POLYPHONIC PIANO TRANSCRIPTION USING NON-NEGATIVE MATRIX FACTORISATION WITH GROUP SPARSITY

Ken O'Hanlon and Mark D. Plumbley

Queen Mary University of London

ABSTRACT

Non-negative Matrix Factorisation (NMF) is a popular tool in musical signal processing. However, problems using this methodology in the context of Automatic Music Transcription (AMT) have been noted resulting in the proposal of supervised and constrained variants of NMF for this purpose. Group sparsity has previously been seen to be effective for AMT when used with stepwise methods. In this paper group sparsity is introduced to supervised NMF decompositions and a dictionary tuning approach to AMT is proposed based upon group sparse NMF using the β-divergence. Experimental results are given showing improved AMT results over the state-of-the-art NMF-based AMT system

Index Terms— Automatic music transcription, non-negative matrix factorisation, group sparsity

1. INTRODUCTION

Automatic Music Transcription (AMT) seeks a pitch-time representation from a musical signal. A popular family of methods for performing this task are based upon Non-negative Matrix Factorisation (NMF) [1], which seeks to approximate a non-negative matrix, $\mathbf{S} \in \mathbb{R}^{M \times N}$ such that

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

where $\mathbf{D} \in \mathbb{R}^{M \times K}$ is a dictionary matrix in which each column is referred to as an atom and $\mathbf{X} \in \mathbb{R}^{K \times N}$ is an activation matrix, in which each row relates the activity of the dictionary atom in the corresponding row. In the context of AMT it is hoped that each atom will represent a note leading to the activation matrix admitting a pitch-time representation.

NMF was first proposed for the purpose of AMT in [2], where the authors note the issue of separability, suggesting that each note may have to be played individually at least once in order that a relevant factor may be learned. Experiments in [3] demonstrate that NMF can learn relevant atoms in musical signals even when the notes are not necessarily isolated. However, performing AMT using unsupervised NMF is still

difficult. Meaningless atoms are often learnt and NMF is sensitive to the learning order, or rank of the factorisation.

Several alternative NMF-based approaches have been proposed in the AMT literature. Improved results are reported when a fixed dictionary is used, referred to as supervised NMF or Non-negative Matrix Decomposition (NMD) [4] [5]. However, the performance of such an approach deteriorates when the dictionary is not relevant to the signal to be decomposed [5]. Harmonically constrained variants of NMF have been proposed [6] [5]. A dictionary is created with one atom for each note, with zeroes in all dimensions other than those representing a harmonic partial of the fundamental frequency of the note in [6]. As NMF methods use multiplicative updates, the zero elements remain inactive. An alternative Harmonic-NMF approach is proposed in [5] using a hierarchical dictionary. Each note is represented by one high-level full spectrum atom, formed from a superposition of several narrowband harmonic atoms of similar pitch, with a semi-supervised NMF algorithm used to learn the superposition of narrowband atoms for each note.

Previously, we have explored the use of group sparsity for AMT using stepwise approaches, based on greedy pursuits [7], and backwards elimination [8], observing that the use of a group of atoms to represent a note affords improved AMT. However, these approaches used fixed dictionaries learnt off-line that may not be adaptable to other signals. The stepwise methods are ultimately constrained Non-Negative Least Squares decompositions and it is considered in NMF-based AMT that cost functions other than Euclidean distance afford better performance [5] [4] [2]. The work presented here incorporates group sparsity into NMD and NMF for the purpose of AMT, using the dictionaries of narrowband atoms proposed in [5]. While these atoms are used to construct a higher-level dictionary in [5], here the atoms are grouped, offering different superpositions at each frame of the spectrogram. Group sparse NMF was previously proposed in [9], however an alternative variant is proposed here, using the generalised β-divergence [10] and ℓ_2-norm based group coefficients. In the rest of this paper, the background on group sparsity, NMF and the harmonic atoms is briefly given, before the proposed approaches are outlined. Experiments and results are then described showing that the proposed approaches outperform the method in [5], before concluding with pointers to future work.

2. BACKGROUND

2.1. Harmonic atoms

The use of harmonic atoms based on a ERB scale is proposed in [5] with several fixed narrowband atoms used to construct one broadband atom. A group of six narrowband atoms used to represent one note is shown in Fig. 1. The details of the construction of the dictionary are omitted here, for which the reader is referred to the original paper [5]. The authors of [5] propose using a semi-supervised NMF approach, in which the individual narrowband atoms are fixed, while the higher-level broadband atoms are adapted by learning the superposition of similarly pitched narrowband atoms.
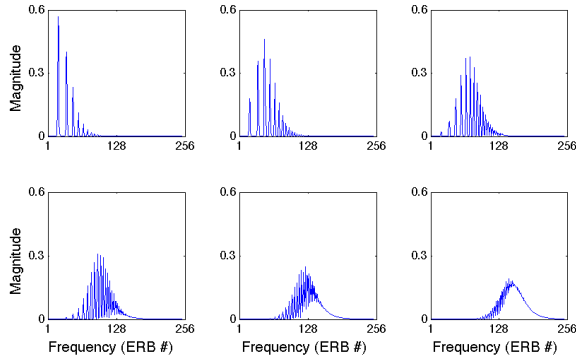

Fig. 1. Narrowband harmonic atoms for one note

2.2. NMF with β-divergence

NMF algorithms have been proposed for many different cost functions [10] [11]. The β-divergence [10] :

$$C_\beta(\mathbf{s}|\mathbf{z}) = \frac{1}{\beta(\beta-1)} \sum_m s_m^\beta + (\beta-1)z_m^\beta - \beta(s_m z_m^{\beta-1}) \quad (2)$$

where $\mathbf{z} = \mathbf{Dx}$, generalises several different cost functions popular in musical signal processing, such as the Euclidean distance ($\beta = 2$). The Kullback-Leibler (KL) divergence ($\beta = 1$) and Itakuro-Saito (IS) divergence ($\beta = 0$) [12] are limit cases of the β-divergence. The cost function (2) can be optimised using alternating multiplicative updates :

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left(\frac{[\mathbf{D}^T(\mathbf{S} \otimes [\mathbf{DX}]^{[\beta-2]})]}{[\mathbf{D}^T[\mathbf{DX}]^{[\beta-1]}]} \right)^{[\varphi(\beta)]} \quad (3)$$

$$\mathbf{D} \leftarrow \mathbf{D} \otimes \left(\frac{[\mathbf{S} \otimes [\mathbf{DX}]^{[\beta-2]}]\mathbf{X}^T}{[\mathbf{DX}]^{[\beta-1]}\mathbf{X}^T} \right)^{[\varphi(\beta)]} \quad (4)$$

where $\varphi(\beta)$ is a parameter that guarantees descent of the cost function [13] [14], division is elementwise, $\otimes$ denotes elementwise multiplication and $\mathbf{X}^{[p]}$ denotes elementwise exponentiation. Recently, it has been reported that superior AMT

results are achieved when $\beta = 0.5$ [5] [4] when a fixed dictionary is used.

2.3. Group sparsity

Group sparsity allows incorporation of prior information of expected co-activity of dictionary atoms into a decomposition. Given the set of tuples

$$\mathcal{L} = \{\mathcal{L}^l\} \quad (5)$$

where $\mathcal{L}^l$ contains the column indices of the lth group leads to the notation for the lth group of the dictionary, $\mathbf{D}[l]$, and of the coefficient vector, $\mathbf{x}[l]$:

$$
\begin{aligned}
\mathbf{D}[l] &= [\mathbf{d}_{\mathcal{L}^l(1)}, ..., \mathbf{d}_{\mathcal{L}^l(|\mathcal{L}^l|)}] \\
\mathbf{x}[l] &= [x_{\mathcal{L}^l(1)}, ..., x_{\mathcal{L}^l(|\mathcal{L}^l|)}]^T
\end{aligned} \quad (6)
$$

where $\mathcal{L}^l(i)$ is the ith member of the lth tuple in the set $\mathcal{L}$, and $\sum_l |\mathcal{L}^l| = N$. Typically, group sparsity is encouraged using a $\ell_{p,q}$ vector norm penalty term [15] [16] where

$$\|\mathbf{x}\|_{p,q} = \|\mathbf{g}\|_q \quad where \quad g_l = \|\mathbf{x}[l]\|_p. \quad (7)$$

The $\ell_{2,1}$ mixed norm is commonly used, favouring few active groups with several active atoms in each group [15]. Other norms, such as $\ell_{1,2}$ which favours many active groups, with few active atoms in each group, can be used [16].

A group sparse variant of NMF was previously proposed in [12] for the purpose of source separation using the Itakuro-Saito divergence with a log-based penalty applied to the ℓ_1 norm of the group coefficients, leading to a cost function

$$\mathcal{C}_{IS(G)}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_{IS}(\mathbf{s}|\mathbf{z}) + \lambda\Psi(\mathbf{x}) \quad (8)$$

where $\Psi(\mathbf{x}) = \sum_l \log(a + \|\mathbf{x}[l]\|_1)$, with an alternating multiplicative update strategy used and atoms constrained to have unit sum.

2.4. Sparse penalisation with an ℓ_p^p norm

The use of an ℓ_1-norm penalty is a popular choice for enhancing sparsity in a signal representation [17] and is also commonly applied as a penalty term in NMF [14]. However, other penalty terms can also be applied. In terms of a penalised Euclidean distance cost function the log based penalty is shown to be related to the reweighted least squares problem [17]. An alternative approach to the log penalty is the ℓ_p^p norm [18]:

$$\|\mathbf{x}\|_p^p = \|\mathbf{x}^{[p]}\|_1 \quad (9)$$

which is shown to be proportional to the log of a Gaussian prior [18]. Monotonic descent of NMF algorithms using the β-divergence with a $\|\mathbf{x}\|_p^p$ sparse penalty term is shown in [19].

3. METHOD

Before the proposed approaches are detailed their relationship to previous work is described. The use of group sparsity has previously been proposed by the authors for the purpose of AMT using stepwise NMD methods [7] [8] and shown to afford improved AMT. However, this previous work used dictionaries that were learnt offline, and it is intended to construct a method that is more adaptable to the signal, and affords the use of group sparsity. The harmonic dictionaries proposed in [5] readily afford a generic musical dictionary with a natural group structure and are adapted for the research presented here. The use of these dictionaries is central to this work and while the purpose here in performing AMT concurs with that of [5], a different approach to the problem is taken here, using group sparsity.

Group sparse NMF has previously been proposed [9] using the cost function (8), and the proposed approaches build upon this research. Modifications are made to the approach of [9]; the IS-divergence in (8) is replaced with $\mathcal{C}_\beta^{(0.5)}$ (2), for which superior AMT results have been recorded [5] [4], while an alternative group sparse penalty, based on the ℓ_p^p-norm approach of [18], is proposed with an ℓ_2-norm also preferred for the group coefficient. Using this modified group sparse cost function, two different approaches to AMT are proposed; the first a group sparse NMD, with a fixed harmonic dictionary; the second is a dictionary tuning approach based on group sparse NMF with modifications to the alternating multiplicative update philosophy and to the dictionary update.

3.1. Group sparse decompositions

In order to circumvent the extra parameter introduced when using the log-based group sparse penalty in (8), an alternative $\ell_{p,q}^q$-norm penalty is proposed for the same purpose

$$\|\mathbf{x}\|_{p,q}^q = \|\mathbf{g}^{[q]}\|_1 \quad where \quad g_l = \|\mathbf{x}[l]\|_p. \quad (10)$$

leading to a group sparse penalised β-divergence cost function

$$\mathcal{C}_{\beta_{GS}}(\mathbf{s}|\mathbf{z}) = \mathcal{C}_\beta(\mathbf{s}|\mathbf{z}) + \lambda\|\mathbf{g}^{[q]}\|_1 \quad (11)$$

which can be estimated using the augmented multiplicative update

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \left(\frac{[\mathbf{D}^T(\mathbf{S} \otimes [\mathbf{DX}]^{[\beta-2]})]}{[\mathbf{D}^T(\mathbf{DX})^{[\beta-1]}] + \lambda\Psi} \right)^{[\varphi(\beta)]} \quad (12)$$

where $k \in \mathcal{L}^l$, and λ is a parameter controlling the sparsity level, and $\varphi(\beta) = 1/(2 - \beta)$ as recommended for penalised approaches in [14] when $0 < \beta \le 1$.

$$\Psi_{k,n} = \frac{dg_{l,n}}{dx_{k,n}} = qg_{l,n}^{q-p}x_{k,n}^{p-1} \quad (13)$$

which reduces to $\frac{qx_{k,n}}{\|\mathbf{x}[l]\|_2^{2-q}}$ when $p = 2$. Monotonic descent of the cost function (11) using the update (12) is provable,

the details of which are left to [20]. This update is generally applicable for non-negative group sparse decompositions.

3.2. Dictionary tuning using group sparse NMF

In order to afford greater adaptation to the signal, a group sparse NMF-based approach is proposed which uses the harmonic dictionary as an initialisation, and in which the atoms maintain their initial pitch labels. The harmonic dictionary proposed in [5] does not set to zero the frequencies in dimensions that do not represent harmonic partials, as proposed in [6]. Therefore, the learning is unconstrained apart from the group sparsity in the activation matrix update (12) and is therefore prone to the problems associated with unsupervised NMF in the context of AMT [6] [5]. Several subtle modifications to the GS-NMF approach are proposed to avoid the problems associated with unsupervised NMF, a dictionary learning algorithm, and this modified approach is referred to as *dictionary tuning*, outlined in Algorithm 1.

The first feature of the dictionary tuning approach is that the dictionary updating does not start until a late stage in the NMF process, when the activation matrix is considered to be in the locality of its final decomposition. In practice, the unpenalised β-NMF activation matrix update (3) is performed until convergence. When the dictionary updating commences, the one-to-one alternating projection strategy typically used in NMF [1] [10] is modified with several iterations of the group sparse activation matrix update (12) performed for each single dictionary update. A modified dictionary update is also proposed :

$$\mathbf{D} \leftarrow \mathbf{D} \otimes \left(\frac{[(\mathbf{S} \otimes (\mathbf{DX})^{[\beta-2]})\mathbf{X}^T] + \epsilon}{[(\mathbf{DX})^{[\beta-1]}\mathbf{X}^T] + \epsilon} \right)^\eta \quad (14)$$

where $\epsilon = 0.1$ stabilises the update when a large gradient is present [10] and a small value of $\eta = 0.4$, rather than the optimal stepsize $\eta = 1$, is used to further reduce the step size

---

Algorithm 1 Dictionary tuning (GS-β-NMF)

Input
    $\mathbf{S}, \mathbf{D}, \lambda, \mathcal{L}, p, \beta$
Initialise
    $\mathbf{X} = \mathbf{D}^T\mathbf{S}$
Decompose using β-NMD (3)
Dictionary updating
    Repeat
        Update dictionary using (14)
        Update activation matrix (12) ($\times 3$)
    Until convergence
Output $\mathbf{X}, \mathbf{D}$

---

4. EXPERIMENTS

AMT experiments were run on the EnStDkcl subset of the MAPS database [21]. The first 30 seconds of each piece in the dataset were used. Each piece is a live electromechanical piano recording with a ground truth supplied. The two proposed approaches, GS-β-NMD and dictionary tuning with GS-β-NMF were compared. GS-β-NMD was performed used the harmonic dictionary [5] and the update (12) performed until convergence defined by the group sparse penalised cost function decreasing by less than 0.5% over 5 iterations. Parameter values of $\beta = q = \lambda = 0.5$ were used, which were seen in early experiments to provide reasonable performance. For the dictionary tuning approach a different value of $q = 0.01$ is used and the stopping condition is defined by either the group sparse penalised cost function (11) or the β-divergence (2) decreasing by less than 1% over 5 iterations of the dictionary updating iteration, that includes several iterations of the coefficient matrix update.

For comparison, the state-of-the-art H-NMF of [5] was also run, using code supplied online by the authors. A further comparison is performed by performing NMD using the β-divergence, with $\beta = 0.5$, using a dictionary learnt offline. This dictionary was constructed with one atom used to represent each note. The atoms were learnt from signals containing isolated notes, also contained in the EnStDkcl subset of the MAPS database. In this sense the atoms can be considered to optimal for the dataset used, having been learnt in the same environment.

The performance of the different approaches is analysed by thresholding the group activation matrices. Different thresholds are selected relative to the maximum value of each activation matrix

$$\mathcal{T} = \delta \times \max \mathbf{X} \qquad (15)$$

by using a range of values of $\delta \in \{15, ...40\}dB$. At each value of δ, the threshold $\mathcal{T}$ is calculated and applied to the activation matrix, leading to a binary representation. The sets of true positives, $\mathcal{TP}$, false positives, $\mathcal{FP}$, and false negatives, $\mathcal{FN}$ are labelled for each value of δ. For each value of δ the Precision, $\mathcal{P} = \frac{|\mathcal{TP}|}{|\mathcal{TP}|+|\mathcal{FP}|}$, Recall, $\mathcal{R} = \frac{|\mathcal{TP}|}{|\mathcal{TP}|+|\mathcal{FN}|}$ and $\mathcal{F}$-measure, $\mathcal{F} = \frac{2|\mathcal{TP}|}{2|\mathcal{TP}|+|\mathcal{FP}|+|\mathcal{FN}|}$ metrics are calculated for all pieces. The values recorded correspond to the maximum value of $\mathcal{F}$-measure, recorded at δ_{opt}.

The results for the different AMT approaches used are given in Table 1. Here it seen that the β-NMD using the dictionary learnt offline performs best, as may be expected. The proposed GS-β-NMD using the fixed dictionary performs almost as well as H-NMF of [5]. The dictionary tuning approach outperforms H-NMF by 3.5%, with $\mathcal{F}$-measure only 0.6% less than the β-NMD approach. A much higher value of δ_{opt} is seen for GS-β-NMF. It is noted that performance flattens below δ_{opt} for GS-β-NMF, unlike other approaches for which deterioration in $\mathcal{F}$-measure is observed beyond δ_{opt}.

	$\mathcal{P}$	$\mathcal{R}$	δ_{opt}(dB)	$\mathcal{F}$
H-NMF [5]	70.3	65.3	29	67.7
GS-β-NMD	70.1	65.1	28	67.5
GSβ-NMF	74.0	68.8	52	71.3
β-NMD	*74.5*	*69.4*	*33*	*71.9*

Table 1. Comparison of AMT performance of the proposed approaches, GS-β-NMD and GS-β-NMF with H-NMF [5]. Italics denote results from a β-NMD decomposition using atoms learnt from isolated notes.

5. CONCLUSIONS AND FUTURE WORK

The use of group sparsity for AMT in the context of NMD and NMF has been explored here, with a particular emphasis on the use of a generic dictionary. Results are promising. The NMD approach demonstrates results similar to those of the semi-supervised NMF method in [5], considered a state-of-the-art NMF approach for AMT, when using the same dictionaries. Further to this, the group sparse NMF approach improves on those results by over 3%.

However, the group sparse approach requires the use of a sparsity parameter λ, and some further work may be required to optimise its use. Some initial experiments indicate similar performance with λ set to half, or twice, the value employed here, in terms of $\mathcal{F}$-measure, with relatively small variations in Recall and Precision noted. It has been demonstrated that larger dimension ERB transforms lead to improved AMT performance [22], and a similar type of dictionary should be developed in these transforms. While the dictionaries used were seen to be fit for purpose, further consideration may also be given to the construction of different generic dictionaries.

6. REPRODUCIBLE RESEARCH

This research benefits from the efforts of other researchers to share their code [5] and dataset [21]. The open availability of these resources is commendable, allowing other researchers to easily and accurately compare methods. The code used in the experiments described in this paper is available at *http://code.soundsoftware.ac.uk/projects/gs_bnmf/*.

7. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS 14)*, Denver, 2000, pp. 556–562.

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

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, Honolulu, 2007, pp. 65–68.

[4] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, 2010, pp. 489–494.

[5] 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, March 2010.

[6] 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 the European Signal Processing Conference (EUSIPCO)*, Glasgow, 2009, pp. 934–938.

[7] K. O'Hanlon, H. Nagano, and M. D. Plumbley, "Structured sparsity for automatic music transcription," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, 2012, pp. 441–444.

[8] K. O'Hanlon, N. Keriven, and M. D. Plumbley, "Structured sparsity using backwards elimination for automatic music transcription," in *Proceedings of IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, 2013.

[9] A. Lefevre, F. Bach, and C. Fevotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[10] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended smart algorithms for non-negative matrix factorization," *Lecture notes in Artificial Intelligence, 8th International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, vol. 4029, pp. 548–562, 2006.

[11] A. Cichocki, S. Cruces, and S. Amari, "Generalized alpha-beta divergences and their application to robust non-negative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, January 2011.

[12] 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, March 2009.

[13] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative update algorithms for nonnegative matrix factorization with the β-divergence," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Kittila, 2010, pp. 283–288.

[14] C. Fevotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, September 2011.

[15] Y. C. Eldar, P. Kuppinger, and H. Bolsckei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[16] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, vol. 68, no. 1, pp. 49–67, February 2006.

[17] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1-minimisation," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, December 2008.

[18] M.A.T Figueiredo and R.D. Nowak, "A bound optimization approach to wavelet-based image deconvolution," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2005, vol. 2, pp. 782–785.

[19] K. O'Hanlon, *Automatic Music Transcription using Structure and Sparsity*, Ph.D. thesis, Queen Mary University of London, 2013.

[20] K. O'Hanlon and M. D. Plumbley, "Non-negative group sparse methods for automatic music transcription," in preparation.

[21] 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, no. 6, pp. 1643–1654, August 2010.

[22] K. O'Hanlon and M. D. Plumbley, "Row-weighted decompositions for automatic music transcription," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.