Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders

Emad M. Grais, Dominic Ward, and Mark D. Plumbley
Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, UK.
Email: grais, dominic.ward, m.plumbley@surrey.ac.uk

***Abstract*—Supervised multi-channel audio source separation requires extracting useful spectral, temporal, and spatial features from the mixed signals. The success of many existing systems is therefore largely dependent on the choice of features used for training. In this work, we introduce a novel multi-channel, multi-resolution convolutional auto-encoder neural network that works on raw time-domain signals to determine appropriate multi-resolution features for separating the singing-voice from stereo music. Our experimental results show that the proposed method can achieve multi-channel audio source separation without the need for hand-crafted features or any pre- or post-processing.**

I. Introduction

In supervised multi-channel audio source separation (MCASS), extracting suitable spectral, temporal, and spatial features is usually the first step toward tackling the problem [1]–[3]. The spectro-temporal information is considered imperative for discriminating between the component sources, while spatial information can be harnessed to achieve further separation [4], [5]. The spectro-temporal information is typically extracted using the short-time Fourier transform (STFT), where there is a trade-off between frequency and time resolutions [6]. Computing the STFT to obtain features with high resolution in frequency leads to features with low resolution in time, and vise versa [6]. Most audio processing approaches prefer an auditory motivated frequency scale such as Mel, Bark, or Log scaling rather than a linear frequency scale [7], [8]. However, it is usually not easy to reconstruct the time-domain signals from those type of features. Another common pre-processing step is to take the logarithm of the spectrograms. Despite this, many source separation techniques focus on estimating the magnitude spectra, using the phase of the mixture to reconstruct the time-domain source signals [5], [9]. Unfortunately, omitting phase estimation for the sources usually results in poor perceptual separation quality [10], [11]. Spatial information can be extracted for example from the magnitude and phase differences of the STFT of different spatial channels [4], [5], or by estimating a spatial covariance matrix [1], [2]. All the aforementioned features are hand-crafted features and most of the time we can not have features that are good in representing all the spectral, temporal, and spatial characteristics of different audio sources. There is usually a trade-off between these features.

Instead of humans deciding which features to extract from the audio signals, recently, different deep neural networks (DNNs) have been used to process the time-domain audio signal directly to automatically extract suitable features for each type of audio signal [12]–[15]. In those papers, convolutional layers in the DNNs were capable of extracting useful features from the raw waveform of the input signal. Each convolutional layer in [12]–[15] has filters with the same size, which extract features with a certain time resolution.

In this paper, we propose a novel multi-channel Multi-Resolution Convolutional Auto-Encoder (MRCAE) neural networks for MCASS. Each layer in MRCAE is composed of sets of filters, where the filters in one set have the same size which is different to the sizes of the filters in the other sets. The large filters extract global information from the input signal while small filters extract the local details from the input signal. The features that capture both global and local (multi-resolution) details can help discriminate between different audio sources, which is an essential issue for source separation. The inputs and outputs of the MRCAE are the mixtures and the estimated target sources respectively in the time-domain. The proposed MRCAE is also multi-channel which captures the information in the different channels of the input signals. We do not perform any pre-processing or post-processing operations on the audio signals.

This paper is organized as follows. In Section II, the proposed MRCAE neural network is presented. In Section III, we show how the proposed MRCAE is used for source separation. The remaining sections present the experiments and conclusion of our work.

II. Multi-resolution convolutional auto-encoder neural networks

The proposed multi-resolution convolutional auto-encoder (MRCAE) neural network is a fully convolutional denoising auto-encoder neural network as in [16], but with each layer consisting of a different set of filters. The MRCAE has two main parts, the encoder and decoder. The encoder is used to extract multi-resolution features from the input mixtures and the decoder uses these features to estimate the sources. The encoder and decoder consist of many convolutional and transpose convolutional layers [17] respectively as shown in Fig 1. Each layer in MRCAE consists of different sets of filters, where the filters in one set have the same size and the filters in different sets have different sizes.

ISBN 978-90-827970-1-5 © EURASIP 2018

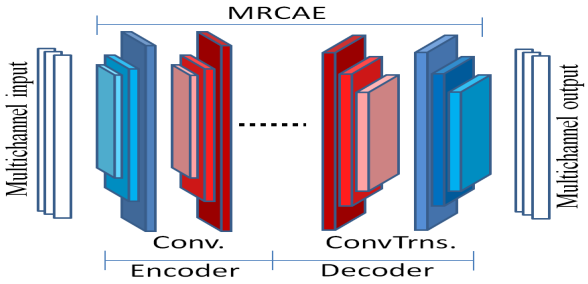

Fig. 1. Overview of the structure of a multi-channel multi-resolution convolutional auto-encoder (MRCAE). "Conv" denotes convolutional layers and "ConvTrns" denotes transpose convolutional layers. Each layer consists of different sets of filters with different sizes.

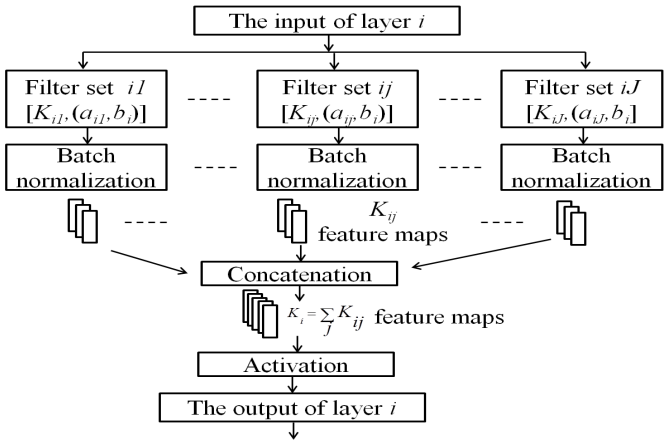

Fig. 2. Overview of the proposed structure of each layer of the MRCAE. Where K_{ij} denotes the number of filters with size $a_{ij} \times b_i$ in set j in layer i, a_{ij} is the length of the filters in the time direction, and b_i is the size of the filters that equals to the number of channels in the input. "Activation" denotes the activation function.

Considering the concept of calculating the STFT of an audio signal, if the STFT window is large, the STFT features capture the frequency components of the signal in high resolution and the temporal characteristics in low resolution [6] and vice versa. STFT can not produce features in high resolution in both time and frequency.

To build a system that is automatically capable of extracting suitable features from the input raw data (time-domain signal) in a suitable time and frequency resolution according to each source in the input mixtures, we propose to use MRCAE, where each layer consists of different sets of filters with different sizes as shown in Fig. 2. This figure shows that at each layer i there are J sets of filters. Each filter set j in layer i has K_{ij} filters with the same size $a_{ij} \times b_i$, where a_{ij} is the filter length and b_i is the number of channels that the input data to layer i has. In each layer i, the value of a_{ij} in set j is different than the value $a_{ij'}$ in set j', but b_i is the same for all sets in the same layer i, because all sets have the same number of channels of the input data to the same layer. Each set j of filters at layer i generates K_{ij} feature maps in a certain resolution and each layer i generates $K_i = \sum_j^J K_{ij}$ feature maps in different resolutions. The K_i is the number of channels for the input data of the next layer.

The long filters with large a_{ij} are good in capturing the global information of the processed signals and the short filters with small a_{ij} can capture the local details. We might think of using long filters as calculating the STFT using a long window, and the short filters as calculating the STFT using a short window. This means using long and short filters together in the same layer produces features with different time-frequency resolutions. This can be very useful for many audio signal processing applications. In MCASS, there are different audio sources in the mixtures and useful information can be extracted for different sources using different time-frequency resolutions that is suitable for different sources [18]. Since the input signal is multi-channel time-domain signal, each filter in the first layer is a multi-dimensional filter to be able to run over the multi-channel input signals.

III. MRCAE FOR MULTI-CHANNEL AUDIO SOURCE SEPARATION

Suppose we have C mixtures each with L sources as $y(t,c) = \sum_{l=1}^{L} s_l(t,c)$, $\forall\ c \in C$, where C is the number of channels and t denotes time. The aim of MCASS is to estimate the sources $s_l(t,c)$, $\forall l, c$, from the mixed signals $y(t,c)$ $\forall c$. In the stereo case, $C = 2$. We work here on the time-domain input and output signals.

In this work, we propose to use a single MRCAE to separate all the target sources from the input mixtures. The inputs for the MRCAE are multi-channel (two channels for the stereo case) segments of the input mixture signals. Each segment has length N of time-domain samples. The corresponding output segments for each target source are also multi-channel with length N samples. The total number of filters in the output layer of the MRCAE should be equal to the number of target sources multiplied by the number of channels for each source. This way we guarantee that the output layer generates feature maps equal to the number of target sources, where each source has its multiple channel components. For example, in the stereo case, if we wish to separate four sources, the number of filters in the output layer should be eight filters.

A. Training the MRCAE for source separation

Let us assume we have training data for the mixed signals and their corresponding target sources. Let $y(t,c)$ be the mixed input signal for channel c and $s_l(t,c)$ be the target source l for channel c. The MRCAE is trained to minimize the following cost function:

$$D = \sum_{t,c,l} |z_l(t,c) - s_l(t,c)| \tag{1}$$

where $z_l(t,c)$ is the actual output of the last layer of the MRCAE for source l and channel c, $s(t,c)$ is the reference

ISBN 978-90-827970-1-5 © EURASIP 2018

target output signal for source l and channel c. The input of the MRCAE is the mixed signals $y(t, c),\ \forall c$.

B. Testing the MRCAE for source separation

The multi-channel mixture is passed through the trained MRCAE. The output of each filter in the last layer is considered to be the time-domain estimate of one of the channels c of one of the sources l.

IV. Experiments

We applied our proposed MRCAE approach to separate the singing-voice/vocal sources from a group of songs from the SiSEC-2016-MUS-task dataset [19], which consists of 100 stereo (two-channels) songs. Each song is a mixture of vocals, bass, drums, and other musical instruments. The first 50 songs in the dataset were used as training and validation datasets, and 46 of the remaining 50 songs were used for testing as four test songs were corrupted. The data were sampled at 44.1kHz.

The quality of the separated vocals was measured using four metrics of the BSS-Eval toolkit [20]: source to distortion ratio (SDR), source image to spatial distortion ratio (ISR), source to interference ratio (SIR), and sources to artifacts ratio (SAR). ISR is related to the spatial distortion, SIR indicates the remaining interference between the sources after separation, and SAR indicates the artifacts in the estimated sources. SDR measures the overall distortion (spatial, interference, and artifacts) of the separated sources, and is usually considered the overall performance evaluation for any source separation approach [20]. Achieving high SDR, ISR, SIR, and SAR indicates good separation performance.

In the training stage of the MRCAE, the time-domain samples of the 50 signals for the input mixtures from the training set were normalized to have zero mean and unit variance. The normalized input mixtures and their corresponding target vocal source were then divided into segments of length 1025 samples. The segments of the input mixtures and the target vocal signals were used to train the MRCAE.

In the test phase, the input signals of each song were divided into 1025 samples with hop size 16 and passed through the trained MRCAE. The outputs of the MRCAE were used with simple shift and add procedures to reconstruct the time-domain signal for the target vocal source. It is worth mentioning that we did not perform any pre- or post-processing on the input or output data other than normalizing the input signals to have zero mean and unit variance.

A. MRCAE structure

The MRCAE consists of two convolutional layers in the encoder part, two transpose convolutional [17] layers in the decoder part, and one output layer as shown in Table I. Table I also shows the number of filter sets, the number of filters in each set, and the length of the filters in each set. The lengths of the filters are analogies for using window sizes of 5, 50, 256, 512, and 1025 in the case of calculating the STFT of the input signal. The short filters capture the local details in high resolution in time, while the long filters capture global information (maybe seen as features with high frequency resolution) of the input signals. Since we separate one source (vocal) with two channels, the output layer of the MRCAE is a transpose convolutional layer with two filters, where each filter generates a feature map corresponding to the estimate of one of the channels of the estimated vocal. Batch normalization was used after each set of filters as shown in Fig. 2. The activation function for all layers is exponential linear unit (ELU) function that allows positive and negative values in its output, which has been shown to speed up the learning in deep neural networks [21]. The length of the input and output segments for the MRCAE was 1025 time-domain samples.

MRCAE model summary. The input/output data with size 1025 samples					
Layer	Encoder		Decoder		Output
1	set 1	Conv[20,(5)]	set 1	ConvTrns[50,(5)]	ConvTrns[2,(1025)]
	set 2	Conv[20,(50)]	set 2	ConvTrns[25,(50)]	
	set 3	Conv[20,(256)]	set 3	ConvTrns[20,(256)]	
	set 4	Conv[20,(512)]	set 4	ConvTrns[20,(512)]	
	set 5	Conv[20,(1025)]	set 5	ConvTrns[20,(1025)]	
2	set 1	Conv[50,(5)]	set 1	ConvTrns[20,(5)]	
	set 2	Conv[25,(50)]	set 2	ConvTrns[20,(50)]	
	set 3	Conv[20,(256)]	set 3	ConvTrns[20,(256)]	
	set 4	Conv[20,(512)]	set 4	ConvTrns[20,(512)]	
	set 5	Conv[20,(1025)]	set 5	ConvTrns[20,(1025)]	

TABLE I
THE NUMBER AND SIZES OF THE FILTERS IN EACH LAYER IN THE MRCAE. FOR EXAMPLE "CONV[20,(5)]" DENOTES CONVOLUTIONAL LAYER WITH 20 FILTERS AND THE LENGTH OF EACH FILTER IS 5. "CONVTRNS" DENOTES TRANSPOSE CONVOLUTIONAL LAYER.

The parameters for the MRCAE were initialized randomly. The MRCAE was trained using backpropagation with gradient descent optimization using Adam [22] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-08$, a batch size 100, and a learning rate of 0.0001, which was reduced by a factor of 10 when the values of the cost function ceased to decrease on the validation set for 3 consecutive epochs. The maximum number of epochs was 20. We implemented our proposed algorithm using Keras with Tensorflow backend [23].

B. Comparison with related works

We compared the performance of the proposed MRCAE approach for MCASS with five different deep neural networks (DNNs) based approaches from the submitted results to the SISEC-2016-MUS challenge [19]. Two of those approaches are the best submitted results in this challenge, known as UHL3 and NUG1 [19], and the three other approaches are known as CHA, KON, and GRA3 in [19]. UHL3 combined different deep feed forward neural networks (FFN) and deep bidirectional long short-term memory (BLSTM) neural networks, with data augmentation from different data set [2]. In UHL3, the spectrogram of the linear combination of the outputs of the models was used to compute spatial covariance matrices to separate the sources from the input mixtures in the STFT domain. The second best approach in the SISEC-2016-MUS challenge was NUG1, which used a deep FFN to find spectrogram estimates for the sources then these estimates were used to compute spatial covariance matrices that were then used to separate the sources in the STFT domain [1].

ISBN 978-90-827970-1-5 © EURASIP 2018

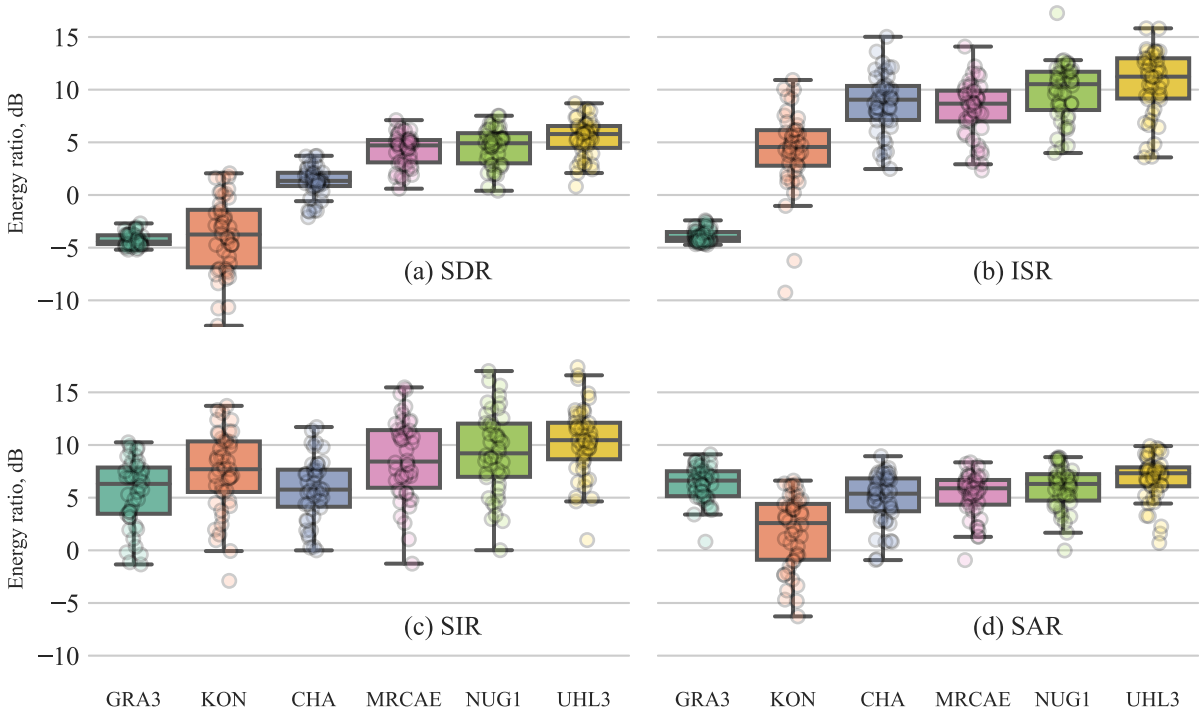

Fig. 3. Boxplots (with individual data points overlaid) of the SDR (a), ISR (b), SIR (c) and SAR (d) BSS-Eval performance measures for our proposed MRCAE and five singing-voice separation systems applied to the SiSEC-2016-MUS test set (46 songs).

NUG1 used the expectation maximization (EM) algorithm to iterate between using the FFN to find spectrogram estimates and updating the spatial covariance matrices to improve the separation quality of the estimated sources. UHL3 and NUG1 stacked numbers of neighbouring frames of the spectrograms of the input mixtures and used principle component analysis (PCA) to reduce the dimensionality of the stacked spectral frames. CHA [24] and KON used deep convolutional neural networks and deep recurrent neural networks respectively to extract the spectrogram of each source from the spectrogram of the average of the two channel input mixtures. GRA3 stacked the magnitude spectrograms of the two channels and used deep FFN to estimate the magnitude spectrograms of the two channels of each source [9].

C. Results

Fig. 3 shows boxplots of the SDR (a), ISR (b), SIR (c) and SAR (d) measures, of the proposed MRCAE method and the aforementioned five other DNN methods from the SISEC-2016-MUS challenge. Considering the SDR as the overall quality measurement, we can see that the proposed MRCAE method, that works by just sending the mixed signals in the time-domain into the trained MRCAE to estimate the time-domain vocal signals, works better than CHA, KON, and GRA3 which all used STFT and different DNNs to estimate the sources. The performance of MRCAE in SDR, SIR, and SAR is not too far from UHL3 and NUG1 methods. The main advantage of our proposed approach over UHL3 and NUG1 is dealing with the raw data without any pre- or post-processing of the input and output signals. In contrast to our method, both UHL3 and NUG1 require many pre- and post processing such as: computing STFT and dealing with complex numbers, stacking numbers of neighbouring spectral frames, using PCA for dimensionality reduction, computing spatial covariance matrices, combining different DNN outputs, data augmentations, and iterative EM algorithm. The results in Fig. 3 shows that our proposed approach of using MRCAE for MCASS is very promising. In our future work, we hope that by refining the MRCAE parameters and exploring other cost functions different to Eq. 1, we can improve the separation quality of our system.

Table II shows the across-song medians of the BSS-Eval measures for the estimated sources using the proposed MRCAE, most of the submitted approaches to SiSEC-2016-MUS challenge [19], and the input mixtures. The order of the methods in Table II is based on the SDR values. DUR [25], KAM [26], OZE [27], RAF3 [28], JEO2 [29], and HUA [30] are blind source separation approaches. STO1 [31] is supervised source separation approach based on feed-forward DNN architecture using patched overlapped STFT frames on input and output. MIX corresponds to the original unprocessed input mixtures. According to the median SDR values, our proposed MRCAE outperforms most of the other approaches except UHL3 and NUG1. The difference in median SDR

ISBN 978-90-827970-1-5 © EURASIP 2018

between MRCAE and UHL3 is -1dB, and between MRCAE and NUG1 is -0.2dB. Audio examples of source separation using MRCAE are available online[1].

Method	SDR	ISR	SIR	SAR
UHL3	5.79	11.23	10.46	7.32
NUG1	4.91	10.52	9.21	6.30
MRCAE	4.71	8.67	8.43	5.89
STO1	4.23	8.07	8.44	5.42
JEO2	4.20	8.76	7.01	5.91
KAM1	2.11	5.98	9.85	1.09
RAF3	1.92	8.60	1.42	6.46
OZE	1.85	5.46	3.75	2.18
DUR	1.36	1.57	5.14	2.86
CHA	1.34	9.05	5.77	5.38
KON	-3.75	4.56	7.70	2.59
HUA	-4.14	15.05	-2.43	7.99
GRA3	-4.43	-4.05	6.31	6.62
MIX	-6.40	31.23	-6.42	248.85

TABLE II
THE MEDIAN VALUES FOR THE BSS-EVAL MEASURES FOR OUR PROPOSED MRCAE, MOST SUBMITTED SYSTEMS TO THE SISEC-2016-MUS, AND THE INPUT UNPROCESSED MIXTURES OF THE TEST SET.

V. CONCLUSION

In this paper, we proposed a new multi-channel audio source separation method based on separating the waveform directly in the time-domain without extracting any hand-crafted features and without any pre- or post-processing. We introduced a novel multi-resolution convolutional auto-encoder neural network to separate the stereo waveforms of the target sources from the input stereo mixed signals. Our experimental results show that the proposed approach is very promising. In future work we will investigate combining the multi-resolution concept with generative adversarial neural networks (GANs) for waveform audio source separation.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[2] 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 *Proc. ICASSP*, 2017, pp. 261–265.

[3] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[4] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–18, 2016.

[5] A. Zermini, Q. Liu, X. Yong, M. Plumbley, D. Betts, and W. Wang, "Binaural and log-power spectra features with deep neural networks for speech-noise separation," in *Proc. International Workshop on Multimedia Signal Processing*, 2017.

[6] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 1, pp. 236–243, 1984.

[7] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using Gammatone filterbank and itakurasaito nonnegative matrix two-dimensional factorizations," *IEEE Trans. on Circuits and Systems I*, vol. 60, no. 3, pp. 662–675, 2013.

[8] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," in *arXiv:1709.04396v1*, 2017.

[9] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Single channel audio source separation using deep neural network ensembles," in *Proc. 140th Audio Engineering Society Convention*, 2016.

[10] M. Dubey, G. Kenyon, N. Carlson, and A. Thresher, "Does phase matter for monaural source separation?" in *Proc. NIPS*, 2017.

[11] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1–10, 2014.

[12] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. InterSpeech*, 2015.

[13] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. ICASSP*, 2014, pp. 6964–6968.

[14] S. Fu, Y. Tsao, X. Lu, and H. Kawais, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," in *arXiv:1709.03658*, 2017.

[15] Y. Hoshen, R. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Proc. ICASSP*, 2015.

[16] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. GlobalSIP*, 2017.

[17] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," in *arXiv:1603.07285*, 2016.

[18] A. J. Simpson, "Time-frequency trade-offs for audio source separation with binary masks," in *arXiv:1504.07372*, 2015.

[19] A. Liutkus, F. Stoter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, 2017, pp. 323–332.

[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–69, Jul. 2006.

[21] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *arXiv:1511.07289*, 2015.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. arXiv:1412.6980 and presented at ICLR*, 2015.

[23] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[24] P. Chandna, M. Miron, J. Janer, and E. Gomez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. LVA/ICA*, 2017, pp. 258–266.

[25] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Trans. on on Selected Topics on Signal Processing*, vol. 5, no. 6, pp. 1118–1133, Oct. 2011.

[26] A. Liutkus, D. FitzGerald, Z. Rafii, and L. Daudet, "Scalable audio separation with light kernel additive modelling," in *Proc. ICASSP*, 2015, pp. 76–80.

[27] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, Oct. 2012.

[28] Z. Rafii and B. Pardo, "REpeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 71–82, Jan. 2013.

[29] I.-Y. Jeong and K. Lee, "Singing voice separation using RPCA with weighted l1-norm," in *Proc. LVA/ICA*, 2017, pp. 553–562.

[30] P. Huang, S. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP*, 2012, pp. 57–60.

[31] F.-R. Stoter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proc. ICASSP*, 2016, pp. 126–130.

[1] https://cvssp.github.io/maruss-website/publications/Grais_2018.html