Proceedings of the 1999 IEEE Speech Coding Workshop, June 20-23, 1999, Porvoo, Finland

ADVANCES IN OBJECTIVE ESTIMATION OF PERCEIVED SPEECH QUALITY

S. Voran

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

ABSTRACT

We present two techniques that can be used to enhance objective estimators of perceived speech quality. Frame normalization and frame-energy plane partitioning are described and applied to a log-spectral-error-based estimator. The resulting estimators are compared with each other and with two established estimators. This is done through correlation with MOS values from 17 formal subjective tests. We find that the proposed techniques significantly improve the log-spectral-error-based estimator.

1. INTRODUCTION

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. The Perceptual Speech Quality Measure (PSQM) [1] has been recognized as a useful estimator for codecs that preserve or nearly preserve waveforms connected by error-free transmission channels [2]. Measuring Normalizing Block (MNB) estimators [3][4] have broader applicability that also includes lower-rate non-waveform codecs and transmission channels that include bit errors and frame erasures [5].

MNB estimators were motivated by the observation that listeners adapt and react differently to spectral deviations that span different time and frequency scales. In a simple model for adaptation and reaction, MNB estimators iterate to measure and remove spectral deviations that cover different time and frequency scales. The iterations start at larger scales and progress to smaller scales. MNB estimators are effective and somewhat intuitive, but their iterative nature makes it difficult to mathematically demonstrate the merit of a single measure-and-normalize step.

In this paper, we provide such a demonstration by adding a single measure-and-normalize step to a log-spectral-error-(LSE) based estimator. In addition, we introduce frame-energy plane partitioning and show how it can further enhance the LSE-based estimator. We then compare the performance of the resulting LSE-based estimators with each other and with the PSQM and MNB estimators.

2. FRAME NORMALIZATION

In this section we describe a simple LSE-based estimator and then augment it with a single, frame-based, measure-and-normalize step. Let $x(t)$ represent a speech signal that is fed into a system under test. Let $y(t)$ be the resulting output. To estimate the perceived speech quality of the system under test, we estimate the perceived distance between $x(t)$ and $y(t)$. Preliminary steps include normalizing the long-term mean and variance of each signal to zero and one respectively and removing any time shifts. Next we form 32 ms time-domain frames (256 samples, 8000 samples/sec, 50% overlap) and transform them to frequency-domain frames using a Hanning window followed by FFT and magnitude-squared operations. (We use an unscaled FFT: if a length N sequence of ones is transformed, the resulting value in the DC bin will be N.) The result is 129 frequency-domain samples from DC to Nyquist. For both signals, these samples ($\{x_i\}$ and $\{y_i\}$) are then logarithmically transformed and clipped according to

$$\tilde{x}_i = \max\left(10 \cdot \log_{10}(x_i), -40\right), \text{ when } 0 < x_i, \quad (1)$$
$$= -40, \text{ otherwise.}$$

Since peak values of $\tilde{x}_i$ are near 50, this clipping limits the log-spectrum dynamic range to about 90 dB, preventing inaudible distortions from dominating the calculations that follow. An LSE can then be calculated as

$$e_t = \frac{1}{u-l+1} \sum_{i=l}^{u} \left| \tilde{y}_i - \tilde{x}_i \right| . \quad (2)$$

As an alternative to (2), we can apply a measure-and-normalize step before the LSE calculation. This involves measuring the mean value of each of the two log-spectra and normalizing the output log-spectrum to force equality between the means of the input log-spectrum and the normalized output log-spectrum:

$$e_c = \frac{1}{u-l+1} \sum_{i=l}^{u} \tilde{y}_i - \frac{1}{u-l+1} \sum_{i=l}^{u} \tilde{x}_i , \quad \hat{y}_i = \tilde{y}_i - e_c . \quad (3)$$

Finally, an LSE is calculated as

$$e_f = \frac{1}{u-l+1} \sum_{i=l}^{u} \left| \hat{y}_i - \tilde{x}_i \right| . \quad (4)$$

One interpretation of this normalization step is that it performs a simple decomposition of total LSE (e_t) into coarse (e_c) and fine (e_f) components. The results in Section 4 are based on two-band implementations of (2) and (3). The low band extends from near DC to near 2 kHz (l=2, u=64) resulting in e_{tl}, and e_{cl}, while the high band extends from 2 kHz to near Nyquist (l=65, u=128) resulting in e_{th} ,and e_{ch} . For the normalization step in (3), $\tilde{y}_i$ is normalized by e_{cl} when 2 ≤ i ≤ 64, and $\tilde{y}_i$ is normalized by e_{ch} when 65 ≤ i ≤ 128.

3. FRAME-ENERGY PLANE PARTITIONING

In frame-based speech quality estimation algorithms, each pair of frames (input, output) can be classified according to the energy in those frames. We calculate frame energy in the frequency domain:

$$e_x = 10 \cdot \log_{10}\left(\sum_{i=2}^{128} x_i \right) \text{dB}, \quad e_y = 10 \cdot \log_{10}\left(\sum_{i=2}^{128} y_i \right) \text{dB}. \quad (5)$$

The most general classification scheme corresponds to an arbitrary partitioning of the (e_x , e_y) frame-energy plane. The classification of each pair of frames can be used to select or modify the estimation algorithm for that pair of frames. The MNB estimators effectively partition this plane into quadrants but use only frames in one of the four quadrants. The PSQM divides the plane into two halves, applies the same estimation algorithm to each half, and then weights the two sets of results differently.

A relatively simple partitioning is shown in Figure 1. Each pair of frames can then be classified by comparing e_x and e_y with the thresholds shown in Figure 1. (By the definition in (5), most of the active speech frames fall into the 10-55 dB range and about 40% of active speech frames fall above 40 dB.) The area of Figure 1 marked "Louder Frames" contains frames of active speech that are largely intact but contain some distortion. We apply the normalized LSE equations (3) and (4) to these frames only. We then average frame results over the entire speech signal (typically two sentences, 6 to 9 seconds total duration). For e_{cl} and e_{ch}, we average the absolute value to arrive at $\overline{e}_{cl}$ and $\overline{e}_{ch}$. The other averaged results are indicated by $\overline{e}_{tl}$, $\overline{e}_{th}$, and $\overline{e}_{f}$.

The area marked "Muted Frames" corresponds to active input speech that was partially or completely muted by the system under test. Potential causes include groups of bits that were lost, erased, or seriously delayed in transmission, and imperfect voice activity detection. The area marked "Noise Frames" corresponds to lower-energy frames of input speech that contain significant energy at the output. Potential causes include noisy analog systems and undetected transmission errors. We apply the most trivial of all algorithms to frame pairs that fall into the "Muted" and "Noise" categories. We simply count the number of frame pairs in each category and normalize by the total number of frames in the input signal, resulting in a muted-frame ratio (r_m) and a noise-frame ratio (r_n). Frame pairs that fall in the area marked "Quieter Frames" are not used at all. As shown in Section 4, even this simple partition and these trivial ratios provide valuable information for speech quality estimation. More intricate partitions and more sophisticated algorithms are likely to provide additional valuable information.

4. EVALUATION AND DISCUSSION

We measure the performance of estimators by the correlation between their estimates and MOS scores from formal subjective tests. We first average all MOS values for a given condition to a single value. Likewise, we average the corresponding estimates to a single value. We then calculate Pearson correlations between these two sets of per-condition means, yielding per-condition correlations. The results given here are drawn from the 17 formal subjective tests summarized in Table 1. Together, these tests contain over 12,000 speech files and cover 320 speech codecs, transmission systems and reference conditions. Bit rates for digital systems range from 2.4 to 64 kb/s. Both clear and errored channels are included. Three of the tests use flat speech and the remainder use IRS filtered speech. The test material includes over 24 hours of speech from about 80 different talkers using four different languages.

We use least-squares to form optimal linear combinations of the calculated quantities (parameters) described in this paper. The target vector contains over 12,000 MOS values so these least-squares problems are highly over-determined, even when 5 parameters are used. Since MOS values come from a finite range, we use a logistic function to map objective estimates into the finite range (0,1). Thus, when parameters p_1, p_2, …, p_M are used, the quality estimate is given by

$$\hat{q} = \left(1 + \exp\left(w_0 + \sum_{i=1}^{M} w_i \cdot p_i \right) \right)^{-1}. \quad (6)$$

Note that $0 < \hat{q} < 1$, and that $\hat{q}$ is positively correlated to MOS.

When $\overline{e}_{tl}$, and $\overline{e}_{th}$ are used together to estimate speech quality, correlations with the 17 sets of MOS results range from 0.36 to 0.95. When $\overline{e}_{cl}$, $\overline{e}_{ch}$, and $\overline{e}_{f}$ are used together, eleven of the correlations increase. Seven of those increases are greater than 0.10. (The largest decrease is 0.06.) The range of correlations is now 0.74 to 0.98. If we then include r_m and r_n to form a five-parameter estimator (LSE+), further increases are seen in eleven of the correlations. Seven of those increases are greater than 0.10. (The largest decrease is 0.03.) The range of correlations is now 0.84 to 0.98. These correlations are shown in Figure 2, along with those of the PSQM [2] and the MNB-2 [4] estimators. For tests 1-8, MNB-2 and PSQM perform similarly. The LSE+ performance is somewhat lower but still remarkably high considering the very low relative complexity of the algorithm. Tests 10-17 include the bulk of the bit-error, frame-erasure, and lower-rate codec conditions. Here MNB-2 and LSE+ show a distinct advantage over PSQM. The weights for LSE+ are given in Table 2.

We conclude that a two-band LSE-based estimator can benefit dramatically from a single measure-and-normalize step. This step separates the coarse and fine spectral errors and allows them to be treated separately in an estimate of speech quality. (In MNB estimators other scales between these two extremes are exploited as well.) Frame-energy plane partitioning offers additional benefits, and these benefits are likely to be more dramatic as we move beyond the very simple version presented here. The resulting estimator shows marked improvements over PSQM in some situations.

5. REFERENCES

[1] Beerends, J. and Stemmerdink J. “A perceptual speech-quality measure based on a psychoacoustic sound representation,” J. Audio Eng. Soc., 87:115-123, 1994.

[2] ITU-T Rec. P.861, “Objective quality measurement of telephone-band speech codecs,” Geneva, 1996.

[3] Voran, S. “Estimation of perceived speech quality using measuring normalizing blocks,” IEEE Speech Coding Workshop, pp. 83-84, Pocono Manor, PA, 1997.

[4] Voran, S. “Objective estimation of perceived speech quality, Part I: Development of the measuring normalizing block technique” and “Objective estimation of perceived speech quality, Part II: Evaluation of the measuring normalizing block technique,” IEEE Trans. on Speech and Audio Processing, July 1999.

[5] ITU-T Rec. P.861, Appendix II, “Objective quality measurement of telephone-band speech codecs using measuring normalizing blocks,” Geneva, 1998.

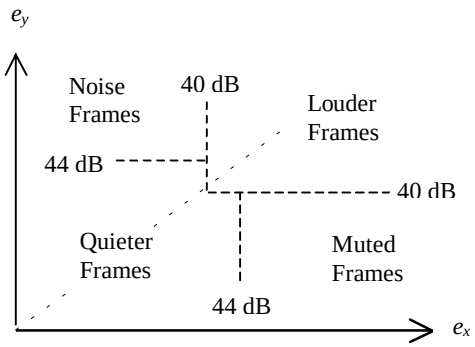

Figure 1. Example frame-energy plane partition.

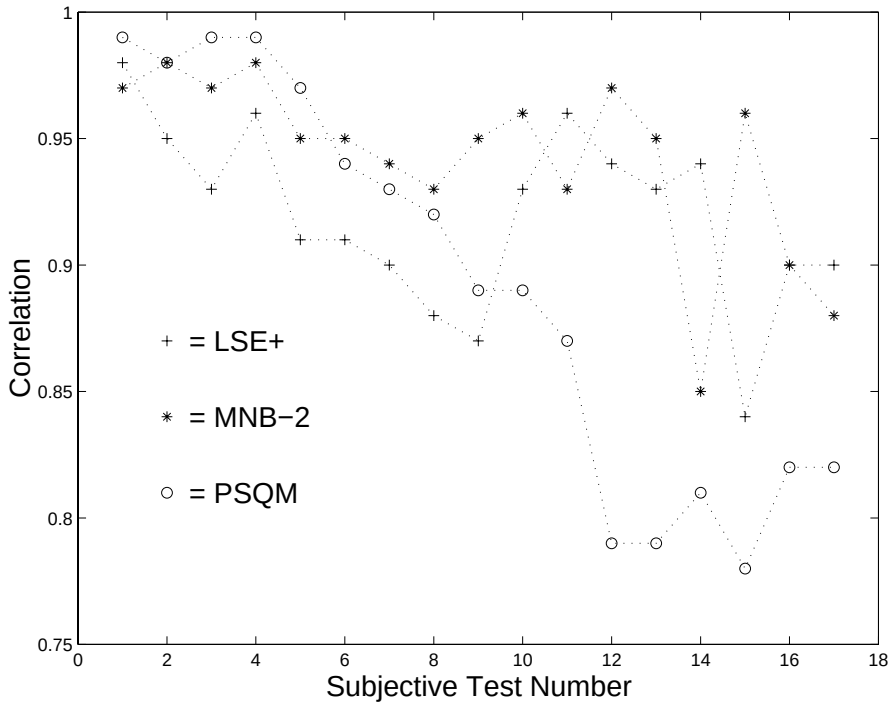

Figure 2. Correlations for 17 subjective tests.

Test Number	Language	Filtering	Number of Conditions	Summary of Conditions	Bit-Rates (kb/s)
1	English	IRS	20	LD-CELP, ADPCM, PCM, MNRU, (Tandems)	16-64
2	Italian	IRS	20	Same as 1	16-64
3	Japanese	IRS	20	Same as 1	16-64
4	English	Flat	19	IMBE, AMBE, VSELP, RPE-LTP, LD-CELP, ADPCM, MNRU, (Mixed Tandems)	6.4-32
5	English	IRS	19	Same as 4	6.4-32
6	English	IRS	35	Multi-Rate CELP, PCM, AMPS, MNRU, (Frame Erasures)	8-64
7	English	IRS	44	ACELP, VSELP, RPE-LTP, LD-CELP, ADPCM, PCM, MNRU, (Mixed Tandems)	8-64
8	Japanese	IRS	44	Same as 7	8-64
9	French	IRS	44	Same as 7	8-64
10	English	IRS	47	Multi-Rate CELP, LD-CELP, MNRU, (Frame Erasures)	13-16
11	Japanese	IRS	27	ACELP, ADPCM, MNRU, (Bit Errors, Frame Erasures)	8-32
12	English	IRS	27	Same as 11	8-32
13	French	IRS	27	Same as 11	8-32
14	Italian	IRS	27	Same as 11	8-32
15	English	Flat	27	LPC, STC, IMBE, CELP, VSELP, CVSD, ADPCM, POTS, MNRU, (Bit Errors)	2.4-32
16	English	Flat	41	CELP, ACELP, VSELP, PCM, AMPS, POTS, MNRU, (Bit Errors, Frame Erasures)	4.8-64
17	English	IRS	41	Same as 16	4.8-64

Table 1. Summary of material in 17 subjective tests.

i	p_i	w_i
0		- 4.29
1	$\overline{e}_{cl}$	0.150
2	$\overline{e}_{ch}$	0.0243
3	$\overline{e}_{f}$	0.449
4	r_m	48.6
5	r_n	52.7

Table 2. Parameters and weights used in LSE+.