INDEPENDENT VECTOR ANALYSIS FOR SOURCE SEPARATION USING AN ENERGY DRIVEN MIXED STUDENT'S T AND SUPER GAUSSIAN SOURCE PRIOR

Waqas Rafique[*], *Suleiman Erateb*[†], *Syed Mohsen Naqvi*[*], *Satnam S. Dlay*[*], *Jonathon A. Chambers*[*]

[*]School of Electrical and Electronic Engineering, Newcastle University, NE1 7RU, UK
{w.rafique2, mohsen.naqvi, satnam.dlay, jonathon.chambers}@newcastle.ac.uk
[†]School of Electronic, Electrical and Systems Engineering, Loughborough University, LE11 3TU, UK

ABSTRACT

Independent vector analysis (IVA) can thoretically avoid the permutation problem in frequency domain blind source separation by using a multivariate source prior to retain the dependency between different frequency bins of each source. The performance of the IVA method is however very dependent upon the choice of source prior. Recently, a fixed combination of the original super Gaussian, previously used in the IVA method, and the Student's t distributions has been found to offer performance improvement; but due to the non-stationary nature of speech, this combination should adapt to the statistical properties of the measured speech mixtures. Therefore, in this work we propose a new energy driven mixed multivariate Student's t and super Gaussian source prior for the IVA algorithm. For further performance improvement, the clique based IVA method is used to exploit the strong dependency between neighbouring frequency components. This new algorithm is evaluated on mixtures formed from speech signals from the TIMIT dataset and real room impulse responses and performance improvement is demonstrated over the conventional IVA method with fixed source prior.

Index Terms— Blind source separation, independent vector analysis, binaural room impulse responses

1. INTRODUCTION

In signal processing, an important tool for blind source separation (BSS) is independent component analysis (ICA) [1]. A major application domain for BSS is the well-known cocktail party problem, in which the desired speaker must be separated from speech mixtures [2, 3]. Due to reverberations in the real room environment, it becomes convolutive BSS (CBSS) [4]. Time domain methods for CBSS are generally computational complex [5]. Therefore frequency domain (FD) methods are preferred [6]. Although this approach reduces the computational cost, it introduces the permutation problem across the frequency bins and various prior/post processing methods have been proposed as solution but they increase latency [7].

Independent vector analysis (IVA) is an algorithmic approach which has been proposed to solve the permutation problem in FD-CBSS [8]. The IVA method theoretically avoids the permutation problem by using a multivariate super Gaussian distribution to retain the dependency between different frequency bins of each source. Such modelling preserves the intra-vector source dependencies whilst delivering inter-vector independence. Thus, the IVA method mitigates the permutation problem in the learning process without any prior or post processing [9]. Moreover, joint dependency modelling in the IVA method assumes that the frequency bins of sources have symmetric distribution and the IVA method is performed on a fully connected clique (range of frequency bins). Since speech signals are typically spherically invariant random processes in the FD [10], such an assumption yields reasonable performance. However, the separation performance of the IVA method is slightly inferior when compared with the FD-ICA followed by perfect permutation correction [9]. Therefore, a better dependency model is still needed for improved separation performance of the IVA algorithm.

In this paper, we propose a new energy driven mixed multivariate Student's t and super Gaussian (as in original IVA) distributions source prior for the IVA algorithm. The Student's t distribution due to its heavy tailed nature can be used to model the high amplitude information in speech signals, while the original super Gaussian distribution can be used to model the other information. The weight of both distributions in the mixed source prior should be automatically adapted according to the energy of the measured speech mixture signals. Importantly, the method is found to be successful only with access to the mixtures not the original sources. Moreover, to further improve the separation performance of the IVA algorithm, fully connected frequency bins are decomposed into many smaller groups because the dependency among the neighbouring frequency bins is generally stronger and much weaker between the distant frequency bins. So the strong dependency between neighbouring frequency bins is exploited by dividing them into smaller cliques whilst retaining considerable overlap between adjacent cliques. Furthermore, the proposed IVA model with energy driven mixed source prior is tested with real room impulse responses (RIRs) [18, 19], instead of commonly used synthetic RIRs. The experimental results show that the adaptation of the proposed

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

source prior with clique based IVA consistently improves separation performance in realistic scenarios.

2. INDEPENDENT VECTOR ANALYSIS

The noise-free model in FD-CBSS is described as:

$$\mathbf{x}(k) = \mathbf{H}(k)\mathbf{s}(k) \tag{1}$$

$$\hat{\mathbf{s}}(k) = \mathbf{W}(k)\mathbf{x}(k) \tag{2}$$

where $\mathbf{x}(k) = [x_1(k), x_2(k) \cdots x_m(k)]^T$ and $\hat{\mathbf{s}}(k) = [\hat{s}_1(k), \hat{s}_2(k) \cdots \hat{s}_n(k)]^T$ is the observed mixture signal vector and estimated signal vector both in the FD, respectively, and $(.)^T$ denotes vector transpose. $\mathbf{H}(k)$ and $\mathbf{W}(k)$ are the mixing matrix and the unmixing matrix respectively. The index k denotes the k-th frequency bin of this multivariate model. In this paper we assume that the number of sources is the same as the number of microphones, i.e. $m = n$.

The Kullback-Leibler divergence between the joint probability density function $p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_n)$ and the product of probability density functions (PDFs) of the individual source vectors $\prod q(\hat{\mathbf{s}}_i)$ is used for IVA [9]. Each source in the IVA method is a multivariate vector and the cost function would be minimised when the vector sources are independent while the dependency between the components of each vector is still preserved. The inter-frequency dependency is modelled by the PDF of the source and the original IVA algorithm exploits a particular multivariate super Gaussian distribution as the source prior, which can be written as:

$$q(\mathbf{s}_i) \propto \exp\left(- \sqrt{\sum_{k=1}^{K} \left| \frac{\hat{s}_i(k)}{\sigma_i(k)} \right|^2} \right) \tag{3}$$

where $\sigma_i(k)$ denotes the standard derivation of the ith source at the kth frequency bin. In the original IVA method, where the cost function is minimised by the gradient descent method, the nonlinear score function for source $\hat{s}_i$ can be obtained as [9]:

$$\varphi(k)(\hat{s}_i(1) \cdots \hat{s}_i(K)) = \frac{\hat{s}_i(k)}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i(k)|^2}} \tag{4}$$

where $\varphi(k)(\hat{s}_i(1) \cdots \hat{s}_i(K))$ is a multivariate score function and is used to retain dependency across the frequency bins and K is the number of frequency bins.

3. MULTIVARIATE MIXED SOURCE PRIOR

In the IVA method, the separation performance depends strongly on the score function which is used to preserve the inter-frequency dependency. In [9], it is stated that the non-linear function is derived based on the PDF of sources; so by selecting a more appropriate source prior for the source speech signals, performance of the IVA method can potentially be improved. Therefore in this paper, instead of modelling all the sources by an identical multivariate super Gaussian distribution, we have considered an energy driven mixture of Student's t distribution and the original super Gaussian distribution as a source prior. The Student's t distribution due to its heavy tailed nature can better model the high amplitude information in speech sources, such as during voiced intervals [11–13] and other information can be modelled by the original super Gaussian distribution. The weight of both distributions in the mixed source prior is adapted according to the energy of the speech mixtures as will be described below.

3.1. Mixed Distribution Source Prior

The cost function for the IVA method can be minimised by adopting the multivariate Student's t distribution as the source prior and the score function can be obtained as [11]:

$$\varphi(k)(\hat{s}_i(1) \cdots \hat{s}_i(k)) \propto \frac{\nu + K}{\nu} \frac{\hat{s}_i(k)}{1 + \frac{1}{\nu} \sum_{k=1}^{K} |\hat{s}_i(k)|^2} \tag{5}$$

where ν is a degrees of freedom parameter that can tune the variance and leptokurtic nature of the distribution. With decreasing ν, the tails of the distribution becomes heavier. We found empirically that $\nu = 4$ is the appropriate value for degrees of freedom parameter in this work. Therefore, the Student's t distribution can account for the high amplitude signals and the super Gaussian distribution can model the other samples in the speech signals. So, the new mixed source prior can be written as:

$$q(\mathbf{s}_i) = (\lambda_d).f_{St} + (1 - \lambda_d).f_G \tag{6}$$

where f_{St} and f_G are the multivariate Student's t distribution and the original multivariate super Gaussian distributions, respectively; $\lambda_d \epsilon [0, 1]$ is a weighting parameter and determines the weight of each distribution in the mixed source prior.

By using the mixed Student's t and the original super Gaussian source prior as given in equations (5) and (4), respectively, with appropriate normalisation, the overall non linear score function for source $\mathbf{s}_i$ can be obtained as:

$$\varphi(k)(\hat{s}_i(1) \cdots \hat{s}_i(K)) \propto (\lambda_d) \left(\frac{\hat{s}_i(k)}{1 + \frac{1}{\nu} \sum_{k=1}^{K} |\hat{s}_i(k)|^2} \right) + (1 - \lambda_d) \left(\frac{\hat{s}_i(k)}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i(k)|^2}} \right) \tag{7}$$

In equation (7), the score function is a multivariate function. This score function can preserve the inter-frequency dependency, as all the frequency bins are accounted for during the learning process. In this paper, the value of λ_d becomes frequency dependent i.e. $\lambda_d(k)$, and the weight of the distributions in the mixed source prior is adapted accordingly.

In our work, the weighting parameter $\lambda_d(k)$ is calculated according to the energy of the measured speech mixture. It is determined as the normalised energy of the speech mixtures in the FD blocks. When frequency bins are divided into non overlapping blocks to calculate the normalised energy of the source mixture, then the energy of a particular block can be obtain as:

$$E_b = \frac{1}{E_t}\left(\sum_{k=f_b}^{l_b} ||\mathbf{x}_p(k)||^2\right) \tag{8}$$

where f_b and l_b are the first and last indices of the block, respectively and $\mathbf{x}_p(k)$ represents the vector of all frequency components k calculated by dividing the entire speech observation into 50% overlapping subblocks index by p, whereas E_b and E_t are the energy of the particular block (clique) and the total energy of the source mixture, respectively, and $||(\cdot)||$ denotes Euclidean norm. In the case of high energy signals more weight is given to the Student's t distribution and lower weight to the super Gaussian distribution in the mixed source prior and vice versa. Hence, this source prior can better model the underlying non-stationary speech signals by adapting to the nature of the measured speech mixtures thereby, improving the separation performance of the IVA method.

3.2. Clique Based IVA Method

In the IVA method, the neighbouring and distant frequency components are assigned the same dependency whereas in real life speech sources the dependency between the neighbouring frequency components is much stronger than that of distant frequency components. Therefore, to further improve the performance of the IVA method, the single and fully connected statistical dependency model of IVA is decomposed into several overlapping cliques of fixed size. The corresponding multivariate PDF can be written as [14]:

$$q(\mathbf{s}_i) \propto \exp\left(-\sum_{c=1}^{C}\sqrt{\sum_{k=f_c}^{l_c}\left|\frac{\hat{\mathbf{s}}_i(k)}{\sigma_i(k)}\right|^2}\right) \tag{9}$$

where C is the number of cliques and f_c and l_c are the first and last indices of the c-th clique, respectively.

As an example, in the case of 1024 frequency bins, in order to consider strong dependency between neighbouring frequency bins, the single clique of IVA is decomposed into 128 cliques, each of fixed size 256 and the clique ranges in equation (8) are $[f_1, l_1] = [0, 255], [f_2, l_2] = [16, 271], \ldots, [f_c, l_c] = [768, 1023]$. This dependency model can make better use of strong dependency between neighbouring frequency bins and improve the separation performance of the IVA method when used with the proposed energy driven mixed source prior.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed energy driven mixed source prior for the IVA method by using three different kinds of RIRs. In all the experiments we randomly select the speech signals from the whole TIMIT dataset [22].

4.1. Separation Performance with Synthetic RIRs

The image method [17] is used to generate the RIRs and the size of the room was $7 \times 5 \times 3m^3$. The DFT length was 1024 and the reverberation time (RT_{60}) was set to 200ms. Two speech signals were randomly chosen from the whole TIMIT database and convolved into two mixtures. The microphone sources were positioned at $[3.48, 2.50, 1.50]m$ and $[3.52, 2.50, 1.50]m$ and the sampling frequency was 8kHz. For repeated simulations, the position of both sources were changed six times. The separation performance was evaluated objectively by the signal-to-distortion ratio (SDR) [20]. Separation performance of the proposed algorithm with Image RIRs is shown in Table 1 and it confirms the proposed energy driven mixed source prior can consistently improve the separation performance of the IVA method.

Source Prior	Mix-1	Mix-2	Mix-3	Mix-4	Mix-5
As in [9]	8.58	9.01	8.65	7.24	8.03
Proposed	9.53	9.93	9.7	8.12	9.09
Improvement	0.95	0.92	1.09	0.88	1.06

Table 1: The table indicates the improvement in separation performance in terms of SDR (dB) for five randomly chosen speech mixtures from TIMIT dataset [22] using the image method [17]. For each mixture the SDR values are averaged for six different positions.

4.2. Separation Performance with Real RIRs

Commonly, the IVA method had generally only been evaluated with RIRs generated by the image method, which are synthetic and can not provide proper evaluation of a BSS algorithms for real life contexts. In this paper, we have evaluated the proposed IVA method with a variety of real binaural room impulse responses (BRIRs). The binaural room impulse responses used in this work come from two sources.

4.2.1. BRIRs from Hummersone [18]:

Table 2: DIFFERENT ROOM ACOUSTICS PROPERTIES.

Room	Type	RT60 (ms)
A	Medium office	320
B	Small class room	470
C	Large lecture room	680
D	Large seminar theatre	890

The first set of BRIRs comes from [18]. As shown in Table 2, the advantage of this BRIRs dataset is that they were measured in different rooms with different acoustic properties, which facilitates evaluation of the performance of algorithms in different real life scenarios. Source location azimuths ranging from $(-90\,^\circ \text{ to } 90\,^\circ)$ relative to the second source were available. We have used source location azimuths from $(15\,^\circ$ to $90\,^\circ$ with a step of $15\,^\circ)$. Separation performance of the proposed energy driven mixed source prior IVA in all four rooms at all angles is shown in Figure 1 and it shows when compared with the conventional IVA method, we obtained improvement in all the rooms when the interfering speech is placed far away from the target source.

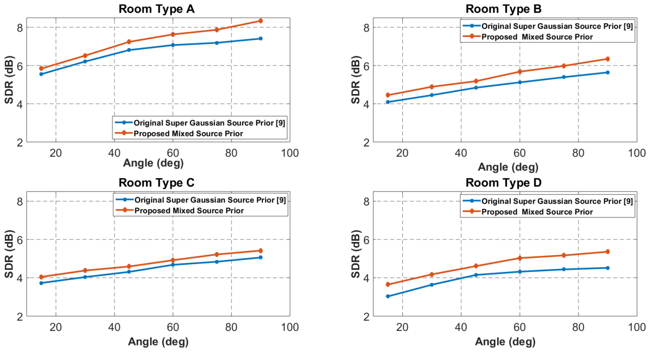

Fig. 1: The graph indicates results at different room types. Real BRIRs from [18] were used. Results were averaged over twelve mixtures at each angle. The original IVA method is shown in blue and the proposed in red.

4.2.2. *BRIRs from Shinn [19]:*

Table 3: DIFFERENT PARAMETERS USED IN EXPERIMENTS.

STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	2.5 s (TIMIT)

The second set of BRIRs comes from [19]. They were recorded in a real classroom with RT_{60} of 565ms, which shows the achieved performance of the algorithm in a difficult and highly reverberant environment. Six different source location azimuths $(15^\circ - 90^\circ)$ relative to the second source were used. Also for reliability all measurements were recorded on three separate occasions. The common parameters used in all experiments using these BRIRs are given in Table 3. Separation performance of the source prior is shown in Figure 2 and it presents the average SDRs of the separated sources in different room settings by using BRIRs [19]. Compared with the conventional IVA method we obtained at least 0.90 dB of average improvement at all the angles.

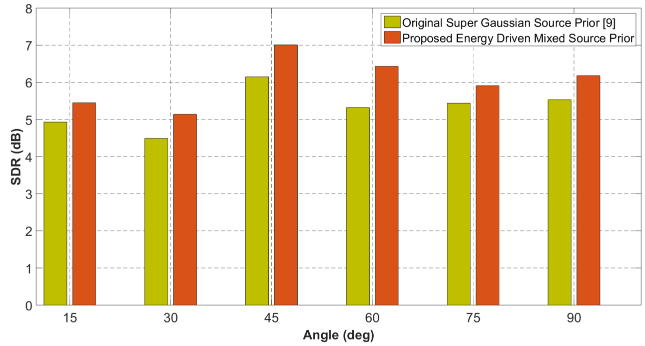

Fig. 2: The graph provides results for the IVA method at different separation angle by using real BRIRs [19]. Results were averaged over eighteen mixtures at each angle. Our proposed energy driven mixed source prior yields a considerable improvement in all room settings.

In addition to objective measures, a subjective measure, perceptual evaluation of speech quality (PESQ) [21] is also used with BRIRs [19] to confirm the superior performance of the proposed mixed source prior and the results are presented in Table 4 which shows the PESQ score for the proposed source prior in the context of an extremely high $RT_{60} = 565ms$ and this subjective measure also confirms the improved separation performance of the proposed source prior.

Source Prior	Mix-1	Mix-2	Mix-3	Mix-4	Mix-5
As in [9]	1.66	2.04	2.09	1.92	2.02
Proposed	1.97	2.27	2.32	2.11	2.21

Table 4: The table shows PESQ values for both source priors with BRIRs [19]. For each mixture PESQ values are averaged over six different angles.

In order to demonstrate further advantage of the energy driven source prior, its performance is compared with the fixed coefficient mixed source prior, in which $\lambda_d = 0.5$ and both distributions have equal weight in the mixed source prior irrespective of the energy of the source mixtures and the results are shown in Figure 3.

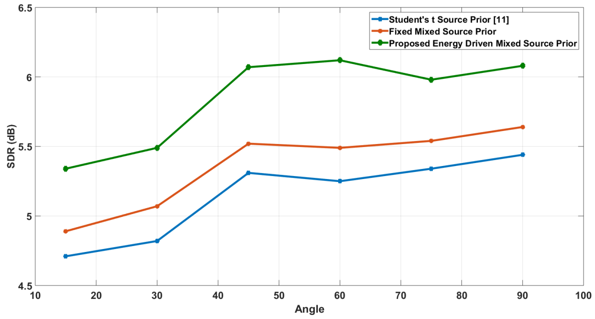

Fig. 3: The graph provides performance of the Student's t, fixed and energy driven source prior using BRIRs [19]. Results were averaged over twelve mixtures at each angle. Our proposed energy driven mixed source prior yields a considerable improvement in all room settings.

Figure 3 confirms that the adaptation of the proposed source prior with the clique based IVA consistently improves the separation performance in a realistic scenario.

5. RELATION TO PRIOR WORK AND CONCLUSIONS

In this paper, an energy driven mixed multivariate Student's t and super Gaussian source prior is used for the first time in the IVA algorithm. This particular source prior can be adopted according to the energy of the speech mixtures and it can better preserve the frequency dependencies as compared to the conventional super Gaussian source prior [9]. The analysis of source priors is also discuss in [15, 16]. In this paper we have also used clique based IVA to make use of strong dependency between the neighbouring frequency components in order to further enhance the separation performance of energy driven mixed source prior. Speech signals are highly random and they can have significantly high as well as low energy components. The mixed energy driven source prior can model both components efficiently by calculating the energies of the mixtures and adjusting weights for either heavy tailed Student's t distribution or super Gaussian distribution in the source prior. The new detailed experimental results using realistic BRIRs show that the IVA method with the new source prior can consistently achieve improved separation performance.

6. REFERENCES

[1] C. Jutten and J. Herault, “Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture,”*Signal Processing*, vol. 24, pp. 1-10, 1991.

[2] C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *The Journal of The Acoustical Society of America*, vol. 25, pp. 975-979, 1953.

[3] S. Haykin and Z. Chen, “The cocktail party problem,” *Neural Computation*, vol. 17, pp. 1875-1902, 2005.

[4] A. Cichocki and S. Amari, “Adaptive Blind Signal and Image Processing,” *John Wiley*, 2002.

[5] E. Bingham and A. Hyvarinen, “A fast fixed point algorithm for independent component analysis of complex valued signals,”*Int. J. Neural Networks*, vol. 10, pp. 1–8, 2000.

[6] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources,”*IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 320–327, 2000.

[7] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, “A survey of convolutive blind source separation methods,” *Springer Handbook on Speech Processing and Speech Communication*, vol. 8, pp. 1-34, 2007.

[8] T. Kim, I. Lee, and T.-W. Lee, “Independent vector analysis: definition and algorithms,” in *Fortieth Asilomar Conference on Signals, Systems and Computers 2006*, (Asilomar, USA), 2006.

[9] T. Kim, H. Attias, S. Lee, and T. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 70-79, 2007.

[10] H. Brehm, and W. Stammler “On the assumption of spherical symmetry and spaseness for frequency-domain speech model,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 70.79, 2007.

[11] W. Rafique, S.M. Naqvi, P.J.B. Jackson and J.A Chambers, “Independent vector analysis with multivariate Student’s t distribution source prior for speech separation in real room environments,”*IEEE ICASSP*, Brisbane, Australia, pp. 474-478, 2015.

[12] D. Peel and G. J. McLachlan, “Robust mixture modelling using the t distribution“, *Satistics and Computing*, vol 10, pp. 339-348, 2000.

[13] I. Cohen, “Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation,” *Speech Communication*, vol. 47, pp. 336-350, 2005.

[14] I. Lee and G. Jang, “Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation,“ *EURASIP Journal on Advances in Signal Processing*, vol. 113,pp. 1-12, 2012

[15] I. Lee and T. W. Lee,“On the assumption of spherical symmetry and sparseness for the frequency-domain speech model,” *IEEE Trans. on Audio, Speech and Language processing*, vol. 15, pp. 1521-1528, 2007.

[16] Y. Liang, S. M. Naqvi and J. A. Chambers, “Independent vector analysis with a multivariate generalized Gaussian source prior for frequency domain blind source separation,” *IEEE ICASSP*, Vancouver, Canada, pp. 6088-6092, 2013.

[17] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustical Society of America*, vol. 65, pp. 943-950, 1979.

[18] C. Hummersone, “A psychopsychoacoustic engineering approach to machine sound source separation in reverberant environments,“ *Ph.D. dissertation*, University of Surrey, 2011

[19] B. Shinn-Cunningham, N. Kopco, and T. Martin, “Localizing nearby sound sources in a classroom: Binaural room impulse responses,“ *Journal of the Acoustical Society of America*, vol. 117, pp. 3100-3115, 2005.

[20] E. Vincent, C. Fevotte, and R. Gribonval, “Performance measurement in blind audio source separation,“ *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462-1469, 2006.

[21] Y. Hu and P.C. Loizou,“Evaluation of Objective Quality Measures for Speech Enhancement,“ *IEEE Transactions on Audio, Speech, and Language Processing*, vol.16, pp. 229-238, 2008

[22] J. S. Garofolo et al., “TIMIT acoustic-phonetic continuous speech corpus,” in *Linguistic Data Consortium*, (Philadelphia), 1993.