NEW SONORITIES FOR JAZZ RECORDINGS: SEPARATION AND MIXING USING DEEP NEURAL NETWORKS

Stylianos Ioannis Mimilakis, Estefanía Cano, Jakob Abeßer and Gerald Schuller

Fraunhofer Institute for Digital Media Technology
Ilmenau, Germany
{mis, cano, abr, shl}@idmt.fraunhofer.de

ABSTRACT

The audio mixing process is an art that has proven to be extremely hard to model: What makes a certain mix better than another one? How can the mixing processing chain be automatically optimized to obtain better results in a more efficient manner? Over the last years, the scientific community has exploited signal processing, music information retrieval, machine learning, and more recently, deep learning techniques to address these issues. In this work, a novel system based on deep neural networks (DNNs) is presented. It replaces the previously proposed steps of pitch-informed source separation and panorama-based remixing by an ensemble of trained DNNs.

1. INTRODUCTION

The main goal of the proposed method is to automatically create new mixes from stereophonic jazz recordings. This work goes hand in hand with the previously proposed system for automatic mixing [1] and the recently introduced deep learning approaches to automatic music production [2].

In [1], a framework for automatic mixing of (old) Jazz recordings was presented. It included two steps: a) an initial decomposition of the original mix into solo, backing, and percussive tracks by means of sound source separation algorithms and b) a remixing process using automatic mixing tools. Recently, DNNs were also investigated for their performance in estimating coefficients for dynamic range compression of music content [2].

In this work we propose a replacement of the source separation and the mixing processes by an ensemble of two trained DNNs. The separation process is constrained to solo and backing track estimation and the mixing is constrained to panoramic gain modelling via one-hot vector encoding.

The structure of the document is as follows. Section 2 provides a description of the proposed system and the underlying methodology. Section 3 describes the experimental procedure for training the system followed by Section 4 which concludes the current work.

2. SYSTEM DESCRIPTION

The proposed system is composed of two modules. The first one observes a two-channel time-domain signal and estimates two stems(groups) of musical instruments, henceforth denoted as solo and accompaniment sources. Then, the estimates are served to the second module which is responsible for remixing them. An illustration of the proposed system is given in Figure 1.

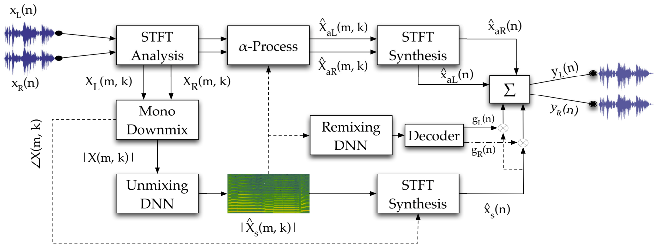

Figure 1: System Overview

More specifically, each channel $x_L(n), x_R(n)$ of a stereophonic mixture is analysed by means of a short-time Fourier transform (STFT). The number of frequency sub-bands(bins) k is set to $N = 4096$ and the computation is performed using a symmetric *bartlett* windowing function covering 2049 samples of the signal to be analysed and 50% overlap between consecutive frames m [1].

The resulting complex representation is down-mixed to one channel, from which the magnitude $|X(m, k)|$ is computed and parsed to the first DNN, which provides estimates for the magnitude spectrum of the solo source $|\hat{X}_s(m, k)|$. The estimation is performed by a series of matrices multiplications, up to the depth of the model, plus transformation operations as described in [3]. The used activation is the ReLU non-linear function, defined as $f(x) = max(0, x)$, a reasonable choice for magnitude spectra. The amount of total layers encapsulated in the current model is equal to 5.

From the estimated $|\hat{X}_s(m, k)|$, the two-channel accompaniment source is estimated through spectral subtraction and generalised Wiener filtering with fractional power spectro-

[1] Since $x_L(n), x_R(n)$ are real signals, their spectra are Hermitian and thus the redundant information is assumed to be discarded. Moreover, each time-frequency sample is treated as an independent random variable.

grams of $\alpha = 1.3$ [4]. The assumption behind this operation, is that the solo source is dominant in both channels. Using the original phase spectra, the time-domain representations of the single-channel solo source $\hat{x}_s(n)$ and the two-channel accompaniment source $\hat{x}_{aL}(n)$, $\hat{x}_{aR}(n)$, are synthesised by the inverse STFT.

With respect to the second module, it accepts as input the reconstructed waveforms and the estimated $|\hat{X}_s(m,k)|$ spectrum. Another *feed-forward* operation, involving matrices multiplications, is attained using $|\hat{X}_s(m,k)|$ and the optimized layers of the second deep model. From the aforementioned operation, a vector is produced for each frame m. This vector consists of probabilities assigned to a codebook. By decoding the codebook two values are derived, one for the panning location in degrees and one for the linear gain applied to the solo source waveform.

It should be noted that the allowed panning and gain values span from $[-45\,^\circ, 45\,^\circ]$ with an increment of 5° degrees, and $[0, 2]$ with an increment of 0.1, respectively. Furthermore, the probabilities are summed with respect to m and the maximum values are selected. The number of total layers existent in this architecture is equal to 3, where the ReLU activation function is used for the first two layers and the *softmax* function for the last one. Finally, the decoded values are used alongside a simple panoramic effect, which applies the corresponding gain functions g_L and g_R to $\hat{x}_s(n)$ to each time domain sample n.

3. EXPERIMENTAL PROCEDURE

For training the above mentioned models, 40 musical compositions from the Jazzomat dataset [2] were used. From each composition the existent fundamental frequency of the solo instrument and the single-channel mixture were collected. The collected data was equally divided in half in order to train the two DNNs equivalently. For the "un-mixing DNN" the approach proposed in [5] was used to acquire estimates of the solo sources. The magnitude spectra of the current estimations were then used through an iterative training procedure, using the back-propagation algorithm, the Euclidean distance as loss function and *adam* as an optimizer for 100 *epochs*. Basically, this procedure can be seen as "de-noising" the mixture spectrogram.

After the training, the first module of the proposed system, which now incorporates the optimized DNN was used to process the rest 20 music tracks. The resulting estimated sources were passed down to an unofficial listening test, in which the participants were asked to provide desired panning locations and mixing gains for remixing the solo source. Using these annotations the second DNN was trained similarly to the first training phase, minimizing the binary *categorical cross-entropy* loss function.

4. CONCLUSIONS

In this work a novel system for automated isolation and remixing of the solo and accompaniment sources from jazz recordings was described. The system is based on state of the art machine learning algorithms, providing a good alternative to cumbersome procedures requiring manual annotations for processing newly observed audio content. Although no formal evaluation took place, auditory examples using segments from jazz mixtures acquired from [6], are available online: `blablabla`. The trained models and the corresponding source code are also available through:
`https://github.com/Js-Mim/aes_wimp`

5. 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. The authors would like to thank the developers of the following Python frameworks : *keras, theano*.

6. REFERENCES

[1] D. Matz, E. Cano, and J. Abeßer, "New sonorities for early jazz recordings using sound source separation and automatic mixing tools," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pp. 749–755, Oct. 2015.

[2] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "Deep neural networks for dynamic range compression in mastering applications," in *Audio Engineering Society Convention 140*, May 2016.

[3] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *CoRR*, vol. abs/1507.06228, 2015.

[4] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brisbane, Australia), IEEE, Apr. 2015.

[5] E. Cano, G. Schuller, and C. Dittmar, "Pitch-informed solo and accompaniment separation towards its use in music education applications," *EURASIP Journal on Advances in Signal Processing*, vol. 23, no. 1, pp. 1–19, 2014.

[6] B. De Man, M. Mora-Mcginity, G. Fazekas, and J. D. Reiss, "The open multitrack testbed," in *Audio Engineering Society Convention 137*, Oct. 2014.

[2] Available from `http://jazzomat.hfm-weimar.de/dbformat/dbcontent.html`