GAMMA HIDDEN MARKOV MODEL AS A PROBABILISTIC NONNEGATIVE MATRIX FACTORIZATION

Nasser Mohammadiha, W. Bastiaan Kleijn, Arne Leijon

KTH Royal Institute of Technology, Department of Electrical Engineering, Stockholm, Sweden

ABSTRACT

Among different Nonnegative Matrix Factorization (NMF) approaches, probabilistic NMFs are particularly valuable when dealing with stochastic signals, like speech. In the current literature, little attention has been paid to develop NMF methods that take advantage of the temporal dependencies of data. In this paper, we develop a hidden Markov model (HMM) with a gamma distribution as output density function. Then, we reformulate the gamma HMM as a probabilistic NMF. This shows the analogy of the proposed HMM and NMF, and will lead to a new probabilistic NMF approach in which the temporal dependencies are also captured inherently by the model. Furthermore, we propose an expectation maximization (EM) algorithm to estimate all the model parameters. Compared to the available probabilistic NMFs that model data with Poisson, multinomial, or exponential distributions, the proposed NMF is more suitable to be used with continuous-valued data. Our experiments using speech signals shows that the proposed approach leads to a better compromise between sparsity, goodness of fit, and temporal modeling compared to state-of-the-art.

Index Terms— Hidden Markov Model (HMM), Nonnegative Matrix Factorization (NMF), Expectation Maximization (EM) algorithm.

1. INTRODUCTION

In recent years, nonnegative matrix factorization (NMF) has attracted the interest of many researchers. As a result, different approaches have been developed to obtain NMF using a variety of criteria [1].

In its basic form, NMF finds a locally optimal and deterministic approximation of a nonnegative matrix $\mathbf{x}$ in the form of a product of two nonnegative matrices $\mathbf{v}$ and $\mathbf{w}$, i.e., $\mathbf{x} \approx \mathbf{vw}$ where $\mathbf{x}$ is a matrix of dimension $K \times T$, $\mathbf{v}$ is a matrix of dimension $K \times I$, and $\mathbf{w}$ is a matrix of dimension $I \times T$, where I is the number of the basis vectors, and usually $I < K$ and $I < T$.

In this paper, we develop an HMM with gamma output density functions and show that it is equivalent to a probabilistic NMF. A number of NMF algorithms have been recently derived in a probabilistic framework [2–5]. In [4], a Bayesian approach was proposed to perform NMF. In this method, it is assumed that data is drawn from an exponential distribution while the rate parameter of the distribution is factorized using an NMF. The model is constructed in such a way that it is possible to infer the optimal number of the NMF basis vectors automatically. However, an important aspect of data (which exists in most of the potential applications), the temporal correlation, is ignored. For example, temporal dependencies are very important in source separation and speech enhancement [6].

Hidden Markov modeling is a strong yet simple approach to capture the temporal aspects of the processes. Methods that combine the NMF paradigm (which in principle ignores the temporal dependencies) and the HMM paradigm have recently been introduced [5, 7]. In [7], the Itakura-Saito NMF was combined with HMM. In this approach, each sample of the complex data was assumed to be a sum of some complex-valued Gaussian components with covariance matrices factorized using NMF. Moreover, a separate Markov chain was considered to govern the transitions between different states of each component independently. Another nonnegative hidden Markov model (NHMM) was proposed in [5] in which the model has a Markov chain with a number of states. In each state, the observed data is assumed to be drawn from a multinomial mixture model. That is, given a state, the data is assumed to be generated by a linear combination of the nonnegative basis vectors corresponding to that state. Nevertheless, due to the assumption of a multinomial distribution, the observed data has to be scaled to be integer. The integer assumption was also used in [2] where the data was assumed to be drawn from a Poisson distribution. Although the observed data might be scaled to be integer in practice, the scaling level will directly affect the assumed noise level in the model, and it might create side problems.

In this paper, we devise an HMM in which the output density functions are assumed to be gamma distributions to cope with nonnegative data. The choice of a gamma distribution provides a more flexible modeling than the exponential distribution considered in [4]. Also, the approach does not need data-scaling as required in [2, 5]. Then, we reformulate the gamma HMM as a probabilistic NMF. Each state of the HMM leads to a basis vector in the NMF representation, and the temporal correlation of the data is imposed through the transition probability between the states. To take care of the time-varying level of the data (e.g., the long-term level of the speech energy), an explicit gamma prior distribution is considered over the gain variable. We propose an efficient EM algorithm to estimate the time-varying and the stationary

model parameters. The main contribution of this work, in addition to the novel HMM structure and the proposed approach to estimate HMM parameters, is the NMF formulation of the developed HMM, which shows how a gamma HMM can be used as a probabilistic NMF. We demonstrate the new NMF and compare it with state-of-the-art by applying it to speech signals.

2. GAMMA HMM AND ITS FORMULATION AS NMF

In this section, the gamma HMM is derived, and the proposed structure to handle the time varying level of the signal is explained. Then, it is shown how the proposed gamma HMM can be used to obtain an NMF representation of a nonnegative matrix. In the following, we represent random variables with capital letters, e.g., $\mathbf{X}_{K\times T} = [X_{kt}]$ denotes a matrix of random variables with elements X_{kt}. The corresponding realizations are shown as $\mathbf{x}$=$[x_{kt}]$. Also, we denote the t^{th} column of matrix $\mathbf{X}$ as $\mathbf{X}_t$ and the t^{th} column of matrix $\mathbf{x}$ is denoted as $\mathbf{x}_t$. Moreover, the conditional distribution $f_{X|Y}(x \mid y)$ is shown as $f(x \mid Y = y)$ or $f(x \mid y)$ for simplicity.

2.1. HMM with Gamma Distributions

The proposed HMM models the multidimensional nonnegative signal $\mathbf{X}$ with a limited number of hidden states. In the field of speech processing, where the short-time magnitude/power spectral vectors are commonly used as the input matrix for NMF, these hidden states can be identified, for example, with different speech sounds (phones). Let us assume that the hidden random variable S_t (at time t) of the HMM can take one of I available discrete values $i = 1, \dots I$. Because of the nonnegative nature of the signal, the output probability density functions of the HMM are modeled as gamma distributions. Hence, the conditional distribution of each element of $\mathbf{X}$ is given as:

$$f\left(x_{kt} \mid S_t = i, G_t = g_t\right) = \frac{x_{kt}^{a_{ki}-1}}{\left(g_t b_{ki}\right)^{a_{ki}} \Gamma\left(a_{ki}\right)} e^{\frac{-x_{kt}}{g_t b_{ki}}}, \quad (1)$$

where G_t is the short-term stochastic gain parameter, k is the dimension index, $\Gamma(\cdot)$ is the Gamma function, and a_{ki} and b_{ki} are state-dependent shape and scale parameters. Thus, the expected value and variance are obtained as $E(X_{kt} \mid S_t = i, G_t = g_t) = a_{ki} g_t b_{ki}$, and $\text{var}(X_{kt} \mid S_t = i, G_t = g_t) = a_{ki}(g_t b_{ki})^2$. In modeling the speech spectra, the choice of the gamma distribution is motivated by the super-Gaussianity of the speech DFT coefficients [8].

We assume that, given the hidden state S_t, different elements of the vector $\mathbf{X}_t$ are independent [8]. Hence, the HMM output density function is given as:

$$f\left(\mathbf{x}_t \mid S_t = i, G_t = g_t\right) = \prod_{k=1}^{K} f\left(x_{kt} \mid S_t = i, G_t = g_t\right). \quad (2)$$

G_t is allowed to take only nonnegative values, and is assumed to have a gamma distribution for the sake of tractability:

$$f\left(g_t\right) = \frac{g_t^{\phi-1}}{\theta_t^{\phi}\Gamma\left(\phi\right)} e^{-g_t/\theta_t}, \quad (3)$$

with ϕ and θ_t being the shape and scale parameters, respectively. In practice, the scale parameters b_{ki} are estimated to describe the signal statistics for different states while θ_t is meant to only model the long-term level changes of the signal. The sequence of hidden states is characterized by a first-order Markov chain, with initial state probability mass vector $\mathbf{p}$, with elements $p_i = f[S_{t=0} = i]$, and a transition probability matrix $\mathbf{q}$, with elements $q_{ij} = f[S_t = j \mid S_{t-1} = i]$. Fig. 1 illustrates this model.

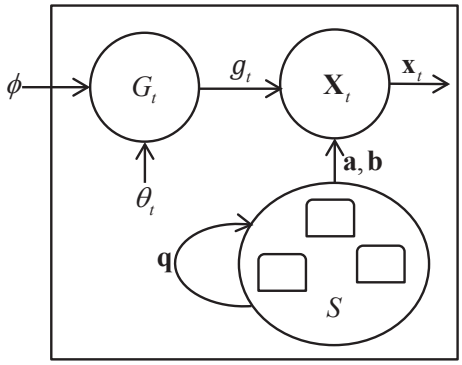

Fig. 1. A schematic representation of the HMM with gain modeling. The shape and scale parameters $\mathbf{a}$ and $\mathbf{b}$ are conditioned on the hidden state.

2.2. Gamma HMM as a Probabilistic NMF

The HMM described in Subsection 2.1 can alternatively be formulated as a probabilistic NMF. A usual approach in probabilistic NMFs [2–4] is to approximate an input matrix with an expected value that is obtained under the model assumptions. Then, the expected value is decomposed into a product of two nonnegative matrices. By doing so, a condensed representation of data can be achieved. Following this approach, we use an expected value of $\mathbf{X}_t$ to derive an NMF representation of the given vector $\mathbf{x}_t$. That is, $\mathbf{x}_t \approx \hat{\mathbf{x}}_t = \mathbf{v}\mathbf{w}_t$ where $\hat{\mathbf{x}}_t$ refers to the expected value. We calculate $\hat{\mathbf{x}}_t$ by considering the posterior distribution of the state and gain variables conditioned on the entire sequence of the observed signal over time, $\mathbf{x}$:

$$\hat{\mathbf{x}}_t = \sum_{i=1}^{I} \int E\left(\mathbf{X}_t \mid S_t = i, g_t\right) f\left(S_t = i, g_t \mid \mathbf{x}\right) dg_t. \quad (4)$$

Let us denote the hidden random states by a one-of-I indicator vector $\mathbf{S}_t$, i.e. $S_{it} = 1$ if the Markov chain at time t is in state i and $S_{jt} = 0$ for $j \neq i$. Accordingly, the realizations are referred to as $\mathbf{s}_t$.

We define the basis matrix $\mathbf{v}_{K\times I} = [v_{ki}]$ as $v_{ki} = a_{ki} b_{ki}$. Using this notation, we can write $\mathbf{v}_t = \mathbf{v}\mathbf{s}_t$. Therefore, recalling the paragraph following (1), we have $E[\mathbf{X}_t \mid \mathbf{s}_t, g_t] =$

$g_t \mathbf{v}\mathbf{s}_t$. Eq. (4) can now be written as:

$$\hat{\mathbf{x}}_t = \mathbf{v} \sum_{\mathbf{s}_t} \mathbf{s}_t f\left(\mathbf{s}_t \mid \mathbf{x}\right) \int g_t f\left(g_t \mid \mathbf{s}_t, \mathbf{x}\right) dg_t. \quad (5)$$

Denoting $\mathbf{w}_t = \sum_{\mathbf{s}_t} \mathbf{s}_t f(\mathbf{s}_t \mid \mathbf{x}) E(g_t \mid \mathbf{s}_t, \mathbf{x})$, we can write: $\hat{\mathbf{x}}_t = \mathbf{v}\mathbf{w}_t$. Thus, we have $\mathbf{x} \approx \mathbf{v}\mathbf{w}$, i.e., $\mathbf{x}$ is factorized into two nonnegative factors, a basis matrix $\mathbf{v}$ and an NMF coefficients matrix $\mathbf{w}$. In an extremely sparse case where $f(\mathbf{s}'_t \mid \mathbf{x}) = 1$ only for one state $\mathbf{s}'_t$, depending on time t, and all the other states have zero probability, we will have $\mathbf{w}_t = \mathbf{s}'_t E(g_t \mid \mathbf{s}'_t, \mathbf{x})$.

The conditional state probabilities $f(\mathbf{s}_t \mid \mathbf{x})$ can be calculated using the forward-backward algorithm [9]. To finish our NMF derivation, we need to evaluate $E(g_t \mid \mathbf{s}_t, \mathbf{x})$. Noting that g_t depends only on the observation at time t, the posterior distribution of the gain variable can be obtained by using the Bayes rule as:

$$f\left(g_t \mid \mathbf{s}_t, \mathbf{x}\right) = \frac{f\left(\mathbf{x}_t \mid g_t, \mathbf{s}_t\right) f\left(g_t\right)}{f\left(\mathbf{x}_t \mid \mathbf{s}_t\right)}. \quad (6)$$

Since the denominator of (6) is constant, using (2) and (3) we get:

$$\ln f\left(g_t \mid \mathbf{s}_t = \mathbf{1}_i, \mathbf{x}\right) \propto$$
$$-\frac{1}{\theta_t} g_t + \left(\phi - 1 - \sum_{k=1}^{K} a_{ki}\right) \ln g_t - \left(\sum_{k=1}^{K} \frac{x_{kt}}{b_{ki}}\right) \frac{1}{g_t}. \quad (7)$$

Eq. (7) corresponds to a generalized inverse Gaussian (GIG) distribution [10] with parameters $\vartheta = \phi - \sum_{k=1}^{K} a_{ki}$, $\rho = \theta_t^{-1}$, and $\tau = \sum_{k=1}^{K} b_{ki}^{-1} x_{kt}$. Hence,

$$\int g_t f\left(g_t \mid \mathbf{s}_t, \mathbf{x}\right) dg_t = E\left(g_t \mid \mathbf{s}_t, \mathbf{x}\right) = \frac{\mathcal{K}_{\vartheta+1}\left(2\sqrt{\rho\tau}\right)\sqrt{\tau}}{\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)\sqrt{\rho}},$$

where $\mathcal{K}_{\vartheta}(\cdot)$ denotes a modified Bessel function of the second kind.

3. PARAMETER ESTIMATION

We propose an EM algorithm to estimate the model parameters, denoted by $\lambda = \{\mathbf{p}, \mathbf{q}, \mathbf{a}, \mathbf{b}, \phi, \theta\}$. These parameters can be obtained online for a given signal $\mathbf{x}$. Alternatively, the time-invariant parameters $\mathbf{p}, \mathbf{q}, \mathbf{a}, \mathbf{b}$, and ϕ can be obtained given a training data set. To obtain the NMF representation of a new vector in this case, the only unknown parameter is θ, which should be estimated online.

In the E step of the EM, a lower bound is obtained on the log-likelihood of data, and in the M step, this lower bound is maximized. Let $\mathbf{Z}$ represent the hidden variables in the model. The EM lower bound takes the form

$$\mathcal{L}\left(f\left(\mathbf{z} \mid \mathbf{x}, \lambda\right), \hat{\lambda}\right) = Q\left(\hat{\lambda}, \lambda\right) + \text{const.}, \quad \text{where} \quad (8)$$

$$Q\left(\hat{\lambda}, \lambda\right) = \int f\left(\mathbf{z} \mid \mathbf{x}, \lambda\right) \ln\left(f\left(\mathbf{z}, \mathbf{x} \mid \hat{\lambda}\right)\right) d\mathbf{z}, \quad (9)$$

where λ includes the estimated parameters from the previous iteration of the EM, and $\hat{\lambda}$ contains the new estimates to be obtained.

For our problem, $\mathbf{Z} = \{\mathbf{S}, \mathbf{G}\}$ in which $\mathbf{S} = \{S_1, \ldots S_T\}$, and $\mathbf{G} = \{G_1, \ldots G_T\}$ where T is the number of data samples, i.e., the number of columns of $\mathbf{x}$. Now, $Q(\hat{\lambda}, \lambda)$ can be written as:

$$Q\left(\hat{\lambda}, \lambda\right) = \hat{Q}\left(\hat{\lambda}, \lambda\right) + \sum_{t,i} \omega_t(i) \int f\left(g_t \mid \mathbf{x}_t, S_t = i, \lambda\right)$$
$$\left(\ln f\left(g_t \mid \hat{\lambda}\right) + \ln f\left(\mathbf{x}_t \mid g_t, S_t = i, \hat{\lambda}\right)\right) dg_t. \quad (10)$$

Here, $\hat{Q}(\hat{\lambda}, \lambda)$ includes the terms for optimizing the Markov chain parameters $\mathbf{p}$ and $\mathbf{q}$, which is done similarly to [9]. The state probabilities $\omega_t(i) = f(S_t = i \mid \mathbf{x}, \lambda)$ are obtained by the forward-backward algorithm in which:

$$f\left(\mathbf{x}_t \mid S_t = i\right) = \int_0^{\infty} f\left(\mathbf{x}_t \mid S_t = i, G_t = g_t\right) f\left(g_t\right) dg_t. \quad (11)$$

Inserting (2) and (3) in (11), and using the definition of the GIG distribution [10] we get:

$$f\left(\mathbf{x}_t \mid S_t = i\right) = \frac{2\tau^{\vartheta/2}\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)}{\rho^{\vartheta/2}\theta_t^{\phi}\Gamma\left(\phi\right)} \prod_{k=1}^{K} \frac{x_{kt}^{a_{ki}-1}}{b_{ki}^{a_{ki}}\Gamma\left(a_{ki}\right)},$$

with $\rho = 1/\theta_t$, $\vartheta = \phi - \sum_{k=1}^{K} a_{ki}$, $\tau = \sum_{k=1}^{K} x_{kt} b_{ki}^{-1}$.

To obtain the new estimate of the rest of the parameters, (10) is differentiated w.r.t. parameters of interest, and the result is set to zero. Obtaining the gradient w.r.t. $\hat{b}_{ki}$ and setting it to zero yields the following estimate:

$$\hat{b}_{ki} = \frac{\sum_t \omega_t(i) x_{kt} E\left(G_t^{-1} \mid \mathbf{x}_t, S_t, \lambda\right)}{\hat{a}_{ki} \sum_t \omega_t(i)} \stackrel{\text{def}}{=} \frac{\mu_{ki}}{\hat{a}_{ki}}. \quad (12)$$

Inserting (12) into (10), and setting the gradient of the objective function w.r.t. $\hat{a}_{ki}$ to zero yields:

$$\varphi\left(\hat{a}_{ki}\right) - \ln\left(\hat{a}_{ki}\right) =$$
$$\frac{\sum_t \omega_t(i)\left(\ln x_{kt} - E\left(\ln G_t \mid \mathbf{x}_t, S_t, \lambda\right) - \ln \mu_{ki}\right)}{\sum_t \omega_t(i)}, \quad (13)$$

where $\varphi(u) = \frac{d}{du}\ln\Gamma(u)$ is the digamma function. Therefore, $\mathbf{a}$ is first estimated using (13), then (12) is solved to obtain $\hat{\mathbf{b}}$. Similarly, $\hat{\phi}$ and $\hat{\theta}$ are obtained by first estimating the shape parameter ϕ as:

$$\varphi\left(\hat{\phi}\right) - \ln\left(\hat{\phi}\right) = \frac{\sum_{t,i} \omega_t(i)\left(E\left(\ln G_t \mid \mathbf{x}_t, S_t, \lambda\right) - \ln \xi_t\right)}{\sum_{t,i} \omega_t(i)},$$

with ξ_t defined as:

$$\xi_t = \frac{\sum_i \omega_t(i) E\left(G_t \mid \mathbf{x}_t, S_t, \lambda\right)}{\sum_i \omega_t(i)},$$

and then using $\hat{\phi}$ to estimate θ_t:

$$\hat{\theta}_t = \xi_t / \hat{\phi}. \quad (14)$$

An alternative estimate for θ_t can be obtained for the situation where the long-term level of the signal remains constant

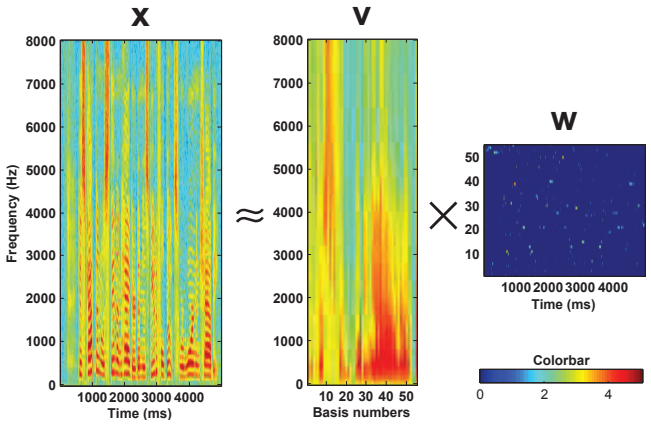

Fig. 2. Demonstration of the proposed NMF using speech signal. The figure shows the equation $\mathbf{x} \approx \mathbf{vw}$ for a speech sequence. On the left is the sequence of speech spectra and on the right are the nonnegative basis matrix $\mathbf{v}$ and the coefficient matrix $\mathbf{w}$. As is shown in the figure, $\mathbf{w}$ is a sparse matrix.

within a duration, say from sample 1 to T_1. Using this knowledge, we obtain a new estimate for θ, which is referred to as θ_{T_1} (in contrast to θ_t in (14)). The new estimate $\hat{\theta}_{T_1}$ is obtained as:

$$\begin{aligned} \xi_{T_1} &= \frac{\sum_{t=1}^{T_1} \sum_i \omega_t(i)\, E\left(G_t \mid \mathbf{x}_t, S_t, \lambda\right)}{\sum_{t=1}^{T_1} \sum_i \omega_t(i)}, \\ \hat{\theta}_{T_1} &= \xi_{T_1}/\hat{\phi}. \end{aligned} \tag{15}$$

Due to the concavity of the logarithm of the gamma density function in a_{ki} and b_{ki} around the stationary points $\mathbf{a}, \mathbf{b}$, these update rules are guaranteed to increase the overall log likelihood score of the parameters. As mentioned earlier, the posterior distribution of the gain variable is a GIG distribution for which $E(G^{-1})$ and $E(\ln G)$ are given in [10].

4. DEMONSTRATION

We applied our proposed probabilistic NMF to sequences of short-term speech spectra to demonstrate its capability. For this purpose, we estimated the stationary parameters of the model using 600 sentences from the training set of the TIMIT database with a sampling rate of 16 kHz.

The speech signal corresponding to each sentence of the database was segmented, windowed, and transformed into the frequency domain by applying the discrete Fourier transform (DFT). As the observation for NMF, the periodogram coefficients (magnitude-squared DFT coefficients) were used. The DFT was implemented using a frame length of 320 samples with 50% overlapped windows using a Hann window. The periodograms were stored as columns in a matrix of dimension $K \times T$ with $K = 161$ and T specified by the length of the sentence. Finally, the training signal was obtained by concatenating the spectra of all the sentences.

We learned 55 basis vectors ($I = 55$) for speech, i.e., $\mathbf{v}$ is a nonnegative matrix of dimension 161×55. Therefore,

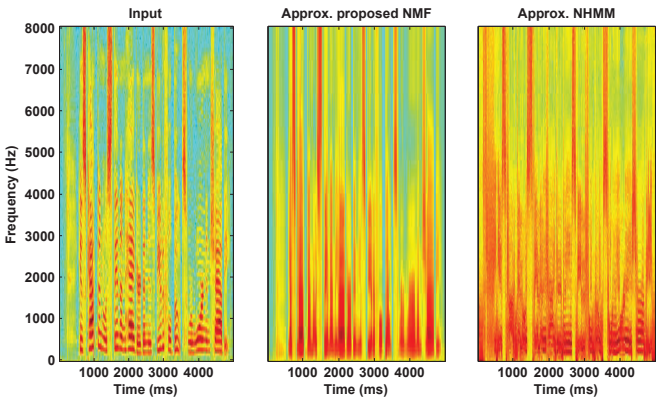

Fig. 3. NMF approximations for the same input as in Fig. 2. The approximations obtained using the proposed NMF and the NHMM approach [5] are shown in the middle and right plots, respectively. Although most of the harmonic structure is gone using the new method, the approximation is less distorted compared to NHMM.

each basis vector corresponds roughly to one phoneme and presents the expected spectrum of that phoneme.

The proposed gamma HMM was used to find an NMF approximation of the periodogram of speech sentences in a predictive manner. Periodograms are used here for the purpose of demonstration, but depending on the application, it might be preferable to apply the NMF on other types of smoothed spectral estimates. The time-varying scale parameter θ_t was estimated online using (15) with ϕ being fixed to the value obtained using the training data.

Fig. 2 shows the power spectrogram of a sample sentence "His captain was thin and haggard and his beautiful boots were worn and shabby" uttered by a female speaker, and its NMF representation. As was expected, the NMF coefficient matrix $\mathbf{w}$ is very sparse, and each basis vector is active for multiple subsequent time frames. The reason for the sparsity is that in a practical scenario, only one of the states is dominant in each time frame (a probability close to 1), which leads to a large coefficient for the corresponding basis vector in the NMF representation.

As Fig. 2 shows, the basis vectors represent a smooth estimate of the power spectra of different phonemes, and they do not have the fine structure in contrast to the detailed representation available in the original spectrogram. This indicates the potentiality of the proposed scheme in dealing with noise and undesired fluctuations when applied to a more realistic application such as speech recognition or enhancement.

The obtained NMF approximation for the same input as in Fig. 2 is shown in Fig. 3. The approximations obtained using the proposed NMF and the NHMM approach ([5], 60 states with 1 basis vector each) are shown in the middle and right plots, respectively. As can be seen in the figure, the new method leads to a cleaner and smoother representation. The NHMM representation, on the other hand, has introduced a lot of distortions but also has preserved the harmonic structure better.

To have a better understanding of the obtained factorization, the derived NMF representation was compared to two state-of-the-art approaches using objective measures. First, we considered two variants of NHMM: a sparse NHMM for which 60 states with 1 spectral component each was learned offline, and a non-sparse NHMM for which an HMM with 40 states with 10 basis vectors per state was learned. Second, we considered Kullback-Leibler divergence based NMF (KL-NMF) [2]. Aimed to compare the goodness of fit as a function of the sparsity level, the NMF coefficient vectors, $\mathbf{w}_t$, were constrained to have a specified l_0 norm (number of non-zero elements). For each value of l_0 in KL-NMF, the training was repeated to have the best possible basis matrix under the sparsity constraint.

The log-spectral distortion was evaluated between the input periodogram and the NMF estimate as the relative figure-of-merit. Moreover, to get a measure of the obtained correlation in the NMF coefficient vectors, the number of the consecutive vectors, $\mathbf{x}_t$, for which the same basis vector had the largest coefficient was computed, and it was normalized by the total number of columns in $\mathbf{x}$. In other words, if the system is in the state i in the current time frame, this measure (denoted as P_{rep}) gives the probability of staying in the same state in the next time frame.

The results averaged over 192 sentences from TIMIT core test set are shown in Table 1. The experiment shows that the proposed approach leads to a sparse representation in which a tradeoff between accurate fitting and temporal modeling is achieved. The accuracy of the approximation using the gamma HMM is comparable to that of KL-NMF with l_0 norm > 10. However, the gamma HMM gives a higher probability to continuously stay in the same state. The NHMM approaches provide the worst fit to the observed data, but the probability of staying in the same state is the highest using this method. By comparing sparse and non-sparse NHMM variants we see that the non-sparse NHMM provides both better temporal modeling and better fit as each state of the NHMM has a greater flexibility to model the observation.

In terms of the computational complexity, gamma HMM is substantially faster than the NHMM. When applied for the online factorization, the computational requirements of the gamma HMM is comparable with that of the KL-NMF, but for the offline training it is more demanding.

5. CONCLUSION

The present theoretical study was aimed at deriving a probabilistic NMF approach. To employ the temporal correlations of the signal, we developed an HMM with gamma output density functions (gamma HMM). Moreover, another gamma distribution was considered to govern the long-term level of the signal. We showed the analogy of the gamma HMM and NMF, and hence, derived a new probabilistic NMF. This work forms a basis for many applications, including speech and image processing. Currently, we are investigating the application of the method in speech enhancement, and will report the findings in a separate publication.

Table 1. Comparison between the proposed NMF and state-of-the-art methods. l_0 norm is the number of non-zero elements in each column of $\mathbf{w}$. The mean and standard deviation of the measures over 192 sentences are shown in the table.

Method	l_0 norm	log SD (dB)	P_{rep}
Proposed NMF	1	9.4± 0.7	0.59± 0.06
Sparse NHMM	1	16.4± 1.8	0.6± 0.05
Non-sparse NHMM	10	14.3 ± 1.9	0.66± 0.05
Sparse KL-NMF	1	12.6± 1.5	0.53± 0.07
Non-sparse KL-NMF	10	10.3± 1.1	0.53± 0.08
Non-sparse KL-NMF	55	7.8± 1	0.51± 0.05

References

[1] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. New York: John Wiley & Sons, 2009.

[2] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009, article ID 785152, 17 pages.

[3] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. European Signal Process. Conf. (EUSIPCO)*, vol. 47, 2009, pp. 1913–1917.

[4] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. Int. Conf. Machine Learning*, 2010, pp. 439–446.

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

[6] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2012, pp. 4561–4564.

[7] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, oct. 2009, pp. 121–124.

[8] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 13, no. 5, pp. 845–856, sep. 2005.

[9] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," U.C. Berkeley, Tech. Rep. ICSI-TR-97-021, 1997.

[10] T. Kawamura and K. Iwase, "Characterizations of the distributions of power inverse Gaussian and others based on the entropy maximization principle," *J. of The Japan Statistical Society*, vol. 33, no. 1, pp. 95–104, 2003.