NTIA Report 98-347

# Objective Estimation of Perceived Speech Quality Using Measuring Normalizing Blocks

**Stephen D. Voran**

**U.S. DEPARTMENT OF COMMERCE**
**William M. Daley, Secretary**

Larry Irving, Assistant Secretary
for Communications and Information

April 1998

This Page Intentionally Left Blank

This Page Intentionally Left Blank

**CONTENTS**

**Page**

This Page Intentionally Left Blank

This Page Intentionally Left Blank

# OBJECTIVE ESTIMATION OF PERCEIVED SPEECH QUALITY USING MEASURING NORMALIZING BLOCKS

Stephen Voran*

Perceived speech quality is most directly measured by subjective listening tests. These tests are often slow and expensive, and numerous attempts have been made to supplement them with objective estimators of perceived speech quality. These attempts have found limited success, primarily in analog and higher-rate, error-free digital environments where speech waveforms are preserved or nearly preserved. How to objectively measure the perceived quality of highly compressed digital speech, possibly with bit errors or frame erasures, has remained an open question. We describe a new approach to this problem, using a simple but effective perceptual transformation, and a hierarchy of measuring normalizing blocks to compare perceptually transformed speech signals. The resulting estimates of perceived speech quality were correlated with the results of nine subjective listening tests. Together, these tests include 219 4-kHz bandwidth speech encoders/decoders, transmission systems, and reference conditions, with bit rates ranging from 2.4-64 kb/s. When compared with six other estimators, significant improvements were seen in many cases, particularly at lower bit rates, and when bit errors or frame erasures were present. These hierarchical structures of measuring normalizing blocks, or other structures of measuring normalizing blocks, may also address open issues in perceived audio quality estimation, layered speech or audio coding, automatic speech or speaker recognition, audio signal enhancement, and other areas.

Key words: audio quality; distance measures; measuring normalizing blocks; objective estimation of audio quality; objective estimation of speech quality; perceptual transformations; speech coding; speech quality; subjective estimation of audio quality; subjective estimation of speech quality

## 1. BACKGROUND

Digital speech encoding and transmission involves a four-way compromise between complexity, delay, bit rate, and the perceived quality of decoded speech. Complexity, delay, and bit rate can often be quantified in fairly straightforward ways, but perceived quality can be more difficult to measure. Subjective listening or conversation tests can be used to gather firsthand evidence about perceived speech quality, but such tests are often fairly expensive, time-consuming, and labor-intensive. These

* The author is with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, 325 Broadway, Boulder, Colorado 80303.

costs are often well-justified, and there is no doubt that the most important measurements of perceived speech quality will always rely on formal subjective tests.

There are also situations where the costs associated with formal subjective tests do not seem to be justified. In particular, much speech coder/decoder (codec) development and optimization work apparently relies on objective estimators of perceived speech quality, along with "informal listening tests." Of 26 codecs described at the 1995 IEEE Workshop on Speech Coding for Telecommunications, only 11 had been tested in formal subjective tests. Segmental signal-to-noise ratio (SNRseg) or SNR was used to estimate perceived speech quality in ten cases, cepstral distance (CD) was used twice, and Bark spectral distortion (BSD) was used once [1]. Codec evaluations presented at the 1997 IEEE Workshop on Speech Coding for Telecommunications relied mainly on informal and formal subjective tests [2].

SNR and SNRseg are simple to implement, have straightforward interpretations, and can provide indications of perceived quality in some waveform-preserving speech systems. Unfortunately, as shown in this report and in [3-5], when they are used to evaluate more general coding and transmission systems, SNR and SNRseg often show little, if any, correlation to perceived speech quality. The continued popularity of these two estimators is likely due to their history, their simplicity, and the lack of a widely tested and accepted replacement. The main body of ITU-T Recommendation P.861 describes a perceived speech quality estimator called noise disturbance (ND), but its scope is limited to higher bit rate speech codecs operating over error-free channels [6]. How to objectively measure the perceived quality of highly compressed digital speech, possibly with bit errors or frame erasures has remained an open question.

Researchers have recently begun to include explicit models for some of the known attributes of human auditory perception in their estimators of perceived speech or audio quality [6-15]. The motivation for this perception-based approach is to create estimators that "hear" speech signals through the same transformations that humans hear them. In principle, this was a significant advance. In practice, when estimators are evaluated, they often show modest improvement, at best. The limitations of the perception-based approach can be traced to two sources. First, while detailed models for the detectability and perceived loudness of many different combinations of tones and narrow bands of noise have been derived, the nonlinear, time-varying nature of human hearing makes aggregating those results into practical models for the processing of more general signals (e.g., speech) a formidable task. Simplifying approximations are often made, resulting in moderately complex models that generally are not tested beyond tones and noise, if they are tested at all. Second, human perception of speech quality involves both hearing and judgment. Extensive efforts to model hearing have often been followed by relatively trivial models for judgment. Our studies have lead us to reverse this emphasis, resulting in a simple, yet effective, model for hearing, and a more sophisticated model for judgment.

A high-level description of our approach is shown in Figure 1. The delay of the device under test is first estimated and removed. The perceptual transformation contains a simple model for hearing, and the

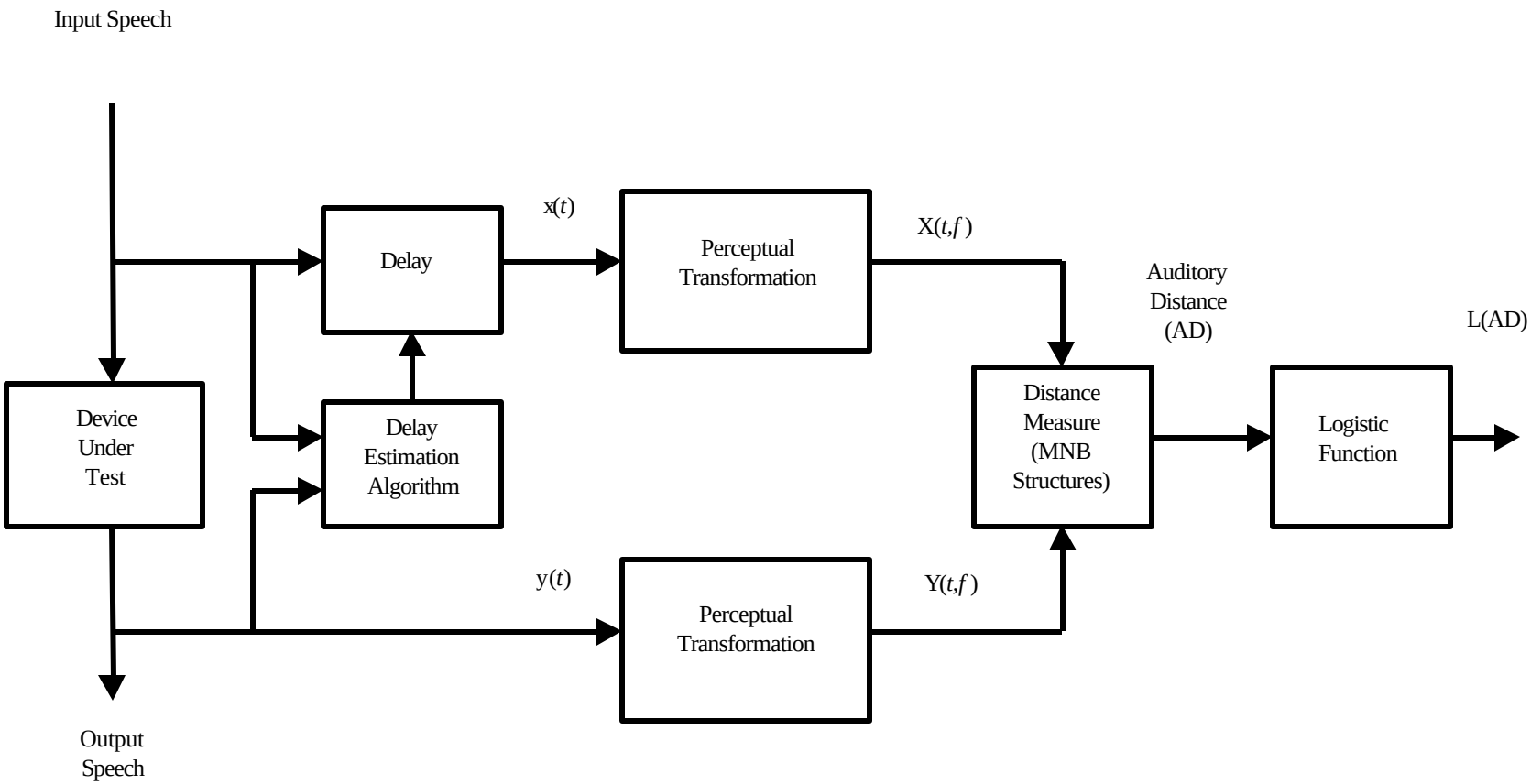


Figure 1. High-level block diagram of the objective estimation approach described in this report.

distance measure models judgment. This partition is an approximation. There is no single clean dividing line between human hearing and judgment. The distance measure generates auditory distance (AD) values. These non-negative values increase as the input speech and output speech signals move apart perceptually. A logistic function can be used to map AD into a finite interval, to better match finite subjective test results. Note that Figure 1 describes an estimation approach based on the comparison of two speech signals. This most closely parallels the subjective tests known as degradation category rating (DCR) tests. In DCR tests, listeners hear the reference and test signals sequentially, and are asked to compare them. In the simpler and more popular absolute category rating (ACR) tests, listeners hear only the test signal and are asked to rate its quality. In spite of the clear parallel to DCR tests, the approach shown in Figure 1 provides useful estimates of perceived speech quality as measured in ACR tests.

In the following sections we describe a delay estimation algorithm and a simple but effective perceptual transformation. We discuss distance measures, and the motivation behind measuring normalizing blocks (MNB's). MNB's are defined, and then combined in hierarchical structures that form distance measures. We provide evaluations of the resulting objective estimators of perceived speech quality through comparison with the results of nine subjective tests. Together, these tests include 219 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit rates ranging from 2.4-64 kb/s. When compared with six other estimators, the MNB-based estimators show significant improvements in many cases, particularly at lower bit rates, and when bit errors or frame erasures are present. Some benchmark objective estimates of perceived speech quality for standardized codecs are provided as well. The estimation algorithms are described in full detail in Appendix A.

## 2. DELAY ESTIMATION

As shown in Figure 1, the delay of the device under test must be estimated and removed prior to the estimation of perceived speech quality. Many speech codecs do not preserve speech waveforms. When waveforms are not preserved, waveform cross-correlation and other waveform-matching techniques give ambiguous or erroneous delay estimates. For this reason we have developed a two-stage delay estimation algorithm that is included in ANSI Standard T1.801.04-1997 [16]. A coarse stage uses speech envelopes, and a fine stage uses speech power spectral densities (PSD's), both of which are approximately preserved by speech codecs.

Speech envelopes are calculated in the coarse stage by rectifying speech samples and low-pass filtering them to an approximate bandwidth of 125 Hz. These envelopes are then subsampled at 250 samples/s, and cross-correlated. The peak in the smoothed cross-correlation function becomes the coarse delay estimate with an uncertainty of ±4 ms. Whenever possible, the fine stage then refines this estimate by cross-correlating the PSD's. This is done at several different times, and the locations of the resulting peaks are checked for consistency. For some speech codecs PSD's are not adequately preserved and fine estimates are not consistent. This indicates that, from a high resolution viewpoint, the delay is not constant. In these situations the coarse delay estimate, along with its inherent 4-ms uncertainty, becomes the total delay estimate.

The two-stage process is efficient because the coarse stage can search a wide range of delay values, but at low resolution. Once the coarse stage has finished its work, its low-resolution estimate provides a starting place for the fine stage that follows. The fine stage needs to search only a narrow range of delay values, consistent with the uncertainty of the coarse estimate.

## 3. PERCEPTUAL TRANSFORMATIONS

Perceptual transformations seek to model human hearing. A useful perceptual transformation will modify the representation of an audio signal in a way that is approximately equivalent to the human hearing process. The goal is to mimic human hearing so that only information that is perceptually relevant is retained. The literature of psychoacoustics is full of experimental results that describe how humans perceive tones and bands of noise. From these results, one finds several prominent properties of human hearing that might be modeled in a perceptual transformation. It is clear that the ear's frequency resolution is not uniform on the Hertz scale. It is also clear that perceived loudness is related to signal intensity in a nonlinear way. The ear's sensitivity is clearly a function of frequency, and absolute hearing thresholds have been characterized. Finally, many studies have demonstrated time- and frequency-domain masking effects.

Much less is known about how humans perceive more complex signals, such as speech. In typical models, complex signals are decomposed into simple stimuli for which human auditory perception is better understood. Internal representations for the simple stimuli are calculated, and then combined in some manner to generate an internal representation for the original signal. For example, if $E_1(f)$ is the cochlear excitation pattern due to simple stimulus 1 and $E_2(f)$ is the cochlear excitation pattern due to simple stimulus 2 then the total cochlear excitation pattern has often been modeled as

$$E_t(f) = \left[ E_1(f)^p + E_2(f)^p \right]^{\frac{1}{p}}. \quad (1)$$

However, different values of $p$ have been selected by various authors. The maximum function "$p = \infty$" is used in [17], $p = 1$ in [18-21], $p = 0.5$ in [22], and $p = 0.48$ in [23]. In [24], $p = 0.4$ is shown to be most useful when $E_t(f)$ is used to estimate the perception of coding distortions, and in [25] values of $p$ between 0.1 and 0.3 provide the best fit to experimental results. A comparative study with $p = 0.25$, 0.5, 1.0, and $\infty$ is given in [26].

We have studied many of the perceptual transformation components that have been proposed to model various attributes of the hearing process [6-15],[17-33]. By observing correlations with subjective test results, we have sought to identify the most effective perceptual transformation components, and the most appropriate level of perceptual transformation detail for perceived speech quality estimation [26,34]. We have found that simpler perceptual transformations can be as effective or more effective than more complex ones. This observation is in general agreement with [9,27]. In particular, we have found that the nonuniform frequency resolution and the nonlinear loudness perception seem to be the most important properties to model.

Thus, we have arrived at a very simple, yet effective perceptual transformation. This perceptual transformation is applied to frequency domain representations of the speech signals. Speech signals are broken into frames, multiplied by a Hamming window, and then transformed to the frequency domain using a Fast Fourier Transform (FFT). Our investigations have not identified any phase measurements

that reliably result in perceptually relevant information. Thus only the squared magnitudes of the FFT results are retained. The results that follow are based on a sample rate of 8000 samples/s, a frame size of 128 samples (16 ms) and a 50% frame overlap. We have experimented with frame sizes of 64 and 256 samples, and found them to be less useful for this application. We have also experimented with the frame overlap value, and have found this to be a less critical parameter.

The nonuniform frequency resolution of the ear is treated by the use of a psychoacoustic frequency scale. Several such scales have been proposed [20,30-33] and we have determined that for this application, the minor differences between them are not particularly significant. We have elected to use a Bark frequency scale. The Hertz scale frequency variable $f$ is replaced with the Bark scale frequency variable $b$ using the relationship

$$b = 6 \cdot \sinh^{-1}\left(\frac{f}{600}\right), \tag{2}$$

which can be found in [30]. Note that $b$ increases approximately linearly with $f$ below about 500 Hz, and $b$ increases according to a compressive nonlinearity above about 500 Hz. This scale was derived to match experimental results on critical bands in human hearing [31]. Roughly speaking, on this Bark scale, equal frequency intervals are of equal perceptual importance. We used this relationship to regroup frequency domain samples that are uniformly spaced on the Hertz scale into bands that have approximately uniform width on this Bark scale.

Many models for loudness perception as a function of signal intensity are available as well [20,24,30,31]. Again, our studies indicate that for this application, the choice of a model is not critical, as long as it contains a compressive nonlinearity. We have chosen to use a logarithm to convert signal intensity to perceived loudness.

We have also implemented models for the inner-outer ear transfer function, absolute hearing thresholds, equal loudness curves, and time- and frequency-domain masking effects. We have elected not to include these models in our perceptual transformation. While these attributes of hearing have all been well-documented in tone and noise experiments, modeling them does not appear to help with the estimation of the perceived quality of 4-kHz bandwidth speech.

## 4. DISTANCE MEASURES

Distance measures seek to measure the perceived distance between two perceptually transformed signals. Unfortunately, many existing conventional distance measures display properties that are clearly inconsistent with human auditory judgment. As an example, consider a distance measure that takes the form

$$D[X(f),Y(f)]=\left[\frac{1}{\Omega}\int\left|X(f)-Y(f)\right|^{g}\,df\right]^{1/g}, \tag{3}$$

where $X(f)$ and $Y(f)$ are frequency-domain representations of the input and output of the device under test, respectively, and the integration is over some band of interest with bandwidth $\Omega$. Such distance measures are invariant to the sign of the difference $X(f)-Y(f)$. This means that the hissy signal $Y_1(f)$ and the muffled signal $Y_2(f)$ in Figure 2 will received the same distance value, which would not generally be a perceptually consistent result.

For a second example, consider the more refined distance measure

$$D[X(f),Y(f)] = \left[\frac{1}{\Omega_p}\int_{Y(f)\geq X(f)} w_p(f)\left(X(f)-Y(f)\right)^{g_p}\,df\right]^{1/g_p} + \left[\frac{1}{\Omega_n}\int_{Y(f)<X(f)} w_n(f)\left(X(f)-Y(f)\right)^{g_n}\,df\right]^{1/g_n}. \tag{4}$$

In (4) the sign of $Y(f)-X(f)$ is acknowledged, with separate integrations, integration exponents $\gamma$, and weighting functions $w(f)$. With the signals $X(f)$, $Y1(f)$, and $Y2(f)$ shown in Figure 3, $D[X(f), Y_1(f)] = D[X(f),Y_2(f)]$. This is unlikely to be a perceptually consistent result, because $Y_1(f)$ has a harsh sound, while $Y_2(f)$ has a hollow sound. Analogous examples exist for undesired time-domain invariances.

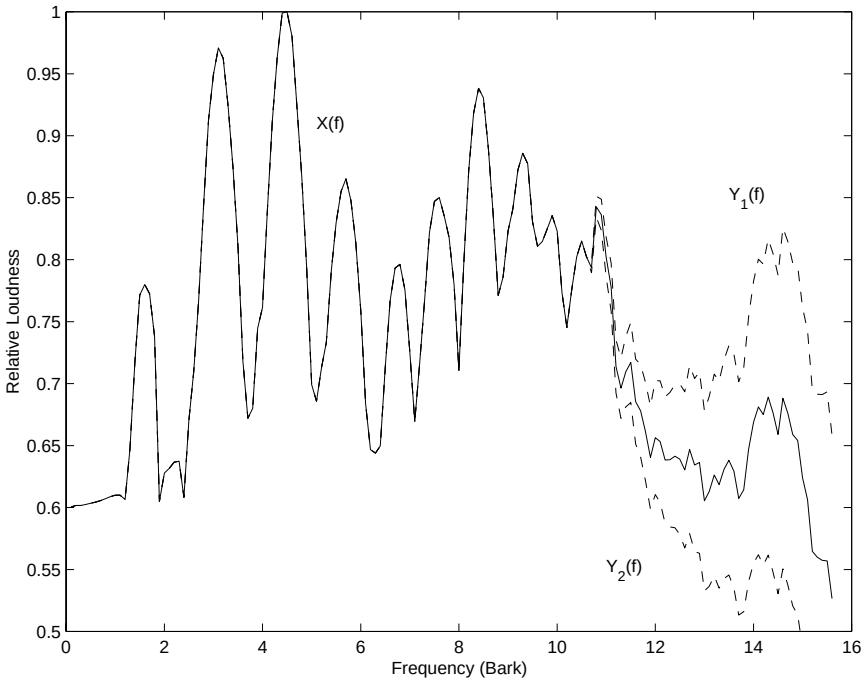


Figure 2. Distance measure invariance example 1.

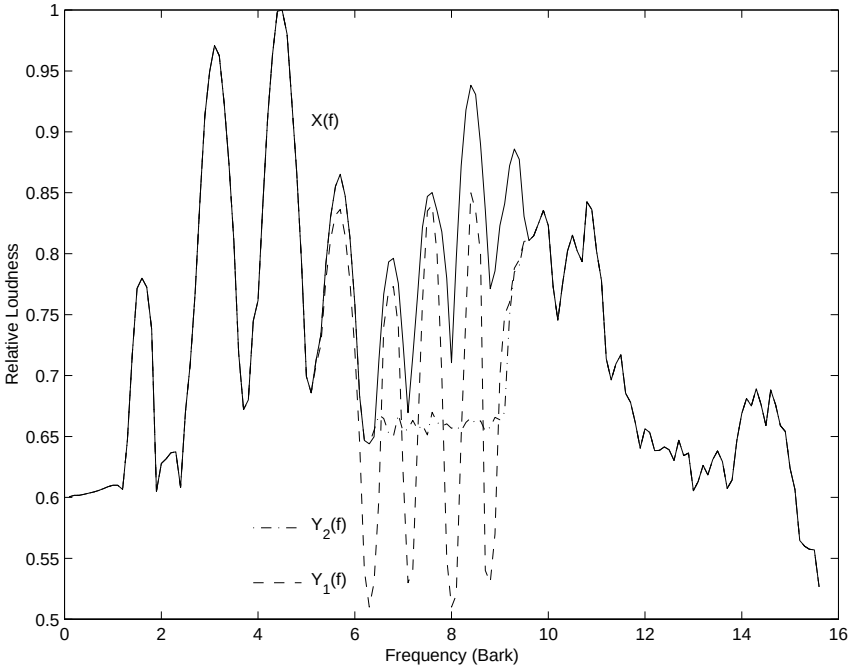


Figure 3. Distance measure invariance example 2.

## 5. MEASURING NORMALIZING BLOCKS

Based on our studies of conventional distance measures, and our understanding of human hearing and judgment, we concluded that listeners adapt and react differently to spectral deviations that span different time and frequency scales. We further observed that for the speech quality estimation application, maximal perceptual consistency over a wide range of distortion types requires a family of analyses that cover multiple frequency and time scales. The spectral deviations at one scale must be removed so they are not counted again as part of the deviations at other scales. We also concluded that working from larger to smaller scales is most likely to emulate listeners' patterns of adaptation and reaction to spectral deviations. In light of these findings, we elected to form a distance measure from a hierarchy of time and frequency measuring normalizing blocks.

A time measuring normalizing block (TMNB) is shown in Figure 4 and a frequency measuring normalizing block (FMNB) is given in Figure 5. Each of these blocks takes perceptually transformed input and output signals (X( *f*,*t*) and Y( *f*,*t*), respectively) as inputs, and returns a set of measurements and a normalized version of Y( *f*,*t*). The TMNB integrates over some frequency scale, then measures differences and normalizes the output 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 output signal at multiple **frequencies**. Finally, the positive and negative portions of the measurements are integrated over frequency.

We now formalize the MNB definitions. The TMNB operating on the band that extends from $fl$ to $fu$ using the measurement time intervals defined by $t_i$, $i$=0 to $N$, normalizes $\mathrm{Y}(f,t)$ to $\tilde{\mathrm{Y}}(f,t)$ and generates $2N$ measurements $\mathbf{m}(j)$:

$$\tilde{\mathrm{Y}}(f,t) = \mathrm{Y}(f,t) - \mathrm{e}(fl,t)\ ,$$

$$\mathbf{m}(2i-1) = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \max(\mathrm{e}(fl,t),0)\,dt\ ,$$

$$\mathbf{m}(2i) = \frac{-1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \min(\mathrm{e}(fl,t),0)\,dt, \quad i = 1 \text{ to } N\ , \tag{5}$$

$$\text{where } \mathrm{e}(fl,t) = \frac{1}{fu - fl} \int_{fl}^{fu} \mathrm{Y}(f,t)\,df - \frac{1}{fu - fl} \int_{fl}^{fu} \mathrm{X}(f,t)\,df\ .$$

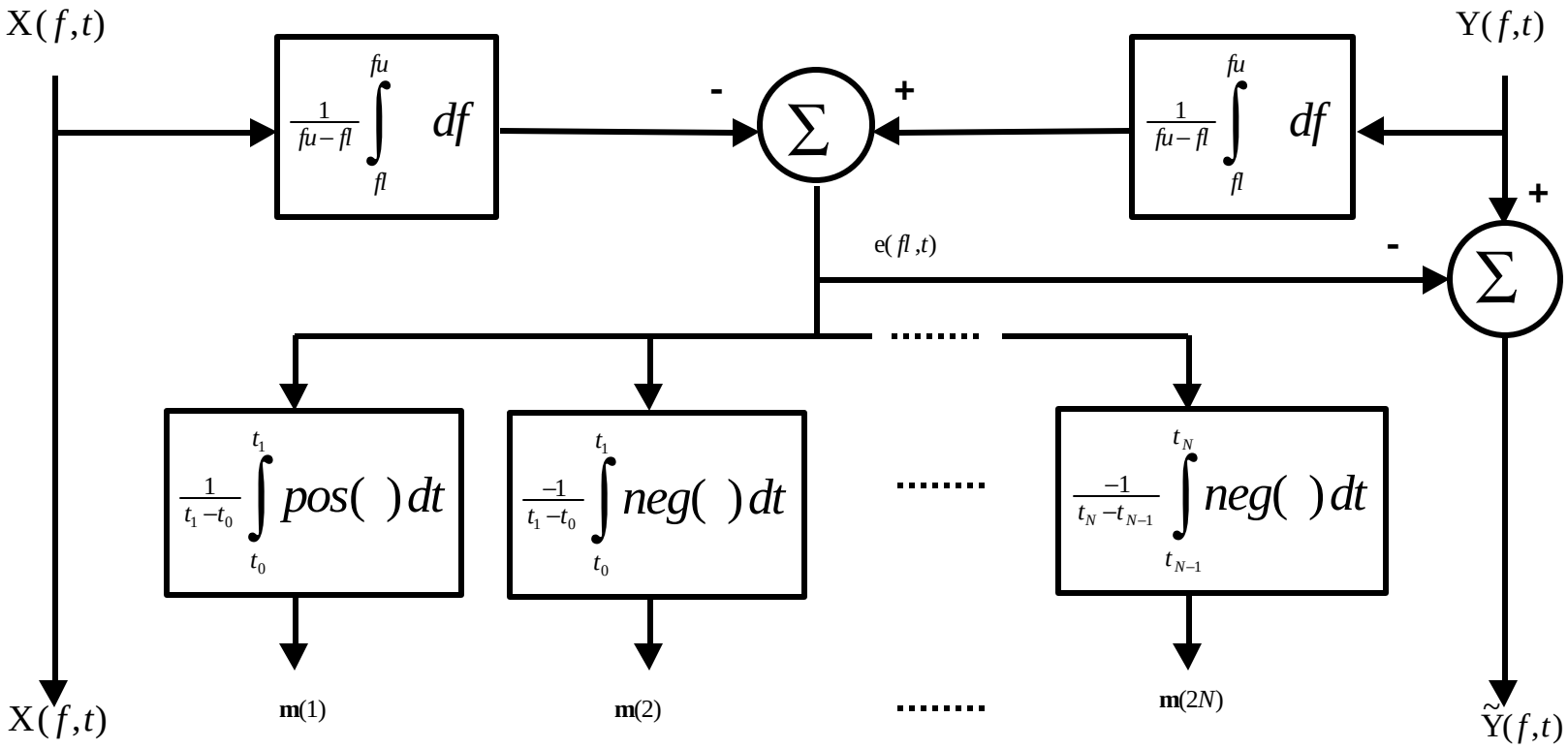


Figure 4. A time measuring normalizing block (TMNB).

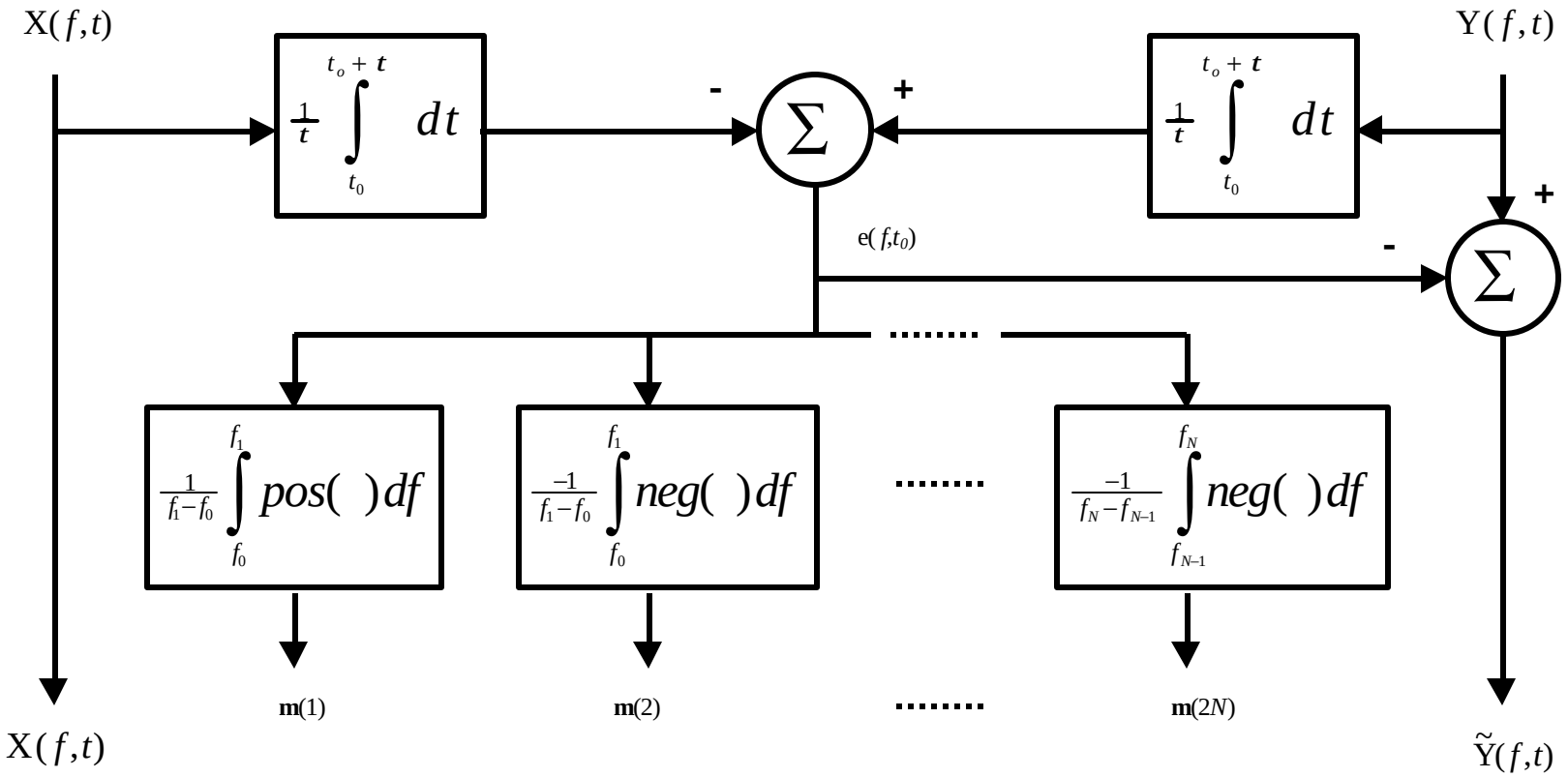


Figure 5. A frequency measuring normalizing block (FMNB).

The FMNB definition is analogous, with the roles of time and frequency exchanged. At time $t_0$ , the FMNB operating over time scale $\tau$, using the measurement bands defined by $f_i$, $i$=0 to $N$, normalizes $Y(f,t)$ to $\tilde{Y}(f,t)$ and generates $2N$ measurements $\mathbf{m}(j)$:

$$\begin{aligned}
&\tilde{Y}(f,t) = Y(f,t) - e(f,t_0)\ , \\
&\mathbf{m}(2i-1) = \frac{1}{f_i - f_{i-1}} \int_{f_{i-1}}^{f_i} \max(e(f,t_0),0)\,df\ , \\
&\mathbf{m}(2i) = \frac{-1}{f_i - f_{i-1}} \int_{f_{i-1}}^{f_i} \min(e(f,t_0),0)\,df\ , \quad i = 1 \text{ to } N\ , \\
&\text{where } e(f,t_0) = \frac{1}{t} \int_{t_0}^{t_0+t} Y(f,t)\,dt - \frac{1}{t} \int_{t_0}^{t_0+t} X(f,t)\,dt\ .
\end{aligned} \tag{6}$$

By design, both types of MNB's are idempotent.

$$\text{If MNB}(X,Y) = (X,\tilde{Y},\mathbf{m}),\ \text{then}\ \ \text{MNB}(X,\tilde{Y}) = (X,\tilde{Y},\mathbf{0}). \tag{7}$$

In other words, a second pass through a given MNB will not further alter the output signal, and the vector of measurements resulting from that second pass will contain only zeros. The idempotency of MNB's allows them to be cascaded and yet they measure the deviation at a given time or frequency scale once and only once.

### 5.1 Distance Measures that Use Measuring Normalizing Blocks

In order to measure spectral deviations at multiple time and frequency scales, we have formed hierarchical structures of TMNB's and FMNB's, that operate at decreasing scales. In these structures, spectral deviations at one time or frequency scale are measured and removed before the next smaller scale is considered. When used as distance measures in conjunction with the simple perceptual transformation described above, this top-down approach appears to do a good job of emulating listeners' patterns of adaptation and reaction to spectral deviations. A generalized diagram of these structures is shown in Figure 6. Each MNB in the structure generates a measurement vector $\mathbf{m}_{i,j}$. Two specific structures are shown in Figures 7 and 8. These are referred to as MNB structure 1 and MNB structure 2, respectively. As always, a complexity-performance trade-off is at work here. These two structures were chosen for their balance of relatively low complexity and high performance as estimators of perceived speech quality across a wide range of conditions and quality levels. Other MNB structures may be more appropriate for more specific speech or audio quality estimation applications. In addition, these structures or other MNB structures may address open issues in perceived audio quality

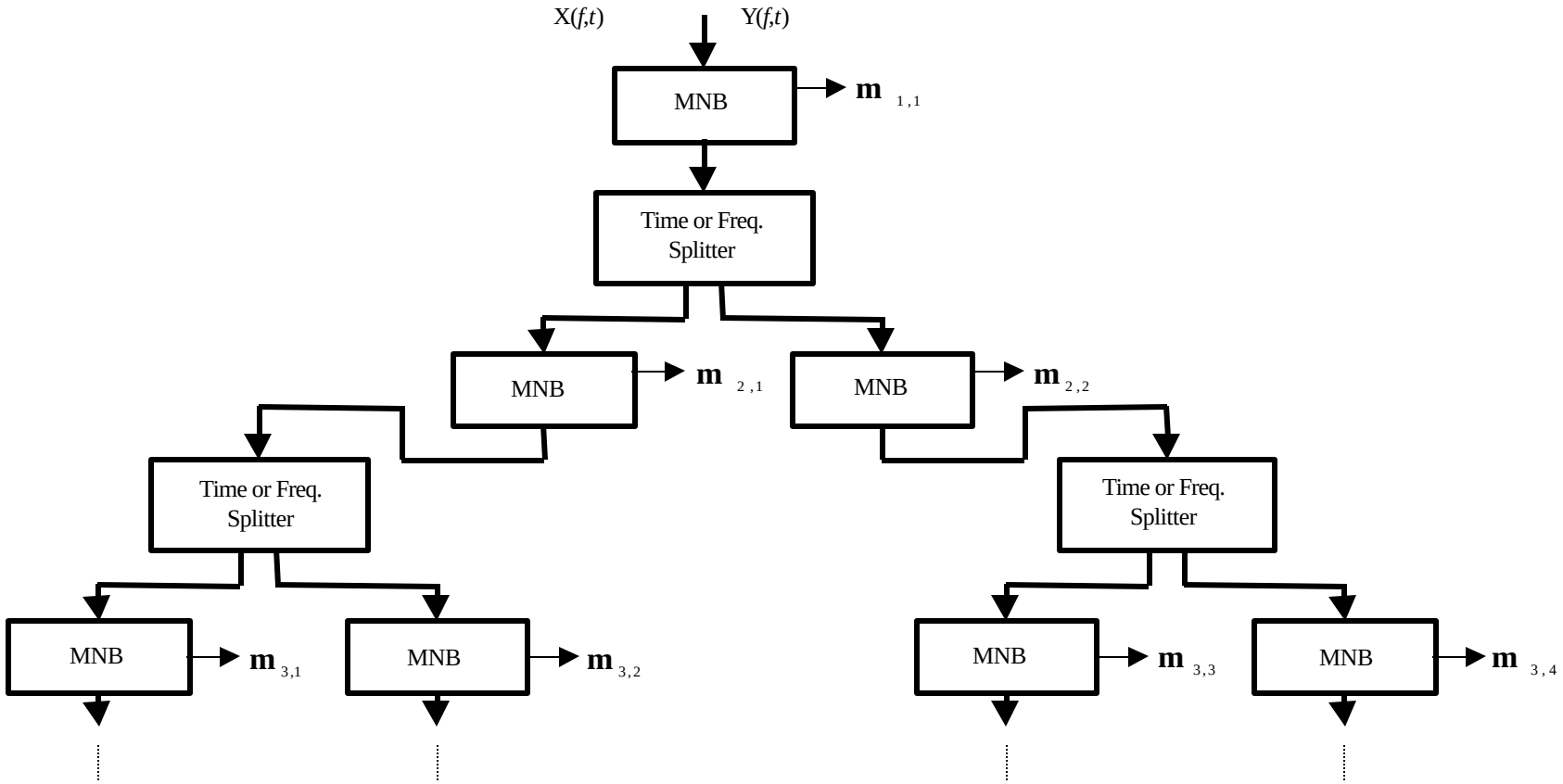


Figure 6. Generalized measuring normalizing block (MNB) structure.

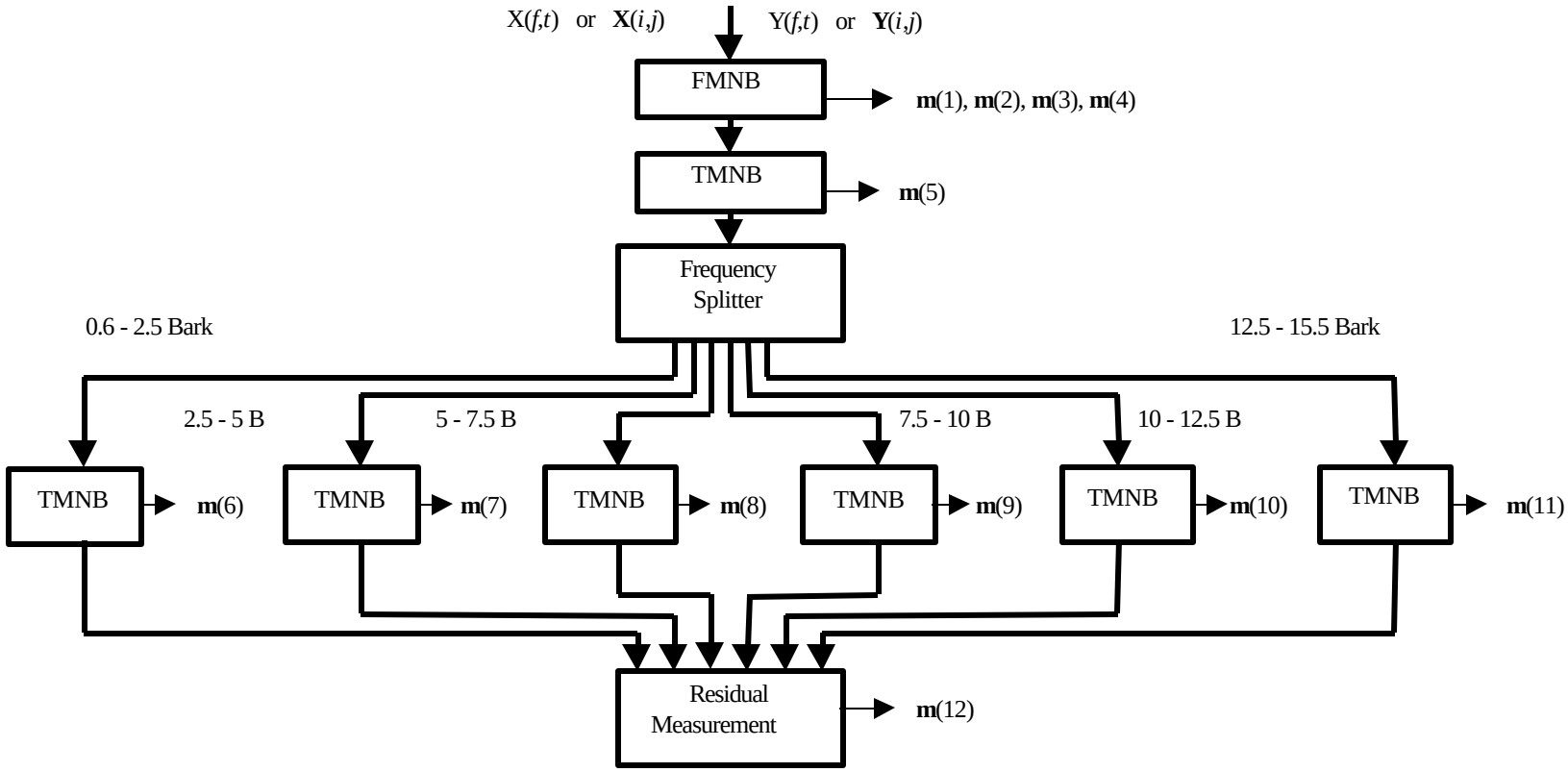


Figure 7. MNB structure 1.

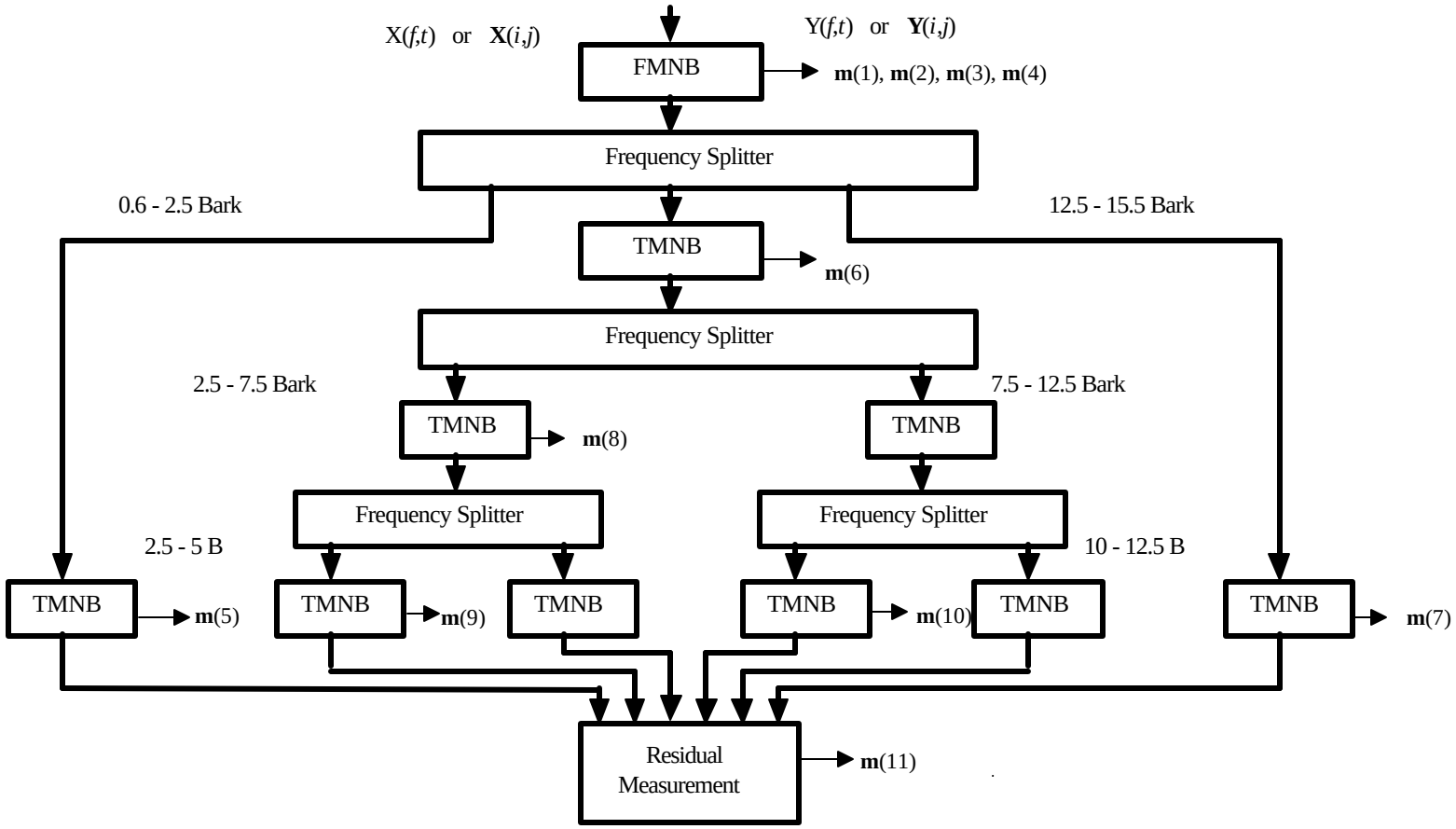


Figure 8. MNB structure 2.

estimation, layered speech or audio coding, automatic speech or speaker recognition, audio signal enhancement, and other areas.

Both MNB structures start with an FMNB that is applied to the input and output signals at the longest available time scale. Four measurements are extracted and stored in the measurement vector **m**. These measurements cover the lower and upper band edges of telephone band speech (0-500 Hz and 3000-3500 Hz.) In MNB structure 1, a TMNB is then applied to the input and output signals at the largest frequency scale (approximately 15 Bark). Six additional TMNB's are then applied at a smaller scale (approximately 2-3 Bark). Finally a residual measurement is made.In MNB structure 2, the middle portion of the band undergoes two 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.

The perceptual transformation and the MNB structures are described together in full detail in Appendix A. The idempotence of the MNB along with the hierarchical nature of MNB structures leads to linear dependence among the MNB measurements. As shown in Figures 7 and 8, only linearly independent measurements are retained. Thus, MNB structure 1 results in 12 measurements, while MNB structure 2 results in 11 measurements. For these two structures, a full set of linearly independent measurements can be formed from just the positive portions of the error functions e( $f,t$). These are the odd-numbered measurements in (5) and (6). Linear combinations of these measurements provide good estimates of the perceptual distance between two speech signals and good estimates of perceived speech quality. The value that results from this linear combination is called auditory distance (AD):

$$\mathrm{AD} = \mathbf{w}^{\mathrm{T}} \cdot \mathbf{m}, \tag{8}$$

where **w** is a length 12 (MNB structure 1) or 11 (MNB structure 2) vector of weights. In practice, AD values are non-negative. When the input and output signals are identical, all measurements are zero and AD is zero. As the input and output signals move apart perceptually, AD increases.

## 6. ESTIMATION OF PERCEIVED SPEECH QUALITY

### 6.1 Logistic Function

MNB structures 1 and 2 were designed to be used as distance measures. The AD distance values they generate were intended to be used to estimate perceived speech quality. Subjective perceived speech quality ratings usually cover finite ranges. The mean opinion score (MOS) scale is often used in ACR tests, while the degradation mean opinion score (DMOS) scale is very popular for DCR tests. Both of these scales cover the interval from 1 to 5. Thus, correlation with these subjective rating scales may be increased by mapping AD values into a finite range. We use the logistic function with asymptotes at 0 and 1:

$$L(z) = \frac{1}{1 + e^{a \cdot z + b}} \quad . \tag{9}$$

When $a > 0$, $L(z)$ is a decreasing function of $z$.

### 6.2 Correlation with Subjective Test Results

To judge the usefulness of the L(AD) values as estimators of relative perceived speech quality, we compared L(AD) and six other established objective estimators of speech quality with the results of formal subjective tests. Nine ACR tests that use the MOS scale tests were available to us, and they are summarized in Table 1. While the objective estimator structure more closely parallels DCR subjective tests, only ACR subjective tests were available for this study. Together, these 9 tests include 219 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit rates ranging from 2.4-64 kb/s, and some analog conditions as well. Both flat and intermediate reference system (IRS)-filtered speech material [35] was included. IRS filtering simulates the sending response of a typical telephone handset. A total of 22 hours of speech from at least 52 different speakers, both male and female, in three different languages was used. This collection of speech files and scores has allowed us to complete one of the most comprehensive tests of objective estimators of perceived relative speech quality.

The six established estimators are SNR [4], SNRseg [4], perceptually weighted SNRseg (PWSNRseg) [36], CD [4], BSD [15], and ND as defined in the main body of ITU-T Recommendation P.861 [6]. To create a uniform comparison, each of these estimators was passed through the logistic function in (9). For each estimator, the constants $a$ and $b$ were selected to maximize the coefficient of correlation between the logistic function output and MOS across the nine subjective tests. The maximizing values of $a$ and $b$ are shown in Table 2. The resulting coefficients of correlation are shown in Table 3.

The correlation values in Table 3 were calculated after averaging all available subjective scores for each condition to a single score for that condition. Similarly, for each condition, all available objective estimates were averaged to generate a single objective estimate for that condition. Thus, we refer to

Table 1. Summary of Material in Nine Subjective Tests

| Test | Number of Conditions | Conditions [1,2] | Filtering of Input Speech | Language | Talkers per Condition | Files | Minutes |
|---|---|---|---|---|---|---|---|
| 1 | 22 | PCM: 64, 48, 40 kb/s<br>ADPCM: 32 kb/s, x1, 2, 3, 4<br>APC: 16 kb/s, 2 versions<br>Proprietary Codec: 16 kbps<br>SELP: 4.8 kb/s, 2 versions<br>LPC: 2.4 kb/s<br>MNRU: 6 levels<br>Narrow-Band MNRU: 3 levels | None | North American English | 4 | 176 | 8 |
| 2 | 35 | PCM: 64 kb/s<br>Proprietary CELP A: 8 kb/s, over 9 RF channels, bit errors and frame erasures<br>Proprietary CELP B: 8 kb/s, over 9 RF channels, bit errors and frame erasures<br>AMPS over 9 RF channels<br>MNRU: 7 levels | IRS filtered | North American English | 6 | 1050 | 100 |
| 3 | 27 | ADPCM: 32 kb/s, clear and bit errors<br>CVSD: 32, 16 kb/s, clear and bit errors<br>VSELP: 8 kb/s<br>CELP: 4.8 kb/s, clear and bit errors<br>IMBE: 4.8, 2.4 kb/s<br>STC A: 4.8, 2.4 kb/s, clear and bit errors<br>STC B: 2.4 kb/s<br>LPC: 2.4 kb/s, clear and bit errors<br>POTS<br>MNRU: 8 levels | None | North American English | 6 | 1994 | 225 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4 | 38 | ADPCM: 32 kb/s, x4<br>LD-CELP: 16 kb/s<br>VSELP: 8 kb/s<br>Proprietary Non-Waveform Codec: 6.4 kb/s<br>Proprietary Non-Waveform Codec: 4 kb/s, 3 input levels<br>Proprietary Non-Waveform Codec: 4 kb/s, x2<br>Proprietary Non-Waveform Codec: 4 kb/s + ADPCM: 32 kb/s<br>Proprietary Non-Waveform Codec: 4 kb/s + VSELP: 8 kb/s<br>Proprietary Non-Waveform Codec: 4 kb/s + RPE-LTP: 13 kb/s<br>Proprietary Non-Waveform Codec: 4 kb/s + LD-CELP: 16 kb/s + LD-CELP: 16 kb/s<br>MNRU: 7 levels | Both IRS filtered and unfiltered | North American English | 8 | 2432 | 264 |
| 5 | 20 | PCM: 64 kb/s, x1, 2, 4, 8, 16<br>ADPCM: 32 kb/s, x1, 2, 4<br>G.728 Candidate 16 kb/s, x1, 2, 4<br>MNRU: 9 levels | IRS filtered | North American English | 4 | 1440 | 206 |
| 6 | 20 | Same as test 5 | IRS filtered | Japanese | 4 | 1440 | 188 |
| 7 | 20 | Same as test 5 | IRS filtered | Italian | 4 | 1440 | 131 |
| 8 | 47 | LD-CELP: 16 kb/s<br>8 CELP Codecs: ≅13 kb/s, frame error rates 0, 1, 2, 3, 5%<br>MNRU: 6 levels | IRS filtered | North American English | 8 | 1360 | 136 |
| 9 | 30 | VSELP: 8 kb/s, 11 simulated radio environments<br>ACELP: 8 kb/s, 11 simulated radio environments<br>PCM: 64 kb/s<br>CELP: 4.8 kb/s<br>POTS<br>MNRU: 5 levels | Both IRS filtered and unfiltered | North American English | 8 | 480 | 54 |

[1] The notation “xN” is used to indicate N passes through the indicated device.

[2] The notation “codec1 + codec2” is used to indicate that two different codecs were tandemed to create a single condition.

Table 2. Optimized Values of Logistic Function Parameters

| Objective Estimator | *a* | *b* |
|---|---|---|
| SNR | -0.0552 | -0.3490 |
| SNRseg | -0.0542 | -0.3927 |
| PWSNRseg | -0.1073 | 0.1910 |
| CD | 0.4175 | -1.8274 |
| BSD | 6.3081 | -0.7434 |
| ND | 0.5567 | -1.7450 |

Table 3. Per-condition Coefficients of Correlation Between Subjective Scores and Objective Estimators

| Test | L(SNR) | L(SNRseg) | L(PWSNRseg) | L(CD) | L(BSD) | L(ND) |
|---|---|---|---|---|---|---|
| 1* | .333 | .381 | .393 | .486 | .825 | .928 |
| 2* | .526 | .522 | .620 | .729 | .731 | .941 |
| 3* | .295 | .494 | .507 | .617 | .368 | .793 |
| 4* | .247 | .221 | .636 | .789 | .863 | .973 |
| 5 | .226 | .267 | .523 | .948 | .919 | .986 |
| 6 | .271 | .313 | .502 | .933 | .850 | .986 |
| 7 | .317 | .340 | .542 | .975 | .892 | .976 |
| 8* | .556 | .381 | .605 | .671 | .801 | .858 |
| 9* | .433 | .326 | .544 | .838 | .712 | .827 |

* These tests include conditions that are outside the defined scope of the ND algorithm.

these correlation values as "per-condition" correlations. A more advanced analysis technique, described in [37], recognizes the importance of the distributions of the objective estimates and the subjective scores for each condition, how they influence confidence intervals, and in turn, the final conclusions that one draws from objective and subjective tests.

Table 3 demonstrates the limitations of SNR, SNRseg, and PWSNRseg as estimators of perceived speech quality. CD and BSD tend to show higher correlations for tests 5, 6, and 7, which contain only conditions that tend to preserve waveforms. L(ND) appears to be the most reliable of these six existing objective estimators, across these nine tests. Since tests 1-4, 8, and 9 contain conditions that are outside of the defined scope of the ND algorithm, we conclude that this algorithm can sometimes make useful estimates outside of its scope. Because L(ND) appears to be the most reliable of these six objective estimators, we use it as the reference against which to compare L(AD).

Table 4 shows per-condition correlation values for L(AD) as calculated by the two MNB structures. Since L(ND) is used as a reference, that column from Table 3 is repeated as column 2 of Table 4 to allow for easy comparisons. Two versions of the estimators were evaluated. These versions differ only in the values of the weights used in (8), and the constants used in (9).

The first version of each estimator was created by optimizing variables in (8) and (9) to maximize correlation between L(AD) and MOS across tests 1 and 2 only. The parameter *a* in (9) was absorbed into the weights in (8), resulting in 13 or 12 free variables. These variables were used to fit 1,226 data points, so the fitting problem was over-determined by an approximate factor of 100. The resulting correlation values are shown in Table 4, columns 3 and 4. These columns show that this limited optimization results in an objective speech quality estimator that generalizes well to the other seven tests. This result is important because it indicates that these estimators do model perception and judgment, rather than inadvertently modeling some specific properties of the conditions in tests 1 and 2.

To create the most effective estimator, one must use all available data. Thus, we created a second version of each estimator by optimizing variables in (8) and (9) to maximize correlation across all nine tests. This involved fitting 11,812 data points, so the fitting problem was over-determined by a factor greater than 900. The resulting correlations are shown in columns 5 and 6 of Table 4. When all nine tests are considered together, MNB structure 2 appears to be slightly more useful than MNB structure 1. Both structures show dramatic improvements over L(ND) on tests 3, 8 and 9, which contain the lower rate speech codecs, bit error, and frame erasure conditions. We have provided four scatter plots to allow for visual interpretations of per-condition correlation values. Each plot shows an objective estimator vs MOS, using a single point per condition. Four cases were selected to display a range of correlation values. Figure 9 shows L(BSD) for test 3 where the per-condition correlation, $\rho$, is .368. Figure 10 shows L(ND) for test 3 where $\rho$=.793. Figure 11 gives L(AD) using the fully optimized MNB structure 2, also on test 3, with $\rho$=.959. Finally, Figure 12 shows L(AD) using the fully optimized MNB structure 1, on test 5, where $\rho$=.986.

Table 4. Per-condition Coefficients of Correlation Between Subjective Scores and Objective Estimators

| Test | L(ND) | L(AD) | | | |
|---|---|---|---|---|---|
| | | MNB-1 | MNB-2 | MNB-1 | MNB-2 |
| | | Weights optimized using only tests 1 and 2. | | Weights optimized using tests 1-9. | |
| 1 | .928 | .931 | .928 | .932 | .956 |
| 2 | .941 | .965 | .963 | .951 | .945 |
| 3 | .793 | .939 | .944 | .935 | .959 |
| 4 | .973 | .964 | .979 | .977 | .976 |
| 5 | .986 | .955 | .963 | .986 | .984 |
| 6 | .986 | .965 | .969 | .983 | .982 |
| 7 | .976 | .967 | .971 | .980 | .984 |
| 8 | .858 | .954 | .953 | .936 | .961 |
| 9 | .827 | .921 | .923 | .910 | .942 |

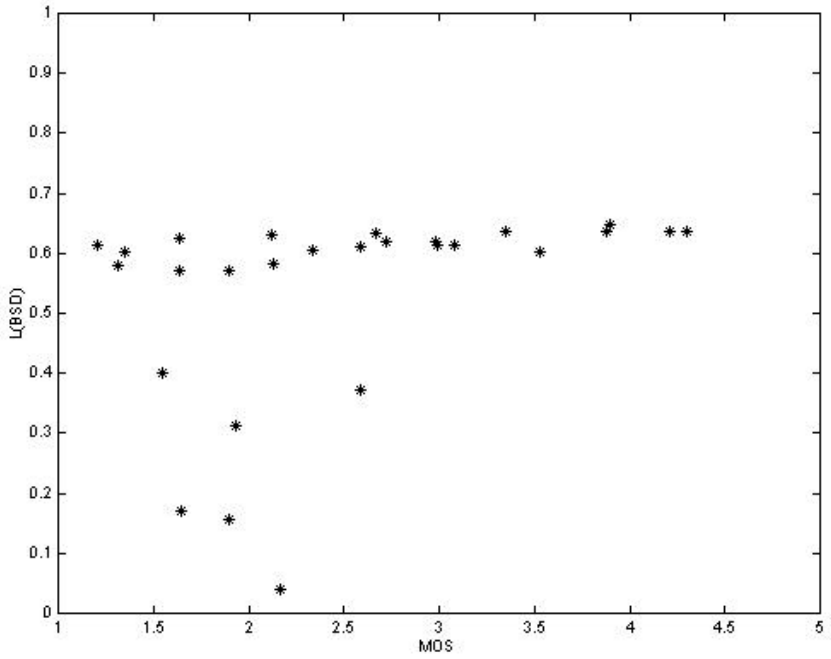


Figure 9. L(BSD) as an estimator of perceived speech quality on test 3, ρ=.368.

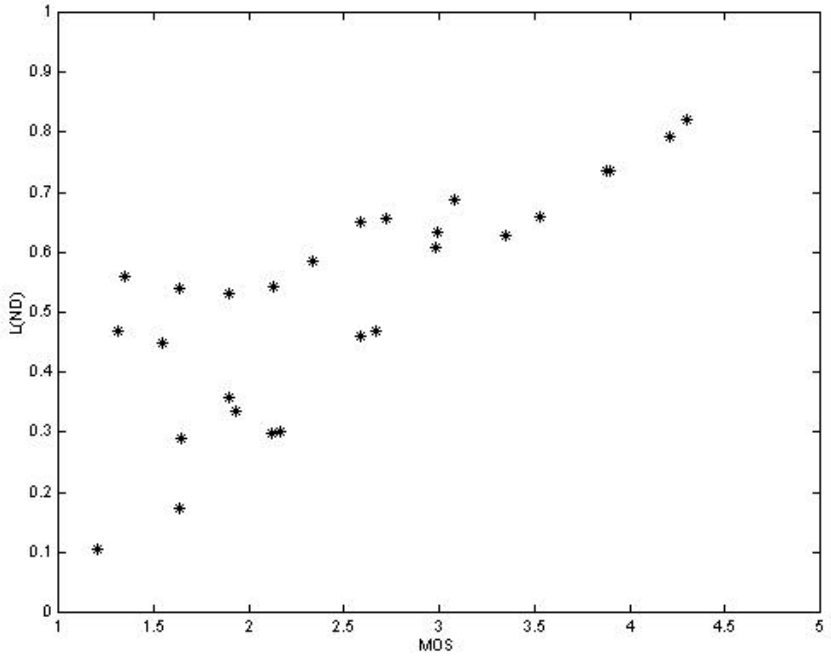


Figure 10. L(ND) as an estimator of perceived speech quality on test 3, ρ=.793.

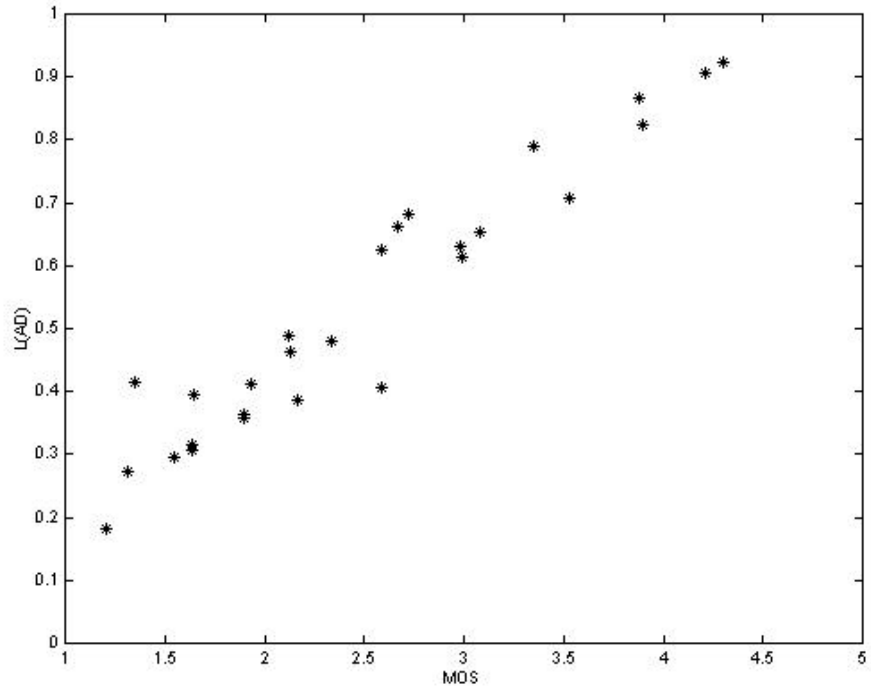


Figure 11. MNB structure 2 as an estimator of perceived speech quality on test 3, $\rho$=.959.

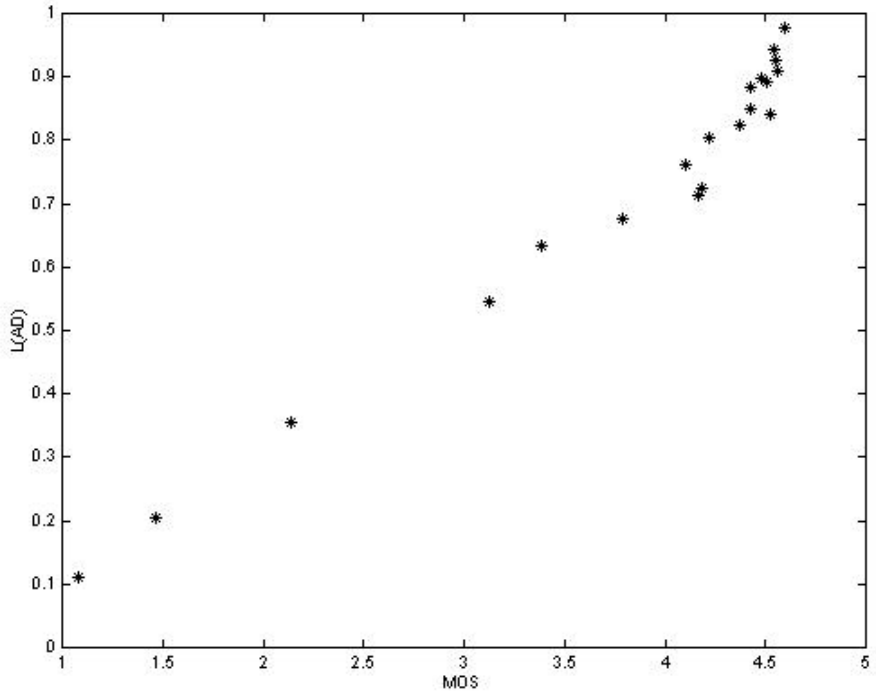


Figure 12. MNB structure 1 as an estimator of perceived speech quality on test 5, $\rho$=.986.

### 6.3 Observations and Discussion

The optimized values of the variables in (8) and (9) are given in Table A-1. Because the measurements have different variances, the weights do not indicate the relative importance of the measurements. Note that one weight in Table A-1 is zero, indicating that the first measurement in MNB structure 2 does not presently provide useful information for this application. We retain this measurement for completeness, and for its potential future utility in this or other applications. In both structures, the first four weights are applied to FMNB measurements taken at the edges of the speech band. For MNB structure 1, $\mathbf{w}(1)$ and $\mathbf{w}(2)$ indicate that to maximize estimated speech quality, energy below 250 Hz (outside the telephony speech passband) should be minimized, but only if energy above 250 Hz can be retained. Similarly, $\mathbf{w}(3)$ and $\mathbf{w}(4)$ indicate that energy above 3250 Hz should be minimized, but not at the expense of energy below 3250 Hz. These data-driven mathematical results agree with our intuitions about in-band speech power and out-of-band noise. As part of a sensitivity analysis, we determined that when the weights in $\mathbf{w}$ are perturbed from their optimal values by 10%, resulting coefficients of correlation are reduced by about 1%. In addition, 1% and 0.1% perturbations in the weights result in 0.1% and 0.01% reductions in correlation, respectively.

Table 4 shows that correlations between fully optimized MNB estimators and subjective scores range from .910 to .986. Given the breadth of conditions covered by the nine tests, these are very encouraging results. In particular, the improved ability to estimate perceived speech quality for lower rate speech codecs, some of which are operating with bit errors or frame erasures, represents an important advance. Based on this improvement, ITU-T Recommendation P.861 has been updated by the inclusion of an MNB algorithm in Appendix II of the Recommendation [38]. The algorithm that appears there is an earlier version of MNB structure 2 described in this report.

For unoptimized software implementations, we found that either of the MNB-based estimators requires approximately 920,000 floating-point operations to process 1 second (8,000 samples) of speech. Because the bulk of these operations is devoted to the FFT, both MNB algorithms can be run at the same time using only 940,000 floating-point operations. Similarly, an unoptimized implementation of the ND algorithm required about 1.21 million floating-point operations to process 1 second (8,000 samples) of speech.

We have also implemented the MNB estimators with the frame overlap reduced from 50% to zero. This reduces the number of computations in the unoptimized implementation by a factor of two but has surprisingly little impact on estimator performance for the conditions described in Table 1. When the frame overlap is reduced to zero and the parameters given in Table A-1 are optimized, the resulting coefficients of correlation shown in Table 4 all change by less than 0.5% from their original values. In spite of this result, we do not recommend implementations with zero frame overlap because the estimator could be extremely vulnerable to certain periodic, frame-synchronous noises and distortions. In addition, 50% overlap of Hamming windows places equal weight on each speech sample but zero overlap does not place equal weight on each speech sample.

### 6.4 Benchmark Values

Tables 5-8 provide benchmark values of AD and L(AD) for both MNB algorithms. Results are given for 11 standardized speech codecs and for 14 modulated noise reference unit (MNRU) [39] conditions. Within each table, AD and L(AD) results generally agree with known results on the perceived quality of these codecs and MNRU conditions. These results provide context for AD or L(AD) measurements made on other codecs or conditions.

Each condition in Tables 5-8 was evaluated using a total of 64 English-language sentence pairs. These 64 sentence pairs come from 4 female and 4 male talkers, each providing 8 different sentence pairs. Together, the 64 sentence pairs last about 400 seconds. Two sets of values were computed. Wideband speech recordings were band limited to 200-3400 Hz using a flat bandpass filter and then passed through the 25 conditions listed in the tables. Values for this "flat speech" experiment are given in Tables 5 and 6. In addition, wideband speech was filtered according to the IRS sending sensitivity characteristic [35] and then passed through the 25 conditions. Values for this "IRS speech" experiment are given in Tables 7 and 8. The tables provide a mean value taken across all 64 sentence pairs, as well as the half-width of the 95% confidence interval about that mean. As indicated in the tables, conditions 1-5 use μ-law compression. The results for A-law compression were computed as well, and in all cases, their confidence intervals overlap those of the μ-law results.

The MNRU is the most common reference condition for subjective and objective speech quality assessments. A common anchoring technique uses MNRU conditions with $Q$ (SNR) values at 5- or 6-dB increments. We have provided benchmark values for MNRU conditions with $Q$ values between 0 and 40 dB, in 5- and 6-dB increments. In addition, (10) through (13) give quadratic fits between $Q$ and AD for the 4 cases that correspond to Tables 5-8.

$$AD \approx 0.0003 \cdot Q^2 - 0.1862 \cdot Q + 8.1859, \quad 0 \le Q \le 40, \quad \text{MNB} - 1, \text{Flat Speech} \tag{10}$$

$$AD \approx 0.0020 \cdot Q^2 - 0.2583 \cdot Q + 7.6220, \quad 0 \le Q \le 40, \quad \text{MNB} - 2, \text{Flat Speech} \tag{11}$$

$$AD \approx 0.0024 \cdot Q^2 - 0.2719 \cdot Q + 8.2523, \quad 0 \le Q \le 40, \quad \text{MNB} - 1, \text{IRS Filtered Speech} \tag{12}$$

$$AD \approx 0.0031 \cdot Q^2 - 0.2846 \cdot Q + 6.9276, \quad 0 \le Q \le 40, \quad \text{MNB} - 2, \quad \text{IRS Filtered Speech} \tag{13}$$

When coupled with (9), these results allow one to relate $Q$ to L(AD). These relationships in turn allow reference to subjective test results that are given in terms of $Q$.

Table 5. MNB Structure 1 Benchmark Values for Flat Speech

| Condition | Mean AD | Half-width of 95% CI on Mean AD | Mean L(AD) | Half-width of 95% CI on Mean L(AD) |
|---|---|---|---|---|
| G.711 PCM, μ-law, 64 kbps | 1.9144 | 0.0645 | 0.9395 | 0.0040 |
| G.726 ADPCM μ-law, 40 kbps | 2.3810 | 0.0545 | 0.9077 | 0.0048 |
| G.726 ADPCM μ-law, 32 kbps | 2.9522 | 0.0543 | 0.8480 | 0.0070 |
| G.726 ADPCM μ-law, 24 kbps | 3.9458 | 0.0571 | 0.6753 | 0.0121 |
| G.726 ADPCM μ-law, 16 kbps | 5.1584 | 0.0745 | 0.3866 | 0.0176 |
| G.728 LD-CELP, 16 kbps | 3.2460 | 0.0710 | 0.8048 | 0.0112 |
| GSM 6.10 RPE-LTP, 13 kbps | 3.3194 | 0.0532 | 0.7949 | 0.0086 |
| TIA/EIA 635 VSELP, 8 kbps | 3.5978 | 0.0531 | 0.7462 | 0.0100 |
| FS1016 CELP, 4.8 kbps | 4.2856 | 0.0532 | 0.5981 | 0.0127 |
| FS1015 LPC, 2.4 kbps | 4.9589 | 0.0684 | 0.4340 | 0.0164 |
| MELP, 2.4 kbps [40] | 4.4928 | 0.0748 | 0.5475 | 0.0182 |
| MNRU, $Q$=40 | 1.5366 | 0.0365 | 0.9586 | 0.0015 |
| MNRU, $Q$=36 | 1.8960 | 0.0522 | 0.9411 | 0.0030 |
| MNRU, $Q$=35 | 2.0097 | 0.0568 | 0.9343 | 0.0036 |
| MNRU, $Q$=30 | 2.7244 | 0.0785 | 0.8728 | 0.0086 |
| MNRU, $Q$=25 | 3.6246 | 0.0933 | 0.7368 | 0.0171 |
| MNRU, $Q$=24 | 3.8173 | 0.0951 | 0.6986 | 0.0189 |
| MNRU, $Q$=20 | 4.6089 | 0.1020 | 0.5182 | 0.0244 |
| MNRU, $Q$=18 | 5.0027 | 0.1059 | 0.4244 | 0.0252 |
| MNRU, $Q$=15 | 5.5805 | 0.1127 | 0.2985 | 0.0236 |
| MNRU, $Q$=12 | 6.1346 | 0.1209 | 0.2013 | 0.0198 |
| MNRU, $Q$=10 | 6.4870 | 0.1272 | 0.1532 | 0.0169 |
| MNRU, $Q$=6 | 7.1354 | 0.1388 | 0.0893 | 0.0115 |
| MNRU, $Q$=5 | 7.2862 | 0.1414 | 0.0783 | 0.0103 |
| MNRU, $Q$=0 | 7.9791 | 0.1497 | 0.0418 | 0.0059 |

Table 6. MNB Structure 2 Benchmark Values for Flat Speech

| Condition | Mean AD | Half-width of 95% CI on Mean AD | Mean L(AD) | Half-width of 95% CI on Mean L(AD) |
|---|---|---|---|---|
| G.711 PCM, μ-law, 64 kbps | 0.8605 | 0.0334 | 0.8997 | 0.0030 |
| G.726 ADPCM μ-law, 40 kbps | 1.1822 | 0.0296 | 0.8669 | 0.0034 |
| G.726 ADPCM μ-law, 32 kbps | 1.6170 | 0.0406 | 0.8078 | 0.0063 |
| G.726 ADPCM μ-law, 24 kbps | 2.4503 | 0.0545 | 0.6465 | 0.0124 |
| G.726 ADPCM μ-law, 16 kbps | 3.6229 | 0.0824 | 0.3665 | 0.0187 |
| G.728 LD-CELP, 16 kbps | 1.8195 | 0.0454 | 0.7743 | 0.0080 |
| GSM 6.10 RPE-LTP, 13 kbps | 1.6594 | 0.0419 | 0.8011 | 0.0066 |
| TIA/EIA 635 VSELP, 8 kbps | 2.1782 | 0.0461 | 0.7060 | 0.0095 |
| FS1016 CELP, 4.8 kbps | 2.7902 | 0.0486 | 0.5667 | 0.0118 |
| FS1015 LPC, 2.4 kbps | 3.8886 | 0.0790 | 0.3084 | 0.0163 |
| MELP, 2.4 kbps [40] | 3.0911 | 0.0959 | 0.4935 | 0.0232 |
| MNRU, $Q$=40 | 0.6219 | 0.0214 | 0.9196 | 0.0016 |
| MNRU, $Q$=36 | 0.8669 | 0.0324 | 0.8991 | 0.0029 |
| MNRU, $Q$=35 | 0.9468 | 0.0359 | 0.8915 | 0.0034 |
| MNRU, $Q$=30 | 1.4778 | 0.0554 | 0.8274 | 0.0076 |
| MNRU, $Q$=25 | 2.2351 | 0.0770 | 0.6915 | 0.0155 |
| MNRU, $Q$=24 | 2.4129 | 0.0818 | 0.6527 | 0.0175 |
| MNRU, $Q$=20 | 3.1958 | 0.1017 | 0.4669 | 0.0243 |
| MNRU, $Q$=18 | 3.6213 | 0.1123 | 0.3686 | 0.0255 |
| MNRU, $Q$=15 | 4.2878 | 0.1272 | 0.2382 | 0.0237 |
| MNRU, $Q$=12 | 4.9660 | 0.1402 | 0.1428 | 0.0187 |
| MNRU, $Q$=10 | 5.4123 | 0.1475 | 0.0991 | 0.0149 |
| MNRU, $Q$=6 | 6.2511 | 0.1596 | 0.0476 | 0.0084 |
| MNRU, $Q$=5 | 6.4478 | 0.1624 | 0.0398 | 0.0072 |
| MNRU, $Q$=0 | 7.3357 | 0.1727 | 0.0173 | 0.0033 |

Table 7. MNB Structure 1 Benchmark Values for IRS Filtered Speech

| Condition | Mean AD | Half-width of 95% CI on Mean AD | Mean L(AD) | Half-width of 95% CI on Mean L(AD) |
|---|---|---|---|---|
| G.711 PCM, μ-law, 64 kbps | 1.6095 | 0.0406 | 0.9554 | 0.0019 |
| G.726 ADPCM μ-law, 40 kbps | 2.6178 | 0.0504 | 0.8863 | 0.0052 |
| G.726 ADPCM μ-law, 32 kbps | 3.2749 | 0.0554 | 0.8018 | 0.0088 |
| G.726 ADPCM μ-law, 24 kbps | 4.1863 | 0.0537 | 0.6214 | 0.0125 |
| G.726 ADPCM μ-law, 16 kbps | 5.3573 | 0.0688 | 0.3413 | 0.0153 |
| G.728 LD-CELP, 16 kbps | 3.2370 | 0.0630 | 0.8070 | 0.0101 |
| GSM 6.10 RPE-LTP, 13 kbps | 3.6603 | 0.0582 | 0.7339 | 0.0112 |
| TIA/EIA 635 VSELP, 8 kbps | 3.8011 | 0.0700 | 0.7049 | 0.0145 |
| FS1016 CELP, 4.8 kbps | 4.3568 | 0.0716 | 0.5803 | 0.0170 |
| FS1015 LPC, 2.4 kbps | 5.2181 | 0.0911 | 0.3743 | 0.0205 |
| MELP, 2.4 kbps [40] | 4.8443 | 0.0701 | 0.4614 | 0.0171 |
| MNRU, $Q$=40 | 1.4121 | 0.0288 | 0.9634 | 0.0011 |
| MNRU, $Q$=36 | 1.6163 | 0.0393 | 0.9552 | 0.0018 |
| MNRU, $Q$=35 | 1.6834 | 0.0431 | 0.9521 | 0.0021 |
| MNRU, $Q$=30 | 2.1452 | 0.0667 | 0.9248 | 0.0052 |
| MNRU, $Q$=25 | 2.8320 | 0.0952 | 0.8583 | 0.0127 |
| MNRU, $Q$=24 | 2.9971 | 0.1001 | 0.8369 | 0.0147 |
| MNRU, $Q$=20 | 3.7362 | 0.1180 | 0.7124 | 0.0246 |
| MNRU, $Q$=18 | 4.1487 | 0.1237 | 0.6255 | 0.0285 |
| MNRU, $Q$=15 | 4.8046 | 0.1278 | 0.4736 | 0.0301 |
| MNRU, $Q$=12 | 5.4782 | 0.1285 | 0.3226 | 0.0261 |
| MNRU, $Q$=10 | 5.9331 | 0.1279 | 0.2357 | 0.0216 |
| MNRU, $Q$=6 | 6.8134 | 0.1248 | 0.1160 | 0.0124 |
| MNRU, $Q$=5 | 7.0248 | 0.1239 | 0.0963 | 0.0105 |
| MNRU, $Q$=0 | 7.9904 | 0.1221 | 0.0395 | 0.0047 |

Table 8. MNB Structure 2 Benchmark Values for IRS Filtered Speech

| Condition | Mean AD | Half-width of 95% CI on Mean AD | Mean L(AD) | Half-width of 95% CI on Mean L(AD) |
|---|---|---|---|---|
| G.711 PCM, μ-law, 64 kbps | 0.7007 | 0.0230 | 0.9135 | 0.0018 |
| G.726 ADPCM μ-law, 40 kbps | 1.4589 | 0.0433 | 0.8309 | 0.0062 |
| G.726 ADPCM μ-law, 32 kbps | 2.0275 | 0.0560 | 0.7354 | 0.0112 |
| G.726 ADPCM μ-law, 24 kbps | 2.8975 | 0.0739 | 0.5405 | 0.0180 |
| G.726 ADPCM μ-law, 16 kbps | 3.9852 | 0.0940 | 0.2904 | 0.0181 |
| G.728 LD-CELP, 16 kbps | 1.9666 | 0.0455 | 0.7477 | 0.0085 |
| GSM 6.10 RPE-LTP, 13 kbps | 1.9071 | 0.0454 | 0.7587 | 0.0083 |
| TIA/EIA 635 VSELP, 8 kbps | 2.4007 | 0.0620 | 0.6572 | 0.0139 |
| FS1016 CELP, 4.8 kbps | 2.8412 | 0.0687 | 0.5536 | 0.0166 |
| FS1015 LPC, 2.4 kbps | 4.1366 | 0.1037 | 0.2622 | 0.0188 |
| MELP, 2.4 kbps [40] | 3.4433 | 0.0863 | 0.4085 | 0.0201 |
| MNRU, $Q$=40 | 0.5631 | 0.0201 | 0.9238 | 0.0015 |
| MNRU, $Q$=36 | 0.7183 | 0.0268 | 0.9120 | 0.0022 |
| MNRU, $Q$=35 | 0.7698 | 0.0291 | 0.9077 | 0.0025 |
| MNRU, $Q$=30 | 1.1213 | 0.0450 | 0.8730 | 0.0052 |
| MNRU, $Q$=25 | 1.6646 | 0.0667 | 0.7982 | 0.0110 |
| MNRU, $Q$=24 | 1.7990 | 0.0710 | 0.7755 | 0.0126 |
| MNRU, $Q$=20 | 2.4236 | 0.0899 | 0.6500 | 0.0203 |
| MNRU, $Q$=18 | 2.7858 | 0.0978 | 0.5662 | 0.0235 |
| MNRU, $Q$=15 | 3.3875 | 0.1084 | 0.4230 | 0.0254 |
| MNRU, $Q$=12 | 4.0407 | 0.1176 | 0.2828 | 0.0226 |
| MNRU, $Q$=10 | 4.4979 | 0.1226 | 0.2034 | 0.0188 |
| MNRU, $Q$=6 | 5.4260 | 0.1309 | 0.0948 | 0.0105 |
| MNRU, $Q$=5 | 5.6576 | 0.1326 | 0.0772 | 0.0089 |
| MNRU, $Q$=0 | 6.7363 | 0.1414 | 0.0285 | 0.0038 |

## 7. CONCLUSION

There is a clear need for estimators of perceived relative speech quality that provide reliable estimates, especially for lower-rate speech codecs, errored transmission channels, and other situations where waveforms are not preserved. Although they are clearly not perceptually consistent, SNR-based estimators are still in common use, probably due to their history, their simplicity, and the lack of a widely tested and accepted replacement. The recent attempts to incorporate models for human auditory perception into these estimators are clearly an important step forward. Unfortunately, it is not clear how simple models for the perception of tones and bands of noise might be best combined to create perceptual transformations that model the perception of more general signals such as speech. In addition, judgment is at least as important as hearing, but many highly refined hearing models have been followed by fairly simplistic judgment models, resulting in estimators that do not perform as reliably as one might hope. Our studies of perceptual transformations and distance measures have lead us to reverse this emphasis, resulting in a simple yet effective model for hearing, and a more sophisticated model for judgment.

Listeners adapt and react differently to spectral deviations that span different time and frequency scales. This motivates the development of a family of analyses that cover multiple frequency and time scales. To best emulate listeners' patterns of adaptation and reaction to spectral deviations, these analyses should proceed from larger scales to smaller scales. Further, spectral deviations at one scale must be removed so they are not counted again as part of the deviations at other scales. To meet these requirements, we have developed time measuring normalizing blocks and frequency measuring normalizing blocks. These idempotent blocks have been combined to form two hierarchical structures that comprise two distance measures. In effect, these structures decompose a codec output signal in a space defined partly by human hearing and judgment, and partly by the codec input signal. The parameters of this dynamic decomposition are combined linearly to form a measure of the perceptual distance between those two signals, which in turn is used to form an estimate of relative perceived speech quality.

Nine ACR subjective tests, using the MOS scale were available for testing objective estimators of perceived speech quality. Together, these 9 tests included 219 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit rates ranging from 2.4-64 kb/s, and some analog conditions as well. This collection of speech files and scores has allowed us to complete one of the most comprehensive tests of objective estimators of perceived relative speech quality. Six established estimators were tested along with the new MNB-based estimators. When the MNB estimators were optimized using only two of the tests, they generalized well to the other seven tests. The correlations between subjective scores and the fully optimized MNB estimators range from .910 to .986. Given the breadth of conditions covered by the nine tests, these are very encouraging results. In particular, the improved ability to estimate perceived speech quality for lower rate speech codecs, some of which are operating with bit errors or frame erasures represents an important advance. The two MNB structures presented here were chosen for their balance of relatively low complexity and high

performance as estimators of perceived speech quality across a wide range of conditions and quality levels. Other MNB structures may be more appropriate for more specific speech or audio quality estimation applications. In addition, these structures or other MNB structures may address open issues in perceived audio quality estimation, layered speech or audio coding, automatic speech or speaker recognition, audio signal enhancement, and other areas.

Formal subjective tests will very likely always provide the final definitive word when codecs and transmission systems are evaluated in major standardization, marketing, and procurement decisions. But objective estimators of perceived relative speech quality have a role to play as well. That role continues to expand as new estimators, like those described here, demonstrate increased reliability across broader ranges of test conditions. Perceptually consistent objective estimators of speech quality can provide a meaningful common language for designers and developers who wish to compare their results, but do not have access to subjective testing facilities. Estimators may also be consulted to aid in design decisions that might otherwise be made on the basis of a single designer's perception and judgment alone. In this situation, a large number of talkers, languages, or other relevant conditions can be tested with little effort in a comparatively short time. Finally, objective estimators are particularly well-suited for continuously monitoring speech transmission and storage systems of interest, and reporting deviations from established baseline quality levels.

## 8. REFERENCES


[1] *Proc. 1995 IEEE Workshop on Speech Coding for Telecommunications*, 1995.

[2] *Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications*, 1997.

[3] J.D. Gibson and W.W. Chang, "Objective and subjective optimization of APC system performance," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1053-1058, Jun. 1990.

[4] N. Kitawaki, "Quality assessment of coded speech," in *Advances in Speech Signal Processing*, S. Furui and M. Sondi, Ed., New York: Marcel Dekker, 1992, pp. 357-385.

[5] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, NJ: Prentice Hall, 1988.

[6] ITU-T Recommendation P.861, "Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs," Geneva, 1996.

[7] A. De and P. Kabal, "Rate-distortion function for speech coding based on perceptual distortion measure," in *Proc. IEEE Globcom `92*, 1992, pp. 452-456.

[8] M. Hansen and B. Kollmeier, "Using a quantitative psychoacoustical signal representation for objective speech quality measurement," in *Proc. IEEE ICASSP `97*, 1997, pp. 1387-1390.

[9] M. Hauenstein, "Comparative study of psychoacoustics-based objective speech-quality measures using markov-SIRPS," in *Proc. Speech Quality Assessment Workshop at Ruhr-Universität Bochum, Germany*, 1994, pp. 30-35.

[10] J. Herre, E. Eberlein, H. Schott, and K. Brandenburg, "Advanced audio measurement system using psychoacoustic properties," presented at 92nd Audio Engineering Society Convention, Vienna, 1992.

[11] M.P. Hollier, M.O. Hawksford, and D.R. Guard, "Characterization of communications systems using a speech-like test stimulus," presented at 93rd Audio Engineering Society Convention, San Francisco, California, 1992.

[12] R.F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, 1993, pp. 125-128.

[13] L.B. Nielsen, "Objective scaling of sound quality for normal-hearing and hearing-impaired listeners," Oticon Internal Report 43-8-4, Snekkersten, Denmark, 1993.

[14] B. Paillard, B. Mabilleau, and S. Morissette, "PERCEVAL: perceptual evaluation of the quality of audio signals," *J. Audio Engineering Society*, vol. 40, pp. 21-31, Jan. 1992.

[15] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. on Selected Areas in Communications*, vol. 10, pp. 819-829, Jun. 1992.

[16] ANSI Standard T1.801.04-1997, "Multimedia Communications Delay, Synchronization, and Frame Rate Measurement," New York, 1997.

[17] G. Theile, G. Stoll, and M. Link, "Low bit-rate coding of high-quality audio signals," presented at 82nd Audio Engineering Society Convention, London, 1987.

[18] ISO/IEC International Standard 11172-3, "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s - Part 3: Audio," Geneva, 1993.

[19] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. on Selected Areas in Communications*, vol. 6, pp. 314-323, Feb. 1988.

[20] M.R. Schroeder, B.S. Atal, and J.L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoustical Society of America*, vol. 66, pp. 1647-1652, Dec. 1979.

[21] D. Sinha and A.H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3463-3479, Dec. 1993.

[22] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, pp. 155-182, Mar. 1979.

[23] R.N.J. Veldhuis, "Bit rates in audio source coding," *IEEE J. on Selected Areas in Communications*, vol. 10, pp. 86-96, Jan. 1992.

[24] J.G. Beerends and J.A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Engineering Society*, vol. 40, pp. 963-978, Dec. 1992.

[25] L. E. Humes and W. Jesteadt, "Models of the additivity of masking," *J. Acoustical Society of America*, vol. 85, pp. 1285-1294, Mar. 1989.

[26] S. Voran, "Observations on auditory excitation and masking patterns," in *Proc. 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 1995.

[27] J.G. Beerends and J.A. Stemerdink,"A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Engineering Society*, vol. 42, pp. 115-123, Mar. 1994.

[28] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustical Society of America*, vol. 87, pp. 1738-1752, Apr. 1990.

[29] R.F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proc. IEEE ICASSP `82*, Paris, France, 1982, pp. 1282-1285.

[30] T.H. Bullock, Ed., "Report of the Dahlem Workshop on Recognition of Complex Acoustic Signals," Life Sciences Report 5, Berlin, Sep. 1976, p. 324.

[31] B.C. Moore, *An Introduction to the Psychology of Hearing*. London: Academic Press, 1989.

[32] G. Fant, *Speech Sounds and Features*. Cambridge, MA: The MIT Press, 1973.

[33] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoustical Society of America*, vol. 68, pp. 1523-1525, Nov. 1980.

[34] S. Voran and C. Sholl, "Perception-based objective estimators of speech quality", in *Proc. 1995 IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, Maryland, 1995, pp. 13-14.

[35] CCITT Recommendation P.48, "Specification for an Intermediate Reference System," Geneva, 1989.

[36] Y. Be'ery, Z. Shpiro, T. Simchony, L. Shatz, and J. Piasetzky, "An efficient variable-bit- rate low-delay CELP coder," in *Advances in Speech Coding*, B.S. Atal, V. Cuperman, and A. Gersho, Eds., Boston: Kluwer Academic Publishers, 1990, pp. 37-46.

[37] S. Voran, "Techniques for comparing objective and subjective speech quality tests," in *Proc. Speech Quality Assessment Workshop at Ruhr-Universität Bochum, Germany*, 1994, pp. 59-64.

[38] ITU-T Recommendation P.861, Appendix II, "Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs Using Measuring Normalizing Blocks (MNBs)," Geneva, 1998.

[39] CCITT Recommendation P.81, "Modulated Noise Reference Unit (MNRU)," Geneva, 1989.

[40] A.V. McCree, et. al., "A 2.4 kbits/s MELP coder candidate for the new U.S. federal standard," in *Proc. IEEE ICASSP `96,* Atlanta, USA, May 1996, pp. 200-203.

## ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| ACELP | algebraic code-excited linear prediction |
| ACR | absolute category rating |
| AD | auditory distance |
| ADPCM | adaptive differential pulse-code modulation |
| AMPS | advanced mobile phone service |
| APC | adaptive predictive coding |
| BSD | Bark spectral distortion |
| CCITT | The International Telegraph and Telephone Consultative Committee (now ITU-T) |
| CD | cepstral distance |
| CELP | code-excited linear prediction |
| CVSD | continuously variable slope delta modulation |
| DCR | degradation category rating |
| FFT | fast Fourier transform |
| FMNB | frequency measuring normalizing block |
| G.728 | ITU-T Recommendation G.728 |
| Hz | Hertz |
| IEEE | The Institute of Electrical and Electronics Engineers, Inc. |
| IMBE | improved multiband excitation |
| IRS | intermediate reference system |
| ITU-T | International Telecommunication Union-Telecommunication Standardization Sector |
| kb/s | kilobits per second |
| kHz | kilohertz |
| LD-CELP | low-delay-code-excited linear prediction |
| LPC | linear predictive coding |
| MNRU | modulated noise reference unit |
| MELP | mixed excitation linear prediction |
| MNB | measuring normalizing block |
| MOS | mean opinion score |
| ms | millisecond |
| ND | noise disturbance |
| P.861 | ITU-T Recommendation P.861 |
| PCM | pulse-code modulation |
| POTS | plain old telephone service |
| PSD | power spectral density |
| PWSNRseg | perceptually weighted segmental signal-to-noise ratio |
| RF | radio frequency |
| RPE-LTP | regular pulse excitation - long-term prediction |
| SELP | self-excited linear prediction |

| | |
|---|---|
| SNR | signal-to-noise ratio |
| SNRseg | segmental signal-to-noise ratio |
| STC | sinusoidal transform coding |
| TMNB | time measuring normalizing block |
| VSELP | vector sum-excited linear prediction |

## APPENDIX A: DESCRIPTION OF MNB ALGORITHMS

This appendix provides complete descriptions of the MNB algorithms at a level of detail that allows for implementation. To implement MNB structure 1, follow steps A.1-A.6 and A.8. To implement MNB structure 2, follow steps A.1-A.5, A.7, and A.8. To avoid a proliferation of variable names, this appendix does not use a unique variable for each intermediate result. Rather, variables are reused, just as they would be in a programming language.

### A.1. Signal Preparation

The input to the algorithm is a pair vectors **x** and **y**. These vectors contain speech samples from the input and output of the speech device under test, respectively. The recommended speech sample precision is at least 16 bits. The assumed sample rate is 8000 samples/s. The vectors must contain at least 1 s of telephone bandwidth speech. (Vectors used in the development of these algorithms ranged from 3 to 9 s in duration.) It is assumed that the two vectors have the same length, and are synchronized. Synchronization may be accomplished as described in reference [16] of the report. The mean value is then removed from each of the *N1* entries in **x** and **y**:

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

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

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

### A.2. Transformation to Frequency Domain

Each vector is next broken into a series of frames, with 128 samples in each frame. The frame overlap is 50%, so each frame begins 64 samples from the start of the previous frame. Any samples beyond the final full frame are discarded. Each frame of samples is multiplied (sample by sample) by the length 128 Hamming window:

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

After multiplication by the Hamming window, each frame is transformed to a 128 point frequency domain vector using the FFT. Scaling in FFT implementations is apparently not well standardized. The FFT used in this algorithm should be scaled so that the following condition is met. When a frame of 128 real-valued samples, each with value 1, is the input to the FFT (no Hamming window), then the complex value in the DC bin of the FFT output must be 128+0·j.

For each transformed 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**.

### A.3. Frame Selection

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

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

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

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

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 vectors do not contain suitable signals and this algorithm is terminated.

The thresholds given above appear to be the most useful for the general problem of estimating perceived speech quality across the conditions given in Table 1 of the report. Other thresholds may be more useful for other, more specific applications. In particular, multiple thresholds that separate a speech or audio signal into several categories (e.g., main signal, background noise, or silence) may be advantageous.

### A.4. Perceived Loudness Approximation

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

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

### A.5. Frequency Measuring Normalizing Block

An FMNB is applied to **X** and **Y** at the longest available time scale, defined by the length (*N1*) of the input vectors. Four measurements are extracted and stored in the measurement vector **m**. These measurements cover the lower and upper band edges of telephone band speech. Positive and negative

portions of the measurements are not separated. Temporary vectors **f1**, **f2**, and **f3** are used for clarity.

$$\mathbf{f1}(i) = \tfrac{1}{N3}\sum_{j=1}^{N3}\mathbf{Y}(i,j) - \tfrac{1}{N3}\sum_{j=1}^{N3}\mathbf{X}(i,j),\ 1 \le i \le 65 \quad \text{(measure)}$$

$$\mathbf{Y}(i,j) = \mathbf{Y}(i,j) - \mathbf{f1}(i),\ 1 \le i \le 65,\ 1 \le j \le N3 \quad \text{(normalize } \mathbf{Y}\text{)}$$

$$\mathbf{f2}(i) = \mathbf{f1}(i) - \mathbf{f1}(17),\ 1 \le i \le 65 \quad \text{(normalize measurement to 1 kHz)}$$

$$\mathbf{f3}(i) = \tfrac{1}{4}\sum_{j=1}^{4}\mathbf{f2}(1 + 4\cdot(i-1) + j),\ 1 \le i \le 16 \quad \text{(smooth the measurement)}$$

$$\left[\mathbf{m}(1)\quad \mathbf{m}(2)\quad \mathbf{m}(3)\quad \mathbf{m}(4)\right] = \left[\mathbf{f3}(1)\quad \mathbf{f3}(2)\quad \mathbf{f3}(13)\quad \mathbf{f3}(14)\right] \quad \text{(save 4 measurements)}$$

### A.6. Structure 1 Time Measuring Normalizing Blocks

In MNB structure 1, a TMNB is applied to **X** and **Y** at the largest frequency scale (approximately 15 Bark). Six additional TMNB's are then applied at a smaller scale (approximately 2-3 Bark). Finally a residual measurement is made. The result is eight additional measurements that are stored in the length 12 column vector **m**. Temporary variables **t0**, **t1**, and **t2** are used for clarity. A graphical representation of MNB structure 1 is given in Figure 7 of the report. The operations are grouped into steps a, b, and c below.

a. Largest Scale TMNB (14.9 Bark wide)

$$\mathbf{t0}(j) = \tfrac{1}{64}\sum_{i=2}^{65}\mathbf{Y}(i,j) - \tfrac{1}{64}\sum_{i=2}^{65}\mathbf{X}(i,j),\ 1 \le j \le N3 \quad \text{(measure)}$$

$$\mathbf{Y}(i,j) = \mathbf{Y}(i,j) - \mathbf{t0}(j),\ 2 \le i \le 65,\ 1 \le j \le N3 \quad \text{(normalize } \mathbf{Y}\text{)}$$

$$\mathbf{m}(5) = \tfrac{1}{N3}\sum_{j=1}^{N3}\max(\mathbf{t0}(j), 0) \quad \text{(save positive portion of measurement)}$$

b. Define the vector of band limits $\mathbf{g} = [2\ \ 7\ \ 12\ \ 19\ \ 29\ \ 43\ \ 66]^T$. Then the six small-scale TMNB's are implemented by the following pseudocode.

for $k = 1$ to 6

$$\mathbf{t1}(j) = \tfrac{1}{\mathbf{g}(k+1) - \mathbf{g}(k)}\sum_{i=\mathbf{g}(k)}^{\mathbf{g}(k+1)-1}\mathbf{Y}(i,j) - \tfrac{1}{\mathbf{g}(k+1) - \mathbf{g}(k)}\sum_{i=\mathbf{g}(k)}^{\mathbf{g}(k+1)-1}\mathbf{X}(i,j),\ 1 \le j \le N3 \quad \text{(measure)}$$

$$\mathbf{Y}(i,j) = \mathbf{Y}(i,j) - \mathbf{t1}(j),\ \mathbf{g}(k) \le i \le \mathbf{g}(k+1) - 1,\ 1 \le j \le N3 \quad \text{(normalize } \mathbf{Y}\text{)}$$

$$\mathbf{m}(5+k) = \tfrac{1}{N3}\sum_{j=1}^{N3}\max(\mathbf{t1}(j), 0) \quad \text{(save positive portion of measurement)}$$

end

c. Residual Measurement

$$\mathbf{t}2(i,\ j) = \mathbf{Y}(i,j) - \mathbf{X}(i,j),\ 1 \le i \le 65,\ 1 \le j \le N3 \quad \text{(measure residual)}$$

$$\mathbf{m}(12) = \frac{1}{N3 \cdot 64} \sum_{i=2}^{65} \sum_{j=1}^{N3} \max(\mathbf{t}2(i,\ j), 0) \quad \text{(save positive portion of residual measurement)}$$

## A.7. Structure 2 Time Measuring Normalizing Blocks

In MNB structure 2, the middle portion of the band undergoes two 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. The result is seven additional measurements that are stored in the length 11 column vector **m**. A graphical representation of MNB structure 2 is given in Figure 8 of the report. Temporary variables **t0**, **t1**, and **m0**, are used for clarity. The operations are grouped into steps a and b below.

a. Define the vectors of band limits $\mathbf{u} = [2\ 7\ 43\ 7\ 19\ 7\ 12\ 19\ 29]^T$ and $\mathbf{v} = [6\ 42\ 65\ 18\ 42\ 11\ 18\ 28\ 42]^T$. Then all TMNB's are implemented by the following pseudocode.

for $k = 1$ to 9

$$\mathbf{t}0(j) = \frac{1}{\mathbf{v}(k) - \mathbf{u}(k) + 1} \sum_{i=\mathbf{u}(k)}^{\mathbf{v}(k)} Y(i,j) - \frac{1}{\mathbf{v}(k) - \mathbf{u}(k) + 1} \sum_{i=\mathbf{u}(k)}^{\mathbf{v}(k)} \mathbf{X}(i,j),\ 1 \le j \le N3 \quad \text{(measure)}$$

$$\mathbf{Y}(i,j) = \mathbf{Y}(i,j) - \mathbf{t}0(j),\ \mathbf{u}(k) \le i \le \mathbf{v}(k),\ 1 \le j \le N3 \quad \text{(normalize } \mathbf{Y}\text{)}$$

$$\mathbf{m0}(k) = \frac{1}{N3} \sum_{j=1}^{N3} \max(\mathbf{t}0(j), 0) \quad \text{(save positive portion of measurement)}$$

end

$$\left[\mathbf{m}(5)\ \ \mathbf{m}(6)\ \ \mathbf{m}(7)\ \ \mathbf{m}(8)\ \ \mathbf{m}(9)\ \ \mathbf{m}(10)\right] = \left[\mathbf{m0}(1)\ \ \mathbf{m0}(2)\ \ \mathbf{m0}(3)\ \ \mathbf{m0}(4)\ \ \mathbf{m0}(6)\ \ \mathbf{m0}(8)\right]$$

b. Residual Measurement

$$\mathbf{t}1(i,\ j) = \mathbf{Y}(i,j) - \mathbf{X}(i,j),\ 1 \le i \le 65,\ 1 \le j \le N3 \quad \text{(measure residual)}$$

$$\mathbf{m}(11) = \frac{1}{N3 \cdot 64} \sum_{i=2}^{65} \sum_{j=1}^{N3} \max(\mathbf{t}1(i,\ j), 0) \quad \text{(save positive portion of residual measurement)}$$

## A.8. Linear Combinations and Logistic Functions

The 12 or 11 measurements from MNB structures 1 and 2, respectively, are next combined linearly to generate an AD value:

$$AD = \mathbf{w}^T \mathbf{m}.$$

Finally the AD value is passed through the logistic function to generate the final algorithm output, L(AD):

$$L(AD) = \frac{1}{1+ e^{a \cdot AD + b}} .$$

The weights and logistic parameters used in these steps are given in Table A-1.

Table A-1. Linear Combination Weights and Logistic Parameters for MNB Structures 1 and 2

| | Structure 1 | Structure 2 |
|---|---|---|
| **w**(1) | 0.0034 | 0.0000 |
| **w**(2) | -0.0650 | -0.0837 |
| **w**(3) | -0.1304 | -0.1199 |
| **w**(4) | 0.1352 | 0.1260 |
| **w**(5) | 0.5931 | 0.1660 |
| **w**(6) | 0.2040 | 0.6387 |
| **w**(7) | 0.5577 | 0.2195 |
| **w**(8) | 0.1008 | 0.0122 |
| **w**(9) | 0.0627 | 1.5544 |
| **w**(10) | 0.0052 | 0.0954 |
| **w**(11) | 0.0107 | 0.1720 |
| **w**(12) | 1.1037 | |
| *a* | 1.0000 | 1.0000 |
| *b* | -4.6877 | -3.0613 |