高時間分解能と高周波数分解能のスペクトログラムの並列NMFによる多重音解析*

☆落合和樹, 中野允裕, 小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1 はじめに

本研究では，自動採譜を目的としたスペクトログラムの非負値行列分解を用いた新しい音楽音響信号解析手法を提案する．自動採譜は，音楽音響信号処理における重要な課題のひとつであり，録音されたものしかない曲を楽譜にすることで演奏に役立てることや，音楽を音符の記号列に変換することで音楽検索に応用することができる．自動採譜のためにはどの高さの音がどのタイミングで鳴り始めどれくらいの時間長で鳴っているかという音符情報の取得と，得られた音符列を音楽的に正しい楽譜に作り上げることが必要である．

多重音から基本周波数や発音時刻を推定する研究が従来から数多くなされており，近年では，多重音解析に有効な手段として非負値行列分解 (Nonnegative Matrix Factorization, NMF) が注目されている [1]．これは，ある単音を1つの基底スペクトルでモデル化しそれが音量のみ変化しているとみなすことと，スペクトログラムがスパースであるという仮定により多重音を単音毎に分解できることを期待している．

自動採譜では音高と発音時刻推定の正確性が同時に求められることから，高周波数分解能であることと高時間分解能であることの両方が必要となるが，従来のNMFでは，時間分解能と周波数分解能の間に存在する不確定性原理によるトレードオフにより，音高発音時刻の推定精度を同時に高めることが課題となっている．

これまでNMFにおいてスペクトログラムの解析フレーム長やフレームシフトが積極的に着目されることはなかったが，解析フレーム長を変えることで時間 (周波数) 分解能が変わることを利用することができると考えられる．フレーム長を短くすると高時間分解能となり細かい時間間隔での解析ができるために発音時刻推定精度を高めることができるが，スペクトルの非定常な変化によりNMFの基底モデルに合わなくなり単音毎への分解が失敗しやすくなってしまう．これを解決するためにフレーム長を長くすると，スペクトルの非定常な変化を吸収して表現できるので音高推定精度を高めることができる．本研究ではこの点に着目し，音高と発音時刻を高精度に推定するために，2つの異なるフレーム長でのスペクトログラムを並列にNMFで分解する新しい音楽音響信号分解手法を提案し，実際に発音検出実験を行い有用性を検討する．

2 非負値行列分解 (NMF)

本研究では，信号を短時間Fourier変換により時間周波数領域へ変換したスペクトログラムを扱う．観測信号の振幅 (もしくはパワー) スペクトログラムを非負値行列 $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega,T} \in \mathbb{R}^{\geq 0,\Omega\times T}$ とし，これが限られた数の基底の重ね合わせで表現されるとすると，NMFによるスペクトログラムの分解表現は，

$$Y_{\omega,t} \simeq X_{\omega,t} = \sum_{i=1}^{I} H_{\omega,i}U_{i,t} \tag{1}$$

というように，基底 $\boldsymbol{H} = (H_{\omega,i})_{\Omega,I} \in \mathbb{R}^{\geq 0,\Omega\times I}$ とアクティベーション $\boldsymbol{U} = (U_{i,t})_{I,T} \in \mathbb{R}^{\geq 0,I\times T}$ の2つの非負値行列を決定することで得られる．ここで，ω, t はそれぞれ周波数と時刻に対応するインデックス，i は基底のインデックスであり，観測スペクトログラムが I 個の基底スペクトルと各基底の音量に相当するアクティベーションの積で表現されるというモデルとなっている (Fig. 1)．ただし，分解スケールの任意性を回避するため，

$$\sum_{\omega} H_{\omega,i} = 1 \quad (i = 1, \cdots I) \tag{2}$$

という仮定を与えておく．

NMFは一般的に観測とモデル間の何らかの距離尺度を目的関数とし，これを最小化する問題として解かれる．距離尺度としてはEuclidean distanceやI-divergenceなどがよく用いられており，いずれにおいても，効率のよい乗法更新アルゴリズムにより非負性の保証された解が得られることがわかっている [1]．

3 提案手法

3.1 問題設定

自動採譜をするためには，発音時刻の推定と鳴っている各音高の正確な推定と分離が必要で

---

*Parallel Nonnegative Matrix Factorization of high-time-resolution and high-frequency-resolution spectrograms for multipitch analysis of music signals. by OCHIAI Kazuki, NAKANO Masahiro, ONO Nobutaka, SAGAYAMA Shigeki (The University of Tokyo)

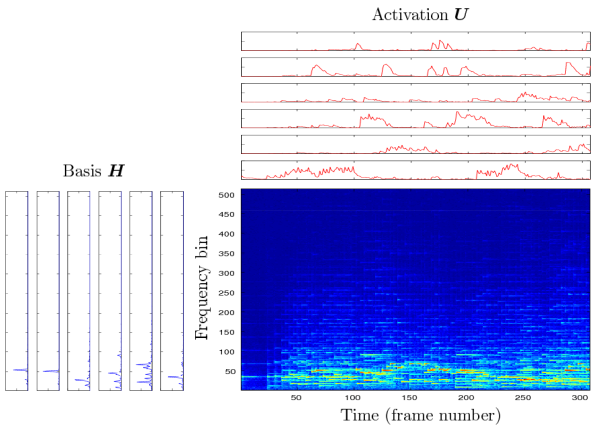

Fig. 1　音楽音響信号に NMF を適用した例．各基底スペクトルとそのアクティベーションに分解される．

ある．楽器音には基本周波数と倍音に強いエネルギー持つという性質があり，NMF では，単音毎に調波成分のみ非零とし乗法更新によりその構造を保持する手法 [2] や，調波成分を複数の調波構造の線形和で表現する手法 [3] が提案されている．この他にも単音を表すモデルとして，基本周波数 ω_i とその倍音 $k\omega_i (k = 2, \cdots, K)$ に小さい分散 σ を持つ正規分布の重み $a_{i,k}$ での混合で表現するという手法が提案されている [4]．本研究においてもこれらの枠組みは利用できると考えられ，単音に分離するために次式のような打ち切りの正規分布の混合を基底の初期値とした．

$$H_{\omega,i} = \begin{cases} \displaystyle\sum_{k=1}^{K} \frac{a_{i,k}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\omega - \log k\omega_i)^2}{2\sigma^2}\right] \\ \qquad \left(2^{-\frac{1}{12}} k\omega_i \leq \omega < 2^{\frac{1}{12}} k\omega_i\right) \\ 0 \qquad \text{(otherwise)} \end{cases} \tag{3}$$

また，楽器のアタック時には調波成分だけでなく広帯域のスペクトルが発生することから，周波数方向に広い分布を持つ基底を与えておくことで単音が鳴り始める際に複数の基底が同時に立ち上がることを抑制できる．しかしながら，アタックの強弱や音高によってその瞬間的なスペクトルは異なり，時間とともに急峻に変化するため，発音時刻の高分解能での推定のために短いフレーム長で解析を行うと，たとえアタック用の基底を用意しても各音高用の基底の立ち上がりを完全に防ぐことはできない (Fig. 2)．また，後処理による平滑化では発音時刻の推定精度が失われてしまう．

そこで，解析フレーム長を長くすると時間分解能が低下し，アタック時における音高用の基底の立ち上がりが平滑化され，各音高の推定精度が向上すると考えられるが，その一方で発音時刻を正確に取ることができなくなる．以上から，発音時刻と各音高の正確な同時推定がトレードオフとなっている．

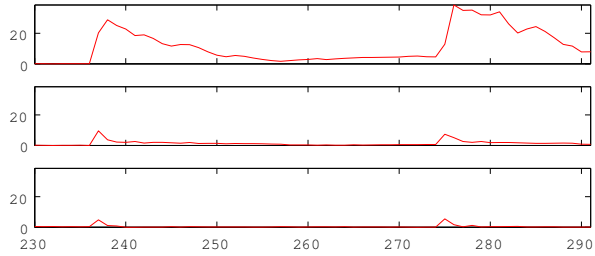

Fig. 2　ピアノの A_4(真ん中のラ) が 2 回鳴っている部分の各基底のアクティベーションの様子 (一部)．A_4 の基底に関する最上段のアクティベーションが立ち上がる瞬間に他の基底も立ち上がっている．解析フレーム長は 64ms.

3.2　スペクトログラムの並列 NMF

NMF によるスペクトログラムの分解において，短い解析フレームでは時間分解能が高く発音時刻を細かく取得しやすく，長いフレームでは音高推定精度が向上するという性質に着目し，前節で挙げた問題を解決するために，長短 2 種類の解析フレーム長のスペクトログラムを併用することを考える．スペクトログラムの NMF においては，長短の解析フレーム両方で相互に妥当な分解が得られていることが期待される．これにはそれぞれの NMF で得られた基底とアクティベーションは形状が似ていることが望ましく，各々で対応する周波数ビン，フレーム間の誤差が小さくなるようにすれば，双方の基底が同じ音高情報を保ったまま解析フレームが短い方のアクティベーションでアタック時に出現する複数基底の誤った立ち上がりを NMF の更新反復毎に抑制していくことができると考えられる．この制約を目的関数に加え NMF の更新により最小化することとする．本研究では実験的に制約項を 2 乗誤差として，2 つの解析フレーム長の比を $2^C (C > 1, C \in \mathbb{N})$ とすると，

$$\mathcal{R}_H(\theta) = \sum_{i}^{I} \sum_{\omega_1}^{\Omega_1} \left| H^{(1)}_{\omega_1,i} - \sum_{\omega_2 \in \omega_1} H^{(2)}_{\omega_2,i} \right|^2 \tag{4}$$

$$\mathcal{R}_U(\theta) = \sum_{i}^{I} \sum_{t_2}^{T_2} \left| U^{(2)}_{i,t_2} - \sum_{t_1 \in t_2} U^{(1)}_{i,t_1} \right|^2 \tag{5}$$

となる．ここで，$H^{(1)}, U^{(1)}, \Omega_1$ はそれぞれ短い解析フレームでの基底とアクティベーション，周波数ビン数を，$H^{(2)}, U^{(2)}, T_2$ は長い解析フレームでの基底とアクティベーション，フレーム数を表し，$\theta = \{\boldsymbol{H}^{(1)}, \boldsymbol{H}^{(2)}, \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}\}$ である．また，スペクトログラムのスパース性に関する制約

項を [5] に従い L_p ノルム $(0 < p \leq 1)$ とすると，

$$\mathcal{S}(\theta) = \sum_{i,t_n,n} \left| U_{i,t_n}^{(n)} \right|^p \tag{6}$$

と書ける．さらに，隣り合った基底すなわち音高が半音異なる音同士の倍音構造は似ていると考えられることから，

$$\mathcal{Q}(\theta) = \sum_{n} \left\| \boldsymbol{H}^{(n)} - \boldsymbol{W}_1^{(n)} \boldsymbol{H}^{(n)} \boldsymbol{W}_2^{(n)} \right\|_2^2 \tag{7}$$

の制約を加える [2]．$\boldsymbol{W}_1^{(n)}, \boldsymbol{W}_2^{(n)}$ は形状を比較するための変換行列で，それぞれ各基底の要素を半音分上げる行列と基底行列を 1 列シフトする行列である．いま，NMF の距離尺度として，I-divergence:

$$\begin{aligned}\mathcal{I}(\theta) = \sum_{\omega,t} \Bigg[& Y_{\omega,t} \log \frac{Y_{\omega,t}}{\sum_i H_{\omega,i} U_{i,t}} \\ & - \left(Y_{\omega,t} - \sum_i H_{\omega,i} U_{i,t} \right) \Bigg]\end{aligned} \tag{8}$$

を考えると，解くべき問題は，
minimize

$$\begin{aligned}\mathcal{J}(\theta) = & \sum_n \mathcal{I}^{(n)}(\theta) + \mu_H \mathcal{R}_H(\theta) \\ & + \mu_U \mathcal{R}_U(\theta) + \lambda \mathcal{S}(\theta) + \eta \mathcal{Q}(\theta)\end{aligned} \tag{9}$$

subject to

$$\begin{aligned}\forall_i \sum_{\omega_n} H_{\omega_n,i}^{(n)} = 1, \quad \forall_{\omega_n,i} H_{\omega_n,i}^{(n)} \geq 0, \\ \forall_{i,t_n} U_{i,t_n}^{(n)} \geq 0 \quad (n = 1, 2), \\ \mu_H,\ \mu_U,\ \lambda,\ \eta \geq 0\end{aligned} \tag{10}$$

となる θ を求める制約付き最適化問題となる．この目的関数を直接最小化することは困難なので，補助関数法 [6] を用いることによりパラメータの更新式を導出することができる．更新式の導出アルゴリズムと更新式は紙面の都合上省略するが，乗法更新により非負性を保ち調波構造を維持できるものとなっている．反復の度に基底を式 (2) を満たすように規格化する，

3.3 音符検出

NMF により得られた基底とアクティベーションから自動採譜をするためには，鳴っている各音とその発音時刻を得ることが必要である．各音は強弱をつけて演奏されるが，どんなに弱い音でもある一定以上の音量が出ていることが考えられる．一度発せられた音の音量は徐々に減衰していき，リリースとともに急激に小さくなる．音高と発音時刻の推定には様々な方法が考えられるが，本稿では単純な方法として，発音消音に関する閾値を用意し，分解能の異なる 2 つのスペクトログラムでの NMF の結果に対し，長い解析フレームでのアクティベーションでは閾値未満の値をすべて 0 にし，短い解析フレームでのアクティベーションでは連続する数フレームで閾値を超えている部分のみ残し他の値を 0 にして，前者で閾値を超えた時に後者で対応するフレームでの値がすべて 0 であればその音高は発音されていないとし，0 でないフレームがあれば発音されたとし，その中で最大値を取るフレームを発音時刻とする．

4 評価実験

4.1 予備実験

提案法の効果を検証するために，用いられた音高が既知であるという条件の下で，RWC クラシック音楽データベース [7] からピアノで演奏された楽曲 (RWC-MDB-C-2001 No.26) を用いて従来の NMF との比較実験を行った．短時間 Fourier 変換には，サンプリング周波数 16kHz，フレーム長 64ms と 256ms，フレームシフト 32ms と 128ms，Hanning 窓という条件を用い，データ長約 10s，基底数 12(うちアタック時の広帯域スペクトルを吸収させる基底 1 つ含む)，各制約項に関するパラメータは実験的に $\mu_H = 0,\ \mu_U = 1,\ \lambda = 1,\ p = 0.5,\ \eta = 0$ とした．アクティベーション $\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}$ の初期値は乱数とし，更新の反復回数は 30 回とした．解析結果の一部を Fig. 3 に示す．F$_\sharp$3(ファ) の音は解析フレームが短い 64ms のときのフレーム数約 230 以降で 2 回鳴っていたが，従来 NMF ではフレーム数 230 以前にも立ち上がりが多数発生してしまっているのに対し，提案法ではそれが抑制できていることが確認できる．解析フレームが長い 256ms のときは従来法での解析同様提案法でも音高の推定ができている．

4.2 発音検出実験

次に，提案法の発音検出における有効性を検証するために，使用された音高が未知の状態で従来の NMF との比較実験を行った．短時間 Fourier 変換における条件は予備実験と同じで，用いた楽曲は RWC クラシック音楽データベースよりピアノ曲 (RWC-MDB-C-2001 No.26, 30) のデータ長約 30s で，基底数 55(うちアタック用 1 つ含む)，各パラメータは $\mu_H = 0.5,\ \mu_U = 2,\ \lambda = 1,\ p = 0.5,\ \eta = 0.5$ とし，反復回数は 60 回とした．音高と発音時刻を推定し F 値を求めたものを Table 1 に，各音高の発音消音時刻をピアノロールで表

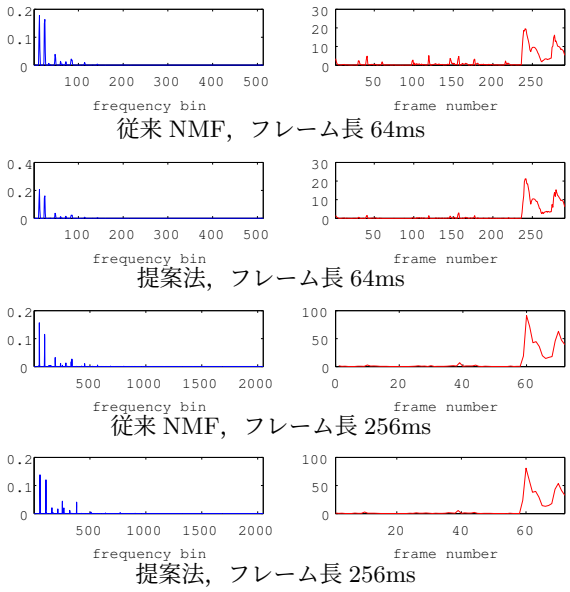

Fig. 3　Mozart: Sonata in A Major, K.331(300i) 冒頭部分の解析結果のうち $F_{\sharp 3}$(ファ) の音に当たる基底とそれに対応するアクティベーションの比較.

Table 1　解析した曲と音符検出における F 値 (%)

Composer	Title	Notes	従来	提案
W. A. Mozart	Sonata in A Major, K. 331(300i)	105	73.0	86.0
F. Chopin	Nocturne in E♭ Major, Op. 9, No. 2	124	60.5	82.4

示したものを Fig. 4 に示す．Mozart のソナタにおける提案手法での誤検出はすべて倍音成分であった．従来手法で誤検出されていた音高推定誤りが減ったことにより，発音時刻推定精度を保ちつつ音高推定精度が向上していることが確認できた．

5　おわりに

本研究では，自動採譜に向けた音高と発音時刻の同時推定のために，高時間分解能と高周波数分解能でのスペクトログラムを並列に NMF で分解することによる新しい多重音解析手法を提案した．音楽音響信号を用いた実験により，従来手法に比べ発音時刻推定精度を保ちつつ音高推定精度が向上したことを確認した．今後は，アタック用の基底を発音検出に用いるなどより正確な発音消音時刻の推定手法の検討や音符出力アプリケーションの作成の検討をしている．

参考文献

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, Oct. 1999.

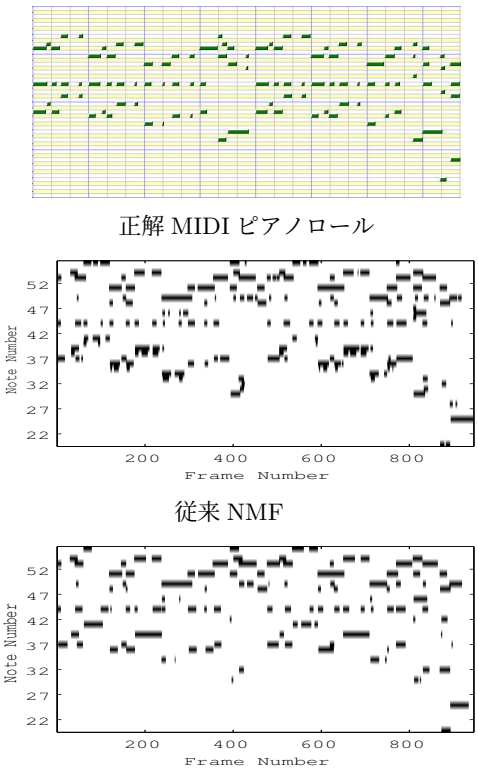

Fig. 4　Mozart: Sonata in A Major, K. 331(300i) を解析しピアノロールとして表示したもの．発音されていると推定された音高とその発音から消音までを黒で表示してある．

[2] S. A. Raczyński, N. Ono, and S. Sagayama, “Multipitch Analysis with Harmonic Nonnegative Matrix Approximation,” *Proc. 8th International Conference on Music Information Retrieval*, pp. 381–386, Sep. 2007.

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

[4] H. Kameoka, T. Nishimoto, and S. Sagayama, “A multipitch analyzer based on harmonic temporal structured clustering,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 982–994, Mar. 2007.

[5] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, “Complex NMF: A new sparse representation for acoustic signals,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009*, pp. 3437–3440, Apr. 2009.

[6] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, pp. 556–562, MIT Press, 2001.

[7] G. Masataka, H. Hiroki, N. Takuichi, and O. Ryuichi, “RWC Music Database : Classical Music Database and Jazz Music Database,” *IPSJ SIG Notes*, vol. 2002, no. 14, pp. 25–32, 2002.