EVALUATION OF OBJECTIVE MEASURES OF LOUDNESS

Gilbert A. Soulodre
Communications Research Centre, Ottawa, Ontario, Canada, K2H 8S2 gilbert.soulodre@crc.ca

1. INTRODUCTION

In many applications it is desirable to measure and control the subjective loudness of typical program material. Examples of this include television and broadcast applications where the nature and content of the audio material changes frequently. In these applications the audio content can continually switch between music and speech, or some combination of the two. The program material can also include sound effects and environmental sounds. These changes in the content of the program material can result in dramatic changes in subjective loudness. Moreover, various forms of dynamic range processing are frequently applied to the signals, which can have a significant effect on the perceived loudness of the signal.

There is currently an effort within broadcast standards organizations (ITU-R, NABA) to identify or develop an objective loudness measure (a loudness meter) that can be used by broadcasters to equalize the perceived loudness of their content. The ultimate goal is to have more consistent broadcast levels across program materials and broadcast stations. The matter is also of great significance to the music industry where dynamic range processing is commonly used to maximize the perceived loudness of a recording. In the present study the performance of various objective loudness measures is evaluated.

2. OVERVIEW OF TESTS

In the first part of the study a series of subjective tests (loudness-matching experiments) were conducted at 5 test sites around the world in order to create a database for evaluating the objective loudness measures. A total of 97 subjects listened to a broad range of typical program material and adjusted the level (in 0.25 dB steps) of each test item until its loudness matched that of a reference signal. The reference signal consisted of English female speech with no background sounds, and was reproduced at a level of 60 dBA.

The program material used in the tests was taken from actual television and radio broadcasts from various locations around the world. The 98 sequences included music, television and movie dramas, sporting events, news broadcasts, sound effects, and advertisements. Included in the sequences were speech segments in several languages.

The test setup employed a single loudspeaker placed directly in front of the listener. The subject could switch instantly between the reference signal and the test items while matching their levels.

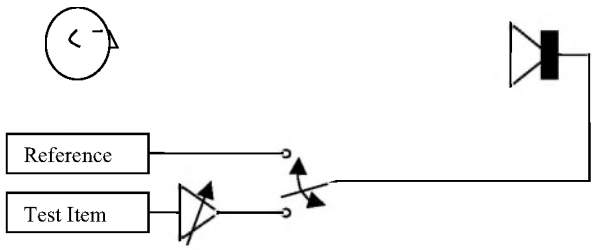

Figure 1: Subjective test setup.

Ten commercially developed loudness meters (labeled A-K in order to hide their identities) were submitted by proponents to be evaluated in their ability to predict the results of the subjective database. In addition, the author contributed two additional basic loudness measures to serve as a performance baseline. One was a simple *Leq* measure, while the other was a frequency-weighted *Leq* using a "*Revised Low-frequency B-weighting*" (referred to as *Leq*(RLB)) [1]. The individual audio sequences of the subjective database were processed through each of the loudness meters and the measured loudness estimates were recorded. These objective readings were then compared against the subjective loudness ratings using a variety of metrics to assess each meter's performance.

In order to assess the performance of the various loudness meters objectively, it was necessary to establish a set of suitable performance metrics that would effectively reflect the requirements of a practical loudness meter. In general, we want the meter to match the relative levels of the database as closely as possible. However, small errors in the meter's predictions are probably acceptable since listeners are unlikely to detect (or be annoyed by) small changes in loudness. Based on previous findings loudness errors of less than 1.25 dB are expected to go largely unnoticed [1]. Therefore, a meter could be considered to be ideal if all of its errors were less than 1.25 dB. Conversely, even a single error beyond some limit (say 10 dB?) could be considered entirely unacceptable, thus disqualifying a given meter from further consideration.

The metrics included, correlation (R), Spearman's rho, the root-mean squared error (RMSE), the maximum absolute

error (MAE), and the average absolute error (AAE).

3. RESULTS AND DISCUSSION

The results of the evaluation are summarized in Table 1 where the performance of the 12 meters is shown in terms of the various performance metrics. The numbers shown in square brackets indicate the relative ranking of the loudness meters for each metric.

Table 1: Performance of the loudness meters for the nine performance metrics. Values in brackets [ ] indicate the relative ranking for each metric.

	R	*Spearman rho*	*RMSE* (dB)	*MAE* (dB)	*AAE* (dB)
A	0.944[10]	0.916 [10]	2.37[11]	6.37[10]	1.88[11]
B [*Leq*(A)]	0.929[11]	0.889 [11]	2.19[10]	6.39[11]	1.77[10]
C	0.955 [9]	0.952 [8]	1.75 [9]	5.76 [9]	1.35 [9]
D	0.976 [3]	0.958 [5]	1.31 [3]	4.70 [5]	0.99 [3]
F	0.965 [8]	0.951 [9]	1.55 [8]	3.61 [1]	1.28 [8]
G [*Leq*(B)]	0.972 [5]	0.952 [7]	1.37 [6]	4.19 [4]	1.07 [5]
H	0.848[12]	0.841[12]	3.33[12]	6.89[12]	2.90[12]
I	0.972 [5]	0.960 [3]	1.36 [5]	4.80 [6]	1.09 [6]
J	0.968 [7]	0.955 [6]	1.51 [7]	4.97 [7]	1.17 [7]
K	0.975 [4]	0.958 [4]	1.33 [4]	5.13 [8]	0.99 [3]
Leq	0.979 [2]	0.971 [1]	1.26 [2]	4.33 [3]	0.93 [2]
Leq(RLB)	0.982 [1]	0.971 [1]	1.15 [1]	3.62 [2]	0.87 [1]

It can be seen that the basic loudness meter *Leq*(RLB) is ranked as the best meter for all of the metrics except the maximum absolute error (MAE). For this metric it is ranked second. However, it can be considered to be effectively equivalent to the first ranked meter for this measure, since its error is only 0.01dB larger. According to the various performance metrics the second best meter is a simple *Leq* measure. Therefore, for the present study, none of the commercially developed loudness meters submitted by the proponents performed as well as *Leq* or *Leq*(RLB).

This finding is quite remarkable given that most of the loudness meters included some form of complex perceptual model. The accuracy of the simple measures is quite impressive. Note that the worst case error for *Leq*(RLB), corresponding to the MAE metric, is only 3.62 dB. It was revealed that Meters B and G were *Leq*(A) and *Leq*(B) respectively.

It is of interest to examine plots of the worst and best performing loudness meters (Meter H and *Leq*(RLB) respectively). The data are plotted in terms of the gain that needs to be applied to a given audio signal in order to match its level to the reference signal. The open circles represent speech-based audio sequences, while the stars are non-speech-based sequences. It should be noted that a perfect objective meter would result in all data points falling on the diagonal line having a slope of 1 and passing through the origin (as shown in the figures). Any data point falling above the diagonal line indicates that the meter overestimated the gain required to match the loudness of that audio sequence to the reference signal. That is, the meter underestimated the perceived loudness of that particular audio sequence. The plots of Figure 2 and 3 clearly demonstrate the difference in performance of these two loudness meters.

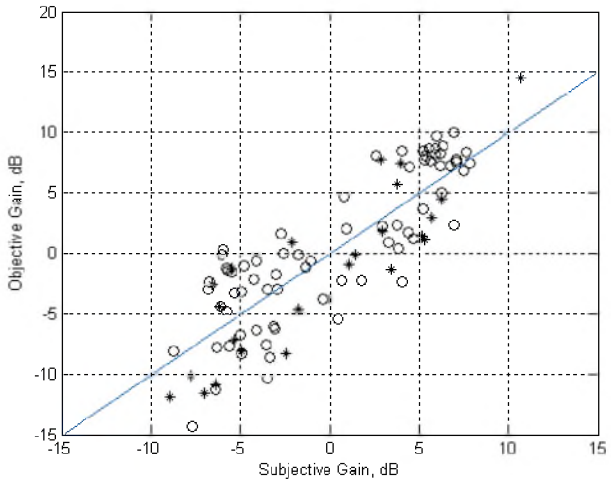

Figure 2: Meter H.

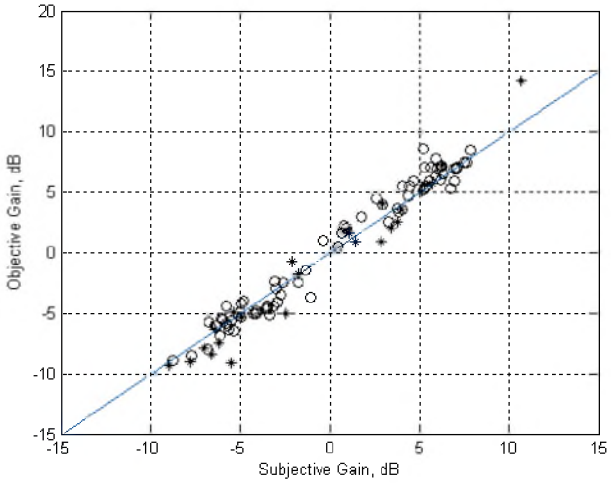

Figure 3: *Leq*(RLB).

The results of the present study indicate that for typical broadcast material, a simple energy-based loudness measure is more robust than more complex measures that may include detailed perceptual models. This finding is supported by the fact that one European broadcaster (TV2/Denmark) has been successfully using high-pass filtered RMS for many years as a measure of loudness.

REFERENCES

[1] Soulodre, G. A. (2004) Evaluation of Objective Loudness Meters, AES 116th Convention, Berlin (preprint 6161).