Harmonic and Percussive Source Separation Using a Convolutional Auto Encoder

Wootaek Lim and Taejin Lee
Realistic AV Research Group
Electronics and Telecommunications Research Institute
218 Gajeong-ro, Yuseong-gu, Daejeon, Korea
wtlim@etri.re.kr

***Abstract*— Real world audio signals are generally a mixture of harmonic and percussive sounds. In this paper, we present a novel method for separating the harmonic and percussive audio signals from an audio mixture. Proposed method involves the use of a convolutional auto-encoder on a magnitude of the spectrogram to separate the harmonic and percussive signals. This network structure enables automatic high-level feature learning and spectral domain audio decomposition. An evaluation was performed using professionally produced music recording. Consequently, we confirm that the proposed method provides superior separation performance compared to conventional methods.**

Keywords*— Source Separation, Deep Learning, Auto-Encoder, Convolutional Neural Networks

I. INTRODUCTION

The separation of harmonic and percussive sources is widely used in various applications, such as active listening, audio sources remixing, and pre-processing, which include an automatic description of a pitched signal through the elimination of percussive components. Likewise, decreasing harmonic components can improve the results of percussive source analysis, such as drum beat detection [1]. As shown in Figure 1, most audio signals are composed of harmonic and percussive components, and the goal of this research on harmonic and percussive source separation (HPSS) is to decompose a mixture into a sum of two components and utilize them in many applications.

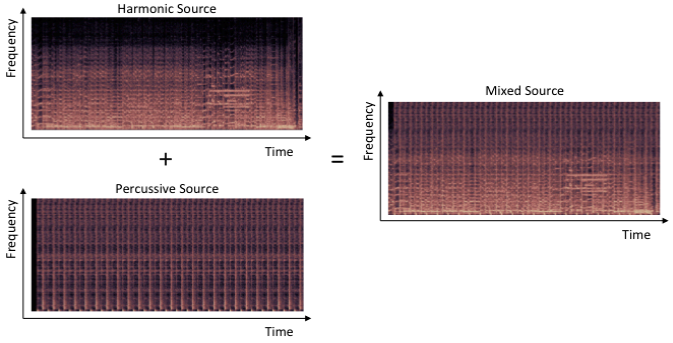


Fig. 1. Examples of harmonic, percussive, and mixed sources.

There have been many previous studies for HPSS, and these studies have assumed that harmonic components have a time continuity characteristic, and percussive components have a frequency continuity characteristic [1, 2]. Ono et al. proposed a simple method for separating monaural audio signals into harmonic and percussive components based on Euclidean distance. This method employed a minimization technique for harmonic and percussive components [3]. They also enhanced their algorithm using an alternative cost function based on the Kullback–Leibler divergence [4]. Another method based on median filtering was proposed in [1]. This approach was fast, simple, and effective in separating the harmonic and percussive components of a monaural audio signal. Furthermore, some improvements in HPSS were reported when using the kernel additive modeling (KAM) method, which generalizes the median filtering method employed in the paper [5, 6]. In addition, a nonlinear filter-based HPSS algorithm was proposed in [7]. However, these previous researches have performance limitations caused by the usage of hand-crafted filters.

The recent development of deep learning algorithms has made significant advances in machine learning technology. Deep learning has become increasingly important because of its significant success in solving complex learning problems. Moreover, these breakthroughs in machine learning affect audio analysis fields, such as speech and music recognition, as well as classification tasks [8, 9]. With this development of deep learning, in this paper, we propose a novel HPSS method based on the convolutional auto-encoder (CAE) that effectively extracts the acoustic features. We also explore the use of CAE for monaural HPSS based on a spectrogram in a supervised setting. In order to evaluate the performance of the proposed method, we compare it to three kinds of conventional HPSS methods. As a result, it has been verified that applying the proposed method to the spectrogram can separate harmonic and percussive signals better than the conventional methods.

The remainder of this paper is organized as follows: Section 2 introduces the proposed architecture of the HPSS system, including a detailed schematic; Section 3 presents the experimental settings and results obtained using a real world music database; and Section 4 draws the conclusions of our paper.

ISBN 978-0-9928626-7-1

II. Proposed Method

We propose a HPSS algorithm that uses a CAE based framework, as shown in Figure 2. First, the spectrogram, which was used as an input for the encoder network, was obtained by applying a Short Time Fourier Transform (STFT) to a mixture audio signal. Next, encoder and decoder networks were simultaneously updated to obtain optimal network coefficients in training step. In the separation step, the CAE network facilitated an acquisition of the masking map for time–frequency (TF) masking. The masked harmonic and percussive signals were restored by applying an inverse STFT (iSTFT).

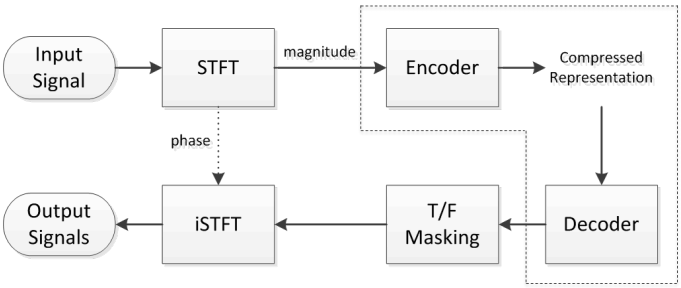


Fig. 2. Proposed scheme for the harmonic and percussive source separation.

A. Convolutional Auto-encoder

The auto-encoder was trained to encode the input data in a compressed representation to reconstruct the original signal [10, 11]. However, a fully-connected auto-encoder does not consider the two-dimensional (2D) structure. On the other hand, the CAE introduced in [12] performs excellently because a convolutional neural network (CNN) learns high-level feature automatically and shows a remarkable performance in computer vision [13] and speech classification tasks [14]. Since the target source is the spectrogram, which is simultaneously indicating time and frequency characteristics, CNN structure is used as a basis for the proposed architecture. In this paper, the encoder network stacks the convolution and max-pooling layers, while the decoder network stacks the convolution and up-sampling layers.

B. Architecture

Figure 3 presents the structure of the CAE employed in our proposed system, which corresponds to the dotted block in Figure 2. The construction of the network architecture must be carefully considered to connect the network from one mixed input audio signal to two separated output signals. As depicted in Figure 3, we used STFT as a 2D representation of an audio signal with a frame size of 1024 and an overlap length of 75%. The spectrogram provided a part of the image measuring 128 × 128 for the input of the network. This 2D block-based acquisition method strides by 32 pixels for every learning sample in both the temporal and frequency axes. The number of node for each layer was 64, 32, and 16. In addition, 2 × 2 max-pooling and up-sampling methods were applied on every convolutional layer of the encoder and decoder networks, respectively. The filter size of the convolution layer was 3 × 3. A rectified linear unit (ReLU) [15] was employed for the nonlinear activation function of the network, and the entire network was optimized using the Adadelta algorithm [16]. The neural network parameters were updated to minimize the squared error between the predicted value $\tilde{y}(t,f)$ and the original spectrogram $y(t,f)$ as shown in Eq. (1), where n is the number of samples. The hyper parameters and settings for the proposed CAE are listed in Table 1.

$$\sum_{i=1}^{n}\left(\left\|\tilde{y}_{h_i}(t,f)-y_{h_i}(t,f)\right\|^2+\left\|\tilde{y}_{p_i}(t,f)-y_{p_i}(t,f)\right\|^2\right) \quad (1)$$

TABLE I. THE HYPER PARAMETERS AND SETTINGS FOR THE CONVOLUTIONAL AUTO-ENCODER

Parameter	Value
Convolution filter size	3 × 3
Max poolong & upsampling size	Vertical = 2, Horizontal = 2
Activation function	ReLU
Loss function	Mean squared error
Optimizer	Adadelta

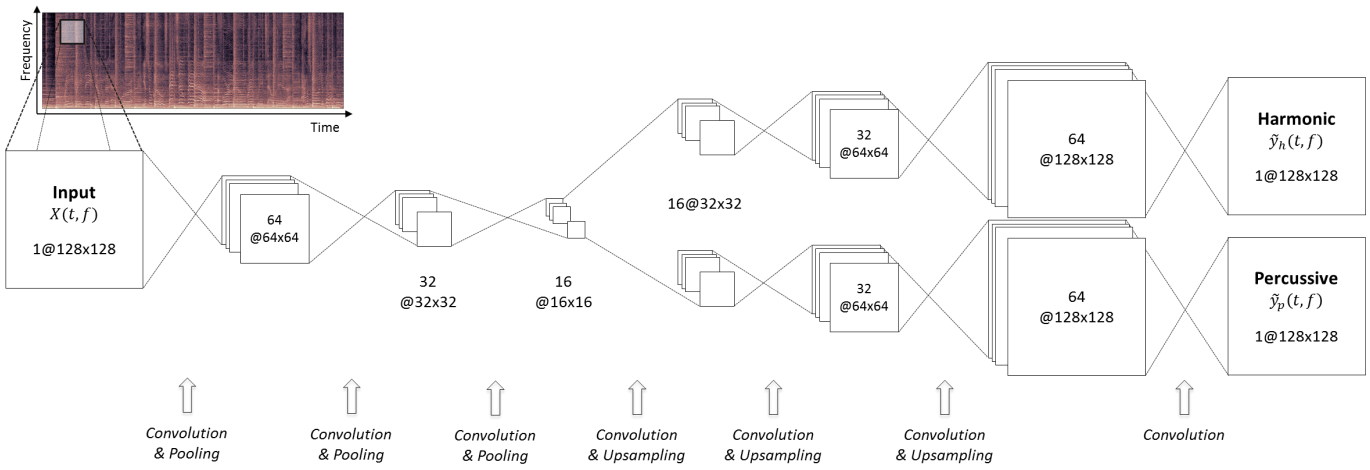


Fig. 3. Proposed neural network architecture using a convolutional auto-encoder.

ISBN 978-0-9928626-7-1 © EURASIP 2017

C. Time-Frequency Masking

Since CAE output values are not equal to the corresponding portions of the original mixture, it is not appropriate to directly use them to separate the harmonic and the percussive components. Therefore, the output signals must be constrained by applying TF masking to the separated signals. In order to ensure proper constraints, we used the soft masking method [17]. The soft TF masking formula is presented as follows:

$$M_h(t,f) = \frac{\left|\widetilde{y}_h(t,f)\right|}{\left|\widetilde{y}_h(t,f)\right| + \left|\widetilde{y}_p(t,f)\right|} \tag{2}$$

where $\widetilde{y}_h(t,f)$ and $\widetilde{y}_p(t,f)$ are the output signals of the harmonic and percussive components from the CAE network, and $M_h(t,f)$ represents the harmonic part masking function of time frame t and frequency f. Then, the magnitude of separated harmonic and percussive sources were acquired from Equations (3) and (4).

$$\widetilde{S}_h(t,f) = M_h(t,f) \otimes X(t,f) \tag{3}$$

$$\widetilde{S}_p(t,f) = (1 - M_h(t,f)) \otimes X(t,f) \tag{4}$$

where $X(t,f)$ is the spectrogram of the mixture audio, $\widetilde{S}_h(t,f)$ and $\widetilde{S}_p(t,f)$ are the estimated separation spectral magnitudes corresponding to the harmonic and percussive sources, respectively. Operator $\otimes$ was an element-wise Hadamard product.

III. PERFORMANCE EVALUATION

In this section, we describe the experimental setting and report the performance of the proposed HPSS based on CAE network. We also compare the performance of the proposed method to that of the three following conventional methods: FitzGerald's Median Filtering based HPSS (MFS) [1], a KAM based HPSS [5, 6], and a conventional stacked deep neural network (DNN) based HPSS.

A. Databases

In order to verify the effectiveness of the proposed CAE network structure, we used two kinds of databases for the experiment. One was the database, which composed of 10 audio clips with mixed harmonic and percussive components from multi-track recordings [18]. In the experiments, it was indicated as DB-1 and we tested the database through the leave-one-sample-out cross-validation.

The other was the De-mixing Secrets Dataset 100 (DSD100), which is a database of professionally recorded music sources, designed to evaluate separation performance for multiple sources. This database consisted of 100 different audio clips, and divided into a development set and a test set. It was indicated as DB-2 in the experiments. We used the development set for training and validation, and the test set for evaluation. All the recorded wave files were approximately 2-7 minutes long [19]. The experiment was repeated five times for each of these two databases.

B. Pre-processing and Network Settings

In order to perform the experiments, audio signals were resampled to 16 kHz sampling rate and mono channel down-mixed. Thereafter, the audio signal was converted into a spectrogram with a frame size of 1024 and an overlap length of 75%. From the spectrogram, we obtained a 128 × 128 image for the input data. Then, each block was positioned in such a way that 75% overlap between adjacent blocks. In other words, the 2D block-based acquisition method strides by N/4 = 32 in both the temporal and frequency axes, as depicted in Figure 4.

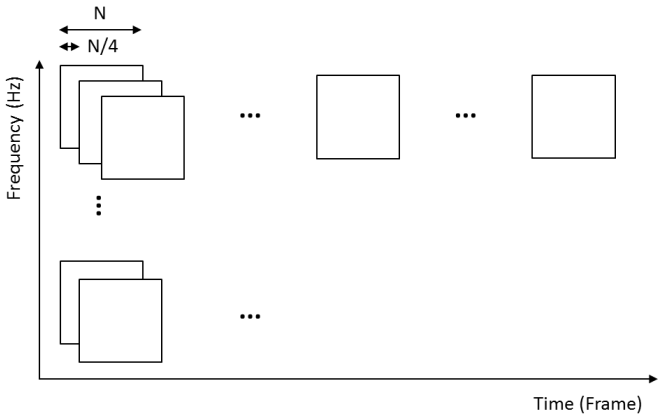
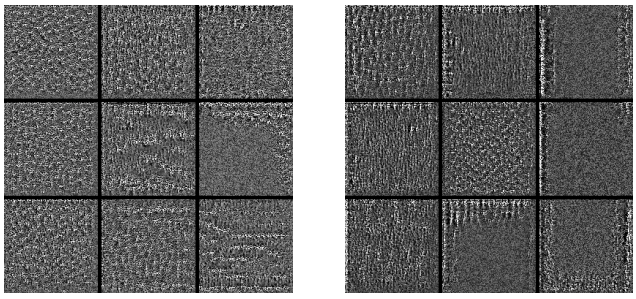


Fig. 4. Overlap structure in the spectrogram for network input.

C. Visualization of the Convolutional Auto-encoder Filters

Filter visualization is one way to ensure that the CNN filters are learning well. In this section, we examined what the CAE actually learns and how it understands the input spectrogram. From the pre-trained network of the CAE, we defined a loss function that seeks to maximize the activation of a specific convolutional filter. The input image was initialized with uniformly distributed random values. Figure 5 presents an exploration of the proposed CAE filters. As shown in Figures 5 (a) and (b), the harmonic layer filters allowed to see horizontal lines, and the percussive layer filters allowed to see vertical lines. This confirms that the CAE filters properly separated the harmonic and percussive components.



(a) Filters on the harmonic layer (b) Filters on the percussive layer

Fig. 5. Exploration of the convolutional auto-encoder filters.

ISBN 978-0-9928626-7-1 © EURASIP 2017

D. Experimental Results

To evaluate the performance of proposed algorithm, we performed quantitative analyses, such as Source to Distortion Ratio (SDR), Source to Artifacts Ratio (SAR), and Source to Interferences Ratio (SIR), as implemented in the bss-eval toolbox 3.0 [20]. The higher SDR, SAR, and SIR values indicate a better separation quality. As aforementioned, we used three conventional methods as the baseline for the performance comparison. All evaluation results are presented in Figures 6 and 7. In experiments for MFS, the median filter size is 31. And the frequency band for the experiment of the KAM method is divided into 100, 1000, 3500, and 8000 Hz. The network structure for DNN is [2500-2000-2000-2000-2500] stacked nodes and flatten the spectrogram as an input signal. Figure 6 shows the results of the HPSS for DB-1. The results illustrated that the proposed CAE method achieved the best average separation performance in SDR, SAR, and SIR. Compared to the KAM method using soft TF masking, the proposed model achieved approximately 1.4 dB, 0.9 dB, and 4.0 dB gains in SDR, SAR, and SIR, respectively. The poor results for the DNN indicated that its network structure did not adequately learn the features of the harmonic and percussive signals.

Fig. 6. Separation performance of the harmonic and percussive sources using the DB-1 database.

Figure 7 presents the performance of the HPSS for database DB-2. Similar to that of the DB-1 database, the proposed CAE method had the best performance for the DB-2 database. Compared to the KAM method using soft TF masking, the proposed model achieved approximately 2.1 dB, 1.3 dB, and 5.4 dB gains in SDR, SAR, and SIR, respectively.

Fig. 7. Separation performance of the harmonic and percussive sources using the DB-2 database.

A cross-database experiment was finally conducted to verify that the proposed method could be generalized. We performed the experiment using the networks trained by the opposite databases from previous experiments. That is, we tested DB-2 with networks trained through DB-1, and tested DB-1 with networks trained through DB-2. All results were calculated as mean values. Figure 8 depicts that the proposed method provided superior performance for cross-databases than the conventional methods.

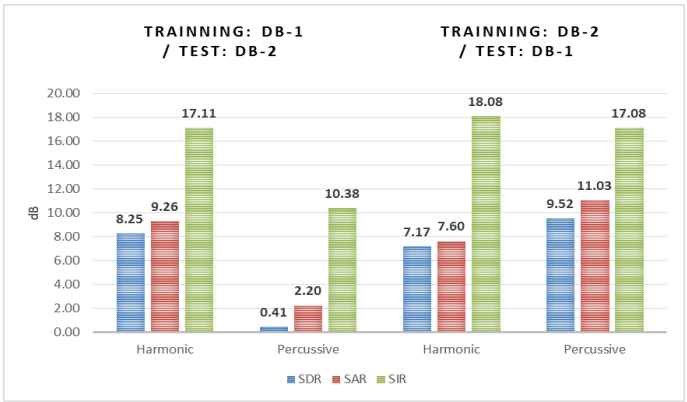


Fig. 8. Cross-DB separation results of the harmonic and percussive sources.

ISBN 978-0-9928626-7-1 © EURASIP 2017

IV. Conclusion

The machine learning field is rapidly expanding because of the innovation in new formalizations of machine learning problems driven by practical applications. This study proposed a novel and effective HPSS algorithm based on a CAE structure. The proposed method achieved automatically generated convolution filters instead of using the conventional hand-crafted filters. The generated filters were learned from the spectrogram as filters that detect well the harmonic and percussive component characteristics. By applying the proposed method to a magnitude of the spectrogram, separated harmonic and percussive sources were obtained and their qualities were analyzed. The proposed method was shown to achieve superior results for the two databases. We also confirmed that the proposed method was generally applicable through the cross-database experiments. Consequently, it was verified that the proposed method provides higher separation quality than the conventional methods.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). [2017-0-00046, Basic Technology for Extracting High-level Information from Multiple Sources Database on Intelligent Analysis]

References

[1] D. FitzGerald, “Harmonic/Percussive separation using median filtering”, Proc. Int. Conf. Digital Audio Effects (DAFx), Graz, Austria, 2010.

[2] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, “Signal processing for music analysis”, IEEE Journal of Selected Topics in Signal Processing., vol. 5, no. 6, pp. 1088-1110, 2011.

[3] N. Ono, K. Miyamoto, J. LeRoux, H. Kameoka, S. Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram”, Proc. European Signal Process Conf. (EUSIPCO), pp. 240-244, 2008.

[4] N. Ono, K. Miyamoto, H. Kameoka, S. Sagayama, “A real-time equalizer of harmonic and percussive components in music signals”, Proc. Int. Conf. Music Inf. Retrieval (ISMIR), pp. 139-144, 2008.

[5] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, L. Daudet, “Kernel additive models for source separation”, IEEE Trans. Signal Processing., vol. 62, no. 16, pp. 4298-4310, Aug. 2014.

[6] D. Fitzgerald, A. Liutkus, Z. Rafii, B. Pardo, L. Daudet, “Harmonic/percussive separation using Kernel Additive Modelling”, Proc. of the 25th IET Irish Signals and Systems Conference, 2014.

[7] A. Gkiokas, V. Papavassiliou, V. Katsouros, G. Carayannis, “Deploying nonlinear image filters to spectrogram for harmonic/percussive separation”, Proc. Int. Conf. Digital Audio Effects (DAFx), 2012.

[8] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning”, Nature, vol. 521, pp. 436-444, May 2015.

[9] Eric J. Humphrey, Juan P. Bello, and Yann LeCun, “Moving beyond feature design: Deep architectures and automatic feature learning in music informatics," Proc. Int. Conf. Music Inf. Retrieval (ISMIR), 2012.

[10] Y. Bengio, “Learning Deep Architectures for AI”, Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[11] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, “Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion”, The Journal of Machine Learning Research., vol. 11, no. 11, pp. 3371-3408, 2010.

[12] J. Masci, U. Meier, D. Cires¸an, J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction”, Proc. Int. Conf. Artificial Neural Networks and Machine Learning (ICANN), pp. 52-59.

[13] A. Krizhevsky, I. Sutskever, G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Proc. Neural Information and Processing Systems, 2012.

[14] H. Lee, P. Pham, Y. Largman, A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks”, Proc. Neural Information and Processing System, 2009.

[15] V. Nair, G.E. Hinton, “Rectified linear units improve restricted boltzmann machines”, Proc. Int. Conf. Machine Learning (ICML), 2010.

[16] M. D. Zeiler, “Adadelta: An adaptive learning rate method”, arXiv preprint arXiv:1212.5701, 2012.

[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, “Deep learning for monaural speech separation”, Int. Conf. Acoustics Speech. and Signal Processing (ICASSP), pp. 1562-1566, 2014.

[18] E. Cano, M. Plumbley, C. Dittmar, “Phase-based harmonic/percussive separation”, Proc. the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2014.

[19] N. Ono and Z. Rafii and D. Kitamura and N. Ito and A. Liutkus, (2015). The 2015 signal separation evaluation campaign. In Latent Variable Analysis and Signal Separation (pp. 387-395). Springer International Publishing.

[20] R. Gribonval, C. Fvotte, E. Vincent, “Performance measurement in blind audio source separation”, IEEE Trans. Speech Audio Language Processing, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

ISBN 978-0-9928626-7-1 © EURASIP 2017