Deep neural network based audio source separation

Alfredo Zermini[1], Yang Yu[2], Yong Xu[1], Mark D. Plumbley[1], and Wenwu Wang[1]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey (UK)
[2]Northwestern Polytechnical University, Xi'an, China
{a.zermini,y.xu,m.plumbley,w.wang}@surrey.ac.uk
nwpuyuy@nwpu.edu.cn

Abstract. Audio source separation aims to extract individual sources from mixtures of multiple sound sources. Many techniques have been developed such as independent component analysis, computational auditory scene analysis, and non-negative matrix factorisation. A method based on Deep Neural Networks (DNNs) and time-frequency (T-F) 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 T-F masks for the recovery/estimation of the individual audio sources.

1 Introduction

Sound source separation techniques aim to extract a speech or a sound source from a mixture of speech or sound signals. While humans have an innate ability to distinguish a single speech or sound, for computers this task is challenging.

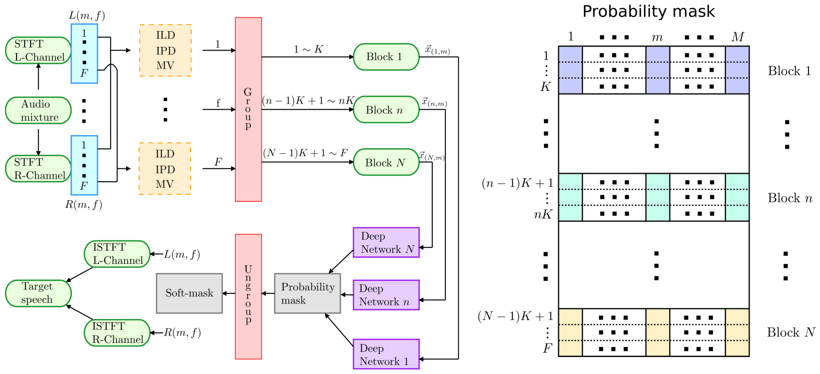
Extensive efforts are currently devoted to developing effective and efficient algorithms for audio source separation, for a variety of applications, such as music, home recording, forensic, cinema, cellphones, just to mention a few. Many methods have been developed such as independent component analysis, computational auditory scene analysis and non-negative matrix factorisation. Recently, the method based on Deep Neural Networks (DNNs) and time frequency masking has been developed for binaural audio source separation, offering state-of-the-art performance compared to the other previously mentioned methods.

The methodology proposed in this paper is based on the work in [1] where a Gaussian mixture model (GMM) framework is used to model the distribution of various cues at each T-F unit, Mixing Vector (MV), Interaural Level Difference (ILD) and Interaural Phase Difference (IPD) [1]. The Expectation-Maximization (EM) algorithm is employed to estimate the model parameters and the probability at each T-F point. In [2], [3] and [4], the GMM and EM are replaced with deep neural networks. The DNNs consist of two sparse autoencoders and a softmax classifier and are used to extract high-level features (i.e. spatial information of the sources) from the T-F representation of the mixtures and to predict the Direction Of Arrival (DOA) of the audio sources with respect to the listener by estimating source occupation probabilities at each T-F point, which are then used to generate soft time-frequency masks for the recovery/estimation of the individual audio sources. This paper performs a further study of these methods and shows some new experimental results.

2 System Overview

Each DNN consists of four stages:

1. extraction of the low-level features (i.e. MV, ILD and IPD) (details in Section 2.1).
2. Training of the deep networks (details in Section 2.2).
3. Estimation of the probabilities that each T-F unit belongs to different sources and generation of the soft-mask (details in Section 2.2).
4. Reconstruction of the target signal from the soft-mask and mixture signal.



(a) Diagram of the system architecture using deep neural networks.

(b) Probability mask.

Fig. 1. The system architecture and the output probability mask.

2.1 Proposed system

Figure 1 (a) 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. The MV, IPD and ILD are then estimated at each time-frequency unit[5]. By putting them together, as in [2]

$$\boldsymbol{x}(m,f) = [\boldsymbol{M}V(m,f), ILD(m,f), IPD(m,f)]^T \tag{1}$$

In the next step, the low-level features are arranged into N blocks, each block containing the information for all the frequency bins:

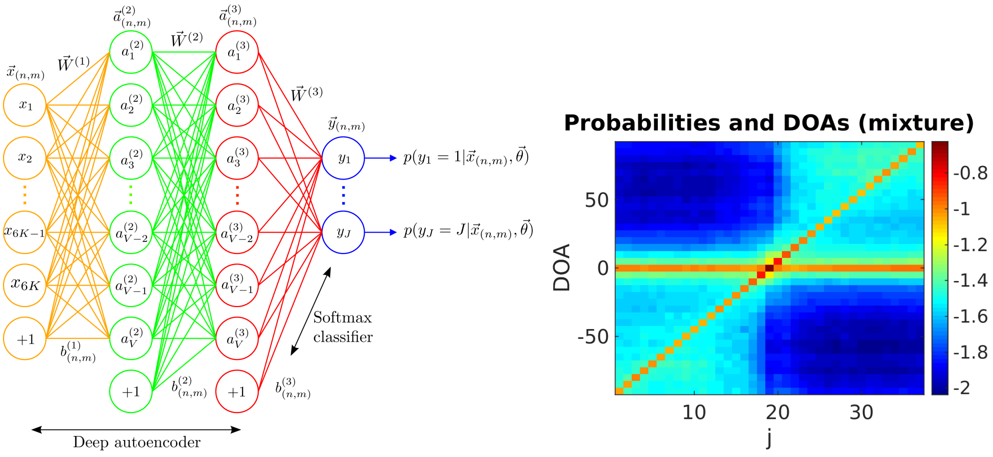
$$\boldsymbol{x}_{(n,m)} = \left[\boldsymbol{x}^T(m,(n-1)K+1), \ldots, \boldsymbol{x}^T(m,K)\right]^T \tag{2}$$

The dimension of each DNN input is $6K = 48$, where 32, 8 and 8 are the dimensions of the MV, ILD and IPD respectively. The block n includes K frequency bins $((n-1)K+1, \ldots, nK)$, where $K = F/N$ and N is the number of DNNs. The high-level features can be obtained with an unsupervised training of the sparse autoencoder and then used as inputs for the output layer of the DNNs, i.e. the softmax regression of the networks. The output of softmax regression for each block is a set of probabilities corresponding to how much a source is likely to come from a specific DOA and it is called the probability mask, as shown in Figure 1 (b). The soft-mask can be obtained by ungrouping the T-F bins and assigning to the K frequency bins inside the block n with the same probability. The soft-mask can be applied to the audio mixtures in order to recover the original sources, after applying the inverse STFT (ISTFT).

2.2 Implementation

This section discusses how the network used in Section 2 is constructed and trained. Figure 2 shows how each deep neural network is built.

– *Training.* Given an unlabeled audio track containing speech, convolved with a room BRIR, the MV, ILD and IPD features can be evaluated as in section 2.1, which are used as the first input layer. The speech tracks are generated by convolving an audio file 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°, with steps of



(a) Complete structure of one single DNN

(b) The DOA tracking for a speech mixture in room A.

Fig. 2. The DNN structure and the DOA tracking.

5°.

The next layers are trained by setting both the inputs and the outputs layers with the same parameters. Once the training of the two deep autoencoders is finished, the parameters of the output layer are used as the input layer of the softmax classifier, in the so called fine-tuning step, which utilises a back-propagation algorithm to minimise the cost-function and to find the global optimized parameters for the whole deep networks. 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, $p(y_j = j|\boldsymbol{x}_{(n,m)}) = 1$ otherwise $p(y_j \neq j|\boldsymbol{x}_{(n,m)}) = 0$.

- *Testing.* After training the whole DNN and the set of parameters $(\boldsymbol{W}, \boldsymbol{b})$, soft-masks can be generated with these training parameters and used to estimate the audio sources from the mixtures. The performance for the two sources scenario is given in section 3.2.

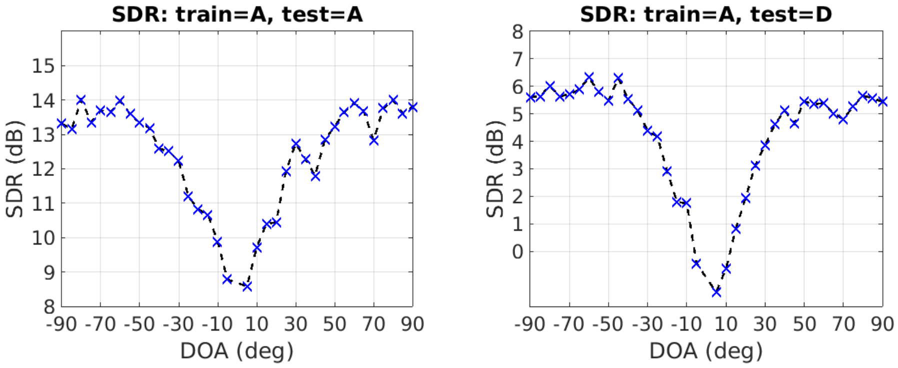
3 Experiments

3.1 Dataset generation

The DNN presented in section 2.2 has been used with audio tracks that are generated by convolving samples from the TIMIT database with room BRIRs measured in rooms with different acoustics properties. The BRIRs dataset has been recorded at the University of Surrey by using a dummy head and torso in four different types of rooms, namely A, B, C, D. The training data is built by randomly selecting 8 sentences from two speakers and then convolving them with the BRIRs from −90° to +90° with a step of 5°. The test data is a mixture of two sets of 8 sentences each.

3.2 Experimental results

Figure 2 (b) shows how the DOA of the two speech mixtures is obtained as a function of the test set j: as expected, one speech source is fixed at 0°, while the other changes from −90° to +90°. The DOA has been used in order to generate the soft-masks and separate the two speech tracks. Figures 3 (a) and (b) shows the SDR evaluation in the case in which the DNN system has been trained with a speech source convolved with the room A BRIR. The pre-trained DNNs are then applied to the same sound mixture convolved with the testing rooms A or D BRIR. Better



(a) Training A, testing A.

(b) Training A, testing D.

Fig. 3. SDR evaluations for different testing rooms.

results can be obtained in testing room A, where the reverberation time is shorter compared to room D, with SDR of the output ratios up to ∼ 13 dB on the wide angles (where the two sources are more spaced from each other, so they are easier to separate with respect to 0°), while in room D the SDR is up to ∼11 dB.

4 Conclusions

A localization-based stereo speech separation system using DNNs has been presented. The current version of the code has been written in MATLAB, though a Python version of the same code has been almost completely rewritten, which should be able to achieve comparable performances in a shorter training time. In the future work the DNN structure will be upgraded and new terms for the cost-function will be improved in order to achieve higher SDRs.

References

1. A. Alinaghi, P. J. B. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation." in *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 9, 2014, pp. 1434–1448.
2. Y. Yu, W. Wang, J. Luo, and P. Feng, "Localization based stereo speech separation using deep networks," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, July 2015, pp. 153–157.
3. 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.
4. 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.
5. A. Alinaghi, W. Wang, and P. J. Jackson, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 684–688.