ATIS-0100005.2006

AUDITORY NON-INTRUSIVE QUALITY ESTIMATION PLUS (ANIQUE+): PERCEPTUAL MODEL FOR NON-INTRUSIVE ESTIMATION OF NARROW-BAND SPEECH QUALITY

AMERICAN NATIONAL STANDARD FOR TELECOMMUNICATIONS

The Alliance for Telecommunication Industry Solutions (ATIS) is a technical planning and standards development organization that is committed to rapidly developing and promoting technical and operations standards for the communications and related information technologies industry worldwide using a pragmatic, flexible and open approach. Over 1,100 participants from over 300 communications companies are active in ATIS' 22 industry committees and its Incubator Solutions Program.

< http://www.atis.org/ >

AMERICAN NATIONAL STANDARD

Approval of an American National Standard requires review by ANSI that the requirements for due process, consensus, and other criteria for approval have been met by the standards developer.

Consensus is established when, in the judgment of the ANSI Board of Standards Review, substantial agreement has been reached by directly and materially affected interests. Substantial agreement means much more than a simple majority, but not necessarily unanimity. Consensus requires that all views and objections be considered, and that a concerted effort be made towards their resolution.

The use of American National Standards is completely voluntary; their existence does not in any respect preclude anyone, whether he has approved the standards or not, from manufacturing, marketing, purchasing, or using products, processes, or procedures not conforming to the standards.

The American National Standards Institute does not develop standards and will in no circumstances give an interpretation of any American National Standard. Moreover, no person shall have the right or authority to issue an interpretation of an American National Standard in the name of the American National Standards Institute. Requests for interpretations should be addressed to the secretariat or sponsor whose name appears on the title page of this standard.

CAUTION NOTICE: This American National Standard may be revised or withdrawn at any time. The procedures of the American National Standards Institute require that action be taken periodically to reaffirm, revise, or withdraw this standard. Purchasers of American National Standards may receive current information on all standards by calling or writing the American National Standards Institute.

Notice of Disclaimer & Limitation of Liability

The information provided in this document is directed solely to professionals who have the appropriate degree of experience to understand and interpret its contents in accordance with generally accepted engineering or other professional standards and applicable regulations. No recommendation as to products or vendors is made or should be implied.

NO REPRESENTATION OR WARRANTY IS MADE THAT THE INFORMATION IS TECHNICALLY ACCURATE OR SUFFICIENT OR CONFORMS TO ANY STATUTE, GOVERNMENTAL RULE OR REGULATION, AND FURTHER, NO REPRESENTATION OR WARRANTY IS MADE OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR AGAINST INFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS. ATIS SHALL NOT BE LIABLE, BEYOND THE AMOUNT OF ANY SUM RECEIVED IN PAYMENT BY ATIS FOR THIS DOCUMENT, WITH RESPECT TO ANY CLAIM, AND IN NO EVENT SHALL ATIS BE LIABLE FOR LOST PROFITS OR OTHER INCIDENTAL OR CONSEQUENTIAL DAMAGES. ATIS EXPRESSLY ADVISES ANY AND ALL USE OF OR RELIANCE UPON THIS INFORMATION PROVIDED IN THIS DOCUMENT IS AT THE RISK OF THE USER.

NOTE - The user's attention is called to the possibility that compliance with this standard may require use of an invention covered by patent rights. By publication of this standard, no position is taken with respect to the validity of this claim or any patent rights in connection therewith. The patent holder has, however, filed a statement of willingness to grant license under these rights on reasonable and nondiscriminatory terms and conditions to applicants desiring to obtain such a license. Details may be obtained from the publisher.

ATIS-0100005.2006, *Auditory Non-Intrusive QUality Estimation Plus (ANIQUE+): Perceptual Model for Non-Intrusive Estimation of Narrow-Band Speech Quality*

Is an American National Standard developed by the **Quality of Service (QoS)** Task Force under the **ATIS Network Performance, Reliability, and Quality of Service Committee (PRQC)**.

Published by

Alliance for Telecommunications Industry Solutions
1200 G Street, NW, Suite 500
Washington, DC 20005

ATIS-0100005.2006

American National Standard for Telecommunications

AUDITORY NON-INTRUSIVE QUALITY ESTIMATION PLUS (ANIQUE+): PERCEPTUAL MODEL FOR NON-INTRUSIVE ESTIMATION OF NARROW-BAND SPEECH QUALITY

Secretariat

Alliance for Telecommunications Industry Solutions

Approved November 2, 2006

American National Standards Institute, Inc.

Abstract

This standard describes a perceptual objective model for non-intrusive estimation of narrow-band speech quality. This standard provides the description of the perceptual objective model, Auditory Non-Intrusive QUality Estimation Plus (ANIQUE+), which estimates the quality of speech without reference speech information.

NOTE - Annex B, *ANSI-C Reference Implementation of ANIQUE+*, of this Standard has also been formatted as a separate plain text file and electronically packaged with this standard.

FOREWORD

The information contained in this Foreword is not part of this American National Standard (ANS) and has not been processed in accordance with ANSI's requirements for an ANS. As such, this Foreword may contain material that has not been subjected to public review or a consensus process. In addition, it does not contain requirements necessary for conformance to the Standard.

Considering that the nature of speech quality is a subjective sensation by human listeners, the most reliable way to measure it is to perform a subjective listening test. Historically, formal subjective listening tests (e.g., Mean Opinion Score (MOS) tests) have been used in evaluating the performance of speech processing and transmission systems such as speech coders. These tests require well-controlled facilities and expertise to enable quality variations to be distinguished reliably from other effects and provide reliable results related only to the quality. Throughout the development of network systems and their deployment, it is highly necessary to investigate the impact of specific system components, combinations of them, and sets of system parameter values on the perceived quality of speech. Since it is difficult to obtain these results promptly and constantly by subjective tests, which are very expensive both in time and cost, it is desirable to have a computational objective model which can reflect subjective quality ratings in reliable manner.

The need for objective models is becoming more prominent, especially in complex modern telecommunication network environments. In addition to existing traditional public switched telephone networks (PSTN), various types of mobile and internet-based networks are being widely used. The resulting interconnected heterogeneous network increases significantly the number of factors that can affect speech quality, and the understanding of the impact of individual system components and combinations of them on the end-to-end speech quality is very difficult. In addition, these newer networks provide a tradeoff between service quality and cost. Thus, the reliable estimation of speech quality over modern telecommunication networks is very important not only for network systems design and development but also for the maintenance of quality of service (QoS).

The need for objective measures for predicting speech quality has long been a goal in the industry and standard bodies such as the American National Standard Institute (ANSI) and the International Telecommunication Union Telecommunication Sector (ITU-T). To satisfy the need, some algorithms were adopted as standard recommendations: ITU-T P.861, P.862, and ANSI T1.518-1998 (R2008). These are intrusive models in which the source speech signal applied as an input to the network under test should be available in order to estimate the quality of speech signal passed through the network. Considering the time-varying nature of modern wireless and internet-based networks, these intrusive models cannot be used for monitoring the quality of individual calls of in-service live networks, where no source speech signals uttered by end users are available. In 2004, the ITU-T adopted Recommendation P.563 for non-intrusive estimation of speech quality as the need for monitoring the speech quality of in-service networks is growing.

This American National Standard (ANS) presents a method an algorithm to estimate the quality of speech in non-intrusive manner -- i.e., without the reference speech information used as an input to the network under test. This standard offers improved performance, approaching that of traditional intrusive methods, and can be used for monitoring the speech quality of in-service live networks, where the reference speech signals from end-users are not available.

The Alliance for Telecommunication Industry Solutions (ATIS) serves the public through improved understanding between carriers, customers, and manufacturers. The Alliance for Telecommunication Industry Solutions (ATIS) serves the public through improved understanding between carriers, customers, and manufacturers. The Network Performance, Reliability, and Quality of Service Committee (PRQC) -- formerly T1A1 -- develops and recommends standards, requirements, and technical reports related to the performance, reliability, and associated security aspects of communications networks, as well as the processing of voice, audio, data, image, and video signals, and their multimedia integration. PRQC also develops and recommends positions on, and foster consistency with, standards and related subjects under consideration in other North American and international standards bodies.

ANSI guidelines specify two categories of requirements: mandatory and recommendation. The mandatory requirements are designated by the word *shall* and recommendations by the word *should*. Where both a mandatory requirement and a recommendation are specified for the same criterion, the recommendation represents a goal currently identifiable as having distinct compatibility or performance advantages.

Suggestions for improvement of this document are welcome. They should be sent to the Alliance for Telecommunications Industry Solutions, Network Performance, Reliability, and Quality of Service Committee, Secretariat, 1200 G Street NW, Suite 500, Washington, DC 20005.

At the time it approved this document, Network Performance, Reliability, and Quality of Service Committee, which is responsible for the development of this Standard, had the following members:

M. Neibert, PRQC Chair
N. Seitz, PRQC Vice-Chair
C. Underkoffler, ATIS Chief Editor
D. Kim, PRQC Technical Editor

Organization Represented	Name of Representative
Alcatel USA Inc.	Ken Biholar
AT&T	Percy Tarapore Charles A. Dvorak (Alt.)
BellSouth Telecommunications	Archie McCain
C.S.I Telecommunications	Michael S. Newman Thomas G. Croda (Alt.)
Department of Defense	Chris Fitzgerald
Ericsson Incorporated	Mustafa Kocaturk Susana Sabater-Maroto (Alt.)
Harris Corporation	Marlis Humphrey
Intelsat	Mark T. Neibert
Lucent Technologies	Stuart O. Goldman
National Communications System	An Nguyen Carol-Lyn Taylor (Alt.)
NTIA	Neal B. Seitz

Organization Represented	Name of Representative
	Arthur Webster (Alt).
Nortel Networks	Joseph A . Zebarth
Qwest	Steve Showell Michael Fargano (Alt.)
Siemens Communications, Inc.	Suhas S. Gandhi David E. Francisco (Alt.)
Sprint LTD	Jack Moonigngam
Sprint Nextel	Mark L. Jones
Telcordia Technologies	Spilios Makris Cliff Halevi (Alt.)
Tellabs Operations, Inc.	William A. Walker Kevin Stodola (Alt.)
VeriSign, Inc.	Anthony M. Rutkowski
Verizon Communications	John Colombo Wendy Pugh (Alt.)

The Quality of Service (QoS) Task Force was responsible for the development of this document.

TABLE OF CONTENTS

1 SCOPE, PURPOSE, & APPLICATION 1

1.1 SCOPE 1
1.2 PURPOSE 1
1.3 APPLICATION 1

2 NORMATIVE REFERENCES 2

3 DEFINITIONS, ACRONYMS, & ABBREVIATIONS 3

3.1 DEFINITIONS 3
3.2 ACRONYMS & ABBREVIATIONS 3

4 CONVENTIONS 3

5 REQUIREMENTS ON SPEECH SIGNALS 3

6 STRUCTURE OF AUDITORY NON-INTRUSIVE QUALITY ESTIMATION PLUS (ANIQUE+) MODEL 4

7 LEVEL NORMALIZATION AND RECEIVE-SIDE MODIFIED IRS FILTERING 5

8 ARTICULATION ANALYSIS 5

8.1 CRITICAL-BAND FILTERBANK 6
8.2 MODULATION SPECTRUM ANALYSIS 6
8.3 FEATURE EXTRACTION 8

9 DETECTION OF ACTIVE SPEECH AND AUDIBLE BACKGROUND NOISE 9

10 FRAME DISTORTION MODEL 10

10.1 OVERVIEW 10
10.2 MULTI-LAYER PERCEPTRON MODEL 12

11 MUTE MODEL 14

11.1 ACTIVITY PROFILE ANALYSIS 14
11.2 MUTE DETECTION 15
11.2.1 Unnatural Abrupt Stops 15
11.2.2 Unnatural Abrupt Starts 17
11.3 MUTE IMPACT MODEL 18

12 NONSPEECH MODEL 18

A BIBLIOGRAPHY 19

B ANSI-C REFERENCE IMPLEMENTATION OF ANIQUE+ 20

TABLE OF FIGURES

FIGURE 1 - OVERVIEW OF ANIQUE+ MODEL 4
FIGURE 2 - PROCESSING OF ARTICULATION ANALYSIS 5
FIGURE 3 - FREQUENCY RESPONSE OF MODULATION FILTERBANK 8
FIGURE 4 - DETECTION OF ACTIVE SPEECH AND AUDIBLE BACKGROUND NOISE 10
FIGURE 5 - FRAME DISTORTION ESTIMATION MODEL 12
FIGURE 6 - MULTI-LAYER PERCEPTRON FOR FRAME DISTORTION MODEL 13
FIGURE 7 - OVERVIEW OF ACTIVITY PROFILE ANALYSIS 14
FIGURE 8 - ILLUSTRATION OF ACTIVITY PROFILE ANALYSIS FOR MUTE DETECTION 15
FIGURE 9 - MFCC FILTERBANK ALONG THE FFT BINS 17

TABLE OF TABLES

TABLE 1 - RELATIONSHIP OF CODING TECHNOLOGIES, EXPERIMENTAL FACTORS AND APPLICATIONS TO THIS STANDARD 2
TABLE 2 - CHARACTERISTIC FREQUENCY AND BANDWIDTH OF CRITICAL-BAND FILTERS 6
TABLE 3 - FILTER COEFFICIENTS FOR HILBERT TRANSFORM 7

American National Standard for Telecommunications –

Auditory Non-Intrusive QUality Estimation Plus (ANIQUE+): Perceptual Model for Non-Intrusive Estimation of Narrow-Band Speech Quality

1 SCOPE, PURPOSE, & APPLICATION

1.1 Scope

The objective speech quality estimated by the Auditory Non-Intrusive QUality Estimation Plus (ANIQUE+) model of this American National Standard (ANS) is the subjective quality of telephone band speech.

The quality estimated or predicted by the ANIQUE+ model is not conversational quality but listening-only quality, which can be originally obtained by subjective auditory tests investigating listening quality in Absolute Category Rating (ACR) scale using a common receiving handset at a standard listening level of 79 dB SPL (See ITU-T P.800 and P.830). This ANS can be used to predict the quality of one-way speech transmission.

1.2 Purpose

The need for objective models for estimating speech quality is becoming more prominent, especially in complex modern telecommunication network environments. In addition to existing traditional public switched telephone networks (PSTN), various types of mobile and internet-based networks are being widely used. The resulting interconnected heterogeneous network increases significantly the number of factors that can affect speech quality, and the understanding of the impact of individual system components and combinations of them on the end-to-end speech quality is very difficult. In addition, these newer networks provide a tradeoff between service quality and cost. Thus, the reliable estimation of speech quality over modern telecommunication networks is very important not only for network systems design and development, but also for the maintenance of quality of service (QoS).

This ANS defines an algorithm which provides acceptable accuracy in estimating the quality of speech processed by telecommunication networks in a non-intrusive manner. For 20 Mean Opinion Scope (MOS) test databases which have never been used in the development of objective models, the average correlation between subjective and objective quality scores is about 0.87 for this ANS, whereas the ITU-T P.563 shows about 0.81 correlation for the same task.

1.3 Application

Table 1 presents a guide to facilitate the determination of the test factors, coding technologies, and applications to which this ANS applies. It includes conditions for which this ANS has demonstrated acceptable accuracy, has demonstrated inaccurate estimation, and has not been validated in full.

Table 1 - Relationship of coding technologies, experimental factors and applications to this standard

Test Factors	Note
Speech input level to a codec	1
Transcoding	1
Transmission channel errors	1
Packet loss and its concealment with CELP codecs	1
Bit rates if a codec has more than one bit rate mode	1
Environmental noise in the sending side	1
Temporal clipping of speech	1
Short-term time warping of audio signal	1
Long-term time warping of audio signal	1
Effect of time-varying delay	1
Characteristics of acoustical interface at the sending side (different handsets and their positions)	1
Lombard effect *	1
Listening levels in subjective experiments	2
Singing voice	2
Music as input to a codec	2
Network information signals as input to a codec	2
Delay in conversational tests	2
Amplitude clipping of speech	3
Bit-rate mismatch between an encoder and a decoder if a codec has multiple bit rates	3
Talker dependencies	3
Multiple simultaneous talkers	3
Coding technologies	
Waveform coders, e.g., G.711, G.726	1
CELP and hybrids ≥ 4 kbit/s, e.g., G.728, G.729, G.723.1, ACELP, VSELP, EVRC, SMV	1
Other codecs, e.g., GSM-FR, GSM-EFR, GSM-HR, AMR	1
CELP and hybrids ≤ 4 kbit/s	3
Applications	
Live monitoring of in-service networks	1
Live network testing without known reference speech	1
NOTES 1. The objective measure has demonstrated acceptable accuracy for this factor. 2. The objective measure is known to provide inaccurate estimation when used in conjunction with this factor, or is otherwise not intended to be used with this factor. 3. The objective measure has not been fully validated for this factor.	

2 NORMATIVE REFERENCES

The following standards contain provisions which, through reference in this text, constitute provisions of this American National Standard. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties to agreements based on this American National Standard are encouraged to investigate the possibility of applying the most recent editions of the standards indicated below.

ITU-T P.56 (1993), *Objective measurement of active speech level*.[1]

* The Lombard effect is the tendency to increase one's vocal intensity in noise.

[1] This document is available from the International Telecommunications Union. < http://www.itu.int/ITU-T/ >

ITU-T P.830 (1996), *Subjective performance assessment of telephone telephone-band and wideband digital codecs*.[1]

3 DEFINITIONS, ACRONYMS, & ABBREVIATIONS

3.1 Definitions

3.1.1 critical band: A frequency range in psychoacoustic experiments for which perception abruptly changes as a narrowband sound stimulus is modified to have frequency components beyond the band.

3.1.2 pre-emphasis: Filtering process in which the frequency response of the filter has emphasis at a given frequency range.

3.2 Acronyms & Abbreviations

For the purpose of this ANS, the following abbreviations are used:

ACR	Absolute Category Rating
ANIQUE+	Auditory Non-Intrusive QUality Estimation Plus
ANSI	American National Standards Institute
ATIS	Alliance for Telecommunications Industry Solutions
CELP	Code Excited Linear Prediction
DBov	Decibel to overload point
DC	Direct Current
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
MOS	Mean Opinion Score
PCM	Pulse Code Modulation
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RXMIRS	Receive-side Modified Intermediate Reference System
SPL	Sound Pressure Level
VoIP	Voice over Internet Protocol

4 CONVENTIONS

Subjective evaluation of speech quality of telephone networks and codecs may be conducted using listening-only or conversational methods of subjective testing. For practical reasons, listening-only tests are the only feasible method of subjective testing during the development of speech codecs, when a real-time implementation of the codec is not available. Also, listening-only tests are often not practical for live network monitoring. This ANS defines an objective measurement technique for estimating subjective quality on the MOS scale obtained in listening-only ACR tests. It should be noted that the maximum value of objective measurement by this ANS is limited to 4.5.

5 REQUIREMENTS ON SPEECH SIGNALS

This ANS is intended for estimating the quality of human-talking speech only, and has not been validated for estimating the quality of non-speech signals such as singing voice, music, noise, and artificial speech.

The format of speech file to be evaluated by this ANS should be 16 bit linear PCM and the sampling frequency should be 8 kHz. The file should contain at least single speech burst preceded by at least 100 msec of silence or background noise. The recommended minimum and maximum lengths of signal are 3 and 20 seconds, respectively.

6 STRUCTURE OF AUDITORY NON-INTRUSIVE QUALITY ESTIMATION PLUS (ANIQUE+) MODEL

Figure 1 shows the overall block diagram of the ANIQUE+ model.

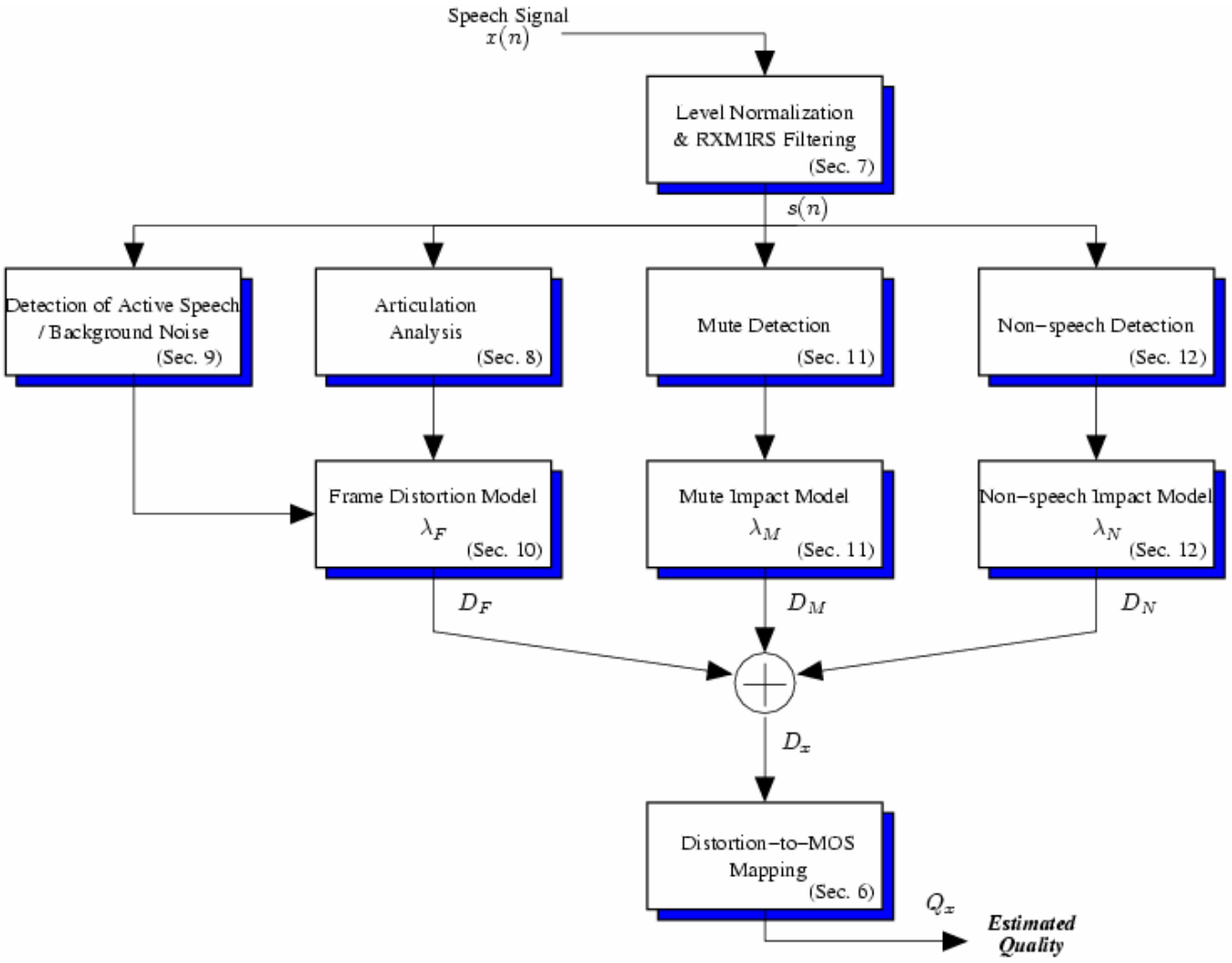

Figure 1 - Overview of ANIQUE+ model

The overall objective distortion D_x ($0 \le D_x \le 1$) of the PCM speech signal $x(n)$ is estimated by the sum of overall frame distortion D_F, mute distortion D_M and non-speech distortion D_N:

$$D_x = \min\{D_F + D_M + D_N, 1.0\}.$$

The frame distortion block decomposes the incoming speech signal into successive time frames (64 msec interval) and estimates perceptual distortion of individual frames to derive the overall frame

distortion. The mute distortion block detects unnatural mutes in speech signals and models their impact on perceived distortion. The non-speech distortion block detects the presence of very annoying non-speech activity and models its impact. Each objective distortion (D_F, D_M, and D_N,) takes the value between 0 and 1.

Finally, the distortion D_x is then mapped onto subjective MOS scale to yield objective speech quality Q_x:

$$Q_x = -3.5\, D_x + 4.5$$

assuming the maximum and minimum value of quality are 4.5 and 1.0, respectively. Thus, the range of objective quality estimated by this ANS is between 1 (lowest quality) and 4.5 (highest quality).

Each functional block in Figure 1 includes a section number in which the algorithmic description is provided.

7 Level Normalization and Receive-Side Modified IRS Filtering

The level of speech signal $x(n)$ is first normalized to -26 dBov using the ITU-T P.56 speech voltmeter. Then the receive-side modified intermediate reference system (RXMIRS) receive filter is applied to reflect the frequency characteristics of the handset used in subjective listening tests, resulting in the preprocessed speech signal $s(n)$.

8 Articulation Analysis

In the articulation analysis, the preprocessed speech signal $s(n)$ is analyzed by functional blocks motivated by human auditory systems at peripheral and central levels, and perceptual feature vectors are extracted to be used in the frame distortion model. Figure 2 shows the processing of articulation analysis.

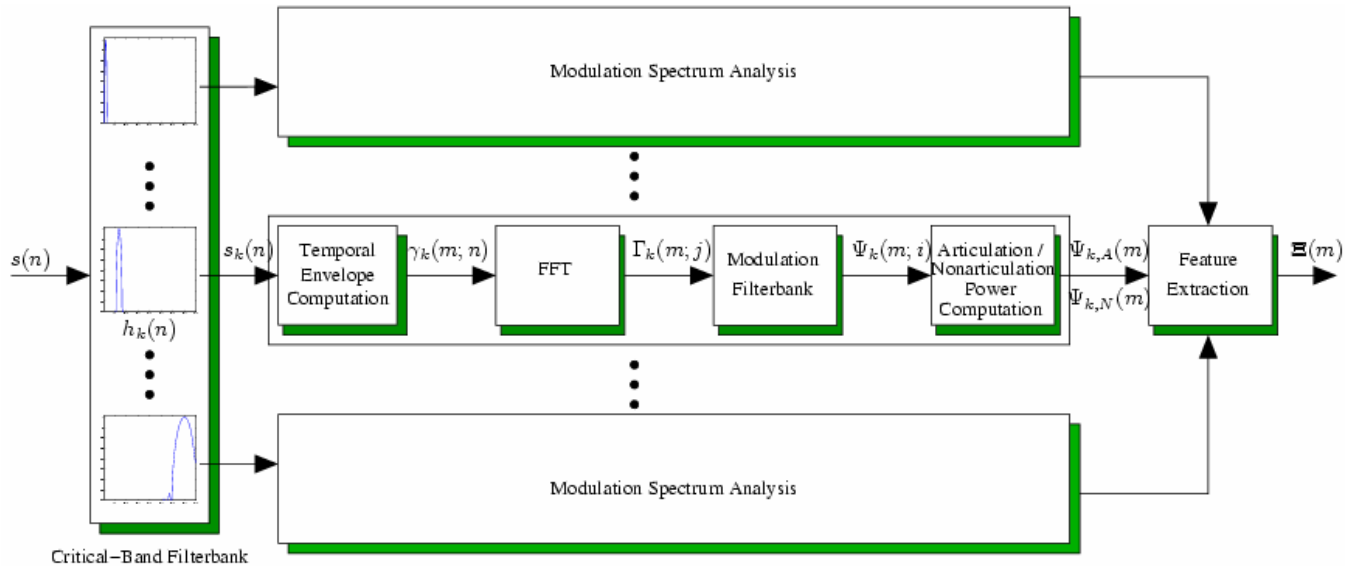

Figure 2 - Processing of articulation analysis

8.1 Critical-Band Filterbank

Simulating the peripheral stage of human auditory system, the level-normalized and RXMIRS-filtered speech signal, $s(n)$, is filtered by a bank of critical-band filters, $h_k(n)$, k = 1, 2, … , N_{cb}, where $h_k(n)$ is the impulse response of the k-th critical-band filter and N_{cb} (=23) denotes the number of critical-band filters. The critical-band filters are implemented as FIR filters, thus the k-th critical band signal is represented by the convolution as:

$$s_k(n) = s(n) * h_k(n).$$

The characteristic frequency of the filters in critical-band filterbank ranges from 125 Hz to 3500 Hz to cover telephone-band speech signals, and the bandwidth of each critical-band filter is characterized by equivalent rectangular bandwidth (ERB):

$$\mathrm{ERB}_k = F_k / Q_{\mathrm{ear}} + B_{\mathrm{min}}$$

where Q_{ear} and B_{min} are set 9.26449 and 24.7, respectively, and F_k is the characteristic frequency of the k-th critical-band filter in Hertz. Table 2 shows the characteristic frequencies and bandwidths of critical band filters.

Table 2 - Characteristic frequency and bandwidth of critical-band filters

Critical-Band Index, k	Characteristic Frequency, F_k [Hz]	Critical Bandwidth [Hz]
0	125.0	38.19
1	165.0	42.51
2	209.5	47.31
3	259.0	52.66
4	314.1	58.61
5	375.5	65.22
6	443.7	72.60
7	519.7	80.80
8	604.3	89.93
9	698.4	100.09
10	803.2	111.40
11	919.8	123.98
12	1049.6	137.99
13	1194.0	153.58
14	1354.8	170.94
15	1533.7	190.25
16	1732.9	211.74
17	1954.5	235.67
18	2201.2	262.30
19	2475.8	291.93
20	2781.4	324.92
21	3121.5	361.63
22	3500.0	402.49

8.2 Modulation Spectrum Analysis

In this ANS, the higher level of auditory pathway is modeled by the modulation filterbank and the analysis of modulation band power. Each critical-band signal, $s_k(n)$, is processed to obtain temporal envelope signal as

$$\gamma_k(n) = \sqrt{s_k^2(n) + \hat{s}_k^2(n)}$$

where $\hat{s}_k(n)$ is the Hilbert transform of $s_k(n)$. The Hilbert transform is approximated by an 35-tap FIR filter, of which coefficients are shown in Table 3.

Table 3 - Filter coefficients for Hilbert transform

Index	0	1	2	3	4	5	6	7
Coefficient	0.038135	0.0	0.024179	0.0	0.032403	0.0	0.043301	0.0
Index	8	9	10	11	12	13	14	15
Coefficient	0.05842	0.0	0.081119	0.0	0.120167	0.0	0.207859	0.0
Index	16	17	18	19	20	21	22	23
Coefficient	0.635163	0.0	-0.635163	0.0	-0.207859	0.0	-0.120167	0.0
Index	24	25	26	27	28	29	30	31
Coefficient	-0.081119	0.0	-0.05842	0.0	-0.043301	0.0	-0.032403	0.0
Index	32	33	34					
Coefficient	-0.024179	0.0	-0.038135					

The temporal envelope at the k-th critical band, $\gamma_k(n)$, is multiplied by the 256 ms Hamming window $Hamm_{256ms}(n)$:

$$Hamm_{256ms}(n) = 0.54 - 0.46\cos(2\pi n/(N-1)), \qquad \text{for } 0 \le n \le N\text{-}1$$

where N=2048. The Hamming window is shifted by 64 ms every frame, in order to obtain $\gamma_k(m;n)$, which is the temporal envelope for the k-th critical band at the m-th frame. The modulation spectrum for each critical band is then estimated by the magnitude of Fourier transform as:

$$\Gamma_k(m,f) = |F\{\gamma_k(m;n)\}|$$

where f denotes modulation frequency and $F\{x\}$ is the Fourier transform of x.

The modulation spectrum is grouped into 7 bands by a modulation filterbank $W_{\text{mod}}(i,f)$, i = 1, 2, …, 7, which is implemented and applied in modulation frequency domain, and one can obtain modulation band power of the k-th critical band at the m-th time frame as:

$$\Psi_k(m,i) = \int \Gamma_k^2(m,f) W_{\text{mod}}^2(i,f)\,df \ .$$

For discrete-time signals, the modulation spectrum is obtained by 2048-point FFT and the modulation band power can be approximated as

$$\Psi_k(m,i) = \sum_{j=0}^{1023} \Gamma_k^2(m,j) W_{\text{mod}}^2(i,j) \ .$$

Figure 3 shows the frequency response of modulation filterbank $W_{\text{mod}}(i,f)$. The quality factor of each filter is set 2 which is defined as the center frequency divided by the bandwidth of filter.

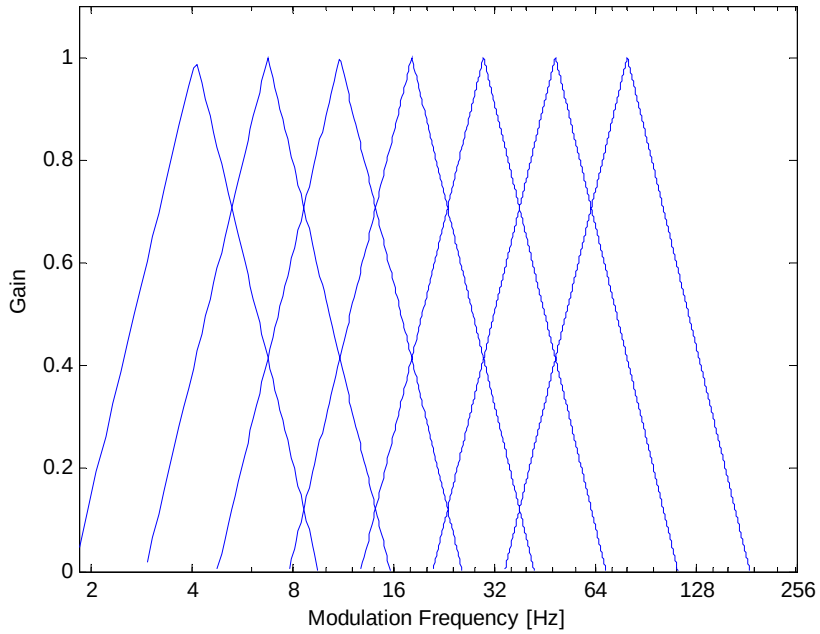

Figure 3 - Frequency response of modulation filterbank

The average articulation power is defined as:

$$\Psi_{k,A}(m) = \frac{1}{L_A}\sum_{i=1}^{L_A}\Psi_k(m,i)$$

and reflects the amount of signal components relevant to natural human speech. In order to cover the frequency range of 2 ~ 30 Hz, corresponding to the movement speed of human articulation system, L_A is set to 4.

On the contrary, the average nonarticulation power represents the amount of perceptually annoying distortions produced at the rates beyond the speed of human articulation systems, and is defined as:

$$\Psi_{k,N}(m) = \frac{1}{L_N(k) - L_A}\sum_{i=L_A+1}^{L_N(k)}\Psi_k(m,i)$$

Where:

$$L_N(k) = \begin{cases} 5, & 0 \le k \le 13 \\ 6, & 14 \le k \le 18 \\ 7 & 19 \le k \le 22 \end{cases}$$

specifies the last modulation filter index to be considered in estimating the nonarticulation power.

8.3 Feature Extraction

The logarithm of the DC-value of modulation spectrum is subtracted from the logarithms of average articulation power and the average nonarticulation power to represent normalized values as:

$$\Psi'_{k,A}(m) = \log(\Psi_{k,A}(m)+1) - \log(\Gamma_k(m,0)+1)$$

And:

$$\Psi'_{k,N}(m) = \log(\Psi_{k,N}(m)+1) - \log(\Gamma_k(m,0)+1).$$

Also,

$$\Gamma'_k(m,0) = \log(\Gamma_k(m,0)+1) \Big/ \max_j \log(\Gamma_j(m,0)+1)$$

is the normalized DC-value of modulation log-spectrum.

The input to frame distortion model at the *m*-th frame is a 69-dimensional feature vector:

$$\Xi(m) = [\Psi_A'(m);\ \Psi_N'(m);\ \Gamma'(m,0)]$$

where $\Psi_A'(m)$, $\Psi_N'(m)$, and $\Gamma'(m,0)$ are row vector representations of $\Psi'_{k,A}(m)$, $\Psi'_{k,N}(m)$, and $\Gamma'_k(m,0)$, respectively.

9 DETECTION OF ACTIVE SPEECH AND AUDIBLE BACKGROUND NOISE

The voice activity detection (VAD) is estimated every 16 msec based on frame power, its time-derivative, and estimated adaptive background noise power. The frame power calculated every 16 msec, $P_{16ms}(m)$, is defined as:

$$P_{16ms}(m) = 10\log_{10}\left(\sum_{n=0}^{255} s^2(m;n)\,Hamm_{32ms}^2(n) + 1\right)$$

where $s(m;n)$ is the *m*-th, 32 msec frame of $s(n)$, advanced in 16 msec steps and $Hamm_{32ms}(n)$ is the 32 msec-long Hamming window:

$$Hamm_{32ms}(n) = 0.54 - 0.46\cos(2\pi n/(N-1)), \qquad \text{for } 0 \le n \le N\text{-}1$$

where $N = 256$.

The time-derivative of $P_{16ms}(m)$ is calculated as:

$$\Delta P_{16ms}(m) = \left[P_{16ms}(m+1) - P_{16ms}(m-1)\right]/2.$$

Figure 4 shows the overview of the VAD procedure. For every frame, a threshold is adaptively determined based on the estimated background noise level and is used to obtain VAD profile -- whether the current frame belongs to active speech or background noise. The VAD profile is then post-processed to obtain the final VAD profile with samples spaced every 16 msec.

In order to use the VAD profile in the frame distortion estimation (Clause 10), the estimated VAD profile with 16 msec sampling is then transformed to a final estimated VAD profile with 64 msec sampling:

$$I_S(m) = \begin{cases} 1, & \textit{for active speech,} \\ 0, & \textit{otherwise.} \end{cases}$$

Among the frames determined as background noise, those of which the value of frame envelope ($P_{env}(m)$ defined in Section 10) exceeds a threshold (= 82) are considered as audible background noise, to obtain the profile for audible background noise:

$$I_B(m) = \begin{cases} 1, & \textit{for audible background noise,} \\ 0, & \textit{otherwise.} \end{cases}$$

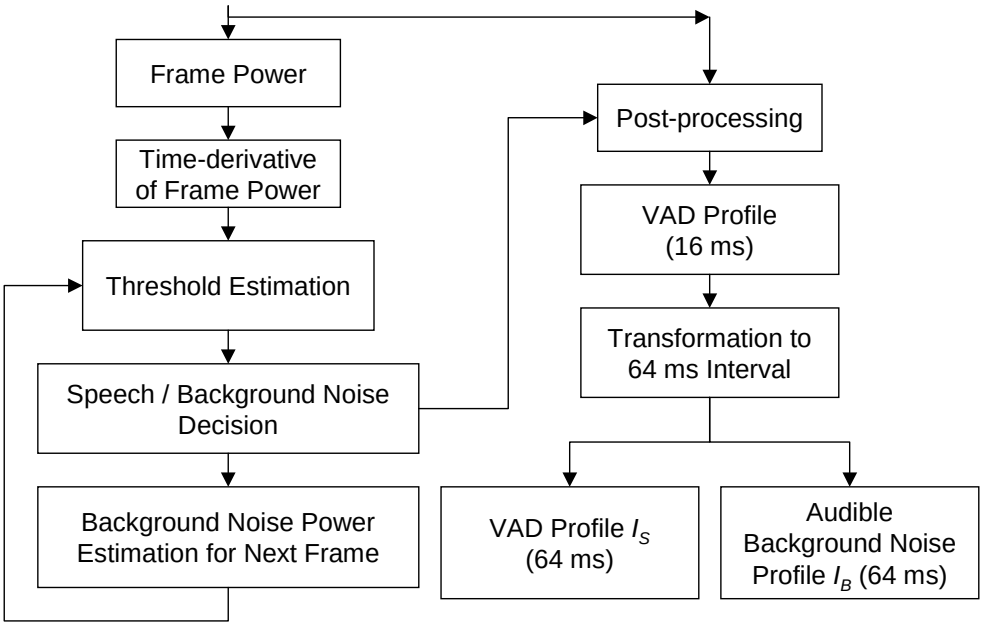

Figure 4 - Detection of active speech and audible background noise

10 FRAME DISTORTION MODEL

10.1 Overview

In the frame distortion model, the perceptual distortion of individual speech frames is estimated every 64 msec and the overall frame distortion D_F is modeled as:

$$D_F = D_S + D_B .$$

Here, D_S is the distortion in speech obtained by accumulating frame distortions over active speech frames ($I_S(m) = 1$ in Clause 9) and then normalizing by the total number of active speech frames T_S as:

$$D_S = \frac{1}{T_S} \sum_{m \in S} \chi(m)$$

where $\chi(m)$ is the output of the frame distortion model ranging from 0 to 1 at the m-th frame in 64 msec interval. D_B is the distortion in background noise and is estimated for audible background noise frames ($I_B(m) = 1$ in Clause 9) as:

$$D_B = \frac{1}{T_B} \sum_{m \in B} \{\alpha_F (P_{env}(m) - P_{th}) + \beta_F\} \chi(m)$$

where T_B is the number of frames determined as audible background noise, P_{th} (= 82.0) is the threshold to determine whether the background noise is audible enough, and $P_{env}(m)$ is the frame envelope defined as

$$P_{env}(m) = 10\log_{10}\sum_{k=1}^{N_{cb}}\Gamma_k^2(m,0).$$

Two parameters α_F (= 0.0064) and β_F (= 0.014696) are weighting factors for the frame envelope.

Figure 5 illustrates how the frame distortion is estimated from a speech signal. The feature vector Ξ(m) at the m-th frame is fed into the frame distortion model as an input to produce the frame distortion $\chi(m)$. If the current frame is classified as active speech frame, the frame distortion value is accumulated to produce the distortion for active speech D_S. If the current frame is classified as audible background noise, the frame distortion is weighted through the frame envelope and accumulated to yield the background distortion D_B. These two distortion values, D_S and D_B, are added to obtain the overall frame distortion D_F.

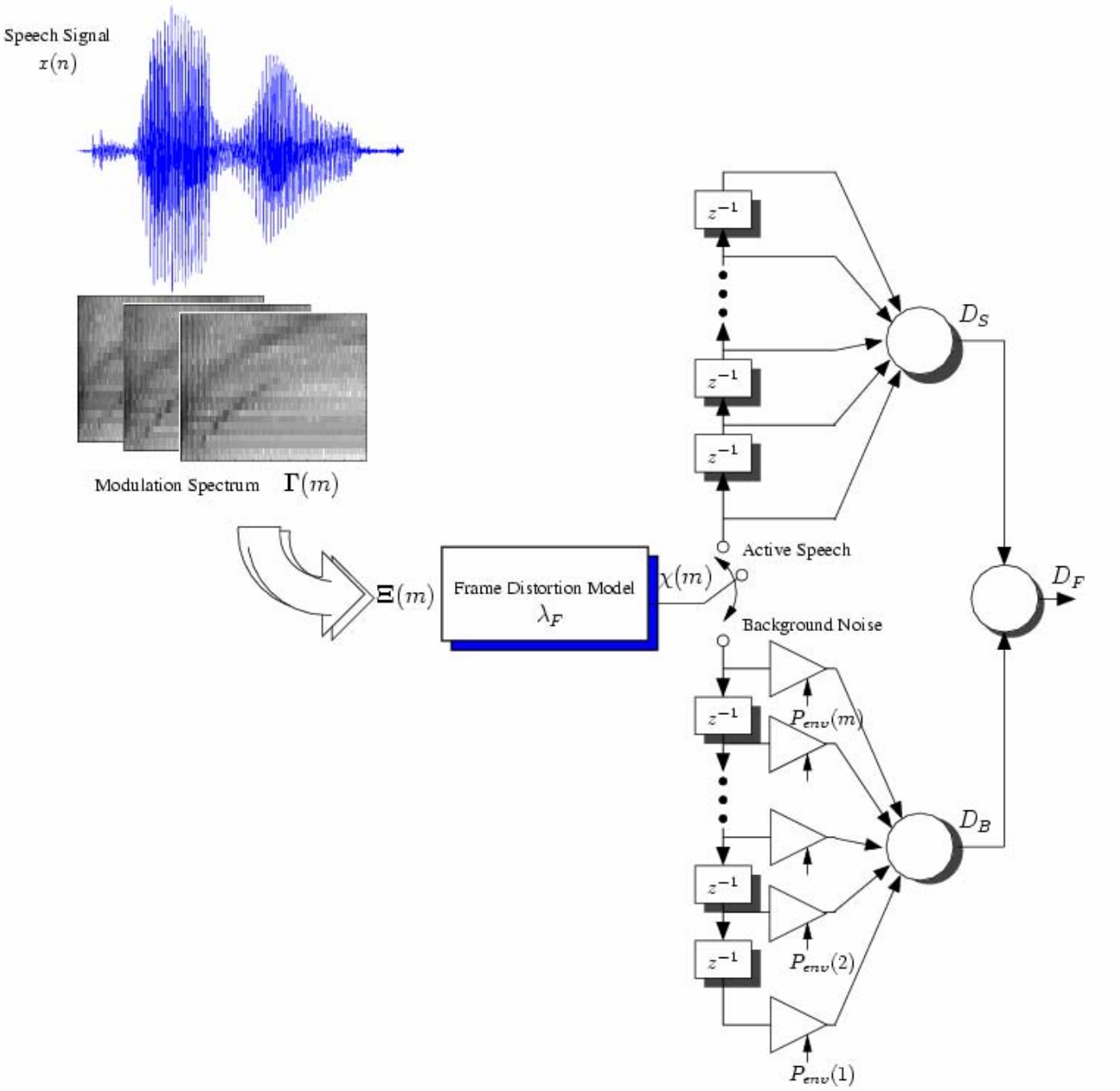

Figure 5 - Frame distortion estimation model

10.2 Multi-Layer Perceptron Model

The frame distortion model, λ_F, is the multi-layer perceptron (MLP) with one hidden layer as shown in Figure 6.

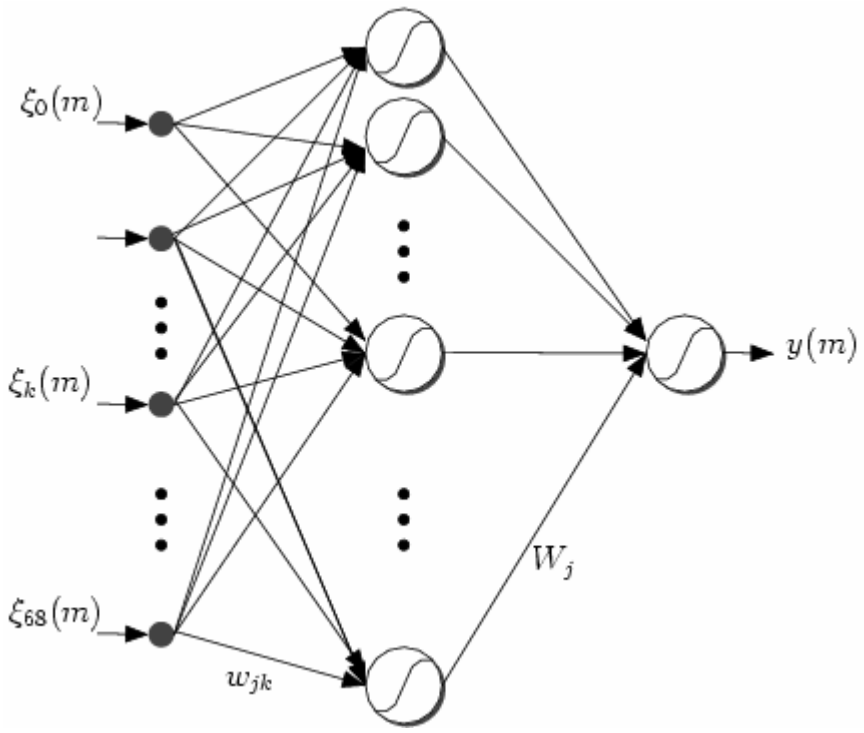

Figure 6 - Multi-layer perceptron for frame distortion model

The MLP consists of 69 input neurons, 120 hidden neurons, one output neuron, and synaptic weights connecting between layers. It can learn the mapping function between input feature vectors and the corresponding perceptual distortions derived from subjective MOS value. The output of MLP is expressed as:

$$y(m) = g\left(\sum_j W_j g\left(\sum_k w_{jk} \xi_k(m) \right) \right)$$

Here, $\xi_k(m)$ is the k-th element of input feature vector $\Xi(m)$, w_{jk} and W_j are synaptic weights for the input and hidden layer, respectively, and g(x) is the nonlinear sigmoid function defined as:

$$g(x) = \frac{2}{1+\exp(-\alpha_{MLP} x)} - 1$$

where a_{MLP} (= 0.3) is the slope of sigmoid function. As the range of output of MLP, $y(m)$, is (-1, 1), the estimated frame distortion can be obtained:

$$\chi(m) = (\tilde{y}(m) + 0.9)/1.8$$

Where:

$$\tilde{y}(m) = \begin{cases} -0.9, & y(m) < -0.9 \\ y(m), & -0.9 \le y(m) < 0.9 \\ 0.9, & y(m) \ge 0.9 \end{cases}$$

Synaptic weights and the parameters α_F and β_F are obtained by error back-propagation learning algorithm.

11 MUTE MODEL

Interruption, such as mutes, is the one of most common distortions observed as a form of packet loss or frame erasure in modern wireless and Voice over Internet Protocol (VoIP) networks. The purpose of the mute model in this section is to detect unnatural mutes in speech signals and estimate its impact on perceived quality.

11.1 Activity Profile Analysis

The activity profile for mute detection is based on frame power, its time-derivative, and adaptive background noise power estimated every 4 msec, as finer resolution is required to detect possible short mutes in speech signals. The overview of this procedure is shown in Figure 7.

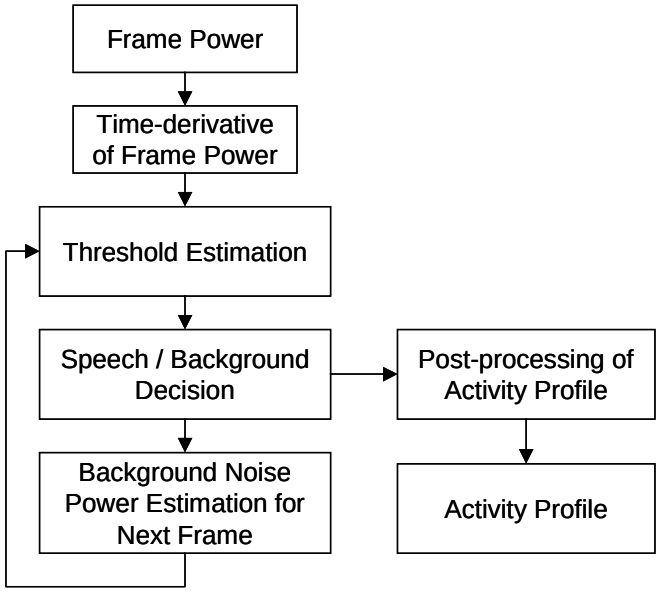

Figure 7 - Overview of activity profile analysis

The frame power $P_{4ms}(l)$ is estimated every 4 msec as:

$$P_{4ms}(l) = 10\log_{10}\left(\sum_{n=0}^{63} s^2(l;n)Hamm_{8ms}^2(n) + 1\right)$$

where $s(l;n)$ is the l-th, 8 msec frame of $s(n)$, advanced in 4 msec steps. $Hamm_{8ms}(n)$ is the 8 msec-long Hamming window defined as:

$$Hamm_{8ms}(n) = 0.54 - 0.46\cos(2\pi\, n/(N-1)), \qquad \text{for } 0 \le n \le N\text{-}1$$

where $N = 64$. The time-derivative of $P_{4ms}(l)$ is calculated as:

$$\Delta P_{4ms}(l) = \left[P_{4ms}(l+1) - P_{4ms}(l-1)\right]/2\,.$$

Figure 8 shows an example of the result of activity profile analysis. In the upper part of the figure, the blue curve is the frame power $P_{4ms}(l)$ and the red curve shows the estimated background noise power. The time-derivative of $P_{4ms}(l)$ is depicted at lower part of the figure together with the estimated activity profile $A(l)$. The value of $A(l)$ is either 0 or 1 (magnified by 10 in Figure 8), and the time instances for the beginning and end of the i-th activity are denoted by u_i (upward) and d_i (downward).

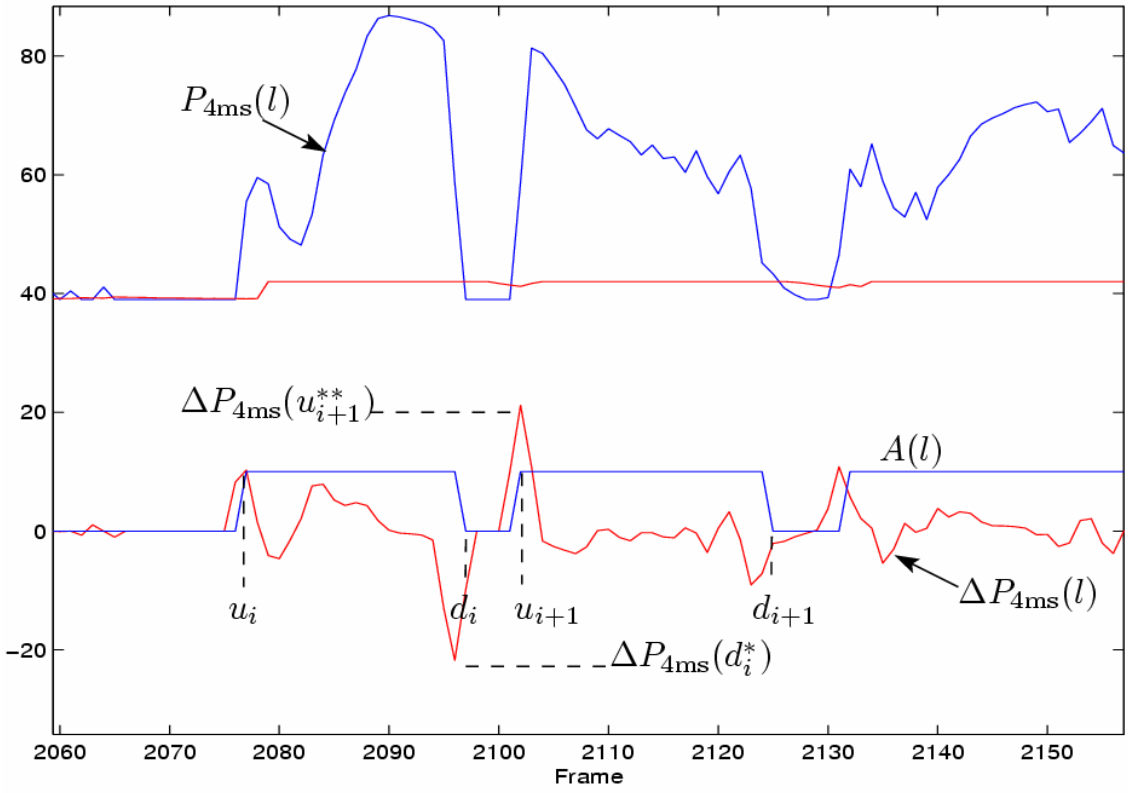

Figure 8 - Illustration of activity profile analysis for mute detection

11.2 Mute Detection

Unnatural mute events in speech signals are categorized into two types - *unnatural abrupt stop* and *unnatural abrupt start* – and they are handled separately in this ANS.

11.2.1 Unnatural Abrupt Stops

Unnatural abrupt stops may occur when speech frames or packets start to be lost during active speech intervals. As these should be distinguished from natural stop sounds in speech such as /p/ and /t/, spectral characteristics as well as the time-derivative of frame power are considered. The detection is started at every possible candidate time instance of the beginning of mute, d_i, by investigating $\Delta P_{4ms}(d_i)$. If the value of $\Delta P_{4ms}(d_i)$ is below –8, a 30-dimensional feature vector, z_{stop}, is extracted for two time instances, d_i and 15 ms prior to d_i. The feature vector z_{stop} is a column vector represented as:

$$z_{stop} = [v(d_i\text{-}15\text{ms})\ \mathbf{cep}(d_i\text{-}15\text{ms})\ v(d_i)\ \boldsymbol{cep}(d_i)\ \Delta P_{4ms}(d_i^*)\ \ P_{4ms}(d_i^*)\ \ \Delta P_{4ms}(u_{i+1}^*)\ \ \overline{P}_{4ms}(d_i)\ ]^t$$

where

$$d_i^* = \arg\min_{d_i - 5 \le l \le d_i + 1} \Delta P_{4ms}(l)$$

is the time instance for minimum ΔP_{4ms} around the beginning of a mute event, and:

$$u_{i+1}^* = \arg\max_{u_{i+1} - 1 \le l \le u_{i+1} + 1} \Delta P_{4ms}(l)$$

indicates the time instance for maximum ΔP_{4ms} around the beginning of activity following a mute event. $\overline{P}_{4ms}(d_i)$ is the averaged frame power obtained between d_i and u_{i+1}. In addition, ***cep***(m) is a row vector consisting of the 12-th order Mel-Frequency Cepstral Coefficients (MFCC) obtained from the $s_p(m;n)$, which is the pre-emphasized (with the coefficient 0.97) and 32-msec Hamming windowed version of speech signal $s(n)$ centered at m. The MFCC is defined as the cosine transform of log-spectrum, and the i-the element of ***cep***(m) can be represented as:

$$cep_i(m) = \frac{1}{N_{MFCC}} \sum_{j=1}^{N_{MFCC}} \log_{10}\left[\sum_k W_{MFCC}(j,k)\left|S_p(m;k)\right|\right] \cos\left[\frac{\pi}{N_{MFCC}} i(j-0.5)\right]$$

where $|S_p(m;k)|$ is the FFT magnitude of $s_p(m;n)$ and N_{MFCC} (= 32) is the number of filters in the filterbank $W_{MFCC}(j,k)$. As shown in Figure 9, the filterbank is constructed according to the Mel-frequency scale by using 13 linearly-spaced filters, followed by 19 log-spaced filters. The linearly-spaced filters have a spacing of 66.67 Hz, and their center freuquencies cover the range from 200 to 800 Hz. The log-spaced filters cover approximately up to 3700 Hz.

The voicing factor $v(m)$ indicates how much periodicity the $s(m;n)$ contains, where $s(m;n)$ is the 30-msec Hamming windowed version of speech signal $s(n)$ centered at time instance m. The voicing factor is defined as the normalized autocorrelation:

$$v(m) = r(m; n^*) \,/\, r(m; 0),$$

where:

$$r(m;n) = \sum_i s(m;i)s(m;n+i)$$

is the autocorrelation function of speech and:

$$n^* = \arg\max_{20 \le n \le 160} r(m;n)$$

is the estimated fundamental frequency in the search range between 50 ~ 400 Hz.

The detector for unnatural abrupt stops is the 2-layer MLP with 30 input neurons, 30 hidden neurons, and one output neuron. The output of the detector is expressed as:

$$y_{stop} = g\left(\sum_j W_{stop,j}\, g\left(\sum_k w_{stop,jk}\, z_{stop}(k)\right)\right)$$

where $z_{stop}(k)$ is the k-th element of a vector $\boldsymbol{z}_{\mathrm{stop}}$, $w_{stop,jk}$ and $W_{stop,j}$ are synaptic weights for the input and hidden layer, respectively. The nonlinear sigmoid function $g(x)$ is defined as:

$$g(x) = \frac{2}{1 + \exp(-\alpha_{MLP,stop} x)} - 1$$

where $a_{MLP,stop}$ (= 2.0) is the slope of sigmoid function. The detection is based on the output value of the detector, y_{stop}. If y_{stop} is above 0.55, the time instance d_i is considered to have an unnatural abrupt stop.

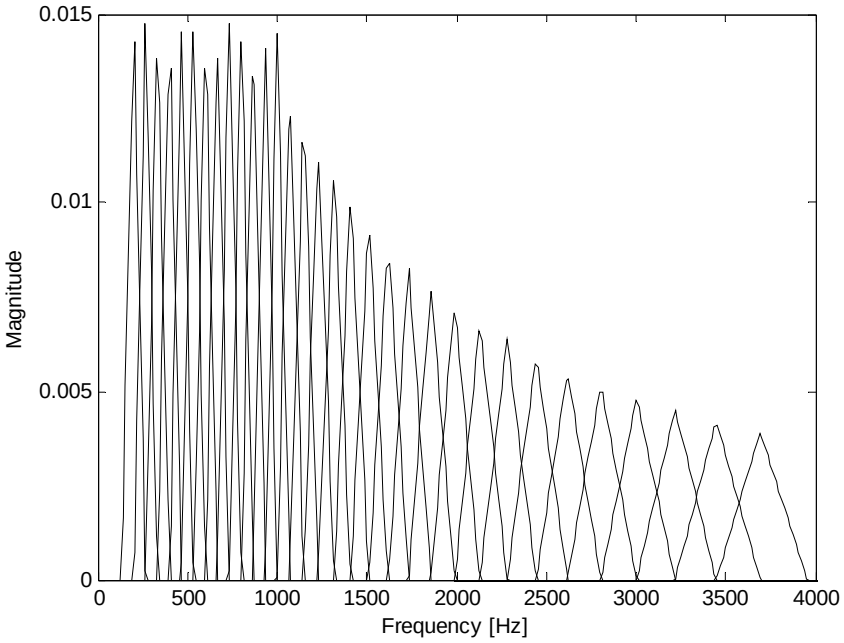

Figure 9 - MFCC Filterbank along the FFT bins

11.2.2 Unnatural Abrupt Starts

Unlike unnatural abrupt stops, the beginning of mute cannot be detected when frames start to be erased during silence even before a speech activity starts. In this case, only the end of mute exists and is termed here *unnatural abrupt start*. Similar to unnatural stop detection, a feature vector is extracted for two time instances, u_i and 15 ms after u_i, where u_i is the candidate time instance for the beginning of unnatural abrupt starts. The feature vector z_{start} is a 31-dimensional column vector represented as:

$$z_{\text{start}} = [\mu(u_i)\ v(u_i)\ \boldsymbol{cep}(u_i)\ \mu(u_i\text{+15ms})\ v(u_i\text{+15ms})\ \mathbf{cep}(u_i\text{+15ms})\ \Delta P_{4ms}(u_i^{*})\ \ P_{4ms}(u_i^{**})\ \ u_i\text{-}d_{i\text{-}1}]^{\text{t}}$$

where:

$$u_i^{**} = \underset{u_i - 1 \le l \le u_i + 1}{\arg\max} P_{4ms}(l)$$

is the time instance for maximum $P_{4\text{ms}}$ around the beginning of the abrupt start. The spectral centroid $\mu(m)$ is defined as:

$$\mu(m) = \frac{\sum_k k\left|S(m;k)\right|}{\sum_k \left|S(m;k)\right|}$$

where $|S(m;k)|$ is the FFT magnitude of speech segment $s(m;n)$.

The detector for unnatural abrupt starts is the 2-layer MLP with 30 input neurons, 20 hidden neurons, and one output neuron. The output of the detector is expressed as:

$$y_{start} = g\left(\sum_j W_{start,j}\, g\left(\sum_k w_{start,jk}\, z_{start}(k)\right)\right)$$

where $z_{start}(k)$ is the k-th element of a vector z_{start}, $w_{start,jk}$ and $W_{start,j}$ are synaptic weights for the input and hidden layer, respectively. The nonlinear sigmoid function $g(x)$ is defined as:

$$g(x) = \frac{2}{1+\exp(-\alpha_{MLP,start}\, x)} - 1$$

where $a_{MLP,start}$ (= 2.0) is the slope of sigmoid function. The detection is based on the output value of the detector, y_{start}. If y_{start} is above 0.55, the time instance u_i is considered to have an unnatural abrupt start.

11.3 Mute Impact Model

The mute impact model estimates the impact of mutes as the combination of abrupt instantaneous distortion followed by decays simulating short-term memory effects in biological systems. Let us suppose a speech signal contains K mutes and t_i (i = 1, 2, ... , K) is the time instance when each mute event ends. Then the objective distortion caused by mutes is modeled as:

$$D_M = \sum_{i=1}^{K} v_i \exp\left[-(T - t_i)/\tau\right] u(T - t_i)$$

where v_i is the instantaneous distortion of the i-th mute at time t_i, $u(x)$ is a unit step function which is 1 for $x \geq 0$ and 0 for x<0, and T is the length of a speech file, assuming the determination of quality rating is done at the end of the presentation of a speech signal in subjective MOS tests. For each mute, perceived distortion is raised by the amount of v_i at the end of mute event and decays over time with the time constant τ (=12 sec).

The instantaneous distortion of the i-th mute is estimated by:

$$v_i = p_1 \log_2(L_i) + p_2$$

Where L_i is the length of i-th mute, and p_1 (=0.038417) and p_2 (=-0.096898) are constants.

12 NONSPEECH MODEL

This module detects very annoying non-speech activities that may occur when bit information within a packet or frame is distorted during transmission but not detected at the speech decoder side, for example. In this case, the speech decoder uses the corrupted bits to generate very annoying non-speech signals.

In the nonspeech detection model, the nonspeech activity which has significantly abrupt changes in frame power is considered. For each activity period between u_i and d_i, positive and negative peaks of $\Delta P_{4ms}(l)$ are marked if the absolute value of $\Delta P_{4ms}(l)$ is above a threshold. And the reciprocal of the time interval between two adjacent marked peaks is accumulated within the activity period. If the accumulated value is above a threshold, the activity period is determined to be non-speech.

The impact of nonspeech is estimated to be proportional to the accumulated maximum frame power between two adjacent peaks of $\Delta P_{4ms}(l)$.

Annex A
(informative)

A BIBLIOGRAPHY

ITU-T P.800, *Methods for subjective determination of transmission quality*.[2]

T1.518-1998 (R2008), *Objective measurement of telephone band speech quality using measuring normalizing blocks (MNBs)*.[3]

ITU-T P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*.[2]

ITU-T P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*.[2]

D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representation by error propagation*, MIT Press: 1986.

ETSI ES 201 108 v1.1.2, *Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm*.[4]

[2] This document is available from the International Telecommunications Union. < http://www.itu.int/ITU-T/ >.

[3] This document is available from the Alliance for Telecommunications Industry Solutions (ATIS), 1200 G Street N.W., Suite 500, Washington, DC 20005. < https://www.atis.org/docstore/default.aspx >

[4] This document is available from the European Telecommunications Standards Institute (ETSI). < http://www.etsi.org/WebSite/Standards/StandardsDownload.aspx >

Annex B
(normative)

B ANSI-C REFERENCE IMPLEMENTATION OF ANIQUE+

This Annex has been formatted as a separate folder containing the C-source code of the ANIQUE+ algorithm, and electronically packaged with this standard.