Objective Estimation of Perceived Speech Quality—Part I: Development of the Measuring Normalizing Block Technique

Stephen Voran

***Abstract*—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. The objective estimation of the perceived quality of highly compressed digital speech, possibly with bit errors or frame erasures has remained an open question. We report our findings regarding two essential components of objective estimators of perceived speech quality: perceptual transformations and distance measures. A perceptual transformation modifies a representation of an audio signal in a way that is approximately equivalent to the human hearing process. A distance measure reflects the magnitude of a perceived distance between two perceptually transformed signals.**

We then describe a new objective estimation approach that uses a simple but effective perceptual transformation and a distance measure that consists of a hierarchy of measuring normalizing blocks. Each measuring normalizing block integrates two perceptually transformed signals over some time or frequency interval to determine the average difference across that interval. This difference is then normalized out of one signal, and is further processed to generate one or more measurements. The resulting new estimators, and several established estimators, are thoroughly evaluated and compared in Part II of this paper. 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.

***Index Terms*—Auditory system, measuring normalizing blocks, speech coding, speech quality estimation, speech quality testing, subjective testing.**

I. Background

DIGITAL speech encoding and transmission involve 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 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 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. Signal-to-noise ratio (SNR) or segmental SNR (SNRseg) 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 Part II of this paper [3] and in [4]–[6], 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 estimation algorithm called perceptual speech quality measure (PSQM), but its scope is limited to higher bit rate speech codecs operating over error-free channels [7], [8]. The objective measurement of 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 [8]–[18]. 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

Manuscript received March 26, 1998; revised September 29, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Kroon.

The author is with the Institute for Telecommunication Sciences, U.S. Department of Commerce, NTIA/ITS, Boulder, CO 80303 USA (e-mail: sv@its.bldrdoc.gov).

Publisher Item Identifier S 1063-6676(99)04630-1.

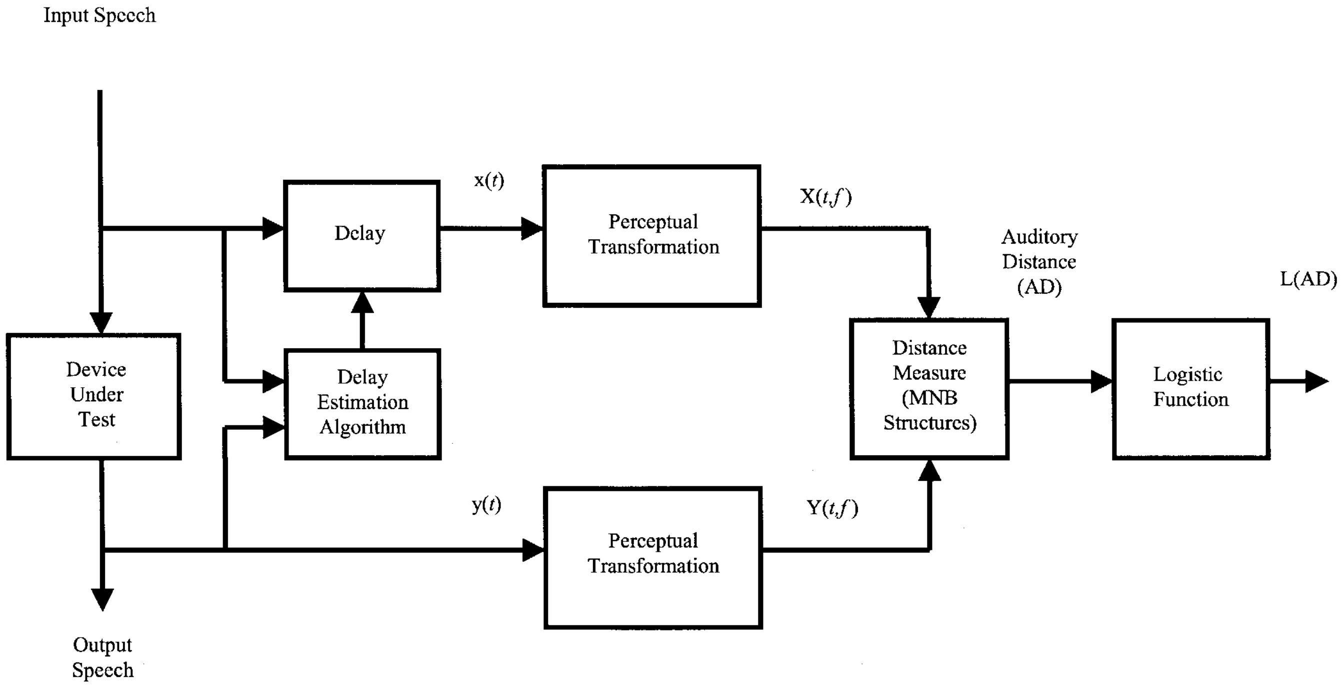


Fig. 1. High-level block diagram of the objective estimation approach.

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 led 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 Fig. 1. The delay of the device under test is first estimated and compensated for. The perceptual transformation contains a simple model for hearing, and the 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. Ideally, these nonnegative values increase in a consistent way 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 Fig. 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, or A-B comparison 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 between Fig. 1 and DCR tests, the results given in Part II show that the approach of Fig. 1 provides useful estimates of perceived speech quality as measured in ACR tests. This is not surprising since listeners may well accomplish the ACR rating task by forming an internal template of a perfect version of the test signal for comparison purposes. Thus, ACR tests might become DCR tests inside the listener. In [18], it is suggested that objective estimators be used to estimate differences between ACR test results.

In the following sections, we describe a delay estimation algorithm and a simple but effective perceptual transformation. We discuss distance measures and identify invariances in conventional distance measures that are clearly not perceptually consistent. Elimination of these undesired invariances motivates the development of measuring normalizing blocks (MNB's). MNB's are defined, and then combined in hierarchical structures that form distance measures. The resulting new MNB algorithms for objectively estimating perceived speech quality are described in full detail in Appendix A. In Part II of this paper we provide evaluations of the resulting objective estimators of perceived speech quality through comparison with the results of nine subjective tests. Part II also contains further observations and discussion, and a set of benchmark objective estimates of perceived speech quality for standardized codecs.

II. Delay Estimation

As shown in Fig. 1, the delay of the device under test must be estimated and compensated for 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. The algorithm is included in [19]. 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 lowpass 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 PSD's. PSD's are calculated from 8 ms segments of speech samples. Each segment is Hamming windowed and transformed using a discrete Fourier transform (DFT) or fast Fourier transform (FFT). The magnitude of the complex transform result is then extracted. The delay estimation algorithm performs PSD correlation multiple times and checks the locations of the resulting peaks 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. The sensitivity of the MNB algorithms to errors in delay estimation is discussed in Part II of this paper.

III. 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. Many references can be found in [20]. 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

$$E_t(f) = [E_1(f)^p + E_2(f)^p]^{1/p}. \tag{1}$$

However, different values of p have been selected by various authors. The maximum function, "$p = \infty$" is used in [21], $p = 1$ in [22]–[25], $p = 0.5$ in [26], and $p = 0.48$ in [27]. In [28], $p = 0.4$ is shown to be most useful when $E_t(f)$ is used to estimate the perception of coding distortions, and in [28] 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 ∞ is given in [30].

We have studied many of the perceptual transformation components that have been proposed to model various attributes of the hearing process [8]–[17], [20]–[35]. 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 [30], [36]. This work was repeated for a collection of different distance measures. We found that simpler perceptual transformations can be as effective or more effective than more complex ones. This observation is in general agreement with [8] and [11]. In particular, we have found that the nonuniform frequency resolution and the nonlinear loudness perception seem to be the most important properties to model. In fact, these are the only two properties that are explicitly modeled in the perceptual transformation described below. Further, we found that correlation results are much more sensitive to the choice of a distance measure than to the details of the perceptual transformation.

We have arrived at a very simple yet effective perceptual transformation that is built from a sequence of already established steps. 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 an 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], [24], [33]–[35] 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 frequency scale variable

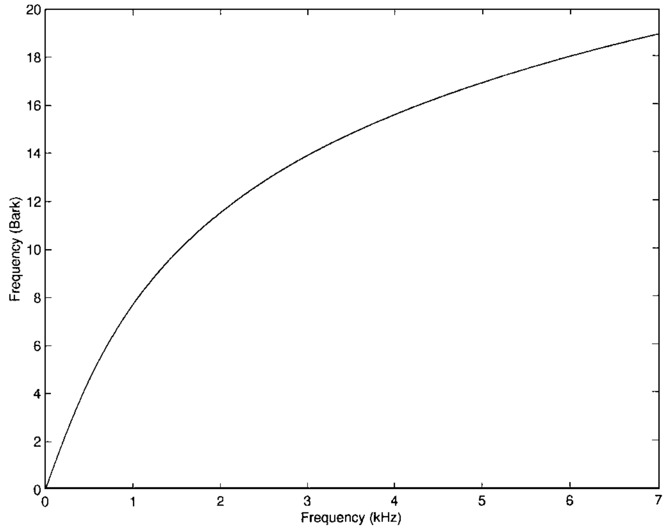


Fig. 2. Hertz to Bark transformation.

b using the relationship

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

which can be found in [33]. This relationship is plotted in Fig. 2. 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 [20]. Roughly speaking, on this Bark scale, equal frequency intervals are of equal perceptual importance. We use 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], [28], [33]. 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.

IV. Distance Measures

Distance measures seek to measure the magnitude of 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

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

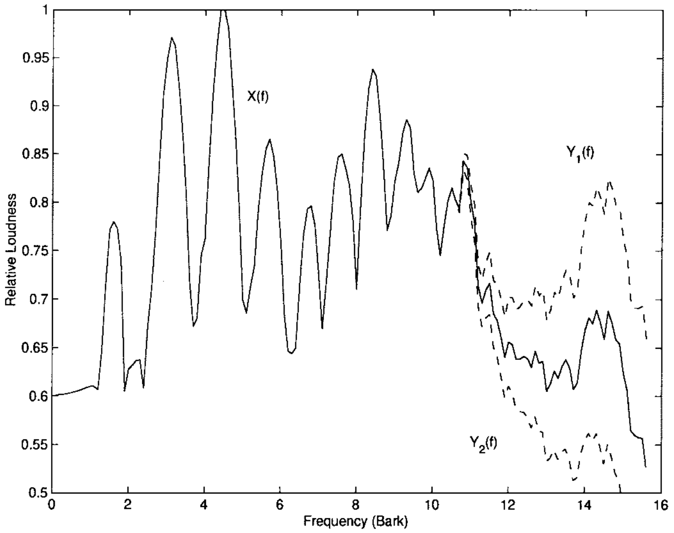


Fig. 3. Distance measure invariance example 1.

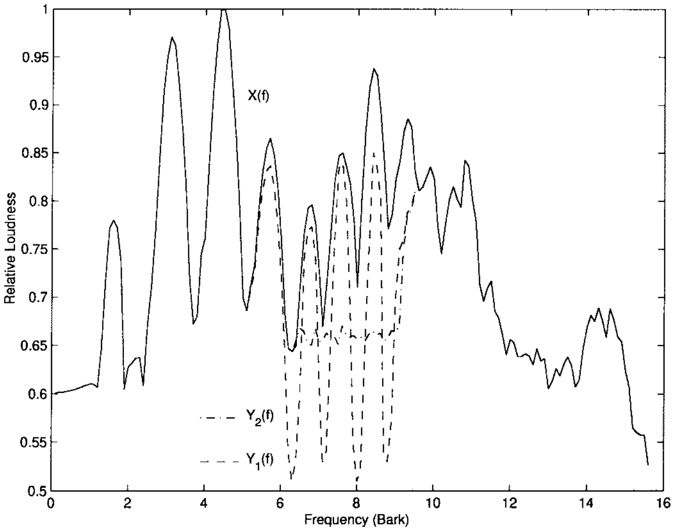


Fig. 4. Distance measure invariance example 2.

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 Ω. 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 Fig. 3 will receive the same distance value, which would not generally be a perceptually consistent result.

For a second example, consider the refined distance measure

$$\begin{aligned} &D[X(f), Y(f)] \\ &= \left[\frac{1}{\Omega_p}\int_{Y(f)\geq X(f)} w_p(f)(X(f) - Y(f))^{\gamma_p}\, df\right]^{1/\gamma_p} \\ &\quad + \left[\frac{1}{\Omega_n}\int_{Y(f)<X(f)} w_n(f)(X(f) - Y(f))^{\gamma_n}\, df\right]^{1/\gamma_n}. \end{aligned} \tag{4}$$

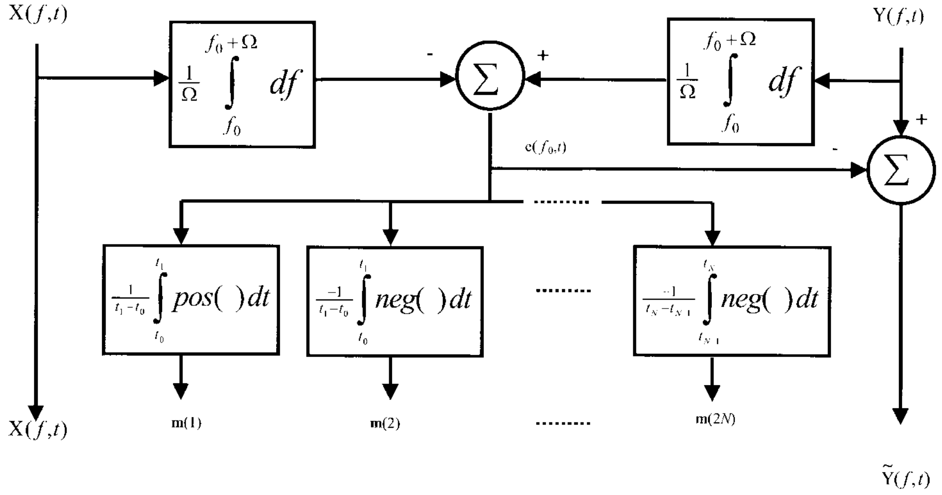

X(f,t)
$\frac{1}{\Omega}\int_{f_0}^{f_0+\Omega} df$
-
+
Σ
$\frac{1}{\Omega}\int_{f_0}^{f_0+\Omega} df$
Y(f,t)
+
e(f₀,t)
-
Σ
$\frac{1}{t_1-t_0}\int_{t_0}^{t_1} pos(\ )dt$
$\frac{-1}{t_1-t_0}\int_{t_0}^{t_1} neg(\ )dt$
$\frac{-1}{t_N-t_{N-1}}\int_{t_{N-1}}^{t_N} neg(\ )dt$
X(f,t)
m(1)
m(2)
m(2N)
Ỹ(f,t)

Fig. 5. Time measuring normalizing block (TMNB).

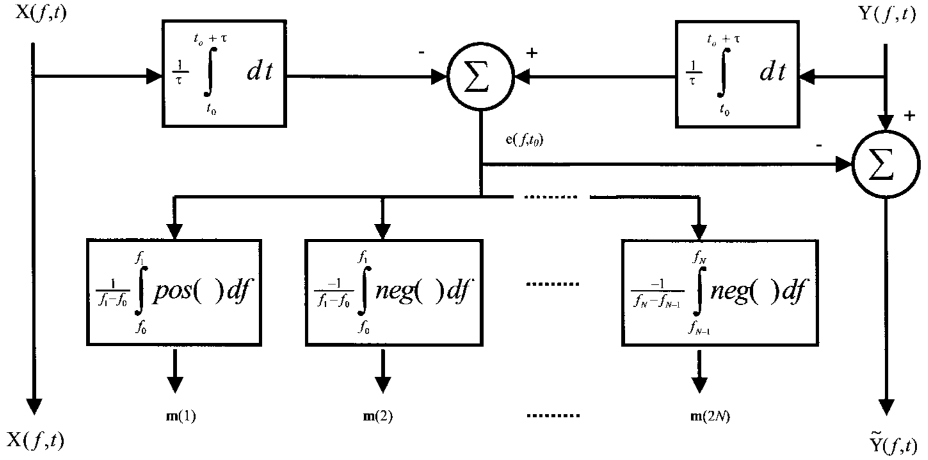

X(f,t)
$\frac{1}{\tau}\int_{t_0}^{t_0+\tau} dt$
-
+
Σ
$\frac{1}{\tau}\int_{t_0}^{t_0+\tau} dt$
Y(f,t)
+
e(f,t₀)
-
Σ
$\frac{1}{f_1-f_0}\int_{f_0}^{f_1} pos(\ )df$
$\frac{-1}{f_1-f_0}\int_{f_0}^{f_1} neg(\ )df$
$\frac{-1}{f_N-f_{N-1}}\int_{f_{N-1}}^{f_N} neg(\ )df$
X(f,t)
m(1)
m(2)
m(2N)
Ỹ(f,t)

Fig. 6. Frequency measuring normalizing block (FMNB).

In (4) the sign of $Y(f)-X(f)$ is acknowledged, with separate integrations, integration exponents γ, and weighting functions $w(f)$. With the signals $X(f), Y_1(f)$, and $Y_2(f)$ shown in Fig. 4, $D[X(f), Y_1(f)] = D[X(f), Y_2(f)]$. 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.

V. Measuring Normalizing Blocks

Section IV provides several simple examples of undesired invariances exhibited by conventional distance measures. These invariances are undesirable because they are not perceptually consistent: differences that are obvious to listeners disappear inside of these conventional distance measures. The problem of undesired invariances in conventional distance measures extends far beyond these examples. In general, distance measures that follow, or are similar to, the forms of (3) and (4) are invariant to the distributions of the differences that they are attempting to measure. Listeners, on the other hand, are often sensitive to distributions of differences.

To address these shortcomings of conventional distance measures, we developed MNB's and then formed distance measures from hierarchies of MNB's. Each MNB treats spectral differences that are distributed over a given scale in time or frequency. The MNB provides simple modeling of the disturbance caused by that spectral difference, and the ability of a listener to adapt to that spectral difference. When multiple MNB's covering multiple time or frequency scales are combined, they allow for simple modeling of the way in which listeners adapt and react to more complex spectral deviations that span different time and frequency scales.

A. Measuring Normalizing Block Definitions and Discussion

A time measuring normalizing block (TMNB) is shown in Fig. 5 and a frequency measuring normalizing block (FMNB) is given in Fig. 6. 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

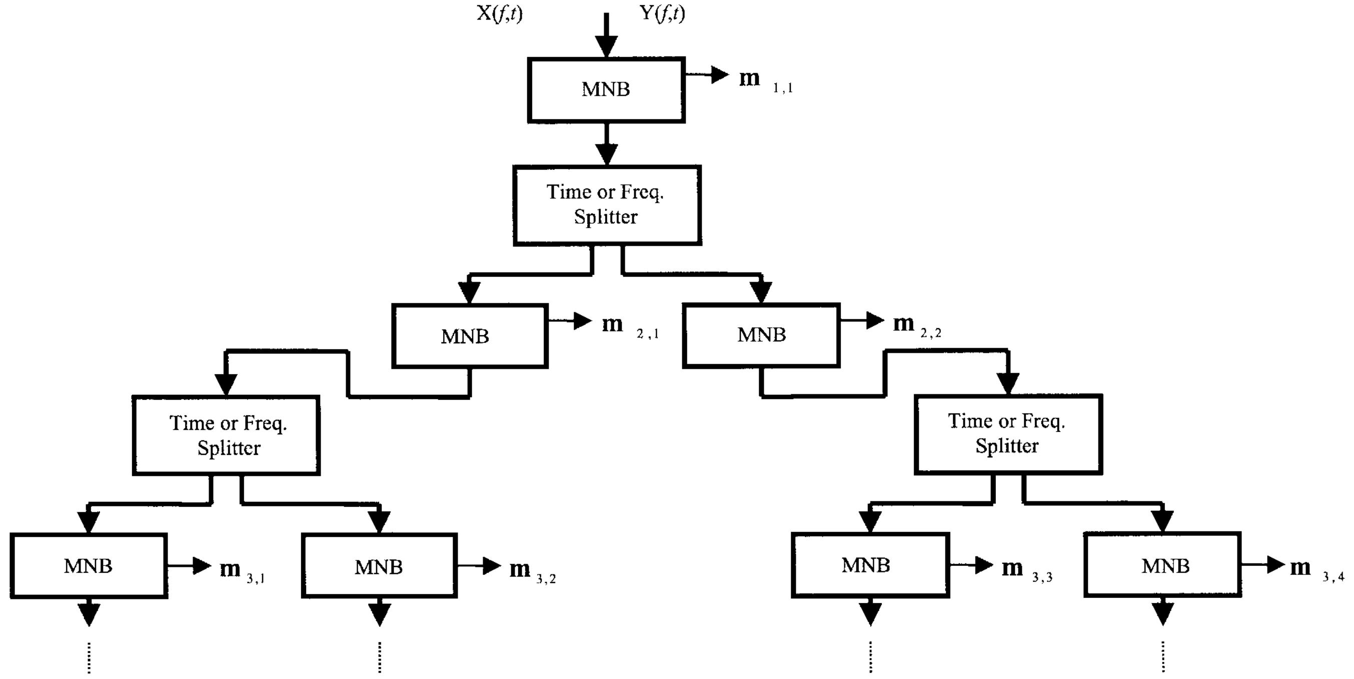


Fig. 7. Generalized measuring normalizing block (MNB) structure.

normalizes the output signal at multiple frequencies. Finally, the positive and negative portions of the measurements are integrated over frequency. Through these steps, each MNB provides a simple modeling of the disturbance caused by a spectral difference at a given scale, and the ability of a listener to adapt to that spectral difference.

We now formalize the MNB definitions. The TMNB operating on the band of width Ω that begins at f_0 using the measurement time intervals defined by $t_i, i = 0$ to N, normalizes $Y(f,t)$ to $\tilde{Y}(f,t)$ and generates $2N$ measurements $\boldsymbol{m}(j)$:

$$\begin{aligned}\tilde{Y}(f,t) &= Y(f,t) - e(f_0,t),\\ \boldsymbol{m}(2i-1) &= \frac{1}{t_i - t_{i-1}}1 \int_{t_{i-1}}^{t_i} \max(e(f_0,t),0)\, dt,\\ \boldsymbol{m}(2i) &= \frac{-1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \min(e(f_0,t),0)\, dt,\\ &\qquad i = 1 \text{ to } N \end{aligned} \tag{5}$$

where

$$e(f_0,t) = \frac{1}{\Omega}\int_{f_0}^{f_0+\Omega} Y(f,t)\, df - \frac{1}{\Omega}\int_{f_0}^{f_0+\Omega} X(f,t)\, df.$$

The FMNB definition is analogous, with the roles of time and frequency exchanged. At time t_0, the FMNB operating over time scale τ, 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 $\boldsymbol{m}(j)$:

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

where

$$e(f,t_0) = \frac{1}{\tau}\int_{t_0}^{t_0+\tau} Y(f,t)\, dt - \frac{1}{\tau}\int_{t_0}^{t_0+\tau} X(f,t)\, dt.$$

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

$$\begin{aligned}&\text{If } \mathrm{MNB}(X,Y) = (X,\tilde{Y},\boldsymbol{m}), \text{ then}\\ &\quad \mathrm{MNB}(X,\tilde{Y}) = (X,\tilde{Y},\mathbf{0}).\end{aligned} \tag{7}$$

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. This property of MNB's is critical as it allows them to be cascaded and yet they measure the deviation at a given time or frequency scale once and only once.

B. Distance Measures that Use Measuring Normalizing Blocks

In order to measure spectral deviations at multiple time and frequency scales, we have formed structures of TMNB's and FMNB's. We hypothesized that hierarchical structures that work from larger time and frequency scales down to smaller time and frequency scales would be most likely to emulate listeners' patterns of adaptation and reaction to spectral differences. By this technique spectral deviations at one time or frequency scale are measured and removed before the next smaller scale is considered. Results in Part II of this paper indicate that this hypothesis is a reasonable one. Specifically, when these hierarchical structures are used as distance measures in conjunction with the simple perceptual transformation described above, this top-down approach generates very useful estimates of perceived speech quality. A generalized diagram of these hierarchical MNB structures is shown in Fig. 7. Each MNB in the structure generates a measurement vector $\boldsymbol{m}_{i,j}$.

There are many alternatives to explore within the framework of Fig. 7. As always, there is a complexity-performance tradeoff at work here. Through a sequence of heuristic explorations, we have identified two MNB structures that offer relatively low complexity and high performance as estimators of perceived speech quality across a wide range of conditions and quality levels. These structures are shown in Figs. 8 and 9. These are referred to as MNB structure 1 and MNB structure 2, respectively. 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.

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 and 3000–3500 Hz.) In essence, this MNB stage measures and equalizes out the long-term frequency response at the edges of the telephone band. In MNB structure 1, a TMNB is then applied to the input and output signals at the largest frequency scale (approximately 15 Bark). This step can be viewed as a short-time, wide-band spectral difference measurement, followed by a fast adaptive gain stage that removes this spectral difference. Six additional TMNB's are then applied at a smaller scale (approximately 2–3 Bark). These TMNB's correspond to additional, narrower-band, spectral difference measurements and gain adaptations. Finally, a residual measurement is made to take account of the spectral differences at all remaining (finer) scales. 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. We can also loosely describe the action of these MNB structures as a dynamic decomposition of a codec output signal. This decomposition proceeds in a space that is defined partly by human hearing and judgment (via the MNB structure) and partly by the codec input signal.

The idempotence of the MNB along with the hierarchical nature of MNB structures leads to linear dependence among the MNB measurements. As shown in Figs. 8 and 9, 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):

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

where $\boldsymbol{w}$ is a length 12 (MNB structure 1) or 11 (MNB structure 2) vector of weights. In practice, AD values are nonnegative. 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.

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 1–5. Thus, correlation with these subjective rating scales may be increased by mapping AD values into a finite range. We use the following logistic function with asymptotes at 0 and 1:

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

When $a > 0, L(z)$ is a decreasing function of z. We selected this function because it maps AD into a finite interval, it exhibits the necessary scale compression at the high and low quality extremes, and it is nearly linear over the intermediate quality range.

Note that MNB's were developed to measure (react) and normalize (adapt) in a way that emulates listeners. The MNB structures, in turn, were developed to perform these steps sequentially at decreasing scales, as we hypothesized listeners might do. Combining the resulting measurements linearly was a purely utilitarian choice, not motivated by properties of perception or judgment. Part II of this paper shows that the structures are successful in the sense that they generate very useful estimates of perceived speech quality. On the other hand, we do not claim any direct, firm, correspondence between the algorithmic steps given above and the process of human audition and judgment. That is, the MNB structures are able to emulate the responses of listeners, but they do not directly explain or explicitly model how listeners arrive at those responses.

The perceptual transformation and the MNB structures are described together in full detail in Appendix A. Part II of this paper provides further discussion, interpretation, and results.

VI. 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

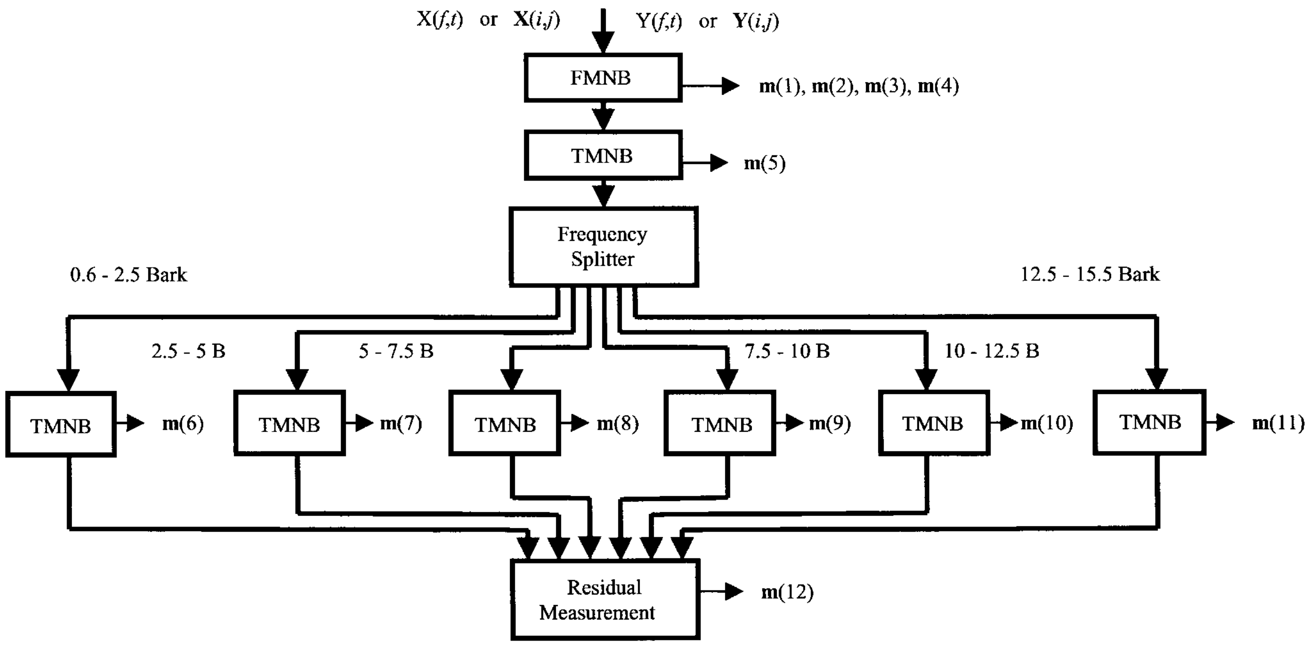


Fig. 8. MNB structure 1.

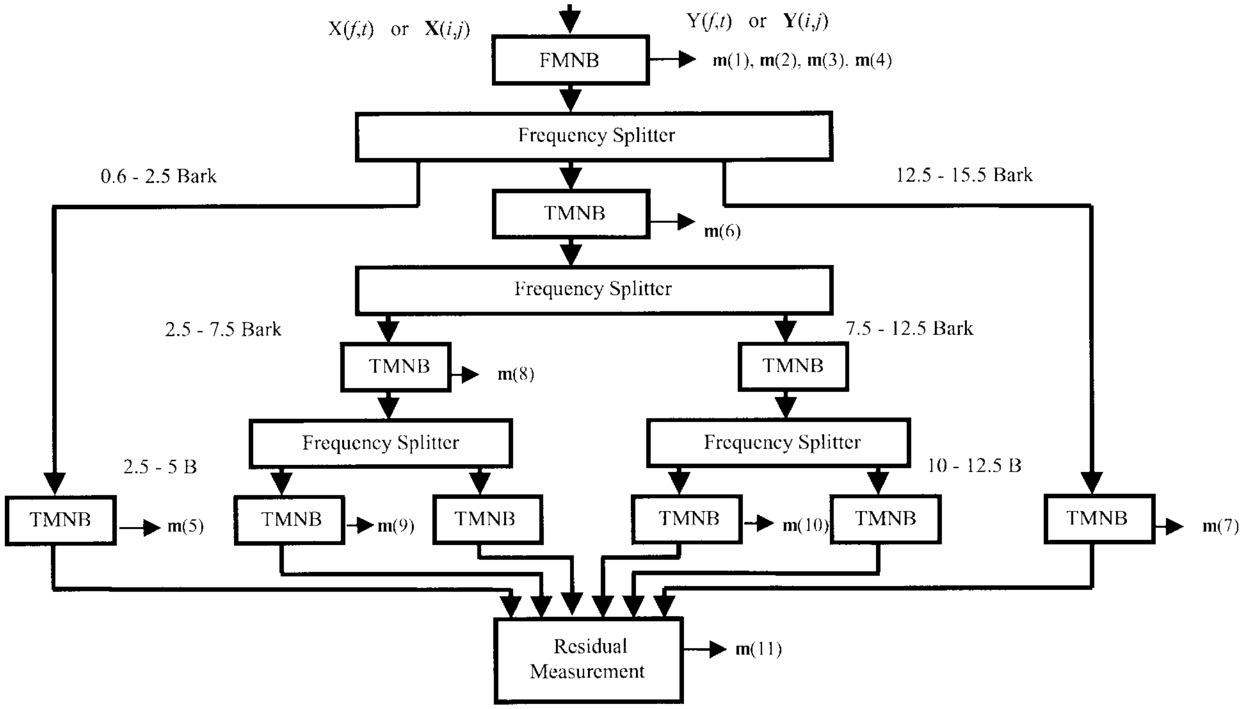


Fig. 9. MNB structure 2.

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 led us to an effective but rather simple perceptual transformation and more sophisticated distance measures built from measuring normalizing blocks.

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. This new technique for objectively estimating perceived speech quality is thoroughly evaluated in Part II of this paper. These structures or other MNB structures 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.

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–F and H. To implement MNB structure 2, follow steps A–E, G, and H. 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. Signal Preparation

The input to the algorithm is a pair vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. 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 3–9 s in duration.) It is assumed that the two vectors have the same length, and are synchronized. Synchronization may be accomplished as described in [18]. The mean value is then removed from each of the $N1$ entries in $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\begin{aligned}\boldsymbol{x}(i) &= \boldsymbol{x}(i) - \frac{1}{N1}\cdot\sum_{j=1}^{N1}\boldsymbol{x}(j),\\ \boldsymbol{y}(i) &= \boldsymbol{y}(i) - \frac{1}{N1}\cdot\sum_{j=1}^{N1}\boldsymbol{y}(j), \qquad 1 \le i \le N1.\end{aligned}$$

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

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

B. 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:

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

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 $128 + 0 \cdot j$. For each transformed frame, the squared-magnitude of frequency samples 1–65 (DC through Nyquist) are retained. The results are stored in the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. These matrices contain 65 rows, and $N2$ columns, where $N2$ is the number of frames that are extracted from the $N1$ original samples in $\boldsymbol{x}$ and $\boldsymbol{y}$.

C. Frame Selection

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

$$\begin{aligned}\mathit{xenergy}(j) &= \sum_{i=1}^{65} X(i,j),\\ \mathit{xthreshold} &= 10^{-15/10}\cdot\max_j\,(\mathit{xenergy}(j)).\end{aligned}$$

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

$$\begin{aligned}\mathit{xenergy}(j) &= \sum_{i=1}^{65} Y(i,j),\\ \mathit{ythreshold} &= 10^{-35/10}\cdot\max_j\,(\mathit{yenergy}(j)).\end{aligned}$$

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

$$\begin{aligned}&\mathit{xenergy}(j) \ge \mathit{xthreshold}\ \text{AND}\\ &\qquad \mathit{yenergy}(j) \ge \mathit{ythreshold} \Rightarrow \text{frame } j \text{ is retained.}\end{aligned}$$

If any frame contains one or more samples that are equal to zero, that frame is eliminated from both $\boldsymbol{X}$ and $\boldsymbol{Y}$. 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 described in Part II of this paper. 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.

D. Perceived Loudness Approximation

Each of the frequency domain samples in $\boldsymbol{X}$ and $\boldsymbol{Y}$ is then logarithmically transformed to an approximation of perceived loudness:

$$\begin{aligned}\boldsymbol{X}(i,j) &= 10 \cdot \log_{10}(X(i,j)),\\ \boldsymbol{Y}(i,j) &= 10 \cdot \log_{10}(\boldsymbol{Y}(i,j)),\\ &\qquad 1 \le i \le 65, \quad 1 \le j \le N3.\end{aligned}$$

E. Frequency Measuring Normalizing Block

An FMNB is applied to $\boldsymbol{X}$ and $\boldsymbol{Y}$ 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 $\boldsymbol{m}$. 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 $\boldsymbol{f1}, \boldsymbol{f2}$, and $\boldsymbol{f3}$ are used for clarity.

$$\begin{aligned}\boldsymbol{f1}(i) &= \frac{1}{N3}\sum_{j=1}^{N3} Y(i,j) - \frac{1}{N3}\sum_{j=1}^{N3} \boldsymbol{X}(i,j),\\ &\qquad 1 \le i \le 65 \quad \text{(measure)}\\ \boldsymbol{Y}(i,j) &= \boldsymbol{Y}(i,j) - \boldsymbol{f1}(i), \qquad 1 \le i \le 65, \quad 1 \le j \le N3\\ &\qquad \text{(normalize } \boldsymbol{Y}\text{)}\\ \boldsymbol{f2}(i) &= \boldsymbol{f1}(i) - \boldsymbol{f1}(17), \qquad 1 \le i \le 65\\ &\qquad \text{(normalize measurement to 1 kHz)}\\ \boldsymbol{f3}(i) &= \tfrac{1}{4}\sum_{j=1}^{4} \boldsymbol{f2}(1 + 4 \cdot (i-1) + j),\\ &\qquad 1 \le i \le 16 \quad \text{(smooth the measurement)}\end{aligned}$$

$$\begin{aligned}&[\boldsymbol{m}(1) \quad \boldsymbol{m}(2) \quad \boldsymbol{m}(3) \quad \boldsymbol{m}(4)]\\ &\quad = [\boldsymbol{f3}(1) \quad \boldsymbol{f3}(2) \quad \boldsymbol{f3}(13) \quad \boldsymbol{f3}(14)]\\ &\qquad \text{(save 4 measurements.)}\end{aligned}$$

F. Structure 1: Time Measuring Normalizing Blocks

In MNB structure 1, a TMNB is applied to $\boldsymbol{X}$ and $\boldsymbol{Y}$ 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 $\boldsymbol{m}$. Temporary variables $\boldsymbol{t0}, \boldsymbol{t1}$, and $\boldsymbol{t2}$ are used for clarity. A graphical representation of MNB structure 1 is given in Fig. 8.

1) Largest scale TMNB (14.9 Bark wide):

$$\begin{aligned}\boldsymbol{t0}(j) &= \frac{1}{64}\sum_{i=2}^{65} \boldsymbol{Y}(i,j) - \frac{1}{64}\sum_{i=2}^{65} \boldsymbol{X}(i,j),\\ &\qquad 1 \le j \le N3 \quad \text{(measure)}\\ \boldsymbol{Y}(i,j) &= \boldsymbol{Y}(i,j) - \boldsymbol{t0}(j), \qquad 2 \le i \le 65, \quad 1 \le j \le N3\\ &\qquad \text{(normalize } \boldsymbol{Y}\text{)}\\ \boldsymbol{m}(5) &= \frac{1}{N3}\sum_{j=1}^{N3} \max(\boldsymbol{t0}(j), 0)\\ &\qquad \text{(save positive portion of measurement)}\end{aligned}$$

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

```
for $k = 1$ to 6
    $\boldsymbol{t1}(j) = \frac{1}{\boldsymbol{g}(k+1) - \boldsymbol{g}(k)} \sum_{i=\boldsymbol{g}(k)}^{\boldsymbol{g}(k+1)-1} \boldsymbol{Y}(i,j) - \frac{1}{\boldsymbol{g}(k+1) - \boldsymbol{g}(k)} \sum_{i=\boldsymbol{g}(k)}^{\boldsymbol{g}(k+1)-1} \boldsymbol{X}(i,j)$,
        $1 \le j \le N3$ (measure)
    $\boldsymbol{Y}(i,j) = \boldsymbol{Y}(i,j) - \boldsymbol{t1}(j)$,
    $\boldsymbol{g}(k) \le i \le \boldsymbol{g}(k+1) - 1, \quad 1 \le j \le N3$
        (normalize $\boldsymbol{Y}$)
    $\boldsymbol{m}(5 + k) = \frac{1}{N3} \sum_{j=1}^{N3} \max(\boldsymbol{t1}(j), 0)$
        (save positive portion of measurement)
end.
```

3) Residual measurement

$$\begin{aligned}\boldsymbol{t2}(i,j) &= \boldsymbol{Y}(i,j) - \boldsymbol{X}(i,j), \quad 1 \le i \le 65, \quad 1 \le j \le N3\\ &\qquad \text{(measure residual)}\\ \boldsymbol{m}(12) &= \frac{1}{N3 \cdot 64}\sum_{i=2}^{65}\sum_{j=1}^{N3} \max(\boldsymbol{t2}(i,j), 0)\\ &\qquad \text{(save positive portion of residual measurement).}\end{aligned}$$

G. 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 $\boldsymbol{m}$. A graphical representation of MNB structure 2 is given in Fig. 9. Temporary variables $\boldsymbol{t0}, \boldsymbol{t1}$, and $\boldsymbol{m0}$, are used for clarity.

1) Define the vectors of band limits $\boldsymbol{u} = [2 \quad 7 \quad 43 \quad 7 \quad 19 \quad 7 \quad 12 \quad 19 \quad 29]^T$ and

TABLE I
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

$\boldsymbol{v} = [6 \;\; 42 \;\; 65 \;\; 18 \;\; 42 \;\; 11 \;\; 18 \;\; 28 \;\; 42]^T$. Then all TMNB's are implemented by the following pseudocode:

```
for $k = 1$ to $9$
    $\boldsymbol{t}0(j) = \frac{1}{\boldsymbol{v}(k) - \boldsymbol{u}(k) + 1} \sum_{i=\boldsymbol{u}(k)}^{\boldsymbol{v}(k)} Y(i,j) - \frac{1}{\boldsymbol{v}(k) - \boldsymbol{u}(k) + 1} \sum_{i=\boldsymbol{u}(k)}^{\boldsymbol{v}(k)} \boldsymbol{X}(i,j), \quad 1 \le j \le N3$ (measure)
    $\boldsymbol{Y}(i,j) = \boldsymbol{Y}(i,j) - \boldsymbol{t0}(j), \quad \boldsymbol{u}(k) \le i \le \boldsymbol{v}(k), \quad 1 \le j \le N3$
        (normalize $\boldsymbol{Y}$)
    $\boldsymbol{m0}(k) = \frac{1}{N3} \sum_{j=1}^{N3} \max(\boldsymbol{t0}(j), 0)$
        (save positive portion of measurement)
end.
```

$$[\boldsymbol{m}(5) \;\; \boldsymbol{m}(6) \;\; \boldsymbol{m}(7) \;\; \boldsymbol{m}(8) \;\; \boldsymbol{m}(9) \;\; \boldsymbol{m}(10)] = [\boldsymbol{m0}(1) \;\; \boldsymbol{m0}(2) \;\; \boldsymbol{m0}(3) \;\; \boldsymbol{m0}(4) \;\; \boldsymbol{m0}(6) \;\; \boldsymbol{m0}(8)].$$

2) Residual measurement

$$\boldsymbol{t1}(i,j) = \boldsymbol{Y}(i,j) - \boldsymbol{X}(i,j), \quad 1 \le i \le 65, \quad 1 \le j \le N3 \quad \text{(measure residual)}$$

$$\boldsymbol{m}(11) = \frac{1}{N3 \cdot 64} \sum_{i=2}^{65} \sum_{j=1}^{N3} \max(\boldsymbol{t1}(i,j), 0) \quad \text{(save positive portion of residual measurement).}$$

H. 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:

$$\text{AD} = \boldsymbol{w}^T \boldsymbol{m}.$$

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

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

The weights and logistic parameters used in these steps are given in Table I.

REFERENCES

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

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

[3] S. Voran, "Objective estimation of perceived speech quality—Part II: Evaluation of the measuring normalizing block technique," this issue, pp. 385–390, July 1999.

[4] J. D. Gibson and W. W. Chang, "Objective and subjective optimization of APC system performance," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1053–1058, June 1990.

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

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

[7] ITU-T Recommendation P.861, "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," Geneva, Switzerland, 1996.

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

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

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

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

[12] J. Herre, E. Eberlein, H. Schott, and K. Brandenburg, "Advanced audio measurement system using psychoacoustic properties," presented at *92nd Audio Engineering Soc. Conv.*, Vienna, Austria, 1992.

[13] 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 Soc. Conv.*, San Francisco, CA, 1992.

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

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

[16] B. Paillard, B. Mabilleau, and S. Morissette, "PERCEVAL: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, pp. 21–31, Jan. 1992.

[17] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 819–829, June 1992.

[18] W. Yang, M. Dixon, and R. Yantorno, "A modified bark spectral distoriton measure which uses noise masking threshold," in *Proc. 1997 IEEE Workshop Speech Coding for Telecommunications*, 1997, pp. 55–56.

[19] ANSI Standard T1.801.04-1997, "Multimedia communications delay, synchronization, and frame rate measurement," 1997.

[20] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. New York: Academic, 1989.

[21] G. Theile, G. Stoll, and M. Link, "Low bit-rate coding of high-quality audio signals," presented at *82nd Audio Engineering Soc. Conv.*, London, U.K., 1987.

[22] 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, Switzerland, 1993.

[23] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[24] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1652, Dec. 1979.

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

[26] E. Terhardt, "Calculating virtual pitch," *Hearing Res.*, vol. 1, pp. 155–182, Mar. 1979.

[27] R. N. J. Veldhuis, "Bit rates in audio source coding," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 86–96, Jan. 1992.

[28] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978, Dec. 1992.

[29] L. E. Humes and W. Jesteadt, "Models of the additivity of masking," *J. Acoust. Soc. Amer.*, vol. 85, pp. 1285–1294, Mar. 1989.

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

[31] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, Apr. 1990.

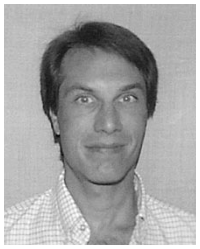
[32] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proc. IEEE ICASSP'82*, pp. 1282–1285.

[33] T. H. Bullock, Ed., *Report of the Dahlem Workshop on Recognition of Complex Acoustic Signals*, Life Sci. Rep. 5, Berlin, Germany, Sept. 1976, p. 324.

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

[35] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, pp. 1523–1525, Nov. 1980.

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



Stephen Voran received the B.A. degree in mathematics from Carleton College, Northfield, MN, in 1985, and the M.S. degree in electrical engineering from the University of Colorado, Boulder, in 1989, with thesis work in quantization effects.

He has been with the Institute for Telecommunication Sciences, Boulder, since 1990. His research interests include speech and audio coding, perception, and quality assessment.