frame for VAD is sampled at the 6.4 kHz frequency and thus it contains 256 samples. Samples of the input frame (s(i)) are divided into sub-bands and level of the signal (level[n]) in each band is calculated. Input for the tone detection function are the normalized open-loop pitch gains which are calculated by open-loop pitch analysis of the speech encoder. The tone detection function computes a flag (tone\_flag) which indicates presence of a signalling tone, voiced speech, or other strongly periodic signal. Background noise level (bckr\_est[n]) is estimated in each band based on the VAD decision, signal stationarity and the tone-flag. Intermediate VAD decision is calculated by comparing input SNR (level[n]/bckr\_est[n]) to an adaptive threshold. The threshold is adapted based on noise and long term speech estimates. Finally, the VAD flag is calculated by adding hangover to the intermediate VAD decision.



Figure 1: Simplified block diagram of the VAD algorithm

### 3.3.1 Filter bank and computation of sub-band levels

The input signal is divided into frequency bands using a 12-band filter bank (Figure 2). Cut-off frequencies for the filter bank are shown in Table 1.

Table 1. Cut-off frequencies for the filter bank

|  |  |
| --- | --- |
| Band number | Frequencies |
| 1 | 0 – 200 Hz |
| 2 | 200 – 400 Hz |
| 3 | 400 – 600 Hz |
| 4 | 600 – 800 Hz |
| 5 | 800 – 1200 Hz |
| 6 | 1200 – 1600 Hz |
| 7 | 1600 – 2000 Hz |
| 8 | 2000 – 2400 Hz |
| 9 | 2400 - 3200 Hz |
| 10 | 3200 – 4000 Hz |
| 11 | 4000 – 4800 Hz |
| 12 | 4800 – 6400 Hz |

Input for the filter bank is a speech frame pointed by the new\_speech pointer of the speech encoder [1]. Input values for the filter bank are scaled down by one bit. This ensures safe scaling, i.e. saturation can not occur during calculation of the filter bank.



Figure 2: Filter bank

The filter bank consists of 5th and 3rd order filter blocks. Each filter block divides the input into high-pass and low-pass parts and decimates the sampling frequency by 2. The 5th order filter block is calculated as follows:

 (1a)

 (1b)

where

x(i) input signal for a filter block

 low-pass component

 high-pass component

The 3rd order filter block is calculated as follows:

 (2a)

 (2b)

The filters ,, andare first order direct form all-pass filters, whose transfer function is given by:

, (3)

where C is the filter coefficient.

Coefficients for the all-pass filters ,, and are COEFF5\_1, COEFF5\_2, and COEFF3, respectively.

Signal level is calculated at the output of the filter bank at each frequency band as follows:

, (4)

where:

n index for the frequency band

 sample i at the output of the filter bank at frequency band n

 =

 =

Negative indices of  refer to the previous frame.

### 3.3.2 Tone detection

The purpose of the tone detection function is to detect information tones, vowel sounds and other periodic signals. The tone detection uses normalized open-loop pitch gains (ol\_gain), which are received from the speech encoder. If the pitch gain is higher than the constant TONE\_THR, tone is detected and the tone flag is set:

if (ol\_gain > TONE\_THR)

tone\_flag = 1

The open-loop pitch search and correspondingly the tone flag is computed twice in each frame, except for mode 6.60 kbit/s, where it is computed only once.

### 3.3.3 VAD decision

The block diagram of the VAD decision algorithm is shown in figure 3.



Figure 3: Simplified block diagram of the VAD decision algorithm

Power of the input frame is calculated as follows:

, (5)

where samples s(i) of the input frame are pointed by the new\_speech pointer of the speech encoder. Variable pow\_sum is sum of the powers of the current and previous frames. If pow\_sum is lower than the constant POW\_TONE\_THR, tone-flag is set to zero.

The difference between the signal levels of the input frame and the background noise estimate is calculated as follows:

, (6)

where:

level[n] signal level at band n

bckr\_est[n] level of background noise estimate at band n

VAD decision is made by comparing the variable snr\_sum to a threshold. The threshold (vad\_thr) is adapted to get desired sensitivity depending on estimated speech and background noise levels.

Average background noise level is calculated by adding noise estimates at each band except the lowest band:

 (7)

If SNR is lower that the threshold (MIN\_SPEECH\_SNR), speech level is increased as follows:

If (speech\_level/noise\_level < MIN\_SPEECH\_SNR)

Speech\_level = MIN\_SPEECH\_SNR \* noise\_level

Logarithmic value for noise estimate is calculated as follows:

 (8)

Before logarithmic value from the speech estimate is calculated, MIN\_SPEECH\_SNR\*noise\_level is subtracted from the speech level to correct its value in low SNR situations.

 (9)

Threshold for VAD decision is calculated as follows:

*Vad\_thr = NO\_SLOPE \* (ilog2\_noise\_level - NO\_P1) + THR\_HIGH + MIN(SP\_CH\_MAX,   
MAX(SP\_CH\_MIN, SP\_CH\_MIN + SP\_SLOPE \* (ilog2\_speech\_level – SP\_P1))),* (10)

where NO\_SLOPE, SP\_SLOPE, NO\_P1, SP\_P1, THR\_HIGH, SP\_CH\_MAX and SP\_CH\_MIN are constants.

The variable vadreg indicates intermediate VAD decision and it is calculated as follows:

if (snr\_sum > vad\_thr)

vadreg = 1

else

vadreg = 0

#### 3.3.3.1 Hangover addition

Before the final VAD flag is given, a hangover is added. The hangover addition helps to detect low power endings of speech bursts, which are subjectively important but difficult to detect.

VAD flag is set to "1" if less that hang\_len frames with "0" decision have been elapsed since burst\_len consecutive "1" decisions have been detected. The variables hang\_len and burst\_len are computed using vad\_thr as follows:

hang\_len = MAX(HANG\_LOW, (HANG\_SLOPE \* (vad\_thr – HANG\_P1) + HANG\_HIGH)) (11)

burst\_len = BURST\_SLOPE \* (vad\_thr – BURST\_P1) + BURST\_HIGH) (12)

The power of the input frame is compared to a threshold (VAD\_POW\_LOW). If the power is lower, the VAD flag is set to "0" and no hangover is added. The VAD\_flag is calculated as follows:

Vad\_flag = 0;

if (pow\_sum < VAD\_POW\_LOW)

burst\_count = 0

hang\_count = 0

else

if (vadreg = 1)

burst\_count = burst\_count + 1

if (burst\_count >= burst\_len)

hang\_count = hang\_len

VAD\_flag = 1

else

burst\_count = 0

if (hang\_count > 0)

hang\_count = hang\_count - 1

VAD\_flag=1

#### 3.3.3.2 Background noise estimation

Background noise estimate (bckr\_est[n]) is updated using amplitude levels of the previous frame. Thus, the update is delayed by one frame to avoid undetected start of speech bursts to corrupt the noise estimate. The update speed for the current frame is selected using intermediate VAD decisions (vadreg) and stationarity counter (stat\_count) as follows:

if (vadreg for the last 4 frames has been zero)

alpha\_up = ALPHA\_UP1

alpha\_down = ALPHA\_DOWN1

else if (stat\_count = 0)

alpha\_up = ALPHA\_UP2

alpha\_down = ALPHA\_DOWN2

else

alpha\_up = 0

alpha\_down = ALPHA3

The variable stat\_count indicates stationary and its purpose is explained later in this subclause. The variables alpha\_up and alpha\_down define the update speed for upwards and downwards, respectively. The update speed for each band "n" is selected as follows:

if ( < )

alpha[n] = alpha\_up

else

alpha[n] = alpha\_down

Finally, noise estimate is updated as follows:

, (13)

where:

n index of the frequency band

m index of the frame

Level of the background estimate (bckr\_est[n]) is limited between constants NOISE\_MIN and NOISE\_MAX.

If level of background noise increases suddenly, vadreg will be set to "1" and background noise is not normally updated upwards. To recover from this situation, update of the background noise estimate is enabled if the intermediate VAD decision (vadreg) is "1" for long enough time and spectrum is stationary. Stationary (stat\_rat) is estimated using following equation:

, (14)

where:

STAT\_THR\_LEVEL a constant

n index of the frequency band

m index of the frame

ave\_level average level of the input signal

If the stationary estimate (stat\_rat) is higher than a threshold, the stationary counter (stat\_count) is set to the initial value defined by constant STAT\_COUNT. If the signal is not stationary but speech has been detected (VAD decision is "1"), stat\_count is decreased by one in each frame until it is zero.

if (5 last tone flags have been one)

stat\_count = STAT\_COUNT

else

if (8 last internal VAD decisions have been zero) OR (stat\_rat > STAT\_THR)

stat\_count = STAT\_COUNT

else

if (vadreg) AND (stat\_count ¹ 0)

stat\_count = stat\_count – 1

The average signal levels (ave\_level[n]) are calculated as follows:

 (15)

The update speed (alpha) for the previous equation is selected as follows:

if (stat\_count = STAT\_COUNT)

alpha = 1.0

else if (vadreg = 1)

alpha=ALPHA5

else

alpha = ALPHA4

#### 3.3.3.3 Speech level estimation

First, full-band input level is calculated by summing input levels in each band except the lowest band as follows:

 (16)

A frame is assumed to contain speech if its level if high enough (MIN\_SPEECH\_LEVEL1), and the intermediate VAD flag (vadreg) is set or the input level is higher than the current speech level estimate. Maximum level (sp\_max) from SP\_EST\_COUNT frames is searched. If the SP\_ACTIVITY\_COUNT number of speech frames is located in within SP\_EST\_COUNT number of frames, speech level estimate is updated by the maximum signal level (sp\_max). The pseudocode for the speech level estimation is as follows:

If (SP\_ACTIVITY\_COUNT > SP\_EST\_COUNT – sp\_est\_cnt + sp\_max\_cnt)

sp\_est\_cnt = 0

sp\_max\_cnt = 0

sp\_max = 0

sp\_est\_cnt = sp\_est\_cnt + 1

if (in\_level > MIN\_SPEECH\_LEVEL1) AND ((vadreg = 1) OR (in\_level > speech\_level))

sp\_max\_cnt = sp\_max\_cnt + 1

sp\_max = MAX(sp\_max, in\_level)

if (sp\_max\_cnt > SP\_ACTIVITY\_COUNT)

if (sp\_max > MIN\_SPEECH\_LEVEL2)

if (sp\_max > speech\_level)

speech\_level = speech\_level + SP\_ALPHA\_UP \* (sp\_max – speech\_level)

else

speech\_level = speech\_level + SP\_ALPHA\_DOWN \* (sp\_max – speech\_level)

sp\_max\_cnt = 0

sp\_max = 0

sp\_est\_cnt = 0

# 4 Computational details

A low level description has been prepared in form of ANSI C-code [1].

Annex A (informative) :  
Change history

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Change history** | | | | | | | | |
| **Date** | **TSG #** | **TSG Doc.** | | **CR** | **Rev** | **Subject/Comment** | **Old** | **New** |
| 03-2001 | 11 | SP-010089 |  | |  | Version 2.0.0 presented for approval |  | 5.0.0 |
| 12-2004 | 26 |  |  | |  | Version for Release 6 | 5.0.0 | 6.0.0 |
| 06-2007 | 36 |  |  | |  | Version for Release 7 | 6.0.0 | 7.0.0 |
| 12-2008 | 42 |  |  | |  | Version for Release 8 | 7.0.0 | 8.0.0 |
| 12-2009 | 46 |  |  | |  | Version for Release 9 | 8.0.0 | 9.0.0 |
| 03-2011 | 51 |  |  | |  | Version for Release 10 | 9.0.0 | 10.0.0 |
| 02-2012 |  |  |  | |  | Edithelp improvements | 10.0.0 | 10.0.1 |
| 09-2012 | 57 |  |  | |  | Version for Release 11 | 10.0.1 | 11.0.0 |
| 09-2014 | 65 |  |  | |  | Version for Release 12 | 11.0.0 | 12.0.0 |
| 12-2015 | 70 |  |  | |  | Version for Release 13 | 12.0.0 | 13.0.0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change history** | | | | | | | |
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2017-03 | 75 |  |  |  |  | Version for Release 14 | 14.0.0 |
| 2018-06 | 80 |  |  |  |  | Version for Release 15 | 15.0.0 |
| 2020-07 | - | - | - | - | - | Update to Rel-16 version (MCC) | 16.0.0 |
| 2022-04 | - | - | - | - | - | Update to Rel-17 version (MCC) | 17.0.0 |
| 2024-03 | - | - | - | - | - | Update to Rel-18 version (MCC) | 18.0.0 |