COMPENSATING FOR GAIN IN OBJECTIVE QUALITY ESTIMATION ALGORITHMS

Stephen Voran

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado 80305, USA
svoran@its.bldrdoc.gov

ABSTRACT

When objectively estimating speech, audio, or video quality, it is often necessary to compensate for a system gain or to "gain match" two or more signals. One can take three views of a system, leading to three different definitions of gain, and three different gain compensation solutions: one that minimizes distortion, one that matches input-output power, and one that maximizes signal-to-distortion ratio. We derive these three solutions, describe the algebraic and geometric relationships between them, and provide a generalized result that subsumes all three. We provide examples showing that these three solutions do differ in practical quality estimation situations. We also report some of the gain compensation choices found in the quality estimation literature.

1. INTRODUCTION

There are numerous engineering situations where it is desirable to estimate and compensate for system gain or to "gain match" two or more signals. Example situations could be drawn from objective quality estimation, system modeling or identification, channel modeling or identification, coding, and other areas. One specific set of examples comes from the objective estimation of perceived speech, audio, or video quality [1]-[6]. In general, a system will distort the signal and will also apply some non-unity gain factor to the signal. Constant, non-unity system gain results in a shift of volume, contrast, brightness, or color balance. Such shifts are often not considered to be part of the distortion introduced by the system since users often routinely adjust volume, contrast, brightness, and color balance to suit individual preferences. Objective estimators of perceived speech, audio, and video quality should give results that agree with human perception. If system gain is not considered to be a distortion component by humans, then it should be likewise ignored by objective estimators. This means that the system gain must be compensated for so that the objective estimator can properly measure the actual distortion components. Fig. 1 describes the typical approach where the system output $\boldsymbol{y}$ is scaled by the reciprocal of the estimated system gain $\tilde{g}$ to create $\tilde{\boldsymbol{y}}$.

In this paper we identify three, distinct, mathematically-motivated solutions for compensating for system gain. We derive each solution, describe the algebraic and geometric relationships between them, and provide a generalized result that subsumes all three. Finally we offer example results taken from a digital speech codec and an analog video recorder/player.

Each of these three solutions for system gain compensation has been used before. Different authors have chosen different solutions but we have found very little discussion regarding the choices that have been made. Each of the solutions is very simple and when presented in isolation from the other two, each solution may initially appear to be the only logical choice, thus creating the appearance that no discussion of the solution is necessary. This paper shows that discussion may be warranted because each solution has a unique mathematical motivation, each solves a unique problem, and each yields a unique result.

2. THREE GAIN COMPENSATION SOLUTIONS

If the system gain g were a well-defined quantity, then for a given estimation criterion there would be a single best estimate of g. But g is not so easy to define, and thus multiple gain compensation techniques (corresponding to multiple definitions of g) exist. In general, we wish to treat the case where the distortion of the system S is not well-modeled by additive noise. Rather we assume that S induces some arbitrary distortion and gain g on signal vector $\boldsymbol{x}$ to create the output signal vector $\boldsymbol{y}$. Given a single input vector $\boldsymbol{x}$ and output vector $\boldsymbol{y}$ we must find a reasonable value of $\tilde{g}$ so that scaling $\boldsymbol{y}$ by $\tilde{g}^{-1}$ will give the compensated output signal vector $\tilde{\boldsymbol{y}}$. Without loss of generality, we assume that all signals have zero mean and non-zero magnitude, and that $\boldsymbol{x}$ and $\boldsymbol{y}$ are not orthogonal.

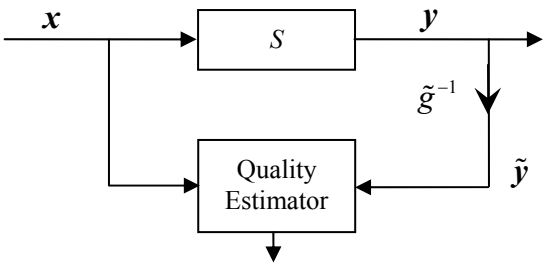

Fig 1. Typical block diagram for gain-compensated quality estimation of the system *S*.

It is certainly reasonable to seek a value of $\tilde{g}$ such that remaining system distortion is minimized:

$$\min_{\tilde{g}} \left|\boldsymbol{x}-\tilde{\boldsymbol{y}}\right|^2 \;\Rightarrow\; \min_{\tilde{g}} \left|\boldsymbol{x}-\tilde{g}^{-1}\boldsymbol{y}\right|^2 \;\Rightarrow\; \tilde{g}_{MD}=\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}, \qquad (1)$$

where we have used conventional least squares to solve for the minimum distortion gain estimate $\tilde{g}_{MD}$.

For many systems absolute distortion values are less relevant than signal-to-distortion ratios. The nature of human auditory and visual perception makes this especially true for audio and video signals. Thus it would also be reasonable to seek a value of $\tilde{g}$ such that remaining system signal-to-distortion ratio (SDR) is maximized:

$$\max_{\tilde{g}} 10\log_{10}\left(\frac{|\tilde{\boldsymbol{y}}|^2}{|\boldsymbol{x}-\tilde{\boldsymbol{y}}|^2}\right) \;\Rightarrow\; \max_{\tilde{g}} \frac{|\boldsymbol{y}|^2}{|\tilde{g}\boldsymbol{x}-\boldsymbol{y}|^2} \qquad (2)$$

$$\Rightarrow\; \min_{\tilde{g}} \left|\tilde{g}\boldsymbol{x}-\boldsymbol{y}\right|^2 \;\Rightarrow\; \tilde{g}_{MS}=\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}},$$

where we have used conventional least squares to solve for the maximum SDR gain estimate $\tilde{g}_{MS}$.

A third intuitive solution is the matched power solution. This solution forces $\boldsymbol{x}$ and $\tilde{\boldsymbol{y}}$ to have the same power and it is also the geometric mean of the two previous solutions:

$$\tilde{g}_{MP}=\sqrt{\frac{|\boldsymbol{y}|^2}{|\boldsymbol{x}|^2}}=\sqrt{\tilde{g}_{MD}\,\tilde{g}_{MS}}\;. \qquad (3)$$

In the decibel domain the geometric mean in (3) becomes an arithmetic mean:

$$\tilde{G}_{MP}=\frac{\tilde{G}_{MD}+\tilde{G}_{MS}}{2}, \quad \text{where } \tilde{G}_x=20\log_{10}\left(|\tilde{g}_x|\right). \qquad (4)$$

Note that $\tilde{g}_{MD}$ and $\tilde{g}_{MS}$ will correctly detect a negative gain value (indicating a phase inversion in *S*) but $\tilde{g}_{MP}$ as defined in (3) will not. Thus we use an intuitive extension to redefine $\tilde{g}_{MP}$ as

$$\tilde{g}_{MP}=\mathrm{sign}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y})\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}, \qquad (5)$$

so that all three solutions will have the same sign.

3. ALGEBRAIC AND GEOMETRIC OBSERVATIONS

If we define the normalized input-output cross correlation (or input-output direction cosine) ρ as

$$\rho=\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}, \qquad -1\le\rho\le 1, \qquad (6)$$

then we can summarize the three solutions as

$$\tilde{g}_{MD}=\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}\rho^{-1},$$

$$\tilde{g}_{MP}=\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}sign(\rho), \qquad (7)$$

$$\tilde{g}_{MS}=\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}\rho^{+1}.$$

Equation (7) makes it clear that all three solutions have the same sign and that their magnitudes can be ordered:

$$\left|\tilde{g}_{MS}\right|\le\left|\tilde{g}_{MP}\right|\le\left|\tilde{g}_{MD}\right|. \qquad (8)$$

In the limit as *S* becomes distortionless (but has a positive or negative non-unity gain), $|\rho|\to 1$ and (7) makes it clear that the three solutions converge to a single solution as expected.

Equation (7) also highlights that each of the three solutions is a special case of the more general solution

$$\tilde{g}(\alpha)=sign(\rho)\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}|\rho|^{\alpha}, \quad -1\le\alpha\le 1, \qquad (9)$$

where α = -1, 0, and +1 correspond to the minimum distortion, matched power, and maximum SDR solutions respectively. Fig. 2 provides an example of the geometric relationships among $\boldsymbol{x}$, $\boldsymbol{y}$, and the three possible gain-compensated outputs $\tilde{\boldsymbol{y}}_{MD}$, $\tilde{\boldsymbol{y}}_{MP}$, and $\tilde{\boldsymbol{y}}_{MS}$.

Note that these three solutions allow for arbitrary distortions in *S* but they also reproduce solutions that come from the minimization of additive input or output noises. The case of input noise leads to the minimum distortion solution. That is, if we solve for $\tilde{g}$ to satisfy $\boldsymbol{y}=\tilde{g}(\boldsymbol{x}+\boldsymbol{n}_{in})$ while minimizing $|\boldsymbol{n}_{in}|^2$, we will arrive at $\tilde{g}=\tilde{g}_{MD}$. The case of output noise leads to the maximum SDR solution. That is, if we solve for $\tilde{g}$ to satisfy $\boldsymbol{y}=\tilde{g}\boldsymbol{x}+\boldsymbol{n}_{out}$ while minimizing $|\boldsymbol{n}_{out}|^2$, we will

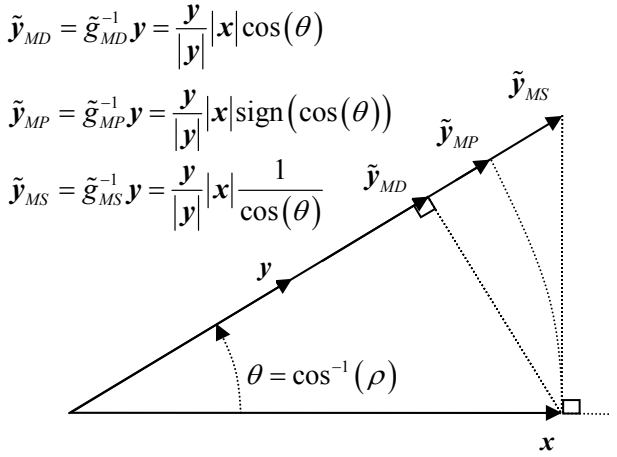

Fig. 2. Example geometry of the three gain-compensation solutions.

arrive at $\tilde{g} = \tilde{g}_{MS}$. The matched power solution $\tilde{g}_{MP}$ assumes that all output power is scaled input power and thus corresponds to the noise-free case $\boldsymbol{y} = \tilde{g}\boldsymbol{x}$.

4. EXAMPLE GAIN COMPENSATION RESULTS

In this section we present gain compensation examples for speech and video systems. In the speech example the input $\boldsymbol{x}$ is a ten-second speech signal and the system *S* is a 5.3 kbps speech codec conforming to ITU-T Recommendation G.723.1 [7]. In the video example the input $\boldsymbol{x}$ is one frame of a video chrominance signal and the system *S* consists of several play and record cycles of an analog video tape recorder/player. Fig. 3 shows the SDR and distortion as a function of $\tilde{G}$ for the speech example. The calculated values of $\tilde{G}_{MS}$, $\tilde{G}_{MP}$, and $\tilde{G}_{MD}$ are marked. As expected, $\tilde{G}_{MS}$ corresponds to the SDR maximum, $\tilde{G}_{MD}$ corresponds to the distortion minimum, and $\tilde{G}_{MP}$ is midway between $\tilde{G}_{MS}$ and $\tilde{G}_{MD}$. In the speech example, the values of the three gain compensation solutions span a range of about 0.8 dB, and the corresponding SDR and distortion values each span a range of about 0.4 dB. In the video example the curves have nearly identical shapes. The values of the three solutions span a range of about 0.2 dB, and the resulting SDR and distortion values span a range of about 0.1 dB. In general, these ranges will depend on both the input $\boldsymbol{x}$ and the system *S*.

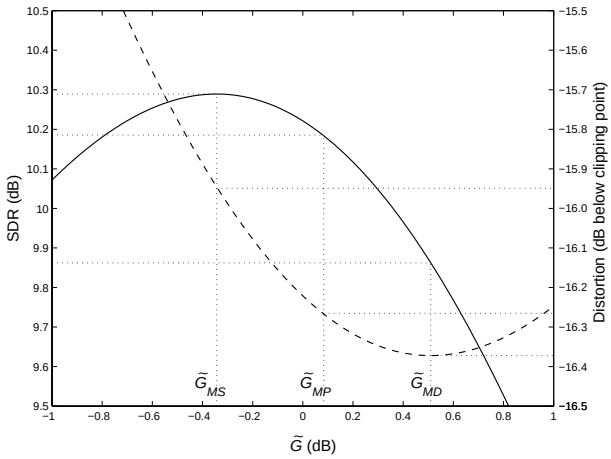

Fig. 3. Example gain-compensated SDR (solid line) and gain-compensated distortion (dashed line) vs. estimated gain ($\tilde{G}$) for G.723.1, 5.3 kbps speech codec.

5. DISCUSSION

We have established that there are multiple well-motivated solutions to the gain compensation problem (corresponding to multiple well-motivated definitions of system gain) described in Fig. 1, and that these solutions can differ significantly in real applications. One might ask which solution should be used in practice, but of course, there is no single answer. Each solution does exactly what its name says it does, so the question of which solution to use boils down to our view of the system *S*: Do we wish to view *S* as a system that minimizes distortion, matches power, or maximizes SDR?

We could also ask the question somewhat differently in terms of $\tilde{g}$. Loosely speaking, we could ask if $\tilde{g}^{-1}$ should describe how to scale the output to make it "as close as possible" to the input (leading to the minimum distortion solution), or if $\tilde{g}$ should describe the output-to-input power ratio (leading to the matched power solution), or if $\tilde{g}$ should describe how to scale the input to make it "as close as possible" to the output (leading to the maximum SDR solution)?

We could ask the question in a third way, again using informal language. Should $\tilde{g}^{-1}$ describe the fraction of the output that "matches" the input (leading to the minimum distortion solution where we project the input vector onto the output vector), or should $\tilde{g}$ describe the output-to-input power ratio (leading to the matched power solution), or should $\tilde{g}$ describe the fraction of the input that "matches" the output (leading to the maximum SDR

solution where we project the output vector onto the input vector)?

As noted above, the nature of human auditory and visual perception may make SDR more relevant than distortion in speech, audio, and video systems. This might lead one towards a maximum SDR solution for the gain compensation problem in the audio and video system context. We note further that "distortion" as we have used it here is waveform distortion. In low rate perception-based coding, waveforms may be severely distorted while the perceived signals are minimally distorted. For these systems the matched power solution may be more appropriate than the minimum waveform distortion solution or the maximum waveform SDR solution. It may be even more appropriate to compensate for system gain by the matching of some estimates of perceived loudness or contrast.

Next we report some of the gain compensation choices found in the objective speech, audio, and video quality estimation literature. The audio quality estimation algorithm in [4] includes a level adaptation stage that effectively calculates $0 < \tilde{g}_{MD}$ for frequency-domain excitation patterns. When $\tilde{g}_{MD} \le 1$ the input is scaled by $\tilde{g}_{MD}$, otherwise the output is scaled by $\tilde{g}_{MD}^{-1}$. The speech quality estimation algorithms given in [1] and [2] effectively perform matched power gain compensation, but they apply different scale factors to both the input and output signals to bring them to a common fixed level (e.g., unit variance). The speech quality estimation algorithm in [3] performs matched power gain compensation and applies a scale factor to the output only. The video quality estimation algorithm in [6] uses an iterative algorithm to find a robust estimate of $\tilde{g}_{MS}$. This algorithm uses weighted least-squares and places smaller weights on samples that have greater distortion so that they will not unduly influence the gain estimate. The resulting scale factor is applied to the output only.

Finally we mention system gain compensation via input scaling as shown in Fig. 4 as an alternative to the output scaling approach given in Fig. 1. If the input scaling approach is adopted, then the minimum distortion and maximum SDR solutions will be identical to each other and will be given by (2). The matched power solution will be unchanged from the solution above, and thus will be given by (5).

6. CONCLUSIONS

There are three mathematically-motivated solutions to the gain compensation problem described in Fig. 1, corresponding to three different definitions of system gain. Ideally, the solution selected in a given situation would reflect a conscious choice based on the system under consideration. Depending on the view of the system, one may choose to seek a gain compensation that minimizes distortion, matches power, or maximizes SDR. We have derived these three solutions, described the algebraic and geometric relationships between them, and provided a generalized result that subsumes all three. We have demonstrated that these solutions can differ significantly in real applications (e.g., 0.8 dB for the G.723.1 speech coder) and have reported some of the gain compensation choices found in the objective speech, audio, and video quality estimation literature.

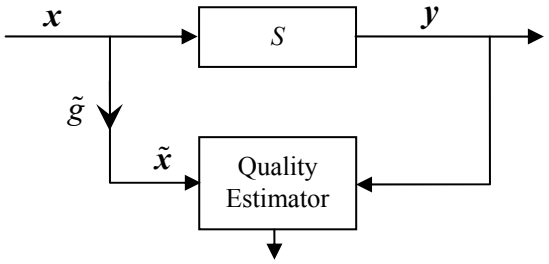

Fig. 4. Block diagram for gain-compensated quality estimation with input scaling.

7. REFERENCES

[1] J.G. Beerends, A.P. Hekstra, A.W. Rix, and M.P. Hollier, "Perceptual evaluation of speech quality (PESQ) – The new ITU standard for end-to-end speech quality assessment, Part II – Psychoacoustic model," Journal of the Audio Engineering Society, vol. 50, no. 10, Oct. 2002.

[2] S.D. Voran, "Objective estimation of perceived speech quality, part II: Evaluation of the measuring normalizing block technique," IEEE Transactions on Speech and Audio Processing, vol. 7, no. 4, pp. 383-390, Jul. 1999.

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

[4] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ—The ITU standard for objective measurement of perceived audio quality," Journal of the Audio Engineering Society, vol. 48, no. 1/2, Jan/Feb 2000.

[5] S. Wolf, "Measuring the end-to-end performance of digital video systems," IEEE Transactions on Broadcasting, vol. 43, no. 3, pp. 320-328, Sep. 1997.

[6] S. Wolf and M. Pinson, "Video quality measurement techniques," NTIA Report 02-392, National Telecommunications and Information Admin., June 2002.

[7] ITU-T Recommendation G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," Geneva, 1996.