FROM SOURCE LOCALIZATION TO BLIND SOURCE SEPARATION: AN INTUITIVE ROUTE TO CONVOLUTIVE BLIND SOURCE SEPARATION

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

[2]bjoern.schoelling@rtr.tu-darmstadt.de
[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

ABSTRACT

Most algorithms for blind source separation (BSS) of convolutive speech mixtures are derived in a deductive way from abstract statistical principles and exploit a combination of three signal properties, i.e. nongaussianity, nonwhiteness and nonstationarity. In this paper we show how a separation system can be build the opposite, inductive, way using basic speech processing building blocks. The main block and starting point of our derivation is a simple generalized cross correlation based localization system with two microphones. The capability of source separation (2 signals and 2 sensors) is added by duplicating the localization structure and adding an inhibition mechanism which suppresses already localized sounds. Finally, experimental results with artificially mixed speech signals are presented.

1. INTRODUCTION

Blind source separation (BSS) with FIR demixing filters is one promising approach to solve the so called cocktail party problem. Here multiple speakers are talking at the same time and a separation of one source from the others using multiple microphone recordings is aimed for. Good results in low reverberant environments have been achieved, e.g. [1], [2], by applying the principles and algorithms for instantaneous mixtures in the DFT domain. However, these methods need some extra repair measures as through the scaling and permutation ambiguity inherent to BSS the separated signals differ from frequency bin to bin and have to be aligned for proper reconstruction [1]. The most common technique is to use localization and similarity information across frequencies after separation. This post processing is however unnatural and to a certain degree error-prone and thus a more exact time domain modeling for convolutive BSS as proposed for example by Buchner et al. [3] seems more appropriate and suitable to deal with long impulse responses. The algorithmic derivation in [3] is statistically motivated and driven by the abstract concepts of exploiting nonwhiteness and nonstationarity. A physical interpretation of the resulting update equations of the separating filters is neither given nor easy to find when deduced this way. In order to gain more insight into BSS in the time domain, we therefore construct in the following a two signal two sensor FIR separation system by extending the well known Generalized Cross Correlation (GCC) method [4] for Time Delay Estimation (TDE). As a special case the natural gradient update equations of the Buchner system are found.

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 & 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

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

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

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

Although designed for a free field signal model, the above method also works in low reverberant environments [5] 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. Fig. 1 depicts the situation. Only one speech signal $s_1(n)$ is active and received

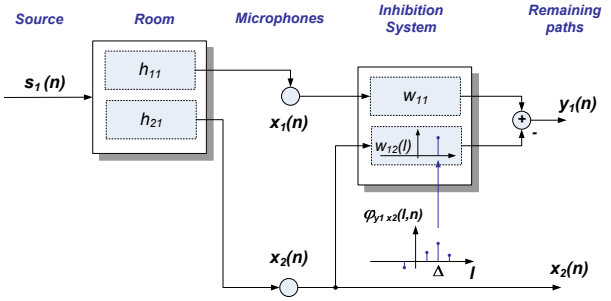

Figure 1: Causal FIR inhibition system for localized speech.

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 $2L+1$ where for the moment 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 $x_2(n)$:

$$w_{12}^{i+1}(l,n) = w_{12}^{i}(l,n) + \mu_1 \cdot \delta(l-(L+\Delta_{y_1x_2}^{i})) \quad (3)$$

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 $\varphi_{y_1x_2}^{i+1}(l)$ 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 $\tilde{H}_{21}$ [4], 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 $s_2(n)$.

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 $h_{y_1y_2}^{\mathrm{Roth}}$ from y_1 to y_2 using the GCC method. For the Roth processor we get in the frequency domain

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

where the last term can be obtained by using the fact that both signals are independent. 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 $H_{y_1y_2}^{\mathrm{Roth}}(e^{j\Omega})$ 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 is how to use the mapping estimate $h_{y_1y_2}^{\mathrm{Roth}} = \mathrm{IDTFT}\left\{H_{y_1y_2}^{\mathrm{Roth}}\right\}$ for the inhibition update rule. The answer is found from simple algebra as the relationship of the virtual mapping filter $h_{y_1y_2}^{\mathrm{Roth}}$ can be expressed in terms of all filters and signals:

$$h_{y_1y_2}^{\mathrm{Roth}} \star y_1^{s_2} = y_2^{s_2} \quad (10)$$

$$h_{y_1y_2}^{\mathrm{Roth}} \star (w_{11} \star x_1^{s_2} - w_{12} \star x_2^{s_2}) = \ldots \quad (w_{22} \star x_2^{s_2} - w_{21} \star x_1^{s_2}) \quad (11)$$

$$h_{y_1y_2}^{\mathrm{Roth}} \star (w_{11} \star h_{12} \star s_2 - w_{12} \star h_{22} \star s_2) = \ldots \quad (w_{22} \star h_{22} \star s_2 - w_{21} \star h_{12} \star s_2) \quad (12)$$

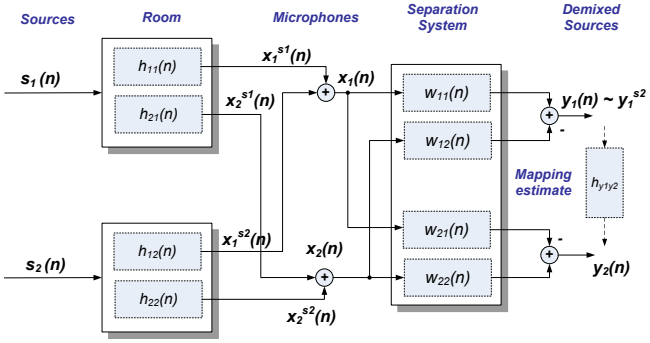

Figure 2: Full 2x2 separation system. For better understanding the upper part is assumed to have converged.

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 ($h^{\text{Roth}}_{y_1y_2}$):

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

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

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

$$\begin{aligned} y_2 &= w^{\text{opt}}_{22}\star x^{s_2}_2 - w^{\text{opt}}_{21}\star x^{s_2}_1 \stackrel{!}{=} 0 && (14)\\ &= w^{\text{opt}}_{22}\star h_{22}\star s_2 - w^{\text{opt}}_{21}\star h_{12}\star s_2 && (15)\\ &= \left(w^{\text{opt}}_{22}\star h_{22} - w^{\text{opt}}_{21}\star h_{12}\right)\star s_2 && (16)\end{aligned}$$

By comparing the terms in Eq. (16) with the ones in (14), as optimal solution $w^{\text{opt}}_{21} = w_{21}+w_{11}\star h^{\text{Roth}}_{y_1y_2}$ and $w^{\text{opt}}_{22} = w_{22}+w_{12}\star h^{\text{Roth}}_{y_1y_2}$ 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):

$$\begin{aligned} w^i_{21} &= w^{i-1}_{21} + \mu_3\cdot w^{i-1}_{11}\star h^{\text{Roth},i-1}_{y_1y_2} && (17)\\ w^i_{22} &= w^{i-1}_{22} + \mu_3\cdot w^{i-1}_{12}\star h^{\text{Roth},i-1}_{y_1y_2} && (18)\\ w^i_{11} &= w^{i-1}_{11} + \mu_3\cdot w^{i-1}_{21}\star h^{\text{Roth},i-1}_{y_2y_1} && (19)\\ w^i_{12} &= w^{i-1}_{12} + \mu_3\cdot w^{i-1}_{22}\star h^{\text{Roth},i-1}_{y_2y_1} && (20)\end{aligned}$$

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 (17)-(20) is that the GCC $h^{\text{Roth}}_{y_1y_2}$ 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:

$$\mathbf{h}^{\text{Roth}}_{y_1y_2} = \mathbf{r}^T_{y_2y_1}\mathbf{R}^{-1}_{y_1y_1}\,, \quad (21)$$

where $\mathbf{r}_{y_2y_1}$ is a $2\cdot L+1$ vector that holds cross correlation values, i.e. $r_{y_2y_1,i} = \text{E}\{y_2(n)y_1(n-L+i))\}$ 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 (21) with the natural gradient update rule in Buchner et al. [3] (equation 35 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 [3] can be related to a special case of our system with the Roth processor, the same reasoning as in [3] holds and the convergence analysis carries over.

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 50 from the CIPIC database [6] was chosen. Finally, spatially uncorrelated white noise was added to the mixture, such that the overall SNR is approximately 12 dB and 20 dB respectively. The GCC was estimated with FFTs of size 2048, the total demixing filter length was 100 and the number of iterative refinements 20. 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 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.

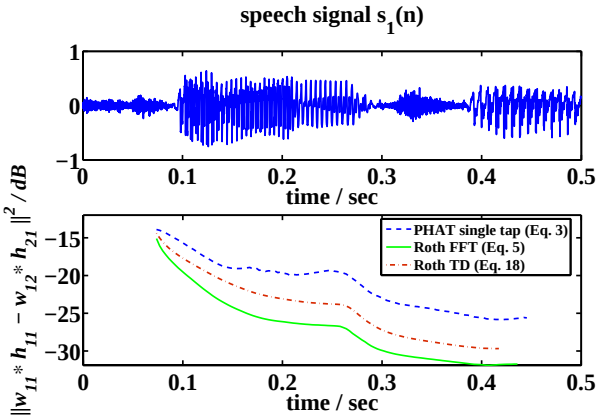

Figure 3: Typical performance of the inhibition system for one active speech signal in white noise SNR $= 12\,\text{dB}$.

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

From the plot it is again evident that the algorithm slows down and has even problems when one of the sources is highly periodic and the corresponding excitation for channel estimation is not full band. 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.

6. 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.

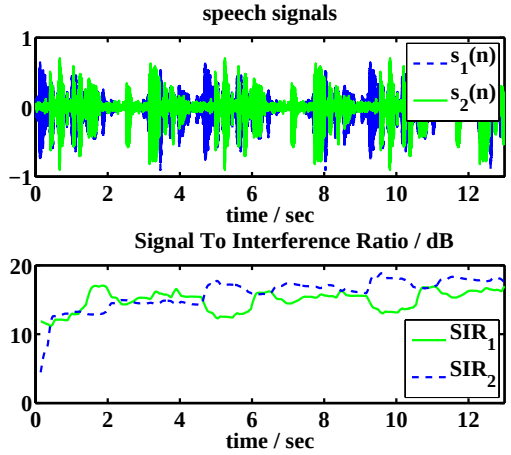

Figure 4: Typical performance of the source separation system for two active speech signals in white noise SNR $= 20\,\text{dB}$ (update via Eqs. 17 - 20)

7. REFERENCES

[1] H. Sawada, R. Mukai, S. Araki, and S. Makino, *Speech Enhancement*, chapter Frequency-Domain Blind Source Separation, pp. 299–327, Signals and Communication Technology. Springer, 2005.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] V. Algazi, R. Duda, and D. Thompson, "The cipic hrtf database," 2001.