International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:2, No:11, 2008

An ICA Algorithm for Separation of Convolutive Mixture of Speech Signals

Rajkishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano

***Abstract*—**This paper describes Independent Component Analysis (ICA) based fixed-point algorithm for the blind separation of the convolutive mixture of speech, picked-up by a linear microphone array. The proposed algorithm extracts independent sources by non-Gaussianizing the Time-Frequency Series of Speech (TFSS) in a deflationary way. The degree of non-Gaussianization is measured by negentropy. The relative performances of algorithm under random initialization and Null beamformer (NBF) based initialization are studied. It has been found that an NBF based initial value gives speedy convergence as well as better separation performance

***Keywords*—** Blind signal separation, independent component analysis, negentropy, convolutive mixture.

I. Introduction

THE goal of Blind Signal Separation (BSS) is to estimate latent sources from their mixed observations without any knowledge of mixing process. This challenging problem has bagged much research attention due to very wide area of applicability such as in speech signal separation, image processing, computer vision, bioinformatics, cosmo-informatics etc. [1]- [3]. In the area of speech signal processing BSS can be supposed as an engineering effort to imitate a very special anthropomorphic capability of focusing hearing attention to a particular speaker in the cacophony of speech signals e.g. listening in a crowd. This is well known as 'Cocktail party problem' in the scientific community [4]. A BSS algorithm can serve the same purpose for an automatic speech recognizer. Mathematically, a BSS problem can be described as the process of estimating R original sources $s(n)=[s_1(n),s_2(n).....s_R(n)]^T$ from their M observed mixed signals $x(n)=[x_1(n),x_2(n).....x_M(n)]^T$ at sensors produced by some unknown mixing function F among the R original sources given as

$$x(n)=F[s(n)], \qquad (1)$$

where n is the time index. The task of BSS is to estimate the optimal $\hat{F}^{-1}$, the inverse of the mixing function, so that the underlying original sources can be optimally estimated, i.e.

$$\hat{s}(t)=[\hat{s}_1(n),\hat{s}_2(n).....\hat{s}_M(n)]^T=\hat{F}^{-1}[x(t)]. \qquad (2)$$

Manuscript received on August 5, 2004. Authors are with Graduate School of Information Science, Nara Institute of Science and Technology 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN.E-mail: {kishor-p, sawatari, shikano}@is.aist-nara.ac.jp

In the simplest case the mixing process F produces instantaneous mixture; however, in this paper we will consider the case of convolutive mixing. The complete lack of knowledge about mixing process makes BSS problem challenging and work is further carried out by bringing into focus the principle of statistical independence of hidden sources. However, due to unknown mixing process observed signals even with spatial distinction are not independent. Thus under the assumption of statistical assumption the task in the BSS is to obtain Independent Components (IC) from the mixed signals and such algorithms are called ICA based BSS algorithms [1]. The independent components are extracted either as maximally non-Gaussian components or looking spectral dissimilarity among the sources [5]. The application of the BSS technique in audio signal separation can be traced back to the work in [6],[7] on the ICA based signal separation algorithms for practical applications. In contrast to the other source separation techniques, such as the organization of hierarchical perceptual sounds [8], formant tracking [9], auditory scene analysis [10] used with single channel processing, delay and sum beamforming, adaptive beam forming (ABF) [11]-[13], and NBF used with multichannel or array signal processing [14], BSS is the unsupervised adaptive filtering for the array processing based on information geometry theory [15],[16].

For the blind separation of convolutive mixture of speech, it was first proposed in [17] that in the frequency domain convoluted mixture is converted into instantaneous mixture in different frequency sub-bands or bins which simplify the demixing process. Recently, many ICA based BSS algorithms have been developed, either separately in the time domain or in the frequency domain or mutualistically combined in both while weighing their pros and cons, for audio source separation [18]-[20]. However, still there hardly exist algorithms for the real world application because separation performance degrades in real acoustic environment with unacceptable computational time [21]. Real-world application requires faster methods to perform on-line separation. To date, the algorithms developed are not sufficiently fast to satisfy real-time requirements. Frequency-domain approaches are relatively faster due to the power of FFT, yet the gradient based FDICA techniques require a larger number of iterations

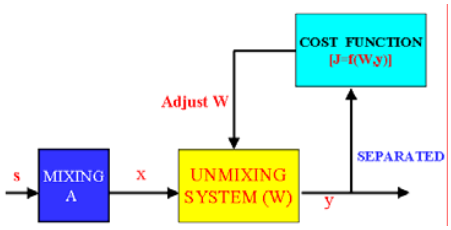


Figure1: Block diagram showing basic working principle of the ICA based BSS algorithms.

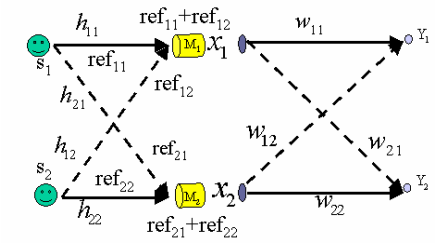


Fig.2 Convolutive mixing and demixing models for speech signal at a linear microphone array (M=R=2)

to converge [13]. The basic functioning of the ICA based BSS algorithm is shown in Fig.1.The observed mixed signals $x(n)=[x_1(n),x_2(n)...x_M(n)]^T=As(n)$ where A is the mixing system, are passed through a tentative initial demixing system W (randomly chosen or based on some heuristic guess and subject to further modification) and then the mutual independence among the estimated independent component signals y is evaluated by some cost function $J(W,y)$, usually based on the statistics of the signal and candidate demixing system. That in turn goes on modifying demixing system unless and until the cost function is not optimized for themaximum mutual independence among the separated ICs. So, paradigmatically, most of the known ICA-based BSS algorithms exhibit such functional similarities, but basic differences occur in the choice of the cost function, the domain of operation and the process of optimization. The mixing process increases Gaussianity of the signal, in the light of Central Limit Theorem (CLT), the non-Gaussianization can yield independent components. Here we will look for the independent components as the most non-Gaussian components and thus our cost function will be based on non-Gaussianity measure as proposed in [22]. In order to solve permutation and scaling problem we will use null beamformer based technique [14].

II. Signal Mixing and Demixing Model

In the real recording environment, signals picked-up by a microphone consist of direct-path signals as well as their delayed (reflected) and attenuated versions and noise signals. Therefore, the speech signal picked up by an M element linear microphone array is modeled as a linear convolutive mixture of R impinging source signals s_i (we exclude here noise signal for the simplicity) such that the M-dimensional observed signal $x(n)=[x_1(n),x_2(n),.....x_M(n)]^T$ is given by

$$x_j(n)=\sum_{i=1}^{R}\sum_{p=1}^{P}h_{ji}(p)\,s_i(n-p+1);\quad (j=1,2,....M), \tag{3}$$

where $s_i(n)=[s_1(n),s_2(n),.....s_R(n)]^T$ represents the original source signals, h_{ji} is the P-point impulse response between the source i and the microphone j. However, in this paper we consider the case of two microphones and two sources, i.e., M=R=2, for which the signal mixing and demixing models are shown in Fig.2. Accordingly, the observed signals $x_1(n)$ and $x_2(n)$ at the microphones are given by

$$\begin{bmatrix}x_1(n)\\x_2(n)\end{bmatrix}=\begin{bmatrix}h_{11}&h_{12}\\h_{21}&h_{22}\end{bmatrix}\otimes\begin{bmatrix}s_1(n)\\s_2(n)\end{bmatrix}=\begin{bmatrix}ref_{11}+ref_{12}\\ref_{21}+ref_{22}\end{bmatrix}, \tag{4}$$

where $ref_{11}=h_{11}\otimes s_1(n)$; $ref_{12}=h_{12}\otimes s_2(n)$; $ref_{21}=h_{21}\otimes s_1(n)$; $ref_{22}=h_{22}\otimes s_2(n)$ are called reference signals and $\otimes$ represents the convolution operation.

In the frequency domain, the same model is represented by taking Short-Time Fourier Transform (STFT) of Eq.(3) and the model in Eq.(4) can be expressed as

$$\begin{bmatrix}X_1(f)\\X_2(f)\end{bmatrix}=H(f)S(f)=\begin{bmatrix}H_{11}(f)&H_{12}(f)\\H_{21}(f)&H_{22}(f)\end{bmatrix}\begin{bmatrix}S_1(f)\\S_2(f)\end{bmatrix}, \tag{5}$$

where symbols in capital denote Fourier transforms of corresponding subjects expressed by small letter symbols. The FDICA separates the signal in each frequency bin independently, and this separation process is given by

$$\begin{bmatrix}\hat{S}_1(f)\\\hat{S}_2(f)\end{bmatrix}=\begin{bmatrix}Y_1(f)\\Y_2(f)\end{bmatrix}=W(f)X(f)=\begin{bmatrix}W_{11}(f)&W_{12}(f)\\W_{21}(f)&W_{22}(f)\end{bmatrix}\begin{bmatrix}X_1(f)\\X_2(f)\end{bmatrix}, \tag{6}$$

where $[Y_1(f),Y_2(f)]^T$ are ICs ; and $W(f)=$ separation matrix in frequency bin f. It is important to note that obtained ICs are not exact replica of original sources.

III. Fixed Point FDICA

FDICA algorithm works on the TFSS of the mixed speech data to sieve out TFSS of the independent components in each frequency bin. The whole process of TFSS generation by the STFT analysis is depicted in Fig.3. It is evident that the time-frequency series consists of speech spectral components of same frequency from all analysis frames in the time succession. Fixed-point ICA was first developed and proposed in [23] for the separation of the instantaneous mixture. The key feature of this algorithm is that it converges

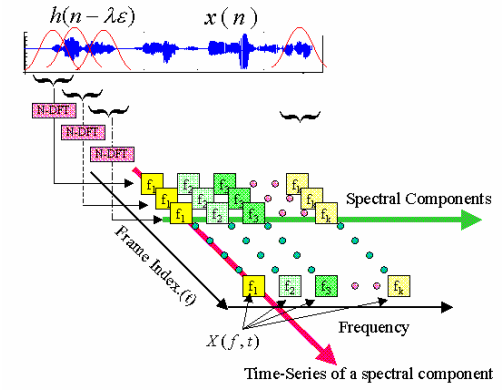


Figure 3: Process of the generation of time-frequency series of speech spectral components by STFT analysis. $h(n)$ is the Hanning window and ε is the step size of the analysis frame of size λ. Each short frame of speech is N-point DFTed and then spectral components of the same frequency bins from different analysis frames are stacked to form TFSS $X(f,t)$.

faster than other algorithms, like natural gradient-based algorithms, with almost same separation quality. However, the algorithm in [23] is not applicable to TFSS as these are complex valued. In [22], fixed-point ICA algorithm of [23]has been extended for the complex-valued signals, however, this algorithm has no strategy for solving the problem of permutation and scaling arising in FDICA for speech signal separation. In [24], Mitianoudis et al. have proposed the application of the fixed-point algorithm for speech signal separation with a time-frequency-model-based likelihood ratio jump scheme as a solution for permutation. The basic functioning of the fixed-point FDICA is shown in Fig. 4. The fixed-point ICA algorithm [23] is based on the heuristic assumption that when the non-Gaussian signals get mixed it becomes more Gaussian and thus its non-Gaussianization can yield independent components. The frequency domain mixing model for the speech signal in Eq.(5) revels that the TFSS in any frequency bin is superposition of spectral contributions of each source. Thus, in the light of CLT,TFSS of mixed speech signal in any frequency bin is more Gaussian than that of any independent source.

Obviously, non-Gaussianization of TFSS can give TFSS of independent sources from which original signals can be reconstructed. The process of non-Gaussianization consists of two-steps approaches, namely, pre-whitening or sphering and rotation of the observation vector as shown in Fig.4. Sphering is half of the ICA task and gives spatially decorrelated signals. The effect of mixing, whitening and rotation on the data is shown in the scatter plots of Fig.5. Whitening of the zero mean TFSS is done using Mahalanobis transform [25]. The whitened

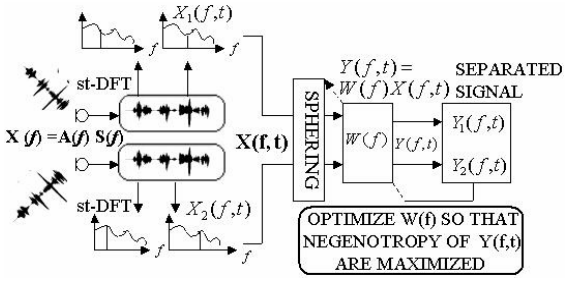


Figure 4: Functioning of the fixed-point FDICA

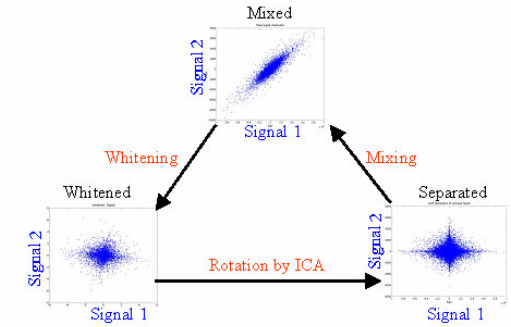


Fig.5 Scatter plots showing effect of mixing, whitening and ICA on the speech data distribution.

signal $X_W(f,t)$ in the *f*th frequency bin is obtained as

$$\mathbf{X}_W(f,t) = Q(f)\mathbf{X}(f,t), \qquad (7)$$

where $Q(f) = \Lambda_x^{-0.5} V_X$ is called whitening matrix; $\Lambda_x = diag\{1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, ..., 1/\sqrt{\lambda_n}\}$ is the diagonal matrix with positive eigenvalues $\lambda_1 > \lambda_1 > ..... > \lambda_n$ of the covariance matrix of $X(f_p,t)$ and V_x is the orthogonal matrix consisting of eigenvectors.

The task remaining after whitening involves rotating the whitened signal vector $X_W(f,t)$ by the separation matrix such that $\mathbf{Y}(f) = W(f)\mathbf{X}_w(f,t)$ equals independent components. The cost function can be based on the various measures, such as kurtosis or negentropy, for measuring the non-Gaussianity. However, negentropy provides better performance as explained in [23]. The negentropy *J(Y)* of the TFSS of the candidate IC, $\mathbf{Y}(f,t)$ is given by (frequency index *f* and frame index *t* are dropped hereafter for clarity)

$$J(\mathbf{Y}) = H(\mathbf{Y}_{gauss}) - H(\mathbf{Y}) \qquad (8)$$

where H(.) is the differential entropy of (.) and $\mathbf{Y}_{gauss}$ is the Gaussian random variable with the same covariance as of *Y*. This definition of negentropy ensures that it will be zero if $\mathbf{Y}(f,t)$ is Gaussian and will be increasing if $\mathbf{Y}(f,t)$ is tending towards non-Gaussianity. Thus negentropy based

contrast function can be maximized to obtain optimally non-Gaussian component. Here we will place derivation of such a deflationary learning rule in which one separation vector ***w*** (any one row of the separation matrix) at a time will be learned. The negentropy can be approximated in terms of non-quadratic non-linear function *G* as follows [23]:

$$J(y) = \sigma[E\{G(y) - E\{G(y_{gauss})\}]^2, \tag{9}$$

where σ is a positive constant. The performance of the fixed-point algorithm depends on the used non-quadratic non-linear function *G*. The choice of the non-linear function *G* depends on the Probability Distribution Function (PDF) of the data. Some of the non-quadratic functions used for complex-valued signal separation are

$$\begin{aligned} G_1(Y) &= \sqrt{a_1 + Y}\ ; a_1 = 0.01, \\ G_2(Y) &= \log(a_2 + Y); a_2 = 0.01, \\ G_3(Y) &= \frac{Y}{|Y|}; \forall Y \neq 0. \end{aligned} \tag{10}$$

The most general form of non-linear function that can be used for speech data (assuming TFSS has super-Gaussian distribution) is G_2. Following findings in [22], we will also use non-quadratic function G_2, hereafter denoted by *G*, whose first and 2nd-order derivatives *g* and *g'*, respectively, are given by

$$g(Y) = \frac{1}{\left(a_2 + Y^2\right)} \text{ and } g'(Y) = \frac{0.5}{\left(a_2 + Y^2\right)^2}. \tag{11}$$

The one unit algorithm for learning the separation matrix ***W(f)*** is obtained by maximizing the negentropy based contrast function. The speech signal is also modeled as a spherically symmetric variable, and as pointed out in [22], for a spherically symmetric variable, modulus-based contrast function can be used to measure non-Gaussianity. Accordingly, we use the same contrast function as in [23] given by

$$J(\mathbf{Y}) = E\{G(|\mathbf{w}^H \mathbf{X}_w|^2)\} \tag{12}$$

where ***w*** is an *M*-dimensional complex vector such that

$$E\{|\mathbf{w}^H \mathbf{X}_w|^2\} = 1 \Rightarrow |\mathbf{w}| = 1. \tag{13}$$

This contrast function may have *M* local or global optimum solutions $\mathbf{w}_i$ (*i*=1,2, …, *M*) for each source. Thus learning each ***w*** calls for the maximization of Eq.(12) under the constraint given in Eq.(13). The maxima of *J*(Y) can be found by solving the Lagrangian function $L(\mathbf{w}, \mathbf{w}^H, \lambda)$ of the above, given as

$$L(\mathbf{w}, \mathbf{w}^H, \lambda) = E\{G(|\mathbf{w}^H \mathbf{X}_w|^2)\} \pm \lambda[E\{\mathbf{w}^H \mathbf{X}_w\} - 1\}, \tag{14}$$

where λ is Lagrangian multiplier. In order to locate maxima of the contrast function, the following simultaneous equations must be solved.

$$\frac{\partial L}{\partial \mathbf{w}} = 0; \frac{\partial L}{\partial \mathbf{w}^H} = 0; \text{ and } \frac{\partial L}{\partial \lambda} = 0 \tag{15}$$

These equations can be obtained from Eq.(12) as follows

$$\frac{\partial L}{\partial \mathbf{w}} = E\{g(|\mathbf{w}^H \mathbf{X}_w|^2)\mathbf{w}^H\} + \lambda \mathbf{w}^H = 0, \tag{16}$$

$$\frac{\partial L}{\partial \mathbf{w}^H} = E\{g(|\mathbf{w}^H \mathbf{X}_w|^2)\mathbf{X}_w^H \mathbf{w}\} + \lambda \mathbf{w} = 0, \tag{17}$$

$$\frac{\partial L}{\partial \lambda} = |\mathbf{w}|^2 - 1 = 0, \tag{18}$$

From here, we proceed further in the light of following two theorems [26]:

THEOREM 1: If function f(z,z) is analytic with respect to z and z*, all stationary points can be found by setting the derivative with respect to either z or z*.*

THEOREM 2: If f(z,z) is a function of the complex-valued variable z and its conjugate, then by treating z and z* independently, the quantity directing the maximum rate of change of f(z,z*) is* $\nabla z^* f(z)$

Accordingly, the final solution using Newton's iterative method is given by

$$\mathbf{w}_{new} = \mathbf{w} - \left[\frac{\partial L}{\partial \mathbf{w}^H}\right]\left[\frac{\partial L}{\partial \mathbf{w}}\left(\frac{\partial L}{\partial \mathbf{w}^H}\right)\right]^{-1}. \tag{19}$$

$$\begin{aligned} \mathbf{w}_{new} = \mathbf{w}(E\{g(|\mathbf{w}^H \mathbf{X}_w|^2) + (|\mathbf{w}^H \mathbf{X}_w|^2)g'(|\mathbf{w}^H \mathbf{X}_w|^2)\}) \\ -E\{g(|\mathbf{w}^H \mathbf{X}_w|^2)(\mathbf{X}_w^H \mathbf{w})\mathbf{X}_w\}. \end{aligned} \tag{20}$$

The stopping criterion for iteration is defined as $\delta = (|\mathbf{w}_{old} - \mathbf{w}_{new}|)^2$, which becomes very small near the convergence. Since each update changes the norm of **w**, after each iteration **w** is normalized to maintain compliance of Eq. (13).

$$\mathbf{w}_{new} = \frac{\mathbf{w}_{new}}{|\mathbf{w}_{new}|} \tag{21}$$

As this is a deflationary algorithm, independent sources are extracted one by one in the decreasing order of negentropy from the mixed signal. Thus after each iteration, it is also essential to decorrelate **w** to prevent its convergence to the previously converged point. In order to achieve this, Gram-

Schmidt sequential orthogonalization can be used, in which components of all previously obtained separation vectors falling in the direction of the current vector are subtracted. Accordingly, the orthogonalized separation vector $\mathbf{w}_i$ for the *i*th source after *j*th iteration is given by

$$\mathbf{w}_i = \mathbf{w}_i - \sum_{j=1}^{i-1} (\mathbf{w}_i^T \mathbf{w}_j)\mathbf{w}_j. \qquad (22)$$

The update Eq.(20) is used to estimate separation vector **w** in each frequency bin from whitened TFSS of mixed signal for each source in the deflationary fashion and separation matrix W(*f*) in any frequency bin *f* is given by

$$W(f) = \begin{bmatrix} \mathbf{w}_1 \\ . \\ \mathbf{w}_R \end{bmatrix} = \begin{bmatrix} W_{11}(f) & .. & W_{1M}(f) \\ . & .. & . \\ W_{R1}(f) & .. & W_{RM}(f) \end{bmatrix} \qquad (22)$$

Each row of this separation matrix uniquely corresponds to a separation vector **w** for each source. Because this separation matrix has been obtained using whitened signals, its pre-multiplication with whitened signals in the frequency domain gives the TFSS $\mathbf{Y}(f,t) = [\mathbf{Y}_1(f,t), \mathbf{Y}_2(f,t), ... \mathbf{Y}_R(f,t)]^T$ of the separated signal, i.e.,

$$\hat{\mathbf{S}}(f,t) = \mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}_W(f,t). \qquad (23)$$

IV. Permutation and Scaling Problem

In order to get separated signal correctly, the order of separation vectors (position of rows) in W(*f*) must be same in each frequency bin. The deflationary algorithm separates original sources in the decreasing order of negentropy. But the order of negentropy for the independent sources does not remain same, due to change in contents, in all frequency bins which in turn leads to the inter-exchange or flipping of rows of ***W(f)*** in an unknown order. This is called permutation problem. The other problem is related with different gain values in each frequency bin, however, for the faithful reconstruction of the signal it should be same. This is called scaling problem. If these two problems are not solved, Eq.(23) will give another mixed signals instead of separated components. There have been developments of several methods to resolve these two inherent problems [27]. However, we will use here Directivity Pattern (DP) based method using null beamformer [28] for the reason explained in the following section. The DP based method requires the Direction of Arrival (DOA) of each source to be known. In the totally blind setup, this cannot be known so it is estimated from the directivity pattern of the separation matrix. The DP $F_R(f,\theta)$ of the microphone array in the *R*th source direction is given by [28]

$$F_R(f,\theta) = \sum_{k=1}^{M} W_{Rk}^{(ICA)}(f) \exp[j2\pi d_k \sin\theta / c], \qquad (24)$$

where $W_{Rk}^{(ICA)}(f)$ is an element of the separation matrix obtained in Eq. (22), *R*=1,2. The DP of the separation matrix contains nulls in each source direction. However, the positions of the nulls vary in each frequency bin for the same source direction. Hence by calculating the null directions in each frequency bin, the DOA of the *R*th source can be estimated as

$$\hat{\theta}_R = \frac{2}{N} \sum_{p=1}^{N/2} \theta_R(f_p), \qquad (25)$$

where $\theta_R(f_p)$ denotes the direction of null in the *p*th frequency bin. For the present case of two sources, these are given by

$$\begin{aligned} \theta_1(f_p) &= \min\left[\arg.\min_{\theta} |F_1(f_p,\theta)|, \ \arg.\min_{\theta} |F_1(f_p,\theta)| \right], \\ \theta_2(f_p) &= \max\left[\arg.\min_{\theta} |F_1(f_p,\theta)|, \ \arg.\min_{\theta} |F_1(f_p,\theta)| \right], \end{aligned} \qquad (26)$$

where $\min[u,v]$ and $\max[u,v]$ are defined to choose minimum and maximum, respectively, from *u* and *v*. Then the separation matrix in each frequency bin is arranged in accordance with the directions of nulls, which sort-out the permutation problem. After estimating DOA, the gain value in each frequency bin is normalized in each source direction. Gain in the *R*th source direction in the *p*th frequency bin is given by

$$\alpha_R(f_p) = \frac{1}{F_R(f_p, \hat{\theta}_R)} \qquad (27)$$

where $\hat{\theta}_R$ is the estimated direction of the *R*th source. Thus, a scaled separation matrix is obtained as

$$\mathbf{W}(f_p) = \begin{bmatrix} \alpha_1(f_p) 0.. & 0 \\ : & :: \ 0 \\ 0 & 0..\alpha_R(f_p) \end{bmatrix} \begin{bmatrix} W_{11}(f_p) & .. & W_{1R}(f_p) \\ . & .. & . \\ W_{M1}(f_p) & .. & W_{MR}(f_p) \end{bmatrix} \qquad (28)$$

This scaled and depermuted matrix is used to separate the signals in each frequency bin. Then by using overlap-add technique [29] time-domain signal is reconstructed from the TFSS of each source. However, in order to use W(f) of Eq. (22) in the time domain to form an FIR filter, it is essential to de-whiten the separation filter as follows:

$$\mathbf{W}(f) = \mathbf{W}(f)(\mathbf{Q}(f))^{-1}. \qquad (29)$$

Then using de-whitened *W(f)*, an FIR filter of length *P* can be formulated to separate the signals directly in the time-domain as follows

$$y(n) = \sum_{r=0}^{P} w(r)x(n-r). \qquad (30)$$

A. Algorithm initialization

The deflationary learning rule for **w** in Eq.(20) is sensitive to the initial value of separation vector **w**. It can be initialized by a random value or some heuristically chosen good guess values such as NBF-based initial value. NBF is a geometrical technique for the speech signal separation in which the separation filter depends on the DOA, frequency of the signal and the geometry of the used microphone array. NBF jams signals from the undesired directions by forming nulls in DP in that directions while setting look direction in the direction of desired signal source. Accordingly, DP in Eq.(24) for the NBF based separation matrix $W^{BF}(f)$ for the look direction $\hat{\theta}_1$ and null direction $\hat{\theta}_2$ should satisfy the following conditions

$$F_1(f,\hat{\theta}_1)=1 \text{ and } F_1(f,\hat{\theta}_2)=0 \tag{31}$$

These simultaneous equations can be solved to give the following solutions for the elements of separation matrix $W^{BF}(f)$

$$W_{11}^{BF}(f)=-\exp[-q_1\sin\hat{\theta}_2]\times\{-\exp[q_1(\sin\hat{\theta}_1-\sin\hat{\theta}_2)\times \exp[q_2(\sin\hat{\theta}_1-\sin\hat{\theta}_2)]\}^{-1} \tag{32}$$

$$W_{12}^{BF}(f)=-\exp[-q_2\sin\hat{\theta}_2]\times\{-\exp[q_1(\sin\hat{\theta}_1-\sin\hat{\theta}_2)\times \exp[q_2(\sin\hat{\theta}_1-\sin\hat{\theta}_2)]\}^{-1} \tag{33}$$

Similarly, for the look direction $\hat{\theta}_2$ and null direction $\hat{\theta}_1$ following conditions are satisfied by the elements of separation matrix $\mathbf{W}^{BF}(f)$

$$F_2(f,\hat{\theta}_1)=0 \text{ and } F_2(f,\hat{\theta}_2)=1 \tag{34}$$

On solving these, the following solutions are obtained

$$W_{21}^{BF}(f)=-\exp[-q_1\sin\hat{\theta}_1]\times\{-\exp[q_1(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]- \exp[q_2(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]\}-1 \tag{35}$$

$$W_{22}^{BF}(f)=-\exp[-q_2\sin\hat{\theta}_1]\times\{-\exp[q_1(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]- \exp[q_2(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]\}^{-1} \tag{36}$$

where $q_1 = j2\pi d_1 f/c$ and $q_2 = j2\pi d_2 f/c$,
c= velocity of sound in the given environment.

The NBF based separation matrix is approximately optimal and is derived for ideal far-field propagation of acoustic wave. However, under the reverberant condition, its separation performance degrades markedly.

B. Objective Evaluation Score

In order to evaluate the performance of the algorithm Noise Reduction Rate (NRR), Spectral NRR (SNRR), and Spectral Correlation Coefficient (SCRF) γ(f) will be used. NRR is defined as ratio of speech signal power (computed from reference signal) to the noise power. SNRR (SNRR) is given as NRR in any frequency bin. SNRR for the *i*th source (here M=R=2) in the *f*th frequency bin is given by

$$SNRR_i(f) = 10\log_{10}\frac{E\{|W_{i1}(f)ref_{1i}(f)+W_{i2}(f)ref_{2i}(f)|^2\}}{E\{|Y_i(f)-W_{i1}(f)ref_{1i}(f)+W_{i2}(f)ref_{2i}(f)|^2\}}, \tag{37}$$

SCRF between ICs Y_1 and Y_2 in a frequency bin *f* is given by

$$\gamma(f)=\frac{\sum_1^m[\{Y_1(f)-\bar{Y}_1(f)\}\{Y_2(f)-\bar{Y}_2(f)\}]}{\sqrt{\sum_1^m|Y_1(f)-\bar{Y}_1(f)|^2}\sqrt{\sum_1^m|Y_2(f)-\bar{Y}_2(f)|^2}}. \tag{38}$$

V. Experiments and Results

The layout of experimental room is shown in Fig.6.The spacing between two microphone was kept at 4 cm. Voices of two male and two female speakers, at the distances of 1.15 meters and from the directions of -30° and 40° were used to generate 12 combinations of mixed signals x_1 and x_2 under the described convolutive mixing model. Mixed signals at each microphone were obtained by adding speech signals *ref*$_{11}$, *ref*$_{12}$, *ref*$_{21}$, *ref*$_{22}$. The speech signals *ref*$_{11}$, *ref*$_{12}$, *ref*$_{21}$, *and ref*$_{22}$ reaching each microphone from each speaker are used as the reference signals. These speech signals were obtained by convolving seed speech with room impulse response, recorded under different acoustic conditions, characterized by a different Reverberation Time (RT), e.g., RT=0 ms, RT=150 ms and RT=300 ms.

First of all STFT analysis of the mixed data is done to obtain TFSS. The STFT analysis conditions are shown in the Table1. The TFSS data in each frequency bin are whitened in accordance with Eq.(7) before being fed into iterative ICA loop. As explained in the previous sections whitening is only half ICA, the whitened data are used to learn separation vector in accordance to Eq.(20). At first the algorithm is initialized using random values of separation vector **w** in each frequency bin**.** Algorithm learns separation vector in each frequency bin

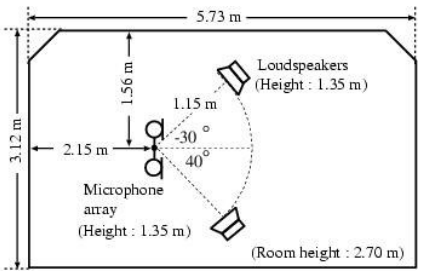


Fig.6 Layout of the experimental setup.

by formed by the separation matrix are shown in Fig.7. The algorithm begins to converge after 20 iterations (less for RT=0 ms) for RT=300 ms and stops when the stopping criterion is satisfied. The stopping criterion δ was fixed at 0.001.

Using directivity-pattern-based methods, DOAs of the sources are estimated. The DOAs of the 1st source S_1 and 2nd source S_2, estimated using Eq.(25), are presented in Table 2 along with true DOAs. The histograms of Direction of Nulls (DON) formed by the separation matrix are shown in Fig.7. It is evident from there that in all frequency bins DON are not in the same direction. In some frequency bins it is swapped with the DOA of other sources indicating that separation matrix is permuted, however, maximum no. of nulls are occurring in a particular source direction (shown as white bar in Fig.8) and hence this can also be used as the DOA information.

Using the estimated source direction, the separation matrix

Table 1. Signal analysis conditions	
Sampling freq.	8000 Hz
Frame Length	20 ms
Step Size ε	10 ms
Window	Hanning
FFT length	512
δ	0.001

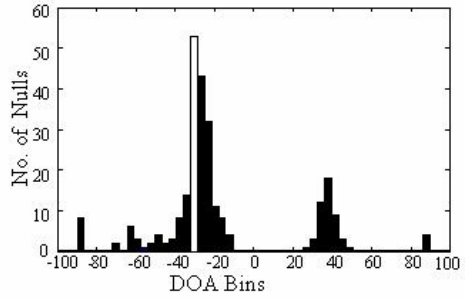
Table 2 DOA Estimation result						
RT	RT= 0 ms		RT=150ms		RT=300 ms	
Methods	S_1	S_2	S_1	S_2	S_1	S_2
Averaging	-31.1	40.0	-32.2	39.0	-28.1	42.1
True DOA	-30	40	-30	40	-30	40

is scaled using Eq.(28). The DP of the separation matrix before and after de-permutation and scaling are shown in Fig. 9. That figure shows how the directional nulls of the separation matrix get blurred with increasing *RT*. After solving the permutation and scaling problem the DP of separation matrix

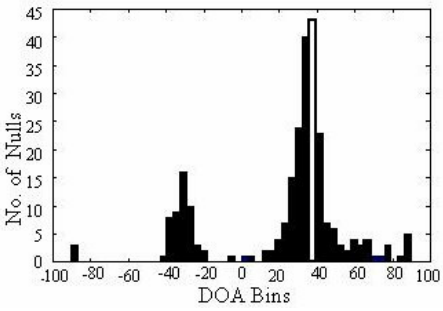
shows unity gain in the estimated look direction and nulls in the direction of source to be rejected.

In order to evaluate the performance of the algorithm with NBF based initialization, the initial value of **w** is generated for every frequency bin using the estimated DOA and Eq. (32, 33, 35 and 36). Using these initial values in each frequency bin, ICA is performed. The NRR results under both initializations are shown in Fig.10. There occurs severe degradation in the separation performance with the increasing reverberation time in both cases. It is also evident from Fig.10 that the NRR improvements for the non-reverberant case are almost same for the both types (NBF based and random value based) of initializations. However, for reverberant conditions, NBF-based guess value shows improvement in the NRR performance as well as in the convergence speed, see Fig. 11, over random initialization In order to study the effect of over-iteration on the separation performance, NRRs for the different number of iterations for both the NBF based initialization and random value based initialization were observed under different RTs. The average NRR versus number of iterations for RT=150 ms and RT=300 ms are shown in Fig.12. The maximum iteration limit was set at 1000. It is evident from that figure that NRR performance is slightly changed by over-learning and NBF based initialization results in better performance than that of the random value based initialization.

The overall separation performance of the algorithm depends on the performance in each frequency bins. As stated before algorithm works independently in each frequency bins, the separation performance measures such as spectral NRR and correlation coefficients between ICs in each frequency bin were observed. Spectral NRR for the male-female speaker combination for RT=0 ms, RT=150ms and RT=300 ms are shown in Fig.13, Fig.14 and Fig.15. Similarly correlation coefficients for RT=300 ms is shown in Fig.16. It is evident that the algorithm does not show similar and good performance in each frequency bin. In some frequency bins it has better performance while in some other frequency bin it has very poor performance, especially with increasing RT. This is indicative of the fact that data in some frequency bins



(a) Nulls of first row of the separation matrix



(b) Nulls of first row of the separation matrix

Figure 7 (a) and (b): Histogram of number of nulls formed by the separation matrix before solving permutation and scaling problem (for RT=150 ms)

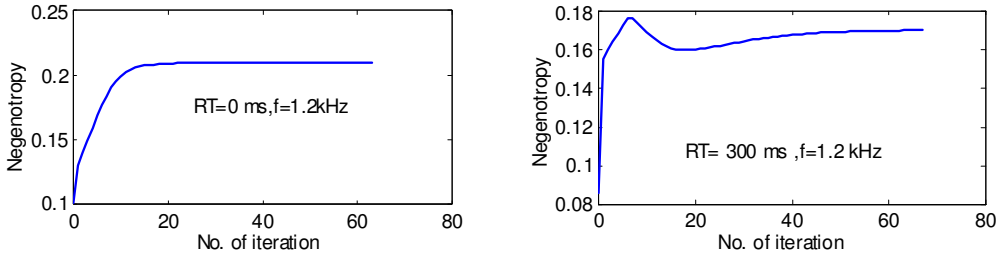


Figure 8: Convergence of the algorithm for the source combination (male and female), *f*=1.2kHz, RT=0 ms (Left), RT=300 ms (Right), δ=. 001

may be ill conditioned. In [30] it has been investigated that the TFSS of speech in each frequency bin does not follow CLT, as a result of which working of the algorithm is hampered in such frequency bin. However, this may be one of the important causes for the poorer performance of thealgorithm in some frequency bins

VI. CONCLUSIONS

In this study, we used Lagrangian multiplier method optimization to derive a fixed-point learning rule for the blind separation of convoluted mixture of speech in the frequency domain. We also used DP method to solve permutation and scaling problems. The use of Null beamformer as the initial value for the algorithm initialization was also studied and results were compared for that of random value based initialization. Also, the histogram-based method for DOA estimation was introduced. The performance of the algorithm under reverberant condition is very poor and need to be improved. However, the convergence speed of the algorithm is much better than that of the gradient based algorithms. We are looking further for the possibility of improving the separation performance of the algorithm. The possibility of combining gradient-based FDICA with fixed-point ICA is also left for future work. The slow convergence near the convergence point of the gradient-based ICA might be improved by switching over to the fixed-point algorithm.

REFERENCES

[1] P. Common, "Independent Component Analysis, A New Concept?," *Signal Processing*,vol.36, pp.287-314, 1994.
[2] A. Hyvarinen, "Survey on independent component analysis,' *Neural Computing Surveys*, vol.2, pp.94-128, 1999.
[3] Cardoso J.F., J. Delabrouille, Guillaume, "Independent component analysis of the cosmic microwave background," *Proc. ICA2003*, 1111-1116, Nara, Japan, 2003.
[4] Cherry, E. Collin, "Some experiments on recognition of speech, with one and with two ears," *Journal of Acoustical Society of America*, 25:975-979, 1953.
[5] Cardoso, J.F., "On the performance of orthogonal source separation algorithms," *Proc. of EUSIPCO-94*, Edinburgh, 1994.
[6] Cardoso, J.F., "Eigenstructure of 4th order cumulants tensor with application to the blind source separation problem," *Proc. ICASSP'89*, 2109-2112, 1989.
[7] Jutten, C., Herault, J, "Blind separation of sources part 1: An adaptive algorithm based on neuromimetic architechture," J. *Signal Processing*, 24, 1-10, 1991.
[8] K.Kashino, K. Nakadai, T.Kinoshita, and H.Tanaka, Organization of hierarchical perceptual sounds," *Proc. 14th Int. conf. On Artificial Intelligence*, vol.1, 158-164,1995.
[9] T.W. Parson, "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am., 60, 911-918, 1976.
[10] M. Unoki, M. Akagai, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, 27, 261-279, 1999.
[11] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, 60, 926-935, 1972.
[12] L.J. Griffiths, C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propag*., 30, 27-34, 1982.
[13] Y. Kaneda, J.Ohga, Adaptive microphone array system for noise reduction,"IEEE Trans. Acoust., Speech and Signal procs., ASSP-34, 27-34,1986.
[14] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*,Vol.2003, No.11, pp.1135—1146, 2003.
[15] T.W.Lee, "*Independent Component Analysis*", Norway, Kluwer Academic Press, 1998.
[16] S. Haykin, "*Unsupervised Adaptive Filtering*, John Wiley and Sons Inc.,NewYork, 2000.
[17] P. Samaragadis, Blind separation of convolved mixture in the frequency domain, *Neurocomputing*, vol.22, pp.21-34, 1998.
[18] Araki, S. et al., "The fundamental relation of frequency domain blind source separation for convolutive mixures of speech,". IEEE Trans. Speech and Audio Processing, Vol. 11, no.2, 109-116, 2003.
[19] S.Ikeda., S. Murata, "A method of ICA in time-frequency domain," *Proc. Workshop Indep. Compon. Anal.Signal Sep.*, 365-367, 1999.
[20] T.Nishikawa, et al., (2003). Blind Source Separation of Acoustic Signals Based on Multistage ICA Combining Frequency-Domain ICA and Time-Domain ICA. IEICE Trans. Fundamentals, Vol.E86-A, pp.846-58, no.4, April, 2003.
[21] K. Torkkola, "Blind Separation for audio signals-are we there yet? *Proc. Workshop on ICA & BSS*, France, 1999.
[22] E. Bingham et al., "A fast fixed point algorithm for independent component analysis of complex valued signal," *Int. J. of Neural System*, 10(1)1: 8, 2000.
[23] Aapo Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks* 10(3):626-634, 1999.
[24] N. Mitianoudis, N. Davies, "New fixed point solution for convolved audio source separation," *Proc. IEEE Workshop on Application of Signal Processing on Audio and Acoustics*, New York. (2001).
[25] Cichocki A., S. Amari, "*Adaptive Blind Signal and Image Processing, Learning Algorithms and Application*," John Wiley & Sons Ltd., 130-131, 2002.

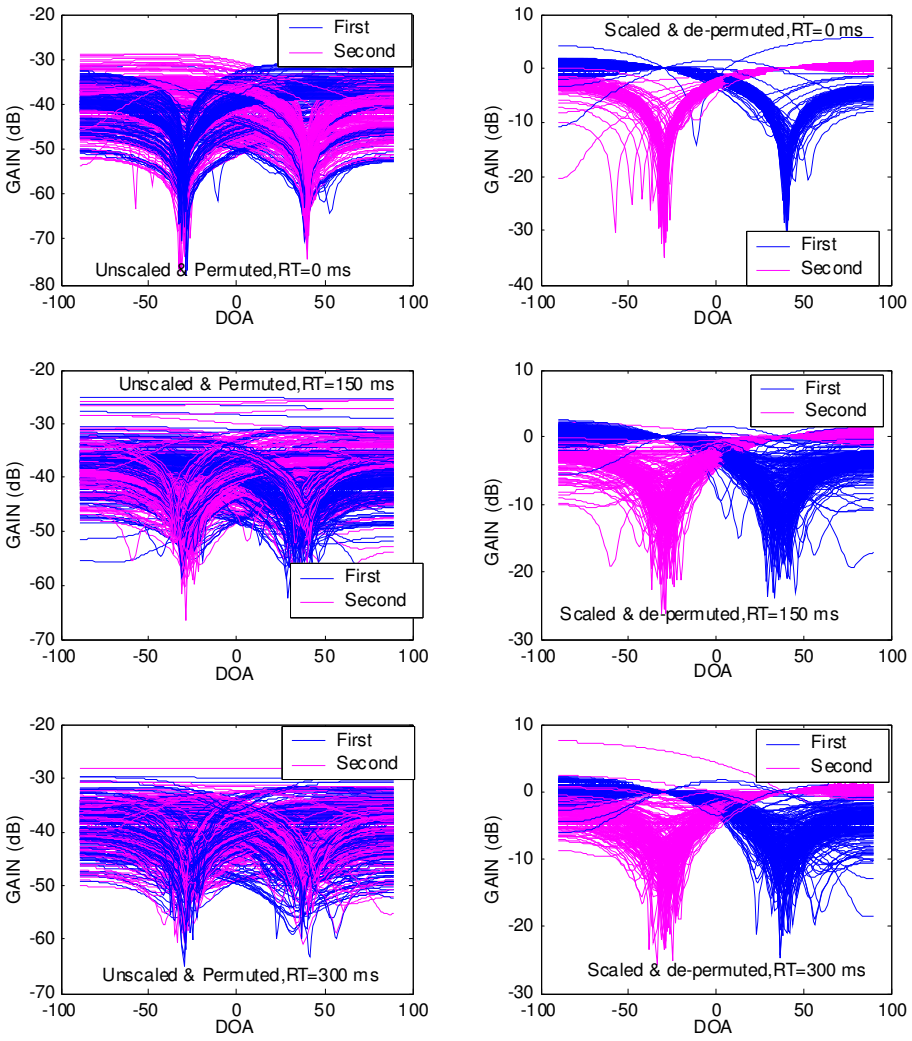


Figure 9: Directivity patterns of the ICA based separation matrix obtained under different reverberation time. The left-hand side is unscaled and permuted and right-hand side figures represent DP for the scaled and permuted separation matrix. Under no reverberation nulls are sharp and clear resulting in good separation. For moderate or high reverberation directional nulls are blurring which results in poor separation

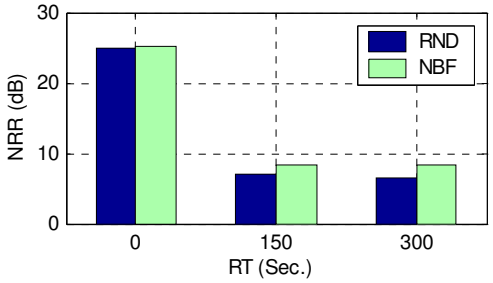


Figure 10: NRR improvement using NBF based and random initial value for **w** in different acoustic environment

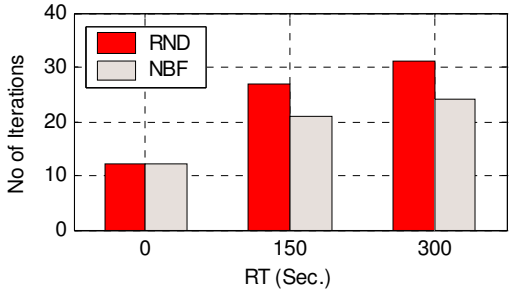


Figure 11: Average no of iteration consumed in extracting both sources under NBF and random (RND) value based initialization for different RT.

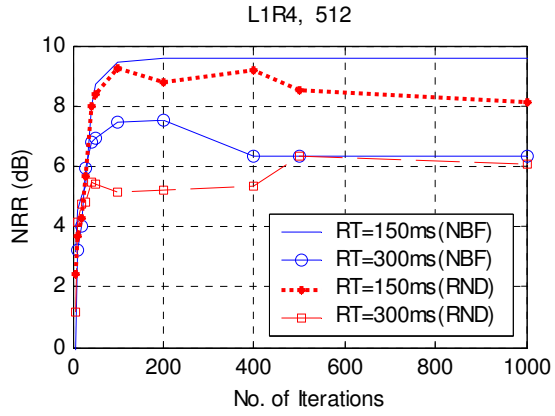


Figure 12: Effect of over-iteration on the NRR performance

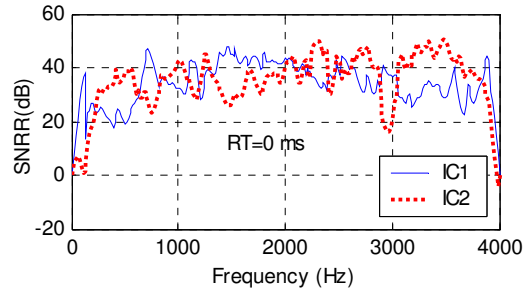


Figure 13: SNRR for RT=0 ms for male female speaker combination

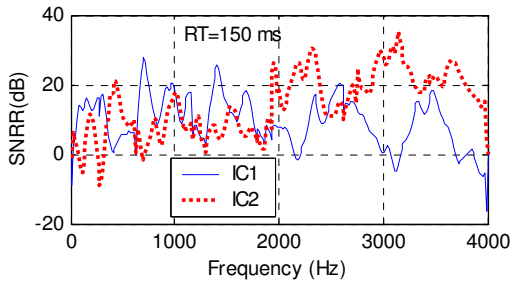


Figure 14: SNRR for RT=150 ms for male female speaker combination

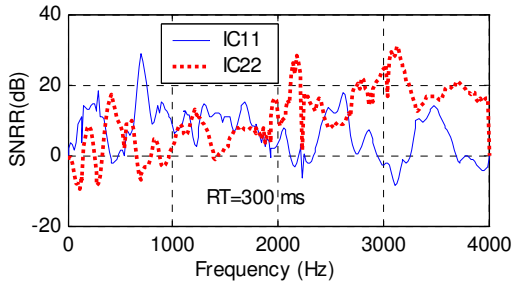


Figure 15: SNRR for RT=300 ms for male female speaker combination

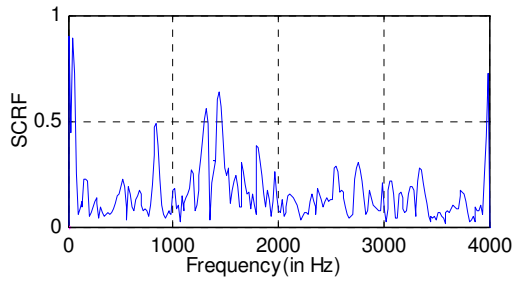


Figure 16: SCRF for RT=300 ms for male female speaker combination

[26] H. Johnson et al, "*Array Signal Processing Concepts and Techniques,*" Prentice Hall, 1993.

[27] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust approach to the permutation problem of frequency-domain blind source separation," *IEEE International Conference on Acoustics, Speech, and Signal (ICASSP2003)*, 381-384, 2003.

[28] S.Kurita,H.Saruwatari,S.Kajita,K.Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant condition," *Proc. ICASSP2000*, vol.5, pp.3140-3143, (2000).

[29] Godsill S.J., P.J.W. Rayner., "*Digital Audio Restoration*," Springer Verlag London, (1998).

[30] Prasad. R. K, H. Saruwatari, K. Shikano, "Problems in blind separation of convolutive speech mixture by negenotropy maximization,", *Proc. IWAENC 2003*, Kyoto, Japan, 287-290. (2003)

Rajkishore Prasad was born in Bihar, India on January 23, 1970. He received B.Sc(H)., M.Sc. (1995) and Ph.D (2000) in Electronics from B.R.A. Bihar University Muzaffarpur, India. He received JRF and eligibility for Lectureship from CSIR, Govt. of India in 1995 and joined University Deptt. of Electronics BRA Bihar University, Muzaffarpur, India as a Lecturer in 1996. He received Japan Government fellowship 2000 under which he has earned degree of D.Eng. in 2005 from Nara Institute of Science and Technology, Japan. His research interests include expert system development, voice-active robots, and blind source separation. He is life-member of IETE, New Delhi, India.

Prof. Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO.,LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include speech processing, array signal processing, nonlinear signal processing, blind source separation, blind deconvolution, and sound field reproduction. He received the Paper Award from IEICE in 2000.He is a member of the IEEE, the IEICE and the Acoustical Society of Japan.

Prof. Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. His major research areas are speech recognition, multi-modal dialog system, speech enhancement, adaptive microphone array, and acoustic field reproduction. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990-1993, he was the executive research scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986-1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984-1986, he was a visiting scientist in Carnegie Mellon University, where he was working on distance measures, speaker adaptation, and statistical language modeling. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001. He is a member of the Institute of Electronics,Information and Communication Engineers of Japan (IEICE), Information Processing Society of Japan, the Acoustical Society of Japan (ASJ), Japan VR Society, the Institute of Electrical and Electronics, Engineers (IEEE), and International Speech Communication Society