Audio source separation with deep neural networks using the dropout algorithm

Alfredo Zermini, Wenwu Wang, Qiuqiang Kong, Yong Xu, Mark D. Plumbley
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK
E-mails: {a.zermini, w.wang, q.kong, yong.xu, m.plumbley}@surrey.ac.uk

***Abstract*—A method based on Deep Neural Networks (DNNs) and time-frequency masking has been recently developed for binaural audio source separation. In this method, the DNNs are used to predict the Direction Of Arrival (DOA) of the audio sources with respect to the listener which is then used to generate soft time-frequency masks for the recovery/estimation of the individual audio sources. In this paper, an algorithm called 'dropout' will be applied to the hidden layers, affecting the sparsity of hidden units activations: randomly selected neurons and their connections are dropped during the training phase, preventing feature co-adaptation. These methods are evaluated on binaural mixtures generated with Binaural Room Impulse Responses (BRIRs), accounting a certain level of room reverberation. The results show that the proposed DNNs system with randomly removed neurons is able to achieve higher SDRs performances compared to the baseline method without the dropout algorithm.**

I. Introduction

Sound source separation techniques aim to estimate speech or sound sources from a mixture of speech or sound signals. The current paper is based on the previous work in [1], [2], [3] and [4], where a DNNs based separation system was proposed for stereo source separation problem. Each DNN is made by two deep autoencoders and one softmax classifier, trained independently with a final fine-tuning stage, for a total of four stages. The DNNs are used to extract high-level features and allow the prediction of the DOAs of the audio sources with respect to the listener, which are used to estimate the source occupation probabilities at each T-F point. These are used to generate soft time-frequency masks for the recovery/estimation of the individual audio sources. In this paper, we study the potential of dropout technique for performance improvement of source separation. More specifically, the dropout algorithm [5] [6] is applied to the hidden layers of the DNNs system, and the impact of dropout percentages on source separation performance is investigated.

II. Dropout algorithm

The dropout algorithm is used to force any neuron to rely less on the output of any other neuron and more on the population behavior of its inputs. This helps reducing overfitting, especially when the data set is small [5] [6]. During training, each neuron and its connections can be randomly deleted with a given probability value, while the remaining weights are trained by back-propagation, resulting in a 'thinned' network. During the testing, the weights of the 'thinned' network are averaged in order to get an 'unthinned' network, which has smaller weights.

III. System Overview

The system is composed by several different DNNs, where a higher number of DNNs increases the T-F mask resolution. Each DNN is composed by several stages. First, the low-level binaural features are extracted. Then, each DNN is trained independently from the others (Figure 1). After that, the probabilities that each T-F unit of the mixture belongs to different sources are estimated. Finally the target signal is reconstructed from the soft-mask and mixture signal.

IV. Proposed system

Figure 2 shows how the system of DNNs works. The inputs for each DNN are the stereo channel mixtures, then the short-time Fourier transform (STFT) is performed on the left and right channels in order to obtain the T-F representation of the input signals, $L(m, f)$ and $R(m, f)$, where $m = 1, \ldots, M$ and $f = 1, \ldots, F$ are the time frame and frequency bin indices respectively. Binaural features such as the Interaural Phase Difference (IPD) and the Interaural Level Difference (ILD) are then estimated at each time-frequency unit [7], put together as in [1] and arranged into $N = 128$ blocks, each block containing the information for $K = 8$ frequency bins. Each of the N blocks is the input of each DNN and the output is a softmax classifier, which gives a set of probabilities corresponding to how much a source is likely to come from a specific DOA. This information is used to generate a soft-mask by ungrouping the T-F bins and then applied to the speech mixtures to separate the single speech tracks.

V. Experiments

Training. Given an unlabeled audio track convolved with a room BRIR, the ILD and IPD features can be evaluated as in section III and are used in the input layer. The speech tracks are generated by convolving 8 sentences from the TIMIT database with the BRIRs of a set of echoic rooms, which consists of several audio samples recorded by a sensor placed around a half-circular grid, in variable positions ranging from -90° to $+90^\circ$, with steps of 10°, as shown in Figure 4.
A back-propagation algorithm is used in order to minimize the cost-function and to find the global optimized parameters for the whole deep network. The ground truth for the softmax classifier is obtained from the orientation information of the unlabeled data: if the individual source in the observed signals belongs to the DOA j, $p(y_j = j|\vec{x}_{(n,m)}) = 1$ otherwise $p(y_j \neq j|\vec{x}_{(n,m)}) = 0$.
Testing. Soft-masks can be generated with the set of training parameters $(\vec{W}, \vec{b})$ and used to estimate the audio sources from the mixtures, which consist of two tracks of 8 sentences from the TIMIT database, different from those used for training.

VI. Results and discussion

Figure 5 shows several SDRs plots in the case in which the same training and testing rooms, labeled 'A', are used. Different dropout percentages, 0% (no dropout), 40%, 50% and 60% are compared. All the three cases with non-zero dropout percentages perform better than the case without dropout, with improvements on the SDR up to ~ 1 dB for the 60% case, which is close to the optimal value of 50% suggested by the literature [6]. This can be explained by the fact that the training set used is relatively small, with ~ 10 minutes of total audio recording, so that the dropout algorithm improves the testing data adaptation to the training data. Higher dropout percentages have been tested and not shown in Figure 5. In fact, removing a very high number of connections leads to a worse DOA detection.

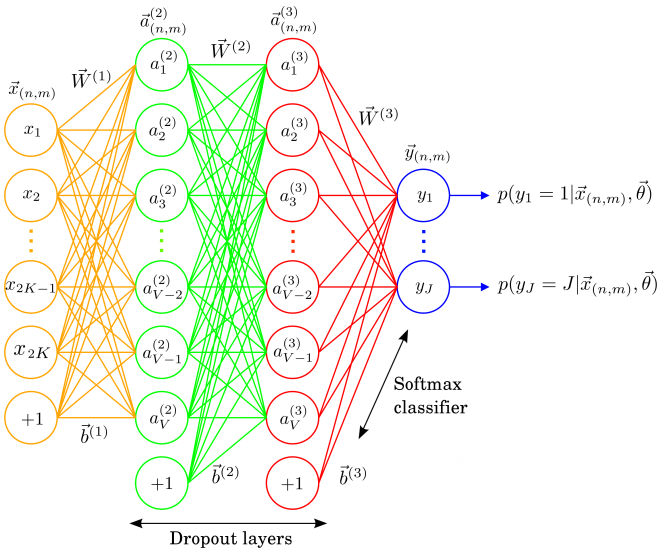

Fig. 1: Complete structure of one single DNN.

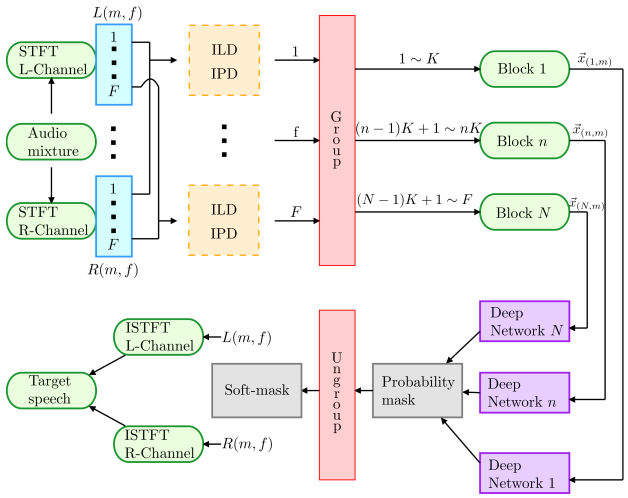

Fig. 2: The system architecture.

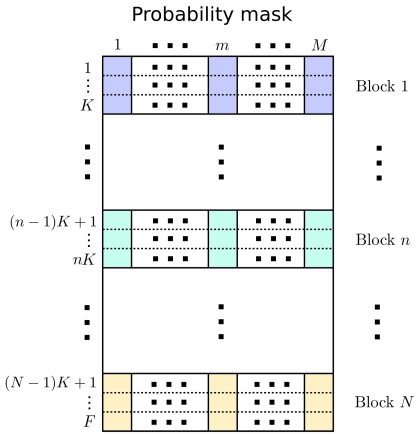

Fig. 3: Probability mask.

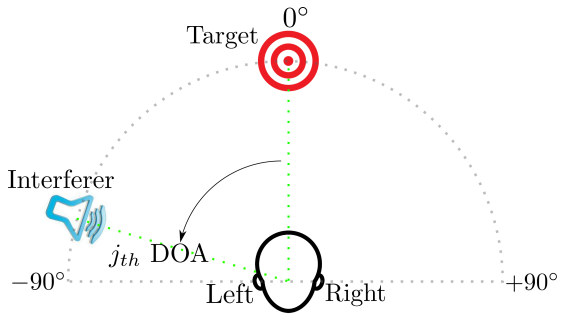

Fig. 4: The experimental setup.

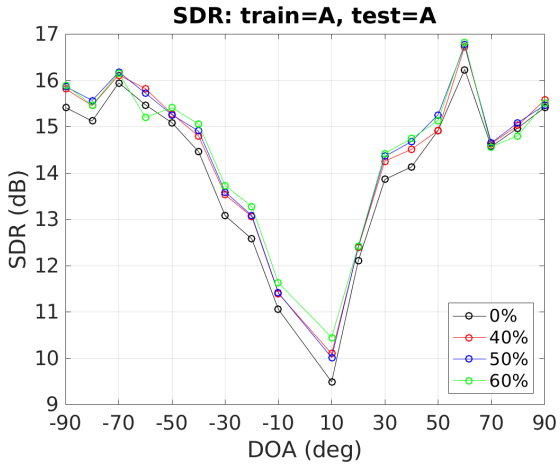

Fig. 5: SDR evaluation for a few dropout percentages: 0%, 40%, 50% and 60%.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement no 607290 SpaRTaN.

References

[1] Y. Yu, W. Wang, J. Luo, and P. Feng, "Localization based stereo speech separation using deep networks," in *Proc. 2015 IEEE International Conference on Digital Signal Processing (DSP)*, July 2015, pp. 153–157.

[2] Y. Yu and W. Wang, "Unsupervised feature learning for stereo source separation," in *Proc. 10th International Conference on Mathematics in Signal Processing (IMA 2014), Birmingham, UK*, Dec 2014, pp. 15–17.

[3] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," in *EURASIP Journal on Audio Speech and Music Processing*, Sept 2016, pp. 7–18.

[4] A. Zermini, Y. Yu, Y. Xu, W. Wang, and M. D. Plumbley, "Deep neural network based audio source separation," in *Proc. 11th International IMA International Conference on Mathematics in Signal Processing*, 2016.

[5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[6] P. Baldi and P. Sadowski, "The dropout learning algorithm," *Artif. Intell.*, vol. 210, pp. 78–122, May 2014.

[7] A. Alinaghi, W. Wang, and P. J. Jackson, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 684–688.