Proceedings of the 1995 IEEE Workshop on Speech Coding, September 20-22, 1995, Annapolis, MD

Perception-based Objective Estimators of Speech Quality

Stephen Voran[1] and Connie Sholl[2]

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

[2]Netrix Corporation, 2402 Clover Basin, Suite A, Longmont, Colorado 80503, USA, csholl@netrix.com

ABSTRACT

Four proposed perception-based techniques for objectively estimating speech quality and three traditional estimators are applied to coded speech samples. Agreement between objective estimates and corresponding subjective test scores is reported. Several observations on key elements of perception-based estimators are offered.

1 Background

Speech coding often involves a four-way compromise among complexity, delay, bit-rate, and the perceived quality of decoded speech. The most critical perceived quality measurements will always rely on formal subjective tests. However, the costs associated with formal subjective tests are not justified in some situations. Specifically, much coder development work relies on objective estimators of perceived speech quality, along with "informal listening tests." For example, of the 30 coders described at the 1993 IEEE Workshop on Speech Coding for Telecommunications, only nine had been tested in formal subjective tests, while several different objective quality estimators were employed [1]. Segmental SNR (SNRseg) was applied in four cases, spectral distortion measures were used in three cases, while SNR, perceptually-weighted SNRseg (PWSNRseg), Bark Spectral Distortion (BSD), and Cepstral Distance (CD) were each used once.

This diversity makes comparisons difficult. Further, it is unclear how reliable some of these estimators are. In particular, SNR and SNRseg are not generally reliable estimators of perceived speech quality, and their continued popularity is likely due to their history and simplicity [2]. A reliable, widely used and accepted objective technique for estimating perceived speech or audio quality would be quite valuable. Much of the recent work in this area has followed a perception-based approach. See [3-11] for examples.

2 The Perception-based Approach

A high-level description of the perception-based approach to objectively estimating the perceived quality of coded speech is given in Figure 1. The basic premise of the approach is that by transforming signals into an appropriate perceptual domain, only perceptually-relevant information is retained. By definition, this information is both necessary and sufficient for the accurate assessment of perceived speech quality, independent of the type of coding and transmission used. The perceptually-transformed speech signals are then compared by a distance measure that estimates the perceived quality of the coded speech.

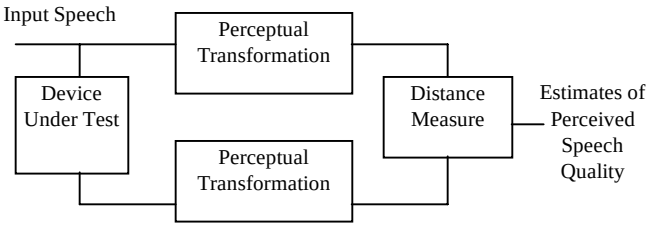

Figure 1. Perception-based Approach to Quality Estimation

To test their reliability, four perception-based estimators were applied to coded speech samples that were also formally evaluated in subjective tests. The four estimators were PWSNRseg [3], BSD [4], PSQM1 [5,6], and PSQM2 [5-7]. Maximization of PWSNRseg is equivalent to the minimization of perceptually-filtered mean-squared coding error done by many analysis-by-synthesis coders. PSQM2 uses an asymmetric distance measure, but is otherwise identical to PSQM1. Three traditional estimators, SNR, SNRseg, and CD were calculated as well.

Three groups of coded speech samples (male-female balanced, North American English) were used. The first explores the impact of radio channels (0-6.4% BER) on Federal Standard 1016 CELP coding (12 conditions, 20 seconds each). A second group includes two proprietary CELP-based coders using radio channels (0-3% BER), analog FM radio, μ-law PCM, and the MNRU (34 conditions, 3 minutes each). The third group includes the MNRU and 13 coders using error-free channels (PCM, ADPCM, APC, SELP, and LPC), (22 conditions, 20 seconds each).

Table 1 summarizes the results with magnitudes of coefficients of correlation, calculated across all conditions, using mean objective and subjective values for each condition. These results indicate that SNR, SNRseg, PWSNRseg and CD are not reliable indicators of perceived speech quality across all the conditions considered here. BSD and the PSQM's show more reliable readings across coding technologies, indicating they may be more successful at emulating human perception. For some groups of coders, higher correlation values have been reported [4,6].

	Group 1	Group 2	Group 3
SNR	.46	.54	.40
SNRseg	.68	.56	.46
PWSNRseg	.34	.55	.50
CD	.97	.74	.51
BSD	.93	.68	.83
PSQM1	.74	.79	.74
PSQM2	.89	.89	.86

Table 1. Measured Magnitudes of Coefficients of Correlation

3 Observations on Perceptual Transformations and Distance Measures

Perceptual transformations are generally based on the results of human perception experiments that use tones or bands of noise. Differing interpretations of these experimental results and differences in how they are applied to speech signals have lead to a range of plausible perceptual transformations. To study how these differences impact speech quality estimates, four perceptual transformation parameters were varied and the resulting signals were passed into three time-averaged distance measures: a normalized mean-squared error measure, a log ratio measure and a direction cosine measure. Trends in the agreement between the resulting objective estimates and subjective scores were observed across the three groups of speech samples. This breadth was intended to free the analysis from the specifics of one coding technology or subjective test.

This study led to the following five observations: (1) Temporal smoothing of spectral representations [5] does not increase and often decreases agreement with subjective scores. (2) Calculating a "level-dependent upward spread of excitation" as in [5,8,9] adds little or no information that is useful for this application. Instead, a significantly simpler "fixed upward spread of excitation" calculation [4,10] might be used. (3) No consistent preference was found for the rule by which excitation components are added [5]. Thus, the simplest rule (adding of power) might be used without sacrificing performance. (4) No consistent preference was found for the compressive nonlinearity that converts power to perceived loudness (Stevens' Law [4], Zwicker's Law, or "highly compressed" [7]). (5) None of the three distance measures is consistently superior to the others. The first three observations are in accordance with [11] while the first and fourth also agree with [6].

From these observations, it appears that highly detailed perceptual transformations are not particularly beneficial as inputs to the distance measures used here. Table 1 shows that PSQM2 is preferred to PSQM1, and this difference is due to the distance measure alone. Results in [11] also point to the value of improved distance measures. Thus, our current efforts are focussed on distance measures that may more fully exploit the information available in the time and frequency distributions of perceptually-transformed speech signals. These efforts may lead to efficient estimators of perceived speech quality that are maximally effective and robust across a wide range of codec types and distortion sources, thus providing cost-effective yet meaningful input to the perceived speech quality dimension of the four-way speech coding compromise.

4 References

1. Proc. IEEE Workshop on Speech Coding for Telecom., Sainte-Adèle, Québec, Canada, Oct. 1993.

2. N. Kitawaki, "Quality Assessment of Coded Speech," in *Advances in Speech Signal Processing*, S. Furui & M. Sondi, Ed., Marcel Dekker, 1992.

3. 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., Kluwer Academic Publishers, 1990.

4. S. Wang, et. al., "An Objective Measure for Predicting Subjective Quality of Speech Coders," IEEE J. on Sel. Areas in Comm., vol.10, pp. 819-829, June 1992.

5. J.G. Beerends & 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.

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

7. J.G. Beerends, “Improving the Perceptual Speech Quality Measure,” Contribution to ITU-T, SG12, SQEG, SQ-17.94, Geneva, Feb. 1994.

8. M.P. Hollier, et. al., “Characterization of Communications Systems Using a Speech-Like Test Stimulus,” *93rd Convention of the Audio Engineering Society*, San Francisco, USA, Oct. 1992.

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

10. B. Paillard, et. al., "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Audio Eng. Soc.*, vol 40, pp. 21-31, Jan. 1992.

11. M. Hauenstein, “Comparative Study of Psychacoustics-Based Objective Speech-Quality Measures Using Markov-SIRPS,” Proc. Speech Quality Assessment Workshop, Bochum, Germany, Nov. 1994.