Journal of Machine Learning Research 7 (2006) 2585-2602 Submitted 3/06; Revised 8/06; Published 12/06

Linear State-Space Models for Blind Source Separation

Rasmus Kongsgaard Olsson RKO@IMM.DTU.DK
Lars Kai Hansen LKH@IMM.DTU.DK
Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Editor: Aapo Hyvärinen

Abstract

We apply a type of generative modelling to the problem of blind source separation in which prior knowledge about the latent source signals, such as time-varying auto-correlation and quasi-periodicity, are incorporated into a linear state-space model. In simulations, we show that in terms of signal-to-error ratio, the sources are inferred more accurately as a result of the inclusion of strong prior knowledge. We explore different schemes of maximum-likelihood optimization for the purpose of learning the model parameters. The Expectation Maximization algorithm, which is often considered the standard optimization method in this context, results in slow convergence when the noise variance is small. In such scenarios, quasi-Newton optimization yields substantial improvements in a range of signal to noise ratios. We analyze the performance of the methods on convolutive mixtures of speech signals.

Keywords: blind source separation, state-space model, independent component analysis, convolutive model, EM, speech modelling

1. Introduction

We are interested in blind source separation (BSS) in which unknown source signals are estimated from noisy mixtures. Real world application of BSS techniques are found in as diverse fields as audio (Yellin and Weinstein, 1996; Parra and Spence, 2000; Anemüller and Kollmeier, 2000), brain imaging and analysis (McKeown et al., 2003), and astrophysics (Cardoso et al., 2002). While most prior work is focused on mixtures that can be characterized as instantaneous, we will here investigate causal convolutive mixtures. The mathematical definitions of these classes of mixtures are given later in this introductory section. Convolutive BSS is relevant in many signal processing applications, where the instantaneous mixture model cannot possibly capture the latent causes of the observations due to different time delays between the sources and sensors. The main problem is the lack of general models and estimation schemes; most current work is highly application specific with the majority focused on applications in separation of speech signals. In this work we will also be concerned with speech signals, however, we will formulate a generative model that may be generalizable to several other application domains.

One of the most successful approaches to convolutive BSS is based on the following assumptions: 1) The mixing process is linear and causal, 2) the source signals are statistically independent, 3) the sources can be fully characterized by their *time variant* second order statistics (Weinstein et al., 1993; Parra and Spence, 2000). The last assumption is defining for this approach. Keeping to second order statistics we simplify computations but have to pay the price of working with time-

variant statistics. It is well-known that stationary second order statistics, that is, covariances and correlation functions, are not informative enough in the convolutive mixing case.

Our research concerns statistical analysis and generalizations of this approach. We formulate a generative model based on the same statistics as the Parra-Spence model. The benefit of this generative approach is that it allows for estimation of additional noise parameters and injection of well-defined a priori information in a Bayesian sense (Olsson and Hansen, 2005). Furthermore, we propose several algorithms to learn the parameters of the proposed models.

The linear mixing model reads

$$\mathbf{x}_t = \sum_{k=0}^{L-1} \mathbf{A}_k \mathbf{s}_{t-k} + \mathbf{w}_t. \tag{1}$$

At discrete time t, the observation vector, $\mathbf{x}_t$, results from the convolution sum of the L time-lagged mixing matrices $\mathbf{A}_k$ and the source vector $\mathbf{s}_t$. The individual sources, that is, the elements of $\mathbf{s}_t$, are assumed to be statistically independent. The observations are corrupted by additive i.i.d. Gaussian noise, $\mathbf{w}_t$. BSS is concerned with estimating $\mathbf{s}_t$ from $\mathbf{x}_t$, while $\mathbf{A}_k$ is unknown. It is apparent from (1) that only filtered versions of the elements of $\mathbf{s}_t$ can be retrieved, since the inverse filtering can be applied to the unknown $\mathbf{A}_k$. As a special case of the filtering ambiguity, the *scale* and the ordering of the sources is unidentifiable. The latter is evident from the fact that various permutation applied simultaneously to the elements of $\mathbf{s}_t$ and the columns of $\mathbf{A}_t$ produce identical mixtures, $\mathbf{x}_t$.

Equation (1) collapses to an *instantaneous* mixture in the case of $L = 1$ for which a variety of Independent Component Analysis (ICA) methods are available (e.g., Comon, 1994; Bell and Sejnowski, 1995; Hyvarinen et al., 2001). As already mentioned, however, we will treat the class of convolutive mixtures, that is $L > 1$.

Convolutive Independent Component Analysis (C-ICA) is a class of BSS methods for (1) where the source estimates are produced by computing the 'unmixing' transformation that restores statistical independence. Often, an inverse linear filter (e.g., FIR) is applied to the observed mixtures. Simplistically, the separation filter is estimated by minimizing the mutual information, or 'cross' moments, of the 'separated' signals. In many cases non-Gaussian models/higher-order statistics are required, which require a relatively long data series for reliable estimation. This can be executed in the time domain (Lee et al., 1997; Dyrholm and Hansen, 2004), or in the frequency domain (e.g., Parra and Spence, 2000). The transformation to the Fourier domain reduces the matrix convolution of (1) to a matrix product. In effect, the more difficult convolutive problem is decomposed into a number of manageable instantaneous ICA problems that can be solved independently using the mentioned methods. However, frequency domain decomposition suffers from *permutation over frequency* which is a consequence of the potential different orderings of sources at different frequencies. Many authors have explored solutions to the permutation-over-frequency problem that are based on measures of spectral structure (e.g., Anemüller and Kollmeier, 2000), where amplitude correlation across frequency bands is assumed and incorporated in the algorithm.

The work presented here forges research lines that treat instantaneous ICA as a density estimation problem (Pearlmutter and Parra, 1997; Højen-Sørensen et al., 2002), with richer source priors that incorporate time-correlation, non-stationarity, periodicity and the convolutive mixture model to arrive at an C-ICA algorithm. The presented algorithm, which operates entirely in the time-domain, relies on a linear state-space model, for which estimation and exact source inference are available. The states directly represent the sources, and the transition structure can be interpreted as describing the internal time-correlation of the sources. To further increase the audio realism of the model,

Olsson and Hansen (2005) added a harmonic excitation component in the source speech model (Brandstein, 1998); this idea is further elaborated and tested here.

Algorithms for the optimization of the likelihood of the linear state-space model are devised and compared, among them the basic EM algorithm, which is used extensively in latent variable models (Moulines et al., 1997). In line with Bermond and Cardoso (1999), the EM-algorithm is shown to exhibit slow convergence in good signal to noise ratios.

It is interesting that the two 'unconventional' aspects of our generative model: the non-stationarity of the source signals and their harmonic excitation, do not change the basic quality of the state-space model, namely that exact inference of the sources and exact calculation of the log-likelihood and its gradient are still possible.

The paper is organized as follows: First we introduce the state-space representation of the convolutive mixing problem and the source models in Section 2, in Section 3 we briefly recapitulate the steps towards exact inference for the source signals, while Section 4 is devoted to a discussion of parameter learning. Sections 5 and 6 present a number of experimental illustrations of the approach on simulated and speech data respectively.

2. Model

The convolutive blind source separation problem is cast as a density estimation task in a latent variable model as was suggested in Pearlmutter and Parra (1997) for the instantaneous ICA problem

$$\mathrm{p}(\mathbf{X}|\theta) = \int \mathrm{p}(\mathbf{X}|\mathbf{S},\theta_1)\mathrm{p}(\mathbf{S}|\theta_2)d\mathbf{S}.$$

Here, the matrices $\mathbf{X}$ and $\mathbf{S}$ are constructed as the column sets of $\mathbf{x}_t$ and $\mathbf{s}_t$ for all t. The functional forms of the conditional likelihood, $\mathrm{p}(\mathbf{X}|\mathbf{S},\theta_1)$, and the joint source prior, $\mathrm{p}(\mathbf{S}|\theta_2)$, should ideally be selected to fit the realities of the separation task at hand. The distributions depend on a set of tunable parameters, $\theta \equiv \{\theta_1,\theta_2\}$, which in a blind separation setup is to be learned from the data. In the present work, $\mathrm{p}(\mathbf{X}|\mathbf{S},\theta_1)$ and $\mathrm{p}(\mathbf{S}|\theta_2)$ have been restricted to fit into a class of linear state-space models, for which effective estimation schemes exist (Roweis and Ghahramani, 1999)

$$\mathbf{s}_t = \mathbf{F}^n\mathbf{s}_{t-1} + \mathbf{C}^n\mathbf{u}_t + \mathbf{v}_t, \quad (2)$$

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{w}_t. \quad (3)$$

Equations (2) and (3) describe the *state/source* and *observation* spaces, respectively. The parameters of the former are time-varying, indexed by the block index n, while the latter noisy mixing process is stationary. The randomness of the model is enabled by i.i.d. zero mean Gaussian variables, $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0},\mathbf{Q}^n)$, and $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0},\mathbf{R})$ The 'input' or 'control' signal $\mathbf{u}_t \equiv \mathbf{u}_t(\psi^n)$ deterministically shifts the mean of $\mathbf{s}_t$ depending on parameters ψ^n. Various structures can be imposed on the model parameters, $\theta_1 = \{\mathbf{A},\mathbf{R}\}$ and $\theta_2 = \{\mathbf{F}^n,\mathbf{C}^n,\mathbf{Q}^n,\psi^n\}$, in order to create the desired effects. For equations (2) and (3) to pose as a generative model for the instantaneous mixture of first-order autoregressive, AR(1), sources it need only be assumed that $\mathbf{F}^n$ and $\mathbf{Q}^n$ are diagonal matrices and that $\mathbf{C}^n = \mathbf{0}$. In this case, $\mathbf{A}$ functions as the mixing matrix. In Section 2.1, we generalize to AR(p) and convolutive mixing.

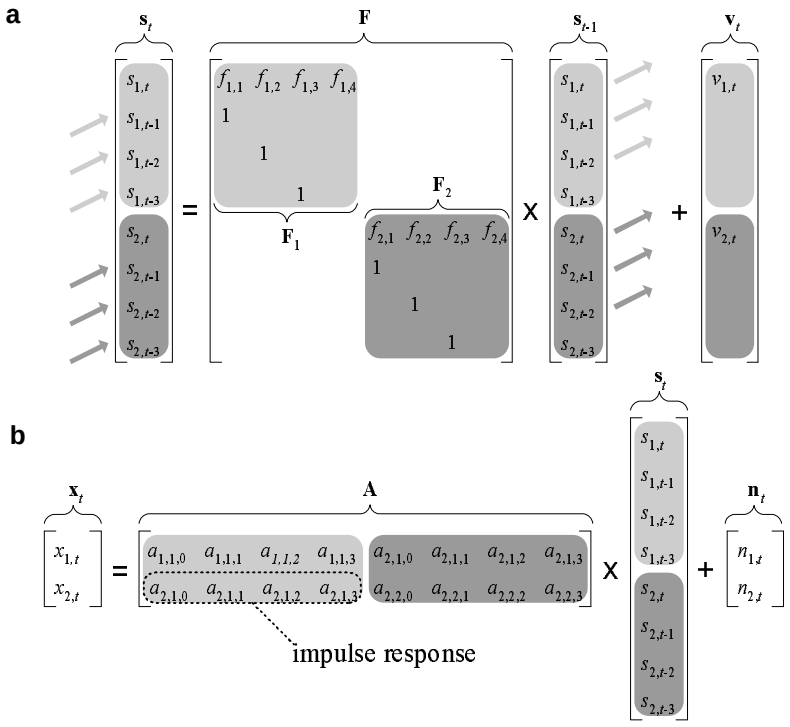

Figure 1: The dynamics of the linear state space model when it has been constrained to describe a noisy convolutive mixture of $P = 2$ autoregressive (AR) sources. This is achieved by augmenting the source vector to contain time-lagged signals. In **a** is shown the corresponding source update, when the order of the AR process is $p = 4$. In **b**, the sources are mixed through filters ($L = 4$) into $Q = 2$ noisy mixtures. Blanks signify zeros.

2.1 Auto-Regressive Source Prior

The AR(p) source prior for source i in frame n is defined as follows,

$$s_{i,t} = \sum_{k=1}^{p} f_{i,k}^{n} s_{i,t-k} + v_{i,t}$$

where $t \in \{1,2,..,T\}$, $n \in \{1,2,..,N\}$ and $i \in \{1,2,..,P\}$. The excitation noise is i.i.d. zero mean Gaussian: $v_{i,t} \sim \mathcal{N}(0, q_i^n)$. It is an important point that the convolutive mixture of AR(p) sources can be contained in the linear state-space model (2) and (3), this is illustrated in Figure 1. The enabling trick, which is standard in time series analysis, is to augment the source vector to include a time history so that it contains L time-lagged samples of all P sources

$$\mathbf{s}_t = \left[\begin{array}{cccc} (\mathbf{s}_{1,t})^\top & (\mathbf{s}_{2,t})^\top & \ldots & (\mathbf{s}_{P,t})^\top \end{array} \right]^\top$$

where the i'th source is represented as

$$\mathbf{s}_{i,t} = \left[\begin{array}{cccc} s_{i,t} & s_{i,t-1} & \ldots & s_{i,t-L+1} \end{array} \right]^\top .$$

Furthermore, constraints are enforced on the matrices of θ

$$\mathbf{F}^n = \begin{bmatrix} \mathbf{F}_1^n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_2^n & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{F}_P^n \end{bmatrix}, \qquad \mathbf{Q}^n = \begin{bmatrix} \mathbf{Q}_1^n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^n & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_P^n \end{bmatrix},$$

$$\mathbf{F}_i^n = \begin{bmatrix} f_{i,1}^n & f_{i,2}^n & \cdots & f_{i,p-1}^n & f_{i,p}^n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \qquad (\mathbf{Q}_i^n)_{jj'} = \begin{cases} (q_i^2)^n & j = j' = 1 \\ 0 & j \neq 1 \bigvee j' \neq 1 \end{cases},$$

$$\mathbf{C}^n = \mathbf{0},$$

where $\mathbf{F}_i^n$ was defined for $p = L$. In the interest of the simplicity of the presentation, it is assumed that $\mathbf{F}_i^n$ has L row and columns. We furthermore assume that $p \leq L$; in the case of $p < L$, zeros replace the affected (rightmost) coefficients. Hence, the dimensionality of $\mathbf{A}$ is $Q \times (p \times P)$,

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{11}^\top & \mathbf{a}_{12}^\top & .. & \mathbf{a}_{1P}^\top \\ \mathbf{a}_{21}^\top & \mathbf{a}_{22}^\top & .. & \mathbf{a}_{2P}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{Q1}^\top & \mathbf{a}_{Q2}^\top & .. & \mathbf{a}_{QP}^\top \end{bmatrix}$$

where $\mathbf{a}_{ij} = [a_{ij,1}, a_{ij,2}, .., a_{ij,L}]^\top$ can be interpreted as the impulse response of the channel filter between source i and sensor j. Overall, the model can described can be described as the generative, time-domain equivalent of Parra and Spence (2000).

2.2 Harmonic Source Prior

Many classes of audio signals, such as voiced speech and musical instruments, are approximately piece-wise periodic. By the Fourier theorem, such sequences can be represented well by a harmonic series. In order to account for colored noise residuals and noisy signals in general, a harmonic and noise (HN) model is suggested (McAulay and Quateri, 1986). The below formulation is used

$$s_{i,t} = \sum_{t'=1}^{p} f_{i,t'}^n s_{i,t-t'} + \sum_{k=1}^{K} \left[c_{i,2k-1}^n \sin(\omega_{0,i}^n t) + c_{i,2k}^n \cos(\omega_{0,i}^n t) \right] + v_{i,t}$$

where $\omega_{0,i}^n$' is the fundamental frequency of source i in frame n and the Fourier coefficients are contained in $c_{i,2k-1}^n$ and $c_{i,2k}^n$. The harmonic model is represented in the state space model (2) & (3)

through the definitions

$$\begin{aligned}
\mathbf{C}^n &= \begin{bmatrix} (\mathbf{c}_1^n)^\top & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (\mathbf{c}_2^n)^\top & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & (\mathbf{c}_P^n)^\top \end{bmatrix}, \\
\mathbf{c}_i^n &= \begin{bmatrix} c_{i,1}^n & c_{i,2}^n & \cdots & c_{i,2K}^n \end{bmatrix}^\top, \\
\mathbf{u}_t^n &= \begin{bmatrix} (\mathbf{u}_{1,t}^n)^\top & (\mathbf{u}_{2,t}^n)^\top & \ldots & (\mathbf{u}_{P,t}^n)^\top \end{bmatrix}^\top,
\end{aligned}$$

where the k'th harmonics of source i in frame n are defined as $(\mathbf{u}_{i,t}^n)_{2k-1} = \sin(k\omega_{0,i}^n t)$ and $(\mathbf{u}_{i,t}^n)_{2k} = \cos(k\omega_{0,i}^n t)$, implying the following parameter set for the source mean: $\psi^n = \begin{bmatrix} \omega_{0,1}^n & \omega_{0,2}^n & \ldots & \omega_{0,P}^n \end{bmatrix}$. Other parametric mean functions could, of course, be used, for example, a more advanced speech model.

3. Source Inference

In a maximum a posteriori sense, the sources, $\mathbf{s}_t$, can be optimally reconstructed using the Kalman filter/smoother (Kalman and Bucy, 1960; Rauch et al., 1965). This is based on the assumption that the parameters θ are known, either a priori or have been estimated as described in Section 4. While the filter computes the time-marginal moments of the source posterior conditioned on past and present samples, that is, $\langle \mathbf{s}_t \rangle_{p(\mathbf{S}|\mathbf{x}_{1:t},\theta)}$ and $\langle \mathbf{s}_t \mathbf{s}_t^\top \rangle_{p(\mathbf{S}|\mathbf{x}_{1:t},\theta)}$, the smoother conditions on samples from the entire block: $\langle \mathbf{s}_t \rangle_{p(\mathbf{S}|\mathbf{x}_{1:T},\theta)}$ and $\langle \mathbf{s}_t \mathbf{s}_t^\top \rangle_{p(\mathbf{S}|\mathbf{x}_{1:T},\theta)}$. For the Kalman filter/smoother to compute MAP estimates, it is a precondition due that the model is linear and Gaussian. The computational complexity is $O(TL^3)$ due to a matrix inversion occurring in the recursive update. Note that the forward recursion also yields the exact log-likelihood of the parameters given the observations, $\mathcal{L}(\theta)$. A thorough review of linear state-space modelling, estimation and inference from a machine learning point of view can be found in Roweis and Ghahramani (1999).

4. Learning

The task of learning the parameters of the state-space model from data is approached by maximum-likelihood estimation, that is, the log-likelihood function, $\mathcal{L}(\theta)$, is optimized with respect to the parameters, θ. The log-likelihood is defined as a marginalization over the hidden sources

$$\mathcal{L}(\theta) = \log \mathrm{p}(\mathbf{X}|\theta) = \log \int \mathrm{p}(\mathbf{X}, \mathbf{S}|\theta) d\mathbf{S}.$$

A closed-form solution, $\theta = \arg\max_{\theta'} \mathcal{L}(\theta')$, is not available, hence iterative algorithms that optimize $\mathcal{L}(\theta)$ are employed. In the following sections three such algorithms are presented.

4.1 Expectation Maximization Algorithm

Expectation Maximization (EM) (Dempster et al., 1977), has been applied to latent variable models in, for example, Shumway and Stoffer (1982) and Roweis and Ghahramani (1999). In essence, EM

is iterative optimization of a lower bound decomposition of the log-likelihood

$$\mathcal{L}(\theta) \quad \geq \quad \mathcal{F}(\theta, \hat{p}) = \mathcal{J}(\theta, \hat{p}) - \mathcal{R}(\hat{p}) \tag{4}$$

where $\hat{p}(\mathbf{S})$ is any normalized distribution and the following definitions apply

$$\begin{aligned} \mathcal{J}(\theta, \hat{p}) &= \int \hat{p}(\mathbf{S}) \log p(\mathbf{X}, \mathbf{S}|\theta) d\mathbf{S}, \\ \mathcal{R}(\hat{p}) &= \int \hat{p}(\mathbf{S}) \log \hat{p}(\mathbf{S}) d\mathbf{S}. \end{aligned}$$

Jensen's inequality leads directly to (4). The algorithm alternates between performing Expectation (E) and Maximization (M) steps, guaranteeing that $\mathcal{L}(\theta)$ does not decrease following an update. On the E-step, the Kalman smoother is used to compute the marginal moments from the source posterior, $\hat{p} = p(\mathbf{S}|\mathbf{X}, \theta)$, see Section 3. The M-step amounts to optimization of $\mathcal{J}(\theta, \hat{p})$ with respect to θ (since this is the only $\mathcal{F}(\theta, \hat{p})$ term which depends on θ. Due to the choice of a linear Gaussian model, closed-form estimators are available for the M-step (see appendix A for derivations).

In order to improve on the convergence speed of the basic EM algorithm, the search vector devised by the M-step update is premultiplied by an adaptive step-size η. A simple exponentially increase of η from 1 was used until a decrease in $\mathcal{L}(\theta)$ was observed at which point η was reset to 1. This speed-up scheme was applied successfully in Salakhutdinov and Roweis (2003). Below follow the M-step estimators for the AR and HN models. All expectations $\langle \cdot \rangle$ are over the source posterior, $p(\mathbf{S}|\mathbf{X}, \theta)$:

4.1.1 AUTOREGRESSIVE MODEL

For source i in block n:

$$\begin{aligned} \mathbf{f}_{i,\mathtt{new}}^n &= \Big[\sum_{t=2+t_0(n)}^{T+t_0(n)} \langle \mathbf{s}_{i,t-1} \mathbf{s}_{i,t-1}^\top \rangle \Big]^{-\top} \Big[\sum_{t=2+t_0(n)}^{T+t_0(n)} \langle s_{i,t} \mathbf{s}_{i,t-1} \rangle \Big], \\ q_{i,\mathtt{new}}^n &= \frac{1}{T-1} \sum_{t=2+t_0(n)}^{T+t_0(n)} \Big[\langle s_{i,t}^2 \rangle - \big(\mathbf{f}_{i,\mathtt{new}}^n \big)^\top \langle s_{i,t} \mathbf{s}_{i,t-1} \rangle \Big], \end{aligned}$$

where $t_0(n) = (n-1)T$. Furthermore:

$$\begin{aligned} \mathbf{A}_{\mathtt{new}} &= \Big[\sum_{t=1}^{NT} \mathbf{x}_t \langle \mathbf{s}_t \rangle^\top \Big] \Big[\sum_{t=1}^{NT} \langle \mathbf{s}_t (\mathbf{s}_t)^\top \rangle \Big]^{-1}, \\ \mathbf{R}_{\mathtt{new}} &= \frac{1}{NT} \sum_{t=1}^{NT} \mathrm{diag}[\mathbf{x}_t (\mathbf{x}_t)^\top - \mathbf{A}_{\mathtt{new}} \langle \mathbf{s}_t \rangle (\mathbf{x}_t)^\top], \end{aligned}$$

where the diag$[\cdot]$ operator extracts the diagonal elements of the matrix. Following an M-step, the solution corresponding to $||\mathbf{A}_i|| = 1 \ \forall i$ is chosen, where $||\cdot||$ is the Frobenius norm and $\mathbf{A}_i = \begin{bmatrix} \mathbf{a}_{i1} & \mathbf{a}_{i2} & \cdots & \mathbf{a}_{iQ} \end{bmatrix}^\top$, meaning that $\mathbf{A}$ and $\mathbf{Q}^n$ are scaled accordingly.

4.1.2 HARMONIC AND NOISE MODEL

The linear source parameters and signals are grouped as

$$\mathbf{d}_i^n \equiv \begin{bmatrix} (\mathbf{f}_i^n)^\top & (\mathbf{c}_i^n)^\top \end{bmatrix}^\top \quad , \quad \mathbf{z}_i \equiv \begin{bmatrix} (\mathbf{s}_{i,t-1})^\top & (\mathbf{u}_{i,t})^\top \end{bmatrix}^\top,$$

where

$$\mathbf{f}_i^n \equiv \left[\begin{array}{cccc} f_{i,1}^n & f_{i,2}^n & \cdots & f_{i,p}^n \end{array} \right]^\top \quad , \quad \mathbf{c}_i^n \equiv \left[\begin{array}{cccc} c_{i,1} & c_{i,2}^n & \cdots & c_{i,p}^n \end{array} \right]^\top .$$

It is in general not trivial to maximize $\mathcal{J}(\theta, \hat{p})$ with respect to $\omega_{i,0}^n$, since several local maxima exist, for example, at multiples of $\omega_{i,0}^n$ (McAulay and Quateri, 1986). However, simple grid search in a region provided satisfactory results. For each point in the grid we optimize $\mathcal{J}(\theta, \hat{p})$ with respect to $\mathbf{d}_i^n$:

$$\mathbf{d}_{i,\mathbf{new}}^n = \left[\sum_{t=2}^{NT} \left\langle \mathbf{z}_{i,t}(\mathbf{z}_{i,t})^\top \right\rangle \right]^{-1} \sum_{t=2}^{NT} \left\langle \mathbf{z}_{i,t}(s_{i,t})^\top \right\rangle .$$

The estimators of $\mathbf{A}$, $\mathbf{R}$ and q_i^n are similar to those in the AR model.

4.2 Gradient-based Learning

The derivative of the log-likelihood, $\frac{d\mathcal{L}(\theta)}{d\theta}$, can be computed and used in a quasi-Newton (QN) optimizer as is demonstrated in Olsson et al. (2006). The computation reuse the analysis of the M-step. This can be realized by rewriting $\mathcal{L}(\theta)$ as in Salakhutdinov et al. (2003):

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \int p(\mathbf{S}|\mathbf{X}, \theta) \frac{d \log p(\mathbf{X}, \mathbf{S}|\theta)}{d\theta} d\mathbf{S} = \frac{d\mathcal{J}(\theta, \hat{p})}{d\theta}. \tag{5}$$

Due to the definition of $\mathcal{J}(\theta, \hat{p})$, the desired gradient in (5) can be computed following an E-step at relatively little effort. Furthermore, the analytic expressions are available from the derivation of the EM algorithm, see appendix A for details. A minor reformulation of the problem is necessary in order to maintain non-negative variances. Hence, the reformulations $\Omega^2 = \mathbf{R}$ and $(\phi_i^n)^2 = q_i^n$ are introduced. Updates are devised for Ω and ϕ_i^n. The derivatives are

$$\begin{aligned}
\frac{d\mathcal{L}(\theta)}{d\mathbf{A}} &= -\mathbf{R}^{-1}\mathbf{A} \sum_{t=1}^{NT} \left\langle \mathbf{s}_t \mathbf{s}_t^\top \right\rangle + \mathbf{R}^{-1} \sum_{t=1}^{N} \mathbf{x}_t \left\langle \mathbf{s}_t^\top \right\rangle , \\
\frac{d\mathcal{L}(\theta)}{d\Omega} &= \Omega^{-3} \sum_{t=1}^{NT} \left[\mathbf{x}_t \mathbf{x}_t^\top + \mathbf{A} \left\langle \mathbf{s}_t \mathbf{s}_t^\top \right\rangle \mathbf{A}^\top - 2\mathbf{x}_t \left\langle \mathbf{s}_t^\top \right\rangle \mathbf{A}^\top \right] , \\
\frac{d\mathcal{L}(\theta)}{d\mathbf{f}_i^n} &= \sum_{t=2+t_0(n)}^{T+t_0(n)} \left[\langle s_{i,t} \mathbf{s}_{i,t-1} \rangle - \left\langle \mathbf{s}_{i,t-1} \mathbf{s}_{i,t-1}^\top \right\rangle \mathbf{f}_i^n / q_i^n \right] , \\
\frac{d\mathcal{L}(\theta)}{d\phi_i^n} &= (1-T)/\phi_i^n + \\
&\quad \phi_i^{-3} \sum_{t=2+t_0(n)}^{T-1+t_0(n)} \left[\left\langle s_{i,t} s_{i,t}^\top \right\rangle + (\mathbf{f}_i^n)^\top \left\langle \mathbf{s}_{i,t-1} \mathbf{s}_{i,t-1}^\top \right\rangle \mathbf{f}_i^n - 2(\mathbf{f}_i^n)^\top \left\langle s_{i,t} \mathbf{s}_{i,t-1}^\top \right\rangle \right] .
\end{aligned}$$

In order to enforce the unit L2 norm on $\mathbf{A}_i$, a Lagrange multiplier is added to the derivative of $\mathbf{A}$. In this work, the QN optimizer of choice is the BFGS optimizer of Nielsen (2000).

4.3 Stochastic Gradient Learning

Although quasi-Newton algorithms often converge rapidly with a high accuracy, they do not scale well with the number of blocks, N. This is due to the fact that the number of parameters is asymptotically proportional to N, and therefore the internal inverse Hessian approximation becomes increasingly inaccurate. In order to be able to efficiently learn θ_2 ($\mathbf{A}$ and $\mathbf{R}$) for large N, a stochastic gradient approach (SGA), (Robbins and Monro, 1951), is employed.

It is adapted here to estimation in block-based state-space models, considering only a single randomly and uniformly sampled block, n, at any given time. The likelihood term corresponding to block n is $\mathcal{L}(\theta_1^n, \theta_2)$, where $\theta_1^n = \{\mathbf{F}^n, \mathbf{C}^n, \mathbf{Q}^n, \psi^n\}$. The stochastic gradient update to be applied is computed at the current optimum with respect to θ_1^n,

$$\begin{aligned} \Delta\theta_2 &= \eta \frac{d\mathcal{L}(\hat{\theta}_1^n, \theta_2)}{d\theta_2}, \\ \hat{\theta}_1^n &= \arg\max_{\theta_1^n} \mathcal{L}(\theta_1^n, \theta_2). \end{aligned}$$

where $\hat{\theta}_1^n$ is estimated using the EM algorithm. Employing an appropriate 'cooling' of the learning rate, η, is mandatory in order to ensure convergence: one such, devised by Robbins and Monro (1951), is choosing η proportional to $\frac{1}{k}$ where k is the iteration number. In our simulations, the SGA seemed more robust to the initial parameter values than the QN and the EM algorithms.

5. Learning from Synthetic Data

In order to investigate the convergence of the algorithms, AR(2) processes with time-varying pole placement were generated and mixed through randomly generated filters. For each signal frame, $T = 200$, the poles of the AR processes were constructed so that the amplification, r, was fixed while the center frequency was drawn uniformly from $\mathcal{U}(\pi/10, 9\pi/10)$. The filter length was $L = 8$ and the coefficients of the mixing filters, that is, the $\mathbf{a}_{ij}$ of $\mathbf{A}$, were generated from i.i.d. Gaussians weighted by an exponentially decaying function. Quadratic mixtures with $Q = P = 2$ were used: the first 2 elements of $\mathbf{a}_{12}$ and $\mathbf{a}_{21}$ were set to zero to simulate a situation with different channel delays. All channel filters were normalized to $||\mathbf{a}_{ij}||_2 = 1$. Gaussian i.i.d. noise was added in each channel, constructing the desired signal to noise ratio.

For evaluation purposes, the signal-to-error ratio (SER) was computed for the inferred sources. The true and estimated sources were mapped to the output by filtering through the direct channel so that the true source at the output is $\tilde{s}_{i,t} = \mathbf{a}_{ii} * s_{i,t}$. Similarly defined, the estimated source at the sensor is $\hat{s}_{i,t}$. Permutation ambiguities were resolved prior to evaluating the SER,

$$\mathrm{SER}_i = \frac{\sum_t \tilde{s}_{i,t}^2}{\sum_t (\tilde{s}_{i,t} - \hat{s}_{i,t})^2}.$$

The EM and QN optimizers were applied to learn the parameters from $N = 10$ frames of samples with SNR $= 20$dB, $r = 0.8$. The algorithms were restarted 5 times with random initializations, $\mathbf{A}_{ij} \in \mathcal{N}(0,1)$, the one that yielded the maximal likelihood was selected. Figure 2 shows the results of the EM run: the close match between the true and learned models confirms that the parameters can indeed be learned from the data using maximum-likelihood optimization. In Table 1, the generative approach is contrasted with a stationary finite impulse response (FIR) filter separator that

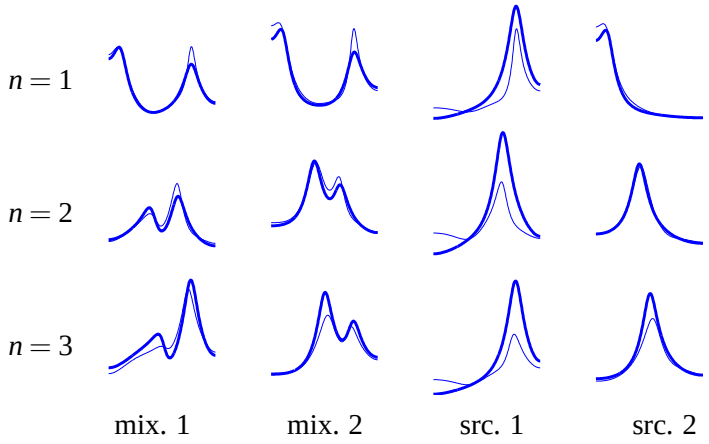

Figure 2: The true (bold) and estimated models for the first 3 frames of the synthetic data based on the autoregressive model. The amplitude frequency responses of the combined source and channel filters are shown: for source i, this amounts to the frequency response of the filter, with the scaling and poles of $\theta_{1,i}$ and zeros of the direct channel $\mathbf{a}_{ii}$. For the mixtures, the responses across channels were summed. The EM algorithm provided the estimates.

	Estimated	Generative	MSE FIR
AR	9.1 ± 0.4	9.7 ± 0.4	7.5 ± 0.2
HN	11.8 ± 0.7	13.2 ± 0.4	7.9 ± 0.5

Table 1: The signal-to-error ratio (SER) performance on synthetic data based on the autoregressive (AR) and harmonic-and-noise (HN) source models. Mean and standard deviation of the mean are shown for 1) the EM algorithm applied to the mixtures, 2) inferences from data and the true model, and, 3) the optimal FIR filter separator. The mean SER and the standard deviation of the mean were calculated from $N = 10$ signal frames, SNR $= 20$dB.

in a supervised fashion was optimized to minimize the squared error between the estimated and true sources, $L_{\text{FIR}} = 25$. Depending on the signal properties, the generative approach, which results in a time-varying filter, results in a clear advantage over the time-invariant FIR filter, which has to compromise across the signal's changing dynamics. As a result, the desired signals are only sub-optimally inferred by methods that apply a constant filter to the mixtures. The performance of the learned model is upper-bounded by that of the generative model, since the maximum likelihood estimator is only unbiased in the limit.

The convergence speed of the EM scheme is highly sensitive to the signal-to-noise ratio of the data, as was documented in Olsson et al. (2006), whereas the QN algorithm is more robust to this condition. In Bermond and Cardoso (1999), it was shown that the magnitude of the update of $\mathbf{A}$ scales inversely with the SNR. By varying the SNR in the synthetic data and applying the EM algorithm, it was confirmed that the predicted convergence slowdown occurs at high SNR. In contrast, the QN algorithm was found to be much more robust to the noise conditions of the data. Figure 3 shows the SER performance of the two algorithms as computed following a fixed number

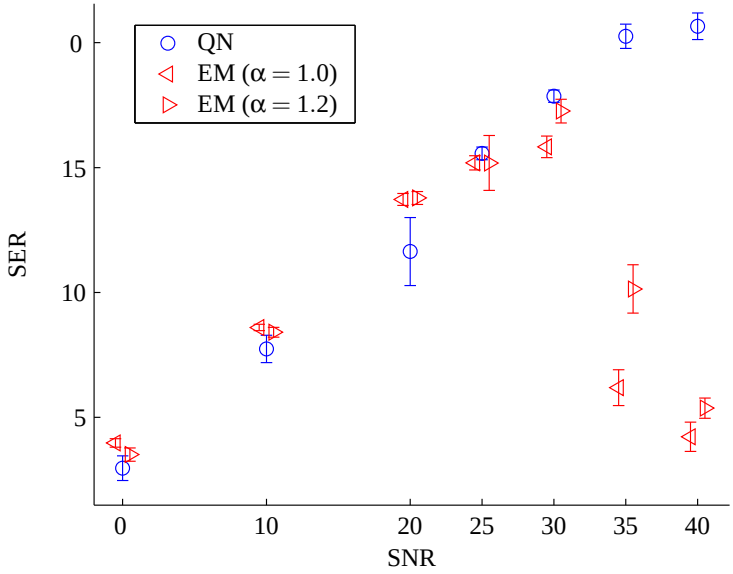

Figure 3: Convergence properties of the EM and QN algorithms as measured on the synthetic data (autoregressive sources). The signal-to-error ratio (SER) was computed in a range of SNR following 300 iterations. As the SNR increases, more accurate estimates are provided by all algorithms, but the number of iterations required increases more dramatically for the EM algorithm. Results are shown for the basic EM algorithm as well as for the step-size adjusted version.

of iterations ($i_{\max} = 300$). It should be noted that the time consumption per iteration is similar for the two algorithms, since a similar number of E-step computations is used (and E-steps all but dominate the cost).

For the purpose of analyzing the HN model, a synthetic data set was generated. The fundamental frequency of the harmonic component was sampled uniformly in a range, see Figure 4, amplitudes and phases, $K = 4$, were drawn from a Gaussian distribution and subsequently normalized such that $||\mathbf{c}_i|| = 1$. The parameters of the model were estimated using the EM algorithm on data, which was constructed as SNR $= 20$dB, HNR $= 20$dB. The fundamental frequency search grid was defined by 101 evenly spaced points in the generative range. In Figure 4, true and learned parameters are displayed. A close match between the true and estimated harmonics is observed.

In cases when the sources are truly harmonic and noisy, it is expected that the AR model performs worse than the HN model. This is due to the fact that a harmonic mean structure is required for the model to be unbiased. The AR model will compensate by estimating a larger variance, q_i, leading to suboptimal inference. In Figure 5, the bias is quantified by measuring the performance gap between the HN and AR models for varying HNR. The source parameters were estimated by the EM algorithm, whereas the mixing matrix, $\mathbf{A}$, was assumed known.

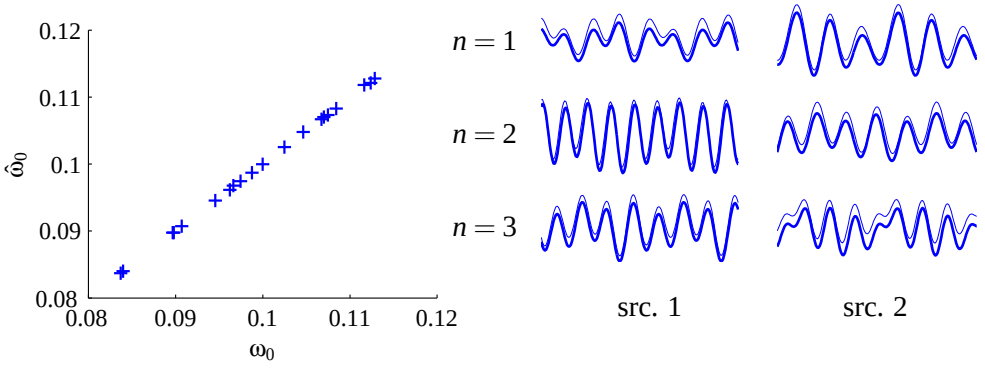

Figure 4: The true and estimated parameters from synthetic mixtures of harmonic-and-noisy sources as obtained by the EM algorithm. Left: fundamental frequencies in all frames. Right: the waveforms of the true (bold) and estimated harmonic components. For visualization purposes, the estimated waveform was shifted by a small offset.

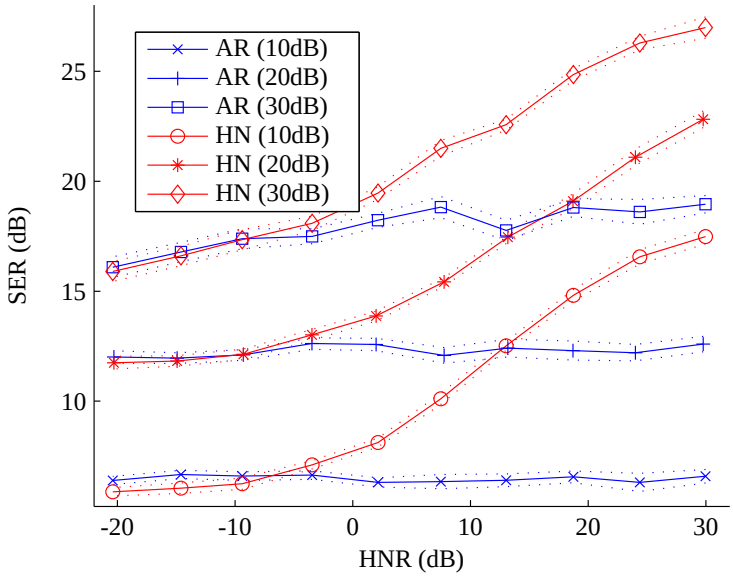

Figure 5: The signal-to-error ratio (SER) performance of the autoregressive (AR) and harmonic-and-noisy (HN) models for the synthetic data set ($N = 100$) in which the harmonic-to-noise ratio (HNR) was varied. Results are reported for SNR $= 10, 20, 30$dB. The results indicate that the relative advantage of using the correct model (HN) can be significant. The error-bars represent the standard deviation of the mean.

6. Speech Mixtures

The separation of multiple speech sources from room mixtures has potential applications in hearing aids and speech recognition software (see, for example, Parra and Spence, 2000). For this purpose,

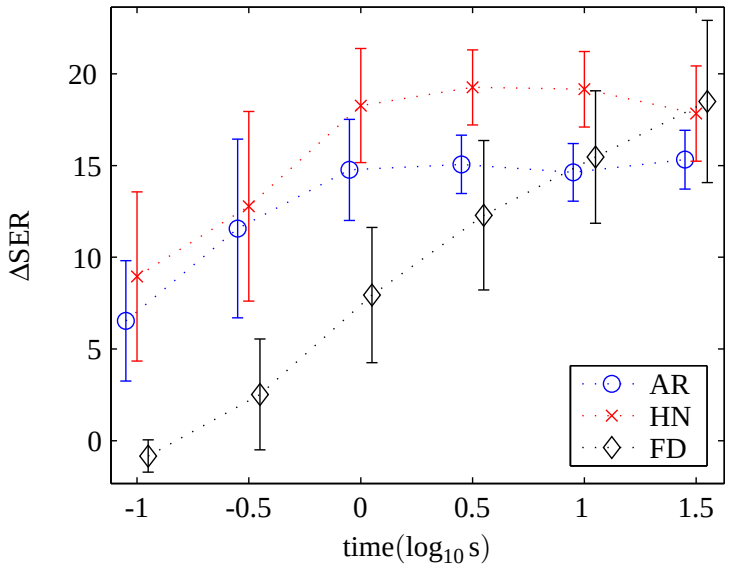

Figure 6: The separation performance (SER) on test mixtures as a function of the training data duration for the autoregressive (AR) and harmonic-and-noisy (HN) priors. Using the stochastic gradient (SG) algorithm, the parameters were estimated from the training data. Subsequently, the learned filters, $\mathbf{A}$, were applied to the test data, reestimating the source model parameters. The noise was constructed at 40dB and assumed known. For reference, a frequency domain (FD) blind source separation algorithm was applied to the data.

we investigate the models based on the autoregressive (AR) and harmonic-and-noisy source (HN) priors and compare with a standard frequency domain method (FD). More specifically, a learning curve was computed in order to illustrate that the inclusion of prior knowledge of speech benefits the separation of the speech sources. In Figure 6 is shown the relationship between the separation performance on test mixtures and the duration of the training data, confirming the hypothesis that the AR and HN models converge faster than the FD method. Furthermore it is seen that the HN model can obtain a larger SER than the AR model.

The mixtures were constructed by filtering speech signals (sampled at 8Hz) through a set of simulated room impulse responses, that is, $\mathbf{a}_{ij}$, and subsequently adding the filtered signals. The room impulse responses were constructed by simulating $Q = 2$ speakers and $P = 2$ microphones in an (ideal) anechoic room, the cartesian coordinates in the horizontal plane given (in m) by $\{(1,3),(3,3)\}$ and $\{(1.75,1),(2.25,1)\}$ for the speakers and microphones, respectively.[1]. This corresponds to a maximum distance of 1.25m between the speakers and the microphones, and a set of room impulse responses that are essentially Kronecker delta functions well represented using a filter length of $L = 8$.

1. A Matlab function, rir.m, implementing the image method (Allen and Berkley, 1979) is available at http://2pi.us/rir.html.

The SG algorithm was used to fit the model to the mixtures and subsequently infer the source signals. The speech data, divided into blocks of length $T = 200$, was preprocessed with a standard pre-emphasis filter, $H(z) = 1 - 0.95z^{-1}$, and inversely filtered prior to the SER calculations. From initial conditions ($q_i^n = 1$, $f_{i,j}^n = 0$, $c_{i,j}^n = 0$ and $a_{i,j,k}$ normally distributed, variance 0.01, for all i, j, n, k except $a_{1,1,1} = 1$, $a_{2,2,1} = 1$; $\omega_{0,i}^n$ was drawn from a uniform distribution corresponding to the interval $50 - 200$Hz), the algorithm was allowed $i_{\max} = 500$ iterations to converge and restarts were not necessary. The source model order was set to $p = 1$ (autoregression order) and in the case of the harmonic-and-noise model, the number of harmonics was set to $K = 6$. The complex JADE algorithm was employed in the frequency domain as the reference method (Cardoso and Souloumiac, 1993). In order to correct the permutations across the 101 frequencies, amplitude correlation between the bands was maximized (see, for example, Olsson and Hansen, 2006).

In order to qualitatively assess the effect of the two priors, a mixture of speech signals was constructed using $P = 2$ speech signals (a female and a male, shown in Figure 7a and b). They were mixed through artificial channels, $\mathbf{A}$, which were generated as in Section 5. Noise was added up to a level of 20dB. The EM algorithm was used to fit the source models to the mixtures. It is clear from Figure 7 c-f that the estimated harmonic model to a large extent explains the voiced parts of the speech signals, and the unvoiced parts to a lesser extent. In regions of rapid fundamental frequency variation, the harmonic part cannot be fitted as well (the frames are too long here). In Figure 7 g and h, the separation performances of the AR and HN models are contrasted. Most often, the HN performs better than the AR model. A notable exception occurs in the case when either speaker is silent, in which case the misfit of the HN model is more severe, suggesting that the performance can be improved by model control.

7. Conclusion

It is demonstrated that careful generative modelling is a viable approach to convolutive source separation and can yield improved results. Noisy observations, non-stationarity of the sources and small data volumes are examples of scenarios which benefit from the higher level of modelling detail.

The performance of the model was shown to depend on the choice of optimization scheme when the signal-to-noise ratio is high. In this case, the EM algorithm, which is often preferable for its conceptual and analytical simplicity, experiences a substantial slowdown, and alternatives must be employed. Such an alternative is a gradient-based quasi-Newton algorithm, which is shown to be particularly useful in low-noise settings. Furthermore, the required gradients are obtained in the process of deriving the EM algorithm.

The harmonic-and-noise model was investigated as a means to estimating more accurately a number of speech source signals from the their mixtures. Although a substantial improvement is shown to result when the sources are truly harmonic, the overall model is vulnerable to overfitting when the energy of one or more sources is locally near-zero. An improvement of the existing framework would be a model control scheme, such as variational Bayes, which could potentially cancel the negative impact of speaker silence. This is a topic for future research.

Acknowledgments

The authors thank the Oticon Fonden for providing financial support.

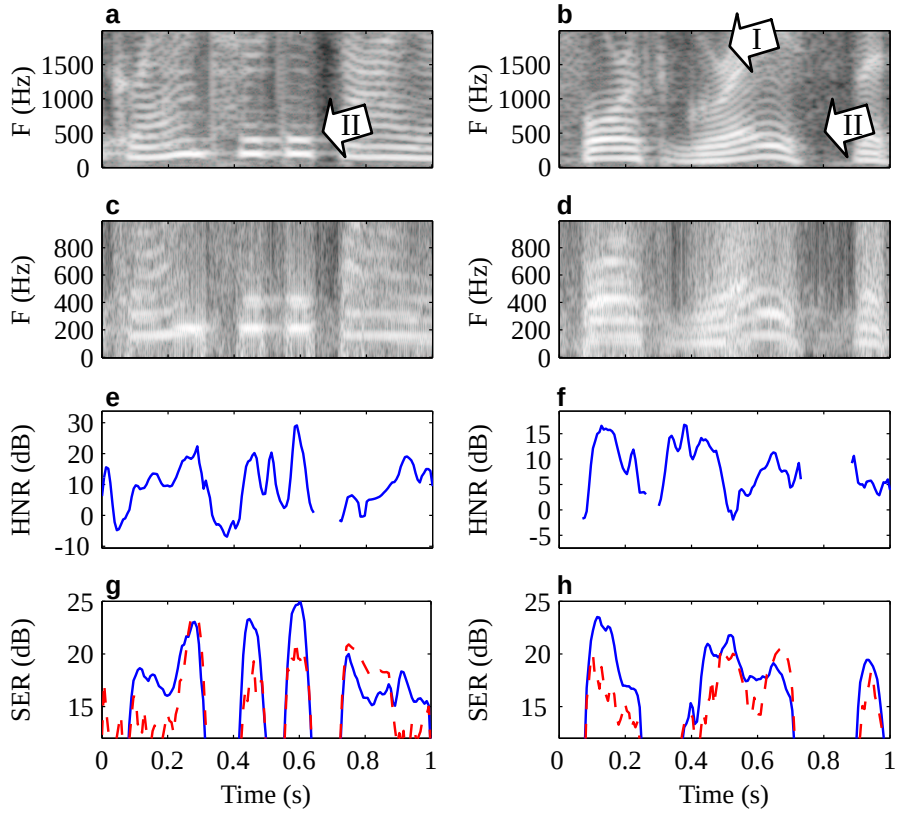

Figure 7: The source parameters of the autoregressive (AR) and harmonic-and-noisy (HN) models were estimated from $Q = 2$ convolutive mixtures using the EM algorithm. Spectrograms show the low-frequent parts of the original female (a) and male (b) speech sources. The appropriateness of the HN model can be assessed in c and d, which displays the re-synthesization of the two sources from the parameters ($K = 6$), as well as e and f, where the estimated ratio of harmonics to noise (HNR) is displayed. Overall the fit seem good, except at rapid variations of the fundamental frequency, for example, at (I), where the analysis frames are too long. The relative separation performance of the AR and HN models, which is shown in g and h for the two sources, confirms that the HN model is superior in most cases, with a notable exception in regions such as (II), where one of the speakers is silent. This implies a model complexity mismatch which is more severe for the more complex HN model.

Appendix A.

Below, an example of an M-step update derivation is shown for $\mathbf{F}^n$. As a by-product of the analysis, the derivative for the gradient-based optimizers appears. Care must be devised in obtaining the derivatives, since $\mathbf{F}^n$ is a structured matrix, for example, certain elements are one and zero. Therefore, the cost-function is expressed in terms of $\mathbf{f}_i^n$ rather than $\mathbf{F}^n$. Since all variables, which are here

indexed by the block identifier, n, are Gaussian, we have that:

$$\begin{aligned}\mathcal{J}(\theta) &= -\frac{1}{2}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\Big\{\log|\Sigma_i^n| + \Big\langle\left(\mathbf{s}_{i,1}^n-\mu_i^n\right)^T\left(\Sigma_i^n\right)^{-1}\left(\mathbf{s}_{i,1}^n-\mu_i^n\right)\Big\rangle\Big\} \\ &\quad +(T-1)\sum_{i=1}^{P}\log q_i^n + \frac{1}{q_i^n}\sum_{t=2}^{T}\sum_{i=1}^{P}\Big\langle\left(s_{i,t}^n-\left(\mathbf{f}_i^n\right)^T\mathbf{s}_{i,t-1}^n\right)^2\Big\rangle \\ &\quad +T\log\det\mathbf{R}+\sum_{t=1}^{T}\Big\langle\left(\mathbf{x}_t^n-\mathbf{A}\mathbf{s}_t^n\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_t^n-\mathbf{A}\mathbf{s}_t^n\right)\Big\rangle\Big].\end{aligned}$$

The vector derivative of $\mathcal{J}(\theta)$ with respect to $\mathbf{f}_i^n$ is:

$$\frac{d\mathcal{J}(\theta)}{d\mathbf{f}_i^n} = \frac{1}{q_i^n}\left[\sum_{t=2}^{T}\Big\langle\mathbf{s}_{i,t-1}^n\left(\mathbf{s}_{i,t-1}^n\right)^\top\Big\rangle\mathbf{f}_i^n-\sum_{t=2}^{T}\left\langle\mathbf{s}_{i,t-1}^n s_{i,t}^n\right\rangle\right].$$

This was the desired gradient, which is directly applicable in a gradient-based algorithm. By equating to zero and solving, the M-step update is derived:

$$\mathbf{f}_{i,\mathtt{new}}^n = \left[\sum_{t=2}^{T}\Big\langle\mathbf{s}_{i,t-1}^n\left(\mathbf{s}_{i,t-1}^n\right)^\top\Big\rangle\right]^{-1}\sum_{t=2}^{T}\left\langle\mathbf{s}_{i,t-1}^n s_{i,t}^n\right\rangle.$$

References

J. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65:943–950, 1979.

J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. ICA 2000*, pages 215–220, 2000.

A J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

O. Bermond and J.-F. Cardoso. Approximate likelihood for noisy mixtures. In *Proc. ICA*, pages 325–330, 1999.

M. Brandstein. On the use of explicit speech modeling in microphone array applications. In *Proc. ICASSP*, pages 3613–3616, 1998.

J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon. Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging. In *Proc. EUSIPCO*, pages 561–564, 2002.

J. F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F*, 140(6):362–370, 1993.

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistics Society, Series B*, 39:1–38, 1977.

M. Dyrholm and L. K. Hansen. CICAAR: Convolutive ICA with an auto-regressive inverse model. In *Proc. ICA 2004*, pages 594–601, 2004.

P. A. Højen-Sørensen, Ole Winther, and Lars Kai Hansen. Mean-field approaches to independent component analysis. *Neural Computation*, 14(4):889–918, 2002.

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering ASME Transactions*, 83:95–107, 1960.

T.W. Lee, A. J. Bell, and R. H. Lambert. Blind separation of delayed and convolved sources. In *Advances of Neural Information Processing Systems*, volume 9, page 758, 1997.

R.J. McAulay and T.F. Quateri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

M. McKeown, L.K. Hansen, and T.J. Sejnowski. Independent component analysis for fmri: What is signal and what is noise? *Current Opinion in Neurobiology*, 13(5):620–629, 2003.

E. Moulines, J. Cardoso, and E. Cassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP*, volume 5, pages 3617–3620, 1997.

H. B. Nielsen. UCMINF - an algorithm for unconstrained nonlinear optimization. Technical Report IMM-Rep-2000-19, Technical University of Denmark, 2000.

R. K. Olsson and L. K. Hansen. A harmonic excitation state-space approach to blind separation of speech. In *Advances in Neural Information Processing Systems*, volume 17, pages 993–1000, 2005.

R. K. Olsson and L. K. Hansen. Blind separation of more sources than sensors in convolutive mixtures. In *International Conference on Acoustics, Speech and Signal Processing*, 2006.

R. K. Olsson, K. B. Petersen, and T. Lehn-Schiøler. State-space models - from the EM algorithm to a gradient approach. *Neural Computation - to appear*, 2006.

L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions, Speech and Audio Processing*, pages 320–7, 5 2000.

B. A. Pearlmutter and L. C. Parra. A context-sensitive generalization of ICA. In *In Advances in Neural Information Processing Systems 9*, pages 613–619, 1997.

H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.

R. Salakhutdinov and S. Roweis. Adaptive overrelaxed bound optimization methods. In *International Conference on Machine Learning*, pages 664–671, 2003.

R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani. Optimization with EM and Expectation-Conjugate-Gradient. In *International Conference on Machine Learning*, volume 20, pages 672–679, 2003.

R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Anal.*, 3(4):253–264, 1982.

E. Weinstein, M. Feder, and A. V. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech and Audio Processing*, 1(4), 1993.

D. Yellin and E. Weinstein. Multichannel signal separation: Methods and analysis. *IEEE Transactions on Signal Processing*, 44(1):106–118, 1996.