AUTOMATIC MUSIC TRANSCRIPTION USING ROW WEIGHTED DECOMPOSITIONS

Ken O'Hanlon *Mark D. Plumbley*

Queen Mary University of London
Centre for Digital Music

ABSTRACT

Automatic Music Transcription (AMT) seeks to understand a musical piece in terms of note activities. Matrix decomposition methods are often used for AMT, seeking to decompose a spectrogram over a dictionary matrix of note-specific template vectors. The performance of these methods can suffer due to the large harmonic overlap found in tonal musical spectra. We propose a row weighting scheme that transforms each spectrogram frame and the dictionary, with the weighting determined by the effective correlations in the decomposition. Experiments show improved AMT performance.

Index Terms— Matrix decompositions, sparse representations, dictionary coherence, automatic music transcription

1. INTRODUCTION

Automatic Music Transcription (AMT) is a musical machine listening problem, in which a pitch-time representation of a musical piece is sought. AMT is often performed using matrix decomposition methods, which seek to decompose a spectrogram $\mathbf{S} \in \mathbb{R}^{M \times N}$ such that

$$\mathbf{S} \approx \mathbf{DT} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{M \times K}$ is a dictionary matrix with a spectral atom $\mathbf{d}_k$ in each column, and $\mathbf{T} \in \mathbb{R}^{K \times N}$ is a coefficient matrix containing the temporal activations of a corresponding atom in each row. When the dictionary atoms are pitch labelled, $\mathbf{T}$ admits a pitch-time representation.

The most common matrix decomposition methods used for AMT are based on Non-negative Matrix Factorisation (NMF), for which algorithms for the Euclidean distance and Kullback-Leibler (KL) divergence cost functions were originally proposed in [1]. NMF was first proposed for AMT in [2], and variations have been proposed for machine listening. Penalty terms for properties such as time persistence [3] [4] and alternative cost functions, such as the Itakuro-Saito (IS) divergence [5], and the generalised β-divergence [6] have led to improved AMT. While NMF was proposed as a dictionary learning method, the best AMT results are found when a fixed dictionary, trained on relevant signals, is used [7].

This research is supported by ESPRC Leadership Fellowship EP/G007144/1

Several problems may be seen when using matrix decompositions for AMT. Due to the harmonic nature of tonal music, each atom in the dictionary is highly overlapping with several other atoms. This can lead to harmonic jumping, where incorrect detections made in the AMT are harmonically related to notes that are known to be active, a problem common to all AMT methods. Using matrix decomposition methods, we have observed that harmonic jumping usually takes the form of truly recognised notes being accompanied by the false recognition of harmonically related "ghost" notes. Low energy elements may be present in the spectrogram e.g. at the tail of a sustained piano note, which may be temporally (and harmonically) coincident with higher energy elements. Typically for AMT using matrix decompositions, a hard thresholding is performed on the coefficient matrix $\mathbf{T}$, with pitch-time points above the threshold deemed to represent active notes. In selecting a threshold, it is hoped to keep low energy signal elements representing active notes whilst omitting as many harmonic "ghost" elements as possible.

In the sparse representations literature, recovery conditions of signals have been proposed based on dictionary coherence [8]. Hence methods have been proposed for preconditioning overcomplete dictionaries [9] [10] [11], by reducing coherence in order to improve recovery. While musical dictionaries, which are often not overcomplete, are seen to fail in terms of recovery conditions, even for two atoms, it is the presence of correlated atoms in the matrix decomposition that is the chief obstacle to AMT.

In the rest of this paper, we briefly introduce dictionary coherence and coherence-reducing methods used for sparse representations. Then, experiments are described showing that using transforms which lead to less coherent dictionaries can improve AMT performance. A method for reducing the effective coherence in a decomposition is proposed, followed by experimental results using this method. We then conclude, showing pointers for further work.

2. DICTIONARY COHERENCE

In the sparse representations literature, conditions on the exact recovery of noiseless signals are given in terms of coherence [8] and the Restricted Isometry Property (RIP) [12]. Coherence is defined as the maximum correlation between atoms

Published in: Proc IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), Vancouver, Canada, 26-31 May 2013. pp. 16-20.

in the dictionary. Assuming that $\|\mathbf{d}_k\| = 1 \; \forall k$ the coherence is defined as

$$\mu = \max_{i \neq j} |\mathbf{d}_i^T \mathbf{d}_j| \tag{2}$$

and a further measure, the cumulative coherence is defined by

$$\mu(k) = \max_{i} \max_{|\mathcal{J}|=k, i \notin \mathcal{J}} \sum_{j \in \mathcal{J}} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| \tag{3}$$

where k is the number of atoms considered in the cumulative coherence, and $\mu(k)$ is the maximum sum of any k off-diagonal elements in one column of the Gram matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$. By definition it is seen that $\mu(1) = \mu$ and $\mu(k) \leq k\mu$. Both the greedy Orthogonal Matching Pursuit algorithm (OMP) [13] and the optimisation based Basis Pursuit (BP) [14], are shown to recover a k-sparse signal representation exactly [8] in the noiseless case when

$$\mu(k) + \mu(k-1) < 1 \tag{4}$$

which is guaranteed when $k\mu < 0.5$.

With the dictionaries and signals of interest to this work, typically it is found that $\mu > 0.7$, as seen in Table 1, and the pessimistic recovery properties based on coherence would not give any assurances of recovery performance. However, some relationship between coherence based measures and transcription performance may be observed.

Several methods for countering the effects of dictionary coherence have been proposed, in particular for use with greedy methods, based on OMP [13]. For example, in [9] a sensing dictionary is derived for use with a modified OMP algorithm. Similarly in [11], the authors propose a data-adaptive sensing dictionary for use with the same modified OMP. Both of these works focus on dictionaries which are relatively incoherent such as unions of bases or Gaussian random matrices, and use quasi-orthogonalisation to perform decoherence. In [10] a dictionary learning method which seeks to learn incoherent dictionaries is shown to achieve improved sparse approximations in audio signals. A non-negative version of OMP (NN-OMP) is proposed in [15]. The authors note that problems with dictionary coherence are innate to non-negative dictionaries, and propose a pre-conditioning which centres the data and dictionary, reporting improved performance for sparse recovery. However, experiments with this method were not shown to enhance AMT performance, possibly due to the large harmonic overlap.

3. SOME AMT EXPERIMENTS

In [16] the Short-Time Fourier Transform (STFT) and the Equivalent Rectangular Bandwidth (ERB) transform were compared for the purpose of AMT, showing similar performance. In [16] an audio sampling rate of $22.05kHz$ was used and spectrograms using an STFT of dimension 1024 with 75% overlap and an ERB transform of dimension 250 interpolated onto a 23ms grid were used for AMT. We extend this exploration, using also an ERB of dimension 512 at the same sampling rate, as well as an STFT of dimension 2048 and an ERB of dimension 1024 both using the higher sampling rate of $44.1kHz$, and also using some matrix measures to compare the dictionaries learnt for each transform.

	$\kappa(\mathbf{D})$	$\|\mathbf{G}-\mathbf{I}\|_F$	μ	$\mathcal{F}$
STFT(1024)	42.87	19.82	0.8695	64.0
STFT (2048)	43.13	19.93	0.8693	64.1
ERB (250)	55.35	20.11	0.8723	66.7
ERB (512)	24.36	15.74	0.8619	68.4
ERB(1024)	17.30	13.85	0.8513	68.6

Table 1. Several matrix measures on dictionaries learnt from the same data for different transforms and AMT results in terms of F -measure

The AMT experiments were performed on the first $30s$ of 30 classical piano pieces, recorded live on a Disklavier piano, from the MAPS database [17]. The dictionaries are learnt offline using Euclidean NMF [1] with one atom learnt per note from signals containing isolated notes, recorded in the same environment as the recorded piano pieces. The same audio signals used for learning atoms across all transforms.

For all transforms, the decomposition was performed on a frame-wise basis using Non-Negative Least Squares (NNLS) [18]:

$$\mathbf{t} = \arg\min_{\mathbf{t}} \|\mathbf{s} - \mathbf{D}\mathbf{t}\|_2^2. \tag{5}$$

In particular, a fast variant of NNLS [19] was used, and it is noted that results are equivalent to using the Euclidean NMF coefficient update. Subsequent thresholding was performed with the threshold, η, adapted to the signal through use of the maximum value of the decomposition $\mathbf{T}$ and the parameter δ

$$\eta = \delta \times \max \mathbf{T}. \tag{6}$$

AMT performance was measured by comparing the resultant decompositions with the ground truths available in MAPS. True positives, tp, false positives, fp, and false negatives, fn, were labelled allowing the precision, $\mathcal{P} = \frac{\#tp}{\#tp+\#fp}$, recall, $\mathcal{R} = \frac{\#tp}{\#tp+\#fn}$ and $\mathcal{F}$-measure $\mathcal{F} = 2 \times \frac{\mathcal{P}\times\mathcal{R}}{\mathcal{P}+\mathcal{R}}$ metrics to be calculated. The metrics were calculated for various $\delta \in \{-15dB, ..., -40dB\}$, and the results are given for the optimal value of δ across all tracks.

Some matrix metrics were calculated for the various dictionary matrices. The matrix condition number, $\kappa(A) = \sigma_{max}(A)/\sigma_{min}(A)$ is the ratio of the largest and smallest singular values of the matrix. To give some measure of all dictionary correlations $\|\mathbf{G} - \mathbf{I}\|_F$ where $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ is used, where $\|\mathbf{X}\|_F$ is the Frobenius norm of $\mathbf{X}$. The coherence value μ (2) and the cumulative coherence $\mu(k)$ (3) for $k \in \{1, ...88\}$ are also calculated.

In Table 1 the AMT results in terms of $\mathcal{F}$-measure are given for each transform, alongside matrix measures of the

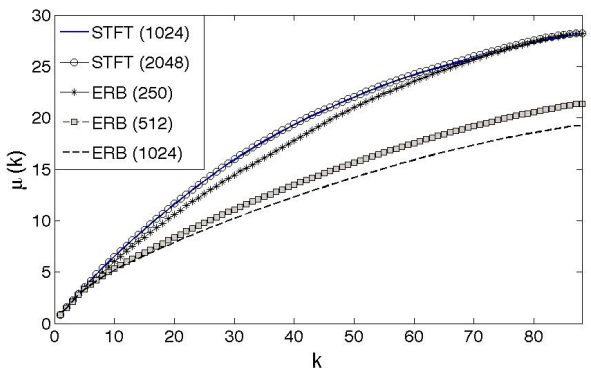

Fig. 1. Cumulative coherence $\mu(k)$ plotted against k for dictionaries learned from the same dataset in several transforms

corresponding dictionaries. In Figure 1 the cumulative coherence values for the dictionary in each transform are plotted. The two STFTs are almost indistinguishable, both in AMT and in dictionary matrix measures. The ERB(250) performs better than the STFT, displaying a lower cumulative coherence in Figure 1 whilst the tabulated measures are relatively worse. However the transforms which perform best, the ERB(512) and ERB(1024), are seen to have better values for all matrix measures relative to the other transforms, and further to this the ERB(1024) which shows the best AMT performance displays the optimum value for all matrix measures. This demonstrates that the AMT performance may be somewhat related to coherence measures

4. CONDITIONING A HARMONIC DICTIONARY

Row weighting is known to effect a least squares solution [20], and is a commonly used approach in methods such as Total Least Squares [21]. We propose a row-weighting scheme for AMT, using a diagonal weighting matrix $\mathbf{W}$ which seeks to reduce the effects of harmonic coherence on musical spectra decompositions. While it would be desirable to find a single weighting matrix which would enhance many decompositions, in practice this is found not to be viable. Instead, based on the NNLS solution $\mathbf{t}_n$ a different weighting matrix, $\mathbf{W}^n$ is derived at each time frame, which is then applied to both the dictionary and the signal giving the alternative approximation :

$$\mathbf{W}^n\mathbf{s}_n \approx \hat{\mathbf{D}}\hat{\mathbf{t}}_n \tag{7}$$

where $\hat{\mathbf{D}} = \mathbf{W}^n\mathbf{D}$. This transformation results in the Gram matrix $\Phi = \hat{\mathbf{D}}^T\hat{\mathbf{D}}$. If the columns of $\hat{\mathbf{D}}$ are normalised such that $\|\hat{\mathbf{d}}_i\|_2 = 1\ \forall i$, the Gram matrix is given by Θ, where

$$\Theta^{[2]} = \Phi^{[2]} \oslash [\mathbf{h}\mathbf{h}^T] \tag{8}$$

where $\oslash$ denotes elementwise division, $\mathbf{X}^{[.]}$ indicates elementwise exponentiation of $\mathbf{X}$ and $\mathbf{h} = \mathrm{diag}(\Phi)$.

In order to find a suitable value for $\mathbf{W}^n$, we propose an effective coherence measure :

$$\mu^e = \mathbf{t}^T[\Theta^{[2]} - \mathbf{I}_K]\mathbf{t} \tag{9}$$

The use of the normalised Gram matrix Θ is necessary in the calculation of μ^e as $\mathbf{t}$ is not updated. The weighting matrix, $\mathbf{W}^n$ is sought that minimises the *effective coherence measure* (9), whilst maintaining data integrity. This is performed using the projected gradient method, with bounds placed on the possible values of $w_m = \mathbf{W}_{m,m}$. Keeping $\mathbf{t}$ constant gives :

$$\frac{\partial\mu^e}{\partial w_m} = \sum_{i\neq j}\mathbf{t}_i\mathbf{t}_j\frac{\partial[\Theta^{[2]}]_{i,j}}{\partial w_m} \tag{10}$$

where $[\mathbf{X}]_{i,j} = x_{i,j}$ denotes the element in the ith row and jth column of X, and

$$\frac{\partial[\Theta^{[2]}]_{i,j}}{\partial w_m} = \frac{2w_m\Phi_{i,j}}{\|\hat{\mathbf{d}}_i\|_2^4\|\hat{\mathbf{d}}_j\|_2^4} \times \{2\|\hat{\mathbf{d}}_i\|_2^2\|\hat{\mathbf{d}}_j\|_2^2 d_{m,i}d_{m,j} - \Phi_{i,j}(\|\hat{\mathbf{d}}_i\|_2^2|d_{m,j}^2 + \|\hat{\mathbf{d}}_j\|_2^2 d_{m,i}^2)\} \tag{11}$$

or alternatively, in matrix form

$$\frac{\partial\Theta^{[2]}}{\partial w_m} = 2w_m\Phi\circ\mathbf{X}\circ[[2\mathbf{A}^m\circ\mathbf{A}^{mT}] - \Phi\circ[\mathbf{Z}^m + \mathbf{Z}^{mT}]] \tag{12}$$

where $\circ$ denotes the Hadamard elementwise multiplication, $\mathbf{X} = [\mathbf{h}^{[2]}\mathbf{h}^{[2]T}]^{[-1]}$ and $\mathbf{A}^m = \mathbf{d}^m\mathbf{h}^T$ and $\mathbf{Z}^m = \mathbf{d}^{m[2]}\mathbf{h}^T$, where $\mathbf{d}^m$ is the mth row of $\mathbf{D}$.

After $\mathbf{W}$ is estimated a solution to the approximation (7) is calculated, again using NNLS (5), giving $\hat{\mathbf{T}}$, the new coefficient matrix, from which the piano roll can be derived.

4.1. Relationship to other work

While decomposition methods are commonplace for the AMT problem, little work has considered the effects of the correlation or coherence of the dictionary, with much of the research emphasis considering structure in the matrix decomposition. The only work of which we are aware that considers dictionary correlation in AMT is that of [3] in which atom coefficients are penalised through multiplication with a coherence penalising matrix in dictionary learning based NMF experiments. Row weighting and scaling is more commonly used in methods such as Total Least Squares [20] [21], which also considers overdetermined problems. To summarise, to the best of our knowledge this is the first work to consider dictionary conditioning for AMT, and also to use coherence as a parameter for row weighting schemes. We note that a similar experimental setup to that used in [7] is used for the transcription experiments.

	NNLS	WNNLS	β-NMF	Wβ-NMF
ERB (250)	66.7	69.7	71.9	73.7
ERB (512)	68.4	72.6	74.9	77.0
ERB(1024)	68.6	73.7	75.2	77.9

Table 2. Transcription results comparing the weighted methods (WNNLS and Wβ-NMF) against the unweighted methods in terms of $\mathcal{F}$-measure

5. SOME FURTHER EXPERIMENTS

Further AMT experiments were run using the row weighting scheme. A similar setup to the earlier experiments was used, however the experiments were now limited to the ERB transforms. For the considered transforms, NNLS was used to decompose the spectrograms using the same dictionaries as in the previous set of experiments, and a subset of the atoms was selected by thresholding at each time frame of the decomposition with $\delta = 0.01$. This subset was used to calculate the weighting matrix $\mathbf{W}^n$. at the nth spectrogram frame, and $\mathbf{W}$ was bounded to have values in the interval $(0.4\ 1.6)$ giving an extremal weighting factor of 4. A NNLS decomposition was performed on the transformed signal $\mathbf{W}^n\mathbf{s}_n$ using the transformed dictionary $\hat{\mathbf{D}}$. Similar to the earlier experiments, results with an optimum relative threshold are reported.

Experiments were also run using the generalised β-NMF [6], which has been shown to enhance transcription performance [7]. Experiments were run using the same dataset and dictionaries as above. The value of $\beta = 0.5$ was used as this setting was seen to be the optimum value for AMT on a similar dataset [7]. Weighted experiments were also run using the β-NMF, using the weightings that were derived for the weighted NNLS experiments.

The results for these experiments are shown in Table 2, where it is seen that the weighted methods modestly outperform their unweighted counterparts. In particular the weighted NNLS experiments show improvements of up to 5.1%, with improvements more marked in the already better performing transforms. In the case of β-NMF, the improvements through weighting are seen to be relatively small, which may in part be due to using a weighting derived from an NNLS decomposition. It is noted that the results for β-NMF without weighting are seen to improve more significantly than NNLS as the dimension of the transform is increased. Overall it is seen that using a more suitable transform and the reweighting scheme results in an improvement of 7% and 6%, for the NNLS and the β-NMF respectively, which is significant in terms of AMT performance. An example data weighting is seen in Figure 2 where the weightings are seen to be often set to extremes, which is common. While significant downwards scaling of the signal can be seen in this example, this is not indicative, as upwards scaling is as likely. However the relative flattening effect seen in Figure 2 is a general phenomenon, as the correlation between two

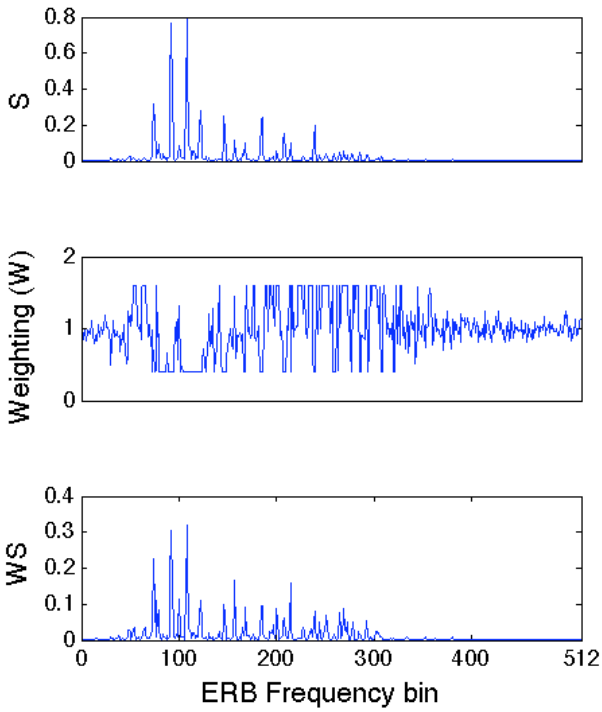

Fig. 2. Example data point $\mathbf{s}_n$ (top), weighting vector $\mathbf{w}^n$ (centre) and weighted datapoint $\mathbf{W}^n\mathbf{s}_n$ (bottom).

atoms is most significantly reduced when coincidentally large elements are scaled down. Closer inspection of the results of the weighted NNLS shows that while the AMT recall may be slightly improved, the precision is more significantly improved, as the number of false positives is reduced. This validates the approach taken, which sought to eliminate some of the false positives found in notes correlated to those that are active.

6. CONCLUSIONS AND FURTHER WORK

It has been shown that current AMT performance may be improved by considering the condition of the dictionary, which may be improved by using appropriate transforms. A row weighting scheme reducing an effective coherence metric was introduced, resulting in modestly improved AMT. However, further investigation may be worthwhile using different coherence metrics and experimenting with weighting bounds. The coefficient vector was fixed throughout the coherence reduction and it may be useful to update this. The reweighting scheme is currently computationally expensive relative to the decomposition methods, an issue we hope to address, possibly using multiplicative updates due to the non-negativity of the problem and warm restarts.

7. REFERENCES

[1] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13*, pp. 556–562, 2001.

[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] S. Raczynski, N. Ono, and S. Sagayama, "Extending non-negative matrix factorisation - a discussion in the context of multiple frequency estimation of musical signals," in *Proceedings of European Signal Processing Conference (EUSIPCO) 2009*, pp. 934–938.

[4] T. Virtanen, "Monaural sound source separation by non-negative matrix factorisation with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[6] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, 2006, pp. 32–39.

[7] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[8] Joel A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions in Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[9] K. Schnass and P. Vandergheynst, "Dictionary preconditioning for greedy algorithms," *IEEE Transactions in Signal Processing*, vol. 56, no. 5, pp. 1994–2002, May 2008.

[10] B. Mailhe, D. Barchiesi, and M. D. Plumbley, "INK-SVD: Learning incoherent dictionaries for sparse representations," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 3573–3576.

[11] A. Huang, G. Guan, Q. Wan, and A. Mehbodniya, "A reweighted algorithm for designing data dependent sensing dictionary," *International Journal of the Physical Sciences*, vol. 6, no. 3, pp. 386–390, February 2011.

[12] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, pp. 4203 – 4415, Dec. 2005.

[13] Y. C. Pati and R. Rezaiifar, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition.," in *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[15] A.M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of non-negative sparse solutions to underdetermined systems of equations," *IEEE Transactions on Information Theory*, vol. 54, pp. 4813–4820, 2008.

[16] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 109–112.

[17] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, pp. 1643–1654, Aug. 2010.

[18] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, 1987.

[19] R. Bro and S. De Jong, "A fast non-negativity-constrained least squares algorithm," *Journal of Chemometrics*, vol. 11, no. 5, pp. 393–401, 1997.

[20] G. H. Golub and C. F. van Loan, *Matrix Computations*, North Oxford Academic, 1983.

[21] I. Markovsky and S. Van Huffel, "Overview of total least-squares methods," *Signal processing*, vol. 87, no. 10, pp. 2283–2302, 2007.