Audio Engineering Society

Convention Paper 8612

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

HELM: High Efficiency Loudness Model for Broadcast Content

Alessandro Travaglini[1], Andrea Alemanno[2], and Aurelio Uncini[3]

[1] Fox International Channels Italy, Rome, I-00138, Italy
info@alessandrotravaglini.it

[2] Department of Information, Electronic and Telecommunication (DIET) - University of Rome "La Sapienza", I-00184 Rome – Italy
andrea.alemanno@live.it

[3] Department of Information, Electronic and Telecommunication (DIET) - University of Rome "La Sapienza", I-00184 Rome – Italy
aurelio.uncini@diet.uniroma1.it

ABSTRACT

In this paper, we propose a new algorithm for measuring the loudness levels of broadcast content. It is called the High Efficiency Loudness Model (HELM) and it aims to provide robust measurement of programs of any genre, style and format, including stereo and multichannel audio 5.1 surround sound. HELM was designed taking into account the typical conditions of the home listening environment and it is therefore particularly good at meeting the needs of broadcast content users. While providing a very efficient assessment of typical generic programs, it also successfully approaches some issues that arise when assessing unusual content such as programs heavily based on bass frequencies, wide loudness range programs and multi-channel programs as opposed to stereo ones. This paper details the structure of HELM, including its channel-specific frequency weighting and recursive gating implementation. Finally, we present the results of a mean opinion score (MOS) subjective test that demonstrates the effectiveness of the proposed method.

1. INTRODUCTION

Over the last few decades, the international scientific communities involved in professional audio and broadcasting have been conducting in-depth research into the assessment of the equivalent loudness levels of programs. Inconsistent levels can be deeply annoying to viewers; therefore this issue was, and still is, considered a very critical technical aspect to deal with when managing the large variety of program genres typically handled by broadcasters nowadays.

This research has aimed to define technical solutions capable of normalizing all programs regardless their genre, mixing style, audio characteristics or format, to a specific yet unique target level in order to provide the audience with a consistent perceived loudness experience. Recently, some algorithms have rapidly received international consensus among the broadcast community (especially ITU-R. BS.1770-2) [1] and have largely proved to be capable of properly assessing program loudness levels under laboratory testing and for the largest majority of content. In particular, BS1770-2 is resulting very effective and popular as it is used as loudness model in many technical documents implemented worldwide such as ITU-R.BS1864 [2], ATSC-A/85 [3] and EBU-R128 [4].

However, at the time of writing it seems still needed to gather more data that can confirm that its goal is fully achieved for all kind of programs under typical home listening conditions. In particular, it seems appropriate to verify the evidence that it is possible to achieve uniform loudness normalization of all kinds of audio mixes and formats, particularly for unusual content such as programs heavily based on bass frequency range, wide loudness range programs, and multi-channel programs as opposed to stereo ones.

The research we present here was born with all these aspects in mind, as well as to verify the performance of ITU-R.BS1770-2 in real typical, specific and unusual TV experience. The result is the design of HELM, a sophisticated loudness model designed to assess the loudness levels of programs of different genres, styles and formats in broadcasting, including stereo and 5.1 surround sound. It originates from the need to verify the BS1770-2 algorithm and to investigate some issues that have been raised by several engineers who have independently spotted some slight yet important lack of robustness in this method, and it only aims to provide the broadcast community with more test data and eventually some possible improvements to the current standard.

The reported misreadings consist in the not always well correlated loudness measurement of specific content such as:

- Very short programs consisting of large parts of background sounds and a small percentage of foreground sound which is then broadcast very loudly (e.g. very dynamic advertisements)
- Content with a heavy bass frequency spectrum
- Multichannel audio 5.1 surround sound program loudness levels not matching with the corresponding downmixed stereo versions

Once we began to work on the subject and started to spot the aspects that appeared to lower the performance of ITU-R.BS1770-2 for specific unusual content, we began designing the amendments that seemed to improve the robustness and the correlation of the algorithm. As our research continued and new findings came to light, the sheer number of amendments led us to create what was essentially a brand new loudness model, sufficiently divergent from the original as to merit its own name.

2. LOUDNESS IN BROADCASTING

In order to properly predict and emulate human loudness perception, it is vital to reproduce as closely as possible not only the biological behavior of the hearing system but also the whole listening environment for which the algorithm is designed, including the reproduction formats, the TV set or loudspeakers set-up, and the playback SPL levels. HELM was designed taking into account all these aspects.

In broadcasting, typical audio formats include 2-track and multichannel audio 5.1 surround sound. Two-track programs (either stereo or dual-mono) consists of two audio channels (left and right) that are reproduced directly via stereo apparels (TV sets, radios, or home-theater) or, more rarely, via mono equipment (obsolete TV or radio sets) by the summation of the two. Two-track programs are usually reproduced "as is" and do not require decoding or downmixing: two channels in – two loudspeakers out.

In the last decade, with the diffusion of HD technologies and TV channels, the distribution of multichannel 5.1 audio (aka 5.1 surround sound) services has increased significantly. This has also sprung from the current universal availability of cinematic content originally produced in that format and subsequently available for home entertainment.

The 5.1 surround sound format consists of six discrete audio channels and can be reproduced in the following ways:

- through a multichannel loudspeaker system (Home Theater) consisting of six independent loudspeakers, each one dedicated to reproducing just one specific audio channel, representing the corresponding content contained in the original 5.1 program. Placement and alignment of the six loudspeakers must comply with the recommendation ITU-RBS775-1 [5];

- when no surround system is available, all 5.1 content can be reproduced through any stereo or mono apparel (TV or radio set) via the downmixing of the original six audio tracks into two streams or one stream respectively. The typical downmixing coefficients implemented to merge the six tracks into two are:

- Left = 0
- Right = 0
- Centre = –3
- LFE = not included / +10
- Left Surround = –6 / –3
- Right Surround = –6 / –3

In order to base the development of HELM on the real listening conditions typically present at home, we measured the frequency responses of several commercial apparels consisting of TV and Home-Theater sets. The results were averaged, producing the following findings. The typical frequency response of TV sets shows a decreasing linearity below 200Hz and a particular poor bass response below 80 Hz, as shown in Figure 1.

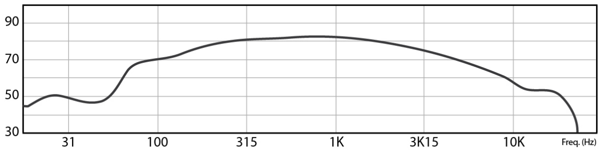


Figure 1 “TV set Frequency Response”

The frequency response is more even for Home-Theater 5.1 sets, also because of the implementation of the Bass Management feature which optionally routes the bass component of each of the 5.0 channels to the subwoofer. Consequently, the typical frequency response of Home-Theater sets is that shown in figure 2.

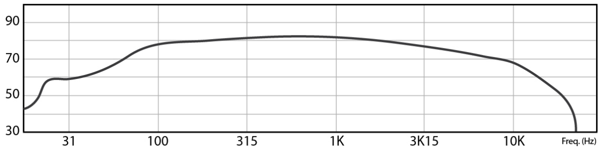


Figure 2 “Home-Theater set Frequency Response”

The frequency response above 10kHz drops in both sets, and in particular for the TV set. By analyzing the frequency responses of the figures 1 and 2, we conclude that the frequency weighting of the algorithm should take into account the average limitation in reproducing the low-end and the high-end that is typical of consumer audio sets.

In terms of the SPL level typically measured for home reproduction of broadcast content, scientific tests report that for stereo presentation through TV sets it averages around 65 dBSPL(A) whilst for Home-Theater 5.1 presentations the typical SPL level is approximately 70dBSPL(A).

3. ALGORITHM DESCRIPTION

In this section we describe the new algorithm HELM (High Efficiency Loudness Model) and how it was designed. In order to develop it, we analyzed all loudness characteristics of content in both their technical and scientific facets. Starting from the structure of ITU-R.BS1770-2 [1], we introduced several key enhancements based on solid foundations acknowledged by the scientific community, as we describe in the following paragraphs. The block diagram of the algorithm is as shown in Figure 3.

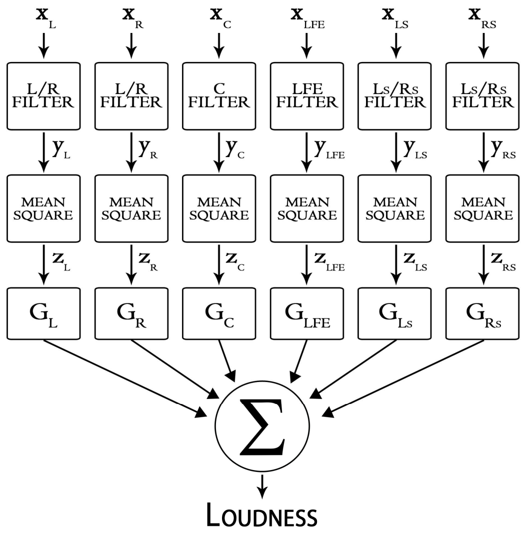


Figure 3 “HELM Block Diagram”

It works both in stereo and multi-channel audio 5.1 surround sound. As the figure clearly shows, the first difference between the HELM algorithm and ITU-R.BS1770-2 is the addition of the LFE channel. Since multi-channel audio content do have sounds reproduced by this channel, it seemed necessary to include it in the overall computation of the loudness levels in order to produce measurements well correlated with the real sound pressure occurring when 5.1 content are reproduced.

4. CHANNELS WEIGHTING

Unlike what is implemented in ITU-R.BS1770-2, in order to reproduce the human auditory system as closely as possible, we decided to optimize the frequency weighting for each type of audio channel included in the 5.1 format. We sought to maintain a high level of robustness without introducing excessive complexity.

As we can see from ITU-RBS775-1 [5], in a multichannel surround reproduction the sources of sound can be played from any of the following channels: Left, Right, Center, LFE, Left Surround, and Right Surround.

Depending on the channel they are played from, and thus the place and direction they occur in the space around the listener, the perceived intensity of sound frequencies changes according to several acoustic phenomena like masking and localization. These aspects, not tackled in ITU-R.BS1770-2 where all channels are equally weighted in terms of spectrum, play an important role in the HELM design. In fact, we worked on differentiating the frequency weighting for each of the 5.1 channel groups as specified below.

4.1. Center Channel

This channel is placed in front of the listener. We based the drawing of the weighting of this channel on the equal-loudness-level contours described in ISO 226-2003 (see Figure 4), since they have been obtained by placing one single loudspeaker right in front of the listener.

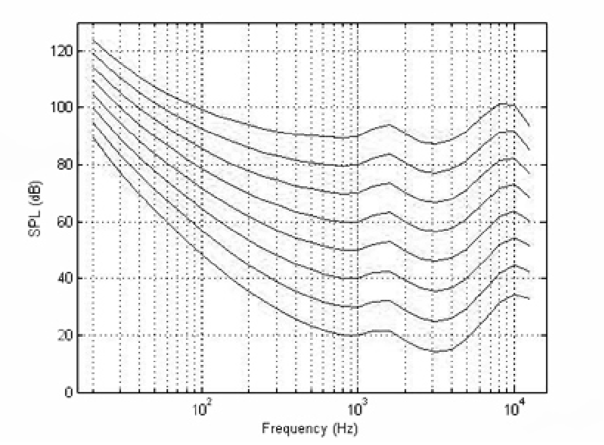


Figure 4 “ISO 226-2003 Standard”

Therefore, the frequency weighting of the Center Channel was based on inverting the effect of the ISO226-2003 curve, reported in Figure 4, measured at 65 phons. This level was chosen taking into consideration the typical Sound Pressure Level (SPL) of home entertainment as mentioned in paragraph 2. The Center Channel frequency weighting curve implemented in HELM is shown in Figure 14.

4.2. Left - Right Channels

The left and right channels in 5.1 surround format are located respectively on –30° and +30° in relation to the frontal axis of the sweet spot. Therefore, to draw the frequency weighting for these channels we referred to the study on spatiality by Blauert [6] and Moore [7]. In their researches, they found that for sounds coming from the same angle as left and right channels, the head effect is less important than for the other channels. At the same time the location cue effect led by the outer ear creates an emphasis effect on frequencies around 8kHz. To represent the decreasing hearing perceptions existing on the highest and lowest edges of the spectrum, the filtering design includes a 1st order High Pass Filter (*Fc=150Hz*) and a 2nd order Low Pass Filter (*Fc=13kHz*). The overall frequency weighting of Left and Right Channels is shown in figure 15.

4.3. Left Surround - Right Surround Channels

In 5.1 surround format, the loudspeakers for the Left Surround (Ls) and the Right Surround (Rs) channels are located respectively between –100° and –120°, and between +100° and 120° in relation to the frontal axis of the listener position. As per the filter for Left and Right channels, we based the design of the surround channels filtering on Blauert [6] and Moore's [7] researches in psychoacoustics.

Moreover, to improve the performance of these filters we followed Tomlinson's indications [8] on how to compensate for the difference between the perception of sounds when reproduced from the surround channels and the perception they would generate if reproduced from the front channels, and vice versa.

As for the other front channels, we implemented the same HPF and LPF to reflect the decreasing hearing perception at the edges of the spectrum.

The Left Surround and Right Surround Channels frequency weighting curve implemented in HELM is shown in Figure 16.

4.4. LFE Channel

The filtering for this channel has been drawn following the same psychoacoustic findings explained above, adapted on the basis of the technical conditions given by ITU-775-1 [5]. The LFE channel is assumed to be played from a subwoofer speaker placed in the frontal area of the loudspeaker set, in front of the listener, between the left and right speakers, possibly in the central zone.

The typical audio characteristics of subwoofers show a fairly linear (+/– 3dB) range for frequencies between 20Hz and 250Hz, above which the curve gradually descends with a 3rd order decay. We also took into account that "best practice" multichannel sound mixing that recommends to apply a 2nd order LPF at 120Hz on the LFE track.

Consequently, the range of frequencies reproduced by the LFE channel should never exceed 120Hz and in any case, because of the technical limitations of the media, they are never above 250Hz. Furthermore, the human hearing system, as discussed earlier, reports a decreasing low sensitivity below 150Hz.

Consequently, the only two filters implemented in the LFE Channel weighting are a 1st order HPF at 150Hz and a 2nd order LPF at 250Hz, as shown in Figure 17.

5. MEAN-SQUARE LOUDNESS ESTIMATION

Proceeding in a similar way to the BS1770-2, the signal is divided in 400ms long frames (aka gating block), using a rectangular running window with 75% overlap. According to previous discussion, let $y_i(t)$ the pre-filtered (by related weighting curve) signal sample the for the i^{th} channel the loudness z_i is defined as

$$z_i = \frac{1}{T}\int_0^T y_i^2(t)dt\,. \qquad (1\,)$$

From definition (1), the level is estimated by the mean-square over the j^{th} gating block of length T_g. Let t_s the running step and t_o the overlap coefficient, such that $t_s = 1 - t_o$, for the i^{th} input channel in the interval T, we can write

$$z_{ij} = \frac{1}{T}\int_{T_g\cdot j\cdot t_s}^{T_g\cdot(j\cdot t_s+1)} y_i^2(t)dt \qquad (2\,)$$

for $j = 0,\ 1,\ ...,\ \dfrac{T-T_g}{T_g\cdot t_s}$

The j^{th} gating block loudness is then defined as:

$$l_j = -1.979 + 10\log_{10}\sum_i G_i\cdot z_{ij} \qquad (3\,)$$

where the value -1.979 is intended to compensate for the total gain of the filters, giving a unified figure when measuring a stereo sine wave at 1kHz.

6. RECURSIVE GATING COMPUTATION

In order to eliminate the issue generated when measuring programs with very wide loudness range the recursive gating is implemented. In fact, it allows to measure Programme Loudness levels more precisely. A first definition of Recursive gating has been introduced in 2010 in the AES Paper "Determining an Optimal Gated Loudness Measurement for TV Sound Normalization" by Grimm et.al [9].

Recursive gating is particularly efficient for programs where the presence of "background sounds" parts is relevant, especially when background loudness levels are significantly lower than the Target Level. If no recursive gating is used in these cases, the "foreground sounds" (like dialogues) are reproduced at a much higher level than the average. This is because the threshold of the relative gating is set according to the first computation of the program's ungated measurement.

Consequently, if the loudness modulation of the program is wide, the ungated level is low and after the normalization of the whole program to Target Level the foreground sounds are set at too high a level.

This problem is shown in Figure 5. Let's consider a short interstitial program, like a 35-second promo, consisting of a first part of background sounds (ambience, a few subtle sound effects, very few musical instruments) lasting 30 seconds, followed by a voice announcing the promoted program (5 seconds of voice). The correct presentation of this content would have the voice being played at the average level (Target Level). If relative gating is used, due to the low ungated level that the 35 seconds content would have, the final voice message would be reproduced at a much higher level, generating annoyance and altering the original creative intent.

Figure 5 shown loudness curve of the content, with the 30 seconds of background sounds followed by the 5-second voice message. This curve is compared with a program with very little loudness modulation consisting of foreground sounds all the way through. For the latter, the foreground sounds would be reproduced at a consistent Target Level. The figure shows that by applying relative gating, the two foreground sound parts of the two pieces of content do not match.

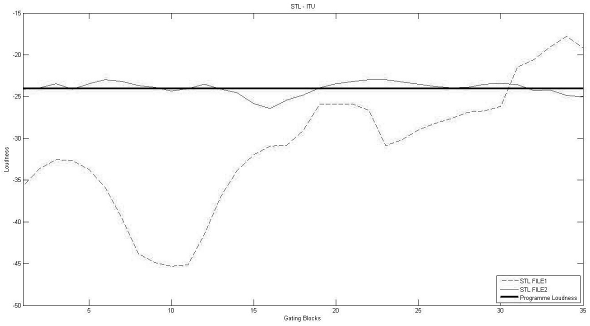


Figure 5 "Short-term Loudness curves of programs normalized according to BS.1770-2 "

By contrast, recursive gating means the computation of the ungated level is repeated in many cycles until the measurement is very accurate. In this way, the quantity and level of foreground sound parts do not influence the Programme Loudness measurement. Consequently, foreground sound parts are aligned to the correct Target level as shown in Figure 6.

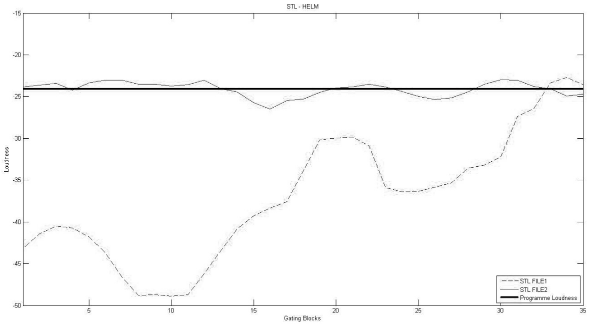


Figure 6 "Short-term Loudness curves of programs normalized according to HELM"

As you can see by applying the recursive gating, the foreground sound of the two tracks overlap and therefore they would result equally loud. On the contrary, if no recursive gating is applied (like in BS1770-2 and R128) the foreground sounds of File 1 would be reproduced several LU louder than the ordinary Target Level (indicated by the straight bold line at –24LUFS).

In order to make an increasingly more precise measurement of the program loudness, HELM includes an iterative process, starting with an absolute threshold (–70 LU) and then employing a relative threshold changing at every iteration. The block diagram of this important part of the algorithm is shown in Figure 7.

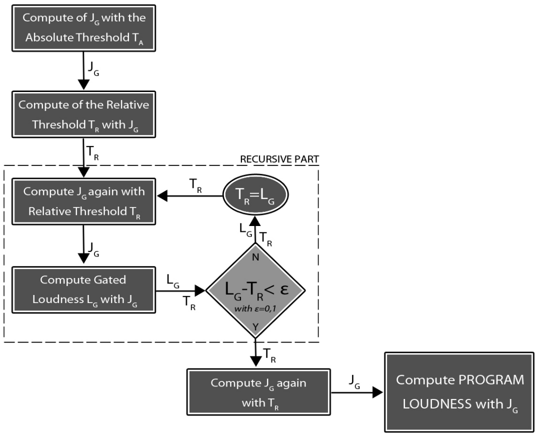

Compute of J_G with the Absolute Threshold T_A
J_G
Compute of the Relative Threshold T_R with J_G
T_R
RECURSIVE PART
Compute J_G again with Relative Threshold T_R
T_R
$T_R=L_G$
L_G
T_R
N
J_G
Compute Gated Loudness L_G with J_G
L_G
T_R
$L_G-T_R< \varepsilon$
with ε=0,1
Y
T_R
Compute J_G again with T_R
J_G
Compute PROGRAM LOUDNESS with J_G

Figure 7 “Scheme for Recursive Gating”

For a gating threshold Γ, there is a set of gating block indices J_g={j:l_j>Γ} where the gating block loudness is above the gating threshold. The number of elements in J_g is $|J_g|$.

The FIRST relative threshold Γ_r is calculated by measuring the loudness using the absolute threshold, $\Gamma_a = -70$ *LUFS* and subtracting *GT* from the result, thus:

$$T_r = -1.979 + 10\log_{10}\sum_i G_i\left(\frac{1}{\left|J_g\right|}\cdot\sum_{J_g} z_{ij}\right) - GT \qquad (4)$$

where J_g={j:l_j>Γ_a} with $T_a = -70$ LUFS and $GT = 7$.

The gating threshold is set at – 7. This value was found through the experiment described in the paragraph 9.1.

It is now possible to start the iterative process. Unlike the original BS1770-2, for the HELM algorithm we decided to use a simple convergence method: we minimized the error between the step n-1 and the step n, which allows us to we keep a constant distance between the gated loudness and the gating threshold during the calculation.

We recalculate the relative threshold Γ_r at every iteration, using this formula:

$$T_r(n) = -1.979 + 10\log_{10}\sum_i G_i\left(\frac{1}{\left|J_g\right|}\cdot\sum_{J_g} z_{ij}\right) - GT \qquad (5)$$

where J_g={j:l_j>$\Gamma_r(n-1)$} and $GT = 7$.

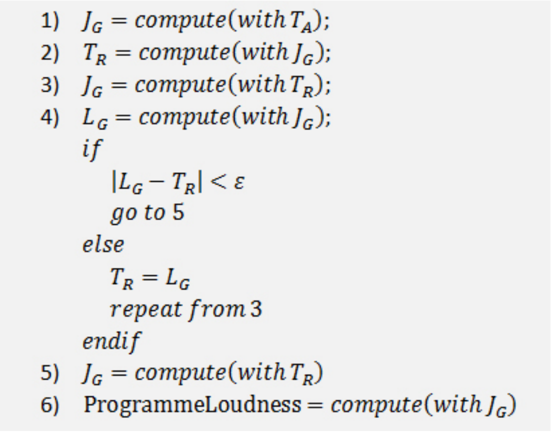
7. PROGRAMME LOUDNESS MEASUREMENT

The Programme Loudness (PL) is computed as

$$PL = -1.979 + 10\log_{10}\sum_i G_i\left(\frac{1}{\left|J_g\right|}\cdot\sum_{J_g} z_{ij}\right) \qquad (6)$$

where J_g={j:l_j>$\Gamma_r(n)$} with *n = last iteration number*.

The algorithm is described in Figure 8.



```
1) $J_G = compute(with\,T_A);$
2) $T_R = compute(with\,J_G);$
3) $J_G = compute(with\,T_R);$
4) $L_G = compute(with\,J_G);$
   $if$
      $|L_G - T_R| < \varepsilon$
      $go\ to\ 5$
   $else$
      $T_R = L_G$
      $repeat\ from\ 3$
   $endif$
5) $J_G = compute(with\,T_R)$
6) $\mathrm{ProgrammeLoudness} = compute(with\,J_G)$
```

Figure 8 “Meta-Language for Recursive Gating”

We also implemente different coefficients to weighting the channel levels. The new vector is: *G={1.0,1.0,1.0,10,1.0,1.0}* or also *{0,0,0,10,0,0}* in dB, where the order of the channel is intended to be *{L, R, C, LFE, Ls, Rs}*.

8. POSITIVE INTERVAL LOUDNESS LEVEL

Besides the main algorithm just described, HELM introduces one more – yet no less important – parameter, named Positive Interval Loudness Level (abbreviated as PILL). The purpose of this parameter is to estimate the consecutive variation of loudness of the program. This is to detect possible fast changes in the short-term loudness that may annoy the listener. It is focused on foreground sounds only as it measures the difference between any Short Loudness Level and the average of the 30 Short Loudness Levels just previously computed (covering a reference integration time of 10 seconds).

The process to compute this parameter is as easy as it is useful. The input to the algorithm is the Programme Loudness (previously calculated) and a vector of loudness levels, computed as specified in ITU-R.BS1770-2, using 3-second sliding blocks. An overlap between consecutive blocks is used to prevent a loss of precision in the measurement of short programs. A minimum overlap of 66% (i.e. a minimum 2-second overlap) between consecutive blocks is required; the exact amount of overlap is implementation-dependent.

The vector is normalized as follows. First of all, a threshold is defined as PILL Threshold = PL-GT (where PL = Program Loudness and GT = Gating Threshold, the same as for HELM). Then, the Short-Term blocks with values higher than the PILL Threshold maintain their values while Short-Term blocks that have a loudness value lower than the PILL Threshold change their values to the PILL Threshold value itself.

Next, the differences between the two figures thus generated are computed as follows:

Difference (n) = Loudness Value (n) – mean (Loudness Values from the n-10 Short-Term)

This descriptor could be used to spot fast changes of loudness levels during the reproduction of a piece of content. More importantly, it highlights the positive interval of a specific short sound event in comparison to an immediately previous part. Therefore, defining a MaxPILL Level could be very useful in assessing whether a sound element is potentially generating annoyance to the viewer as its value is continuously updated and synchronized with the event being played.

9. PERFORMANCE ANALYSIS: OBJECTIVE AND SUBJECTIVE TEST

We carried out many tests to ensure that the algorithm was robust enough to correlate very closely with human hearing, not only for generic content but also for unusual content such as programs with heavy bass frequencies, wide loudness material, multichannel audio, music and speech programs. The test consisted of two parts: objective tests and subjective tests (MOS).

9.1. OBJECTIVE TESTS

Before the subjective test, we performed many objective tests. These tests included many parameters of the algorithm in order to set them for the subjective test.

The main parameters evaluated in this part were the Gating Threshold, algorithm performance on MCA 5.1 Surround Sound vs. STEREO contents, algorithm performance on Music vs. Speech contents, and algorithm performance on Low Frequency (tracks with special contents on the low frequency range) vs. Average Spectrum contents.

For all these categories we used audio tracks gathered from the official EBU-PLOUD database. In order to assess the specific performance of HELM, we compared the results with ITU-RBS1770-2 [1].

9.1.1.Gating Threshold

The gating threshold is set to –7. This value was found through the following experiment.

We gathered 49 original TV program mixes, the same ones used to define the gating in the EBU-PLOUD tests, consisting of programs of different genres (including drama, feature film, music) and different formats (including stereo and 5.1) provided by several members of the group, and including:

- WLR (Wide Loudness Range): characterized by a large loudness range
- NLR (Narrow Loudness Range): characterized by a small loudness range
- MXD: characterized by both music and speech contents
- MUS: characterized just by music contents
- SP: characterized just by speech contents

Each program was labeled with the suffix FULL to indicate that they were presented in their whole original length. Each was accompanied by a very short excerpt, consisting in the foreground sounds as selected by the professional expert who provided PLOUD with the samples. Those parts were labeled with the suffix ANCHOR. As described in ATSC-A/85 [3], an "anchor element" is "the perceptual loudness reference point or element around which other elements are balanced in producing the final mix of the content, or that a reasonable viewer would focus on when setting the volume control". Speech is a typical foreground sound.

Since the ANCHOR parts represent the element used by viewers to set the volume control, the ideal gating method would be able to provide an integrated measurement of the FULL program as close as possible to the one focused on the ANCHOR part only. Indeed, our experiment consisted of comparing the integrated loudness measurements of the FULL programs with the integrated loudness measurements of the ANCHOR parts. The closer the two measurements, the more robust the gating method. Different thresholds were selected, starting from –12 up to –5 and the best one resulted in the –7 recursive.

9.1.2 Experts subjective alignment

To verify the performance of HELM in assessing the program loudness levels of specific content, we asked a team of 9 professional mixers to subjectively align the following tracks:

- Music vs. Speech
- Multichannel Audio vs. Stereo
- Low Frequency vs. Average Spectrum

A total of 36 tracks were used for this test. We took an average of the mixers' alignments and the resulting programs levels were measured using both HELM and ITU-R.BS1770-2.

9.2. SUBJECTIVE LISTENER TEST

Finally, an intense subjective test was carried out in order to evaluate the effective correlation and robustness of the new algorithm HELM. Results were also compared with ITU-R.BS1770-2.

To perform this test we used a small variation of the Mean Opinion Score (MOS) procedure. The MOS test has been used over the last few decades in telephone networks to obtain a "human view" of the network quality. In the field of multimedia (audio, voice, phone, video), especially when codecs are used to compress the bandwidth, MOS provides a numerical indication of the user's perceived quality of the downstream of a conversion. The MOS value is a single number between 1 and 5, where 1 indicates the lowest quality perceived and 5 the highest.

The MOS test for the voice was taken from ITU-T in the P-800 recommendation [11]. MOS is generated by averaging the results of a set of standard, subjective tests whereby a number of listeners rate the heard audio quality of test sentences read aloud by both male and female speakers over the communications medium being tested. A listener is required to give each sentence a rating using the scheme in Table 1.

MOS	QUALITY	IMPAIRMENT
5	Excellent	Imperceptible
4	Good	Slightly perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 1 "Rating Scheme for MOS Test"

The final MOS value is the arithmetical mean of all the individual scores and can range from 1 (worst) to 5 (best).

Our MOS version used to perform the subjective test differs from the original only in terms of the questions posed to the tester and the meaning attributed to his/her answers.

First, an introductory track was played to train the subject. This track was also used by each subject to set the volume level of the test in order to reproduce his/her typical conditions of home TV viewing. Then the tester heard one form at a time; each form contained a pair of stimuli, for a total of 28 forms (or 28 pairs of stimuli), of which 14 pairs were normalized with the HELM algorithm and 14 pairs were normalized with ITU-RBS1770-2.

In addition we also included 4 pairs of generic content only normalized by HELM. This was meant to verify the correlation of this algorithm in assessing ordinary programs loudness levels.

Each pair presented an “unusual” piece of content (an “unusual” content is characterized either by heavy bass frequency, or wide loudness range, or multichannel mix, or any combination of them, see par. Stimuli) and a generic ordinary program (ordinary narrow loudness range mix with music, sound effects and voice at consistent level),

For each form, we asked the subject this question (see Table 2):

“HOW DO YOU ASSESS THE VOLUME OF THESE TWO TRACKS?

Indicate your answer with a cross in the corresponding box. The boxes range from 1 to 5, where 1 shows the tracks were played at completely different volumes and 5 shows the tracks were played at exactly the same volume.”

1	The two tracks are at completely different volumes
2	The two tracks are at very different volumes
3	The two tracks are not at the same volume
4	The two tracks are at similar volumes
5	The two tracks are at exactly the same volume

Table 2 “Possible answers for the MOS test”

The final result was calculated using the MOS mode, taking the arithmetical mean.

9.2.1.Stimuli

Let’s analyze now the tracks chosen to conduct the tests. To select the tracks, we first took into account “what we had to verify”.

The choices fell into 5 main categories:

- Low Frequency: Tracks characterized by unusual contents at low frequencies.

- Gating: Tracks to verify the gating threshold and gating process.

- Music vs. Speech: Tracks to verify the correct measurement of musical and speech contents.

- MCA 5.1 Surround Sound vs. Stereo: Tracks reflecting the need to perform correct objective measurements of the correlation between different audio formats like Stereo and Surround 5.1.

- Generic: Finally, this category was used to test the HELM algorithm only, to verify its effectiveness on generic contents.

We chose 4 pairs of tracks per category, normalizing them with HELM and then copying the same 4 pairs normalized with the ITU-RBS1770.2 algorithm. The exception to this was for “Music vs. Speech” where there were just 2 pairs for HELM and 2 for ITU-R.BS1770-2. The tracks pairs were shuffled so that the order in which they were presented to the users was completely random. This test was performed in double blind mode: neither those giving the test nor those taking it knew the answers.

9.2.2.Subjects Statistics

The test was performed on 30 subjects aged between 18 and 69 years old, fairly divided between male and female (60% male and 40% female).

In addition to average users, some of those taking the test were people who usually work with music in the audio field such as musicians, dancers, choreographers, sound designers, etc.

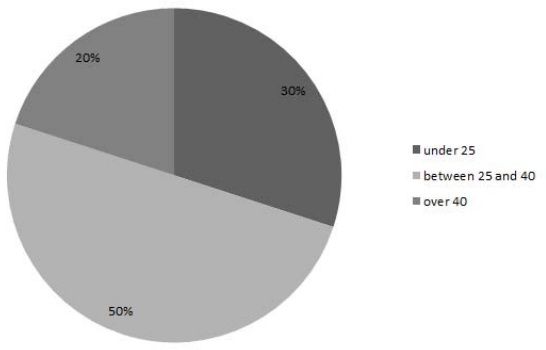


Figure 9 “Subjects age statistics of the MOS test”

9.2.3.Test

The test was performed in the Electric Light Studio in Rome from November 18–20, 2011. The mixing room used for the test was a proper sound proofing 5.1 surround sound studio and was ideal for reproducing the typical size of an average living room (6x5 meters). It was equipped with full range 5.1 professional loudspeakers and a reference near-field stereo pair aligned according to ITU-R.BS775-1 [5]. On average the tracks were played at was 65dBSPL(A) (or 70dBSPL(C)) with a background noise of 40dBSPL(A) (or 45dBSPL(C)) and a RT60 = 260ms. Every subject filled in a form in order to provide the statistics discussed above. Each subject performed the test independently using a PowerPoint presentation. Each subject followed these steps:

1) Introducing to the test

2) Completing form with background information

Age:	– –
Gender:	– –
Do you have "normal" hearing in both ears?	– –
Have you recently had a cold or flu?	– –
Have you had a hearing test in the last 5 years?	– –
If yes, were any significant problems detected?	– –
Are you often exposed to very loud music?	– –
If yes, please describe briefly in what situation and what kind of music: –	

Table 3 "Subjects' form of the MOS test"

3) Regulating the volume so as to simulate watching a TV program according to the subjective judgment of the subject.

4) Starting the test, allowing the user to play back each pair of tracks independently

The total time of the test varied from subject to subject but never exceeded 60 minutes. The test was performed individually; one person entered the room at a time, completed the test, then the next subject entered and so on. Only in two occasions subjects took part to the test in group of 3.

10. EVALUATION OF THE RESULTS

This section shows the results for all the tests we performed. For each test, we will analyze the results comparing the HELM algorithm with the ITU-R.BS1770-2 algorithm.

10.1. OBJECTIVE TESTS

As discussed above, we performed many computer simulations. Here, we will analyze the result using a gating threshold of –7 recursive for HELM, which was the highest performer of all our tests.

10.1.1. Gating Threshold

We evaluated the absolute difference between the FULL and the corresponding ANCHOR version for all the analyzed tracks. We obtained the statistics shown in Table 4 and Table 5 (ITU-R.BS1770-2 implements the official –10 relative gating threshold).

HELM		
	MEDIAN	MEAN
WLR	1.10	1.349
MUS	1.41	1.386
MXD	2.82	3.198
NLR	0.22	0.411
SP	0.81	1.191
TOT.	0.83	1.223

Table 4 "Results for Gating Threshold Test HELM"

ITU-R. BS. 1770-2		
	MEDIAN	MEAN
WLR	1.48	1.912
MUS	1.28	1.426
MXD	3.54	3.710
NLR	0.45	0.431
SP	1.41	1.364
TOT.	1.23	1.912

Table 5 "Results for Gating Threshold Test BS1770-2"

Graphically, this gives the results shown in Figure 10 where the good performance of HELM is confirmed by median and mean values lower than BS1770-2.

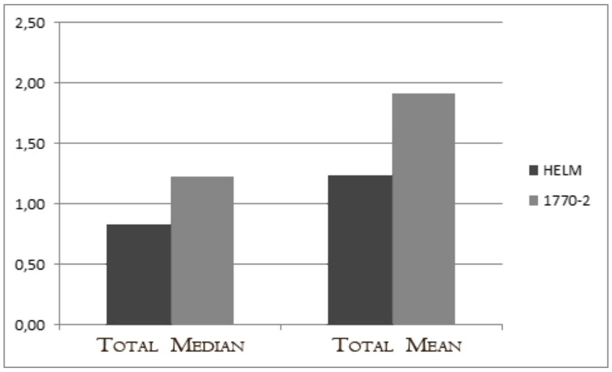


Figure 10 “Bar Graph for Gating Threshold Test”

10.1.2. Music vs. Speech

We used HELM and ITU-R.BS1770-2 to evaluate the Program Loudness of all 8 tracks aligned subjectively. To evaluate the results, we measured the standard deviation because it provides a good description of what we are looking for: that is, a perceptive alignment that provides minimally dispersed Program Loudness values. The results shows a Standard Deviation value of *1.590* for HELM and *1.888* for ITU-R S.1770-2.

This indicates that the HELM algorithm provides results that are closer each other than the BS1770-2, thereby validating the perceptive alignment. This result is further strengthened by the findings of the subjective experiment, as we will explore below.

10.1.3. Low Frequency vs. Average Spectrum

We measured the Program Loudness with both HELM and BS1770-2 for all the 22 tracks perceptively aligned as illustrated before. To evaluate these results, we also used standard deviation. The value for HELM is *1.480* while the value for ITU-RBS.1770-2 is *2.041*.

This is further confirmation that the HELM algorithm seems to provide better correlation of programs with heavy bass content than BS1770-2.

10.1.4. MCA 5.1 vs. Stereo

We used 7 MCA tracks and the 7 corresponding Stereo versions, generating the results shown in Figure 11 and Figure 12 in terms of absolute difference between the MCA 5.1 Surround Sound track and corresponding stereo downmix and in terms of mean and median of the absolute differences.

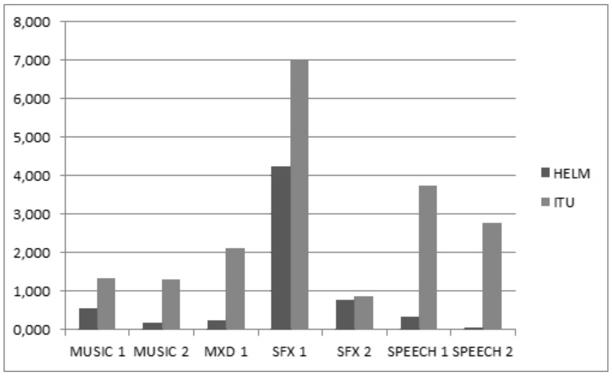


Figure 11 “Results for MCA 5.1 vs. STEREO test”

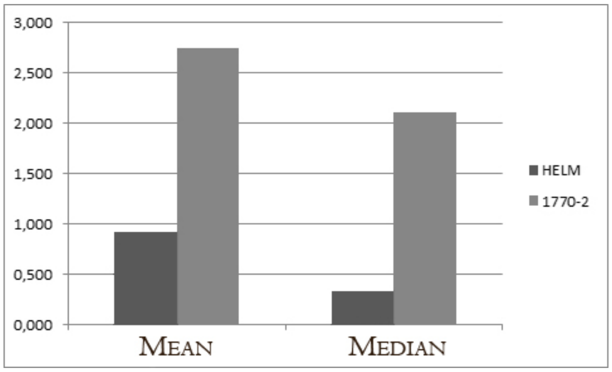


Figure 12 “Overall results for MCA vs. STEREO test”

The new HELM algorithm seems to outperform the ITU-R.BS1770-2 algorithm: Figure 12 clearly shows an effective improvement in the assessment of MCA 5.1 Surround Sound vs. STEREO content.

10.2. SUBJECTIVE TEST

The MOS final scores resulting from the subjective tests are shown in Table 6, Table 7, and Figure 13.

	HELM	BS1770-2
LOW FREQ.	3.808	3.342
MCA 5.1 vs. STEREO	3.992	3.317
MUSIC vs. SPEECH	3.433	2.983
GATING	3.667	2.508
GENERIC	3.967	Not assessed
TOTAL	3.811	3.045

Table 6 “Results for MOS Test”

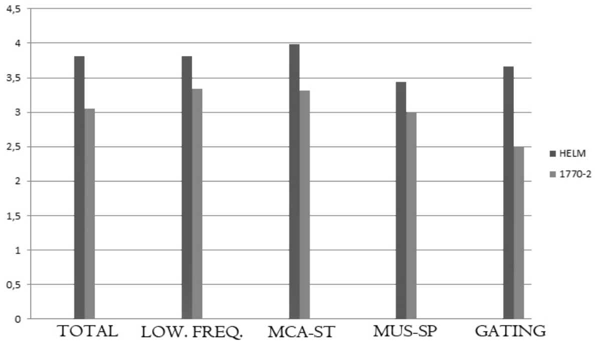


Figure 13 “Bar Graph for MOS Test Results”

Note that even if we omitted the generic results from the total mean calculation, the HELM MOS value would still be higher compared with ITU-R-BS.1770-2. Moreover, we can appreciate the performance of the two algorithms in Table 7, where the MOS scores gathered for each one in regard to the pairs’ assessments are compared.

Pair	MOS HELM	MOS BS1770-2
GATING 1	4.17	2.77
GATING 2	3.67	3.17
GATING 3	3.27	2.27
GATING 4	3.57	1.83
LOW FREQ. 1	3.67	3.23
LOW FREQ. 2	3.63	2.83
LOW FREQ. 3	3.60	2.97
LOW FREQ. 4	4.33	4.33
MCA 5.1 vs. ST 1	4.37	4.17
MCA 5.1 vs. ST 2	4.43	3.90
MCA 5.1 vs. ST 3	3.67	2.37
MCA 5.1 vs. ST 4	3.50	2.83
MUSIC vs. SPEECH 1	3.57	3.17
MUSIC vs. SPEECH 2	3.30	2.80

Table 7 “MOS Test Results – Algorithms comparison”

Even in this case, the HELM algorithm has a higher MOS value for every pair than the BS1770-2 algorithm, except in one pair where the value for the two algorithms is the same (LOW FREQ. 4).

11. CONCLUSIONS

HELM (High Efficiency Loudness Level), a new algorithm for measuring the loudness levels of broadcast content, has been designed to correlate well with human hearing, encompassing all usual kinds of broadcast programs, genres and formats.

To achieve this, we developed the algorithm according to scientific evidence on the spatialization of sound. The algorithm is specifically designed to properly represent the typical listening conditions that occur at the broadcast home presentation of both Stereo and Multichannel 5.1 surround sound content. It implements recursive gating with a -7 recursive threshold. Even so, the algorithm design has been optimized to avoid any redundant complexity.

A large number of tests were run in order to verify the design of HELM, including objective mathematical measurements and subjective MOS tests. The results were measured and compared with ITU-R.BS1770-2. All tests gave very encouraging results indicating a very high grade of correlation between the subjective perception of loudness and the loudness level provided by the objective measurement. The MOS test indicates an overall value of *3.811* representing a good correlation and a slightly perceptible but not annoying subjective perception of level differences across all content.

The same subjective tests performed with unusual programs aligned according to ITU-R.BS1770-2 gives a MOS value of *3.045* representing a fair correlation and a slightly annoying subjective perception of difference between various pieces of content.

In all cases HELM seemed to represent an improvement compared with ITU-R.BS1770-2, and in some specific correlation tests (such as in the gating test and the multichannel vs. stereo test) it outperformed the other algorithm, as shown in previous paragraphs. Furthermore, HELM seems to result very effective in assessing the loudness levels of program’s modulations, especially by implementing the descriptor PILL, as described in the AES Paper “Defining the Listening Comfort Zone in Broadcasting through the Analysis of the Maximum Loudness Levels", Travaglini et al. (2012) [10].

In conclusion, we believe that this research could offer a valid base upon which to begin new studies aimed at improving current standards in the loudness measurement of broadcast content and that the implementations included in HELM are capable of providing a very good correlation between objective and subjective loudness assessments, especially for unusual broadcast content. Tests have confirmed that it seems to competently assess the loudness levels of all kinds of programs, regardless of their genre, mixing style, audio spectrum or format.

Furthermore, we think HELM meets all technical requirements that current real-time and file-based meters present.

Moreover, we believe that HELM can be implemented successfully in any broadcast scenario and that it can coexist with ITU-R.BS1770-2 as it works equally well for normalizing generic typical content and it seems to represent a significant improvement in aligning unusual program material.

12. ACKNOWLEDGEMENTS

We would like to thank Fox International Channels Italy for supporting this research. We would also like to thank Zoë Hallington for her assistance in editing this paper and Stuart Mabey at Electric Light Studio in Rome for providing the technical facilities necessary to perform the subjective tests. Finally, we thank everyone who contributed to this research by taking part in the tests.

13. REFERENCES

[1] ITU-R.BS1770-2 "Algorithms to measure audio programme loudness and true-peak audio level" (2011)

[2] EBU Technical Recommendation R 128 "Loudness normalisation and permitted maximum level of audio signals" (2010)

[3] ATSC Recommended Practice – Document A/85 . "Techniques for Establishing and Maintaining Audio Loudness for Digital Television" (2011)

[4] ITU-RBS1864 "Operational practices for loudness in the international exchange of digital television programmes" (2010)

[5] ITU-RBS775-1 – "Multichannel Stereophonic System with and without accompanying picture SYSTEM" (1992-1994)

[6] JENS BLAUERT – "Spatial Hearing: The Psychophysics of Human Sound Localization" (1997)

[7] BRIAN C. J. MOORE – "An introduction to the Psychology of Hearing" - Fifth Edition (2004)

[8] TOMLINSON HOLMAN – "Surround Sound Up and Running" – Second Edition (2008)

[9] AES Convention Paper 8154 – Eelco Grimm, Esben Skovenborg and Gerhard Spikofski – "Determining an Optimal Gated Loudness Measurement for TV Sound Normalization" (2010)

[10] AES Convention Paper – Alessandro Travaglini, Andrea Alemanno, Fabrizio Lantini – "Defining the Listening Comfort Zone in Broadcasting through the Analysis of the Maximum Loudness Levels" (2012)

[11] ITU-T Recommendation P.800 – "Methods for subjective determination of transmission quality" (1996)

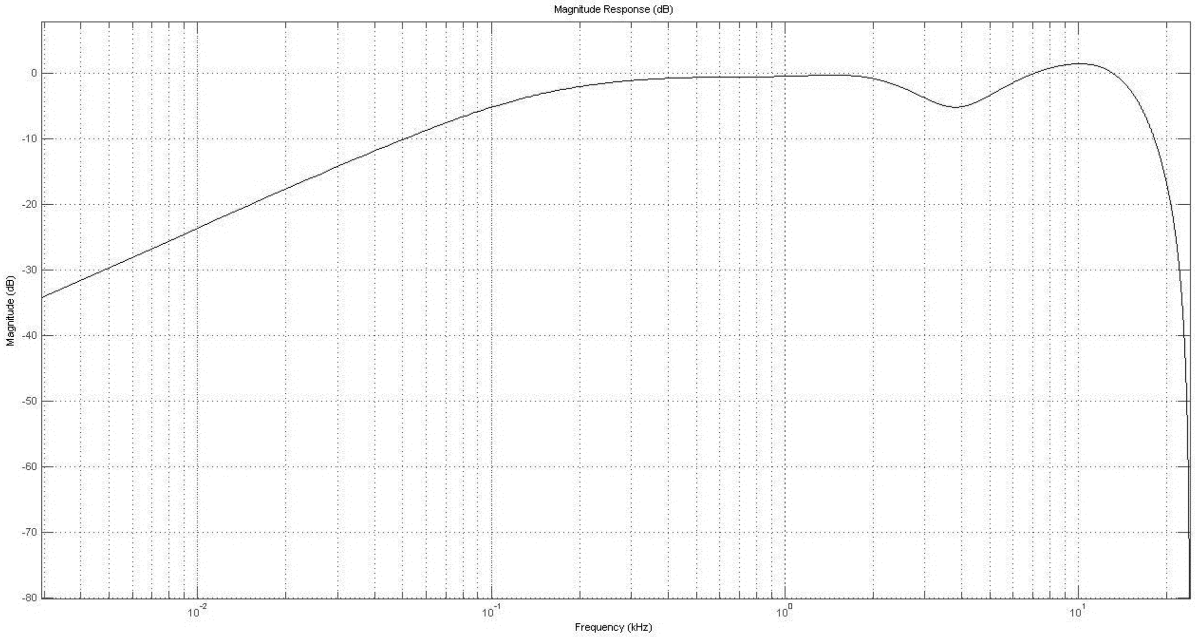


Figure 14 “Filter Response for Center Channel”

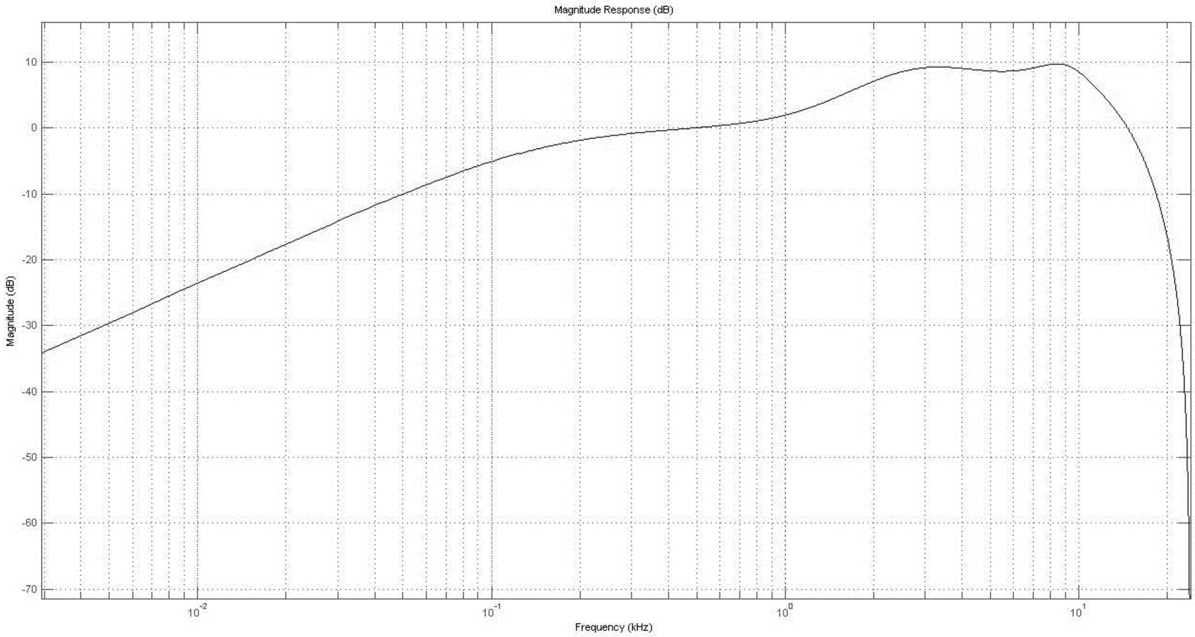


Figure 15 “Filter Response for Left and Right Channels”

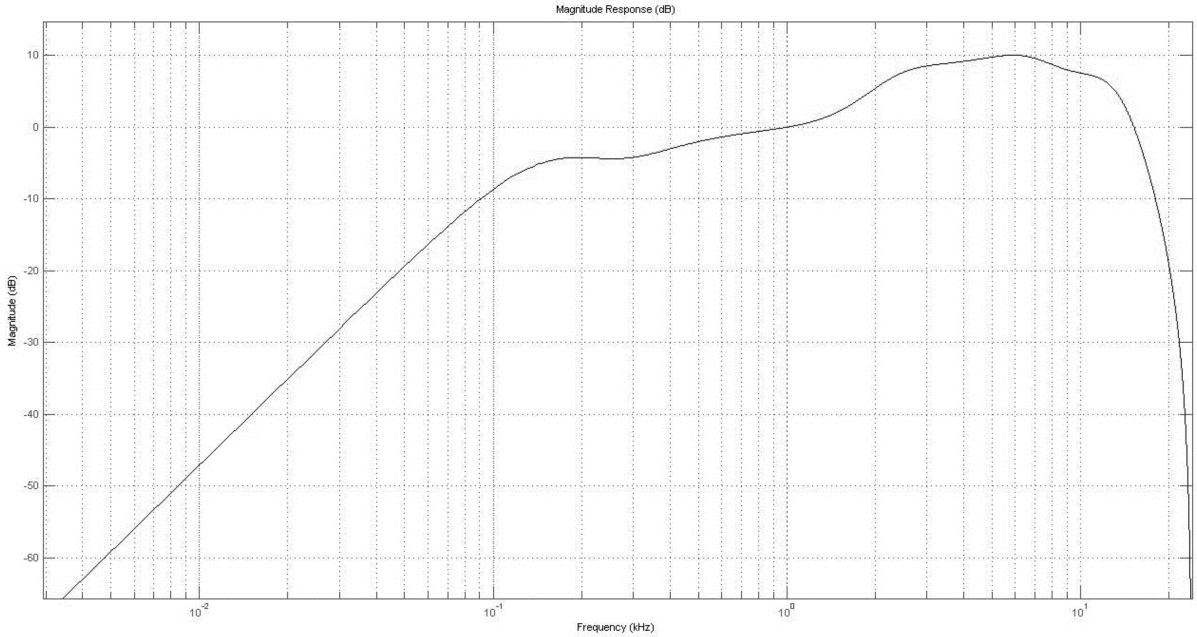


Figure 16 “Filter Response for Left Surround and Right Surround Channels”

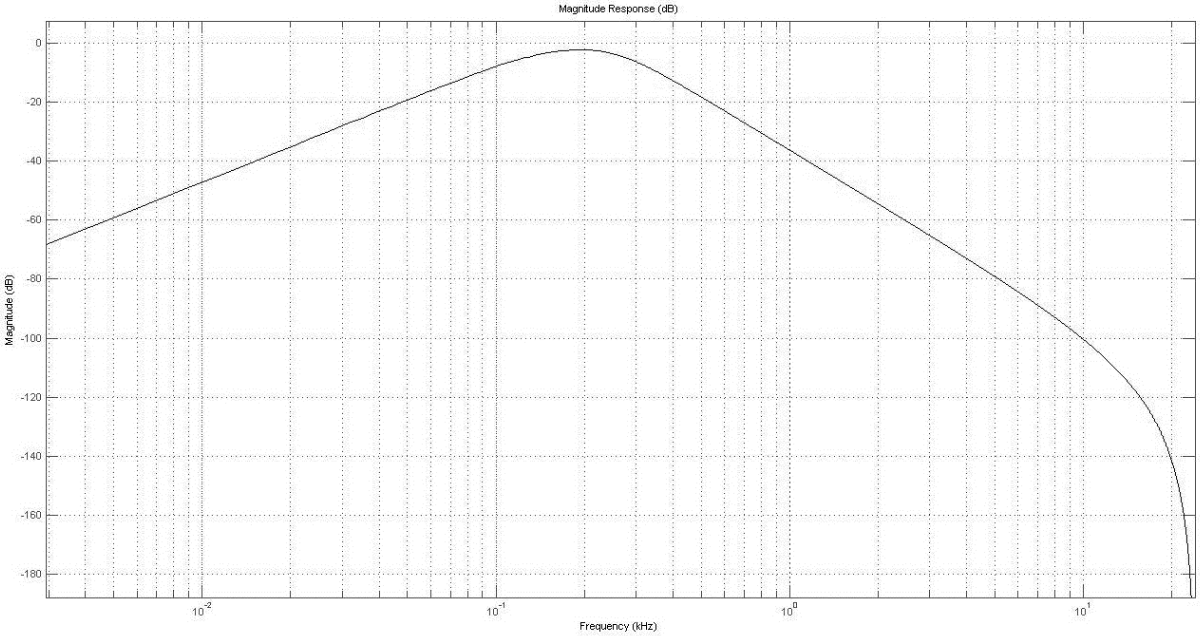


Figure 17 “Filter Response for LFE Channel”