REAL-TIME INDEPENDENT VECTOR ANALYSIS WITH STUDENT'S T SOURCE PRIOR FOR CONVOLUTIVE SPEECH MIXTURES

Jack Harris[⋆†], *Bertrand Rivet*[⋆], *Syed Mohsen Naqvi*[†], *Jonathon A. Chambers*[†], *Christian Jutten*[⋆]

[†]School of Electronic, Electrical and Systems Engineering, Loughborough University, UK
{j.harris, s.m.r.naqvi, j.a.chambers}@lboro.ac.uk
[⋆]GIPSA-Lab, CNRS UMR 5216, Université de Grenoble, France.
{bertrand.rivet, christian.jutten}@gipsa-lab.grenoble-inp.fr

ABSTRACT

A common approach to blind source separation is to use independent component analysis. However when dealing with realistic convolutive audio and speech mixtures, processing in the frequency domain at each frequency bin is required. As a result this introduces the permutation problem, inherent in independent component analysis, across the frequency bins. Independent vector analysis directly addresses this issue by modeling the dependencies between frequency bins, namely making use of a source prior. An alternative source prior for real-time (online) natural gradient independent vector analysis is proposed. A Student's t probability density function is known to be more suited for speech sources, due to its heavier tails, and is incorporated into a real-time version of natural gradient independent vector analysis. In addition, the importance of the degrees of freedom parameter within the Student's t distribution is highlighted. The final algorithm is realized as a real-time embedded application on a floating point Texas Instruments digital signal processor platform, where simulated recordings from a reverberant room are used for testing. Results are shown to be better than with the original (super-Gaussian) source prior.

Index Terms— source separation, independent vector analysis, embedded application, real-time, multivariate distribution

1. INTRODUCTION

The cocktail party problem is a well-known problem within the signal processing community; which was originally proposed in [1]. This is a typical blind source separation (BSS) problem (i.e. the mixing filters are unknown) and is often addressed with independent component analysis (ICA) [2, 3, 4].

Realistic audio signals measured at microphones are generally convolutive due to the reverberant nature of real world environments; thus ICA algorithms which address the audio BSS problem are commonly implemented in the frequency domain [5]. A drawback of frequency domain ICA is that the calculated unmixing filters may permute the sources at each frequency bin (known as the permutation problem), due to the permutation ambiguity inherent in ICA. Various methods have been suggested to mitigate this effect, in [6] smoothing over adjacent frequency bins is suggested as a way of addressing the issue. In addition [7] suggests limiting the length of the filter in the time domain. Also in [8] video tracking of sources is suggested as an approach to address the permutation problem.

However [9] introduces independent vector analysis (IVA). This directly addresses the permutation problem by maintaining the dependencies between the frequency bins in the algorithmic formulation by using a dependent multivariate super-Gaussian distribution as the source prior, instead of a univariate distribution used in ICA style methods.

Previously an online (thus real-time) version for natural gradient IVA (NG-IVA) was formulated in [10] and mentioned in [11], along with various implementations of real-time (online) ICA [12, 13, 14] (which all have to address the permutation problem by means of various post-processing techniques). Furthermore, a batch version of Auxiliary IVA is implemented on an embedded system in the form of a smartphone application [15].

In this paper we build on previous work [16], by introducing a Student's t source prior model and incorporating it into the NG-IVA algorithm. Distributions with heavier tails are more suited to speech, as they better model the higher amplitude data points in a frequency domain speech signal [17]. This differs from the original multivariate super-Gaussian distribution used in [9]. We implement this alternative source prior within NG-IVA as an embedded application on a Texas Instruments digital signal processing platform and show that the new source prior performs well in terms of separation performance when compared to the original NG-IVA algorithm. The importance of choosing a suitable value for the degrees of freedom is also discussed.

The mixing and unmixing process formulation is described in Section 2.1; the real-time natural gradient IVA algorithm is described in Section 2.2. In Section 2.3 the

alternative Student's t pdf is introduced. We describe the experimental setup and results from a realistic binaural environment in Sections 3 and 4. Finally, a discussion and conclusion can be found in Section 5.

2. METHOD

2.1. Problem Formulation

The observation at each sensor of a microphone array can be modeled in the general case in the frequency domain as a convolutive mixture from each source of the form:

$$x_i^{(k)} = \sum_{j=1}^{N} h_{ij}^{(k)} s_j^{(k)} + \sigma_i^{(k)} \tag{1}$$

where s_j is the speech signal generated by the j-th source, h_{ij} is the filter that models the effect of the environment between the j-th source and the i-th sensor, k is the frequency bin index, σ_i is additive zero mean noise uncorrelated with the speech signals, x_i is the detected signal at the i-th sensor, and N is the number of sources. The noise, $\sigma_i^{(k)}$, can be considered as an extra source and for brevity is dropped for the remainder of the paper.

The frequency domain unmixing model is considered as:

$$\hat{s}_j^{(k)}[n] = \sum_{j=1}^{N} g_{ji}[n]^{(k)} x_i^{(k)}[n] \tag{2}$$

where n is the block index, $\hat{s}_j$ is the estimated signal for the j-th source and g_{ji} is the frequency domain unmixing filter to find the estimation of the j-th source from the i-th observation.

2.2. Real-Time Natural Gradient Independent Vector Analysis

The cost function ($\mathcal{C}$) of the IVA algorithm uses the Kullback-Lieber divergence (denoted by $\mathcal{KL}(\cdot)$) between the joint probabilities and the product of the marginal probabilities as a measure of independence:

$$\mathcal{C} = \mathcal{KL}\big(p(\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_L) || \prod_{i=1}^{L} q(\hat{\mathbf{s}}_i)\big) \tag{3a}$$

$$= \int p(\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_L) \log \frac{p(\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_L)}{\prod_{i=1}^{L} q(\hat{\mathbf{s}}_i)} d\hat{\mathbf{s}}_1 \dots \hat{\mathbf{s}}_L \tag{3b}$$

$$= \int p(\mathbf{x}_1 \dots \hat{\mathbf{x}}_M) \log p(\mathbf{x}_1 \dots \mathbf{x}_M) d\mathbf{x}_1 \dots \mathbf{x}_M \tag{3c}$$

$$- \sum_{k=1}^{K} \log \big| \det G^{(k)} \big| - \sum_{i=1}^{L} \int p(\hat{\mathbf{s}}_i) \log q(\hat{\mathbf{s}}_i) d\hat{\mathbf{s}}_i$$

$$= const. - \sum_{k=1}^{K} \log |\det G^{(k)}| - \sum_{i=1}^{L} E\big[\log q(\hat{\mathbf{s}}_i)\big] \tag{3d}$$

where $E[\cdot]$ denotes the mathematical expectation and L denotes the number of sources. To minimize the cost function ($\mathcal{C}$) we take a natural gradient approach by taking the partial derivatives with respect to the separating filter co-efficients ($g_{ij}^{(k)}$), the gradients for the filter co-efficients are given by:

$$\Delta g_{ij}^{(k)} = -\frac{\partial \mathcal{C}}{\partial g_{ij}^{(k)}} = g_{ij}^{(k)-1} - E\big[\varphi^{(k)}\big(\hat{s}_1^{(1)} \dots \hat{s}_1^{(K)}\big)\big] x_j^{(k)*} \tag{4}$$

where $(\cdot)^*$ denotes the complex conjugate. Then by multiplying by the scaling matrices [9], we obtain

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{L} \big(I_{il} - E[\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) \hat{s}_i^{(k)*}]\big) g_{lj}^{(k)} \tag{5}$$

where $I_{il} = 1$ when $i = l$, and zero otherwise. The expectation in equation (5) is dropped to form the block wise algorithm and thus yields:

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{L} \big(I_{il} - \varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) \hat{s}_l^{(k)}\big) g_{lj}^{(k)} \tag{6}$$

which is the major difference between the original (batch) NG-IVA and the real-time version in this paper.

The non-linear score function (φ) which maintains the dependencies between frequency bins is given by

$$\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = -\frac{\partial \log q\big(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}\big)}{\partial \hat{s}_i(k)} \tag{7}$$

A nonholomonic constraint is also implemented as in [10], therefore (6), becomes:

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{L} \big(\Lambda_{il} - \varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) \hat{s}_l^{(k)}\big) g_{lj}^{(k)} \tag{8}$$

where $\Lambda_{ii}^{(k)} = \Re_{ii}^{(k)}$, $\Re_{il}^{(k)} = \varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) \hat{s}_l^{(k)}$ and zero otherwise. The block-wise update equation, which includes a gradient normalization as in [10], for the separating filter co-efficients is given by:

$$g_{ij}^{(k)}[n+1] = g_{ij}^{(k)}[n] + \mu \sqrt{(\xi^{(k)}[n])^{-1}} \Delta g_{ij}^{(k)} \tag{9}$$

for which the normalization factor ($\xi^{(k)}$) is defined as:

$$\xi^{(k)}[n] = \beta \xi^{(k)}[n-1] + (1-\beta) \sum_{i=0}^{L} x_i^{(k)}[n]^2 / L \tag{10}$$

where β is the smoothing factor. The following section introduces the alternative Student's t source prior.

2.3. Alternative Student's t Source Prior

The main contribution of this paper is suggesting an alternative source prior to the original super-Gaussian source prior. A Student's t multivariate pdf is proposed to model the high amplitude data points in a frequency domain speech signal more accurately than the original super-Gaussian source prior, due to its heavier tails.

Derived from a multivariate super-Gaussian distribution by setting the mean to zero and the covariance matrix to the identity matrix, the non-linear score function derived from the source prior in [9] is given as:

$$\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}} \tag{11}$$

Previous work [16], introduces a new multivariate Student's t probability density function as the source prior as an alternative to the original super-Gaussian source prior. A multivariate Student's t distribution takes the form:

$$q(s_i) \propto \left(1 + \frac{(\mathbf{s}_i - \mu_i)^H \mathbf{\Sigma}_i^{-1} (\mathbf{s}_i - \mu_i)}{v}\right)^{(v+K/2)} \tag{12}$$

The degrees of freedom parameter (v) controls the leptokurtic nature of the pdf. As v decreases the tails become heavier and as they increase the pdf becomes more Gaussian-like.

By assuming zero mean ($\mu_i = \mathbf{0}$) and setting the covariance matrix ($\mathbf{\Sigma}_i$) to the identity matrix, a new non-linear score function is derived and replaces equation (11):

$$\varphi_{new}^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{1 + (1/v)\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \tag{13}$$

The choice of the degrees of freedom (v) becomes important in the real-time version of NG-IVA as shown in the results section.

3. EXPERIMENTAL SETUP

3.1. Floating point TI TMS320C6713 platform

The real-time version of IVA is implemented on a Texas Instruments TMS320C6713 floating point digital signal processing platform (TI DSP) (Fig. 1) [18]. Features of the board include a TI C6713 digital signal processor, an AIC23 codec, 16 MB of external memory, line-in/out socket and headphone in/out socket. NG-IVA was implemented in C using fast Fourier transform (FFT) code provided by TI. Not including the FFT code, the approximate time to execute the update equations (2), (8) - (10) and 13 for one time block is 0.063 seconds (approx 14.4 million instructions cycles), which is easily realized on the TI DSP.

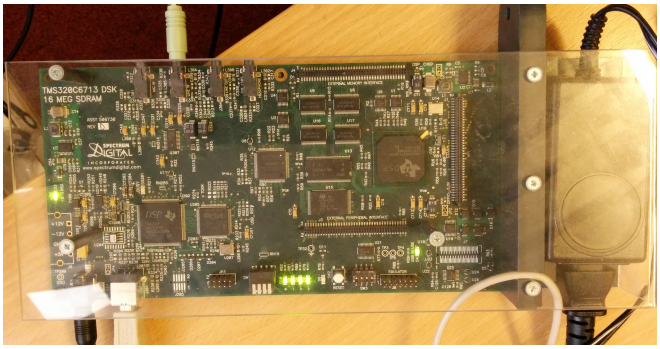

Fig. 1: Texas Instruments TMS320C6713 floating point digital signal processing platform.

3.2. Methodology and Room Layout

NG-IVA was tested on a two-speaker, two-sensor scenario. To recreate a realistic room environment, binaural room impulse responses of 565ms as used in [19] were convolved with speech files from fourteen randomly selected individual speakers from the TIMIT database [20]. The sampling frequency for the system is 8kHz. Male and female speakers were swapped between two positions which where both 40cm away from the microphone array at 0° and 45° relative to the center of the array. Utterances from each speaker across different accents were selected to form the clean speech sources, the utterances were then concatenated to form longer speech signals, up to 240s (i.e. each speaker was repeating what they were saying with the full range of utterances available for that speaker) . These speech mixtures were then played via a PC sound card into the line in of the TI DSP for processing, the separated sources were audible via headphones attached to the headphone out jack of the TI DSP.

Unmixing matrices (G) were saved at every five seconds, the results given are based on the unmixing matrices obtained and simulated (unrepeated) speech mixtures.

Performance of the separated mixtures are based on two measurements, the signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) [21] as the original speech sources are available. SIR takes into account the interfering sources affecting an estimated source, the SDR also considers interfering sources and in addition takes into account any artifacts (e.g. filtering effects) and any noise within an estimated source.

4. RESULTS

Results are given in SDR (Figure 2) and SIR (Figure 3) and show good performance over a period of approximately 2.5 minutes. However convergence time is not as good as that in [9] (where convergence is less than 20 seconds), there are two reasons for this; realistic reverberant binaural room impulse responses are used in our experimental setup, rather

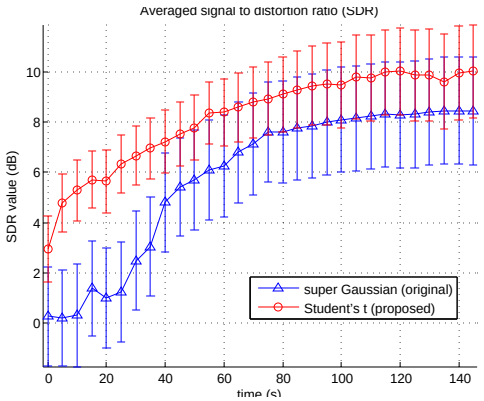

Fig. 2: Convergence of NG-IVA as averaged SDR over 17 mixtures. Note $T_{60} = 565$ms.

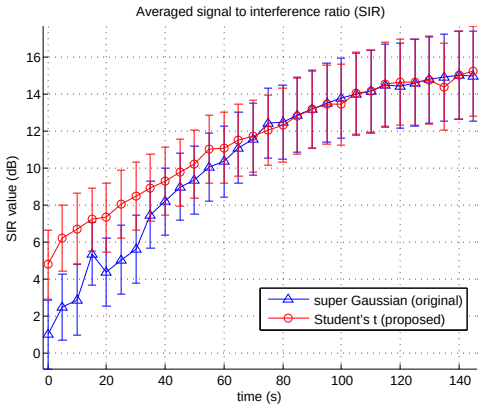

Fig. 3: Convergence of NG-IVA as averaged SIR over 17 mixtures. Note $T_{60} = 565$ms.

than room impulse responses generated by the image method [22]. Secondly, in an attempt to ensure that there are enough frequency bins to cover the length of the time-domain room impulse response, more frequency bins are used for the unmixing filters (2048, compared to 256), thus it takes longer for all the unmixing filters to converge for all frequency bins.

Also included is a comparison between values for the degrees of freedom (v). Averaged SDR convergence plots for different values of v are shown (Fig. 4), where μ was chosen for relatively fast convergence for a range of mixtures and values of v. The value for μ was kept constant for all Student's t plots ($\mu = 0.6$, with a scaling factor of $1{\times}10^6$). For comparison a typical performance curve for the super-Gaussian source prior is given.

5. CONCLUSION

A real-time (online) algorithm for NG-IVA has been presented with an alternative source prior, based on a multivariate Student's t distribution, this gives an improved model for high amplitude data points in a frequency domain speech signal due to the heavier tails of the Student's t distribution. Reverberant mixtures used are more realistic, thus more of a challenge to separate, than those used in some previous studies. In addition, the importance of the degrees of freedom within the Student's t distribution was highlighted.

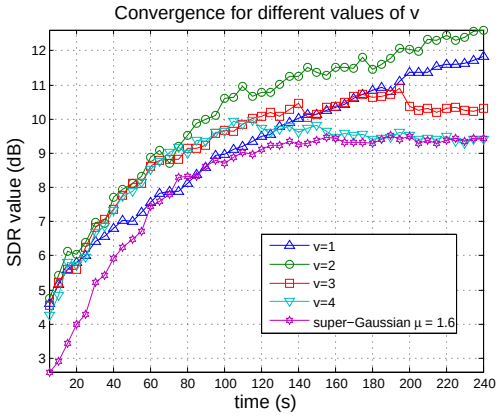

Fig. 4: SDR convergence for different values of v. μ=0.6 except for the super-Gaussian plot where it is 1.6. Plots have been averaged over nine mixtures which include male and female speakers. The SIR plot shows similar performance and for brevity is not included.

Results show improved performance in terms of SDR and SIR when compared to the original NG-IVA, which has a source prior based on a multivariate super-Gaussian distribution. Real-time NG-IVA is easily implemented as an embedded application on a TI TMS320C6713 DSP platform, a common floating-point DSP platform, due to its lower complexity when compared to the batch version.

Future work will involve incorporated video signals and developing real-time NG-IVA on a field programmable gate array (FPGA). Another angle for future research is incorporating video cues to aid the source separation and convergence speed of real-time IVA.

Acknowledgements: J. Harris is funded by a DGA/Dstl PhD scholarship. This work is partially funded by 2012-ERC-AdG-320684 CHESS. GIPSA-lab is a partner of the LabEx PERSYVAL-Lab (ANR–11-LABX-0025).

6. REFERENCES

[1] E.C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *J. Acoust. Soc. Am*, vol. 25, no. 5, pp. 975–979, 1953.

[2] P. Comon, "Independent Component Analysis, a New

Concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, 2001.

[4] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[5] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 1065–1084, 2007.

[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[7] L. Parra and C. Spence, "Convolutive Blind Separation of Non-Stationary Sources," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 320–327, 2000.

[8] S.M. Naqvi, M. Yu, and J.A. Chambers, "A Multimodal Approach to Blind Source Separation of Moving Sources," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 895–910, 2010.

[9] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 70–79, 2007.

[10] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 57, no. 7, pp. 1431–1438, 2010.

[11] J. Hao, I. Lee, T. Lee, and T. Sejnowski, "Independent vector analysis for source separation using a mixture of gaussians prior," *Neural computation*, vol. 22, no. 6, pp. 1646–1673, 2010.

[12] S. Ding, J. Huang, D. Wei, and A. Cichocki, "A near real-time approach for convolutive blind source separation," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 53, no. 1, pp. 114–128, 2006.

[13] L.N. Oliva-Moreno, J.A. Moreno-Cadenas, L.M. Flores-Nava, and F. Gomez-Castaneda, "DSP implementation of extended infomax ICA algorithm for blind source separation," in *Electrical and Electronics Engineering, 2006 3rd International Conference on*. IEEE, 2006, pp. 1–4.

[14] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Real-time blind source separation for moving speakers using blockwise ICA and residual crosstalk subtraction," in *Proc. ICA*, 2003, pp. 975–980.

[15] N. Ono, "Blind Source Separation On iPhone In Real Enviroment," *Proc. EUSIPCO*, 2013.

[16] Y. Liang, G. Chen, S.M.R. Naqvi, and J.A. Chambers, "Independent vector analysis with multivariate student's t-distribution source prior for speech separation," *Electronics Letters*, vol. 49, no. 16, 2013.

[17] I. Cohen, "Speech enhancement using super-gaussian speech models and noncausal a priori SNR estimation," *Speech communication*, vol. 47, no. 3, pp. 336–350, 2005.

[18] R. Chassaing and D. Reay, *Digital Signal Processing and Applications with the TMS320C6713 and TMS320C6416 DSK*, John Wiley & Sons, 2nd edition, 2008.

[19] B. G. Shinn-Cunningham, N. Kopco, and T. J Martin, "Localizing Nearby Sound Sources in a Classroom: Binaural Room Impulse Responses," *J. Acoust. Soc. Am.*, vol. 117, pp. 3100, 2005.

[20] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[21] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide," Tech. Rep. 1706, IRISA Technical Report 1706, Rennes, France, 2005.

[22] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.