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 for source separation. In this method, the DNNs are used to predict the DOA of the audio sources with respect to the listener, which are used to estimate the high-level features and allow the prediction of the DOAs of the audio sources in the observed signals belongs to the DOA 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 $-90^\circ$ to $+90^\circ$, with steps of $10^\circ$, 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 $j$, $p(y_j = j|x_{(n,m)}) = 1$ otherwise $p(y_j \neq j|x_{(n,m)}) = 0$.

Testing. Soft-masks can be generated with the set of training parameters $(\hat{W}, \hat{b})$ and used to estimate the audio sources from the mixtures, which consist of two tracks of 8 sentences from the TIMIT database, with different from those used for training.

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 $-90^\circ$ to $+90^\circ$, with steps of $10^\circ$, 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 $j$, $p(y_j = j|x_{(n,m)}) = 1$ otherwise $p(y_j \neq j|x_{(n,m)}) = 0$.

Testing. Soft-masks can be generated with the set of training parameters $(\hat{W}, \hat{b})$ and used to estimate the audio sources from the mixtures, which consist of two tracks of 8 sentences from the TIMIT database, with 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 $\sim 1$ 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 $\sim 10$ 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