Deep Neural Networks for Dynamic Range Compression in Mastering Applications

4 authors:

Konstantinos Drossos
Tampere University
50 PUBLICATIONS **279** CITATIONS
SEE PROFILE

Tuomas Virtanen
Tampere University
270 PUBLICATIONS **8,277** CITATIONS
SEE PROFILE

Gerald Schuller
Technische Universität Ilmenau
112 PUBLICATIONS **1,197** CITATIONS
SEE PROFILE

Stylianos Ioannis Mimilakis
Fraunhofer Institute for Digital Media Technology IDMT
31 PUBLICATIONS **193** CITATIONS
SEE PROFILE

Deep Neural Networks for Dynamic Range Compression in Mastering Applications

Stylianos Ioannis Mimilakis[1], Konstantinos Drossos[2], Tuomas Virtanen[2], and Gerald Schuller[1]

[1] *Fraunhofer IDMT, Ilmenau, Germany*
[2] *Audio Research Group, Dept. of Signal Processing, Tampere University of Technology, Tampere, Finland*

ABSTRACT

The process of audio mastering often, if not always, includes various audio signal processing techniques such as frequency equalisation and dynamic range compression. With respect to the genre and style of the audio content, the parameters of these techniques are controlled by a mastering engineer, in order to process the original audio material. This operation relies on musical and perceptually pleasing facets of the perceived acoustic characteristics, transmitted from the audio material under the mastering process. Modelling such dynamic operations, which involve adaptation regarding the audio content, becomes vital in automated applications since it significantly affects the overall performance. In this work we present a system capable of modelling such behaviour focusing on the automatic dynamic range compression. It predicts frequency coefficients which allow the dynamic range compression, via a trained deep neural network, and applies them to unmastered audio signal served as input. Both dynamic range compression and the prediction of the corresponding frequency coefficients take place inside the time-frequency domain, using magnitude spectra acquired from a critical band filter bank, similar to human's peripheral auditory system. Results from conducted listening tests, incorporating professional music producers and audio mastering engineers, demonstrate on average an equivalent performance compared to professionally mastered audio content. Improvements were also observed, when compared to relevant and commercial software.

I Introduction

Audio production often includes a final stage of process which is placed just before the stage of replication and commercial distribution of the audio material. It is entitled mastering and involves a series of audio signal processing algorithms, aiming to provide an overall audio enhancement in order to link the professional audio with the hi-fidelity / home-entertainment industries [1].

Mastering consists of two main signal processing methods: i) equalisation of the frequency content, and ii) dynamic range control. These two operations require a considerable amount of parameters that have to be defined and controlled, in order to process the audio signals. Main ambition of this processing is to aesthetically enhance perceived acoustic characteristics of the signals [2]. The selection and the adjustment of these parameters relies solely on a continuous interaction between the audio / mastering engineer and the apparatus that handles the audio signals.

During the above interaction takes place an acoustic monitoring of the processed audio, driven through the mastering apparatus. The aforementioned parameters are adjusted until convergence to the desired result, based on auditory feedback and a set of subjective criteria, which are dependent on musical facets of the audio corpus. On one hand this fact imposes an extensive human effort but, on the other, it is the essence of a successful procedure. Consequently, these criteria have been proved to be substantial in audio production [3] and especially in the design of intelligent systems that automatically perform various tasks in different stages of audio and music production [4], i.e. audio mixing or mastering.

There have been published works concerned with providing automated solutions to the above-mentioned time consuming routines [2]. They aim to unveil a correlation between various audio signal features contained inside the original audio material and the one processed by the engineer [3, 5]. In most cases though, the focus is in automated processes of audio mixing where observations of the independent channels and the target mixture signals are available [3, 4, 5].

For automated procedures in audio mastering, where only the original (unmastered) and processed (mastered) audio mixtures are available, only two approaches exist. The first tries to exploit statistical properties of the tracked fundamental frequency of the audio content, in order to derive a set of frequency bands that will be enhanced [6]. In that case the fundamental frequency was extracted from the time-domain representation of the unmastered audio signals. The extracted information was then used to compute histograms and the most prominent observations of frequencies were served as information to second order peaking-type filters, boosting these particular frequency regions.

The second focuses on statistical properties of audio signals which are used to control parameters for dynamic range compression [7]. In more detail, it takes into account that dynamic range compression significantly modifies the probability density function (PDF) of the root mean square energy of the audio signal. Thus, by minimizing the difference of the PDFs between the mastered and unmastered audio signals, in short time frames, parameters for the dynamic range control can be acquired [7].

These two approaches can be understood as an operation of simulating the process of audio mastering by a recording or audio mastering engineer. It is non trivial to define a feature space which will model such complex and adaptive operations. Neither fundamental frequency nor basic statistical properties could sufficiently yield enough information for complex modelling purposes, especially when the prior knowledge of the audio corpus is limited, i.e. the observed two channel mixtures before and after the processing.

A solution to the imposed difficulty from the limited knowledge of the feature space could be given by factorization techniques and especially non-negative matrix factorization (NMF) [8]. Its application to observed mixtures of audio magnitude spectral representations can provide decompositions of the individual components consisted inside the mixture. In addition to this, the signal representation obtained by NMF also allows various implementations of audio signal processing techniques [9]. Nevertheless, in the case of audio mastering, where much dynamic range compression and gain processes are usually applied [1], different probability distributions should be assumed in the NMF model, resulting into a much more complicated model [10].

Deep neural networks (DNNs) seem to offer a straightforward method that encompasses the benefits from the above techniques [11, 12]. Especially with their capabilities in learning non-linear mappings from low-level features to high-level ones [11]. More specifically, a fundamental architecture of DNNs entitled *autoencoders* is capable of establishing various associations of the presented data in an unsupervised fashion similarly to NMF, while these auto-associated representations can be served as features that provide predictions or solutions to a specific problem [12].

In this work we try to expand the existing technologies for automated mastering process by proposing a novel system for off-line automated dynamic range compression. Our system is based on a DNN formed by two pre-trained fully connected autoencoders. In particular, we try to map low-level, magnitude features to dynamic range compression factors, in such a way that it simulates the aesthetics of dynamic range processing in audio mastering. This mapping is performed by a trained DNN and is later used to compute gain factors that modify the input magnitude spectra.

The rest of this paper is organized as follows. Section II gives a detailed overview of the proposed system. Section 3 describes the experimental procedure followed for training the DNN. Obtained results are presented and discussed in Section 4. Section 5 concludes the paper and proposes possible feature directions of research.

II Proposed System

The proposed system consists of two components. The first one is responsible for spectral analysis and synthesis of the input audio signal while the second is responsible for the prediction and utilization of necessary factors that will be used to transform the original spectra.

The stages of analysis and synthesis consist of short-time Fourier transformation (STFT) and its inverse (ISTFT), followed by an overlap and add operation. From the output of the analysis stage, the magnitude information is given to the second component, while the phase is kept for the re-synthesis stage.

The second component utilizes the imported magnitude information, warps its linear frequency resolution by a filter bank and then drives it through a DNN, yielding exponent coefficients, which are then used to transform the warped spectra. Both transformed and unprocessed warped spectra are being interpolated to their original linear scale resolution, with their ratio providing estimations of gain factors.

Finally, the gain factors are used to transform the complex spectra captured by the first component and then proceed with the time-domain synthesis of the corresponding signal. An illustration of the proposed system is being given in Figure 1.

The detailed processing in the proposed system operates as follows. A time domain signal $x(n)$, with discrete samples n, is transformed into a two-dimensional time-frequency representation $X_[m,k]$, under the assumption that $x(n)$ is stationary inside m short time frames and independent over k sub-band channels / frequency bins. To do so, a STFT is used by evaluating Equation 1 for $0 \leq k < N$:

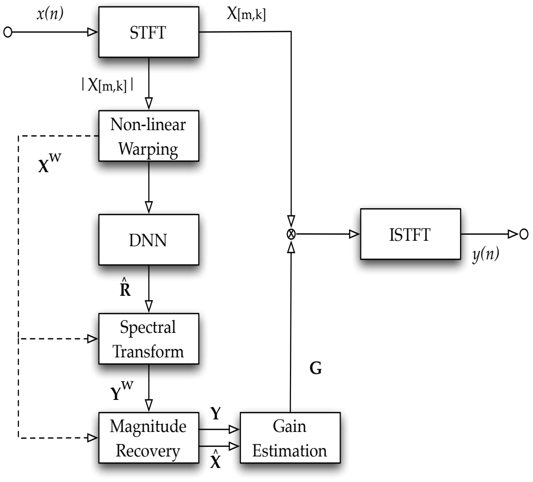

Fig. 1: System Overview.

$$X_{[m,k]} = \sum_{n=-N/2}^{N/2-1} w(n)x(n-mR)e^{-j2\pi kn/N}. \tag{1}$$

In the above equation N denotes the number of samples for the discrete Fourier transformation (DFT), R the analysis step size and $w(n)$ a *hanning* windowing function. The resulting representation has a linear frequency resolution. Our interest is to investigate and model a perceptual process. For these reasons, the magnitude information $|X_{[m,k]}|$ is warped to a non-linearly scaled frequency resolution, denoted as X^w. This scaled frequency resolution, includes information of critical frequency bands, similar to human's peripheral auditory system.

It has to be noted that we are concerned with an offline process, thus X^w is a matrix containing all time frames m over the warped sub-bands c, derived from the input audio signal. As for the warping procedure, it is performed in two steps: i) compute triangular frequency responses for each sub-band of the linear frequency resolution, that form a matrix W and ii) perform a matrix muliplication between the basis functions and the magnitude spectra defined as :

$$X^w = W|X|. \tag{2}$$

The dimensions of matrix W are $C \times M$, with C and M being the total number of sub-bands and short time frames, respectively. The center frequencies and bandwidths employed for the basis functions, according to the human's peripheral auditory system and [13], are given by:

$$b_c = 0.108 f_c + 24.7\text{Hz}, \tag{3}$$

where

$$f_c = 229[10^{(a_1 c + a_0)/21.4} - 1] \tag{4}$$

and c is an integer denoting the sub-band index and $c = 0, 1, 2, \ldots, C-1$. $a_0 = 1.5$ and $a_1 = 0.79$ are constants that determine the centre frequency of the lowest band and the band density in critical bandwidth units, respectively.

Then X^w is used as an input to the trained DNN which outputs estimations of the exponent factor $\hat{R}$. The latter will be utilized in next stage for transforming X^w. More specifically, the estimations are performed by simply feed-forwarding the warped spectra, leading to a series of matrix vector multiplications defined as :

$$\begin{aligned} h^l{}_{ij} &= g(X_i^w W^l{}_{ij} + b^l{}_j) \\ \hat{R}_{ij} &= g(h^l{}_{ij} W^L{}_{ij} + b^L{}_j) \end{aligned} \tag{5}$$

where l is an index of the corresponding layer of the network ($l = 1, \cdots, L$), g an activation function, which in this work is the rectified linear (ReLU), and W^l and b^l are the weights and biases of each layer l, respectively. The index i corresponds to a vector containing short time frames, matching the input dimensions of the DNN and j the dimensions of the hidden layer representation h^l.

The predicted coefficients $\hat{R}$ are in a matrix form of the same dimensions as X^w. Then, the transformation is performed by raising all the elements of X^w to the power of $\hat{R}$. For computing the gain factors G both warped spectra X^w and Y^w must be transferred to the original linear scale. This can be performed using Equation 6.

$$\begin{aligned} Y &= W^T Y^w \\ \hat{X} &= W^T X^w \end{aligned} \tag{6}$$

Gain factors G can now be computed by the element-wise division of the above quantities, leading to:

$$G = f_s(\frac{Y}{\hat{X}}) \tag{7}$$

where f_s is a bounding sigmoid function, which will ensure a distortion free reconstructions, defined as:

$$f_s(x) = \frac{2}{1+\exp(-bx)} - 1, \text{ for } b = 2. \tag{8}$$

Finally, an element-wise multiplication between the computed gains and the original complex spectra is performed followed by the ISTFT and overlap-add synthesis procedure. In case of multichannel audio input, the prediction is performed using the average, over the number of channels, magnitude spectra while the gain is applied to all input channels.

III Experimental Procedure

The experimental procedure is divided in two stages. The first one is concerned with the training procedure of the DNN, including training data preparation, network topologies and the strategies followed, in order to perform the mapping from low level acoustic features to the factors R. The latter stage, consists of the preparation of another audio corpus, containing processed files from various operations including professional ones and from commercial software.

III.a Training Procedure

The overall training process is performed in three steps. The first two incorporate an unsupervised learning approach and the third one, henceforth called fine-tuning, is done in a supervised fashion. During the fine-tuning step, the input and target functions are matrices of the same dimensions that contain the warped spectra X^w and true estimates R, respectively. These are given as objectives to the DNN.

In order to acquire the target function we implemented an iterative analysis of the training dataset which was acquired from an online dataset [14]. The latter contains both mastered and unmastered versions of audio tracks from various genres. Thus, for each version, i.e. mastered and unmastered, of all the audio tracks we computed X^w with the described methodology. By having analysed pairs of unmastered and mastered audio signals, their logarithmic ratio $R = log10(Y^w)(log10(X^w))^{-1}$ can provide the dynamic range factor for the corresponding frequency subbands [15, 9].

In practice, mapping X^w to the dynamic range factor resulted in a poor function fitting. In addition to this, it was experimentally observed that time fluctuations of magnitude spectra would also penalise the fitting procedure in an undesired manner. For dealing with the mapping issue, two prior steps of unsupervised learning relying on autoencoders were introduced. With this technique the initial parameters for the DNN, in fine-tuning stage, can be learned and thus resulting a better convergence to the desired result.

As for the time fluctuation, the matrices used in the objective of each training instance were reshaped so each column contained five short time frames of X^w. The training procedure consists of the following procedures:

1. Train a deep autoencoder, with four layers of 260 fully connected, ReLU, nodes using X^w as input and target functions.
2. Train a deep autoencoder, with three layers of 260 fully connected nodes using R as input and target functions. For the first two layers, the ReLU activation function g is used. The number of nodes of the hidden layer representation is equal to 350.
3. Construct a new DNN with seven layers in total, using the same dimensions and activation functions as above. Initialize this DNN with the pre-trained parameters W^l and b^l, acquired from the first steps. Train this network with X^w as input and R as target functions, respectively.

Each of the above training procedure was performed over a 150 iterations, i.e. epochs, through the dataset while the parameters updates where performed in a small batch size of 20 matrix rows i. For all the layers l during the first two steps, a uniform distribution was selected to pseudo-randomly initialize all the parameters. The optimization technique used is described in [16] with its criterion set to the mean squared error (MSE).

Finally, both autoencoders, i.e. ones from procedures 1 and 2, where trained using the dropout technique [17]

Table 1: Employed system parameters.

Parameter	Quantity
Window size ($w(n)$)	2049 samples
DFT size (N)	4096 samples
Step size (R)	1025 samples
Number of critical bands (C)	52

with a probability of 0.3 for a neural unit to stop contributing to the training at each epoch. The selection of the aforementioned parameters and techniques was based on informal experimentation and empirical observations. A comprehensive overview of the parameters used throughout all the described procedure can be found in Table 1.

III.b Audio Corpus Preparation and Subjective Evaluation

For the evaluation of the proposed system we utilized a different dataset obtained from an online source [18]. This consisted of different unmastered audio tracks in a multi-channel form, which can be categorized to various music genres, e.g. jazz, pop, rock, ethnic, electronic etc. Each audio track was mixed by the authors by the usage of a typical digital audio workstation (DAW). The mixing process yielded four stems (groups) of the aforementioned multiple channels such as vocals, percussion, bass and other.

From these stems we exported two versions of the eight audio tracks. One version contained the mixture of the stems alongside a professional mastering procedure, following guidelines and best practices for dynamic range compression and equalization described in [1, 2]. For the second version only the mixing process was considered. Table 2 demonstrates the utilised apparatus for mixing and mastering the audio corpus.

Table 2: Utilised apparatus

Usage of apparatus	Brand & model
Monitoring System	Audio Technica ATHM40FS
I/O Interface	N.I. Komplete Audio 6
DAW	Pro Tools First

The version which contained only the mixture, was served as input to the proposed system and to one commercial software that is acknowledged to perform automated procedures in audio mastering [19]. In more detail, the software from [19] denoted as AAMS, performs spectral equalization and dynamic range compression for audio mastering purposes, by defining the music genre of the input audio signal. After the genre definition based on descriptions of [18], the automatic procedure took place and the outcome was stored in an uncompressed format.

From the above procedure the three resulting versions, i.e. professionally mastered, processed by the proposed method and by the AAMS software, were segmented into instances of 30 seconds. The segmentation was performed for each individual audio track, but same time regions for all the versions of each track were selected. The criteria for segmentation was the contribution of all the stems to the mixture. In addition to this, all the versions were normalized to have an equal RMS energy, since loudness is outside the scope of this research.

Nine experienced and professional music producers, mixing and mastering engineers, with relevant studies participated in a subjective evaluation experiment. The main objective was to grade each version according to their subjective preference, assuming 1 as the lowest grading point denoting poor performance, and in contrast 10, best performance. All grades were given with respect to the dynamic range and spectral balance of the audio material. A random shuffling of the versions was performed before the experiment, while the amount of playback repetitions and the used monitoring/audio reproducing hardware was subject to each participant. The only requirement was the usage of studio quality headphones.

IV Results & Discussion

Results from the subjective evaluation are illustrated in Figure 2. The lower and upper quartiles are depicted with the lower and upper horizontal lines of each box. Red line indicates median value of grading, while cross denotes an outlier in the observations.

By observing Figure 2 it can been seen that the proposed system performs worse than professional mastering operations, but on average equally well with the AAMS commercial system that we utilized. A closer inspection on the results' figure can also reveal that although our and the AAMS system have an equal median rating, the former exhibits more higher ratings than the latter. The difference of the upper quartiles is at the order

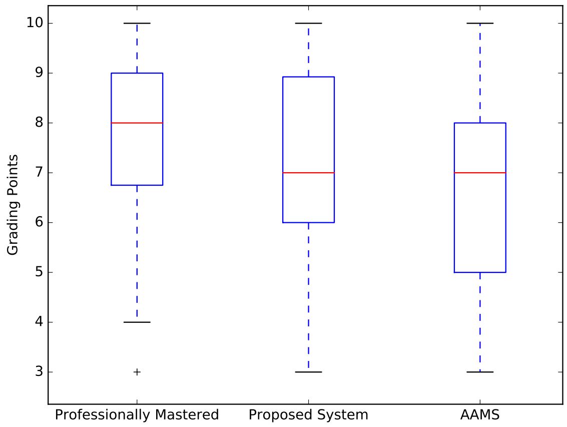

Fig. 2: Variation Analysis of Subjective Grading for the three versions yielded from the corresponding systems.

of one degree in the employed rating scale. This fact clearly depicts that the 25% of the upper ratings were significant higher than the ones of the reference system. A similar trend can be seen on the lower quartile where the proposed system exhibits greater minimum ratings than the AAMS one. The difference of the lower quartile values between the proposed and the reference systems is at the order of 1.5 points in the used rating scale.

Finally, one more interesting observation is that the upper quartile value in Figure 2 for the proposed system is almost the same as the one from the ratings that professional mastered versions had and the lower quartile is less than one rating degree lower from the corresponding one of the professionally mastered versions. This fact clearly demonstrates the improvement in the resulting dynamic range compression and spectral balance from the proposed system over the existing state of the art where the reference system had lower upper quartile at the order of one rating degree and almost two rating degrees smaller value of the lower quartile.

V Conclusions

In the work at hand we focused on automated audio signal processing for audio mastering applications. We utilized DNNs relied on the useful initialization provided by autoencoders, for predicting dynamic range compression and spectral balance enhancement parameters. The latter were automatically applied to unmastered audio tracks. The resulting automated mastered audio material was compared to professionally mastered versions of the same musical compositions. In addition, we also created automated mastered versions, again of the same audio tracks, with another and commercial system for automated mastering.

In order to evaluate our system we compared the abovementioned mastered versions, i.e. the professionally mastered one, from the proposed system and from the reference one, by implementing subjective evaluation tests. In the latter were participating currently active professional master and recording engineers. The results of the subjective evaluation tests depicted that the proposed system achieves an average rating same as the reference one and less than the professionally mastered versions. Nevertheless, the proposed system clearly received more higher ratings than the reference one, as illustrated at the resulting box plots of the subjective evaluation.

Nevertheless, there are significant improvements to be implemented at the existing automated mastering systems in order to achieve a subjective rating similar to the one that a professional mastering engineer would have.

VI Acknowledgements

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement no. 642685 MacSeNet.

References

[1] Owsinski, B., *The Mastering Engineer's Handbook: The Audio Mastering Handbook*, Artistpro, 2nd edition, 2007.

[2] Bob, K., *Mastering Audio: The Art and the Science*, Focal Press, 2nd edition, 2007.

[3] De Man, B., Leonard, B., King, R., and Reiss, J. D., "An Analysis and Evaluation of Audio Features for Multitrack Music Mixtures," in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[4] Reiss, J. D., "Intelligent systems for mixing multichannel audio," in *17th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, Corfu, Greece, 2011.

[5] Ma, Z., De Man, B., Pestana, P. D. L., Black, D. A. A., and Reiss, J. D., “Intelligent Multitrack Dynamic Range Compression,” *J. Audio Eng. Soc*, 63(6), pp. 412–426, 2015.

[6] Mimilakis, S.-I., Drossos, K., Floros, A., and Katerelos, D., “Automated Tonal Balance Enhancement for Audio Mastering Applications,” in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[7] Hilsamer, M. and Herzog, S., “A Statistical Approach to Automated Offline Dynamic Processing in the Audio Mastering Process,” in *Proc. of the 17th International Conference on Digital Audio Effects (DAFx-14)*, pp. 35–40, Erlangen, Germany, 2014.

[8] Févotte, C., Bertin, N., and Durrieu, J.-L., “Nonnegative Matrix Factorization with the Itakura-saito Divergence: With Application to Music Analysis,” *Neural Comput.*, 21(3), pp. 793–830, 2009, ISSN 0899-7667.

[9] Sarver, R. and Klapuri, A., “Application of Non-Negative Matrix Factorization to Signal-Adaptive Audio Effects,” in *Proc. of the 14th Conference on Digital Audio Effects (DAFx-11)*, volume 45, pp. 249–252, Paris, France, 2011.

[10] Simsekli, U., Liutkus, A., and Cemgil, T., “Alpha-Stable Matrix Factorization,” *IEEE Signal Processing Letters*, p. 5, 2015.

[11] Bengio, Y., “Learning deep architectures for AI,” *Foundations and Trends in Machine Learning*, 2(1), pp. 1–127, 2009.

[12] Smaragdis, P., “NMF? Neural Nets? It’s all the same...” `http://youtube.com/watch?v=wfmpViJIjWw`, November, 2015, presentation; Accessed December-2015.

[13] Moore, B. C., editor, *Hearing (Handbook of Perception and Cognition*, Academic Press, San Diego, California, 2nd edition, 1995.

[14] Dimensions, A., “Mastering Audio Samples- before and after mastering.” 2015, online; Accessed December-2015.

[15] Zoelzer, U., *Digital Audio Signal Processing*, John Wiley & Sons, 2nd edition, 2008.

[16] Kingma, D. P. and Ba, J., “Adam: A Method for Stochastic Optimization,” *CoRR*, abs/1412.6980, 2014.

[17] Srivastava, N., Geoffrey, H., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, 15, pp. 1929–1958, 2014.

[18] Senior, M., *Mixing Secrets For the Small Studio*, Focal Press, 2011, online Dataset; `http://cambridge-mt.com/ms-mtk.htm`, Accessed November-2015.

[19] Curioza, S. F., “AAMS: Auto Audio Mastering System,” `http://curioza.com`, 2011.