A TIKHONOV REGULARIZATION METHOD FOR SPECTRUM DECOMPOSITION IN LOW LATENCY AUDIO SOURCE SEPARATION

*Ricard Marxer, Jordi Janer**

Music Technology Group, Universitat Pompeu Fabra,
Roc Boronat 138, Barcelona
ricard.marxer@upf.edu

ABSTRACT

We present the use of a Tikhonov regularization based method, as an alternative to the Non-negative Matrix Factorization (NMF) approach, for source separation in professional audio recordings. This method is a direct and computationally less expensive solution to the problem, which makes it interesting in low latency scenarios. The technique sacrifices the non-negativity constraint that characterizes NMF in exchange for a closed-form solution to the problem of spectrum factorization. We quantitatively evaluated it in terms of reconstruction and separation quality on a dataset of excerpts of professionally recorded songs with singing voice. Results show that the the proposed approach achieves similar quality to that of NMF.

Index Terms— Source separation, Harmonic analysis

1. INTRODUCTION

Spectrum decomposition has often been used in audio transcription and source separation tasks. It consists in modelling the spectral representation of a signal as a combination of a set of spectral components.

Some techniques such as Harmonic Temporal Clustering (HTC) [1] propose spectrum components with parameterized frequency and temporal envelopes and with a fixed harmonic structure. Similarly Wu et al. [2] consider components for the modelling of transients. In both cases the parameters are found using iterative Expectation Maximization update rules.

Non-negative Matrix Factorization (NMF) has received a lot of attention in the past few years. NMF was first introduced in the context of music transcription in [3]. The main strengths of such methods are the non-negativity constraints on the component gains, the ability to learn the components and its flexibility in adding additional cost terms. Raczyński et al. [4] use a harmonic initialization of the components and musically inspired penalties on the factorization. Durrieu et al. [5] propose an NMF method to decompose a signal using a source-filter model and then performing NMF on the residual. Ozerov et al. [6] present a source separation framework in which priors on the distributions of the spectral components can be introduced in a hierarchical way. In all cases the decomposition is performed iterating over a set of multiplicative rules.

Existing spectrum decomposition methods have proven useful in audio source separation tasks, however their iterative nature carries a high computational cost. Here we present an alternative method based on Tikhonov regularization that sacrifices the flexibility and the non-negativity constraints of NMF or the generality of other methods in exchange for a direct and rapid solution with a much lower computational cost.

*This research has been partially funded by Yamaha Corp. (Japan) for funding.

2. SIGNAL DECOMPOSITION MODEL

The main assumption of our spectrum decomposition method is that the short-term Fourier transform (STFT) of our audio signal, Y is a linear combination of N_C elementary spectra, also named basis components. This can be expressed as $Y = BG$ where $Y \in \mathbb{R}^{N_S \times 1}$ is the spectrum at a given frame m, N_S being the size of the spectrum. $B \in \mathbb{R}^{N_S \times N_C}$ is the matrix whose columns are the basis components, it is also referred to as the basis matrix. $G \in \mathbb{R}^{N_C \times 1}$ is a vector of component gains for the current frame.

Our focus is on low latency, unsupervised applications which require the decomposition of each spectrum frame to be done very quickly. Therefore, we will only consider solutions in which the basis components B are constant and fixed a priori.

It is obvious that the choice of the basis matrix has a large influence on the decomposition results. It is not in the scope of this article to study the effect of the basis matrix, but rather to propose a computationally cheap method to perform the decomposition given a suitable basis matrix.

As in many other NMF based [7, 5] approaches we set the basis matrix to be composed of a set of N_P single pitch multiple-harmonic spectra. However in order to model harmonic sources of different timbres we must allow different spectral envelopes. This is done by filtering the single pitch components with a filterbank of N_F filters. This results in a total of $N_P \cdot N_F$ harmonic basis components.

Modeling only harmonic sources is often not enough to explain all the possible observed spectra. In [2] the authors propose modelling wideband components to reconstruct transient sounds or background noise. We take a similar approach by adding to our basis matrix the spectra of the filters in our filterbank as wideband components. This results in a total of $N_C = (N_P + 1) \cdot N_F$.

The spectra components can be defined as:

$$\begin{aligned} \varphi[i,n] &= 2\pi f_l H N_P \frac{2^{\frac{iH-F/2+n}{HN_P}} - 1}{S_r \ln(2)} \\ E_i[\omega] &= \sum_{n=0}^{F} w[n] \left(\sum_{h=1}^{N_h} sin\left(h\varphi[i,n]\right) \right) e^{-j\omega n} \\ B_{i,k}[\omega] &= \begin{cases} U_k[\omega] E_i[\omega] & \text{if } i \leq N_P \\ U_k[\omega] & \text{if } i = N_P + 1 \end{cases} \end{aligned} \quad (1)$$

with $H = (1 - \alpha)F$. Where α is a coefficient to control the frequency overlap between the components, F is the frame size, S_r the

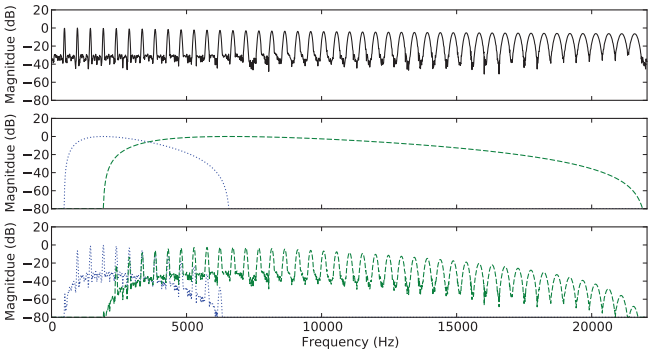

Fig. 1. Two components of our basis matrix B. Top shows $E_i[\omega]$ for a frequency of 480Hz. Middle shows $U_k[\omega]$ for two consecutive values of k. Bottom shows $B_{i,k}[\omega]$ for the selected $E_i[\omega]$ and $U_k[\omega]$.

sample rate, $w[n]$ is the analysis window, N_h is the number of harmonics of our components, $B_{i,k}$ is the spectrum of the component of i^{th} pitch filtered by k^{th} filter. U_k is the spectrum of the k^{th} filter in our filterbank. U_k is constructed as a sequence of N_F Hann windows, linearly distributed in the Mel scale and with a 50% overlap.

The column vectors $B_{i,k}$ are stacked horizontally to form the matrix B. This results in the spectrum $B_{i,k}$ of the component of i^{th} pitch and k^{th} filter being the column vector B_{iN_F+k}.

3. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) has been widely used in audio source separation tasks [5, 6, 8]. The NMF-based approach to solving our spectrum decomposition problem $Y = BG$ consists in finding the best non-negative estimate of the component gains $\hat{G}$ that minimizes a given objective function. We consider the following objective functions:

$$\Phi_{euc}(G) = \sum_{k=1}^{N_S} \frac{1}{2} \left([BG]_k - [Y]_k\right)^2 \quad (2)$$

$$\Phi_{kl}(G) = \sum_{k=1}^{N_S} [Y]_k log \frac{[Y]_k}{[BG]_k} - [Y]_k + [BG]_k \quad (3)$$

$$\Phi_{is}(G) = \sum_{k=1}^{N_S} \frac{[Y]_k}{[BG]_k} - log \frac{[Y]_k}{[BG]_k} - 1 \quad (4)$$

where $[X]_k$ is the k^{th} element of vector X. It is well known [8] that the solution to the non-negative factorization problem given these objective functions results in the following multiplicative update rule:

$$\hat{G}_n^{NMF} = \hat{G}_{n-1}^{NMF} \otimes \frac{B^t \left(\left(B\hat{G}_{n-1}^{NMF} \right)^{[\beta-2]} \otimes Y \right)}{B^t \left(B\hat{G}_{n-1}^{NMF} \right)^{[\beta-1]}} \quad (5)$$

where $\otimes$ is the Hadamard product (an elementwise multiplication of the matrices), all divisions are elementwise and $0 \leq \beta \leq 2$ is the coefficient that will define the objection function that is being minimized. $\beta = 2$ for the Euclidean distance (NMF_{euc}) $\beta = 1$ for the Kullback-Leibler divergence (NMF_{kl}) and $\beta = 0$ for the Itakura-Saito divergence (NMF_{is}). Finally n is the iteration of the solution and $\hat{G}_0^{NMF}$ is a random positive vector.

4. TIKHONOV REGULARIZATION

The condition number of the basis matrix B defined in Equation 1 is very high ($\kappa(B) \approx 5.9 \cdot 10^{17}$), therefore we may assume that our problem is ill posed. This could be due to the harmonic structure and correlation between the components in our basis matrix.

We propose using the Tikhonov regularization (TR) method [9] to find an estimate of the compontents gains vector $\hat{G}$ given the spectrum Y. This consists in the minimization of the following objective function:

$$\Phi_{TR}(G) = \sum_{k=1}^{N_S} \left([BG]_k - [Y]_k\right)^2 + \left([\Gamma G]_k\right)^2 \quad (6)$$

where Γ is the Tikhonov matrix that defines the preference among all possible solutions. In this study we set $\Gamma = \lambda W$ where $W \in \mathbb{R}^{N_C \times N_C}$ is a singular matrix that allows weighting the *a priori* probabilities of the solutions. λ is a positive scalar hyperparameter. This parameter controls the effect of the regularization on the estimated solution.

We have decided to give preference to solutions with a low norm while compensating for biases due to energy differences between components of different pitch. This is known as Weighted Minimum Norm Estimate (WMNE) and it can be acheived by defining W as a diagonal matrix such that:

$$diag(W)_{iN_F+k} = \sqrt{\sum_{\omega=1}^{N_S} \sum_{k=1}^{N_F} B_{i,k}^2[\omega]} \quad (7)$$

where $i = 1...N_P + 1$ and $k = 1...N_F$. The main reason for such a choice is that we assume that the basis components correspond quite well to the sources in the audio signal and only a few sources are simultaneously present in the audio. Therefore the gains of the components should have few high values and low small values, leading to a small norm.

The TR method, results in the following closed-form solution $\hat{G}^{TR} = RY$ where $\hat{G}^{TR}$ is the estimated components gains, and R is defined as:

$$R = (W^t W)^{-1} B^t [B(W^t W)^{-1} B^t + \lambda I_{N_S}]^+ \quad (8)$$

where $[X]^+$ denotes the MoorePenrose pseudoinverse of X. The calculation of R is computationally costly, however this operation is independent of the input spectra and can be performed before the analysis of the audio signal. The R matrix only depends on B and Γ. As we saw in section 2 the B only needs the parameters of the analysis process, therefore the only operation that is performed at each frame is $\hat{G}^{TR} = RY$.

Compared to the NMF method, the TR approach does not constrain the component gains to be non-negative. However, as we will show in the experiments, this assumption has little impact on the performance of the reconstruction and source separation tasks.

5. EVALUATION

The main goal of the article is to compare the TR closed-form solution and NMF solution in the general context of source separation. The comparison will be made on two main factors:

- How faithful is the factorization to the data?
- How well does the factorization separate the data?

In order to evaluate the factorization quantitatively, we simply compare the Signal to Noise Ratio (SNR) of the reconstruction without modifying the factors (components and gains). The reconstruction is computed as $\hat{Y} = B\hat{G}$. And the SNR calculation is performed in the frequency domain:

$$SNR = 10 \cdot \log_{10} \frac{\sum Y^2}{\sum \left|\hat{Y} - Y\right|^2} \quad (9)$$

To quantitatively evaluate how well the factorization separates the data, we perform a simple separation of the vocal track on a set of audio recordings. The separation produces two versions of the excerpt, one with only the voice track (foreground) and another with all but the voice track (background). We follow the same procedure as in [10] for the separation. We reconstruct the spectrum selecting the candidates in $\hat{G}$ that correspond to the voice. We have run two different tests: a supervised test in which the pitch of the vocal track is estimated in a previous stage using the well known Yin method on the vocal track in isolation, and an unsupervised test in which the pitch is estimated using $\hat{G}$:

$$i^{f_0} = \arg \max_{i=1\ldots N_P} \left(\sum_{k=1\ldots N_F} \max\left(\hat{G}_{i,k}, 0\right) \right) \quad (10)$$

where i^{f_0} is the index corresponding to the f_0 at a given frame. Due to correlations between pitches with harmonic relations, we also remove pitches that are at intervals Θ_f (Θ_i in pitch index units) from the predominant pitch.

$$i_{sel} = \{i^{f_0} + o | o \in \Theta_i\} \quad (11)$$

Since the voice often presents pitch fluctuations a series of adjacent basis components will also be selected. In our experiments, we select Δf semitones (Δi in pitch index units) around the selected pitches. This results in the following set of selected indices:

$$C_{sel} = \{(i \pm j)N_F + k | i \in i_{sel},\ j \leq \frac{\Delta i}{2},\ k \leq N_F\} \quad (12)$$

where $j \geq 0$ and $k \geq 1$. The estimate of the foreground and background spectra are computed using a binary mask $M \in \mathbb{R}^{N_C \times 1}$ on the component gains:

$$M_l = \begin{cases} 1 & \text{if } l \in C_{sel} \\ 0 & \text{else} \end{cases} \quad (13)$$

$$\hat{Y}_f = \gamma(M \otimes \hat{G})B \qquad \hat{Y}_b = ((1 - M) \otimes \hat{G})B \quad (14)$$

where $\gamma > 1$ is a gain on the foreground estimation. This is needed because part of the target source energy is actually spread in other pitch components that share harmonic relations, such as fifths and octaves.

Once we have the spectra estimates we calculate the actual foreground and background Discrete Fourier Transform (DFT) signals using Wiener filtering:

$$\hat{S}_f = \frac{\hat{Y}_f^2}{(\hat{Y}_f^2 + \hat{Y}_b^2)} S \qquad \hat{S}_b = \frac{\hat{Y}_b^2}{(\hat{Y}_f^2 + \hat{Y}_b^2)} S \quad (15)$$

where S is the original DFT of the mix signal. Note that eventhough the mask applied to the component gains is binary, the final mask applied to the DFT frames is actually a soft mask, resulting from the Wiener filtering. To go back to the time domain we apply a simple overlap and add technique. Finally we evaluate the performance of the separation computing the Signal to Distortion Ratio (SDR) with the popular audio source separation evaluation toolbox BSS_{EVAL} [11]. We compared each method to a baseline obtained with the oracle separation [12]. The values used in our experiments are the difference between the measure of each algorithm and the oracle estimation measure, averaged for all audio examples in the dataset. The evaluation material consists of a dataset of 11 multitrack recordings with vocals, compiled from publicly available resources (MASS[1], SiSEC[2], BSS Oracle[3]).

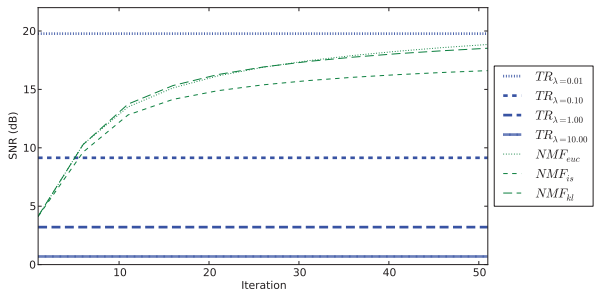

Fig. 2. Reconstruction SNR versus the factorization method and number of iterations.

6. RESULTS

The STFT analysis is performed with a 92ms Blackman-Harris window ($F = 4096$ for signals at sample rate $S_r = 44100$Hz), a hop size of 46ms ($H = 2048$) and a DFT size of 4096 which results in $N_S = 2049$. As in other pitch estimation techniques, we apply whitening to the spectrum to enhance the high harmonics by applying a compression factor of $\eta = 0.75$ so that $Y = |S|^\eta$. We also apply this process to the components spectra of matrix B. Regarding the parameters of the B matrix, we have set the number of filters $N_F = 12$, the lowest pitch frequency $f_l = 27.5$, the frequency overlap $\alpha = 0.5$, 60 pitches per octave covering a total of 6 octaves ($N_P = 60 \cdot 6 = 360$) and a maximum number of harmonics per component $N_h = 120$. This leads to a total number of components $N_C = 4332$. The factorization has been performed using the presented NMF solution (5) for the three objective functions in Eq. 2 and the proposed TR method with $\lambda = 10, 1, 0.1, 0.01$. Audio examples from our experiments are available online [4].

In Fig. 2 we observe the evolution of the SNR with relation to the number of iterations of the NMF approaches. On the same figure we plot the SNR of the TR methods. The results for the NMF behave as expected, constantly growing with the iteration count. The SNR results of the TR approaches demonstrates reconstruction equivalent to NMF-based methods depending on the value of λ. As expected, lower values of λ lead to better reconstruction results. Methods to find optimal λ values will be considered in future work. We have tested the following separation parameter values: source estimation gain $\gamma = 1, 2, 4, 8$, component gains mask width $\Delta f = 0.1, 0.2, 0.4, 0.8$ and intervals for the mask $\Theta_f = \{0\}, \{0, -12, 12\}$. For each factorization method the best

[1] http://www.mtg.upf.edu/static/mass
[2] http://sisec.wiki.irisa.fr/
[3] http://bass-db.gforge.inria.fr/bss_oracle/
[4] http://www.mtg.upf.edu/~rmarxer/papers/icassp12

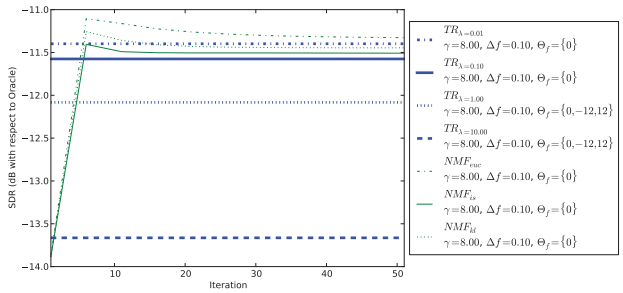

Fig. 3. Separation SDR for the background source (non vocals track) in the supervised test where the pitch is extracted from the vocal track in isolation.

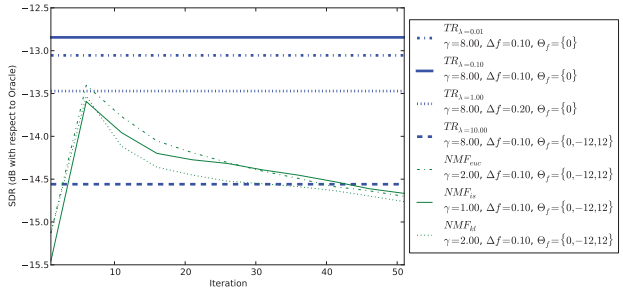

Fig. 4. Separation SDR for the background source (non vocals track) in the unsupervised test where the pitch has been estimated from $\hat{G}$.

parameter combination has been selected for the plots and comparaisons. In Figures 4 and 3 we show the results of our separation tests. As we can see the difference between TR and NMF methods is relatively small ($<$ 2dB). In the supervised scenario of Fig. 3 we can observe a slightly better performance of NMF with respect to TR. However in the case where the pitch is estimated from $\hat{G}$ the TR method performs better, this could be due to NMF finding and separating better other non-predominant pitches. We must keep in mind that the TR method is of much lower computational cost and is a closed-form solution that does not require iterations. This makes it much more attractive for low-latency and computation-limited contexts. Taking a closer look to the TR method we observe that in contrast to the SNR case, lower values of λ do not necessarily lead to better separation.

7. CONCLUSIONS

We present a new spectrum model and factorization method with applications in source separation. This method, based on a TR approach to the spectrum decomposition problem, offers a direct and closed-form solution at a significantly lower computational cost than NMF-based methods. We also present a comparative study between the TR approach and the NMF approach in the context of spectrum reconstruction and source separation. The study shows how TR can perform similarly to NMF with the proposed basis matrix.

In the current study the comparaison has been limited to one single basis matrix. In future work we should compare the TR method to NMF-based approaches using different basis matrices. Furthermore the flexibility of NMF should be taken into account when comparing the computational cost, for instance source-filter models for the basis matrix could lead to a significant lower number of components. NMF with sparsity constraints [4] should also be taken into account. Another direction for future research consists in exploring the choice of the Tikhonov matrix Γ. Finally other measures (SIR and SAR) should also be evaluated for a more complete comparaison.

8. REFERENCES

[1] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[2] J. Wu, E. Vincent, S. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Multipitch estimation by joint modeling of harmonic and transient sounds," in *IEEE Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 3, no. 3, pp. 177–180, 2003.

[4] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *in ISMIR 2007, 8th International Conference on Music Information Retrieval*, 2007, pp. 381–386.

[5] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, & Language Processing*, vol. 18, no. 3, pp. 564 –575, 2010.

[6] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *9th International Conference on Latent Variable Analysis & Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sep. 2010.

[7] 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.

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

[9] A. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," in *Soviet Math. Doklady*, vol. 4, 1963, pp. 1035–1038.

[10] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *IEEE Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, 2009, pp. 105–108.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[12] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.