Sindh Univ. Res. Jour. (Sci. Ser.) Vol.46 (1) 29-32 (2014)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

A Novel Approach for Blind Separation of Convolutive Noisy Speech Mixtures Based on Block Thresholding

S. IMRAN, T.U. JAN[++], A. JEHANGIR, S. HAQ*

Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan

Received 6th December 2013 and Revised 18th February 2014

Abstract: A novel algorithm for blind source separation of convolutive noisy speech mixtures based on block thresholding, independent component analysis (ICA), and ideal binary mask along with cepstral smoothing will be proposed in this work. The proposed method consists of four stages. First, two microphone recordings (in presence of noise) has been processed using block thresholding to suppress the effects of noise. The noise considered here will be white Gaussian noise .Then independent components analysis (ICA) has been employed to achieve the separated signals. The ICA works on the principle of statistical independence. As ICA is not very effective in the presence of room reverberations; therefore in the next step ideal binary mask has been estimated from the outputs obtained via ICA and then applied to the mixtures to enhance the separation quality. Ideal binary mask is one of new approach proposed in CASA for speech segregation. Finally, cepstral smoothing is utilized and applied to the separated signals to reduce the musical noise generated due to estimation error caused by masking. The proposed method is evaluated using signal-to-noise ratio approach and results show the enhanced performance.

Keywords: ICA, Block Thresholding, Cepstral Smoothing, Musical Noise.

1. INTRODUCTION

Blind source separation (BSS) is one of advance technique proposed for cocktail party problem. Cocktail party problem arises in most of the environment where we have multiple talkers and limited no of sensors (Cardoso *et al.,* 1998). Segregation of source signals is easy for the humans but difficult in terms of machine to exhibit the same property. Blind source separation segregates the source signals with no prior knowledge about the source signals and mixing process. Many algorithms are proposed for BSS convolutive mixtures in time domain (Cichocki *et al*., 2002)(Jan *et al*., 2009)(Jihua *et al.,* 2009)(Johan *et al*., 2006) or in frequency domain (Wang *et al*., 2005) (Madhu *et al.,* 2008) (Smaragdis *et al*., 1998) (Araki *et al.,* 2003). The solution to this problem will help in Automatic speech recognition, cochlear implantation, hearing impaired devices, interference suppression in CDMA channel and medical imaging. ICA is one of common method employed for blind source separation of convolutive mixtures but its performance is limited when dealing with rooms having reverberations and noise (Jihua *et al.,* 2009)(Johan *et al.,* 2006). Block thresholding is proposed by Cai and Silverman in statistical mathematics to improve the diagonal thresholding estimators and its very effective in removing the noise components (Goushen *et al.,* 2008)(Sreekanth *et al*., 2010). In (Goushen *et al*., 2008) an adaptive block thresholding is proposed which adapts the block thresholding parameters according to signal variation. The attenuation factor is calculated using non-diagonal method by minimizing the stein risk of estimate. The block parameters are adjusted accordingly to minimizing the stein risk of estimate (Stein *et al*., 1980), and stein risk of estimate is calculated from noisy signal. The technique in (Goushen *et al*., 2008) is robust, show improve SNR value and adaptable to signal variation as compared with state of the art methods for audio de noising.(Goushen *et al.,* 2008) (Stein *et al*., 1980) (Sreekanth *et al*., 2010).

IBM proposed in computational auditory scene analysis (CASA) (Wang *et al*., 2005) show high performance in suppressing the interference effects and improving the separated signals quality. IBM represents the signals into time-frequency units and estimated their masks by comparing the energy of target signal with background interference at each unit with 1 is assigned to those units where target signal energy is dominant otherwise zero. But IBM estimation is however difficult without having clean target signal and interference signal (Li *et al*., 2008)

In this paper an algorithm is proposed for blind source separation of linear convolutive noisy mixtures based on block thresholding (Cai *et al.,* 2001) (Goushen *et al*., 2008). In the first stage block thresholding is applied on noisy convolutive mixture to remove the noise component which is considered as white Gaussian noise. Non-diagonal estimation of attenuation factor causes fewer amounts of musical noise but it is introduced (Sreekanth *et al*., 2010). ICA in (Wang *et al*., 2005) is applied to segregate target signals from the linear convolutive mixtures. The output signals are then estimated using IBM to enhance the quality and remove the interference effects. But the estimation errors of IBM create isolated time-frequency structures perceived as musical noise (Wang *et al*., 2005). For the solution of

[++]corresponding author : Tariqullah Jan tariqullahjan@nwfpuet.edu.pk +92-9216498

* Institute of Physics and Electronics, University of Peshawar, Pakistan

musical noise problem due to block thresholding and IBM, cepstral smoothing is applied. Different spectral smoothing levels are applied in cepstral domain on binary estimated signals and musical noise is removed (Madhu *et al*., 2008)

The paper is organized as follows in section II proposed algorithms is presented along with details of each stage Experimental setup and results are presented in section III followed by conclusion and future work.

2. MAERIAL AND METHODS

Proposed Algorithm

In cocktail party environment a convolutive mixture is considered with N no of speech signals observed at the sensors when M no of microphones are used as describe in equation (1). This convolutive mixture contain noise component considered as zero mean Gaussian process independent of signal. The proposed method is employing various steps to segregate the components. Different steps used in the method are discussed here.

2.1 Block thresholding for noise removal

Block thresholding was introduced by Cai and Silverman (Cai *et al*., 2001) in statistics to improve the asymptotic decay of diagonal thresholding estimators. Signal is transformed into time-frequency representation using short-time Fourier transform and resulting coefficient are processed to attenuate the noise component (Smaragdis *et al*., 1998). Time-frequency representation reveals the structures easily discriminated as noise and removed by multiplying with attenuation coefficient. Adaptive block thresholding non-diagonal estimator approach is used as it's automatically adapts the parameters according to signal properties. This approach needs no prior knowledge about the audio signal and works on computing risk of estimate (Goushen *et al*., 2008)

$$X_j(\mathrm{n}) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1)+\epsilon(\mathrm{n}) \quad (\mathrm{j}=1,\ldots,\mathrm{M}) \qquad (1)$$

Equation (1) can be represented as

$$X_j(\mathrm{n}) = \mathrm{S(n)} + \epsilon(\mathrm{n}) \quad \mathrm{n}= 0,1,2\ldots..\mathrm{N}\text{-}1. \qquad (2)$$

While S(n) represents the original mixture signal without noise component and ϵ(n) is the white Gaussian noise element. h_{ji} represents the impulse response of a room (Wang *et al*., 2005). Two signals recorded by two microphone is considered so two input two output system is taken, i.e M = N = 2. Audio signal is decomposed into time-frequency index with g{l,k} atoms where l represents the time and k represents the frequency (Goushen *et al.*, 2008). The time-frequency block thresholding regularize the power subtraction regularization by estimating single attenuation factor over each block. Where * denotes the conjugate.

$$\mathrm{X[l,k]} = \langle X, g_{l,k}\rangle = \sum_{n=0}^{N-1} X_j(n)\, g^*{}_{l,k}[n] \qquad (3)$$

Where $g_{l,k}[n]$ is the short-time fourier atoms and its equal to w[n-lu]exp(i2πkn/K) and w[n-lu] represents the Hanning window with a step function of K. A time-frequency plane of signal in eq (2) is segmented in I blocks B_i and their shapes may be chosen arbitrarily. The original audio mixture $\hat{S}$(n) is estimated by calculating attenuation factor a_i over each block B_i of noisy data X_i(n).(Sreekanth *et al*., 2010)

$$\hat{S}(\mathrm{n}) = \sum_{i=1}^{I}\sum_{(i,k)\epsilon\, Bi} a_i \mathrm{X[l,k]}\, g_{l,k}[\mathrm{n}] \qquad (4)$$

The attenuation factor a_i is calculated over each block by relating an estimate of risk to the frame energy conservation to obtain

$$\mathrm{r} = \mathrm{E}\{\| \mathrm{S} - \hat{S}\|^2\}$$
$$\leq \frac{1}{A}\sum_{i=1}^{I}\sum_{(l,k)\epsilon\, B_k} E\{\, |a_i\, \mathrm{X[l,k]} - \mathrm{S[l,k]}\,|^2\,\} \qquad (5)$$

By choosing $a_i = 1 - \frac{1}{\xi_i+1}$ minimize the upper bound of equation (4) while the $\xi_i = \bar{S}_i^2 / \bar{\sigma}_i^2$ is the averaged a priori SNR over each blocks of B_i and its simply computed from the averaged signal energy and averaged noise energy. But this calculation of attenuation factor from a priori SNR is not possible because pure signal is unknown. Cai and silverman in (Cai *et al*., 2001) introduced a block thresholding estimators for estimation of ξ_i over each blocks B_i. It is calculated by averaging the noisy signal energy over each block B_i.(Goushen *et al*., 2008)

$$\hat{\xi}_i = \frac{\bar{X}_i^2}{\bar{\sigma}_i^2} \qquad (6)$$

Where

$$\bar{X}_i^2 = \frac{1}{B_i^{\#}}\sum_{(l,k)\epsilon B_i} |X[l,k]|^2 \qquad (7)$$

Resulting attenuation factor a_i is calculated using power subtraction estimator

$$a_i = \left(1 - \frac{\lambda}{\hat{\xi}_i+1}\right)_+ \qquad (8)$$

Block thresholding estimator used in this algorithm is non-diagonal (Sreekanth *et al.*, 2010) because its computed from averaged SNR estimations over each block and attenuation factor is calculated from each coefficient in a block so it regularizes the time-frequency coefficient estimation. The estimated signal after removal of noise is

$$Y_j(\mathrm{n}) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1) \qquad (9)$$

2.2 ICA of convolutive mixture for source separation

ICA is applied in frequency domain to separate the source signals from the convolutive mixture. Time domain convolutive mixtures are converted into multiple instantaneous problem into frequency domain (Wang *et al*., 2005)(Araki *et al*., 2003)(Parra *et al*.,2000) utilizing equation (9). It is obtain by applying short time Fourier transform on equation (8) and applying the matrix notation we get

$$\mathrm{Y(k,n)} = \mathrm{H(k)\ S(k,n)} \qquad (10)$$

Signal is transformed into discrete Fourier transform using k for representation of frequency index and n for

representation of discrete time index (Smaragdis *et al.*, 1998). Source signals can be easily estimated by applying an unmixing filter matrix W(k) equivalent to H(k) where H(k) is considered invertible and time invariant

$$Z(k,n)=W(k)Y(k,n) \quad (11)$$

Many algorithms are proposed for estimating the unmixing filter W(k) in (Araki *et al*., 2003)(Parra *et al.*, 2000).In this research a convolutive constrained algorithm of ICA in (Wang *et al*) is used. Z(k) represents the separated source signals from the mixtures $Y_j(n)$ but the separated source signals quality is still limited by interference in (Wang *et al*., 2005). The signal quality degrades especially when rooms having reverberations and interference effects. IBM in (Wang *et al*., 2005) is proposed by CASA for estimating the binary masks and its very effective in improving the separated signal quality.

2.3 Binary masking of ICA segregated source signals

Segregated source signals are converted back into time domain by applying inverse Fourier transform (Wang *et al*., 2005)(Li *et al.*, 2008)

$$z(n)= [z_1(n) \quad z_2(n)]^T \quad (12)$$

Where T represents the transpose. Scaling is utilized to get the normalized output $\bar{z}_1(n)$ and $\bar{z}_2(n)$ and STFT is applied to transform the signal again into frequency domain. Two binary masks are estimated from comparing the energy of each time-frequency unit of each two outputs as

$$[\bar{z}_1 \quad \bar{z}_2] = [\bar{Z}_1(k,n) \quad \bar{Z}_2(k,n)] \quad (13)$$

$$M_1(k,n) = \begin{cases} 1 \;\; if \; |\bar{Z}1(k,n)| > \tau |\bar{Z}2(k,n)|, \\ 0 \qquad otherwise \qquad \forall\, k,n \end{cases} \quad (14)$$

$$M_2(k,n) = \begin{cases} 1 \;\; if \; |\bar{Z}2(k,n)| > \tau |\bar{Z}1(k,n)| \\ 0 \;\; otherwise \qquad \forall\, k,n \end{cases} \quad (15)$$

τ is a threshold for defining the sparsity of source signal and in this experiment its value is taken 1 in reference to (Jan *et al*., 2009). The original source signal are estimated by applying the binary masks to time-frequency representation microphone recording as follow

$$W_i(k,n) = M_i(k,n).Z_i(k,n) \; i=1,2,\ldots\ldots\ldots N. \quad (16)$$

Simply by taking the inverse Fourier transform of $W_i(k,n)$ source signal are converted into time domain. Experimental results show that IBM improves the segregated signal quality considerably by decreasing the interference effects as compared with section 2.2. IBM shows improve results but its mask estimation errors produce fluctuation known as musical noise (Madhu *et al*., 2008). To remove this musical noise cepstral smoothing in cepstral domain is applied on separated signal (Jan *et al*., 2009).

2.4 Cepstral smoothing of estimated signals

Outputs from equation (12) (13) are converted into cepstral domain and variable smoothing levels are used to suppress the musical noise and then transformed back to time-frequency domain as in (Madhu *et al.*, 2008). The binary estimated masks are converted into cepstral domain as

$$M_i^c(\mathbf{l},n) = DFT^{-1}\{ \ln(M_i(\mathbf{k},n)) \mid_{k=0,\ldots K-1}\} \quad (17)$$

Where **l** is quefrency bin index, **k** is frequency bin index and DFT^{-1} represents the inverse Fourier transforms and *K* is its total length. The resulting smoothing masks are obtained by

$$\bar{M}_i^s(\mathbf{l},n)=\lambda_l \bar{M}_i^c(l,n-1)+(1-\lambda_l)\bar{M}_i^c(l,n) \; i=1,.N. \quad (18)$$

Parameter λ is used for regulating the smoothing level. Depending on different value of l smoothing parameter λ is selected as follows

$$\lambda_l = \begin{cases} \lambda_{env} & if \; l \in \{0,\ldots . l_{env}\} \\ \lambda_{pitch} & if \;\; l = l_{pitch} \\ \lambda_{peak} & if \; l \in \{(l_{env}+1),\ldots\ldots,K\} \backslash l_{pitch} \end{cases}$$

Where $0 \leq \lambda_{env} < \lambda_{pitch} < \lambda_{peak} \leq 1$, where λ_{env} is the factor for spectral envelope of the estimated masks M(**k,n**) and λ_{pitch} gives the information of pitch harmonics in masks M(**k,n**). For different smoothing these values are selected to eliminates the musical noise and retain the signal structure. Finally the smoothed estimated masks are calculated as

$$\bar{M}_i(\mathbf{k},\mathbf{n}) = \exp(DFT\{\bar{M}_i^s(\mathbf{l},\mathbf{n}) \mid_{l=0,\ldots K-1}\}) \quad (19)$$

These smoothed masks are again used to extract the segregated signals to get the final output signals

$$S_i(k,n)= \bar{M}_i(k,n).W_i(k,n) \quad (20)$$

3. RESULTS AND DISCUSSION

In this segment the performance of proposed approach is evaluated using simulations based on artificially mixed signal and real room recordings. The metric used for evaluation is signal-to-noise ratio as noise removal is the main aim of algorithm along with separated signal quality enhancement.

3.1 Experimental setup

The parameters for cepstral smoothing are set as l_{env}=8.0, l_{low}=16.0, l_{high}=120.0, λ_{env}= 0, λ_{pitch}=0.4, and λ_{peak}=0.8. Audio mixtures from 12 different speaker are selected both male and female, are used in (Jan *et al* 2009). Overlap factor of .75 is used for hamming window. The signals are selected for 5 seconds with a sampling rate of 10KHz. The performance is evaluated on the basis of signal-to-noise ratio described in three values as $mSNR_i$, $mSNR_O$ and ΔSNR. Where SNR_i and SNR_o will give us average results for fifty random tests. SNR_i gives the ratio among the source signal and the interference signal. Ratio among the source signal as (outputs of IBM) and the difference of source signals to estimated ones is given and $mSNR_o$. (Jan *et al.*, 2009).

3.2 General evaluation and comparison

A number of experiments have been performed along with changing parameters to evaluate the algorithm performance. First, different window lengths of 256.0,

512.0, 1024.0 and 2048.0 are considered to separate two source signals from 50 different noisy convolutive mixtures and the results are presented into Table.1. It is evaluated that highest ΔSNR is secured for window size of 512.0.

The performance is also evaluated on considering variable FFT frame length and results are presented in Table.2. SNR improves considerably by increasing the FFT frame lengths as shown in the results. The algorithm shows high performance at frame length of 2048. While Table.3 shows the performance of algorithms in presence of noise and its shows that its performance degrades as noise level is increased. The performance evaluation of proposed algorithm is studied in comparison with algorithm in (Jan *et al*., 2009) and it shows high quality of separated source signal even in presence of white Gaussian noise, its SNR is still better and it shows the enhanced performance of the proposed method in comparison to state of the art method (Jan, *et al.*) specially in presence of noise. Also it shows that whenever noise is present in the convolutive mixtures, then there is a need of method to tackle the noise component separately in order to enhance its performance.

Table.1 Results for different window lengths at the noise level fixed at -10db

Window Length	mSNRi	mSNRo	ΔSNR
256	1.20	6.83	5.63
512	1.09	7.35	6.26
1024	1.21	6.94	5.73
2048	1.22	6.04	4.82

Table.2 Results for variable FFT frame lengths at the noise level fixed 10db

NFFT	mSNRi	mSNRo	ΔSNR
512	1.10	6.61	5.51
1024	1.20	6.82	5.62
2048	1.09	7.35	6.26

Table.3 Results at different noise levels and comparison with state of the art method (Jan *et al.*)

Noise	mSNRi	mSNRo	ΔSNR	Jan *et al.*method
-10 dB	1.09	7.35	6.26	5.81
-20 dB	1.10	7.54	6.44	6.33
-30 dB	1.10	7.55	6.45	6.34
-40 dB	1.10	7.55	6.45	6.34

4. CONCLUSION AND FUTURE WORK

A novel algorithm of BSS based on block thresholding is proposed for segregation of source signals from noisy convolutive mixtures. Proposed algorithm removes the noise component by applying block thresholding and then applies ICA to segregate the two source signal recorded by two microphones. Source signals quality is further improved by estimating the ideal binary masks from the separated signal and applied to T-F representation of original mixture. T-F estimation of source signal induces musical noise, so cepstral smoothing in cepstral domain is applied to remove this musical noise without distorting the source signals. The observed source signals show high quality in terms of signal-to-noise ratio. Future work involves the segregation of source signals in presence of every type of noises.

REFERENCES:

Araki. S., R. Mukai, S. Makino, and H. Saruwatari. (2003) The fundamental limitation of frequency domain blind source separation. IEEE Trans. Speech audio process., vol. **(11)**: 109-116.

Parra. L. and C. Spence.(2000) Convolutive blind separation of non stationary sources. IEEE Trans. Speech audio process., vol. **(8):** 320-327.

Cichocki. A., S. Amari (2002) Adaptive blind signal and image processing. Wiley press. USA

Cai. T. and B.W.Silverman.(2001) Incorporation information on neighboring coefficients into wavelet estimation. Sankhya. vol. **(63):** 127-148.

Goushen., Y., S. Mallat. and E. Bacry. (2008) Audio denoising by time-frequency block thresholding. IEEE Trans on signal processing., vol. **(56):**5. 1830-1839.

Jan. T., W. Wang. and D. Wang.(2009). A multistage approach for blind separation of convolutive speech mixtures. IEEE ICASSP. Taiwan, 2009, 1713-1716.

Jihua.C, and J. Liu. (2009) A new algorithm of blind source separation based on ICA. World computer science and information engineering. Cardoso. J. F. (1998) Blind signal separation: statistical principles., Proc.I EEE. vol **(9):** 2009–2025.

Li.Y. and D. Wang. (2008) On the optimality of ideal binary time-frequency masks. IEEE ICASSP, Las Vegas, 2008, 3501-3504.

Madhu. N., C. Breithaupt. and R. Martin. (2008) Temporal smoothing of spectral masks in the cepstral domain for speech separation. Proc. IEEE ICASSP., 45-48, 2008, Las Vegas.

Stein. C.(1980) Estimation of the mean of a multivariate normal distribution. Ann statist., vol **(9):** 1135-1151.

Sreekanth. S., P. D. Khanna.and, P. Uma (2010) An efficient noise reduction by using diagonal and non diagonal estimation technique. proceeding of the international conference on comm and computational intelligence.,393-398, India

Smaragdis.P (1998) Blind separation of convolved mixtures in the frequency domain," Neurocomput., vol. 22, 21–34.

Thomas.J., Y. Deville. S. Hosseini (2006) Time-Domain Fast Fixed-Point Algorithms for Convolutive ICA IEEE. Signal processing letters. Vol **(13):**4. 228-231.

Wang. D. L. (2005) On ideal binary mask as the computational goal of auditory scene analysis. In Divenyi P.(ed). Speech separation by humans and machines., 181-297.

Wang. W., S. Sanei. and J. A. Chambers.(2005) Penalty function based joint diagnolization approach for convolutive blind separation of nonstationary sources. IEEE trans signal process., vol. **(53):** 1654–1669.