SCORE GUIDED AUDIO RESTORATION VIA GENERALISED COUPLED TENSOR FACTORISATION

Umut Şimşekli, Y. Kenan Yılmaz, A. Taylan Cemgil

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

ABSTRACT

Generalised coupled tensor factorisation is a recently proposed algorithmic framework for simultaneously estimating tensor factorisation models where several observed tensors can share a set of latent factors. This paper proposes a model in this framework for coupled factorisation of piano spectrograms and piano roll representations to solve audio interpolation and restoration problem. The model incorporates temporal and harmonic information from an approximate musical score (not necessarily belonging to the played piece), and spectral information from isolated piano sounds. The performance of the proposed approach is evaluated on the restoration of classical music pieces where we get about 5dB SNR improvement when 50% of data frames are missing.

Index Terms— Audio Restoration, Coupled Tensor Factorisation

1. INTRODUCTION

Audio modelling based on factorisation has become popular along with the rapid development of computational power and statistical modelling techniques. This modelling paradigm has found place in various audio applications related to music information retrieval and content analysis, such as transcription or source separation.

Pioneering work on Nonnegative Matrix Factorisation (NMF) for audio processing [1] has demonstrated that, this modelling paradigm leads to practical and useful algorithms. For polyphonic transcription and source separation, which are the main applications of this model, various extensions and improvements have been proposed [2].

Apart from polyphonic transcription and source separation, audio restoration is another popular audio processing application where the aim is to interpolate/restore the missing parts in the audio. Many audio restoration methods have been proposed in the literature, to name a few [3, 4, 5]. The majority of these methods propose different models that are assumed to capture the underlying process of how the audio signals are generated. Impressive results have been reported in these studies, however, these methods have at least one of the two major problems: The first one is that, it is not straightforward to introduce domain specific information to these methods, i.e. the methods that are proposed in [3, 4] both require heavy computational needs. Upgrading these methods would slow down the estimation process while requiring more complex inference schemes. The second problem is, as the case in [5], some methods cannot restore the missing parts if entire frames of audio are missing.

In this paper, we present a model for piano spectrogram restoration by using the Generalised Coupled Tensor Factorisation (GCTF) framework [6]. The main idea of our model is to incorporate different kinds of musical information while estimating the missing parts of the audio: the reconstruction will be aided by an approximate musical score, not necessarily belonging to the played piece, and spectra of isolated piano sounds. A similar model was presented in [6] in order to illustrate the usage of the framework. In this study, we focus on this particular model in detail and investigate the capabilities and the limits of the model by simulating a challenging real-world application.

2. GENERALISED COUPLED TENSOR FACTORISATION

The Generalised Coupled Tensor Factorisation (GCTF) framework [6] is a generalisation of the Probabilistic Latent Tensor Factorisation (PLTF) framework [7] where the PLTF model is given as a natural extension of the NMF model:

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

where $\alpha = 1, ...|\alpha|$. In this framework, the goal is computing an approximate factorisation of a given a multiway array 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 indices

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.

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 factorisation is latent. The optimisation problem is the minimisation of $d(X, \hat{X})$, where d is a divergence (a quasi-squared-distance) typically taken as Euclidean (EUC), Kullback-Leibler (KL) or Itakura-Saito (IS). In order to illustrate the framework, we can define the NMF model of [8] 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 modelling via PLTF can be found in [9].

The Generalised Coupled Tensor Factorisation (GCTF) model takes the PLTF model one step further where in this case we have multiple observed tensors X_ν that are supposed to be factorised 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:

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

Note that, as distinct from the PLTF model, there are multiple visible index sets ($V_{0,\nu}$) in the GCTF model. In order to illustrate the GCTF framework, we can give the following example:

$$\hat{X}_1(i,j,k) = \sum_r A(i,r)B(j,r)C(k,r) \tag{5}$$

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

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

where we employ the symbols $A : E \equiv Z_{1:5}$. Here, we have three observed tensors, therefore three simultaneous factorisation problems. In this case, we have the following R matrix with $|\alpha| = 5$, $|\nu| = 3$

$$R = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{with} \quad \begin{aligned} \hat{X}_1 &= \textstyle\sum A^1 B^1 C^1 D^0 E^0 \\ \hat{X}_2 &= \textstyle\sum A^0 B^1 C^0 D^1 E^0 \\ \hat{X}_3 &= \textstyle\sum A^0 B^1 C^0 D^0 E^1 \end{aligned} . \tag{8}$$

Note that, the factor B is shared by all models.

Table 1. Update rules for different p values

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

2.1. Inference

The inference, i.e., estimation of the latent factors Z_α can be achieved via iterative optimisation (see [6]). For non-negative data and factors, one can obtain the following compact fixed point equation where each Z_α is updated in an alternating fashion fixing the other factors $Z_{\alpha'}$ for $\alpha' \neq \alpha$

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

where $\circ$ is the Hadamard product (element-wise product) and M_ν is a $0-1$ mask array where $M_\nu(v_{0,\nu}) = 1$ ($M_\nu(v_{0,\nu}) = 0$) if $X_\nu(v_{0,\nu})$ is observed (missing). Here p determines which cost function to be used, i.e. for $p = \{0,1,2\}$ correspond to the β-divergence [10] that unifies Euclidean, Kullback-Leibler, and Itakura-Saito cost functions, respectively. In this iteration, the key quantity 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 $A = M_\nu \circ \hat{X}_\nu^{1-p}$. As an example, it is easy to verify that the update equations for the KL-NMF problem (for $p = 1$) are obtained as a special case of Equation 3. Further cases are summarised in Table 1. A key observation is that the $\Delta_{\alpha,\nu}$ function is computing a product of tensors and collapses this product over indices not appearing in Z_α, which is algebraically equivalent to computing a marginal sum.

3. SCORE GUIDED AUDIO RESTORATION

In this section, by using the GCTF framework, we will form a model where we reconstruct missing parts of an audio spectrogram of a piano piece $X_1(f,t)$, that represents the short time Fourier transform coefficient magnitude at frequency bin f and time frame t. This is a difficult matrix completion problem since entire time frames (columns of X_1) can be missing,

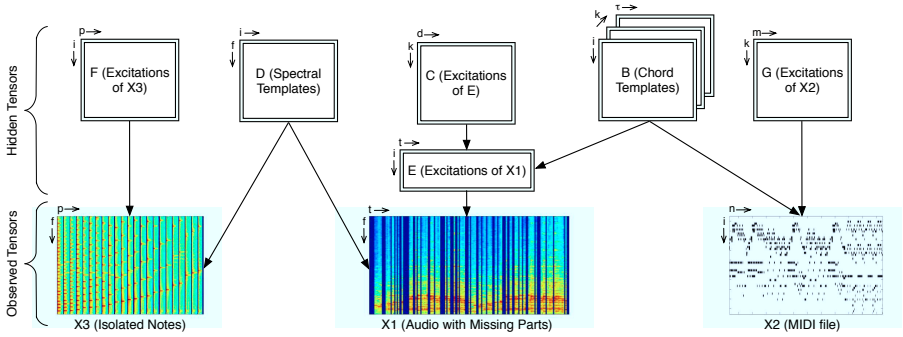


Fig. 1. General sketch of the proposed approach. The idea is to incorporate information from the recordings of the instrument and a score of the same genre. The blocks visualise the tensors that are defined in the model and the relation between them. The lower-case letters and arrows near the blocks represent the indices of a particular tensor.

low rank reconstruction techniques are likely to be ineffective. Besides, this kind of missing data patterns arise often in practice, e.g., when packets are dropped during digital communication.

It has been demonstrated that [1], when an audio spectrogram of music is decomposed using NMF as in Equation 2, the computed factors D and E tend to be semantically meaningful and correlate well with the intuitive notion of spectral templates (harmonic profiles of musical notes) and a musical score (reminiscent of a piano roll representation such as a MIDI file). However, as time frames are modelled conditionally independently, it is impossible to reconstruct audio with this model when entire time frames are missing.

In order to restore the missing parts in the audio, we form a model that incorporates musical information of chords structures and how they evolve in time. In order to achieve this, we hierarchically decompose the excitation matrix E as a convolution of some basis matrices and their weights and come up with a model for E which is similar to the model that is proposed in [11]: $E(i,t) = \sum_{k,\tau} B(i,\tau,k)C(k,t-\tau)$. Here the basis tensor B encapsulates both vertical and temporal information of the notes that are likely to be used in a musical piece; the musical piece to be reconstructed will share B, possibly played at different times or tempi as modelled by G. After replacing E with the decomposed version, we get the following model (Equation 11):

$$\begin{aligned}\hat{X}_1(f,t) &= \sum_{i,\tau,k} D(f,i)B(i,\tau,k)C(k,\overbrace{t-\tau}^{d}) \\ &= \sum_{i,\tau,k,d} D(f,i)B(i,\tau,k)C(k,d)Z(d,t,\tau) \quad (11)\end{aligned}$$

$$\begin{aligned}\hat{X}_2(i,n) &= \sum_{\tau,k} B(i,\tau,k)G(k,\overbrace{n-\tau}^{m}) \\ &= \sum_{\tau,k,m} B(i,\tau,k)G(k,m)Y(m,n,\tau) \quad (12)\end{aligned}$$

$$\hat{X}_3(f,p) = \sum_{i} D(f,i)F(i,p)T(i,p) \quad (13)$$

where X_2 is a score matrix, which can be possibly obtained from a MIDI file and X_3 contains the isolated piano recordings where it is constructed by concatenating isolated recordings corresponding to different notes. Here, we have introduced new dummy indices d and m, and new (fixed) factors $Z(d,t,\tau) = \delta(d-t+\tau)$ and $Y(m,n,\tau) = \delta(m-n+\tau)$ to express this model in our framework. Besides, T is a $0-1$ matrix, where $T(i,p) = 1(0)$ if the note i is played (not played) during the time frame p and F models the time varying amplitudes of the isolated notes. Figure 1 visualises the general structure of the model. The coupling matrix R for this model is defined as follows:

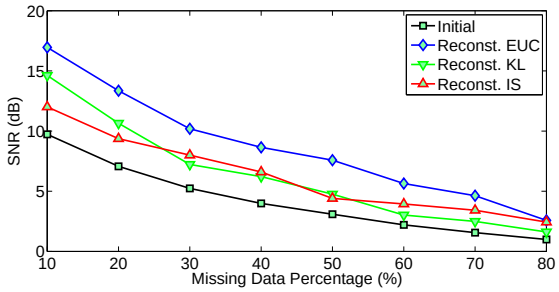
$$R = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad (14)$$

4. RESULTS

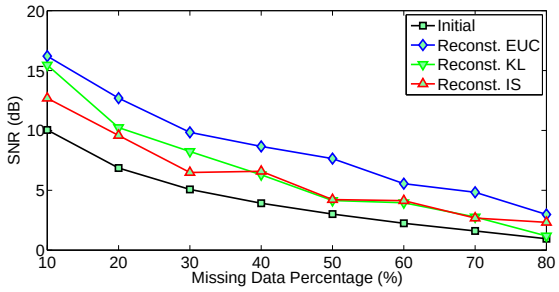
In order to evaluate our model, we have conducted several experiments. We have used the MIDI Aligned Piano Sounds (MAPS) piano database [12]: 16 bit 44.1 kHz piano samples are down-sampled to 11.025 Hz and the test files are corrupted by erasing big chunks of samples. In all our experiments the audio is subdivided into frames of 93 milliseconds.

In the experiments, we have used the first 20 seconds of 6 different recordings of 3 pieces from J. S. Bach. In 2 of these 6 different recordings, the piano samples (X_3) are available for each isolate note. The remaining 4 recordings are from different pianos. In order to obtain the restored version of the corrupted spectra we have simply combined the observed parts of X_1 and the estimated parts of $\hat{X}_1$: $M_1 \circ X_1 + (1 - M_1) \circ \hat{X}_1$, where M_1 is the $0-1$ mask that is introduced in Equation 9.

In our first experiment, after synthetically corrupting the test files, we have restored them by using their *own* transcrip-



(a) First experiment



(b) Second experiment

Fig. 2. Results of the experiments. As side information (X_2), we used a) own transcriptions of the test files, b) *different* transcriptions of other test files. Initial SNR is computed by substituting 0 as missing values.

tions as the side information. In the second experiment, we have used transcriptions of *different* pieces. Figure 2 illustrates the performance the model for different missing data percentages and different cost functions. For both cases the Euclidean cost function seems to perform better than the others. It can also be observed that, the results of both experiments are similar. One interpretation of this observation is that as long as the musical score (X_2) reflects the chord structure and its temporal evolution of corrupted the audio, it does not necessarily belong to the same piece as $X1$.

In order to assess the quality of our reconstructions, we measure the SNR between the true and the reconstructed spectrograms. In both cases, we get about 5 dB SNR improvement where 50% of the data is missing; gracefully degrading from 10% to 80% missing data. We believe that the results are encouraging as quite long portions of audio are missing.

5. CONCLUSION AND FUTURE WORK

In this study, a method for audio data restoration is presented. The restoration operation is aided by an approximate musical score and spectra of isolated piano sounds. The GCTF framework enables the model can be defined in a compact way and once the model is defined in this framework, making inference on the model becomes straightforward. The proposed model is evaluated on a challenging audio application, where big chunks of audio frames are missing.

A possible improvement for the model can be using convolutive models that can capture the temporal evolution of the spectral dictionary. This might come up with more realistic outputs due to better modelling of the frequency structure of the instrument.

6. REFERENCES

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

[2] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[3] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *EUSIPCO*, 2005.

[4] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *ICASSP*, 2005, pp. 517–520.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for time-frequency representations of audio signals," *JSPS*, vol. 10, pp. 1–10, 2010.

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

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

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

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

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

[11] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *ICA*, 2004, pp. 494–499.

[12] V. Emiya, R Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE TASLP*, vol. 18, no. 6, pp. 1643–1654, 2010.