COMMITTEE T1 - TELECOMMUNICATIONS STANDARDS CONTRIBUTION

DOCUMENT NUMBER: T1A1.5/96-123
T1A1.7/96-037

T1BBS FILE: 6a151230.doc, 6a170370.doc

DATE: October 29, 1996

STANDARDS PROJECT: T1Q1-12, T1Y1-20

TITLE: Proposed objective quality measure for the audio portion of an audio-visual session

SOURCE: Institute for Telecommunication Sciences
National Telecommunications and Information Administration
U. S. Department of Commerce

AUTHORS: D.J. Atkinson and Stephen Voran

CONTACT: D.J. Atkinson
NTIA/ITS.N3
325 Broadway
Boulder, CO 80303-3328
Tel +1 303 497 5281
Fax +1 303 497 5323
email dj@its.bldrdoc.gov

DISTRIBUTION: T1A1.5, T1A1.7, T1BBS

ABSTRACT: This contribution provides an overview of ITS-developed objective audio quality measures which use measuring normalizing blocks (MNBs), and reports the performance of the MNB algorithms when using data from tests containing a wide variety of audio conditions. It begins by providing resources for understanding the decisions that went into the development of the MNB algorithms, and information on the perceptual transformation selected for use with them. The performance of the MNB algorithms is then compared with that of six other established objective audio quality measures over seven sets of subjectively scored speech data. After the results are summarized, a look forward at the potential development process for a standard based on the MNB algorithms is presented.

Introduction

In 1996, T1 approved standard performance parameters for the objective assessment of video [1] as a task under project T1Q1.12. This standard was a breakthrough in that it was the first ANSI standard adopted that provided a perception-based performance parameter for the measurement of audio-visual quality. It seems a logical conclusion, based on the wording of the standards project, to supplement this standard with a related (i.e., perception-based) objective measure for audio quality. The Institute for Telecommunication Sciences (ITS) has been developing such an objective audio quality measure for the past several years and believes that it can provide T1A1.5 with a measure that is suitable for the objective measurement of quality for the audio component of an audio-visual session.

This contribution provides an overview of ITS-developed objective audio quality measures which use measuring normalizing blocks (MNBs), and reports the performance of the MNB algorithms when using data from tests containing a wide variety of audio conditions. It begins by providing resources for understanding the decisions that went into the development of the MNB algorithms, and information on the perceptual transformation selected for use with them. The performance of the MNB algorithms is then compared with that of six other established objective audio quality measures over seven sets of subjectively scored speech data. After the results are summarized, a look forward at the potential development process for a standard based on the MNB algorithms is presented.

Developmental Background

An overview of the ITS research on perception-based objective audio quality assessment tools is provided in [2]. This research is motivated by a need for objective audio and speech quality assessment tools that show significantly improved invariance to the technology employed in an audio or speech device under test. The recently approved ITU-T Recommendation P.861 describes a speech quality assessment tool that has demonstrated accuracy for higher bit-rate speech codecs operating over error-free channels[3]. The emergence of audio-visual conferencing, digital wireless, internet, and other communications options has created an increased interest in the quality assessment of speech produced by lower bit-rate coders that operate under the higher bit-error rate and packet loss conditions associated with some RF transmission paths and data networks.

We have adopted a perception-based approach in our research on objective audio quality assessment tools. The basic premise of this approach is that by transforming audio signals into an appropriate perceptual domain, only information that is perceptually relevant is retained. That information is, by definition, both necessary and sufficient for the accurate assessment of audio quality, independent of the coding and transmission technology employed by the device under test. A more complete accounting of our work to date can be found in [2,4-7]. In the following, we focus on 4 kHz bandwidth speech. We intend to extend this work to include 7 kHz speech, and ultimately to encompass general audio signals with 20 kHz bandwidth.

A high-level description of the ITS approach is shown in Figure 1 The following discussion is limited to the components inside the dashed-line box in Figure 1. We have studied many of the perceptual transformations that have been proposed to model the human hearing process, and have drawn some conclusions about what level of detail is appropriate for this application [5-7]. Some of the transformations and distance measures used in our studies are based on those described in [8-14], and some are original. The results of these studies underscore the important role that the distance measure plays. As a result, we have spent significant effort to develop original distance measures that compare original and decoded speech signals in a way that is as consistent with human judgment as possible. This development work has revealed that a family of frequency domain analyses at several different time and/or frequency resolutions can be very advantageous.

In the following sections, we describe a very simple yet effective perceptual transformation. We then present some observations on conventional distance measures, and describe how those observations motivated the development of new distance measures. These new distance measures are formed from hierarchies of measuring normalizing blocks. The properties and performance of the resulting objective audio quality assessment algorithms are then discussed.

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 respond to tones and bands of noise. From these results, one finds several properties of human hearing that might be modeled in a perceptual transformation. It is clear that the ear's frequency resolution not uniform on Hertz scale. It is also clear that loudness perception is related to signal power in a non-linear 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 we respond to more complex signals, such as speech. Some form of linear or non-linear "addition" is often invoked to combine known responses to simple signals to generate approximate responses to more complex signals. We have studied many of the perceptual transformations that have been proposed to model the human hearing process, and have drawn some conclusions about what level of detail is appropriate for this application[5-7]. 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 [15,16]. In particular, we have found that the non-uniform frequency resolution and the non-linear loudness perception seem to be the most important properties to model.

Thus we have arrived at a very simple, yet effective perceptual transformation. The non-uniform frequency resolution of the ear is treated by the use of a psychoacoustic frequency scale. The Hertz scale frequency variable, f is replaced with the Bark frequency variable b, scale using the relationship,

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

which can be found in [14]. Roughly speaking, on the Bark scale, equal frequency intervals are of equal perceptual importance. Second, signal intensity is converted to perceived loudness by taking a logarithm. These operations are most conveniently carried out in the frequency domain, resulting in what one might call "Bark Scale Loudness Distributions," or "Bark Scale Log PSDs (power spectral densities)."

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. For example, a perceptually-consistent distance measure should not be invariant to reversal of the test and reference signals shown in figure 1. Such an invariance equates excess loudness at a particular frequency with reduced loudness at that frequency. If we use a high frequency example, then a hissy audio signal will be equivalent to a muffled audio signal. While it may be possible to find a region where this equivalence is perceptually consistent, this equivalence is clearly not a general property of human auditory judgment. Also, a perceptually-consistent distance measure cannot be invariant to scrambling of samples along the time and frequency axis, yet many distance measures show this invariance.

Based on our experiences with conventional distance measures, and our understanding of human hearing and judgment, we conclude that listeners adapt and react differently to spectral deviations that span different time and frequency scales. We further observe that for the speech quality estimation application, maximal perceptual consistency over a wide range of distortion types requires a family of analyses at multiple frequency and time scales. 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 conclude 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.

We developed measuring normalizing blocks (MNBs) in response to the observations above. Two types measuring normalizing blocks are considered here. The first is the time measuring normalizing block (TMNB) and the second is the frequency measuring normalizing block (FMNB). Each of these blocks takes perceptually transformed reference (R(t, f)) and test (T(t, f)) signals as inputs and returns them and a set of measurements as outputs. These two building blocks are defined by Figures 2 and 3 respectively. The TMNB integrates over some frequency scale, then measures differences and normalizes the test signal at multiple **times**. Finally, the positive and negative portions of the measurements are integrated over time. In an FMNB the converse is true. An FMNB integrates over some time scale, then measures differences and normalizes the test signal at multiple **frequencies**. Finally, the positive and negative portions of the measurements are integrated over frequency. By design, both types of MNBs are idempotent.

This important property is illustrated in Figure 4 and simply says that a second pass through a given MNB will not further alter the test signal, and that second pass will result in a measurement vector of zeros. The idempotency of MNBs allows them to be cascaded and yet they measure the deviation at a given time or frequency scale once and only once.

In order to measure spectral deviations at multiple time and frequency scales, we have formed hierarchical structures of TMNBs and FMNBs, operating at decreasing scales. When used as distance measures in conjunction with the perceptual transformation described above, these structures appear to do a good job of emulating listeners' patterns of adaptation and reaction to spectral deviations. Two of these structures are described by Figures 5 and 6. Note that as always, a complexity-performance trade-off exists. The two structures presented were are chosen for their balance of relatively low complexity and relatively high performance. Structure 1 results in 13 measurements, while structure 2 results in 12 measurements. Because of the hierarchical nature of these structures, measurements from other than the top layer mean little individually, but a linear combination of the measurements has been found to be a good indicator of the perceptual distance between the two signals. The value that results from this linear combination is called auditory distance (AD):

$$AD = \sum_{i=1}^{N} w_i \cdot m_i \ .$$

Auditory distance is a positive quantity. When the reference and test signals are similar, AD is small. As the reference and test signals move apart perceptually, AD increases. A logistic function or some other "limiter function" can be used to map AD into a finite interval. This allows AD to correlate better with subjective quality or impairment judgments, which usually cover a finite range. When applied as shown in Figure 1, AD correlates with subjective quality or impairment judgments made on the device under test. The weights w_i can be selected to maximize this correlation. In the following section, we will see that the 13 (structure 1) or 12 (structure 2) weights are used to fit either 1226 or 9972 data points, resulting in fitting problems that are over-determined by factors of approximately 100 or 800. We have found that weights selected in this way tend to agree with our intuitions about human hearing and judgment.

Performance

To judge the usefulness of the AD values calculated by these two MNB structures as indicators of relative speech quality, we compared AD and six other objective estimators of speech quality with the results of formal subjective tests. Data from seven tests were available to us, and they are summarized in Table 1. For each of these seven tests, the coefficient of correlation between mean opinion score (MOS) and several objective estimators of speech quality was calculated. The resulting values for six existing objective estimators of speech quality are shown in Table 2. The correlation values were calculated after averaging all available subjective values for each condition to a single value for that condition. Similarly, for each condition, all available objective values were averaged to generate a single objective value for that condition. Thus, we refer to these correlation values as "per-condition" correlations.

From Table 2, P.861 appears to be the most reliable of these six objective estimators, across these seven tests. Since tests 1-4 contain conditions that are outside of the defined scope of P.861, we conclude that P.861 can sometimes make useful measurements outside of its scope. Because P.861 appears to be the most reliable of these six objective estimators, we use it as a benchmark to compare AD with.

Table 3 shows per-condition correlation values for AD as calculated by the two MNB structures. Since P.861 is used as a benchmark, that column from Table 2 is repeated as column 2 of Table 3 to allow for easy comparisons. Columns 3 and 4 of Table 3 show correlation values when the weights of the linear combination are optimized using only subjective scores from tests 1 and 2. These columns show that those weights result in an objective speech quality estimator that generalizes well to the other five tests. This generalization indicates that this approach does model hearing and perception, rather than inadvertently modeling some specific properties of tests 1 and 2.

To create the most effective estimator possible, we also optimized the weights using subjective scores from all seven tests. The resulting correlations are shown in columns 5 and 6 of Table 3. Across these seven tests, the second MNB structure appears slightly more useful than the first MNB structure. Both structures show reliable improvements over P.861 on tests 1-4. In unoptimized software implementations, it was found that either structure required about 1.2 million floating point operations to process 1 second (8000) samples of speech. Similarly, an unoptimized implementation of P.861 required about 1.7 million floating point operations.

Summary

ITS staff have been conducting research on perception-based objective audio quality assessment tools for several years[2,4-7]. The research is motivated by a need for objective audio and speech quality assessment tools that show significantly improved invariance to the technology employed in an audio or speech device under test. The basic premise of this approach is that by transforming signals into an appropriate perceptual domain, only information that is perceptually relevant is retained. That information is, by definition, both necessary and sufficient for the accurate assessment of audio quality, independent of the coding and transmission technology employed by the device under test. Results described here apply to an important and pervasive class of audio signals: 4 kHz bandwidth speech signals.

Our work has shown that for the speech quality estimation application, simpler perceptual transformations work as well or better than more complex ones. The use of a psychoacoustic frequency scale and logarithmic loudness growth function together provide a simple yet effective perceptual transformation. Perceptual consistency requirements suggest that a distance measure should be sensitive to the sign and the time scale or frequency scale of the spectral deviations between two audio signals. The time measuring normalizing block (TMNB) and frequency measuring normalizing block (FMNB) provide means of building such a distance measure. Because they are idempotent, these blocks can be cascaded and still measure the deviation at a given time or frequency scale once and only once.

Two hierarchical structures of mixtures of TMNB's and FNMBs, operating at decreasing scales have been developed. In effect, these structures decompose the reference signal in a space defined partly by human hearing and judgment, and partly by the test signal. The parameters of this dynamic decomposition are combined linearly to form a measure of the perceptual distance between those two signals. These structures have been evaluated relative to seven subjective tests. The results they produce correlate well with subjective scores, even when non-waveform coders and errored channels are considered.

Moving Forward

ITS staff believes that the results presented in this contribution show that the AD warrants further consideration from T1A1.5 as a candidate for standardization under standards project T1Q1.12. As a part of this process, ITS is prepared to provide an algorithmic description of the AD to T1A1.5 so that members can implement the measure themselves and perform the validation that their companies require for acceptance of this measure. In parallel with the validation conducted by the members of T1A1.5, a draft standard will be written based on the algorithmic description provided.

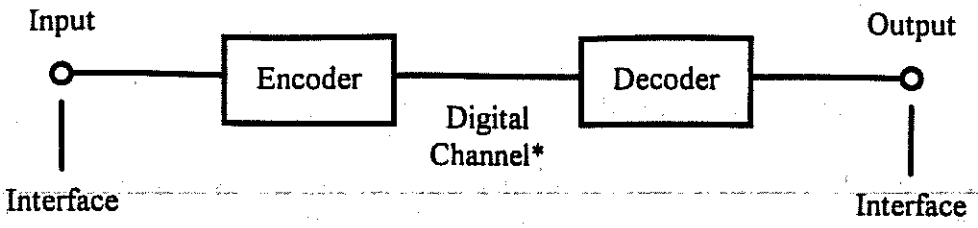
A draft scope for such a document could be as follows:

1 Scope, purpose, and application

1.1 Scope

This standard covers the operational assessment of one-way audio systems utilizing digital transport facilities to provide audio-visual services. It gives measurement parameters that may be used to detect changes in the current status of a system when used in comparison with a set of reference measurements on the same system made under initial provisioning circumstances. Additionally, the parameter specified in this standard may be utilized to characterize performance aspects of one-way audio signals.

This standard specifies methods of measurement for audio performance for one-way audio transmission service channels that employ digital transport, as shown in Figure A. The audio performance parameter identified within this standard is defined for the end-to-end transmission quality between the interfaces shown. Those interfaces are between the one-way audio transmission service providers and end users.



*A digital channel is implemented on a network composed of digital telecommunications components.

Figure A - One-way audio transmission service channel

The parameter given here is intended to be used for the specific applications identified in the clauses that follow. This document will be updated to include additional applications and measurements as research and development continues. Committee T1 will consider new methods and applications and add them here when there is industry consensus that they are appropriate. The following applications are beyond the scope of this standard:

a) Measuring the following performance aspects of an audio system: audio-visual interaction, and any implications of full duplex working (e.g., round-trip delay that reduces the conversational spontaneity);

b) Determination of the performance of audio signals with bandwidth greater than 4 kHz;

c) Measuring the performance aspects of one-way audio systems where the input and output interfaces shown in Figure A are not accessible;

d) Prediction of user opinion (e.g., mean opinion or degradation mean opinion scores) of a particular system or the applicability of a particular system to a particular application.

1.2 Purpose

The purpose of this standard is to assure the uniform application of, provide a framework for, and provide definitions of standard audio performance parameters for one-way audio signals transported digitally, in conjunction with video signals, on a portion of the telecommunications network. This standard is intended to be especially useful as a basis for comparing the present operational readiness of a system with the same system's past performance. This standard is intended to provide a common understanding by manufacturers, carriers, and their customers.

1.3 Application

The initial application for the standard is:

Determining the audio signal quality, relative to reference measurements, of one-way audio systems utilizing digital transport facilities.

Committee T1 recognizes that, in the first applications of this standard, accuracy may depend heavily on the expertise and objectivity of the technical staff performing the measurements. It is expected that, over time, industry experience in applying the standard will reduce the need for a highly-trained staff to correctly apply the standard.

References

[1] ANSI T1.801.03-1996, "Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment," ANSI, Inc., New York, 1996.

[2] Voran, S., "An Overview of ITS Research on Perception-Based Audio Quality Assessment," Contribution to T1A1.6, Contribution Number T1A1.6/94-030, March 1994.

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

[4] Voran, S. & Sholl, C., "An Update on ITS Perception-Based Audio Quality Assessment Research," Contribution to T1A1.7, Contribution Number T1A1.7/94-051, September 1994.

[5] Voran, S., "May 1995 Update on ITS Perception-Based Audio Quality Assessment Research," Contribution to T1A1.7, Contribution Number T1A1.7/95-045, May 1995.

[6] Voran, S. "Observations on Auditory Excitation and Masking Patterns," *Proceedings of the 1995 Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York, October 1995.

[7] Voran, S. & Sholl, C. "Perception-based Objective Estimators of Speech Quality", *Proceedings of the 1995 IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, MD, September 1995.

[8] Beerends, J.G. & Stemerdink, J. A., "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 40, Dec. 1992, pp. 963-978.

[9] Hollier, M. P., Hawksford, M. O. & Guard, D. R., "Characterization of Communications Systems Using a Speech-Like Test Stimulus," *93rd Convention of Audio Engineering Society*, San Francisco, USA, Oct. 1992.

[10] Humes, L.E. and Jesteadt, W., "Models of the Additivity of Masking," *J. Acoust. Soc. Am.*, vol. 85, pp. 1285-1294, March 1989.

[11] Nielsen, L. B., "Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners," *Oticon Internal Report no. 43-8-4*, Snekkersten, Denmark, 1993.

[12] Paillard, B., Mabilleau, B. & Morissette, S., "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Audio Eng. Soc.*, vol. 40, no. 1, Jan. 1992.

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

[14] Wang, S., Sekey, A. & Gersho, A., "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. on Sel. Areas in Comm.*, vol.10, no.5, Jun. 1992.

[15] Beerends, J.G. & Stemerdink, J. A. "A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 42, pp. 115-123, March 1994.

[16] Hauenstein, M. "Comparative Study of Psychacoustics-Based Objective Speech-Quality Measures Using Markov-SIRPS," *Proc. Speech Quality Assessment Workshop*, Bochum, Germany, November 1994.

[17] Kitawaki, N. "Quality Assessment of Coded Speech," *in Advances in Speech Signal Processing*, Furui, S., et. al., Ed., Marcel Dekker, 1992.

[18] Be'ery, Y., et. al. "An Efficient Variable-Bit-Rate Low-Delay CELP Coder," *in Advances in Speech Coding*, Atal, B.S., et. al., Ed., Kluwer Academic Publishers, 1990.

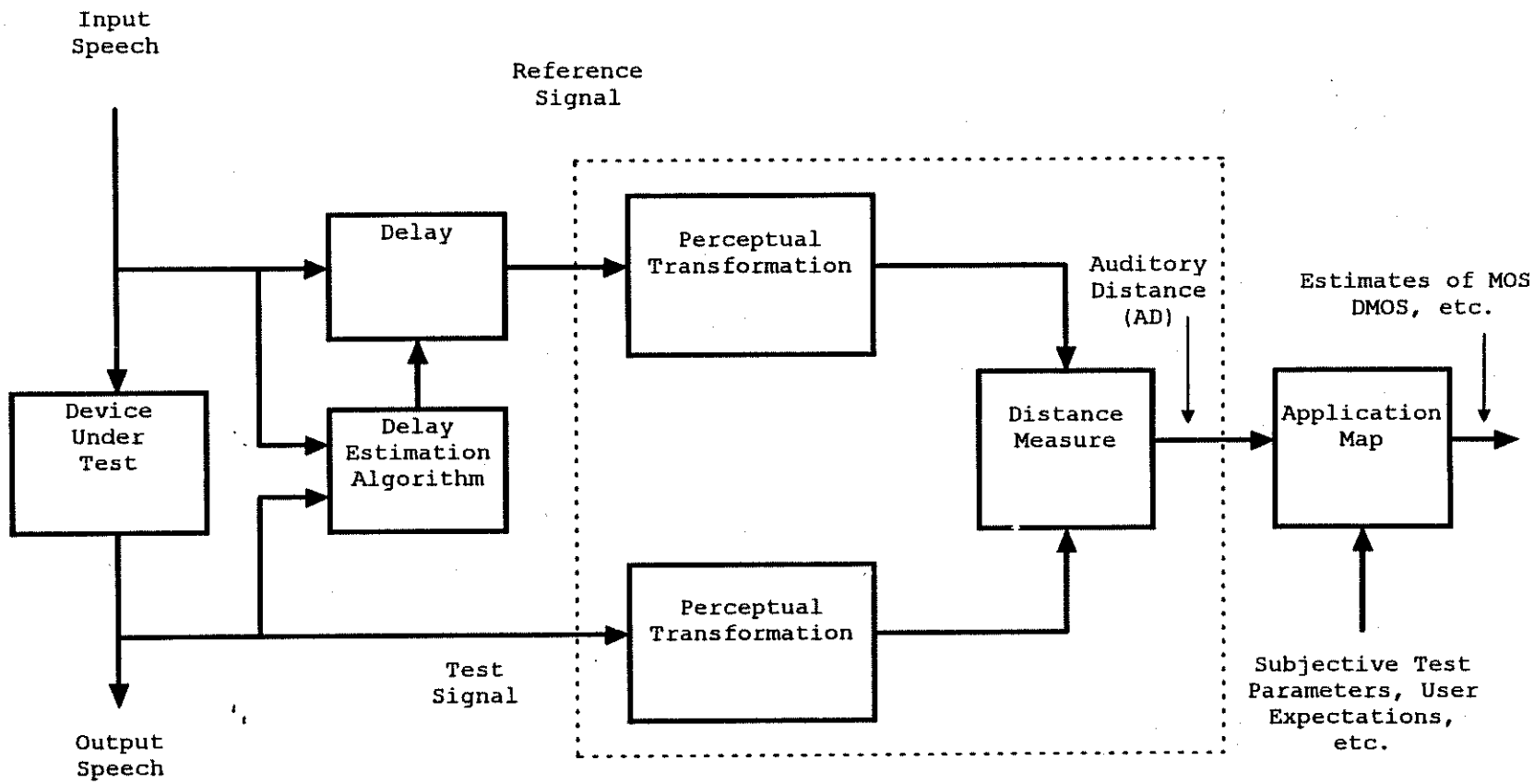


Figure 1: High-Level Description of the ITS Approach

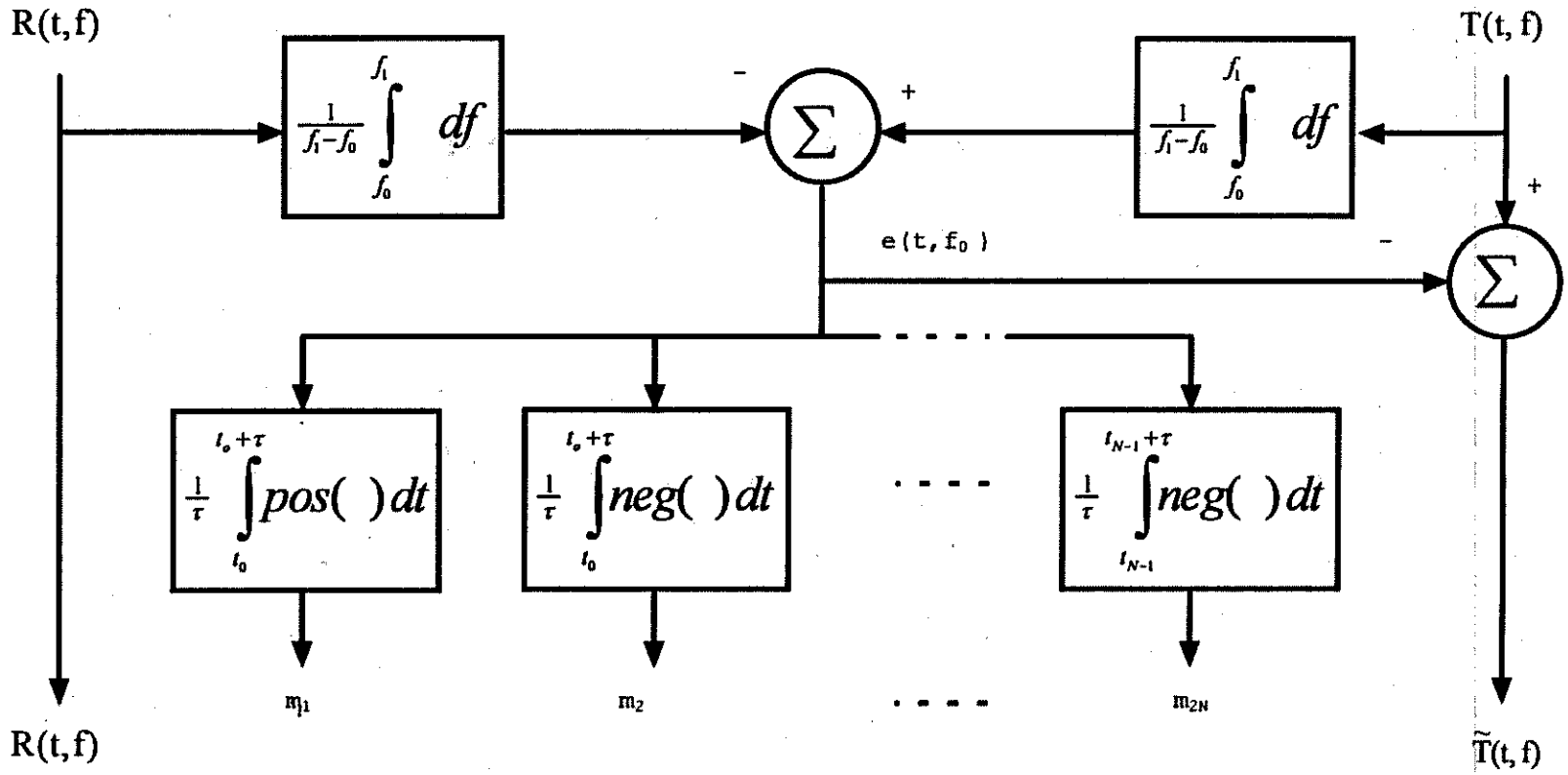


Figure 2: Time Measuring Normalizing Block (TMNB)

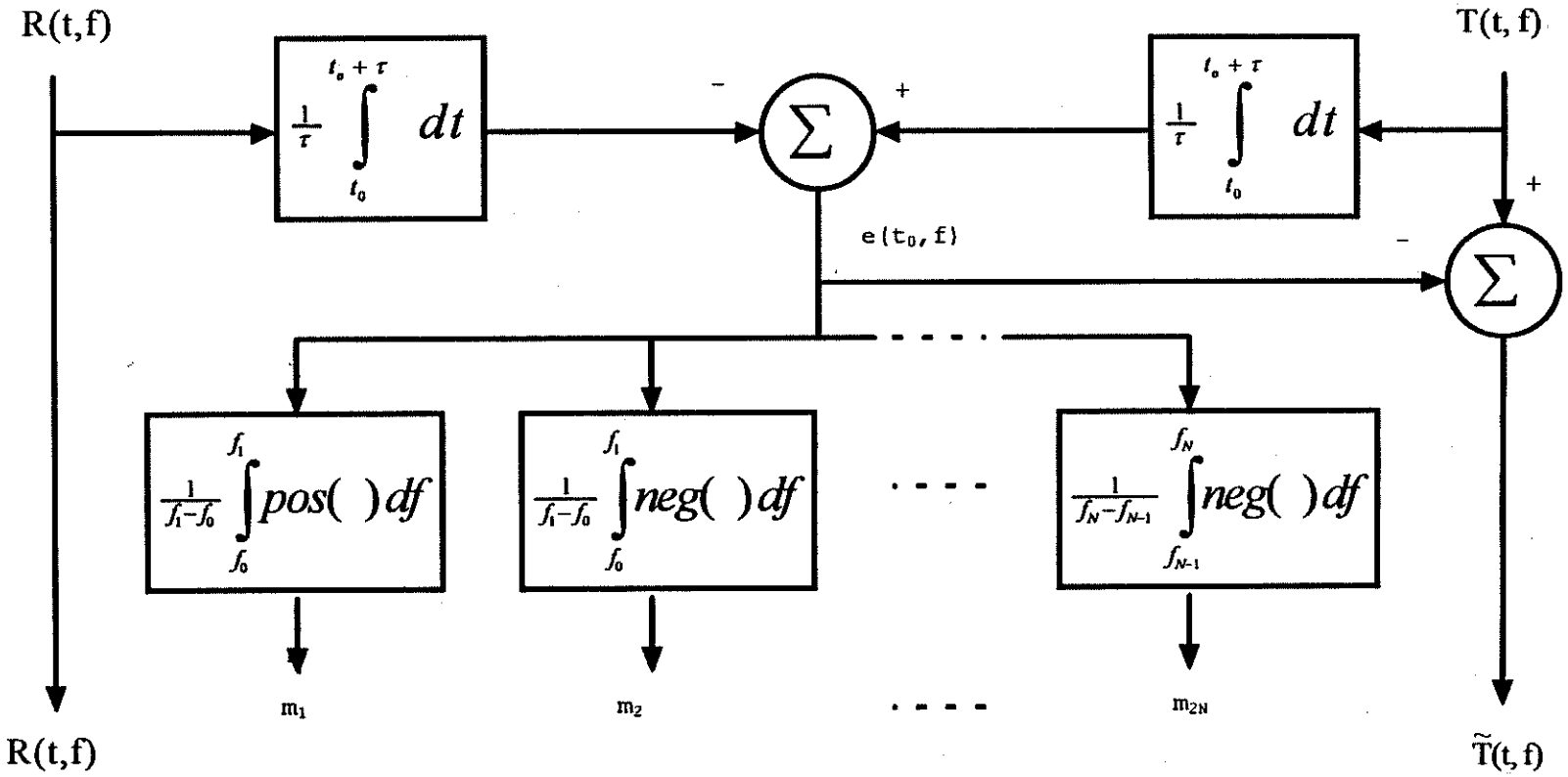
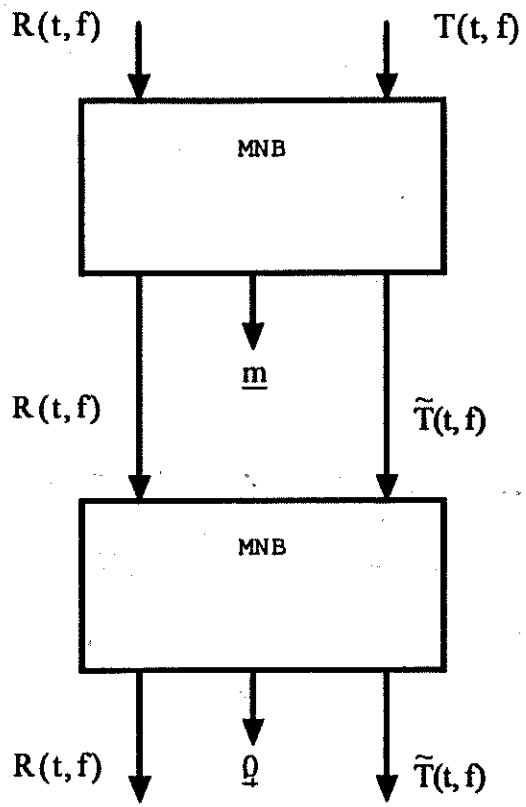


Figure 3: Frequency Measuring Normalizing Block (FMNB)



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

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

Figure 4: MNBs are Idempotent

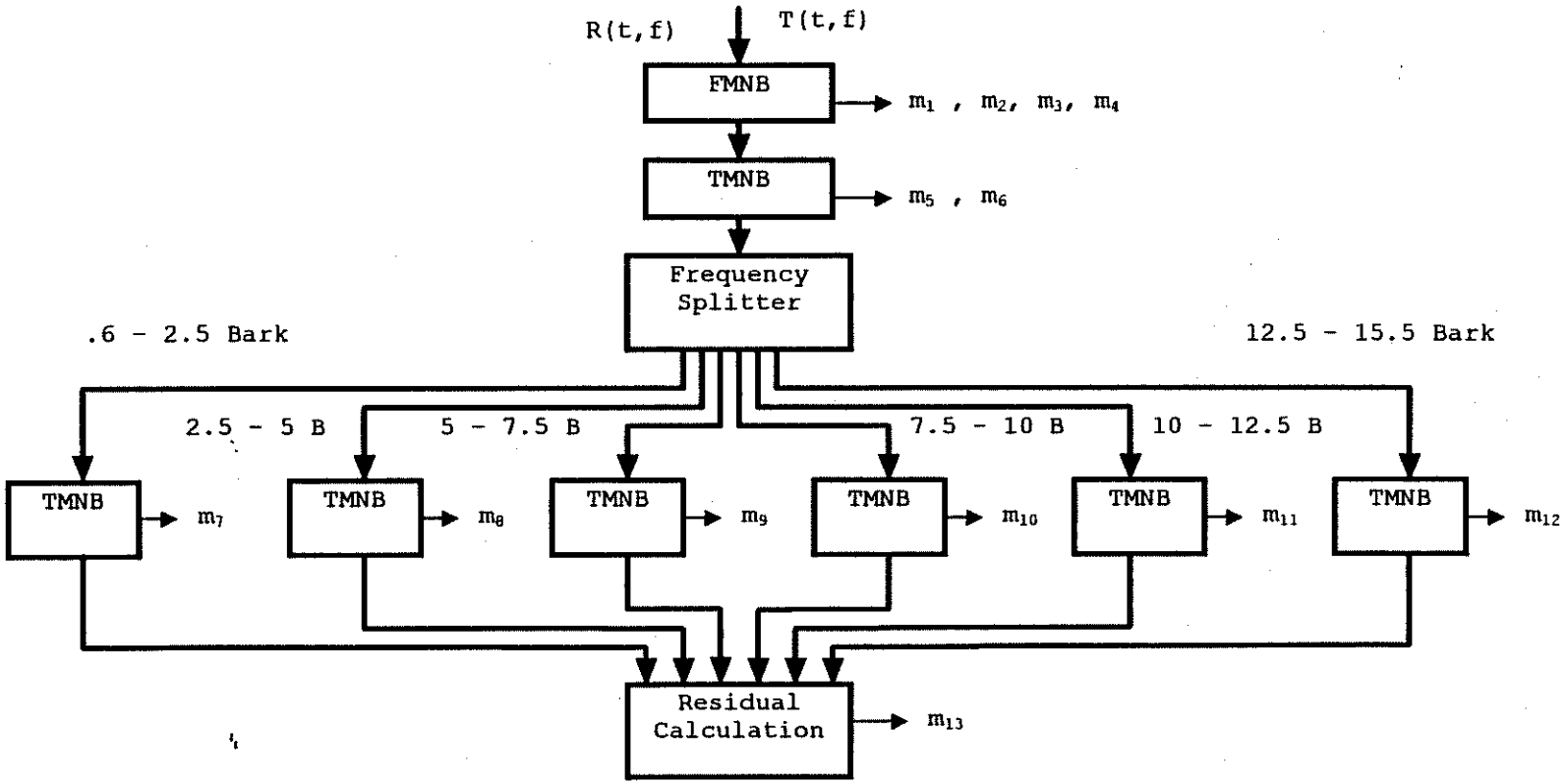


Figure 5: MNB Structure 1

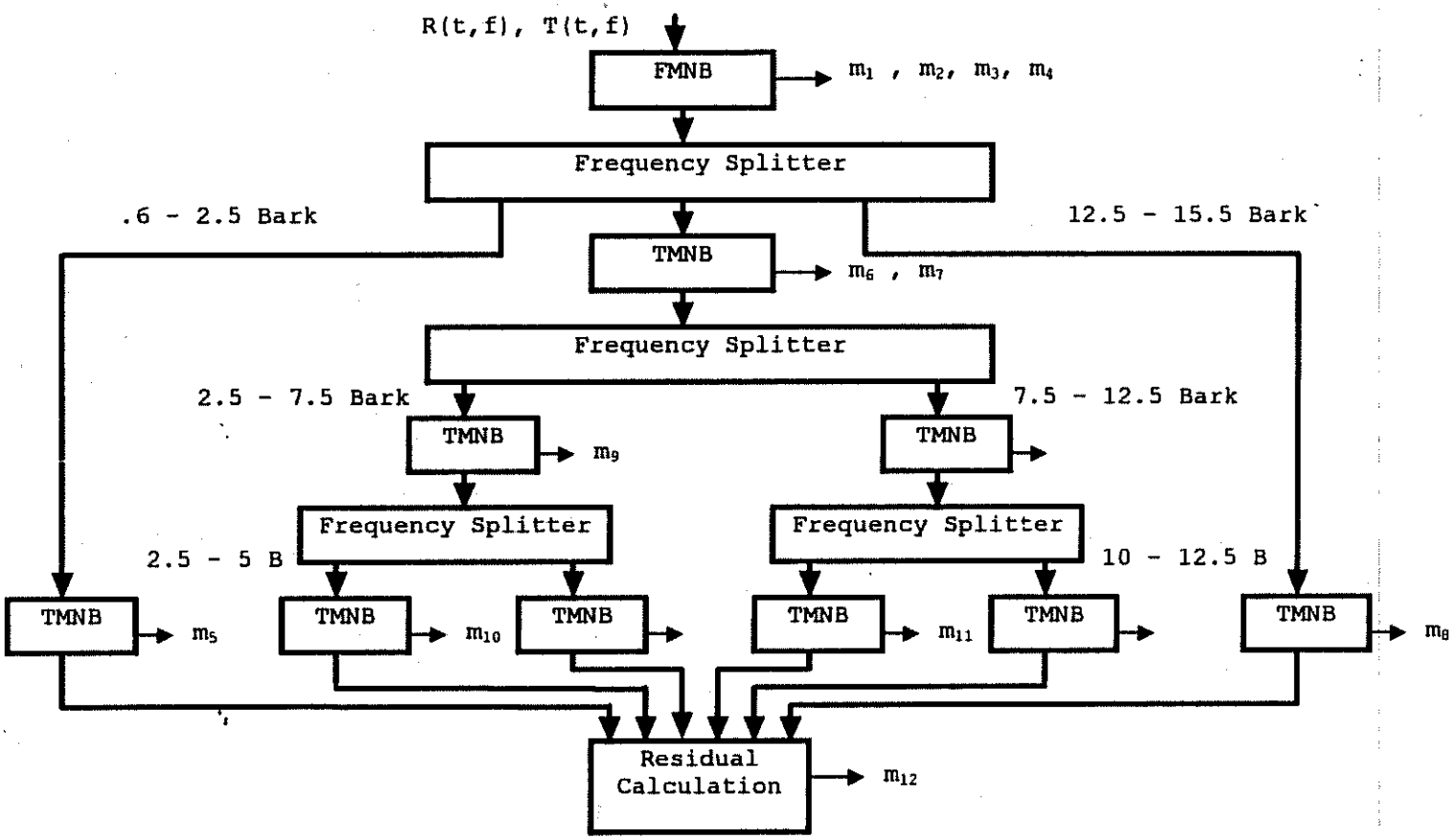


Figure 6: MNB Structure 2

Test	Number of Conditions	Conditions	Language	Talkers per Condition	Files	Minutes
1	22	PCM@64,48,40 ADPCM@32 (x1, 2, 3, 4) APC@16 (2 versions) Proprietary Codec@16 SELP@4.8 (2 versions) LPC@2.4 MNRU (6 levels) Narrow Band MNRU (3 levels)	North American English	4	176	8
2	35	PCM@64 Proprietary CELP-1@8 over 9 RF channels (bit errors) Proprietary CELP-2@8 over 9 RF channels (bit errors) AMPS over 9 RF channels MNRU (7 levels)	"	6	1050	100
3	27	ADPCM@32 (clear and errored) CVSD@32,16 (clear and errored) VSELP@8 (clear channel) CELP@4.8 (clear and errored) IMBE@4.8,2.4 (clear channel) STC-1@4.8,2.4 (clear and errored) STC-2@2.4 (clear channel) LPC@2.4 (clear and errored) POTS MNRU (8 levels)	North American English	6	1994	225

Test	Number of Conditions	Conditions	Language	Talkers per Condition	Files	Minutes
4	38	ADPCM (x4) G.728@16 VSELP@8 Proprietary Non-Waveform Codec@6.4 Prop. Non-Waveform Codec@4 (3 input levels) Prop. Non-Waveform Codec @4 (x2) Prop. Non-Waveform Codec @4 + G.726 Prop. Non-Waveform Codec @4 + VSELP Prop. Non-Waveform Codec @4 + GSM Prop. Non-Waveform Codec @4 + G.728 + G.728 MNRU (7 levels) IRS and Flat for all conditions	"	8	2432	264
5	20	G.711 (x1, 2, 4, 8, 16) G.726@32 (x1, 2, 4) G.728 Candidate (x1, 2, 4) MNRU (9 levels)	"	4	1440	206
6	"	"	Japanese	4	"	188
7	"	"	Italian	4	"	131

The notation "xN" is used to indicate N tandems or passes through the indicated device.
The notation "codec1 + codec2" is used to indicate that two different codecs were tandemed to create a single condition.

Table 1: Summary of Material in Seven Subjective Tests

Test	SNR[17]	SNRseg[17]	PWSNRseg[18]	CD[17]	BSD[14]	P.861[3]
1	.347	.387	.384	.488	.825	.929*
2	.523	.521	.621	.730	.732	.941*
3	.295	.494	.507	.615	.367	.795*
4	.247	.221	.637	.789	.862	.973*
5	.226	.267	.525	.947	.919	.985
6	.271	.313	.503	.933	.851	.986
7	.318	.340	.543	.976	.892	.976

* These tests include conditions that are outside the defined scope of ITU-T Recommendation P.861.

Table 2: Per-condition Coefficients of Correlation between Subjective Scores and Objective Estimators

Test	P.861	MNB-1	MNB-2	MNB-1	MNB-2
		Weights optimized using only tests 1 and 2.		Weights optimized using tests 1-7.	
1	.929*	.933	.954	.934	.952
2	.941*	.955	.950	.952	.946
3	.795*	.951	.929	.947	.938
4	.973*	.956	.969	.977	.980
5	.985	.950	.959	.982	.981
6	.986	.962	.968	.979	.981
7	.976	.967	.970	.976	.983

* These tests include conditions that are outside the defined scope of ITU-T Recommendation P.861.

Table 3: Per-condition Coefficients of Correlation between Subjective Scores and Objective Estimators