MUSIC SOURCE SEPARATION USING WEAKLY LABELLED DATA

Qiuqiang Kong, Yong Xu, Wenwu Wang, Mark D. Plumbley
Center for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK, GU2 7XH
{q.kong, yong.xu, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

Music source separation (MSS) aims to separate vocal and accompaniment sources from a mixed music recording. Supervised MSS methods need clean separated vocal and accompaniment for training. Instead, we propose a source separation method trained only on weakly labelled audio data, that is, only the presence or absence of vocal or accompaniment is known. The proposed method contains two mappings. The first mapping is modeled by a convolutional neural network (CNN) from the input log Mel spectrogram to the segmentation masks of vocal and accompaniment. The second mapping is from each segmentation mask to the weakly labelled audio tag of vocal and accompaniment. The separated waveform of the vocal and accompaniment sources can be obtained from the segmentation masks. Results show the proposed method achieves average SDR, SIR and SAR of 0.91 dB, 5.99 dB and 4.50 dB in vocal and accompaniment source separation.

1. INTRODUCTION

Music source separation (MSS) [3] aims to separate vocal and accompaniment sources in a mixed music recording. Music source separation can be classified into supervised and unsupervised methods. Unsupervised methods, such as non-negative matrix factorizations (NMFs) [1], learn dictionaries for vocal and accompaniment for source separation. Recently many supervised methods apply deep neural networks to learn a regression from music to vocal and accompaniment [3]. However, these supervised methods need ideal binary masks (IBMs) for training [3]. In this paper, we propose a MSS method trained on *weakly labelled data*, that is, only the presence or absence of the vocal or accompaniment sources are known.

2. PROPOSED SEPARATION FRAMEWORK

We only use weakly labelled audio tags to train the MSS model. To begin with, the log Mel spectrogram of a music clip is used as input feature. The proposed framework contains two parts. The first part is a mapping from the input feature x to the time-frequency (T-F) segmentation masks $h = [h_1, h_2]$, $g_1 : x \mapsto h$, where h_1 and h_2 represents the segmentation masks for music and vocal, respectively. The second part is a mapping from each segmentation mask h_k to the tags $y_k \in \{0, 1\}$, $g_2 : h_k \mapsto y_k$, where k is the index of vocal and accompaniment. In the training phase, the model can be trained end-to-end, that is, from the input feature to the tags of vocal and accompaniment directly.

In the separation phase, the log Mel spectrogram of an unseen music clip is calculated and fed to the trained network to obtain the T-F segmentation masks. Then the T-F segmentation masks are multiplied the spectrogram of the music clip to yield an estimate of the separated source. We then apply the inverse Fourier transform to obtain the separated vocal and accompaniment audio waveforms.

3. EXPERIMENTS

We experiment on the MIR-1K dataset [2], containing 1,000 vocal and accompaniment recording clips. We remixed the dataset so that there are three combinations of audio, (1,0), (0,1), (1,1), representing the presence of only accompaniment, the presence of only vocal and both accompaniment and vocal are present. Training and testing data are split to 80% and 20%. Fig. 1 shows the segmentation masks of a mixed music.

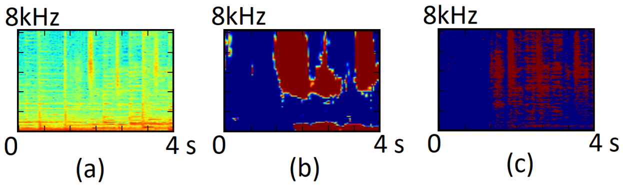

Figure 1. (a) Log Mel spectrogram of a mixed music clip. (b) Learned segmentation vocal mask. (c) Ground truth ideal binary mask (IBM) of vocal.

Table 1 shows the averaged SDR, SIR, SAR [4] of the separated accompaniment and vocal sources. SDR, SIR and SAR without separation and with IBM are provided.

4. CONCLUSION

In this paper, a source separation method only using weakly labelled data is proposed. The segmentation masks

Attribution: Qiuqiang Kong, Yong Xu, Wenwu Wang, Mark D. Plumbley. "Music source separation using weakly labelled data", Extended abstracts for the Late-Breaking Demo Session of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

Table 1. Average SDR, SIR, SAR of separated accompaniment and vocal sources.

	SDR	SIR	SAR
w/o separation	0.12	0.12	58.67
proposed	0.91	5.99	4.50
IBM	12.44	22.15	13.13

of vocal and accompaniment can be learned from the end-to-end training to recover the separated vocal and accompaniment.

5. ACKNOWLEDGEMENT

This research is supported by EPSRC grant EP/N014111/1 "Making Sense of Sounds" and Research Scholarship from the China Scholarship Council (CSC).

6. REFERENCES

[1] E. Grais and H. Erdogan. Single channel speech music separation using nonnegative matrix factorization and spectral masks. In *International Conference on Digital Signal Processing (DSP), 2011*, pages 1–6. IEEE, 2011.

[2] C. Hsu and J. Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, 2010.

[3] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*, pages 1562–1566. IEEE, 2014.

[4] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.