

Article

An Expectation–Maximization-Based IVA Algorithm for Speech Source Separation Using Student’s t Mixture Model Based Source Priors

Waqas Rafique ^{1,*} , Jonathon Chambers ² and Ali Imam Sunny ¹

¹ Department of Informatics, King’s College London, London WC2B 4BG, UK; ali.imam@kcl.ac.uk

² Department of Engineering, University of Leicester, Leicester LE1 7RH, UK; jonathon.chambers@leicester.ac.uk

* Correspondence: waqas.rafique@kcl.ac.uk

Received: 23 November 2018; Accepted: 29 December 2018; Published: 10 January 2019



Abstract: The performance of the independent vector analysis (IVA) algorithm depends on the choice of the source prior to better model the speech signals as it employs a multivariate source prior to retain the dependency between frequency bins of each source. Identical source priors are frequently used for the IVA methods; however, different speech sources will generally have different statistical properties. In this work, instead of identical source priors, a novel Student’s t mixture model based source prior is introduced for the IVA algorithm that can adapt to the statistical properties of different speech sources and thereby enhance the separation performance of the IVA algorithm. The unknown parameters of the source prior and unmixing matrices are estimated together by deriving an efficient expectation maximization (EM) algorithm. Useful improvement in the separation performance in different realistic scenarios is confirmed by experimental studies on real datasets.

Keywords: blind source separation; student’s t mixture model; independent vector analysis; real room impulse responses

1. Introduction

The process of automated separation of acoustic sources from measured mixtures is known as acoustic blind source separation (BSS) [1]. The typical application of blind source separation is to handle the cocktail party problem, which is the process of focusing on one particular acoustic source of interest in the presence of multiple sound sources [2–4]. Human beings can easily pay attention to one of the speakers in the presence of multiple active speakers; however, it is much more difficult to replicate the same ability in machines [5]. In the past few decades, much research has been conducted to study different aspects of the cocktail party problem. This research includes the study of the geometry of the microphone array [6], room impulse response identification [7], localisation of speech sources [8] and statistical estimation of speech sources [9]. Independent component analysis (ICA) is one of the fundamental techniques to solve the cocktail party problem. The ICA algorithm was proposed by Herault and Jutten [10,11]; however, it has limitations such as permutation and scaling problems [12–14]. The IVA algorithm is an extension of the ICA algorithm which was proposed to theoretically mitigate the permutation problem of the ICA method that is inherent to most of the BSS algorithms [15].

The IVA algorithm is based on a dependency model which retains inter-frequency dependencies within each source vector. The dependent sources are arranged together as a multivariate variable in the frequency domain components of a signal. When the IVA algorithm is compared with the ICA algorithm, the inter-frequency dependencies within each source depend on the modified prior of

the source signal. In the ICA algorithm, independence for each frequency component is measured separately at each frequency bin. On the other hand, the IVA method formulates the problem by considering that the dependencies exist between frequency bins rather than assuming the independence for frequency bins. The source priors in conventional algorithms were defined as independent priors; however, within the IVA algorithm each source prior is defined as a multivariate super-Gaussian distribution. Therefore, the cost function for the IVA algorithm is minimised only when the dependency between the source vectors is removed but the dependency among the components of each vector is preserved [16]. Hence it measures the dependence across the whole multivariate source and it can retain the higher order inter-frequency dependencies and structure of frequency components. Therefore, choosing an appropriate dependency model for the IVA algorithm is crucial to the performance of the algorithm.

Statistically the process of human speech production is highly complex [4,17,18] and the human speech signal is non-stationary in nature. Furthermore, the human speech signal is difficult to model with one fixed model as there can be wide variations in human speech, i.e., properties of natural speech vary from person-to-person and depend on which language is being spoken as the pronunciation rates and phonemes can be completely different in different parts of the world. Moreover, recorded speech is dependent on variations in room acoustics and microphone characteristics, e.g., different rooms will have different reverberation effects and different microphones will have variable frequency responses [3]. All of these factors can change the observed human speech signal and thereby different speech signals generally have different statistical properties. Hence, it is important that the BSS algorithms adapt their statistical structure according to the characteristics of the observed speech signals.

The IVA algorithm preserves the inter-frequency dependency between the individual sources in the frequency domain. The IVA method uses the score function and its form is crucial to the performance of the IVA algorithm. The score function is derived by statistical modelling of the speech sources by selecting an appropriate source prior. Speech signals are often characterized with single distribution statistical models that does not change according to the nature of speech signals. In the original IVA [15] method all the speech sources were modelled by identical multivariate Laplacian distributions. Different sources can have different statistical properties and modelling all the sources with identical distribution may not be the most appropriate solution. As a result a novel potential approach is to adopt the Student's t mixture model (SMM) as a source prior for the IVA algorithm, instead of the conventional identical multivariate distributions. The probability density function of the Student's t mixture model (SMM) has heavier tails as compared to other super Gaussian models and therefore it can model outliers in the data [19–21]. As human speech is highly random, the spread of samples can be very wide, and the SMMs, due to its heavier tails, can generally model high amplitude information in human speech more accurately [22].

The new framework of the expectation-maximization (EM) algorithm is implemented efficiently for the proposed IVA algorithm to estimate the unmixing matrices. The EM algorithm is a two step iterative approach which efficiently estimates the unknown parameters of the source prior and unmixing matrices. The EM method overcomes non-analytically solvable problems and it has been commonly used in the field of statistics, signal processing and machine learning [23]. By using SMMs as the source prior and implementing the new framework of EM, the proposed IVA algorithm shows performance improvement when compared with previous approaches [15,21,24,25]. To the best of our knowledge, there are no other studies that are using SMMs as source prior for the IVA algorithm to achieve robust and improved performance for different speech mixtures in realistic scenarios.

The rest of the paper is organized as follows. We begin by explaining the related work in the next section. We begin by explaining the IVA algorithm and related work in Section 2. It is followed by detailed description of the proposed EM framework for the IVA algorithm with SMMs as source priors in Section 3. Experimental results in realistic scenarios and comparisons between the proposed

approach and other state-of-the-art methods are presented in Section 4. Finally, some concluding remarks are included in Section 5.

2. Related Work

In order to implement the IVA algorithm for the convolutive BSS, the short time Fourier transfer (STFT) is used to convert the problem from the time domain to the frequency domain as it eases the computational complexity of the time domain method. The basic noise free BSS model for the IVA method in the frequency domain can be defined as follows:

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k) \quad (1)$$

where \mathbf{A} is a mixing matrix of dimensions $m \times n$. The index k represents the k -th frequency of this multivariate method. In order to separate the source signals from the observed mixtures, an unmixing matrix must be estimated to retrieve the estimate of the original sources, as

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

where $\hat{\mathbf{s}}(k)$ is the estimated source signal, $\hat{\mathbf{s}}(k) = [\hat{s}_1, \hat{s}_2(k) \cdots \hat{s}_n(k)]^T$, $\mathbf{W}(k)$ is the unmixing matrix of dimensions $n \times m$. In this paper, focus is on the exactly determined case, so the number of sources is considered equal to the number of microphones, i.e., $n = m$.

In order to model the independence between sources, the IVA method uses the Kullback-Leibler divergence. So a cost function can be derived as follows [15]:

$$J_{IVA} = \mathcal{KL}(p(\hat{\mathbf{s}}) || \prod q(\hat{\mathbf{s}})) \quad (3)$$

$$= \int p(\hat{s}_1 \cdots \hat{s}_n) \log \frac{p(\hat{s}_1 \cdots \hat{s}_n)}{\prod q(\hat{\mathbf{s}})} d\hat{s}_1 \cdots d\hat{s}_n \quad (4)$$

$$= \text{const} - \sum_{k=1}^K \log |\det(\mathbf{W}^{(k)})| - \sum_{i=1}^n E[\log q(\hat{s}_i)] \quad (5)$$

where $\det(\cdot)$ represents the matrix determinant and $E(\cdot)$ shows the expectation operator. All the sources in the cost function of the IVA algorithm are multivariate and the cost function will be minimised when different vector sources become independent of each other and the dependency within each source vector is retained. Hence this cost function can be used to eliminate the dependency between the vector sources and preserve the frequency dependency within each vector source.

Previously, in the IVA method, speech signals have been modelled with various superGaussian distributions, e.g., the Laplacian distribution [15] or generalized Gaussian distribution [26] but speech signals can have very high and low amplitudes and other superGaussian distributions may not be able to accurately model the high amplitudes in the speech signals [27–29]. Therefore, new source priors are still needed to improve the performance of the IVA algorithm.

3. Proposed Method

In order to model the speech signals with low and high amplitude, the Student's t distribution is adopted as a source prior for the IVA method. The multivariate Student's t distribution is given as: A K -dimensional random separated source vector $\mathbf{s} = (s(1), \dots, s(K))^T$ can have a K -variate t distribution with degree of freedom ν , precision matrix $\mathbf{\Lambda}$ and mean $\boldsymbol{\mu}$, if its joint pdf is given by [27]:

$$St(\mathbf{s} | \boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \frac{\left(\frac{\nu+d}{2}\right) |\mathbf{\Lambda}|^{1/2}}{\sqrt{\nu\pi} \left(\frac{\nu}{2}\right)} \left(1 + \frac{(\mathbf{s} - \boldsymbol{\mu})^T \mathbf{\Lambda} (\mathbf{s} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}} \quad (6)$$

In the joint pdf of the Student's t distribution, the leptokurtic nature and the variance of distribution can be adjusted by tuning the degrees of freedom parameter ν [30]. When the ν parameter is set to a lower value, the tails of the distribution becomes heavier and if ν is increased to infinity, the Student's t distribution tends to a Gaussian distribution [19,20]. Since different sources can have different statistical properties, so instead of using identical Student's t source prior for all sources, the Student's t mixture model (SMM) is adopted as a source prior in this work. By adopting the SMM as a source prior, different speech sources can be modelled with different SMMs thereby enabling the IVA method to adapt to the statistical properties of different types of signals. Assuming the sources are statistically independent, for a 2×2 case, an SMM can be represented as:

$$p(\mathbf{s}(1) \cdots \mathbf{s}(K)) = \prod_{i=1}^2 p(s_i(1) \cdots s_i(K)) \quad (7)$$

$$p(s_i(1) \cdots s_i(K)) = \sum_{q_i} p(q_i) \prod_k S_i(\mathbf{s}_i(k) | \boldsymbol{\mu}_i(k), \boldsymbol{\Lambda}_i(k))$$

where q_i is the weight of the mixture component of the SMM source prior for source i and K represents the total number of frequency bins in the multivariate model. The precision matrix $\boldsymbol{\Lambda}$ is defined as the inverse of the covariance matrix and its ik -th element satisfies $1/\Lambda_i(k) = E\{|\mathbf{s}_i(k)|^2 | q_i\}$. With appropriate normalisation and zero mean assumption, the Student's t distribution can be written as:

$$St(\mathbf{s}_i(k) | 0, \boldsymbol{\Lambda}_i(k)) = \frac{\Lambda_i(k)}{\pi} \left(1 + \frac{\Lambda_i(k) |\mathbf{s}_i(k)|^2}{\nu} \right)^{-\frac{\nu+d}{2}} \quad (8)$$

When the vector of frequency components is considered from the same source i , the interdependency between these frequency components is preserved whereas the vectors that originate from different sources are independent of each other. Therefore, by adopting this inter-frequency dependency model, the IVA method prevents the permutation problem that is inherent to most BSS methods [4].

In the IVA algorithm, the scaling of mixture signal $\mathbf{x}(k)$ and mixing matrix $\mathbf{A}(k)$ cannot be determined by the separated source signals $\mathbf{s}(k)$, therefore observations are prewhitened. Because of the prewhitening process, both the mixing matrix $\mathbf{A}(k)$ and the unmixing matrix $\mathbf{W}(k)$ are unitary matrices. In this work, the 2×2 case has been considered, so the Cayley Klein parameterizations [31] for the unitary matrix $\mathbf{W}(k)$ are as follows:

$$\mathbf{W}(k) = \begin{pmatrix} a_k & b_k \\ -b_k^* & a_k^* \end{pmatrix} \therefore |\mathbf{W}(k)| = a_k a_k^* + b_k b_k^* = 1 \quad (9)$$

In the next section, the maximum likelihood estimate is derived for the IVA algorithm.

3.1. Maximum Likelihood Estimation of SMM

The maximum likelihood estimate is a well-known method that is usually used to estimate the mixture parameters. Based on the maximum likelihood method, the mixture parameters can be effectively estimated iteratively via the EM algorithm [23]. The log likelihood function for t components mixture of Student's t distributions is considered and it is given as:

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^t \log p(\mathbf{x}_i(1), \cdots, \mathbf{x}_i(K)) = \sum_{i=1}^t \log \left(\sum_{q_i} \prod_{k=1}^K p(\mathbf{x}_i(k) | q_i) p(q_i) \right) \quad (10)$$

where $\theta = \{\mathbf{W}_i, \boldsymbol{\Lambda}_i, p(q_i)\}$ consists of the model parameters for the log likelihood function; $p(\mathbf{x}_i(1), \cdots, \mathbf{x}_i(K))$ is the PDF of the observed source mixture signals which is an SMM as it is generated by the SMM source priors. The \mathbf{W}_i denotes the unmixing matrix, $\boldsymbol{\Lambda}_i$ represents the precision matrix and

$q_i = [q_1, q_2]$ is the collective mixture index of the SMMs for the source prior. In the maximum likelihood estimation, the best fitting model helps to estimate parameters that can maximize the log-likelihood function, which is usually performed by using the EM algorithm [23]. Therefore, the model parameters set $\theta = \{\mathbf{W}_i, \Lambda_i, p(q_i)\}$ can be estimated by training the SMM and maximizing the log likelihood function by using the EM algorithm. The detailed method for estimating the model parameters by the EM algorithm is explained in the next section.

3.2. The Expectation-Maximization Algorithm

The EM algorithm is suitable in finding latent parameters in probabilistic models by using an iterative optimization technique [23]. The EM algorithm is implemented by introducing discrete random variables $z(q_i)$ which are dependent on the observations $(\mathbf{x}_i(1), \dots, \mathbf{x}_i(K))$ and the model parameter set θ . The log likelihood function with these variables is given by

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \theta) &= \sum_{i=1}^t \log \left(\sum_{\mathbf{q}_i} \prod_{k=1}^K p(\mathbf{x}_i(k) | \mathbf{q}_i) p(q_i) \right) \\ &= \sum_{i=1}^t \log \left(\sum_{q_i} \frac{z(q_i) \prod_{k=1}^K p(\mathbf{x}_i(k) | q_i) p(q_i)}{z(q_i)} \right)\end{aligned}\quad (11)$$

and can be used to optimise the model parameters. In the case of an increasing log likelihood function, a lower bound is formed on the increasing log likelihood for the observations $(\mathbf{x}_i(1), \dots, \mathbf{x}_i(K))$. So the new parameters that increase the log likelihood function of the complete data with respect to current parameters, can be found. Hence there is an increase in the expected log likelihood of the complete data with respect to current parameters and it is produced by the updated parameters. Therefore, an auxiliary function can be used to represent the expected log likelihood function. There will be a definite increase in the log likelihood function when the auxiliary function is optimised but it does not necessarily yield a maximum likelihood solution [23]. Therefore, it is important to iteratively calculate and maximize the auxiliary function until convergence. Hence a local approximation is made which is the lower bound to the objective function. By using the Jensen's inequality [32], the lower bound for the log likelihood function in Equation (11) can be calculated as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \theta) &\geq \sum_{i=1}^t z(q_i) \log \left(\frac{\prod_{k=1}^K p(\mathbf{x}_i(k) | q_i) p(q_i)}{z(q_i)} \right) \\ &= \mathfrak{F}(z, \theta)\end{aligned}\quad (12)$$

The EM algorithm will run until convergence and it will iteratively maximize $\mathcal{L}(\mathbf{s}, \theta)$ in two steps. The first step is the expectation step in which the posterior probability of the hidden variable $\mathfrak{F}(z, \theta)$ is calculated over $z(x_i)$ and in the second step, the θ is updated.

3.3. The Expectation Step

In the expectation step, θ is fixed and $\mathfrak{F}(z, \theta)$ is maximised over $z(q_i)$. In order to maximise $\mathfrak{F}(z, \theta)$, the derivative of the log-likelihood equation with respect to $z(\mathbf{q}_i)$ is calculated as follows:

$$\frac{\partial}{\partial z(q_i)} (\mathcal{L}(\mathbf{x}, \theta)) = \frac{\partial}{\partial z_i} \left(\sum_{i=1}^t z(q_i) \log \frac{\prod_{k=1}^K p(\mathbf{x}_i(k) | \mathbf{q}_i) p(q_i)}{z(q_i)} \right)\quad (13)$$

In order to maximize θ for fixed $\mathfrak{F}(z, \theta)$, Equation (13) is equalised to zero and with appropriate normalization,

$$z(q_i) = \frac{\prod_{k=1}^K p(\mathbf{x}_i(k) | q_i) p(q_i)}{p(\mathbf{x}_i(1), \dots, \mathbf{x}_i(K))}\quad (14)$$

Now by using $\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k)$,

$$p(\mathbf{s}_i(k)|\mathbf{q}_i) = St(\mathbf{s}_i(k)|0, \Lambda_i(k)) \tag{15}$$

and the precision matrix for the 2×2 case can be written as:

$$\Lambda_{ik} = \mathbf{W}(k)^\dagger \Phi_i(k) \mathbf{W}(k); \Phi_i(k) = \begin{pmatrix} v_1(k) & 0 \\ 0 & v_2(k) \end{pmatrix} \tag{16}$$

As $\mathbf{W}(k)$ is a unitary matrix, therefore $\det(\Lambda_i(k)) = v_1(k)v_2(k)$ and from Equation (10), the function $f(x_i)$ can be defined as:

$$f(q_i) = \sum_k \log p(\mathbf{x}_i(k)|\mathbf{q}_i) + \log p(q_i) \tag{17}$$

By using Equation (14), function $f(x_i)$ can be rewritten as $z(x_i) \propto e^{f(x_i)}$, therefore:

$$\begin{aligned} j_i &= \sum_{s_i} e^{f(x_i)} \\ z(x_i) &= \frac{1}{j_i} e^{f(x_i)} \end{aligned} \tag{18}$$

Next, the maximization step is considered.

3.4. The Maximization Step

In the maximization step (M-step) the parameters $\theta = \{\mathbf{W}_i, \Lambda_i, p(q_i)\}$ can be estimated by maximising the cost function. In this step, each parameter is estimated separately. In the first step, the maximisation of \mathbf{W}_i over the unitary constraint is considered. In order to maximize the \mathbf{W}_i , the precision matrix for the 2×2 case can take the following form:

$$\Phi_{ik} = \begin{pmatrix} v_1(k) - v_2(k) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} v_2(k) & 0 \\ 0 & v_2(k) \end{pmatrix} \tag{19}$$

When Equation (15) is rearranged and $p(\mathbf{x}_i(k)|\mathbf{q}_i)$ is replaced in the log likelihood Equation (12), it will take the following form:

$$= \sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ (-\nu/2 - d/2) \log \left(1 + \frac{\Lambda_i(k) | \mathbf{x}_i(k) |^2}{\nu} \right) \right\} \tag{20}$$

By using the log approximation $\log(1 + a) \approx a$, where a is a small value, the above mentioned equation can take the following form, wherein equality is assumed for convenience.

$$= \sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ (-\nu/2 - d/2) \left(\frac{\Lambda_i(k) | \mathbf{x}_i(k) |^2}{\nu} \right) \right\} \tag{21}$$

By further manipulating the above mentioned equation, details of which are given in Appendix A, the parameters of the unmixing matrix can be calculated as follows:

$$\begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} = \frac{1}{\sqrt{1 + \left(\frac{\beta_k - M_{11}}{M_{12}}\right)^2}} \begin{pmatrix} 1 \\ \frac{\beta_k - M_{11}}{M_{12}} \end{pmatrix} \tag{22}$$

Since the unmixing matrix $\mathbf{W}_i(\mathbf{k}) = \begin{pmatrix} a_k & b_k \\ -b_k^* & a_k^* \end{pmatrix}$, it can be estimated by using the above mentioned analytical solution. It is an efficient method to estimate the unmixing matrix as the above mentioned method avoids the matrix calculations.

The model parameters $\theta = \{\mathbf{W}(k), \Lambda(k), p(q_i)\}$ are estimated by maximizing the log likelihood function. Therefore, now $\mathfrak{F}(z, \theta)$ will be maximized over Λ_{iK} and $p(q_i)$ and they are given as follows.

$$\frac{1}{v_{ik_{j=r}}} = \left(\frac{-v/2 - d/2}{v} \right) \frac{\left[\sum_{i=1}^t z(q_{ij=r}) (\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \mathbf{W}(k) \mathbf{x}_i(k)) \right]_{jj}}{\sum_{i=1}^t z(q_{i=jr})} \tag{23}$$

where $[.]_{jj}$ denotes the (j, j) element of the matrix. So $\mathfrak{F}(z, \theta)$ over Λ_i using the above mentioned solution.

Now, maximisation of $\mathfrak{F}(z, \theta)$ over $p(q_i)$ is performed and it is given as:

$$p(q_i = r) = \frac{\sum_{i=1}^t z(q_i = r)}{t} \tag{24}$$

Hence the weighting parameter can be calculated by using the above mentioned equation. Detailed derivation for maximization of θ is included in Appendix A.

It can be seen that the EM algorithm effectively estimates all the model parameters $\theta = \{\mathbf{W}_i, \Lambda_i, p(q_i)\}$. The E-step updates the $z(q_i)$, while the M-step effectively estimates the model parameters. In the EM algorithm the degrees of freedom parameter v is fixed in advance for all the sources, then the M-step exists in the closed form (Algorithm 1). The value for the degrees of freedom can be estimated empirically for different source signals. The complete EM framework for the SMIVA algorithm is summarized as follows.

Algorithm 1 EM algorithm for Student's t Mixtures

Require: Given a Student's t mixture model, the aim is to maximize the log likelihood function with respect to the parameters $\theta = \{\mathbf{W}_i, \Lambda_i, p(q_i)\}$.

- 1: Initialize the model parameters, the unmixing matrix \mathbf{W}_i , the precision Λ_i and the weight coefficients $p(q_i)$ and evaluate the initial value of the log likelihood.
- 2: **Expectation Step:** Evaluate the probabilities using the current parameter values

$$z(x_i) = \frac{\prod_{k=1}^K p(\mathbf{x}_i(k)|q_i)p(q_i)}{p(x_i(1), \dots, x_i(K))} \tag{25}$$

- 3: **Maximization Step:** Using the current probabilities, re-estimate the parameters

- Coefficients of the unmixing matrices \mathbf{W}_i are estimated by:

$$\begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} = \frac{1}{\sqrt{1 + \left(\frac{\beta_k - M_{11}}{M_{12}}\right)^2}} \begin{pmatrix} 1 \\ \frac{\beta_k - M_{11}}{M_{12}} \end{pmatrix} \tag{26}$$

- Coefficients of the precision matrix Λ_i are estimated by

$$\frac{1}{v_{ik_{j=r}}} = \left(\frac{-v/2 - d/2}{v} \right) \frac{\left[\sum_{i=1}^t z(x_{ij=r}) (\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \mathbf{W}(k) \mathbf{x}_i(k)) \right]_{jj}}{\sum_{i=1}^t z(x_{i=jr})} \tag{27}$$

- The weighting coefficients can be estimated as

$$p(q_i = r) = \frac{\sum_{i=1}^t z(q_i = r)}{t} \tag{28}$$

Algorithm 1 *Cont.*

4: Evaluate the log likelihood

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^I \log \left(\sum_{x_i} \prod_{k=1}^K p(\mathbf{x}_i(k) | q_i) p(q_i) \right) \quad (29)$$

and check for convergence of the log likelihood function, if the criterion for convergence is not fulfilled, return to step 2.

The separation performance of this EM framework for the SMIVA method will be evaluated in the next section.

4. Experimentations and Results

In this section, the separation performance of the SMIVA algorithm will be tested in three different experimental setups. Firstly, the new framework for the IVA algorithm is tested in a simulated environment and then in order to evaluate the performance in real scenarios, it is tested with real room impulse responses (RIRs), which can depict the performance of the proposed method in changing realistic settings. The results from all three sets of experiments for the proposed algorithm will be compared with the original IVA algorithm with different source priors.

4.1. Case I: Simulations with the Image Method

Firstly, the proposed method will be tested with RIRs that were generated by using the image method. The speech signals were selected randomly from the whole of the TIMIT dataset [33] and the length of speech signals was approximately 4 s. A 2×2 case was considered and the room has the $RT_{60} = 200$ ms and it provides a good setup for comparing the separation performance of different algorithms. The positions of microphones in the room were set to [3.44, 2.50, 1.50] and [3.48, 2.50, 1.50] with azimuth angles of 60° and 30° , respectively with reference to the normal of the microphone position. The STFT length is 1024 and sampling frequency is 8 kHz. The separation performance of the algorithm was evaluated with the objective measure of SDR [34]. The common parameters used in these experiments are summarised in Table 1.

Table 1. Summary of parameters used in experiments.

Sampling rate	8 kHz
STFT frame length	1024
Reverberation time	200 ms
Room dimensions	7 m \times 5 m \times 3 m
Source signal duration	4 s (TIMIT)
Room impulse responses	Image method
Objective measure	Signal to Distortion Ratio (SDR)

The speech signals were convolved into mixtures in the above mentioned room settings. These speech mixtures were then separated by using the proposed SMIVA method and the separation results for different mixtures were compared with the separation performance of the original IVA method with the original super Gaussian source prior [15] and also with the IVA method with Student's t source prior [21] and the results are shown in Table 2 and all the values shown for SDR are in dB. For each mixture SDR performance shown in the Table 2 is the average of two speech signals. It is evident from the Table 2 that when the SMM is adopted as a source prior, the average SDR improvement is approximately 1.1 dB for all the mixtures as compared to the original super Gaussian source prior for the IVA method. It is evident from Table 2 that the SMM source prior based SMIVA algorithm enhances the separation performance of the IVA method with a single distribution source prior, such as the Student's t distribution and also the original super Gaussian.

Table 2. SDR (dB) values for different source priors for the IVA method with an image room impulse response [35]. The SMM source prior shows improvement for all mixtures.

	Original Super Gaussian	Student's t Distribution	SMM Source Prior
Set-1	9.09	9.84	10.27
Set-2	8.98	9.72	10.24
Set-3	9.26	10.11	10.87
Set-4	9.02	9.95	10.49
Set-5	9.53	10.21	10.62
Set-6	9.51	10.14	10.74
Set-7	8.91	9.67	10.09
Set-8	9.86	10.48	11.05
Set-9	9.94	10.66	11.24
Set-10	10.02	10.56	11.01

For the derivation purposes, the unmixing matrix \mathbf{W} is assumed to be unity in this study. In order to measure the effect of this assumption on the performance of the SMIVA algorithm, its separation performance is compared with the algorithms that do not place the same restriction on the solution such as [36,37]. Since the latent variables are modelled with a single distribution in these methods, the SMIVA algorithm is also adjusted to a uni-model case by using a single Student's t distribution. For the sake of consistency in results, same experimental settings are used as in the previous case and the SDR measure is used to estimate the separation performance of the algorithms and the results are shown in Figure 1. It is clear from results that the SMIVA algorithm with the assumptions placed on the solution space still consistently perform better than the algorithms that does not have the same assumptions on the solution space. These results shows the significance of the modelling of the high and low amplitude information within the speech signals by using the Student's t distribution.

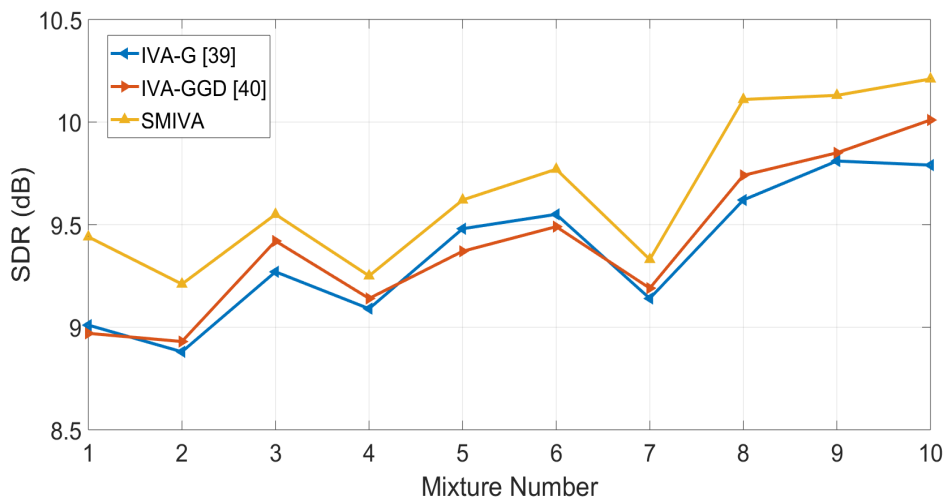


Figure 1. SDR (dB) values for algorithms without restriction on the solution space and SMIVA algorithm. The SMIVA algorithm shows better separation for all mixtures.

In order to further investigate the separation performance of the SMIVA algorithm, its separation results are compared with the other mixture model source prior such as Gaussian mixture model [24] in the next section with real RIRs.

4.2. Case II: Simulations with Real RIRs

In the second set of experiments, the proposed SMIVA algorithm is tested with real RIRs. These real RIRs are obtained from [38] and these are recorded in different rooms with different

acoustic properties. Three different room types (A, B,D) have been used with RT_{60} of 320 ms, 470 ms and 890 ms, respectively. By using these RIRs the proposed method can be tested with the range of reverberation time. Therefore, these simulations show the performance of the proposed algorithm in real life scenarios as the RT_{60} can vary drastically in realistic environments [39]. Different source location azimuth angles are available, which range from 15° to 90° , relative to the second source.

Firstly, the proposed algorithm is tested in Room A, which is a typical medium sized office and it has the RT_{60} of 320 ms, which is relatively small for a medium size office. In the experiments two speech signals are randomly chosen from the whole of the TIMIT dataset and the source location azimuth angles are set to be from (15° to 90° with a step of 15°). The mixed sources are separated by using the proposed SMIVA method and the separation performance in terms of SDR is compared with the IVA using the identical Student's t source prior [21] and also with the original super Gaussian source prior based IVA method [15]. The separation performance for both methods is evaluated for six different angles varying from 15° to 90° with a step of 15° . At all the angles separation performance is averaged over six different speech mixture and the results are presented in Figure 2.

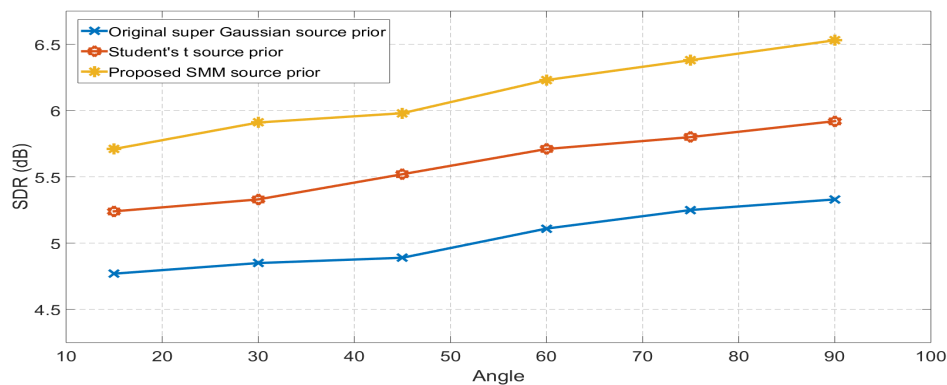


Figure 2. Comparison between original IVA with original super Gaussian source prior, Student's t source prior and EM framework IVA with SMM source prior for Room-A ($RT_{60} = 320$ ms). The separation performance at each angle is averaged over six different speech mixtures. The proposed mixture model IVA performs better than a single Student's t distribution at all the separation angles.

It is evident from Figure 2 that when proposed algorithm is used to separate the mixtures and the performance is compared with identical distribution source prior for the IVA, it consistently has a better separation performance at all the selected azimuths angles and approximately 1 dB of improvement in SDR values is recorded at all the angles as compared to the original IVA method [15].

The same set of experiments are then repeated for room-B, which evaluate the performance of the algorithms in more challenging and realistic scenarios as it is a medium size classroom with RT_{60} of 470 ms. The separation performance in terms of SDR of both methods for six different azimuth angles is showed in Figure 3. It is evident from the result that the EM framework SMIVA performs better than the identical source priors for the original IVA method at all separation angles in this reverberant real room environment.

Finally, the separation performance of the proposed EM framework for the IVA method is evaluated in a highly reverberant realistic environment that can depict the performance of the algorithm in the real life scenarios. For the highly reverberant environment, Room D was used which is a medium size seminar and presentation hall with a very high ceiling. The RT_{60} for this seminar hall is 890 ms, which is high reverberation time and therefore it provides a good insight into the algorithm's performance in an extremely difficult real life situations. The experimental setup in this highly reverberant room D is similar to the previous two rooms. The mixtures were separated with the IVA method with different source priors and the separation performance in terms of SDR for all methods is shown in Figure 4 for all six angles varying from 15° to 90° . The SDR values in room D

is lower in comparison with the SDR values for Room-A and Room-B, it is mainly because the RT_{60} for Room D is really high as compared to the other two rooms. Also, it is evident from the Figure 4 that even in this highly reverberant environment the IVA method with SMM source prior performs better than the identical distribution source priors for the original IVA with Student's t source prior and improves the average SDR performance of the original IVA method by approximately 0.8 dB.

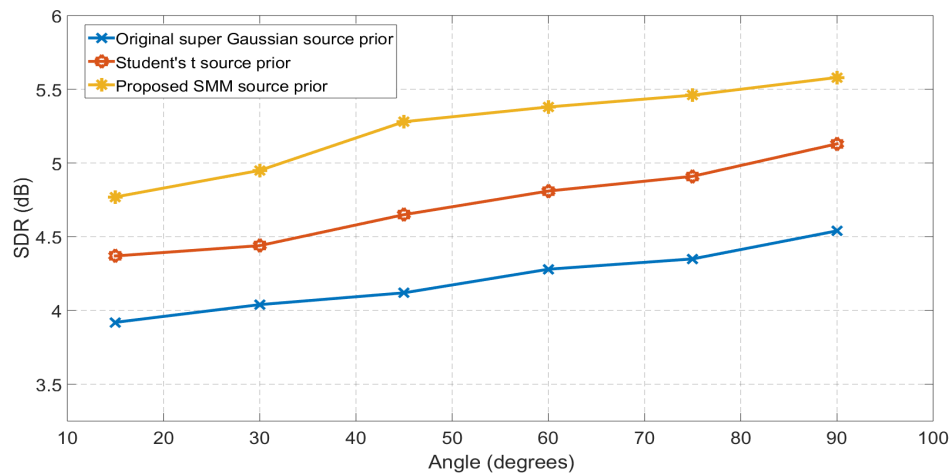


Figure 3. Comparison between original IVA with Student's t source prior and EM framework IVA with SMM source prior for Room-B ($RT_{60} = 470$ ms). The separation performance at each angle is averaged over six different speech mixtures. The proposed mixture model IVA perform better than single Student's t distribution at all the separation angles.

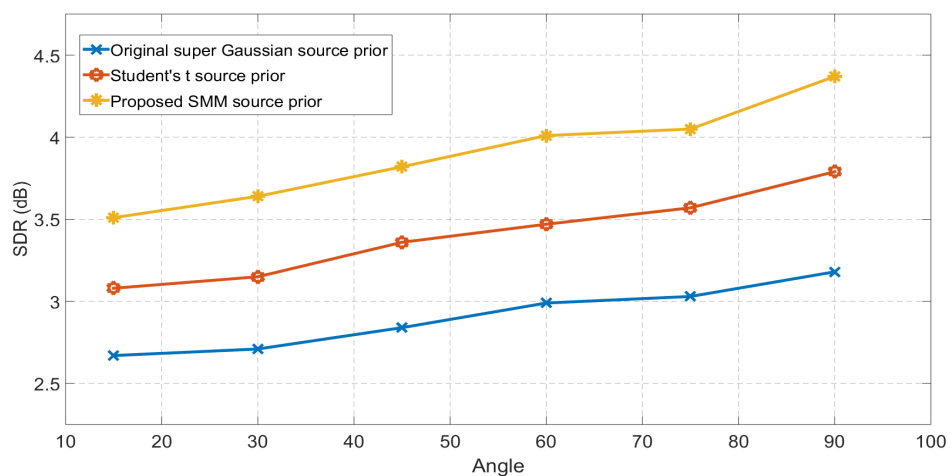


Figure 4. Comparison between Original IVA with Student's t source prior and EM framework IVA with SMM source prior for Room-D ($RT_{60} = 890$ ms). The separation performance at each angle is averaged over six different speech mixtures. The proposed mixture model IVA performs better than single Student's t distribution at all the separation angles.

The separation performance of the proposed EM framework for the IVA algorithm with SMMs as a source prior is also compared with the IVA algorithm with a Gaussian mixture model (GMM) as a source prior [24]. Since the mixture model is adapted as a source prior for the IVA algorithm, the comparison with other mixture models, i.e., GMM can provide better understanding of the separation performance of the proposed source prior. In this set of experiments, a mixture of two Gaussians is adapted as source prior for the IVA algorithm and the rest of the parameters are adjusted similar to the SMIVA algorithm and simulations are performed in the real room settings. For the experimental setup, same settings for Room A, B and D are used as in the previous case and speech

signals are randomly chosen from TIMT dataset. Initially, experiments are performed in room A, which has RT_{60} of 320 ms and it is repeated for six different source location varying from 15° to 90° . Similarly, the same experimental setup is used for room B with RT_{60} of 470 ms and for room D with RT_{60} of 890 ms. In all the rooms mixtures are separated by using the EM framework IVA with both SMM and GMM source priors and the separation performance in terms of SDR is compared with the proposed method at six different source azimuth angles varying from 15° to 90° . All the SDR values at all the angles are the average of separation performance of six different mixtures. The separation performance of both methods for all three rooms with the range of RT_{60} is shown in Figure 5 and it is evident that the IVA method with SMM as a source prior has better separation performance than IVA with GMM as a source prior.

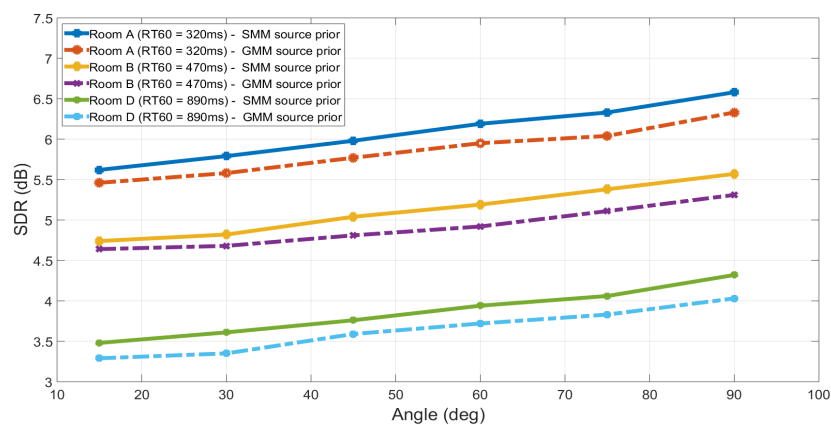


Figure 5. Comparison between EM framework IVA with SMM and GMM source prior for three different rooms (Room-A, Room-B, Room-D). The separation performance at each angle is averaged over six different speech mixtures. The EM framework IVA algorithm with proposed SMM source prior performs better than GMM source prior at all the separation angles.

4.3. Case III: Simulations with Binaural Room Impulse Responses

The proposed algorithm is further tested with binaural room impulse response (BRIRs) obtained from [40]. These BRIRs are recorded in a real classroom which roughly has dimensions of $5 \times 9 \times 3.5 \text{ m}^3$. The six source location azimuths ($15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$) relative to the right of listener were used for the experimentation. Also the distance between the source were changed three times (0.15, 0.40 and 1 m). The measurements for the BRIRs are taken at four different listener locations (back, ear, corner and center) and the distance between the floor and ears was approximately 1.50 m. In these experiments only center location is used and the RT_{60} at the center location for the classroom was 565 ms. All the measurements are repeated on three different occasions by taking down the equipment and reassembling it which improves the reliability of the measurements. Therefore, these BRIRs have been used in the experiments as they are reliable and also provide accurate estimate of the separation performance of the BSS algorithms in highly reverberant room environments. A summary of different parameters used in this set of experiments is given in Table 3.

Table 3. Summary of parameters used in experiments.

Sampling rate	8 kHz
STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m \times 5 m \times 3.5 m
Source signal duration	3.5 s (TIMIT)

The 2×2 case was considered for the experiments and speech signals were randomly chosen from the whole TIMIT dataset and mixtures were created by using BRIRs. The length of the speech signals was approximately four seconds. The speech signals were then separated from the mixtures by using the proposed EM framework for the IVA algorithm with SMM as source prior. The separation performance of the proposed algorithm is compared with the separation performance of the IVA with GMM as source prior for the IVA algorithm. It provides a good estimate for the separation performance of the proposed algorithm and source prior as comparison is drawn with mixture model source priors. The separation performance in terms of SDR is shown in Table 4 for the six different source location (15° , 30° , 45° , 60° , 75° , 90°). All the experiments are repeated three times and at each source location six different speech mixtures are separated. In order to improve the reliability of results, all the SDR values are the average of separation performance of the algorithms over eighteen different speech mixture.

Table 4. Comparison between SMM source prior and GMM source prior for the EM framework IVA algorithm with BRIRs ($RT_{60} = 565$ ms). The separation performance at each angle is averaged over eighteen different speech mixtures. The IVA algorithm with proposed mixture model Student's t source prior perform significantly better at all the separation angles than the GMM source prior.

	GMM Source Prior	SMM Source Prior	Percentage Improvement
Angle- 15°	4.51	4.82	6.87%
Angle- 30°	4.62	4.97	7.56%
Angle- 45°	4.77	5.09	6.70%
Angle- 60°	4.97	5.32	7.04%
Angle- 75°	4.91	5.28	7.73%
Angle- 90°	4.77	5.12	7.34%

From Table 4 it is evident that when the SMM is used as a source prior for the IVA algorithm it performs better as compared with the GMM as a source prior. Since speech signals are highly non-stationary in nature and there can be many useful samples in outliers which might not be properly modelled with the Gaussian mixtures but Student's t mixtures because of its heavy tails can model the outlier information and therefore enhance the separation results of the IVA method. When the SMM is adopted as a source prior for the IVA method, at all the source location azimuths it improves the average separation performance for the IVA method by more than 6% for all the angles as shown in Table 4.

Furthermore, the separation performance is evaluated with the subjective measure of PESQ [41]. This subjective measure compares the original signals and separated signals and gives a score from 0 to 4.5, 0 for the poor separation performance and 4.5 being the excellent separation performance. This measure therefore provides a good estimate about the similarity between the original and separated sources. So the speech mixtures made with BRIRs are separated with the proposed SMM source prior for the EM framework IVA and also with the GMM source prior IVA and the PESQ score is calculated for both the methods. The PESQ score for the IVA method with both source priors is shown in Table 5 and the IVA method with SMM source prior consistently has the better PESQ score as compared with the GMM source prior for the IVA algorithm. The PESQ score was generally low in this experiment because of difficult room environment. However, it is evident from the table that when SMM is adapted as a source prior, it improves the separation performance for the IVA method.

Table 5. PESQ score for GMM and SMM source prior for the IVA algorithm.

	GMM Source Prior	SMM Source Prior
Set-1	1.85	2.02
Set-2	1.98	2.11
Set-3	1.96	2.13
Set-4	2.02	2.19
Set-5	1.93	2.14
Set-6	2.08	2.21

Finally, the separation performance of the proposed EM framework for the IVA method with SMM as source prior is compared with the original IVA with identical source priors. BRIRs with RT_{60} of 565 ms are used to evaluate the algorithms in highly reverberant environment that can depict the performance of the algorithms in the realistic scenarios. The Same experimental settings are used as in the first experiments and the source location is varied six times from ($15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$). All the measurements are repeated three times and six different speech mixtures are separated at each angle by using IVA method with SMM as source prior and the results are compared with the separation performance of IVA method with multivariate Student's t distribution as source prior, the IVA method with original multivariate super Gaussian source prior and also with IVA method with the mixed Student's t and original super Gaussian source prior. This provides an overall comparison of the separation performance of different source prior and the framework for the IVA method. The results in terms of SDR (dB) for six different source location are shown in Figure 6.

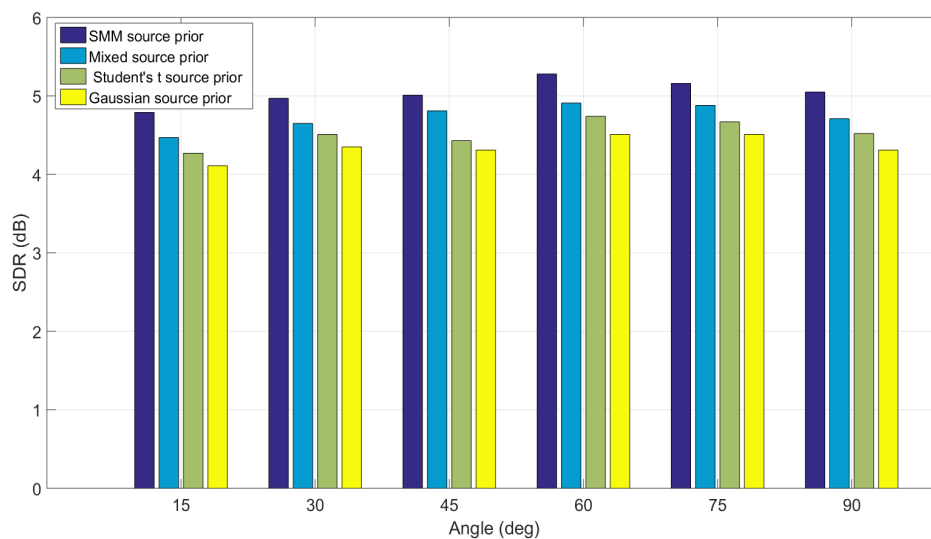


Figure 6. Comparison between different source priors for the IVA algorithm for BRIRs ($RT_{60} = 565$ ms). The separation performance at each angle is averaged over eighteen different speech mixtures. The IVA algorithm with proposed mixture model Student's t source prior perform better at all the separation angles in comparison to identical source prior for all the sources.

It is evident from Figure 6 that the mixture model source prior performs better than the identical distribution source prior at all the source locations. Since different speech sources can have different statistical properties and the mixture model such as SMM source prior can model different sources with different Student's t distribution in the mixture model while identical source prior model all the sources with the identical distribution and therefore their separation performance suffers as compared to the mixture model source priors.

5. Conclusions

This work presented the EM framework for the IVA method that uses the mixture of Student's t distribution as a source prior in order to better model the different statistical properties in different speech sources. The mixture of Student's t distribution source prior made use of the heavy tails nature of the Student's t distribution to effectively model the high amplitude information in the speech signal. The complete EM framework was derived efficiently to estimate the model parameters for the IVA method. The separation performance for the proposed method was tested with image room impulse method and it confirms the advantage of using the proposed framework for the IVA method as it shows the SDR improvement of approximately 1 dB as compared to the original IVA method. Further experiments were conducted in real room environments with different reverberation times. In varying reverberant environment, new SMIVA consistently performed better as compared to other source priors and provides average SDR improvement of 0.8 dB. In order to further test the new SMIVA algorithm, it was tested with BRIRs and the SMM source prior improves the separation performance as compared to the GMM source prior and in all case, improvement of more than 6% in SDR performance was recorded. This improvement in the separation performance can be further verified and used by implementing the algorithm in real time practical scenarios, which remain the topic for future work in this study. All the simulations performed with different real room environments confirmed that the proposed EM framework for the IVA algorithm that make use of the SMM source prior improves the separation performance even in highly reverberant real room environments.

Author Contributions: W.R.—Methodology, investigation, experimentation, validation, writing, reviewing and editing; J.C.—Methodology, reviewing, supervision and resources; A.I.S.—writing, reviewing and editing.

Funding: This research received no external funding

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The EM framework for SMIVA

The log likelihood function for t components mixture of Student's t distributions is given in Equation (10) as:

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^t \log p(\mathbf{x}_i(1), \dots, \mathbf{x}_i(K)) = \sum_{i=1}^t \log \left(\sum_{q_i} \prod_{k=1}^K p(\mathbf{x}_i(k)|q_i)p(q_i) \right) \quad (\text{A1})$$

Expectation step is already derived and is given by Equation (18):

$$\begin{aligned} j_i &= \sum_{s_i} e^f(x_i); \\ z(x_i) &= \frac{1}{j_i} e^f(x_i) \end{aligned} \quad (\text{A2})$$

Next the maximization step is derived in detailed. The MATLAB implementation for the SMIVA algorithm is available online (<https://github.com/wr27/The-IVA-algorithm-with-SMM-source-prior>).

The maximization step (M-step) the parameters $\theta = \{\mathbf{W}_i, \Lambda_i, p(q_i)\}$ are estimated by maximising the cost function. Firstly, the maximisation of \mathbf{W}_i over the unitary constraint is considered. The precision matrix can be written in the following form

$$\Phi_{ik} = \begin{pmatrix} v_1(k) - v_2(k) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} v_2(k) & 0 \\ 0 & v_2(k) \end{pmatrix} \quad (\text{A3})$$

Now Equation (15) can be rearranged as:

$$p(\mathbf{x}_i(k)|\mathbf{q}_i) = S_t(\mathbf{x}_i(k)|0, \Lambda_i(k)) = \frac{\Lambda_i(k)}{\pi} \left(1 + \frac{\Lambda_i(k) \|\mathbf{x}_i(k)\|^2}{\nu} \right)^{-\nu/2-d/2} \tag{A4}$$

Now by replacing the value of the precision $\Lambda_i(k) = \mathbf{W}(k)^\dagger \Phi_i(k) \mathbf{W}(k)$ in the above equation

$$= \sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ (-\nu/2 - d/2) \left(\frac{\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \Phi_{ik} \mathbf{W}(k) \mathbf{x}_i(k)}{\nu} \right) \right\} \tag{A5}$$

The above equation can be rewritten by replacing the value of Φ_{ik} from the Equation (A3) as follows:

$$= - \sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ \frac{(\nu/2 + d/2)}{\nu} \left(\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \begin{pmatrix} v_1(k) - v_2(k) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} v_2(k) & 0 \\ 0 & v_2(k) \end{pmatrix} \mathbf{W}(k) \mathbf{x}_i(k) \right) \right\} \tag{A6}$$

After appropriate manipulation and ignoring the constant terms, Equation (A6) takes the following form:

$$= - \sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ \frac{(\nu/2 + d/2)}{\nu} (v_1(k) - v_2(k)) \left(\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \mathbf{W}(k) \mathbf{x}_i(k) \right) \right\} + \beta_k (a_k a_k^* + b_k b_k^* - 1) \tag{A7}$$

Now by replacing the value of $\mathbf{x}_i(k)$ and $\mathbf{W}(k)$ for the 2×2 case:

$$= - \sum_{i=1}^t z(q_i) \{ \lambda \} \frac{(\nu/2 + d/2)}{\nu} (v_1(k) - v_2(k)) \left\{ \begin{pmatrix} \mathbf{x}_1(k) & \mathbf{x}_2(k) \end{pmatrix} \begin{pmatrix} a_k & -b_k^* \\ b_k & a_k^* \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a_k & b_k \\ -b_k^* & a_k^* \end{pmatrix} \begin{pmatrix} \mathbf{x}_1(k) \\ \mathbf{x}_2(k) \end{pmatrix} \right\} + \beta_k (a_k a_k^* + b_k b_k^* - 1) \tag{A8}$$

After the matrix multiplication, the previous equation takes the following form:

$$= - \sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ \frac{(\nu/2 + d/2)}{\nu} (v_1(k) - v_2(k)) (a_k \mathbf{x}_1(k) + b_k \mathbf{x}_2(k))^2 \right\} + \beta_k (a_k a_k^* + b_k b_k^* - 1) \tag{A9}$$

Now by taking the derivative of above mentioned equation with respect to a_k and equalising to zero.

$$\sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ \frac{(\nu/2 + d/2)}{\nu} (v_1(k) - v_2(k)) \begin{pmatrix} \mathbf{x}_i(k) & \mathbf{x}_i(k) \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} \right\} = a_k^* \beta_k \tag{A10}$$

Likewise, taking the derivative with respect to b_k and equating it to zero

$$\sum_{i=1}^t z(q_i) \{ \lambda \} \left\{ \frac{(\nu/2 + d/2)}{\nu} (v_1(k) - v_2(k)) \begin{pmatrix} \mathbf{x}_i(k) & \mathbf{x}_i(k) \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} \right\} = b_k^* \beta_k \tag{A11}$$

Assuming $\mathbf{M}_{ik} = \sum_{i=1}^t z(s_i) \frac{(v/2 + d/2)}{v} (v_1(k) - v_2(k)) \mathbf{x}_i(k) \mathbf{x}_i(k)^\dagger$ and by Equations (A10) and (A11):

$$\mathbf{M}_{ik} \begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} = \beta_k \begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} \tag{A12}$$

where vector $(a_k, b_k)^\dagger$ is the eigenvector of \mathbf{M}_{ik} with the smaller eigenvalue. This can be found by replacing \mathbf{M}_{ik} in Equation (A7) and taking trace of the equation:

$$-Tr \left\{ \mathbf{M}_{ik} \begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} \begin{pmatrix} a_k & b_k \end{pmatrix}^\dagger \right\} = \beta_k \tag{A13}$$

where $Tr(x)$ denotes the trace of the matrix. Whereas the eigenvectors associated with the smaller eigenvalues will give the higher value of the cost function. Therefore, $(a_k, b_k)^\dagger$ is the eigenvector of \mathbf{M}_{ik} with the smaller eigenvalue. In order to calculate the eigenvalues associated with the \mathbf{M}_{ik} for the 2×2 case, \mathbf{M}_{ik} can be written as:

$$\mathbf{M}_{ik} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \tag{A14}$$

where M_{11}, M_{22} are real and $M_{21} = M_{12}^*$, because \mathbf{M}_{ik} is Hermitian. Eigenvalues in this case can be calculated as:

$$\det \left\{ \begin{pmatrix} M_{11} - \lambda & M_{12} \\ M_{21} & M_{22} - \lambda \end{pmatrix} \right\} = 0 \tag{A15}$$

$$(M_{11} - \lambda)(M_{22} - \lambda) - (M_{12})(M_{21}) = 0 \tag{A16}$$

Since the above equation is a quadratic equation, therefore the quadratic formula can be used to find the eigenvalues which are $\frac{M_{11} + M_{22}}{2} \pm \sqrt{\frac{(M_{11} - M_{22})^2}{4} + |M_{12}|^2}$, so the smaller eigenvalue can be written as:

$$\beta_k = \frac{M_{11} + M_{22}}{2} - \sqrt{\frac{(M_{11} - M_{22})^2}{4} + |M_{12}|^2} \tag{A17}$$

Hence, by using this analytical approach, computational complexity can be reduced as this approach avoids any matrix calculation.

In the next step, $\mathfrak{F}(z, \theta)$ will be maximized over Λ_{iK} Firstly $p(\mathbf{x}_i(k)|\mathbf{q}_i)$ is replaced in Equation (12) as follows:

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^t z(q_i) \log \left(\frac{\prod_{k=1}^K \frac{|\Lambda_i(k)|}{\pi} \left(1 + \frac{\mathbf{x}_i(k)^\dagger \Lambda_i(k) \mathbf{x}_i(k)}{v} \right)^{-v/2-d/2} p(q_i)}{z(q_i)} \right) \tag{A18}$$

$$= \sum_{i=1}^t z(q_i) \left\{ \log \left(\frac{|\Lambda_i(k)|}{\pi} \right) + \left\{ (-v/2 - d/2) \log \left(1 + \frac{\mathbf{x}_i(k)^\dagger \Lambda_i(k) \mathbf{x}_i(k)}{v} \right) \right\} \right\} \tag{A19}$$

Again, by using the log approximation $\log(1 + a) \approx a$, where a is a small value, the above equation can take the following form, wherein equality is consider for convenience.

$$= \sum_{i=1}^t z(q_i) \left\{ \log \left(\frac{|\Lambda_i(k)|}{\pi} \right) + \left\{ (-v/2 - d/2) \left(\frac{\mathbf{x}_i(k)^\dagger \Lambda_i(k) \mathbf{x}_i(k)}{v} \right) \right\} \right\} \tag{A20}$$

By replacing the value of $\Lambda_i(k)$ in the Equation (A20) and by taking the derivative with respect to v_{k1} yields

$$= \sum_{i=1}^t z(q_i) \left\{ \left(\frac{1}{v_1(k)} \right) - \left\{ \frac{(-v/2 - d/2)}{v} \left(\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \mathbf{W}(k) \mathbf{x}_i(k) \right) \right\} \right\} \quad (\text{A21})$$

After appropriate manipulation, coefficient of precision matrix are estimated as follows.

$$\frac{1}{v_{ik_{j=r}}} = \left(\frac{-v/2 - d/2}{v} \right) \frac{\left[\sum_{i=1}^t z(q_{ij=r}) (\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \mathbf{W}(k) \mathbf{x}_i(k)) \right]_{jj}}{\sum_{i=1}^t z(q_{ij=r})} \quad (\text{A22})$$

where $[\cdot]_{jj}$ denotes the (j, j) element of the matrix. So $\mathfrak{F}(\mathbf{z}, \theta)$ over Λ_i using the above mentioned solution.

Now in the final step, maximisation of $\mathfrak{F}(\mathbf{z}, \theta)$ over $p(q_i)$ is performed. The lower bound of the log likelihood equation is:

$$\mathfrak{F}(\mathbf{z}, \theta) = \sum_{i=1}^t z(q_i) \log \left(\frac{\prod_{k=1}^K p(\mathbf{x}_i(k)|q_i) p(q_i)}{z(q_i)} \right) \quad (\text{A23})$$

If \mathbf{q}_i can take s possible states, $p(q_i) = r$ has to satisfy $\sum_r p(\mathbf{q}_i = r) = 1$. So $p(q_i) = r$ does not have s degrees of freedom instead it has $q - 1$ free parameters. So, the Lagrange multiplier is used in this case. Therefore, the cost function can be described as:

$$\sum_{i=1}^t z(q_i) \log \left(\frac{p(\mathbf{x}_i(k)|x_i) p(q_i)}{z(q_i)} \right) + \beta \left(1 - \sum_{i=1}^t (p(q_i = r)) \right) \quad (\text{A24})$$

Now taking the derivative of the above mentioned equation with respect to $p(q_i = r)$ and equating it to zero yields:

$$\sum_{i=1}^t z(q_i = r) \left(\frac{1}{p(q_i = r)} \right) - \beta = 0 \quad (\text{A25})$$

$$p(q_i = r) = \frac{\sum_{i=1}^t z(q_i = r)}{\beta} \quad (\text{A26})$$

Now $p(q_i = r) = 1$ and $\sum_{i=1}^t z(q_i = r) = 1$, therefore the above equation can be rewritten as:

$$1 = \frac{\sum_{i=1}^t 1}{\beta} \Rightarrow 1 = \frac{t}{\beta} \Rightarrow \beta = t \quad (\text{A27})$$

Hence the weighting parameter can be calculated by using the above mentioned equation. It can be seen that the EM algorithm effectively estimates all the model parameters $\theta = \{\mathbf{W}_i, \Lambda_i, p(q_i)\}$. The E-step updates the $z(q_i)$, while the M-step effectively estimates the model parameters. The above mentioned E-step and M-step will keep iterating until convergence criteria is fulfilled.

References

1. Haykin, S. *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*; Wiley: Hoboken, NJ, USA, 2000.
2. Cherry, C. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **1953**, *25*, 975–979. [[CrossRef](#)]
3. Haykin, S.; Chen, Z. The cocktail party problem. *Neural Comput.* **2005**, *17*, 1875–1902. [[CrossRef](#)] [[PubMed](#)]
4. Cichocki, A.; Amari, S. *Adaptive Blind Signal and Image Processing*; John Wiley: Hoboken, NJ, USA, 2002.
5. McDermott, J.H. The cocktail party problem. *Curr. Biol.* **2009**, *19*, R1024–R1027. [[CrossRef](#)] [[PubMed](#)]
6. Wang, D.; Brown, G. Fundamentals of computational auditory scene analysis. In *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*; John Wiley and Sons: Hoboken, NJ, USA, 2006; pp. 1–44.
7. Adali, T.; Anderson, M.; Geng-Shen, F. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. *IEEE Signal Process. Mag.* **2014**, *31*, 18–33. [[CrossRef](#)]

8. Parra, L.; Alvino, C. Geometric source separation: merging convolutive source separation with geometric beamforming. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 352–362. [[CrossRef](#)]
9. Pedersen, M.S.; Larsen, J.U.; Kjems, U.; Parra, L.C. A survey of convolutive blind source separation methods. *Springer Handb. Speech Process. Speech Commun.* **2007**, *8*, 1–34.
10. Jutten, C.; Herault, J. Blind Separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process.* **1991**, *24*, 1–10. [[CrossRef](#)]
11. Jutten, C.; Comon, P. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*; Academic Press: Cambridge, MA, USA, 2010.
12. Lee, T.W. *Independent Component Analysis: Theory and Applications*; Kluwer Academic: Norwell, MA, USA, 2000.
13. Hyvriinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **1999**, *10*, 626–634. [[CrossRef](#)]
14. Parra, L.; Spence, C. Convolutive blind separation of non-stationary sources. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 320–327. [[CrossRef](#)]
15. Kim, T.; Attias, H.; Lee, S.; Lee, T. Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 70–79. [[CrossRef](#)]
16. Kim, T.; Lee, I.; Lee, T.W. Independent vector analysis: Definition and algorithms. In Proceedings of the Fortieth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 29 October–1 November 2006.
17. Simonyan, K.; Ackermann, H.; Chang, E.F.; Greenlee, J.D. New developments in understanding the complexity of human speech production. *J. Neurosci.* **2016**, *36*, 11440–11448. [[CrossRef](#)] [[PubMed](#)]
18. Cooke, M.; Ellis, D. The auditory organization of speech and other sources in listeners and computational models. *Speech Commun.* **2001**, *35*, 141–177. [[CrossRef](#)]
19. Sun, Y.; Rafique, W.; Chambers, J.A.; Naqvi, S.M. Underdetermined source separation using time-frequency masks and an adaptive combined Gaussian-Student's t probabilistic model. In Proceedings of the 2017 IEEE ICASSP, New Orleans, LA, USA, 5–9 March 2017; pp. 4187–4191.
20. Sundar, H.; Seelamantula, C.S.; Sreenivas, T. A mixture model approach for formant tracking and the robustness of Student's t distribution. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 2626–2636. [[CrossRef](#)]
21. Rafique, W.; Naqvi, S.M.; Jackson, P.J.B.; Chambers, J.A. IVA algorithms using a multivariate Student's t source prior for speech source separation in real room environments. In Proceedings of the IEEE ICASSP, South Brisbane, QLD, Australia, 19–24 April 2015; pp. 474–478.
22. Rafique, W. Enhanced Independent Vector Analysis for Speech Separation in Room Environments. Ph.D. Thesis, Newcastle University, Newcastle upon Tyne, UK, 2017.
23. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer-Verlag: Secaucus, NJ, USA, 2006.
24. Hao, J.; Lee, I.; Lee, T.W.; Sejnowski, T. J. Independent Vector Analysis for Source Separation Using a Mixture of Gaussians Prior. *Neural Comput.* **2010**, *22*, 1646–1673. [[CrossRef](#)]
25. Rafique, W.; Erateb, S.; Naqvi, S.M.; Dlay, S.S.; Chambers, J.A. Independent vector analysis for source separation using an energy driven mixed Student's t and super Gaussian source prior. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 29 August–2 September 2016.
26. Liang, Y. Enhanced Independent Vector Analysis for Audio Separation in a Room Environment. Ph.D. Thesis, Loughborough University, Loughborough, UK, 2013.
27. Peel, D.; McLachlan, G.J. Robust mixture modelling using the t distribution. *Stat. Comput.* **2000**, *10*, 339–348. [[CrossRef](#)]
28. Rafique, W.; Naqvi, S.M.; Chambers, J.A. Speech source separation using the IVA algorithm with multivariate mixed super Gaussian Student's t source prior in real room environment. In Proceedings of the IET Conference Proceedings, London, UK, 1–2 December 2015.
29. Rafique, W.; Naqvi, S.M.; Chambers, J.A. Mixed source prior for the fast independent vector analysis algorithm. In Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), Rio de Janeiro, Brazil, 10–13 July 2016; pp. 1–5.
30. Aroudi, A.; Veisi, H.; Sameti, H.; Mafakheri, Z. Speech signal modeling using multivariate distributions. *EURASIP J. Audio Speech Music Process.* **2015**, *2015*, 35. [[CrossRef](#)]

31. Bauchau, O.A.; Trainelli, L. The vectorial parametrization of rotation. *J. Nonlinear Dyn.* **2003**, *32*, 71–92. [[CrossRef](#)]
32. Dragmor, S.S.; Goh, C.J. Some counterpart inequalities in for a functional associated with Jensen’s inequality. *J. Inequal. Appl.* **1997**, *1*, 311–325. [[CrossRef](#)]
33. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L.; Zue, V. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
34. Vincent, E.; Fevotte, C.; Gribonval, R. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469. [[CrossRef](#)]
35. Allen, J.B.; Berkley, D.A. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [[CrossRef](#)]
36. Andreson, M.; Adali, T.; Li, X.L. Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis. *IEEE Trans. Signal Process.* **2012**, *60*, 1672–1682. [[CrossRef](#)]
37. Boukouvalas, Z.; Fu, G.-S.; Adali, T. An efficient multivariate generalized Gaussian distribution estimator: Application to IVA. In Proceedings of the 2015 49th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, 18–20 March 2015.
38. Hummersone, C. A Psychophysicoacoustic Engineering Approach to Machine Sound Source Separation in Reverberant Environments. Ph.D. Thesis, University of Surrey, Guildford, UK, 2011.
39. *ISO 3382-2: 2008. Acoustics. Measurements of Room Acoustics Parameters, Part 2*; ISO: Geneva, Switzerland, 2008.
40. Shinn-Cunningham, B.; Kopco, N.; Martin, T. Localizing nearby sound sources in a classroom: Binaural room impulse responses. *J. Acoust. Soc. Am.* **2005**, *117*, 3100–3115. [[CrossRef](#)]
41. Hu, Y.; Loizou, P.C. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 229–238. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).