Sindh Univ. Res. Jour. (Sci. Ser.) Vol.46 (2): 153-158 (2014)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Blind separation of convolutive speech mixtures with background interference employing a Hybrid approach With ICA & PCA.

U. ALI, K. M. YAHYA, T. JAN[++], A. JEHANGIR, S. ALI, S. R. HASSNAIN

Department of Electrical Engineering UET Peshawar, Pakistan

Received 4th January 2014 and Revised 7th March 2014

Abstract: This paper presents a comparative analysis of Principal component analysis (PCA) and a rapidly emerging novel method, that is, the independent component analysis (ICA).An algorithm is designed in order to segregate the convolutive mixtures of speech with background noise utilizing two microphones recordings. The efficiency of the target speech has been analyzed by utilizing principle component analysis (PCA) and ideal binary mask (IBM), proceeding by post-filtering in cepstral domain. The segregation of the speech signal with background noise is achieved by three steps algorithm. Initially PCA algorithm is applied on mixture of source signals received by two microphone recording to segregate them. In the next step, the segregated sources obtain by PCA is used to assess the IBM with the comparison of the energy of corresponding time-frequency (T-F) units. Finally, T-Fmasking using cepstral smoothing is used to decrease musical noise**.** Then the results achieved using PCA based approach has been compared with ICA based approach.

Keywords*:* Principal component analysis (PCA), Independent component analysis (ICA), Ideal binary mask (IBM), Time - Frequency (T-F), cepstral smoothing, musical noise.

1. INTRODUCTION

The problem with a single speaker you want to extract from composite signal, frequently related in a problem known as cocktail party (Cherry, 1953) (Haykin, 2005). Humans listen to the main notable well in unfavorable surroundings, but the legibility of human speech suffers especially in high level of sound. To separate speech sources separately with their complex mixtures is a bit difficult job when the information known is much less about the sources and the process of their mixing.

This problem has been studied for several years. To overcome this problem different techniques have been introduced. BSS which was first suggested by the French researchers (Ans *et al.*, 1985) (Herault *et al.*, 1986) in the mid-1980s is a technology for separating the superimposed sound signal from each other. PCA algorithm is applied to a linear convolutive model with background noise, to segregate a complex signals to their superimposed waves in transformed domain or in time-domain. However, there is a large space vacant for the betterment of segregation of numerous developed algorithms which are yet limited. This is particularly true when it deals with noisy and reverberant signals. Also the ICA technique has been employed to the complex mixture with background noise and then the results achieved has been compared.

A newly strategy, known Ideal Binary Mask (IBM), comes from computational auditory scene analysis computational auditory scene analysis (CASA) (Wang, 2005), reflects auspicious outcomes in subduing interference and enhancing the quality of the desired signal. Typically IBM is acquire in contrast with the T-F representations of the desired signal and backdrop interference, with "1" is entrusted to T-F unit where the desired output is greater than interference energy and else "0". Though, it is an effortful task to assess a rigorous IBM directly without the fair desired signal and distract sound from the convoluted mixtures.

This paper presents how to fetch the desired signal and backdrop sound individually from the convoluted signals using a Principle Component Analysis (PCA) algorithm, the results of which are then used to estimate the IBM. This method can efficiently solve the above problem in connection with the individual methods. But the cumulative error resulted during estimation of IBM give rise to the segregated T-F units and thus results in fluctuating artifacts, commonly known as the musical sound (Madhu *et al.*, 2008).To gain control of this issue, the spectral smoothing is imposed to the assessed binary mask in the cepstral domain. In contrast to the method (Madhu *et al.*, 2008), at various frequencies, different levels of smoothing are used for a binary mask, calculated by the pitch data and is estimated directly from the separated signals. Then the results obtained from PCA based approached discussed above has been compared with the ICA based method to assess their performance.

[++]Corresponding Author: tariqullahjan@nwfpuet.edu.pk, Ph. 091-9216498

The paper is organized as following: Section 2 contains the proposed algorithm, Section 3 consists of the experimental discussions and results, Section 4 consists of the comparison proposed algorithm with the existing algorithm in (Jan *et al.*, 2011), and Section 5 concludes the paper.

2. MATERIAL AND METHOD

The section 2 contains two main sub sections consisting of previous method based on ICA approach which describes the Jan *et al.*, method and current method based on PCA approach.

2.1 Previous Method.

In literature, a method discussed in (Jan *et al.*, 2011), is the baseline method for this approach. Therefore we will discuss Jan et al. method here. There are three steps in Jan et al. method.

2.1.1 Bss of Convolutive Mixtures in the Frequency Domain.

N sound signals are captured by M microphones in cocktail party surroundings, portrayed mathematically by,

$$z_j(n) = \sum_{i=1}^{n} \sum_{p=1}^{p} q_{ji}(p) s_i(n-p+1) \qquad \text{j=1,..,M} \qquad \text{(i)}$$

Where $s_i$represents the source and $z_j$represents the mixture signal, q_{ji} is a P-point room impulse response (Allen *et al.*, 1979). Additionally a simple framework consisting of two input two output is used also known as two input two output (TITO) framework, i.e., N=M=2. The BSS issue for convoluted signals in time domain is changed over to various instantaneous issues in frequency domain (Cherry, 1953), (Ans *et al.*, 1985) with equation (ii). The short time Fourier transform (STFT) is applied to equation (i), and using matrix notations we obtain,

$$Z(k,n) = Q(k)S(k,n) \qquad \text{(ii)}$$

where k&n indicates the frequency & discrete time indexes respectively. The matrix $Q(k)$is a mixing matrix, likely to be time invariant and invertible. The unmixing filter $\beta(k)$is applied to the convolutive mixtures of sources to find the sources.

$$U(k,n) = \beta(k)Z(k,n) \qquad \text{(iii)}$$

where $U(k,n)$ is assessed source signals. $\beta(k)$is acquired when the assessed sources come out to be statistically independent. Numerous algorithms have been produced for this reason (Jan *et al.*, 2011) (Araki *et al.*, 2003) (Cichocki *et al.*, 2002), (Parra *et al.*, 2000). Jan et al. used a constrained convolutive ICA approach (Wang *et al.*, 2005) for the segregation at this level. Like numerous existing ICA techniques, e.g. (Parra *et al.*, 2000), the segregation performance of (Wang *et al.*, 2005), particularly the performance of the segregated speech is yet bound due to interference.

The quality of the desired signal declines by increasing of reverberation time (RT). The IBM technique from the computational auditory scene analysis (CASA) space is employed to enhance the performance of the segregated speech signals.

2.1.2 Combining Convolutive Ica And Binary Masking for The Segregation Of Speech Signals.

$U(k,n)$can be converted back into time space by applying inverse fourier transform, denoted as:

$$u(n) = [u_1(n) \quad u_2(n)]^T \qquad \text{(iv)}$$

normalized outputs $\tilde{u}_1(n)$ and $\tilde{u}_2(n)$ is obtained by applying scaling on the $u_1(n)$ and $u_2(n)$, transformed into the T-F space utilizing STFT.

$$\tilde{u}_1(n) \rightarrow \tilde{U}_1(k,n) \qquad \text{(v)}$$
$$\tilde{u}_2(n) \rightarrow \tilde{U}_2(k,n) \qquad \text{(vi)}$$

The estimation of the two binary masks placed by comparing the energy of each T-F unit of the above spectrograms as,

$$M_1^f(k,n) = \begin{cases} 1 \; if \; |\tilde{U}_1(k,n)| > \alpha|\tilde{U}_2(k,n)|, \\ 0 \qquad\qquad otherwise \; \forall k,n. \end{cases} \qquad \text{(vii)}$$

$$M_2^f(k,n) = \begin{cases} 1 \; if \; |\tilde{U}_2(k,n)| > \alpha|\tilde{U}_1(k,n)|, \\ 0 \qquad\qquad otherwise \; \forall k,n. \end{cases} \qquad \text{(viii)}$$

Where α is the factor for controlling the scattering of the mask, and α = 1is used in Jan et al. method. The source signals are fetched up by applying masks to the T-F representation of the source signals taken by two microphone recordings as under below.

$$U_i^f(k,n) = M_i^f(k,n).Z_i(k,n) \; i\text{= 1,….,N} \qquad \text{(ix)}$$

Now original sources are retrieved in time domain utilizing STFT.

$$U_i^f(k,n) \rightarrow u_i^t(n) \qquad \text{(x)}$$

The results in (Jan *et al.*, 2011) reveals that, the assessed IBM significantly enhance the segregation performance by subduing the noise to a much lower degree, leading to the segregated sound signals with massively enhanced the quality of the desired output gained in previous section. Though, a common issue with the binary T-F masking is the development of the bugs in the estimation of the masks causing unstable musical noise (Madhu, *et al.*, 2008). Acepstral smoothing technique (Madhu *et al.*, 2008), is utilized to alleviate this issue, as discussed in the next section.

2.1.3 Cepstral Smoothing Of the Binary Mask.

The estimated IBM is transformed into the cepstral space and is converted back into the T-F space, utilizing different smoothing levels on the transformed mask. It has been observed, the musical artifacts in the signal decreases, without affecting the broadband structure and pitch data simultaneously due to smoothing operation (Madhu*et al.,* 2008)(Oppenheim *et al.,* 1975). By transforming the equations (vii) and (viii) in the cepstral space is as under below,

$$M_i^c(\omega,n) = DFT^{-1}\left\{ln\left(M_i^f(k,n)\right)|k=0,\dots,K-1\right\} \quad \text{(xi)}$$

Where "ω" is the quefrency bin index and "k" is the frequency bin index (Madhu*et al.,* 2008), $DFT$is the discrete fourier transform. $ln$and K is the natural logarithm and length of the DFT respectively. The resultant smoothed mask is obtained by utilizing smoothing is as under below.

$$\overline{M_i^s}(\omega,n) = \sigma\,\overline{M_i^s}(\omega,n-1) + (1-\sigma)M_i^c(\omega,n) \quad i=1,..,N \text{(xii)}$$

The factor σ is critically important, used for handling the smoothing level.The criteria for the selection of σ with respect to different values of ω is given as.

$$\sigma = \begin{cases} \sigma_{env} & \text{if } \omega \in \{0,\dots,\omega_{env}\} \\ \sigma_{pitch} & \text{if } \omega = \omega_{pitch} \\ \sigma_{peak} & \text{if } \omega \in \{(\sigma_{env}+1),\dots,K\}/\omega_{pitch} \end{cases} \quad \text{(xiii)}$$

Where$0 \le \sigma_{env} < \sigma_{pitch} < \sigma_{peak} \le 1$, ω_{env} is the quefrency bin index, shows the spectral envelope of the mask $M^f(k,n)$ and ω_{pitch} is the quefrency bin index indicates the structure of the pitch harmonics in$M^f(k,n)$. The principle engaged for this range of σis demonstrated as follows.

$M^c(\omega,n), \omega \in \{0,\dots,\omega_{env}\}$essentially represents the spectral envelope of the mask $M^f(k,n)$. The value chosen for σ in this region is moderately low to ignore the interference in the envelope. Likewise, low smoothing is applied if $\omega = \omega_{pitch}$, with the goal that the harmonic structure of the signal is maintained. In order to decrease the noise high smoothing is applied on the last range selected for σ, while the structure and the pitch information of the spectral envelope remains safe. Diverse from (Madhu *et al.,* 2008), pitch frequency has been computed in (Jan *et al.,* 2011), utilizing by speech signal achieved in previous section. Computed pitch frequency is given as:

$$\omega_{pitch} = \text{argmax}_\omega\left\{\text{sig}^c(\omega,n)\middle|\omega_{low} \le \omega \le \omega_{high}\right\} \text{(xiv)}$$

$\text{sig}^c(\omega,n)$shows the cepstral domain representation of the isolated speech signal $u^t(n)$. The range $\omega_{low}, \omega_{high}$is picked up to provide a set of human speech frequencies having range between 50Hz to 500Hz. The resultant smoothed mask is as under below:

$$\overline{M_i^f}(k,n) = exp\left(DFT\left\{\overline{M_i^s}(\omega,n)|t=0,\dots,K-1\right\}\right) \text{(xv)}$$

The output of the isolated speech signal acquire in above sectionis applied to the resultant smoothed mask, as under,

$$\overline{U_i^f}(k,n) = \overline{M_i^f}(k,n).U_i^f(k,n) i=1,\dots,N \text{ (xvi)}$$

2.2 Current Method.

In our proposed method, initially PCA algorithm is applied to a linear convolutive model with background noise in order to fetch the desired output. As it is common with numerous other algorithms (Karhunen *et al.,* 1994) (Asano *et al.,* 2000) (Karhunen *et al.,* 1995) (Oja, 2008) (Karhunen *et al.,* 1998), by this step the resultant segregated target speech holds a lot of noise from other sources. The performance of the model becomes poor with the increment of reverberation time constantly. IBM is estimated with the comparison of the energy of the corresponding T-F units obtained from the outputs of the PCA algorithm, by doing this it helped to suppress the noise within the desired speech, thereafter the estimated IBM is applied to the original mixtures to fetch the desired output and interference sources. By doing so, the segregated performance improves by reducing the interference at very low range. Somehow, a problem occurs with the T-F masking that it creates bugs in the estimation of the masks.

To overcome this problem, the estimated IBM is processed utilizing cepstral smoothing. The IBM is transformed into a cepstral domain, and smooth the transformed mask over time frames utilizing the overlap and add technique. The smoothed mask is converted back into the T-F space, is then applied to the outputs of the last step to diminish the musical artifacts.

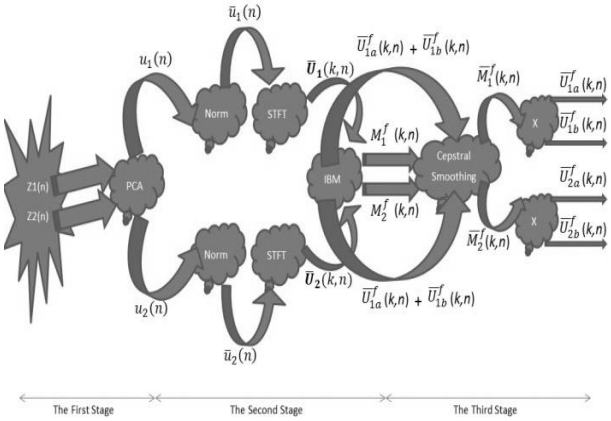

Fig. 1: Block Diagram of the Model

The proposed algorithm is basically multistage algorithm, as shown by a block diagram in **(Fig. 1).** for two microphones recordings. The theme of this work is to perform comparative analysis of the multistage algorithm based on principle component analysis (PCA) with Jan et al. method in (Jan *et al.,* 2011) based on independent component analysis (ICA). The evaluation results will be discussed in the next section.

3. EXPERIMENTAL DISCUSSION AND RESULTS.

This section briefly explains the performance of the proposed technique using simulations.

3.1 Experimental Arrangement.

For the assessment of the proposed method 12 different voice signals consist of 6 male and 6 female with 11 different languages have been recorded (Jan *et al.,* 2011). The time span of each of the signals, 5 seconds and the sampling frequency is $f_s = 10KHZ$. The volume level of all the voice signals has same. The hamming window is used with an overlap factor of 0.75. Remaining parameters σ_{env}, σ_{pitch}, σ_{peak}, ω_{env}, ω_{low} and ω_{high} are setting as: 0, 0.4, 0.8, 8,16 and 120 respectively. Moreover, to check the model performance, criteria namely signal to noise ratio *(SNR)*, is used (Jan *et al.,* 2011). $mSNR_i$, *mSNRo, and* ΔSNR are also used for the assessment. SNR_i is the ratio between the desired signal and the interfering signal taken from the mixture, SNR_o is the ratio between the desired signal resynthesized from IBM to the difference of the desired resynthesized signal and the estimated signal (Jan *et al.,* 2011). $mSNR_i$ and $mSNR_o$ are the average of 50 random test. ΔSNR is given as, $\Delta SNR = mSNR_o - mSNR_i$.

3.1 General Assessment.

Initially a mixture of voice signal is obtained using model as in (Jan *et al.,* 2011). A set of experiments were performed to assess the proposed technique by changing the parameters. Using above parameter arrangement, experiments have been performed for *RT* = 100 ms. Different frame length have been used for the analysis e.g. that is 256, 512, 1024 and 2048.

The results are given in **Table 1-1 and 1-2**. However table 1-1 shows the results from Jan et al. method (Jan *et al.,* 2011) and table 1-2 shows the results of proposed algorithm. The mean behavior shows for 50 different convolutive mixtures each consisting of 2 voice sources picked up in a random way from a set of 12 voice signals (Jan *et al.,* 2011). From the above experiment it reflects that the highest ΔSNR is gained of 512 frame length. Hence this frame length is fixed for succeeding experiments.

Table 1-1 Results for Different Window Lengths Based on ICA

Window Length	$mSNR_i$	$mSNR_o$	ΔSNR
256	1.10	7.11	6.01
512	1.10	7.44	6.34
1024	1.10	7.11	6.01
2048	1.12	6.32	5.20

Table 1-2 Results for Different Window Lengths Based on PCA

Window Length	$mSNR_i$	$mSNR_o$	ΔSNR
256	1.10	3.12	2.02
512	1.10	3.22	2.12
1024	1.10	3.11	2.01
2048	1.12	2.54	1.42

Table 2-1 and table 2-2 show the efficiency of the Jan et al. method and proposed technique for different FFT frame length. It has been observed that the efficiency of the Jan et al. method is higher than the proposed technique by increment in terms of SNRby varying frame length from 512 to 2048.

Table 2-1 Results for Different FFT Frame Lengths Based on ICA

NFFT	$mSNR_i$	$mSNR_o$	ΔSNR
512	1.10	7.17	6.06
1024	1.10	7.40	6.30
2048	1.10	7.44	6.34

Table 2-2Results for Different FFT Frame Lengths Based on PCA

NFFT	$mSNR_i$	$mSNR_o$	ΔSNR
512	1.10	3.01	1.91
1024	1.10	3.08	1.98
2048	1.10	3.22	2.12

Table 3-1 and Table 3-2 shows the mean values of $\Delta SNR$of Jan et al. method and proposed technique with respect to different values of reverberation time. It can be seen that with the rise of reverberation time, the efficiency of the ΔSNRdecreases.

Table 3-1 Results for Different RT Based on ICA

RT	$mSNR_i$	$mSNR_o$	ΔSNR
40	1.13	13.22	12.08
60	1.15	10.94	9.79
80	1.14	9.42	8.27
100	1.10	7.44	6.34
120	1.03	6.30	5.26
140	0.94	5.48	4.53
150	0.90	5.29	4.39

Table 3-2 Results for Different RT Based on PCA

RT	$mSNR_i$	$mSNR_o$	ΔSNR
40	1.13	5.44	4.31
60	1.15	4.43	3.28
80	1.14	3.92	2.78
100	1.10	3.22	2.12
120	1.03	3.11	2.08
140	0.94	2.99	1.94
150	0.90	2.33	1.43

It can be seen that the experiments perform as above has noiseless mixture. Now by adding noise to the mixture we performed the experiments which show that the efficiency of the above parameters decrease with the rise of noise, likewise to (Jan *et al.*, 2011), the algorithm can bear the level of noise up to -20dB. **Table 4-1 and table 4-2** show the mean value of $\Delta SNR$of Jan *et al.*, method and proposed technique with respect to different values of noise level.

4. COMPARISON

The performance of the current proposed approach is compared with the Jan *et al.* method as discussed above (Jan *et al.*, 2011). The comparison is done by simulation in Mat lab.Overall the results given in tables show that the ICA based approach is giving better results than the PCA based approach.

Table 4-1 Results for Different Noise Levels Based on ICA

Noise	$mSNR_i$	$mSNR_o$	ΔSNR
-10dB	1.09	6.91	5.81
-20dB	1.10	7.43	6.33
-30dB	1.10	7.44	6.34
-40dB	1.10	7.45	6.34

Table 4-2 Results for Different Noise Levels Based on PCA

Noise	$mSNR_i$	$mSNR_o$	ΔSNR
-10dB	1.09	2.82	1.73
-20dB	1.10	3.23	2.13
-30dB	1.10	3.24	2.14
-40dB	1.10	3.24	2.14

5. CONCLUSION

A multistage technique is used for the separation of convolutive mixtures of speech signal using two microphones recordings with background noise. Initially the convolutive mixtures of speech sources are segregated using PCA algorithm. The IBM is estimated using segregated speech signals, further applied to the time frequency representation of the original source signals. In the end the musical noise is decreases by T-F masking, cepstral smoothing is applied on the estimated IBM. The proposed algorithm can be used in automatic speech recognition systems, in cochlear implantation and also be used for the suppression of interference and designing of acoustic systems.

REFERENCES:

Araki. S., R. Mukai, S. Makino, and H. Saruwatari. (2003) The fundamental limitation of frequency domain blind source separation for convolutive mixture of speech, IEEE Trans. Speech Audio Process., vol. **(11):** 109–116.

Asano. F., Y. Motomura, H. Asoh, T. Matsui, P. Pajunen and J. Karhunen. (2000) Effect of PCA in blind source separation, in Proceedings of the Second International Workshop on ICA and BSS, P.Pajunen and J. Karhunen, Eds., Helsinki, Finland, 57–62

Ans. B., J. H´erault, and C. Jutten. (1985) Adaptive neural architectures Detection of primitives in Proceedings of COGNITIVA'85, Paris, France, 593-597.

Cherry. E. C. (1953) Some experiments on the recognition of speech, with one and two ears, The Journal of the Acoustical Society of America., vol. **(25):** 975–979.

Cichocki. A and S. Amari. (2002) Adaptive Blind Signal and Image Processing, John Wiley& Sons Press, New York.

Allen. J and D. Berkley (1979) Image Method for Efficiently Simulating Small Room Acoustics, Journal of the Acoustical Society of America, vol. **(65)**: 4 943-950.

Haykin. S. and Z. Chen. (2005) The cocktail party problem, Neural Computation., vol. **(17):** 1875–1902.

H´erault. J and C. Jutten. (1986) Space or time adaptive signal processing by neural networks models, Intern. Conf. on Neural Networks for Computing, Snowbird (Utah, USA), 206–211.

Jan. T., W. Wang, D. L. Wang.(2009)A multistage approach for blind separation of convolutive speech mixtures. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1713-1716.

Karhunen. J. and J. Joutsensalo. (1994) Representation and separation of signals using nonlinear PCA type learning, Neural Networks, Elsevier, vol. **(7):** 1, 113–127.

Karhunen. J., L. Wang and R. Vigario (1995) Nonlinear PCA type approaches for source separation and independent component analysis, IEEE, Perth, WA, vol. **(2):** 995-1000.

Karhunen. J., P. Pajunen, E. Oja (1998) The nonlinear PCA criterion in blind source separation: Relations with other approaches. Neurocomputing, **(22)**: 520-524.

Madhu. N., C. Breithaupt, and R. Martin. (2008) Temporal smoothing of spectral masks in the cepstral domain for speech separation, in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA, 45-48.

Oppenheim. A. V. and R. W. Schafer. (1975)Digital Signal Processing, Prentice Hall, New Jersey.

Oja. E (1995) The nonlinear PCA learning rule and signal separation – mathematical analysis. Helsinki University of Technology, Laboratory of Computer and Information Science, Finland, Report A26.

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

Wang. D. L. (2005) On ideal binary mask as the computational goal of auditory scene analysis, in P. Divenyi (Ed.), "Speech Separation by Humans and Machines," Chapter 12, Kluwer Academic, Norwell MA. 181-197,

Wang. W., S. Sanei and J. A. Chambers. (2005) Penalty function based joint diagnolization approach for convolutive blind separation of no stationary sources, IEEE Trans. Signal Process., vol. **(53):** 1654–1669.