BLIND DETECTION OF CLT DISOBEYING FREQUENCY BINS FOR AUDIO SOURCESEPARATION BY FIXED-POINT ICA

Rajkishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Sc. & Tech, Takayama-cho, Ikoma City, Nara, Japan 6300101.
E-mail- {kishor-p, sawatari, shikano}@is.aist-nara.ac.jp

Abstract

This paper peeps into the problem of disobeyance of Central Limit Theorem (CLT) by the Time-Frequency Series of the Speech (TFSS) data. For the blind separation algorithms, based on the non-gaussianization of mixed signal e.g. signal separation by negentropy or kurtosis maximization, such behavior by TFSS is problematic in the sense that it is against the working spirit of algorithm. The detection of such CLT failure bin is important. This paper describes (explores) such failure in depth and prescribes a blind method, based on the statistics TFSS that can detect 70-75% of the CLT disobeying frequency bins. The proposed method is blind in the sense that it relies only on the statistics of observed data.

1. Introduction

Recently, application of the Independent Component Analysis (ICA) based Blind Signal Separation (BSS) techniques for the audio signal separation has fascinated much research attention. However, due to the computational complexities and poor performance, final goals are yet to be achieved [1]. For the segregation of the real world audio signal in real time, the fixed-point ICA algorithm [2] is one of the promising candidates because of faster convergence. The ICA algorithms described in [3,4,5] for the convoluted mixture of audio source separation works under the compliance of the Central Limit Theorem (CLT) by the mixed speech data, picked-up by a microphone array. Compliance to CLT implies that the gaussianity of the mixed signals is more than that of the hidden speech signals from individual sources. Thus non-gaussianisation of the mixed signal can yield individual signals. Frequency-domain fixed-point ICA by negentropy maximization is one of such algorithms which too requires obeyance of CLT by the speech data in each frequency sub-band. However, the mixed speech data in each sub-band fails to comply CLT [6]. There may be different causes for such obnoxiousness that are unexplored, to the best of our knowledge, till date. Unavailability of speech signals from all speakers at every instant of time and in every frequency bin or spectral sparseness seems one of the most appealing causes for such failure, however, at the same time other factors such as role of room acoustics and natural pauses in speech cannot be given clean chit. But in this paper we have attempted to assess the role and contribution of spectral sparseness of independent sources only. The dereliction to CLT by the speech spectral components results in the poor separation performance of the algorithm in the corresponding frequency sub-bands. If such frequency bins are detected then some alternative methods, e.g., null-beamforming can be used for such bins by stopping ICA for them. This can save computational load and improve separation performance. In this paper we also provide a statistical method for the blind detection of the CLT disobeying frequency sub-bands based on the statistics of the TFSS.
The rest of the paper is organized as follows. In the second section signal pick-up by the linear microphone is presented. In the third section we provide discussion on the CLT failure by TFSS and in the fourth section we provide method of blind detection .The last section deals with the experimental results followed by references.

2. Signal mixing and demixing model

We consider here the case of two speakers and two microphones. The signal mixing and demixing model for this case is shown in the Fig.1.The real world mixing model is best approximated by the 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

$$x = h \otimes s \Rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \otimes \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} ref_{11} + ref_{12} \\ ref_{21} + ref_{22} \end{bmatrix}, \quad (1)$$

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

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 FDICA algorithm works on TFSS $Z(\lambda, f) = [X(1, f), X(2, f) \ldots\ldots X(\lambda, f)]^t$, where λ=frame no., and separates ICs in each frequency bin independently. The TFSS data in the fixed-point ICA is first decorrelated by sphering, using pca technique and then separation filter is learned from the data using ICA algorithm in deflationary or symmetric fashion.

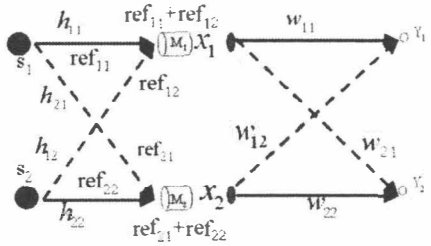

Fig.1 Convolutive mixing and demixing model for speech signal.

The separated ICs $Y_1(f)$ and $Y_2(f)$ are given by

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

where $W(f)$ represents separation matrix.

The inherent problem of permutation and scaling is solved by directivity pattern method [9].

3. CLT disobedience by TFSS

The ICA based BSS algorithm extracting ICs by non-gaussianization of the mixed signal e.g. ICA by negentropy or kurtosis maximization relies on the basic fact that the speech signal is gaussianized, in accordance with CLT, during the mixing process and thus non-gaussinisation of the mixed signal can yield non-gaussian hidden components. Our FDICA algorithm [3] for audio source works on the time-frequency series of the data and separates signal by negentropy, a measure of non-gaussianity, maximization using fixed-point iteration algorithm. Thus it is expected from the TFSS that they obey the CLT. However in our previous research [6] it was shown that the TFSS in every frequency bin does not follow CLT, i.e. in some frequency bins mixed signal does not gain gaussianity more than the individual speech signal. Spectral Kurtosis (SK), defined below in Eq. (4),

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

and for the sphered data

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

can be used to check the validity of CLT in each frequency sub- band by verifying the following criteria for the CLT compliance

$$SK_{m1}(f) < \min\{SK_{ref_{11}}(f), SK_{ref_{12}}(f)\},$$
$$SK_{m2}(f) < \min\{SK_{ref_{21}}(f), SK_{ref_{22}}(f)\}. \quad (5)$$

where SK_{mi}=SK of mixed signal at Mic. i.

Interestingly, it was reported that in about 30-40% frequency bins CLT compliance by TFSS was denied and separation performance was found deteriorated. Since TFSS is generated by the short-time Fourier transform method, it can be inferred that unless there is no big pause in speech it cannot contribute large number of samples in TFSS in any frequency bin. In the presence of moderate reverberation pause period may be modified by the reflected speech. Such reflected part increases only correlation among the samples of TFSS. Thus the spectral content of the signal remains same even under high reverberation but if there is any role of pauses in the CLT failure it will be modified by the reverberation. From Eq. (2) it is obvious that the mixed signal in any frequency bin is addition of signal contribution from each independent source in the same frequency. Thus absence of the spectral component in any frequency bin in any speakers can lead to zero contribution and mixed signal then contains contribution from either of them or no signal if both sources have zero contribution. Such spectral sparseness is one of the strongest factors that may be responsible for the CLT disobedience. Here we will provide study on such CLT disobedience with different size of DFT and in different reverberation condition.

4. Blind detection of CLT disobeying TFSS

The idea of blind detection of the CLT disobeying TFSS is based on the fact that when the speech signal get mixed its gaussianity increases. So by measuring gaussianity of the TFSS and comparing it with some threshold it can be said whether the mixed TFSS will follow CLT or not. Such measure can be derived from the statistical model of the TFSS. In [7] we have shown that the TFSS can be better modeled by the Generalized Gaussian distribution (GGD) which is parameterized by the mean, scale parameter and shape parameter (say β). As a measure of the gaussianity kurtosis is used. The kurtosis of the GGD can be given in terms of β by

$$K(\beta) = \left[\Gamma\left(\frac{5}{\beta}\right)\Gamma\left(\frac{1}{\beta}\right)\right]\left[\Gamma\left(\frac{3}{\beta}\right)^2\right]^{-1} \quad (6)$$

where $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$=Gamma distribution.

Thus if a TFSS of the mixed signal is fully gaussian its SK will corresponds to $SK_G = K(2) = 3$ in eq.(6) and if it is not mixed signal , speech will be at least Laplacian or strongly Laplacian for which SK corresponds to $SK_L = K(2) = 6$. For the strongly Laplacian case, which is more accurate as shown in [6], kurtosis will be higher than 6. The SK of TFSS can be directly computed from the data using (4). Thus if SK of TFSS, calculated from (4), lies above SK_L it will represent Laplacian or strongly Laplacian signal and TFSS will contain

contribution only from the single source. So the 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 signal and so it will comply CLT. Thus change in kurtosis can be related with the change in the shape parameter β and some threshold value of it can be used to detect CLT disobeying sub-bands. The acoustic channel too gaussianizes speech signal, so the gaussinity of true speech is less than that of received by the microphones. Thus the threshold corresponding the $\beta \leq 1$ can work well.

5. Experimental condition and results

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, at the distances of 1.15 meters and from the directions of -30° and 40° are used to generate 12 combinations of mixed signals x_1 and x_2 under the described convolutive mixing model. Mixed signals at each microphone were obtained by adding 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 the different reverberation time (RT), 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. The CLT compliance test result, in accordance with Eq.(5), performed for the speech data for the six combinations of mixed data is shown in the Fig.2. The percentage of CLT disobeying TFSS is almost independent of DFT size and there are no significant changes with the change in reverberation time. However, for increasing value of RT significant difference in the percentage of CLT failing sub-bands has been found, shown in Fig.3 and Fig.4, for both microphones. This is indicative of the fact that room acoustic is also influential in the said CLT disobedience.

As the DFT size increases the number of CLT disobeying TFSS increases, however, they remain clustered, which is due to increase in the frequency resolution. For the higher value of the DFT size failure bins are densely clustered as is shown in Fig.5. In order to explain the role of sparseness in the spectrum, the spectral content of the mixed signal and reference signals were examined in the CLT disobeying frequency bin and in the nearest CLT complying frequency bin. For that the plots of magnitude of spectral contribution from each reference signals and mixed signal were examined and one such plot is shown in the Fig.6. In that figure temporal contribution of each source in the given frequency sub-band is shown. It is evident from this figure that in the CLT failing frequency bin

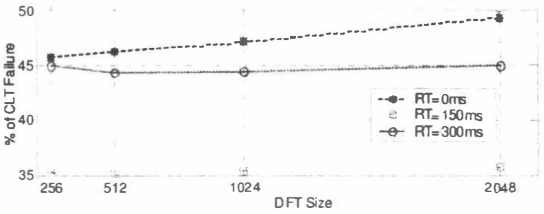

Fig.2 CLT disobeying bins for different DFT size and reverberation time at Mic1. Shown values are averaged for 6 mixed speech data.

contribution from the first speaker is not available at all instances, however, in the CLT passing bin its temporal contribution is relatively better. It is also evident that in the CLT obeying bands both sources have rich contribution but in the CLT disobeying bin either one has very rare contribution or no contribution which in accordance with Eq.(2) results in a mixed signal with content from either one sources. It is interesting to note that there have been development of ICA algorithm which exploits such temporal absence and existence of signal from different speakers for the blind source separation [8] in the anechoic environment, however, spectral sparseness is problematic for ICA by non-gaussianization and to the best of our knowledge its use in audio signal separation in the realistic environment has not been reported yet. Almost similar results have been found for the other CLT obeying and disobeying frequency bins. In order to determine threshold for the blind detection of CLT disobeying bin the relation between CLT disobeying and SK can be

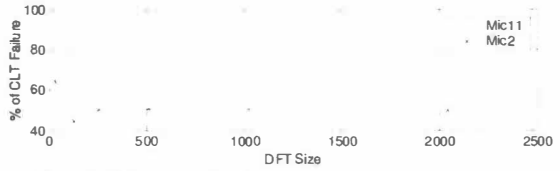

Fig.3 CLT failure at both microphones at RT=0 ms.

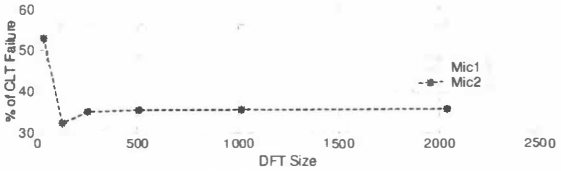

Fig.4 CLT failure at both microphones at RT=300 ms.

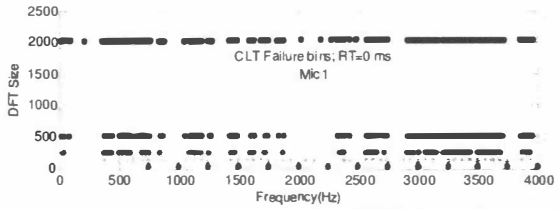

Fig 5 Clustering of CLT disobeying TFSS for different DFT size for speech signal picked-up by Mic1.

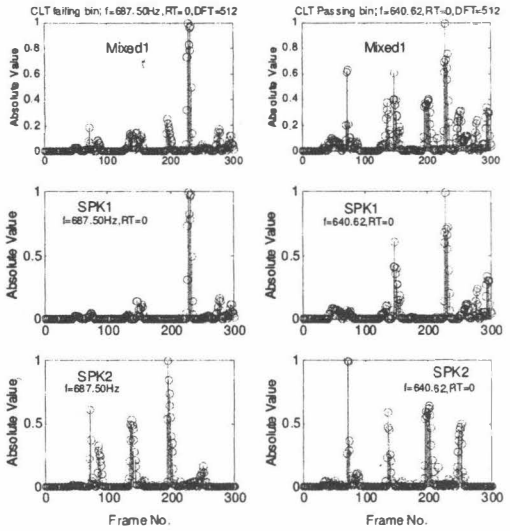

Fig.6 Role of spectral sparseness in CLT disobeyance. Left side represents CLT failing bin at f=687.50 Hz and right side represents CLT complying bin at f=640.62 Hz. DFT size=512. SPK1 and SPK2 represent plots for spectral contribution from speaker1 and speaker2 respectively.

observed in the Fig.7. The right part represents kurtosis variation of GGD with β and left part show CLT failing bins (gray colored vertical lines in the background) and plot of SK computed using Eq.(4). It is evident that SK is high for CLT disobeying bin and is relatively low for the CLT obeying bins. The dashed horizontal lines across the plots in Fig.7 show different threshold values for the different values of β. The blind detection result and true detection result are shown in the Fig.8. The term true detection represents the result obtained by the verification of the of conditions stated in Eq.(4) which needs reference signal from each speaker.

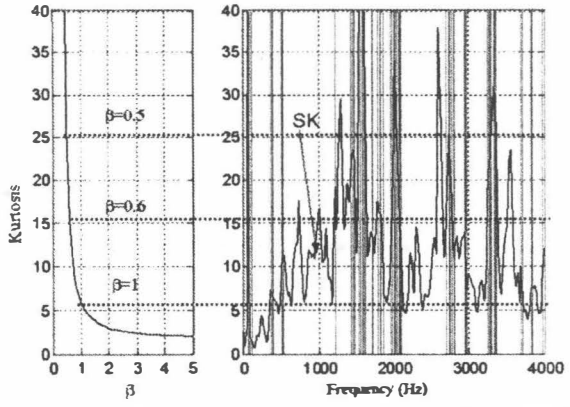

Fig.7 Threshold determination for the blind detection of CLT disobeying TFSS.

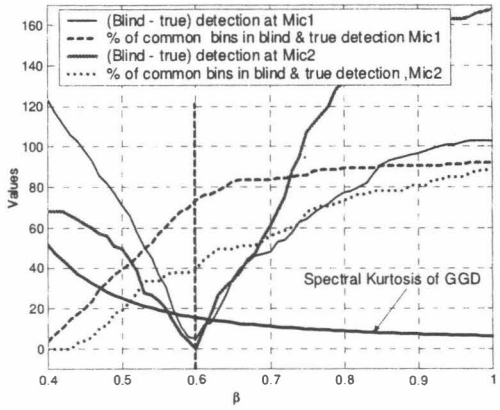

Fig. 8 Comparison of blind and true detection. Error in detection is shown by line with legend Blind-true which is minimum for β=0.6 and 70-80% bins are correctly detected.

However, in the real application these reference signals are unavailable. Plots in Fig.8 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, Blind method falsely detects some bins as CLT failing bin while giving clean chit to some really failing bins. However, for the threshold around $\beta = 0.6$, 70-80 % bins can be correctly detected. As it is evident from Fig.7 that the slight change in β produces large change in SK, so the slight change in the threshold can lead to highly affect the detection accuracy.

Acknowledgements

First author likes to express his gratitude to MONBUSHO, Japan for providing Doctoral fellowship. We also acknowledge valuable discussion with Prof. Scott C. Douglas, of DEE, SMU, Texas, 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] Hyvarinen et al., "Independent component analysis," John Wiley & Sons, 2001.

[3] 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.

[4] 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.

[5] J.LeBlanc and P. De Leon, "Speech separation by kurtosis maximization", Proc. ICASSP 1998, Seattle, Washington.

[6] Prasad. R. K, H. Saruwatari, A. Lee, K. Shikano, "Problems in blind separation of convolutive speech mixture by negenotropy maximization", Proc. IWAENC 2003, pp.287-290, Kyoto, Japan.

[7] Prasad. R. K, H. Saruwatari, A. Lee, K. Shikano, ``Probability Distribution of Time-Series of Speech Spectral Components," IEICE Trans. Fundamentals, March 2004 (in printing).

[8] Scott Rickard et al., "Real-time time-frequency based blind source separation", Proc. of ICA2001, Dec. 9-13, San Diego, CA, 2001.

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