Audio Engineering Society

Convention Paper 8588

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

Implementation and Evaluation of Autonomous Multi-track Fader Control

Stuart Mansbridge[1], Saoirse Finn[1] and Joshua D. Reiss[1]

[1] *Centre for Digital Music, Queen Mary University of London, Mile End Road, London, E1 4NS, UK*
stuart.mansbridge@eecs.qmul.ac.uk
saoirse.finn@eecs.qmul.ac.uk
josh.reiss@eecs.qmul.ac.uk

ABSTRACT

A new approach to the autonomous control of faders for multi-track audio mixing is presented. The algorithm is designed to generate an automatic sound mix from an arbitrary number of monaural or stereo audio tracks of any sample rate, and to be suitable for both live and post-production use. Mixing levels are determined by the use of the EBU R-128 loudness measure, with a cross-adaptive process to bring each track to a time-varying average. A hysteresis loudness gate and selective smoothing prevents the adjustment of intentional dynamics in the music. Real-time and off-line software implementations have been created. Subjective evaluation is provided in the form of listening tests, where the method is compared against the results of a human mix and a previous automatic fader implementation.

1. INTRODUCTION

Producing a balanced audio mixture from multi-track content requires the considered choice of fader levels. Previous proposals for the automation of this procedure suggest either a machine-learning method [1][2], or the extraction of perceptual attributes (specifically loudness) to allow the emulation of real-time decisions made by a sound engineer [3]. This paper provides a description, analysis and evaluation of a new flexible implementation of the latter approach. The basis of the method is to achieve optimal inter-channel intelligibility, with the assumption that this is achieved by the adjustment of all inputs to a dynamic average perceptual loudness. The focus throughout this paper is on the development of a new detailed, versatile and reliable real-time, low latency algorithm.

The chosen method for loudness estimation is from the EBU R-128 recommendation [4], which calculates a loudness value of unit LUFS (equivalent to dBFS) by a mean-square energy calculation over a frame of audio samples, using two bi-quadratic IIR filters to provide a frequency weighting for the psycho-acoustic model. This has the advantage over the method utilised in [3] in being an officially recognised measure, independent of listening level, and being significantly more efficient to process.

The algorithm has been designed to be flexible regarding inputs and suitable for live or post-processing use. The program will adapt to process any number of mono or stereo tracks, and at any sample rate. Filter coefficients for the loudness estimation are calculated pre-mix, to ensure the correct frequency response for the given sample rate. The user also has the option to specify channels to which a gain boost will be applied, for example in the case of a lead vocal that needs to be placed above the rest of the mix.

In addition to fader control, a pre-amp gain control is included in the signal chain to normalise the level of each input track.

1.1. Exponential moving average filter

Due to the sample-based and short-frame processing employed in the algorithm, an efficient and reliable long-term average measure is necessary throughout to produce useful and smoothly varying data variables. Exponential moving average (EMA) filters are used extensively to fulfil this role.

The EMA filter is a 1st order IIR filter, with the following difference equation:

$$y[n] = (1-\propto) \cdot x[n] + \propto \cdot\, y[n-1] \qquad (1\ )$$

The factor $\propto$ determines the degree of filtering between adjacent samples; the higher the value of $\propto$ the less the level of decay. In terms of signal processing all that is required for the calculation is the storage of the preceding average value $y[n-1]$, two multiplies and an addition.

A value of $\propto$ equivalent to an equally weighted simple moving average (SMA) can be obtained by equating the average values of the step functions of both filters, where the step function is defined as:

$$x[n] = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases} \qquad (2\ )$$

The SMA filter difference equation is:

$$y[n] = \frac{1}{W} \sum_{i=0}^{W-1} x[n-i] \qquad (3\ )$$

Where W is the window size in samples. The step response for the SMA filter is therefore:

$$y[n] = \begin{cases} n/W & n < W \\ 1 & n \geq W \end{cases} \qquad (4\ )$$

While the step response of the EMA filter is:

$$y[n] = 1 - \propto^{n} \qquad n \geq 0 \qquad (5\ )$$

The average values, over W samples, of the step responses from Equations 4 and 5 are calculated and set equal:

$$\frac{W-1}{2W} = 1 - \frac{1-\propto^{W}}{W(1-\propto)} \qquad (6\ )$$

Resulting in the final approximation:

$$\propto \sim \frac{W-1}{W+1} \qquad (7\ )$$

Weighting can then be applied by increasing or decreasing this value appropriately.

Finally, to ensure a constant EMA filter response at any sample rate, the alpha value should be allowed to adjust while keeping a fixed time constant τ_W, where:

$$\tau_W = {}^{W}\!/_{f_s} \qquad (8\ )$$

1.2. Notation

Throughout this document some constant notation is used. A mix is considered to contain M channels, with individual channel number *m*. The current time instance or frame number is indicated by *n*.

1.3. Full system

A block diagram depicting the full system is given in Figure 1. $La1_m[n]$ and $La2_m[n]$ are smoothed loudness values, derived from the same loudness estimation but passed through different EMA filters for use in the pre-amp and fader value calculations. $G_m[n]$ is the pre-amplifier gain and $F_m[n]$ the fader gain.

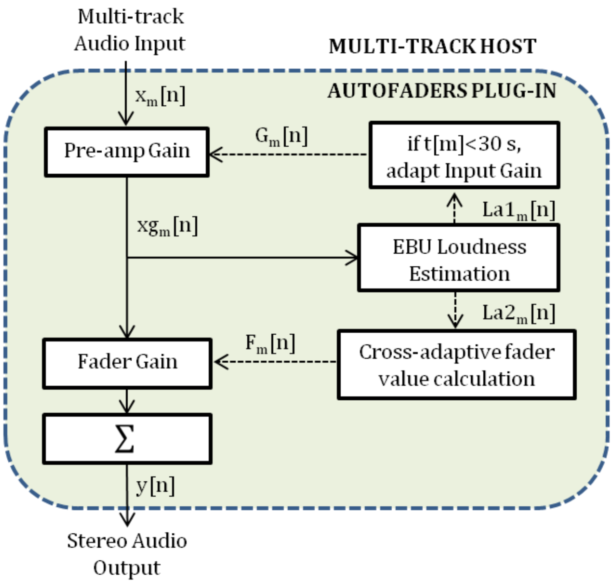


Figure 1: System block diagram

The overall system equation is:

$$y[n] = \sum_{m=1}^{M} x_m[n] \cdot G_m[n] \cdot F_m[n] \qquad (9)$$

2. PRE-AMPLIFIERS

To optimise the dynamic range of each track there is a requirement for each to have approximately normalised levels, where the maximum amplitude level before distortion is 1, or 0dBFS. In a traditional live sound mixing scenario this is achieved by analysing the peak meters and adjusting the pre-amplifier gain, whilst allowing sufficient head room so spikes in the input signal do not distort [5]. This also ensures the good practice of avoiding excessive gain being required from the faders [6]. The algorithm is designed to emulate this process, but using the loudness value (passed through a slowly decaying smoothing filter) to determine what gain is required. Whilst mixing in the digital domain doesn’t present the same constraints on signal level as with an analogue mixer, the process is necessary for automatic mixing primarily so that fixed loudness thresholds can be effectively employed.

As a part of the signal chain each input track $x_m[n]$ is multiplied by pre-amp gain $G_m[n]$, set initially to unity. The gain value for each track is then adjusted according to its loudness over a 30 second period of activity, when the track is deemed to be above the absolute loudness threshold of -70LUFS. This means that the system will wait until a track has become active and its loudness value has become relevant before adjusting the gain.

The automatic gain can be used in two ways; either considered to be equivalent to a sound check process in which the pre-amp levels are set during a preliminary run, or allowed to adjust gain in real-time as a part of the mix. While it is preferable to have the gain levels fixed at the start of the mix, the faders will constantly adapt to the new incoming signal level and any detrimental effect on the mix is minimal. Within the setup period, gain is adjusted according to the following formulae:

Gain change factor: $\gamma = 0.005.$

$$G_m[n] = \begin{cases} (1+\gamma).G_m[n-1] \,, & -70 < La1_m[n] < -20 \\ (1-\gamma).G_m[n-1] \,, & La1_m[n] > -10 \\ G_m[n-1] \,, & otherwise \end{cases} \qquad (10)$$

The gain is then applied to each input signal:

$$x_m[n] = G_m[n].x_m[n] \qquad (11)$$

3. ALGORITHM

The basis of the algorithm is the assumption made in [3], that each track should be brought to a dynamic average perceptual loudness in order to achieve optimal inter-channel intelligibility.

The cross-adaptive processing is controlled by a noise gate with hysteresis, which determines from a track’s loudness whether it can be considered to be active. When active the track enters the cross-adaptive processing stage. When inactive all values relating to that track are kept stationary, preventing periods of ambient noise or intentionally low level sounds from being amplified and affecting the system.

Furthermore, a fader level which changes too rapidly will smooth desirable variations in dynamic range and thus act in a similar fashion to a dynamic range compressor. Exponential moving average smoothing filters are used on both loudness and fader values, with the frame number of each track entry point stored to commence the process and avoid any undesirable fade-in effects.

Additionally, a gain boost can be applied to chosen tracks to be placed above the rest of the mix.

3.1. Loudness Estimation

A function is required to provide the system with values representing the perceptual loudness level of each track. For the purposes of a real-time automatic mixing algorithm, the loudness estimate needs to be accurate, reactive and efficient to process. The EBU loudness R-128 standard was chosen to satisfy the criteria. Due to the long window size required to gain a loudness measure the method was adapted for use in real-time processing to allow for regular updates in loudness value.

3.1.1. Calculation

The calculation of the EBU loudness is essentially an energy measurement on a filtered signal. Pre and RLB filters process the signal to provide a psycho-acoustic model to emulate the frequency response of the human ear, as portrayed by the ISO226 contours [7]. A mean square energy calculation is then performed on the filtered signal.

The estimation is based on the short-term loudness measure, which requires the processing of three seconds of audio. Due to the impracticality of processing that number of samples in real-time, and the inherent length of time between updates, two versions of the loudness estimation were created. The first processes an energy calculation over three seconds of audio as specified in EBU documentation [4]. The second is a sample-based version with the mean-square energy accumulated at the end of every frame. An exponential moving average filter is applied to the signal energy to provide an estimation of the energy over a 3 second period, which updates with every new host frame. For stereo tracks the system sums the energy calculation of both channels.

To provide the final loudness reading, the energy is then converted to a dB scale, and a correction constant of -0.691dB applied to account for the non unity gain of the combined frequency response at 1 kHz. The overall equation is shown below:

$$L_m[n] = 0.691 * 10log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} xg_m^2[n]\right) \qquad (12\ )$$

The unit of loudness is LU, and acts as an equivalent measure to dB. Ultimately loudness is measured with units LUFS (Loudness Unit with reference to digital Full scale), where 0 LUFS is the maximum possible level.

3.1.2. Filter coefficients from sample rate

To maintain the pre and RLB filter frequency responses the filter coefficients need to be calculated for different sample rates.

The general bi-quadratic formula in the Laplace domain is defined below:

$$H(s) = \frac{V_H s^2 + V_B \frac{\omega}{Q} s + V_L \omega^2}{s^2 + \frac{\omega}{Q} s + \omega^2} \qquad (13\ )$$

When converted into the z-domain using the bilinear transform and frequency transformation ($s = \frac{z-1}{z+1}$ and $\omega \rightarrow \Omega = \tan\left(\pi \frac{f_c}{f_s}\right)$), this yields the transfer function:

$$H(z) = \frac{\left(V_L\Omega^2 + V_B \Omega/_Q + V_H\right) + 2\left(V_L\Omega^2 - V_H\right)z^{-1} + \left(V_L\Omega^2 - V_B \Omega/_Q + V_H\right)z^{-2}}{\left(\Omega^2 + \Omega/_Q + 1\right) + 2\left(\Omega^2 - 1\right)z^{-1} + \left(\Omega^2 - \Omega/_Q + 1\right)z^{-2}}$$

(14)

Filter coefficients are provided by the EBU for 48kHz, and so the values of the constants can be computed and hard-coded: $V_H \approx 1.58$, $V_B \approx 1.26, V_B = 1$, $Q \approx 0.71$ and $f_c \approx 1681.97$.

All that remains is for angular frequency Ω to be calculated from the given sample rate. The coefficients of the filter from Equation 14 can then be calculated and stored at the start of every new mix. Figure 2 shows the equal frequency response of the combined filters at common sample rates.

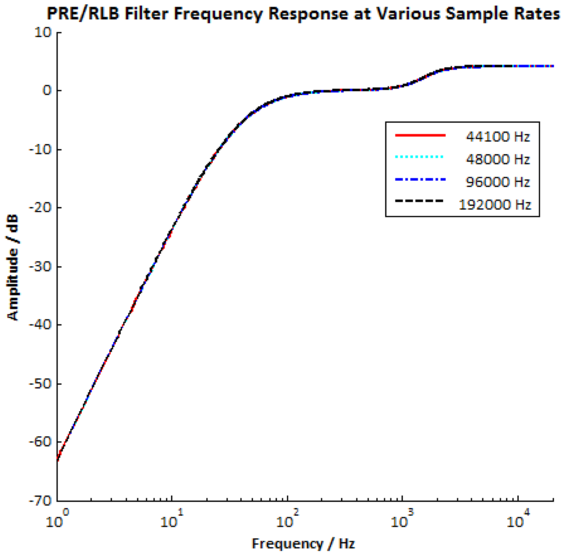


Figure 2: Combined Pre/RLB filter frequency response at various sample rates. The different sample rates give virtually identical frequency response and hence are not easily distinguished.

3.2. Hysteresis Noise Gate

Critical to the operation of the algorithm is the correct determination of whether or not each track is active, without which low level ambient noise will be boosted. Additionally, the algorithm is based on the principle that a track must be active for it to contribute to the cross adaptive calculation, and if not all values are kept at their previous level.

A noise gate is required to decide for every frame whether or not each track is in an active state. Hysteresis thresholds at -25 and -30 LUFS are used to help prevent excessive switching of states. After crossing one threshold the gate maintains its current state until the loudness level moves below or above the other threshold. This is a concept employed in the Schmitt trigger [8], and is portrayed in Figure 3.

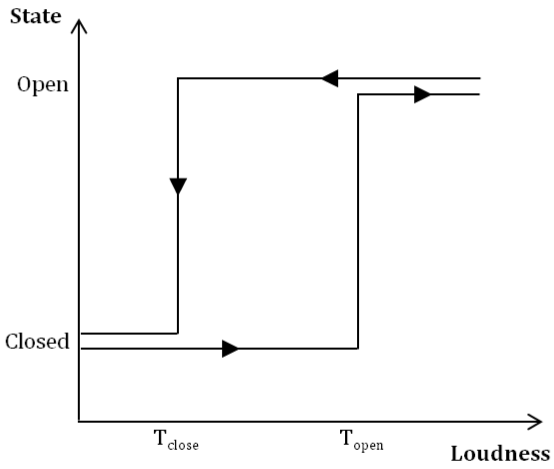


Figure 3: Noise gate hysteresis operation

In addition, the frame at which a track first becomes active is monitored. Any smoothing of variables occurs after this point, preventing initial periods of inactivity from affecting the variable values and the speed of reactivity of the faders.

3.3. Average Loudness

A count of currently active channels is kept per frame, as determined by the noise gate. The total absolute loudness of all tracks is summed and divided by the channel count, to produce a dynamic target loudness that allows for intended fluctuations in the overall mix signal level. An EMA filter is used to provide a smoothly varying value.

3.4. Fader calculation

Fader values are then calculated as a ratio of the track loudness to the average loudness:

$$F_m[n] = 10^{\frac{Lav[n]-La2_m[n]}{20}} \qquad (15\ )$$

As with other variables, the fader values are smoothed using an EMA filter.

3.5. Lead track

In some cases there is a requirement for one or more tracks to be given a greater presence in the mix. This is most commonly beneficial for placing a vocal track above backing instruments.

This functionality is provided within the software. A gain increase in dB can be chosen, and tracks selected to apply the gain. The application of this extra gain to the faders is smoothed to prevent a quick change in volume if a track is selected during the mix. The equation is shown below, where *B* is the chosen gain change in dB.

$$F_m[n] = F_m[n] \times 10^{\left(\frac{B}{20}\right)} \qquad (16\ )$$

4. SOFTWARE

Two real-time software implementations of the algorithm have been created in C++, one using long frame processing and the other using the sample-based approach. Both operate on a frame by frame basis using a host/plug-in structure, where consecutive frames of chosen length are provided by the host. Buffering of audio data where necessary occurs within the plug-in.

A screenshot of the user interface is shown in Figure 4. Fader levels are updated regularly and displayed on the interface. Also included are loudness meters for each channel, pre-amp gain values, and the option to apply a gain boost for particular channels. Track names can be added for reference. An output master gain control automatically adjusts the overall mix level to prevent clipping from occurring.

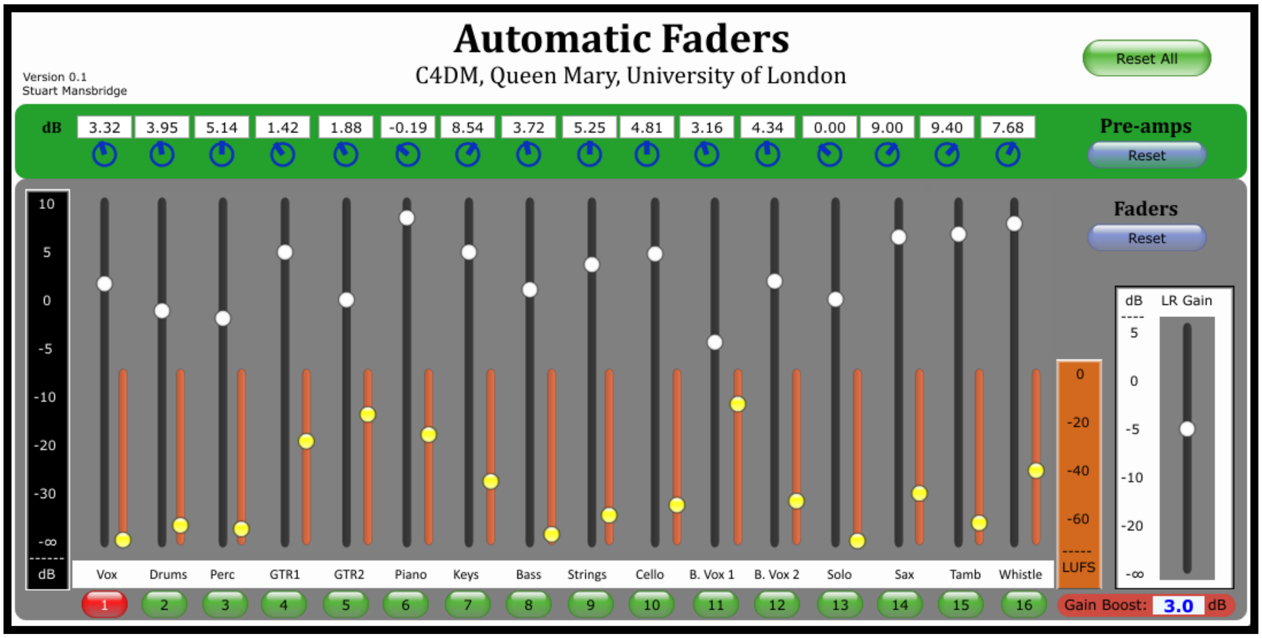


Figure 4: Graphical user interface for the autonomous fader control system

5. LISTENING TEST

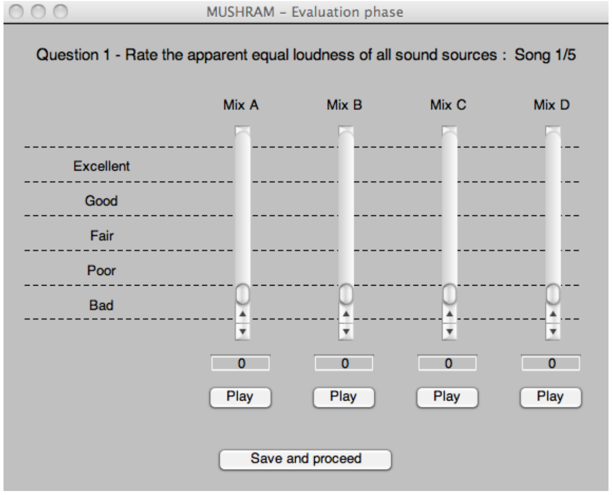
A listening test was carried out to evaluate the automatic faders system for a variety of multi-track content.

5.1. Method

Each experiment consisted of blind comparisons of either four or five alternative mixes: one with fixed fader levels (‘Standard’), a manually produced mix by a sound engineer (‘Manual’), a mix created using a previous automatic faders implementation from [3] (‘EPG’), a mix by the automatic faders implementation in this paper (‘Auto’) and, for the second experiment only, an automatic mix with a 2dB gain boost applied to the lead vocal track (‘2dB’). The audio material used was taken from a wide variety of musical genres, to demonstrate the application of the system under different circumstances. The number of input sources per mix ranged between 5 and 8, due to the limit on the number of tracks from the previous implementation. The overall level of each mix was normalised to ensure a fair comparison.

The test was based on the MUSHRA framework [9] in which each audio sample is rated on a scale of 0-100, split up into five descriptors: ‘Bad’, ‘Poor’, ‘Fair’, ‘Good’ and ‘Excellent’. Due to the subjective nature of the tests it was decided that the use of a fixed reference would be inappropriate, and thus overall it can be considered a semantic differential test. The ‘Unmixed’ and ‘Manual’ samples were taken to be anchors of low and high quality. To ensure the use of the whole of the scale, participants were requested to rate the best mix between 80-100, and the worst mix between 0-20. A similar evaluation procedure was used in [10].

The experiments were performed using the MUSHRAM [11] interface in Matlab, and took approximately 30-45 minutes to complete. Control measures included the use of a listening room, the same headphones and a constant output level. A screenshot of the listening test interface is shown in Figure 5.



Question 1 - Rate the apparent equal loudness of all sound sources : Song 1/5

Figure 5: Screenshot of listening test user interface

In total 16 people participated in the test. Before commencing, subjects were required to answer a few questions about themselves. The results are displayed in Table 1.

Gender	**Male**	14
	Female	2
Listening test experience?	**No**	2
	Some	4
	Yes	10
Hearing impairment?	**No**	15
	Yes	1

Table 1: Results of preliminary questions to test subjects

The table shows that the vast majority of subjects had at least some experience in critically analysing audio and no hearing impairment. This information was only used to gain an insight into the background of the participants and no individual’s results were excluded.

Two experiments were run, the first asking to rate the apparent equal loudness of the input sources, and the second to rate the overall quality of the mix.

5.2. Results

5.2.1. Question 1

For the first experiment participants were asked to rate each mix on the apparent equal loudness of all sound sources. The question was chosen because both of the automatic mixing systems under test make the assumption that each source should be equally audible in the mix. Therefore an indication would be received on how well the automatic mixes achieve this goal, and also whether the 'Manual' mix follows the same assumption.

Five songs of different genres were selected for analysis. Figure 6 shows the mean with error bars displaying the 85% confidence intervals using the T-distribution.

It can be seen that the 'Manual' mix gets the most consistently high rating overall, but is closely matched by the 'Auto' mix, which out-performs it on one song. The 'EPG' mix also rates highest for one song, but shows poor consistency overall. The 'Unmixed' mix gives typically poor results, with the exception of one song which was reasonably well normalised to begin with. Confidence intervals are largely uniform throughout, with the exception of a few mixes which were rated equally poorly by the majority of participants.

5.2.2. Question 2

In the second experiment participants were asked to rate each mix on the overall quality of the mix. The responses are therefore entirely and necessarily subjective.

Nine songs from different genres were selected for analysis, each with an additional '2dB' mix. Figure 7 shows the mean and 85% confidence intervals.

The 'Auto' mix is consistently highly rated. The '2dB' mix is more variable, providing higher ratings along with lower ones, but with a marginally higher overall mean. The other mixes are far less consistent, the 'Manual' mix generally rates reasonably highly, and the 'EPG' and 'Unmixed' mixes have a wide range of results but tend to rate poorly. Confidence intervals again show little variation with a few exceptions.

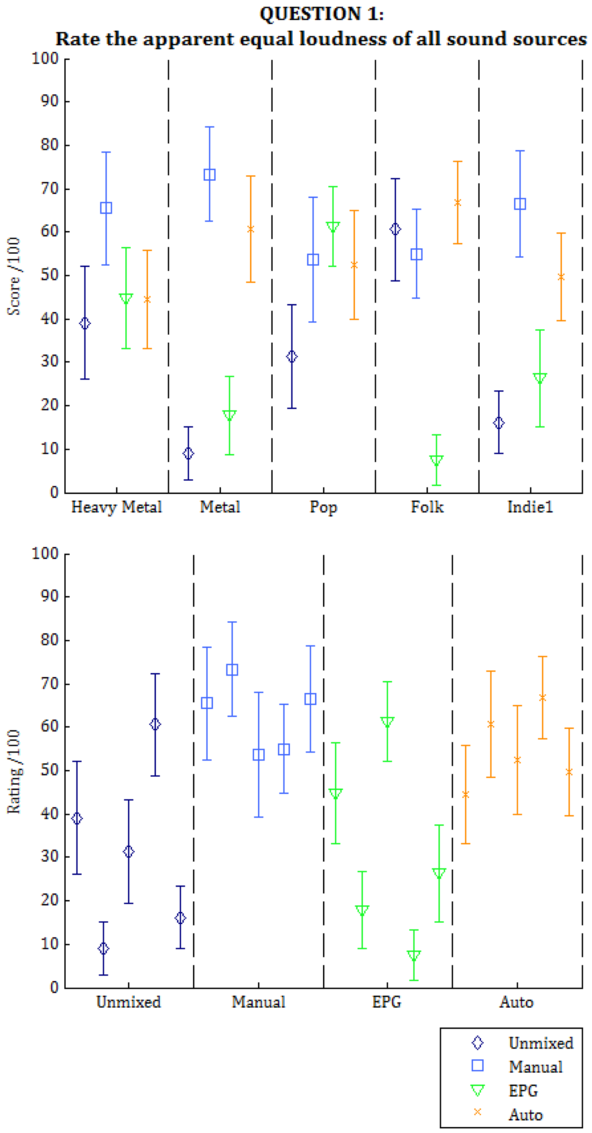


Figure 6: Mean and 85% confidence interval results for Question 1, arranged by song and mix type.

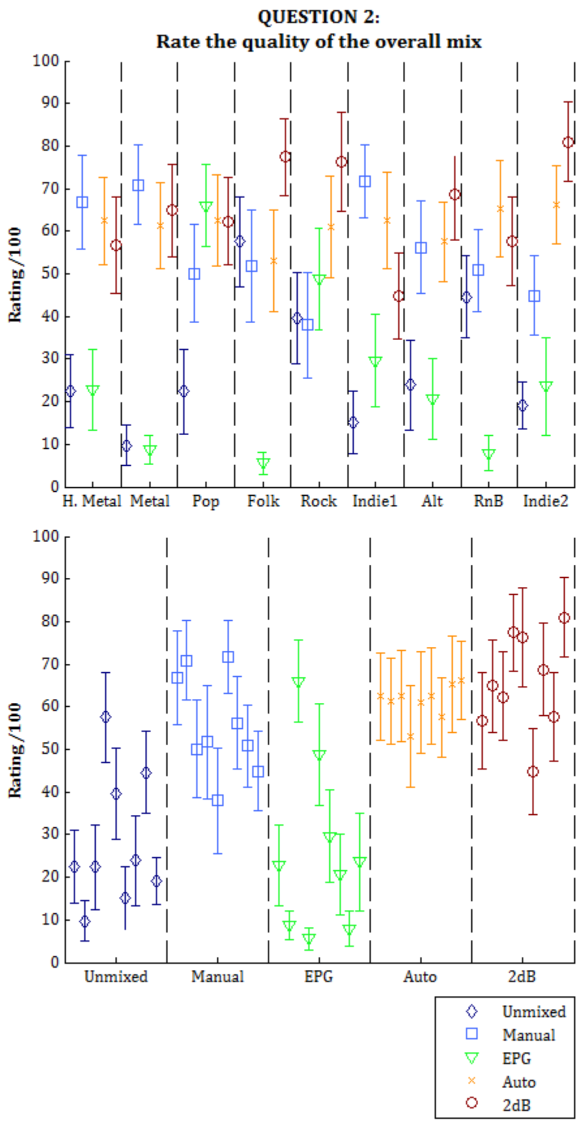


Figure 7: Mean and 85% confidence interval results for Question 2, arranged by song and mix type.

Averaged mean and median results for all songs are displayed in Figure 8 for each mix type, and for both experiments. These give a clearer depiction of the overall performance of each mix type across all music genres, where the median is less influenced by anomalous results. The 'Manual' mix can be seen to perform best overall for Question 1, and the '2dB' mix for Question 2.

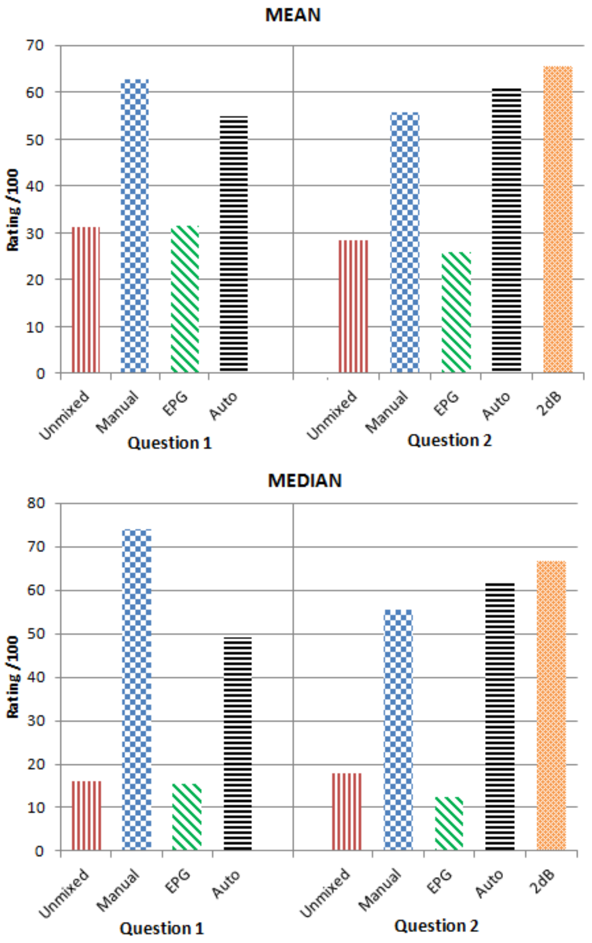


Figure 8: Overall mean and median results for both experiments.

5.3. Evaluation

Overall the results are very positive for the new automatic faders implementation. For Question 1 the automatic mix closely matched the 'Manual' mix in terms of equal loudness of input sources. For Question 2, the consistency of the automatic mix's performance regardless of song is particularly important. This indicates success with making a versatile multi-purpose algorithm. Furthermore, the ‘2dB’ mix on occasion provides a significant further improvement. This matches the expectation that a vocal gain boost would be advantageous in some cases.

Despite the success of the listening test it is recognised by the authors that further and better controlled testing needs to be performed. There are some questionable outcomes from the test, for example the relatively poor performance of the ‘Manual’ mix in Question 2. It was assumed beforehand that the ‘Manual’ mix would rate highly throughout, and so there is an implication that for some of the songs a better mix could have been created. Alternatively it could be due to it being a subjective test with a relatively small number of participants. Overall, results should be considered to be indicative but not definitive.

A number of limitations of the system have been established. From the results of Question 1 it is evident that the assumption of equal loudness for all sources is fair, but there are exceptional cases where this is not the case. For example, a song containing a single tambourine track was discarded from the results after having it at the same loudness as other tracks meant it dominated the mix. A tambourine is particularly overpowering due to the wide range of frequencies produced. It is proposed that all individual percussive instruments should be automatically identified, and then mixed down into a single percussion track prior to the automatic mix. Secondly, the manual application of a gain boost allows an aspect of user input into an automatic mix. There is potential for vocal tracks to be identified and boosted automatically, at a gain factor determined by the number of backing instrument tracks. Finally, further research should be done into the use of variable loudness thresholds. Currently the fixed loudness threshold method is heavily reliant on the correct normalisation of all input sources by the automatic pre-amplifier gain.

A major unexpected outcome of the listening test was the poor performance of the previous fader implementation in all but a few cases in both experiments. Although source code was unavailable for the demonstration program used to generate the ‘EPG’ mixes for analysis of the algorithm, it is possible to speculate on reasons for this from background knowledge and the information in [3]. Firstly, the demonstration was designed largely as a proof of concept, and hence may not be an optimal implementation of the algorithm, and it was tailored for use on a small number of multi-tracks. In comparison, the new approach has been designed to work in a variety of circumstances. Secondly, it is unknown to the authors whether the previous method adapts to the number of inputs present, or whether a full count of eight is assumed. This has the potential to cause a miscalculation of the target loudness and directly affect all fader positions. A casual look at the results seems to disprove this however, as both its highest and lowest ratings in Question 2 were created from 8-track mixes. Finally, doubts had been raised in [3] about the suitability of the loudness estimation function. In particular the estimation is based on fixed input levels, and therefore the absence of normalised input sources would lead to inaccuracies in the loudness estimation.

6. CONCLUSION

A new approach to the automatic fader control for mixing multi-track audio has been described, implemented, and evaluated with comparison to a number of manual and automatic mixes. The algorithm has been designed to be flexible regarding the number of inputs of monaural or stereo type, and suitable for live or off-line applications. The new approach scored highly in the listening tests over a wide range of musical genres.

Future work should concentrate on the appropriate classification and amalgamation of percussive tracks to prevent them being adjusted independently and dominating the mix. Automatic gain boost of vocal tracks, alternative loudness models, and variable loudness thresholds are also potential areas for further research.

7. ACKNOWLEDGEMENTS

This work was supported by the FP7 European Project DigiBIC. The authors would also like to thank all volunteers from Queen Mary, University of London and elsewhere who participated in the listening tests.

8. REFERENCES

[1] Scott, J., and Kim, Y. E. "*Analysis of acoustic features for automated multi-track mixing*", in Proc. of International Society for Music Information Retrieval (ISMIR), 2011.

[2] Scott, J., and Kim, Y. E. "*Automatic Multi-track Mixing Using Linear Dynamical Stems*", in Proc. of the SMC 2011 – 8th Sound and Music Computing Conference, 2011.

[3] Perez-Gonzalez, E., and Reiss, J. D. "*Automatic Gain and Fader Control for Live Mixing*", in Proc. of IEEE WASPAA Workshop, p. 1-4, 18-21 October, 2009.

[4] International Telecommunication Union. Rec. ITU-R BS.1770-2, *"Algorithms to measure audio programme loudness and true-peak audio level"*. Geneva, 2011.

[5] P. White, *"20 Tips on Mixing"*, June 1998, http://www.soundonsound.com/sos/jun98/articles/20tips.html.

[6] M. Senior, "*Building the Raw Balance*", in Mixing Secrets for the Small Studio, 1st Ed. United Kingdom: Focal Press, 2011.

[7] International Organization for Standardization, *"Normal Equal-Loudness Level Contours"*, IS0226, 2003.

[8] M. Filanovsky, H. Baltes, *"CMOS Schmitt Trigger Briefs",* in IEEE Trans. Circuits Syst. I, Fundamental Theory and Applications 1994, Vol 41, No. 1, pp. 46-49

[9] International Telecommunication Union, *"Multiple Stimuli with Hidden Reference and Anchor",* ITU-R BS.1534-1, 2003.

[10] M. Zaunschirm, J. D. Reiss and A. Klapuri, *"A sub-band approach to musical transient modification"*, to appear in Computer Music Journal, 2012.

[11] E. Vincent, M. G. Jafari and M. D. Plumbley. *"Preliminary guidelines for subjective evaluation of audio source separation algorithms"*, in Proc. of the ICA Research Network International Workshop, pp. 93-96, 18-19 September 2006.