A Comparison of Some Loudness Measures for Loudspeaker Listening Tests*

RONALD M. AARTS, *AES Member*

Philips Research Laboratories, 5600 JA, Eindhoven, The Netherlands

Simple weighting methods and the ISO loudness models are compared with listener-adjusted loudness levels. For a loudness level of about 80 phons, B-weighting appeared to be the best method while A-weighting is unreliable.

0 INTRODUCTION

In listening tests, opinions that are formed about sound quality and stereo imaging are influenced by many factors in addition to the one that may be of specific interest; see Toole [1] for a brief overview. One of the sources of variability is loudness. The loudness balancing of loudspeakers during listening tests is considered to be very important. Among a variety of publications [1]–[8] it was recently noted by Gabrielsson et al. [8] that an increase in sound level will increase the perceived fullness, and spaciousness, and will give a better clarity and fidelity.

In a previous paper [9] the calculation of the loudness of loudspeakers was discussed. Some standardized loudness calculations were compared with the traditional method relying on A-weighted sound levels and with subjective loudness measurements obtained through listening tests. One of the conclusions of that paper was that the A-weighted sound-level method was not recommended for accurate loudness balancing and that the loudness differences between the loudspeakers are hardly influenced by the program choice.

The recommended measures (ISO 532) correlated well with the subjective ratings of the various subjects. However, many consider these methods to be cumbersome and too complicated for everyday use. It is the aim of the present paper to extend the comparison of loudness measures to include other simple ones, such as the B-, C-, and D-weighting functions. It should be noted that the A-, B-, C-, and D-weighting curves are only intended for rank ordering of noises according to loudness and not for measuring absolute loudness, while the ISO 532 methods are based on psychoacoustical data of human ears and can be used to measure absolute loudness. In the following sections it will be discussed why A-weighting is not recommended, and a better alternative will be examined.

1 EQUAL LOUDNESS

In the equal-loudness-level contours for pure tones, plotted in Fig. 1, two psychoacoustic phenomena may be observed: 1) the contours are heavily frequency de-

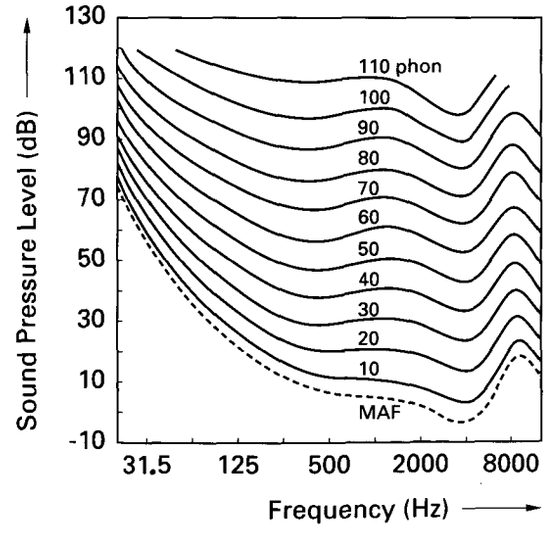


Fig. 1. Normal equal-loudness level contours for pure tones (binaural free-field listening, frontal incidence). From [12].

* Manuscript received 1991 June 17; revised 1991 November 15.

pendent and 2) the curves are level dependent. The latter is illustrated by Fig. 2. Fig. 2 shows that the normalized differences between the 80-phon curve and the 20-, 40-, 60-, and 100-phon curves are increasing for decreasing frequency below 200 Hz. The shapes of equal-loudness contours have been used in the design of sound-level meters, which attempt to give an approximate measure of the loudness of complex sounds. Such meters contain weighting networks so that the meter does not simply sum the power at all frequencies but, instead, weights the power at each frequency according to the shape of the equal-loudness contours. At low sound levels low-frequency components contribute little to the total loudness of a complex sound, so A-weighting is used to reduce the contribution of low frequencies to the final meter reading. At high levels all frequencies contribute more or less equally to the loudness sensation, so that a more nearly linear weighting characteristic, the C network, is used [10]. B-weighting is used for intermediate levels, while D-weighting is used for very high levels, such as aircraft noise. A-weighting, which is traditionally used for general purposes, is supposed to be an approximation of the 40-phon contour. This level is much too low for loudspeaker listening tests. When the 80-phon contour is used to obtain a weighting function by normalizing it to 0 dB at 1 kHz, the curve labeled 80-phon weighting will result, as shown in Fig. 3. As a reference, A-weighting and B-weighting are also plotted in Fig. 3. It appears, however, that at low frequencies A-weighting is too strong while B-weighting is a reasonable approximation of the 80-phon weighting curve. Another way to demonstrate the weakness of a simple weighting in general and A-weighting in particular is the following. When a subject listens to a pure tone at 200 Hz or 2 kHz, each with the same sound-pressure level of 60 dB, each tone will give about the same loudness (Fig. 4). The A-weighted value of the 200-Hz tone does not

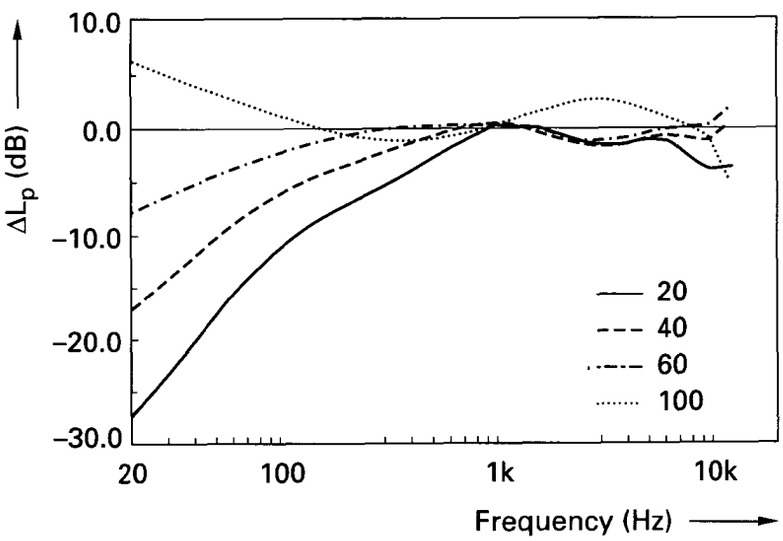


Fig. 2. Differences between 80-phon curve and 20-, 40-, 60-, and 100-phon curves, respectively. Difference curves have been normalized to 0 dB at 1 kHz.

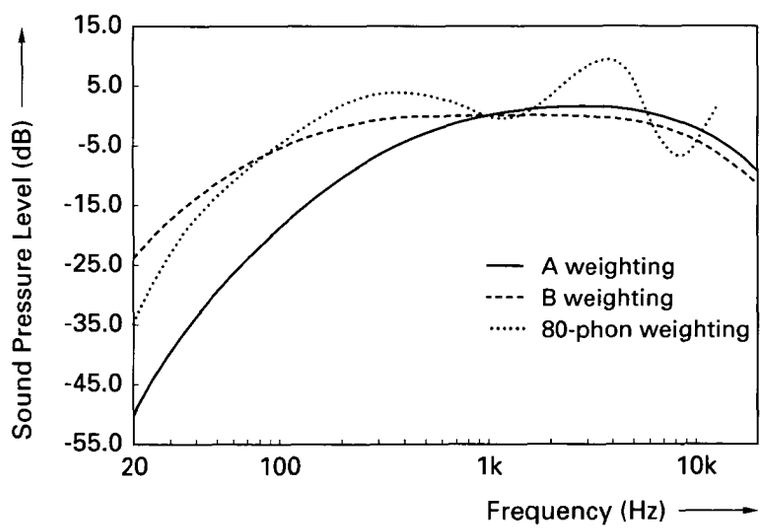


Fig. 3. A-, B-, and 80-phon (free-field) weighting functions.

reflect the perceived strength, however.

If the subject listens to the two tones simultaneously, with a frequency separation of more than a critical band, the perceived loudness will increase by about 10 phons (GD) with respect to a single tone. (The suffix GD stands for *group* and *diffuse* field; see [9] or [11].) The A-weighted level will remain the same as for the 2-kHz tone presented alone. This is due to the too rigorous weighting at low frequencies (at higher levels), and because the addition of signals has different effects in psychoacoustics than for electrical signals. The addition rules for loudness are incorporated in the more advanced loudness measurements, as discussed in [9]. However, if loudspeakers under test are similar, then there is no serious objection to a simpler weighting.

For reference purposes, the A-, B-, C-, and D-weighting functions are plotted together in Fig. 5. In the Appendix a computer procedure to compute these functions is presented. In the following sections these global comparisons will be tested against listener-adjusted loudness levels.

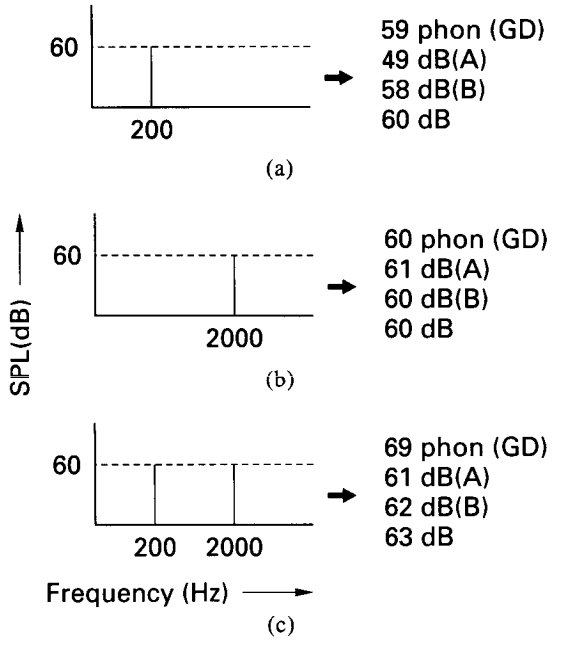


Fig. 4. Levels of tones. (a) At 200 Hz. (b) At 2 kHz. (c) For both tones simultaneously.

2 SUBJECTIVE LOUDNESS MEASUREMENTS

To test the usefulness of the objective loudness measures, these values will be compared with listener-adjusted loudness levels. These subjective values were obtained by an experiment that is summarized here; the details can be found in [9]. Ten subjects, one at a time, listened to six different loudspeakers LS_1–LS_6 (the same as those used in [9]), including the standard, or reference, LS_1, at a distance of 3.5 m. The loudspeakers were of different brands and covered wide ranges of price and quality. They exhibited very dissimilar frequency responses and different efficiencies, especially at low frequencies. The listening room was a soundproof room arranged and equipped as a normal living room. The loudspeakers could not be seen by the subjects, due to an acoustically transparent but visually opaque screen. They were connected to a switching facility which contained a set of high-quality relays, remotely controlled by the subjects. Variable attenuators were placed in the signal path from the CD player to the power amplifier. Each loudspeaker could be attenuated by the experimenter by adjusting the knob corresponding to the loudspeaker that was playing. The stimuli were presented by reproducing pink noise via the six different loudspeakers LS_1–LS_6. The subjects could compare loudspeakers LS_2–LS_6 to the reference loudspeaker LS_1 as often as they desired. The loudspeakers LS_2–LS_6 were to be matched by the subjects so that they perceived a loudness level equal to that of the standard. The subjects gave a signal to the experimenter to lower or raise the volume of the loudspeaker under test. When the subject was satisfied with all loudness levels (which took approximately 10 min),

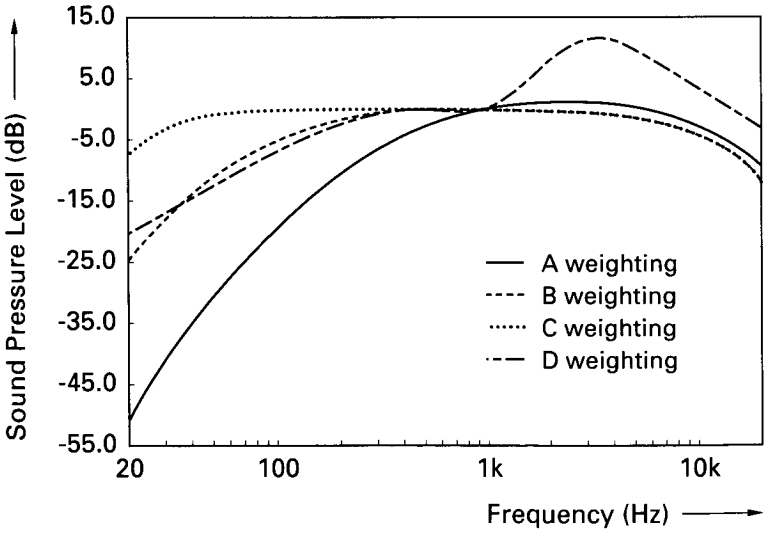


Fig. 5. A-, B-, C-, and D-weighting functions.

the level of each attenuator was stored. This was done once per subject. These values were averaged (over the 10 subjects) and are hereafter referred to as $L_{\text{sub}\,j}$. In a previous experiment [9] it was shown that the subjects could reproduce this task, even after a retention period of 15 min, with good accuracy. The reference loudspeaker was used as an anchor or standard. Its volume setting remained constant during all the tests, resulting in a loudness level of 80 phons (GD) for pink noise. See Table 1 for a comparison with other measures.

2.1 Results

To compare the results of the various methods for the loudspeakers tested the error was calculated as

$$\Delta_{\text{m}\,j} = (L_{\text{m}\,1} - L_{\text{m}\,j}) - L_{\text{sub}\,j} \quad (1)$$

where $L_{\text{m}\,1}$ is the sound level of the reference loudspeaker (LS_1) using method m, $L_{\text{m}\,j}$ is the sound level of the jth loudspeaker using method m, and $L_{\text{sub}\,j}$ is the averaged relative level adjusted by the subjects for the jth loudspeaker. The results of the listening test, using Eq. (1), are summarized in Table 2. (The first column is the unweighted sound-pressure level.) The entry HT^2 is Hotelling's T^2 [15], given as

$$T^2_{\text{m}} = \delta^T_{\text{m}}\,\text{cov}^{-1}\,\delta_{\text{m}} \quad (2)$$

where δ_{m} is the vector of differences of method m (the columns of Table 2) and cov is the covariance matrix of the subjects' ratings. A large value of T^2 indicates a large deviation from the subjects' ratings. Clearly, Table 2 shows that D-weighting is not applicable. The second worst method is the A-weighted sound level. The simple B-weighting is surprisingly the best in this test. The statistical significance of the data presented in Table 2 can be tested against the following zero hypothesis: "The differences between the various methods (unweighted; A-, B-, C-, and D-weighted; ISO-A and ISO-B) and the subjective ratings are due to random variations only." Using the T^2 values from Table 2, one cannot reject the zero hypothesis for B-weighting and ISO 532B. The hypothesis is rejected for the other five methods. The entry α in Table 2 denotes the level of significance of the T^2 test, which is the probability of making the decision to reject the zero hypothesis when in fact it is true (type I error). The power of this test cannot be calculated explicitly. However, it can been shown that the power of the present test is much higher than the power of a one-dimensional test and is sufficient to reject some methods.

One may conclude that the B-weighting and ISO 532B methods provide results similar to those of the subjective assessments. The five other methods are not consistent with the subjective ratings. It should be noted that the ISO 532 method is intended for general absolute loudness measures applicable for various levels and sound sources, while in this present case only relative loudness measures of comparable sound sources are of interest.

3 CONCLUSIONS

Experimental evaluations were made of seven measurement techniques to identify those that would be useful for the adjustment of loudness levels of loudspeakers for listening tests, at a level of 80 phons. The most satisfactory results were obtained by the use of a B-weighted measure of sound level. This provided results similar to those derived from subjective adjustments by a population of 10 subjects. The elaborate ISO 532B method also gave good results. The A-weighted measure yielded poor results and therefore is not recommended for accurate loudness balancing.

4 REFERENCES

[1] F. E. Toole, "Subjective Evaluation: Identifying and Controlling the Variables," in *Proc. AES 8th Int. Conf.* (Washington, DC, 1990 May 3–6, pp. 95–100.

Table 1. Comparison of some loudness measures for pink-noise source with SPL of 48 dB in each one-third octave in the range of 20 Hz to 20 kHz.

SPL	Sound level				ISO 532A	ISO 532B
(dB)	(dBA)	(dBB)	(dBC)	(dBD)	[phons (OD)]	[phons (GD)]
62.91	59.93	60.31	61.69	67.04	75.71	80.33

Table 2. Difference between objective and subjective measurements.

	SPL	Sound level				ISO 532A	ISO 532B
	(dB)	(dBA)	(dBB)	(dBC)	(dBD)	[dB (OD)]	[dB (GD)]
LS_2	0.530	−1.140	−0.300	0.490	−1.690	−0.050	−0.460
LS_3	1.235	−1.175	0.405	1.235	−1.845	−0.555	−0.355
LS_4	0.130	−1.360	0.080	0.290	−1.910	−1.130	−0.690
LS_5	1.135	−1.875	−0.185	1.165	−2.785	−1.295	−0.995
LS_6	−0.450	−1.530	−1.580	−0.670	−0.690	−0.030	−0.750
HT^2	13.62	48.22	4.16	14.58	74.35	18.90	6.54
α	$<10^{-5}$	$<10^{-15}$	~0.7	$<10^{-6}$	$<10^{-15}$	$<10^{-10}$	~0.1

[2] F. E. Toole, "Listening Tests—Turning Opinion into Fact," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 30, pp. 431–445 (1982 June).

[3] F. E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *J. Audio Eng. Soc.*, vol. 33, pp. 2–32 (1985 Jan./Feb.).

[4] F. E. Toole, "Loudspeaker Measurements and Their Relationship to Listener Preferences, Parts 1 and 2," *J. Audio Eng. Soc.*, vol. 34, pp. 227–235 (1986 Apr.); pp. 323–348 (1986 May).

[5] A. Illényi, and P. Korpássy, "Correlation between Loudness and Quality of Stereophonic Loudspeakers," *Acustica*, vol. 49, pp. 334–336 (1981 Dec.).

[6] Y. Tannaka and T. Koshikawa, "Correlations between Soundfield Characteristics and Subjective Ratings on Reproduced Music Quality," *J. Acoust. Soc. Am.*, vol. 86 (1989 Aug.).

[7] A. Gabrielsson and B. Lindström, "Perceived Sound Quality of High-Fidelity Loudspeakers," *J. Audio Eng. Soc.*, vol. 33, pp. 33–53 (1985 Jan./Feb.).

[8] A. Gabrielsson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived Sound Quality of Reproductions with Different Frequency Responses and Sound Levels," *J. Acoust. Soc. Am.*, vol. 88, pp. 1359–1366 (1990 Sept.).

[9] R. M. Aarts, "Calculation of the Loudness of Loudspeakers during Listening Tests," *J. Audio Eng. Soc.*, vol. 39, pp. 27–38 (1991 Jan./Feb.).

[10] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic Press, London, 1982).

[11] ISO 532-1975(E), "Acoustics—Method for Calculating Loudness Level," 1st ed. 1975, Int. Standards Organization (1977).

[12] ISO 226-1987(E), "Acoustics—Normal Equal-Loudness Level Contours," Int. Standards Organization (1987).

[13] IEC 651, "Sound Level Meters," Int. Electrotechnical Commission, Geneva, Switzerland (1979).

[14] IEC 537, "Frequency Weighting for Measurements of Aircraft Noise (D-Weighting)," Int. Electrotechnical Commission, Geneva, Switzerland (1976).

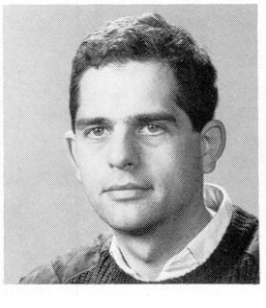
[15] B. J. Winer, *Statistical Principles in Experimental Design* (McGraw-Hill, New York, 1962).

APPENDIX
COMPUTATION OF A–D WEIGHTING FUNCTIONS

A computer procedure is listed in order to compute the A-, B-, C-, and D-weighting functions. The time constants for the filters for A-, B-, and C-weighting are from [13], those for D-weighting from [14].

```
PROCEDURE abcdweight(f: REAL; VAR aw, bw, cw, dw: REAL);
(* f = freq. (Hz) *)
CONST
  ca = 8.0002266419162E-01; cb = 9.8767069950664E-01;
  cc = 6.6709544848173E-09; cd = 6.8966888496476E-05;

  p1c = 20.6; p2c = 12200.0;

  p1a = 107.7; p2a = 737.9;

  p1b = 158.5;

  p1d  = 282.7; p2d  = 1160.0; p3d1 = 1712.0;
  p3d2 = 2628.0; z1d1 = 519.8; z1d2 = 876.2;

  ps1c = sqr(p1c); ps2c = sqr(p2c);
  ps1a = sqr(p1a); ps2a = sqr(p2a); ps1b = sqr(p1b);

  ps1d  = sqr(p1d ); ps2d  = sqr(p2d ); ps3d1 = sqr(p3d1);
  ps3d2 = sqr(p3d2); zs1d1 = sqr(z1d1); zs1d2 = sqr(z1d2);

VAR f2, hw: REAL;
BEGIN
  f2:= sqr(f);
  cw:= f2 / (f2 + ps1c) / (f2 + ps2c) / cc;
  aw:= cw * f2 / sqrt((f2 + ps1a) * (f2 + ps2a)) / ca ;
  bw:= cw * f / sqrt(f2 + ps1b) / cb ;
  hw:= 1 / (sqr(ps3d1 + ps3d2 - f2) + 4 * f2 * ps3d1);
  hw:= hw * (sqr(zs1d1 + zs1d2 - f2) + 4 * f2 * zs1d1);
  dw:= f * sqrt(hw/(f2 + ps1d) / (f2 + ps2d)) / cd;
END;
```

THE AUTHOR



Ronald M. Aarts was born in Amsterdam, The Netherlands, in 1956. He received a B.Sc. degree in electrical engineering in 1977, then joined the optics group of Philips Research Laboratories where he was engaged in research into servos and signal processing for use in both video long-play players and Compact Disc players. In 1984 he joined the acoustics group of the Philips Research Laboratories and was engaged in the development of CAD tools for loudspeaker systems.

Mr. Aarts has published a number of technical papers and reports and holds several patents in his field. He is a member of the AES and the ASA.