To appear in: Proceedings of the 1997 IEEE Workshop on Speech Coding, September 7-10, 1997.

ESTIMATION OF PERCEIVED SPEECH QUALITY USING MEASURING NORMALIZING BLOCKS

Stephen Voran

Institute for Telecommunication Sciences, U.S. Department of Commerce, NTIA/ITS.N3
325 Broadway, Boulder, Colorado 80303, sv@bldrdoc.gov

ABSTRACT

We describe a new approach to the estimation of perceived speech quality. The approach uses a simple, but effective, perceptual transformation to emulate hearing and a hierarchy of Measuring Normalizing Blocks (MNB's) to emulate auditory judgment. The resulting estimates were correlated with the results of seven subjective listening tests. Together, these seven tests include 182, 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit-rates ranging from 2.4 to 64 kbps. When compared with six other estimators, the MNB approach offers significant improvements in many cases, particularly at lower bit-rates, and when bit errors or frame erasures are present.

1. BACKGROUND

Perceived speech quality is measured most directly by subjective listening tests. Because these tests often are slow and expensive, numerous attempts have been made to supplement them with objective estimators of perceived speech quality. Of 26 codecs described at the 1995 IEEE Workshop on Speech Coding for Telecommunications, 11 had been tested in formal subjective tests[1]. Segmental SNR (SNRseg)[2] or SNR[2] was used to estimate perceived speech quality in 10 cases, Cepstral Distance (CD)[2] was used twice, and Bark Spectral Distortion (BSD)[3] was used once. ITU-T Recommendation P.861 describes a perceived speech quality estimator, but its scope is limited to higher bit-rate speech codecs operating over error-free channels[4]. As shown below, when a broad range of speech coding and transmission conditions is considered, none of these estimators is as reliable as one might wish.

Researchers have recently begun to include explicit models for some of the known attributes of human auditory perception in estimators of perceived speech or audio quality[3-7], sometimes resulting in modest improvements. The limited success of the perception-based approach might be traced to two sources. First, while detailed models for the detectability and perceived loudness of combinations of tones and narrow bands of noise are well established, the nonlinear, time-varying nature of hearing makes properly aggregating those models into practical models for the processing of more general signals (e.g. speech) a very difficult task. Second, the perception of speech quality involves both hearing and judgment, but detailed models for hearing have often been followed by less insightful models for judgment.

2. MEASURING NORMALIZING BLOCK APPROACH

Our studies[7,8] have led us to reverse the traditional emphasis, resulting in a simpler model for hearing, and a more sophisticated model for judgment. We have studied the outer-middle ear transfer function, absolute hearing thresholds, equal loudness curves, and time and frequency domain masking effects. None of these appear to be helpful in the estimation of the perceived quality of 4-kHz bandwidth speech. These results agree with [5,6]. Our hearing model contains only a frequency mapping from Hertz to Bark, and a logarithmic transformation from power to approximate perceived loudness.

Distance measures seek to emulate auditory judgment by comparing two perceptually transformed signals (e.g. coder input and decoder output). Many existing conventional distance measures display properties that are clearly inconsistent with auditory judgment. We observe that listeners adapt and react differently to spectral deviations that span different time and frequency scales. Thus, the speech quality estimation problem can benefit from a family of analyses at multiple frequency and time scales, where spectral deviations at one scale are measured and removed so that they are not counted again as part of the deviations at other scales. Working from larger to smaller scales is most likely to emulate listeners' patterns of adaptation and reaction. In light of these observations, we elected to form a distance measure from a hierarchy of Measuring Normalizing Blocks (MNB's).

Each of these blocks takes perceptually transformed reference and test signals, $R(t,f)$ and $T(t,f)$, as inputs, and returns a set of measurements and a normalized version of $T(t,f)$. A Time MNB (TMNB) integrates over some frequency scale, measures differences, and normalizes the test signal at multiple *times*. The positive and negative portions of the measurements are then integrated over time. The TMNB operating on frequencies from fl to fu using the measurement time intervals defined by t_i, i=0 to N, normalizes $T(t,f)$ to $\tilde{T}(t,f)$ and generates $2N$ measurements m_j:

$$e(t,fl) = \frac{1}{fu-fl}\int_{fl}^{fu} T(t,f)\,df - \frac{1}{fu-fl}\int_{fl}^{fu} R(t,f)\,df$$

$$\tilde{T}(t,f) = T(t,f) - e(t,fl)\ , \quad m_{2i-1} = \frac{1}{t_i - t_{i-1}}\int_{t_{i-1}}^{t_i} \max(e(t,fl),0)\,dt$$

$$m_{2i} = \frac{-1}{t_i - t_{i-1}}\int_{t_{i-1}}^{t_i} \min(e(t,fl),0)\,dt, \quad i = 1 \text{ to } N.$$

The Frequency MNB (FMNB) definition is analogous, with the roles of time and frequency exchanged. By design, both types of MNB's are idempotent. If $\mathrm{MNB}(R,T) = (R,\tilde{T},\ \underline{m})$, then $\mathrm{MNB}(R,\tilde{T}) = (R,\tilde{T},\ \underline{0})$. The idempotency of MNB's allows them to be cascaded in hierarchies and still measure the deviation at a given time or frequency scale once and only once.

We have formed hierarchical structures of TMNB's and FMNB's, operating at decreasing scales. Two structures, MNB-1 and MNB-2, are described below. They were chosen for their balance of relatively low complexity and high performance as estimators of perceived speech quality. Other useful structures exist, and may address open issues in audio quality estimation, automatic speech or speaker recognition, layered coding, signal enhancement, or other areas.

Both MNB structures start with an FMNB at the longest available time scale, resulting in four measurements that cover the lower and upper edges of the telephone speech band (0-500 Hz and 3-3.5 kHz). In MNB-1, a TMNB is then applied at the largest frequency scale (≈15 Bark). Six additional TMNB's are then applied at smaller scales (2-3 Bark). Finally, a residual measurement is made. In MNB-2, the middle portion of the band undergoes two levels of binary band splitting, resulting in bands that are 2-3 Bark wide. The extreme top and bottom portions of the band are each treated once by separate TMNB's, and a residual measurement is made. MNB-1 generates 13 linearly independent measurements, and MNB-2 generates 12. A linear combination of these measurements has been found to be a good estimator of perceived speech quality.

3. RESULTS AND OBSERVATIONS

We have correlated the MNB estimates and six other estimates with the results of seven absolute category rating subjective listening tests. These seven tests, detailed in Table 1, use 18.7 hours of flat and IRS-filtered source speech, passed through 182, 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit-rates ranging from 2.4 to 64 kbps, and some analog conditions. The coefficients of correlation in Table 2 demonstrate the limitations of SNR, SNRseg, and perceptually-weighted SNRseg[9] as estimators of perceived speech quality. CD, BSD, and P.861 show mixed results.

Columns 8 and 9 of Table 2 show correlation values for MNB-1 and MNB-2 when the linear combinations of measurements are optimized using only Tests 1 and 2 (13 or 12 variables are used to fit 1226 data points). This limited optimization results in estimators that generalize well to the other five tests. To create the most effective estimator, one must use all the available data. Columns 10 and 11 show correlation values when all seven tests are used to optimize the linear combination (13 or 12 variables, 9972 data points). The resulting optimized weights tend to agree with our intuitions about human hearing and judgment. Both MNB-1 and MNB-2 show improvements over P.861 on Tests 1-4, which contain lower rate codecs, bit error, and frame erasure conditions.

REFERENCES

[1] Proc. 1995 IEEE Workshop on Speech Coding for Telecom., Annapolis, MD, Sep. 1995.

[2] N. Kitawaki, "Quality assessment of coded speech," in *Advances in Speech Signal Processing*, S. Furui & M. Sondi, Ed., New York: Marcel Dekker, 1992.

[3] S. Wang, A. Sekey & A. Gersho, "An objective measure for predicting subjective quality of speech coders," IEEE J. on Sel. Areas in Comm., vol. 10, no. 5, pp. 819-829, Jun. 1992.

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

[5] J.G. Beerends & J.A. Stemerdink, "A perceptual speech quality measure based on a psychoacoustic sound representation," J. Audio Eng. Soc., vol. 42, Mar. 1994.

[6] M. Hauenstein, "Comparative study of psychoacoustics-based objective speech-quality measures ..." Proc. Speech Quality Assessment Workshop, Bochum, Germany, Nov. 1994.

[7] S. Voran & C. Sholl, "Perception-based objective estimators of speech quality," pp. 13-14, in [1].

[8] S. Voran, "Observations on auditory excitation and masking patterns," Proc. 1995 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, Oct. 1995.

[9] Y. Be'ery, et al. "An efficient variable-bit-rate low-delay CELP coder," in *Advances in Speech Coding*, B.S. Atal, et al., Ed., Boston: Kluwer Academic Publishers, 1991.

Test*	1	2	3	4	5
Num. Conds.	22	35	27	38	20
Cond. List	PCM, ADPCM, APC, SELP, LPC, MNRU (Tandems)	PCM, CELP, AMPS , MNRU (Frame Erasures)	ADPCM, CVSD, VSELP, CELP, IMBE, STC, LPC, POTS, MNRU (Bit Errors)	ADPCM, CELP, VSELP, IMBE, AMBE, MNRU, (Mixed Tandems)	PCM, ADPCM, CELP, MNRU (Tandems)
Rates (kbps)	2.4-64	8-64	2.4-32	6.4-32	16-64
Talker/Cond.	4	6	6	8	4
Num. Files	176	1050	1994	2432	1440

*Tests 1-5 are in English, Tests 6 and 7 are identical to Test 5, but are in Japanese and Italian, respectively.

Table 1. Summary of Speech Material in Subjective Listening Tests

Test*	SNR	SNRseg	PWSNRseg	CD	BSD	P.861	MNB-1†	MNB-2†	MNB-1‡	MNB-2‡
1	.347	.387	.384	.488	.825	.929	.933	.954	.934	.952
2	.523	.521	.621	.730	.732	.941	.955	.950	.952	.946
3	.295	.494	.507	.615	.367	.795	.951	.929	.947	.938
4	.247	.221	.637	.789	.862	.973	.956	.969	.977	.980
5	.226	.267	.525	.947	.919	.985	.950	.959	.982	.981
6	.271	.313	.503	.933	.851	.986	.962	.968	.979	.981
7	.318	.340	.543	.976	.892	.976	.967	.970	.976	.983

*Tests 1-4 contain conditions outside the scope of P.861. †Optimized using only Tests 1 and 2. ‡Optimized using Tests 1-7.

Table 2. Per-condition Coefficients of Correlation with Subjective Scores