DEEP LEARNING AND INTELLIGENT AUDIO MIXING

Marco A. Martínez Ramírez, Joshua D. Reiss

Centre for Digital Music
Queen Mary University of London
{m.a.martinezramirez,joshua.reiss}@qmul.ac.uk

ABSTRACT

Mixing multitrack audio is a crucial part of music production. With recent advances in machine learning techniques such as deep learning, it is of great importance to conduct research on the applications of these methods in the field of automatic mixing. In this paper, we present a survey of intelligent audio mixing systems and their recent incorporation of deep neural networks. We propose to the community a research trajectory in the field of deep learning applied to intelligent music production systems. We conclude with a proof of concept based on stem audio mixing as a content-based transformation using a deep autoencoder.

1. INTRODUCTION

Audio mixing essentially tries to solve the problem of unmasking by manipulating the dynamics, spatialisation, timbre or pitch of multitrack recordings.

Automatic mixing has been investigated as an extension of adaptive digital audio effects [1] and content-based transformations [2]. This is because the processing of an individual track depends on the content of all the tracks involved, thus audio features are extracted and mapped for the implementation of cross-adaptive systems [3].

The most informed expert knowledge framework for automatic mixing applications can be found in [4] and [5]. In which the potential assumptions for technical audio mixing decisions were validated through various strategies.

In the same way, audio features have been analysed to gain a better understanding of the mixing process or to perform different tasks within an automatic mixing framework. In [6], low-level features were extracted from a set of mixing sessions and their variance was analysed between instruments, songs and sound engineers. [7] extracted features from 1501 mixes and proposed that high-quality mixes are located in certain areas of the feature space.

Likewise, machine learning techniques have been applied to the estimation of mixing coefficients through linear dynamical systems [8] and least-squares optimization models [9]. [10] used random forests classifiers and agglomerative clustering for automatic subgrouping of multitrack audio. Also, by means of an interactive genetic algorithm, [11] proposed an exploration of a subjective mixing space.

Through the different listening tests of these models, the systems proved to perform better than amateur sound engineers, but experienced sound engineer mixes were always preferred. In addition, most of the models were aimed at a technically clean mix, yet audio mixing often includes different types of objectives such as transmitting a particular feeling, creating abstract structures, following trends, or enhancing [12]. Also, the models cannot be extended to complex or unusual sounds. Thus past attempts fall short, since a mix that only fulfils basic technical requirements cannot compete with the results of an experienced sound engineer.

"The intelligence is in the sound" [1]

Therefore, machine learning techniques together with expert knowledge, may be able to guide towards the development of a system capable of performing automatic mixing or to assist the sound engineer during the mixing process. We propose to the community a research outline within the framework of automatic mixing and deep neural networks.

The rest of the paper is organised as follows. In Section 2 we summarise the relevant literature related to deep learning and music. We propose a research outline in Section 3 and in Section 4 we present a proof of concept. Finally, we conclude with Section 5.

2. BACKGROUND

2.1. Deep learning and music

In recent years, *deep neural networks* (DNN) applied to music have experienced a significant growth. A large percentage of current research has been devoted to extract information and understand its content. Most of the examples are in the fields of music information retrieval [13, 14], music recommendation [15, 16], and audio event recognition [17].

Nonetheless, deep learning applied to the generation of music has become a growing field with emerging projects such as *Magenta* and its *Nsynth* [18]: A *Wavenet* [19] *autoencoder* that generates audio sample by sample and allows instrument morphing from a dataset of short musical notes. This was achieved using an end-to-end architecture, where raw audio is both the input and the output of the system. Similarly, [20] obtained raw audio generation without the need of handcrafted features and [21] accomplished singing voice synthesis.

Also, as an alternative to working directly with audio samples other architectures have being explored. [22] trained

a Long Short Term Memory (LSTM) recurrent neural network (RNN) architecture together with Reinforcement Learning (RL) to predict the next note in a musical sequence. [23] used a character-based model with a LSTM to generate a textual representation of a song.

2.1.1. Deep learning and music production

Recent work has demonstrated the feasibility of deep learning applied to intelligent music production systems. [24] used a convolutional DNN and supervised learning to extract vocals from a mix. [25] used DNNs for the separation of solo instruments in jazz recordings and the remixing of the obtained stems. Similary, [26] used a DNN to perform audio source separation in order to remix and upmix existing songs. [27] relied on pre-trained *autoencoders* and critical auditory bands to perform automatic dynamic range compression (DRC) for mastering applications. This was achieved using supervised and unsupervised learning methods to unravel correlations between unprocessed and mastered audio tracks.

Thus, it is worth noting the immense benefit that generative music could obtain from intelligent production tools and vice versa.

3. RESEARCH OUTLINE

We seek to explore *how can we train a deep neural network to perform audio mixing as a content-based transformation* without using directly the standard mixing devices (*e.g.* dynamic range compressors, equalizers, limiters, etc.). Following an end-to-end architecture, we propose to investigate whether a DNN is able to learn and apply the intrinsic characteristics of this transformation.

In addition, we investigate *how can a deep learning system use expert knowledge* to improve the results within a mixing task. In this way, based on a set of best practices, we explore how to expand the possibilities of a content-based audio mixing system.

Similarly, we research *whether the system can perform a goal-oriented mixing task* and if we can guide it with features extracted from a mixdown. This, taking into account statistical models used in style-imitation or style-transfer [28, 29]. We seek to investigate these principles within an audio mixing framework.

We explore if the system can act alongside human action as a *technical assistant*, and also *whether we can integrate user interaction as a final fine-tuning of the mixing process*. We investigate a tool capable of guaranteeing technical criteria while learning from the user. Thus, by making the user an integral part of the system, we take advantage of both the memory and processing power of the machine as well as the intuitive and synthesizing ability of the user [30].

4. PROOF OF CONCEPT

Processing individual stems from raw recordings is one of the first steps of multitrack audio mixing. We investigate stem processing as a content-based transformation, where the frequency content of *raw* recordings (input) and *stems* (target) leads us to train a *deep autoencoder* (DAE). Thus, we are exploring whether the system is capable of learning a transformation that applies the same chain of audio effects.

The raw recordings and individual processed stems were taken from [31], mostly based on [32] and following the same structure; a song consists of the mix, *stems* and *raw* audio. The dataset consists of 102 multitracks which correspond to genres of commercial western music. Each track was mixed and recorded by professional sound engineers.

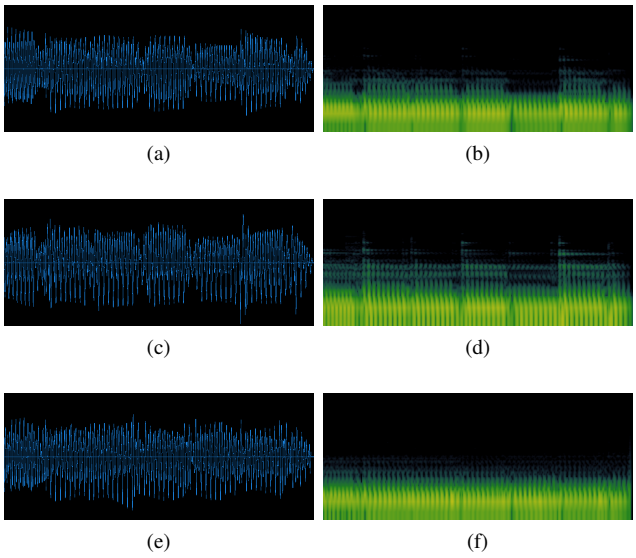
All tracks have a sampling frequency of 44.1 kHz, and we proceeded to find the 10 seconds with the highest energy for each stem track. Thus, the corresponding raw tracks were then analysed and the one with the highest energy in the same 10 second interval was chosen.

The selected segments were downmixed to mono and loudness normalisation was performed using [33]. Data augmentation was implemented by pitch shifting each raw and stem track by ± 4 semitones in intervals of 50 cents. The frequency magnitude was computed with frame/hop sizes equal to 2048/1024 samples. The test dataset corresponds to 10% of the raw and stem segments and data augmentation was not applied.

Table 1: Number of raw/stem tracks and augmented segments.

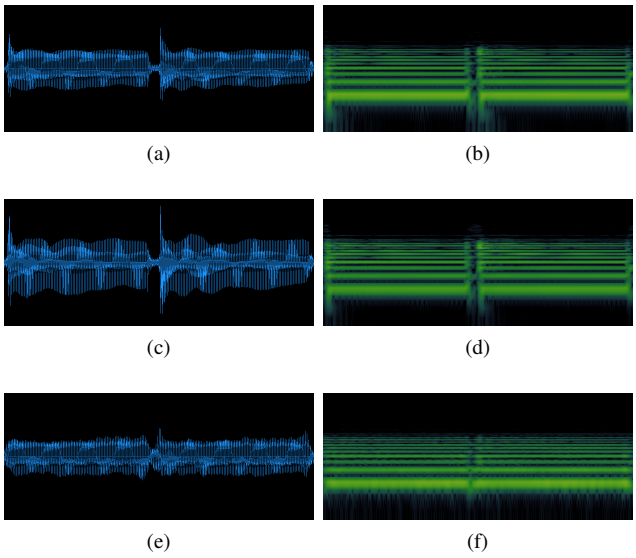
Group	Instrument Source	Raw	Stem	Augmented Raw/Stem
Bass	electric bass	96	62	1020
	synth bass	12	6	
Guitar	clean electric guitar	112	36	1224
	acoustic guitar	55	24	
	distorted electric guitar	78	20	
	banjo	2	2	
Vocal	male singer	145	36	969
	female singer	61	22	
	male rapper	12	2	
Keys	piano	113	38	884
	synth lead	51	17	
	tack piano	27	7	
	electric piano	3	3	

The DAEs consist of feed-forward stacked *autoencoders*. Each hidden layer was trained using the greedy layer-wise approach [34], dropout with a probability of 0.2, *Adam* as optimizer, *reLu* as activation function, and mean absolute error as loss function. In total, each DAE has 3 hidden layers of 1024 neurons and input and output layers of 1025 samples.



(a) (b) (c) (d) (e) (f)

Figure 1: Bass, waveforms and spectrograms. (a), (b) Raw input. (c), (d) Stem target. (e), (f) DAE's output .

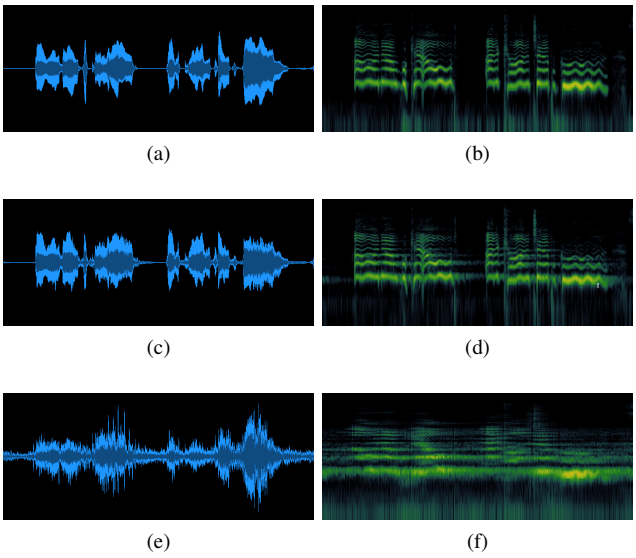


(a) (b) (c) (d) (e) (f)

Figure 2: Guitar, waveforms and spectrograms. (a), (b) Raw input. (c), (d) Stem target. (e), (f) DAE's output .

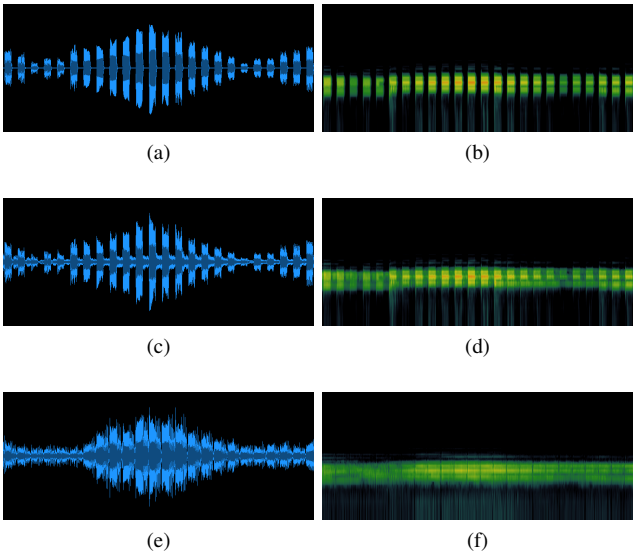
The DAEs were trained independently for each instrument group. After 100 learning iterations, Figs. 1-4 show the results of the DAEs when tested with a raw recording from the test dataset. The waveform was reconstructed with the original phase of the raw segments.

It can be seen that the output contains artefacts and noise introduced by the DAEs, however the main harmonics and envelope has been maintained. Among the groups, the DAEs performed better for Bass and Guitar than for Vocal and Keys. This is because vocal sounds are much more complex than the other instruments groups. Also, the Keys' instruments have a higher variance of roles within the different genres. Further analysis is required as well as listening and similarity tests to correctly measure the performance of the system.



(a) (b) (c) (d) (e) (f)

Figure 3: Vocal, waveforms and spectrograms. (a), (b) Raw input. (c), (d) Stem target. (e), (f) DAE's output .



(a) (b) (c) (d) (e) (f)

Figure 4: Keys, waveforms and spectrograms. (a), (b) Raw input. (c), (d) Stem target. (e), (f) DAE's output .

5. CONCLUSION

The current research is at an early stage and the DAE architecture is considerably simple. We plan to incorporate already successful end-to-end architectures to achieve a sys-

tem capable of performing stem audio mixing as a content-based transformation.

Intelligent music production systems have the potential to benefit from deep learning techniques applied to music generation and vice-versa. We encourage the community to investigate the proposed research questions. In this way, an intelligent system capable of perform automatic mixing or to assist the sound engineer during the mixing process could be possible.

6. REFERENCES

[1] V. Verfaille, U. Zölzer, and D. Arfib, "Adaptive digital audio effects (A-DAFx): A new class of sound transformations," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1817–1831, 2006.

[2] X. Amatriain *et al.*, "Content-based transformations," *Journal of New Music Research*, vol. 32, no. 1, pp. 95–114, 2003.

[3] J. D. Reiss, "Intelligent systems for mixing multichannel audio," in *17th International Conference on Digital Signal Processing*, pp. 1–6, IEEE, 2011.

[4] P. D. Pestana and J. D. Reiss, "Intelligent audio production strategies informed by best practices," in *53rd Conference on Semantic Audio: Audio Engineering Society*, 2014.

[5] B. De Man, *Towards a better understanding of mix engineering*. PhD thesis, Queen Mary University of London, 2017.

[6] B. De Man *et al.*, "An analysis and evaluation of audio features for multitrack music mixtures," in *15th International Society for Music Information Retrieval Conference*, 2014.

[7] A. Wilson and B. Fazenda, "Variation in multitrack mixes: analysis of low-level audio signal features," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 466–473, 2016.

[8] J. Scott *et al.*, "Automatic multi-track mixing using linear dynamical systems," in *8th Sound and Music Computing Conference*, 2011.

[9] D. Barchiesi and J. Reiss, "Reverse engineering of a mix," *Journal of the Audio Engineering Society*, vol. 58, no. 7/8, pp. 563–576, 2010.

[10] D. Ronan *et al.*, "Automatic subgrouping of multitrack audio," in *18th International Conference on Digital Audio Effects*, 2015.

[11] A. Wilson and B. Fazenda, "An evolutionary computation approach to intelligent music production informed by experimentally gathered domain knowledge," in *2nd AES Workshop on Intelligent Music Production*, vol. 13, 2016.

[12] E. Deruty, "Goal-oriented mixing," in *2nd AES Workshop on Intelligent Music Production*, vol. 13, 2016.

[13] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 6959–6963, IEEE, 2014.

[14] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 5, pp. 927–939, 2016.

[15] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in Neural Information Processing Systems*, pp. 2643–2651, 2013.

[16] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *22nd International Conference on Multimedia*, pp. 627–636, ACM, 2014.

[17] H. Lee *et al.*, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, pp. 1096–1104, 2009.

[18] J. Engel *et al.*, "Neural audio synthesis of musical notes with wavenet autoencoders," *34th International Conference on Machine Learning*, 2017.

[19] A. v. d. Oord *et al.*, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[20] S. Mehri *et al.*, "Samplernn: An unconditional end-to-end neural audio generation model," in *5th International Conference on Learning Representations*, ICLR, 2017.

[21] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *Interspeech 2017*.

[22] N. Jaques *et al.*, "Tuning recurrent neural networks with reinforcement learning," in *5th International Conference on Learning Representations*, 2017.

[23] B. Sturm, J. F. Santos, and I. Korshunova, "Folk music style modelling by recurrent neural networks with long short term memory units," in *16th International Society for Music Information Retrieval Conference*, 2015.

[24] A. J. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 429–436, Springer, 2015.

[25] S. I. Mimilakis *et al.*, "New sonorities for jazz recordings: Separation and mixing using deep neural networks," in *2nd AES Workshop on Intelligent Music Production*, vol. 13, 2016.

[26] G. Roma *et al.*, "Music remixing and upmixing using source separation," in *2nd AES Workshop on Intelligent Music Production*, September 2016.

[27] S. I. Mimilakis *et al.*, "Deep neural networks for dynamic range compression in mastering applications," in *140th Audio Engineering Society Convention*, 2016.

[28] G. Hadjeres, J. Sakellariou, and F. Pachet, "Style imitation and chord invention in polyphonic music with exponential families," *arXiv preprint arXiv:1609.05152*, 2016.

[29] F. Pachet, "A joyful ode to automatic orchestration," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, p. 18, 2016.

[30] D. Reed, "A perceptual assistant to do sound equalization," in *5th International Conference on Intelligent User Interfaces*, pp. 212–218, ACM, 2000.

[31] B. De Man *et al.*, "The open multitrack testbed," in *137th Audio Engineering Society Convention*, 2014.

[32] R. M. Bittner *et al.*, "Medleydb: A multitrack dataset for annotation-intensive mir research," in *15th International Society for Music Information Retrieval Conference*, vol. 14, pp. 155–160, 2014.

[33] ITU, *Recommendation ITU-R BS.1770-3: Algorithms to measure audio programme loudness and true-peak audio level*. Radiocommunication Sector of the International Telecommunication Union, 2012.

[34] Y. Bengio *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.