Underdetermined Blind Separation of Convolutive Mixtures of Speech by Combining Time-frequency Masks and ICA

Shoko Araki † *Shoji Makino* † *Audrey Blin* †‡ *Ryo Mukai* † *Hiroshi Sawada* †

† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp
‡ Université du Québec, INRS-EMT
800 de la Gauchetière Ouest, Suite 6900, Montréal, Québec, H5A 1K6, Canada

Abstract

We propose a method for separating speech signals when sources outnumber sensors. Several methods have already been proposed for solving the underdetermined blind separation problem for convolutive mixtures, and they utilize the sparseness of speech signals. Some methods employ binary masks to extract the signals, and therefore, their extracted signals contain loud musical noise. To overcome this problem, we propose utilizing both a binary mask and independent component analysis (ICA). First, using sparseness, we estimate the time points when only one source is active. Then, we remove this single source from the observations and apply ICA to the remaining mixtures. Experimental results show that our proposed method can separate signals with little distortion even in reverberant conditions of T_R=130 and 200 ms.

1. Introduction

Blind source separation (BSS) is an approach that estimates original source signals $s_i(n)$ only from observations $x_j(n)$ without source or mixing process information.

In this paper, we consider the BSS of speech signals observed in a real environment, i.e., the BSS of convolutive mixtures of speech. Recently, many methods have been proposed to solve the BSS problem of convolutive mixtures (e.g., [1]). However, most of these methods consider the determined or overdetermined case, i.e., the number of sensors is equal to or greater than the number of signals. In contrast, we focus on the underdetermined BSS problem where source signals outnumber sensors.

It is our understanding that there are two approaches to realize the underdetermined BSS. Both approaches rely on the sparseness of source signals. One is the clustering of time-frequency points with binary masks [2], and the other is based on ML estimation, where the sources are estimated after the mixing matrix estimation [3–5]. Because separation in a real environment has been tried with the former method, we have decided to watch a binary masks approach [2]. If the signals are sufficiently sparse, that is, most of the samples of a signal are almost zero, we can assume that the sources rarely overlap. [2] uses this assumption and extracts each signal using a time-frequency binary mask. However, due to these binary masks, their method results in too much discontinuous zero-padding of the extracted signals, and so the extracted signals are severely distorted.

To overcome this problem, we propose utilizing both a binary mask and independent component analysis (ICA). First, using a binary mask, we estimate the time points when only one source is active. Then, we remove this single source from the observations and apply ICA to the remaining mixtures in order to separate the signals. This single source removal does not cause severe zero-padding of the separated signals, therefore we can improve their sound quality. Experimental results show that our method can separate signals with little distortion even in real reverberant environments of T_R=130 and 200 ms.

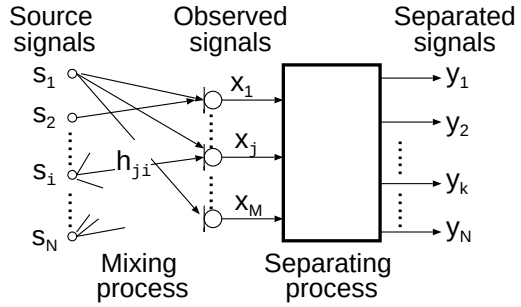


Figure 1: Block diagram of underdetermined BSS. $N > M$.

2. Problem description

In real environments, N signals observed by M sensors are modeled as convolutive mixtures $x_j(n) = \sum_{i=1}^{N} \sum_{k=1}^{P} h_{ji}(k)\, s_i(n-k+1)$ $(j = 1, \cdots, M)$, where s_i is the signal from a source i, x_j is the signal observed by a sensor j, and h_{ji} is the P-taps impulse response from a source i to a sensor j (see Fig. 1). Here, we consider the underdetermined case $N > M$. In this paper $N = 3$ and $M = 2$. Moreover, in this paper, the sources are speech signals, i.e., the sources are assumed to be mutually independent and sufficiently sparse in the time-frequency domain.

This paper employs a time-frequency domain approach because speech signals are more sparse in the time-frequency domain than in the time-domain [5] and convolutive mixture problems can be converted into instantaneous mixture problems in each frequency. In the time-frequency domain, mixtures are modeled as $\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m)$, where $\boldsymbol{H}(\omega)$ is a 2×3 mixing matrix whose j-i component is a transfer function from a source i to a sensor j, $\boldsymbol{S}(\omega, m) = [S_1(\omega, m), S_2(\omega, m), S_3(\omega, m)]^T$, $\boldsymbol{X}(\omega, m) = [X_1(\omega, m), X_2(\omega, m)]^T$ and $\boldsymbol{Y}(\omega, m) = [Y_1(\omega, m),$

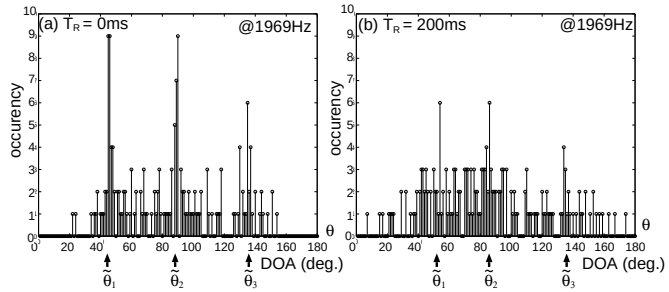


Figure 2: Example histogram. (a) T_R = 0 ms and (b) T_R = 200 ms. A male-male-female combination with DFTsize T = 512.

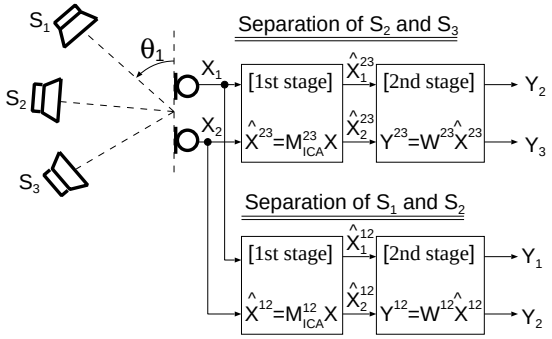


Figure 3: System setup

$Y_2(\omega,m)$,$Y_3(\omega,m)]^T$ show Fourier transformed source, observed and separated signals, respectively. ω is the frequency and m is the frame index.

Our objective is to estimate separated signals $\boldsymbol{Y}(\omega,m)$ using only the information provided by observations $\boldsymbol{X}(\omega,m)$.

3. Conventional methods: with binary masks only

Standard ICA cannot be applied to underdetermined cases because it assumes that a mixing matrix is invertible. Several methods have been proposed (e.g., [2–5]) for solving an underdetermined BSS problem, and they utilized source sparseness.

If most of the samples of a signal are almost zero, we say that this signal is sparse. When signals are sufficiently sparse, we can assume that the sources overlap at rare intervals. For a detailed analysis of sparseness, see [6].

Some conventional methods use the sparseness assumption and extract each signal using time-frequency binary masks. Because we can assume that sources do not overlap very often, we can extract each source by selecting the time points at which there is only one signal. One way of estimating such time points is to use the level difference of the observations and the phase difference between the observations. In this paper, we utilize omnidirectional microphones, therefore we use the phase difference $\varphi(\omega,m)=\angle\frac{X_1(\omega,m)}{X_2(\omega,m)}$ between the observations.

Using $\varphi(\omega,m)$, we estimate the direction of arrival (DOA) for each time point m by calculating $\theta(\omega,m)=\cos^{-1}\frac{\varphi(\omega,m)c}{\omega d}$, where c is the speed of sound and d is the microphone spacing, and make a histogram of the DOA $\theta(\omega,m)$. Each peak corresponds to each source in the histogram for each frequency. Let these peaks be $\tilde{\theta}_1,\tilde{\theta}_2$ and $\tilde{\theta}_3$ where $\tilde{\theta}_1\le\tilde{\theta}_2\le\tilde{\theta}_3$ (Fig. 2), and the signal from $\tilde{\theta}_\xi$ be $\tilde{S}_\xi$ ($\xi=1,2,3$).

We can extract each signal with a binary mask

$$M_\xi(\omega,m)=\begin{cases}1 & \tilde{\theta}_\xi-\Delta\le\theta(\omega,m)\le\tilde{\theta}_\xi+\Delta\\ 0 & \text{otherwise}\end{cases}\quad(1)$$

by calculating $Y_\xi(\omega,m)=M_\xi(\omega,m)X_j(\omega,m)$ where j=1 or 2. Here, Δ is an extraction range parameter: if Δ is small the separation performance is good but the distortion is large, in contrast, if Δ is large the musical noise problem is reduced but the separation performance deteriorates.

Although we can extract each signal using this binary mask (1), such extracted signals are discontinuously zero-padded by the binary masks, and therefore, we hear musical noise in the extracted output.

4. Proposed Method: Combination of binary mask and ICA

To overcome this musical noise problem, we propose using both a binary mask and ICA. Our method has two stages (Fig. 3). In the first stage, unlike the conventional approach, we *remove* one source from mixtures using the signals' sparseness. By this removal, it is expected that their zero-padding of the extracted signals to be less trouble because we extract more time-frequency points than the conventional approach. Moreover, because we can expect the remaining mixtures to consist of only two signals, we can apply a standard ICA to these remaining mixtures in the second stage. Because these separated signals are not highly zero-padded, we can expect less musical noise.

[1st stage] One source removal:

Instead of extracting each source as in conventional approaches, we remove only one source from the mixtures with a binary mask

$$M^{pq}_{\text{ICA}}(\omega,m)=\begin{cases}1 & \theta_{min}\le\theta(\omega,m)\le\theta_{max}\\ 0 & \text{otherwise}\end{cases}\quad(2)$$

by calculating

$$\hat{\boldsymbol{X}}^{pq}(\omega,m)=M^{pq}_{\text{ICA}}(\omega,m)\boldsymbol{X}(\omega,m).\quad(3)$$

In (2), θ_{min} and θ_{max} are extraction range parameters, and in (3), $\hat{\boldsymbol{X}}^{pq}(\omega,m)=[\hat{X}^{pq}_1(\omega,m),\hat{X}^{pq}_2(\omega,m)]$ are expected to be mixtures of $\tilde{S}_p$ and $\tilde{S}_q$. For instance, if $\tilde{S}_1$ can be removed from the observations with a mask M^{23}_{ICA} we can use ICA to separate $\tilde{S}_2$ and $\tilde{S}_3$ in the next stage. In this case θ_{min} and θ_{max} in (2) can be $\tilde{\theta}_1<\theta_{th1}=\theta_{min}<\tilde{\theta}_2,\theta_{max}=180^\circ$ (see Fig. 4), where $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are estimated in the same way as in the previous section. Similarly, when $\tilde{S}_3$ is to be removed from the observations with a mask M^{12}_{ICA}, $\theta_{min}=0^\circ,\tilde{\theta}_2<\theta_{th2}=\theta_{max}<\tilde{\theta}_3$.

Because our system has only two outputs, both removals should be performed to obtain three separated signals (see Fig. 3).

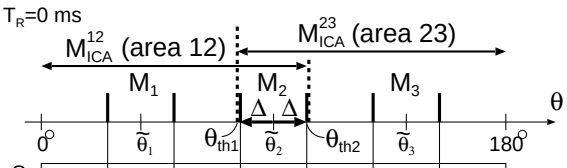
[2nd stage] Separation of remaining sources by ICA:

Because the remaining signals $\hat{\boldsymbol{X}}^{pq}$ are expected to be mixtures of two signals, we can use 2×2 ICA to separate $\hat{\boldsymbol{X}}^{pq}$. The separation process can be formulated as

$$\boldsymbol{Y}^{pq}(\omega,m)=\boldsymbol{W}^{pq}(\omega)\hat{\boldsymbol{X}}^{pq}(\omega,m),\quad(4)$$

where $\hat{\boldsymbol{X}}^{pq}$ is the masked observed signal obtained by (3), $\boldsymbol{Y}^{pq}(\omega,m)=[Y_p(\omega,m),Y_q(\omega,m)]^T$ is the separated output signal, and $\boldsymbol{W}^{pq}(\omega)$ represents a (2×2) separation matrix. $\boldsymbol{W}^{pq}(\omega)$ is determined so that $Y_p(\omega,m)$ and $Y_q(\omega,m)$ become mutually independent.

(a) T_R=0 ms



S_1	91	99	58	5	8	1	7
S_2	7	1	37	91	32	2	2
S_3	2	0	5	4	60	97	91
area 23	removed			S_1: 3	S_2: 47	S_3: 50	
area 12	S_1: 50	S_2: 48	S_3: 2	removed			[%]

(b) T_R=200 ms

S_1	87	86	66	13	13	3	5
S_2	4	5	24	80	43	11	6
S_3	9	9	10	7	44	86	89
area 23	removed			S_1: 10	S_2: 58	S_3: 32	
area 12	S_1: 36	S_2: 56	S_3: 8	removed			[%]

Figure 4: Source power in each area (%). A male-male-female combination.

5. Experiments

5.1. Experimental conditions

In our experiments, we utilized the set-up shown in Fig. 5. For tests of T_R = 0 ms, we simulated a recording using the mixing matrix $H_{ji}(\omega) = \exp(j\omega\tau_{ji})$, where $\tau_{ji} = \frac{d_j}{c}\sin\theta_i$, d_j is the position of the j-th microphone, and θ_i is the direction of the i-th source.

For the reverberant tests, we used speech data convolved with impulse responses recorded in a real room whose reverberation time was T_R = 130 and 200 ms.

As the original speech, we used three Japanese sentences spoken by three male and three female speakers. We investigated three combinations of speakers: male-male-female, male-male-male, and female-female-female.

The DFT frame size T was 512 and the frame shift was 256 at a sampling rate of 8 kHz. The Δ value in the conventional method's binary masks (1) was 15° in DOA. We used $\theta_{th1} = \tilde{\theta}_2 - \Delta$ for M^{23}_{ICA} (area 23), and $\theta_{th2} = \tilde{\theta}_2 + \Delta$ for M^{12}_{ICA} (area 12), where Δ was also 15°. With this Δ value the conventional method and our proposed method provided compatible signal to noise ratio (SIR).

The adaptation rule of ICA utilized in our experiments was $\boldsymbol{W}_{i+1}(\omega) = \boldsymbol{W}_i(\omega) + \eta\left[\text{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\text{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\text{H}}\rangle\right] \cdot \boldsymbol{W}_i(\omega)$, where $\Phi(\boldsymbol{y}) = \phi(|\boldsymbol{y}|)\cdot e^{j\cdot\angle(\boldsymbol{y})}$, $\phi(x) = \tanh(gx)$ and $g = 100$ [7]. To solve the permutation problem of frequency domain ICA, we employed the DOA and correlation approach [8], and for solving the scaling problem of frequency domain ICA, we used the minimum distortion principle [9].

5.2. Performance measures

We used the signal to interference ratio (SIR) as a measure of separation performance, and the signal to distortion ratio (SDR) as a measure of sound quality:

$$\text{SIR}_i = 10\log\frac{\sum_n y^2_{is_i}(n)}{\sum_n(\sum_{i\neq j} y_{is_j}(n))^2} \tag{5}$$

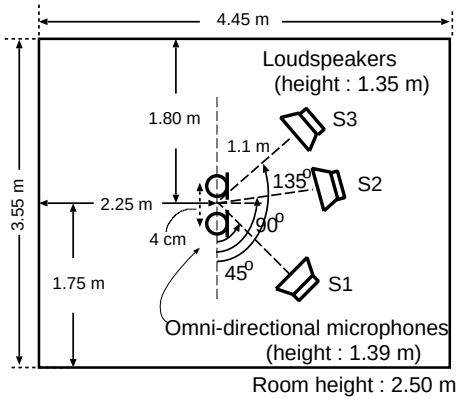


Figure 5: Room for reverberant tests.

$$\text{SDR}_i = 10\log\frac{\sum_n x^2_{ks_i}(n)}{\sum_n(x_{ks_i}(n) - \alpha y_{is_i}(n-D))^2}. \tag{6}$$

The permutation is solved before calculating SIR and SDR, i.e., y_i is the estimation of s_i, and y_{is_j} is the output of the whole separating system at y_i when only s_j is active, and x_{ks_j} is the observation obtained by microphone k when only s_j is active. α and D are parameters to compensate for the amplitude and phase difference between x_{ks_i} and y_{is_i}.

5.3. Experimental results

5.3.1. Sparseness assessments

The histograms we used to design the binary masks are shown in Fig. 2. Because of the sparseness property of speech signals, the signals are well localized and the histogram has sharp peaks for T_R = 0 ms [Fig. 2 (a)]. However, in a reverberant case the peaks are not so sharp [Fig. 2 (b)]. This shows that signals overlap each other in a reverberant case and it becomes more difficult to use the sparseness assumption than when T_R = 0 ms.

We can also see this overlap in Fig. 4, which shows the power content by percentage of each signal when $\Delta = 15°$. When T_R = 0 ms [Fig. 4 (a)], S_1, S_2 and S_3 are dominant in the areas of M_1, M_2 and M_3, respectively. When T_R = 200 ms [Fig. 4 (b)], however, the other signals' overlaps increase in each area. It is difficult for the sparseness assumption to held in a reverberant case.

The degree of sparseness affects the performance of the 2nd stage of our method. Figure 4 also shows the percentage of each source power in the areas of M^{12}_{ICA} and M^{23}_{ICA}. The contributions of the *third* signal are small and two signals are dominant even when T_R=200 ms. Therefore, we can say that we can use ICA in the 2nd stage of our method.

5.3.2. Effect of one source removal

Table 1 shows the signal power eliminated by the zero-padding $\frac{\sum_n x_{1s_i}(n)^2 - \sum_n \hat{s}_i(n)^2}{\sum_n x_{1s_i}(n)^2}$ caused by binary masks, where $\hat{s}_i(n) = \text{IDFT}[M(\omega,m)X_{1s_i}(\omega,m)]$. In the sparseness only case [i.e., $M(\omega,m) = M_i(\omega,m)$], a large part of the signal power was eliminated by the binary mask. By contrast, with our proposed method, signal power eliminated by $M^{pq}_{\text{ICA}}(\omega,m)$ was inferior. This result convinces us that the adverse effect of zero-padding was mitigated by using our method.

Table 1: Power lost by binary masks (in %). (a) T_R=0 ms, (b) T_R=130 ms, (c) T_R=200 ms. A male-male-female combination,

	mask	M_1	M_2	M_3	M_{ICA}^{12}		M_{ICA}^{23}	
(a)	output	Y_1	Y_2	Y_3	Y_1	Y_2	Y_2	Y_3
	[%]	15	8.7	15	1.4	3.9	4.9	0.9
	mask	M_1	M_2	M_3	M_{ICA}^{12}		M_{ICA}^{23}	
(b)	output	Y_1	Y_2	Y_3	Y_1	Y_2	Y_2	Y_3
	[%]	39	6.0	28	1.6	2.5	3.6	4.4
	mask	M_1	M_2	M_3	M_{ICA}^{12}		M_{ICA}^{23}	
(c)	output	Y_1	Y_2	Y_3	Y_1	Y_2	Y_2	Y_3
	[%]	48	9.2	37	4.1	4.5	4.7	8.8

5.3.3. Separation results

Table 2 shows the experimental results we obtained for T_R = 0 ms. The first row shows the results obtained solely using binary masks, and the second and third rows show the results obtained with our proposed method. With only a binary mask, the SDR values were unsatisfactory, and a large musical noise was heard. In contrast, with our proposed method, we were able to obtain high SDR values without destroying the separation performance SIR and the musical noise was reduced.

Moreover, Tables 3 and 4 show the results of reverberant tests when T_R = 130 and T_R = 200 ms, respectively. In reverberant cases, due to the decline of sparseness, the performance with both methods was worse than when T_R = 0 ms. However, we were able to obtain higher SDR values with our proposed method than with the conventional method even in reverberant environments. Some sound samples can be found at our web site [10].

6. Conclusion

We proposed utilizing both a binary mask and ICA for BSS when speech signals outnumber sensors. Our method avoids excessive zero-padding, and therefore, can separate the signals with little distortion in reverberant environments of T_R=130 and 200 ms.

7. References

[1] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] S. Rickard and O. Yilmaz, "On the W-Disjoint orthogonality of speech," *Proc. ICASSP2002*, vol.1, pp. 529-532, 2002.

[3] F. J. Theis, C. G. Puntonet, E. W. Lang, "A histogram-based overcomplete ICA algorithm," *4th International Symposium on ICA and BSS (ICA2003)*, pp. 1071-1076, 2003.

[4] L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda and J. C. Principe, "Underdetermined blind source separation in a time-varying environment," *Proc. ICASSP2002*, pp. 3049-3052, 2002.

[5] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," *2nd International Workshop on ICA and BSS (ICA2000)*, pp. 87-92, 2000.

[6] A. Blin, S. Araki and S. Makino, "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix combination," *Proc. IWAENC2003*, pp. 211-214, 2003.

[7] H. Sawada, R. Mukai, S. Araki and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *Proc. ICASSP2002*, pp. 1001-1004, 2002.

[8] H. Sawada, R. Mukai, S. Araki and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *4th International Symposium on ICA and BSS (ICA2003)*, pp. 505-510, 2003.

[9] K. Matsuoka and S. Nakashima, "A robust algorithm for blind separation of convolutive mixture of sources," *4th International Symposium on ICA and BSS (ICA2003)*, pp. 927-932, 2003.

[10] http://www.kecl.ntt.co.jp/icl/signal/araki/underdeterminedBSSdemo.html

Table 2: Results of T_R=0 ms simulations. 'Conv.': with conventional method, 'area 12' and 'area 23': with our method. Δ=15°.

male-male-female						[dB]
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	18.4	11.6	17.3	9.0	11.5	9.4
area 12	15.0	8.4	-	18.7	14.4	-
area 23	-	7.9	14.6	-	13.9	18.5
male-male-male						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	14.4	5.5	17.4	5.5	9.2	6.0
area 12	9.2	2.5	-	16.8	14.7	-
area 23	-	3.7	11.3	-	10.5	12.4
female-female-female						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	21.3	9.7	20.5	9.3	13.9	9.6
area 12	13.6	6.8	-	18.9	16.6	-
area 23	-	6.8	13.2	-	16.3	21.3

Table 3: Results of reverberant tests. T_R=130 ms.

male-male-female						[dB]
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	10.4	6.8	10.7	4.8	13.3	6.2
area 12	9.5	6.1	-	8.8	15.9	-
area 23	-	6.4	8.9	-	14.4	9.5
male-male-male						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	10.6	2.9	8.6	3.7	12.5	4.4
area 12	7.7	2.5	-	5.9	14.3	-
area 23	-	2.2	8.9	-	12.7	7.9
female-female-female						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	16.0	9.2	13.7	6.4	16.0	6.9
area 12	12.2	7.9	-	8.6	17.6	-
area 23	-	7.9	9.7	-	16.5	10.6

Table 4: Results of reverberant tests. T_R=200 ms.

male-male-female						[dB]
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	8.3	5.9	8.6	3.4	11.3	4.9
area 12	8.1	5.4	-	6.0	12.6	-
area 23	-	5.5	7.5	-	13.6	6.5
male-male-male						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	7.7	1.9	5.4	1.1	10.6	4.0
area 12	6.6	2.3	-	4.2	11.6	-
area 23	-	1.7	2.4	-	10.5	5.4
female-female-female						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Conv.	10.2	6.0	8.2	3.3	13.5	4.8
area 12	9.6	5.5	-	5.8	14.0	-
area 23	-	5.6	9.0	-	13.6	7.1