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 methods from 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 (historic / 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 predicting 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 isolation and the mixing is constrained to panoramic gain modelling through a designed codebook, to which angle positions and gain values for mixing the solo source are assigned.

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 this 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 estimated sources are served to the second module which is responsible for synthesizing and re-mixing the sources. An illustration of the proposed system is given in Figure 1.

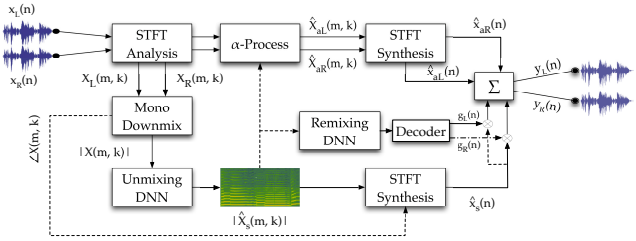

Figure 1: System Overview

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 with approximately 75% overlap between consecutive frames m [1].

The resulting complex-valued representation is mixed down to mono, by averaging the two channels. Then, the magnitude $|X(m,k)|$ is computed and parsed to the first DNN, which estimates the magnitude spectrum of the solo source $|\hat{X}_s(m,k)|$. The estimation is performed by a series of matrix multiplications and transformations, up to the depth of the model, as described in [3]. The used activation is the ReLU non-linear function. The current model has 5 total layers.

Employing $|\hat{X}_s(m,k)|$ and the complex-valued representation of each channel $|X_L(m,k)|$ and $|X_R(m,k)|$, the two-channel accompaniment source is computed via spectral subtraction and generalised Wiener filtering with fractional power spectrograms of $\alpha = 1.3$ [4]. 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 inverse STFT.

Focusing on the mixing procedure, another operation involving matrix multiplications is attained using $|\hat{X}_s(m,k)|$ and the optimized layers of the second DNN. The number

[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.

of total layers existent in this DNN 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.

The output of the second DNN is a vector for each frame m, which consists of probabilities assigned to a manually generated codebook. By decoding the codebook, using the computed probabilities, two values are derived one for the panning location in degrees and one value for the linear gain. These two values are given to a panoramic effect processor, which applies the corresponding gain functions g_L and g_R to each time domain sample n of $\hat{x}_s(n)$. Finally, the stereophonic sources are mixed together.

3. EXPERIMENTAL PROCEDURE

For training the above mentioned models, 40 jazz recordings from the Jazzomat dataset [2] were used. From each composition the existent fundamental frequency of the solo instrument and the single-channel mixture were gathered. The collected data was equally divided in half in order to train the two networks equivalently.

For the first DNN, the approach proposed in [5] was used to derive spectral estimates of the solo sources. These estimates alongside the single-channel mixture signals were then used in an iterative optimization procedure. This procedure included the *back propagation* algorithm and an adaptive gradient descent method (*adam*) [6], for minimizing the Euclidean distance between the predicted, by the DNN, magnitude spectra and the target estimates.

Afterwards, the first module of the proposed system was utilized to process the remaining 20 music tracks. For each one of the separated sources additional annotations were generated by the authors, with respect to the desired panning locations and mixing gains applicable to the the solo source. Using these annotations a codebook was designed and the second network was trained similarly, by minimizing the binary *categorical cross-entropy* loss function this time. The considered panning locations and linear gain values span from $[45°, 45°]$ with an increment of $5°$ and $[0, 2]$ with an increment of 0.1, respectively.

By further processing supplementary audio content from the Jazzomat and the open multi-track [7] datasets with the proposed system, a series of informal listening tests were performed. Results from the listening tests suggest that a good quality of automated mixing for solo instruments can be achieved. Interested readers are welcome to audition examples generated by the described system through: `https://js-mim.github.io/aes_wimp/`.

4. CONCLUSIONS

In this work a novel system for automated separation and mixing of the solo and accompaniment sources from jazz recordings was presented. It is based on state of the art machine learning methods and provides a good alternative to cumbersome procedures that require manual annotations, for processing newly observed audio content. The system was evaluated via informal listening tests, which suggest that an overall good performance for audio re-purposing tasks can be achieved. The corresponding source code can be accessed from: `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 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] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[7] 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`