Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

Loudness Normalization In The Age Of Portable Media Players

Martin Wolters[1], Harald Mundt[1], and Jeffrey Riedmiller[2]

[1] Dolby Germany GmbH, Nuremberg, Germany
martin.wolters@dolby.com, harald.mundt@dolby.com

[2] Dolby Laboratories Inc., San Francisco, CA, USA
jcr@dolby.com

ABSTRACT

In recent years, the increasing popularity of portable media devices among consumers has created new and unique audio challenges for content creators, distributors as well as device manufacturers. Many of the latest devices are capable of supporting a broad range of content types and media formats including those often associated with high quality (wider dynamic-range) experiences such as HDTV, Blu-ray or DVD. However, portable media devices are generally challenged in terms of maintaining consistent loudness and intelligibility across varying media and content types on either their internal speaker(s) and/or headphone outputs.

This paper proposes a nondestructive method to control playback loudness and dynamic range on portable devices based on a worldwide standard for loudness measurement as defined by the ITU. In addition the proposed method is compatible to existing playback software and audio content following the Replay Gain (www.replaygain.org) proposal. In the course of the paper the current landscape of loudness levels across varying media and content types is described and new and nondestructive concepts targeted at addressing consistent loudness and intelligibility for portable media players are introduced.

1 INTRODUCTION

The movie industry was probably the first to identify and solve the problem of varying mixing & playback levels of content. Both, content producers as well as movie theaters had a genuine interest in finding a solution that would re-produce audio content at a level pleasant to consumers. Since then the corresponding SMPTE recommendations guarantee a rather consistent playback level across theaters and across different content; a true win-win situation for consumers, movie theaters and content owners alike.

The situation in broadcast was already more challenging, given that the individual playback systems in the homes are not controlled by technicians and due to the more complex distribution channels and networks for broadcast. With the introduction of digital broadcast the industry established the concept of time-varying-metadata which enables to control gain-values at the

receiving end to tailor content to a specific listening environment. An example is the metadata included in AC-3 / E AC-3 / HE AAC[1] which includes general loudness normalization information (called "dialnorm" in AC-3 and E AC-3, and "program_reference_level" in HE AAC) as well as gain-words to reduce the dynamic range of a program (called "dynrng" and "compr" in AC-3 and E AC-3).[1] Such systems are specifically powerful for situations where the operating modes in a receiver are also specified, like the so called "line mode" and "RF mode" for AC-3 and E AC-3. These technologies are also part of today's AV discs like DVD and BluRay.

The most important distribution channel for audio content is still the CD which contains 16-bit PCM data unfortunately without any loudness or dynamic range metadata. The peak-normalization typically used for CD's is said to be the main reason for the so called "loudness war" which has led to extremely reduced dynamic range of audio content with high average levels [2]. Many authors, mixing engineers and audio enthusiasts alike, have complained about this situation and several proposals have been made to improve the situation. Solutions range from voluntary metering and leveling practices on the production side similar to the mentioned SMPTE recommendations (for example [3]) up to branding of content on the consumer side (for example [4]).

At the same time, consumer behavior changed over recent years with coded content (e.g. content in data-reduced formats such as mp3) becoming more popular and important for content distribution and storage. Such formats allow for virtually unlimited dynamic range which content owners and audio enthusiasts would like to take advantage of. In addition, the increasing popularity of mobile phones as personal media players has created new challenges in designing high quality playback devices that meet customer expectations of consistent leveling and best audio quality under various listening conditions. The large number of content in personal music collections (often exceeding thousands of files) as well as the broad range of audio formats such as mp3, HE AAC, OGG, WMA, AC-3, or E AC-3 further complicate the matter.

[1] AC-3 is also known as Dolby Digital, E AC-3 as Dolby Digital Plus and HE AAC as Dolby Pulse.

2 GENERAL CONSIDERATIONS FOR LOUDNESS NORMALIZATION IN PORTABLE MEDIA PLAYERS

Today there are already many proposals, components, ideas, and approaches available for loudness normalization in consumer devices. However, none of these proposals is widely supported in portable media players and none specifically addresses the challenges in such devices. Before explaining the proposed solution, the key stakeholders and their expectations with regard to loudness normalization are identified and the basic building blocks of a loudness normalization system are explained.

2.1 Identifying stakeholders and their expectations

2.1.1 Consumers

In general consumers would like to be presented with audio content in a form that matches their listening environment, whether it is being played back on a high-end multi-channel setup or on the mono speaker of their mobile phone. The listening environment influences the desired output level as well as the maximum dynamic range tolerance [5]. For portable devices this most often means: "loud" to address the physical limitations of the playback device and little dynamic variations to adapt to the environmental noise.

The average consumer also does not want to understand technical details such as the specific codec used for data compression. Instead, the playback experience should be equally well for all of the content. It might be acceptable to use different versions of the same content tailored to the different listening environments. However, this requires careful trade-offs with other requirements such as file size (e.g. number of hours of content on a limited storage) and download time. With content moving freely between devices, a single version of the same content is certainly the preferred solution.

2.1.2 Content owners and distributors

There are probably still content owners out there that would like their content to playback the loudest. However, in general the desire matches more that of consumers: Playback all content at the same desired output level and with desired dynamic range not being limited by the distribution format. Of course, variations in level must be possible. For complete albums some

tracks might specifically be recorded above or below the album average, and certain parts of a song or movement might be louder or quieter than the average for the complete song. An automatic gain control (AGC) in the distribution channel is usually less preferred as it also overrides such intended level changes. Certainly, content owners are aware of the varying dynamic range tolerances of consumers and are willing to take these limitations into account. However, they would like to be able to control the final output to the consumer. Thus, they prefer a predictable signal chain for their content so that they can mix, setup parameters. and monitor accordingly.

Both, content owners as well as distributors do favor a limited number of versions of the same content. A single version is preferred but several exceptions are common: High vs. low-bitrate versions, sometimes specifically downmixed versions vs. multi-channel originals, radio edit vs. album version etc.

2.1.3 Device manufacturers

Obviously, the device manufacturer's heart is closest to the consumer who buys the devices. Therefore, meeting the customer's desire for playback at the appropriate level and within the tolerable dynamic range has highest priority. Single-sided solutions are an interesting option as they also provide the opportunity to differentiate own products from competitive offerings. Compatibility to a large share of available content is high on the priority list.

Device manufacturers are also concerned about computational complexity. Hence, off-loading some of the signal processing to the encoding stage is preferred. Overall system design becomes a critical part of product design as well. Ensuring that the user experience and user interface is independent of the content format is a challenge. Simple, codec-agnostic solutions support the device manufacturer in addressing the consumer's desire for systems that work seamlessly across all their content. The complexity of the system design also increases with the versatility of today's devices: A mobile phone with built-in speakers, headphone output, and the ability to connect to a high-end AV system by means of a cradle demands a sophisticated control of the audio output stage that needs to be aware of the current listening environment.

Last but not least, appropriate test materials to ensure interoperability between encoding and decoding systems are an essential requirement for device manufacturers. Recent investigations (not yet published) have shown that many open-standard-codecs and features suffer from the lack of appropriate test vectors and procedures, resulting in numerous incompatibilities between devices.

2.2 Identifying technical components

These requirements lead to the following components and design restrictions for a loudness normalization system on portable media players:

2.2.1 The preference for a non-destructive solution

The authors strongly believe that the desire to limit the number of different versions of content together with the various listening environments that need to be supported demand a non-destructive solution for the problem at hand. This seems to be in the best interest of consumers, content owners, distributors, and device manufacturers alike. Non-destructive means no changes to the actual PCM signal (or payload) prior to the decoding stage and hence requires the use of metadata, potentially combined with single-ended solutions, instead.

2.2.2 Applying gains to match the desired target output

The basic solution is obvious: Define a target output level, determine the actual level of the content, and apply a matching gain value during playback. It also seems obvious that a solution needs to take into account potential clipping in cases where the content needs to be boosted to match the target.

In order to address the stakeholder's expectations, the definition of a target output level becomes a critical piece in defining a solution. The average consumer would like to be presented with "the best" solution, not worrying about selecting from too many options. The content owners prefer a predictable behavior – too many options require monitoring content with too many different setups. Device manufacturers also favor fewer options (which in turn means a simpler system design). However, the specific output target reference level might be highly dependent on the device's capabilities which can in turn lead to a large variation across devices.

Given the fact that the proposed solution will rely on metadata to be present in the content one also needs to take into account, how the system behaves for content without the required metadata present, e.g. legacy content.

2.2.3 Conveying the loudness level

Systems for streaming and broadcasting like AC-3, E AC-3, or HE AAC typically rely on transmitting a time-varying[2] value representing the loudness level of the current program to the decoding device (e.g. the "dialnorm" parameter). This allows content owners to easily control the complete signal chain and reduces the computational complexity on the decoding device.

For file-based systems such a value typically does not change for a given file. While it is still possible to utilize loudness levels encoded into the payload, systems have been designed that rely on a single value per file. iTunes is using such a value, called "Sound Check". Little is known about the specific algorithms used by Apple. However, it is known that only certain file types are supported: "Sound Check works with .mp3, .AAC, .wav, and .aiff file types. It does not work with other file types that iTunes can play." [6]

Another proposal from 2001 which got some traction across formats, implementations, and devices is called "Replay Gain". [7] It does specify an algorithm to compute a gain value to normalize loudness across tracks and albums and proposes a storage format. However, the actual storage format used was not formally standardized so far and no test vectors are available to date. Replay Gain is implemented for example in the freely available software *aacgain* [8] (version 1.8.0 was used throughout our investigations) or in the *WINAMP* plug-in *ml_rg.dll*. Matlab code is given on www.replaygain.org.

2.2.4 Determining loudness

Loudness itself is a highly subjective quantity (and as such, cannot be measured directly) that involves psychoacoustic, physiological, and other factors which have been and continue to be thoroughly studied to this day. Hence, this highly subjective quantity (loudness) often results in substantial differences in loudness perception between listeners, making a single measurement method that considers all of the above factors – for all individuals – incredibly complex. This is borne out by real-world experience, as there is often no single loudness level that will satisfy all listeners (or even a single listener) all of the time.

At best, we are only able to approximate the loudness of sounds by artificial means. One study performed by Dolby Laboratories concluded that even when audio programming has been "normalized" by a group of people "by ear" the "normalized" programs do not completely satisfy a different group of listeners 100 percent of the time. In fact, the different groups only agreed approximately 86 percent of the time. Given this level of uncertainty among groups of listeners, any loudness measure we utilize will not guarantee that we satisfy 100% of the listening audience.

However, what we can achieve is to agree on a common loudness measurement such that different encoding tools at least behave consistently. This in turn will maximize consumer satisfaction since content from different sources will behave similarly.

3 THE PROPOSED SOLUTION

Given the considerations above, the authors recommend to implement loudness normalization in portable media players

- based on a worldwide accepted standard for loudness measurement as defined by the ITU
- by conveying such loudness values as Replay Gain equivalents specifically for formats that do not support loudness metadata in the actual payload, and
- by setting a new target output reference level for portable devices while controlling the dynamic range either with single-sided algorithms or combining a single-sided limiter with dynamic range control derived from already available metadata.

Following some more details and explanations about aspects of the the proposal:

[2] While this time-varying loudness value is carried in every bitstream syncframe, it is recommended to set this value on a program-by-program basis whereby the value itself is 'static' for the duration of each (single) program and/or file.

3.1 Loudness Measurement & Normalization according to ITU-R BS.1770

As stated earlier, over the past few years consumers have continued to demand support for a wider range of media source types on their portable devices (i.e. mobile phones) including music as well as multimedia files such as TV, movie content and their own user generated content. Moreover, the convergence of several media sources/types creates new challenges for portable device manufacturers, service operators and content creators alike as media sources like those listed above are typically created with different production and/or normalization philosophies (since they are destined for very different playback environments/devices, etc).

In July 2006, after several years of work, the ITU-R approved and published a new recommendation for estimating the loudness of broadcast programs entitled, ITU-R Rec BS.1770 "Algorithms to measure audio programme loudness and true-peak audio level". [9]

ITU-R BS.1770 defines a royalty-free and low-complexity method to estimate the overall loudness of an audio program by computing the frequency weighted energy average over time. Most importantly, it yields a single loudness value that represents the ***overall*** loudness of an **entire** program (i.e. AV or Music track file) thereby making automated loudness estimation and normalization of media files much simpler and (in many cases) non-destructive without (or with a much lower) impact to artistic intent when compared to real-time loudness controllers and dynamic range processing.

The worldwide acceptance of BS.1770 has grown considerably since its introduction and is currently referenced in documentation published by well known groups such as the Advanced Television Systems Committee (ATSC) [10] and the European Broadcasting Union (EBU). Furthermore, the use of BS.1770 is a mandatory provision within a significant (and growing) number of program delivery specifications for many of the world's most influential broadcasters and content providers (for example [11]).

One of the most significant benefits of the BS.1770 method is its ability to work very well with estimating the loudness across a wide variety of content types. Including, common channel configurations such as mono, stereo and multichannel programming as well as content such as music, television, movies, sport and sound effects. Thus, making it a perfect candidate for use in estimating & normalizing the loudness of a wide variety of media files destined for portable media players.

During the ITU-R study, subjective testing showed that the BS.1770 algorithm yielded the best performance among several other methods including algorithms based on psychoacoustic models.

Prior to the release of BS.1770 the ITU-R study group conducted formal subjective tests in order to create a subjective database (containing perceived loudness ratings for each test item) that would be utilized to evaluate candidate loudness measurement algorithms and their ability to predict the results from the subjective tests. The subjective database itself included a wide range of audio material including television & movie dramas, sporting events, news broadcasts, sound effects and music. Several candidate algorithms were submitted for consideration and in the end, the results demonstrated that a reasonably simple weighted mean-square measure[3] worked extremely well on mono, dual mono, stereo, as well as multichannel audio programming and yielded an overall correlation of $r = 0.977$ across 336 test sequences from the subjective database.

A bit more on the mathematical machinery under the hood of ITU-R Rec. BS.1770, the algorithm was derived from, and is based on the Leq(RLB) algorithm described by Soulodre in [12] to support mono, stereo and multichannel audio signals while retaining a very low computational complexity. Thus, allowing it to be easily implemented and/or adopted by many equipment manufacturers at a very low cost.

Figure 1 (page 6) shows a high-level functional block diagram of the ITU-R algorithm taken from ITU-R Rec BS.1770.

The BS.1770 algorithm can be scaled to accommodate any number of audio channels contained within programs, however for multichannel 5.1 audio programs, it is assumed that the multichannel signals conform to the ITU-R BS.775-1 5.1 channel configuration. As a side note, the LFE channel is NOT included in the measurement.

[3] Ultimately became ITU-R Rec. BS.1770 Loudness Algorithm.

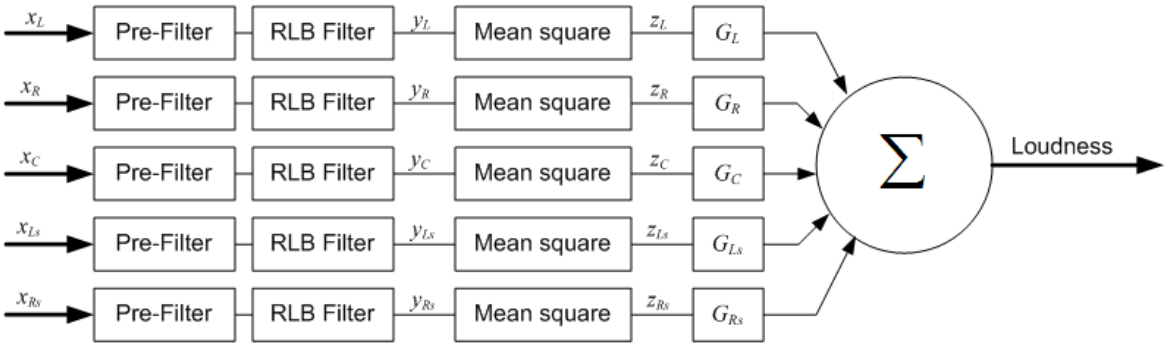

Figure 1 - ITU-R Loudness Meter/Algorithm Functional Block Diagram

Within the BS.1770 algorithm, each of the individual audio channels being measured is first passed through two filters in cascade. The pre-filter, which has a high-frequency shelving characteristic, is to account for the acoustical effects of the human head (modeled as a rigid-sphere) while the RLB filter (**R**evised **L**ow-frequency **B**) is a modified version of the standard IEC B-weighting curve and has a characteristic where the low frequency response falls between the IEC C and B weighting curves. Figure 2 shows the combined response compared to A-weighting.

Once the input signal for each channel has been filtered, the mean-square energy for each channel is computed for the entire measurement interval (time). The individual channel powers are then weighted in accordance to the angle of arrival [Refer to Table 3 in [9]] and then linearly summed to provide the overall (final) loudness value as follows:

$$\text{ITU Loudness} = -0.691 + 10\log_{10}\sum_{i}^{N} G_i \bullet z_i, \text{ in dB} \quad (1)$$

The weighting blocks (G_i) in Figure 1 above for the Left, Center and Right channels are always 1.0 (0dB) where the weightings for the Left Surround and Right Surround channels are always 1.41 (~ + 1.5dB) each. The emphasis on the surround channels acknowledges the fact that sounds arriving from behind (the listener) could conceivably be perceived as being louder relative to those arriving from the frontal direction (see [9] for details).

The result is the loudness given in LKFS (LKFS indicates a K-weighted level relative full scale). Note, the constant value (-0.691) in the equation above is a calibration constant that addresses the combined effects of the Pre-filter and RLB filter at 1kHz. (Note the small filter *gain* (positive) at 1kHz in) This constant ensures that a mono full scale sine wave is assigned a loudness of -3.01 LKFS.

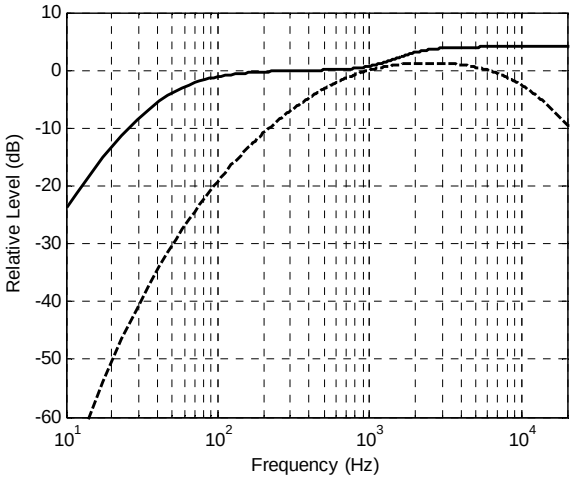

Figure 2 - ITU-R BS.1770 Frequency Weighting (solid) vs. A-weighting (dashed)

3.2 Conveying loudness values as Replay Gain equivalents

In general, the basic Replay Gain proposal as well as the way it is supported in various products today are very much in line with the requirements identified during our investigations. Instead of coming up with yet another syntax to transmit loudness values for file-based content, the authors suggest to embrace the current Replay Gain format in order to speed up adoption in the market. However, given that the authors are in favor of the ITU-R BS.1770 algorithm – which

was not available at the time of the original Replay Gain proposal and which probably would have been chosen otherwise – the authors suggest matching the Replay Gain semantics to ITU-R BS.1770 results based on a statistically derived linear equation.

3.2.1 Matching the Replay Gain syntax

Replay Gain consists of two values: "peak signal amplitude" and "gain adjustment". It can be calculated on a track-by-track basis, or album-by-album basis. Track-based values are more suited for use cases and playlists where tracks from different albums are mixed. Album-based values are more suited for use cases where all tracks of an album are played consecutively.

The originator of the Replay Gain values can be specified as the engineer/artist/producer, user, or automatically determined.

Since the authors were not able to identify a formal specification of the actual syntax used to store Replay Gain values, several files with Replay Gain support have been analyzed. Based on these results the following syntax is proposed: Content stored in files compliant to the MPEG-4 file standard shall use iTunes-style metadata. Other formats shall store Replay Gain values in ID3v2 tags.

The Replay Gain adjustments shall be between -18 dB and +13 dB (corresponding to a range of loudness values from 0 to -31 LKFS). Values outside this range shall be clamped to –18 dB and +13 dB.

3.2.1.1 Replay Gain in iTunes-style metadata

- Replay Gain metadata shall be added as an extension box of type '----', conforming to standard iTunes-style metadata.
- A 'mean' box shall be present within the '----' box and contain the meaning "org.hydrogenaudio.replaygain"
- A 'name' box shall be present within the '----' box and contain the name of the value :
  - replaygain_track_gain
  - replaygain_track_peak
  - replaygain_album_gain
  - replaygain_album_peak
- A 'data' box shall be present within the '----' box and contain the value in the following formats:
  - The gain adjustments shall be written as a dB floating-point value with 2 decimal places and a leading -/+. (e.g. "`-4.65 dB").`
  - The peak signal amplitudes shall be written as a floating-point value (e.g. `"0.860931396"`). The peak signal amplitude may (and often is) over 1.0
- Players should match only on the value in the 'name' box and ignore the value in 'mean' box for compatibility.

This specification defines additional iTunes-style metadata for the Replay Gain originator code:

- Replay Gain may include "originator code" information.
- A 'mean' box shall be present within the '----' box and contain the meaning "org.hydrogenaudio.replaygain"
- A 'name' box shall be present within the '----' box and contain the name "replaygain_originator_code".
- The following originator codes shall be used:
  - 000 = Replay Gain unspecified
  - 001 = Replay Gain pre-set by artist / producer / mastering engineer
  - 010 = Replay Gain set by user
  - 011 = Replay Gain determined automatically
- The 'data' box shall contain a text string representing the concatenation of the 3-bit originator codes for the Replay Gain values in the following order:
  - replaygain_track_gain
  - replaygain_track_peak
  - replaygain_album_gain

- replaygain_album_peak

For example, “011011000000” maps to automatically generated values for track gain and peak and unspecified values for album gain and peak)

At a minimum a file with Replay Gain metadata shall include a track gain value.

Appendix B provides an example for a complete Replay Gain data set in iTunes style metadata.

3.2.1.2 Replay Gain in ID3v2 tags

Replay Gain values are stored in ‘TXXX’ fields which follow the following syntax:

```
<Header for 'User defined text
information frame', ID: "TXXX">
Text encoding $xx
Description <string according to
encoding> $00 (00)
Value <string according to encoding>
```

Each Replay Gain parameter is contained in its own specific ‘TXXX’ element. To distinguish parameters, the “Description” string takes the same values as written in the iTunes ‘name’ box (see above)

- replaygain_track_gain
- replaygain_track_peak
- replaygain_album_gain
- replaygain_album_peak
- replaygain_originator_code

The parameter value is stored in the “Value” field. It uses the same format as described in the iTunes section above.

3.2.2 Matching the Replay Gain semantics

In the Replay Gain proposal [7] a gain based on a loudness measure is derived. This gain is designed to adjust the music loudness to the loudness of pink noise at -20 dB relative full scale played back over stereo loudspeakers. The associated sound pressure level is 83dB SPL. No formal study on Replay Gain performance for loudness alignment is available. The Replay Gain procedure comprises the following steps:

- Frequency weight all channels according to an average inverse equal loudness curve

-Compute power for non-overlapping blocks of 50 ms lengths and average over channels for stereo content

-Compute the block power which is exceeded by 5% of all blocks per music track and derive the *loudness* in dB relative full scale

The gain is computed in dB as [*ReferenceLoudness-loudness*] so that the loudness of the pink noise reference is matched.

In order to support existing Replay Gain metadata in addition to ITU-based loudness for both the proposed playback solution as well as for existing Replay Gain compatible audio players, a reversible conversion is required. The goal is to achieve consistent loudness across files potentially with both types of loudness information and to benefit from ITU-based loudness where available.[4] To achieve this goal both leveling approaches have been investigated statistically by means of a music data base which is described in section 3.2.2.2.

3.2.2.1 Comparing ITU-R BS.1770 and Replay Gain

Both leveling approaches measure a frequency weighted power. The main differences are the filter characteristics and the statistical power analyses from which the loudness is derived. Based on the loudness measure and a fixed target loudness a gain value is derived in the Replay Gain procedure. No target loudness is defined in the ITU recommendation.

While ITU-R BS.1770 applies a high-pass filter characteristic the Replay Gain filter looks more like a band-pass[5] as shown in Figure 3.

[4] ITU-R BS.1770 is described in more detail in section 3.1

[5] Care has to be taken with ITU-R BS.1770 if signal energy is present above the frequency hearing range

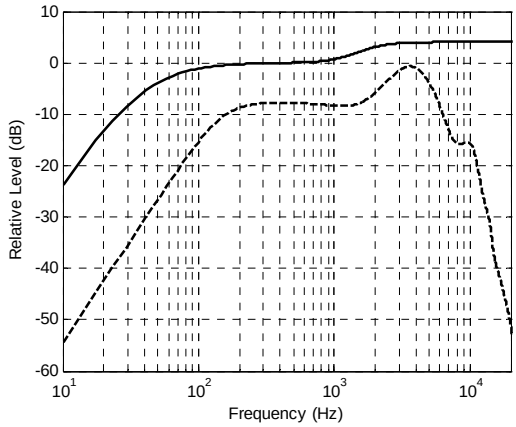

Figure 3 - ITU-R BS.1770 Frequency Weighting (solid) vs. Replay Gain Frequency Weighting (dashed)

In ITU-R BS.1770 signal energy is averaged over the complete music track potentially including silence which typically does not contribute to the subjective loudness. In our investigations silence is therefore excluded from the measurement. However since the music data base does not exhibit a significant amount of silence the impact is limited.

Replay Gain on the other hand measures the frame power that is exceeded by only 5% of all frame powers which is near the maximum frame power. As will be shown in section 3.2.2.2 the relation between this near maximum power and the long term power as applied in ITU-R BS.1770 is non-linear especially for higher loudness ranges where the dynamic range of the music tends to be reduced.

3.2.2.2 The Music Data Base

The music data base consists of 21220 stereo files originating from different private music collections. The data base is considered as statistically relevant for today's typical stereo music content for playback on portable devices. Compression formats are mp3 and AAC at various bitrates and sample rates between 32 and 48 kHz. Replay Gain has been calculated for all files using the freely available software *aacgain* (version 1.8.0) [8]. Loudness according to ITU-R BS.1770 has been computed using an internally developed software implementation (excluding silence[6]).

[6] Silence is identified when the maximum peak level relative to full scale remained below -60 dBFS for more than one second and less than ten seconds

On average 0.6% of the music track lengths are identified as silent. Files at low loudness seem to have longer silent periods than louder files (0.3% at -5LKFS, 1.2% at -30LKFS). Still 95% of all files around -30 LKFS have less than 5.3% silence.

The data base genre focus is on Pop, Rock and Alternative music with some Jazz and Electronic and very little Classical music. Figure 4 shows the distribution of genres according to metadata associated with the files[7].

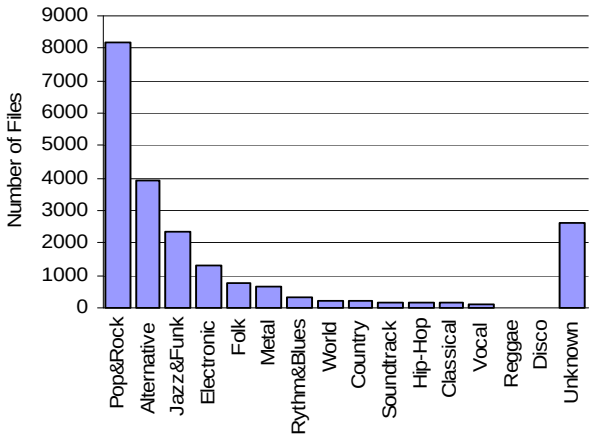

Figure 4 - Distribution of Genres in the Music Data Base

Loudness varies depending on genre for up to 10 LKFS on average. Classical music has the lowest average loudness level at about -20 LKFS as can be seen in Figure 5.

[7] The metadata was either part of files acquired through online distribution channels such as Amazon, or was added when ripping CDs by means of databases like CDDB.

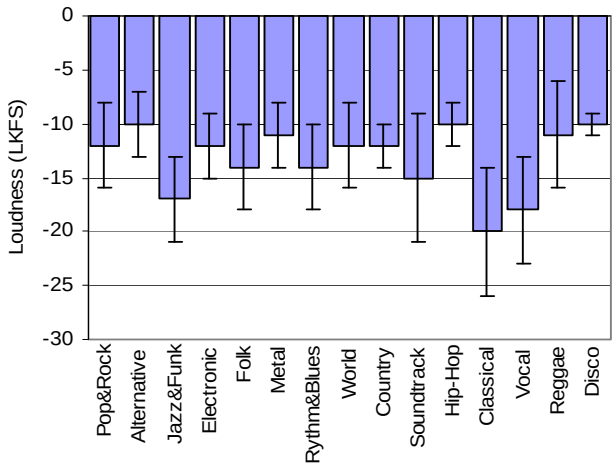

Figure 5 - Mean Loudness (ITU-R BS.1770) and Standard Deviation for different Music Genres.

According to the available metadata about half of all music files are no older than year 2001 as can be seen in Figure 6. There is a tendency for increasing loudness from the early 90's on until today as can be seen in Figure 7.[8]

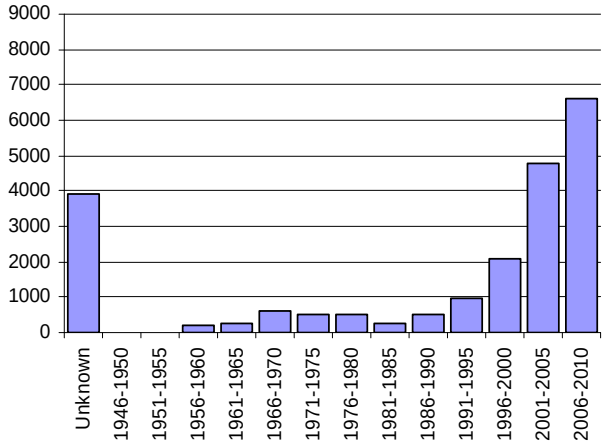

Figure 6 - Distribution of Year Information

The overall loudness histogram and distribution are shown in Figures 8 and 9. The data base also includes the Top 75 Bestseller mp3 downloads from www.amazon.de as of December 2009. The mean loudness for this content is -8.5 LKFS.

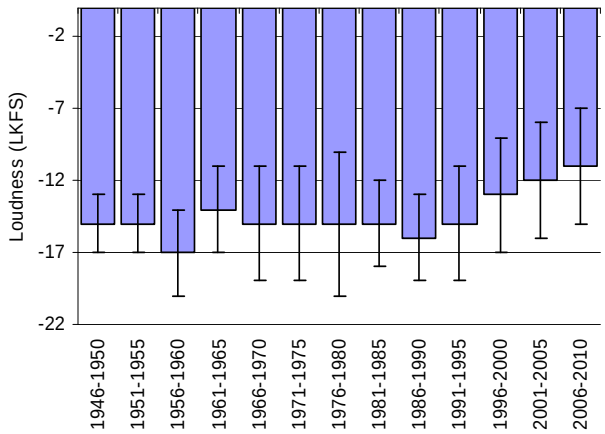

Figure 7 - Mean Loudness (ITU-R BS.1770) and Standard Deviation vs. Year.

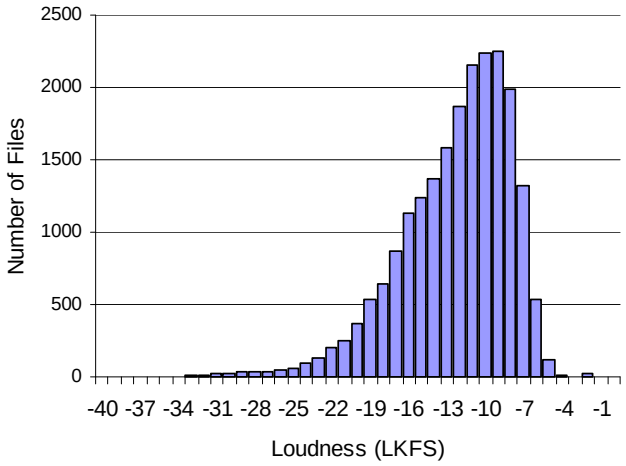

Figure 8 - Loudness Histogram for the Music Data Base

[8] As with all the metadata used during this study, it can not be guaranteed that it always reflects reality. In case of the "Year" metadata, this typically reflects the year of the original recording. However, some CD's might have been re-mastered at a later point in time with actual changes to the overall loudness level.

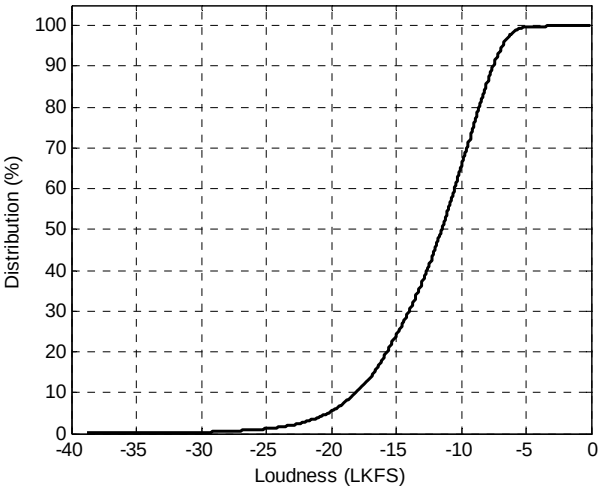

Figure 9 - Loudness Distribution for the Music Data Base

3.2.2.3 Conversion between Replay Gain and ITU-based loudness

Although Replay Gain and ITU-based loudness share common types of processing such as frequency weighting and power measurement their relation is highly complex and content dependent. Therefore a relation shall be derived statistically from the investigated music data base. Since Replay Gain describes a gain in dB and ITU-based loudness describes a level relative to full scale the basic relation is inverse. In order to maintain the nature of the Replay Gain proposal the relation shall ideally be modeled as an inverse linear relation. This way a loudness change of for example -10 LKFS will translate into an estimated Replay Gain change of +10 dB.

A Replay Gain compatible player would then directly benefit from the ITU-based Replay Gains by providing a loudness matching solution based on an industry wide accepted standard. Therefore the task is to find an ideal offset for a straight line fit between both types of loudness data with a slope of -1 which is shown in Figure 10 as a red solid line. For comparison the general least mean squares fit is also shown as a red dashed line with an estimated slope of -0.81[9]. The standard deviation of the Replay Gain estimation error for this fit is 1.0 dB. Since more than 75% of all music files have loudness above -15 LKFS (see Figure 9), the fit with the slope of -1 is considered a reasonable approximation to the general least mean squares fit with respect to the observed error variance.

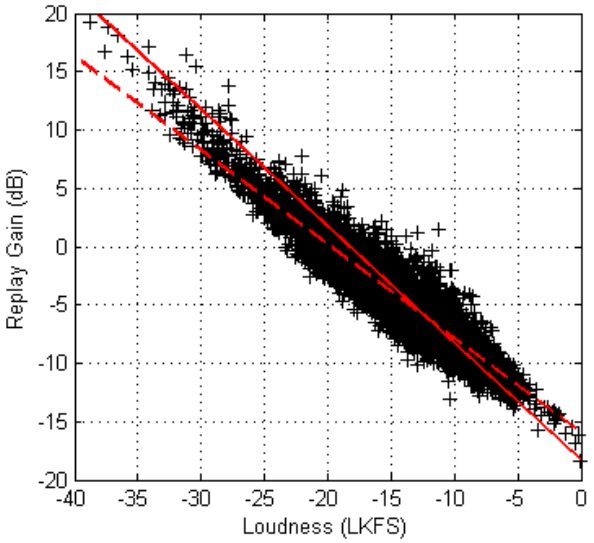

Figure 10 - Replay Gain vs. ITU-R BS.1770 Loudness for 22120 music files.

Therefore the following reversible conversion between Replay Gain and ITU-based loudness is proposed:

$$RG = -18dB - L \qquad (2),$$

where RG is the estimated Replay Gain and L is the Loudness according to ITU-R BS.1770 in LKFS (dB relative full scale).

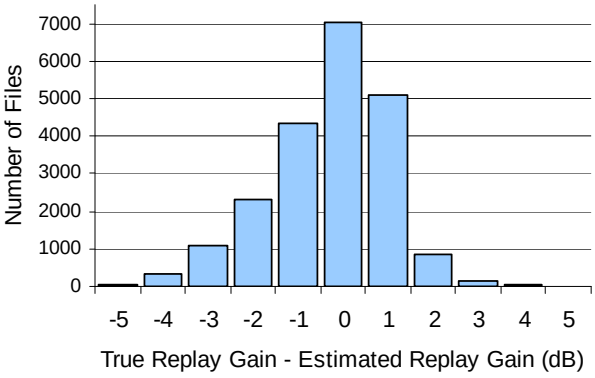

Figure 11 - Replay Gain Estimation Error Histogram for the Conversion according to (2) and Loudness >= -31 LKFS

The optimum slope of -0.81 indicates a dependency on loudness. This can be explained by the loudness

[9] Linear regression has been applied to the most relevant data between -31 and 0 LKFS.

dependent level difference between the near maximum frame power (Replay Gain) and the long term power average (ITU-R BS.1770). Due to dynamic range processing and limiting this difference tends to get smaller for louder music. This observation is supported by an experiment where in the Replay Gain procedure the near maximum frame power analysis was replaced by a long term power average. Then the least squares fit between this modified Replay Gain and the ITU-based loudness has indeed a slope of -1 as can be seen in Figure 12.

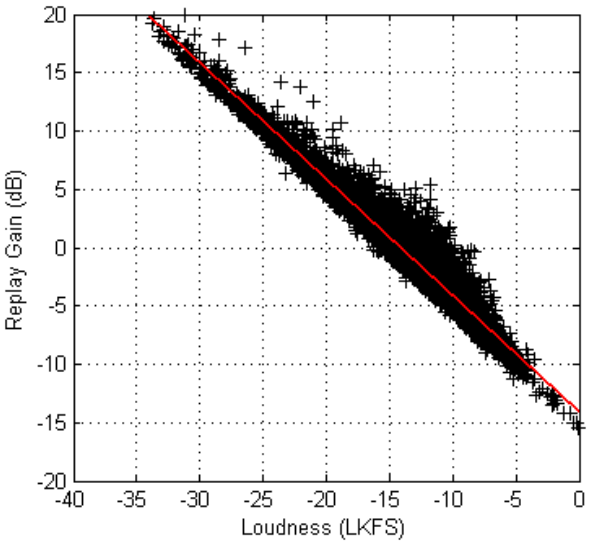

Figure 12 - Modified Replay Gain (use long term power) vs. Loudness (ITU-R BS.1770) for all music files. The red line is the general least mean squares straight line fit and has a slope of -1.0.

3.3 Recommended playback system

In order to implement a playback system supporting the proposed method for loudness normalization, three principal design decisions need to be made:

1. What is the desired target level?
2. How is the dynamic range controlled?
3. How are files treated that do not include any loudness metadata?

3.3.1 Loudness Distributions for Music & AV Content

As stated earlier, consumers are continuing to diversify the types of content they carry on their portable media players. Considering this fact, the authors also acquired 116 full-length (audio/video) television programs and movies for analysis and comparison against the music database. The majority of the AV content was acquired directly from the programmers/studios (prior to transmission encoding) and included both stereo and multichannel content. The red distribution in Figure 13 displays the loudness distribution of the 116 programs in terms of ITU-R BS.1770 Loudness (LKFS).

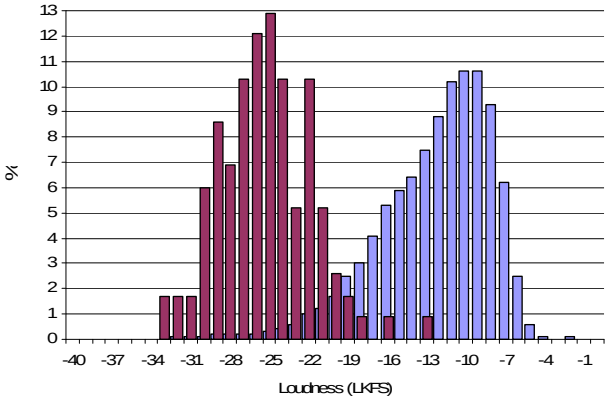

Figure 13 - Loudness Distributions from 116 Full-length Television Program & Movie Files (red, left) & 22120 Music Files (blue, right)

The loudness range of the AV content does span a very wide range of 20dB with the majority of items spanning a 7dB range from ~ -22 LKFS to ~ - 29 LKFS. In contrast, the blue "music" distribution in Figure 13 displays the loudness distribution for the 22120 music files used in our research. Given that there is very little overlap between the loudness distributions of the two data sets it is quite clear that consumers would be forced to adjust the volume (significantly) when switching between AV and music content. One option to reduce the need for frequent listener adjustment would be to apply a static attenuation of ~ 14dB to the music items (in the figure above) only. Using this approach, the two data sets would no longer have a bimodal characteristic and thus more likely to be pleasing to the listeners. However, utilizing this approach for portable devices would likely yield a significant reduction in output level drive to the internal speakers and/or headphones. Moreover, considering that these devices will likely be used in noisy environments limits (and likely eliminates) the practical usability of this approach. Hence, it is the AV loudness (content) distribution that must be adjusted (in the positive direction) to overlap the music distribution to address this issue. Considering that the production philosophies of television and movie

content typically produce a lower average loudness (leaving headroom for music and effects to make a dramatic impact) a new common reference level must be established for devices with severe level and dynamic range limitations that does not require portable device specific content preparation and simulcast for these devices. Hence, playback devices of this type can easily take advantage of either coded audio bitstream and/or container metadata to normalize any/all content to this newly defined and common reference loudness level. Furthermore, if the content can be played back via a higher quality reproduction system (directly from the portable device) a different reference loudness level can be selected by the user and/or automatically (default) by the decoder in the higher quality reproduction system. This approach would allow wider and/or the original dynamic range to be presented to the listener and more importantly a loudness and dynamic range that is typical for this type of system/device, etc.

3.3.2 Choosing an appropriate target level

Portable media players (like mobile phones, dedicated personal music players, or laptops) often need to support up to three listening environments:

- built-in speakers
- headphone output
- line output used in combination with a cradle (either analog or digital; newer devices might even support multi-channel output)

For the later use case – which most often means that the device is connected to Hifi-equipment – a lower target level of -31 LKFS as specified for example in "line mode" for AC-3 and E AC-3 is most appropriate, enabling full dynamic range capabilities.

For noisy environments a higher target playback level in combination with high quality limiting and suitable dynamic range compression will provide a better listening experience. At the same time level jumps from content with varying loudness are avoided. The dynamic range compression profile may be adapted to the specific use-case. Built-in speakers or open headphones in noisy environments may require moderate dynamic range compression. Spoken content such as in audio books or podcasts under noisy environmental conditions may require a more aggressive compression profile for optimum speech intelligibility.

In order to find a useful target playback level, the loudness for all music files (see section 3.2.2.2 for details) has been analyzed statistically (see Figures 8, 9, and 13). The goal is to find a target playback level which a) does not require excessive analog gain for sufficient loudness, b) needs only moderate limiting for typical content. A natural choice is the average level of typical content without any additional leveling or dynamic range compression. It turns out that this choice also fulfills both requirements mentioned above. As a side effect if the leveling system is switched off severe loudness jumps are unlikely since many files already are at or near the target playback level. In order to be adaptable to the actual environment and playback device, the authors propose 3 different target levels around the measured median loudness namely -8, -11 (the recommended default), and -14 LKFS.

3.3.3 Combining loudness normalization and dynamic range control

Given that the lowest supported loudness value is -31 LKFS, all target levels above -31 LKFS need to support clipping prevention through dynamic range control. Hence, at a minimum portable media players supporting this proposal need to provide a limiter.

Formats that support metadata for dynamic range control such as AC-3, E AC-3, or HE AAC can also apply such metadata prior to the signal being fed into the limiter. For example, running an AC-3 or E AC-3 decoder in RF-mode with a target level of -20 LKFS will then require an additional boost of 9 dB followed by a limiter.

This combination results in a new reference decoder level of -11 LKFS and thus defines a new decoder operating mode for metadata based coding systems. This new mode is referred to as Portable Mode and is an addition to the two legacy decoder operating modes within the AC-3 and E AC-3 systems. (i.e. Line & RF Modes) and could easily accommodate the use of a 3rd set of dynamic range control gains (computed in the encoder) and carried in the bitstream for clip protection while in Portable Mode.

Obviously, the quality of the single-sided limiter is crucial. Dolby Laboratories for example developed a look-ahead limiter with signal-dependent attack and release times, which is able to prevent clipping even for critical (e.g. dynamic) content without any audible artifacts.

3.3.4 Integrating files without any loudness metadata

When preparing for playback of a file, a device first needs to check, whether or not a Replay Gain value is available. In cases where a complete album is being played back, the authors recommend to prefer the album-gain over track-gain. Otherwise, using the track-gain should be the default.

In cases where no Replay Gain value is available the system should check for the presence of a format-dependent loudness value such as the "dialnorm" parameter in AC-3 and E AC-3 or the program reference level in MPEG HE AAC and then use that value instead.

If neither Replay Gain nor format-dependent loudness values are available, the authors recommend using a default loudness value of -11 LKFS for stereo music content (the median of all content measured during this study) and -23 LKFS for AV and multi-channel content.

3.3.5 Overview of the complete playback system

The following page provides a graphical overview of the complete system (Figure 14) as well as a pseudo-code implementation of the playback control to further illustrate our proposal.

4 SUMMARY AND CONCLUSIONS

In this paper the authors motivated a new system for loudness normalization in portable media players which satisfies the expectations of key stakeholders. The building blocks of the system have been described in detail and design decisions were explained.

First products supporting the new method – both on encode as well as decode side – have been released by Dolby Laboratories already[10]. The authors hope to convince more players on the market to support this approach. In the end this will increase satisfaction of consumers, content owners, and device manufacturers alike and will solve one of today's industry wide problems: Loudness normalization in portable media players.

[10] Dolby Media Generator is a file-based content creation tool supporting the proposed Replay Gain syntax and semantic. Dolby Mobile 3 includes support for the so called "portable mode" which implements loudness normalization in line with this proposal.

5 ACKNOWLEDGEMENTS

The authors would like to acknowledge David Robinson, the author of the original Replay Gain proposal. We applaud his efforts towards resolving a very common problem with consistent music playback among large music libraries on personal computers.

6 REFERENCES

[1] http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/18_Metadata.Guide.pdf, last visited 2010-02-04

[2] Katz, B., Mastering Audio, The Art and The Science, 2nd edition. Focal Press, 2007

[3] Katz, B., "Integrated Approach to Metering, Monitoring, and Leveling Practices", AES Journal, Vol 48, No. 9, 2000 September

[4] http://www.pleasurizemusic.com/, last visited 2010-02-02

[5] Lund, T., Control of Loudness in Digital TV, NAB Proceedings 2006

[6] http://support.apple.com/kb/HT2425, last visited 2010-02-02

[7] http://www.replaygain.org, last visited 2010-02-02

[8] http://altosdesign.com/aacgain/, last visited 2010-03-05

[9] ITU-R Rec BS.1770, "Algorithms to measure audio programme loudness and true-peak audio level".

[10] Advanced Televisions Systems Committee, Inc.: "ATSC Recommend Practice: Techniques for Establishing and Maintaining Audio Loudness for Digital Television". Document A/85: 2009, 4 November 2009.

[11] http://www.cablelabs.com/specifications/MD-SP-VOD-CEP2.0-I03-100129.pdf (l.ast visited 2010-03-10)

[12] Soulodre, G. A., Evaluation of Objective Loudness Meters, 116th AES Convention, May 2004

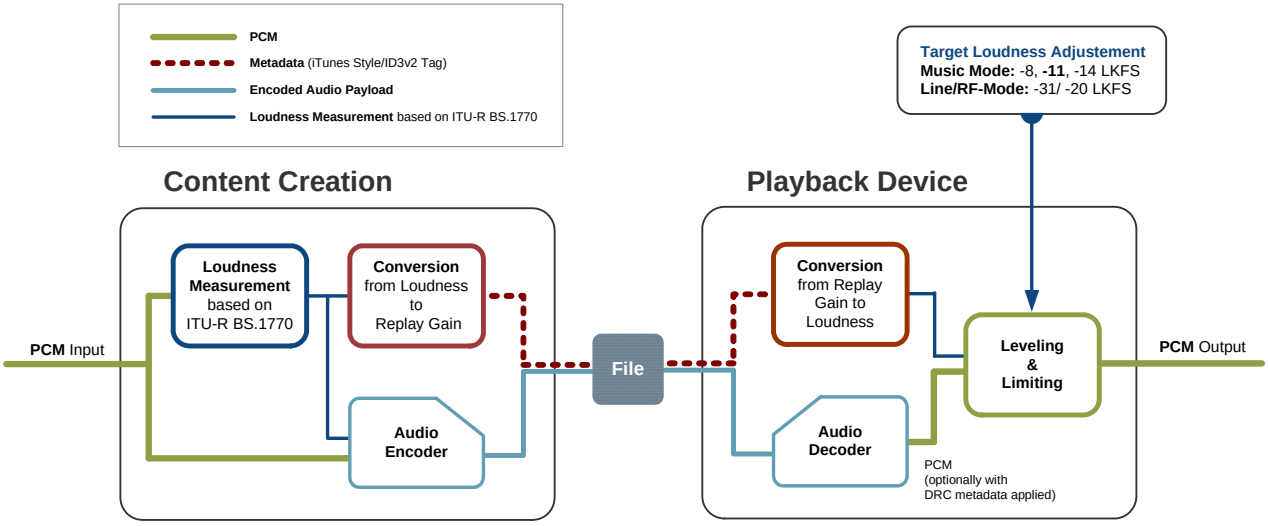

Figure 14 - Overview of the complete system proposal

```
////////////////////////////////////
// Leveling and Limiting pseudo control
// code in the playback device
////////////////////////////////////

// Determine target level (see 3.3.2):

Switch ( currentListeningEnvironment() ) {
    Case builtInSpeaker:
        TargetLevel = -8; // LKFS
        break;
    Case lineOut:
        TargetLevel = -31; // LKFS
        break;
    default:
        TargetLevel = -11; // LKFS;
        // -14 for high-end phones with
        // good amplifiers
        break;
}

// Determine reference level (see 3.3.4):

If ( replayGainPresent() ) {
    Switch ( currentPlaybackMode() ) {
        Case albumPlayback:
            If ( albumGainPresent() ) {
                ReplayGain = getAlbumGain();
            } else {
                ReplayGain = getTrackGain();
            }
            Break;
        default:
            If ( trackGainPresent() ) {
                ReplayGain = getTrackGain();
            } else {
                ReplayGain = getAlbumGain();
            }
            Break;
    }
    ReferenceLevel =
        -ReplayGain-18; // see 3.2.2.2
} else {
    If ( levelInfoInPayload() )
        ReferenceLevel =
            getLevelInfoFromPayload();
    } else {
        // Default levels
        Switch ( currentMusicFormat ) {
            Case stereoMusic:
                ReferenceLevel = -11; //LKFS
                Break;
            default:
                ReferenceLevel = -23; //LKFS
                Break;
        }
    }
}

// The following signal processing
// functions may be combined in a single
// gain-application-stage. The pseudo-code
// only provides a conceptual view.

// Apply dynamic range compression
// if available(see 3.3.3)

If ( metadataAvailable() &&
        TargetLevel >= -20) {
    applyMetadata(RF_MODE);
    // e.g. AC-3 or E AC-3
    ReferenceLevel = -20; // LKFS
 // data is leveled to -20 LKFS in RF-mode
}

// Apply final gain/limiter

Gain = TargetLevel - ReferenceLevel;
applyGain();

If ( Gain > 0 ) {
    applyLimiterAndGain();
}
```

7 APPENDIX A – DECODER OPERATING MODES

AC-3 & E AC-3 Decoder - Line Mode Operation:
Line Mode operation generally applies to the baseband line level outputs from two-channel set-top decoders, two-channel digital televisions and multichannel Home Theatre decoders. It is important to note that Line Mode operation is a requirement for all digital set top boxes that have analog baseband (line level) outputs. With respect to consumer type applications, a decoder's outputs operating in this mode will typically be connected to a much higher quality sound reproduction system than that found in a typical television set. In this mode, dialogue normalization is enabled and applied in the decoder at all times. Furthermore, in this mode the normalized level of dialogue is reproduced at a level of -31 LKFS, but **ONLY** when the transmitted *dialnorm* value has been correctly adjusted/provisioned for the particular program. In general, with the reproduced dialogue level at –31 LKFS, this mode allows wide dynamic range programming to be reproduced without any peak limiting and/or compression applied as may be intended by the original program producers. And since the AC-3 and E AC-3 systems can provide more than 100dB of dynamic range, there is no technical reason to encode the dialogue peaks at or near 100% as is commonly practiced in analog television broadcasts. This allows the AC-3 and E AC-3 systems to meet one of its goals of being able to deliver high impact cinema type sound to the digital subscriber's living room.

AC-3 and E AC-3 Decoder - RF Mode Operation:
RF Mode is intended for products such as terrestrial, cable and satellite set-top boxes that generate a monophonic and/or downmixed signal for transmission via the channel 3/4 remodulated output that feeds the RF (antenna) input of a television set. This mode was specifically designed to match the average reproduced dialogue levels and dynamic range of digital sources to those of existing analog sources such as NTSC and analog cable TV broadcasts. In this operating mode dialogue normalization is enabled and applied in the decoder at all times. However, the dialogue level in this mode is reproduced at a level of -20 LKFS **ONLY** when the transmitted *dialnorm* value is valid for a particular program. In this mode, AC-3 and E AC-3 decoders introduce an +11 dB gain shift and thus the maximum possible peak to dialog level ratio is reduced by 11 dB. This is achieved by compression and limiting internal to the AC-3 and E AC-3 decoders (but is calculated in the encoder). It is important to note that digital set top boxes which include an RF modulator are required to provide and default to RF Mode operation.

Figure 15 compares the reference loudness level relationships in a decoder operating in Line, RF & and the newly defined Portable operating modes. Notice the reproduced loudness level and dynamic range available each mode.

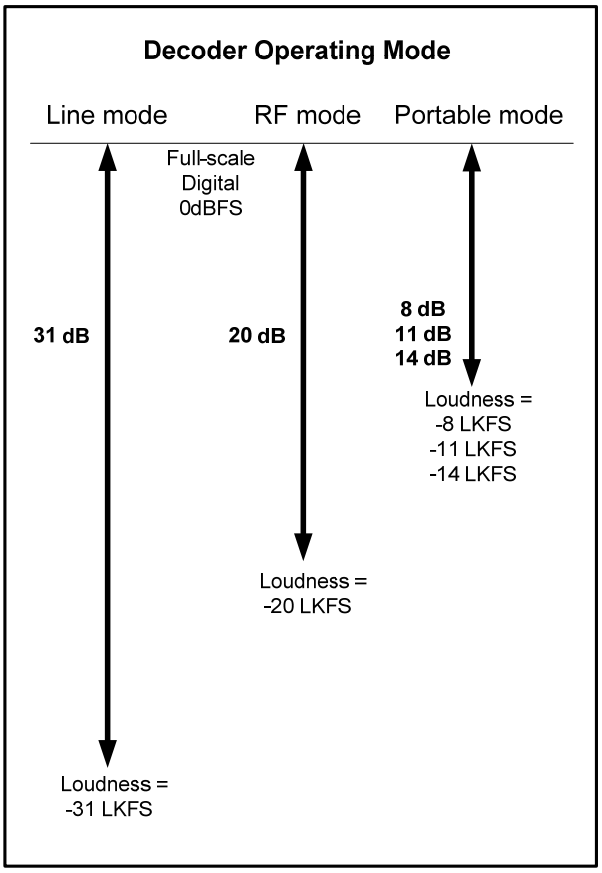

Figure 15 - Decoder Operating Mode(s) & Target Reference Levels

8 APPENDIX B – ITUNES-STYLE REPLAY GAIN METADATA

Below is an example for the Replay Gain compatible iTunes metadata to further illustrate the proposed syntax:

```
0000:0000 00 00 02 16 69 6c 73 74 00 00 00 72 2d 2d 2d 2d ....ilst...r----
0000:0010 00 00 00 28 6d 65 61 6e 00 00 00 01 6f 72 67 2e ...(mean....org.
0000:0020 68 79 64 72 6f 67 65 6e 61 75 64 69 6f 2e 72 65 hydrogenaudio.re
0000:0030 70 6c 61 79 67 61 69 6e 00 00 00 26 6e 61 6d 65 playgain...&name
0000:0040 00 00 00 01 72 65 70 6c 61 79 67 61 69 6e 5f 6f ....replaygain_o
0000:0050 72 69 67 69 6e 61 74 6f 72 5f 63 6f 64 65 00 00 riginator_code..
0000:0060 00 1c 64 61 74 61 00 00 00 01 00 00 00 00 30 31 ..data........01
0000:0070 31 30 31 31 30 30 30 30 30 30 00 00 00 69 2d 2d 1011000000...i--
0000:0080 2d 2d 00 00 00 28 6d 65 61 6e 00 00 00 01 6f 72 --...(mean....or
0000:0090 67 2e 68 79 64 72 6f 67 65 6e 61 75 64 69 6f 2e g.hydrogenaudio.
0000:00a0 72 65 70 6c 61 79 67 61 69 6e 00 00 00 21 6e 61 replaygain...!na
0000:00b0 6d 65 00 00 00 01 72 65 70 6c 61 79 67 61 69 6e me....replaygain
0000:00c0 5f 74 72 61 63 6b 5f 67 61 69 6e 00 00 00 18 64 _track_gain....d
0000:00d0 61 74 61 00 00 00 01 00 00 00 00 2d 35 2e 35 38 ata........-5.58
0000:00e0 20 64 42 00 00 00 69 2d 2d 2d 2d 00 00 00 28 6d  dB...i----...(m
0000:00f0 65 61 6e 00 00 00 01 6f 72 67 2e 68 79 64 72 6f ean....org.hydro
0000:0100 67 65 6e 61 75 64 69 6f 2e 72 65 70 6c 61 79 67 genaudio.replayg
0000:0110 61 69 6e 00 00 00 21 6e 61 6d 65 00 00 00 01 72 ain...!name....r
0000:0120 65 70 6c 61 79 67 61 69 6e 5f 74 72 61 63 6b 5f eplaygain_track_
0000:0130 70 65 61 6b 00 00 00 18 64 61 74 61 00 00 00 01 peak....data....
0000:0140 00 00 00 00 30 2e 39 36 33 32 31 32 00 00 00 28 ....0.963212...(
0000:0150 6d 65 61 6e 00 00 00 01 6f 72 67 2e 68 79 64 72 mean....org.hydr
0000:0160 6f 67 65 6e 61 75 64 69 6f 2e 72 65 70 6c 61 79 ogenaudio.replay
0000:0170 67 61 69 6e 00 00 00 21 6e 61 6d 65 00 00 00 01 gain...!name....
0000:0180 72 65 70 6c 61 79 67 61 69 6e 5f 61 6c 62 75 6d replaygain_album
0000:0190 5f 67 61 69 6e 00 00 00 18 64 61 74 61 00 00 00 _gain....data...
0000:01a0 01 00 00 00 00 2d 34 2e 35 38 20 64 42 00 00 00 .....-4.58 dB...
0000:01b0 69 2d 2d 2d 2d 00 00 00 28 6d 65 61 6e 00 00 00 i----...(mean...
0000:01c0 01 6f 72 67 2e 68 79 64 72 6f 67 65 6e 61 75 64 .org.hydrogenaud
0000:01d0 69 6f 2e 72 65 70 6c 61 79 67 61 69 6e 00 00 00 io.replaygain...
0000:01e0 21 6e 61 6d 65 00 00 00 01 72 65 70 6c 61 79 67 !name....replayg
0000:01f0 61 69 6e 5f 61 6c 62 75 6d 5f 70 65 61 6b 00 00 ain_album_peak..
0000:0200 00 18 64 61 74 61 00 00 00 01 00 00 00 00 30 2e ..data........0.
0000:0210 39 37 33 32 31 32                              973212
```