DESIGN OF A CONVOLUTIONAL NEURAL NETWORK FOR SEPARATING SINGING VOICE FROM MONAURAL POP MUSIC

Kin Wah Edward Lin **Enyan Koh** **David Grunberg** **Simon Lui**
Information Systems Technology and Design
Singapore University of Technology and Design, Singapore, 487372
{edward_lin, enyan_koh}@mymail.sutd.edu.sg, {david_grunberg, simon_lui}@sutd.edu.sg

ABSTRACT

We present a specially-designed Convolutional Neural Network (CNN) for separating the singing voice from monaural recordings of pop music. The design principle is that we strive to match the pre-and-post processing procedures with what the network architecture is originally designed for. Our result shows that our CNN outperforms two state-of-the-art Singing Voice Separation systems and the rPCA baseline by $1.7973 \sim 6.1506$ dB GNSDR gain on the iKala dataset. This result encourages a further study of our CNN design on another datasets.

1. MOTIVATION

Although deep learning has been shown to benefit singing voice separation (SVS) systems in two major evaluation campaigns, namely the MIREX [1] and the SiSEC [6], we believe further improvement can be made. We note that though multiple SVS systems used in the aforementioned campaigns include deep learning networks, those networks were often designed for other problems and were later altered for use in a SVS task. We propose that, by instead using a deep learning network which was designed from the ground-up to be suitable for SVS, we can create a system which outperforms the state-of-the-art.

2. OUR CNN DESIGN

2.1 Pre-Processing

The network input for most current SVS is a magnitude spectrogram, either linear or mel-scaled, which is treated as an image. Thus, as the objective of the SVS is to extract the singing voice spectrogram from the mixture spectrogram, we treat the SVS task as the image segmentation.

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

Layer	Configuration	Number of Trainable Parameters (Xavier's initializer [2])
Input	$(9 \times 2049) = 18,441$	
Convolution	$32@(3 \times 12)$, Stride 1 Zero Pad, ReLU [4]	$(3 \times 12) \times 32 + 32 = 1,184$
Convolution	$16@(3 \times 12)$, Stride 1 Zero Pad, ReLU [4]	$(3 \times 12) \times 32 \times 16 + 16 = 18,448$
Max-Pooling	Non-Overlap (1×12) reshapes input size to $(9 \times 171) = 1,539$	
Convolution	$64@(3 \times 12)$, Stride 1 Zero Pad, ReLU [4]	$(3 \times 12) \times 16 \times 64 + 64 = 36,928$
Convolution	$32@(3 \times 12)$, Stride 1 Zero Pad, ReLU [4]	$(3 \times 12) \times 64 \times 32 + 32 = 73,760$
Max-Pooling	Non-Overlap (1×12) reshapes input size to $(9 \times 15) = 135$	
Dropout [7] with probability 0.5		
Fully-Connected	2,048 Neurons, ReLu [4]	$135 \times 32 \times 2,048 + 2,048 = 8,849,408$
Dropout [7] with probability 0.5		
Fully-Connected	512 Neurons, ReLu [4]	$2,048 \times 512 + 512 = 1,049,088$
Output	18,441 Neurons, Sigmoid	$512 \times 18,441 + 18,441 = 9,460,233$
Objective Function: Cross Entropy		Total: $19,489,049$

Table 1. Network Architecture of our CNN

Since SVS is an image segmentation task, we use a Convolutional Neural Network (CNN) proposed in [4], which has been shown to benefit this task. Some state-of-the-art systems instead use a Recurrent Neural Network(RNN) [3] or a bi-directional Long Short Term Memory (BLSTM) Network [8], but this is unnecessary; these systems are designed to capture temporal change, but such change can more simply be captured by providing several consecutive frames as the network input. As such, rather than pass a single frame of the spectrogram at a time into the system like current algorithms do, we pass an excerpt of 9 frames at a time. Our prior work has shown that 9 frames and $4\times$ zero padding factor on FFT Size provide for sufficient temporal and spectral cues [5].

2.2 Network Architecture

Table 1 shows our CNN architecture along with the configuration. We set the objective function to be the cross entropy with sigmoid, and we define the training target as the Ideal Binary Mask (IBM) of the singing voice [10]

2.3 Post-Processing

The network output of current state-of-the-art SVS is assumed to be the estimated magnitude spectrogram of the ground truth singing voice. This assumption is due to the objective function such systems use, namely the Magnitude Spectrum Approximation (MSA). However, this

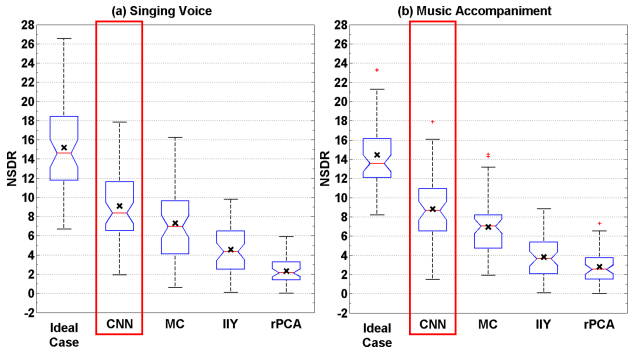

Figure 1. The NSDRs distribution of each SVS algorithm. The marks x indicate the GNSDRs of each SVS algorithm. The ideal GNSDR of the singing voice and the music accompaniment are 15.1944 dB and 14.4359 dB respectively.

makes the usual post-processing steps, such as applying a Wiener Filter to create the ratio mask of the estimated singing voice spectrogram, incomprehensible.

Instead, we define the output of our systems' CNN to be the ratio mask of the estimated singing voice spectrogram. After we reshape the CNN output from $(1 \times 18{,}441)$ to $(9 \times 2{,}049)$, we perform element-wise multiplication between the ratio mask and the mixture magnitude spectrogram, to obtain the estimated singing voice spectrogram mV. To combine these mV into a single spectrogram for a test song, we concatenate the central frames of this mV.

3. EXPERIMENTAL RESULT

We tested our system on the popular iKala dataset [1]. We first took 252 clips of music from that dataset and randomly assigned 152 of them to a training set, 50 to a validation set, and 50 to a testing set. We then used the BSS Eval Ver 3.0 [9] to compare our system to several state-of-the-art algorithms. Quality was measured in global Normalized Source to Distortion Ratio (GNSDR), with higher values of the GNSDR indicating better separation quality.

The results of our experiment are shown in Figure 1. Our CNN achieves the highest GNSDRs for both singing voice and music accompaniment at 9.1045 dB and 8.8042 dB respectively. For the singing voice, our CNN achieve 1.7937 dB higher than MC, 4.6179 dB higher than IIY, 5.4342 dB higher than rPCA. For the music accompaniment, our CNN achieve 1.9362 dB higher than MC, 5.5121 dB higher than IIY, 6.1506 dB higher than rPCA. To further justify that our CNN outperforms the others, we perform a one-way ANOVA. Table 2 summaries the results. It shows our CNN achieves a statistically significant GNSDR difference (< 0.05) compared with the other systems.

4. FUTURE WORK

We will evaluate our CNN design on DSD100 dataset [2] and present the details of dataset usage, audio samples and spectrogram plots at http://people.sutd.edu.sg/∼1000791.

[2] http://sisec17.audiolabs-erlangen.de

Pair	Singing Voice		Music Accompaniment	
	F(1,98)	p-value	F(1,98)	p-value
CNN,MC	5.3481	0.0228	9.2806	0.0030
CNN,IIY	47.9577	4.5855×10^{-10}	76.0115	7.2677×10^{-14}
CNN,rPCA	50.8079	1.7456×10^{-10}	122.6471	5.7306×10^{-19}
MC,IIY	17.9755	5.0706×10^{-5}	35.8675	3.4918×10^{-8}
MC,rPCA	22.838	6.1939×10^{-6}	66.96450	1.0299×10^{-12}
IIY,rPCA	1.5871	0.2107	1.5620	0.2143

Table 2. One-way ANOVA result for the significant difference of GNSDR between each pair of the SVS systems.

5. REFERENCES

[1] T.S. Chan, T.C. Yeh, Z.C. Fan, H.W. Chen, L. Su, Y.H. Yang, and R. Jang. Vocal activity informed singing voice separation with the ikala dataset. In *ICASSP*, pages 718–722, Apr 2015.

[2] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[3] P-S Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *ISMIR*, 2014.

[4] A. Krizhevsky, I. Sutskever, and G.E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[5] K.W.E. Lin, H. Anderson, C. So, and S.Lui. Sinusoidal partials tracking for singing analysis using the heuristic of the minimal frequency and magnitude difference. In *Interspeech*, pages 3038–3042, 2017.

[6] A. Liutkus, F-R Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave. The 2016 signal separation evaluation campaign. In *LVA/ICA*, 2017.

[7] N .Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[8] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In *ICASSP*, March 2017.

[9] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 14(4):1462–1469, Jul 2006.

[10] DeLiang Wang. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*, pages 181–197. Springer US, 2005.

This work is supported by the MOE Academic fund AFD 05/15 SL.