WARWICK
THE UNIVERSITY OF WARWICK

**Manuscript version: Author's Accepted Manuscript**
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**
http://wrap.warwick.ac.uk/114641

**How to cite:**
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

**warwick.ac.uk/lib-publications**

# Dirichlet Latent Variable Model: A Dynamic Model Based on Dirichlet Prior for Audio Processing

Anurendra Kumar, Tanaya Guha, *Member, IEEE*, Prasanta Kumar Ghosh, *Senior Member, IEEE*

***Abstract*—We propose a dynamic latent variable model for learning latent bases from time varying, non-negative data. We take a probabilistic approach to modeling the temporal dependence in data by introducing a dynamic Dirichlet prior - a Dirichlet distribution with dynamic parameters. This new distribution allows us to assure non-negativity and avoid intractability when sequential updates are performed (otherwise encountered in using Dirichlet prior). We refer to the proposed model as the *Dirichlet latent variable model (DLVM)*. We develop an expectation maximization algorithm for the proposed model, and also derive a maximum a posteriori estimate of the parameters. Furthermore, we connect the proposed DLVM to two popular latent basis learning methods - probabilistic latent component analysis (PLCA) and non-negative matrix factorization (NMF). We show that (i) PLCA is a special case of our DLVM, and (ii) DLVM can be interpreted as a dynamic version of NMF. The usefulness of DLVM is demonstrated for three audio processing applications - speaker source separation, denoising, and bandwidth expansion. To this end, a new algorithm for source separation is also proposed. Through extensive experiments on benchmark databases, we show that the proposed model outperforms several relevant existing methods in all three applications.**

***Index Terms*—Latent variable model, Dirichlet distribution, time varying, non negative, NMF, exponential family distributions.**

## I. Introduction

Learning effective generative models to achieve rich and compact representation of signals is critical to many signal processing and modeling tasks. Latent variable models (LVMs) form a class of generative models that associate a set of unobserved (latent) variables to the observed variables, where the latent variables are assumed to be the underlying cause of the observations. LVMs that are commonly used to model *non-negative* data are probabilistic latent component analysis (PLCA) [1] (an extension of the probabilistic latent semantic analysis (PLSA) [2]) and latent Dirichlet allocation (LDA) [3]. The wide success of LVMs is noted in many applications, such as source separation [4], topic modeling [3] and biomedical signal processing [5].

Another popular data modeling approach that is closely related to the LVMs is the non-negative matrix factorization (NMF) [6], [7]. The objective of both LVMs and NMF is to learn the underlying 'building blocks' in data, often called the *latent bases*. Both assume that the data is inherently low rank, and represent each observation as a linear combination of the latent bases. Given a data matrix, these models aim to learn a basis matrix and its corresponding coefficient matrix. It has been shown that for certain cost functions, LVMs converge to NMF [7], [6]. Thus, LVMs can be thought of as the probabilistic counterpart of NMF. The advantages of probabilistic methods (such as LVMs) over non-probabilistic approaches (such as NMF) are that the probabilistic approaches can be easily generalized to higher dimensions, and they also allow imposing constraints with suitable prior distributions [7].

The LVMs and NMF in their basic forms do not take into account the temporal correlation in the data. The basic (static) models assume that each data point is independent. However, signals like speech exhibit strong temporal dependence, and an effective strategy is needed to capture such temporal dependence. Efforts have been put towards learning dynamic models by imposing temporal constraints on the bases as well as on their coefficients. The dynamic models include sparse and dynamic variant of LVM/NMF [8], [9], [10], [11], convolutive NMF [12], [13], [14] and non-negative hidden Markov model (NHMM) [15].

In this paper, we take a probabilistic approach to modeling the temporal dependence and propose a dynamic LVM for learning latent bases from time varying, non-negative data [0]. We model the temporal dependence in data by introducing a *dynamic Dirichlet* prior i.e., a Dirichlet distribution with dynamic parameters. Earlier the Dirichlet prior (without dynamic parameters) has been shown to be useful for only static non-negative data [3], and to yield non-negative updates when applied to dynamic non-negative data [17]. For the likelihood function, we use a mixture multinomial as it is well known to capture the structure of non-negative data [3], [2] and often yields simple and closed form solutions. We develop an expectation maximization (EM) algorithm for the proposed model, and derive a maximum a posteriori (MAP) estimate of the parameters. We show that the expected log-likelihood function is concave and can be solved by standard convex optimization methods. We refer to the proposed model as the *Dirichlet latent variable model (DLVM)*.

We also establish strong connections between the proposed DLVM and the two well known basis learning methods - PLCA and NMF. We show that (i) the PLCA model is a special case of the proposed DLVM, and (ii) DLVM is a dynamic version of NMF. We also show that our model is generic and suitable for both count data (e.g., word count data) and non-count data (e.g., speech data). Unlike other dynamic LVMs, the proposed model does not have any free parameter (other

A. Kumar was with the Department of Electrical Engineering, Indian Institute of Technology (IIT), Kanpur. T. Guha is with the Department of Computer Science, University of Warwick. P. K. Ghosh is with the Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore.

[0]A preliminary version of this work has been published in ICASSP'18 [16].

than the number of latent bases). The effectiveness of the proposed model is demonstrated through three applications: speaker source separation, bandwidth expansion and speech-noise separation. Extensive experiments on the TIMIT [18] and the signal processing information base (SPIB) [19] databases show that the proposed model outperforms several relevant existing methods in all three applications. The contributions of this work are as follows:

- The main contribution of the paper lies in proposing a new *dynamic Dirichlet* prior - a Dirichlet distribution with dynamic parameters - which yields non-negative updates for dynamic non-negative data. Subsequently, using this new prior for our model, we develop a suitable EM algorithm, and derive MAP estimates of the parameters.
- We show that (a) DLVM is a dynamic version of NMF, and (b) the popular PLCA model is a special case of DLVM.
- We proposed a source separation algorithm utilizing the latent bases learned using DLVM.
- The proposed DLVM has been successfully applied to three speech processing tasks (speaker source separation, bandwidth expansion and speech-noise separation) to achieve superior results.

## II. Related Work

One of the first latent variable models, the PLSA, was developed for addressing natural language processing tasks [2]. Later, a similar model, popularly known as the PLCA [1], was proposed to analyze audio spectrograms, and was successfully applied to audio source separation [1]. The PLCA constructs a generative story of the observed data using latent variables with a mixture multinomial distribution as the likelihood. Another model, the LDA [3], was developed by extending the PLSA framework that imposed Dirichlet distribution as a prior. Dirichlet distribution provides an intuitive understanding of the corresponding multinomials as pseudocounts [3].

However, the corresponding EM algorithm becomes intractable because the likelihood is a mixture multinomial instead of a multinomial [3]. Several sampling strategies, such as Markov chain Monte Carlo (MCMC) and variational Bayes method are used to resolve the intractability issue [3]. Even though LDA has shown promising results in natural language processing [3], it has not been very successful in audio processing tasks, such as source separation [4]. Similar to the probabilistic models, a large number of NMF algorithms involving different cost functions exist in the literature [20] [9] [21]. Most of these algorithms use an alternating maximization method to learn the basis and their coefficients. It has been shown that for certain cost functions, LVMs converge to NMF [7], [6], and can be thought of as the probabilistic counterpart of NMF.

As discussed earlier, LVMs and NMF in their basic forms do not take into account the temporal correlation in the data. Extensions have been proposed to incorporate temporal dependencies in static LVMs and NMF. Shift invariant PLCA [12], [22] captures the temporal structure by imposing constraints on the basis matrix. It models the observed data as a convolutive mixture of latent bases, and has the property of shift invariance. Another natural extension of this idea is to have temporal constraints on the coefficient matrix while the basis matrix is constant [23]. HMM is a popular method for modeling temporal data with discrete states. An attempt has been made to connect HMM and NMF resulting into a non-negative HMM [15]. This model is useful for data with discrete number of states.

Another line of approach to model time varying data is to use Kalman filter [24]. Kalman filters and its nonlinear variants [25] [26] [27] have been widely used for the estimation of continuous states. Such models assume a Gaussian distribution as the likelihood, and is not well suited for modeling non-negative data. Recently, there has been a significant amount of work on learning continuous state representations [28], [29], [17] for non-negative data. One such work used the Gamma distribution as the likelihood function [28]. This model can be viewed as a dynamic counterpart of NMF with Itakura-Saito divergence [17]. Both the Gamma and the mixture multinomial distributions are suitable for modeling non-negative data, where the preference of one over the other is application-specific [23]. Another work has extended the basic PLCA model by combining ideas from state space models and Kalman filtering [17]. This work used an exponential distribution as a prior. In contrast to the past literature, the work in this current paper proposes a new prior - a Dirichlet distribution whose parameters are dynamic. We derive the corresponding update equations, which are a generalized form of the static PLCA.

## III. Dynamic Dirichlet Latent Variable Model

In this section, we develop the DLVM along with its variant - the *bidirectional* DLVM. Our objective is to model a time varying signal $x(t)$ by learning its latent bases from its spectral distributions. We represent $x(t)$ spectrographically by taking its short time Fourier transform (STFT) and retaining its scaled magnitude spectrogram

$$\mathbf{N} = \gamma |STFT(x(t))| = \gamma \mathbf{X} \tag{1}$$

where $\gamma$ is a large integer that ensures that all elements in $\mathbf{N}$ are integers [1]. Now, the observed (spectral) data matrix $\mathbf{N}$ can be seen as count data, where $N_{ft}$ corresponds to the count of the frequency $f$ at a time instant $t$. Each column of the matrix $\mathbf{N}$ thus corresponds to a spectral distribution at a particular time instant. With each frequency count $f \in \{1, 2, ..., F\}$, we associate an unknown latent variable $\mathbf{z}$ of dimension $K$ with one of the entries as 1 and the rest as zero, $\mathbf{z} = [z_1, z_2, ....z_K]$. Here, $z_i$ denotes the $i^{th}$ latent basis described by a spectral distribution $P(f|z_i)$.

LVMs assume that the underlying cause of an observed variable $f$ is a set of unobserved latent variables $z_k$ where $k \in \{1, 2, ...K\}$. Marginalizing over the latent bases $\mathbf{z}$, the spectrogram ($\mathbf{N}$) at time $t$ is a mixture of the $K$ hidden distributions, where $K$ is a known positive integer

$$P_t(f) = \sum_{k=1}^{K} P_t(f, z_k) = \sum_{k=1}^{K} P_t(z_k) P(f|z_k) \tag{2}$$

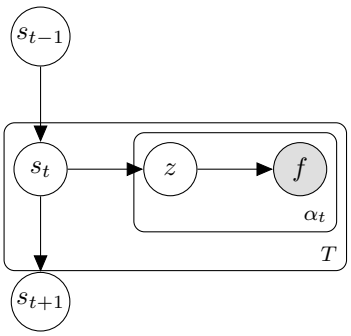


Figure 1: Plate notation for DLVM ($T$: total number of time instants).

where $P_t(f)$ is the probability of frequency $f$ at $t$, $P(f|z_i)$ is a multinomial distribution (similar to that used PLCA [1]), and the coefficients of the mixture are $P_t(z_i)$, $i \in \{1, 2, ....K\}$.

Our model assumes that the latent bases $P(f|z_k)$ are the same at all time instants. The bases are source-specific and are viewed as the spectral signatures of the sources. On the other hand, the coefficients $P_t(z_k)$ vary over time, and they describe the probability of each latent base at a given time $t$.

Let $\mathbf{n}_t$ denote the observation vector at time instant $t$ (the $t^{th}$ column of the matrix $\mathbf{N}$). Let us now define a state, $\mathbf{s}_t$, of the observation vector $\mathbf{n}_t$ as follows

$$\begin{aligned}\mathbf{s}_t &= [P_t(z_1), P_t(z_2), ....., P_t(z_k)]^{\mathrm{T}} \\ &= [s_t(1), s_t(2), ....., s_t(K)]^{\mathrm{T}}\end{aligned} \tag{3}$$

In general, static LVMs assume that the mixture coefficients $P_t(z_k)$ (hence, the states $\mathbf{s}_t$) are independent at all time instants. However, this assumption limits the effectiveness of the LVMs for modeling time varying signals. Our model addresses this limitation by imposing a Markovian dependence between states using a Dirichlet distribution. Note that the support of the states $\mathbf{s}_t$ of dimension $K$ lies on a $K-1$ dimensional simplex. The Dirichlet distribution has been widely used as a distribution on simplex and is the conjugate of multinomial. Also, it belongs to the exponential family and has finite dimensional sufficient statistics [3]. These properties lead to intuitive and efficient parameter estimation discussed in Section IV. The proposed model is described below in detail.

### A. DLVM

To model the temporal dependence between states, we propose a Dirichlet distribution with time-varying parameters

$$\begin{aligned}P(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{D}) &= \mathrm{Dir}(\alpha_{t-1}\mathbf{D}\mathbf{s}_{t-1} + \mathbf{1}) \\ \text{where } \alpha_t &= \sum_f N_{ft} \\ P(\mathbf{s}_1) &= \mathrm{Dir}(\mathbf{1})\end{aligned} \tag{4}$$

where, 'Dir' denotes the Dirichlet distribution, $\alpha_t$ denotes the total number of observations at time $t$, $\mathbf{1}$ is an all-one vector, and $\mathbf{D}$ is a temporal dependence matrix defined as follows

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & & \dots & d_{1K} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{K1} & d_{K2} & d_{K3} & \dots & d_{KK} \end{bmatrix} \tag{5}$$

where, $d_{ij} \in \mathbb{R}^+$ denotes the temporal dependence between states at two consecutive time instants for the $i^{th}$ and $j^{th}$ latent basis. A higher value of $d_{ij}$ indicates higher temporal dependence.

The Dirichlet-multinomial conjugacy allows us to have an intuitive understanding of the parameters of the Dirichlet distribution as pseudo observations. Let us define the pseudo observation for the $k^{th}$ basis at time $t$ as $m_{tk} = \alpha_{t-1}(\mathbf{D}\mathbf{s}_{t-1})(k)$. Therefore (4) can be rewritten as follows

$$P(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{D}) = \frac{\Gamma(\sum_k (m_{tk}+1))}{\prod_k \Gamma(m_{tk}+1))} \prod_k s_t(k)^{m_{tk}} \tag{6}$$

where, $\Gamma(x)$ is the Gamma function. Note that the hyperparameters of the Dirichlet distribution are dynamic. Hence, we refer to it as the *dynamic Dirichlet distribution* in the rest of this paper.

Let $\zeta_t$ be a $\alpha_t$-dimensional vector, where $\zeta_t(l) \in \{z_1, z_2, ...z_K\}$ contains the active latent basis of the $l^{th}$ count of the observation vector $\mathbf{n}_t$. According to our model, the generative process of $\mathbf{N}$ is as follows:

- Sample $\mathbf{s}_t \sim \mathrm{Dir}(\alpha_{t-1}\mathbf{D}\mathbf{s}_{t-1} + \mathbf{1})$
- Sample frequency $f$, $\alpha_t$ times as follows:
  - Choose a latent basis $\zeta_t(l) \sim \mathrm{Mult}(\mathbf{s}_t)$
  - Choose a frequency $f \sim \mathrm{Mult}(P(f|\zeta_t(l)))$.
- Repeat the above process $T$ times.

where, $T$ is the total number of time instants, 'Mult' denotes the multinomial distribution. We observe that each sample is a realization of a mixture multinomial (for simplicity, we use the notation of the categorical distribution, also the conjugate of Dirichlet, for the multinomial distribution. It is a common practice [3], and has no effect on parameter estimation). We assume that the counts in an observation vector at a time instant are independent and identically distributed. Fig. 1 presents a graphical model for the proposed generative process.

The proposed dynamic Dirichlet distribution prior has the following appealing properties which provides an intuitive understanding:

- The generative process (with mixture multinomial as the likelihood) allows us to view the spectrogram at time $t$ as an observed count data over $K$ bases. Static models such as PLCA uses this observation data to estimate the states at each time instant. The dynamic Dirichlet prior allows us to have $m_{tk}$ extra pseudo observations for each basis $k$ at time instant $t$, which is the result of the multinomial-Dirichlet conjugacy [30]. This observation will become clear in (16) later. A higher number of observations at previous time instant ($\alpha_{t-1}$) or higher value of temporal dependence ($d_{jk}$) leads to a higher count of pseudo observations for $k^{th}$ basis.

- The mode of the distribution lies at the normalized pseudo observations

$$\max_k(s_t(k)|\mathbf{s}_{t-1}) = \frac{m_{tk}}{\sum_k m_{tk}}$$

- The variance of each entry of the vector $\mathbf{s}_t$ can be obtained from the properties of the Dirichlet distribution [30]

$$Var(s_t(k)|\mathbf{s_{t-1}}) \propto \frac{1}{(\sum_k m_{tk} + K)^2(\sum_k m_{tk} + K + 1)}$$

which decreases as the total number of observations at previous time instant i.e., $\alpha_{t-1}$ increases. A higher value of $\alpha_{t-1}$ indicates that more prior information (more pseudo observation) is available. This trend in variance is expected because the distribution should have less variance when there is more prior information from previous time instant.

The proposed DLVM can be interpreted as a three-level hierarchical Bayesian model similar to the LDA [3]. In our model, $P(f|\mathbf{z})$ is the source level parameter (sampled once for each source), while $\mathbf{s}_t$ is the time level (sampled once at every time instant), and $z_k$ and $f$ are the frequency level (sampled once per frequency) parameters.

### *B. PLCA as a special case of DLVM*

The relationship between the proposed DLVM and the well known PLCA model is particularly interesting. When the temporal dependence matrix $\mathbf{D}$ is reduced to a zero matrix, the distribution in (4) becomes a symmetric Dirichlet distribution $\mathrm{Dir}(\mathbf{1})$. Note that the symmetric Dirichlet distribution $\mathrm{Dir}(\mathbf{1})$ is nothing but a uniform distribution, and thus, the formulation in (4) is equivalent to the static PLCA. This can also be intuitively understood as the fact that in the absence of prior information, there is no prior preference of any state over the others. Writing the parameters of dynamic Dirichlet as:

$$\alpha_{t-1}\mathbf{D}\mathbf{s}_{t-1} + \mathbf{1} = \alpha_{t-1}\mathbf{D}(s_{t-1} + \mathbf{1}) + (\mathbf{I} - \alpha_{t-1}\mathbf{D})\mathbf{1}$$

where, $\mathbf{I}$ is an identity matrix. The first term contains the prior information from the previous time instant. The second term contains the content of uniform distribution. When $\alpha_{t-1}\mathbf{D}$ is an identity matrix, the prior has no component of uniform distribution and is completely decided by the past information. The amount of past information is controlled by the total number of observed count at previous time instant ($\alpha_{t-1}$).

### *C. Bidirectional DLVM*

We assumed that $\mathbf{s}_t$ depends only on the immediate past state $\mathbf{s}_{t-1}$. This degree of dependence can be relaxed, and more states can be included to account for longer temporal dependence. For example, a natural extension would be to include the temporal dependence both in the past and in the future states.

Let us denote the forward dependence (i.e., dependence on the past states) matrices as $\mathbf{D}_1^+, \mathbf{D}_2^+, ...\mathbf{D}_l^+$ and the backward dependence (i.e., dependence on the future states) matrices as $\mathbf{D}_1^-, \mathbf{D}_2^-, ...\mathbf{D}_l^-$, where the model order $l$ is a positive integer.

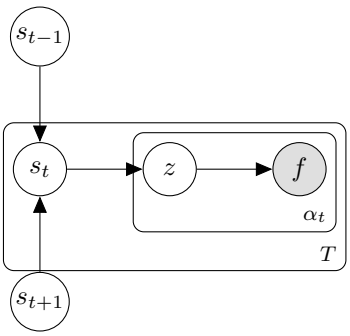


Figure 2: Plate notation for bi-DLVM (order = 1).

$\mathbf{D}_l^+$ denotes the temporal dependence between $\mathbf{s}_t$ on $\mathbf{s}_{t-l}$ while $\mathbf{D}_l^-$ denotes the temporal dependence between $\mathbf{s}_t$ on $\mathbf{s}_{t+l}$. The dynamic Dirichlet distribution in this case takes the following form

$$\begin{aligned} &P(\mathbf{s}_t|\mathbf{s}_{t-l}, ..., \mathbf{s}_{t+l}, \mathbf{D}_1^+, \mathbf{D}_1^-, ..., \mathbf{D}_l^+, \mathbf{D}_l^-) = \\ &\mathrm{Dir}(\sum_{j=1}^{l} \alpha_{t-j}\mathbf{D}_j^+\mathbf{s}_{t-j} + \alpha_{t+j}\mathbf{D}_j^-\mathbf{s}_{t+j} + \mathbf{1}) \end{aligned} \tag{7}$$

where, $l$ denotes the maximum degree of temporal dependence in the model. To keep the equations simple, we consider $l = 1$. Let us denote $\mathbf{D}_1^+$ and $\mathbf{D}_1^-$ as the forward and the backward dependence matrices. The single order bi-DLVM (referred to as the bi-DLVM in rest of the paper) is given by

$$\begin{aligned} &P(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{s}_{t+1}, \mathbf{D}_1^+, \mathbf{D}_1^-) = \\ &\mathrm{Dir}(\alpha_{t-1}\mathbf{D}_1^+\mathbf{s}_{t-1} + \alpha_{t+1}\mathbf{D}_1^-\mathbf{s}_{t+1} + \mathbf{1}) \end{aligned} \tag{8}$$

The corresponding graphical model is presented in Fig. 2. Note that the proposed bi-DLVM (see (8)) reduces to the earlier proposed DLVM for $\mathbf{D_1}^- = 0$.

## IV. Parameter Estimation

In this section, we describe the parameter estimation steps for the proposed DLVM. We derive an EM algorithm for the same.

Let us denote the state matrix $\mathbf{S} = [\mathbf{s}_1, ...\mathbf{s}_t, .....\mathbf{s}_T]$, $\beta$= $\{P(f|\mathbf{z}), \mathbf{D_1}^+, \mathbf{D_1}^-\}$, $\Lambda = \{\beta, \mathbf{S}\}$ and $\zeta = [\zeta_1, ..\zeta_t, .., \zeta_T]$. Let us denote the $(i,j)^{th}$ element of $\mathbf{D_1}^+$ and $\mathbf{D_1}^-$ as $d_{ij}^+$ and $d_{ij}^-$ respectively.

The joint probability of the observed and the latent variables given the parameter $\beta$ is as follows

$$\begin{aligned} &P(\mathbf{S}, \zeta, \mathbf{N}|\beta) = P(\mathbf{S}|\beta)P(\mathbf{N}, \zeta|\mathbf{S}, \beta) = \\ &= \prod_t \left( P(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_{l=1}^{\alpha_t} (P(f|\zeta_t(l))P_t(\zeta_t(l)) \right) \end{aligned} \tag{9}$$

This is obtained using Markovian dependence between states at different time instants, and the assumption that given $\mathbf{S}$ the columns of $\mathbf{N}$ (i.e., $\mathbf{n}_t$) are independent of each other.

The likelihood of the observed spectral data matrix $\mathbf{N}$ is obtained by marginalizing $\zeta$ and $\mathbf{S}$

$$P(\mathbf{N}|\beta) = \prod_t \left( \int_{\mathbf{s}_t} P(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_f \left( \sum_k P(f|z_k) s_t(k) \right)^{N_{ft}} \mathrm{d}\mathbf{s}_t \right) \tag{10}$$

Our aim is to estimate $\hat{\Lambda}$ so as to maximize the above marginalized likelihood (see 10). EM is a common approach to maximize log-likelihood in presence of latent variables. It consists of two iterative updates: i) an expectation (E) step, where the posterior distribution of the latent variables is computed; and ii) a maximization (M) step, where the expected log-likelihood is maximized with respect to the posterior distribution. The posterior distribution of latent variables is given by

$$P(\mathbf{S}, \zeta|\mathbf{N}, \beta) = \frac{P(\mathbf{S}, \zeta, \mathbf{N}|\beta)}{P(\mathbf{N}|\beta)} \tag{11}$$

However, the denominator (see (10)) is computationally intractable [3] [31]. Sampling techniques, such as MCMC, or approximate inference techniques, such as variational Bayes inference can be employed to address the issue. However, for simplicity, we perform a MAP estimate of the states instead of a fully Bayesian inference. Therefore, we maximize the following

$$\begin{aligned} \hat{\Lambda} = \{\hat{\beta}, \hat{\mathbf{S}}\} &= \underset{\Lambda}{\operatorname{argmax}} P(\mathbf{N}, \mathbf{S}|\beta) \\ &= \underset{\Lambda}{\operatorname{argmax}} \prod_t \left( P(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_f \left( \sum_k P(f|z_k) s_t(k) \right)^{N_{ft}} \right) \end{aligned} \tag{12}$$

The steps in our EM algorithm are described below.

### A. Expectation step

The posterior distribution of $\mathbf{z}$ is given by

$$P_t(z_k|f) = \frac{P_t(z_k) P(f|z_k)}{\sum_k P_t(z_k) P(f|z_k)} \tag{13}$$

### B. Maximization step

We intend to maximize the following MAP function

$$\begin{aligned} \mathcal{L}_{MAP} &= \mathop{\mathbb{E}}_{P_t(\mathbf{z}|f)} \log(P(\mathbf{N}, \mathbf{S}, \zeta|\beta)) \\ &= \mathop{\mathbb{E}}_{P_t(\mathbf{z}|f)} \log(P(\mathbf{N}, \zeta|\mathbf{S}, \beta)) + \log(P(\mathbf{S}|\beta)) \\ s.t., \ &\sum_f P(f|z_k) = 1, \ \sum_k s_t(k) = 1, \ 0 < d_{ij}^+, d_{ij}^- \ \forall \ i, j \end{aligned} \tag{14}$$

The objective function $\mathcal{L}_{MAP}$ is concave with respect to each of the parameters $(\mathbf{S}, P(f|\mathbf{z}), \mathbf{D_1}^+, \mathbf{D_1}^-)$ provided others are fixed. [1].

[1]Our proof of concavity: https://tinyurl.com/yxm9gqy7

*1) Update of $P(f|\mathbf{z})$:* Maximizing the above constrained expected log-likelihood in (14) with respect to $P(f|z_k)$ yields the following

$$P(f|z_k) = \frac{\sum_t N_{ft} P_t(z_k|f)}{\sum_f \sum_t N_{ft} P_t(z_k|f)} \tag{15}$$

Note that this update for the latent basis $P(f|z_k)$ is the same as that for PLCA [4].

*2) Update of $\mathbf{S}$:* Let us now define pseudo observation from the previous and the next time instants as $m_{tk}^+$ and $m_{tk}^-$ for a basis $k$ as follows

$$\begin{aligned} m_{tk}^+ &= \alpha_{t-1}(\mathbf{D}_1{}^+ \mathbf{s}_{t-1})(k) \\ m_{tk}^- &= \alpha_{t+1}(\mathbf{D}_1{}^- \mathbf{s}_{t+1})(k) \end{aligned}$$

We perform a sequential update for the states $\mathbf{S}$ in forward direction starting from the first time instant. While estimating $\mathbf{s}_t, \mathbf{s}_{t-1}$ appears inside the Gamma function which has already been estimated. Therefore, the proposed updates are in closed form unlike that in other models based on estimating parameters of Dirichlet (e.g., Latent Dirichlet allocation, Hierarchical Dirichlet process).

Maximizing $\mathcal{L}_{MAP}$ with respect to $s_t(k)$ while keeping $\mathbf{D}_1^+$ and $\mathbf{D}_1^-$ fixed yields

$$s_t(k) = \frac{\sum_f N_{ft} P_t(z_k|f) + m_{tk}^+ + m_{tk}^-}{\sum_k (\sum_f N_{ft} P_t(z_k|f) + m_{tk}^+ + m_{tk}^-)} \tag{16}$$

We see that the updates of the states contain additional terms $(m_{tk}^+, m_{tk}^-)$ as compared to those in PLCA. We call them as *pseudo observation* for each basis $k$. This update is similar to the updates of the states in Kalman filtering [24].

*3) Updates of $\mathbf{D}_1^+$, $\mathbf{D}_1^-$:* The update of $\mathbf{D}_1^+$ and $\mathbf{D}_1^-$ is dependent on scaling factor $\gamma$. However, we ignore its effects and justify the assumption by empirical results.
Maximizing $\mathcal{L}_{MAP}$ with respect to $\mathbf{D}_1^+$ (keeping $\mathbf{S}$ and $\mathbf{D}_1^-$ fixed) does not have any closed form solution.

$$\begin{aligned} \mathbf{D}_1^+ = \underset{\mathbf{D}_1^+}{\operatorname{argmax}} \ &\sum_t \Big( \log\Gamma(\sum_k (m_{tk}^+ + m_{tk}^- + 1)) - \\ &\sum_k \log\Gamma(m_{tk}^+ + m_{tk}^- + 1) + \sum_k m_{tk}^+(k) \log(s_t(k)) \Big) \\ s.t., \quad &0 < \ d_{ij}^+, \quad \forall \quad i, j. \end{aligned} \tag{17}$$

However, the maximizing function is concave since Dirichlet distribution belongs to the exponential family of distributions [1]. Therefore, the function has a unique maxima, which can be obtained via gradient ascent

$$\begin{aligned} \frac{\partial \mathcal{L}_{MAP}}{\partial d_{ik}^+} = \sum_t \alpha_{t-1} s_{t-1}(i) \Big( &\psi(\sum_j (m_{tj}^+ + m_{tj}^- + 1)) \\ &- \psi(m_{tk}^+ + m_{tk}^- + 1) + \log(s_t(k)) \Big) \end{aligned} \tag{18}$$

where, $\psi$ is the digamma function. $\mathbf{D}_1^-$ is updated similarly. However, we find the above equations to be computationally expensive. Therefore, we restrict the dependence matrices ($\mathbf{D}_1^+$ and $\mathbf{D}_1^-$) to be diagonal matrices, which yields similar results in our experimentations. Note that the updates of $P(f|\mathbf{z})$,

and $\mathbf{S}$ are independent of the scaling factor $\gamma$. Therefore, the proposed algorithm is applicable to both count and non-count data, and we may replace $\mathbf{N}$ by $\mathbf{X}$ (refer (1)) in all the update equations.

## V. DLVM as a Dynamic NMF

In this section, we show that the proposed DLVM can be viewed as a dynamic version of NMF. It has been shown in the literature that an LVM can be interpreted as an NMF for specific cost functions [32]. The update equations for PLCA and NMF with Kullback–Leibler (KL) divergence have been shown to be the same [32]. The NMF interpretation of PLCA leads to very compact and fast update equations. Similarly, we can interpret the DLVM (dynamic counterpart of PLCA) as a dynamic version of NMF. Below, we present the update equations for our bi-DLVM as a dynamic version of NMF. The updates of DLVM as a dynamic version of NMF can be obtained by setting $\mathbf{D_1}^-$ to a zero matrix.

**Algorithm 1** Bi-DLVM as a NMF

**Input:** $\mathbf{X}$

**Output:** $\mathbf{W}, \mathbf{S}, \mathbf{D}_1^+, \mathbf{D}_1^-$

Randomly initialize $\mathbf{W}, \mathbf{S}$, $\mathbf{D}_1^+$, $\mathbf{D}_1^-$

**while** *not converged* **do**

$$\left.\begin{aligned} W_{fk} &= W_{fk} \sum_t \frac{X_{ft}}{(WS)_{ft}} S_{kt} \\ W_{fk} &= \frac{W_{fk}}{\sum_f W_{fk}} \end{aligned}\right\}^2$$

**while** *Not converged* **do**

$$\left.\begin{aligned} S_{kt} &= S_{kt} \sum_t W_{fk} \frac{X_{ft}}{(WS)_{ft}} + m_{tk}^+ + m_{tk}^- \\ S_{kt} &= \frac{S_{kt}}{\sum_t S_{kt}} \end{aligned}\right\}^3$$

Update $\mathbf{D}_1^+$ and $\mathbf{D}_1^-$ using (17)

**end**

**end**

DLVM learns the latent bases and the states for an observed data matrix via the following factorization

$$P_t(f) = \sum_{k=1}^{K} P_t(z_k) P(f|z_k)$$

Multiplying both sides of the above equation by $\alpha_t$, we rewrite the equation in matrix-vector form as $\mathbf{n}_t = \mathbf{W}\mathbf{s}_t\alpha_t$, where, $\mathbf{W}$ is a matrix whose columns are latent bases $P(f|\mathbf{z})$. Concatenating observation vector $\mathbf{n}_t$ for all time instants, we can write the observed data matrix $\mathbf{X}$ as,

$$\mathbf{X}_{F\times T} = \mathbf{W}_{F\times K}\mathbf{S}_{K\times T}\mathbf{G}_{T\times T} = \mathbf{W}_{F\times K}\mathbf{H}_{K\times T}$$

where, $\mathbf{W}$ is the basis matrix, $\mathbf{S}$ is the state matrix, $\mathbf{G}$ (normalization matrix) is a diagonal matrix with $\alpha_t$ as the diagonal elements, and the subscripts denote dimensions of the matrices. The $(f,k)^{th}$ element in $\mathbf{W}$ is denoted as $W_{fk}$ and the $(k,t)^{th}$ element in $\mathbf{S}$ is denoted as $S_{kt}$. It is evident that all matrices ($\mathbf{X}$,$\mathbf{W}$, $\mathbf{S}$, $\mathbf{G}$) are non-negative. Therefore, we can view the proposed DLVM as a dynamic version of NMF with iterative updates for $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{D}$ (see Algorithm 1). In Algorithm 1, the outer loop corresponds to the EM iteration, while the inner loop corresponds to the block-wise update of variables in the maximization step of the EM algorithm.
[4]

## VI. Applications to Audio Processing

In this section, we demonstrate the usefulness of the proposed DLVM model and its variant for three audio processing tasks: (i) speaker source separation, (ii) denoising, and (iii) bandwidth expansion. To this end, we also propose a new algorithm for dynamic source separation.

### A. Source separation

Source separation is a long standing problem in signal processing, which aims to recover the constituent source signals from a given mixture signal. It has wide applications in speaker recognition, speech enhancement, music editing and audio information retrieval [33], [34]. In this section, we develop an algorithm for dynamic source separation using the bases learned from bidirectional DLVM with dependence matrices as $\mathbf{D}_1^+$ and $\mathbf{D}_1^-$.

We assume that the given mixture signal is a linear combination of a known number of speaker signals. The spectral distribution of a mixture signal is given by

$$P_t(f) = \sum_a P_t(a) P_t(f|a) = \sum_a P_t(a) \sum_{z_k \in \mathbf{z}^a} P_t(z_k|a) P(f|z_k) \tag{19}$$

where, $P_t(a)$ denotes the apriori probability of the $a^{th}$ source. The parameters associated with the $a^{th}$ source are denoted as $\Lambda^a = \{S^a, \beta^a\}$.

The graphical model of the mixture signal is presented in Fig. 3. Our objective is to separate the constituent sources from a mixture signal. Following the supervised paradigm, we learn the parameters $\beta^a$ for all $a$ from the training data. These parameters are later used to separate the sources in the separation stage.

Let us consider a mixture spectrogram $\mathbf{X}$. The parameters $P_t(a)$ and $P_t(\mathbf{z}|a)$ are learned from $\mathbf{X}$ via an EM algorithm. In the expectation step, we estimate the posterior distribution and the expected number of total observation, $\alpha_t^a$, for each source. In the maximization step, we maximize the expected log-likelihood

$$\mathcal{L}_{MAP} = \mathop{\mathbb{E}}_{P_t(\mathbf{z},a|f)} \log(P(\mathbf{N}, \mathbf{S}, \zeta, a|\beta)) \tag{20}$$

where, $\beta$, $\mathbf{S}$ and $\zeta$ contains latent variables and parameters for all the sources. The steps in the EM algorithm are as follows

[2]From (13) and (15)

[3]From (13) and (16)

[4]Code and data: https://github.com/anurendra/dlvm

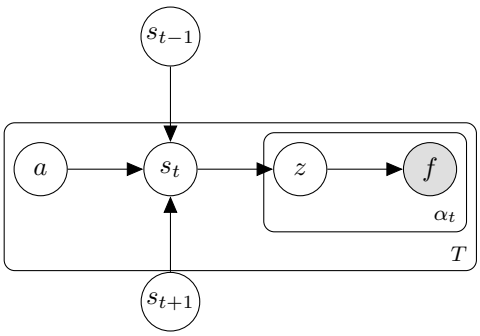


Figure 3: Plate notation of the mixture signal modeled using bi-DLVM, $a$ denotes an audio source.

*1) Expectation step:*

$$P_t(a, z_k|f) = \frac{P_t(a)P_t(z_k|a)P^a(f|z_k)}{\sum_{a'} P_t(a') \sum_{z_k \in \mathbf{z}^{a'}} P_t(z_k|a')P^{a'}(f|z_k)} \tag{21}$$

$$\alpha_t^a = \left( \sum_f X_{ft} \right) P_t(a) \tag{22}$$

*2) Maximization step:*

$$m_{tk}^{a+} = \alpha_{t-1}^a \sum_j d_{kj}^{a+} P_{t-1}(z_j|a)$$

$$m_{tk}^{a-} = \alpha_{t+1}^a \sum_j d_{kj}^{a-} P_{t+1}(z_j|a)$$

$$P_t(z_k|a) = \frac{\sum_f X_{ft} P_t(a, z_k|f) + m_{tk}^{a+} + m_{tk}^{a-}}{\sum_{z_k \in \mathbf{z}^a} \left( \sum_f X_{ft} P_t(z_k, a|f) + m_{tk}^{a+} + m_{tk}^{a-} \right)} \tag{23}$$

$$P_t(a) = \frac{\sum_{z_k \in \mathbf{z}^a} \sum_f X_{ft} P_t(a, z_k|f)}{\sum_{a'} \sum_{z_k \in \mathbf{z}^a} \sum_f X_{ft} P_t(a', z_k|f)}$$

After the above EM algorithm converges, the reconstructed spectral vector for each source is obtained as the expected value of $X_{ft}$ over all sources as follows.

$$P_t(f|a) = \sum_{z_i \in \mathbf{z}^a} P_t(z_i|a) P^a(f|z_i)$$

$$\hat{X}_{ft}(a) = E(X_{ft}(a)) = \frac{P_t(a)P_t(f|a)X_{ft}}{\sum_{a'} P_t(a')P_t(f|a')}$$

Finally, the phase of the mixture signal is combined with the reconstructed magnitude spectrogram ($\hat{X}_{ft}(a)$) to recover each source signal [1].

### B. Denoising

We consider a speech denoising scenario, where the speech signal is degraded by an additive noise. We follow a speaker dependent approach i.e., we assume that the identity of the speaker is known. We also assume that the noise type (e.g., babble, factory) is known and training data for each noise type is available. This assumption is practical in many scenarios where classification techniques can be employed for detecting the noise type and speaker identity [17]. We consider the noise and the speaker's data as two distinct sources, and learn the latent bases separately for the them. We learn the parameters for each speaker and each noise type from the training data. The speech is separated from the noise using the source separation algorithm described in Section VI-A.

### C. Bandwidth expansion

In this section, we develop an algorithm for bandwidth expansion for a band limited signal that utilizes the latent bases learned using DLVM. Bandwidth expansion of a signal may be required in different scenarios, e.g., for signals that are sampled at a low sampling rate, high frequency components may be lost, or, for signals incurring distortion in some frequency bands, or, when the signal acquisition system is incapable of capturing frequencies beyond a particular range.

We address the problem of bandwidth expansion of a narrow band ($0-4$ KHz) speech signal. Using the bases learned by our model, we predict the higher frequency components ($4-8$ KHz) from a given narrow band speech signal. Let us denote the observed frequencies as $f_o \in \{0\text{-}4KHz\}$ and unobserved frequencies as $f_u \in \{4\text{-}8KHz\}$. First, we learn the parameters for a speech signal ($\beta$) for all frequencies from the training data. We use these parameters to estimate states and the total number of draws ($\alpha_t$) using only the observed frequencies. The following equation is used iteratively to estimate $s_t(k)$ from the band limited signal $\mathbf{X}$

$$s_t(k) = \frac{\sum_{f_o} X_{ft} P_t(z_k|f) + m_{tk}^+ + m_{tk}^-}{\sum_k \left( \sum_{f_o} X_{ft} P_t(z_k|f) + m_{tk}^+ + m_{tk}^- \right)}$$

Once the above iterative updates converge, we estimate $\alpha_t$ using only the observed frequencies. Finally, the unobserved frequencies are predicted.

$$P_t(f_u) = \sum_{k=1}^{K} P_t(z_k) P(f_u|z_k)$$

$$\alpha_t = \frac{\sum_{f_o} X_{ft}}{\sum_{f_o} P_t(f)}$$

$$X_{f_u t} = \alpha_t P_t(f_u)$$

The phase values at the unobserved frequencies are estimated separately. Since, phase also contains negative values, our algorithm is not appropriate for phase prediction. We therefore assume that the phase of the unobserved frequencies $\Phi_u$ are a linear transformation of phase of the observed frequencies $\Phi_o$ i.e., $\Phi_u = \mathbf{A}\Phi_o$ [35]. The transformation matrix $\mathbf{A}$ is learned from the phase of the training data $\Phi^{tr}$. The least square estimate of $\mathbf{A}$ is given by $\mathbf{A} = \Phi_u^{tr}(\Phi_o^{tr})^\dagger$, where $(\cdot)^\dagger$ denotes pseudoinverse. Finally, the phase and the magnitude spectrogram are multiplied, and converted to time domain signal using STFT.

## VII. Experimental Validation

In this section, we present various experimental results demonstrating the performance of the proposed DLVM for the

three audio processing tasks described in the previous section. We also compare the performance of the proposed model with several other existing methods.

### A. Experimental setup

We use the TIMIT database [18] and the signal processing information base (SPIB) [19] to carry out our experiments. The TIMIT database contains broadBand recordings of 630 speakers sampled at 16 KHz, each speaker reading ten phonetically rich sentences. The SPIB database contains noise data of 15 different noise types acquired with a sampling frequency of 19.98 KHz, an analog to digital converter (A/D) with 16 bits, an anti-aliasing filter, and without a pre-emphasis stage.

All audio signals were downsampled to 16 KHz. The spectrograms are obtained by performing STFT using a 64ms window with 16ms overlap. The resulting magnitude spectrograms are then processed and analyzed further. The phase spectrograms are analyzed separately, agnostic to the algorithm, using standard methods [4]. We have used 250 iterations for the outer loop, and 10 iterations for the inner loop in DLVM (Algorithm 1). The dependence matrices $(\mathbf{D_1}^+, \mathbf{D_1}^-)$ were fixed to 0 for first 50 iterations.

**Evaluation metrics:** In order to evaluate the source separation and denoising performance, we use the following evaluation metrics (i) signal to noise ratio improvement (SNRI) [36], (ii) source to distortion ratio (SDR), (iii) source to interference ratio (SIR), and (iv) source to artifact ratio (SAR) [34] [37]. The later three provide perceptual evaluation of the source separation results. We have used the BSS-EVAL TOOLBOX [38] for evaluation.

Let $\mathbf{X}$ and $\phi$ represent the magnitude and the phase of a mixture signal. Let $\mathbf{X}^o$ and $\mathbf{X}^r$ represent the original and reconstructed signal spectrogram from the mixture respectively. The SNR improvement (SNRI) of a speaker is calculated by incorporating the phase information and by comparing the improvement in terms of SNR. Define a gain function for a spectrogram $\mathbf{Y}$

$$g(\mathbf{Y}) = 10\log_{10}\frac{\sum_{f,t}(X^o)_{ft}^2}{\sum_{f,t}|(X^o)_{ft}\exp(j\phi_{ft}) - Y_{ft}\exp(j\phi_{ft})|^2}$$

SNRI is defined as SNRI $= g(\mathbf{X}^r) - g(\mathbf{X})$.

In order to evaluate bandwidth expansion, we use generalized KL divergence (GKL) and Itakuro-Saito (IS) divergence [36], [39]. Both the metrics have been widely used as cost functions in NMF, and are appropriate for computing the distance between two data distributions which are the scaled versions of any probability distributions [36], [39].

**Comparison:** The performances of DLVM and bi-DLVM are compared with those of four existing methods: PLCA [1], PLCA with dynamic filtering [29], PLCA with dynamic smoothing [29] and dynamic NMF with exponential prior [17]. These methods were chosen because they all offer probabilistic interpretations and were developed primarily for source separation. For PLCA with dynamic filtering and smoothing, we report the best results obtained after hyperparameter tuning in our experiments. No hyperparameter tuning is required for dynamic NMF and proposed methods.

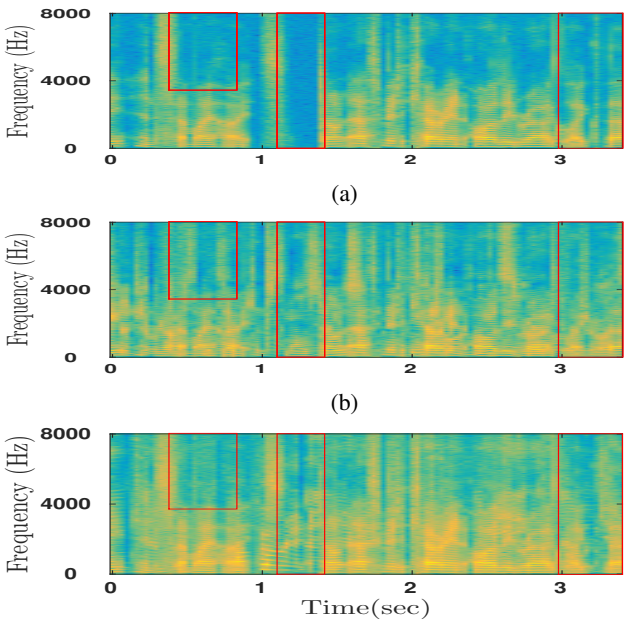


Figure 4: Sample result for speaker source separation: (a) original source, (b) recovered source using PLCA, (c) recovered source using DLVM.

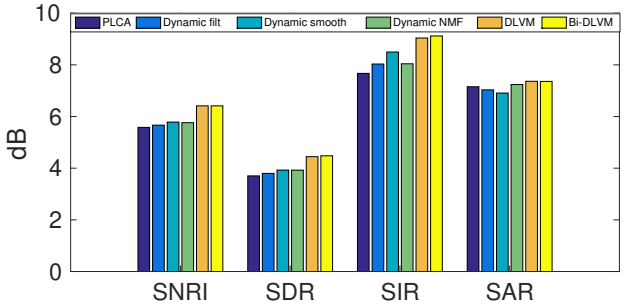


Figure 5: Performance comparison for speaker source separation in terms of four evaluation metrics.

### B. Speaker source separation

We follow an experimental setup similar to that described in the literature of source separation using PLCA and its variants [40], [7]. We have used $\sim$ 25 seconds of speech (8 to 9 sentences) from 20 speakers (10 male, 10 female) in the TIMIT database. To model each speaker (source), the first $\sim$ 17 seconds of the speech is used. The remaining 5 seconds was used to create 190 synthetic mixtures by adding the speech from two speakers. The speech signals were normalized to zero mean and unit variance prior to addition. We learn $K = 30$ latent bases in each case. Finally the spectral vector for each source is reconstructed using steps outlined in section VI-A.

Source separation experiments were performed on 190 mixtures using the proposed DLVM. Fig. 4 presents a qualitative

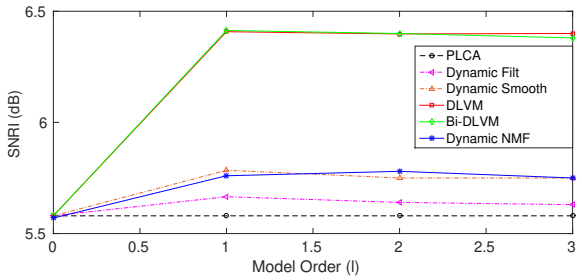


Figure 6: Source separation performance with varying model order.

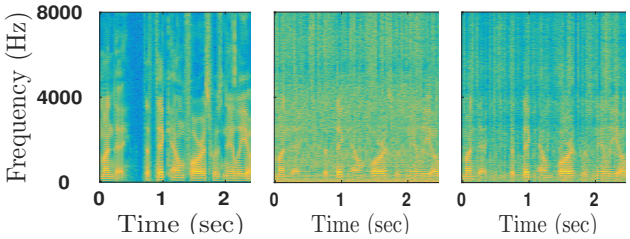


Figure 7: Sample denoising result: (left) original source (left), (center) source recovered by PLCA , and (right) DLVM at 6dB SNR.

Table I: Performance comparison for denoising

| | Babble | Factory | White | Pink | Cockpit | Military | Avg |
|---|---|---|---|---|---|---|---|
| | Average SDR | | | | | | |
| PLCA [1] | 5.92 | 4.01 | 5.75 | 3.55 | 3.97 | 4.26 | 4.58 |
| Dynamic filt [29] | 5.89 | 3.87 | **5.68** | 3.31 | 3.81 | 4.28 | 4.47 |
| Dynamic smooth [29] | 5.95 | 3.22 | 5.69 | 2.75 | 3.61 | 3.96 | 4.19 |
| Dynamic NMF [17] | 5.86 | 5.42 | 5.44 | 3.77 | 4.09 | 3.65 | 4.71 |
| **DLVM** | **6.08** | **5.99** | 5.40 | 5.49 | 4.56 | 4.50 | **5.34** |
| **Bi-DLVM** | 5.71 | 4.56 | 4.59 | **5.63** | **4.59** | **4.77** | 4.98 |
| | Average SAR | | | | | | |
| PLCA [1] | 6.69 | 8.20 | 8.93 | 7.98 | 8.27 | 8.44 | 8.08 |
| Dynamic filt [29] | 6.44 | 7.94 | 8.67 | 7.32 | 8.01 | 8.53 | 7.81 |
| Dynamic smooth [29] | 5.65 | 6.93 | 8.45 | 6.25 | 7.70 | 7.94 | 7.15 |
| Dynamic NMF [17] | **8.30** | 9.09 | 9.09 | 8.27 | 8.27 | 8.30 | 8.55 |
| **DLVM** | 7.22 | 9.38 | **9.51** | **8.94** | **9.18** | 8.78 | **8.83** |
| **Bi-DLVM** | 7.20 | **9.27** | 8.70 | 8.69 | 8.84 | **9.04** | 8.62 |

results on source separation, Fig. 4a shows the original spectrogram while Fig. 4b and Fig. 4c present the reconstructed spectrograms of the given source recovered (from a mixture) using PLCA and DLVM respectively. Notice that DLVM recovers a smoother or better spectrogram (areas exhibiting significant differences are highlighted).

Fig. 5 shows the average values across 190 test cases for all methods. As seen in Fig. 5, DLVM and bi-DLVM perform better than or comparable to the existing methods in terms of all metrics. DLVM outperforms dynamic NMF (with exponential prior) by 0.65 dB in terms of SNRI, 0.52 dB in SDR, 0.99 dB in SIR, and 0.12 dB in SAR. The improvement in terms of SAR implies that the artifacts introduced by DLVM are less compared to other models. Usually, there is a trade-off between removing noise (measured by SDR and SNRI) and introducing artifacts (measured by SAR). The simpler dynamic models ([29]), while improving SDR often introduce more artifacts, which lead to a degraded SAR. However, DLVM and its variant show simultaneous improvement in terms of both SDR and SAR. This indicates an overall better modeling ability of DLVM, which leads to better source separation.

Fig. 6 shows the variation of output SNRI with the model order $l$. Note that $l = 0$ corresponds to the static PLCA. In our experiments, we observe that $l = 1$ is sufficient to capture the temporal dependencies efficiently for dynamic and bi-DLVMs. No further improvement in output SNRI is observed for $l > 1$. The inability of better performance for $l > 1$ can be attributed to the fact that the model only performs temporal smoothing on the coefficient matrix, and longer-term temporal smoothing does not contribute in this case due to the non-stationary nature of the data. We believe that DLVM with $l = 1$ (which has low complexity) should be preferred over other models if MAP estimation is to be performed.

### C. Denoising

We choose six noise types (babble, factory, white, pink, cockpit and military vehicle noise) for this experiments. This noise is added (one at a time) to the speech data of the 20 speakers used in the speaker source separation experiments. Both the noise and the speech are first normalized to have zero mean and unit variance. The noisy mixtures are obtained by adding the noise to each speaker signal. The amount of training data and number of latent bases ($K$) for each speaker and noise are identical to those used in Section VII-B.

Fig. 7 presents a sample (qualitative) denoising result, where the source was corrupted with pink noise at 6 dB SNR. Observe that DLVM yields better reconstruction compared to PLCA - which loses almost all structures at higher frequencies. The performance of DLVM (averaged over 20 mixtures at 6 dB SNR) is presented in Table I along with results from existing methods. DLVM shows an improvement (on average) of 0.59 dB in terms of SDR and 0.28 dB in SAR as compared to dynamic NMF. Interestingly, DLVM performs better than all methods for all noise types, except white noise. This can be explained by the fact that white noise is stationary and has no temporal structure, which DLVM attempts to capture. Nevertheless, our model performs better in all those cases where the noise is non-stationary, as it is able to learn the temporal dependencies in the data and the noise. DLVM shows 0.75 dB SAR improvement on average for all noise types over PLCA. This observation supports our earlier claim that the proposed model introduces less artifacts as compared to other models.

Fig. 8 shows the denoising performance of the proposed and existing methods at different SNR levels. DLVM (or its variant) outperforms all methods for all noise types except white noise at every input SNR level (as observed earlier).

### D. Bandwidth expansion

We obtain speech data of 20 speakers from the TIMIT database, and remove the higher frequencies ($4 - 8$KHz) to generate narrowband ($0 - 4$KHz) speech signals. For each speaker we learn the parameters $(P(f|z), \mathbf{D})$ from the training

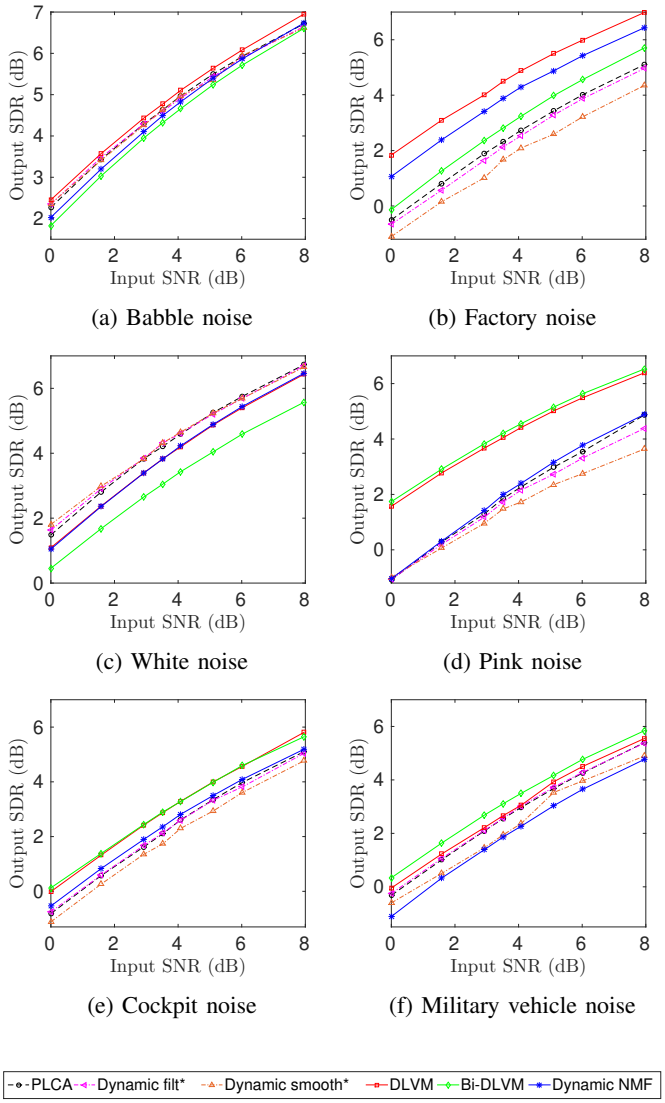


Figure 8: Denoising performance with varying SNR

Table II: Performance comparison for bandwidth expansion in terms of average GKL and IS divergence

| Methods | GKL | IS |
|---|---|---|
| PLCA [1] | 349.3 | 482.5 |
| Dynamic filt [29] | 289.1 | 334.0 |
| Dynamic smooth [29] | 248.8 | 2394.9 |
| Dynamic NMF [17] | 279.5 | 297.6 |
| **DLVM** | **237.8** | **214.7** |
| **Bi-DLVM** | 263.4 | 338.7 |

data. Following the earlier works in the literature, the value of $K$ was chosen to be 100 for better prediction [35]. All other experimental details for the training stage are identical to those described in Section VII-B.

Fig. 9 shows a sample bandwidth expansion result using PLCA and DLVM. The latter yields a smoother spectrum compared to PLCA (difference areas highlighted). Recall that we utilize DLVM only to reconstruct the magnitude spectrogram. The phase spectrogram was predicted separately using least square solution [35]. For quantitative comparison

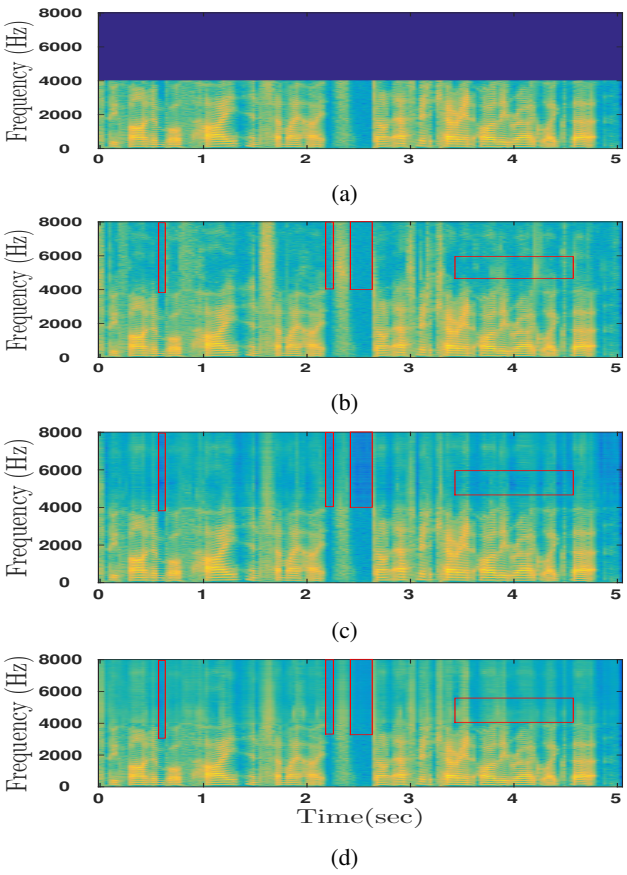


Figure 9: Sample result for bandwidth expansion: (a) bandlimited signal, (b) original signal, (c) bandwidth expansion using PLCA, and (d) bandwidth expansion using DLVM.

of different algorithms for bandwidth expansion we used GKL divergence and IS divergence as metrics. Results averaged over 20 speakers are shown in Table II.

Table II shows that DLVM-based solution has the least KL divergence and IS divergence with respect to the ground truth. This shows that DLVM has better prediction ability, which is due to the better modeling capability of DLVM. Also, we observe that the DLVM outperforms bi-DLVM. This can be attributed due to the i)fact that the model only performs temporal smoothing on the coefficient matrix, and longer-term temporal smoothing does not contribute in this case due to the non-stationary nature of the data, ii)The dependencies in our data are mostly unidirectional.

### *E. Compactness of the representation*

Along with the various results, we would also like to highlight an important observation regarding the state estimates ($\mathbf{s}_t$) obtained using DLVM and its variant. Fig. 10 shows the state estimates for a speaker using PLCA, DLVM and bi-DLVM. It can be seen that the majority of the bases are active in other models, while fewer are active in the case of dynamic and bi-DLVM. Similar trends are observed for other speakers too. We note that imposing temporal dependencies has led to a sparser estimate of the states. Sparser results are often desired and considered to be a better representation

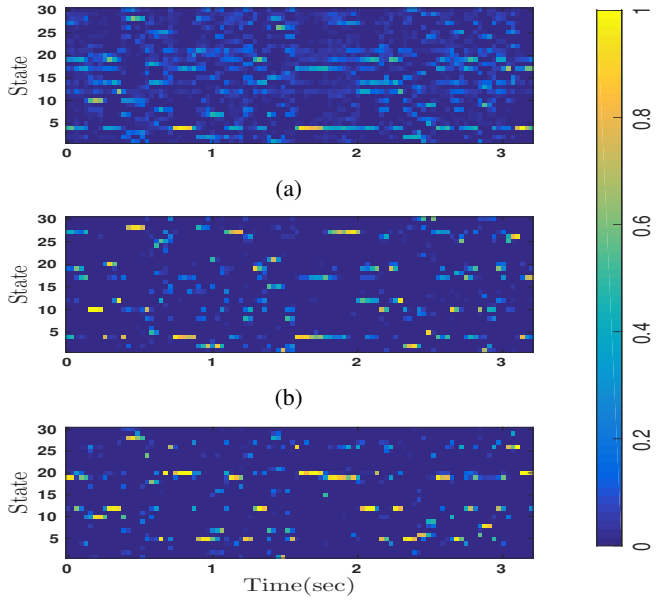


Figure 10: Compactness of representation indicated by state activation in (a) PLCA, (b) DLVM, and (c) bi-DLVM.

Table III: Sparsity of the state matrix $\mathbf{S}$

| Methods | $\ell_0$ | $\ell_{0.1}$ | Entropy |
|---|---|---|---|
| PLCA [1] | 21.07 | $1.94 \times 10^{12}$ | 2.24 |
| Dynamic filt [29] | 20.2 | $1.16 \times 10^{12}$ | 2.14 |
| Dynamic smooth [29] | 17.74 | $5.02 \times 10^{11}$ | 1.94 |
| Dynamic NMF [17] | 19.75 | $2.64 \times 10^{12}$ | 2.10 |
| **DLVM** | 5.59 | $\mathbf{1.64 \times 10^{9}}$ | 0.82 |
| **Bi-DLVM** | **5.58** | $2.13 \times 10^{9}$ | **0.81** |

[41]. To quantify this sparsity, we use Shannon's entropy, $\ell_{0.1}$ and $\ell_0$ norm [7], [42]. We compute the $\ell_0$ norm after hard thresholding the values at $0.001$. Table III shows the sparsity of the state matrix for different methods. The $\ell_0$ norm shows that the average number of active latent bases is the smallest ($5.58$) in the bidirectional case. The results show that the proposed model naturally learns more compact representation, as we have not used any sparsity prior explicitly in the model. The Dirichlet prior induces sparsity only when the values of each of its parameters are less than 1. However, the parameters in the proposed dynamic Dirichlet prior are defined to be greater than 1, which promotes a dense distribution [3]. This leads us to hypothesize that the sparsity observed in the proposed model is due to its ability to learn better representation. In contrast, exponential prior in dynamic NMF ([17]) has an average of 19.7 active bases. Also in Fig., 10, we observe that the posterior estimates of states (although, we have imposed smooth prior on states,) are able to capture sharp transition better than PLCA.

## VIII. Conclusion

We proposed a dynamic latent variable model, called the Dirichlet latent variable model, for learning latent bases from time varying non-negative data. To capture the temporal structures in data efficiently, we introduced a dynamic Dirichlet prior - a Dirichlet distribution with dynamic parameters. A major contribution of this work is to introduce the dynamic Dirichlet prior for non-negative data. We showed that the expected log-likelihood function is concave and can be solved by standard convex optimization methods. This property arises because of (a) Dirichlet-multinomial conjugacy and (b) Dirichlet and multinomial are member of exponential family distributions. The proposed DLVM can be interpreted and implemented as a dynamic version of NMF. We also showed that the popular PLCA model is a special case of DLVM. An EM algorithm was developed for the parameter estimation of DLVM. Through extensive experiments, we demonstrated that DLVM outperforms several existing methods for three applications: speaker source separation, denoising and bandwidth expansion. Unlike existing dynamic models (which contain annealing hyperparameter), DLVM does not require any free parameters to be set by the user, other than the number of bases to be learned. The updates of latent bases and states is independent of scaling factor. Therefore, DLVM can handle non-count data as well. Although the current work in this paper involves modeling magnitude spectra of audio signals (non-count data), DLVM is suitable for modeling other types of non-negative data, such as word count data which appears widely in natural language processing. We also showed that the dynamic Dirichlet prior leads to sparse states (which is well proven to be better representation). Also, we observe that the posterior estimates of states can have sharp transition and has been captured by our model efficiently.

Future work will be directed towards developing faster updates for dependence matrices ($\mathbf{D_1}^+$ and $\mathbf{D_1}^-$ in this paper), and a fully Bayesian inference algorithm instead of a MAP estimate of states.

## Acknowledgement

The authors would like to thank Prof. Ketan Rajawat, IIT Kanpur for various technical discussions related to this work.

## References

[1] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing, NIPS*, vol. 148, pp. 8–1, 2006.

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] B. Raj, M. V. Shashanka, and P. Smaragdia, "Latent dirichlet decomposition for single channel speaker separation," in *Acoustics, Speech and Signal Processing, ICASSP Proceedings. IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[5] A. Hadgu and Y. Qu, "A biomedical application of latent class models with random effects," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 4, pp. 603–616, 1998.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] M. Shashanka. (1999) Results and demos. [Online]. Available: http://cns.bu.edu/~mvss/courses/speechseg/

[8] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of machine learning research*, vol. 5, no. Nov, pp. 1457–1469, 2004.

[9] Z. Chen and A. Cichocki, "Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints," *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep*, vol. 68, 2005.

[10] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[11] S. Hiroya, "Non-negative temporal decomposition of speech parameters by multiplicative update rules," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2108–2117, 2013.

[12] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[13] W. Wang, A. Cichocki, and J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2858–2864, 2009.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational em algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[15] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 140–148.

[16] A. Kumar, T. Guha, and P. Ghosh, "A dynamic latent variable model for source separation," in *Acoustics, Speech and Signal Processing, 2018. ICASSP 2018 Proceedings. 2018 IEEE International Conference on*.

[17] N. Mohammadiha, P. Smaragdis, G. Panahandeh, and S. Doclo, "A state-space approach to dynamic nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 949–959, 2015.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[19] U. K. Speech Research Unit (SRU) at Institute for Perception-TNO, Netherlands. Signal processing information base (SPIB). [Online]. Available: http://spib.linse.ufsc.br/noise.html

[20] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[21] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[22] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis, tech report," 2007.

[23] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[24] M. S. Grewal, "Kalman filtering," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 705–708.

[25] S. J. Julier and J. J. LaViola, "On kalman filtering with nonlinear equality constraints," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2774–2784, 2007.

[26] D. Simon, "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms," *IET Control Theory & Applications*, vol. 4, no. 8, pp. 1303–1318, 2010.

[27] A. Carmi, P. Gurfil, and D. Kanevsky, "Methods for sparse signal recovery using kalman filtering with embedded pseudo-measurement norms and quasi-norms," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2405–2409, 2010.

[28] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3158–3162.

[29] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in nmf," in *Acoustics, Speech and Signal Processing (ICASSP), International Conference on*. IEEE, 2013, pp. 873–877.

[30] K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons, 2011, vol. 888.

[31] J. M. Dickey, "Multiple hypergeometric functions: Probabilistic interpretations and statistical uses," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 628–637, 1983.

[32] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational intelligence and neuroscience*, 2008.

[33] X. Yu, D. Hu, and J. Xu, *Blind source separation: theory and applications*. John Wiley & Sons, 2013.

[34] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[35] B. Raj, R. Singh, M. Shashanka, and P. Smaragdis, "Bandwidth expansionwith a pólya urn model," in *Acoustics, Speech and Signal Processing, ICASSP . IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–597.

[36] M. Shashanka, "Latent variable framework for modeling and separating single-channel acoustic sources," Ph.D. dissertation, BOSTON UNIVERSITY Boston, 2007.

[37] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[38] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide–revision 2.0," 2005.

[39] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," 2009.

[40] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2005, pp. 17–20.

[41] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: computation in neural systems*, vol. 7, no. 2, pp. 333–339, 1996.

[42] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE signal processing magazine*, vol. 24, no. 4, pp. 118–121, 2007.


**Anurendra Kumar** completed his B.tech-M.tech in Electrical Engineering from IIT Kanpur in 2017. His interest is in developing robust, scalable and efficient machine learning algorithms for real world applications. Currently, he is a co-founder in a start-up project, where he leads the team focused on designing mathematical techniques for futuristic financial services and risk management.

**Tanaya Guha** is an Assistant Professor in the department of Computer Science, University of Warwick, UK. Prior to joining Warwick, she worked in IIT Kanpur as an assistant professor, and in University of Southern California as a postdoctoral research fellow. She has received her PhD in Electrical and Computer Engineering from the University of British Columbia (UBC), Vancouver. She was a recipient of Mensa Canada Woodhams memorial scholarship, Google Anita Borg scholarship and Amazon Grace Hopper celebration scholarship. Her current research interests include multimodal signal processing and multimedia analysis.

**Prasanta Kumar Ghosh** received the B.E.(ETCE) in electronics from Jadavpur University, Kolkata, in 2003, M.Sc.(engineering) in electrical communication engineering from Indian Institute of Science (IISc), Bangalore, in 2006, and the Ph.D. in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2011. He is an Assistant Professor in the department of Electrical Engineering, IISc. His research interests include nonlinear signal processing methods with applications to speech and audio, audio–visual signal processing, and biomedical signal processing.