MUSIC REMIXING AND UPMIXING USING SOURCE SEPARATION

Gerard Roma, Emad M. Grais, Andrew J. R. Simpson, Mark D. Plumbley

Centre for Vision, Speech and Signal Processing
University of Surrey
{g.roma, grais, andrew.simpson, m.plumbley}@surrey.ac.uk

ABSTRACT

Current research on audio source separation provides tools to estimate the signals contributed by different instruments in polyphonic music mixtures. Such tools can be already incorporated in music production and post-production workflows. In this paper, we describe recent experiments where audio source separation is applied to remixing and upmixing existing mono and stereo music content.

1. AUDIO SOURCE SEPARATION USING DEEP NEURAL NETWORKS

Audio source separation algorithms have progressed a long way in recent years, moving on to algorithms that exploit prior information in order to estimate time-frequency masks [1]. For example Deep Neural Networks (DNN), are used in a supervised setting that strongly depends on available training data. In exchange, using supervised training frees them from assumptions needed in other algorithms, such as having recordings from multiple microphones or dealing with repetitive music structures. DNNs are trained to estimate time-frequency masks which still rely on the assumption that energy from different sound sources does not overlap in the time-frequency plane. While applying hard (binary) masks to spectrograms achieves good separation, many noticeable artifacts are introduced. Soft masks produce better sounding results, but imperfect separation. Results from soft masks can still be recombined in remixing and upmixing applications. In this paper we describe two recent prototypes that allow repurposing of musical audio using popular instrument classes. While perceptual evaluation is still pending, both can be used to provide convincing results.

2. REPURPOSING MUSICAL AUDIO

The general idea is to use time-frequency masks estimated from DNN models [2] to upmix and remix musical audio. This means that we are able to make audio content interactive by providing the user with controls for remixing or upmixing, not unlike using an intelligent equalizer that knows about the instrument sounds in the mixture. Our prototypes use models trained using the dataset from the SiSEC MUS challenge [3], where sources have been consistently annotated according to common popular music instrument categories (*vocals*, *bass*, *drums*, *other*). Figure 1 shows a diagram of our remixing prototype, presented at the 2nd Web Audio Conference [4]. In this case, one DNN was trained for each instrument. The predictions of each model are used as probability estimates of the corresponding instrument in each time-frequency bin. One slider controls a global minimum threshold for the estimates. The user can then re-scale the magnitude of each bin with instrument-specific sliders. While total soloing or muting of specific instruments cannot be achieved without artifacts, it is possible to obtain good quality remixes for some parameter settings. Figure 2 shows a diagram of our upmixing prototype, which was demonstrated at the 2nd AES Sound Field Control conference. The demo was delivered in a 22.2 system [5]. Here, a single DNN model was trained to predict soft masks for each instrument. The resulting channels are sent to a Vector Base Amplitude Panning (VBAP) object-based sound spatialization engine (under development by the S3A project [6]) and the user interface allows locating different instruments in 3D space. Informal listening tests revealed

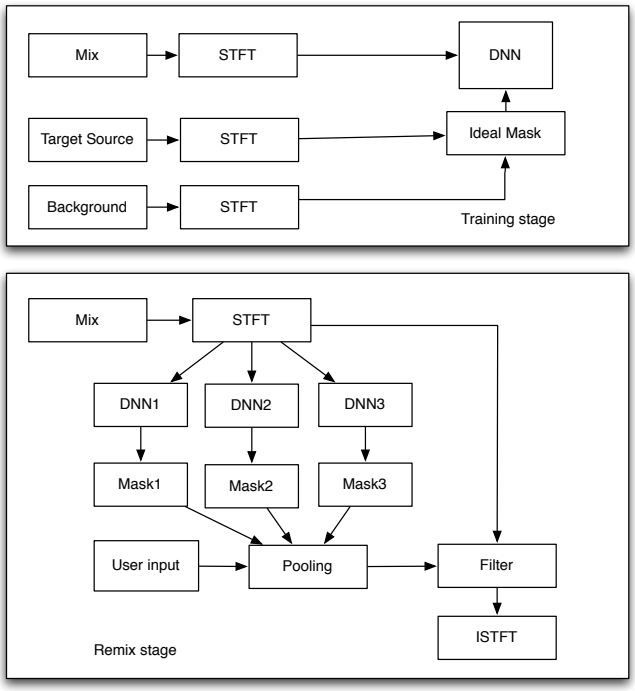

Figure 1: Block diagram of the remixing system.

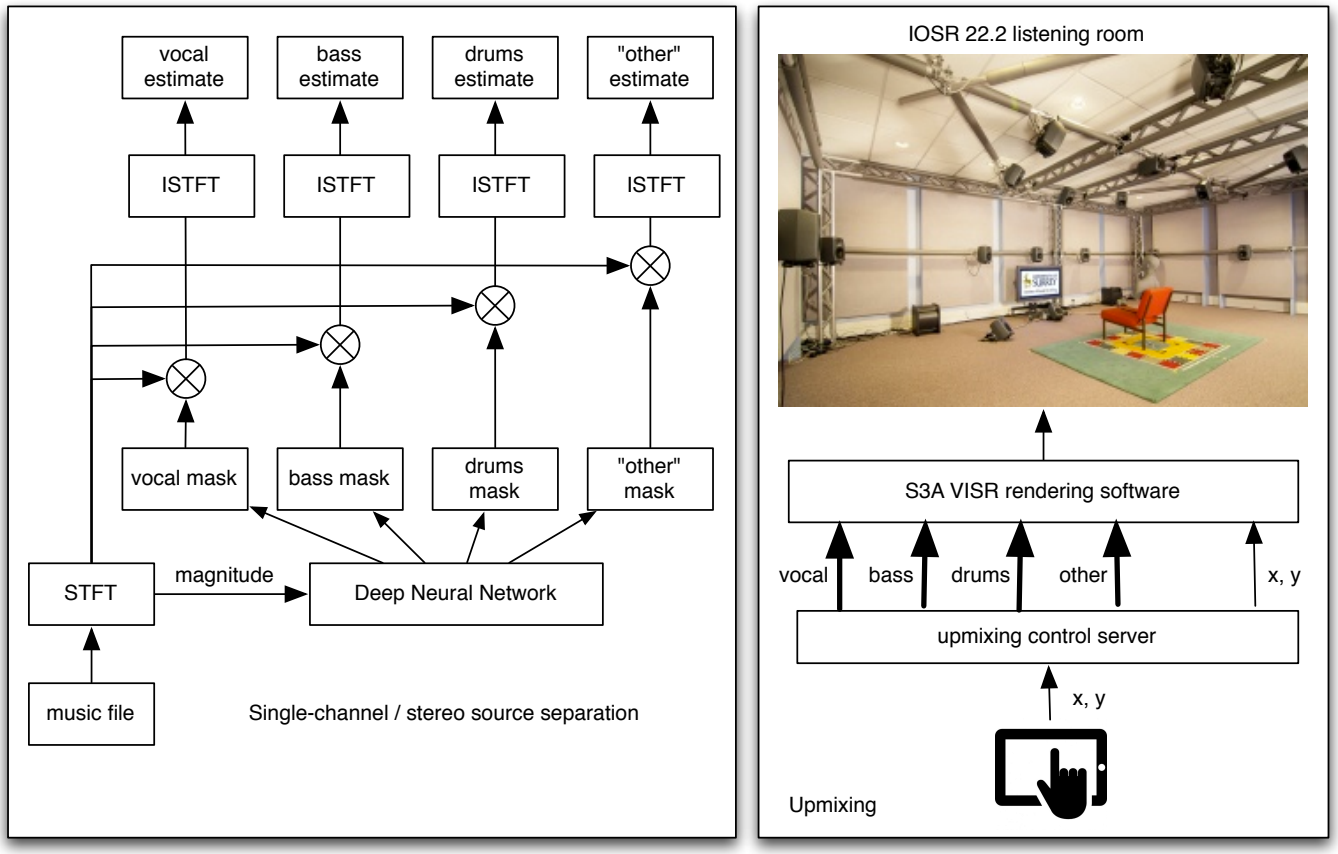

Figure 2: Block diagram of the upmixing system.

that no artifacts were perceived during typical use.

3. CONCLUSIONS

We have described two prototypes for repurposing musical audio. Both systems rely on supervised source separation models trained with a specific set of instrument categories, and thus only work for music with similar instrumentation. Although the quality of separation depends on the difficulty of the mixture, it is possible to constrain the user interface to produce good results by avoiding extreme configurations.

4. ACKNOWLEDGEMENTS

This work was supported by grants EP/L027119/1 and EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).

5. REFERENCES

[1] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[2] 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.

[3] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 387–395, Springer, 2015.

[4] G. Roma, A. J. Simpson, E. M. Grais, and M. D. Plumbley, "Remixing musical audio on the web using source separation," in *Proceedings of the 2nd Web Audio Conference*, 2016.

[5] R. Mason, "Installation of a flexible 3d audio reproduction system into a standardized listening room," in *Audio Engineering Society Convention 140*, May 2016.

[6] "S3A Future Spatial Audio in the Home." http://www.s3a-spatialaudio.org/.