Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Structuring Time Domain Blind Source Separation Algorithms for CASA Integration

Björn Schölling, Martin Heckmann, Frank Joublin, Christian Goerick

2006

Preprint:

This is an accepted article published in Proceedings of the ISCA Tutorial and Research Workshop on Statist. and Percept. Audition (SAPA 2006). The final authenticated version is available online at: https://doi.org/[DOI not available]

Structuring Time Domain Blind Source Separation Algorithms for CASA Integration

Björn Schölling[2]*,Martin Heckmann*[1]*,Frank Joublin*[1]*,Christian Goerick*[1]

[1] Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany
[2] Control Theory and Robotics Lab, Darmstadt University of Technology, D-64283 Darmstadt, Germany
bjoern.schoelling@rtr.tu-darmstadt.de

Abstract

Most algorithms based on Computational Auditory Scene Analysis (CASA) for binaural speech separation do not have the ability to inhibit already localized and for a long time present sources in the auditory scene. This has the major drawback that the auditory cues of weaker and new sources are subject to interference from already localized and perceived signals and the separation performance is worse if the signals overlap in their processing domain. In this paper we outline how one can build intuitively a separation system that has this inhibition feature. The main block and starting point of our derivation is a simple cross correlation based localization system with two microphones. The inhibition is achieved by feeding back localization results to a filter and sum structure that cancels localized sounds. Interestingly, our intuitive approach leads to a special case of a well known time domain blind source separation algorithm which was derived from a statistical signal processing viewpoint and exhibits good convergence even in reverberant environments. Finally, we discuss how the insights gained from building a blind source separation this way can be used to integrate CASA techniques.

1. Introduction

A common strategy of CASA based speech separation algorithms, e.g. [1, 2, 3, 4] is to use time-frequency masks for the extraction of different sound signals. The masks are derived from different auditory cues, i.e. localization, pitch, and are applied to the incoming signal in the time-frequency plane. The separation output is obtained by converting the time-frequency representation to a normal time domain signal. This processing achieves in general good and robust results for strong and sparse non-overlapping sources, however it has problems dealing with weak and overlapping sources. Annoying artifacts and no distinct separation are the consequences.

An alternative to overcome this lies in modeling the convolutive mixing process of the sound sources. An inversion of the model yields then the desired speech sources. This approach has the advantage that in theory perfect separation is possible and no musical artifacts occur. However, most known blind source separation algorithms and methods, e.g. [5], so far have problems with strong reverberated sources and fail to separate sources sufficiently as a huge number of mixing parameters have to be identified and tracked over time.

A promising approach to speed up and regularize the parameter estimation is to integrate techniques used in CASA systems into the pure signal processing motivated algorithms. However, the integration proves to be difficult as most blind source separation algorithms are derived from abstract cost functions and the resulting updates lack intuitive meaning or the meaning is hard to see in the equations when derived this way.

Therefore, we will take a first step into this direction by showing how to build intuitively a source separation system using two basic building blocks, i.e. source localization and inhibition. Tackling the problem this way helps to structure the blind source separation problem and offers insights in which parts of the processing block speech signal properties and perceptual knowledge in general might be helpful to increase performance.

2. Building Blocks

In this section the two building blocks of the system are described. For their motivation and derivation we first assume the presence of only one source. Later on we will lift this restriction.

2.1. Generalized Cross Correlation (GCC) & Localization

The key component of the system is the generalized cross correlation as it provides reliable estimates for signal time delay between microphones when only one speaker is active. The general definition of the correlation can be written as

$$\varphi_{x_1x_2}(l) = \mathrm{IDTFT}\left\{G(e^{j\Omega})\Phi_{x_1x_2}(e^{j\Omega})\right\} \tag{1}$$

where x_1, x_2 denote the two microphone signals, $\Phi_{x_1x_2}(e^{j\Omega})$ the cross power spectrum of x_1 and x_2 and $G(e^{j\Omega})$ is a weighting filter. In practice the above equation is replaced by a DFT and as weighting filter $G(e^{j\Omega}) = 1/|\Phi_{x_1x_2}(e^{j\Omega})|$ is used to sharpen the cross correlation function $\varphi_{x_1x_2}(l)$. Eq. 1 describes then the so called Phase Transform (PHAT) and reliable estimates for the time delay $\Delta_{x_1x_2}$ can be obtained by maximization, that is

$$\Delta_{x_1x_2} = \arg\max_l\{\varphi_{x_1x_2}(l)\}\ . \tag{2}$$

Although designed for a free field signal model, the above method also works in low reverberant environments [6] and we thus are able to identify the relative delay of the main paths of both impulse responses from the signal to both microphones and in direct consequence the direction of arrival.

2.2. Inhibition of known directions

With this knowledge of the main path delay of one signal we are now able to inhibit the signal by delaying and subtracting the microphone signals. Instead of detecting the delay once with the help of the GCC $\varphi_{x_1x_2}$ between the two input signals x_1 and x_2, we use feedback by considering the correlation $\varphi_{y_1x_2}$ between one

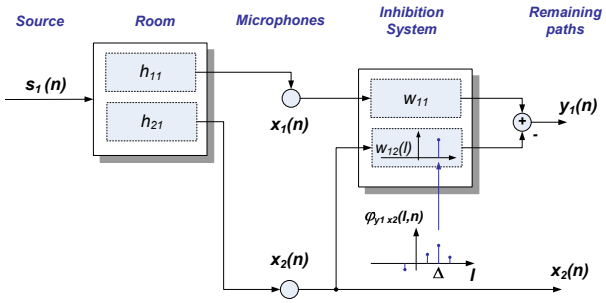

Figure 1: *Causal FIR inhibition system for localized speech. Instead of applying the cross correlation between the two signals x_1 and x_2 the correlation is now computed between the inhibited y_1 and input signal x_2. The benefit is that error feedback is used to identify other cross path delays.*

input, e.g. x_2, and the suppressed output y_1. Fig. 1 depicts the situation. Only one speech signal $s_1(n)$ is active and received by two microphones. Due to the spatial offset and echoes in the room the signals $x_i(n)$ at the microphones can be written as convolution $x_i(n) = h_{i1}(n) \star s_1(n)$ where h_{i1} are the room impulse responses from source 1 to microphone i. The inhibition is performed by a FIR filter and sum structure with length $2L+1$. The filter $w_{11}(l,n) = \delta(l-L)$ is held constant such that a signal delay of L taps for causal filtering is realized. Filter $w_{12}(l,0)$ is initialized at time $n=0$ with all zeros and adapted according to the time delay estimate $\Delta_{y_1x_2}$ of the GCC $\varphi_{y_1x_2}(l)$ between the output of the inhibition system $y_1(n)$ and the unprocessed received signal $x_2(n)$:

$$w_{12}^{i+1}(l,n) = w_{12}^{i}(l,n) + \mu_1 \cdot \delta_{l-(L+\Delta^i_{y_1x_2})} \varphi_{y_1x_2}(\Delta^i_{y_1x_2}) \quad (3)$$

The consequence of this adaption with step size μ_1 is that the system will suppress the detected main room impulse response path by subtracting the correct aligned sensor signals $x_1(n)$ and $x_2(n)$. A repetition of the update rule in Eq. 3, denoted by i, allows then to explain and identify other prominent delays in the inter sensor transfer function $\tilde{H}_{21}(z) = H_{11}(z)/H_{21}(z)$ that maps $X_2(z)$ to $X_1(z)$ as the new correlation $\varphi^{i+1}_{y_1x_2}(l)$ at step $i+1$ takes place between the inhibited/filtered signal

$$y_1^{i+1}(n) = x_1(n-L) - w_{12}^{i+1} \star x_2 \quad (4)$$

and $x_2(n)$.

The above inhibition can therefore be interpreted as a channel estimation method and works best on sparse channels. However, we can also extend the method to adaptation of all taps when we drop the maximum search and adapt all filter weights proportional to the GCC. Of special importance in this case is the version with the weighting function $G(e^{j\Omega}) = 1/\Phi_{x_2x_2}(e^{j\Omega})$ resulting in the so called Roth processor which estimates the linear filter mapping from x_2 to y_1 and provides therefore by itself an estimate of the inter sensor transfer function $\tilde{H}_{21}$ [7], which is then averaged and refined through multiple iterations i. The complete formula with DFT implementation of the cross correlation reads for the full update

$$\mathbf{w}_{21}^{i}(n) = \mathbf{w}_{21}^{i-1}(n) + \mu_2 \mathbf{B} \cdot \mathbf{F}^{-1} \left(\hat{\mathbf{\Phi}}_{y_1x_2} \oslash \hat{\mathbf{\Phi}}_{x_2x_2} \right) \quad (5)$$

where $\mathbf{F}^{-1}$ is an inverse FFT matrix of size $N \times N$, $\oslash$ denotes element wise division of vector elements and $\hat{\mathbf{\Phi}}_{y_1x_2}$ resp. $\hat{\mathbf{\Phi}}_{x_2x_2}$ are vector DFT estimates of the cross and normal power spectrum. The shift & window matrix $\mathbf{B}$ of size $(2L+1) \times N$ extracts the needed filter coefficients from the longer inverse FFT vector by swapping the FFT halves and shortening the correlation.

Through the inhibition we are now able to separate a later impinging signal $s_2(n)$ from $s_1(n)$ as $y_1(n)$ was trained to cancel $s_1(n)$ and thus contains only the other active signals which is in this case only $s_2(n)$. Another positive effect of inhibition is that the localization accuracy increases over time as the channel is estimated more and more precisely. This is a great advantage in reverberant scenarios in comparison to single snap shot localization systems that integrate localization measurements with some model over time and do not feed back information on the channel to refine and correct their direction of arrival estimate.

3. Combining Blocks

For recovery of $s_1(n)$ we have to add another copy of the above blocks to the system as shown in Fig. 2. Under the assumption that the inhibition system for s_1 has already converged, a good filtered reference $y_1^{s_2}$ of s_2,

$$\begin{aligned} y_1(n) &= y_1^{s_1}(n) + y_1^{s_2}(n) && (6)\\ &\approx y_1^{s_2}(n) && (7)\\ &= w_{11} \star x_1^{s_2}(n) - w_{12} \star x_2^{s_2}(n) && (8) \end{aligned}$$

is available at the output y_1. The superscript ^{s_2} denotes the portion of the corresponding signal in the mixture signal. With this reference we can then estimate parts or the full virtual linear cross filter $h^{\text{Roth}}_{y_1y_2}$ from y_1 to y_2 using the GCC method. For the Roth processor we get in the frequency domain

$$H^{\text{Roth}}_{y_1y_2}(e^{j\Omega}) = \frac{\Phi_{y_2y_1}(e^{j\Omega})}{\Phi_{y_1y_1}(e^{j\Omega})} \approx \frac{\Phi_{y_2^{s_2}y_1^{s_2}}(e^{j\Omega})}{\Phi_{y_1^{s_2}y_1^{s_2}}(e^{j\Omega})} \quad (9)$$

where the last term can be obtained by using the fact that both signals are independent and the system has perfectly suppressed signal s_1 in y_1. In practice a further ϵ is added to the denominator in Eq. (9) to avoid division by zero. A closer look at the above equation (9) shows that the cross correlation $H^{\text{Roth}}_{y_1y_2}(e^{j\Omega})$ can be interpreted as virtual optimum channel estimation in the Wiener sense between the filtered version of signal s_2 in y_1, i.e. $y_1^{s_2}$, and the filtered version at output 2, that is $y_2^{s_2}$. The open question that remains

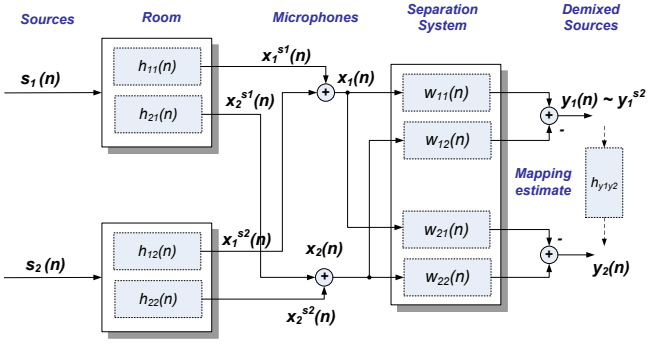

Figure 2: *Full 2x2 separation system. For better understanding the upper part is assumed to have converged.*

is how to use the mapping estimate $h_{y_1y_2}^{\text{Roth}} = \text{IDTFT}\left\{H_{y_1y_2}^{\text{Roth}}\right\}$ for the inhibition update rule. The answer is found from simple algebra as the relationship of the virtual mapping filter $h_{y_1y_2}^{\text{Roth}}$ can be expressed in terms of all filters and signals:

$$h_{y_1y_2}^{\text{Roth}} \star y_1^{s_2} = y_2^{s_2} \tag{10}$$

$$h_{y_1y_2}^{\text{Roth}} \star (w_{11} \star x_1^{s_2} - w_{12} \star x_2^{s_2}) = \ ... \\ (w_{22} \star x_2^{s_2} - w_{21} \star x_1^{s_2}) \tag{11}$$

$$h_{y_1y_2}^{\text{Roth}} \star (w_{11} \star h_{12} \star s_2 - w_{12} \star h_{22} \star s_2) = \ ... \\ (w_{22} \star h_{22} \star s_2 - w_{21} \star h_{12} \star s_2) \tag{12}$$

Rearranging terms for the unknown room impulse responses h_{12} and h_{22} yields then the desired inter sensor relationship in terms of the known current demixing and virtual filters ($h_{y_1y_2}^{\text{Roth}}$):

$$(w_{21} + w_{11} \star h_{y_1y_2}^{\text{Roth}}) \star h_{12} = (w_{22} + w_{12} \star h_{y_1y_2}^{\text{Roth}}) \star h_{22} \tag{13}$$

$$(w_{21} + w_{11} \star h_{y_1y_2}^{\text{Roth}}) \star h_{12} - (w_{22} + w_{12} \star h_{y_1y_2}^{\text{Roth}}) \star h_{22} = 0 \tag{14}$$

The above equality can then be used to find the new optimum separating solution w_{21}^{opt}, w_{22}^{opt} for the filters w_{21}, w_{22}.

$$\begin{aligned} y_2^{s_2} &= w_{22}^{\text{opt}} \star x_2^{s_2} - w_{21}^{\text{opt}} \star x_1^{s_2} \stackrel{!}{=} 0 && (15)\\ &= w_{22}^{\text{opt}} \star h_{22} \star s_2 - w_{21}^{\text{opt}} \star h_{12} \star s_2 && (16)\\ &= \left(w_{22}^{\text{opt}} \star h_{22} - w_{21}^{\text{opt}} \star h_{12}\right) \star s_2 && (17) \end{aligned}$$

By comparing the terms in Eq. (17) with the ones in (14), as optimal solution $w_{21}^{\text{opt}} = w_{21} + w_{11} \star h_{y_1y_2}^{\text{Roth}}$ and $w_{22}^{\text{opt}} = w_{22} + w_{12} \star h_{y_1y_2}^{\text{Roth}}$ is found.

In practice the mapping estimate is not exact as we have leakage from signal s_1 into y_1, additional sensor noise and approximation errors in the computation of the GCC, such that a direct computation of the optimal coefficients is not robust. We therefore fallback to our iterative step wise inhibition as introduced for the single signal case (Eq. 3):

$$\begin{aligned} w_{21}^i &= w_{21}^{i-1} + \mu_3 \cdot w_{11}^{i-1} \star h_{y_1y_2}^{\text{Roth},i-1} && (18)\\ w_{22}^i &= w_{22}^{i-1} + \mu_3 \cdot w_{12}^{i-1} \star h_{y_1y_2}^{\text{Roth},i-1} && (19)\\ w_{11}^i &= w_{11}^{i-1} + \mu_3 \cdot w_{21}^{i-1} \star h_{y_2y_1}^{\text{Roth},i-1} && (20)\\ w_{12}^i &= w_{12}^{i-1} + \mu_3 \cdot w_{22}^{i-1} \star h_{y_2y_1}^{\text{Roth},i-1} && (21) \end{aligned}$$

In comparison to the previous mentioned one signal case we also relaxed the constant delay constraint on the diagonal filters w_{11}, w_{22}. The reason for this is that we need a compensation for the filtering introduced by the cross filters and adapting the diagonal filters is the easiest way to solve this.

4. Relation to Other Approaches

An interesting finding when looking at the full update equations in (18)-(21) is that the GCC $h_{y_1y_2}^{\text{Roth}}$ with Roth weighting is the Wiener filter that optimally tries to estimate y_2 from y_1. If we assume FIR structure for the $2 \cdot L + 1$-tap filter, we can also compute its equivalent time domain solution with correlation matrices:

$$\mathbf{h}_{y_1y_2}^{\text{Roth}} = \mathbf{r}_{y_2y_1}^T \mathbf{R}_{y_1y_1}^{-1} \ , \tag{22}$$

where $\mathbf{r}_{y_2y_1}$ is a $2 \cdot L + 1$ vector that holds cross correlation values, i.e. $r_{y_2y_1,i} = \text{E}\left\{y_2(n)y_1(n - L + i)\right\}$ and the autocorrelation matrix with a $2 \cdot L + 1$ data vector $\mathbf{y_1}^T = [y_1(n)\ y_1(n-1)\ ...\ y_1(n - 2 \cdot L + 1)]$ is defined as $\mathbf{R}_{y_1y_1} = \text{E}\left\{\mathbf{y}_1(n)\mathbf{y}_1(n)^T\right\}$. A comparison of our update with the above channel estimate in (22) with the natural gradient update rule in Buchner et al. [8] (equation 31 on page 125) shows that both updates are structurally identical. Furthermore, this finding sheds new light on the fast convergence of the algorithm in comparison to other updates which result from different cost functions. It seems that the good convergence results from the fact that the virtual channel from y_1 and y_2 is estimated in an "optimum" way and its adaptation is fastest when only one signal is active as the inverse matrices scale the step sizes of the corresponding inhibition filters. In addition the inverse matrices can be interpreted as being responsible for removing time structure, i.e. periodicity of voiced speech and correlation in speech over time in general, from the normal cross correlation. This removal is very beneficial for good convergence as periodicity in the cross correlation leads to strong misadaptions in the demixing filters and some time is needed for averaging out this effect.

A major open point of our intuitive approach so far was the operation behavior at the beginning when neither system has converged. With the above link that the robust natural gradient update equations from the Buchner et al. system [8] can be related to a special case of our system with the Roth processor, the same reasoning as in [8] holds and the convergence analysis carries over.

Results on the convergence and performance of the single inhibition block in section 2.2 are also available as the full inhibition using the Roth processor (Eq. 5) can be identified as a block adaptive FIR filter. Youn et al. use the same structure for time delay estimation of sonar signals and apply a sample by sample LMS update [9]. In general the structure is known as *Adaptive Noise Canceler* [10] and is often used for noise cancellation where a filtered reference signal is available. If the constant delay assumption on the direct filter $w_{11}(l)$ is dropped and also made adaptive, the inhibition system resembles a SIMO blind channel identification system and the filter coefficients can be updated by any algorithm that exploits the cross relation among channels, see [11, 12, 13] for details.

5. Simulations

In order to demonstrate the working principle of the building blocks, we performed simulations with artificially convolved speech data sampled at 16 kHz. As impulse responses a low demand scenario with measured Head Related Transfer Functions of tap length 60 from the CIPIC database [14] was chosen. Finally, spatially uncorrelated white noise was added to the mixture, such that the overall SNR is approximately 12 dB and 30 dB respectively. The GCC was estimated with FFTs of size 2048, the total demixing filter length L was 100 and the number of iterative refinements 10. After one frame was processed the data was shifted 50 samples.

Fig. 3 compares the performance of full and main path inhibition for a male speaker, cf. Eq. (3) and (5). The step sizes have been chosen empirically and are $\mu_1 = 0.003$ (PHAT single tap) and $\mu_2 = 0.001$ (Roth FFT, Roth TD). In order to avoid fluctuations due to strong time structure in the cross correlation, divisions by small values in the frequency domain are suppressed by adding a small epsilon of 0.01 to the denominator in Eq. 9. In the time domain update the inverse auto correlation matrix is regularized by a diagonal loading of $0.01\mathbf{I}$. The effect of structure in the signals on the system performance can be clearly seen in the performance plot for the inhibition of one signal which measures the total energy of the resulting filter $a(n)$ from s_1 to y_1 at each processed frame

of length $t = 0.128\,\text{sec}$, i.e. $||a||^2 = ||w_{11} \star h_{11} - w_{12} \star h_{21}||^2$. At frame instances where only voiced speech is available, the adaptation is slow when unvoiced parts are present the channel can be identified much more reliable. The performance of the full source

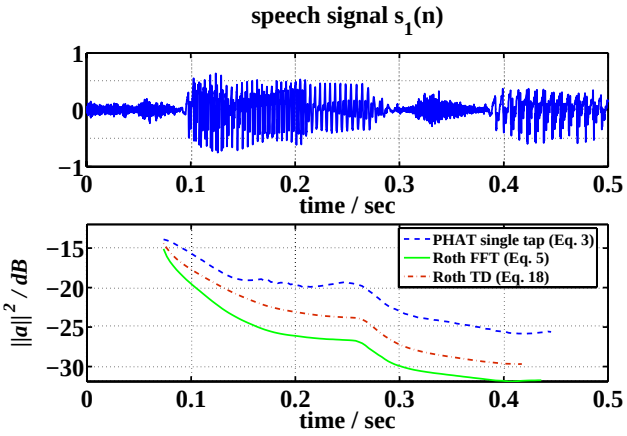

Figure 3: *Typical performance of the inhibition system for one active speech signal in white noise* SNR $= 12\,dB$. *The signals are mixed with 60 tap HRTFs.*

separation system is depicted in Fig. 4. The plot shows now the Signal To Interference Ratio for each output y_i with normalized input signals s_i, i.e.

$$\text{SIR}_1 = \frac{\text{var}\{y_1^{s_2}\}}{\text{var}\{y_1^{s_1}\}} = \frac{||w_{11} \star h_{12} - w_{12} \star h_{22}||^2}{||w_{11} \star h_{11} - w_{12} \star h_{21}||^2}. \qquad (23)$$

A weighting with the power of the sources is omitted as both sources have been normalized for the experiment. From the plot it is again

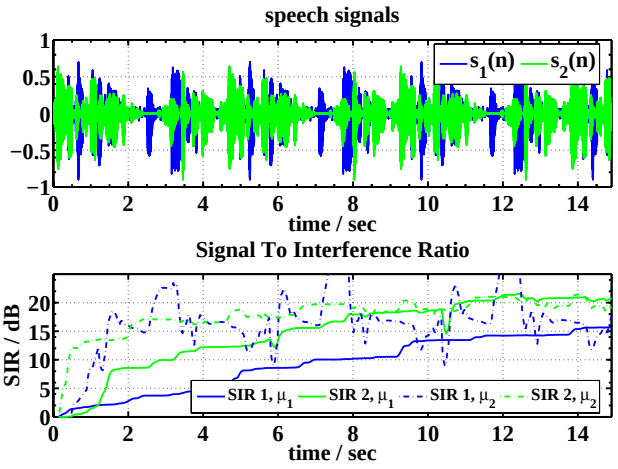

Figure 4: *Typical performance of the source separation system for two speech signals in white noise* SNR $= 30\,dB$ *(*$\mu_1 = 10^{-4}$ *and* $\mu_2 = 8 \cdot 10^{-4}$*, update via Eq. 18). For convenience the same speech signals have been repeated 4 times and two different step sizes are shown*

evident that the algorithm slows down and has even local problems where the severity of the breakdown depends on the stepsize of the update. To solve this problem better strategies for regularization of the rank deficient auto correlation matrix in the time domain or division by zero handling in the frequency domain are needed. The overall separation quality of the system is however very good. Due to the additive noise, only one speaker is audible and linear filtering artifacts are small. The convergence is very fast if a suitable step size is chosen, e.g. $\mu_2 = 0.08$.

To evaluate the performance in more realistic scenarios we conducted a second experiment. The impulse responses are now 500 taps of measured responses between a loudspeaker and 3.5 m apart microphones. The recordings took place in a normal room (6 m $\times$ 4 m $\times$ 2.5 m). In Fig. 5 results for channel estimation in the presence of only one source are presented. The GCC is estimated again with FFTs of size 2048, the total demixing filter length L is now 600 and the number of iterative refinements 10. After one frame was processed the data was shifted 512 samples. Convergence is in this case much slower and occurs after 2 seconds. The complete source separation system works also in this difficult

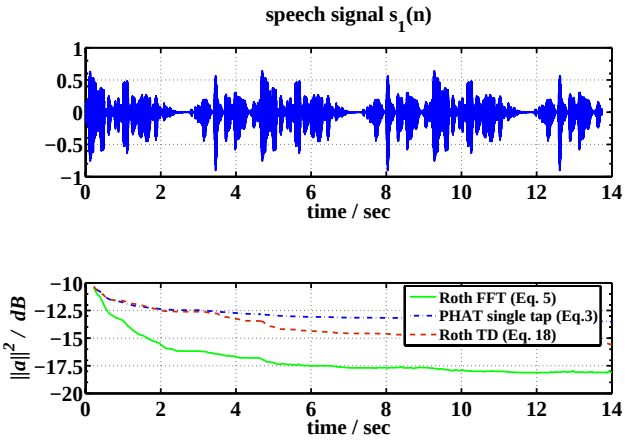

Figure 5: *Typical performance of the inhibition system for one active speech signal in a reverberant environment (500 tap impulse response).*

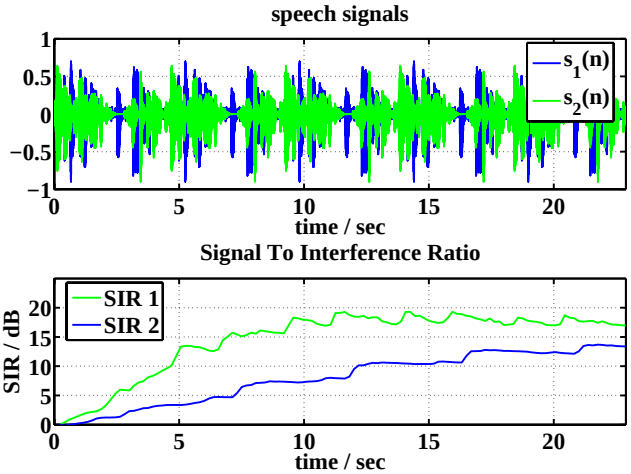

Figure 6: *Typical performance of the source separation system for two active speech signals in a reverberant environment in white noise* SNR $= 30\,dB$ *(*$\mu_2 = 0.001$*, update via Eq. 18)*

scenario as Fig. 6 shows. The overall performance is not as good as

for short impulse responses. However, a significant enhancement of one target source in comparison to the other is achieved.

6. Discussion

Our intuitive source separation system has the ability to separate sources. However, its performance depends on the number of mixing parameters to be estimated and only in "low demand" scenarios a "true" separation is possible. In order to achieve better results for real reverberant environments, strategies are needed that improve the recursive channel estimation in Eq. 9 which suffers from two problems. The first is that the other unwanted signal, e.g. s_1 leaks into the output y_1 and the transfer function estimation from $y_1^{s_2}$ to $y_2^{s_2}$ is impaired. The other problem is that the solution must be regularized for non full-band signals, e.g. periodic sequences in speech. In mathematical terms both problems can be seen when we expand the Wiener filter exactly in (9):

$$H_{y_1 y_2}^{\mathrm{Roth}}(e^{j\Omega}) = \frac{\Phi_{y_2 y_1}(e^{j\Omega})}{\Phi_{y_1 y_1}(e^{j\Omega})} \quad (24)$$

$$= \frac{\Phi_{y_2^{s_1} y_1^{s_1}}(e^{j\Omega}) + \Phi_{y_2^{s_2} y_1^{s_2}}(e^{j\Omega})}{\Phi_{y_1^{s_1} y_1^{s_1}}(e^{j\Omega}) + \Phi_{y_1^{s_2} y_1^{s_2}}(e^{j\Omega})} \quad (25)$$

Leakage is identified in the terms $\Phi_{y_1^{s_1} y_1^{s_1}}$ and $\Phi_{y_2^{s_1} y_1^{s_1}}$. Excitation problems occur when the denominator approaches small values for some frequencies Ω.

Having identified the problems, one can look for solutions with the help of techniques known from Computational Auditory Scene Analysis. CASA could for example come into play as it can be used to identify only parts of the channel frequency response where the energy contribution of the target signal is much higher than that of the interference. Other parts of the frequency response are not updated at this time step. When the signal ratios at different frequencies change over time, the missing parts are updated. This processing strategy is similar to the one of Nakatani et al. [15] who uses partial time instance frequency response estimation for blind single channel dereverberation.

7. Conclusions

An intuitive way to convolutive blind source separation has been presented. Instead of deriving update equations from an abstract cost function, the update rule was developed from source localization and inhibition principles. Furthermore, it was shown that the GCC and especially the Roth processor play an important role in designing fast converging systems. With the new insight how channel estimation between the outputs is linked to inhibition, a promising way to improve convergence has been opened. The introduced processing blocks structure the source separation problem and a control and replacement of the algorithms can happen this way more easily. We especially aim at integrating Computational Auditory Scene Analysis (CASA) ideas into the system.

8. Acknowledgments

The authors would like to thank Andreas Walstra for helpful discussions in the development of this work.

9. References

[1] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, pp. 297–336, 1994.

[2] Martin Cooke and Daniel P.W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 33, pp. 141–177, 2000.

[3] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. On Neural Networks*, vol. 10, pp. 684–697, 1999.

[4] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. On Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[5] L. Parra and C. Spence, "Convolutive blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[6] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

[7] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[8] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, January 2005.

[9] D.H. Youn, N. Ahmed, and G.C. Carter, "On using the LMS algorithm for time delay algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 5, pp. 798–801, October 1982.

[10] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, December 1975.

[11] G. Dong and R. Liu, "Adaptive blind channel identification," in *Proc. 1996 IEEE Int. Conf. Communications (ICC 96)*, 1996, pp. 828–831.

[12] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 1951–1968, October 1998.

[13] Jacob Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 384–391, January 2000.

[14] V. Algazi, R. Duda, and D. Thompson, "The CIPIC HRTF database," 2001.

[15] T. Nakatani, M. Miyoshi, and K. Kinoshita, *Speech Enhancement*, chapter Single-Microphone Blind Dereverberation, pp. 247–270, Signals and Communication Technology. 2005.