2N-1

楽譜情報を用いた NMF による音楽音響信号の音源分離

糸山 克寿　　駒谷 和範　　尾形 哲也　　奥乃 博
京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

グラフィックイコライザなどのツールを使って音楽鑑賞するユーザの要求には,「Hi-Fi な音を楽しみたい」「高度な臨場感を再現したい」などの録音時の忠実な音を再現するというものと,「ベースパートを抑制して，ギターパートを強調したい」「ボーカルだけにエフェクトをかけたい」などの楽器パートごとに操作を行って音楽を鑑賞するというものの 2 つがあるのではないだろうか．再生された音楽をただ聴くだけでなく，後者のように音楽を聴くという行為に対して積極的に働きかけることを能動的音楽鑑賞と呼ぶ．イコライザは，本来の設計では周波数帯域を操作し，残響コントロールなどと組み合わせることで，音響環境の変化を実現するための技術である．これらのツールが楽器パートごとの強調・抑制などの目的に使われる場合もあるが，本質的に後者の要求を満たすには不十分である．

このようなユーザの要求に応えられる技術として，吉井らは Drumix [1] を開発している．ユーザは Drumix を使ってドラムスの音色を置換え，また，ドラムパターンを編集できるので，能動的な音楽鑑賞という新たな鑑賞法が可能となった．ただ，Drumix は楽曲中のドラムスだけを対象としていた．

我々は，トップダウンな情報として利用可能な楽譜とボトムアップな信号分離技術である Non-negative Matrix Factorization (NMF) [2] を組み合わせた音源分離処理を開発した。この音源分離処理を用いて，Drumix の目指した能動的な音楽鑑賞をドラムスだけでなく，ピアノやギターなどの一般的な楽器にまで対象を拡張することが本研究の最終的な目的である．

混合音からなる楽曲のそれぞれの楽器パートごとに対して操作を行うためには，まず混合音からそれぞれの楽器パートを分離して，楽器パートごとの音響信号を作る必要があり，あらかじめ混合音中の楽器音を認識しておくことが不可欠である．Drumix ではドラム音認識に基づいたドラムスパート分離を行っていたが，混合音中の一般楽器音認識システムの精度は分離に直接適用できるまでには至っていないため，本稿では楽譜情報として利用可能な MIDI ファイルを用いる．

2. 音源分離処理の課題

我々が開発した音源分離処理について述べる。

2.1 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) とは，$n \times m$ 非負行列 V を $V \simeq WH$ となるような $n \times r$ 行列 W と $r \times m$ 行列 H に分解する手法である．r は通常 n, m よりも小さい数を与えるので，W, H は V を圧縮した行列とみなすことができる．$\boldsymbol{w}_l, \boldsymbol{h}_l$ $(l = 1, \ldots, r)$ はそれぞれ n, m 次元の列，行ベクトルである．

W と H は反復的な推定によって求められる．V と WH との間にどの程度の差があるかを表す指標に Kullback-Leiblar Divergence (KLD) を用いると，W, H の各要素を

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}}$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}}$$

と更新することで，V と WH と間のの KLD が増加しないように W, H を更新することができる．

2.2 Non-negative Tensor Factorization

Non-negative Tensor Factorization (NTF) [3] は，行列しか扱えなかった NMF を一般化し，階数 d，$n_1, n_2, \ldots, n_d$ 次元のテンソル V を階数 2，n_1, r 次元のテンソル $W^{(1)}$，階数 2，n_2, r 次元のテンソル $W^{(2)}$，…，階数 2，n_d, r 次元のテンソル $W^{(d)}$ に分解する手法である．

階数 3 のテンソル $V = (v_{ijk})$ を $W^{(1)} = (w^{(1)}_{ir})$, $W^{(2)} = (w^{(2)}_{jr})$, $W^{(3)} = (w^{(3)}_{kr})$ へと分解する場合，NMF と同様に KLD を (v_{ijk}) と $(\sum_r w^{(1)}_{ir} w^{(2)}_{jr} w^{(3)}_{kr})$ との差を表す指標に用いると，$W^{(1)}$ の各要素を

$$w^{(1)}_{ir} \leftarrow w^{(1)}_{ir} \frac{\sum_{j,k} v_{ijk} w^{(2)}_{jr} w^{(3)}_{kr} / w^{(1)}_{ir} w^{(2)}_{jr} w^{(3)}_{kr}}{\sum_{j,k} w^{(2)}_{jr} w^{(3)}_{kr}}$$

と更新することで，NMF と同様に非負制約を用いた分解が行える．$W^{(2)}, W^{(3)}$ についても同様である．

2.3 NMF, NTF による音源分離

時間 t $(T_0 \le t \le T_1)$，周波数 f $(F_0 \le f \le F_1)$ で定義されるパワースペクトル $p(f,t)$ において，r 個の楽器音が含まれているとする．ここで解くべき問題は，観測された $p(f,t)$ を r 個のパワースペクトル $p_1(f,t), \ldots, p_r(f,t)$ に分解することである．

パワーの加算性が成り立つとし，さらに l 番目の楽器音のパワースペクトルは

- 時間的に不変である周波数基底 $w_l(f)$
- 周波数的に不変であるパワーの時間変化 $h_l(t)$

の積で表されているものとする．ここで，f, t を適当な間隔で離散化すると，$p(f,t)$ と $\sum_l w_l(f) h_l(t)$ との KLD を最小化する r 個の $(w_l(f), h_l(t))$ の組を求めることは，V との KLD を最小化する W, H を求めることと等価であるので，この問題に NMF を適用することができる．

さらに，入力信号が複数チャンネルからなっており，各チャンネルにおける観測信号が遅延や残響を含まない各楽器音の瞬時混合であるときは，楽器音ごとに，全てのチャンネルでの信号を分離することが要求される．この場合は，各楽器音のチャンネルごとのゲイン $a_l(c)$ という新たなパラメータを導入することで，分離問題は観測されたパワースペクトル $p(c,f,t)$ を最もよく近似する r 個の $(a_l(c), w_l(f), h_l(t))$ を推定するという問題に帰着される．この問題を解くには，複数チャンネル $c = 1, \ldots, d$ での観測パワースペクトル $p(c,f,t)$ から r 個の $(a_l(c), w_l(f), h_l(t))$ の組を求めればよく，シングルチャネルの場合と同様の離散化を行うことで，NTF を用いて解くことができる．

Sound source separation for polyphonic musical signal based on NMF using score information: *Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

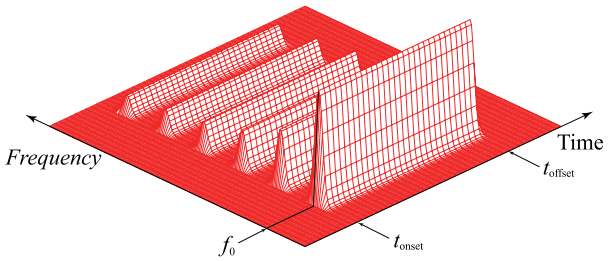

図 1: 調波構造モデル

2.4 NMF による音源分離の問題点

NMF や NTF による音源分離では，

1. ランダムに初期化された W や H を更新する
2. あらかじめ調波構造を表す基底を多数 W として与えておいて H のみを更新する

などの手法で分離を試みるが，

1. 必ずしも一つ一つの基底ベクトルの組が一つ一つの楽器音に対応しない
2. どの基底ベクトルの組がどの楽器音に対応するのかを推定するのが困難

といった問題点がある．

MIDI ファイルを用いることで，事前に楽器音認識が行われていることを仮定できるのでこれらの問題点は解決される．しかし，実信号で使われている楽器音の正確な音色は分からない，同じ楽器，同じ音高・音長で演奏していても，ビブラートなどの奏法の違いといったそれぞれの音ごとのずれが存在する，といった問題がある．そこで，実信号の楽器音の音色を近似し，それぞれの音ごとの違いに応じた分離を行うため，NMF を用いて分離を行う．

MIDI ファイルを用いた音源分離は以下のようになる．

1. MIDI ファイルから音符単位で MIDI メッセージを抽出し，調波構造のモデルを作成する
2. 調波構造モデルを基底ベクトルの初期値として与え，実信号の基底ベクトルを NMF で推定する

調波構造のモデルは，

- 周波数基底は，F0 と倍音成分の周波数にピークを持つくし型構造
- パワーの時間変化の基底は，オンセットからオフセットまで一様

になるように作成した．これを図示すると図 1 のようになる．$f_0, t_{\text{onset}}, t_{\text{offset}}$ はそれぞれ対象とする音の F0，オンセット，オフセットを表す．

3. 実験

本稿で述べた音源分離手法の効果を確かめるため，実験を行った．分離対象の音響信号には MIDI ファイルの楽曲を MIDI 音源で演奏した信号を用いた．これは，分離結果を定量的に評価するために各楽器パートをミックスする前の正確な音響信号を得る必要があるためである．

3.1 実験条件

評価のため，“RWC 音楽データベース：ポピュラー音楽 (RWC-MDB-P-2001)[4]” から#001–#010 の 10 曲を選んだ．これらの楽曲の MIDI ファイルを用いて各 MIDI チャンネルごとに 30 秒の音響信号を録音し，それらを加算合成したものを評価に用いた．

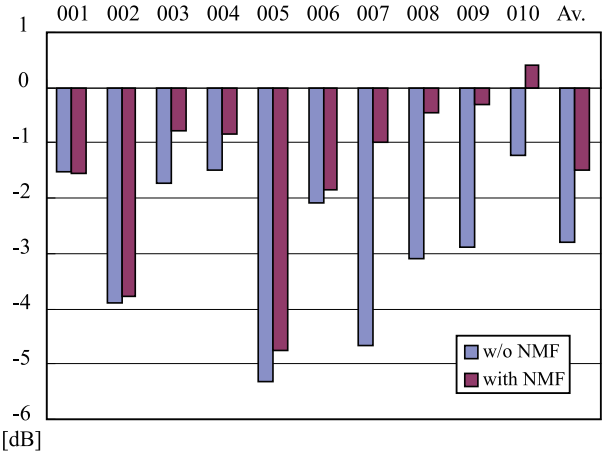

図 2: 実験結果

3.2 NMF による調波構造・パワー変化推定の評価

本実験の目的は，NMF によって調波構造とパワーの時間変化の基底ベクトルを推定することの有効性を評価することである．調波構造モデルをそのまま分離に用いた場合と，調波構造モデルを NMF の初期値として与え，調波構造とパワーの時間変化を推定した後に分離を行った場合の 2 通りの条件で実験を行った．図 2 に本実験の結果を示す．

#001 を除く 10 曲中 9 曲について，NMF を用いて基底ベクトルを推定することで SNR が向上することを確認した．楽器パートごとにみると，多くの楽器パートについては SNR が向上していたものの，電子楽器などでは SNR が低下する傾向が見られた．電子楽器のパワースペクトルは純粋に調波構造だけを含むため，初期値として与えた調波構造モデルを用いると非常によく適合するが，混合音に対して NMF で調波構造の推定を行うと本来その楽器には含まれないはずの非調波成分が混ざるためと考えられる．#001 においても，多くの楽器パートでは SNR が向上していたものの，電子楽器パートで特に大きな SNR の低下があった．

4. おわりに

本稿では，混合音からなる楽曲の各楽器パートの音量を操作可能なオーディオプレイヤーを実現するための NMF を用いた音源分離手法を開発した．今後は，音源分離手法の改良などを行っていく予定である．特に，本手法では音高と音長のみから決まる調波構造モデルを用いているため，複数の楽器が同じパートを演奏している場合，それらを分離することはできない．このような信号を扱うためにはどの楽器で演奏されているかを知ることが不可欠であるため，楽譜情報に加えて楽器情報も用いた分離手法の開発が重要な課題である．

謝辞 本研究の一部は，科研費，21 世紀 COE，CREST-Muse の支援をうけた．

参考文献

[1] 吉井他，“Drumix: ドラムパートのリアルタイム編集機能付きオーディオプレイヤー”，インタラクション 2006，207–208.

[2] D. D. Lee et al, “Algorithms for Non-negative Matrix Factorization”, Advances in Neural Information Porcessing Systems 13: 556–562, MIT Press, 2001.

[3] D. FitsGerald et al, “Sound Source Separation using Shifted Non-negative Tensor Factorization”, *ICASSP* 2006, 653–656.

[4] 後藤他，“RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース”，情処論，Vol. 45，No. 3，728–738，2004.