

# A NEURAL NETWORK FOR SOUND SOURCE SEPARATION

JÖRN ANEMÜLLER AND TINO GRAMSS (†)

*Graduate School in Psychoacoustics and Department of Physics  
Carl von Ossietzky University, Oldenburg, Germany  
e-mail: ane@uni-oldenburg.de*

## 1 Introduction

The brain possesses the remarkable capability to filter incoming signals of multiple speakers in such a way that the subject's attention can be focused on a single sound source, the other sources being suppressed.

Much effort has been spent in order to mimic this behaviour by machines or clever algorithms which try to reconstruct the separated sources. A key issue in such attempts is the question 'Which is the quantity to optimise', i.e., which function tells us whether the original sources have been separated. An appealing answer to this question might be given by the concept of Independent Component Analysis. Here, the mutual statistical independence of the different source signals is exploited. I.e., one of the source signals does not give us information about any of the other sources.

However, when different mixtures of the source signals are detected, e.g., at the two ears of a human subject, the mixtures are strongly correlated, i.e., *not* independent of each other. Therefore, in source separation one attempts to filter recorded mixtures in such a way that mutually independent signals are obtained which resemble the original source signals.

Since learning rules for this task exist which employ simple, 'neuronal' operations only, it is conceivable that the brain exploits the principle of mutual independence of the sources in order to separate them.

## 2 Source Separation in the time-domain

In this section, we sketch a simple algorithm for separating an instantaneous mixture of independent signals. For a review refer to, e.g., [1].

Imagine  $N$  original source signals,  $\mathbf{s}(t) = (s_1(t), \dots, s_N(t))$ , which are emitted from, say, different speakers. These source signals are superimposed by the invertible mixing matrix  $\mathbf{A}$ , resulting in  $N$  recorded mixtures

$$\mathbf{m}(t) = (m_1(t), \dots, m_N(t)), \quad \mathbf{m}(t) = \mathbf{A}\mathbf{s}(t). \quad (1)$$

The goal is to find a weight-matrix  $\mathbf{W}$  such that the reconstructed signals

$$\mathbf{x}(t) = (x_1(t), \dots, x_N(t)), \quad \mathbf{x}(t) = \mathbf{W}\mathbf{m}(t)$$

are mutually statistically independent. It can be shown that  $\mathbf{x}(t)$  resembles  $\mathbf{s}(t)$  up to a constant permutation and rescaling of the sources.

A neural learning rule finds  $\mathbf{W}$  by first computing a non-linear function of the reconstructed signals:

$$u_i(t) = \tanh(x_i(t)), \quad i = 1, \dots, N. \quad (2)$$

Afterwards, the weight-matrix is changed by an increment

$$\Delta \mathbf{W}^{(t)} \propto \mathbf{W}^{(t)} - \mathbf{u}(t)\mathbf{x}^T(t)\mathbf{W}^{(t)}, \quad \mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \Delta \mathbf{W}^{(t)}. \quad (3)$$

This learning scheme is iterated at each time-step  $t$ .

### 3 Sound Source Separation in the frequency-domain

In the case of acoustic superposition of sound sources we are dealing with echoes and reflections. Hence, the multiplication in eq. (1) has to be replaced by the convolution of the source signals with the room's impulse responses. However, in the frequency domain, we rediscover multiple versions of the instantaneous source separation problem, one for each frequency band  $f = f_1, \dots, f_M$ :

$$\hat{\mathbf{m}}(f, T) = \hat{\mathbf{A}}(f)\hat{\mathbf{s}}(f, T).$$

By  $\hat{\mathbf{m}}(f, T)$  and  $\hat{\mathbf{s}}(f, T)$  we denote short-time spectra obtained from  $\mathbf{m}(t)$  and  $\mathbf{s}(t)$ , respectively, at consecutive times  $T = T_1, T_2, \dots$ . Hence, the goal is to solve  $M$  source separation problems of the form of eq. (1), resulting in weights-matrices  $\hat{\mathbf{W}}(f_1), \dots, \hat{\mathbf{W}}(f_M)$ .

Since the spectral components  $\hat{\mathbf{m}}(f, T)$  and  $\hat{\mathbf{s}}(f, T)$  are complex valued, the learning algorithm (eqs. 2, 3) has to be modified. From a maximum likelihood approach [2] we derive the learning rule

$$\hat{u}_i(f, T) = \frac{\hat{x}_i(f, T)}{\|\hat{x}_i(f, T)\|} \tanh(\|\hat{x}_i(f, T)\|), \quad i = 1, \dots, N,$$

$$\Delta \hat{\mathbf{W}}(f) \propto \hat{\mathbf{W}}(f) - \hat{\mathbf{u}}(f, T)\hat{\mathbf{x}}^H(f, T)\hat{\mathbf{W}}(f), \quad f = f_1, \dots, f_M.$$

The operator  $\mathbf{x}^H$  denotes the complex conjugate transpose of vector  $\mathbf{x}$ , and  $\|x\|$  denotes the norm of the complex number  $x$ .

Without additional precautions, reconstruction of the source signals in the time-domain is hindered by the inherent permutation invariance of source separation algorithms, i.e., we can't tell which network output resembles which source. E.g., at frequency  $f_1$  the reconstructed signal  $\hat{x}_1(f_1, T)$  might belong to source  $s_1$  while

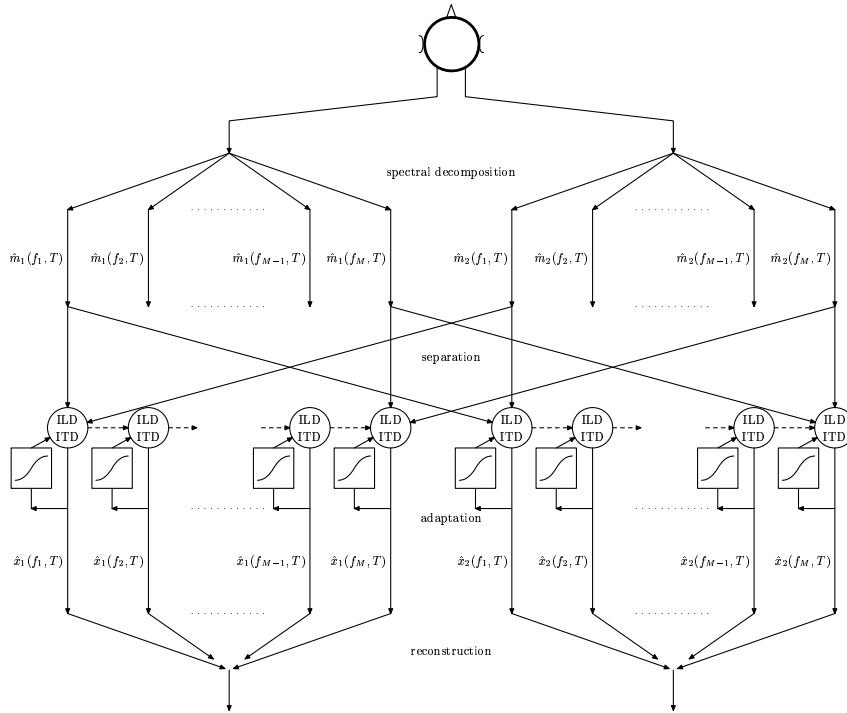


Figure 1: The processing stages of our source separation algorithm.

at frequency  $f_2$  the component  $\hat{x}_1(f_2, T)$  belongs to the opposite source,  $s_2$ . We solve this problem for the case where the mixing process introduces time- and level-differences only, which is fulfilled in an ideal anechoic chamber. In this situation, the coefficients  $\hat{A}_{ij}(f)$  of the mixing matrices at different frequencies can be derived from the propagation-times  $\tau_{ij}$  and -levels  $A_{ij}$  from source  $j$  to microphone  $i$  via the formula

$$\hat{A}_{ij}(f) = A_{ij} e^{-2\pi i f \tau_{ij}}.$$

The algorithm starts adaptation at low frequencies and propagates time- and level-differences towards higher frequencies, see also fig. 1. Hence, a rough estimate of the sources' parameters is obtained at low frequencies while the information obtained in high frequency components improves on this estimate. For more details refer to [3].

Finally, we resolve the scaling invariance by fixing the diagonal elements of the estimated *mixing* matrices  $\hat{\mathbf{W}}^{-1}(f_1), \dots, \hat{\mathbf{W}}^{-1}(f_M)$  to unity.

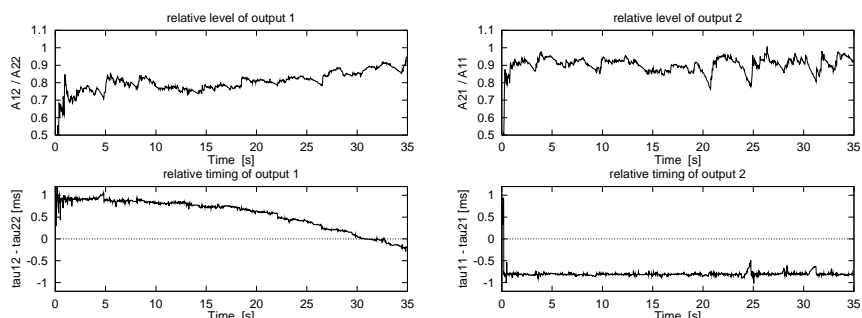


Figure 2: Result of a separation experiment in an anechoic chamber. The figure shows the time-course of the relative timing and level for both reconstructed signals. It is clearly visible that initial convergence of the algorithm takes only about one second of the recorded signals. The algorithm's estimate for the relative timing ('ITD') is accurate and displays the movement of one speaker. The estimates for the relative level ('ILD') show larger fluctuations.

#### 4 Moving sources in an anechoic chamber

Signals from a moving and a standing speaker were recorded in stereo in the anechoic chamber of the University of Oldenburg. The proposed algorithm was used to separate the mixed signals.

The experimental setup was as follows: Two microphones were placed 35 cm apart. The stationary speaker was standing in a distance of 3 m at 60 degrees left of the mid-perpendicular of the microphones. The moving speaker started at a distance of 4 m at 70 degrees to the right. He walked in a straight line parallel to the microphones until he reached a position at about 30 degrees left. The algorithm started off from the (wrong) assumption that no mixing of the signals occurs.

Figure 2 illustrates the outcome of the experiment.

#### References

1. Jean-Pierre Nadal and Nestor Parga. Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches. *Neural Computation*, 9:1421–1456, 1997.
2. D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis, draft 3.7. URL: <ftp://wol.ra.phy.cam.ac.uk/pub/-mackay/ica.ps.gz>, Dec. 1996.
3. Jörn Anemüller and Tino Gramss. On-line blind separation of moving sound sources. In *ICA '99*, Aussois, France, Jan. 1999.