A DYNAMIC LATENT VARIABLE MODEL FOR SOURCE SEPARATION

Anurendra Kumar[1]*, Tanaya Guha*[1]*, Prasanta Ghosh*[2]

[1]Indian Institute of Technology (IIT), Kanpur. [2]Indian Institute of Science (IISC), Bangalore.

ABSTRACT

We propose a novel latent variable model for learning latent bases for time-varying non-negative data. Our model uses a mixture multinomial as the likelihood function and proposes a Dirichlet distribution with dynamic parameters as a prior, which we call the *dynamic Dirichlet* prior. An expectation maximization (EM) algorithm is developed for estimating the parameters of the proposed model. Furthermore, we connect our proposed *dynamic Dirichlet latent variable model* (dynamic DLVM) to the 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 the dynamic DLVM, and (ii) dynamic DLVM can be interpreted as a dynamic version of NMF. The effectiveness of the proposed model is demonstrated through extensive experiments on speaker source separation, and speech-noise separation. In both cases, our method performs better than relevant and competitive baselines. For speaker separation, dynamic DLVM shows 1.38 dB improvement in terms of source to interference ratio, and 1 dB improvement in source to artifact ratio.

Index Terms— Latent variable model, Dirichlet distribution, non-negative matrix factorization, source separation

1. INTRODUCTION

Modeling time-varying, non-negative data is critical for many signal processing problems. One such important problem is audio source separation, where time varying, non-negative data arise in the form of magnitude spectra [1]. Source separation is a long standing problem in signal processing, which has widespread applications in speaker recognition, speech enhancement, music editing and audio information retrieval [2, 3].

This paper addresses the problem of modeling time varying non-negative data, looking particularly at the problem of supervised source separation. In the case of supervised source separation, we assume the availability of training data for each source [1, 4]. The training data is used to learn the underlying building blocks i.e., the *latent bases* for each source. These latent bases are later used to separate the sources from the mixture. Two techniques that have been prominent in this field for learning latent bases are: latent variable model (LVM) [4] and non-negative matrix factorization (NMF) [5, 6]. One of the most popular LVM methods is the probabilistic latent component analysis (PLCA), which has widespread application in acoustic modeling [1]. NMF, on the other hand, is a non-probabilistic approach to learn latent basis with extensive applications to text, image and audio analysis. For certain cost functions, LVM is known to converge to NMF [7], and can be thought of as the probabilistic counterpart of NMF.

The LVM and the NMF in their basic forms do not take into account the temporal correlation in the data. However, many signals, such as music and speech exhibit strong temporal dependence. To address this issue, dynamic variants of LVM and NMF have been developed [8, 9, 10]. Most of these dynamic models capture the temporal dependence in data by imposing temporal constraints on the latent bases and their coefficients [8, 9]. A dynamic variant of PLCA, called the Convolutive PLCA (CPLCA) [8], was proposed to capture the temporal structure in data by assuming the likelihood to be a convolutive mixture. Another dynamic version of PLCA involves dynamic filtering and smoothing, where the coefficient matrix was smoothened by using a vector autoregression (VAR) method [11]. A recent work developed a dynamic NMF (an extension of PLCA) by using an exponential prior [10]. Apart from the LVMs and NMF methods, the hidden Markov model (HMM) has also been extended to model temporal non-negative data [12].

In this paper, we present a novel dynamic LVM for learning latent bases for time varying, non-negative data. Our model uses a mixture multinomial as the likelihood function, and proposes to use a Dirichlet distribution with dynamic parameters as a prior (referred to as the *dynamic Dirichlet* prior). The mixture multinomial likelihood function is chosen because it is known to yield superior results in source separation and topic modeling [13, 14]. The Dirichlet distribution is the conjugate of multinomial, and has been successfully used (*without* dynamic properties) as a prior in text modeling [13]. We propose a dynamic variant of the Dirichlet prior in this work, which is particularly suitable for non-negative data. We develop an expectation maximization (EM) algorithm for the proposed model, and derive a maximum a posteriori (MAP) estimate of the parameters. This leads to simple and intuitive update equations due to multinomial-Dirichlet conjugacy. We refer our model as the dynamic Dirichlet latent variable model (dynamic DLVM) in the rest of the paper. Furthermore, we show that (i) the PLCA model is a special case of the dynamic DLVM, and (ii) our model is also a dynamic version of NMF. The effectiveness of the dynamic DLVM is demonstrated through extensive experiments on speaker source separation and speech-noise separation using the SPIB [15] and the TIMIT database [16]. In all cases, our method performs better than several relevant existing methods.

2. PROPOSED DYNAMIC DIRICHLET LATENT VARIABLE MODEL

Let us consider a time varying signal $x(t)$. We represent $x(t)$ spectrographically by taking its short time Fourier transform (STFT), and retaining its scaled magnitude spectrogram $\mathbf{N}$

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

where, $\mathbf{X}$ is magnitude spectrogram, γ is a large integer which ensures that all the elements in $\mathbf{N}$ are integers [1]. The observation spectral data matrix $\mathbf{N}$ can be seen as count data, where $\mathbf{N}_{ft} \in \mathbf{N}$ corresponds to the count of frequency f, at a time instant t. Each column of the matrix $\mathbf{N}$ thus corresponds to the spectral distribution at each 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 rest as zero, $\mathbf{z} = [z_1, z_2, ...z_K]$. z_k acts as an indicator for the k-th latent basis, which is described by a spectral distribution $P(f|z_k)$.

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on August 24,2022 at 11:07:50 UTC from IEEE Xplore. Restrictions apply.

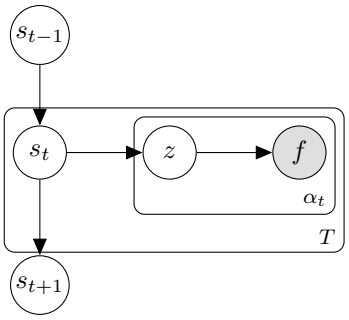

Fig. 1: Plate notation for the proposed dynamic DLVM.

Latent variable models assume that the underlying cause of an observed variable f, is a set of unobserved latent variables $z_k, 1 \leq k \leq K$. Marginalizing over the latent bases $\mathbf{z}$, the spectrogram at each time instant t, is a mixture of K hidden distributions, where K is a known positive integer

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

where, $P_t(f)$ is the probability of frequency f at time instant t, $P(f|z_k)$ is a multinomial distribution similar to PLCA [1] and the coefficients of mixtures are $P_t(z_k)$. Let us now define a state of the data matrix $\mathbf{N}$ at time t as $\mathbf{s}_t$ as follows:

$$\mathbf{s}_t = [P_t(z_1), ..., P_t(z_K)]^{\mathrm{T}} = [s_t(1), s_t(2)..., s_t(K)]^{\mathrm{T}} \tag{3}$$

We impose a Markovian dependence between states, which uses a Dirichlet distribution. This is described below in detail.

2.1. Dynamic DLVM

We propose to model the temporal dependence between states using a Dirichlet distribution with time varying parameters

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

Here, "Dir" denotes the Dirichlet distribution [17], $\mathbf{1}$ is an all-one vector, α_t is the total number of observations at time instant t, and $\mathbf{D}$ is a temporal dependence matrix. For simplicity, we assume $\mathbf{D}$ to be a diagonal matrix with $\mathbf{D}_{kk} = d_k$, $1 \leq k \leq K$, where $d_k \in \mathbb{R}^+$ denotes the temporal dependence between two consecutive time instants for the k-th latent basis. Let us now define pseudo-observation from the previous time instant for each basis k as $\mathbf{m}_{tk} = \alpha_{t-1}\mathbf{d}_k\mathbf{s}_{t-1}(k)$. Therefore Eq. (4) can be rewritten as follows:

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

where, Γ is the gamma function. Note that the the hyperparameters of the Dirichlet distribution are dynamic, hence, we refer to it as the *dynamic Dirichlet distribution* in the rest of this paper. The proposed dynamic Dirichlet distribution prior has several appealing properties with intuitive understanding: (i) The generative process (with mixture multinomial as likelihood) allows us to view the spectrogram at time t as 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 $\mathbf{m}_{tk}$ extra pseudo-observations for each basis k (see Eq. (8)), which is the result of multinomial-Dirichlet conjugacy. (ii) The mode of the distribution lies at the normalized pseudo-observations

$$P_{\mathbf{max}}(\mathbf{s}_t(k)|\mathbf{s}_{t-1}) = \frac{\mathbf{d}_k\mathbf{s}_{t-1}(k)}{\sum_k \mathbf{d}_k\mathbf{s}_{t-1}(k)}$$

(iii) The variance of each entry of the vector $\mathbf{s}_t$ can be obtained from the properties of Dirichlet distribution [17]

$$Var(\mathbf{s}_t(k)|\mathbf{s}_{t-1}) \propto \frac{1}{(\sum_k \mathbf{m}_{tk} + K)^2(\sum_k \mathbf{m}_{tk} + K + 1)}$$

which decreases as total number of observations at previous time instant increases. It is also intuitive, since we expect the distribution to have less variance when we have more prior information from previous time instant. Thus *dynamic DLVM* assumes the following generative process of the spectrogram $\mathbf{N}$:

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

where, T is the total number of time instants, "Mult" denotes the multinomial distribution. Fig. 1 presents the graphical model for the generative process.

2.2. PLCA as a special case of dynamic DLVM

The relationship between the proposed dynamic DLVM and the well known PLCA [1] model is noteworthy. When the temporal dependence matrix $\mathbf{D}$ is reduced to a zero matrix, the distribution in Eq. (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 is equivalent to a static PLCA [18]. This can also be intuitively understood as a fact that in the absence of prior information, there exists no preference of any state over others.

3. PARAMETER ESTIMATION

In this section, we describe the parameter estimation steps for the proposed dynamic DLVM. We obtain point estimates for $\mathbf{s}_t$ by performing a maximum a posteriori (MAP) estimate.

3.1. Expectation step

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

$$P_t(z_k|f) = \frac{P_t(z_k)P(f|z_k)}{\sum_k P_t(z_k)P(f|z_k)} \tag{5}$$

3.2. Maximization step

Let us denote the state matrix $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ...\mathbf{s}_t]$. Let β= $\{P(f|\mathbf{z}), \mathbf{D}\}$. We want to maximize the following MAP objective function

$$\mathcal{L}_{MAP} = \mathop{\mathbb{E}}_{P_t(\mathbf{z}|f)}(\log P(\mathbf{N}, \mathbf{S}|\beta)) = \mathop{\mathbb{E}}_{P_t(\mathbf{z}|f)} \log P(\mathbf{N}|\mathbf{S}, \beta) + \log P(\mathbf{S}|\beta)$$
$$s.t., \sum_f P(f|z_k) = 1, \sum_k s_t(k) = 1, 0 < d_k \tag{6}$$

The objective function $\mathcal{L}_{\mathcal{MAP}}$ is concave with respect to each parameters $(\mathbf{S}, P(f|\mathbf{z}), \mathbf{D})$ when others are fixed[1]. Therefore, we update the parameters in an alternating fashion [19].

[1]Our proof of concavity: https://tinyurl.com/y7qclrc6

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on August 24,2022 at 11:07:50 UTC from IEEE Xplore. Restrictions apply.

Update of $P(f|\mathbf{z})$

Maximizing the above constrained expected log-likelihood in Eq. (6) with respect to $P(f|z_k)$ yields the following

$$P(f|z_k) = \frac{\sum_t \mathbf{N}_{ft} P_t(z_k|f)}{\sum_f \sum_t \mathbf{N}_{ft} P_t(z_k|f)} \tag{7}$$

Update of S

Similarly, maximizing w.r.t. $\mathbf{s}_t(k)$, while keeping $\mathbf{D}$ fixed yields

$$\mathbf{s}_t(k) = \frac{\sum_f \mathbf{N}_{ft} P_t(z_k|f) + \mathbf{m}_{tk}}{\sum_k (\sum_f \mathbf{N}_{ft} P_t(z_k|f) + \mathbf{m}_{tk})} \tag{8}$$

Update of D

Let us define $\mathbf{d} = [d_1, ..., d_K]^{\mathrm{T}}$. Maximizing $\mathcal{L}_{MAP}$ w.r.t. $\mathbf{d}$, keeping $\mathbf{S}$ fixed does not have a closed form solution

$$\mathbf{d} = \underset{\mathbf{d}}{\operatorname{argmax}} \sum_{t=2}^{T} (\log\Gamma(\sum_k (\mathbf{m}_{tk}+1)) - \sum_k \log\Gamma(\mathbf{m}_{tk}+1) + \sum_k \mathbf{m}_{tk} \log(\mathbf{s}_t(k)) \quad s.t., 0 < \mathbf{d}_k. \tag{9}$$

However, the maximizing function is concave since the Dirichlet distribution is a member of the exponential family[1]. Therefore the local minimum of the function is also a global minimum, which can be obtained via gradient ascent

$$\frac{\partial \mathcal{L}_{MAP}}{\partial \mathbf{d}_k} = \sum_{t=2}^{T} \alpha_{t-1}\mathbf{s}_{t-1}(k)(\psi(\sum_i \mathbf{m}_{ti} + K) - \psi(\mathbf{m}_{tk}+1) + \log(\mathbf{s}_t(k))).$$

where, ψ is the di-gamma function. It is to be noted that updates of $P(f|\mathbf{z})$ and $\mathbf{S}$ are independent of scaling factor γ. Therefore, we will replace $\mathbf{N}$ with $\mathbf{X}$ in all the update equations. The complete algorithm is presented in the next section.

4. DYNAMIC DLVM AS DYNAMIC NMF

The proposed dynamic DLVM learns the latent bases and the states for a data matrix via factorization in Eq. (2). Multiplying both sides of the equation by α_t, we rewrite Eq. (2) in matrix form as $\mathbf{v}_t = \mathbf{W}\mathbf{s}_t\alpha_t$, where, $\mathbf{W}$ is a matrix whose columns are latent bases $P(f|\mathbf{z})$, $\mathbf{v}_t$ is the observation vector at time instant t. Concatenating observation vector 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 basis matrix, $\mathbf{S}$ is a state matrix and $\mathbf{G}$ is a diagonal matrix with α_t as the diagonal elements. Therefore, we can view dynamic DLVM as a dynamic NMF with the iterative updates for $\mathbf{W}$ and $\mathbf{S}$ (see Algorithm 1). The novelty of dynamic DLVM lies in the way $\mathbf{S}$ is constrained. The columns of $\mathbf{S}$ are assumed to be realization from *dynamic Dirichlet* distribution. In the algorithm, 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.

Algorithm 1 Dynamic DLVM as Dynamic NMF

Input: $\mathbf{X}$
Output: $\mathbf{W}, \mathbf{S}, \mathbf{d}$
Randomly initialize $\mathbf{W}, \mathbf{S}$,$\mathbf{d}$
while *Not converged* **do**
  $\mathbf{W}_{fk} = \mathbf{W}_{fk} \sum_t \frac{\mathbf{X}_{ft}}{(\mathbf{WS})_{ft}} \mathbf{S}_{kt}$
  $\mathbf{W}_{fk} = \mathbf{W}_{fk} / \sum_k \mathbf{W}_{fk}$
  while *Not converged* **do**
    $\mathbf{m}_{tk} = \alpha_{t-1}\mathbf{d}_k\mathbf{s}_{t-1}(k)$
    $\mathbf{S}_{kt} = \mathbf{S}_{kt} \sum_f \mathbf{W}_{fk} \frac{\mathbf{X}_{ft}}{(\mathbf{WS})_{ft}} + \mathbf{m}_{tk}$
    $\mathbf{S}_{kt} = \mathbf{S}_{kt} / \sum_t \mathbf{S}_{kt}$
    Update $\mathbf{d}$ using Eq. 9
  end
end

5. APPLICATION TO SOURCE SEPARATION

We demonstrate the effectiveness of the proposed dynamic DLVM through its ability to perform supervised source separation. We first learn the basis matrix $\mathbf{W}$ for each source, and then separate the sources by employing a standard source separation algorithm following an earlier work [4]. Note that we only use the latent bases (and not the dependency matrix $\mathbf{D}$) during the source separation. We perform two types of source separation experiments: (i) speaker source separation, where mixtures contain speech sources from two speakers, and (ii) noise-source separation, where mixtures contain speech and noise.

Experimental details

We follow an experimental setup similar to that described in past works on source separation using PLCA and its variants [20, 21]. The source separation experiments use samples from the TIMIT database [16], and noise data from SPIB [15]. The magnitude spectrograms are obtained by performing STFT on a 64ms window with 16ms overlap. We learn $K = 30$ latent basis in each case, and use the maximum number of iterations (250 for outer loop, 8 for inner loop) as the convergence criterion for Algorithm 1.

To evaluate the performance on source separation, we use four evaluation metrics: signal-to-noise ratio improvement (SNRI) [22], source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) [23, 24, 25]. The later three metrics measure perceptual quality of the separated sources. The SNR improvement (SNRI) of a speaker is calculated by incorporating the phase information and by comparing the improvement in SNR with that of the mixture signal as defined in literature [22, 4].

Speaker source separation

Following the experimental set up of the past literature [20, 21], we have used ~ 25 seconds of speech (8 to 9 sentences) from 10 speakers (5 male, 5 female) in the database for our experiments. To model each speaker (source), the first ~ 17 seconds of the speech is used. The remaining 5 to 7 seconds of speech were used to create 45 synthetic mixtures by adding the speech from two speakers. The speech signals were normalized to zero mean and unit variance prior to addition. Source separation experiments were performed on 45 mixtures using the proposed dynamic DLVM. Fig. 2 presents a qualitative result of source separation. Fig. 2b and Fig. 2c present the reconstructed spectrograms of a given source recovered from a mixture

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on August 24,2022 at 11:07:50 UTC from IEEE Xplore. Restrictions apply.

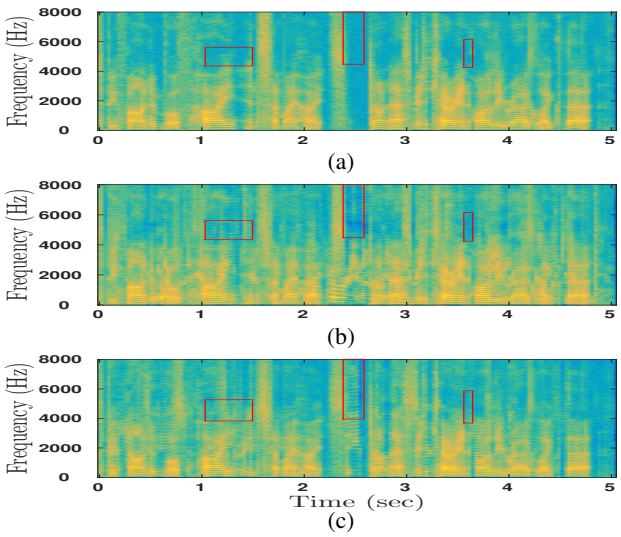

Fig. 2: (a) Original source, (b) recovered source using PLCA, and (c) recovered source using dynamic DLVM.

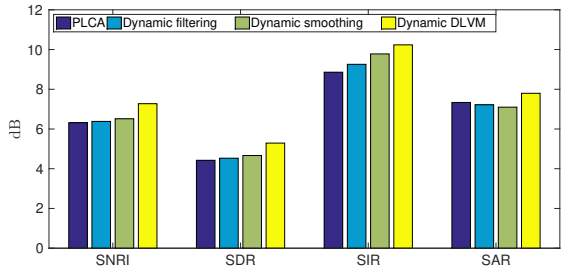

Fig. 3: Results on speaker separation: Dynamic DLVM compared with three existing techniques in terms of four evaluation metrics.

using PLCA and dynamic DLVM. Notice that the dynamic DLVM recovers a smoother spectrogram (areas of significant differences are highlighted). The performance of dynamic DLVM is evaluated in terms of the four evaluation metrics mentioned earlier (see Fig. 3). The performance of dynamic DLVM is compared against those of three baseline methods – PLCA [1], PLCA with dynamic filtering [11] and PLCA with dynamic smoothing [11]. Dynamic DLVM performs better than or comparable to the baseline methods in terms of all evaluation metrics. Our model outperforms PLCA by 0.96 dB in SNRI, 0.87 db in SDR, 1.38 db in SIR, and 0.46 db in SAR. The improvement in terms of SAR implies that the artifacts introduced by dynamic DLVM is lesser than the other models. Usually, there is a trade-off between removing noise (measured by SDR and SNRI) and introducing artifacts (measured by SAR). The existing dynamic models [11, 10, 26, 12] while improving SDR often introduce artifacts, which leads to a degraded SAR. However, the proposed model shows simultaneous improvement in SDR and SAR. This indicates an overall better modeling ability of dynamic DLVM, and consequently, a better source separation.

Speech and noise separation

We consider a speech denoising scenario where prior information about the noise types and the associated training data is available. Both the noise and the speech are first normalized to have zero mean and unit variance. The noisy mixtures were obtained by adding noise

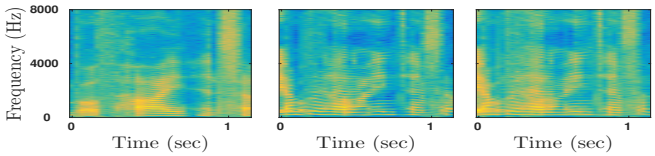

Fig. 4: Original source (left); recovered source using PLCA (center) and dynamic DLVM (right) from a noisy signal.

Table 1: Comparison of different methods for noise separation

	Average SNRI				
	Babble	Factory	White	Pink	Cockpit
PLCA [1]	5.63	2.60	5.07	2.04	2.78
Dynamic filtering [11]	4.93	2.87	**5.83**	2.06	2.70
Dynamic smoothing [11]	4.30	2.99	5.36	2.14	2.38
Dynamic DLVM	**5.83**	**5.30**	3.90	**4.60**	**3.03**
	Average SAR				
PLCA [1]	6.69	8.14	8.30	7.82	7.84
Dynamic filtering [11]	6.44	7.73	5.25	5.97	4.36
Dynamic smoothing [11]	5.65	7.73	3.98	7.44	3.21
Dynamic DLVM	**7.22**	**8.75**	**9.92**	**8.66**	**9.13**

(one at a time) to each speaker signal, resulting into a signal to noise ratio of 0 dB. We experiment with five noise types: babble, factory, white, pink and cockpit [15], and speech from 10 speakers used in the speaker source separation experiments. The latent bases are learned for each speaker and noise from their respective training data (same parameter values as before). Fig. 5 presents a sample qualitative result of denoising a source corrupted with babble noise. The performance of dynamic DLVM averaged over 10 mixtures is listed in Table 1, and compared with the baselines. Dynamic DLVM, on average, shows an improvement of 0.9 dB. Note that it performs better than other 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. Nevertheless, for non-stationary noise, the proposed model is able to learn the temporal dependencies in data/noise, which results in better separation. As observed earlier, dynamic DLVM shows 1dB SAR improvement for all noise types as compared to PLCA. This observation supports our earlier claim that our model introduces less artifacts compared to other dynamic models in literature [11].

6. CONCLUSION

We proposed a latent variable model, called the dynamic DLVM, for modeling time varying non-negative data. We introduced a new prior (dynamic Dirichlet distribution) and used a multinomial as likelihood for this model. An EM algorithm was proposed accordingly for parameter estimation. We showed that the popular PLCA model is a special case of our model. A major contribution of this paper is to introduce this dynamic Dirichlet prior for non-negative data. The existing dynamic variant of Dirichlet can not be used under non-negativity constraints as it yields negative updates. Due to the proposed dynamic Dirichlet prior, the dynamic DLVM transforms to a dynamic version of NMF. Unlike other dynamic latent variable models, our model does not require any free parameter (except the number of latent bases). Although this work involves modeling magnitude spectra, the proposed model is generic and suitable for modeling other types of non-negative data, for example, word count data that appears widely in natural language processing.

7. ACKNOWLEDGEMENT

The authors would like to thank IIT Kanpur (grant: IITK/EE/2015052) and Pratiksha Trust for their support.

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on August 24,2022 at 11:07:50 UTC from IEEE Xplore. Restrictions apply.

8. REFERENCES

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

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

[3] 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.

[4] B. Raj, M. V. Shashanka, and P. Smaragdia, "Latent dirichlet decomposition for single channel speaker separation," in *ICASSP*, vol. 5. IEEE, 2006.

[5] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *Proc. ICA Research Network International Workshop*, 2006, pp. 17–20.

[6] 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.

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

[8] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE TASLP*, vol. 15, no. 1, pp. 1–12, 2007.

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

[10] 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.

[11] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *ICASSP*. IEEE, 2013, pp. 873–877.

[12] 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.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, no. Jan, pp. 993–1022, 2003.

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

[15] 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

[16] 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*, vol. 93, 1993.

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

[18] M. Girolami and A. Kabán, "On an equivalence between plsi and lda," in *Proc. of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 433–434.

[19] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[20] 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.*, 2005, pp. 17–20.

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

[22] ——, "Latent variable framework for modeling and separating single-channel acoustic sources," Ph.D. dissertation, BOSTON UNIVERSITY, 2007.

[23] C. Févotte, R. Gribonval, and E. Vincent. (2005) BSS_EVAL toolbox user guide–revision 2.0. [Online]. Available: https://hal.inria.fr/inria-00564760/

[24] 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.

[25] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 734–741.

[26] 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.

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on August 24,2022 at 11:07:50 UTC from IEEE Xplore. Restrictions apply.