2007:262 CIV

MASTER'S THESIS

Improvements of the Voice Activity Detector in AMR-WB

Andreas Ekeroth

Luleå University of Technology

MSc Programmes in Engineering
Arena, Media, Music and Technology
Department of Computer Science and Electrical Engineering
Division of Signal Processing

2007:262 CIV - ISSN: 1402-1617 - ISRN: LTU-EX--07/262--SE

Improvements of the voice activity detector in AMR-WB

Andreas Ekeroth

November 21, 2007

Abstract

In speech coding one can make use of the speech inactivity to reduce the average bit-rate of the encoded signal. This demands a process commonly referred to as Voice Activity Detection (VAD) that separates the speech frames from the frames that only contains background noise. The purpose of the VAD is to tell the speech encoder to stop or reduce the data flow when no speech is present. The goal with such a process is to lower the average bit-rate without affecting the perceived speech quality.

This work is an investigation and evaluation of possible improvements of the voice activity detector in the Adaptive Multirate Wideband (AMR-WB) speech coder. The purpose of the work was to reduce the sensitivity to babble background noise and improve the performance for detection of music. In the report there is a brief introduction to the theory of speech coding and VAD followed by the outline of the AMR-WB speech coder. The main part of this thesis discuss possible improvements of the detector starting with recent findings in the Adaptive Multirate Narrowband (AMR-NB) algorithm.

Based on the limited material used for evaluation in this work the modifications proposed for the AMR-NB VAD showed good results also for AMR-WB. It turned out however that additional modifications should be done in order to ensure reliable detection of high level non-stationary noises. A music hangover solution was also suggested for better handling of music when the suggested modifications are implemented. The solution suggested for reduction of the sensitivity to babble noises offers a compromise between voice activity and speech clipping that can be tuned to desired performance.

The results and conclusions in this thesis are based on objective tests of limited material and contain no formal subjective testing. The conclusions should therefore be treated as guidance for further studies but indicates that the solutions proposed will reduce the AMR-WB VADs sensitivity to non-stationary background noises.

Preface

This master thesis was carried out at Ericsson Research in Luleå, Sweden and is the final part of my Master of Science degree at Luleå University of technology. The task was to investigate possible improvements of the voice activity detector in the AMR-WB speech coder.

I would like to take the opportunity to thank all of you that have supported me in writing this thesis. First of all I would like to thank Martin Sehlstedt and Stefan Håkansson at Ericsson, Martin for being my supervisor and Stefan for giving me this opportunity. Also, special thanks to my thesis student colleagues and employees at Ericsson how have been supportive and kind throughout this work. Last but not least I would like to thank Anna Carin for being encouraging in troublesome times.

Contents

1 Introduction

1.1 Background

In wireless communications the channel over which speech and audio are delivered has a limited bandwidth. It is therefore important to compress the signal before it is transmitted. To decrease the required bandwidth speech is encoded with coders that use the properties of speech production and perception to compress the signal. In modern speech coders the speech can be compressed approximately 10 times without significantly affecting the quality of the perceived speech. Beyond this point the compression starts to affect the quality and to reduce the bit-rate further speech coders take the activity of speech into account. Since a typical conversation alternate between two or more speakers it is not necessary to continuously send information. By using a voice activity detector information can be transmitted only when speech is present. In these cases it is important to have a reliable algorithm that can detect the activity of speech.

1.2 Problem Formulation

Even for good speech detectors the reliability depends on the characteristics of the background noise. If the noise has characteristics similar to speech or if the relative level difference between noise and speech is small the noise can easily be detected as speech. The problem is evident in the presence of non-stationary noise such as situations when there are multiple people talking in the background. An additional problem that arises when dealing with non-stationary noises is the presence of music. Music is typically a non-stationary signal and requires a continuous encoding not to be degraded in quality.

This thesis deals with the problem of distinguishing between speech and complex background noises in the AMR-WB VAD. The objective is to identify possible improvements and evaluate these by implementing them in a fixed point implementation of the AMR-WB codec. As a starting point for the thesis recent findings in the VAD algorithm for AMR-NB were investigated.

1.3 Scope

This project starts with a background study on production and perception of speech. It then deals with concepts used for encoding of speech and the Code Excited Linear Prediction (CELP) algorithm and the voice activity detection functionality is introduced. The outline and function of the AMR-WB algorithm is then discussed with focus on the voice activity detection part. Next there is an introduction to the improvements suggested for AMR-NB and the difference between the VADs in AMR-NB and AMR-WB are discussed. The rest of the project is dedicated to testing and evaluation of possible improvements of the AMR-WB VAD algorithm.

The experiments and simulations were performed on a C-code implementation of the AMR-WB codec distributed by 3GPP. For analysis and visualization data was exported and processed in Matlab. Throughout the report there have also been subjective evaluations of the improvements based on informal listening to critical sound samples.

2 Basics

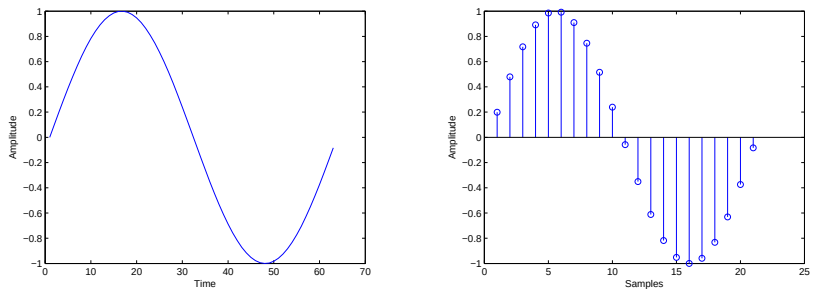
This chapter covers the basic concepts of sampling and quantization as well as a brief introduction to the physical generation and perception of speech. It is not an extensive review and contains only a brief introduction to some signal processing concepts. For further information and a more thorough examination of these concepts refer to books on signal and speech processing such as *Discrete-time Signal Processing* [1] and *Digital Processing of Speech Signals* [2].

2.1 Sampling

Since a computer only work with discrete values a continuous-time signal need to be sampled before it can be processed by a computer. The most widely used method to achieve this is through periodic sampling. By taking the value of a continuous-time signal at defined points in time the discrete signal is created. The sampling of the continuous signal $x_c(t)$, can be expressed as

$$x[n] = x_c(nT), \tag{1}$$

where n is an integer, T is the sampling period defined as $T = \frac{1}{f_s}$ and f_s is the sampling frequency. Figure 1 show a continuous-time signal of a sinusoid alongside a discrete version of the signal represented as a train of impulses. In the general case the continuous-time signal $x_c(t)$ can't be reconstructed from its samples $x[n]$. In order to remove this ambiguity the continuous-time signal needs to be band limited before it is sampled. The Nyquist criterion states that as long as the sampling frequency is at least twice the highest frequency to be sampled the signal can be reconstructed from its samples [1]. In implementations the band limiting is a low-pass filter with a cutoff frequency below $f_s/2$ that is applied to the signal before sampling.



(a) Continuous (b) Discrete

Figure 1: Continuous and discrete time representation of a sinusoid.

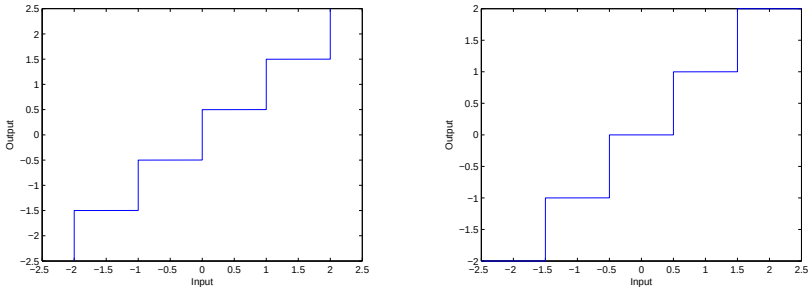
2.1.1 Quantization

In the ideal continuous to discrete-time converter the amplitude of the impulses in the discrete-time signal $x[n]$ would be known with infinite precision. Since computers work with a finite number of bits to represent data the sampled signal

need to be quantized. This means that every value is rounded to a number that exist in a discrete set. The more levels you use for each impulse the better the discrete signal resembles the continuous.

2.1.2 Scalar Quantization

The simplest form of quantization is the uniform scalar quantization where every impulse in $x[n]$ is quantized separately and all intervals Δ are of the same size. There are two common alternatives for the uniform quantizer; the midrise and the midtread shown in Figure 2. The midrise quantizer do not have zero as one of its output levels while the midtread include zero as one of the quantization levels.



(a) Uniform midrise quantizer (b) Uniform midtread quantizer

Figure 2: Uniform midrise and midtread quantizers with $\Delta = 1$.

All quantization give rise to distortion because of the round off errors, this distortion is often referred to as quantization noise. The quantization noise depends on the number of levels and decreases as the number of levels increases. Since the memory needed depends on the number of quantization levels used a compromise between memory usage and quantization noise is necessary.

In order to reduce the number of bits without increasing the overall quantization noise the signal can be non-uniformly quantized. This is done by taking the likelihood of individual amplitude levels into account. Levels that occur often have smaller quantization intervals than the ones occurring more seldom. This results in smaller quantization errors for levels that occur frequently and reduces the overall quantization noise. In speech coding logarithmically spaced quantization levels are commonly used to match the distribution of the speech signal. Two coders that rely on this concept are the A-law and μ-law quantizers. They use a quazi-logarithmic distribution that is linear for small values and logarithmic for large values [2].

Since the subsequent samples in speech signals are highly correlated, a differential quantizer that encodes the difference between adjacent samples instead of the original signal can be used. The differential quantization yields a better quality than uniform at the same bit rate as long as the samples are correlated [3].

Another way to deal with the tradeoff between memory usage and quantization noise is to adapt the step size and levels depending on the input. This is

referred to as adaptive quantization and has gotten its name from the fact that the quantizer adapts to the statistical properties of the input signal [4].

2.1.3 Vector Quantization

For the scalar quantizer every output represents a single sample of the input signal. For the vector quantizer the samples are grouped together in blocks or vectors before they are fed to the quantizer. The idea behind vector quantization was stated by Shannons's rate distortion theory which says that better performance can be achieved by coding vectors instead of scalars [4]. A drawback with the vector quantization however is the amount of computations needed to find the closest reproduction vector in the codebook.

The vector quantization works as follows: First the samples are grouped in vectors and sent to the quantizer. The incoming vector is compared to a codebook of vectors and the codebook-vector which most closely resembles the input is chosen. The index of the chosen vector is then sent to the decoder where the vector is extracted from the codebook and then unblocked to create the output samples.

2.2 Speech Production

In order to understand some of the concepts of speech compression this section brifely explains the acoustic theory of speech production. Figure 3 is a simplified model of the human vocal system showing the most important parts in the generation of sound, the vocal cords and the vocal tract. In the most generalized way speech is arranged in two classes, the voiced and the unvoiced sounds. The voiced sounds, as in the case of *"a"* and *"e"* are produced by the vibration of the vocal cords. The tension of the vocal cords dictates at which rate they vibrate and this in turn determines the pitch of the sound. Unvoiced sounds, as in the case off *"s"* and *"f"* on the other hand are created by forcing air through a constriction in the vocal tract which produces turbulence. Both the voiced and the unvoiced sounds then propagate through the vocal tract which acts like a resonator, in the same way as a pipe in an organ. The resonances produces peaks in the spectrum and the frequencies where they appear are referred to as formant frequencies and depend on the size and shape of the vocal tract. By altering the shape of the vocal tract the formant frequencies can change location and different sounds with different spectral properties can be generated. The voiced and the unvoiced sounds have certain different characteristics associated with them. Voiced sounds tend to have a periodic structure and a rather high energy while the unvoiced sounds have a noise like structure and lower energy content.

2.3 Speech Perception

The way in which humans are able to perceive and interpret speech is often defined as the process of speech perception. This process is discussed in the field of psychoacoustics [5] where the human auditory perception is investigated. The applications of psychoacoustic based models are widely used when encoding speech and music and are incorporated in commercial products such as MP3 and AAC [6], [7].

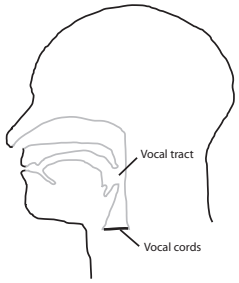


Figure 3: Simplified model of the human vocal system.

2.3.1 Limits of Hearing

The human auditory system is normally said to be able to detect sounds in the frequency range from 20Hz to 20kHz. With age however the upper limit decreases and for most adults it ranges only to around 16kHz. The levels or intensities of sounds that the auditory system can detect depend on frequency, where the lower limit is determined by the absolute threshold of hearing and the upper limit is the threshold of pain [6]. Both the intensity and the frequency resolution tend to have a logarithmic effect on the human ear. Intensity is normally expressed in Sound Pessure Level (SPL) a standard metric that defines the level of sound pressure in dB relative to a defined reference. The threshold of hearing is approximately 0 dB SPL and the threshold of pain in the vicinity of 120 dB SPL. The logarithmic behavior of frequency turn up in musical notation, where the frequency is devided into octaves and semitones. Two frequencies one octave apart are related as 2:1 and each octave is divided into 12 semitones which constitutes the basis for the western music notation.

2.3.2 Masking

When two sounds are present at the same time or closely spaced in time one sound can make the other inaudible. This effect is referred to as masking and is a temporary distortion of the threshold of hearing, making loud sounds mask quieter sounds.

Simultaneous masking sometimes called frequency masking is a process that makes one sound mask another occurring at the same time. The effect of masking is strongest if the two sounds are spaced closely together in frequency and decreases as the sounds fall further apart. Masking is much more prevalent when the masked sound has a frequency greater than the masker.

Masking can also be found between sounds that do not occur at the same point in time. This type of masking is called temporal masking and is mostly prevalent when the masker occurs before the masked sound. Surprisingly however a sound can also be masked by a louder sound occurring at a later point in time.

The concept of masking is widely used in coding of audio since it makes it possible to spend most of the coding effort where it is believed to be of great perceptual importance.

3 Speech Coding

Speech coding or speech compression is the name of the different techniques used to obtain digital representations of speech signals. The objective with speech coding is to represent speech with as few bits as possible without degrading the perceptual quality. In most cases, however, there is always to some extent a loss in quality and a compromise between bit-rate and quality needs to be made.

Coding of speech can be divided into two main classes, the waveform coders and the parametric coders. The waveform coders rely merely on the concepts of sampling and quantization to represent the signal in the digital domain and the digital representation tries to mimic the input waveform. The parametric coders on the other hand rely on a model of the source for which the different parameters are estimated and sent to the decoder. At the decoder side they are then used to reconstruct a signal similar in perception to the original. For this approach to be successful the model of the source need to be well known. Since the concept of speech production is well investigated and documented parametric coding has proven to be an effective method in coding of speech.

As an extension of the two main categories mentioned above there is a mixture of the two classes that combine the low bit-rates possible with the parametric coding with the high quality of waveform coding. These are referred to as hybrid coders.

3.1 Waveform Coding

Waveform coders try to produce a reconstructed signal that is as close to the input waveform as possible. This is done without any consideration of how the signal was generated and can therefore be used with good results on all kinds of sounds. The simplest coder that uses this technique is Pulse Code Modulation (PCM) which is made up of sampling and uniform quantization. It is the PCM encoding technique that is used to code the data in the Compact Disc (CD) format. In order to reduce the bit-rate for the waveform coders, other quantization techniques such as non uniform and adaptive have been proposed and used for encoding. More information on different waveform coding techniques can be found in [8], [2] and [4].

3.2 Parametric Coders

Unlike the waveform coders, the parametric coders do not attempt to reconstruct a representation of the actual waveform, but reproduces a signal that is similar in perception. This is done by considering how the actual waveform is produced and perceived.

3.2.1 Vocoders

The first description of a vocoder was published in 1939 by Homer Dudley [9] and introduced the basic concepts used in parametric coding. The term vocoder is short for voice coder and is a speech synthesizer network that makes up a generalized model of the physical vocal system depicted in Figure 3. Typically a vocoder is based on the source-filter speech model consisting of excitation signals and a time varying filter [3]. A simple source-filter vocoder model is

depicted in Figure 4. The excitation signal which is a model of the lungs and the vocal chords is either a periodic pulse train (for voiced sounds) or white noise (for unvoiced sounds). For the pulse source the period between the pulses is an estimation of the fundamental frequency in voiced speech. The vocal tract filter is time varying and functions as a model for the resonances enforced by the vocal tract. Parameters that need to be evaluated and transmitted for the model are the filter parameters, the choice if the speech is voiced or unvoiced and in the case of a voiced sound also the pitch. This reduces the data transmitted to a fraction of the original speech signal but increases the computational load when encoding and decoding. Another problem is that vocoders have a tendency to sound artificial and often perform badly on non speech signals since the model is optimized for speech.

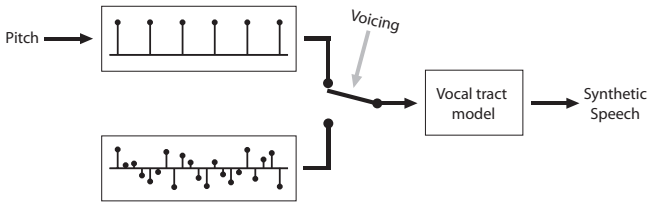


Figure 4: Block diagram of a source-filter speech production model.

3.2.2 Linear Predictive Coding

The most well known vocoder is the Linear Predictive Coder (LPC) which estimates the vocal tract filter using linear prediction. For the source-filter model in Figure 4, a linear predictive filter as a vocal tract model would be given by the difference equation

$$s[n] = \sum_{k=1}^{P} a_k s[n-k] + x[n], \tag{2}$$

where $x[n]$ is the input to the filter, either an impulse train for voiced speech or a random noise sequence for unvoiced speech, $s[n]$ is the synthetic speech output and P is the order of the predictor. The system function of the filter can be expressed as

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}. \tag{3}$$

The system function of a $P : th$ order forward linear predictor is defined as

$$P(z) = \sum_{k=1}^{P} a_k z^{-k}, \tag{4}$$

and the system function for the error of the predictor becomes

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k}. \tag{5}$$

Thus, the system $H(z)$ can be rewritten as

$$H(z) = \frac{1}{A(z)}. \tag{6}$$

The difficulty that arises when using the LPC is then to find the predictor coefficients in $H(z)$ that will yield the optimal predictor. Since the typical speech signal is time-varying the predictor coefficients will change over time and must be estimated from short time segments. In order to achieve a time varying filter, speech is analyzed in time segments, usually called frames. For each frame the predictor coefficients are estimated by minimizing the errors between the predicted segment and the one being analyzed. In most implementations the coefficients are calculated based on minimization of the mean squared error. There are two different solutions for estimating the predictor coefficient values called the autocorrelation method [10] and the covariance method [2].

3.3 Hybrid Coding

The hybrid coders uses concepts from both the parametric and the waveform coders to produce low bit-rate speech with high perceptual quality. These coders use a filter to model the vocal tract (as in parametric coders) and then choose an excitation signal so that the synthesized speech matches the original as close as possible (as in waveform coders). The majority of these coders are based on analysis by synthesis techniques and use a closed-loop approach to find the decoder parameters by minimizing an error signal. In order to perform the search for the optimal parameters a version of the decoder is included in the encoder.

The encoding is performed by perceptually optimizing the synthesized speech so that the difference between the input speech and the decoded speech is minimized. In order to "optimize" the coder for the human ear the difference is perceptually weighted by a filter, so that the quantization noise appears mostly in the formant regions where it is masked by the high energy of speech. In Figure 5 a LPC-based generalized hybrid encoder is depicted.

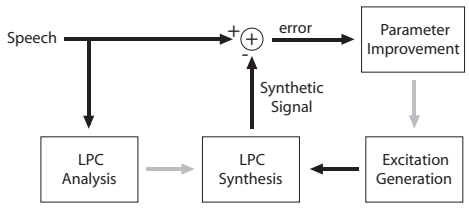


Figure 5: Block diagram of a generalized Analysis by Synthesis encoder.

3.3.1 Code Excited Linear Prediction

Code Excited Linear Prediction (CELP) is an analysis by synthesis method that was originally proposed by M.R Schroeder and B.S Atal in 1984 and is, along with its variants, the currently most widely used speech coding algorithm [4]. In Figure 6 a block diagram describing the original CELP encoder is shown.

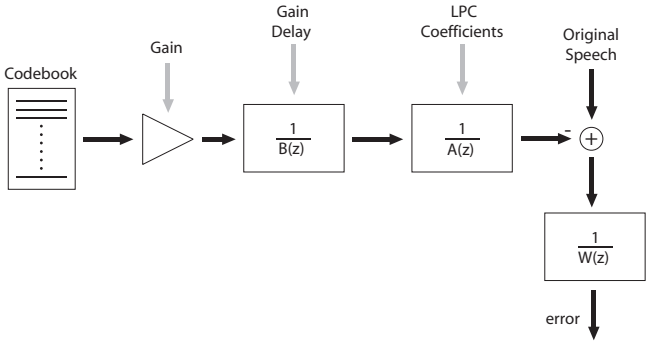


Figure 6: Functional description of the synthesis part of the CELP encoder

The synthesis part of the encoder starts by sending a scaled excitation originating from a codebook populated by gaussian sequences through the long and short term predictors. The Long Term Predictor (LTP) introduced serves to model the pitch period of voiced speech by repeating the excitation sequence. With the LTP included the pulses in the excitation are not all spent on modeling the pitch and can as a result be used to model other structures in the excitation. The LTP is implemented as a one- or three-tap filter and has a system function of the form

$$\frac{1}{B(z)} = \frac{1}{1 - \sum_{k=-K}^{K} b_k z^{-(\tau+k)}}, \tag{7}$$

where K is either zero for a one-tap filter or one for a three-tap, b_k the coefficients and τ the delay. The short term predictor that also substitutes the vocal tract filter is obtained and characterized by the LPC coefficients described previously. Once the synthetic speech is estimated it is compared to the original speech and the difference is fed to a perceptual weighting filter $W(z)$. By perceptually weighting the error signal the spectral contribution of the formants can be reduced. If this is not done the high energy in the formants will dominate the error and coding effort is spent on regions where the quantization noise is partially masked by speech. For this reason the perceptual weighting filter depends on the linear predictor as:

$$W(z) = \frac{A(z)}{A(z/\gamma)}, \tag{8}$$

and the constant γ creates a filter similar to the inverse of the speech spectra. The coding process is then repeated for all excitation signals in the codebook and the excitation is chosen such that the perceptually weighted error is minimized in the mean square error sense.

The information sent to the decoder is the entry of the excitation in the codebook, the gain, the LTP coefficients and the LPC coefficients. The decoding is then performed in the same way as the synthesis part of the encoder.

3.4 Voice Activity Detection

A voice activity detector is used to indicate whether a signal contains speech or not. The decision made by the voice activity detector is then used to facilitate

the processing, coding and transmission of speech signals.

Since a typical conversation between two persons alternate between the speakers, it is not necessary to continuously code the speech signal. By incorporating a voice activity detector it is possible to use this fact and tell the coder only to transmit information when speech is present. As the typical speech coder works on blocks of data one can make a voice activity decision for each frame, indicating whether the frame contains speech or not. The knowledge of speech activity can then be used by the speech coder to lower the average bit rate of the encoded speech and also to lower the average power consumption in mobile handsets. In applications such as VoIP the speech pauses can be used to send other data information.

3.4.1 Functional Description

A functional description of a voice activity detector in a speech coding situation can be found in Figure 7. For this description the incoming speech is assumed to be divided into frames as is common for most speech coders. For each frame a decision is made by the VAD whether the incoming signal is speech or not. The decision can be made directly by analyzing the input signal to the VAD but in many cases the VAD also make use of parameters calculated by the speech encoder. Based on the decision made by the VAD the frame is then coded either with the speech or the noise encoder and sent through the channel.

When the VAD has indicated presence of speech the frame is coded with the speech encoder and the frames are transmitted to the decoder in the normal manner. If the VAD on the other hand indicates that no speech is present the noise encoder will send information to the receiver in a special frame telling the decoder to switch to noise decoding. When the VAD later indicates speech again the normal speech frames will be sent and recognized by the receiver.

It is important to understand that the improvement is dependent on the actual speech activity, if the actual speech activity for example is 50% one could reduce the average bit rate to almost half without degrading the perceptual speech quality. This on the one hand demands a good recognition of speech by the VAD and a decision such that no background noise is classified as speech. On the other hand if speech is classified as noise that speech segment is lost and this in turn deteriorates the speech quality. In all VAD implementations it is therefore a trade-off between speech misclassified as noise and noise misclassified as speech. This is especially difficult in low SNR conditions and when the noise is non-stationary.

3.4.2 Comfort Noise

For the example above when the VAD has indicated that no speech is present there are no parameters sent to the receiver. The result of this may give unwanted effects on the listener at the receiver side since no information is decoded and the listener hears nothing. If total silence is perceived the listener may hang up since he or she believes that the connection has been lost. To resolve this problem a noise is added at the decoder when no speech is present, this is called comfort noise. The noise can be either randomly generated in the decoder or based on properties of the actual background noise at the transmit side. If the background noise at the transmit side is used, a noise encoder similar to

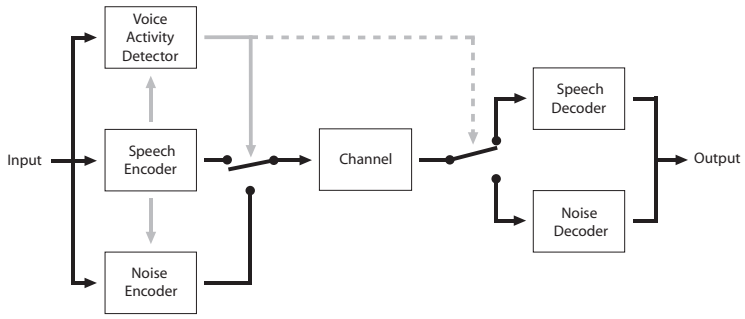


Figure 7: Generalized voice activity detection model

the speech encoder sends information for the non-active speech frames. The bit rate at which the noise encoder sends information must be lower than the speech encoders rate otherwise the VAD serves for nothing.

3.4.3 Speech Detection

There exists a variety of VAD algorithms that differ both in complexity and quality and these can be divided into two main categories: the time and the frequency domain techniques [11], [12].

For the time domain algorithms the most important part in the VAD decision is the energy content and the zero crossings for each frame. The energy content is used since the speech energy is assumed to be higher than the background noise in most situations. When using the zero crossing rate it is assumed that speech in noise has a lower rate than noise on its own which is true as long as the SNR is high and the noise is white in structure.

When a frequency algorithm is used it is the energy distribution and the variance of the spectrum that is of interest.

4 Adaptive Multirate Wideband

Adaptive Multi-Rate Wideband (AMR-WB) is a speech coder based on the Algebraic Code Exited Linear Prediction (ACELP) technology and is standardized by 3GPP as the mandatory codec for wideband telephony [13]. The codec works at a sampling rate of 16 kHz on blocks of 20 ms and consist of nine source rates ranging from 6.60 kbit/s to 23.85 kbit/s. There is also a low bit-rate mode for encoding background noise. The configuration of the processing functions in the AMR-WB codec is given in Figure 8. The three speech processing functions in the codec consist of the multi-rate speech encoder, a voice activity detector and a comfort noise generation system. These are fed to either the Discontinuous Transmission (DTX) functionality in GSM or the Source Controlled Rate (SCR) functionality in 3G. The function of the DTX and SCR is to reduce the average bit rate by taking speech inactivity into account. The general specification of AMR-WB is described in [14].

The input to the speech encoder is a 14-bit uniform PCM signal left justified in a 16-bit word, sampled at 16 kHz. The bit-rate for this raw input is 256 kbit/s which is to be compressed to a maximum of 23.85 kbit/s. In order to achieve the compression the encoder works with an ACELP vocoder, which is based on the CELP vocoder technique but uses a special algebraic codebook structure.

The function of the voice activity detector is to decide whether the 20ms speech frame contains speech or not. This decision is then used by the DTX/SCR functionality to encode the frame either with the ACELP coder or the comfort noise system. The comfort noise system sends information at a much lower rate than the ACELP coder and is used when no speech is expected to be present.

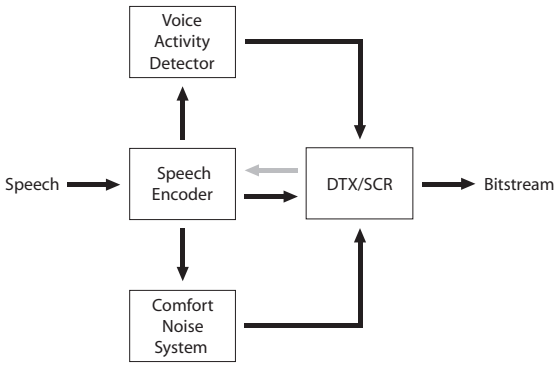


Figure 8: A functional description of the processing blocks in a AMR-WB encoder.

4.1 ACELP

The coder works on two frequency bands, 50-6400 Hz and 6400-7000 Hz, which are encoded separately. Linear Prediction (LP) analysis is performed on the lower band once per 20ms frame and the search of fixed an adaptive codebooks is done every 5ms. The higher frequencies are reconstructed in the decoder by filtering a random excitation through a LP filter derived from the LP coefficients

of the lower band. A Functional description of the synthesis part in the ACELP encoder is found in Figure 9.

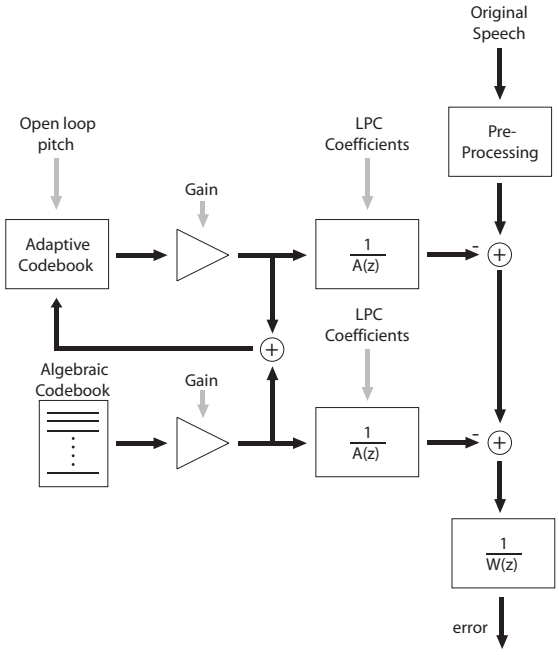


Figure 9: Functional description of the synthesis part in the ACELP encoder.

4.1.1 Pre-processing

Before the actual coding starts some preprocessing of the signal is done. First the signal is down sampled to a rate of 12.8 kHz. Thereafter the signal is high-pass filtered and scaled down by a factor of 2 to avoid overflow in fixed-point implementations. The high-pass filter has a cut off frequency of 50Hz and its purpose is to avoid feeding the coder with undesired low frequency components. In order to enhance the resolution of the LP analysis the energy of higher frequencies is raised. This is done in the last step of the preprocessing where the signal is pre-emphasized with a first order high-pass filter given by

$$H_{pre-emp}(z) = 1 - 0.68z^{-1} \tag{9}$$

4.1.2 Linear Prediction

LP analysis if performed once per 20 ms speech frame with an overhead of 5 ms in both directions. The 30 ms frame is windowed by a asymmetric window that has its weight concentrated around the 4:th sub frame and autocorrelations of the windowed speech is calculated and a 60 Hz bandwidth expansion is performed. The LP coefficients are obtained from the modified autocorrelations by solving the equations using the recursive Levinson-Durbin algorithm [2].

When the LP filter coefficients are estimated they are converted to the Immittance Spectral Pairs (ISP) representation for quantization and interpolation. The reason for using the ISP representation is that quantizing, transmitting and interpolating the coefficients directly may result in unstable prediction filters. The ISP decomposition is performed by splitting the LP-filter $A(z)$ in to two new polynomials

$$P(z) = A(z) + z^{-(N+1)}A(z^{-1}) \tag{10}$$

and

$$Q(z) = A(z) - z^{-(N+1)}A(z^{-1}). \tag{11}$$

While the polynomial $A(z)$ has complex roots anywhere $P(z)$ and $Q(z)$ alternate each other around the unit circle and quantization and interpolation of these are done without the risk associated with working directly on $A(z)$. The coefficients are then quantized in the ISP representation using split-multistage vector quantization.

This is a short explanation on the LP- analysis and quantization for more extensive information refer to [15] and references there in.

4.1.3 Perceptual Weighting

Due to the extended bandwidth in wideband speech coding the traditional perceptual weighing filter $W(z)$ has shown limitations and a modified version of it is therefore used, given as

$$W(z) = A(z/\gamma_1)H_{de-emp}, \tag{12}$$

where

$$H_{de-emp} = \frac{1}{1 - 0.68z^{-1}} \tag{13}$$

and A(z) is the LP filter calculated based on the preemphasized input signal.

4.1.4 Pitch Analysis and the Adaptive Codebook

The pitch analysis in AMR-WB is implemented in three steps. In the first step an open loop pitch search is performed once or twice per frame depending on the coder mode. This result in the open-loop pitch lags and restricts the closed-loop search implemented by the adaptive codebook. In the second step a closed-loop search is performed around the open-loop pitch lag at a range of ± 7 samples. When the optimum integer pitch lag is found the third step goes through fractional lags around the integer value. The resolution of the adaptive codebook is 1/4 of a sample and the fractional pitch search is performed by interpolation. When the lag is found the last thing to do is to search for the gain that minimizes the error signal.

The lags are absolute coded for sub-frames one and three and delta coded around the previous sub-frame for the second and fourth sub-frames. Since the harmonic structure does not necessarily cover the whole spectrum an optional low-pass filter can be applied to give a high frequency attenuation of the periodicity.

4.1.5 Algebraic Codebook

The algebraic codebook uses interleaved single-pulse permutation design based on a code vector which has 64 positions. The code vectors are divided in to 4 tracks with 16 positions in each track. For each of these tracks a certain number of signed impulses with the value $+1$ or -1 are inserted. The number of pulses per track is dependent on the mode of the coder and ranges from 1 to 6. The codeword then represents the signs and positions of the pulses in each track and no codebook storage is needed since the index itself describes the excitation vector.

The use of the sparse algebraic codebook can give rise to artifacts and for this reason an adaptive prefilter $F(z)$ is used to improve the subjective quality. $F(z)$ consists of two parts, the first part being a periodicity enhancement filter based on the integer pitch lag T as $1/(1-0.85z^{-T})$ and the second part is a tilt filter $1-\beta z^{-1}$ where β is related to the voicing of the previous frame.

4.2 Voice Activity Detector

The function of the VAD is to decide whether a frame contains speech or just background noise. The decision made by the VAD algorithm is then sent to the DTX/SCR which code the frame either with the ACELP encoder or the comfort noise system depending on the decision.

The input to the VAD is the 20ms speech frame sampled at 12.8kHz. For each input frame the signal is divided in to 12 sub bands and the level for each sub band is calculated. A tone detection function based on the open-loop pitch gains calculated by the speech encoder is used for indication of strongly periodic signals, such as music or voiced speech. In each frame speech and noise level estimations are calculated. For speech only one estimation is done for all bands while the noise is estimated separately in each sub-band. The background noise estimations are then used to calculate an input SNR. An intermediate decision is made by comparing the SNR to an adaptive threshold that is based on noise and speech estimations. In order not to code low energy endings of speech segments as noise the final VAD decision is made by adding a hangover to the intermediate decision.

4.2.1 Filter Bank and computation of sub-band levels

The filter bank divides the signal in two 12 sub bands by using 3rd and 5th order filters combined in accordance with Figure 10. To ensure that no saturation occurs in the filter bank calculations, all input values are divided by two before filtering starts. Each filter block is a combination of first order direct form all-pass filters and divides the input into high-pass and low-pass outputs. The signal level for each output is calculated by summing the absolute values of the samples in a specified range. For details on filters and signal level calculations the reader can refer to [16] and [17].

4.2.2 Tone Detection

The tone detection function uses the open-loop pitch gains calculated by the ACELP encoder to detect periodic signals. This is done by comparing the open-loop pitch gain to a constant that functions as a threshold. If the value

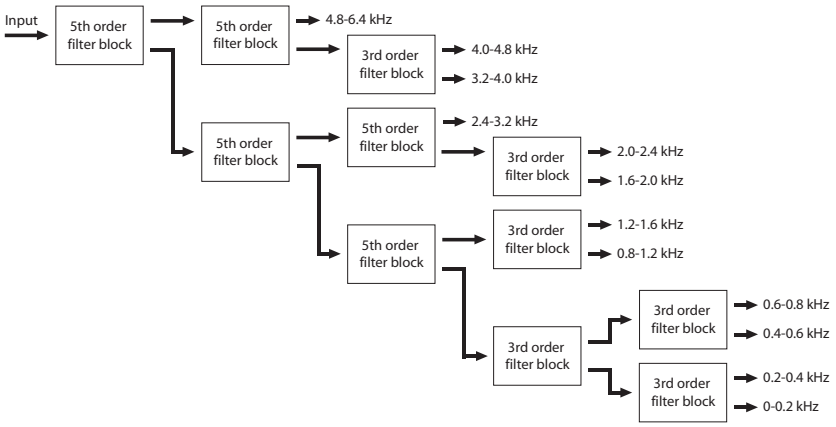


Figure 10: Block diagram of the filters in the filterbank.

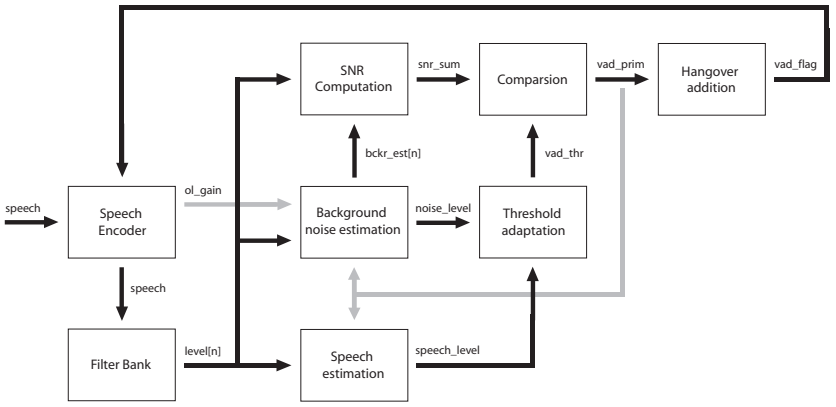


Figure 11: VAD decision block diagram.

of the open-loop gain is larger than the threshold a tone is declared and a tone flag is set to one. Additionally there is a calculation of the total power of the unfiltered speech frame as

$$frame\_pow = \sum_{i=0}^{LEN} s[i]^2 \tag{14}$$

where $s[i]$ are the samples and LEN is the length of the frame. If the power of the current and the previous speech frame is less than a particular threshold the tone-flag is set to zero.

4.2.3 VAD Decision

The decision whether the frames contain speech or not is made by comparing the estimated SNR to an adaptive threshold and then add a hangover. A simplified block diagram of the VAD decision algorithm is shown in Figure 11.

Background Noise Estimation Background noise estimations are calculated for each sub band and is based on previous VAD-decisions, signal stationarity and tone decision. The update of the estimate is based on previous estimations and input levels and is updated as

$$bckr\_est_m[n] = (1 - \alpha[n])bckr\_est_{m-1}[n] + \alpha[n] * level_{m-2}[n], \quad (15)$$

where n is the index of the frequency band and m is the frame index. In order to limit the background estimates not to take on too large or too small values the estimate is restricted between two constants.

The update speed for each sub band is set by the variable α as can be seen in equation 15. This α in turn depends on previous intermediate VAD decisions and a stationary counter. The purpose of the stationary counter is to recover from a situation where the background noise suddenly increases. The update speed for the variable α is selected from a set of constants as follows:

- If the intermediate VAD decision for the last four frames have been zero the update speed α is set to normal update both downward and upwards.
- If the intermediate VAD decision for the last four frames have not all been zero only a reduced update speed for α is used.
- If a variable called stationary counter is set no update upwards and only a reduced update downwards is allowed.

The final decision whether the update should be upwards or downwards depends on how the previous background noise estimate relates to the input level two frames back. If the last background noise estimate is less then the level two frames back α is chosen such that the background estimate is updated upwards otherwise it is updated downwards.

The stationary counter that enables the background noise estimate from updating upwards is set if either of the three following conditions apply. If the last five tone flags have been zero, if the last 8 intermediate VAD decisions have been zero, or the variable that estimates the stationarity is larger then a threshold. The variable defining the stationarity is estimated as follows

$$stat\_rat = \sum_{n=1}^{12} \frac{MAX(STAT\_THR, MAX(ave\_level_m[n], level_m[n]))}{MAX(STAT\_THR, MIN(ave\_level_m[n], level_m[n]))}, \quad (16)$$

where $STAT\_THR$ is a constant and ave_level is an estimated average value of the input signal [16].

There is also a calculation of the sum of the background noise estimations for use in the calculation of the adaptive thresholds. This is defined as

$$noise\_level = \sum_{n=2}^{12} bckr\_est[n]. \quad (17)$$

SNR Computation The signal to noise ratio is calculated once per frame by taking the difference between the input level and the background noise estimate

for each sub band. The quotient is then raised by a power of two and a summing of all individual sub band SNR:s is calculated as

$$snr\_sum = \sum_{n=1}^{12} \left(\frac{level[n]}{bckr\_est[n]} \right)^2, \tag{18}$$

where $level[n]$ is the signal level at sub band n and $bckr\_est[n]$ is the background noise estimate for that same sub band.

Speech level estimation The speech estimate for each frame is calculated based on the sum of all but the lowest sub band. If the input level defined as

$$in\_level = \sum_{n=2}^{12} level[n] \tag{19}$$

is higher then a threshold and the intermediate VAD decision indicates speech activity, or if the input level is larger then the current speech estimate the frame is assumed to contain speech and the speech level is updated if the level stays high for a number of consecutive frames. For details on the implementation refer to the pseudocode for the estimation found in [16].

Intermediate VAD decision and Threshold adaptation The Intermidiate VAD decision is made by comparing the SNR sum to an adaptive threshold. If the SNR sum is larger then the threshold, the frame is declared as speech and the intermediate VAD-flag for that frame is set to active. The calculation of the threshold is performed once per frame and is based on the speech and noise estimates. Details on the calculation of the threshold can be found in [16].

Hangover Addition In order not to lose speech frames with low energy a hangover is added to the intermediate VAD decision before the final VAD decision is sent. If the power of the input frame is lower then a threshold the VAD-flag is set to zero and no hangover is added. For all power levels larger than the threshold a hangover, computed based on the adaptive threshold is added [16].

4.3 Discontinuos Transmission

The decision made by the VAD is used by the SCR/DTX functionality to switch between speech and noise encoding [18]. This feature reduces the average bitrate of the encoded signal which saves power in the user equipment and reduces the overall load in the network. The SCR/DTX mechanisms allows the transmission to be switched off most of the time during speech pauses and reduces the computations in the ACELP encoder so that only the computations necessary for the comfort noise estimation are performed.

4.3.1 Comfort Noise

The comfort noise is generated by sending parameters describing the energy and spectral properties of the background noise. These parameters that define the comfort noise is evaluated from the ACELP speech encoder described in section

4.1. The parameters that are sent to the receiver are averaged over the 8 most recent frames and include a weighted average of the spectral parameters, the average of the logarithmic signal energy and a stationarity flag. This information is then encoded in a Silence Descriptor (SID) frame and sent to the decoder.

At the decoder the parameters are used to synthesize the comfort noise by sending a pseudo random noise scaled by the logarithmic signal energy through a synthesis filter generated from the average of the spectral parameters. If the background noise is non-stationary dithering of the energy and spectral parameters is employed.

Spectral parameters The spectral parameters are calculated from the unquantized linear prediction parameters in the Immittance Spectral Frequency domain (ISF) where the ISF-vector $\vec{f}$ is calculated by the speech encoder.

Before the averaging take place a replacement of at most two of the ISF-vectors is done in order to remove parameters that are not characteristic of the background noise. By calculating the spectral distance ΔS_i for all ISF parameters in the 8 current frames the ISF parameter vector with the smallest spectral distances is chosen and denoted f_{med} [19]. If there are any ISF parameter vectors $f(i)$ that makes the quotient $\frac{\Delta S_j}{\Delta S_{med}}$ exceed a threshold the two of these vectors with the largest quotient are replaced by f_{med}.

The average spectral parameters are then calculated by averaging over the 8 frames

$$f^{mean}(n) = \frac{1}{8}\sum_{i=0}^{7} f(n-i). \tag{20}$$

Frame energy To get a gain parameter for the pseudo random noise generator in the decoder the average frame energy is estimated from the 8 frames making up the SID frame. For each frame the energy of the high-pass filtered input signal is calculated as

$$en_{log}(m) = \frac{1}{2} log_2 \left(\frac{1}{N} \sum_{i=0}^{N-1} s[i]^2 \right), \tag{21}$$

where $s[i]$ denotes the input speech samples. The average energy over the 8 frames is then calculated as

$$en_{log}^{mean}(m) = \frac{1}{8}\sum_{n=0}^{7} en_{log}(m-n). \tag{22}$$

5 Improvements in AMR-NB VAD

The voice activity detector in the AMR-NB speech coder has shown limitations when exposed to non stationary background noise types. For this reason modifications of the VAD algorithm that reduces this problem has been presented in [20]. In this section two modifications proposed for the NB-VAD are discussed.

The AMR-NB coder has two implementations of the voice activity detector, NB-VAD1 and NB-VAD2. Details on the implementations can be found in [21]. Since the VAD for AMR-WB has a structure similar to NB-VAD1 it is probable that the solutions for the NB coder can be used in the WB coder. The modifications suggested in [20] are done in the NB-VAD1 algorithm and for this reason it will be discussed here. The decision making algorithm for NB-VAD1 is depicted in Figure 12 and the major differences between the NB-VAD1 and WB-VAD algorithm are summarized here:

- WB-VAD operates on a broader spectrum up to 6400Hz in contrast to NB-VAD1 which operates on frequencies below 4000Hz.
- The WB-VAD algorithm has, because of the broader spectrum, also more sub-bands, 12 instead of 9 in the NB-VAD1. The cut off frequencies for the different bands for both the wide and narrow band detectors are shown in Table 1.
- The speech estimate that is used in the VAD threshold calculation in the wide band implementation is not included in NB-VAD1.
- Additional functions to detect tones and correlated signals are used in NB-VAD1.

Table 1: Cut-off frequencies for VAD filter banks

(a) Cut-off frequencies for WB-VAD filter bank

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

(b) Cut-off frequencies for NB-VAD filter bank

Band number	*Frequencies*
1	0-250Hz
2	250-500Hz
3	500-750Hz
4	750-1000Hz
5	1000-1500Hz
6	1500-2000Hz
7	2000-2500Hz
8	2500-3000Hz
9	3000-4000Hz

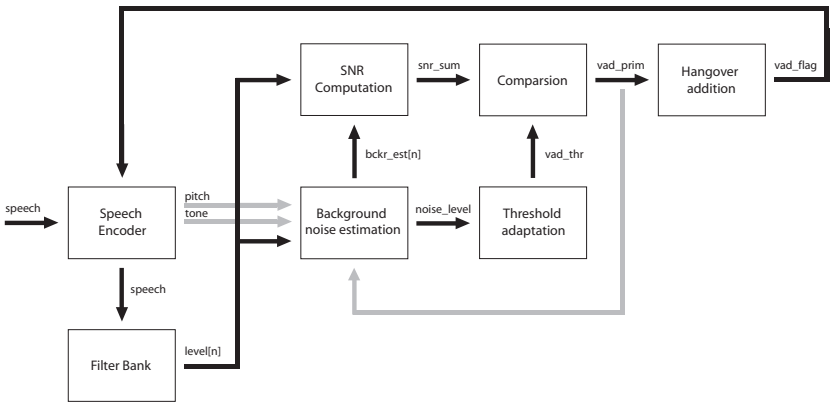


Figure 12: The voice activity detector for AMR-NB

5.1 Significance Thresholds

Insertion of significance thresholds is discussed in [20] and introduces a non linearity in the calculation of the sub band SNR values. The aim with the process is to reduce the risk that many low sub band levels sums up and make the SNR exceed the VAD threshold. The name significance thresholds originates from the fact that a threshold is used to control the SNR required for a sub-band to be considered significant in the SNR summation.

Since the voice activity detector for AMR-NB has the same outline as the AMR-WB coder, except for the details just discussed, the significance threshold modification will be presented here without any deeper discussion on the AMR-NB functionality. The details of the SNR computation block in Figure 12 is depicted in Figure 13.

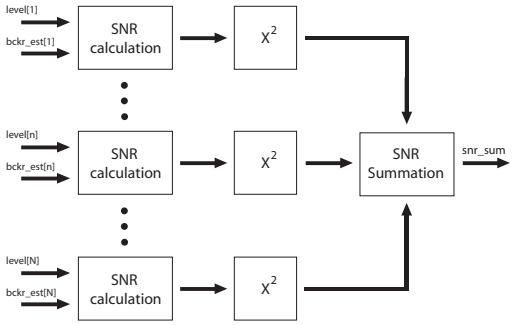


Figure 13: Signal flow diagram of the SNR calculation and summation.

The calculation of the signal to noise ratio for each sub-band is done in the same way for both the WB-VAD and the NB-VAD1, that is

$$snr[n] = \frac{level[n]}{bckr\_est[n]}. \tag{23}$$

A primary voice activity decision is made by summing the squares of all the

sub-band SNRs,

$$snr\_sum = \sum_{n=1}^{N} snr[n]^2 \tag{24}$$

and comparing the SNR sum to the VAD thresholds described in [21] and [16] respectively. The non-linear function that constitutes the modification results in

$$snr\_sum = \sum_{n=1}^{N} \begin{cases} sign\_floor^2 & if \quad sign\_floor < snr[n] < sign\_thr \\ snr[n]^2 & otherwise \end{cases} . \tag{25}$$

A modified version of the signal flow diagram for the SNR calculation, including the significance thresholds can be found in Figure 14.

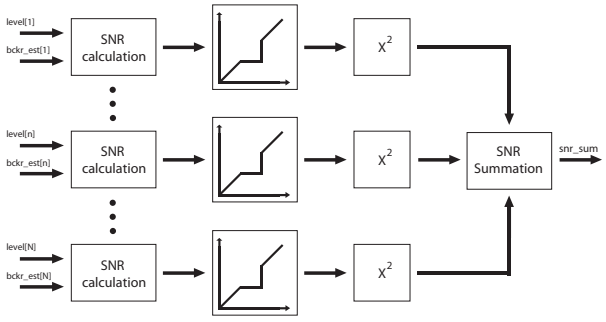


Figure 14: Signal flow diagram of the SNR calculation and summation with the non linearity inserted.

The $sign\_floor$ and $sign\_thr$ values in equation 25 can be altered in order to achieve desired performance. Note however that by setting $sign\_floor$ and $sign\_thr$ equal the non-linearity is removed and the summation is the same as for the original VAD in equation 24.

5.2 Difference Threshold

This next modification is an extension of the significance thresholds just discussed. The idea is to compare a low pass filtered version of the primary VAD activity made by the "aggressive" significance threshold VAD to a low pass filtered version of the original VAD activity. If the difference between the two is larger than a threshold the background noise update is reduced or halted when updating upwards.

The process is divided into three parts. In the first part a short term activity is calculated for the two different VADs based on the sum of the latest 32 primary VAD flags. Secondly the long term activities are calculated by a low pass filter with system function

$$VAD\_AR(z) = \frac{1-\alpha}{1-\alpha z^{-1}}. \tag{26}$$

In the third step the difference between the long term activities are compared to a threshold and a decision is made whether the difference is significant or

not. The result is a difference flag that is sent to the background estimator and used in the decision of the update of background noise levels.

6 Observed WB-VAD Behaviour

In order to understand how the VAD responds to different types of input signals this section considers the behavior of different elements in the VAD algorithm. It also contains a presentation of the test samples and the metrics used for evaluation throughout the report.

6.1 Test Samples

During the investigations a selection of critical test samples have been used to identify problems and to evaluate possible improvements. In order to facilitate for the reader only a subset of these are used and presented in the report. The test samples were grouped in four different categories during the investigation and for each group one or two samples representing the typical behavior of the group was chosen to represent the category throughout this report. All samples used have a sample rate of 16 kHz and are in signed 16 bit PCM format.

The four categories are surmmarized here with a short description of the samples chosen to represent them.

White noise White noise is used as a reference since the VAD has the quality of handling large levels of this particular noise type without making false voice decisions.

Car noise Since the VAD was originally designed to achieve reliable decisions when exposed to car background noise a suggested modification should deteriorate this aspect as little as possible. The sample used is punto_1 from the NTT-AT Multi lingual database and is referred to as car noise in plots and figures.

Babble noise To represent babble noise two different samples were chosen. The first is an actual recording from an office where it is possible to hear what is being said. The sample is as the previous from the NTT-AT database and named offi_i_2. The second babble sample is a synthetic babble combined from 32 different individual speakers in the NTT-AT database. For this sample it is not possible to discern what is being said by a specific speaker. The recorded babble noise is labeled babble and the synthetic is labeled babble 32 throughout the report.

Speech The speech sample is based on a sample of a single speaker from the NTT-AT database. The file used is am_1 and this sample contains a male English speaker. This sample is labeled speech.

6.2 Metrics

Performance measures of the different VAD:s in this report are based on two different metrics. The first is the Voice Activity Factor (VAF) calculated as

$$VAF = \frac{\#\ of\ active\ speech\ frames}{Total\ \#\ of\ frames}. \tag{27}$$

This factor can be compared to a reference VAF to see how different SNR:s, different noise types and different settings affect the VAF.

The second metric is used to find speech misclassified as noise, this is referred to as clipping and measures the loss of speech frames compared to a reference VAD. In this report the clipping is based on a metric introduced in [20] and considers only speech clipping in the case where the relative loudness between speech and noise is high enough not to have the speech masked by noise [22]. This results in the following formulation

$$Clipping = \frac{\#\ non\ active\ frames\ with\ loudness\ >\ 0}{Total\ \#\ reference\ speech\ frames}. \tag{28}$$

The reference used for the calculation is throughout the report based on the encoding of clean speech with a level of -26 dBov and the loudness is calculated on a frame by frame basis as

$$L_{sp}(n) = \left(\frac{max(0, sp(n) - 0.25no(n))}{1 + \left(\frac{no(n)}{sp(n)}\right)^2} \right)^{0.3} \tag{29}$$

where sp(n) and no(n) is the energy in speech and noise respectively.

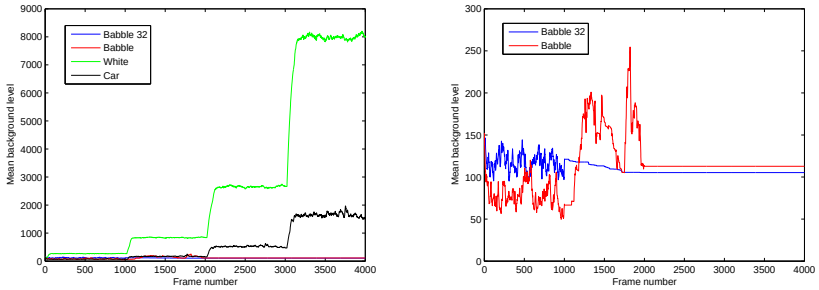
6.3 Background Noise Level Estimate

The update speed of the background noise estimates depend on two things, the intermediate VAD decisions and the stationary counter. If the level increases suddenly these two aspects will halt the background noise from normally updating upwards and as long as speech is detected by the VAD only a reduced update speed is allowed. If the variations in the levels are large the upwards update is halted and the background noise can only be updated downwards.

In order to understand how the background noise estimation responds to different noise inputs the encoder was fed with different noise types and the background noise levels where exported to Matlab for analysis. A simulation showing the background noise levels for different noise input signals can be found in Figure 15. The noise levels of the input signals were scaled to -56 dBov RMS for the first 1000 frames increasing to -26 dBov RMS in 10 dB steps for each 1000 frames. For clarity Figure 15(b) show only the levels for babble noise inputs. Note from the two figures that the background noise level is not updated properly for the two babble noises, while the white and car noise input give the expected response.

6.4 Speech Level Estimate

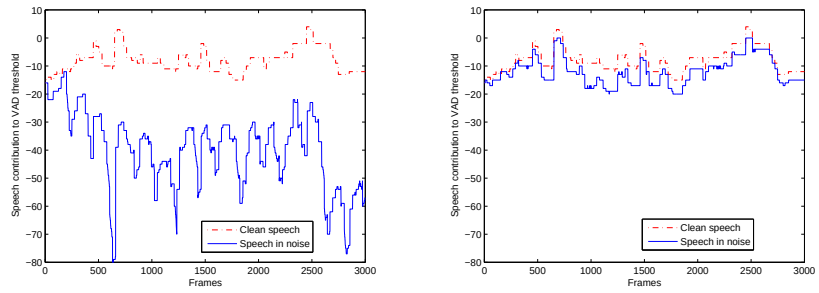
The speech estimate is used in the calculation of the adaptive threshold to raise the threshold if the speech level is high and to lower it if the speech level is low. Since the speech level is expected to correspond to the speech level in the input it is desirable that noise in the input doesn't affect the speech estimation. Figure 16 shows a simulation of the influence of noise in the speech estimation. The SNR for the speech in noise cases are 10 dB and the speech level is -26 dBov. Scaling of the speech level is the speech estimate contribution to the VAD threshold. Note the large variations caused by the babble background noise which is not evident for the car noise or when encoding clean speech.



(a) Background noise levels for all four noise types

(b) Background noise levels for babble and babble 32

Figure 15: Background noise levels as a function of time. The noise input levels are -56 dBov RMS for the first 1000 frames increasing with 10 dB for every 1000 frames up to -26 dBov for frames 3001 - 4000.



(a) Speech level for speech in babble noise.

(b) Speech level for speech in car noise.

Figure 16: Speech levels for clean speech and speech in background noise 10 dB SNR. The speech Level is -26 dBov and the noise types are babble 32 and car.

6.5 Threshold Adaptation

The adaptation of the threshold depends on two things, the noise level and the speech level. Noise and speech levels are calculated as described in section 4.2 and for both levels the lowest band is ignored in the calculation. In Figure 17 one can se how the threshold adapts to different kind of noise inputs. The noise types and input levels are as for the background level in Figure 15 starting at -56 dBov RMS, increasing to -26 dBov RMS in 10 dB steps for each 1000 frames. The plot show that the threshold adapts as intended for the white and car noise inputs, that is a higher noise input generates a lower threshold. For the babble noise inputs on the other hand the behavior is not as expected. The step like decrease and increase for these noise types for the two highest noise levels -36 dBov and -26 dBov arise from the use of noise in the speech level estimation.

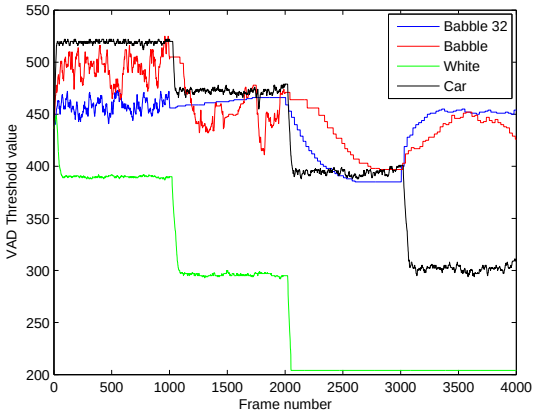


Figure 17: VAD threshold values for four different noise inputs as a function of time.

7 Proposed Modifications

7.1 Background Noise Level Estimate

Investigations of the background noise levels in section 6.3 showed limitations in the noise estimation for high level babble inputs. The possible reason for the halted background noise update is activation of the stationary counter. As described in section 4.2.3 the stationary counter is used to prevent the background estimate from updating upwards. This is done in order to prevent that speech is being used in the background noise estimation. There are three situations that activate the stationary counter:

- A tone have been detected for the 5 last frames.
- The 8 last VAD decisions have been zero.
- The stationary estimate is higher then a threshold.

An investigation of how the stationary counter responded to babble noise inputs was done to establish the reason for the halted update. The simulations showed that the main reason for the halted background noise update was that the stationary estimate *stat_rat* exceeded the threshold. Further investigation indicated that a higher stationarity threshold ($STAT\_THR$) could reduce the problem without any considerable increase in clipping. The suggested improvement is therefore to increase $STAT\_THR$ and based on simulations an increase of 30% is acceptable. In the fixed point implementation [17] this means an increase of the threshold from 1000 to 1300.

7.2 Speech Level Estimate

The speech estimation described in section 4.2.3 states that the speech estimate is updated if *in_level* exceeds a threshold and the VAD has detected speech for long enough time. This generates a problem for the case when the speech is high and the SNR is low. At the end of a speech burst the VAD decision is set to one and an input level higher than the speech threshold is therefore treated as speech. The noise that has a lower energy then the previous speech frames then updates the speech estimate downwards and the lowered speech estimate in turn lowers the VAD threshold. A reduction of the problem is here introduced by a non linearity in the level calculation. The idea is to exclude or suppress sub-band levels under a certain threshold in a way similar to the significance thresholds suggested for use in the SNR estimation. The implementation was done in a straight forward way by a condition in the summation of the sub-band levels described in equation 19. The modification results in

$$in\_level = \sum_{n=2}^{12} \begin{cases} 0 & if \quad level[n] < level\_thr \\ level[n] & otherwise \end{cases} . \tag{30}$$

7.3 Significance Thresholds

As explained in section 4.2.3 the primary voice activity decision is made by comparing a calculated SNR to the adaptive threshold. If the SNR exceeds

the threshold the frame is declared as voiced by the primary detector. The investigation of the AMR-NB coder discussed in section 5 suggested an insertion of a non-linearity in the calculation of the sub-band SNR. The motivation for such a modification was to prevent several moderate SNRs to sum up and make the total SNR exceed the threshold. As discussed previously the function of the WB-VAD and NB-VAD1 are similar, that is the WB-VAD is an extension of the NB-VAD1. For this reason the nonlinearity was implemented in the same way as in the NB-VAD with only minor changes to work with the additional sub-bands. For details of the implementation refer to section 5.1 and equation 25.

7.4 Difference Threshold

When using the difference thresholds it is necessary to run two primary voice activity detectors in parallel. The first one, which for this case also is the one that makes the actual decision, is based on the SNR calculated with the significance thresholds inserted. The second does not use significance thresholds in the summation and corresponds actually to the original primary detector. Note however that the actual decision made by this second detector do not correspond exactly to the original primary VAD since the update of the noise and speech estimates are controlled by the VAD including the significance thresholds. The structure of the VAD algorithm including this additional primary detector is depicted in Figure 18.

In Figure 18 one can also see the inserted difference calculation which is based on differences between the two VAD decisions in long term. The long term activities for the two detectors are based on the filter described in equation 26 with an α value of 0.99. The input to the filter is the short term VAD activity which is the number of active decisions in the last 32 frames. The long term activities thus correspond to time varying voice activity factors for the two different VADs and it is these activity factors that are compared in the final stage. If the difference between the two factors is larger then a threshold for 5 consecutive frames the background noise update speed is set to the reduced one discussed in section 4.2.3.

7.5 Music Hangover

Since music requires an almost continuous encoding not to be degraded in quality it is important not to encode frames with music as noise. The modification discussed next take this into account and adds a hangover in order to reduce the clipping of music. This is important, especially when the significance thresholds are used, since these increases the clipping of non-stationary signals such as babble noise and music.

The idea behind the approach is to add a hangover to the signal if the primary VAD decisions have indicated speech for long enough time. The music hangover is added on top of the primary detector if speech is indicated for 10 consecutive frames. The hangover time added is 200 frames which correspond to 4 seconds for the frame size of 20ms. There are two situations that disable the music hangover:

- If the frame power falls under a low music power threshold for 5 consecutive frames.

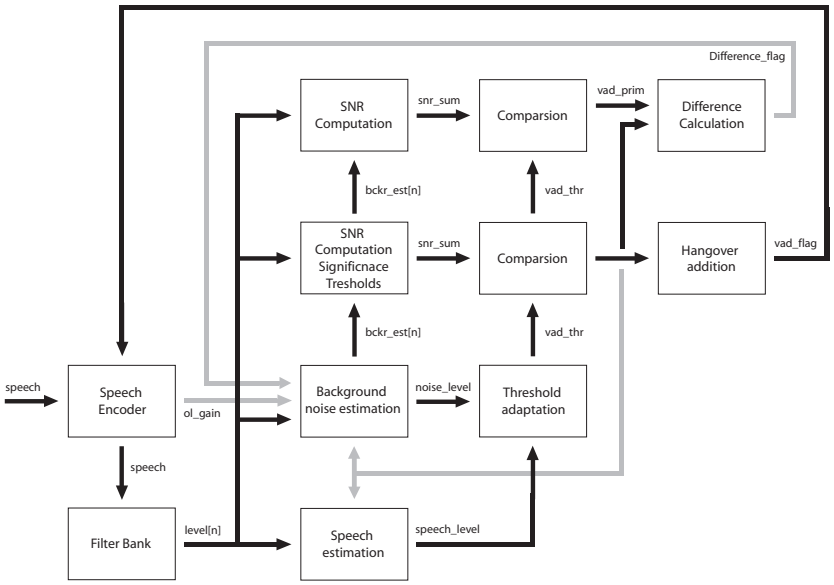


Figure 18: The modified VAD decision algorithm including additional primary dtector and difference calculation.

- If the voice activity factor for the long term activity of the original primary VAD falls under 0.85.

The reason for using the activity of the original primary VAD as a condition is based on the fact that it has a good recognition when exposed to stationary noise types but a tendency to encode non-stationary signals as speech. It therefore releases the music hangover even if a stationary noise generates a level higher than the low music power threshold but will not have a large effect on non-stationary noises or music.

The frame power calculated is actually based on two frames, the current and the previous one. For each frame the power is calculated as described by equation 14 on page 18, then added together and compared to the low music power threshold.

8 Results

8.1 Background Noise Level Estimate

The influence of the increased $STAT\_THR$ value on the background noise updates can be found in Figure 19. A comparison between the plot in Figure 19 and the one for the original $STAT\_THR$ in Figure 15 shows that the raise of the threshold has no noticeable effect on the white and car noise whereas the babble noise inputs are updated for all noise levels for the modified threshold value.

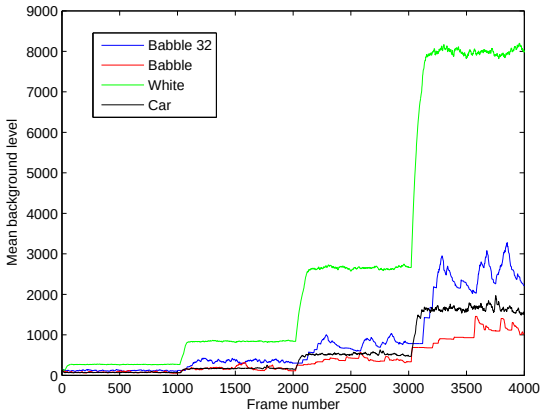


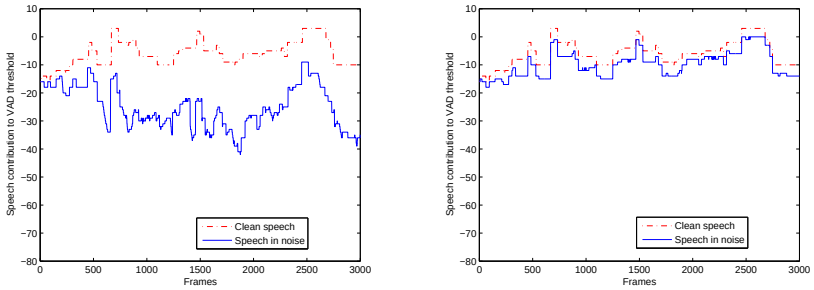
Figure 19: Mean backround noise estimate for the four different noise inputs.

8.2 Speech Level Estimate

In Figure 20 one can see the speech levels VAD-threshold contribution for a threshold value of 3000. The scaling used is the same as in the fixed point implementation in [17]. Table 2 summarizes the mean and standard deviations for the standard and modified speech estimator for clean speech and speech in noise with a SNR of 10 dB. Note the decrease in standard deviation and the increase in mean for the modified version when noise is present. From the table and the figures one can also see that the modification seem to have the largest impact on speech in babble noise and only a small effect on the speech estimate for clean speech and speech in car background noise.

8.3 Significance Thresholds

The purpose with the introduction of the significance thresholds is to reduce the voice activity for speech in non-stationary noise. A reduced voice activity on the other hand increases the possibility to encode speech frames as background noise and may have a large effect on the speech quality. The plots in Figure 21 show how the speech clipping increases with the threshold value for two different speech files and different SNR conditions. In the same figure one can also see the effect on the voice activity caused by the thresholds.



(a) Estimated level for clean speech and speech in babble noise.

(b) Estimated level for clean speech and speech in car noise.

Figure 20: Estimated speech level contribution to the VAD threshold for clean speech and speech in background noise 10dB SNR. The speech Level is -26dBov and the noise types are babble 32 and car.

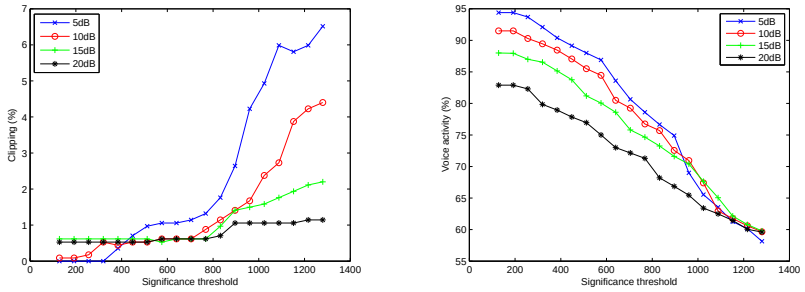
Table 2: Mean and standard deviation for the speech levels in Figure 16 and 20

VAD-type	*file*	*mean*	*standard deviation*
original	clean speech	0.2	2.5
original	speech in babble noise	-40.0	13.6
original	speech in car noise	-12.1	4.9
modified	clean speech	0.4	2.5
modified	speech in babble noise	-25.2	7.6
modified	speech in car noise	-9.8	4.7

Files used in the simulation are speech with a level of -26dBov in babble and car background noises with SNRs of 5, 10, 15 and 20 dB. For both cases the modified $STAT\_THR$ value is used and $sign\_floor = 128$. The clipping is calculated based on the loudness compensated clipping presented in section 6.2. The significance threshold values ranges from 128, which is the same as running the VAD without the significance thresholds, up to 1280, which corresponds to a SNR quotient of 5 since unity for the fixed point implementation is 256 [17]. Note that the thresholds have a much larger impact on voice activity for speech in babble noise than for speech in car noise.

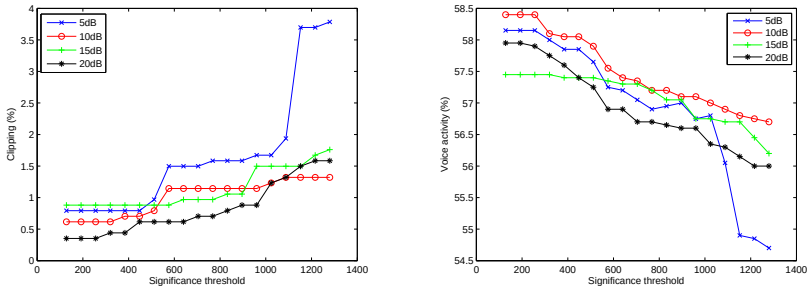
8.4 Difference Threshold

For the difference threshold modification the number of tunable parameters increase and the complexity of choosing the optimum parameters becomes evident. Simulations, however, showed that the parameters that have the largest impact on the voice activity factor and clipping are the actual difference and significance thresholds. To illustrate the operation of the VAD for different settings of these thresholds Figure 22 show the clipping and voice activity factor for a number of settings. The sample used is speech at a level of -26 dBov in babble noise with 10 dB SNR. The sig_floor value is set to 128 and the significance



(a) Speech clipping for speech in babble noise

(b) Voice activity for speech in babble noise



(c) Speech clipping for speech in car noise

(d) Voice activity for speech in car noise

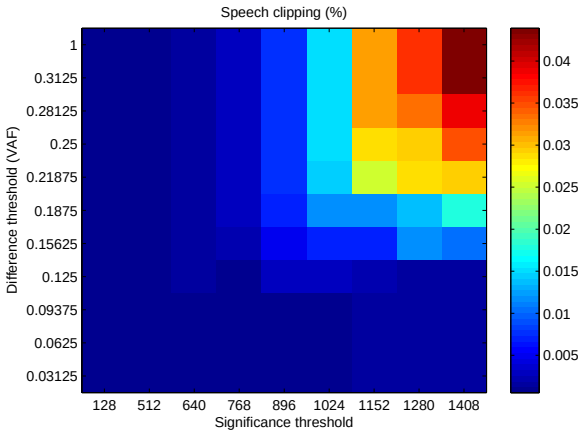
Figure 21: The figures show how clipping and voice activity is affected by different significance threshold values. The plots show the results for a speech level of -26 dBov in car and babble background noise. SNR levels are 5, 10, 15 and 20 dB. The scaling of the thresholds is the same as in the fixed point implementation, unity equal to 256 and $sig\_floor = 128$.

thresholds ranges from 128 up to 1408. Note that the first column in the two plots represents the behavior of the original VAD with no impact either by the significance or the difference thresholds. Also Note that the first row of the plots represents the clipping and voice activity if the difference threshold is equal to 1 which gives the same result as omitting it, the VAD decision is the same as when just using the significance thresholds.

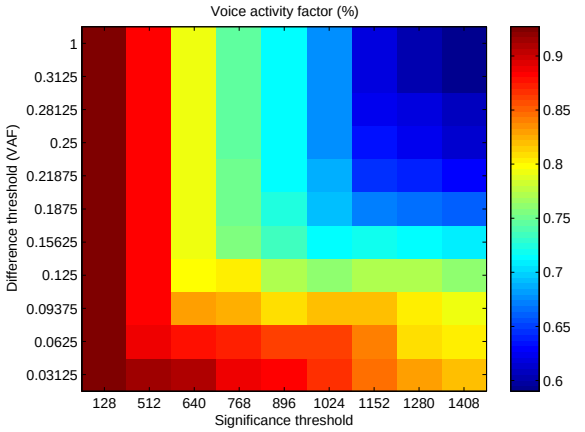
8.5 Music Hangover

Since the significance thresholds deteriorate the VADs capability of handling music in the desired fashion the music hangover becomes more important when these are used. Based on that fact the music hangover was primarily tested and trimmed to work well with the modified versions of the VAD. In Figure 23 one can se how the different VADs respond to music input, the music file is the song Basket case by Green Day scaled so that the input is -26 dBov RMS. From top to bottom the VADs are:

- Original VAD.



(a) Speech clipping.



(b) Voice activity factor.

Figure 22: The voice activity factor and clipping for different settings of the significance and difference thresholds.

- Original VAD with the modified $STAT\_THR$ value.
- VAD with a significance threshold of 640 and a floor of 128.
- VAD with a difference threshold of 0.2, a significance threshold of 960 and a floor of 128.
- Same as the VAD above but with music hangover added.

The effect on the voice activity for the different modifications shows that the introduction of the difference threshold improves the VADs handling of music when significance thresholds are used. It also indicate that the higher $STAT\_THR$ value seem to have no effect on encoding of music and that the music hangover have the desired effect.

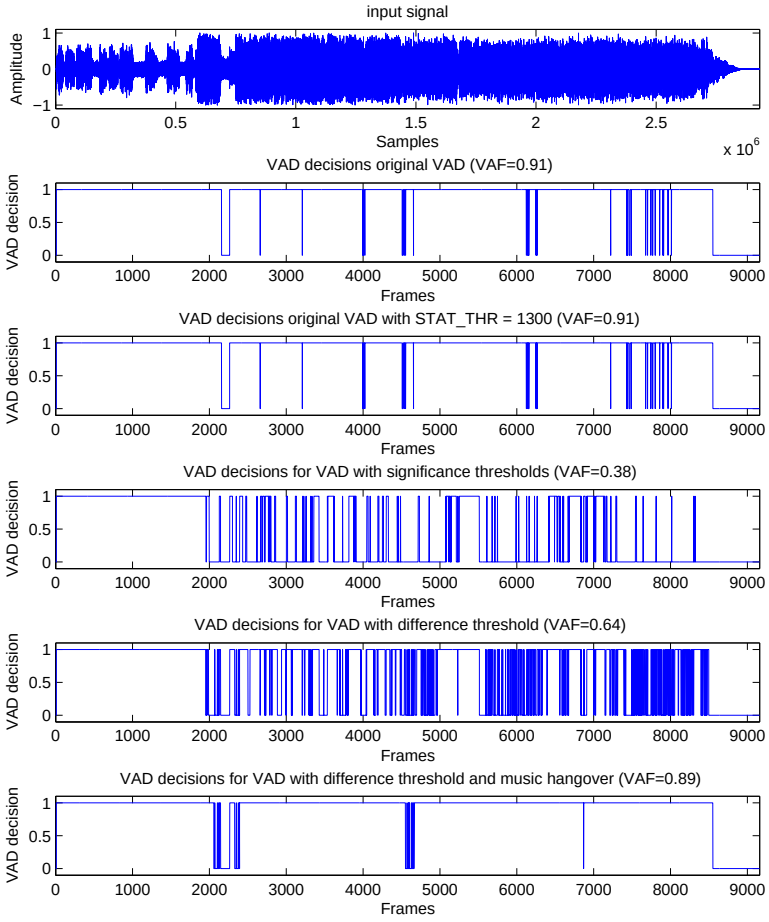


Figure 23: The response for different VADs when encoding music. The file encoded is the song Basket case by Green Day with a level normalized to -26dBov RMS. The VADs are from top to bottom the original VAD, the original VAD with the modified $STAT\_THR$ value, the VAD with a significance threshold of 640 and a floor of 128, the VAD with a difference threshold of 0.2, a significance threshold of 960 and a floor of 128, and the last one is the same as the VAD above but with music hangover added.

8.6 Final Results

This section summarizes the modifications made and their effect with regard to the voice activity factor. In Figure 24 the voice activity factors for four different modifications are shown along with the activity factor for the original VAD. For this case the samples are speech in babble noise at three different levels -16,

-26 and -36 dBov with four different SNR conditions 10, 20, 30 dB and infinity. Figure 25 show the same thing but for speech in car noise. The different VAD modifications are summarized here in order of appearance:

Orig_VAD The original VAD with no modifications

STAT_THR The original VAD but with $STAT\_THR = 1300$

SIG_THR VAD with $STAT\_THR = 1300$, significance threshold set to 640 and floor set to 128.

DIF_THR VAD with a difference threshold included and set to 0.2. $STAT\_THR = 1300$ and the significance threshold set to 960 and a floor set to 128.

Speech_THR VAD with $STAT\_THR = 1300$ and a speech level threshold set to 3000.

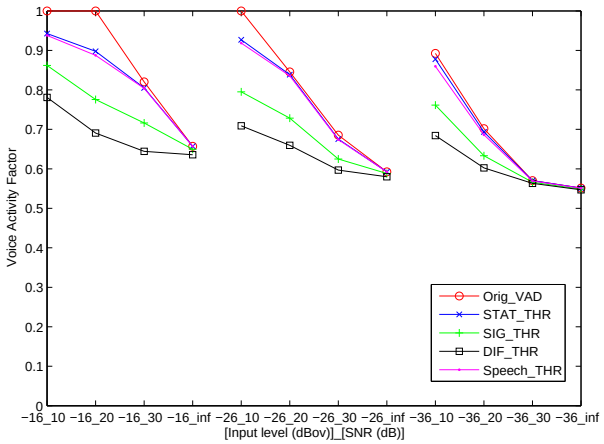


Figure 24: The voice activity factor for speech in babble background noise for standard VAD and modified VADs.

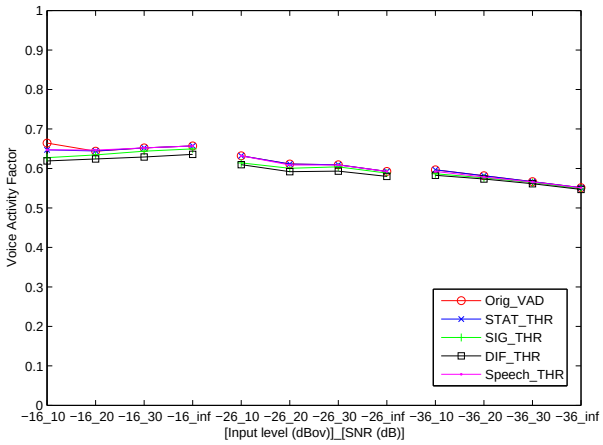


Figure 25: The voice activity factor for speech in car background noise for standard VAD and modified VADs.

9 Discussion

9.1 Stationarity Threshold

As was shown in the results the increase of the stationarity threshold solved the problem with the background noise update for high level babble noise inputs. The solution is simple and yields no increase in computational complexity since it only demands a change of a threshold value. During the investigation the increased threshold value also proved to be crucial for high noise levels in the significance threshold implementations. The simulations did not show any obvious disadvantages with the higher threshold value and it was therefore tested and evaluated in conjunction with the other modifications with good results.

9.2 Speech Level Estimate

The insertion of the non-linearity in the speech level estimation did to some extent resolve the problem with noise disturbing the update. As indicated by the result presented in section 8.2 the large variations caused by babble noise were suppressed in the presence of the threshold. The modification did however not show any significant improvement in voice activity for the tested samples. The reason for the small gain is that the speech estimate is a small contribution to the overall VAD threshold.

9.3 Significance Thresholds

The significance thresholds lead to a reduced voice activity but come at the expense of a higher clipping of speech. This results in a compromise when choosing the thresholds as could be seen in Figure 21. The choices for the thresholds were made such that a low clipping was considered more important than a reduced voice activity. As was mentioned earlier the higher stationarity threshold value proved to be crucial in conjunction with the significance thresholds. This is due to the fact that a low background estimate will generate high sub-band SNRs that may not correspond to the actual input. Since the background noise estimate is not updated properly for high level babble noise inputs the SNRs are estimated to high with the lower stationarity threshold. The consequence of this is that the SNRs never fall in the range of the significance thresholds and the only way to overcome that problem is by thresholding so hard that an unacceptable amount of clipping occurs. One obvious drawback associated with the modification is the decreased activity when encoding music that could be seen in Figure 23.

A possible solution for the thresholds that offers a good compromise between lowered activity and speech clipping is achieved by setting the floor to 128 and the threshold to 640 when the higher stationarity threshold is used.

9.4 Difference Threshold

Introduction of the difference thresholds significantly improves the handling of music if the same voice activity factors that could be achieved with the significance thresholds alone are considered. The modification also tends to generate a lower clipping of speech at the same activity rates and therefore

yields a better compromise between clipping and activity. Since the algorithm need to be modified and new processing functionality is inserted the complexity of the VAD increases with the insertion of the difference threshold. This could be a problem if the computational resources are limited.

A good compromise between clipping and reduction of voice activity for babble noises is achieved by setting the difference threshold to 0.2, using a significance threshold value of 960 with a floor of 128 and $STAT\_THR = 1300$.

9.5 Music Hangover

In use with the difference and significance threshold modifications the music hangover shows a clear improvement when encoding music. The modification can also be used on the original VAD but in this case high level stationary noises can generate problems if no calculation of the long term activity is added. As the decision whether to remove the hangover or not is partially based on the level of the signal the method is not effective if the music level is too low. For the suggested parameters the problem becomes evident if the level falls under approximately -30 dBov.

10 Conclusions

In this thesis several approaches to improve the handling of complex background noise types in the WB-VAD have been tested, evaluated and discussed. The modifications proposed for the NB-VAD that were the start of the investigation did show good results also for the WB-VAD with minor changes. It is important, however, to realize that the modifications involving the significance thresholds only performs well in conjunction with the increased stationarity threshold.

Concerning the encoding of music both the significance and difference threshold modifications have a negative effect and should be used with the music hangover to ensure good results. The addition of difference threshold increases the activity for music encoding but the performance is still worse than the original VAD.

The speech level modification introduced is simple but do not show any significant improvements in voice activity and is therefore not to be considered as a solution.

For all of the suggested solutions the lower activity comes at the expense of an increase in clipping of speech. For the higher stationarity threshold the clipping can be neglected since it is small. In the other solutions a compromise between voice activity and clipping is necessary. The solutions suggested in this report are tunable with one endpoint being the original VAD. The results presented and discussed offer some guidance in the choice of the significance and difference thresholds. Note however that the subjective quality is not considered and listening test evaluating the perceived quality should be performed to ensure that the clipping is not audible.

All modifications made showed only small degradation in the handling of white and car noise input when the threshold settings were kept reasonable. The music hangover did initially cause an increase in activity for the stationary noises. This was however handled by the condition introduced in the release of the hangover based on the activity of the original primary VAD detector.

Based on the results presented in this report an improvement of the VAD should include the higher stationarity threshold together with the difference threshold and the music hangover. A simple and low complexity solution that did not show any obvious disadvantages would be to only increase the value of the stationarity threshold.

11 Further Studies

Since the evaluations made in this report are based only on a selection of samples and no formal subjective testing have been performed there are still some issues that need to be resolved before a final conclusion of the proposed improvements can be drawn. The results, however, have shown some guidance and here are some ideas for further studies.

- This report considers only the activity indicated by the VAD and does not take the hangover added by the DTX/SCR system in to account. This should not have a large effect on the performance but should be investigated to ensure that so is the case.
- Encoding of speech, noise and music should be done on larger material and on signals captured in real life scenarios.
- There have not been any discussion on adaptation of the thresholds in this report. This could be investigated since there seem to be possibilities to adapt the thresholds depending on the stationarity and the levels of the background noise.
- Results of the objective tests need to be confirmed with formal subjective testing to ensure that the clipping that occurs does not have a significant effect on the perceived quality.

References

[1] A. W. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Prentice-Hall, 1999.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[3] R. Goldberg and L. Riek, *A practical handbook of speech coders*. CRC Press LCC, 2000.

[4] S. Khalid, *Introduction to Data Compression*, 2nd ed. Morgan Kaufmann Publishers, 2000.

[5] H. Fastl and E. Zwiker, *Psychoacoustics: facts and models*. Springer, 1999.

[6] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," in *Processings of the IEEE*, vol. 88, no. 4, April 2000, pp. 451–513.

[7] B. Edwards, "Application of psychoacoustics to audio signal processing," in *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, vol. 1, November 2001, pp. 814–818.

[8] A. S. Spanias, "Speech coding: A tutorial review," in *Procedings of the IEEE*, vol. 82, no. 10, October 1994, pp. 1541–1582.

[9] H. Dudely, "The vocoder," Bell Labs Record, Tech. Rep., 1939.

[10] J. Makhoul, "Linear Prediction: A tutorial review," in *Procedings of the IEEE*, vol. 63, no. 4, April 1975, pp. 561–580.

[11] R. V. Prasad *et al.*, "Comparison of Voice Activity Detection Algorithms for VoIP," in *ISCC 2002. Seventh International Symposium on Computers and Communications*, 2002, pp. 530–535.

[12] F. Beritelli *et al.*, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," in *Signal processing letters, IEEE*, vol. 9, no. 3, Mars 2002, pp. 85–88.

[13] R. Salami *et al.*, "The adaptive multirate-rate wideband speech codec (AMR-WB)," in *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, November 1980, pp. 620–636.

[14] 3GPP, "Adaptive Multi-Rate Wideband (AMR-WB) speech codec; General description," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.171 (v7.0.0), 2005-07. [Online]. Available: http://www.3gpp.org

[15] ——, "Adaptive Multi-Rate Wideband (AMR-WB) speech codec; Transcoding functions," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.190 (v6.1.1), 2005-07. [Online]. Available: http://www.3gpp.org

[16] ——, "Adaptive Multi-Rate Wideband (AMR-WB) speech codec; Voice Activity Detector (VAD)," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.194 (v6.0.0), 2004-12. [Online]. Available: http://www.3gpp.org

[17] ——, "Adaptive Multi-Rate (AMR) speech codec; ansi-c code," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.173, 2006-06. [Online]. Available: http://www.3gpp.org

[18] ——, "Adaptive Multi-Rate Wideband (AMR-WB) speech codec; Sourece controlled rate operation," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.193 (v6.1.0), 2006-06. [Online]. Available: http://www.3gpp.org

[19] ——, "Adaptive Multi-Rate Wideband (AMR-WB) speech codec; Comfort noise aspects," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.192 (v6.0.0), 2004-12. [Online]. Available: http://www.3gpp.org

[20] Martin Sehlstedt, "VAD Improvement report," ERICSSON, Tech. Rep. EAB-06:003186 Uen, 2006-12.

[21] 3GPP, "Adaptive Multi-Rate (AMR) speech codec; Voice Activity Detector (VAD)," 3^{rd} Generation Partnership Project, Tech. Rep. TS 26.094 (v6.1.0), 2006-06. [Online]. Available: http://www.3gpp.org

[22] M. R. Schroder *et al.*, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustic Society of America*, vol. 66, pp. 1647–1652, 1979.