Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 Jun 22 – Jun 24, Erlangen, Germany

Intelligent Multitrack Reverberation Based on Hinge-Loss Markov Random Fields

Adán L. Benito[1] and Joshua D. Reiss[1]

[1]*Queen Mary University of London, Mile End Road, London E14NS, United Kingdom*

Correspondence should be addressed to Adán L. Benito (adan.benito@qmul.ac.uk)

ABSTRACT

We propose a machine learning approach based on hinge-loss Markov random fields to solve the problem of applying reverb automatically to a multitrack session. With the objective of obtaining perceptually meaningful results, a set of Probabilistic Soft Logic (PSL) rules has been defined based on best practices recommended by experts. These rules have been weighted according to the level of confidence associated with the mentioned practices based on existent evidence. The resulting model has been used to extract parameters for a series of reverb units applied over the different tracks to obtain a reverberated mix of the session.

1 Introduction

There are no strong rules established for the task of applying reverberation to a multitrack session in order to turn it into a mix, as different engineers have their own processes and techniques. Recent studies have tried to understand and analyse these practices in order to find common ground [1] [2]. Although this research may shed light on how different parameters of the reverberation are set by practitioners, a set of universal ground rules is yet to be established.

Different intelligent mixing techniques have been developed in order to help musicians and producers either alleviate their workload or improve the quality of their creative work. Most effort has been focused on solving automatic tasks [3] that can be interpreted in terms of defined goals that a machine could easily solve. A series of models for autonomous mixing and cross-adaptive digital audio effects (DAFx) [4] [5] have been developed for different areas of the mixing process [6] such as panning [7], equalisation [8] and compression [9].

The only effort for an automatic mixing tool that applies artificial reverberation is limited to individual tracks [10]. However, in a multitrack domain, the characteristics of one track or the whole mix may influence the decisions taken on another, so many of the approaches suggested for single track audio cannot be scaled to solve this problem.

2 Problem Formulation

In order to create an intelligent system capable of working independently in a meaningful way, it is necessary to analyse how humans (in this case experts) conduct different tasks [11]. Since a set of universal fixed rules has not yet been found that governs the application of reverberation to a multitrack session, a system that allows knowledge-informed prediction with different levels of confidence needs to be researched. A summary of best practices extracted from the analysis of previous research on the use and application of reverberation, alongside the document they have been extracted from, can be found in Table 1. To translate these into logical and arithmetic statements, we have parametrized them according to different audio features and characteristics and decomposed them into smaller units.

Designing an automatic multitrack reverberator requires a model capable of either extracting a joint probability distribution from data or a structure that represents all the dependencies for a domain with a high level of interconnection. Due to the subjective characteristics of the use of reverberation, the model should be flexible and interpretable. The approaches that have been taken in intelligent mixing before either use hard constraints (fixed rules) and curve fitting models, linear dynamical systems [6], or learn the rules from a dataset, making them unsuitable for solving our problem.

No.	Practice	Extracted from
1	Percussive instruments require shorter and denser reverbs than sustained sounds	[12] [13][14] .
2	Speech and voiced sounds may demand an increase in density and length of early reverberation but the reverberation tail should be kept short.	[15] [12]
3	It is, in general, better to send less low-frequency elements to a reverberator.	[13]
4	Tracks that present higher spectral centroid allow for higher amount of reverb.	[13]
5	Tracks with lower loudness and/or lower spectral flatness also can be more reverberated.	[13]
6	Clarity may be increased when the spectral occupation is high.	[13]
7	There is a suggestion that ties faster tempo to shorter decay times.	[13] [1]
8	Reverb time is correlated with a measure of the autocorrelation of the signal.	[13]
9	Reverb time is inversely correlated with both spectral flux and track's tempo (in bpm) and this effect is even stronger when applying a logarithmic transformation to both features.	[13]
10	It is recommendable to keep the pre-delay just past the Haas zone.	[13] [1]
11	Some engineers look for the closest subdivision of tempo above the Haas zone to set the predelay.	[13] [1]
12	Sparse mixes allow, in general, for greater reverberation times.	[14] [1]
13	Mixes with -9 dB of relative reverb loudness are rated as too reverberant.	[2]
14	It is preferred to have too little reverb rather than too much.	[2]
15	Bright reverbs may be prefered for dull sounds and vice versa.	[13]
16	High fidelity reverbs may be used with "trashy" sounds and vice versa.	[13]
17	Reverb brightness usually increases with reverb time.	[13]

Table 1: Summary of the different best practices gathered during our literature research and used in the PSL template, alongside the document they have been extracted from.

Furthermore, models typically used to solve relational learning and structured prediction problems do not fit our requirements, either because they lack expressivity [16], result in a slow convergence [17] or need to be trained on large amounts of data.

We make use of a probabilistic graphical model that enables efficient inference and prediction in complex structured domains, the *Hinge-loss Markov Random Fields* (HL-MRF) graphical model [18]. The different elements of HL-MRF models can be defined in terms of *Probabilistic Soft Logic* (PSL) [19], a general-purpose probabilistic programming language, unveiling a powerful framework for structured prediction. HL-MRFs and PSL have been previously used to solve relational problems, but, as far as we are aware, have not been applied before to any mixing or audio processing task.

2.1 Effect Architecture

Our proposed architecture follows a scheme similar to that of other adaptive and cross-adaptive DAFx designs [4] but with the introduction of PSL as a way to infer the different control parameters in such a complex domain.

The structure of the intelligent multitrack reverberator is presented in Figure 1. This scheme has been simplified to show only two tracks, but it is scalable to *N* tracks, so can be used in almost any multitrack mixing process.

This process could be segmented in five distinct stages:

- **Stage 1**: Tracks are analysed and the important features extracted individually.
- **Stage 2**: Different features are processed (scaled, normalised, etc) and the cross-adaptive features computed.

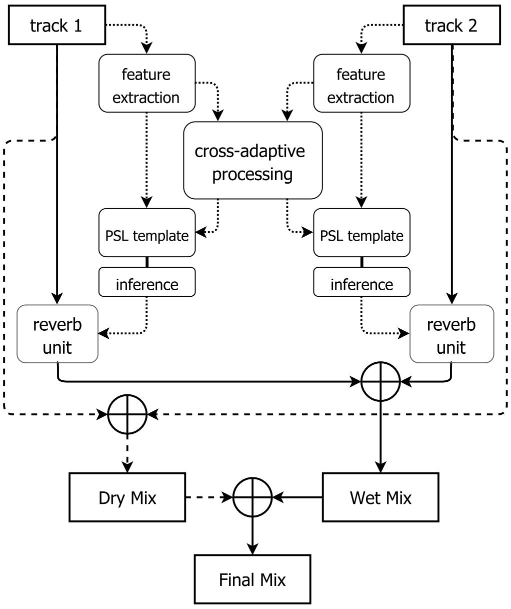

Fig. 1: Architecture of our intelligent multitrack reverberation (simplified for two tracks).

- **Stage 3**: A first PSL template is created to obtain semantic labels in terms of the different features via HL-MRF.
- **Stage 4**: A second PSL template is created using the rules extracted from the gathered knowledge instantiated to the features of this specific mix and the inferred semantic descriptors from the previous stage.
- **Stage 5**: The reverberation parameters are inferred by computing the HL-MRFs for the rules listed on the template.

- **Stage 6**: As not every set of parameters can be implemented, an optimisation method obtains the combination that better approximate the desired target.

3 Machine Learning for Structured Prediction

3.1 Markov Random Fields

For structure prediction, MRFs enable the possibility of using logical relationships to define probabilities. They do this by means of different scores (or weights) that are applied to probability distributions using potentials. These potentials model the behaviour of a domain by defining a probability density function in terms of logical clauses expressing variable relations hard to model otherwise.

3.2 Hinge-Loss Markov Random Fields

Hinge-loss Markov random fields define probability density functions over n continuous variables [19] with a domain equal to $[0,1]^n$ such that a maximum a posteriori is reached as a solution to a MAX SAT problem[1] associated with Lukasiewicz logic[19].
This optimisation problem can be expressed in terms of weighted distances to satisfaction that penalise how far the linear constraint is from being satisfied. *Constrained hinge-loss energy functions* constructed in terms of these weights allow the specification of either *hard* (must be satisfied) or *relaxed* linear constraints. HL-MRFs are then expressed over these functions so that states with lower energy are more probable [18].

3.3 Probabilistic Soft Logic

Probabilistic Soft Logic (PSL) is defined as "general-purpose framework for joint reasoning about similarity in relational domains" [20] and provides an intuitive interface for HL-MRFs as it allows to create templates for potentials and constraints using first-order logic rules as well as linear and quadratic constraints [19].

PSL uses *predicates* to specify relationships in the input data. When the predicate is combined with a defined input, it is called an *atom*, and each substitution of the input in the atom is called a *ground atom*. Ground atoms represent observations, can take values in [0,1] [19] and are the base of the implementation of PSL models.

PSL predicates can be defined either as *closed* (all the atoms are observations over the data) or *open* (some of the atoms are unknown).

HL-MRF templates are created using *logical* (disjunctive clauses of atoms or negations of atoms) or *arithmetic* (linear combination of atoms) PSL rules that can have an associated weight. Unweighted rules represent hard linear constraints, whereas weighed rules are used to penalise the satisfaction of the corresponding rule. Weights can potentially be learned from data using different methods (e.g. maximum likelihood estimation (ML)).

Once each of the atoms has been grounded, rules are translated into linear constraints and potentials and mapped to HL-MRFs.

4 Methodology and Implementation

4.1 A Reverberation Algorithm with Perceptual Control

The algorithm implemented for each of the reverberation units in our architecture is based on the work presented in [21]. This reverberator can be controlled by measures of reverberation : reverberation time (T_{60}), density (D), clarity (C), central time (T_C) and spectral centroid (S_C). These measures are then mapped to the parameters of the reverberator by grounding their definition to the parameterized response of the system.

However, not all combinations of measures can be implemented. A solution to this problem is provided in [22] via an optimisation process presented as a numerical problem.

Our particular implementation includes some small modifications to adapt the algorithm to our purpose. We first adapted the network to produce a mono output and then allowed the use of a rudimentary pre-delay via a delay line before the comb filters. A representation of the wet signal flow of the final reverberator is shown in Figure 2.

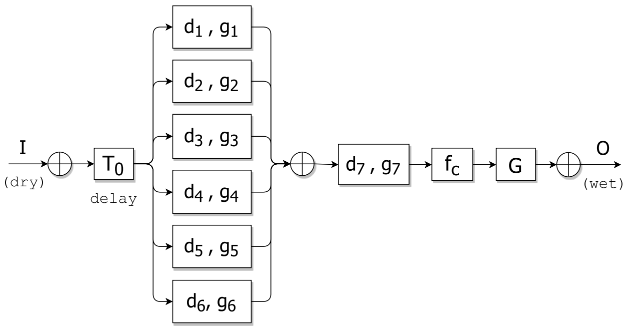

Fig. 2: Diagram of our modified version of Rafii and Pardo's reverberator [21]. T_0 represents the amount of pre-delay (in ms), d_i and g_i are the delay and gain factors of the comb-filters ($i = 1...6$) and the allpass filter ($i = 7$), f_c the cut-off frequency of the low-pass filter and G a gain parameter.

[1] A MAX SAT problem is such that "maximizes the weight of satisfied clauses in a knowledge base" with a boolean assignment [19].

4.2 Feature Extraction Process

In order to define the rules in terms of characteristics of the different tracks, different features need to be extracted from the tracks to characterise them in a meaningful way.

13 different features were chosen based on previous use and analysis and their relationship with perceptual descriptors, of which 6 correspond to spectral characteristics, 2 to temporal characteristics, 2 to dynamics. Additionally, three different cross-adaptive features were also extracted from the data: *tempo* of the dry down-mix, *percussivity weights* that indicate how percussive a track is with respect to the rest of the multitrack, and the *relative loudness* of each track with respect to the dry down-mix. All these features are presented in Table 2).

Feature	Related to	References
Spectral Centroid	brightness	[25]
Brightness	brigthness, presence	[23]
Spectral Roll-off	bandwith, voiced/unvoiced sounds	[23] [26]
Spectral Flatness	noisiness	[23]
Spectral Flux	stationarity, pitch variation	[23] [25]
Roughness	dissonance, “buzziness”	[23]
Autocorrelation	tempo	[26] [25]
Zero-crossing Rate	noisiness, high-frequency content	[26] [25]
Tempo	tempo	[23] [13]
Relative Loudness	loudness	[13]
Crest Factor	percussivity, dynamic range	[13] [9]
Low Energy	sustained/transient sounds	[13] [9]
Percussivity Weights	percussivity	[9]

Table 2: List of the different features and their relationship to perceptually meaningful descriptors

The feature extraction process was performed in Matlab and non-adaptive features were extracted using the *MIRtoolbox* [23]. Each track was previously downsampled to 22050 kHz to reduce the computational cost. Afterwards, a structured segmentation strategy was carried out, as it has proven to provide better accuracy for multitrack contexts [27]. The segmentation frame size was based on the tempo of the down-mix. Each different frame was also segmented into smaller 23 ms fragments. The value for each feature for the tempo-based fragments was given as the average of all subframes.

Once all features were obtained, a re-scaling process was carried out to map all the values to a [0,1] range for their use in the PSL template. Features were scaled using a linear minimum/maximum process with limits based on observations from different multitracks. Logarithmic scaling was applied over some features to provide a better approximation to the response of the human auditory system.

4.3 Designing the PSL Templates

In order to model dependencies between features and reverb parameters, two PSL templates were laid out defining the rules used in the inference process. The *best practices* studied from the literature were laid out as either logical clauses or linear combinations. A weight dependent on the amount of evidence existent for the related practice is used to reward those that have been contrasted or penalise others.

4.3.1 Universal Constants

This first input represents all the elements over which the model will be grounded. We propose four different types to define the model's universe:

- **Track**, an identifier for each track (the name of the track).
- **Feature**, corresponding to all the relevant features that have been extracted from the different tracks.
- **Property**, refers to a set of semantic descriptors that will characterise each of the tracks in terms of different features. In this case, we used six different properties that are pair-wise complementary: bright/-dark, percussive/sustained and voiced/unvoiced.
- **Parameter**, a list of the different reverb parameters to be controlled: pre-delay, T_{60}, density, clarity, central time and spectral centroid [21].

4.3.2 Predicates

Predicates are used to form a relationship between two types of constants. Our model is built around three different predicates.

- **Feature** is a closed predicate that has two different arguments: *Track* and *Feature*. The observation of each singular atom is given by the value obtained during the analysis of the track for a given feature normalised to [0,1].
- **Property** defines the extent to which each track relates to a semantic descriptor. Properties are pair-wise complementary [2]. The value of the HL-MRF random variable that will be induced to each atom indicates how it will be satisfied when specified for a given track.
- **Parameter** takes *Track* and *Parameter* as arguments. The values associated with the different random variables will define the settings of our reverberator units.

[2] For each property defined there is another one that represents the negation of its atoms.

4.3.3 Complement Definition

Since we want to define some properties as complementary, it is required to support this relationship with rules. We use unweighted rules that assure that the inequality is always satisfied. There are two approaches that can be taken, logical or arithmetic rules (the latter being more simple to use).

Properties are structured in pairwise complementary opposites, meaning that each different signal will be defined as a combination of both (for example, a bass track could be 90% "dark" and 10% "bright", but not 80% "dark" and 40% "bright"). Although this increases the computational cost of the PSL model, it enables a more intuitive syntax when defining the rules.

4.3.4 Semantic Labels as a Function of the Signal's Features

Labels have a meaning only if we establish dependencies. The first PSL template is related to the mappings between features extracted from the signals and the semantic properties defined in the model that will be used to create the rules described in Table 1.

Bright vs Dark The brightness of an audio signal has always been associated with higher values of the spectral centroid [28]. However, another measure of the perception of timbre is the so-called brightness estimator, which calculates the percentage of spectral energy up to a cutoff frequency (usually 1 to 3 kHz). Moreover, the spectral roll-off, as a representation of the signal's bandwidth, can give an idea of the high-frequency content. Based on previous research, one could argue that both roll-off and centroid have the same effect on the perceived timbre [29]. However, we do not know of any cross-comparison with the high-frequency content measure.

Percussive vs Sustained Several methods have been used in Music Information Retrieval to account for the percussive or transient qualities of a signal. The most used measure of percussivity is the crest factor, but the low-energy ratio seems to better correlate to this quality. In addition, cross-adaptive percussivity weightings can be used within a mix context [9].

Voiced vs Unvoiced To distinguish between voiced and unvoiced sounds (less tonal and noisier), the zero-crossing rate has been proposed as an idea of the distribution of energy over frequency (higher rates imply higher frequencies) [30]. Another effective parameter in voice detection is spectral flatness, which indicates the noisiness of the signal (flat spectrum) or its tonal character (non-flat spectrum) [31]. Finally, the spectral roll-off may also be correlated, as the greater part of the spectral energy of voiced sounds is located in the lower bands. More specifically, frequency bands under 3kHz correspond with voiced sounds whereas unvoiced sounds accumulate stronger energy in the 3 to 4 kHz bands [32]. Therefore, the brightness measure, when specified at 3 kHz, can also provide information about the voiced quality of a signal.

4.3.5 Knowledge-based Rules

In the second PSL template rules related to the gathered expert knowledge provide the interconnection between the extracted data and the reverb parameters that will be used to apply the effect to different tracks.

The range of weights assigned to the rules was determined based on experimentation with different templates. In a range from 0 to 10, higher weights were associated with rules that have the potential for being grounded but not yet defined as linear constraints, while lower weights correspond to particular suggestions that do not hold enough evidence to be defined as common practices. The rest of the weights were assigned either based on the comparative amount of evidence in the literature or on correlation values extracted from previous research.

If a relationship between different atoms needs to be defined, it has to be decomposed into rules that cover all different explanations via the definition of complementary operations.

5 Results

We first tested the whole system over a simplified dataset comprised of four selected tracks extracted from a real complete session[3]. Simplified version comprised of lead voice, keys, drums and bass from the Open Multitrack Testbed [34]. The selection was made to evaluate the labelling process and to see how the reverberation parameters were obtained.

5.1 Label Assignment and Mapping

As seen in Figure 3, the system works well for labelling percussive instruments and to discern bright instruments. If we focus on the percussive qualities, the singing voice is labelled as the less percussive (and therefore more sustained) of all the tracks and the drums get a high rating for this property. Brightness is less easy to understand, but the keys and the bass are associated with a lower amount (0.2 or 20%) as may be expected. Discriminating between voiced and unvoiced tracks seems to also be satisfactory, as drums are associated with the lower value for 'voiced'. Nevertheless, it can be appreciated how this property is clustered in the upper range, indicating a possible bias.

From the analysis of the average properties assigned to 148 tracks from 6 different sessions[3] available on the Open Multitrack Testbed in Figure 4 we can see

[3] Supporting materials including audio examples from the reference sessions mentioned in this document are available at: https://code.soundsoftware.ac.uk/projects/multitrackreverb

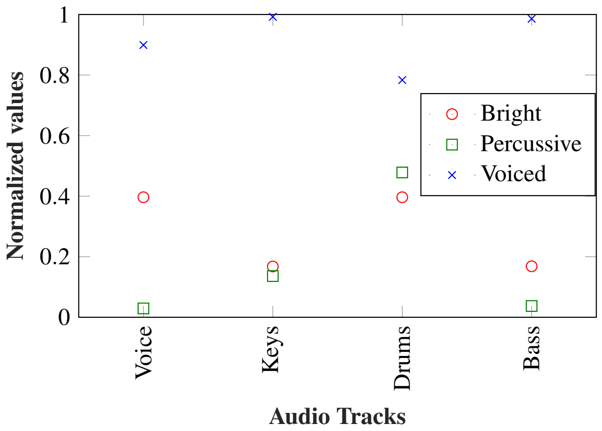

Fig. 3: Association of the different properties to each track of our simplified multitrack session.

how the system is biased towards rating tracks as highly voiced (over 80%) and less percussive (under 20%). This seems to be a result of how the features have been scaled before filling the PSL template and not a problem of the template itself. Limits used for feature scaling may also be affecting negatively. In an ideal scenario, the distribution of the different features over a large number of tracks should be normal. However, a quick look at Figure 6 shows that features such as *percussivity weights*, *spectral centroid* are closely distributed around more extreme values, introducing another bias in the template.

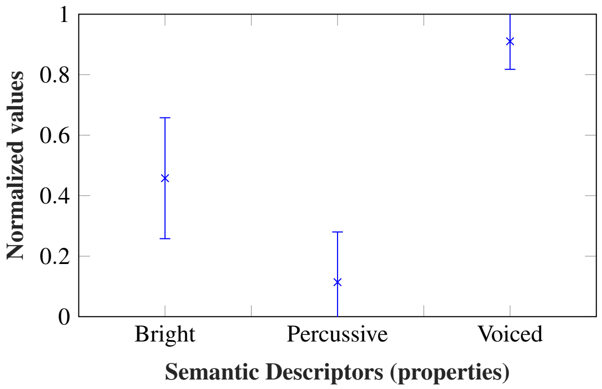

Fig. 4: Mean values for the normalised properties (labels) assigned by the model for 148 tracks. Error bars represent the standard deviation of the data.

5.2 Parameters and Rule Satisfaction

By performing a cross-comparison between results presented in Figures 3 and 5, we can extract to which level our model has succeeded in terms of satisfying the different practices.

If we focus on the reverberation time, RT_{60} we can appreciate how it is higher for the tracks labelled as non-percussive, like voice or keys, and lower for more percussive tracks, like the drums. This observation correlates quite well with rules 1 and 2 in Table 1. The reverb time has been kept higher for sustained sounds and lower for percussive sounds, but also low for darker sounds (as rules 3 and 4 on Table 1 may imply). Additionally, the fact that the density is higher for more percussive sounds is also evidence to support rule number 1 (see Table 1).

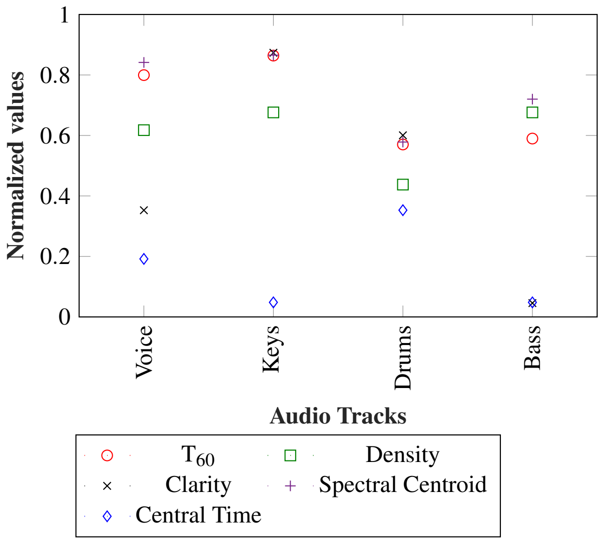

Fig. 5: Association of the different parameters to each track of our simplified multitrack session.

Moreover, reverberation applied to voiced sounds should have a lower central time to represent the higher length of the early reflections. Our results agree with this rule since the central times vary in inverse proportion to those tracks marked as voiced.

If we refer to the brightness of the tracks in terms of their spectral centroid (as it is the most influencing factor) we should have a correlation between reverberation time and brighter tracks (rule 4), but this seems to be an exception for the keys as the relationship between the RT_{60} and the reverb's spectral centroid also has some influence over this (rule 17).

5.3 Perceptual Qualities

It is quite complex to establish the limits of what a "good" reverberator is in terms of its potential use in music production. In [2], a perceptual evaluation of reverberation was conducted based on the *Relative Reverb Loudness* (RRL). Although this parameter was not sufficient to explain the perception of reverberation, perceived excess of reverberation typically had a more detrimental effect on subjective preference than a perceived lack of reverb.

To evaluate the quality our reverberation in terms of this rule (already included in our PSL model), we computed the relative reverb loudness over a the 6 multitrack sessions mentioned before after being processed by

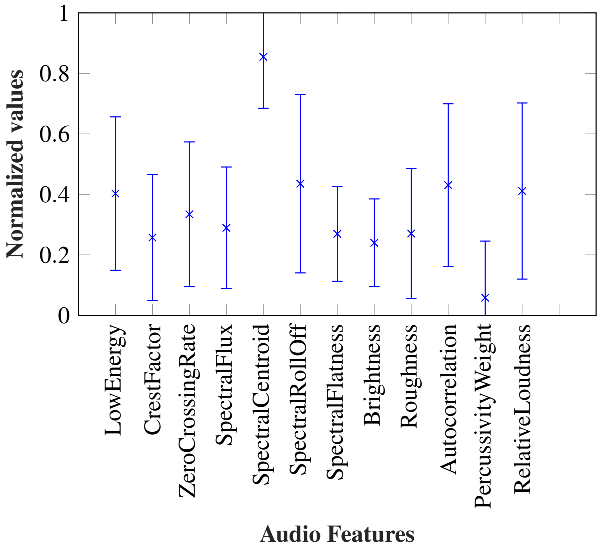

Fig. 6: Mean values for the extracted features for 148 tracks. Error bars represent the standard deviation of the data.

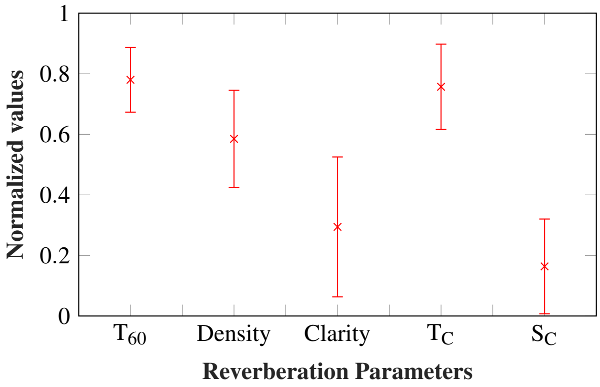

Fig. 7: Mean values for the inferred reverb parameters for 148 tracks. Error bars represent the standard deviation of the data.

our reverberator also on a series mixes of these sessions created by different experts (extracted from the Open Multitrack Testbed) [33].

While the average RRL for professionally mixed tracks was on average -12 LU, our algorithm produced mixes with RRL values around 4.5 LU, way beyond the limit. This was not surprising, using reverberation as an insert effect for each track will always result in higher overall reverberation. In addition, if we look at how the inferred parameters are distributed for a higher number of tracks (Figure 7), we are systematically assigning a high RT_{60} to all tracks but also a higher central time (longer early reverberations), which would naturally result in more perceived reverberation.

However, if we attenuate the wet mix by means of a gain factor to counteract this effect and obtain a more balanced final mix (while still retaining the relationship between the different elements of the wet mix), our results get much closer to the target RRL with an average of f -15 LU for a 10/100 wet/dry mix ratio (in %).

6 Discussion

A multitrack reverberator has been developed that is capable of identifying differences and relationships among input signals and that applies the effect according to a set of rules that have been pre-specified. These rules do not need to be grounded and can respond to different levels of confidence, plus they can be interpreted as understandable arithmetic and logic statements.

The results show that our algorithm is able to label instruments according to different levels of percussivity with great accuracy within the PSL template and then use the results in the prediction process.

Although the resulting mix may not be comparable to a human-made mix in terms of perceptual qualities, we have to take into account that we have not performed any pre-processing of the individual tracks [4] and that the way we applied the reverberation and the algorithm itself will always result in higher levels of relative reverberation loudness than standard practices.

However, this research explores the potential of HL-MRFs and PSL in intelligent mixing practices. When combined together, these two frameworks offer a model with an intuitive interface and fast inference process capable of modelling complex relationships between the data. In addition, the use of PSL templates enables the introduction of semantic data in the model, which has proven to be useful for a variety of multitrack mixing practices [35] [36].

7 Limitations and Further Work

The weights for the different rules were assigned experimentally based on the amount evidence found in the literature. However, these weights can be learned from any specified dataset of mixes using different techniques [18]. As the rules are defined in advance, the amount of data required for training is substantially less than for other techniques (such as NNs or SVMs) and reduces the risk of biasing the results to the data.

The PSL parser we used is still a work in progress and, therefore, has several syntax limitations. Our work was hindered by the lack of support from arithmetic rules with relaxed constraints [19], aggregates and squared potentials.

Considering the reverberation algorithm we have employed on this design, it presents some limitations regarding parameter control and does not represent the current state of the art on reverberation algorithms [37]. Furthermore, in order to achieve better results, pre and post-processing of the different tracks may be necessary.

[4] Level adjustment, equalisation, panning, compression, etc.

8 Aknowledgments

We would like to thank Emmanouil Theofanis Chourdakis for his valuable help and advice during the course of this research.

References

[1] Pestana, P. D. et al., "User preference on artificial reverberation and delay time parameters with applications to automatic multitrack mixing," *J. Audio. Eng. Soc. (to appear)*, 2017.

[2] De Man, B. et al., "Perceptual evaluation and analysis of reverberation in multitrack music production," *J. Aud. Eng. Soc.*, 2017.

[3] De Man, B. and Reiss, J. D., "A semantic approach to autonomous mixing," *Journal on the Art of Record Production (JARP)*, Issue 8, 2013.

[4] Verfaille, V. et al., "Adaptive digital audio effects (A-DAFx): A new class of sound transformations," *IEEE Trans. Audio Speech Lang. Proc.*, 14(5), pp. 1817–1831, 2006.

[5] Matz, D., Cano, E., and Abeßer, J., "New Sonorities for Early Jazz Recordings Using Sound Source Separation and Automatic Mixing Tools." *16th Intl. Soc. Music Infirmation Retrieval (ISMIR)*, pp. 749–755, 2015.

[6] Scott, J. et al., "Automatic multi-track mixing using linear dynamical systems," in *Sound and Music Computing Conf.*, 2011.

[7] Perez_Gonzalez, E. and Reiss, J., "A real-time semi-autonomous audio panning system for music mixing," *EURASIP J. Adv. Sig. Pro.*, 2010.

[8] Hafezi, S. and Reiss, J. D., "Autonomous multitrack equalization based on masking reduction," *J. Aud. Eng. Soc.*, 63(5), 2015.

[9] Ma, Z., De Man, B., Pestana, P. D., Black, D. A., and Reiss, J. D., "Intelligent multitrack dynamic range compression," *J. Audio Eng. Soc.*, 63(6), pp. 412–426, 2015.

[10] Chourdakis, E. T. and Reiss, J. D., "Automatic Control of a Digital Reverberation Effect using Hybrid Models," in *60th Intl. AES Conf.: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, 2016.

[11] De Man, B. and Reiss, J. D., "A knowledge-engineered autonomous mixing system," in *Aud. Eng. Soc. Conv. 135*, 2013.

[12] Case, A. U., *Mix smart*, Focal Press, 2011.

[13] Pestana, P. D. L. G., *Automatic mixing systems using adaptive digital audio effects*, Ph.D. thesis, Universidade Católica Portuguesa, 2013.

[14] Izhaki, R., *Mixing audio: concepts, practices and tools*, Taylor & Francis, 2013.

[15] Deutsch, D., *Psychology of music*, Elsevier, 2013.

[16] De Raedt, L., *Logical and relational learning*, Springer Science &amp; Business Media, 2008.

[17] Blockeel, H. and Uwents, W., "Using neural networks for relational learning," in *ICML-2004 Workshop on Statistical Relational Learning and its Connection to Other Fields*, 2004.

[18] Bach, S., Huang, B., London, B., and Getoor, L., "Hinge-loss Markov random fields: Convex inference for structured prediction," *arXiv preprint arXiv:1309.6813*, 2013.

[19] Bach, S. H., Broecheler, M., Huang, B., and Getoor, L., "Hinge-loss Markov random fields and probabilistic soft logic," *arXiv preprint arXiv:1505.04406*, 2015.

[20] Brocheler, M., Mihalkova, L., and Getoor, L., "Probabilistic similarity logic," *arXiv preprint arXiv:1203.3469*, 2012.

[21] Rafii, Z. and Pardo, B., "Learning to Control a Reverberator Using Subjective Perceptual Descriptors." in *10th Intl. Soc. Music Information Retrieval Conf., Kobe, Japan*, 2009.

[22] Chourdakis, E. T. and Reiss, J. D., "A machine learning approach to application of intelligent artificial reverberation," 2017.

[23] Artillot, O., "MIRtoolbox 1.6, User's Manual," 2014.

[24] Committee, E. T. et al., "Loudness Recommendation EBU R128," 2011.

[25] Lerch, A., *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons, 2012.

[26] Peeters, G., "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," 2004.

[27] Hargreaves, S., Klapuri, A., and Sandler, M., "Structural segmentation of multitrack audio," *IEEE Trans. Aud. Speech Lang. Proc.*, 20(10), pp. 2637–2647, 2012.

[28] Schubert, E. and Wolfe, J., "Does timbral brightness scale with frequency and spectral centroid?" *Acta acustica united with acustica*, 92(5), pp. 820–825, 2006.

[29] Brent, W., "Cepstral analysis tools for percussive timbre identification," in *Proceedings of the 3rd International Pure Data Convention, Sao Paulo, Brazil*, 2009.

[30] Bachu, R. et al., "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Amer. Soc. Eng. Edu. (ASEE) Zone Conf.*, 2008.

[31] Moattar, M. and Homayounpour, M., "A simple but efficient real-time voice activity detection algorithm," in *17th European IEEE Sig. Proc. Conf. (EUSIPCO)*, pp. 2549–2553, 2009.

[32] Turkmen, H. I. and Karsligil, M. E., "Dysphonic Speech Reconstruction," *IEEE Eng. in Medicine and Biology Mag.*, p. 136, 2010.

[33] De Man, B. et al., "An analysis and evaluation of audio features for multitrack music mixtures," *5th Int. Soc. for Music Information Retrieval Conf. (ISMIR 2014)*, 2014.

[34] De Man, B. et al., "The open multitrack testbed," in *Audio Eng. Soc. Conv. 137*, 2014.

[35] Wilmering, T. et al., "High-Level Semantic Metadata for the Control of Multitrack Adaptive Digital Audio Effects," in *Audio Eng. Soc. Conv 133*, 2012.

[36] Stables, R. et al., "SAFE: A system for the extraction and retrieval of semantic audio descriptors," 2014.

[37] Välimäki, V. et al., "More Than 50 Years of Artificial Reverberation," in *60th Intl. AES Conf.: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, 2016.

[38] Ronan, D. et al., "Automatic subgrouping of multitrack audio," *18th Int. Conf. on Digital Audio Effects (DAFx-15)*, 2015.