SPARSE MUSIC DECOMPOSITION ONTO A MIDI DICTIONARY DRIVEN BY STATISTICAL MUSIC KNOWLEDGE

Boyang Gao, Emmanuel Dellandréa and Liming Chen
Université de Lyon, CNRS,
Ecole centrale de Lyon, LIRIS, UMR5205, F-69134, France.
E-mail: {Boyang.Gao, Emmanuel.Dellandrea, Liming.Chen}@ec-lyon.fr

ABSTRACT

The general goal of music signal decomposition is to represent the music structure into a note level to provide valuable semantic features for further music analysis tasks. In this paper, we propose a new method to sparsely decompose the music signal onto a MIDI dictionary made of musical notes. Statistical music knowledge is further integrated into the whole sparse decomposition process. The proposed method is divided into a frame level sparse decomposition stage and a whole music level optimal note path searching. In the first stage note co-occurrence probabilities are embedded to generate a sparse multiple candidate graph while in the second stage note transition probabilities are incorporated into the optimal path searching. Experiments on real-world polyphonic music show that embedding music knowledge within the sparse decomposition achieves notable improvement in terms of note recognition precision and recall.

1. INTRODUCTION

Large amounts of digitalized music available drive the need for the development of automatic music analysis, for example automatic genre classification, mood detection and similarity measurement. Most of the tasks rely on effective features extracted from music signals. Among various features, music notes, denoted by MIDI notes in this paper, provide the most comprehensive information, since music is indeed sound poetry comprised of notes played by instruments. If notes are accurately recovered from music signal, automatic music analysis can be greatly improved. However, mixing different instrument playing is trivial while decomposing is quite challenging due to the intrinsic complexity of polyphonic music.

Recovering notes from a music wave signal is usually referred to multiple F0 estimation. The approaches in literature can be roughly sorted into two categories: parameterized like statistical model based methods and non-parameterized like non-negative matrix factorization (NMF) based methods. Parameterized approaches usually assume that multiple F0 can be described by particular models with a small number of free parameters that can be estimated from the signal. For example, in [1] Kameoka et al. propose a multi-pitch analyzer named the harmonic temporal structured clustering (HTC) method that jointly estimates pitch, intensity, onset and duration. HTC decomposes the power spectrum time series into distinct clusters such that each cluster has originated from a single source modeled by a Gaussian Mixture Model (GMM). The parameters of the source model are computed thanks to maximum a posteriori (MAP) estimation. In [2], Wu et al. extend Kameoka's work to propose a flexible harmonic temporal timbre model to decompose the spectral energy of the signal in the time-frequency domain into individual pitched notes. Each note is modeled with a 2-dimensional Gaussian kernel. Parameters of Gaussian mixtures are then estimated by expectation maximization (EM) algorithm with a global Kullback–Leibler (KL) divergence cost function.

Unlike parameterized approaches, non-parameterized methods like NMF focus on recovering pitch combinations from the signal data itself without presuming any underlying model forms. For example, NMF [3] based methods try to decomposes the multiple pitch spectrum matrix X into two matrices W and H [4]. W contains various harmonic patterns and H consists of activation behaviors so that $X = WH$. In [5], Hoyer extends the original NMF by adding a regulation term to make H sparse. Sparseness property is quite helpful especially for music note estimation, since a short period music can only contain a few notes played together, compared with all possible notes.

NMF is such an extensible framework that it largely dominates non-parameter methods. For example, in [6] Zafeiriou adds a linear discriminant analysis (LDA) stage to the activities extracted by NMF. In [6-8], fisher-like discriminant constraints are embedded inside the decomposition. In [9], Lewandowski proposes a supervised method with two discriminative criteria that maximize inter-class scatter and quantify the predictive potential of a given decomposition. In order to extract features that enforce the separability between pitch labels, pitch information present in time-aligned musical scores is fused in sparse NMF. In [10], Sakaue combines Bayesian inference with NMF to propose a Bayesian non-negative harmonic-temporal factorization (BNHTF). BNHTF models the harmonic and temporal structures separately with Gaussian mixture models. In [11], a music sparse decomposition approach is proposed using high quality MIDI

dictionary. This work is a variant of sparse NMF and uses non-negative matching pursuit to solve sparse NMF. Unlike NMF that processes the entire signal, this work constructs the activity matrix H column by column. It is still worth mentioning the work in [12] where Leveau et al. propose to learn instrument specified note atoms with a modified matching pursuit and a tracking of the played instrumental notes by searching an optimal path with respect to the reconstruction error.

Thus, previous works in the literature have demonstrated the effectiveness of various approaches in multiple-F0 estimation, especially NMF based methods. However, under the NMF framework, the entire music spectrum series X are treated as a whole object to be reconstructed. Most of the algorithms focus on reducing the spectrum reconstruction error so as to overlook the compatibility in concurrent and consecutive notes. This batch processing style makes it hard to fuse note co-occurrence and transition information to guide note detection during the matrix factorization. Indeed, after the signal spectrum matrix is factorized, W and H are new represents of the music, which have lost signal context information for post-processing to correct possible error. Even in [12], the Viterbi algorithm is used to search the optimal path only with respect to a minimum reconstruction error and neglects the underlying note relations. Nevertheless correlation between concurrent and consecutive notes contains significant heuristics that can help to correct the decomposition error introduced by a signal level analysis.

Therefore, to employ note statistical information to help music sparse decomposition, we propose in this paper a two-stage sparse decomposition approach integrated with music knowledge. In frame level decomposition stage, note co-occurrence probabilities are embedded to guide atom selection in modified matching pursuit algorithm with a MIDI dictionary. A sparse multiple candidate graph is then constructed to provide backup choices for later selections. In the global optimal path searching stage, note transition probabilities are incorporated together with a goodness measure of frame decomposition. Its principle is to guide the local sparse music decomposition with co-occurred notes information and decode the global optimal decomposition path with consecutive note knowledge. Due to the Gabor limit, time and frequency resolution cannot be well satisfied at the same time. Thus, we emphasize the frequency resolution aspect rather than the exact time location, since correct note recognition is more important for our following classification task.

The rest of this paper is organized as follows: Section 2 introduces our two-stage approach in detail. Section 3 shows experimental results on real-world music signals. The conclusion is drawn in the final section.

2. SPARSE DECOMPOSITION WITH NOTE STATISTICS

Our proposed method consists of two main steps. In the first step, the entire music signal is framed and a modified orthogonal matching pursuit algorithm is performed on each frame to generate decomposition candidates. In the second step, decomposition candidates are connected to form a directed graph. An optimal path is then constructed to produce the final decomposition result.

2.1 Frame Level Sparse Decomposition

Former study shows that elaborating an appropriate musical dictionary is a key issue since this set of atoms has to be rich enough to characterize the varieties of real word music. Although [12] has developed sophisticated method to learn atoms from instrument recordings, it is still impractical to apply to a large instrument set. Therefore, in order to get adequate instrument note sounds, we propose to make use of a MIDI synthesizer. Logic Pro 9 is employed in our approach to generate the MIDI note dictionary because of its huge instrumental library and the high sound quality. Unlike pre-installed MIDI synthesizer with sound card, Logic Pro 9 uses a large number of real instrument recordings to make synthesized wave signal as natural as possible.

To build the MIDI dictionary, we choose the first 80 realistic instruments as in general MIDI level 1 set and 31 percussion instrument sets including 1860 percussion sounds. For each instrument, we keep 60 notes from note 31 to note 90. The 60 notes span 5 octaves from low to high, covering most instrumental playing range. The duration of each note is set to 186ms, which are 4096 samples under a sample rate of 22050Hz. This duration is long enough to hold one attack-decay-sustain-release (ADSR) envelope and leads to 5.38Hz in terms of frequency resolution, which is sufficient to discriminate adjacent notes in piano roll. Our MIDI note wave is then converted into a single-sided power spectrum obtained by applying the short time Fourier transform (STFT) with a Hamming window. The final MIDI dictionary thus contains 6660 2048-dimensional vectors.

The adoption of the sparse representation is based on the hypothesis that during a 186ms time slot, there will not be many notes played together. Therefore, concurrent notes are sparse within one frame. Sparse representation [13] is originated from finding the solution x^* of an underdetermined linear system $Dx = y$ so that x^* contains as few non-zero components as possible. In most cases $Dx = y$ is hard to satisfy, thus in practice $\|x\|_0$ and $\|Dx - y\|_2$ are minimized simultaneously instead.

Armed with our MIDI dictionary, the classical matching pursuit algorithm like orthogonal matching pursuit (OMP) [17] must be modified because the single-sided power spectrum words in MIDI dictionary impose an inherent positive constraint on sparse solutions. In other words, any negative component of a sparse solution is prohibited, as negative appearance of certain notes is impossible. To solve this problem we adopt a positive constraint matching pursuit (PCMP) algorithm that is mentioned in [11][14]. The difference between OMP and

PCMP is in updating a provisional solution step: for OMP, least mean square (LMS) suffices to solve the minimization resulting in residual signals orthogonal to support set. For PCMP, however, after the positive constraint minimization, orthogonality is not always guaranteed, thus the algorithm is turned to a weak orthogonal matching pursuit.

When scrutinizing the decomposition results of PCMP within one frame, we found a number of irregular note combinations. This is due to PCMP's over-fitting target signals without considering any compatibility of concurrent notes. In fact, atom selection in each iteration of orthogonal matching pursuit algorithm is very important. OMP guarantees that expending support set with any linear independent atoms will decrease the reconstruction error and at the same time keep the residual signal orthogonal to the new expanded support set. Any atom selected in the support set will permanently reside. Therefore previously selected atoms have a great influence on following ones and alter the overall OMP performance. Although PCMP does not always hold orthogonal property, the principle remains the same.

Selecting a new atom in dictionary is thus the very place where concurrent note heuristic information should be embedded. To formulate concurrent note information, Bayes model is employed in our approach to approximate the posterior probability of potential note given observed notes

$$P(X|\boldsymbol{O}) = \frac{\prod_{i=1}^{N} P(O_i|X)P(X)}{\sum_Y \prod_{i=1}^{N} P(O_i|Y)P(Y)} \quad (1)$$

where $\boldsymbol{O} = \{O_1, O_2 \cdots O_N\}$ denotes N observed notes obtained by first N PCMP iterations, X represents a potential co-occurred note with $\boldsymbol{O}$. The note prior probability $P(X)$ and the note co-occurrence posterior probability $P(X|Y)$ are estimated from our classical music MIDI database. To obtain $P(X|Y)$, a joint distribution $P(X,Y)$ is firstly estimated by accounting the frequency with overlap degree of the concurrent note X and Y. Then $P(X|Y)$ is obtained by normalizing $P(X,Y)$ over Y. Although equation (1) provides instructive information to help select appropriate note combinations, it is still risky to only consider the best note decomposition, since the second best one may be more appropriate in adjacent note context. To avoid the one best bias, we propose to preserve multiple candidates to give top-N best decompositions chances to recover in optimal path searching.

Orthogonal matching pursuit is a greedy algorithm. In each iteration only the best atom will be added into support set. This can be risky in some cases, since once a "bad" atom is selected, this error cannot be corrected in the future. In [15], it has been shown that it is possible to select "bad" atom initially so as to trap OMP from reconstructing target signals. Methods like OCMP in [19] are proposed to overcome the problem. However, in music decomposition the same note in different octave or from the same kind of instruments shares the similar harmonic pattern. Therefore it is hazardous to rule out a suboptimal decomposition too early before adjacent note compatibility is checked.

To overcome this drawback of OMP, we propose to keep N best candidates in each iteration instead of only one. To measure the goodness of frame decomposition we define $goodness = \alpha \log(P(\boldsymbol{X})) - \|\boldsymbol{r}\|_2^2$, where $\boldsymbol{X}$ is sparse note decomposition vector, $\boldsymbol{r}$ is decomposition residual signal, $P(\boldsymbol{X}) = \sum_Y \prod_{i=1}^{N} P(X_i|Y)P(Y)$ denotes note concurrent probability, α is a free parameter that balances concurrent probability term and reconstruction error term.·As an example shown in Figure 1 we keep the top 3 decomposition candidates in every iteration. In the first iteration (C), (E), (G) are kept. In the second iteration, (C, D), (E, F) and (E, G) are obtained according to the reconstruction error and concurrent probability. Note that (G) selected in the first iteration is eliminated because its descendant combinations (G,*) are inferior to others'. After 3 iterations, combinations of (C, D, E), (C, D, G) and (E, F, B) survive, as shown in orange.

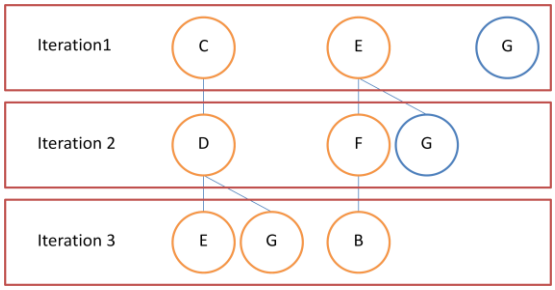

Figure 1. Multiple candidate selection example

When sparse decomposition terminates, the top N note candidates are derived for every signal frame. The best one can be treated as the decomposition result of the current frame. Besides, all candidates are preserved for constructing the optimal decomposition path when we further investigate inter-frame relations. The multiple candidate PCMP algorithm that we propose is summarized in Algorithm 1.

2.2 Global Level Optimal Note Path Searching

All previous steps in section 2.1 focus on improving sparse note decomposition within one signal frame. When further scrutinizing the PCMP decomposition between consecutive frames, we can still find a number of discontinuous note decompositions, in which the note sequence has sudden abnormal jumps in adjacent frames, including octave shift or sharp/flat drift. This is due to a lack of note transition regulation and because the sparse decomposition only minimizes reconstruction error in current frame without considering any neighbor frame contexts.

Besides the co-occurred ones, consecutive notes bear strong correlations which convey various melody, temporal and dynamic information of music. It is reasonable to incorporate such sequential knowledge of notes as to suppress the discontinuous note error.

Task: Approximate the solution of problem: $min_{\boldsymbol{x}}\|\boldsymbol{x}\|_0$, subject to $\boldsymbol{Ax} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}$.
Input: Dictionary $\boldsymbol{A}$, signal $\boldsymbol{b}$, max iteration number I, top N candidates to keep, balance parameter α, note posterior probability $P(X|Y)$ and error threshold ϵ_0.
Output: sparse solution: $\boldsymbol{x}^*$

Initialization:
Initial residual: $\boldsymbol{r}_c^0 = \boldsymbol{b}$.
Initial support: $S_c^0 = \emptyset$.
Initial candidate queue: $Q^0 = \{S_1^0, S_2^0 \cdots S_N^0\}$

Main iteration:
for $k = 1$ to I
 for $c = 1$ to N
 - Compute $P(j|S_c^{k-1})$ according to equation (1)
 - Compute error: $\epsilon(j) = \min_{z_j}\left\|\boldsymbol{a}_j z_j - \boldsymbol{r}_c^{k-1}\right\|_2^2 - \alpha \cdot \log\left(P(j|S_c^{k-1})\right)$ all j using the optimal choice $z_j^* = \boldsymbol{a}_j^T \boldsymbol{r}_c^{k-1} / \left\|\boldsymbol{a}_j\right\|_2^2$.
 - Find top N minimizers of $\epsilon(j)$ to form $J = \{j_1, j_2 \cdots j_N\}$ such that $J \cap S_c^{k-1} = \emptyset$, and push $S^k = \left\{S_c^{k-1} \cup \{j_1\}, S_c^{k-1} \cup \{j_2\} \cdots S_c^{k-1} \cup \{j_N\}\right\}$ into Q^k

 end
 for each $S_i^k \in Q^k$
 - Compute $\boldsymbol{x}$ that minimizes $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2^2$ subject to $support\{\boldsymbol{x}\} = S_i^k, \boldsymbol{x} \geq \boldsymbol{0}$.
 - Compute residual $\boldsymbol{r}_i^k = \boldsymbol{b} - \boldsymbol{Ax}$.

 end
 Ascendingly sort Q^k according to $\left\|\boldsymbol{r}_i^k\right\|_2^2 - \alpha \cdot \log\left(P\left(S_i^k\right)\right)$ and keep the first N items.
 If any $\left\|\boldsymbol{r}_i^k\right\|_2^2 < \epsilon_0$ break.
end
Output result: $Q^* = Q^k$.

Algorithm 1. Positive constraint matching pursuit producing multiple candidates

We thus apply transition probabilities to model relations between two decomposition candidates in adjacent frames. To formulate note transitions, Bayes model is adopted so that conditional probability can be approximated from individual note pairs. Since at most N candidates remain in one frame the posterior probability of candidate j in frame t given candidate i in frame $t-1$ is calculated as

$$P\left(\boldsymbol{X}_t^{(j)} \middle| \boldsymbol{X}_{t-1}^{(i)}\right) \stackrel{\text{def}}{=} \frac{\prod_{k=1}^{D} P\left(X_{k,t}^{(j)} \middle| \boldsymbol{X}_{t-1}^{(i)}\right)}{\sum_l \prod_{k=1}^{D} P\left(X_{k,t}^{(l)} \middle| \boldsymbol{X}_{t-1}^{(i)}\right)} \quad (2)$$

where $\boldsymbol{X}_t^{(j)} = \{X_{1,t}^{(j)}, X_{2,t}^{(j)} \cdots X_{D,t}^{(j)}\}$ denotes the decomposition candidate j in frame t containing D notes. $P\left(X_t | \boldsymbol{X}_{t-1}\right)$ is calculated similarly as in equation (1).

Thanks to the multiple decomposition candidates generated by the modified PCMP previously, an inter-decomposition directed graph is further constructed to help determining the optimal decomposition path through all frames, as illustrated in Figure 2. In this directed graph, each decomposition candidate forms a node and outgoing edge denotes the transition probability computed by equation (2). The nodes are disconnected within the same frame indexed with t.

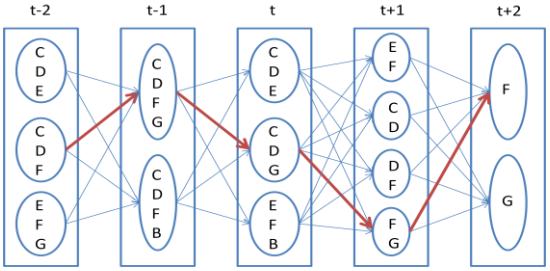

Figure 2. Optimal path decoding example

In order to connect transition probabilities with the sparse decomposition candidates, the decomposition goodness measure is converted into corresponding probabilities as $P(\boldsymbol{X})^{\alpha} \cdot \exp(-\|\boldsymbol{r}\|_2^2)$, since the frame signal $\boldsymbol{b}$ and atoms in $\boldsymbol{A}$ have been normalized to unit vectors. The conversion also reflects a reasonable assumption that the reconstructed signal is approximately Gaussian distributed around original one. Treating the decomposition candidates as hidden states of a Hidden Markov Model (HMM), Viterbi algorithm decodes the optimal decomposition candidate path $\boldsymbol{p}^*$:

$$\boldsymbol{p}^* = \underset{\boldsymbol{p}}{\text{maxarg}} \prod_{t=1}^{F} P(\boldsymbol{X}_{p_t}^{(t)})^{\alpha} P\left(\boldsymbol{X}_{p_t}^{(t)} \middle| \boldsymbol{X}_{p_{t-1}}^{(t)}\right)^{\beta} e^{-\left\|\boldsymbol{r}_{p_t}^{(t)}\right\|_2^2} \quad (3)$$

where t is the frame index, F is the total number of frames, β is a balance parameter to adjust emphasis, p_t denotes decomposition candidate index in frame t along path $\boldsymbol{p}$. Initially $P(\boldsymbol{X}_{p_1}^{(1)} | \boldsymbol{X}_{p_0}^{(1)})$ =1.

3. EXPERIMENT AND RESULTS

To evaluate the decomposition quality of the proposed PCMP with note statistics, a multi-timbral music with time domain note reference has been used, which is provided in Mirex2007 multiF0 development data [16]. The music is a recording of the fifth variation from *L. van Beethoven Variations from String Quartet Op.18 N.5*, lasting for 54s. 5 instruments are included into the music. Each instrument was recorded separately and then mixed to a mono 44.1 kHz 16 bits wave file. The whole music is tested by our system (PCMP with multi-candidate and Viterbi) against its ground truth MIDI file.

Another widely used dataset adopted in our experiments is MUS, provided in MAPS [18]. MUS contains 270 pieces of classical and traditional music, recorded in different conditions which vary in piano instruments and surroundings. For each piano music piece, as in [9], first 30 seconds are tested by our system against ground truth MIDI files.

Figure 3 shows precision and recall scatter diagram of the proposed decomposition that improve original PCMP, noted as PCMPMC and PCMPMCV. Table 1 displays the

comparisons in terms of precision, recall and F-measure between our proposed method and the state of the art results. F-measure is defined as the harmonic mean of precision and recall. Statistic of note recognition precision and recall has been made upon consecutive 186ms frames. For ground truth MIDI, if 70% of some note lies in the frame the note is accounted and there is no frequency tolerance. Threshold for drawing the diagram is imposed on sparse solution vector in each frame to filter insignificant note detection according to its sparse solution value. Different thresholds result in scatter points in Figure 3. Two free parameters α and β are set to 0.8 and 1.3 to balance reconstruction error and note statistics.

	Prec. (%)	Rec. (%)	F-meas. (%)
NMF[4]	41.1	46.6	45.3
HTC[1]	57.4	51.3	54.2
JHT[2]	**59.7**	61.4	**60.5**
PCMPMCV	51.8	**72.0**	60.3

Table 1. Average multiple pitch estimation performance on MIREX2007 dataset.

From Figure 3 we can see that when co-occurrence note information is integrated into PCMP, the precision increases about 6% while recall increases by 2%~3%. When the note transition information is fused and the optimal path decoding is applied, the precision and recall are further improved by 5% and 2% approximately. From Table 1 and Figure 3 we can find that if no threshold is imposed on the sparse solution of PCMPMCV, 72% of the notes can be recalled while the precision is 51.8% resulting in an F-measure of 60.3 %. The recall of our best configuration outperforms state of the art result in [2] by more than 10% while the precision is 8% lower, resulting in an F-measure 0.2% lower than that reported in [2].

	Prec. (%)	Rec. (%)	F-meas.(%)
Spectral constraints [20]	71.6	65.5	67.0
Isolated note spectra [20]	68.6	66.7	66.0
DNMF-LV[9]	68.1	65.9	66.9
DNMF-AE[9]	66.8	68.7	67.8
SONIC[21]	**74.5**	57.6	63.6
PCMPMCV	60.7	**77.3**	**68.0**

Table 2. Average multiple pitch estimation performance on MUS dataset.

Table 2 shows the precision, recall and F-measure results on MUS data set. All parameters and setups are the same as used in previous experiment except for the conditional probability estimation. In this experiment the rest data other than first 30 seconds are used to estimate conditional probabilities $P(X|Y)$. From Table 2 we can observe that the proposed approach achieves the highest recall and F-measure of 77.3% and 68%, although obtains the lowest precision of 60.7%.

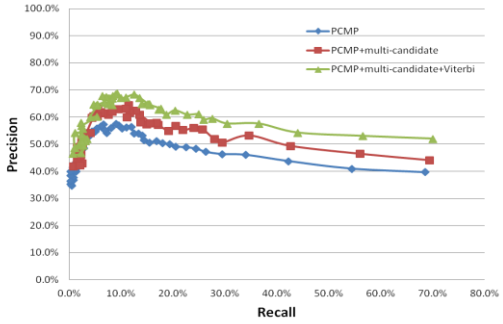

Figure 3. Note precision vs. recall of the two improvements

From the two experiments, we can find that with statistical musical knowledge sparse decomposition is improved in terms of both precision and recall. The proposed approach tends to obtain superior recall and F-measures but lower precisions compared with variant NMF and other methods. Higher recall means the more information is preserved in the decomposition results. Since our final aim of the decomposition is to provide decent features for music classifications, the performance of our system is actually preferred. Our higher recalls and F-measures are attributed to the quality of MIDI dictionary as well as statistical music knowledge fused in sparse decomposition. Longer analysis window is another important factor.

When comparing decomposition with the ground truth, we found numbers of instrument errors even with correct note detections, which is likely caused by mismatches between the MIDI dictionary and the real-world data. In some cases concurrent and transition probability of notes can even make incorrect compensation to original PCMP, which is probably due to the limitation of the naive Bayes model. To overcome these drawbacks, dictionary adaptation techniques and sophisticated graphical models will be proposed and investigated in our future work.

4. CONCLUSION

We have proposed in this paper a novel sparse music decomposition approach driven by music knowledge. It employs note statistical information to improve sparse decomposition with a MIDI dictionary. In the frame level, music signals are decomposed onto a MIDI dictionary with a note co-occurrence heuristic. Transition probabilities are then computed between adjacent decomposition candidates through the whole frame sequence. The final optimal decomposition path is then constructed by the Viterbi algorithm. Experimental results show that embedding concurrent note statistics in PCMP and applying a note sequence heuristic allows improving the note recognition precision and recall.

5. ACKNOWLEDGMENT

This work is partly supported by the French ANR under the project VideoSense ANR-09-CORD-026.

6. REFERENCES

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

[2] J. Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Multipitch estimation by joint modeling of harmonic and transient sounds," in *Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 25–28, 2011.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556-562, 2001.

[4] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[5] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557– 565, 2002.

[6] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 683–695, 2006.

[7] N. Guan, D. Tao, Z. Luo and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2030–2048, 2011.

[8] Y. Wang, Y. Jia, C. Hu and M. Turk, "Fisher non-negative matrix factorization for learning local features," in *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2004.

[9] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative non-negative matrix factorization for multiple pitch estimation," in *Proceedings of 13th International Conference on Music Information Retrieval (ISMIR)*, 2012

[10] D. Sakaue, T. Otsuka, K. Itoyama, and H.G. Okuno, "Bayesian non-negative harmonic-temporal factorization and its application to multipitch analysis," in *Proceedings of 13th International Conference on Music Information Retrieval (ISMIR)*, 2012.

[11] B. Gao, E. Dellandréa, and L. Chen, "Music sparse decomposition onto a midi dictionary of musical words and its application to music mood classification," in *Proceedings of 10th IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1‑6, 2012.

[12] P. Leveau, E. Vincent, G. Richard and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 116-128, 2008.

[13] M. Elad, *"Sparse and Redundant Representations from Theory to Applications in Signal and Image Processing,"* Springer-New York, 2010.

[14] A.M. Bruckstein, M. Elad, and M. Zibulevsky, "Sparse non-negative solution of a linear system of equations is unique," in *Proceedings of 3rd IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 762–767, 2008.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[16] http://www.music-ir.org/mirex/wiki/2007:Multiple_Fundamental_Frequency_Estimation_%26_Tracking

[17] Y. C. Pati, R. Rezaiifar, P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition." In *Proceedings of IEEE Signals, Systems and Computers*, vol. 1, pp. 40-44, 1993

[18] 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, 2010.

[19] G. Rath, and C. Christine, "A complementary matching pursuit algorithm for sparse approximation," in *Proceedings of European Signal Process. Conf.*, Lausanne, Switzerland. 2008.

[20] 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.

[21] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.