ESTIMATING THE LOUDNESS BALANCE OF MUSICAL MIXTURES USING AUDIO SOURCE SEPARATION

Dominic Ward, Hagen Wierstorf, Russell D. Mason[†]*, Mark D. Plumbley, Chris Hummersone*[†]

CVSSP / IoSR[†]
University of Surrey
Guildford, UK
dw0031@surrey.ac.uk

ABSTRACT

To assist with the development of intelligent mixing systems, it would be useful to be able to extract the loudness balance of sources in an existing musical mixture. The relative-to-mix loudness level of four instrument groups was predicted using the sources extracted by 12 audio source separation algorithms. The predictions were compared with the ground truth loudness data of the original unmixed stems obtained from a recent dataset involving 100 mixed songs. It was found that the best source separation system could predict the relative loudness of each instrument group with an average root-mean-square error of 1.2 LU, with superior performance obtained on vocals.

1. INTRODUCTION

One aspect of Intelligent Music Production (IMP) is the development of tools to assist and understand the operations of audio engineers, producers and musicians. By analysing both the multitrack recordings and the digital audio workstation settings of music producers, researchers are able to gain insight into music production practices and establish rules for intelligent mixing systems [1]. Loudness features in particular have received a great deal of interest in previous studies investigating the level-balancing and prioritisation of instruments within a mix [2–6], especially for addressing common assumptions of knowledge-driven automatic mixers. Much of this investigative work involves conducting mixing tasks with human engineers, allowing the experimenter to study mixing trends at the track level. However, such analyses generally require ground truth data such as the original multitrack recordings and mixer settings. An alternative but supplementary approach is to utilise audio source separation as an analytical tool for measuring features from existing mixtures. Such a device would be useful, for example, to compare the balance of specific instruments in commercial songs mixed by professional engineers.

In the audio source separation community, algorithms are typically assessed [7] using blind source separation (BSS) evaluation metrics, notably ‘BSS Eval’ [8], that quantify the amount of interference, distortion and artefacts present in the extracted source. However, it is difficult to relate the magnitude and relationships of these measures to the success of a source separation system when used to study IMP. As such, this paper takes an application-driven approach to system evaluation, with the focus being multitrack audio feature extraction.

The primary goal of this paper is to measure the accuracy of state-of-the-art audio source separation algorithms when used to extract the relative-to-mix loudness levels of four musical sources of a recent multitrack dataset involving 100 semi-professionally mixed songs. In addition to system evaluation, our initial analysis of 100 human mixes contributes to the growing knowledge of music production practices, e.g. [3], and has been made freely available for other IMP researchers to reproduce and expand upon.[1]

2. METHODOLOGY

2.1. Stimuli

The Demixing Secret Database (DSD100) is a set of 100 songs of different musical genres, with each song comprising four stereo sources (bass, drums, other and vocals) that sum to realistic mixtures [9]. Songs from this dataset comprise a mix of genres such as hip-hop and heavy metal, though the majority fall into rock and pop classes. DSD100 was used as both the training and testing data for the ‘MUS’ task of the 2016 Signal Separation Evaluation Campaign (SiSEC), where 23 systems were evaluated using BSS Eval metrics applied to the estimated sources. Algorithm proponents used half of these songs for model development, and 46 for testing (four of the remaining 50 songs submitted were excluded due to file corruption). The submission data, as used to evaluate the algorithms in this paper, were kindly provided by Fabian-Robert Stöter.[2] Of the submitted audio files, only the test set were analysed, converting mono files to stereo where needed.

2.2. Separation Algorithms

Of the 23 SiSEC separation algorithms, 12 targeted the extraction of all four sources from each mixture: CHA, GRA2,

[1] https://code.soundsoftware.ac.uk/hg/wimp17-ward-et-al.

[2] Audio excerpts and further information is available at https://www.sisec17.audiolabs-erlangen.de.

GRA3, KON, NUG1, NUG2, NUG3, NUG4, OZE, UHL1, UHL2 and UHL3. These 12 constituted the systems under test (see [9] for further details).

2.3. Measurements

Although different approaches to loudness estimation exist [10], the ITU-R BS.1770 procedure [11] is most commonly used to investigate the level-balancing paradigm for multitrack audio [3–6]. The loudness measurement takes the form of a sliding 400 ms window that integrates signal power following a moderate frequency weighting that reflects the frequency sensitivity of the ear. Absolute and relative gating of these short-term energy measurements is used to arrive at the final programme loudness. Note that programme loudness is expressed on a decibel scale in Loudness Units (LU), and that this algorithm, though useful for loudness matching, was not designed to predict subjective loudness ratios, e.g. sound A is twice as loud as sound B.

The loudness balance b was extracted for each song by measuring the relative-to-mix loudness levels:

$$b_i = L_i - L_{\text{mix}}, \tag{1}$$

where L_i and L_{mix} is the programme loudness of source i and of the mix, respectively. In addition to being perceptually motivated [4], making the levels relative to the loudness of the mix circumvents any error in the approximations introduced by an overall level offset, i.e. absolute levels are not deemed important here. For a given source i in song j, the error between the predicted balance $\hat{b}$ and the reference balance b was measured using

$$e_{i,j} = \hat{b}_{i,j} - b_{i,j}. \tag{2}$$

3. RESULTS

3.1. Human Balances

Figure 1 shows, for each source, the estimated probability distribution, with a supplementary boxplot overlay, of the relative-to-mix loudness levels across the 100 songs of the DSD100 dataset. Note that it is possible for individual sources to be measured as louder than the mix, as background elements can lower the integrated mix loudness, relative to the loudest components. It can be seen that the vocals show the least inter-song variation, as the mix engineers generally prioritise this instrument group in the mix. However, the estimated probability density of the vocals is indicative of bimodality, which may be attributed to differences across genres. This is reflected by the vocal levels in seven metal songs (white circles), which all cluster in the bottom 25% of the distribution. Additionally, the guitars (within ‘other’) tend to be placed slightly higher in the mix compared to the average for this genre. Although the bass instruments are generally the lowest in the mix, in agreement with [5], there is considerable inter-song variation and no obvious representative value, as reflected by the flatter probability distribution.

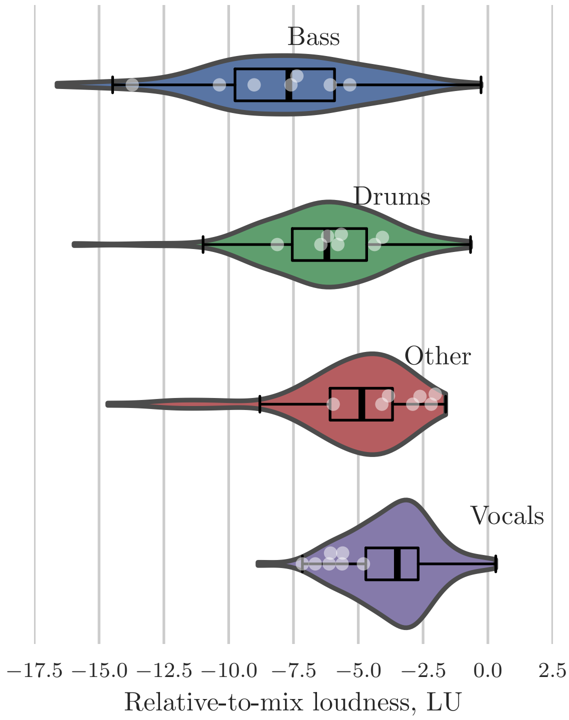

Figure 1: Violin plots, with boxplots overlaid, of the ITU-R BS.1770 relative-to-mix loudness levels measured on the entire DSD100 dataset (100 songs). The violins have been truncated to the range of the data. The black thick line of each boxplot shows the median of each distribution. The white circles highlight data points from seven metal songs.

For the pop and rock songs, which constitute the majority of DSD100 (62/100), the mean relative-to-mix loudness levels (and standard deviation; 95% confidence interval) of the bass, drums, other and vocals was -8.1 (SD 3.0; CI_{95} ±0.8), -6.5 (SD 2.5; CI_{95} ±0.6), -5.0 (SD 1.9; CI_{95} ±0.5) and -3.6 LU (SD 1.3; CI_{95} ±0.3), respectively. The relative level of the vocals is similar to the value of -3 LU reported in [3–5], although the remaining estimates are slightly higher. This might be due to a modified version of the loudness model used in those studies, and/or a smaller sample of songs.

3.2. System Evaluation

The 12 audio source separation algorithms were assessed in terms of loudness balance prediction. Boxplots of the prediction errors (across songs) by algorithm and source reflected a systematic interaction of these two variables. This is demonstrated in Figure 2 which shows the errors by source, for the worst (GRA3) and best (NUG3) algorithm in terms of the average song root-mean-square error (RMSE). NUG3 is most consistent across the four sources, while GRA3 shows

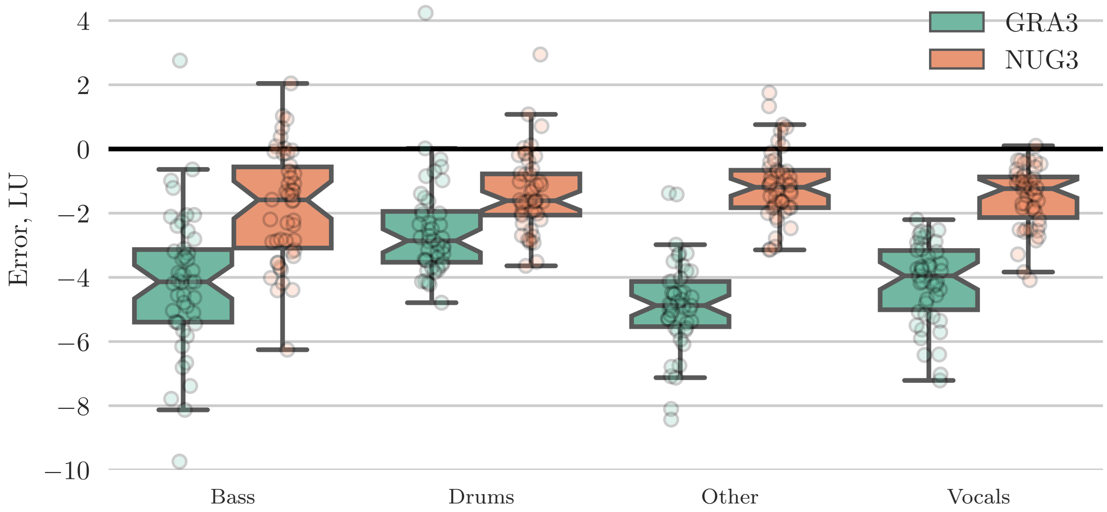

Figure 2: Boxplots, with data points overlaid, of the relative-to-mix loudness level errors by instrument, for the worst (GRA3) and best (NUG3) performing algorithm. The box notches are the 95% confidence intervals for the medians (black horizontal lines).

systematic deviations in average error. It can be seen that the two algorithms produce relative levels that are systematically below the reference data, a trend that was observed in all 12 systems. The underestimation of source-to-mix loudness is likely caused by the presence of cross-source interference that reduces the relative gating threshold used by ITU-R BS.1770, thereby lowering the absolute programme loudness. This effect was not observed for the mixture, however, as the sum of interferers contributes little additional increase in mix energy.

Algorithm	CI_{95} RMSE_μ	$\overline{e}$	$\text{SD}_{\text{source}}$	SD_{song}
NUG3	**[1.72, 2.03]**	-1.46	**1.23**	**1.26**
NUG4	[1.82, 2.13]	-1.56	1.24	1.27
UHL1	[1.99, 2.46]	-1.58	1.68	1.66
NUG2	[2.09, 2.40]	-1.88	1.29	1.33
NUG1	[2.10, 2.41]	-1.89	1.29	1.33
GRA2	[2.02, 2.77]	**-0.76**	2.49	2.04
UHL3	[2.28, 2.77]	-1.84	1.84	1.77
UHL2	[2.39, 2.90]	-1.90	1.99	1.89
CHA	[3.15, 3.58]	-2.50	2.42	1.78
KON	[3.23, 3.63]	-2.16	2.92	1.78
OZE	[3.67, 4.22]	-2.24	3.61	1.92
GRA3	[4.16, 4.52]	-3.88	1.84	1.76

Table 1: CI_{95} for the average song RMSE, mean error $\overline{e}$ and error variation across sources and songs for the 12 separation algorithms. All values are in LU, with lower values indicating better performance (minimum per column displayed in boldface).

Table 1 provides, for each method, the CI_{95} for average of the 46 song RMSEs, and three other measures used to characterise the error: the mean error $\overline{e}$ quantifies the average deviation (and direction) from the reference levels; $\text{SD}_{\text{source}}$ expresses the average variation in error across sources (whether systematic or not); SD_{song} represents the average variation in error across songs. The algorithms have been sorted by ascending RMSE, with NUG3 performing best and GRA3 worst. The four NUG-based models and UHL1 ranked higher than the other systems, in agreement with the BSS Eval based assessments of [7, 9], although there are some notable differences between the ordering of the lower ranking models which may be attributed to the sensitivity of the loudness model to signal interference. Indeed, the mean errors are all negative, which, in the case of GRA3, inflates the RMSE. GRA2, CHA, KON and OZE show notably larger variation in error between sources than between songs, suggesting greater source dependency in separation quality.

NUG3, the best performing algorithm, is a multichannel audio source separation procedure that employs two deep neural networks to estimate source spectra [12]. To give insight into the expected performance of this algorithm on each of the four sources, repeated five-fold cross-validation was used to estimate the RMSE for each instrument group, with the systematic bias, as shown in Figure 2, removed. In short, this involved estimating the level offset from *all* sources of 37 randomly selected songs, and measuring the within-source RMSE on the remaining nine. After averaging the RMSEs from the five folds, this procedure was then repeated 1000 times. The estimated cross-validation RMSEs were 1.7 (bass), 1.2 (drums), 1.1 (other) and 0.9 LU (vocals), giving a mean source RMSE of 1.2 LU (SD 0.3). Using a similar procedure, the overall average song RMSE was 1.2 LU. Thus, the predictions of NUG3 can be improved by applying a correction of 1.46 LU (see Table 1).

Figure 3 shows violin plots of the reference relative-to-mix loudness levels for each source, and those estimated by NUG3. The inner broken lines show the three quartiles of the distributions (middle line being the median). These data were generated using the 46 songs constituting the test set of DSD100, so the reference distributions are not

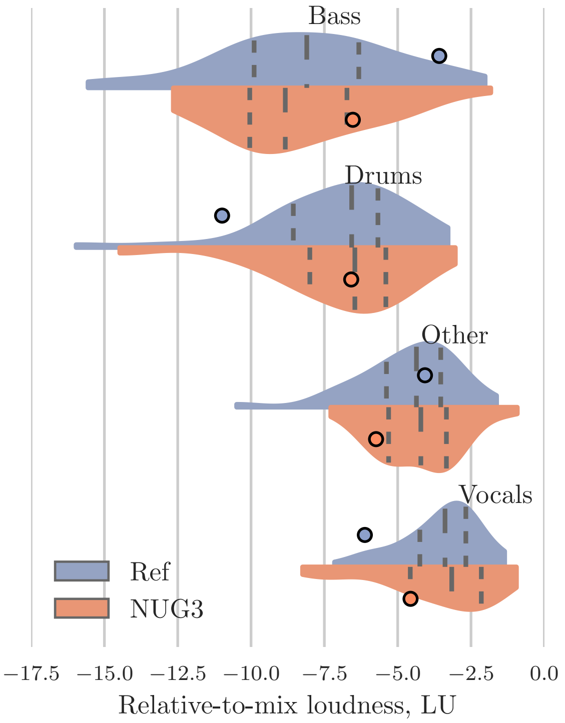

Figure 3: Violin plots showing the distribution of the reference (Ref) and estimated (NUG3) relative-to-mix loudness levels for each source. The violins have been truncated to the range of the data. The broken lines show the three quartiles of each distribution. The circles highlight the levels for the song with the largest RMSE as estimated using NUG3.

identical to those shown in Figure 1. The systematic offset of -1.46 LU was removed from the predictions of NUG3, and so the estimated medians and interquartile ranges are better aligned with those of the reference distributions. This suggests that this separation algorithm may be used to investigate mixing trends over a large corpus of songs, especially for vocal recordings. There are, however, notable differences between the shapes of the distribution for the ‘other’ source, which may reflect poor separation for specific (uncategorised) instruments. In addition, the algorithm fails to capture the quietest elements, especially bass, which can be attributed to poor separation of sources with low signal-to-mix energy ratios. Finally, the overlaid circles emphasise that large prediction errors can still occur at the individual song level, where, in this example, the maximum song RMSE was 2.9 LU. Again, maximum discrepancies are less extreme for the vocals.

4. CONCLUSION

The present study assessed the accuracy of 12 audio source separation algorithms when predicting the relative-to-mix loudness levels of four instrument groups of a musical mixture. The reference loudness balance data, obtained from 100 semi-professionally mixed songs, show that relative loudness is dependent on the source, but also suggests that the balance is genre dependent. The system evaluation indicates that algorithms based on a deep neural network which harnesses spatial information can be used successfully as an analytical tool, allowing researchers to investigate the level-balancing trends of commercial rock and pop music, especially for vocals. Future work should address the accuracy of these algorithms on other datasets involving more specific genres and test their performance on other multitrack features useful for understanding music production techniques.

5. REFERENCES

[1] J. D. Reiss, “Intelligent systems for mixing multichannel audio”, in *Proc. 17th Int. Conf. Digital Signal Processing*, 2011, pp. 1–6.

[2] M. J. Terrell, “Perceptual mixing for musical production”, PhD thesis, Queen Mary University of London, UK, 2013.

[3] P. Pestana and J. D. Reiss, “Intelligent audio production strategies informed by best practices”, in *Proc. 53rd Int. AES Semantic Audio Conf.*, 2014.

[4] B. De Man, B. Leonard, R. King and J. D. Reiss, “An analysis and evaluation of audio features for multitrack music mixtures”, in *Proc. 15th Int. Soc. Music Information Retrieval Conf.*, 2014.

[5] A. Wilson and B. M. Fazenda, “Navigating the mix-space: Theoretical and practical level-balancing technique in multitrack music mixtures”, in *Proc. 12th Sound and Music Conf.*, 2015.

[6] G. Wichern, A. S. Wishnick, A. Lukin and H. Robertson, “Comparison of loudness features for automatic level adjustment in mixing”, in *Proc. 139th AES Conv.*, 2015.

[7] A. J. R. Simpson, G. Roma, E. M. Grais, R. D. Mason, C. Hummersone, A. Liutkus and M. D. Plumbley, “Evaluation of audio source separation models using hypothesis-driven non-parametric statistical methods”, in *Proc. 24th EUSIPCO*, 2016.

[8] E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in blind audio source separation”, *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[9] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono and J. Fontecave, “The 2016 signal separation evaluation campaign”, in *Proc. 13th Int. Conf. LVA/ICA*, 2017, pp. 323–332.

[10] D. Ward and J. D. Reiss, “Loudness algorithms for automatic mixing”, in *Proc. 2nd AES Workshop on Intelligent Music Production*, 2016.

[11] ITU-R BS.1770, “Algorithms to measure audio programme loudness and true-peak audio level”, International Telecommunication Union, Tech. Rep. 4, 2015.

[12] A. A. Nugraha, A. Liutkus and E. Vincent, “Multichannel audio source separation with deep neural networks”, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

This work is supported by grant EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).