SCORE GUIDED MUSICAL SOURCE SEPARATION USING GENERALIZED COUPLED TENSOR FACTORIZATION

Umut Şimşekli, A. Taylan Cemgil

Boğaziçi University
Dept. of Computer Engineering
34342, Bebek, Istanbul, Turkey

ABSTRACT

Providing prior knowledge about sources to guide source separation is known to be useful in many audio applications. In this paper we present two tensor factorization models for musical source separation where musical information is incorporated by using the Generalized Coupled Tensor Factorization (GCTF) framework. The approach is an extension of Non-negative Matrix Factorization where more than one matrix or tensor object is simultaneously factorized. The first model uses a temporally aligned transcription of the mixture and incorporates spectral knowledge via coupling. In contrast of using a temporally aligned transcription, the second model incorporates harmonic information by taking an approximate, incomplete, and not necessarily aligned transcription of the musical piece as input. We evaluate our models on piano and cello duets where the experiments show that instead of using a temporally aligned transcription, we can achieve competitive results by using only a partial and incomplete transcription.

Index Terms— Informed Source Separation, Coupled Tensor Factorization, Non-negative Matrix Factorization

1. INTRODUCTION

Audio source separation is one of the key problems in computer music and acoustic processing. The aim is to estimate individual sources from an audio mixture. This problem is called underdetermined if the number of channels is less than the number of sources in the mixture.

The first approaches to solve underdetermined source separation involved blind source separation techniques [1]. However, audio signals are highly complex, and blind methods fall short in exploiting useful domain specific knowledge. Incorporating domain specific information via signal models yields to informed source separation methods and there exists several studies that make use of different kinds of information. In the case where the original source signals are known beforehand, [2] presented a method which is based on extracting side information from the original sources and using this information at the source separation process. Another informed source separation method is proposed in [3] which makes use of a temporally aligned transcription of the audio mixture.

In this study, we present two models for musical source separation by using the Generalized Coupled Tensor Factorization (GCTF) framework [4]. Our first model makes use of a temporally aligned MIDI file and incorporates spectral information via coupling with isolated note recordings. The second model also incorporates spectral information, however instead of using a temporally aligned transcription, this model incorporates harmonic information by taking an approximate, incomplete, and not necessarily aligned transcription of the piece as input. Arguably, this approach is clearly more practical as often a complete score of a musical piece is simply not available.

The rest of the paper is organized as follows. We describe the GCTF framework and the inference algorithm in Section 2. We present our models for score guided source separation in Section 3. In Section 4, we provide the evaluation results of the models. Finally, Section 5 concludes this paper.

2. GENERALIZED TENSOR FACTORIZATION

The Generalized Coupled Tensor Factorization (GCTF) framework [4] is a direct generalization of the Probabilistic Latent Tensor Factorization (PLTF) framework to factorize simultaneously more than one multiway array (tensor or matrix). The PLTF model can be viewed as a natural extension of the matrix factorization [5]. The signal model of PLTF is

$$X(v_0) \approx \hat{X}(v_0) = \sum_{\bar{v}_0} \prod_{\alpha} Z_{\alpha}(v_{\alpha}), \qquad (1)$$

where $\alpha = 1, ...|\alpha|$ is the factor index. In this framework, the goal is computing an approximate factorization of a given tensor X in terms of a product of individual factors Z_{α}, some of which are possibly fixed. Here, we define V as the set of all

Funded by the scientific and technological research council of Turkey (TÜBİTAK) grant number 110E292, project Bayesian matrix and tensor factorizations (BAYTEN). Umut Şimşekli is also supported by a Ph.D. scholarship from TÜBİTAK.

indices in a model, V_0 as the set of visible indices, V_α as the set of indices in Z_α, and $\bar{V}_\alpha = V - V_\alpha$ as the set of all indices not in Z_α. We use small letters as v_α to refer to a particular setting of indices in V_α.

Since the product $\prod_\alpha Z_\alpha(v_\alpha)$ is collapsed over a set of indices, the factorization is latent. The optimization problem is to minimize the discrepancy between the observation X and the model output $\hat{X}$, as given by a *divergence* function $d(X, \hat{X})$. This divergence is a quasi-squared-distance and is typically taken as Euclidean (EUC), Kullback-Leibler (KL) or Itakura-Saito (IS). To illustrate our nonstandard notation, we define the nonnegative matrix factorization (NMF) model of [6] in the PLTF notation as follows:

$$X(f,t) \approx \hat{X}(f,t) = \sum_i D(f,i)E(i,t) \tag{2}$$

where $Z_1 \equiv D$, $Z_2 \equiv E$, and the index sets $V = \{f,t,i\}$, $V_0 = \{f,t\}$, $V_1 = \{f,i\}$, and $V_2 = \{i,t\}$. A detailed study on audio modeling via PLTF can be found in [7].

The Generalized Coupled Tensor Factorization (GCTF) model takes the PLTF model one step further where in this case we have multiple observed tensors X_ν that are factorized simultaneously:

$$X_\nu(v_{0,\nu}) \approx \hat{X}_\nu(v_{0,\nu}) = \sum_{\bar{v}_{0,\nu}} \prod_\alpha Z_\alpha(v_\alpha)^{R^{\nu,\alpha}} \tag{3}$$

where $\nu = 1, ...|\nu|$ and R is a *coupling matrix* that is defined as follows:

$$R^{\nu,\alpha} = \begin{cases} 1 & X_\nu \text{ and } Z_\alpha \text{ connected} \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

The coupling matrix $R^{\nu,\alpha}$ specifies factors α that effect the ν'th observed tensor. Note that, as opposed to the PLTF model, GCTF model contains multiple visible index sets $(V_{0,\nu})$. We can give the following example in order to illustrate the GCTF framework:

$$\hat{X}_1(i,j,k) = \sum_r Z_1(i,r)Z_2(j,r)Z_3(k,r) \tag{5}$$

$$\hat{X}_2(j,p) = \sum_r Z_2(j,r)Z_4(p,r) \tag{6}$$

$$\hat{X}_3(j,q) = \sum_r Z_2(j,r)Z_5(q,r) \tag{7}$$

Note that the factor Z_2 is shared among all the observations. Here, we have three observed tensors, therefore three simultaneous factorization problems. In this case, we have the following R matrix with $|\alpha| = 5$, $|\nu| = 3$

$$R = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} . \tag{8}$$

Table 1. Update rules for different p values. EU, KL, and IS correspond to Euclidean, Kullback-Leibler, and Itakura-Saito divergences, respectively.

p	Cost Function	Multiplicative Update Rule
0	EU	$Z_\alpha \leftarrow Z_\alpha \circ \frac{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ X_\nu)}{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ \hat{X}_\nu)}$
1	KL	$Z_\alpha \leftarrow Z_\alpha \circ \frac{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ \hat{X}_\nu^{-1} \circ X_\nu)}{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu)}$
2	IS	$Z_\alpha \leftarrow Z_\alpha \circ \frac{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ \hat{X}_\nu^{-2} \circ X_\nu)}{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ \hat{X}_\nu^{-1})}$

2.1. Estimation

Estimation of the latent factors Z_α can be achieved via iterative methods, by fixing all factors $Z_{\alpha'}$ for $\alpha' \neq \alpha$ but one Z_α and updating in an alternating fashion(see [4]). For nonnegative data and factors the update has a simple form

$$Z_\alpha \leftarrow Z_\alpha \circ \frac{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ \hat{X}_\nu^{-p} \circ X_\nu)}{\sum_\nu R^{\nu,\alpha}\Delta_{\alpha,\nu}(M_\nu \circ \hat{X}_\nu^{1-p})}. \tag{9}$$

where $\circ$ is the element-wise product (Hadamard product) and M_ν is a binary *mask* that specifies observed and missing elements: $M_\nu(v_{0,\nu}) = 1$ ($M_\nu(v_{0,\nu}) = 0$) if $X_\nu(v_{0,\nu})$ is observed (missing). In source separation this array is typically just one but the approach allows for missing data as well. The parameter p determines the cost function to be used: for $p = \{0,1,2\}$ correspond to the β-divergence [8] that unifies Euclidean, Kullback-Leibler, and Itakura-Saito cost functions, respectively. The key quantity in the above update equation is the $\Delta_{\alpha,\nu}$ function that is defined as follows:

$$\Delta_{\alpha,\nu}(A) = \left[\sum_{v_{0,\nu} \cap \bar{v}_\alpha} A(v_{0,\nu}) \sum_{\bar{v}_0 \cap \bar{v}_\alpha} \prod_{\alpha' \neq \alpha} Z_{\alpha'}(v_{\alpha'})^{R^{\nu,\alpha'}} \right] \tag{10}$$

For updating Z_α, we need to compute this function twice for arguments $A = M_\nu \circ \hat{X}_\nu^{-p} \circ X_\nu$ and $A = M_\nu \circ \hat{X}_\nu^{1-p}$. The individual cases are summarized in Table 1. While the definition looks complicated, a key observation is that the $\Delta_{\alpha,\nu}$ function is just computing a product of tensors and collapses this product over indices not appearing in Z_α, which is algebraically equivalent to computing a generalized matrix product.

3. SCORE GUIDED SOURCE SEPARATION

In this section, we present two tensor factorization models for musical source separation. Both models are based on NMF and the basic idea in our models follows the notion of decomposing the magnitude spectrum of the mixture (X_1) as

Table 2. Evaluation results of the proposed models. M1 denotes the first model and $M2_d$ denotes the second model where d is the duration of the transcription in seconds. The best results are shown in bold.

	SIR			SAR			SDR		
	$p=0$	$p=1$	$p=2$	$p=0$	$p=1$	$p=2$	$p=0$	$p=1$	$p=2$
M1	12.09	14.82	18.76	9.62	10.23	8.54	7.03	8.55	8.01
$M2_{45}$	7.85	20.08	**21.02**	6.31	**13.25**	6.72	2.06	**12.34**	6.42
$M2_{30}$	7.53	14.51	18.18	6.83	10.97	6.21	2.66	8.66	5.36
$M2_{10}$	6.39	11.17	14.91	8.12	9.35	6.25	2.57	5.37	4.82

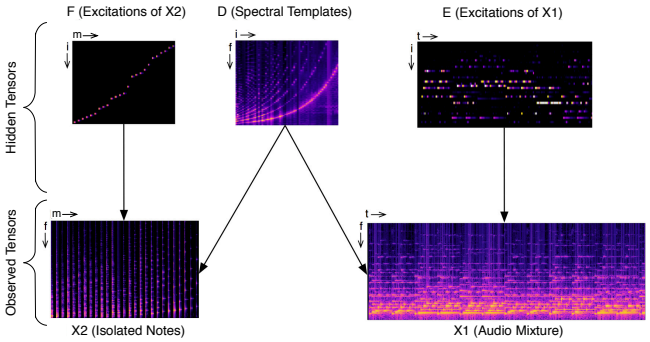

Fig. 1. General sketch of the first model. The idea is to incorporate information from the recordings of the instruments. The excitation matrix is also restricted by a temporally aligned transcription. The blocks visualize the tensors and the arrows denote the relation between them. The lower-case letters and small arrows near the blocks represent the indices of a particular tensor. Note that N and T matrices are masks that are applied on E and F, respectively.

the multiplication of a spectral dictionary (D) and the corresponding excitations (E) as firstly demonstrated in [9]. By this approach, the sources can be separated by Wiener filtering after the factors D and E are estimated.

3.1. Model I

In our first model, we combine two different NMF models that share the dictionary matrix D. The aim in this model is to incorporate spectral information by coupling the observed mixture with isolated note recordings. Here, the excitation matrix E is further restricted by a temporally aligned transcription of the mixture (N). The model is defined as follows:

$$\hat{X}_1(f,t) = \sum_i D(f,i)E(i,t)N(i,t) \tag{11}$$

$$\hat{X}_2(f,m) = \sum_i D(f,i)F(i,m)T(i,m) \tag{12}$$

where f is the frequency index, t and m are time frame indices, and i is the index of the spectral templates. Under Euclidean or KL divergences, X_1 is the magnitude spectrum of the audio mixture and X_2 is the magnitude spectrum of concatenation of isolated recordings corresponding to different notes. If IS divergence is chosen, both X_1 and X_2 are power spectra. Besides, N is the temporally aligned transcription of the mixture where $N(i,t) = 1(0)$ if the note i is played (not played) during the time frame t. Similarly, T is also a $0-1$ matrix, where $T(i,m) = 1(0)$ if the note i is played (not played) during the time frame m and F models the time varying amplitudes of the isolated notes. Figure 1 visualizes the general structure of the model. The coupling matrix R for this model is defined as follows:

$$R = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{13}$$

A similar model to this model was proposed in [10] for drum source separation in polyphonic music signals. In that model, the spectral templates are coupled between a polyphonic audio recording and a collection of drum recordings in order to obtain better drum separation performance.

3.2. Model II

In our second model, we hierarchically factorize the excitation matrix E as multiplication of a chord dictionary matrix B and its weights C as follows:

$$E(i,t) = \sum_k B(i,k)Z(i,k)C(k,t). \tag{14}$$

Here the basis matrix B encapsulates the harmonic structure of the music and incorporates additional information to the source separation system. The basic idea behind factorizing the excitation matrix E is to capture the repeated harmonic patterns in the music and form a harmonic basis for the musical piece.

After replacing E with the decomposed version, we get the following model:

$$\hat{X}_1(f,t) = \sum_{i,k} D(f,i)B(i,k)Z(i,k)C(k,t) \tag{15}$$

$$\hat{X}_2(f,m) = \sum_i D(f,i)F(i,m)T(i,m) \tag{16}$$

$$\hat{X}_3(i,n) = \sum_k B(i,k)Z(i,k)G(k,n)Y(k,n) \tag{17}$$

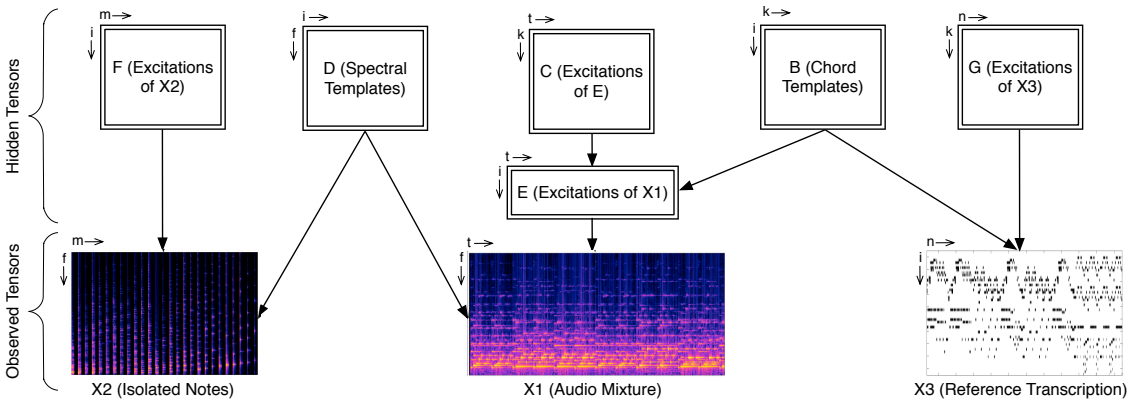

Fig. 2. General sketch of the second model. The idea is to incorporate spectral information from the recordings of the instruments and harmonic information from an approximate score which is not necessarily aligned. The blocks visualize the tensors and the arrows denote the relation between them. The lower-case letters and small arrows near the blocks represent the indices of a particular tensor.

where X_3 is a score matrix, which can be possibly obtained from a MIDI file: $X_3(i,n)$ is set to a constant value if the i^{th} note is active at time frame n. X_1 and X_3 do not necessarily belong to the same piece, however, in this study we select X_3 as a transcription of X_1.

Furthermore, Z and Y are $0-1$ matrices that allow the model to handle audio mixtures with multiple instruments. $Z(i,k) = 1$ if i^{th} note and k^{th} chord template belong to the same instrument. Similarly, $Y(k,n) = 1$ if the instrument that k^{th} chord template belongs to is active at time n. G models the time varying amplitudes of the chord templates. Figure 2 visualizes the general structure of the model. The coupling matrix R for this model is defined as follows:

$$R = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \quad (18)$$

Note that similar models to this model were proposed in order to solve audio restoration [4, 11] and polyphonic transcription problems [12] where encouraging results were obtained.

4. RESULTS

In order to evaluate our models, we have conducted several experiments. We have synthesized 3 piano and cello duets by using RWC Musical Instrument Sound database [13] and a simple concatenative synthesis algorithm and then we have selected 3 excerpts of 45 seconds from random parts of each piece yielding 9 test cases. In all our experiments the audio is subdivided into frames of 186 milliseconds where the audio spectrum is computed via Modified Discrete Cosine Transform (MDCT).

Since our second model needs only an approximate transcription, we also tested this model on different transcription durations: we used the first 10, 30, and 45 seconds of the transcriptions during the tests.

We have run the inference algorithms for $50-75$ iterations for both models and we have used 134 spectral templates that correspond to 88 piano and 46 cello notes. We have also used 80 chord templates for the second model; 50 templates for piano and 30 templates for cello. The factors B, C, D, E, F, and G are initialized randomly and updated during the estimation process. The other factors are clamped to their initial values.

In order to measure the performance of our models, we compute the signal to interference ratio (SIR), signal to artifact ratio (SAR), and signal to distortion ratio (SDR) by using the BSS_{EVAL} toolbox (v3.0) [14]. The evaluation results are given in Table 2. It can be observed that, despite the first model uses the temporally aligned transcription, the second model yields a similar performance and it even performs better than the first model for all metrics when KL divergence is chosen. We can also observe that increasing the duration of the transcription that is used in the second model improves the performance of the system which validates the idea behind the model. Some audio examples can be found in `http://www.cmpe.boun.edu.tr/~umut/eusipco2012/`.

5. CONCLUSION

In this study, two NMF-based models for musical source separation are presented. The first model uses a temporally aligned MIDI file and incorporates spectral information by using isolated note recordings. As opposed to the first model, the second model incorporates harmonic information by using an approximate and not necessarily aligned MIDI file. The GCTF framework enables these models to be defined in a compact notation. Besides, once the models are defined in this framework, the inference algorithm is readily available.

Experiments show that instead of using a temporally aligned transcription, we can achieve competitive and sometimes even better results by using an approximate transcription. This suggests that assuming a perfectly aligned score to the music is over-constraining the model and by relaxing this assumption we may get better separation results. We conclude by mentioning that the signal model can be enhanced by using convolutive structures and the computation time can be reduced by using parallel matrix computations. Both topics are subject to further research.

6. REFERENCES

[1] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[2] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937 – 1949, 2012.

[3] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *ICASSP*, Prague, Czech Republic, may 2011.

[4] Y. K. Yilmaz, A. T. Cemgil, and U. Simsekli, "Generalised coupled tensor factorisation," in *NIPS*, 2011.

[5] Y. K. Yilmaz and A. T. Cemgil, "Probabilistic latent tensor factorization," in *LVA/ICA*, 2010, pp. 346–353.

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

[7] A. T. Cemgil, U. Şimşekli, and Y. C. Subakan, "Probabilistic tensor factorization framework for audio modeling," in *WASPAA*, 2011.

[8] A. Cichoki, R. Zdunek, A.H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorization*, Wiley, 2009.

[9] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *WASPAA*, 2003, pp. 177–180.

[10] Jiho Yoo, Minje Kim, Kyeongok Kang, and Seungjin Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *ICASSP*, 2010, pp. 1942–1945.

[11] U. Simsekli, Y. K. Yilmaz, and A. T. Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *ICASSP*, 2012.

[12] U. Simsekli, Y. K. Yilmaz, and A. T. Cemgil, "Coupled tensor facorization models for polyphonic music transcription," in *10th IEEE Conference on Signal Processing and Communications Applications*, 2012.

[13] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "Rwc music database: Music genre database and musical instrument sound database," in *ISMIR 2003*, 2003, pp. 229–230.

[14] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.