# UNSUPERVISED SINGLE-CHANNEL SOURCE SEPARATION USING BAYESIAN NMF

*Onur Dikmen*,* *A. Taylan Cemgil*†

Boğaziçi University
Computer Engineering Department
Istanbul, Turkey
{onuro, taylan.cemgil}@boun.edu.tr

## ABSTRACT

We propose a prior structure for single-channel audio source separation using Non-Negative Matrix Factorisation. For the tonal and percussive signals, the model assigns different prior distributions to the corresponding parts of the template and excitation matrices. This partitioning enables not only more realistic modelling, but also a deterministic way to group the components into sources. This also prevents the possibility of not detecting/assigning a component and remove the need for a dataset and training. Our method only needs the number of components of each source to be set, but this does not play a crucial role in the performance. Very promising results can be obtained using the model with too few design decisions and moderate time complexity.

***Index Terms***— Non-negative Matrix Factorisation, Single-Channel Source Separation, Gamma Markov Chains, Gibbs Sampler, Metropolis-Hastings

## 1. INTRODUCTION

Non-negative matrix factorisation (NMF), proposed for decomposition of non-negative data [1], is a popular method for multivariate data analysis. The goal is to approximate a $W \times K$ non-negative matrix, $\mathbf{X}$, as the product of two non-negative matrices, $\mathbf{T}$ and $\mathbf{V}$, of sizes $W \times I$ and $I \times K$, respectively. This is done via minimising the dissimilarity between $\mathbf{X}$ and $\mathbf{TV}$

$$(\mathbf{T}^*, \mathbf{V}^*) = \arg\min_{\mathbf{T},\mathbf{V}} D(\mathbf{X}\|\mathbf{TV})$$

where the dissimilarity can be defined as the Kullback-Leibler divergence

$$D(\mathbf{A}\|\mathbf{B}) = -\sum_{\nu=1}^{W}\sum_{\tau=1}^{K}\left(A_{\nu,\tau}\log\frac{A_{\nu,\tau}}{B_{\nu,\tau}} + A_{\nu,\tau} - B_{\nu,\tau}\right) \quad (1)$$

KL divergence is always non-negative and is equal to zero when $\mathbf{X} = \mathbf{TV}$. The minimisation problem is effectively solved using variational bound optimisation in [1].

In audio processing, spectrogram decomposition via NMF is successfully applied to transcription [2] and single-channel audio source separation [3, 4]. Non-negative factorisation of the magnitude spectrogram of an audio signal provides a compact representation where the rows of $\mathbf{T}$ correspond to the frequency bins and the columns show dominant spectral structures of the spectrogram. These columns can be thought of as a codebook of spectra or basis vectors. The matrix $\mathbf{V}$ contains the excitations of these basis vectors along the time frames.

NMF that minimises the KL divergence in Equation 1 assumes the elements of $\mathbf{T}$ and $\mathbf{V}$ a priori independent. However, this does not reflect the physical properties of musical signals. Incorporating prior information, such as the harmonicity of tonal signals and their continuity in time or the short-duration behaviour of transients, into the model is important to obtain more realistic templates and better quality estimates. In [4], in addition to the KL divergence between $\mathbf{X}$ and $\mathbf{TV}$, the objective function also contains terms such that temporal continuity and sparseness of the excitation vectors are satisfied. In [5], the NMF model is defined in the Bayesian framework and the temporal continuity is incorporated through Gamma Markov chains (GMC) [6]. Such priors are shown to be more successful than the previous NMF methods [5].

In single-channel source separation with NMF, a general approach is to train the spectral templates from a dataset of audio classes [7, 8]. In this supervised approach, all of the learnt templates comprise a fixed $\mathbf{T}$ matrix and the excitation matrix, $\mathbf{V}$, is estimated. The training set can also be used to estimate the parameters of prior distributions of templates [5, 9], rather than learning and fixing the templates. Another issue with this approach is the assignment of the estimated components to the sources. This can be accomplished by clustering the components and assigning each cluster to a source. Assignment can be done by considering the similarity to the examples in the training corpus.

In this paper, we propose a Bayesian NMF model to separate tonal and percussive signals from a single-channel audio signal. The template and excitation matrices, $\mathbf{T}$ and $\mathbf{V}$, are divided into two partitions and assigned different prior distributions such that they encode a tonal and a percussive signal. The components obtained from the tonal partitions of the $\mathbf{T}$ and $\mathbf{V}$ matrices comprise the tonal signal, while the remaining ones constitute the percussive signal. The original contribution in this article is that all parameters and hyperparameters of the model are estimated during inference, so there is no need for an additional training step in order to learn the template vectors or their parameters. In this respect, the approach is fully unsupervised.

The remaining part of this paper is organised as follows: In Section 2, we explain our model and the details of the inference procedure. We will demonstrate the performance of our model on some single-channel audio source separation simulations in Section 3. A discussion on the model and possible future work will be presented in Section 4.

*Supported by the Scientific and Technological Research Council of Turkey (TUBITAK) and the State Planning Organization of Turkey (DPT), under Grants 107E050 and DPT 07K120610.

†Supported by Boğaziçi University Research Fund under the project number BAP 09A105P.

## 2. THE MODEL

The statistical interpretation of NMF can be derived by seeking a maximum likelihood solution to following model

$$s_{\nu,i,\tau} \sim \mathcal{PO}(s_{\nu,i,\tau}; t_{\nu,i} v_{i,\tau}) \quad (2)$$

$$x_{\nu,\tau} = \sum_{i}^{I} s_{\nu,i,\tau} \quad (3)$$

where $\mathbf{S}_i = \{s_{\nu,i,\tau}\}$ are latent sources and $\mathcal{PO}(\cdot)$ denotes the Poisson distribution. In the presence of these latent variables, the solution can be obtained using the EM algorithm. This approach leads to the same update rules as the original NMF minimising the information divergence between $\mathbf{X}$ and $\mathbf{TV}$ [10].

In order to obtain template and excitation matrices satisfying some properties, we can define prior distributions on $\mathbf{T}$ and $\mathbf{V}$, such as

$$\mathbf{T} \sim p(\mathbf{T}|\mathbf{\Theta}^t)$$
$$\mathbf{V} \sim p(\mathbf{V}|\mathbf{\Theta}^v)$$

where $\mathbf{\Theta}^t$ and $\mathbf{\Theta}^v$ are the (hyper)parameters of these distributions. Then, $\mathbf{T}$ and $\mathbf{V}$ can be estimated by the maximum a posteriori solution or Bayesian inference.

The topology of our model is designed to separate the underlying tonal and percussive sources from an audio signal. This is accomplished through assigning different prior structures to different parts of the template and excitation matrices, $\mathbf{T}$ and $\mathbf{V}$. Spectral templates of tonal signals have high values for the fundamental frequency and the harmonics of the notes that are being played. The other values are close to zero. These templates are excited for the duration that the notes are audible. However, a percussive hit excites a band of frequencies at the same time. These excitations are generally for short time intervals, except for the bass drum hits.

Our model makes use of GMC priors for columns of $\mathbf{T}$ or rows of $\mathbf{V}$ to enable continuity along those vectors and independent Gamma priors to have sparse values with occasional peaks. So, the tonal vectors of $\mathbf{T}$ are modelled with independent Gamma distributions for sparsity, whereas the vectors for percussions are modelled with GMCs. In contrast, excitation vectors for tonal components are modelled with GMCs to enforce continuity in time. Excitation vectors of percussive sources have independent Gamma priors which are suitable for short-time excitations. $\mathbf{T}$ and $\mathbf{V}$ matrices for one tonal and one percussive components is presented in Figure 1. The choice of Gamma and Gamma Markov chains as priors is mainly for the sake of simplicity. Gamma distribution is the conjugate prior for the Poisson observation model and this enables us to use faster and more convenient inference methods such as the Gibbs sampler or variational Bayes. In addition, we can incorporate the above mentioned requirements of tonal and percussive sources into the model using these prior distributions.

The density of a Gamma distributed random variable, $x \in \Re_+$, with shape and scale parameters, $a$ and $b$ is given by

$$\mathcal{G}(x; a, b) = \exp((a-1)\log x - x/b - \log\Gamma(a) - a\log b).$$

The mean of this distribution is $ab$ and the variance is $ab^2$. With small $ab$ and a larger $b$, the distribution will be sparse, i.e. mainly close to zero but with a heavy tail.

A Gamma Markov chain [6] is a prior structure for a chain of positive variables, where the correlation between consecutive variables is positive. In addition, each variable is conditionally conjugate, i.e. their prior and full conditional distributions are Gamma.

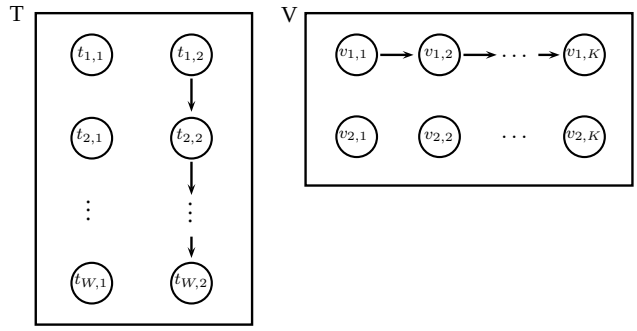



Figure 1: $\mathbf{T}$ and $\mathbf{V}$ matrices for one tonal and one percussive components.

In the Poisson observation model, this conjugacy is preserved. A GMC of $v_{1:K}$ can be defined as

$$v_1 \sim \mathcal{G}(v_1; a_v, b/a_v)$$
$$z_i|v_i \sim \mathcal{IG}(z_i; a_z, a_z v_i), \quad i = 1..K-1$$
$$v_{i+1}|z_i \sim \mathcal{G}(v_{i+1}; a_v, z_i/a_v), \quad i = 1..K-1$$

where $a_v$, $a_z$, $b$ are the hyperparameters of the chain and $z_{1:K-1}$ are auxiliary variables introduced to have positive correlation and conjugacy properties simultaneously. $a_v$ and $a_z$ are the coupling hyperparameters and they determine the degree of correlation between variables. Prior and full conditional distributions of $z_{1:K-1}$ are inverse Gamma and consecutive $z$ variables have positive correlation between them. This interpretation of GMCs is slightly different and more general from [6], but they are actually the same.

Denoting the number of tonal components with $I_{ton}$ and percussive components with $I_{perc} = I - I_{ton}$, the overall NMF model can be written as

$$t_{\nu,i} \sim \mathcal{G}(t_{\nu,i}; a_t^i, b_t^i/a_t^i), \quad i = 1..I_{ton}, \nu = 1..W$$
$$t_{1:W,i} \sim \text{GMC}(t_{1:K,i}; a_{tv}^i, a_{tz}^i, b_t^i), \quad i = I_{ton}+1..I$$
$$v_{i,1:K} \sim \text{GMC}(v_{i,1:K}; a_{vv}^i, a_{vz}^i, b_v^i), \quad i = 1..I_{ton}$$
$$v_{i,\tau} \sim \mathcal{G}(v_{i,\tau}; a_v^i, b_v^i/a_v^i), \quad i = I_{ton}+1..I, \tau = 1..K$$

The observation model is again given as in Equations 2 and 3.

Because of the conditional conjugacy, the full conditional distribution of each variable in the model is a standard distribution: Gamma for $t_{\nu,i}$ and $v_{i,\tau}$, multinomial for the latent sources $s_{\nu,i,\tau}$ and inverse Gamma for the auxiliary variables of the GMCs. This makes it feasible to use the Gibbs sampler or variational Bayes to infer about the variables.

The optimisation of the hyperparameters of the model can be performed using an EM algorithm which makes use of the posterior distribution estimated during the inference: samples drawn by the Gibbs sampler or the sufficient statistics estimated by variational Bayes. In this paper, we assume a uniform distribution for the hyperparameters and estimate them by sampling from their full conditional distributions using the Metropolis algorithm.

### 2.1. An Extension To The Model

As mentioned before, bass drums have a hybrid behaviour: they excite a band of frequencies as the other percussive sources but the duration is longer. This causes the bass drum and tonal instrument

components to get mixed. As a remedy, we added another partition of size $I_{bass}$ to the $\mathbf{T}$ and $\mathbf{V}$ matrices. A template vector in this partition has high values until a change point $\lambda_i$ and very low values afterwards.

$$t_{1:W,i} \sim \prod_{\nu=1}^{\lambda_i} \mathcal{G}(t_{\nu,i}; a_B^i, b_B^i/a_B^i) \prod_{\nu=\lambda_i+1}^{W} \mathcal{G}(t_{\nu,i}; a_b^i, b_b^i/a_b^i)$$

$$v_{i,1:K} \sim \text{GMC}(v_{i,1:K}; a_{vv}^i, a_{vz}^i, b_v^i), \quad i = I_{ton} + I_{perc} + 1..I$$

where $a_B^i$ and $b_B^i$ are selected such that the mean of distribution is high and variance low, in contrast, $a_b^i$ and $b_b^i$ ensure that the distribution is highly sparse. Here, $t_{\nu,i}$ and $v_{i,\tau}$ variables again have Gamma full conditional distributions. $\lambda_i$ is discrete and its full conditional distribution can be evaluated at each $W$ values it can get. So, this extended model can again be inferred using the Gibbs sampler. The pseudocode of the overall method is given in Appendix A.

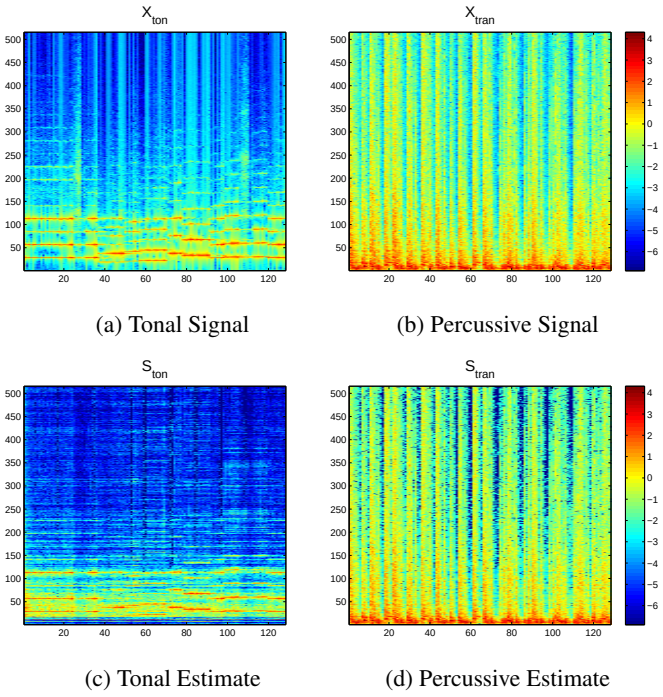
## 3. SIMULATION EXPERIMENTS

In our experiments, we used recordings of 6-10 seconds duration, sampled at 16kHz. Magnitude spectrograms are obtained using STFT, with non-overlapping windows of length 1024. Consequently, we work on spectrograms with 513 frequency bins and roughly 120-140 time frames. Phases of the original signal are stored and added to each estimated source before reconstruction.

The unsupervised NMF method infers the posterior distributions of the $\mathbf{T}$, $\mathbf{V}$ and $\mathbf{S}_i$, $i = 1 : I$ matrices using the Gibbs sampler. The hyperparameters of the model are also estimated during the inference, using the Metropolis algorithm with Gaussian proposal distributions. The only input to the model, apart from $\mathbf{X}$, are the number of components for each source: $I_{ton}$, $I_{perc}$ and $I_{bass}$. The model is based on Poisson observations and needs integer-valued $\mathbf{X}$ matrices. Magnitude spectrograms of audio signals have a large number of elements between zero and one. In order to decrease the effect of round-off error, we multiply the $\mathbf{X}$ matrix with a constant $C$ and round. Estimated components are divided to $C$ accordingly.

We made use of both manually mixed percussive and tonal signals and original recordings where we do not have the individual sources. First type of examples enables us to assess the performance using objective criteria such as signal to distortion ratio (SDR), signal to interference ratio (SIR) or signal to artefacts ratio (SAR) [11]. For the latter type, we judged on the performance perceptually. In Figure 2, we present the spectrograms of the separated sources from a mixed signal of flute and drums. On the top row, spectrograms of the original sources are given. Below them are the corresponding estimation obtained by our extended NMF model. In this experiment we used ten components for the tonal source ($I_{ton} = 10$), six components for the percussive sources ($I_{bass} = I_{perc} = 3$).

We compared the performances of our two models with other unsupervised models in Tables 1 and 2. GMC and GMRF are two models that couple the variances of time-frequency coefficients using Gamma Markov chains and random fields [6, 12]. They only make use of horizontal dependencies for tonal sources and vertical for the percussives. Apart from this, they do not need any kind of training or clustering to assign components to sources. The results show that our extended model (UNMF-e) performs better separation than the other models. GMRF results are also successful and




(a) Tonal Signal (b) Percussive Signal

(c) Tonal Estimate (d) Percussive Estimate

Figure 2: Sources estimated from a mixture of flute and drums recording.

has the highest SAR values in one of the experiments. According to the objective performance criteria, our simpler model (UNMF) performs very poorly. However, by listening to the reconstructed signals, we see that the problem mainly lies in assigning the bass drum to the wrong source.

| | $\hat{\mathbf{s}}_{ton}$ | | | $\hat{\mathbf{s}}_{tran}$ | | |
|---|---|---|---|---|---|---|
| | SDR | SIR | SAR | SDR | SIR | SAR |
| GMC | -4.23 | -2.42 | 4.82 | 1.34 | 13.13 | 1.85 |
| GMRF | -0.85 | 3.5 | 2.74 | 7.67 | 10.61 | 11.11 |
| UNMF | -5.02 | -4.40 | 9.44 | -1.58 | 13.67 | -1.26 |
| UNMF-e | -0.32 | 5.84 | 1.88 | 7.46 | 13.53 | 8.89 |

Table 1: Single channel source separation results on a mixture of guitar and drums.[1]

Original and reconstructed signals presented in this section can be listened to at `http://www.cmpe.boun.edu.tr/~dikmen/waspaa09/`.

## 4. DISCUSSION AND CONCLUSIONS

In this work, we proposed a model to separate percussive and tonal sources from single-channel audio signals via partitioning the spectrogram using NMF. The model makes use of some basic

[1] A mixture of 6-second excerpts from "Matte Kudasai" by King Crimson and "Territory" by Sepultura sampled at 16kHz.

[2] A mixture of 8-second excerpts from "Vandringar I Vilsenhet" by Änglagård and "Moby Dick" by Led Zeppelin sampled at 16kHz.

| | $\hat{\mathbf{S}}_{ton}$ | | | $\hat{\mathbf{S}}_{tran}$ | | |
|---|---|---|---|---|---|---|
| | SDR | SIR | SAR | SDR | SIR | SAR |
| GMC | -7.74 | -6.19 | 4.62 | -1.14 | 16.62 | -0.97 |
| GMRF | -4.27 | -1.61 | 3.0 | 5.59 | 19.82 | 5.8 |
| UNMF | -13.82 | -13.48 | 11.11 | -7.26 | -2.69 | -0.84 |
| UNMF-e | 6.03 | 15.50 | 6.67 | 15.72 | 24.15 | 16.41 |

Table 2: Single channel source separation results on a mixture of flute and drums.[2]

properties of the spectral behaviour of musical instruments. The separation process is totally unsupervised, i.e. there is no need to learn the template vectors from training data or manual assignment of each component to sources. The only parameters that should be set are the number of components each source will have. However, this is not a critical decision. Setting a higher number of components to a source than that is actually needed does not change the performance very much.

The inference of the parameters of the model is carried out using the Gibbs sampler. Good results can be obtained even using 100 MCMC steps. Since the hyperparameters are estimated using Metropolis algorithm, it is better to use more steps if the rejection rate is high. Our method does not include model selection, e.g. for the number of components. Estimation of marginal likelihood can be costly with the Gibbs sampler. Variational Bayes can be used for faster inference. In that case, model selection can be carried out using the variational lower bound of the marginal likelihood.

Our model is based on very basic properties of audio signals. Its separation performance will decrease in the presence of more complicated signals such as those generated by vibraphones, xylophones, etc. These signals contain both harmonic and transient structures. A possible strategy to tackle such problems can be introducing *ad hoc* template and excitation vectors for them along with indicator variables determining the existence of such a source.

## 5. REFERENCES


[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[3] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using nonnegtive matrix factorization and support vector machine," in *European Signal Processing Conference*, Istanbul, Turkey, 2005.

[4] 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.

[5] T. Virtanen, A. T. Cemgil, and S. J. Godsill, "Bayesian extensions to nonnegative matrix factorisation for audio signal modelling," in *Proc. of IEEE ICASSP 08*, Las Vegas, 2008.

[6] A. T. Cemgil and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources," in *ICA 2007, 7th International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 697–705.

[7] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *Proceedings of Interspeech 2008*, Brisbane, Australia, 2008.

[8] M. N. Schmidt and R. Olsson, "Single-channel speech separation using sparse non- negative matrix factorization," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, 2006.

[9] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *ICA 2009*, 2009.

[10] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," University of Cambridge, Tech. Rep. CUED/F-INFENG/TR.609, July 2008.

[11] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL Toolbox User Guide," IRISA, Rennes, France, Tech. Rep. 1706, 2005.

[12] O. Dikmen and A. T. Cemgil, "Gamma Markov random fields for audio source modelling," *IEEE Transactions on Audio, Speech and Language Processing*, Special Issue on Signal Models and Representation of Musical and Environmental Sounds, (to appear).


## A. THE PSEUDOCODE OF OUR METHOD

Below we give the pseudocode of our method. $\mathbf{\Theta}$ denotes the vector of all hyperparameters of the model, $\mathbf{T}_Z$ and $\mathbf{V}_Z$ represents the auxiliary variables of the GMCs in the template and excitation models. $\mathbf{S}_{ton}$ and $\mathbf{S}_{tran}$ are the estimated magnitude spectrograms of the tonal and percussive sources.

**Algorithm 1** UNMF-e ($\mathbf{X}$, $I_{ton}$, $I_{bass}$, $I_{perc}$, $N_{\text{samples}}$)

$I = I_{ton} + I_{bass} + I_{perc}$
Set $N_{\text{burn\_in}}$
Initialise $\mathbf{T}$, $\mathbf{T}_Z$, $\mathbf{V}$, $\mathbf{V}_Z$ and hyperparameters, $\mathbf{\Theta}$
**for** n = 1:$N_{\text{samples}}$ **do**
  Draw $\mathbf{S}^n$, $\mathbf{T}^n$, $\mathbf{T}_Z^n$, $\mathbf{V}^n$ and $\mathbf{V}_Z^n$ from full conditionals
  **for** each hyperparameter $\Theta$ **do**
    Propose $\Theta'$, calculate acceptance probability $a_\Theta$
    Accept $\Theta'$ with probability $a_\Theta$
  **end for**
**end for**
**for** i = 1:$I$ **do**
  $\hat{\mathbf{S}}_i = \sum_{i=N_{\text{burn\_in}}+1}^{N_{\text{samples}}} \mathbf{S}_i^n / (N_{\text{samples}} - N_{\text{burn\_in}})$
**end for**
$\mathbf{S}_{ton} = \sum_{i=1}^{I_{ton}} \hat{\mathbf{S}}_i$
$\mathbf{S}_{tran} = \sum_{i=I_{ton}+1}^{I} \hat{\mathbf{S}}_i$