PROBLEMS IN BLIND SEPARATION OF CONVOLUTIVE SPEECH MIXTURES BY NEGENTROPY MAXIMIZATION

Rajkishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN
E-mail: {kishor-p, sawatari, shikano}@is.aist-nara.ac.jp

ABSTRACT: This paper aims to examine suitability of the marginal statistics based contrast function e.g. negentropy for the separation of convolutive speech mixtures picked up by a linear microphone array. For this study we choose our frequency domain fixed-point ICA algorithm, based on negentropy maximization of the independent components. This algorithm is based on the heuristic assumption, in accordance with the Central Limit Theorem (CLT), that the gaussianity of mixed speech signal is more than that of unmixed individual. This assumption is true for long segment of speech in the time domain and the same is expected to hold even for small segments of speech in time domain and for every spectral bin for frequency sub-banded speech. In this paper we examine this assumption by estimating spectral kurtosis on the frequency time series of the signal obtained by taking Discrete Fourier Transform (DFT) of quasi-stationary segments of speech. It has been found that in more than 35% of the frequency bins, speech signal fails to comply the CLT assumption, which in turn badly affects the separation performance of the fixed-point algorithm.

1. INTRODUCTION

Blind signal separation (BSS), a very hot topic of research among digital signal processing groups since a decade, is the general framework to estimate signal contribution of latent sources only from their observed mixtures without knowing the mixing process. It represents very true copy of the real world "Cocktail party" problem. Recently, several Independent Component Analysis (ICA) based algorithms, both in the time domain and in the frequency domain, under the general framework of BSS have been proposed to solve this problem. In the list of solutions for such separation algorithms fueled by the principle of statistical independence of the sources, popularly known as ICA-based BSS algorithms, have been dominating due to emergence of several successful algorithms separately in the time domain and frequency domain or combined in both [1,2]. In such algorithms, as shown in Fig.1, the observed mixed signals are passed through a tentative initial demixing system (randomly chosen or based on some heuristic guess and are subject to further modification) and then the mutual independence among the estimated Independent Component (IC) signals is evaluated by some cost function, usually based on the statistics of the signal and candidate demixing system. This in turn goes on modifying unmixing system unless and until the cost function is not optimized for the maximum mutual independence among the separated components. So, paradigmatically, most of the ICA-based BSS algorithms show such functional similarity, but basic difference occurs in the choice of the cost function, 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

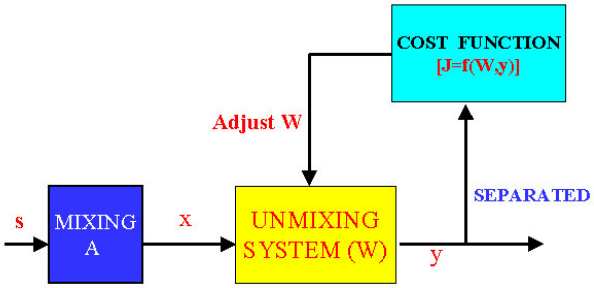


Fig.1 General framework of ICA based BSS.

Kullback-Liebler divergence metric, which measures deviation between the joint distribution of the signal and some pre-assumed source distribution. This cost function is statistically efficient and there have been development of several excellent algorithms based on them [1,2]. However, prior knowledge of source distribution is not always feasible. Second category of the cost functions exploit 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 [3]. On the same line we also proposed such an algorithm in combination with the null-beamformer for the separation of convolutive mixture of speech [4]. In this paper our aim is to check suitability of such cost function based algorithm for the convolutively mixed speech signal separation. Our proposed algorithm uses non-gaussianity based contrast function, e.g., negentropy of the signal, which is optimized by fixed-point iterative algorithm [5]. The rationale behind this study is many folds. First the algorithm works in the frequency domain on the DFT of pseudo-stationary speech segments. So it is essential to test compliance of the CLT by data in the each frequency bin. This compliance depends on the nature of statistical distribution of the spectral data. If the spectral components of speech signal have stable distribution they will fail to comply CLT and then such algorithms cannot be used to separate them.

2. SPEECH SIGNAL SEPARATION

We consider here the case of two speakers and two microphones. The signal mixing and unmixing model for this case is shown in the Fig.2.The real world mixing model is best approximated by convolution of source to sensor transfer function and source signal. Accordingly, observed signals $x_1(t)$ and $x_2(t)$ at microphones are given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \otimes \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} \text{ref}_{11}+\text{ref}_{12} \\ \text{ref}_{21}+\text{ref}_{22} \end{bmatrix}, \quad (1)$$

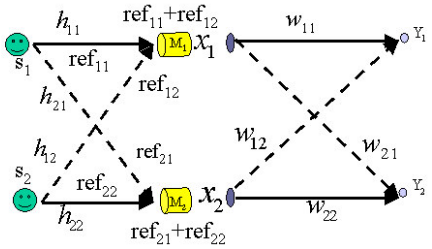


Fig.2 Convolutive mixing and demixing model for speech signal.

$$\Rightarrow x = h \otimes s,$$

where $ref_{11}=h_{11}\otimes s_1$, $ref_{12}=h_{12}\otimes s_2$; $ref_{21}=h_{21}\otimes s1$, $ref_{22}=h_{22}\otimes s_2$; $\otimes$ represents convolution.

In the frequency domain the same is represented as:

$$\begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} = \begin{bmatrix} A_{11}(f) & A_{12}(f) \\ A_{21}(f) & A_{22}(f) \end{bmatrix} \begin{bmatrix} S_1(f) \\ S_2(f) \end{bmatrix}, \quad (2)$$

$$\Rightarrow X(f) = A(f)S(f).$$

The separated ICs are given by

$$\begin{bmatrix} Y_1(f) \\ Y_2(f) \end{bmatrix} = \begin{bmatrix} W_{11}(f) & W_{12}(f) \\ W_{21}(f) & W_{22}(f) \end{bmatrix} \begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} \quad (3)$$

$$\Rightarrow Y = W(f)X(f).$$

Our BSS algorithm is based on higher order statistics of the data.It has ultimate bearings on the assumption of the CLT which says that unstable distribution gains gaussianity under linear combination. The non-gaussianity based contrast function negentropy J(y) has been used to ensure mutual independency among ICs[5]. In this algorithm sphered data X(f) in every frequency bin is used to learn orthonormal separation vector $w_1=[w_{11}\ w_{12}]^T$ and $w_2=[w_{21}\ w_{22}]^T$ in deflationary fashion, by following equation:

$$w_{new} = w(E\{g(|w^H x|^2) + (|w^H x|^2) g'(|w^H x|^2)\}) - E\{g(|w^H x|^2)(x^H w)x\}. \quad (4)$$

where g and g' represents 1st and 2nd order derivative of a non-quadratic and non-linear function G, used to approximate J(y) by

$$J(y) = \sigma[E\{G(y) - E\{G(y_{guass})\}]^2, \quad (5)$$

where σ is a positive constant.

The learning process is stopped by deciding stopping factor $\delta = w_{new} - w_{old}$. Permutation and scaling problems are resolved using directivity pattern method.

3.WHAT IS EXPECTED FROM MIXED DATA

Our fixed-point algorithm is based on the heuristic idea, motivated from CLT, that a mixed speech signal tends to be more gaussian than individual signals. So the estimated signal can represent the independent components if these are obtained by non-gaussianization of the mixed signal. Thus the power to sieve out ICs comes in the algorithm due to validity of the following logical fact that

The gaussianity of the mixed speech signal> Gaussianity of the independent constituent speech signals.

If gaussianity is measured by kurtosis then the above mentioned logical touchstone for the data may be mathematically reproduced as :

$$K_{m1}(f) < \min[K_{s1}(f), K_{s2}(f)] \quad \& \quad K_{m2}(f) < \min[K'_{s1}(f), K'_{s2}(f)] \quad (6)$$

where $K_{mn}(f)$ =Kurtosis of mixed signal at *n(=1,2)*th mic in the frequency bin f; K_{sp} & K'_{sp} =Kurtosis of *p(=1,2)th* independent source at 1st and 2nd microphones respectively. In the case if these mandatory conditions are not complied, algorithms based on marginal statistics may not be able to separate ICs. It has been found that the long segment of speech in time domain exactly comply CLT but short time segments may fail to follow CLT due to temporal correlation of the speech [6]. As mentioned above the data used in the above algorithm is frequency time series gleaned from N-point DFT of very short quasi-stationary segments of speech. Each DFT coefficient is a linear weighted sum of speech samples. So, they form complex normal distribution and thus CLT is complied [7]. The same is also expected from each spectral bin of the speech data. Following [8], it can also be shown that speech data in each frequency bin are Complex Circular Random Variable (CCRV), as they are independent of complex rotation.

4. CLT TEST IN FREQUENCY BINS

As a measure to check the obeyance of CLT condition, stated in eq.(6) which is mandatory for separation, by the spectral component of speech we used Spectral Kurtosis(SK). SK is defined as the ratio of the fourth order central moment to the second order moment given by

$$SK(f)=\frac{C_4\{S^*,S^*,S^*,S^*\}}{[C_2\{S^*,S^*\}]^2}, \quad where\ S^* \in \{X(f), X^H(f)\}. \quad (7)$$

Following [8,9] and assuming spectral component of speech as CCRV simplified expression for spectral kurtosis is given by

$$SK(f)=\frac{E\{|X(f)|^4\} - 2E^2\{|X(f)|^2\}}{[E\{|X(f)|^2\}]^2}. \quad (8)$$

As in our fixed point algorithm data are sphered so this equation further simplifies to

$$SK(f)=E\{|X(f)|^4\} - 2. \quad (9)$$

Using this expression for SK, the validity of the CLT can be tested in each frequency bin. The important requisition for CLT compliance by the spectral speech data is that it must not have stable statistical distribution because such distributions are closed under linear combination [10]. In order to examine the nature of distribution of spectral component, χ^2 tests of goodness of fit [11] for following three null hypotheses for complex valued spectral speech data have been performed (i) it follows Gaussian distribution, (ii) it follows Laplacian distribution, and (iii) it follows Generalized Gaussian distribution (GGD) with the estimated parameters. 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 number of iterations required to reach convergence in each frequency bin were measured. These performance indices are given as follows:

SNRR is defined as the ratio of signal power to noise power in a frequency bin and is given by:

$$\text{SNRR(f)}=10\log_{10}\frac{E\{|X_S(f)|^2|\}}{E\{|X_N(f)|^2|\}}. \quad (10)$$

Accordingly, SNRR for the first and the second ICs are given by

$$\text{SNRR}_1(\text{f})=10\log_{10}\frac{\text{E}\{|\text{W}_{11}(\text{f})\text{Ref}_{11}(\text{f})+\text{W}_{12}(\text{f})\text{Ref}_{21}(\text{f})|^2\}}{\text{E}\{|\text{Y}_1(\text{f})-\text{W}_{11}(\text{f})\text{Ref}_{11}(\text{f})+\text{W}_{12}(\text{f})\text{Ref}_{21}(\text{f})|^2\}}, \quad (11)$$

and

$$\text{SNRR}_2(\text{f})=10\log_{10}\frac{\text{E}\{|\text{W}_{21}(\text{f})\text{Ref}_{12}(\text{f})+\text{W}_{22}(\text{f})\text{Ref}_{22}(\text{f})|^2\}}{\text{E}\{|\text{Y}_2(\text{f})-\text{W}_{21}(\text{f})\text{Ref}_{12}(\text{f})+\text{W}_{22}(\text{f})\text{Ref}_{22}(\text{f})|^2\}}. \quad (12)$$

The SCRF in a frequency bin f is given by

$$\gamma(\text{f})=\frac{\sum_1^m[\{X_1(f)-\bar{X}_1(f)\}\{X_2(f)-\bar{X}_2(f)\}]}{\sqrt{\sum_1^m|X_1(f)-\bar{X}_1(f)|^2}\sqrt{\sum_1^m|X_2(f)-\bar{X}_2(f)|^2}}. \quad (13)$$

5. EXPERIMENT & RESULTS

In the experiments, we used simulated data for a two-element linear microphone array with inter-element spacing of 4 cm. Voices of two speakers (male and female), at the distances of 1.15 meters and from the directions of -30° and 40° are used to generate mixed signals x_1 and x_2 under the described convolutive mixing model. Mixed signals at each microphone were obtained by adding the convolved speeches ref_{11}, ref_{12}, ref_{21}, ref_{22}. These convolved speeches are obtained by convolving seed speech with room impulse response, recorded under different acoustic conditions, characterized by different reverberation times (RTs), e.g., RT=0 ms, RT=150 ms and RT=300 ms. The speech signals reaching at each microphone from each speaker are used as reference signals. CLT test was performed in each frequency bin of the mixed data w.r.t individual sources. Result is shown in Fig.3. It is evidenr from these pie–charts that spectral speech data fail to comply CLT in the every frequency bin. This raises question on distribution of the spectral component of speech. The histograms of the real and imaginary parts of the speech, as shown in Fig.4, look very spiky and strongly hints existence of Laplacianity in the distribution. The χ^2- test was performed in every frequency bin to check it. For the GGD, first parameters (mean, scaling & shape) were independently calculated using maximum likelihood approach for the real and imaginary parts [12]. Shape parameter β decides the shape of the GGD, which is shown in Fig.5 for every frequency bin. Then χ^2-test was performed independently on the real part and imaginary part of the frequency domain speech data arriving at microphone and final score was obtained by simply adding them. The χ^2-score for the gaussian, Laplacian and GGD are shown in Fig 6. The χ^2-test indicates that spectral data of speech are strongly Laplacian, as the test score is minimum for $\beta<1$ which hints strong Laplacianity in the spectral distribution. The stopping factor δ was fixed at 0.0001.Computed SNRR, CRF, γ(f) are shown in successive figures. These performance indices noticeably show poor performance of the algorithm in the CLT failure bins. The spectral component of speech does not have stable distribution, however, surprisingly it does not support CLT in the every frequency bin. This nature of speech data raises question on the efficient working of the marginal distribution based ICA algorithm for the separation of convolutively mixed speech. The number of CLT failure bins is also not strongly correlated with the sparcity of the spectral components.

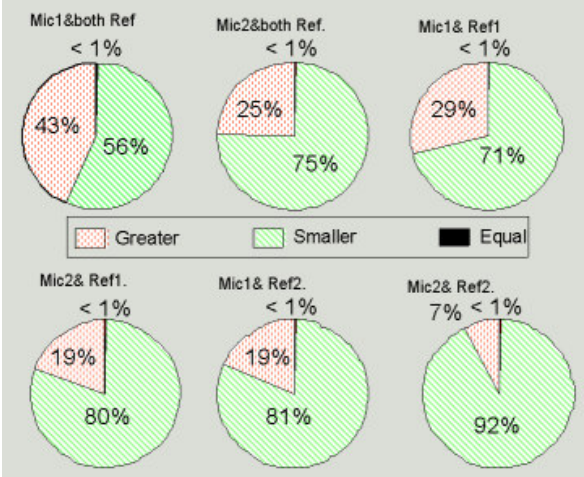


Fig.3 CLT failure bins for signals at both microphones (RT=300 ms, Voices-male and female speakers)

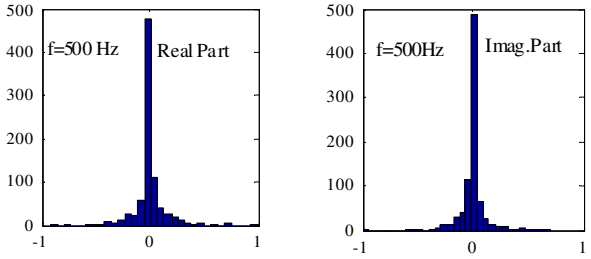


Fig.4 Histogram of real and imaginary parts of speech signa from male speaker at Mic1.

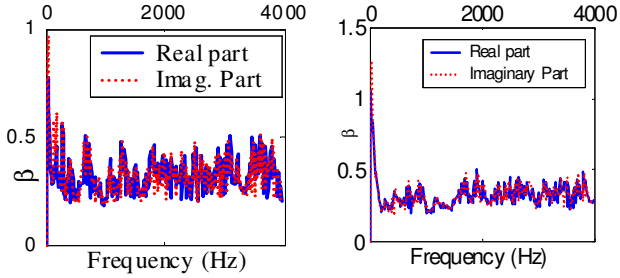


Fig.5 Estimated value of shape parameter β for individual signals at Mic1 [male(Right) and female(Left)].

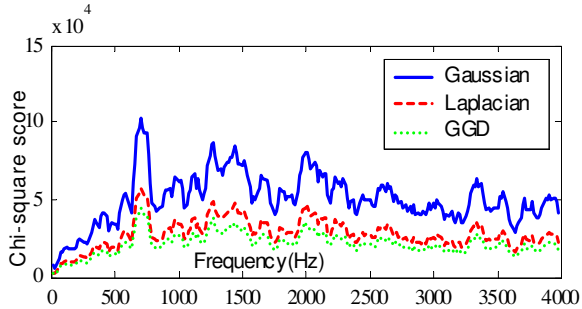
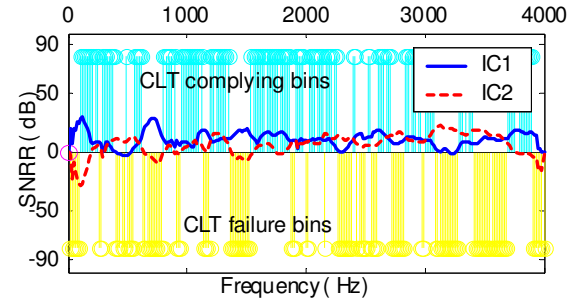


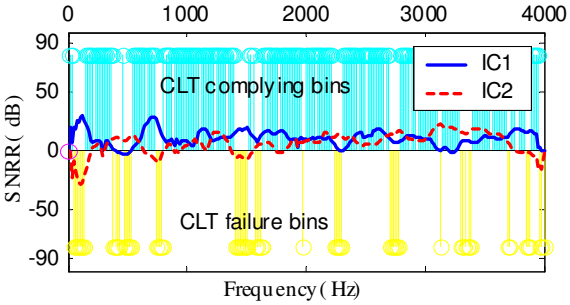
Fig.6 χ2 score for Gaussian,Laplacian and GGD for speech signal from male speaker at Mic1.

6. CONCLUSIONS AND FUTURE WORK

It is concluded from the results that speech signal fails to comply CLT in the accountable number of frequency bins, despite strong Laplacianity of the spectral components. This compels algorithm to achieve poor separation performance in such frequency bins. We are investigating some methods for blind detection of such frequency bins so that in such bins separation process can be switched over to other techniques, like beamforming, whose functioning is independent of such constraint.

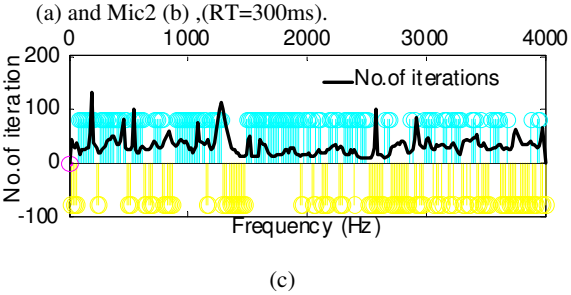


(a)

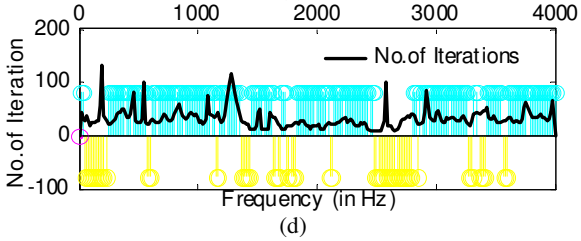


(b)

Fig.7 Spectral NRR w.r.t. CLT test for mixed signals at Mic1 (a) and Mic2 (b) ,(RT=300ms).



(c)



(d)

Fig.8 No. of iteration taken w.r.t CLT test for mixed signals at Mic1(c) and Mic2(d),(RT=300ms, speakers both male).

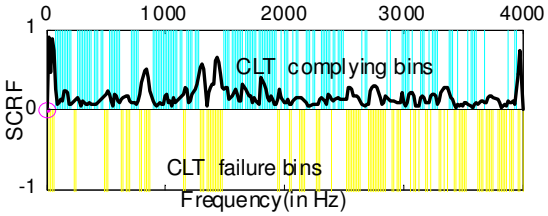


Fig.9 Spectral CRF between separated ICs w.r.t CLT test for mixed signal at Mic1(RT=300ms).

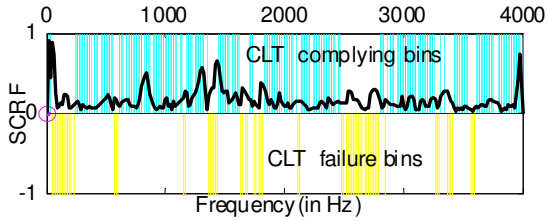


Fig.10 Spectral CRF between separated ICs w.r.t CLT test for mixed signal at Mic2(RT=300ms, speakers both male).

ACKOWLEDGEMENT

The first author expresses his gratitude with utmost respect to Monbusho, Japan for providing doctoral scholarship. We also acknowledge valuable discussion with Prof. Scott C. Douglas, of DEE, SMU, Texas, Prof. Richard M. Stern of ECE, CMU and Dr. H. Sawada from NTT. Co. Ltd., Japan for this study.

REFERENCES

1. K. Torkkola, “Blind separation for audio signals-are we there yet?” Proc. Workshop on ICA & BSS, France, 1999.
2. T.Nishikawa, et al., "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," IEICE Trans. Fundamentals, Vol.E86-A, pp846-58, No.4, April, 2003.
3. 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.
4. Prasad.R.K, 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.
5. Hyvarinen et al., “Independent component analysis,” John Wiley & Sons, 2001.
6. J.LeBlanc and P. De Leon, “Speech separation by kurtosis maximization”, Proc. ICASSP 1998, Seattle, Washington.
7. Gazor S. & Zhang W., “Speech probability distribution”, IEEE Signal Processing Letters, Oct.2002.
8. P.O. Amblard et al, “ Statistics for complex variable and signals-part I & II”, Signal Processing, vol-53,pp-1-25,1996.
9. Chrysostomos L.,et al.,“Signal processing with higher order spectra”, IEEE Signal processing magazine,pp-10-37,July1993.
10. Kismode P., “Alpha-stable distributions in signal processing of audio signals”, Proc.SIM2000, Technical University Denmark, 2000.
11. Papoulis A. & Pillai S.,“Probability, random variables and stochastic process”, McGraw Hill, 2002.
12. Varanasi M.K. et al, “Parametric generalized gaussian distribution”, J.Acoust.Am. 86(4), pp-1404-15,Oct.1989.