SEMI-BLIND SOURCE SEPARATION USING HEAD-RELATED TRANSFER FUNCTIONS

Michael Syskind Pedersen,
Ulrik Kjems, Karsten Bo Rasmussen

Oticon A/S,
Strandvejen 58
DK-2900 Hellerup, Denmark
{msp,uk,kbr}@oticon.dk

Lars Kai Hansen

Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads, Building 321
DK-2800 Kongens Lyngby, Denmark
lkh@imm.dtu.dk

ABSTRACT

An online blind source separation algorithm which is a special case of the geometric algorithm by Parra and Fancourt [1] has been implemented for the purpose of separating sounds recorded at microphones placed at each side of the head. By using the assumption that the position of the two sounds are known, the source separation algorithm has been geometrically constrained. Since the separation takes place in a non free-field, a head-related transfer function (HRTF) is used to simulate the response between microphones placed at the two ears. The use of a HRTF instead of assuming free-field improves the separation with approximately 1 dB compared to when free-field is assumed. This indicates that the permutation ambiguity is solved more accurate compared to when free-field is assumed.

1. INTRODUCTION

The human auditory system is often challenged by sound environments in which several people speak simultaneously. The auditory system copes with this problem by several strategies including use of directional and binaural features, combination of visual and auditory cues, and knowledge of the speech content and context. With limited access to salient sound features multi-agent sound environments are extremely hard to navigate for the hearing impaired, hence, the separation problem is fundamental to hearing aid design.

A number of techniques have been proposed to separate mixed speech signals. Computational Auditory Scene Analysis (CASA) aims to mimic human sound processing by extracting features from the signal using processing steps inspired by the human auditory system. Both monaural and binaural cues can be invoked. Directional cues are invoked by array processing or *beamforming*. If the positions of the sound signals are known, a separation filter can be optimized so that it amplifies signals that arrive from specific directions while signals arriving from other directions may be cancelled out.

Blind source separation techniques based on assumed statistical properties of the source signals are investigated. In so-called independent component analysis (ICA) it's assumed that the source signals are statistically independent. The simplest ICA model is instantaneous mixing

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \tag{1}$$

Here $\mathbf{s}(t)$ is the vector of the source signals, $\mathbf{A}$ is the *mixing matrix* and $\mathbf{x}(t)$ is the observed signals. Usually only $\mathbf{x}(t)$ is known, so that both $\mathbf{s}(t)$ and $\mathbf{A}$ have to be found – hence the word *blind*. An emitted sound signal usually travels along different paths so that the signal arrives at different times. Hence, the model given by (1) does not comply. Instead a convolutive mixture model is used

$$x_i(t) = \sum_{j=1}^{N} \sum_{k=0}^{P-1} a_{ij}(k) s_j(t-k). \tag{2}$$

Here, the multi-path environment is described as a finite impulse response (FIR) convolutive mixture, where N is the number of source signals and P is the length of the FIR filter. A way of simplifying this problem is by mapping the convolutive mixture into the frequency domain

$$\mathbf{X}(\omega, t) = \mathbf{A}(\omega)\mathbf{S}(\omega, t). \tag{3}$$

Convolutive blind source separation has been investigated by many. Often algorithms are based on (assumed) knowledge of the probability distribution functions of the source signals. These methods are refereed to as maximum likelihood (ML) techniques. That approach has been used by e.g. Torkkola [2], Lee et al. [3], Amari et al. [4], Attias and Schreiner [5] and Douglas and Sun [6]. Higher order statistics can as well be used for separation e.g. Comon et al. [7]. By using second order statistics and additional information on speech signals, Parra et al. have developed efficient algorithms for speech separation [8], [9]. Further, Parra has combined blind source separation with beamforming. This concept is known as *geometric source separation* [10]. This algorithm has been investigated in the free-field, in which it works quite well. We expect that the Parra approach can be further adapted by using more specific knowledge of the environment. In particular, in the context of hearing aids it is of interest to investigate the non free-field situation created by the human head. Hence, we investigate a simple model of an environment consisting of, a head placed in between two microphones. Further, we have chosen to invoke the head-related transfer function, for the online-algorithm proposed by Parra and Fancourt [1], and we demonstrate that the separation indeed can be improved by a more realistic geometry.

2. ONLINE GRADIENT DESCENT ALGORITHM

Consider the convolutive case given by (3). The goal is to separate M recorded signals into an estimate of the N source signals. Here we consider only the situation $N = M = 2$. Instead of finding the elements of the mixing matrix $\mathbf{A}(\omega)$, an *unmixing matrix* $\mathbf{W}(\omega)$ will be estimated such that

$$\mathbf{Y}(\omega,t)=\mathbf{W}(\omega)\mathbf{X}(\omega,t), \tag{4}$$

where $\mathbf{Y}(\omega,t)$ is an estimate of the original unmixed source signals in the frequency domain.

2.1. Separation by Second Order Statistics

Assuming that the source signals are independent, the correlation between the output signals $y_i(t)$ and $y_j(t)$ is zero. This criterion is necessary but not sufficient for separation, because the correlation criterion only yields as many decorrelation conditions as there are pairs of sources, i.e. $N(N-1)/2$, less than half the constraints needed to determine the $N\times N$ elements of the separation matrix $\mathbf{W}(\omega)$. As pointed out in [8], the additional use of the property that speech signals are non-stationary signals yields more conditions, hence, more information for estimating the separation matrix $\mathbf{W}(\omega)$.
The source separation algorithm is a gradient descent algorithm based on minimizing a cost function given by the *coherence function* [11]

$$C_{Y_iY_j}(\omega,t)=\frac{S_{Y_iY_j}(\omega,t)}{\sqrt{S_{Y_iY_i}(\omega,t)S_{Y_jY_j}(\omega,t)}} \tag{5}$$

Here $S_{Y_iY_j}(\omega,t)$ is the cross-power density spectrum of the outputs – the Fourier transform of the cross-correlation $R_{yy}(\tau,t)=E[\mathbf{y}(t)\mathbf{y}(t+\tau)^T]$. In matrix form, the cost function is given as

$$\begin{aligned} J &= \sum_t \|\mathbf{C}_{YY}(\omega,t)\|^2 && (6)\\ &= \sum_t \mathrm{tr}(\mathbf{C}^H_{YY}(\omega,t)\mathbf{C}_{YY}(\omega,t)) && (7)\end{aligned}$$

This can be rewritten to

$$J = \sum_t \mathrm{tr}(\Lambda^{-1}_{YY}S_{YY}\Lambda^{-1}_{YY}S_{YY}), \tag{8}$$

where $\Lambda_{YY}(\omega,t)$ is the diagonal matrix of $S_{YY}(\omega,t)$. The complex derivative of (8) is found with respect to $\mathbf{W}(\omega)$ [1]. This gradient update yields

$$\Delta\mathbf{W}=-4\eta(\Lambda^{-1}_{YY}S_{YY}\Lambda^{-1}_{YY}-\mathrm{diag}(\Lambda^{-2}_{YY}S_{YY}\Lambda^{-1}_{YY}S_{YY}))S_{YX}, \tag{9}$$

where η is the learning rate and S_{YX} is the cross-power spectral density between the outputs and the inputs. We aim at an on-line algorithm, hence the power density spectra S_{YY} and S_{YX} are updated as follows

$$\mathbf{S}_{YY}(\omega,t)=\gamma\mathbf{S}_{YY}(\omega,t-T)+(1-\gamma)\mathbf{Y}(\omega,t)\mathbf{Y}^H(\omega,t) \tag{10}$$

$$\mathbf{S}_{YX}(\omega,t)=\gamma\mathbf{S}_{YX}(\omega,t-T)+(1-\gamma)\mathbf{Y}(\omega,t)\mathbf{X}^H(\omega,t). \tag{11}$$

Here, γ is the forgetting factor. To ensure stability, the forgetting factor is constrained to $0<\gamma<1$. The algorithm then consists of the four equations (4), (9), (10) and (11), finally, we use the overlap-save method [12] in the implementation.

2.2. The Geometry Constraint

The derived algorithm only solves the problem up to a permutation ambiguity. One does not know which of the output channels that contains the desired signal. Further, the permutation matrix may not be the same at all frequencies. The permutation ambiguity can be overcome by adding an additional penalty term to the cost function. The particular geometry that we will invoke can be viewed as invoking a weak audio-visual cue, hence, the title 'semi-blind source separation'. We will assume that the hearing impaired and/or the hearing aid have access to directional information, e.g., directions for the preferred speaker and the most important distractor, hence, we want to separate signals using these known directions. The constraint is constructed so that it yields unit response in the direction of the desired signal and zero response in the direction of the interfering signals [10] for a specific output channel. This can be expressed by the following form

$$\mathbf{W}(\omega)\mathbf{D}(\omega,\mathbf{Q})=\mathbf{I}. \tag{12}$$

Here $\mathbf{D}(\omega,\mathbf{Q})$ is a sensor response matrix, where each of the columns, $\mathbf{d}_i(\omega,\mathbf{q})$ consists of the response between the receivers with relation to the ith source signal. $\mathbf{Q}$ is a matrix, where each column vector $\mathbf{q}_i$ contains information of the position of the ith source signal. The cost function from this constraint will be of the form,

$$J_C=\|\mathbf{W}(\omega)\mathbf{D}(\omega,\mathbf{Q})-\mathbf{I}\|^2. \tag{13}$$

The gradient update from this cost function is given by the complex derivative of (13) with respect to $\mathbf{W}$. This yields

$$\frac{\partial J_C(\mathbf{W})}{\partial\mathbf{W}}=2((\mathbf{W}(\omega)\mathbf{D}(\omega,\mathbf{Q})-\mathbf{I})\mathbf{D}^H(\omega,\mathbf{Q})). \tag{14}$$

The gradient update from the penalty term is then given by

$$\Delta\mathbf{W}_C=-2\lambda(\omega)\eta(\mathbf{W}(\omega)\mathbf{D}(\omega,\mathbf{Q})-\mathbf{I})\mathbf{D}^H(\omega,\mathbf{Q})), \tag{15}$$

where η is the learning rate. Besides the learning rate, the update step is in addition weighted by the weight term $\lambda(\omega)$. At some frequencies, the sensor response matrix $\mathbf{D}(\omega,\mathbf{Q})$ may be singular. When the sensor response matrix is close to singular, the additional penalty term should not be included. The weight term $\lambda(\omega)$ is thus found as the inverse of the condition number

$$\lambda(\omega)=(\mathrm{cond}(\mathbf{D}(\omega,\mathbf{Q})))^{-1} \tag{16}$$

If $\mathbf{D}(\omega,\mathbf{Q})$ is close to singular, the weight $\lambda(\omega)$ will be close to zero. Otherwise, $\lambda(\omega)$ can at most be one.

3. NON FREE-FIELD SENSOR RESPONSE

In Parra and Alvino [10] the microphones are placed in a linear array in the free-field and the source signals are assumed to be in the far-field. Hereby, the sensor response matrix only depends on the frequency and the arriving angles of the incoming signals. Since free-field is assumed, the magnitude of the sensor response will be equal to one. The only difference between the received signals at the microphones will be a phase difference. Such a sensor response will be of the form

$$\mathbf{d}(\omega,\theta)=e^{-j\frac{d}{c}\omega\sin(\theta)}, \tag{17}$$

where c is the speed of sound, d is the distance between the head and θ is the arrival angle.

If the microphones are placed at each side of a head, i.e one at each ear, there might be a difference in the magnitude response as well as the phase response of the signals received at the two microphones. This is because the head attenuates the sound signal, when it has to pass around the head. Therefore, when the free-field assumption is assumed to solve the permutation ambiguity, some permutations may be incorrect. In order to estimate the sensor response difference between the two microphones, the head is modelled as a sphere. In Duda and Martens [13] a HRTF-model based on a sphere model has been proposed. An elevation angle has further been included in this model by Brungart and Rabinowitz [14]. In [13] and [14], the ratio between the sound pressure that would be in the center of the sphere in the free-field and the sound pressure that actually is developed at the surface of the sphere is found. Here, the relation between two points placed at the surface of the sphere is needed. The two points are placed at the left and the right side of the sphere, respectively. The head-related transfer function from the left side to the right side of the sphere is given by the following equation derived from the results in Duda and Martens [13].

$$H(\omega, a, r, \theta, \varphi) = \frac{\sum_{m=0}^{\infty}(2m+1)P_m(\cos(\alpha))l_m(\omega, a, r)}{\sum_{m=0}^{\infty}(2m+1)P_m(\cos(\beta))l_m(\omega, a, r)}, \tag{18}$$

where a is the sphere radius, r is the distance between the center of the sphere and the sound source and $P_m(\cdots)$ is the Legendre polynomial of degree m. $l_m(\omega, r, a) = \frac{h_m(r\omega/c)}{h'_m(a\omega/c)}$, where $h_m(\cdots)$ is the spherical mth order Hankel function, $h'_m(\cdots)$ is the first order derivative of the spherical mth order Hankel function and c is the sound velocity. Further, $\alpha = \arccos(\sin(\theta + \pi)\cos(\varphi))$ and $\beta = \arccos(\sin(\theta)\cos(\varphi))$, where θ is the azimuth angle and φ is the elevation angle. These are defined as in Fig. 1. At the bottom of the figure, the magnitude response of the HRTF is plotted for an incoming signal originating from a distance of $r = 170$ cm, an azimuth angle of $\theta = 270°$ (i.e. a signal originating from the direction of the front of the right ear) and an elevation angle of $\varphi = 33°$. The sphere radius is $a = 8.5$ cm. As it can be seen, the response differs significantly from the free-field. Especially at higher frequencies, the head becomes significant.
By using this HRTF, the sensor response matrix in the case of two receivers and two source signals is given as

$$\mathbf{D}(\omega, a, \mathbf{r}, \theta, \varphi) = \begin{bmatrix} H_{\text{denom}}(\omega, a, r_1, \theta_1, \varphi_1) & H_{\text{denom}}(\omega, a, r_2, \theta_2, \varphi_2) \\ H_{\text{num}}(\omega, a, r_1, \theta_1, \varphi_1) & H_{\text{num}}(\omega, a, r_2, \theta_2, \varphi_2) \end{bmatrix}, \tag{19}$$

where $H_{\text{denom}}(\omega, a, r, \theta, \varphi)$ is the denominator and $H_{\text{num}}(\omega, a, r, \theta, \varphi)$ is the numerator of (18), respectively. The two indices $_1$ and $_2$ indicate the two incoming signals.

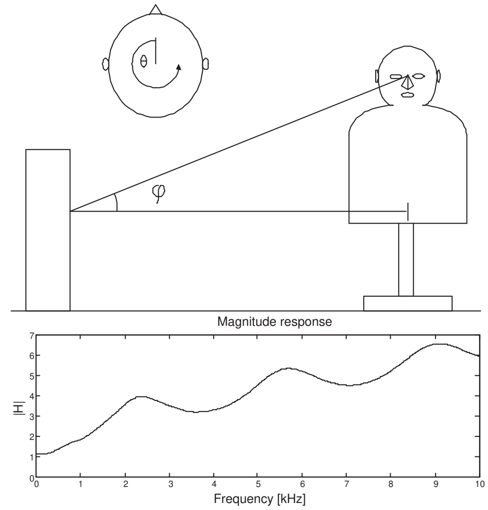

Fig. 1. A sound signal arrives at a dummy head from a distance r with an azimuth angle θ and an elevation angle φ. An azimuth angle of $\theta = 0°$ corresponds to an incoming signal arriving from the front of the head. Below, the magnitude response of the head-related transfer function (18) is shown. The HRTF is calculated for an incoming signal originating from a distance of $r = 170$ cm, an azimuth angle of $\theta = 270°$ and an elevation angle of $\varphi = 33°$. At low frequencies the magnitude response is close to one while the head attenuates the sound signal much more at higher frequencies.

4. EVALUATION

Sound signals have been recorded from 8 different positions. As shown in Fig. 1, the sounds are recorded at a dummy head and torso (B & K *Head and Torso Simulator* Type 4128). A microphone is placed at each ear and the sounds originate from eight loudspeakers equally distributed around the head and torso. The distance between the loudspeaker and the center of the head is 170 cm, the radius of the head is set to 8.5 cm, the elevation angle is estimated to 33°. The used FIR filter has a length of 512 taps, the forgetting factor has been set to 0.6, and the learning rate has been set to 0.003125. Two signals consisting of two different male speakers have been recorded at the microphones. Afterwards, these have been mixed. The recordings have taken place in a damped but not anechoic room, which means that the sounds do not only arrive from the predicted directions. The desired signal is emitted from the loudspeaker at the front of the head ($\theta = 0°$) while the interfering signal arrives from one of the other seven equally distributed directions. The sounds have been separated with two different cases of the sensor response matrix – the case of free-field, where there is no attenuation between the two microphones, and the case, where the sensor response matrix is given by (19). In the free-field, the distance d between the two microphones has been set to 22.1 cm, since it yields the best result. The signal-to-interference ratio has been found in the two cases. Because the separation of speech signals is in the area of interest, the SIR is weighted by an *articulation index* [15]. Hereby, some

Table 1. Frequency band importance function as given in Pavlovic [15]. Each frequency band has a width of $\frac{1}{3}$ octave. For the center frequency of each band, the importance weight is given.

CF [Hz]	160	200	250	315	400	500
Weight	0.008	0.010	0.015	0.029	0.044	0.058
CF [Hz]	630	800	1000	1250	1600	2000
Weight	0.065	0.071	0.082	0.084	0.088	0.090
CF [Hz]	2500	3150	4000	5000	6300	8000
Weight	0.087	0.084	0.077	0.053	0.036	0.019

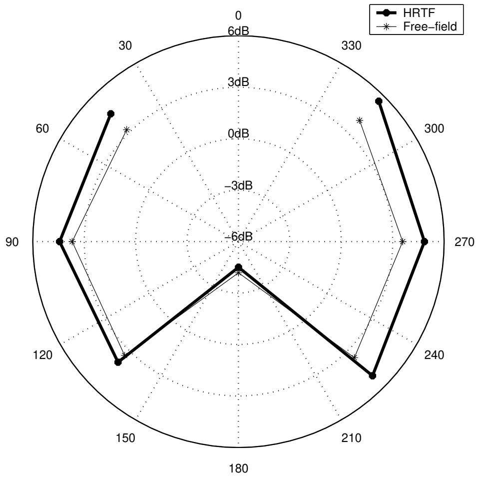

Fig. 2. Experimental results: The desired signal arrives from an angle of 0° while the interfering sound signal arrives from one of the seven angles: 45°, 90°, 135°, 180°, 225°, 270° and 315°. The thin line shows the improvement of the signal-to-interference ratio, where the free-field sensor response matrix is used. The thick line shows the improvement of the signal-to-interference ratio, where the sensor response based on the HRTF (18) has been used. As it can be seen, the signal-to-interference ratio is improved when the HRTF has been taken into account.

frequency bands are weighted more than others. These weights are shown in the table. The SIR has been plotted as function of the arrival angle of the interfering signal in Fig. 2. As it can be seen, the signal-to-interference ratio is improved when the head-related transfer function is used instead of the free-field. The improvement in SIR is about 3.5 dB when the free-field sensor response is used, while the SIR-improvement is about 5 dB when the HRTF-sensor response is used. This yields an improvement of approximately 1.5 dB, which indicates that the permutation problem is solved better when the HRTF is used instead of the free-field. It is important to notice that it is the *improvement* in SIR which has been measured. Depending on the arrival angles and the frequency, the shadowing effects of the head provides a natural separation SIR of up to approximately 6 dB. To find the SIR-improvement, these effects from the head have been subtracted from the SIR between the separated signals. As it can be seen, the algorithm doesn't work if the interfering signal arrives from behind. This is because only two microphones are used. Hence it's not possible to distinguish between signals arriving from the front and the behind. The separation of signals arriving from the right is better than the separation of the signals arriving from the left. This is probably due to the asymmetric influence from the room. The algorithm is very robust. There is no significant difference whether the mixed sound signals consist of male or female voices. Some of the separated sounds are available on-line at `http://www.imm.dtu.dk/~msp/`.

5. CONCLUSION

The geometric source separation algorithm has been extended for a non free-field case. A head-related transfer function based on a sphere model of a head has been used to estimate the response between the two microphones – one placed at each ear. An experiment shows that the separation of two speech signals is further improved, when the HRTF has been used compared to when a free-field assumption has been used. If a more accurate HRTF-model than a sphere is used, the results may be improved further.

6. REFERENCES

[1] Lucas Parra and Craig Fancourt, *An Adaptive Beamforming Perspective on Convolutive Blind Source Separation*, CRC Press LLC, 2002, Book chapter in: Noise Reduction in Speech Applications, Ed. Gillian Davis.

[2] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, September 4-6 1996, pp. 423–432.

[3] Te-Won Lee, Anthony J. Bell, and Russell H. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems*. 1997, vol. 9, pp. 758–764, The MIT Press.

[4] Shun ichi Amari, Scott C. Douglas, Andrzej Cichocki, and Howard H. Yang, "Novel on-line algorithms for blind deconvolution using natural gradient approach," in *SYSID-97*, Kitakyushu, Japan, July 8–11 1997, pp. 1057–1062.

[5] H. Attias and C. E. Schreiner, "Blind source separation and deconvolution: The dynamic component analysis algorithm," *Neural Computation*, vol. 11, pp. 803–852, 1998.

[6] Scott C. Douglas and Xiaoan Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication, Elsevier*, vol. 39, pp. 65–78, 2003.

[7] Pierre Comon, Éric Moreau, and Ludwig Rota, "Blind separation of convolutive mixtures: A contrast-based joint diagonalization approach," in *ICA2001*, San Diego, CA, December 9–13 2001.

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

[9] Lucas Parra and Clay Spence, "On-line convolutive source separation of non-stationary signals," *Journal of VLSI Signal Processing*, vol. 26, no. 1/2, pp. 39–46, August 2000.

[10] Lucas Parra and Christopher Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transaction on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, September 2002.

[11] Craig L. Fancourt and Lucas Parra, "The coherence function in blind source separation of convolutive mixtures of non-stationary signals," in *IEEE Workshop on Neural Networks for Signal Processing*, 2001, pp. 303–312.

[12] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing*, Prentice Hall, New Jersey, 3rd edition, 1996.

[13] Richard O. Duda and William L. Martens, "Range dependence of the response of a spherical head model," in *J.Acoust. Soc. Am. 104(5)*, November 1998, pp. 3048–3058.

[14] Douglas S. Brungart and William M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions.," in *J.Acoust. Soc. Am. 106(3)*, 1999, pp. 1465–1479.

[15] Chaslav V. Pavlovic, "Derivation of primary parameters and procedures for use in speech intelligibility predictions," in *J.Acoust. Soc. Am. 82(2)*, August 1987, pp. 413–422.