Acoust. Sci. & Tech. **26**, 6 (2005)

TECHNICAL REPORT

Effect of Central Limit Theorem non-compliance on blind separation of speech by negentropy maximization

Rajkishore Prasad*, Hiroshi Saruwatari† and Kiyohiro Shikano‡

Graduate School of Information Science, Nara Institute of Science and Technology,
Takayama-Cho, Ikoma, 630–0101 Japan

(*Received 22 July 2004, Accepted for publication 15 April 2005*)

Abstract: In this paper the blind separation of speech signals from their convoluted mixtures using frequency domain fixed-point independent component analysis algorithm, based on negentropy maximization, is presented. We also discuss fundamental problems of fixed-point ICA by negentropy maximization arising in the separation of the speech signal due to disobedience of the Central Limit Theorem (CLT) by the mixed speech data in the frequency domain. The experimental evidences show that CLT failure is happening due to the spectral sparseness of sources. We also present a blind method to mitigate the negative effects of this by combining null beamforming with the ICA. This combination gives a good result under the low reverberation conditions.

Keywords: Blind signal separation, Independent component analysis, Negentropy, Central limit theorem, Speech signal, Microphone array

PACS number: 43.60.Lq [DOI: 10.1250/ast.26.511]

1. INTRODUCTION

Blind signal separation (BSS), a very hot topic of research among digital signal processing groups since a decade, is a statistical framework to estimate signal contribution of independent sources only from their observed mixtures, with no knowledge about the mixing process i.e. the geometry of sources and sensors is implicit. Thus in the BSS problem we are given with the observations $\boldsymbol{x}(n) = [x_1(n), x_2(n) \ldots x_M(n)]^{\mathrm{T}}$ of M sensors produced by some unknown interaction function F among the R original sources $\boldsymbol{s}(n) = [s_1(n), s_2(n), \ldots s_R(n)]$ given as

$$\boldsymbol{x}(n) = F[\boldsymbol{s}(n)]. \tag{1}$$

The task of BSS is to estimate the optimal $\hat{F}^{-1}$, the inverse of the interaction function, so that the underlying R or M original sources can be optimally estimated, i.e.,

$$\hat{\boldsymbol{s}}(n) = [\hat{s}_1(n), \hat{s}_2(n), \ldots \hat{s}_M(n)] \tag{2}$$

The interaction function depends on the physical situation such as on the geometry and number of sources and sensors, and the source to sensor transfer function. Hereafter, we will refer to the interaction function as the mixing system and inverse interaction function as the demixing system. For the simplest condition F can generate linear instantaneous mixture. However, in this paper we will consider for the case of convolutive mixing system. The complete lack of a mixing process in the estimation of the original sources is compensated by pivoting computation on the assumption of the statistical independence of each source. However, the observed mixtures of signals are not statistically independent due to the unknown mixing process. The principle of statistical independence is brought into play by looking for either non-Gaussianity of or spectral dissimilarity among the sources [1]. The process of taking out hidden sources as the most independent components of the mixed data is called Independent Component Analysis (ICA) [2] and there have been developments of numerous ICA-based BSS algorithms in the different areas of practical applications e.g. see [3,4]. In the area of speech signal processing,researchers are taking BSS as one of the strongest aspirant techniques for the practical solution to the Cocktail party problem [5] for an artificial speech recognizer. The origin of the BSS technique in audio signal separation can be traced back to the contributions of Cardoso [6] and Jutten [7] for practical signal separation algorithms based on the aforesaid principle of statistical independence of the sources. The estimated ICs do not represent exact replica of the individual source signals hidden in their observed hotch-

*e-mail: kishor-p@naist.jp
†e-mail: sawatari@naist.jp
‡e-mail: shikano@naist.jp

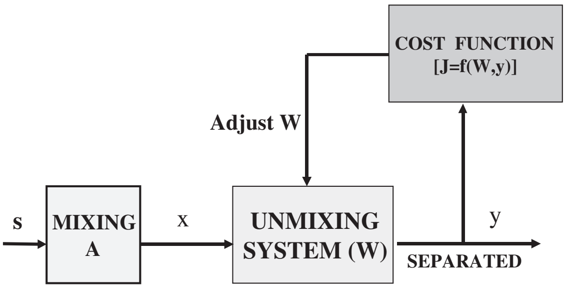

Fig. 1 Block diagram of ICA based BSS algorithm.

potch. Recently, there have been development of many excellent algorithms, in the time domain and in the frequency domain or mutualistically combined in both while weighing their pros and cons, for audio source separation based on ICA [8–10]. In fact, in the list of BSS methods for the audio source separation, ICA-based BSS algorithms have been dominating due to the emergence of several algorithms. However, due to their computational complexities and slow convergence there hardly exists any algorithm that can handle the general class of BSS problems for real world applications in real time [11]. The basic functioning of the ICA based BSS algorithms are shown in Fig. 1. The observed mixed signals $\boldsymbol{x}(n) = [x_1(n), x_2(n) \ldots x_M(n)]^{\mathrm{T}} = \boldsymbol{As}(n)$, where $\boldsymbol{A}$ is the mixing system, are pas sed through a tentative initial demixing system $\boldsymbol{W}$ (randomly chosen or based on some heuristic guess and subject to further modification) and then the mutual independence among the estimated independent component signals $\boldsymbol{y}$ is evaluated by some cost function $J(\boldsymbol{W}, \boldsymbol{y})$, 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 the maximum 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 cost function may be based on the joint distribution or the marginal distribution of the signal. The most popular example of the first category is the Kullback-Liebler Divergence (KLD) metric, which measures deviation between the joint distribution of the signal and a pre-assumed source distribution. However, prior knowledge of the joint distribution of sources is not always feasible. The second category of cost functions exploits only statistical properties of the marginal distribution and non-Gaussianity of the data and are statistically less efficient. A lot of algorithms using such cost functions have also been developed and is main concern of this paper. One of the examples of algorithms based on such cost functions and non-Gaussianization of the signals are fixed-point ICA by the kurtosis or negentropy maximization in [12] for the separation of the instantaneousmixture for real valued signal and in [13] for complex-valued signals; however, this al gorithm has no strategy for solving the problems of permutation and scaling arising in speech signal separation in the frequency domain. The fixed-point algorithm for audio source separation can be found in [14,15]. The fixed-point Frequency Domain ICA (FDICA) algorithm for audio source separation work on the Time-Frequency Series of Speech (TFSS), and thus assumes obeyance of CLT from the TFSS in each frequency bin. However, in [16] it has been shown that TFSS of the mixed speech signal fails to follow CLT in every frequency bin and also the separation performance of the algorithm falls in such frequency bins. In general, any ICA algorithm based on the non-Gaussianization of the signal in the light of CLT can face a similar adverse situation and may fight to loose its performance in the same way because of non-compliance with CLT by the TFSS. Such disobedience of CLT by the TFSS pops up many hooked-up questions such as regarding suitability of negentropy based method for speech signal separation, why does such failure occur and how to get rid of it? These novel points are the focal topics of discussion in this paper. For this study we have used fixed-point FDICA based on negentropy maximization, as described in [15]. Here, We have investigated in details event of CLT non-compliance and proposed a method of blind detection of CLT disobeying bins for combining Null beamformer and FDICA.

The rest of this paper is organized as follows. In the next section a signal mixing and demixing model of the microphone array is presented. Section 3 and its subsection provide a brief overview of functioning of fixed-point FDICA based on the negentropy maximization. Section 4 deals with the obeyance of CLT by the TFSS and their testing using frequency domain kurtosis. Section 5 and its subsections present a different experimental result, followed by the conclusions and references.

2. SIGNAL MIXING AND DEMIXING MODEL

In the real recording environment, signals reaching each microphone are not only direct-path signals, but also delayed and attenuated versions of the source signals. Therefore, the real world mixing model is best approximated by the convolution of the source to sensor transfer function and the source signal. Accordingly, the speech signal picked up by a microphone array with M microphones is modeled as a linear convolutive mixture of R impinging source signals such that the M-dimensional signal captured by the array is given by

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

where $\boldsymbol{s}_i(n) = [s_1(n), s_2(n) \ldots s_R(n)]$ represents the original

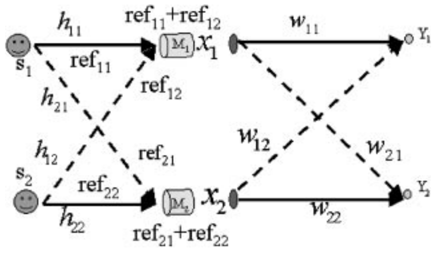

Fig. 2 Convolutive mixing and demixing models for speech signal at two element microphone array.

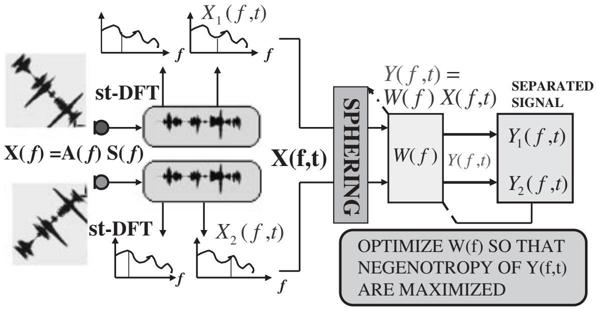

Fig. 3 Block diagram showing basic working principle of the ICA based BSS algorithms.

source signals, h_{ij} is the P-point impulse response between the source i and the microphone j, and n is the time index. 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. A block diagram of the proposed system is illustrated in Fig. 2. Accordingly, the observed signals $x_1(n)$ and $x_2(n)$ at the microphones are given by

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

where $ref_{11}(n) = h_{11} * s_1(n)$, $ref_{12}(n) = h_{12} * s_2(n)$, $ref_{21}(n) = h_{21} * s_2(n)$, and $ref_{22}(n) = h_{22} * s_2(n)$ are here called reference signals and '$*$' represents the convolution operation. In the frequency domain, the same model is represented by taking Short-Time Fourier Transform (STFT) of Eq. (4) as:

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

The FDICA separates the signal in each frequency bin independently. The separation process is given by

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

where $[Y_1(f) \quad Y_2(f)]^{\mathrm{T}}$ are ICs in frequency bin f.

3. FREQUENCY DOMAIN FIXED-POINT ICA

The fixed-point FDICA by negentropy maximization works, like other FDICA, on the time-frequency series obtained by STFT analysis of the speech data. From Eq. (5), it is obvious that the mixed signal in any frequency bin is a summation of contribution of each source in the same frequency bin. Thus, in the frequency domain, assumption of CLT should be obeyed by the TFSS in every frequency bin and its non-Gaussianization in proper way can give optimally non-Gaussian components as ICs or demixed signal [12]. The whole process of fixed-point FDICA is depicted in Fig. 3. It consists of two major operations, namely, whitening and rotation. Whitening or sphering of TFSS is the first half of the ICA task and this process transforms the mixed signal into a spatially decorrelated form. Whitening of the zero mean TFSS can be done using Mahalanobis transform [17]. Accordingly, the whitened signal in the pth frequency bin f_p is obtained as

$$\boldsymbol{X}_w(f_p, t) = \boldsymbol{Q}(f_p)\boldsymbol{X}(f_p, t) \tag{7}$$

where $\boldsymbol{Q}(f_p) = \Lambda_x^{-0.5} V_x$ is called whitening matrix; $\Lambda_x = diag\{1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \ldots 1/\sqrt{\lambda_n}\}$ is the diagonal matrix with positive eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ of the covariance matrix of $\boldsymbol{X}(f_p, t)$ and $\boldsymbol{V}_x$ is an orthogonal matrix consisting of eigenvectors of covariance matrix. The remaining task involves rotating the whitened signal vector $\boldsymbol{X}_w(f_p, t)$ by the separation matrix such that $\boldsymbol{Y}(f_p) = \boldsymbol{W}(f_p)\boldsymbol{X}_w(f_p, t)$ equals TFSS of the independent components in the pth frequency bin. Negentropy is used as cost function to measure degree of non-Gaussianization. The negentropy $J(\boldsymbol{Y})$ of the TFSS of the candidate IC, is given by (frequency index f and frame index t are dropped hereafter for clarity) [12].

$$J(\boldsymbol{Y}) = H(\boldsymbol{Y}_{\text{Gauss}}) - H(\boldsymbol{Y}), \tag{8}$$

where $H(\cdot)$ is the differential entropy of the $(\cdot)$ and $\boldsymbol{Y}_{\text{Gauss}}$ is the Gaussian random variable with the same covariance as of $\boldsymbol{Y}$. This definition of negentropy ensures that it will be zero if $\boldsymbol{Y}(f, t)$ is Gaussian and will be positive if $\boldsymbol{Y}(f, t)$ is tending towards non-Gaussianity. The negentropy can be approximated in terms of non-quadratic non-linear function G as follows [18]:

$$J(\boldsymbol{Y}) = \sigma[E\{G(\boldsymbol{Y})\} - E\{G(\boldsymbol{Y}_{\text{Gauss}})\}]^2, \tag{9}$$

where σ is a constant. The choice of the non-linear function G depends on the Probability Distribution Function (PDF) of the data. The most general form of non-linear function that can be used for speech data (assuming TFSS has super-Gaussian distribution) separation is given as

$$G(\boldsymbol{Y}) = \log(a_2 + \boldsymbol{Y}), \tag{10}$$

where $a_2 = 0.1$. In the deflation type algorithm, the rotation step consists of a one-unit ICA which is used to estimate one separation vector $\boldsymbol{w}$ (any one row of the separation matrix) at a time and is obtained by maximizing negentropy given by

$$J(Y) = E\{G(|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{X}_w|^2)\}, \tag{11}$$

where $\boldsymbol{w}$ is an M-dimensional complex vector such that

$$E\{|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{X}_w|^2\} = 1 \Rightarrow |\boldsymbol{w}| = 1. \tag{12}$$

Using Lagrangian method, the optimization of the above contrast function gives following iterative learning equation for $\boldsymbol{w}$ [13,15].

$$\begin{aligned}\boldsymbol{w}_{\text{new}} &= \boldsymbol{w}(\boldsymbol{E}\{\boldsymbol{g}(|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{X}_w|^2) + (|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{X}_w|^2)\boldsymbol{g}'(|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{X}_w|^2)\}) \\ &\quad - E\{g(|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{X}_w|^2)(\boldsymbol{X}_w^{\mathrm{H}}\boldsymbol{w})\boldsymbol{X}_w\},\end{aligned} \tag{13}$$

where g and g' are first and second order derivatives of G. The stopping criteria for iteration is defined as $\delta = E\{|\boldsymbol{w}_{\text{old}} - \boldsymbol{w}_{\text{new}}|^2\}$, which becomes very small near the convergence. Since each update changes the norm of $\boldsymbol{w}$, after each iteration $\boldsymbol{w}$ is normalized to maintain compliance of Eq. (12). Also, it is essential to decorrelate $\boldsymbol{w}$, after each iteration to forbid convergence to the previously converged point. To achieve this, Gram-Schmidt sequential orthogonalization has been used by which the orthogonalized separation vector for the ith source after jth iteration is given by

$$\boldsymbol{w}_i = \boldsymbol{w}_i - \sum_{j=1}^{i-1}(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j)\boldsymbol{w}_j \tag{14}$$

After finishing ICA in every frequency bin for N sources (less than or equal to the number of microphones), a separation matrix is obtained in every frequency bins for every sources which is given by

$$\boldsymbol{W}(f) = \begin{bmatrix}\boldsymbol{w}_1 \\ \cdot \\ \boldsymbol{w}_2\end{bmatrix} = \begin{bmatrix} W_{11}(f) & \dots & W_{1M}(f) \\ \dots & \dots & \dots \\ W_{R1}(f) & \dots & W_{RM}(f)\end{bmatrix} \tag{15}$$

The deflationary algorithm extracts the TFSS of independent sources one by one in decreasing order of negentropy, however, it is not granted that the ranking of the negentropy of latent sources is the same in the each frequency bin. It depends on the content of the signal. If it is different in different frequency bins, this leads to the inter-exchange or flipping of rows of $\boldsymbol{W}(f)$, given in Eq. (15), in each frequency bin in an unknown order which is called permutation problem. There is no known way, to the best of our knowledge, which can compel algorithm to learn the separation matrix in the same order of sources in each frequency bin. Also, the gain value in each frequency bin is not the same. However, it needs to be the same in each frequency bin for the faithful reconstruction of the signal. This is called the scaling problem. If these two problems are not solved after Eq. (15), further processing with Eq. (16) will give another mixed signals instead of separated components. There have been developments of several methods to resolve these two inherent problems [19]. However, we will use here directivity pattern based method using null beamformer as described in the [15,20]. The Direction of Arrival (DOA) of each source is estimated from the DPs of the separation matrix [19] and is used in solving permutation and scaling problems. The estimated DOA can be further used to do NBF based initialization to algorithm. The depermuted and scaled matrix is used to separate the independent components. The separation matrix has been obtained using whitened signals, its premultiplication with whitened signals in the frequency domain gives the TFSS $\boldsymbol{Y}(f,t) = [\boldsymbol{Y}_1(f,t), \boldsymbol{Y}_2(f,t)\dots \boldsymbol{Y}_R(f,t)]^{\mathrm{T}}$ of the separated signal, i.e.,

$$\hat{\boldsymbol{S}}(f,t) = \boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}_w(f,t). \tag{16}$$

The separated signal is reconstructed using well known add and overlap method. However, in order to use $\boldsymbol{W}(f)$ of Eq. (15) in the time domain to form an FIR filter, it is essential to de-whiten the separation filter as follows:

$$\boldsymbol{W}_{\mathrm{d}}(f) = \boldsymbol{W}(f)\boldsymbol{Q}(f)^{-1}. \tag{17}$$

Then using de-whitened $W_{\mathrm{d}}(f)$, an FIR filter of length P can be formulated to separate the signal.

3.1. Algorithm Initialization

The deflationary learning rule for $\boldsymbol{w}$ in Eq. (13) is sensitive to the initial value of separation vector $\boldsymbol{w}$. The NBF-based initial value for $\boldsymbol{w}$ can be used as one of the good guess values. NBF is 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. In this technique the signal from the undesired direction is jammed by forming null in that direction while look direction is set towards the desired signal source. The details of calculation can be found in [20], however, we place here NBF based separation matrix from [20] for convenience. The elements of NBF based separation matrix are given as

$$\begin{aligned}
W_{11}^{\mathrm{BF}}(f) &= -\exp(-q_1\sin\hat{\theta}_2)\big[-\exp\{q_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)\} \\
&\qquad + \exp\{q_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)\}\big]^{-1}, \\
W_{12}^{\mathrm{BF}}(f) &= -\exp(-q_2\sin\hat{\theta}_2)\big[-\exp\{q_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)\} \\
&\qquad + \exp\{q_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)\}\big]^{-1}, \\
W_{21}^{\mathrm{BF}}(f) &= -\exp(-q_1\sin\hat{\theta}_1)\big[-\exp\{q_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)\} \\
&\qquad - \exp\{q_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)\}\big]^{-1}, \\
W_{22}^{\mathrm{BF}}(f) &= -\exp(-q_2\sin\hat{\theta}_1)\big[-\exp\{q_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)\} \\
&\qquad - \exp\{q_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)\}\big]^{-1}.
\end{aligned} \tag{18}$$

where $q_1 = 2\pi j d_1/c$ and $q_2 = 2\pi j d_2/c$, $c =$ velocity of sound in air. The NBF based separation matrix is approximately optimal and is derived for ideal far-field propagation of acoustic wave. Under the reverberant condition, its separation performance degrades markedly.

4. TFSS AND CENTRAL LIMIT THEOREM COMPLIANCE

The fixed point FDICA by negentropy maximizations extracts TFSS of independent sources by the non-Gaussianization. For the effective functioning of the fixed-point FDICA it is essential that the TFSS of the mixed speech signal should be more Gaussian than that of the ICs. It is evident from Eq. (5) that the TFSS of mixed signal in any frequency bin is a superposition of the spectral contributions of all mixing signals in the same frequency bin. This is the mathematical reason for the Gaussianization of the mixed signal. Thus the power, to separate ICs, comes in the algorithm due to the validity of the logical fact that the Gaussianity of the mixed speech signal is more than that of the independent speech signals. If the above fact is not followed, it will be against the very basic working principle of the algorithm and will hamper the performances of algorithm. The validity of CLT can be checked by computing and comparing the kurtosis of the TFSS of the mixed signal and reference signals in each frequency bin, as it is difficult to approximate true negentropy. The kurtosis of spectral component in each frequency bin, denoted hereafter as Spectral Kurtosis (*SK*), is given as the ratio of the fourth order central moment C_4 to the second order moment C_2 [21]. Accordingly, $SK(f)$ in a frequency bin f is given by

$$SK(f) = \frac{C_4\{S^*, S^*, S^*, S^*\}}{[C_2\{S^*, S^*\}]^2}, \tag{19}$$

where $S^* \in \{\boldsymbol{X}(f,t), \boldsymbol{X}^{\mathrm{H}}(f,t)\}$. This definition varies with the placement of conjugates but following [22,23] and assuming spectral component of speech as complex circular random variable, simplified expression for *SK* is given by

$$SK(f) = \frac{E\{|\boldsymbol{X}(f)|^4\} - 2E^2\{|\boldsymbol{X}(f)|^2\}}{[E\{|\boldsymbol{X}(f)|^2\}]^2}. \tag{20}$$

As in the fixed point algorithm, data are sphered so that Eq. (20) further simplifies to

$$SK(f) = E\{|\boldsymbol{X}(f)|^4\} - 2. \tag{21}$$

The aforementioned condition for Gaussianity of the mixed data can be satisfied by verifying the following conditions in terms of *SK*

$$SK_{m1} < \min\{SK_{ref_{11}}(f), SK_{ref_{12}}(f)\} \tag{22}$$

$$SK_{m2} < \min\{SK_{ref_{21}}(f), SK_{ref_{22}}(f)\}, \tag{23}$$

where $SK_{mi} = SK$ of mixed signal at ith Microphone (Mic). Using the expressions for *SK* in Eq. (20), the validity of the CLT can be tested in each frequency bin. However, this method is not blind because it requires reference signals which are not available in the real applications.

4.1. Blind Method of CLT Obedience Testing

In order to gain Gaussianity in mixing process, TFSS should not belong to alpha stable PDF family as these are closed under any linear operations. In [24] it has been shown that the PDF of TFSS of the unmixed speech signal can be better approximated by the Generalized Gaussian Distribution (GGD) which is parameterized by the mean, scale and shape parameter, say β. The value of shape parameter β decides shape of the distribution. GGD represents a Gaussian PDF for $\beta = 2$, Laplacian PDF for $\beta = 1$, and highly parsimonious PDF for $0 < \beta < 1$. Since the CLT obeyance or disobeyance is logically related to the Gaussianization of the mixed signal, the change in β and *SK* of the TFSS can be used to detect CLT obeyance or disobeyance. The shape parameter and *SK* can be computed from data and their threshold value can be determined by looking into change in Gaussianity of the mixed signal. The relation of β and kurtosis $K(\beta)$ of GGD is given in Eq. (24) below.

$$K(\beta) = \left[\Gamma\left(\frac{5}{\beta}\right)\Gamma\left(\frac{1}{\beta}\right)\right]\left[\Gamma\left(\frac{3}{\beta}\right)\right]^{-1} \tag{24}$$

The variation of kurtosis with the value of β is shown in the Fig. 15. Thus if a TFSS of the mixed signal is fully Gaussian, its *SK* will correspond to $SK_{\mathrm{G}} = K(\beta = 2) = 3$ in Eqs. (22) and (23), and if it is not mixed signal, the speech signals will have at least Laplacian or strongly Laplacian PDF for which *SK* corresponds to $SK_{\mathrm{L}} = K(\beta = 1) = 6$. For the strongly Laplacian case, which is more accurate as shown in [24], kurtosis will be higher than 6. The *SK* of TFSS can be directly computed using Eq. (20). Thus if *SK* of TFSS lies above SK_{L} it will represent a Laplacian or strongly Laplacian signal and related TFSS will fail to comply CLT, however, if *SK* is below SK_{L}, it means signal has gained some Gaussianity due to mixing with other speech signals and so it will comply with CLT. Thus change in kurtosis can be related to the change in the shape parameter β and some threshold value of it can be used to detect CLT obeying and disobeying sub-bands. The acoustic channel too Gaussianizes speech signal, so the Gaussianity of true speech is less than that of received by the microphones. However, mixing of two-speech signals is bigger effect than the Gaussianization by the channel. Thus the threshold less than 1 can work well.

4.2. Objective Evaluation Score

As a performance measure of the algorithm in each frequency bin we define and use spectral noise reduction rate (*SNRR*), spectral correlation coefficient (*SCRF*) $\gamma(f)$, and the number of iterations required to reach convergence in each frequency bin. *SNRR* is the Noise Reduction Rate (*NRR*) defined for TFSS in each frequency bin and is given as the ratio of signal power to noise power in a frequency bin. *SNRR* for the ith source (assuming $M = R = 2$) in the fth frequency bin is given by:

$$SNRR_i(f) = 10\log_{10}\frac{E\{|\boldsymbol{X}_{\text{speech}}(f)|^2\}}{E\{|\boldsymbol{X}_{\text{noise}}(f)|^2\}} = 10\log_{10}\frac{E\{|W_{i1}(f)ref_{1i} + W_{i2}(f)ref_{2i}|^2\}}{E\{|Y_i(f) - W_{i1}(f)ref_{1i} + W_{i2}(f)ref_{2i}|^2\}} \tag{25}$$

SCRF $\gamma(f)$ is easier to calculate and is a good approximation for directly measuring independence. In any frequency bin f it is given by

$$\gamma(f) = \frac{\sum_1^m[\{X_1(f) - \bar{X}_1(f)\}\{X_2(f) - \bar{X}_2(f)\}]}{\sqrt{\sum_1^m |X_1(f) - \bar{X}_1(f)|^2}\sqrt{\sum_1^m |X_1(f) - \bar{X}_2(f)|^2}}. \tag{26}$$

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup and Conditions

In the experiment, we used a two-element linear microphone array with inter-element spacing of 4 cm for the simulated speech data generation. Voices of two male and two female speakers, sampled at 8,000 kHz, situated 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 at both microphones under the described convolutive mixing model. Mixed signal at each microphone was obtained by adding the convolved speech signals ref_{11}, ref_{12}, ref_{21}, ref_{22} according to Eq. (4). These convolved speech signals are 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. The speech signals ref_{11}, ref_{12}, ref_{21}, and ref_{22}, reaching each microphone from each speaker, are used as the reference signals. First TFSS of the mixed speech data were generated by the STFT analysis. For the STFT analysis hanning window of 20 ms with shift size 10 ms, and DFT size of 512 and 1,024 were used. The TFSS of data in each frequency bin is whitened in accordance with Eq. (7) before being fed into iterative ICA loop.

5.2. Separation Results

In the first phase of the experiment, ICA algorithm was initialized with the random values of the separation matrix in each frequency bin and then fixed-point FDICA algorithm was used to compute the separation matrix for $RT = 0$ ms, $RT = 150$ ms and $RT = 300$ ms data. The stopping criterion δ was fixed at 0.001. Using directivity-pattern-based methods, DOAs of the sources were estimated [15,25]. In order to evaluate the performance of the algorithm with NBF based initialization, the initial value of $W(f)$ is generated for every frequency bin using the estimated DOA in Eq. (18). Using these initial values in each frequency bin, ICA was performed. The separation performances for the cases of NBF-based initial values and random initial values of the separation vector have been studied under different acoustic conditions. The *NRR* results are shown in Fig. 4. The performance of the algorithm dramatically degrades with increasing *RT* in the both cases. In order to study the effect of different DFT size and frame shift sizes, further experiments were performed with random and NBF based initialization. The analysis frame size was fixed at 20 ms, which contains 160 samples of data at a sampling frequency of 8,000 Hz, and the frame shift size has been varied from 10% to 80% of the analysis frame size. The results of achieved *NRR* and consumed computation power (number of iterations consumed for fixed δ) are shown in the Fig. 5. The obvious benefit of the NBF based initialization over random value based initialization is rapid convergence. Since the algorithm separates the signals independently in each frequency bin, the separation performance in each frequency bin is important.

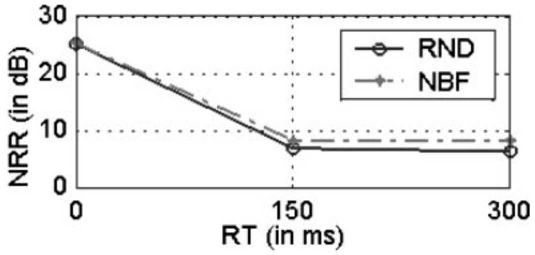

Fig. 4 *NRR* for different *RT* using random initialization and NBF based initialization of the algorithm.

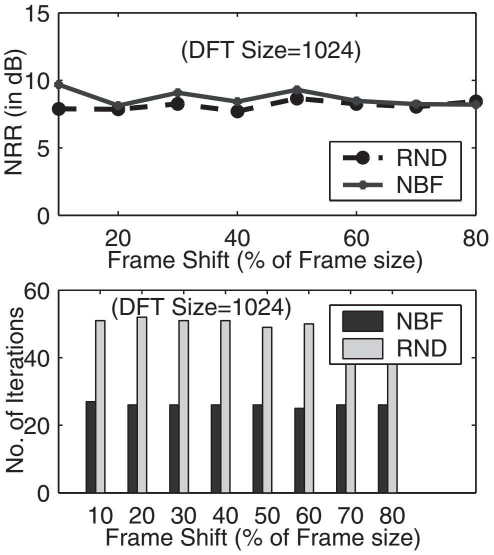

Fig. 5 *NRR* and no. of iterations for different values of frame shift size. RND(NBF) indicates random (NBF based) initialization of the algorithm, $RT = 150$ ms.

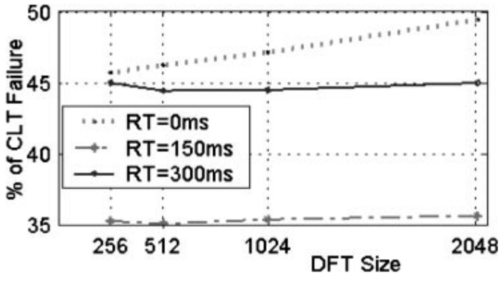

Fig. 6 CLT disobeying frequency bins for different DFT size and *RT* at the first microphone. Shown values are averaged for 6 combination of mixed speech signal.

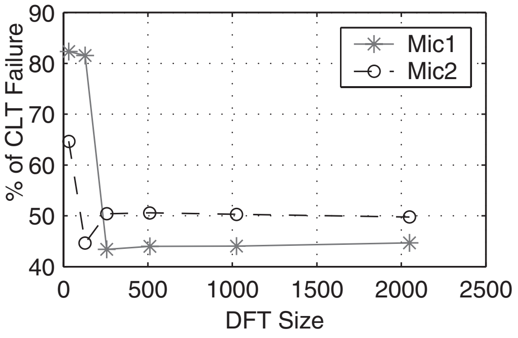

Fig. 7 CLT-disobeying bins at both microphones for different DFT size and $RT = 0\,\mathrm{ms}$.

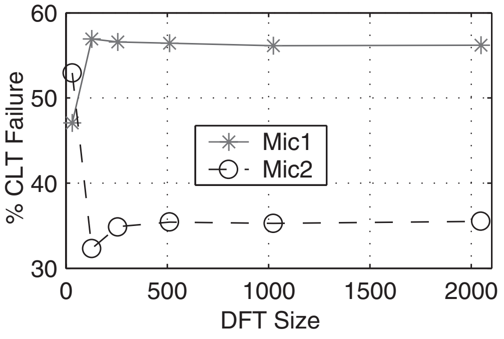

Fig. 8 CLT-disobeying bins at both microphones at $RT = 300\,\mathrm{ms}$.

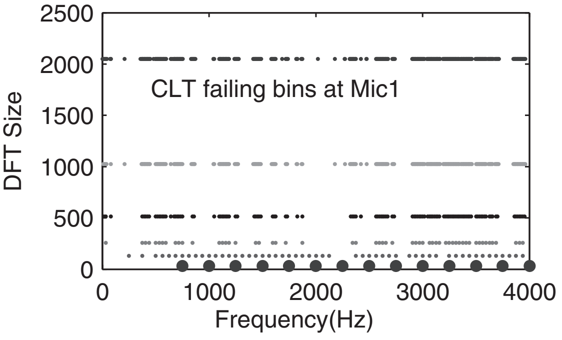

Fig. 9 Clustering of CLT-disobeying bins for different DFT sizes of the speech signal picked up by Mic.1, $RT = 0\,\mathrm{ms}$.

In Figs. 11 and 12 *SNRR*, in Fig. 13 *SCRF* and in Fig. 14 average number of consumed iterations for different speaker combinations are shown. It is evident that the algorithm does not show similar performance in each frequency bin, factors responsible for this may be the difference in nature of data, as while the other experimental conditions are same, in different frequency bins. Among the other statistics of the TFSS Gaussianity of the TFSS of mixed data is important for its proper working. This is discussed next along with more experimental results.

5.3. CLT Obeyance Test

The validity of the CLT in the TFSS of any frequency bin can be checked by verifying the relation given in Eqs. (22) and (23) for the Gaussianity test. That test was performed for the six combination of the mixed speech data for different DFT sizes and reverberation times. Results are shown in Fig. 6. It is interesting to note that the TFSS does not follow the CLT in all frequency bins. The percentage of CLT disobeying TFSS is almost independent of the DFT size and there are no significant changes with the change in reverberation time. However, for the higher values of *RT* a significant difference in the percentage of CLT-failing sub-bands has been found, as shown in Figs. 7 and 8, for both microphones. This is indicative of the fact that the room acoustics is also influential in the disobedience of CLT by the TFSS. As the DFT size increases, the number of CLT-disobeying bins do increase, however, they remain clustered, shown in Fig. 9, due to increase in the frequency resolution for higher DFT sizes. In order to explain this interesting phenomenon we take into consideration the contribution of each signal source in the mixing process, as it is evident from Eq. (5) that TFSS in each frequency bin is a superposition of spectral contribution from each mixing source and this is the cause of Gaussianization. For this the spectral content of the mixed signal and reference signals were examined in the CLT-disobeying frequency bins and in the nearest CLT-obeying frequency bins. In order to measure the spectral contribution, plots of the magnitude of the spectral contribution from each of the reference signals and the mixed signal were examined, and one of such plots is shown in the Fig. 10. In that figure, the temporal contributions of each source in the given frequency sub-band are shown.

It is evident from Fig. 10 that in the shown CLT-failing frequency bin, the contribution from the first speaker is not available at all instances, however, in the nearest CLT-obeying frequency bin its temporal contribution is relatively better. It is also evident from the shown histograms of the temporal contributions that in the CLT obeying or passing bins both sources make very rich contributions but

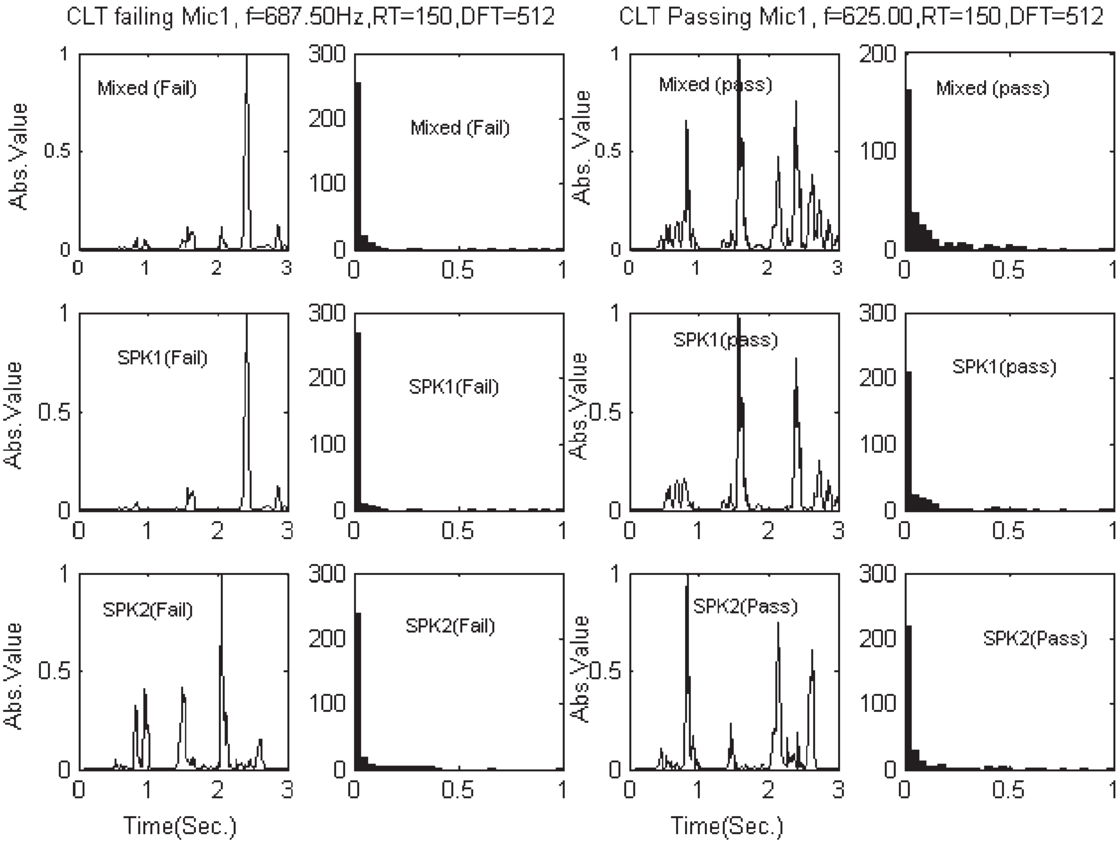

Fig. 10 Role of spectral sparseness in CLT disobedience. Left side, in two columns, represents CLT failing bin at $f = 687.50\,\text{Hz}$ and right side represents CLT failing bin at $f = 687.50\,\text{Hz}$ while same in the right is for nearest CLT passing bin. Subplots in first two columns show temporal contribution and their histograms in the CLT failing bins for mixed and individual speakes. SPK1 and SPK2 represent plots for spectral contributions from the first speaker and second speaker, respectively. Similar things for nearest CLT passing bins are shown in the right two columns (Used speaker combination is male-female).

in the CLT-disobeying bin either one makes a very rare contribution or no contribution, which in accordance with Eq. (5) results in a mixed signal with content from either source or ill-mixed signal. The resulting TFSS, thus in reality, represents a signal from a single source and thus fails to comply with CLT. It is, therefore, concluded that the sparseness in the spectrum has an important role in relation to the CLT non-compliance by the TFSS. It is also important to note that only spectral sparseness cannot be considered to be the sole cause of CLT disobedience. The role of other causes such as room acoustics, natural pauses (this also results in spectral sparseness in the temporal queue of TFSS) cannot be denied. Since TFSS is generated by the STFT analysis it can be inferred that unless there are no long pauses in the speech, it cannot contribute a large number of dumb samples to the TFSS in any frequency bin. In the presence of moderate reverberation, the pause periods may be modified by the reflected speech. Such a reflected speech signal increases correlation among the samples of TFSS, and the spectral content of the signal remains the same even under high reverberation. But if there is any role of pauses in the CLT failure it will be modified by the reverberation. However, such possibilities are still unexplored and are left for further study. In order to show the effect of the CLT disobedience by the TFSS on the separation performance, spectral *NRR* and *SCRF* were observed for different source combinations. Such results for one of the source combinations are shown in Figs. 11–14. It is evident from these figures that in the CLT-disobeying frequency bins *SNRR* is low and *SCRF* is high. This occurs because TFSSs in such frequency bins do not comply with CLT.

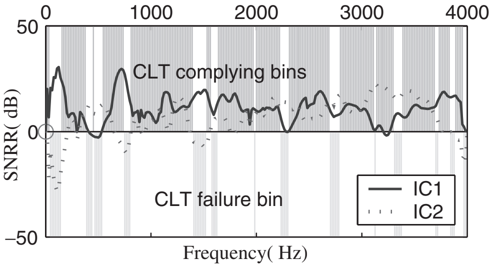

Fig. 11 *SNRR* with CLT disobeying frequency bins at the second Mic. for $RT = 300\,\text{ms}$.

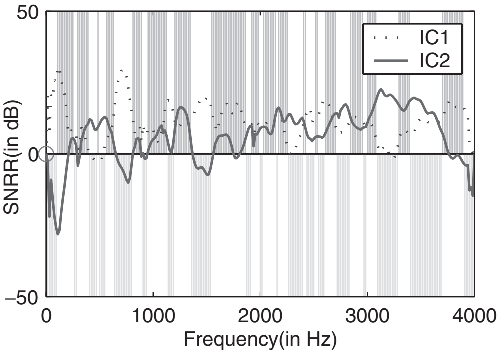

Fig. 12 *SNRR* with CLT disobeying frequency bins at first Mic. for $RT = 300\,\text{ms}$ (for male and female speaker).

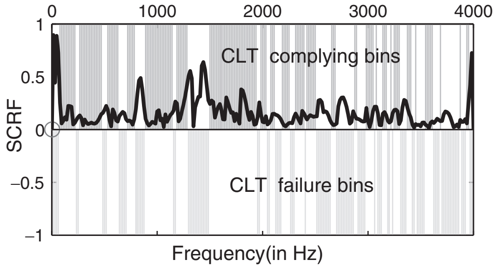

Fig. 13 *SCRF* between separated ICs with CLT disobeying frequency bins at first Mic. for $RT = 300\,\text{ms}$ (for male speakers).

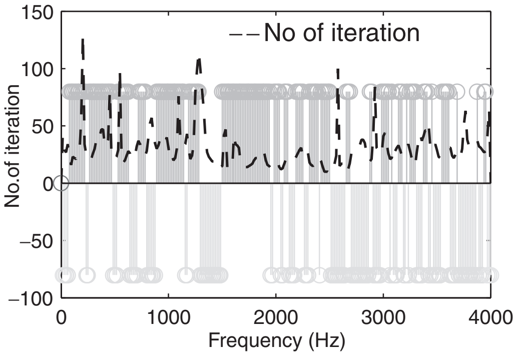

Fig. 14 Average number of iteration taken for convergence for $RT = 300\,\text{ms}$ (for male speakers).

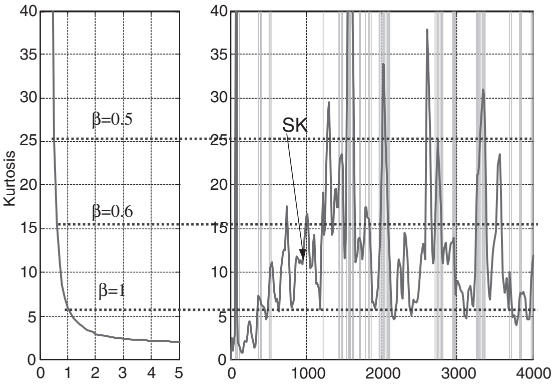

Fig. 15 Threshold determination for the blind detection of CLT-disobeying TFSS.

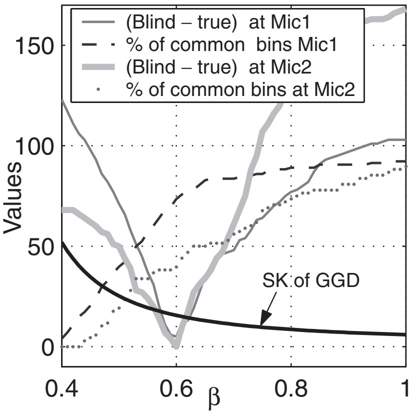

Fig. 16 Comparision of blind and true detection method. Error in the detection is shown by legend 'Blind-true.'

Almost similar results have been found for the other CLT-obeying and disobeying frequency bins. Obviously, CLT compliance is vital for ICA algorithm working under the assumption of Gaussianization of the mixed data under the CLT principle. As the cause, sparseness of spectrum of speech signal, of CLT failure by speech is inherent so its happening cannot be stopped. The only way is to use the algorithms independent from such constraints, or combine some other methods such as NBF having no such problem in the CLT-failing frequency bins. However, this requires the blind detection of CLT obeying and disobeying frequency bins. As discussed in subsection 4.1, a threshold value of *SK* or β can be determined for the blind detection of CLT disobeying bins. The relation between CLT disobeying bins and *SK* can be observed in the right-hand side of Fig. 15.

The plot in the left-hand side represents variation of kurtosis of GGD with β and the right-hand side shows CLT failing bins (gray colored vertical lines in the background) and a plot of *SK* computed using Eq. (20). It is evident that *SK* is high for the CLT-disobeying bins and is relatively low for the CLT-obeying bins. The dashed horizental lines across the plots in Fig. 15 show different threshold values for the different values of β. With the Gaussianization of signal, the value of β shifts towards 2 while for single unmixed speech it is less than 1. The results of blind and true detection are shown in Fig. 16. The term 'true-detection' represents the result obtained by the verifications of the of conditions stated in Eqs. (22) and (23) which need reference signals from each speaker. However, in the real application, these reference signals are unavailable. Plots in Fig. 16 show effect of different value of β on the detection accuracy of the blind method. The plot marked (Blind-true) represents the number of dissimilar frequency bins in the detection, which has minimum value for the threshold around $\beta = 0.6$. Evidently, the blind method falsely detects some bins as the CLT failing, while giving clean chit to a number of frequency bins, which actually fail. However, for the threshold around $\beta = 0.6$, 70% to 80% of bins can be correctly detected. As it is evident from Fig. 15 that the slight change in β produces large change in SK, the slight change in the threshold thus can significantly affect the detection accuracy. An experiment to examine sub-band based combination for null beamformer and fixed-point FDICA was carried out. The combination strategy for the ICA filter and the NBF filter is complex due to occurrence of CLT-failure in different or same frequency bins at both microphones. Thus there are several ways to combine NBF with ICA. However, in our experiment we replaced the ICA filter by that of NBF if CLT failure in any frequency bin is occurring at either microphone. The separation performance, averaged for six mixed signals, is shown in the Fig. 17. It is evident from the figure that the combination shows a significant improvement in the *NRR* for $RT = 0$ ms, and fails to improve for $RT = 150$ ms and $RT = 300$ ms. The reason for this can be explained with the help of Figs. 18 and 19. These figures show, the spectral *NRR* under $RT = 0$ ms, $RT = 150$ ms and $RT = 300$ ms, respectively, for ICA only, NBF only and their combinations. It is evident from these figures that the NBF has a better spectral separation performance under the non-reverberant condition. The performance of NBF too degrades as RT is increased. Under the high reverberation condition, the spectral performance of the NBF is not better than that of the ICA. The spectral performance of both NBF and ICA follow the similar (not exactly same) trend and the overall performance of NBF is worse than that of the ICA. Thus if NBF has a poorer performance than ICA and if the separation filters are exchanged in CLT-failing bins their combination cannot give any improvement instead it may further degrade the performance. Thus the replacement of the ICA filter by the NBF filter results in poor or unimproved performance. However, in some cases it does improve and in other cases its performance was found to be worse than that of ICA.

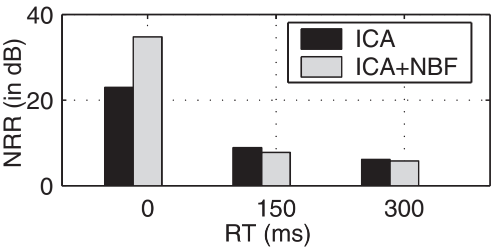

Fig. 17 Overall *NRR* averaged for four combinations of the four mixed speech data.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented in details a peculiar phenomena of CLT disobedience by the TFSS signal, picked up by a linear microphone array in the multiple speaker environment. We have investigated its effect on the blind separation of the speech signal by a fixed-point ICA algorithm based on negentropy maximization. We explored the performance of the algorithm and effect of CLT-failure under different acoustic conditions. We have also presented spectral sparseness of the speech signal as one of the possible causes of noncompliance to CLT by the TFSS. We also proposed a blind detection method for the CLT-

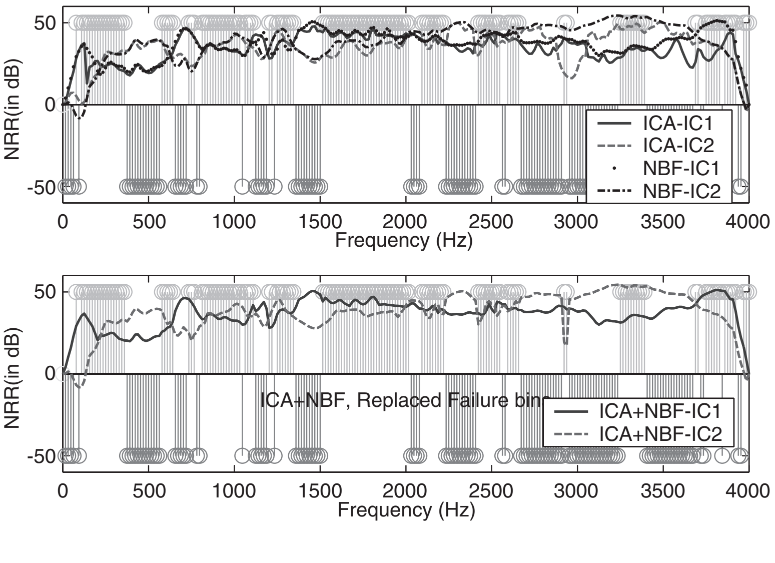

Fig. 18 Spectral *NRR* obtained using ICA, NBF, and ICA+NBF. The positive stems show CLT passing bins while negative stems show CLT failing bins, $RT = 0$ ms (for male and female speakers).

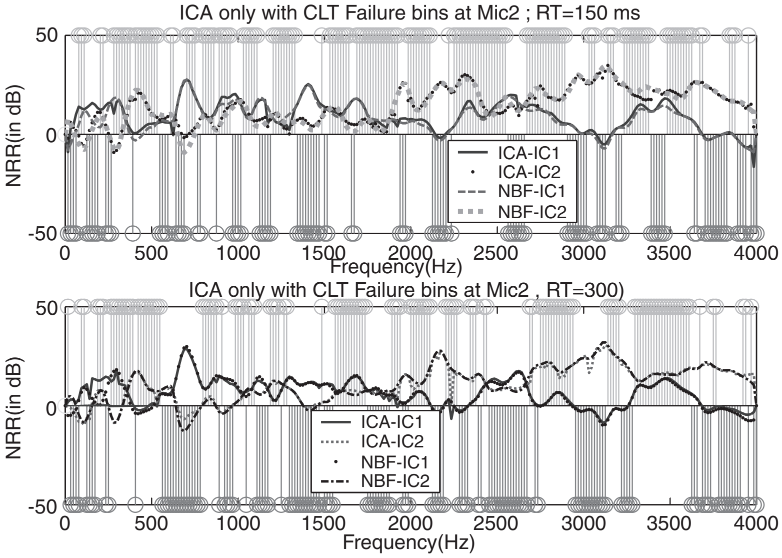

Fig. 19 Spectral *NRR* for $RT = 150$ ms and $RT = 300$ ms. CLT-disobeying and CLT-obeying frequency bins are shown with negative and positive stems, respectively (both speakers are male).

disobeying frequency bins. As a solution to this problem, a new algorithm using ICA and NBF has also been proposed, however, this combination has been found to be effective only in the low reverberation condition. We are further interested in investigating CLT failure from different perspectives and in finding some solution to the problem arising due to this for the fixed-point FDICA algorithm for the audio source separation in reverberant conditions.

REFERENCES

[1] J. F. Cardoso, "On the performance of orthogonal source separation algorithms," *Proc. EUSIPCO-94*, Edinburgh, pp. 776–779 (1994).

[2] P. Comon, "Inependent component analysis: A new concept?," *Signal Process.*, **36**, 287–314 (1994).

[3] A. Hyvarinen, "Survey on independent component analysis," *Neural Comput.*, 94–128 (1999).

[4] J. F. Cardoso, J. Delabrouille and G. Patanchon, "Independent component analysis of the cosmic microwave background," *Proc. ICA 2003*, pp. 1111–1116 (2003).

[5] N. Cherry and E. Collin, "Some experiments on recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, **25**, 975–979 (1953).

[6] J. F. Cardoso, "Eigenstructure of 4th order commulant tensor with application to the blind source separation problem," *Proc. ICASSP '89*, pp. 2109–2112 (1989).

[7] C. Jutten and J. Herault, "Blind separation of sources part 1: An adaptive algorithm based on neuromimetic architechture," *J. Signal Process.*, 24, 1–10 (1991).

[8] S. Araki, R. Mukai, S. Makino, T. Nishikawa and H. Saruwatari, "The fundamental relation of frequency domain blind source separation for convolutive mixures of speech," *IEEE Trans. Speech Audio Process.*, **11**, 109–116 (2003).

[9] R. S. Ikeda and S. Murata, "A method of ICA in time-frequency domain," *Proc. Workshop Indep. Compon. Anal. Signal Sep.*, pp. 365–367 (1999).

[10] T. Nishikawa, H. Saruwatari and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Trans. Fundam.*, **E86-A**, 846–858 (2003).

[11] T. K. Torkkola, "Blind separation for audio signals-are we there yet?," *Proc. Workshop on ICA & BSS*, France, (1999).

[12] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, **9**, 1483–1492 (1997).

[13] E. Bingham and A. Hyvarinen, "A fast fixed point algorithm for independent component analysis of complex valued signal," *Int. J. Neural Syst.*, **10**, 1–8 (2000).

[14] W. N. Mitianoudis and N. Davies, "New fixed-point solution for convolved audio source Separation," *Proc. IEEE Workshop Application of Signal Processing on Audio and Acoustics*, New York (2001).

[15] R. K. Prasad, H. Saruwatari, A. Lee and K. Shikano, "A fixed point ICA algorithm for convoluted speech separation," *Proc. Int. Sym. ICA & BSS*, pp. 579–584, Nara, Japan (2003).

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

[17] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing* (John Wiley & Sons, New York, 2002) pp. 130–131.

[18] A. Hyvarinen, *et al.*, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).

[19] H. Sawada, R. Mukai, S. Araki and S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," *IEEE Int. Conf. Acoust., Speech Signal* (*ICASSP2003*)*, Process.* pp. 381–384 (2003).

[20] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa and K. Shikano, "Blind source eparation combining independent component analysis and beamforming," *EURASIP J. Appl. Signal Process.*, **2003**, 1135–1146 (2003).

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

[22] J. M. Mendel, "Tutorial on higher order statistics (spectra) in signal processing and system theory: Theoretical results and some applications," *Proc. IEEE*, **79**, 277–305 (1991).

[23] P. O. Amblard, M. Gaeta and J. L. Lacoume, “Statistics for complex variable and signals-part I & II,” *Signal Process.*, **53**, 1–25 (1996).

[24] R. K. Prasad, H. Saruwatari and K. Shikano, “Probability distribution of the time-series of speech spectral component,” *J. IEIEC Trans. Fundam.*, **E87-A**, 2292–2300 (2004).

[25] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, “Evaluation of blind signal separation method using directivity pattern under reverberant condition,” *Proc. ICASSP 2000*, Vol. 5, pp. 3140–3143 (2000).