Oracle Analysis of Sparse Automatic Music Transcription

Ken O'Hanlon ⋆, Hidehisa Nagano ⋆†, and Mark D. Plumbley⋆ ⋆

⋆Queen Mary University of London
†NTT Communication Science Laboratories, NTT Corporation
{keno,nagano,Mark.Plumbley}@eecs.qmul.ac.uk

Abstract. We have previously proposed a structured sparse approach to piano transcription with promising results recorded on a challenging dataset. The approach taken was measured in terms of both frame-based and onset-based metrics. Close inspection of the results revealed problems in capturing frames displaying low-energy of a given note, for example in sustained notes. Further problems were also noticed in the onset detection, where for many notes seen to be active in the output transcription an onset was not detected. A brief description of the approach is given here, and further analysis of the system is given by considering an oracle transcription, derived from the ground truth piano roll and the given dictionary of spectral template atoms, which gives a clearer indication of the problems which need to be overcome in order to improve the proposed approach.

Keywords: Automatic Music Transcription, Sparse representations

1 Introduction

Automatic Music Transcription (AMT) is the attempt for machine understanding of musical pieces. Many methods proposed for AMT use atomic decompositions of a spectrogram with spectral basis atoms representing musical notes. The atoms may be learned online, using methods such as Non-negative Matrix Factorisation (NMF) [6] or sparse dictionary learning [8]. Alternatively a dictionary may be learnt offline, and the decomposition performed using methods like P-LCA [9] or sparse coding [4].

Often the output from AMT systems is displayed and understood through a piano roll, a pitch time representation relating the onsets and offsets of pitched note events. AMT performance is measured by comparing a computed piano roll with a given ground truth. Often the performance measures are frame-based, with true positives, false negatives and false positives denoted in the derived piano roll and several metrics have been proposed which use these annotations. An alternative perspective to measuring AMT performance is an event-based

⋆ This research is supported by ESPRC Leadership Fellowship EP/G007144/1 and EU FET-Open Project FP7-ICT-225913 "SMALL".

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

analysis [5]. Event-based metrics compare AMT performance in terms of the number of notes for which a correct onset is found within a time-based tolerance.

We have previously proposed an AMT system using structured sparse representations [7] which produced promising results for both frame- and event-based transcription. Visual inspection of the resultant energy-based piano rolls suggests that this approach performs well, capturing much of the energy in the signal, while some limitations are noticed. Often it is found that the energy in the early part of a note is captured, while later sustained elements may be missed, effecting the frame-based analysis. Errors are also noted in the event-based analysis, for which a simple threshold-based onset detection system was used.

These observations lead us to perform an oracle analysis of the system, in order to investigate the causes of these errors, which could possibly reside in either the dictionary used, the transcription system or in the onset detection system. As the system is ultimately based on a (non-negative) least squares analysis, an oracle transcription can be derived by decomposing the signal at each point in time using non-negative least squares (NNLS) with only the atoms representing the notes active, as given by the ground truth, at that time. In the rest of this paper, we describe briefly the AMT system used and the oracle transcription, before analysing the results given by the oracle transcription.

2 Transcription Using Structured Sparse Representations

Sparse representations seek to form the approximation $\mathbf{s} \approx \mathbf{D}\mathbf{t}$ where $\mathbf{s}$ is a signal vector, $\mathbf{D}$ is a dictionary of atoms, and $\mathbf{t}$ is a coefficient vector which is sparse, having few non-zero coefficients. Algorithms for solving sparse representation problems include Orthogonal Matching Pursuit (OMP) [11] which selects, iteratively, the atom most correlated with the residual error and adds this to the support, or collection of selected atoms. At each iteration the supported atoms are backprojected onto the initial signal, giving interim coefficients and a new residual error. Another approach to sparse approximation is the Basis Pursuit (BP) [12], for which many algorithms can be used to solve the optimisation

$$\min_t \|\mathbf{s} - \mathbf{D}\mathbf{t}\|_2^2 + \lambda\|\mathbf{t}\|_1 \tag{1}$$

where the second term is a penalisation term which promotes sparsity.

Music transcription can be thought of as an inherently sparse problem, as only a few notes are active at a given time. In this work non-negative sparse representations are required to decompose the magnitude spectogram. In group or block sparse representations, it is assumed that certain atoms tend to be active together. This assumption can be leveraged for transcription purposes, as in the previous work [7], allowing several atoms to be used together to represent a note, thereby affording the possiblilty to capture better the dynamics of the frequency spectrum of a note, and hopefully reducing the error in the transcription system. In this prior the block of atoms used to represent each note was made of a fixed number, P, of atoms which were adjacent in the dictionary $\mathbf{D} \in \Re^{M \times K}$. Here

$K = L \times P$ where L is the number of groups, thereby defining a set of indices G for the group-based dictionary:

$$G = \{G_l | G_l = \{P \times (l-1) + 1, ..., P \times l\}\} \forall l \in \{1, ..., L\}.$$

In [7] a variant of the Non-negative Basis Pursuit (NN-BP) algorithm [1] was proposed which we call NN-BP(GC). This variant differs from the NN-BP algorithm only through the calculation of a group coefficient, on which the thresholding step is performed, and is outlined in Algorithm 1. Transcriptions using this method had high recall, as many true positives were recovered, while displaying low accuracy as many false positives were also found, though many of the false positives were seen to be of low energy. This poor performance may be due partially to the lack of explicit group penalisation in this method.

A non-negative group version of OMP called Non-negative Nearest Subspace OMP (NN-NS-OMP) was also proposed. This was seen to suffer from a failure to capture low energy atoms, and harmonic jumping was seen to have a negative effect on time continuity in note events in the piano roll. As the method is iterative, a stopping condition needs to be selected, and it was found that selection of an apt stopping condition was tricky.

Algorithm 1 NN-BP(GC)

Input
$\mathbf{D} \in \Re_+^{M \times K}$, $\mathbf{S} \in \Re_+^{M \times N}$, δ, $\mathbf{T}^0 = \mathbf{D}^{\mathrm{T}}\mathbf{S}$, $\Gamma = 1^{L \times N}$
repeat
$t_{k,n} \longleftarrow t_{k,n} \dfrac{[\mathbf{D}^{\mathrm{T}}\mathbf{S}]_{k,n}}{[\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{T}]_{k,n} + \lambda}$
until a fixed number of iterations
$\mathbf{GC}_{l,n} = \sum \mathbf{T}_{G_l,n} \ \forall (l,n)$
$\mathbf{GC}_{l',n'} = 0$; $\Gamma_{l',n'} = 0 \ \ \forall \{l', n'\}$ *s.t.* $\mathbf{GC}_{l',n'} < \delta \times \max \mathbf{GC}$

Molecular sparsity [2] was proposed as an extension of greedy sparse algorithms, in which several atoms related through proximal structure were selected together at each iteration, based on a coefficient system which considered several atoms simultaneously. This approach has the advantage of favouring structure in the decomposition. For example in the Molecular Matching Pursuit (MMP) [2], a molecule of time-persisting tonal elements were extracted from the spectrogram at each iteration by performing tracking through time from an initially selected atom until the onset and offset of the tonal element were found, and all interim atoms were selected.

Initial attempts to build a molecular transcription system were seen to fail when polyphony grew as it became difficult to track pitched atoms (or groups of atoms), due to high projection values being present beyond the onset and offset points of a note, in particular when notes which were similarly pitched or harmonically related were active there. This led to a two-step approach. As previously mentioned the NN-BP(GC) displayed high recall and it was observed

that notes displayed time continuity in otherwise very noisy transcriptions, and it was proposed to first decompose the spectrogram using the NN-BP(GC). Isolated atom supports were pruned and clustering of time-persisting atoms into molecules was performed on the sparse support $\mathbf{\Gamma}$. The molecules were then input to a greedy method called Molecular Non-negative Nearest Subspace OMP (M-NN-NS-OMP) which selects at each iteration one predetermined molecule.

Algorithm 2 M-NN-NS-OMP

Input
$\mathbf{D} \in \Re_+^{M \times K}$, $\mathbf{S} \in \Re_+^{M \times N}$, $\mathit{\Gamma} \in \{0,1\}^{L \times N}$, G, α
Initialise
$i = 0$; $\mathit{\Phi} = 0^{L \times N}$; $B = \{\beta_n | \beta_n = \{\} \forall n \in \{1, .., N\}\}$
repeat
$i = i + 1$
Get group coefs Θ and smoothed coefs $\bar{\Theta}$
$\mathbf{x}_{G_l,n} = \arg\min_x \|\mathbf{r}_n^i - \mathbf{D}_{G_l}\mathbf{x}\|_2^2 \; s.t. \; \mathbf{x} \geq 0 \; \forall l \in \mathit{\Gamma}_n$
$\Theta_{l,n} = \|\mathbf{x}_{G_l,n}\|_1$; $\quad \bar{\Theta}_{l,n} = \sum_{n'=n}^{n+\alpha-1} \Theta_{l,n'}/\alpha$
Select initial atom and grow molecule
$\{\hat{l}, \hat{n}\} = \arg\max_{l,n} \bar{\Theta}_{l,n}$
$n_{min} = \min \bar{n} \; s.t. \; \mathit{\Gamma}_{\hat{l},\Xi} = 1, \; \Xi = \{\bar{n}, ..., \hat{n}\}$
$n_{max} = \max \bar{n} \; s.t. \; \mathit{\Gamma}_{\hat{l},\Xi} = 1, \; \Xi = \{\hat{n}, ..., \bar{n}\}$
$\beta_n = \beta_n \cup \hat{l} \; \forall n \in \Xi = \{n_{min}, ..., n_{max}\}$
Calculate current coefficients and residual
$\mathbf{t}_{G_{\beta_n},n} = \min_t \|\mathbf{s}_n - \mathbf{D}_{G_{\beta_n}}\mathbf{t}\|_2^2 \; \forall n \in \Xi$
$\mathbf{r}_n^{i+1} = \mathbf{s}_n - \mathbf{D}_{G_{\beta_n}}\mathbf{t}_{G_{\beta_n}} \; \forall n \in \Xi$
until stopping condition met

The M-NN-NS-OMP algorithm returns a sparse group coefficient matrix, $\mathbf{T}$, and the transcription performance using this approach was measured with both frame-based and onset-based analysis. The frame-based analysis is performed by comparing a ground truth and the derived transcription. Each frame which is found to be active in both the ground truth and the transcription denotes a *true positive - tp* while frames which are active only in the ground truth and transcription denote *false negatives -fn* and *false positives - fp*, respectively.

For event-based analysis, onset detection was performed on $\mathbf{T}$. An simple threshold-based onset detector was used, based upon the one used in [10] which registered an onset when a threshold value was surpassed and subsequently sustained for a given number of successive frames for a note in the coefficient matrix $\mathbf{T}$. A *tp* was registered when the onset was detected within one time bin of a similarly pitched onset in the ground truth. Similar to the frame-based analysis, an onset found only in the ground truth registered a *fn*, and an onset found only in the transcription registered a *fp*.

Using these markers the following metrics are defined for both frame- and event-based transcription; $Acc = tp \times 100/(tp + fp)$ relates the accuracy of the system in finding correct frames; the recall $Rec = tp \times 100/(tp + fn)$ defines the performance in terms of the amount of correct frames found relative to the number of active frames in the ground truth; $F = 2 * Acc * Rec/(Acc + Rec)$ defines overall performance, considering both false positives and negatives in the measure.

2.1 Experimental Results

Transcription experiments were run using the molecular approach on a set of pieces played on a Disklavier piano from the MAPS [3] database which includes a midi-aligned ground truth. A subdictionary was learnt for each midi note in the range $21 - 108$ from isolated notes also included in the MAPS database, and $\mathbf{D}$ was formed by concatenating these subdictionaries. Transcription was performed using the two-step NN-BP(GC) followed by M-NN-NS-OMP approach.

	Onset-based			Frame-based		
P	Acc	Rec	F	Acc	Rec	F
1	78.3	74.3	76.3	69.1	73.6	71.3
2	78.8	76.2	77.5	69.0	76.4	72.5
3	77.6	77.1	77.4	69.5	78.7	73.8
4	78.8	77.3	78.1	71.8	79.3	75.3
5	78.6	77.8	78.2	72.9	80.0	76.3

Table 1. Frame-based and onset-based transcription results for the proposed molecular approach, relative to the block size, P

We can see from the table of results the performance for both onset-based and frame-based metrics improves with the group size P, thereby validating the use of group sparse representations for this purpose. The experiments were run with a common value used as the stopping condition. Further experiments have shown that improved performance is possible using different values for each group size. In particular, an F-measure greater than 80% was achieved for frame-based transcription for $P = 5$.

3 Transcription Oracle for Sparse Methods

An oracle for transcription performance is proposed. OMP-based methods use a backprojection of the selected atoms onto the signal to produce the final coefficients, thereby gives a (non-negative) least squares error solution with a given support. As the MAPS [3] database comes with a standardised ground truth, we consider an oracle transcription for a given dictionary, given the ground truth

support. At each time bin we calculate the non-negative least squares solution using only the groups of atoms G_n^{oracle}, known from the ground truth to be active at the time bin n.

$$\mathbf{t}_{G_n^{oracle}} = \min_t \|\mathbf{s}_n - \mathbf{D}_{G_n^{oracle}}\mathbf{t}\|_2^2 \; s.t. \mathbf{t} > 0 \; \forall n \in \{1, ...N\} \tag{2}$$

The oracle group coefficient matrix $\mathbf{E}$ is formed by summing the coefficients of the indiviual group members

$$\mathbf{E}_{l,n} = \sum \mathbf{T}_{G_l,n}^{oracle} \; \forall\{l, n\} \tag{3}$$

4 Oracle Analysis

Using this oracle, we can probe the effectiveness of the approach taken to AMT. Interesting observations were made with relation to two aspects of the transcription system; often there is very low energy in supported atoms in $\mathbf{E}$, which may explain how the thresholding in the NN-BP(GC) effected the possible recall rate; secondly, using the oracle transcription provides an insight into the effectiveness of the onset detection system used.

4.1 Energy Based Thresholding

In the NN-BP(GC) algorithm, a thresholding factor δ is used, which is multiplied by the maximum value of the group sparse coefficients $\mathbf{GC}$. For the experiments in [7], a value of $\delta = 0.01$ was used. Using this value for δ it was found that the recall rate of the NN-BP(GC) algorithm in these experiments was 87%, and closer analysis showed that often the false negatives existed at the tail of sustained notes, were it is expected that low energy is displayed. This recovery rate effectively sets an upper bound on the possible recall rate of the M-NN-NS-OMP.

The oracle energy matrix $\mathbf{E}$ was calculated for each piece from the MAPS dataset used in the previous experiments for both ERB and STFT decompositions, both of which used dictionaries learnt from the same dataset of isolated notes in MAPS as used in the previous work [7]. The signals were undersampled to $22.05kHz$, and the ERB spectrogram used 256 frequency bin scale with a $23ms$ time window. The STFT used a 1024 frequency bin spectrogram, with a a 75% overlap, in order to use the same time resolution as the ERB. The NN-BP(GC) was also run for both tranforms to compare the effects of δ thresholding.

The results are displayed in Table 2, where it seen that Rec^{oracle}, the percentage of frames in the oracle transcription $\mathbf{E}$ with higher coefficients than the signal dependent threshold, $th = \delta \times \max \mathbf{E}$ is very similar in both transforms, across all values of delta. A similar pattern is also seen for Rec, the recall rate using the NN-BP(GC), which is smaller than Rec^{oracle}, but again is similar across the transforms, which suggests that the problem here is energy related, and not related to the dictionaries. It can be seen that while the recall rate increases as

	STFT			ERB		
δ	Acc	Rec	Rec^{oracle}	Acc	Rec	Rec^{oracle}
0.1	88.6	38.2	44.1	84.4	37.3	44.6
0.01	38.5	84.6	90.7	36.4	85.0	90.6
0.001	19.2	92.8	96.5	17.7	93.5	96.4
0.0001	12.7	95.2	97.1	12.2	95.6	97.0

Table 2. Analysis of effect of δ on Acc and Rec of NN-BP(GC) and the oracle

δ decreases, the accuracy of the NN-BP(GC) is greatly reduced. Using a smaller value of accuracy might negatively interfere with the final transcription, by introducing oversized molecules and may also effect on the computational load using the current approach as the M-NN-NS-OMP will require more projections.

4.2 Onset Analysis

In the prior work, a simple threshold-based onset detection system was used, which triggered an onset when a threshold value was surpassed and sustained for a minimum length of time. A true positive was flagged when this trigger happened within one time frame of a ground truth onset of the same note. Using the optimal transcription **E** we can test the effectiveness of this onset detection system. Experiments were run using the same parameters as in [7] and the results are presented in Table 3.

P	1	2	3	4	5
Rec	76.2	78.5	79.5	80.1	80.1
Acc	86.4	87.1	87.0	87.3	86.8

Table 3. Onset analysis of oracle transcription **E** for different values of P

The results are not promising given that an oracle transcription is given to the onset detector. Closer inspection of the individual results reveal systematic flaws in the onset detection. False positives are often found when a sustained note is retriggered by oscillation around the threshold value, behaviour which is often found in the presence of other note onsets and may be due to transient signal elements effecting the smoothness of the decomposition across time. Several common types of false negative were found. It is found that a note replayed with minimal time between the offset of the original event and the onset of the following event may produce a false negative where the observed coefficient has not already fallen below the threshold value. When several notes onset simultaneously, onsets may not be detected for all of these notes. A tendency for lower pitched notes not to trigger an onset event in the detection system is also no-

ticed. Further to this we also find some timing errors, where a false negative and a false positive are closely spaced.

5 Conclusion

We have previously proposed an AMT system based on group sparse representations which is relatively fast and shows promising results. An oracle transcription has been presented here, which gives some insight into the some weaknesses in the AMT system, as currently exists. Further work will focus on improving the AMT system, by incorporating a more sophisticated onset detection system and possibly using a new algorithm to perform the decomposition.

References

1. Aharon, M., Elad, M., Bruckstein, A. M.: K-SVD and its non-negative variant for dictionary design. In: Proc. of the SPIE conference wavelets, 2005, pp. 327-339
2. Daudet, L.: Sparse and structured decompositions of signals with the molecular matching pursuit. In: IEEE Transactions on Audio, Speech and Language Processing, 2006, pp. 1808-1816
3. Emiya, V., Badeau, R., David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. In: IEEE Transactions on Audio, Speech and Language, 2010, pp. 1643-1654
4. Leveau, P., Vincent, E., Richard, G., Daudet, L.: Instrument-Specific Harmonic Atoms for Mid-Level Music Representation. In: IEEE Transactions on Audio, Speech and Language, 2008, pp. 116-128
5. Poliner, G., Ellis, D.: A discrimative model for polyphonic piano transcription. In: EURASIP Journal Advances in Signal Processing, no. 8, 2007, pp. 154-162
6. Smaragdis, P., Brown, J. C.: Non-negative matrix factorization for polyphonic music transcription. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003
7. O'Hanlon, K., Nagano, H., Plumbley, M. D.: Structured Sparsity for Automatic Music Transcription. In: IEEE Int. Conference on Audio, Speech and Signal Processing 2012.
8. Abdallah, S.A., Plumbley, M. D.: Polyphonic transcription by non-negative sparse coding of power spectra. In: Proceedings ISMIR 2004, pp. 318-325
9. Benetos, E., Dixon, S.: Multiple-Instrument polyphonic music transcription using a convolutive probabilistic model. In: Proceedings of the Sound and Music Computing Conference 2011
10. Bertin, N., Badeau, R., Vincent, E.: Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. In: IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538549, Mar 2010.
11. Pati, Y. C., Rezaiifar, R.: 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.
12. Chen, S. S., Donoho, D. L., Saunders, M. A.: Atomic decomposition by Basis Pursuit. In: SIAM Journal on Scientific Computing, vol. 20, pp. 33-61, 1998.