APPROXIMATING NEGENTROPY OF TIME-FREQUENCY SERIES OF SPEECH FOR FIXED POINT ICA BY NEGENTROPY MAXIMIZATION

Rajkishore Prasad[1], Hiroshi Saruwatari[2], Kiyohiro Shikano[2]
1. PG Deptt. of Electronics, BRA Bihar University, Muzaffarpur, Bihar, India. 2.Nara Institute of Science and Technology, Takayama-cho, Nara, Japan.

ABSTRACT

This paper presents study on the approximation of negentropy of Time-Frequency Series of Speech (TFSS) using generalized Higher Order Statistics (HOS) of the non-quadratic non-linear functions in light of their efficient usability in the frequency domain blind signal separation algorithms for the separation of convolutive mixture of speech. We also propose a new non-linear function based on the statistical modeling of TFSS by exponential power functions. The results of standard error and bias, estimated using sequential delete-one Jackknifing method, in the approximation of negentropy of TFSS by different non-linear functions along with their signal separation performances show the superlative power of the exponential power based non-linear function.

1. INTRODUCTION

The techniques of Blind Signal Separation (BSS) have emerged as one of the potential solutions for the extraction or segregation of hidden signals only from their observed mixtures [1]. Because the method is blind and unsupervised in functioning, it has gained wide areas of applicability. 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. This has further pivotal implication in the creation of capability of steering hearing attention, similar to anthropomorphic ability known as Cocktail party effect, in the artificial audition systems. The problem of BSS, in general, can be mathematically formulated as the estimation of R latent signals $s(n)=[s_1(n),s_2(n).....s_R(n)]^T$ only from their M mixed versions, $x(n)=[x_1(n),x_2(n).....x_M(n)]^T$. The mixed signals are produced by some unknown interactions F among them as follows

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

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.

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

In the simplest case the mixing process F produces instantaneous mixture; however, in this paper we will consider the case of convolutive mixing. The Frequency Domain ICA (FDICA) works on the Time Frequency Series of Speech (TFSS) and separates signals independently in each frequency bin [2]. The FDICA algorithms are based on joint or marginal distribution of the signal. Our concern in this paper is with the FDICA algorithms based on the marginal distribution which looks for Independent Component (IC) as the maximally non-Gaussian components in the mixed signal. One of the most successful algorithms in this family is the fixed-point ICA by negentropy maximization [3] in which negentropy of the data, approximated using generalized Higher Order Statistics (HOS) of the non-quadratic non-linear function, is used as a measure of non-Gaussianity. In this paper we examine performance of such conventional non-quadratic non-linear function for the TFSS and propose a new non-linear function based on the approximation of Probability Density Function (PDF) of TFSS by the Generalized Gaussian Distribution (GGD) function.

Rest of this paper is organized as follows. In the section II mixing and demixing model is presented. Section III deals with fixed point FDICA and approximation of negentropy. Section IV deals with the experimental results that is followed by conclusions 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 convolutive mixture of the impinging source signals and impulse response between 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

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} ref_{11}+ref_{12} \\ ref_{21}+ref_{22} \end{bmatrix}, \qquad (3)$$

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

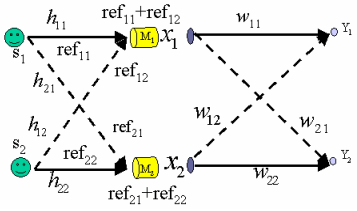

Fig.1. Convolutive mixing and demixing models for speech signal at a linear microphone array.

each frequency bin independently, and this separation process in any frequency bin *f* is given by

$$\begin{bmatrix} \hat{S}_1(f) \\ \hat{S}_2(f) \end{bmatrix} = \begin{bmatrix} Y_1(f) \\ Y_2(f) \end{bmatrix} = W(f)X(f) = \begin{bmatrix} W_{11}(f) & W_{12}(f) \\ W_{21}(f) & W_{22}(f) \end{bmatrix} \begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} \quad (4)$$

where $[Y_1(f),\ Y_2(f)]^T$ are TFSS of ICs ; and $W(f) =$ separation matrix in frequency bin *f*.

2. FIXED-POINT FDICA

The fixed-point algorithm for the ICA, known for its faster convergence speed, based on the negentropy maximization was proposed by Hyverinen for the real valued signals [3]. The FDICA algorithms for the separation of convoluted mixture work on the complex valued TFSS of the mixed speech data to sieve out TFSS of the ICs in each frequency bin. So the algorithm proposed in [3] is not directly applicable for FDICA, however, extension of the same for complex valued data can be applied for FDICA for the speech signal separation [4, 5, 6]. The functioning of fixed-point ICA algorithm is based on Central Limit Theorem (CLT) which states that the mixing of plural number of non-Gaussian signals results in increase in the Gaussianity of the mixed signal and thus its non-Gaussianization can yield independent components. TFSS of mixed signal 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 sources. Obviously, non-Gaussianization of TFSS can give TFSS of independent sources from which original signals can be reconstructed. In the fixed-point ICA the process of non-Gaussianization consists of two-steps namely, pre-whitening or sphering and rotation of the observed signal. Sphering is half of the ICA task and gives spatially decorrelated signals $\mathbf{X}_W(f,t) = [\mathrm{X}_{1W}(f,t), \mathrm{X}_{1W}(f,t)]^T$. The whitened signal in the *f*th frequency bin is obtained using Mahalanobis transform as follows [7]

$$\mathbf{X}_W(f,t) = Q(f)\mathbf{X}(f,t), \quad (5)$$

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 $\mathbf{X}(f,t)$ and V_x is the orthogonal matrix consisting of their eigenvectors. The whitened signal vector $\mathbf{X}_W(f,t)$ is then rotated by the separation matrix such that $\mathbf{Y}(f) = W(f)\mathbf{X}_w(f,t)$ equals independent components. The appropriate separation matrix is learned from the whitened data by optimizing some cost function used to measure degree of non-Gaussianity.

2.1. Negentropy approximation:

As a measure of non-Gaussianity, negentropy provides better performance as explained in [7]. The term negentropy actually represents negative of entropy. The negentropy *J(y)* of the TFSS of the random variable *y(=**Y**(f))*, is given by

$$J(y) = H(y_{gauss}) - H(y), \quad (6)$$

where H(.) is the differential entropy of (.) and 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 y is Gaussian and will be increasing if y is becoming non-Gaussian. This implies that it is always positive. Thus negentropy based contrast function can be maximized to obtain optimally non-Gaussian component. However, estimation of true negentropy, as in Eq.(6), is difficult and it requires knowledge of probability distribution function of the data. Thus several approximations for negentropy estimation have been used and proposed. As a very raw, loose and rough approximation, kurtosis has been used for it [8]. The other approximation has been based on the generalized Higher Order Statistics (HOS) which uses some non-linear non-quadratic functions *G*. In terms of such function the most widely used approximation of negentropy is given by [7]

$$J(y) = \sigma[E\{G(y) - E\{G(y_{gauss})\}]^2, \quad (7)$$

where σ is a positive constant and y_{gauss} is a Gaussian random variable with same covariace as that of *y*. The optimally non-Gaussian component can be obtained by maximizing Eq.(7) for $y = |\mathbf{w}^H\mathbf{X}_w|^2$ as the complex valued samples of TFSS of speech spherically symmetric. The one unit algorithm for learning the separation vector $\mathbf{w}$ (any row of the of the separation matrix ***W(f)***) is given by [4,5]

$$\begin{aligned} \mathbf{w}_{new} = \mathbf{w}(E\{g(|\mathbf{w}^H\mathbf{X}_w|^2) + (|\mathbf{w}^H\mathbf{X}_w|^2)g'(|\mathbf{w}^H\mathbf{X}_w|^2)\}) \\ -E\{g(|\mathbf{w}^H\mathbf{X}_w|^2)(\mathbf{X}_w^H\mathbf{w})\mathbf{X}_w\}, \end{aligned} \quad (8)$$

where first and 2nd-order derivatives of *G(y)*, have been denoted by *g(y)* and *g'(y)*, respectively.

The performance of the fixed-point algorithm depends on the used non-quadratic non-linear function *G*. It is desirable that the function G should provide robustness toward outlier values in the data and should provide better approximation to true negentropy. Better robustness to outliers can be ensured by choosing G with slow variation with respect to change in data and at the same time very close approximation of negentropy can be expected if statistical characteristics of *G* inherits PDF of the data. The statistically efficient and optimal *G* that can accommodate maximum information about HOS of the data is chosen as the function that can minimize trace of the asymptotic variance of $\mathbf{w}$ and can be approximated by [7]

$$G = c_1 \log p(y), \quad (9)$$

where c_1 is an arbitrary constant and *p(Y)* represents PDF of *Y*. Keeping in view these facts many non-linear functions have been proposed for G. However, for the super-Gaussian signals following functions has been recommended [7] and have been used in the speech signal separation [5, 6]

$$G_1(y) = \log(a_1 + y); a_1 = 0.01, \quad (10)$$

$$G_2(y) = \sqrt{a_2 + y}\ ; a_2 = 0.01. \quad (11)$$

In [9] extensive studies have been made on the PDF of TFSS and has been shown that GGD function can provide better approximation to PDF of the TFSS signal. The GGD function is a parametric function defined in terms of location parameter μ, scale parameter α, and shape parameter β. The GGD PDF for a random variable z is given by

$$f_{GG}(z;\mu,\alpha,\beta) = A\exp\left(-[|z-\mu|/\alpha]^{\beta}\right), \quad (12)$$

where, $A = \dfrac{\beta}{2\Gamma\alpha(1/\beta)}$, $\dfrac{1}{\alpha} = \dfrac{1}{\sigma}\sqrt{\dfrac{\Gamma(3/\beta)}{\Gamma(1/\beta)}}$; σ=Stdv.

$\Gamma(x) = \int_0^{\infty} e^{-t}t^{x-1}dt$=Gamma PDF; $-\infty<z<\infty$; $\alpha>0$; $\beta>0$;

The optimal function $G_3(y)$ based on the GGD can be obtained by using Eq.(12) in Eq. (9) and is given by

$$G_3 = \alpha^{-\beta}|y|^{\beta} + \log A. \quad (13)$$

3. JACKKNIFE METHOD FOR ERROR ESTIMATION

In order to judge the relative suitability of these non-linear functions we will evaluate their performances for the negentropy approximation and robustness to outliers, and capacity of signal separation. The statistical technique of Jackknifing can be used to evaluate relative error in the approximation of negentropy and robustness to outliers [10]. Jackknife is one of the powerful tools for the data partitioning and can be used to estimate bias and standard error occurring in the negentropy approximation by non-linear functions $G_k(for\ k = 1,2,3)$ from Jackknife replicates. The Jackknife replicates for the negentropy are obtained by approximating negentropy of Jackknife samples which are created by omitting, in turn, one data sample from the original TFSS. Let us consider the TFSS in any frequency bin f consisting of U samples. The ith Jackknife replicate for negentropy approximation by function G_k is given by

$$J_k^{(-i)}(f) = G_k([y(f,1), y(f,2)...y(f,i-1), y(f,i+1)...y(f,U)]), \quad (14)$$

and this is carried out independently in each frequency bin for each samples. The bias $J_k^B(f)$ in the negentropy approximation by function G_k is given by

$$J_k^B(f) = (N-1)\{\bar{J}_k^T(f) - J_k(f)\}, \quad (15)$$

$$\text{where } \bar{J}_k^T(f) = \frac{1}{N}\sum_{i=1}^{R} J_k^{(-i)}(f). \quad (16)$$

The standard error in negentropy approximation by G_k is given by

$$\hat{J}_k^{SE}(f) = \left[\frac{(N-1)}{N}\sum_{i=1}^{R}\{J_k^{(-i)}(f) - \bar{J}_k^T(f)\}^2\right]^{0.5}. \quad (17)$$

This represents standard deviation of the Jackknife replication, however, it is unbiased due to the presence of factor *(N-1)/N* [11]. Since TFSS in each frequency bin are assumed to be independent the above estimates for bias and standard error can be averaged over the no. of frequency bins and can be given by

$$\bar{J}_k^B = \frac{2}{P}\sum_{i=1}^{P/2} J_k^B(f) \text{ and } \bar{J}_k^{SE} = \frac{2}{P}\frac{2}{P}\sum_{i=1}^{P/2}\hat{J}_k^{SE}(f). \quad (18)$$

The separation performance of each non-linear functions will be judged using the deflationary learning rule given in Eq.(8). Obviously that requires first and second order derivatives of the non-quadratic functions G which are given by

$$g_1(y) = (a_1 + y)^{-1} \text{ and } g_1'(y) = -(a_1 + y)^{-2}, \quad (19)$$

$$g_2(y) = 0.5(a_2 + y)^{-0.5} \quad g_2'(y) = -0.25(a_2 + y)^{-3/2}, \quad (20)$$

$$g_3(y) = -\beta\alpha^{-\beta}\left[|y|^{\beta-1}\, sign(y)\right], \quad (21)$$

$$g_3'(y) = -\beta\alpha^{-\beta}\left[|y|^{\beta-2} + y^2(\beta-2)|y|^{\beta-4}\right]. \quad (22)$$

In order to avoid singularity of derivatives of $G_3(y)$ at y=0, it is replaced by very small (10^{-4}) number. The parameters of GGD are estimated using maximum likelihood approach as is described in [9].

4. EXPERIMENTS AND RESULTS

In the experiment a two element linear microphone array with inter-element spacing of 4 cm was used. Voices of two male and two female speakers (sampled at 8kHz) [12], 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 convolute mixing model for different Reverberation Time (RT), e.g., RT=0 ms, RT=150 ms and RT=300 ms. The experiments were carried out in two parts separately for the jackknifing and blind separation. The TFSS of the speech data were generated by doing P(=512)-point STFT analysis of hanning windowed segments of 20 ms with 50% overlapping. In order to estimate bias and standard error occurring in negentropy approximation by G_k six unmixed speech signals, as in the

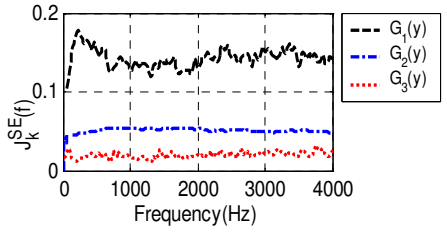

Fig. 2. Averaged SE ($\hat{J}_k^{SE}(f)$) for different G(y).

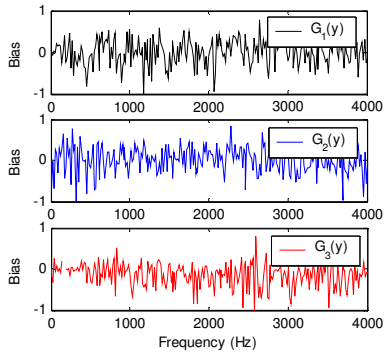

Fig. 3. Normalized bias $J_k^B(f)$ for different G(y).

separation algorithm G_k ultimately approximates negentropy of the separated signal, from different speakers were used. The bias and standard error in the negentropy approximation by each of G_k were estimated in each frequency bin using Eq.(15) and Eq.(17) for sequential delete–one Jackknife method. The estimated standard error, averaged for 6 speech signals including male and female speakers are compared in Fig.2 for each G_k. The averaged bias estimate $J_k^B(f)$, for different non-linear functions are shown in the Fig.3. It is evident from these figures that the standard error and bias is minimum for the GGD based non-linear function which implies that its robustness and closeness to true negentropy of the TFSS signal is better than that of G2(y) and G1(y). The separation performances of the fixed-point FDICA with the use of these three non-linearities were also studied under different RTs. The stopping criterion for algorithms was set at $\delta = |w_{new} - w_{old}|^2 < .0001$. The Noise Reduction Rate (NRR), which is defined as the ratio of signal power and power of residual interference in the separated signals, and no. of iteration taken to converge up to δ were used as the performance measures. The learning rules of Eq.(8) was initialized using null-beam former based value of the separation vector [5]. The results of NRR and number of iterations, averaged for the combination of 12 pair of mixed speech data are shown in Fig.4 and Fig.5 respectively. The value of parameter of the GGD function is estimated after each iteration; however, the shape parameter was fixed to $\beta = 0.9$ following the results reported in [9]. It is evident from these figures that there occurs no significant difference in the achieved NRR, however, significant difference occurs in the number of iterations consumed by different non-linear functions. In this respect, the GGD based non-linear function outperforms the other two with a handsome margin.

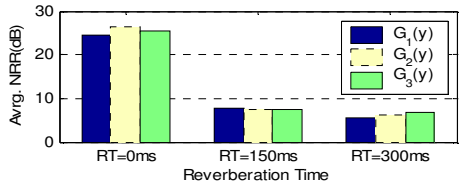

Fig. 4. Averaged(for 12 pairs) NRR for different G(y) under different RT.

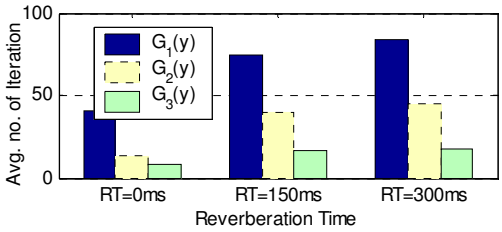

Fig.5. Averaged (for 12 pairs) NRR for different G(y) under different RT.

5. COCLUSIONS

It can be concluded that as the GGD function can better represent statistical model of TFSS, the GGD based non-linear function can incorporates much information about HOS of the TFSS. Due to this it provides better results than the conventional non-linear functions. It is further needed to explore the separation performance of the algorithm for different values β and why separation performances are poor for higher RTs.

6. REFERENCES

[1] P. Comon "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287-314, 1994.

[2] P. Samaragadis, "Blind separation of convolved mixture in the frequency domain," Neurocomputing, vol.22, pp.21-34. 1998.

[3] Aapo Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks* 10(3):626-634, 1999.

[4] 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.

[5] R.K Prasad., H.Saruwatari, A.Lee, K.Shikano, "A fixed point ICA algorithm for convoluted speech separation", *Proc. International Symposium on ICA &BSS,* pp-579-584, Nara, Japan, 2003.

[6] 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).

[7] A. Hyvarinen et al., "*Independent Component Analysis*," John Wiley & Sons, 2001.

[8] S. Haykin ,ed., " *Unsupervised Adaptive Filtering, Vo. 1: Blind Source Separation,l*" John Wiley and Sons,2000.

[9] R.K Prasad., H. Saruwatari, A. Lee, K.Shikano, "Probability distribution of Time-Series of Speech Spectral Components," *IEICE Trans. Fundamental of Elec.,Com., CS*., Vol.E87- A,No.3, 584-597,2004.

[10] J. Shao , D.Tu, *"The Jackknife and Bootsrap,"* Springer, New York,1995.

[11] B. Efron , R.J. Tibshirani, " *An Introduction to Bootstrap*," Chapman and Hall, London, 1993.

[12] T.Kobayashi, S.Itabashi, S.Hayashi, and T. Takewaza, " ASJ continuous corpus for research," *J. Acoustic Soc. Jpn*., vol.48, no.12, pp-888-893, 1992.