Blind Separation of Speech by Fixed-Point ICA with Source Adaptive Negentropy Approximation

Rajkishore PRASAD (a), Nonmember, Hiroshi SARUWATARI (b), and Kiyohiro SHIKANO (c), Members

SUMMARY This paper presents a study on the blind separation of a convoluted mixture of speech signals using Frequency Domain Independent Component Analysis (FDICA) algorithm based on the negentropy maximization of Time Frequency Series of Speech (TFSS). The comparative studies on the negentropy approximation of TFSS using generalized Higher Order Statistics (HOS) of different nonquadratic, nonlinear functions are presented. A new nonlinear function based on the statistical modeling of TFSS by exponential power functions has also been proposed. The estimation of standard error and bias, obtained using the sequential delete-one jackknife method, in the approximation of negentropy of TFSS by different nonlinear functions along with their signal separation performance indicate the superlative power of the exponential-based nonlinear function. The proposed nonlinear function has been found to speed-up convergence with slight improvement in the separation quality under reverberant conditions.

key words: blind separation of speech, frequency domain independent component analysis, generalized Gaussian distribution, negentropy maximization

1. Introduction

The technique of Blind Signal Separation (BSS) refers to the process of extraction or segregation of signals only from their observed mixtures [1]. The method is called blind because it does not need any information about the mixing process to invert the same. The method is also unsupervised in functioning and has gained research attentions from many fields of practical applications such as Cosmo-informatics, bio-informatics, image processing, digital vision, and speech processing. In the area of speech signal separation it provides one of the feasible solutions for the extraction of speech signal from the cacophony of the sounds. Such signal separation algorithms have potential applications in the area of machine audition where the BSS technique can be used to equip a machine with the capability of steering hearing attention, similar to the anthropomorphic ability known as the cock tail party effect [2], towards a speaker of interest and as a preprocessing stage for enhancing speech signals entering into the artificial speech recognizer. The problem of BSS, in general, can be mathematically formulated as the estimation of R latent signals \( s(n) = [s_1(n), s_2(n), \ldots, s_R(n)]^T \) from only their \( M \) mixed versions, \( x(n) = [x_1(n), x_2(n), \ldots, x_M(n)]^T \). The mixed signals are produced by some unknown interaction \( F \) among them as

\[
x(n) = F[s(n)],
\]

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

\[
s(t) = [s_1(t), s_2(t), \ldots, s_R(t)]^T = F^{-1}[x(n)].
\]

In the simplest case, the mixing process \( F \) produces instantaneous mixture, however, in this paper we will consider the case of convolutive mixing of speech signals. Of course, different principles have been applied to estimate the latent speech signal \( s(t) \), however, the algorithms based on the Independent Component Analysis (ICA) are dominating in number as well as in performance. In general, ICA is a statistical technique for finding out independent components of the data and is potentially applicable in source separation under the assumption that the contribution of each speaker in the mixed signal is statistically independent. This further translates the problem of source separation into the estimation of independent components hidden in the mixed data. Furthermore, there also occur variations among ICA-based BSS algorithms, as there are different approaches to find independent components. However, the main concern of this paper is with the methods based on the non-Gaussianization of the mixed signal by optimizing some non-Gaussianization measure, such as negentropy, of the data. The technique of ICA by non-Gaussianization works by reversing the effect of Gaussianization produced, in light of the Central Limit Theorem (CLT), by the mixing process.

As proposed in [3], the convoluted mixture of speech is converted into the instantaneous mixture of speech in the Discrete Fourier Transform (DFT) domain in each frequency bin. Then the signal can be separated independently in each frequency bin. Accordingly, many algorithms have been proposed in the frequency domain exploiting marginal or joint statistics of the Time Frequency Series of Speech (TFSS). Our concern in this paper is the FDICA algorithms with a marginal- distribution-based contrast function to extract an Independent Component (IC) as the maximally non-Gaussian component from the mixed signal. One of the most successful algorithms of this kind is the fixed-point
ICA by maximization of negentropy [4], in which negentropy of the data, approximated using generalized Higher Order Statistics (HOS) of the nonquadratic nonlinear function, is used as a measure of non-Gaussianity. The choice of the nonlinear function for negentropy approximation is crucial task and has very active links with the Probability Density Function (PDF) of the TFSS of the data [5], [6]. Nevertheless, many general-purpose nonlinear functions have been proposed and also used in speech signal separation [9], [10]. In one of our previous studies [7], we showed that the statistical distribution of TFSS in each frequency bin is not the same and can be better approximated by Generalized Gaussian Distribution (GGD) functions against the most commonly used PDF of Laplacian Distribution (LD) and Gaussian Distribution (GD) functions. Based on that study, our aims in the present paper have been focused on novel research questions such as whether negentropy approximation by different nonlinear functions can influence the separation performance of the algorithm, and if the GGD-based function is a better approximation of the underlying PDF of TFSS, do the nonlinear functions based on it also show superiority in separation? Accordingly, in this paper we examine the performance of the conventional nonquadratic nonlinear function for TFSS and propose a new nonlinear function based on the approximation of PDF of TFSS by the GGD function for the FIDICA algorithm by negentropy maximization.

The rest of this paper is organized as follows. In Sect. 2 mixing and demixing models are presented. Section 3 and its subsections deal with fixed-point FIDICA and the approximation of negentropy by different nonlinear functions. It also deals with the estimation of standard error and bias, using jackknife method, in negentropy approximation using different nonlinear functions. Section 4 deals with experimental results, which is followed by conclusions in Sect. 5 and references.

2. Signal Mixing and Demixing Models

In the real recording environment, the speech signal picked up by a linear microphone array is modeled as a linear convoluted mixture of the impinging source signals and impulse response between the source and sensors. Here, we consider the case of two microphones and two sources for which the signal mixing and demixing models are shown in Fig. 1. Accordingly, the observed signals $x_1(n)$ and $x_2(n)$ at the microphones are given by

$$
\begin{bmatrix}
    x_1(n) \\
    x_2(n)
\end{bmatrix} = \begin{bmatrix}
    re_{f_1} + re_{f_2} \\
    re_{f_1} + re_{f_2}
\end{bmatrix},
$$

where signals $re_{f_1} = h_{11} \otimes s_1(n); re_{f_2} = h_{12} \otimes s_2(n); re_{f_1} = h_{21} \otimes s_1(n); re_{f_2} = h_{22} \otimes s_2(n)$ are called reference signals and $\otimes$ represents the convolution operation. In the frequency domain, the same model is represented by taking the $P$-point Short-Time Fourier Transform (STFT) of Eq. (3) as

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

(4)

An FIDICA algorithm separates the signal in each frequency bin independently, and this separation process in any frequency bin $f$ is given by

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

(5)

where $[Y_1(f), Y_2(f)]^T$ are TFSS of ICs; and $W(f)$ is the separation matrix in frequency bin $f$. Any row of the separation matrix is called a separation vector.

3. Fixed-Point FIDICA

The fixed-point ICA algorithm by negentropy maximization was first proposed by Hyvarinen for real valued signals [4] and its further extension for complex-valued signal by Bhingham et al., in [8]. Its application, with a solution for scaling and permutation, for the separation of speech in the frequency domain can be found in [9], [10]. The functioning of the fixed-point ICA algorithm is based on the compliance of CLT by TFSS, which states that the mixing of a plural number of non-Gaussian signals results in an increase in the Gaussianity of the mixed signal, and hence its non-Gaussianization can yield independent components. It can be imbued from Eq. (4) that TFSS of a mixed signal in any frequency bin is the superposition of spectral contributions of each source. Since TFSS in any frequency bin forms a complex circular random variable [7], PDF of spectral components of the mixed signal, at the $i$th microphone, can be given as the convolution of PDFs of spectral components of each source [12] as follows,

$$
f_{S_i,f}(s_i) = \int_{-\infty}^{\infty} f_{S_1,f}(s_1)f_{S_1,f}(s_1 - s_1)ds_1.
$$

(6)

The effect of such operation on PDF of a strongly LD random variable with zero mean and unit variance, where TFSS are also strongly LD [7], are shown in Fig. 2.

This figure reflects how the density function of a new
variable resulting from repeated additions of strongly Laplacian random variables moves towards the reproducible GD function. This is the point from where came the idea of non-Gaussianization of TFSS for the separation of mixed speech signals into independent components and it works effectively. Obviously, if there is a contribution of each source in all frequency bins, TFSS of mixed speech signals in any frequency bin will be more Gaussian than that of any independent source alone and ICs can be obtained by non-Gaussianization. However, when the contribution from each source is not available in each frequency bin, it is problematic [13],[14]. In the fixed-point ICA, the process of non-Gaussianization is performed in two steps namely, prewhitening or smoothing and rotation of the whitened spectral components of the observed signal. Sphering is performed half of the ICA task by producing spatially decorrelated signals. The whitened signal $X(f,t)$ in the $f$th frequency bin is obtained, after centering the data, using the Mahalanobis transform as follows [15]:

$$X_w(f,t) = Q(f)X(f,t),$$

where $Q(f) = \Lambda_f^{-0.5}V_f$, is called the whitening matrix; $\Lambda_f = \text{diag}[1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \ldots, 1/\sqrt{\lambda_n}]$ is the diagonal matrix with positive eigenvalues $\lambda_1 > \lambda_2 \cdots > \lambda_n$ of the covariance matrix of $X(f,t)$ and $V_f$ is the orthogonal matrix consisting of their eigenvectors. The whitened signal vector $X_w(f,t)$ is then rotated by the separation matrix $W(f)$ such that $Y(f) = W(f)X_w(f,t)$ equals independent components.

The appropriate separation or rotation matrix is learned from the whitened data by optimizing negentropy, which represents the degree of non-Gaussianity, of the candidate separated signal $y = |Y(f)|^2$.

3.1 Approximation of Negentropy of TFSS

As a measure of non-Gaussianity, negentropy provides better performance than others such as kurtosis, as explained in [16]. The term negentropy means the negative of entropy. Negentropy $J(y)$ of the random variable $y$, is given by

$$J(y) = H(y_{\text{gauss}}) - H(y),$$

where $H(.)$ is the differential entropy of $(.)$ and $y_{\text{gauss}}$ is the Gaussian random variable with the same covariance as that of $y$. As among the distributions of given covariance, a Gaussian distribution represents the distribution of maximum entropy, the definition of negentropy in Eq. (8) ensures that it will be zero (minimum) if $y$ is Gaussian and will be increasing as $y$ is becoming more non-Gaussian. Thus the negentropy-based contrast function can be maximized to obtain optimally non-Gaussian components. However, the estimation of true negentropy, as in Eq. (8), is difficult and it requires knowledge of the probability density function of the data. However, it is possible to use some approximation of it and several approximations for negentropy estimation have been proposed and used. The moment-based approximation of negentropy is given as [1],[16]

$$J(y) = \frac{1}{12}E(y^3)^2 + \frac{1}{48} kurt(y)^2 + \ldots,$$

where $kurt(.)$ represents the kurtosis of $(.)$. However, this approximation is equivalent to kurtosis which is a very raw, loose and rough approximation; however, it is also extensively used as a non-Gaussianization measure [17]. The other more accurate approximations have been based on the use of generalized HOS of some nonlinear nonquadratic functions $G(y)$. In terms of such a function, the most widely used approximation of negentropy is given by [16]

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

where $\sigma$ is a positive constant. For the separation of speech signal, an optimally non-Gaussian component can be obtained by maximizing Eq. (10) for $y = \{w^H X_n\}^2$, because the distribution of complex valued samples of TFSS are assumed to be spherically symmetric. The one-unit algorithm for learning the separation vector $w$ (any row of the separation matrix $W(f)$, under the constraint $|w|^2 = 1$, is given by [8],[10]

$$w_{\text{new}} = (wE[g(w^H X_n)^2]) - E[g(w^H X_n)^2]X_n^H w),$$

where first and second-order derivatives of $G(y)$, are denoted by $g(y)$ and $g'(y)$, respectively. The learned separation vector $w$ is normalized after each iteration to comply with the constraint $|w|^2 = 1$. In order to extract many sources, the same learning rule is used with Gram-Schmidt orthogonalization of the learned separation vector to avoid convergence to the same solution in successive iterations. The other inherent problem of permutation and scaling is solved by the directivity-pattern-based method as described in [22].

The separation performance of the fixed-point algorithm depends on the nonquadratic nonlinear function $G(y)$ used. It is desirable that the function $G(y)$ should provide robustness toward outlier values in the data and a better approximation of true negentropy. For better robustness to
outliers, $G(y)$ should show slow variation with respect to change in data and, at the same time a very close approximation of negentropy can be expected if statistical characteristics of $G(y)$ inherit PDF of the data. The statistically efficient and optimal $G(y)$ that can accommodate maximum information about HOS of the data is chosen as the function that can minimize the trace of the asymptotic variance of $w$. The trace of the asymptotic variance of $w$ for the estimation of independent component source $y$, is given by

$$V_G = C \frac{E \left[ g^2(y) \right] - (E \left[ y g(y) \right])^2}{(E \left[ y g(y) - g'(y) \right])^2},$$  

(12)

where constant $C$ depends on the mixing matrix. As shown in [16], $V_G$ is minimized if the nonlinear function $G(y)$ is of the form

$$G(y) = c_1 \log p(y),$$  

(13)

where $c_1$ is an arbitrary constant and $p(y)$ represents PDF of $y$. Keeping these in mind many nonlinear functions have been proposed for $G(y)$. However, for the super-Gaussian signals, the following functions have been recommended [16] and used in the speech signal separation [9],[10]:

$$G_1(y) = \log(\alpha_1 + y), \quad \alpha_1 = 0.01, \quad (14)$$

$$G_2(y) = \sqrt{\alpha_2 + y}, \quad \alpha_2 = 0.01. \quad (15)$$

3.2 PDF for TFSS

In [7], extensive studies have been made on the statistical modeling of TFSS using GD, LD and GGD functions, and it has been reported that the GGD function can provide better and flexible approximation of PDF of the TFSS signal. The GGD function is a parametric function defined in terms of location parameter $\mu$, scale parameter $\alpha$, and shape parameter $\beta$. The GGD function for an arbitrary random variable $z$ is given by

$$f_{GG}(z; \mu, \alpha, \beta) = A \exp \left\{ -\left| z - \mu \right|/\alpha \right\}^{\beta},$$

where

$$A = \frac{\beta}{2\alpha \Gamma(1/\beta)} \quad 1/\alpha = \frac{1}{\sigma} \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)},$$

$$\sigma = \text{Stdv}, \quad \Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt = \text{Gamma PDF};$$

$$\infty < z < \infty, \quad \alpha > 0, \quad \beta > 0. \quad (16)$$

Shape parameter $\beta$ determines the shape of the distribution. For $\beta=1$, the distribution is Laplacian, for $\beta=2$ the distribution is Gaussian and tends to become uniform as $\beta \rightarrow \infty$. The optimal function, say $G_3(y)$, based on GGD can be obtained by using Eq. (16), as the PDF model for $y$, in Eq. (13) and is given by

$$G_3(y) = \alpha^{-\beta} |y|^\beta + \log A. \quad (17)$$

The statistical characteristics of function $G_3(y)$ depend on shape parameter $\beta$ and scale parameter $\alpha$. The nonlinear function in Eq. (17) has been plotted in Fig. 3 for different values of $\beta$. The plotted values of function $G_3(y)$ are normalized. Its smoothness changes with changing shape parameter such that it is less smooth for lower values of $\beta$, but its robustness to outliers in data decreases as its slope increases. Also, the nearness to true PDF of TFSS decreases. It is interesting to observe that in Eq. (17) if $\beta=2$, the nonlinear function becomes the same as that of kurtosis. The kurtosis based contrast function has also been extensively used as a measure of non-Gaussianity; however, it provides poor robustness to outlier values [16],[17]. For $\beta=1$, the nonlinear function in Eq. (17) give measure of second order statistics of the separated signal and is not appropriate to find solution as there is no extreme in the function and can not give separation matrix. However, this can be avoided by making $\beta$ slightly less than or more than unity.

3.3 Error in Negentropy Estimation

We described two conventional nonlinear functions, Eq. (14) and Eq. (15), and a new nonlinear function, Eq. (17), based on statistical modeling of TFSS by the GGD function. In order to further investigate their relative suitability for FDICA, we will evaluate their performance of negentropy approximation and robustness to outliers, and their capacity of signal separation. The first two investigations can be accomplished by estimating error and bias in negentropy approximation employing above functions for which statistical jackknifing technique can be used [18]. Jackknife is one of the powerful tools for data partitioning and can be used to estimate and compare the relative errors in the approximation of negentropy and robustness to outliers in negentropy approximation using nonlinear functions $G_k$ (for $k=1,2,3$) from Jackknife replicates. The jackknife replicates for negentropy are obtained by approximating negentropy of jackknifed samples which are created by omitting, in turn, one data sample from the original TFSS. Let us consider that $Y(f,1), Y(f,2), \ldots, Y(f,i-1), Y(f,i), Y(f,i+1), \ldots, Y^*(f,U)$ is the TFSS, in any frequency bin $f$ consisting of spectral components from $U$ frames. The $i$th jackknife replicate for negentropy approximation by function $G_k$ is given by

![Fig. 3 GGD based nonlinear functions for different values of the shape parameter $\beta$. For the lower values of $\beta$ the nonlinear behavior shown by function is less smooth.](image-url)
\[ J_k^{(i)}(f) = G_k(\{Y(f, 1), Y(f, 2), \ldots, Y(f, i-1), Y(f, i+1), \ldots Y(f, U)\}) \]

and the same process is carried out independently in each frequency bin for each sample. The bias \( J_k^B(f) \) in the negentropy approximation by function \( G_k(y) \) is given by

\[ J_k^B(f) = (N - 1)\left[ J_k^E(f) - J_k(f) \right], \]

where

\[ J_k^E(f) = \frac{1}{N} \sum_{i=1}^{K} J_k^{(i)}(f). \]

The standard error in negentropy approximation using \( G_k(y) \) for frequency bin \( f \) is given by

\[ J_k^E(f) = \left[ \frac{(N - 1)}{N} \sum_{i=1}^{K} [J_k^{(i)}(f) - J_k(f)]^2 \right]^{0.5}. \]

This represents the standard deviation of the jackknife replication. However, it is unbiased due to the presence of factor \((N - 1)/N\) [19]. Since spectral components in frequency bins are assumed to be independent, the above estimates for bias and standard error can be averaged over the number of frequency bins and can be given by

\[ J_k^B = \frac{2}{P} \sum_{i=1}^{P/2} J_k^B(f) \text{ and } J_k^E = \frac{2}{P} \sum_{i=1}^{P/2} J_k^E(f). \]

One of the most important expectations from any contrast function is its separation capacity. However, if the contrast function inherits maximum statistical information of the data, it may provide better separation [5]. The separation performance of each nonlinear function will be judged using the deflationary learning rule given in Eq. (11). Obviously, this requires first and second order derivatives of the nonlinear functions \( G_k(y) \) which are given by

\[ g_1(y) = (a_1 + y)^{-1} \quad \text{and} \quad g_1'(y) = -(a_1 + y)^{-2}, \]
\[ g_2(y) = 0.5(a_2 + y)^{-0.5} \quad \text{and} \quad g_2'(y) = -0.25(a_2 + y)^{-3/2}, \]
\[ g_3(y) = -\beta a^\gamma y^{\beta - 1} \text{sign}(y), \]
\[ g_3'(y) = -\beta a^\gamma [y^{\beta - 2} + \beta(y - 2)y^{\beta - 3}]. \]

In order to avoid singularity of derivatives of \( G_1(y) \) at \( y = 0 \), \( y \) is replaced by very small \((10^{-2})\) non-zero numeric value. The parameters of GGD are estimated using the maximum likelihood approach, as is described in [7].

In order to evaluate signal separation performance, Noise Reduction Rate (NRR), Spectral Correlation Coefficient (SCRF) \( \gamma(f) \), and number of iterations taken to reach convergence under the given stopping criterion, defined as \( \delta = |w_{old} - w_{new}|^2 \), for the learning rule in Eq. (11) will be used. NRR is defined as the ratio of the speech signal power (computed from the reference signals) to the noise power and can be formulated as the ratio of the reference signal power to the residual noise power in the estimated IC as

\[ NRR = 10 \log_{10} \frac{\text{reference signal power}}{(\text{Estimated} - \text{reference})\text{signal power}}. \]

SCRF between ICs \( Y_1(f) \) and \( Y_2(f) \) in a frequency bin \( f \) is given by

\[ \gamma(f) = \frac{\sum_{i=1}^{m} |Y_1(f) - \bar{Y}_1(f)|Y_2(f) - \bar{Y}_2(f)|}{\sqrt{\sum_{i=1}^{m} |Y_1(f) - \bar{Y}_1(f)|^2 \sum_{i=1}^{m} |Y_2(f) - \bar{Y}_2(f)|^2}}. \]

The number of iterations consumed and the nature of convergence to reach the final solution in iterative Eq. (11) for a given \( \delta \) depend on the existence of higher order derivatives of the non linear function \( G(y) \) used. It can be shown that the values of samples, other than one non-zero element, in separation vector \( q = H^T(f)Q'(f)w \), denote by \( q' \), after one iteration is given by [16]

\[ q_i' = \frac{1}{2}E[s^2]E[q''(s_1)]q_i^3 + \frac{1}{6}\text{kurt}(s_1)E[q''(s_1)]q_i^5 + \ldots, \]

for \( i > 1 \).

This equation includes higher order derivatives of \( G(y) \) as the coefficient of error terms. It can be imbued that if the 3rd-order derivative \( g''(s_1) \) of \( G(y) \) vanishes, i.e., \( E[q''(s_1)] = 0 \), the convergence becomes cubic and is governed by the value of the 4th-order derivative \( g'''(s_1) \), and so on. The 3rd and 4th-order derivatives of the nonlinear functions used are given by

\[ g_1''(y) = 2(a_1 + y)^{-3}, \quad g_1'''(y) = -6(a_1 + y)^{-4}, \]
\[ g_2''(y) = 3.8(a_1 + y)^{-2.5}, \quad g_2'''(y) = -95(a_1 + y)^{-3.5}, \]
\[ g_3''(y) = K_1 \left[ y^{\beta - 3} + 2y^{\beta - 5} + (\beta - 4)y^{\beta - 4} \text{sign}(y) \right], \]
\[ g_3'''(y) = K_1 \left[ y^{\beta - 2} + (\beta - 4)y^{\beta - 3} \text{sign}(y) + 2y^{\beta - 6} + 2(y - 6)y^{\beta - 6} \text{sign}(y) + 3(\beta - 4)y^{\beta - 4} + (\beta - 6)y^{\beta - 5} \text{sign}(y) \right], \]

where \( K_1 = -\beta(\beta - 2)a^\gamma \). It is important to note that for \( \beta = 2 \), which corresponds to the kurtosis based contrast function, the third order derivative in Eq. (32) vanishes for whitened signals and convergence of algorithm is always cubic.

4. Experiments and Results

The layout of the experimental room is shown in Fig. 4. In the experiment, a two-element linear microphone array with interelement spacing of 4 cm was used. Voices of two male and two female speakers (sampled at 8 kHz) [20], at the distance 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 convolute mixing model.
Fig. 4 Layout of the experimental setup.

Table 1 Signal analysis conditions.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Freq.</td>
<td>8000 Hz</td>
</tr>
<tr>
<td>Frame Length</td>
<td>20 ms</td>
</tr>
<tr>
<td>Step Size (ε)</td>
<td>10 ms</td>
</tr>
<tr>
<td>Analysis Window</td>
<td>Hanning</td>
</tr>
<tr>
<td>FFT length</td>
<td>512</td>
</tr>
<tr>
<td>δ</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**Fig. 5** Averaged SE ($J^SE(f)$) for different $G_i(y)$ in different frequency bins (Averaged over 12 speech signals).

for different Reverberation Times (RTs), e.g., RT=0 ms, RT=150 ms and RT=300 ms. Mixed signals at each microphone were obtained by adding speech signals $ref_{11}$, $ref_{12}$, $ref_{21}$, and $ref_{22}$. The speech signals $ref_{11}$, $ref_{12}$, $ref_{21}$, and $ref_{22}$, that 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 different RTs, e.g., RT=0 ms, RT=150 ms and RT=300 ms. The experiments were carried out in two separate parts for jackknifing and blind separation. The TFSS of the speech data were generated by STFT analysis of the mixed signals under the signal analysis conditions shown in Table 1. In order to estimate bias and standard error occurring in approximation of negentropy by $G_k(y)$ ($k = 1, 2, 3$), 12 sets of unmixed speech signals from different male and female speakers were used. In this analysis, unmixed signals were used because, in the separation algorithm, $G_k(y)$ must ultimately approximate negentropy of the separated signal which ideally should be clean and unmixed. The bias and standard error in the negentropy approximation by each $G_k(y)$ were estimated in each frequency bin using Eq. (19) and Eq. (21) for the sequential delete— one jackknife method. The estimated standard errors, averaged over 12 combinations of the speech signals, are compared in Fig. 5 for each $G_k(y)$. The averaged estimate of bias $J^SE(f)$, for different nonlinear functions are shown in Fig. 6. It is evident from these figures that the standard error and bias is minimum for the GGD based nonlinear function, which implies that its robustness and closeness to true negentropy of the TFSS signal is better than that approximated by $G_1(y)$ and $G_2(y)$. In the Jackknifing process the value of shape parameter $β$ was fixed at 0.86 in light of results reported in [7]. The separation performances of the fixed-point FDICA with the use of these three nonlinear functions were also studied under different RTs. The stopping criterion for algorithms was set at $δ = |w_{new} - w_{old}|^2 < 0.0001$. First of all, the separation performance for different values of $β$ with nonlinear function of Eq. (17) were studied. NRR, which is defined in Eq. (27), SCRF $γ(f)$ of Eq. (28), and number of iterations required to converge for satisfaction of $δ$ were used as the performance measures. The learning rule of Eq. (11) was initialized using the null-beam former based value of the separation vector [9]. The results of NRR, $γ(f)$, and number of iterations, averaged over the six combinations of mixed speech data are shown in Fig. 7, Figs. 8 and 9, respectively.
In Fig. 7 the separation performance is found to be optimum for $\beta$ values between 0.8 and 0.98 for the reverberant and nonreverberant acoustical conditions, however, NRR is very low under the reverberant conditions. A similar trend can be observed in Fig. 8 for SCRF plots. It is important to point out that NRR and SCRF performance figures are good for the values of $\beta$ that are approximately equal to the values of shape parameters corresponding to PDF modeling of TFSS [7]. This is supports the fact that to achieve optimal separation, the nonlinear function used should be in possession of maximum statistical information about the data [5]. It is evident from Fig. 9 that the required number of iteration is also varying with the shape parameter. For the moderately lower values of the shape parameter the number of iterations taken increases highly but with the increasing value of $\beta$, number of iterations required decreases. The result is interesting in the sense that for $\beta=2$, corresponding to GD, number of iterations consumed by algorithm is lower than in the case of $\beta$ values representing the PDF of TFSS. However, achieved NRR becomes low. Under this condition, the contrast function $G_3(y)$ is the same as that of kurtosis. Thus the kurtosis-based contrast function provides speedy convergence but inferior separation quality. The speedy convergence is due to the fact that convergence is always cubic as noted before. The reason behind change in number of consumed iterations with change in $\beta$ can be understood with the help of Eq. (29), which shows how the fixed-point algorithm converges to the optimum separation vector. The coefficients of the 2nd and the 3rd terms of Eq. (29) have been plotted in Fig. 10 for different values of $\beta$ for nonlinear function $G_3(y)$, as well as for data for which the shape parameter has been denoted as $\beta_n$. The data used were artificially generated by fitting GGD parameters with zero mean and unit variance. It is evident that the annihilations of third-order and higher order derivatives are start earlier and faster for larger values of the shape parameters. This ensures higher convergence speed with increasing $\beta$. It is also evident from Fig. 10 that for the less Gaussianized signals, the number of iterations required for fixed $\beta$ is large. This result supports the results of the previous investigations [13], [14] where we showed that due to disobedience of CLT in some frequency bins, large number of iterations was consumed but separation performance was still poor. Further experiments were carried out to compare the performance of function $G_3(y)$ and that of the Natural Gradient (NG) based algorithm reported in [21]. The learning curves shown in Fig. 11 indicate superiority of GGD-based nonlinear function $G_3(y)$ over $G_1(y)$ and $G_2(y)$. In comparison to the NG algorithm, $G_3(y)$ is faster but the NRR achieved is slightly lower. The NRR values shown in Fig. 11 are averaged over both the separated sources. It is important to note here that in the deflationary algorithm separation performance decreases with increasing number of extracted sources. This happens due to the orthogonalization process after each iteration [15], [16].

The same learning curves for each source are shown in Fig. 12 for RT=150 ms and in Fig. 13 for RT=300 ms. It is evident from these two figures that the performance, in both convergence speed and achieved NRR, of $G_3(y)$ for first source is not only superior to those of $G_1(y)$ and $G_2(y)$ but also to that of NG algorithm. However, for the sec-
Fig. 11  Comparison of learning curves for nonlinear function $G_1(y)$ and the NG learning rule for RT=150 ms and RT=300 ms. Results are averaged over both extracted sources from 12 combinations of the mixed speech signals. $\beta$ for nonlinear function was fixed at 0.9.

Fig. 12  Comparison of learning curves for first source and second source for RT=150 ms. Results have been averaged over 12 combinations of mixed speech signals. NRR performances for the NG algorithm are almost the same for both sources, so averaged NRR has been plotted in both figures. $\beta$ for nonlinear function was fixed at 0.9.

Fig. 13  Learning curves for first source and second source with RT=300 ms. Results have been averaged for 12 combinations of mixed speech signals. NRR performances for the NG algorithm are almost the same for both sources, so averaged NRR has been plotted in both figures. $\beta$ for nonlinear function was fixed at 0.9.

Fig. 14  Averaged (over 6 pairs) NRR for different $G(y)$ under different RT.

signal separation; however, to the best of our knowledge, no analytical reasoning for it has been reported yet.

In order to compare the separation performance of all three nonlinear functions, another experiment was performed by changing the nonlinear functions in learning Eq. (11). The value of the parameter of the GGD function is estimated after each iteration, however, the shape parameter was fixed at $\beta = 0.9$ following the above results. The algorithm was initialized by the null-beam-former-based initial values of the separation vectors. The averaged NRR and number of iterations required, for 6 combinations of mixed signals are plotted in Fig. 14 and Fig. 15, respectively. It is evident from these figures that there is no significant difference in the achieved NRR. However, a significant difference occurs in the number of iterations required by different nonlinear functions. In this respect, the GGD based nonlinear function significantly outperforms the other two under both the reverberant and non-reverberant conditions. The GGD-based function shows higher convergence speed because the third order derivative and 4th-order derivative for it are much
5. Conclusions

In this paper, we presented the details of approximation of negentropy by HOS of different nonlinear functions to enable their effective use in fixed-point FDICA. We also proposed a new nonlinear function based on the statistical modeling of TFSS by GGD functions. The proposed nonlinear function is adaptive in the sense that it depends on the parameters of the data and accordingly provides nonlinear behavior. The signal separation performances of conventional and proposed nonlinear functions were also compared with those of the NG algorithm under different acoustic conditions. Favorable results were obtained for the proposed nonlinear functions. It can be concluded that as the GGD function can better represent statistical model of TFSS, the GGD based nonlinear function can incorporate much information about HOS of the TFSS. Due to this it provides better results than the conventional nonlinear functions.

Acknowledgements

The first author acknowledges the financial assistance of MONBUKAGAKUSHO, Govt. of Japan, in the form of Doctoral fellowship. Authors are also thankful to anonymous reviewers for improving the quality of paper in every way.

References


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 Junior Research Fellowship (JRF) and eligibility for Lectureship from UGC-CSIR, Govt. of India in 1995 and joined University Dept. of Electronics BRA Bihar University, Muzaffarpur, India as a Lecturer in 1996. He received Japan Government fellowship 2000. In March, 2005 he earned degree of D.Eng from Nara Institute of Science and Technology, Japan. His research interests include expert system development, voice-active robots, speech signal processing, and blind source separation. He is life-member of IEETE, New Delhi, India.

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 and the Acoustical Society of Japan.

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 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.