ATIS-0100518.1998(R2013)

Objective Measurement of Telephone Band Speech Quality Using Measuring Normalizing Blocks (MNBs)

AMERICAN NATIONAL STANDARD FOR TELECOMMUNICATIONS

As a leading technology and solutions development organization, ATIS brings together the top global ICT companies to advance the industry's most-pressing business priorities. Through ATIS committees and forums, nearly 200 companies address cloud services, device solutions, emergency services, M2M communications, cyber security, ehealth, network evolution, quality of service, billing support, operations, and more. These priorities follow a fast-track development lifecycle – from design and innovation through solutions that include standards, specifications, requirements, business use cases, software toolkits, and interoperability testing.

ATIS is accredited by the American National Standards Institute (ANSI). ATIS is the North American Organizational Partner for the 3rd Generation Partnership Project (3GPP), a founding Partner of oneM2M, a member and major U.S. contributor to the International Telecommunication Union (ITU) Radio and Telecommunications sectors, and a member of the Inter-American Telecommunication Commission (CITEL). For more information, visit < 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 OFMERCHANTABILITY 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, AND IN NO EVENT SHALL ATIS BE LIABLE FOR LOST PROFITS OR OTHER INCIDENTAL OR CONSEQUENTIAL DAMAGES. ATIS EXPRESSLY ADVISES THAT ANY AND ALL USE OF OR RELIANCE UPON THE 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 whether use of an invention covered by patent rights will be required, and if any such use is required no position is taken regarding the validity of this claim or any patent rights in connection therewith. Please refer to [http://www.atis.org/legal/patentinfo.asp] to determine if any statement has been filed by a patent holder indicating a willingness to grant a license either without compensation or on reasonable and non-discriminatory terms and conditions to applicants desiring to obtain a license.

ATIS-0100518.1998(R2013), *Objective Measurement of Telephone Band Speech Quality Using Measuring Normalizing Blocks (MNBs)*

Is an American National Standard developed by 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

American National Standard
for Telecommunications –

Objective Measurement of Telephone Band Speech Quality Using Measuring Normalizing Blocks (MNBs)

1 Purpose, scope and application

1.1 Purpose

The desirability of using objective measures to estimate user opinions of speech quality has long been a goal of the American National Standards Institute (ANSI) and the International Telecommunications Union, Telecommunication Standardisation Sector (ITU-T). Toward that end, the ITU-T adopted Recommendation P.861 in 1996. This Recommendation provides an algorithm that can be used under a limited set of conditions. These conditions are limited to waveform and CELP speech coding technologies under clear-channel conditions. With the proliferation of mobile telephone systems, it has become increasingly important to be able to predict user opinion under conditions of transmission channel errors and lower-rate speech coding. This American National Standard (ANS) defines an algorithm that provides acceptably accurate predictions in the same areas as Recommendation P.861, as well as in additional important conditions, such as transmission channel errors and lower-rate speech coders.

1.2 Scope

Subjective quality assessment of speech codecs can be made in listening-only (one-way) tests or in conversational (two-way) tests. The objective quality measurement described in this American National Standard (ANS) estimates the subjective quality in listening-only tests of telephone band speech.[1)]

To demonstrate the subjective performance of a codec, the effects of a variety of quality factors should be investigated (see ITU-T Recommendation P.830). The accuracy of the objective quality measurement described in this ANS has not been verified for all of the factors specified in Recommendation P.830. Subclause 1.3 defines those conditions from P.830 to which this ANS applies.

When comparing a codec with another codec or with a reference condition based on subjective experimental results, statistical tests that take the distributions of subjective votes into account are often used. Since the objective measurement in this ANS estimates only the mean of subjective votes (e.g., MOS, DMOS), such statistical tests cannot be applied to the results of objective measurement. Prediction of percent poor or worse (%PoW) and percent good or better (%GoB) are currently under study.

1.3 Application

Table 1 is a guide to facilitate determination of the test factors, coding technologies, and applications to which this ANS applies.

[1)] For purposes of this standard the passband is assumed to be 300-3400 Hz (-2 dB at 300 Hz; -3 dB at 3400 Hz), the same as shown in ITU-T Recommendation G.712 for PCM-derived channels between two 2-wire analog interfaces.

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

Test factors	**Note**
speech input levels to a codec	1
listening levels in subjective experiments	2
talker dependencies	1
multiple simultaneous talkers	2
transmission channel errors	1
bit rates if a codec has more than one bit rate mode	1
transcodings	1
bit-rate mismatching between an encoder and a decoder if a codec has more than one bit rate mode	2
environmental noise in the sending side	2
network information signals as input to a codec	2
music as input to a codec	2
delay	3
short-term time warping of audio signal	2
long-term time warping of audio signal	2
temporal clipping of speech	2
amplitude clipping of speech	2
Coding technologies	
waveform	1
CELP and hybrids ≥ 4 kbit/s	1
CELP and hybrids < 4 kbit/s	1
VOCODERs	1
speech activity detectors/silence suppression systems	2
other coders	1
Applications	
coder optimization	1
coder evaluation	1
coder selection	2
network planning	4
live network testing	5
in-service non-intrusive measurement devices	3

NOTES

1) The objective measure has demonstrated acceptable accuracy in the presence of this variable.

2) Insufficient information is available about the accuracy of the objective measure with regard to this variable.

3) The objective measure is known to provide inaccurate predictions when used in conjunction with this variable, or are otherwise not intended to be used with this variable.

4) With caution, the objective measure might be used for some network planning purposes. The reader should note that there are important factors in network planning to which this ANS is not applicable (see the “Test factors” section of this table).

5) With caution, the objective measure might be used for some live network testing. The reader should note that there may be factors or technologies in a live network connection to which this ANS is not applicable (see the “Test Factors” section of this table).

Note that this document contains sufficient information to implement this algorithm in a computer programming language. Implementations can be validated by using information available from ftp://ftp.its.bldrdoc.gov/dist/voice/verify.zip.

2 Normative references

The following standards contain provisions that, through reference in this text, constitute provisions of this American National Standard. At the time of publication, the editions indicated were valid. All references 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 edition of the standards listed below.

ANSI T1.801.04-1997, *Telecommunications – Multimedia Communications Delay, Synchronization, and Frame Rate*

ITU-T Recommendation G.712 – (11/96) – *Transmission performance characteristics of pulse code modulation channels*[2)]

ITU-T Recommendation P.800 – (05/1996) – *Methods for subjective determination of transmission quality*[2)]

ITU-T Recommendation P.830 – (09/1995) – *Subjective performance assessment of telephone-band and wide-band digital codecs*[2)]

3 Abbreviations

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

ACR	Absolute Category Rating
AD	Auditory Distance
ANS	American National Standard
ANSI	American National Standards Institute
CELP	Code Excited Linear Prediction
DCR	Degradation Category Rating
DMOS	Degradation Mean Opinion Score
DUT	Device Under Test
FMNB	Frequency Measuring Normalizing Block
GoB	Good or Better
ITU-T	International Telecommunications Union – Telecommunication Standardization Sector
MOS	Mean Opinion Score
MNB	Measuring Normalizing Block
PoW	Poor or Worse
TMNB	Time Measuring Normalizing Block

4 Definitions

For the purpose of this ANS, the following definitions apply:

Bark: a perception-based unit of frequency equivalent to the width of a critical band. As frequency increases, the width of the critical hearing band (in Hertz) increases. On the Bark scale, equal frequency intervals are of equal perceptual importance.

———

[2)] Available from the American National Standards Institute, 11 West 42nd Street, New York, NY 10036.

reference vector: audio vector used as input, or source, for a device under test.

test vector: output vector from device under test, compared against the reference vector in the measurement process.

5 Conventions

Subjective evaluation of speech 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. This ANS defines an objective measurement technique for estimating subjective quality obtained in listening-only tests.

6 Summary of objective measurement procedure

Objective quality measurement of speech codecs requires a number of steps:

1) Preparation of reference vectors, i.e., recording of talkers and/or generation of the artificial voices conforming to ITU-T Recommendation P.50;
2) Selection of experimental parameters that will exercise the salient features of the codec and are able to be tested by objective measurement;
3) Production of reference and test speech vectors;
4) Calculation of the objective speech quality based on measuring normalizing blocks (MNBs) using reference and test speech vectors;
5) Transformation from the objective quality scale to the subjective quality scale, if necessary;
6) Analysis of results.

Steps 1, 2, 4, 5, and 6 are discussed below.

7 Source speech material preparation

Reference vectors for objective measurement may be real voices or the artificial voices specified in ITU-T Recommendation P.50, depending on the goals of the experiment.

Since the artificial voices defined in ITU-T Recommendation P.50 reproduce the mean characteristics of human speech over various languages, they are useful in objectively estimating the mean subjective quality of a codec over these languages. When the talker-dependency of a codec or the performance of a codec for particular languages is concerned, it is recommended that real voices be used. In either case, no environmental noise should be added.

7.1 Real voices

When real voices are used in objective measurement, they should be produced, recorded, and level-equalized in accordance with section 7 of ITU-T Recommendation P.830.

It is recommended that a minimum of two male talkers and two female talkers should be used for each testing condition. If talker dependency is to be tested as a factor in its own right, it is recommended that more talkers be used as follows:

- 8 male;
- 8 female;
- 8 children.

7.2 Artificial voices

When the artificial voices conforming to ITU-T Recommendation P.50 are used in objective measurement, it is recommended that both male and female artificial voices be used. These vectors should be passed through a filter with appropriate frequency characteristics to simulate sending frequency characteristics of a telephone handset, and level-equalized in the same manner as real voices (see ITU-T Recommendation P.830).

8 Selection of experimental parameters

To demonstrate the performance of a codec, the effects of various quality factors on the performance of the codec should be examined. ITU-T Recommendation P.830 provides guidance on subjectively assessing the following quality factors:

1) speech input levels to a codec;
2) listening levels in subjective experiments;
3) talkers (including multiple simultaneous talkers);
4) errors in the transmission channel between an encoder and a decoder;
5) bit rates if a codec has more than one bit-rate mode;
6) transcodings;
7) bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode;
8) environmental noise in the sending side;
9) network information signals as input to a codec;
10) music as input to a codec.

Note that objective measurement for quality factors other than those specifically noted as applicable in this standard (see Table 1) is still under study. Therefore, these factors should be measured only after the accuracy of an objective measure is verified in conjunction with subjective tests conforming to ITU-T Recommendation P.830.

In addition to the codec conditions, ITU-T Recommendation P.830 recommends the use of reference conditions in subjective tests. These conditions are necessary to facilitate the comparison of subjective test results from different laboratories or from the same laboratory at different times. Also, when expressing the objective test results in terms of equivalent-Q values, reference conditions using the narrow-band Modulated Noise Reference Unit (MNRU) as specified in ITU-T Recommendation P.810 should be tested.

Note that including other standard codecs such as G.711 64-kbit/s PCM, G.726 48-, 32-, 24- and 16-kbit/s ADPCM, G.728 16-kbit/s LD-CELP, and G.729 8-kbit/s CS-ACELP as well as MNRU in objective quality measurement tests may help demonstrate the relative performance of the codec under test and standardized codecs.

Detailed explanations of these experimental parameters are found in ITU-T Recommendation P.830.

9 Computation of the objective measure

This clause describes the computation of an objective measure based on measuring normalizing blocks (MNBs). MNBs were developed in response to the observations that listeners adapt and react differently to spectral deviations that span different time and frequency scales. Thus, for the speech quality estimation application, maximal perceptual consistency over a wide range of distortion types requires a family of analyses at multiple frequency and time scales. As spectral deviations are measured, the deviations at one scale must be removed so that they are not counted again as part of the deviations at other scales. It is also observed that working from larger to smaller scales is most likely to emulate listeners' patterns of adaptation and reaction to spectral deviations. This observation has led to a hierarchical structure of MNBs.

Two types of measuring normalizing blocks are considered here. The first is the time measuring normalizing block (TMNB) and the second is the frequency measuring normalizing block (FMNB). Each of these blocks takes per-

ceptually transformed reference (R(t,f)) and test (T(t,f)) signals as inputs and returns them and a set of measurements as outputs. These two building blocks are defined by Figures 1 and 2, respectively. The TMNB integrates over some frequency scale, then measures differences and normalizes the test signal at multiple times. Finally, the positive and negative portions of the measurements are integrated over time. In an FMNB the converse is true. An FMNB integrates over some time scale, then measures differences and normalizes the test signal at multiple frequencies. Finally, the positive and negative portions of the measurements are integrated over frequency. By design, both types of MNBs are idempotent. This important property is illustrated in Figure 3 and simply means that a second pass through a given MNB will not further alter the test signal, and that second pass will result in a measurement vector of zeros. The idempotency of MNBs allows them to be cascaded and yet still measure the deviation at a given time or frequency scale once and only once.

In order to measure spectral deviations at multiple time and frequency scales, a hierarchical structure of TMNBs and FMNBs, operating at decreasing scales has been formed (Figure 4). When used as a distance measure in conjunction with a perceptual transformation (described below), this structure appears to do a good job of emulating listeners' patterns of adaptation and reaction to spectral deviations. The structure shown results in 12 measurements. Because of the hierarchical nature of these structures, measurements from other than the top layer mean little individually, but a linear combination of the measurements has been found to be a good indicator of the perceptual distance between the two signals. The value that results from this linear combination is called auditory distance (AD):

$$\mathrm{AD} = \sum_{i=1}^{12} m(i) \cdot weight_i .$$

Auditory distance is a positive quantity. When the reference and test signals are similar, AD is small. As the reference and test signals move apart perceptually, AD increases. A logistic function or some other "limiter function" can be used to map AD into a finite interval. This allows AD to correlate better with subjective quality or impairment judgments, which usually cover a finite range. The weights, $weight_i$, have been selected to maximize this correlation.

9.1 Input-Output Specifications

The input to the algorithm is a pair of speech vectors called reference and test. The vector called reference contains a digital representation of the reference signal, which is typically the input to the device under test (DUT), and is referred to as *x* in the equations. The vector called test contains a digital representation of the test signal, which is typically the output from the DUT and is referred to as *y* in the equations. The sample rate is 8000 samples per second, and the recommended precision is at least 16 bits per sample. Lower precision may be used, if the user is willing to accept the associated loss of sensitivity. In addition, higher precision and higher sample rates (e.g., 16000 samples per second) may be used if care is given to ensure that the transformation to the frequency domain produces the results specified in 9.4 of this standard. Also, if higher-sample-rate files are available, they can be down-sampled using the programs available in the Software Tools Library maintained by ITU-T SG 16 and published in ITU-T Recommendation G.191. The input files must contain at least one second of telephone bandwidth speech. (Files that contain only pauses in a natural conversation are not useful.) Files used in the development of these algorithms ranged from 3 to 9 seconds in duration.

The algorithm generates a single, non-negative output value called Auditory Distance (AD). AD is an estimate of the perceptual distance between the reference and test signals. Thus, when the DUT and the test set-up are transparent, the reference and test signals will be identical, and AD will be zero. As the DUT introduces more and more distortion, the reference and test signals will move apart perceptually, and AD will increase.

9.2 Time Delay

It is assumed that the two files have the same length, and are synchronized. That is, any delay in the DUT, or the test set-up has been removed. If these delays are known a priori, they may be removed by proper timing during data acquisition. If these delays are not known a priori, they may be estimated using the technique described in clause 7 of ANSI T1.801.04, and then removed by editing one or both of the files. If delay cannot be estimated and removed, this standard does not apply (see Table 1).

9.3 Signal Preparation

The contents of reference are read into the vector *x*, and the contents of test are read into the vector *y*. The mean value is then removed from each of the *N1* entries in each of these vectors:

$$x(i) = x(i) - \frac{1}{N1} \cdot \sum_{j=1}^{N1} x(j),\ \ y(i) = y(i) - \frac{1}{N1} \cdot \sum_{j=1}^{N1} y(j),\ 1 \le i \le N1.$$

This eliminates any DC component that may be present in the test and reference signals. (The DC component of a signal is inaudible.)

Note that the notation *Array(i) = Array (i) +/-/•/÷ Normalization Factor* is used throughout this document. While not mathematically consistent, it is indicative of how variables might be reused when this algorithm is implemented in computer code. The original array is modified (normalized) in the manner indicated, and the modified value is used in future computations.

Next, each of the vectors is normalized to a common RMS level:

$$x(i) = x(i) \cdot \left[\frac{1}{N1} \sum_{j=1}^{N1} x(j)^2 \right]^{-1/2},\ y(i) = y(i) \cdot \left[\frac{1}{N1} \sum_{j=1}^{N1} y(j)^2 \right]^{-1/2},\ 1 \le i \le N1.$$

This approximately removes any fixed gain in the DUT or the test set-up. Thus, a fixed gain will not influence the values of AD produced by this algorithm.

9.4 Transformation to Frequency Domain

The signals are then transformed to the frequency domain using the FFT. The frame size is 16 ms (128 samples for speech sampled at 8000 Hz), and the frame overlap is 50%. Any samples beyond the final full frame are discarded. Each frame of samples is multiplied (sample by sample) by the length 128 Hamming window:

$$w(i) = 0.54 - 0.46 \cos\left(\frac{2\pi(i-1)}{127} \right),\ 1 \le i \le 128.$$

After multiplication by the Hamming window, each frame is transformed to a 128-point frequency domain vector using the FFT. For each frame, the squared-magnitude of frequency samples 1 through 65 (DC through Nyquist) are retained. The results are stored in the matrices *X* and *Y*. These matrices contain 65 rows, and *N2* columns, where *N2* is the number of frames that are extracted from the *N1* original samples in *x* and *y*.

Note that in these matrices, rows represent the frequency axis (indexed by *i* in the algorithmic description), and columns represent the time axis (indexed by *j*).

Because FFT scaling is not standardized, care should be taken to ensure that the FFT used in this algorithm is scaled so that the following condition is met: When a frame of 128 real-valued samples, each with value 1 is input to the FFT without windowing, then the complex value in the DC bin of the FFT output must be 128+0•j.

9.5 Frame Selection

Only frames that meet or exceed energy thresholds in both *X* and *Y* are used in the calculation of AD. For *X*, that energy threshold is set to 15 dB below the energy of the peak frame in *X*:

$$xenergy(j) = \sum_{i=1}^{65} X(i, j),\ xthreshold = 10^{\frac{-15}{10}} \cdot \max_j(xenergy(j)).$$

For *Y*, the energy threshold is set to 35 dB below the energy of the peak frame in *Y*:

$$yenergy(j) = \sum_{i=1}^{65} Y(i, j),\ ythreshold = 10^{\frac{-35}{10}} \cdot \max_j(yenergy(j)).$$

Frames that meet or exceed both of these energy thresholds are retained:

$$\{xenergy(j) \geq xthreshold\} \text{AND} \{yenergy(j) \geq ythreshold\} \Rightarrow \text{frame } j \text{ is retained.}$$

If any frame contains one or more samples that are equal to zero, that frame is eliminated from both *X* and *Y*. These matrices now contain 65 rows, and *N3* columns, where *N3* is the number of frames that have been retained. If *N3*=0, the input files do not contain suitable signals and the algorithm is terminated.

Note that the threshold levels are different for *X* and *Y*. This is to provide for the inclusion of frames in *Y* that have reduced power due to a perceptually significant artifact (e.g., temporal clipping), yet still retain enough power to be successfully compared to the original speech material.

9.6 Perceived Loudness Approximation

Each of the frequency domain samples in *X* and *Y* are now logarithmically transformed to an approximation of perceived loudness:

$$X(i, j) = 10 \cdot \log_{10}(X(i, j)),\ Y(i, j) = 10 \cdot \log_{10}(Y(i, j)),\ 1 \leq i \leq 65, 1 \leq j \leq N3.$$

9.7 Frequency Measuring Normalizing Block (FMNB)

An FMNB is applied to *X* and *Y* at the longest available time scale, defined by the length of the input files. Four measurements are extracted and stored in the measurement vector *m*. These measurements cover the lower and upper band edges of telephone band speech (where it is easiest to detect noise and changes in frequency response). Temporary vectors *f1*, *f2*, and *f3* are used:

$$f1(i) = \frac{1}{N3} \sum_{j=1}^{N3} Y(i, j) - \frac{1}{N3} \sum_{j=1}^{N3} X(i, j), 1 \leq i \leq 65, \text{ Measure}$$

$$f2(i) = f1(i) - f1(17), 1 \leq i \leq 65, \text{ Normalize measurement to 1kHz}$$

$$Y(i, j) = Y(i, j) - f2(i), 1 \leq i \leq 65, 1 \leq j \leq N3, \text{ Normalize } Y$$

$$f3(i) = \frac{1}{4} \sum_{j=1}^{4} f2(1 + 4 \cdot (i - 1) + j), 1 \leq i \leq 16, \text{ Smooth the measurement}$$

$$m(1) = f3(1),\ m(2) = f3(2),\ m(3) = f3(13),\ m(4) = f3(14), \text{ Save 4 measurements}$$

9.8 Computing Time Measuring Normalizing Blocks

In the MNB structure, the middle portion of the band undergoes two additional levels of binary band splitting, resulting in bands that are approximately 2-3 Bark wide. The extreme top and bottom portions of the band are each treated once by a separate TMNB. Finally, a residual measurement is made. There are a total of 9 TMNBs in the structure and a graphical representation is given in Figure 4. The MNB structure generates 8 measurements in addition to those generated by the initial FMNB (described in 9.7). Temporary variables *t0*, *t1*, ..., *t9* are used.

TMNB-0 (Bottom of band, 1.9 Bark Wide)

$$t0(j) = \frac{1}{5} \sum_{i=2}^{6} Y(i, j) - \frac{1}{5} \sum_{i=2}^{6} X(i, j), 1 \leq j \leq N3, \text{ Measure}$$

$$Y(i, j) = Y(i, j) - t0(j), 2 \leq i \leq 6, 1 \leq j \leq N3, \text{ Normalize } Y$$

$$m(5) = \frac{1}{N3} \sum_{j=1}^{N3} \max(t0(j), 0), \text{ Save positive portion of measurement}$$

TMNB-1 (Middle of band, top layer, 10 Bark wide)

$$t1(j) = \frac{1}{36}\sum_{i=7}^{42} Y(i,j) - \frac{1}{36}\sum_{i=7}^{42} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t1(j), 7 \le i \le 42, 1 \le j \le N3, \text{NormalizeY}$$

$$m(6) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t1(j),0), \text{Savepositiveportionof measurement}$$

$$m(7) = -\frac{1}{N3}\sum_{j=1}^{N3} \min(t1(j),0), \text{Savenegativeportionof measurement}$$

TMNB-2 (Top of Band, 3 Bark wide)

$$t2(j) = \frac{1}{23}\sum_{i=43}^{65} Y(i,j) - \frac{1}{23}\sum_{i=43}^{65} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t2(j), 43 \le i \le 65, 1 \le j \le N3, \text{NormalizeY}$$

$$m(8) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t2(j),0), \text{Savepositiveportionof measurement}$$

TMNB-3 (Middle of band, middle layer, 5 Bark wide)

$$t3(j) = \frac{1}{12}\sum_{i=7}^{18} Y(i,j) - \frac{1}{12}\sum_{i=7}^{18} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t3(j), 7 \le i \le 18, 1 \le j \le N3, \text{Normalize } Y$$

$$m(9) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t3(j),0), \text{Save positive portion of measuremen t}$$

TMNB-4 (Middle of band, middle layer, 5 Bark wide)

$$t4(j) = \frac{1}{24}\sum_{i=19}^{42} Y(i,j) - \frac{1}{24}\sum_{i=19}^{42} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t4(j), 19 \le i \le 42, 1 \le j \le N3, \text{NormalizeY}$$

TMNB-5 (Middle of band, bottom layer, 2.5 Bark wide)

$$t5(j) = \frac{1}{5}\sum_{i=7}^{11} Y(i,j) - \frac{1}{5}\sum_{i=7}^{11} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t5(j), 7 \le i \le 11, 1 \le j \le N3, \text{NormalizeY}$$

$$m(10) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t5(j),0), \text{Savepositiveportionof measurement}$$

TMNB-6 (Middle of band, bottom layer, 2.5 Bark wide)

$$t6(j) = \frac{1}{7}\sum_{i=12}^{18} Y(i,j) - \frac{1}{7}\sum_{i=12}^{18} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t6(j), 12 \le i \le 18, 1 \le j \le N3, \text{NormalizeY}$$

TMNB-7 (Middle of band, bottom layer, 2.5 Bark wide)

$$t7(j) = \frac{1}{10}\sum_{i=19}^{28} Y(i,j) - \frac{1}{10}\sum_{i=19}^{28} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t7(j), 19 \le i \le 28, 1 \le j \le N3, \text{Normalize} Y$$

$$m(11) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t7(j), 0), \text{Save postive portion of measurement}$$

TMNB-8 (Middle of band, bottom layer, 2.5 Bark wide)

$$t8(j) = \frac{1}{14}\sum_{i=29}^{42} Y(i,j) - \frac{1}{14}\sum_{i=29}^{42} X(i,j), 1 \le j \le N3, \text{Measure}$$

$$Y(i,j) = Y(i,j) - t8(j), 29 \le i \le 42, 1 \le j \le N3, \text{Normalize} Y$$

Residual Measurement

$$t9(i,j) = Y(i,j) - X(i,j), 1 \le i \le 65, 1 \le j \le N3, \text{Measure residual}$$

$$m(12) = \frac{1}{N3 \cdot 64}\sum_{i=2}^{65}\sum_{j=1}^{N3} \max(t9(i,j), 0), \text{Save postive portion of residual measurement}$$

Note that if two measurements (positive part and negative part) were retained for each of the 9 TMNBs in the structure, a total of 18 measurements would result. The hierarchical nature of the MNB structure, along with the idempotency property of the MNB leads to linear dependence among these 18 measurements. Only 7 linearly independent MNB measurements are available. These combine with the single residual measurement and the 4 FMNB measurements for a total of 12 measurements.

9.9 Linear Combination of Measurements for MNB Structure

The 12 measurements now are combined linearly to generate an AD value. The weights used in this linear combination are given in Table 2:

$$\text{AD} = \sum_{i=1}^{12} m(i) \cdot weight_i \ .$$

Table 2 - Weights for MNB Structure Linear Combination

i	$Weight_i$
1	0.0000
2	-0.0023
3	-0.0684
4	0.0744
5	0.0142
6	0.0100
7	0.0008
8	0.2654
9	0.1873
10	2.2357
11	0.0329
12	0.0000

Note that when all 12 measurements are zero, AD is zero.

Note also that it may seem inefficient to compute measurement values and multiply them by a weighting factor of zero (i.e., $weight_1$ and $weight_{12}$). However, there are two reasons for maintaining the computation of these measurements. First, the measurement vector, *m1* to *m12*, represents the complete set of linearly independent measurement values as a part of the MNB structure (see 9.8). Second, this weighting vector, $weight_i$, represents the best weightings for the applications and conditions set forth in Table 1. For other applications and conditions, a different weighting vector may be required that has non-zero values for these measurements.

10 Transformation from an objective quality scale to a subjective quality scale

The output of the algorithm described in clause 9, which is called the auditory distance (AD), indicates the degree of subjective quality degradation due to speech coding. Therefore, when estimation of subjective quality on a specific scale is not necessary, e.g., in optimizing parameters of a codec or in simply comparing the performance of codecs, the AD value itself is quite useful. To estimate subjective quality Mean Opinion Score (MOS) or equivalent-Q scales, however, the AD value is transformed as described below.

10.1 Mean opinion scores

In subjective assessment of the performance of codecs, the ACR method using the Listening Quality scale specified in ITU-T Recommendation P.800 is often used, giving subjective quality in terms of MOS. Since the relationship between the MOS and AD values is not necessarily the same for different languages or even for different subjective tests within a language, it is difficult to determine a unique function which transforms the AD value to the estimated MOS value. In practice, therefore, it is necessary to derive such transformation functions for individual languages and individual subjective tests in advance.

Note that the absolute value of the MOS depends on the context of the subjective experiment. The estimated MOS obtained by a pre-determined transformation function predicts subjective quality in the context of the subjective experiment used in deriving the transformation function.

When the results are presented in the estimated-MOS domain, the transformation function from the AD value to the MOS value should be reported.

10.2 Equivalent-Q values

It is difficult to compare the MOSs obtained in different subjective experiments since subjective judgment is affected by the experimental settings, e.g., the range of speech quality in the experiment. Therefore, the equivalent-Q value is sometimes used as a subjective quality scale. The equivalent-Q value is the Q value of MNRU (defined in ITU-T Recommendation P.810) for which the MOS is equivalent to that of coded speech.

In the objective measurement, the equivalent-Q value can be estimated directly from the AD values for coded speech and MNRU conditions, without transforming the AD value to the MOS domain (see Figure 5). When the results are presented in the estimated equivalent-Q domain, the Q vs. AD-value characteristics illustrated in Figure 4 of ITU-T Recommendation P.861 should be reported.

Note that the equivalent-Q value becomes relatively unreliable in the regions of high- and low-Q value because the Q vs. AD curve becomes almost flat in these regions. Accordingly, care should be taken when working in the Q domain with very high- and very low-quality speech.

11 Analysis of results

The analysis of objective measurement results should be carried out based on the AD value, the estimated MOS, or the estimated equivalent-Q. For each testing condition, the mean scores over male talkers, female talkers, and their average should be calculated separately and reported.

Calculation of separate standard deviations for each testing condition is not recommended. Confidence limits should be evaluated by taking into account the variation of objective quality over talkers and sentences and significance tests performed by conventional analysis-of-variance techniques.

Note that the statistical analysis described here is different from those in subjective assessment where the means of subjective quality are statistically evaluated by taking into account the variations over subjects as well as talkers

and sentences. Since the objective measure in this standard cannot estimate the distributions of subjective votes but only the mean of them, it is impossible to perform the analysis over subjects. Estimating the distributions of subjective votes is still under study. Therefore, when the analysis over subjects is necessary, subjective experiments conforming to ITU-T Recommendation P.830 should be conducted.

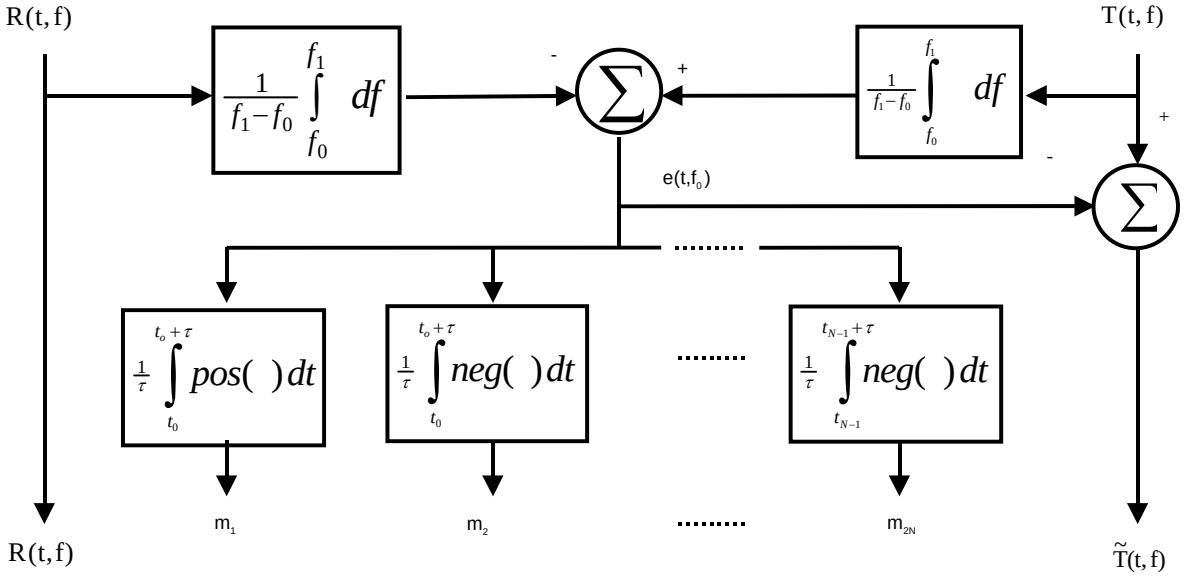


Figure 1 – Time Measuring Normalizing Block (TMNB)

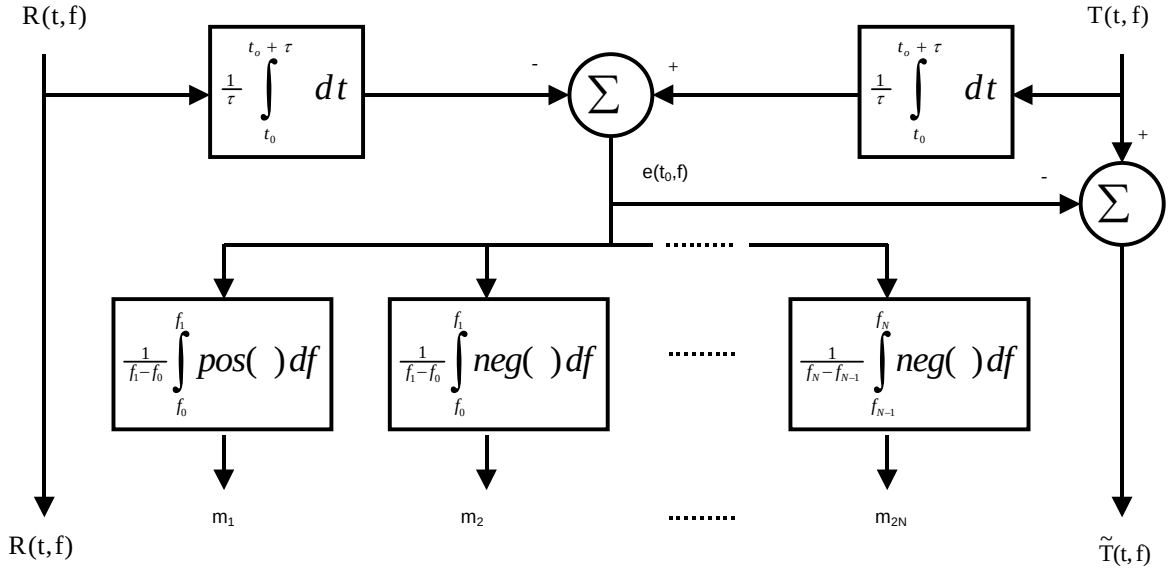
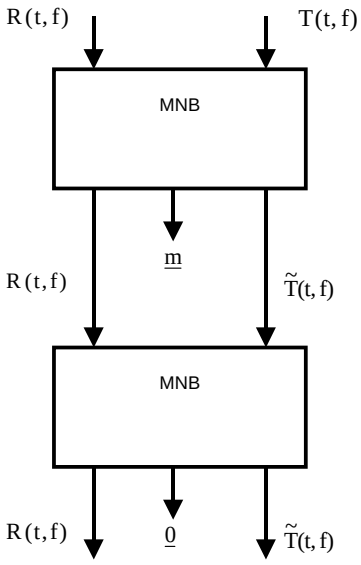


Figure 2 – Frequency Measuring Normalizing Block (FMNB)



If $MNB(R(t,f),T(t,f)) = (R(t,f),\tilde{T}(t,f),\underline{m})$,

then $MNB(R(t,f),\tilde{T}(t,f)) = (R(t,f),\tilde{T}(t,f),\underline{0})$.

Figure 3 – MNBs Are Idempotent

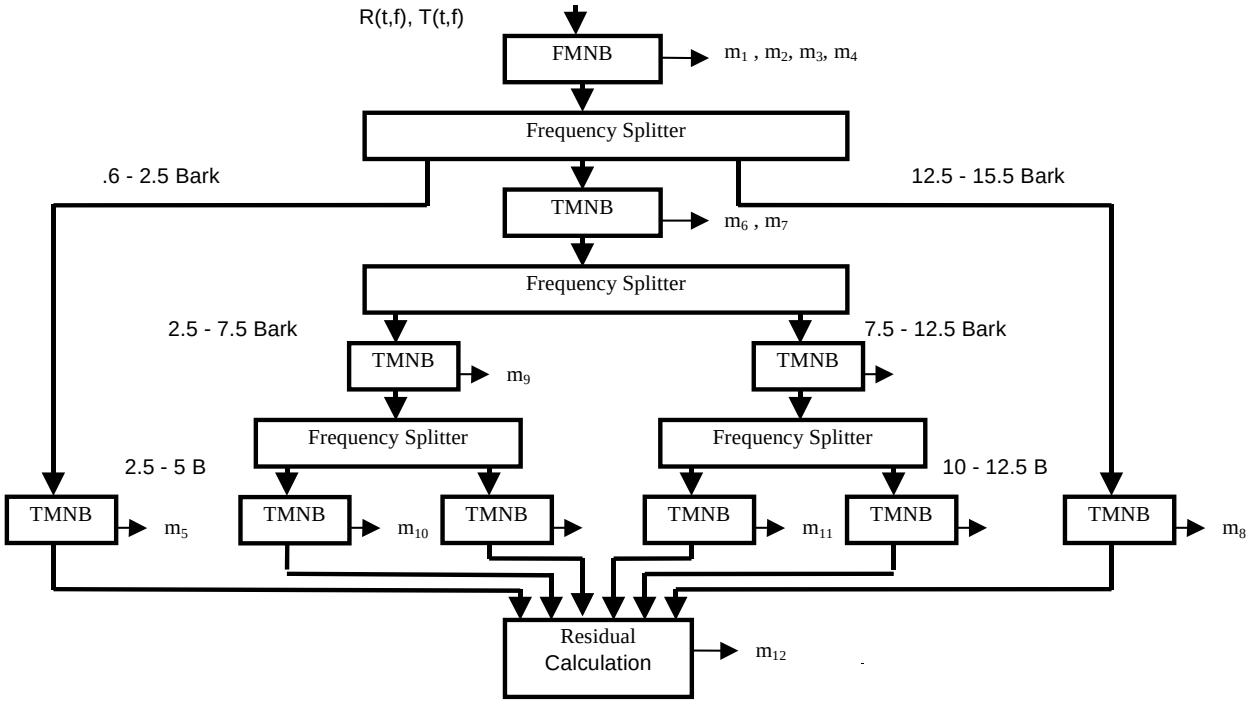


Figure 4 – MNB Hierarchical Structure

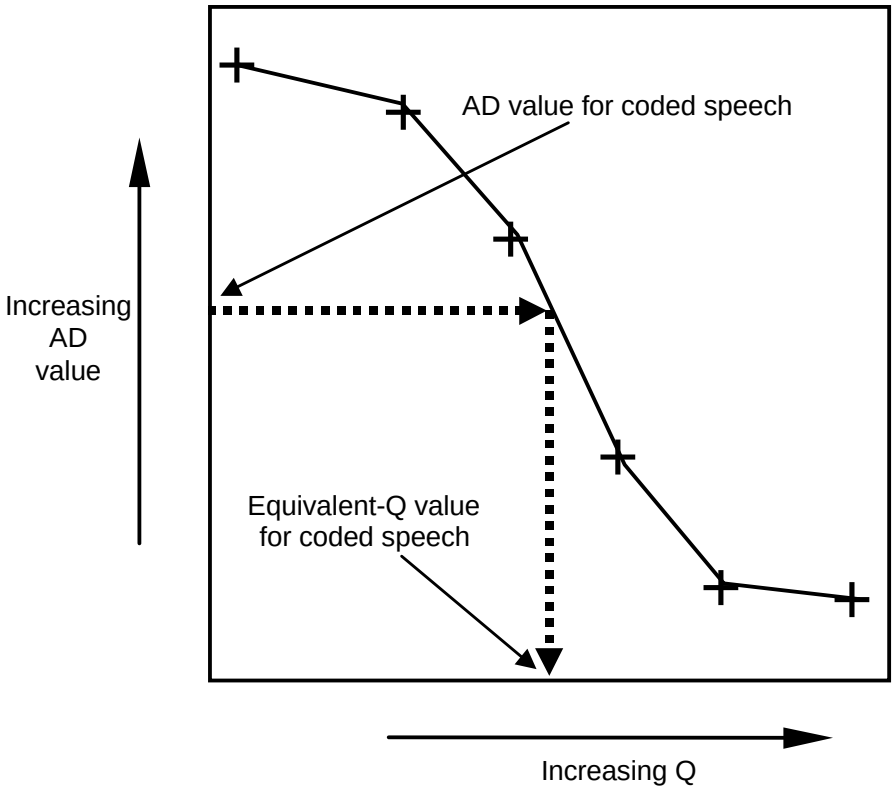


Figure 5 – Determination of equivalent-Q value of coded speech

Annex A
(informative)

Bibliography

ITU-T Recommendation G.191 – (11/1996) – *Software tools for speech and audio coding standardization*[2)]

ITU-T Recommendation G.711 – (1988) – *Pulse code modulation (PCM) of voice frequencies*, Blue Book Fasc. III. 4[2)]

ITU-T Recommendation G.726 – (12/1990) – *40, 32, 24, and 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*[2)]

ITU-T Recommendation G.728 – (09/1992) – *Coding of speech at 16 kbit/s using low-delay code excited linear prediction*[2)]

ITU-T Recommendation G.729 – (11/1995) – *Coding of speech at 8 kbit/s using Conjugate Structure Algebraic Code-Excited Liear-Prediction (CS-ACELP)*[2)]

ITU-T Recommendation P.50 – (03/1993) – *Artificial voices*[2)]

ITU-T Recommendation P.810 – (09/1995) – *Modulated Noise Reference Unit (MNRU)*[2)]

ITU-T Recommendation P.861 – (8/96) – *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*[2)]

ITU-T P-series Recommendations Supplement No. 13 – (1988) – *Noise spectra,* Blue Book, Volume V, 1988[2)]