非負値行列分解における時間包絡の単旋律性と基底の類似性に基づく音楽音響信号の楽器音分離*

○村尾一真 (東大・工), 中野允裕, 北野佑, 小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1　はじめに

本稿では，非負値行列分解（Nonnegative Matrix Factorization, NMF）[1] の後処理として楽器音クラスタリングを行うことによる，モノラル混合音楽音響信号からの単旋律楽器音分離の手法を提案する．

CD に代表されるような，録音された音楽音響信号を考えると，例えばカルテットによる録音であれば 4 つの楽器による音から構成されるように，これらの音楽の多くは，複数の楽器や人の声を含んでいると言える．このような録音信号を混合音楽音響信号と呼ぶことにする．いま，仮に各楽器からセンサ（マイク）へ到達した信号の個々が存在したとすると，混合音楽音響信号はこれらの足し合わせによって得られると考えられるが，逆に，混合音楽音響信号を分割して個々の楽器音信号を得る操作は非常に難しい問題とされる．

NMF は，モノラル音響信号中に混在する構成音を事前知識なしで分離抽出できる可能性をもっているとして期待されている手法である．これは，観測信号が有限の基底系によってモデル近似できるという考えに拠ったものであり，限られた数の音階の要素から構成される音楽音響信号の特性と相性が良いと考えられる．NMF は音響信号のスペクトログラムを，楽曲におけるノート（単一楽器の単一音程）に相当することが期待される基底に分解する．

しかし，NMF によって分解された基底から楽器音分離を達成するためには，楽器音クラスタを求めるパーミュテーション問題（順列組み合わせ問題）を解く必要がある．本稿では，基底を楽器音に相当するクラスタに適当にクラスタリングすることによる，パーミュテーション問題解法の提案を行う．

2　非負値行列分解

観測音響信号から短時間フーリエ変換や Wavelet 変換等の音響解析によって，時間周波数領域における振幅スペクトログラム $\mathbf{X} = [X_{f,t}]_{f,t}$ を得る．f は周波数または対数周波数に，t は時刻に対応するインデックスである．NMF は，非負値行列 $\mathbf{X}$ を，行列 $\mathbf{B}$ と $\mathbf{G}$ の行列積によって近似的に表すことができるというモデルに基く．すなわち，観測 $\mathbf{X}$ に対し，

$$X_{f,t} \approx \sum_i B_{f,i} G_{i,t} \tag{1}$$

なる $\mathbf{B}$，$\mathbf{G}$ を推定する．i は基底に対応するインデックスである．

このモデルでは，観測信号が有限の基底系によって構成されるという仮定がおかれている．言い換えると，例えばフルートの A4 音が別々の時刻に発音されたとして，これらの信号の形状は厳密には一致しないが，これらを類似したスペクトルをもつパターンであるとみなし，ひとつの基底に相当するものと考える．従って，$\mathbf{B}$ と $\mathbf{G}$ は，ある音に対応するスペクトルを示す基底行列（Basis Matrix）と，その発音時刻を示すゲイン行列（Gain Matrix）を示す．分解のスケールの任意性を除くため，

$$\sum_f B_{f,i}^2 = 1 \tag{2}$$

と定めることにする．

$\mathbf{B}$ と $\mathbf{G}$ は，非負性に基いた乗法更新アルゴリズムによって反復更新して学習される．[1]

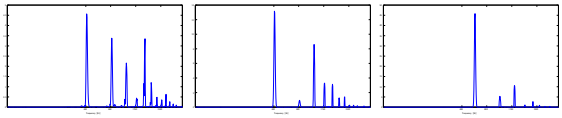

Fig. 1　ヴァイオリン:440 Hz（左），クラリネット:440 Hz（中），クラリネット:660 Hz（右）のスペクトル．クラリネット同士の調波構造の類似性が大きい

3　単旋律楽器音クラスタリング

3.1　着眼点

NMF の乗法更新によって得られた各基底のスペクトルとゲインの組をクラスタリングし，楽器音分離を達成するためのパーミュテーション解決基準として，同じ楽器音に属する基底同士は類似していること，単旋律楽器を仮定すると同一楽器に属する基底同士は同時発音しないこと，という 2 つの尺度を導入する．2 つの尺度によって，基底同士の相関行列を定義し，群平均法（Group Average Method, GAM）[3] によるクラスタリングを行うことで，楽器音分離を行う．

3.2　調波構造による基底スペクトル同士の相関

ある種類の楽器音らしさ，を与える特徴のひとつが，ピッチと呼ばれる基本周波数（f0）に対する倍音成分の比率である．[Fig.1 参照] この倍音比率を調波構造と呼ぶことにすると，同一種の楽器においては，基本周波数が異なる場合でも調波構造は類似している．[4],

いま，前節で得られた基底 $\mathbf{B}$ から，i 番目の基底インデックスによる列ベクトル $\mathbf{B}_i$ のように，横軸に対数周波数をとった基底スペクトルをみると，基本周波数に対する倍音成分の間隔は $\log 2$，$\log \frac{3}{2}$，$\log \frac{4}{3}$，と高倍音になるほど狭くなる一方で，基本周波数が変化しても，この配置は変わらず平行移動するだけである．

従って，基底のもつスペクトルの同士の調波構造による相関として，同一楽器による基底同士の相関を大きく，別楽器による基底同士の相関を小さくとることができるように

$$C_1(i,j) = \max_l \sum_k B_{k+l,i} B_{k,j} \tag{3}$$

を定める．

---

*Music Signal Separation based on Monophonicity and Common Harmonicity in Nonnegative Matrix Factorization. by MURAO Kazuma, NAKANO Masahiro, KITANO Yu, ONO Nobutaka, and SAGAYAMA Shigeki (University of Tokyo)

3.3 単旋律性によるゲイン同士の相関

人間が楽器音を識別する際に，情報として用いられるのは，スペクトル調波構造だけではない．例えば，ピアノやギターといった，和音を発音できる楽器が演奏されていないことが予め分かっている場合，同時に 2 音以上が鳴っている箇所があれば，その音のそれぞれは別の楽器によって発音されていることが分かる．言い換えれば，この知見は，それぞれの楽器が単旋律しか演奏しない，という仮定に基く．

前節で得られた行列 $\mathbf{G}$ の i 番目の基底インデックスによる行ベクトル $\mathbf{G}_i$ は，それぞれの音量変化を示していると言える．よって，ある 2 基底の音量を時刻を変化させながら観察してゆき，両基底が同時に発音される，つまり音量が大きな値をとることがあれば，これらの 2 基底は別の楽器による基底であると考えられる．これは基底同士の内積をとる操作に他ならず，ゲイン同士の相関として

$$C_2(i,j) = \left(1 - \frac{\mathbf{G}_i \cdot \mathbf{G}_j}{|\mathbf{G}_i||\mathbf{G}_j|}\right) \tag{4}$$

を定義すると，同時発音しない基底同士の相関は，ある微小値 ϵ として $1-\epsilon$ より大きな値をとる．

3.4 階層クラスタリング

上記で定義した C_1，C_2 の各要素をは [0 1] の範囲の値をとる．いま C_2 によるバイナリフィルタを設計し，2 つの尺度による相関行列

$$C(i,j) = \lceil C_2(i,j) + \epsilon \rceil C_1(i,j) \tag{5}$$

を検討する．$\lceil \cdot \rceil$ によって値の小数部分を切り捨てる．これを基底同士の相関行列として，階層クラスタリングの手法のひとつである，GAM を用いてクラスタリングを行う．階層クラスタリングとは，相関の大きい要素同士を再帰的に反復結合してクラスタを生成する手法である．予めクラスタ数（楽器数）を与えておくことで反復を終了するため，各クラスタに含まれる基底数を与える必要がない．GAM においては，各反復ごとに結合されたクラスタごとの相関を計算しなおすため，間違った基底が連鎖的にクラスタリングされるチェイニング効果と呼ばれる現象を防止することが可能である．

4 シミュレーションによる分離実験

提案法の有効性を確認するため，シミュレーションによる楽器音分離の実験を行った．単旋律楽器による 3 つの録音データ（信号長： 9.0 s，サンプリング周波数： 16 kHz) を用い，これらの加算によって得られた混合音楽音響信号に対して提案法を行うことで 3 つの推定分離結果を得た．$\mathbf{X}$ は Wavelet 変換（解析スケール： 16.0 ms, 14.0 cent, ）によって計算し，時間連続性とスパース性をコストとした NMF の手法 [2] を用いた．$\mathbf{B}$，$\mathbf{G}$ の初期値はいずれも乱数生成により設定し，反復回数は 200 回とした．また前節のクラスタリング（$\epsilon = 0.05$）により，基底インデックスに対するある楽器クラスタ $\{i | i \in I_n\}$（n は楽器インデックス）を定め，この基底を用いて，推定スペクトログラム

$$\hat{Y}_{f,t}^{(n)} = \sum_{i \in I_n} B_{f,i} G_{i,t} \tag{6}$$

を求めた．

Fig.2 に，3 種類の元信号，混合信号，分離推定した信号それぞれのスペクトログラムを示した．また表 1 に，正解信号に対する混合信号と分離信号の SNR を示した．推定信号と対応する元信号のスペクトログラムを $\mathbf{Y}$ とし，SNR は

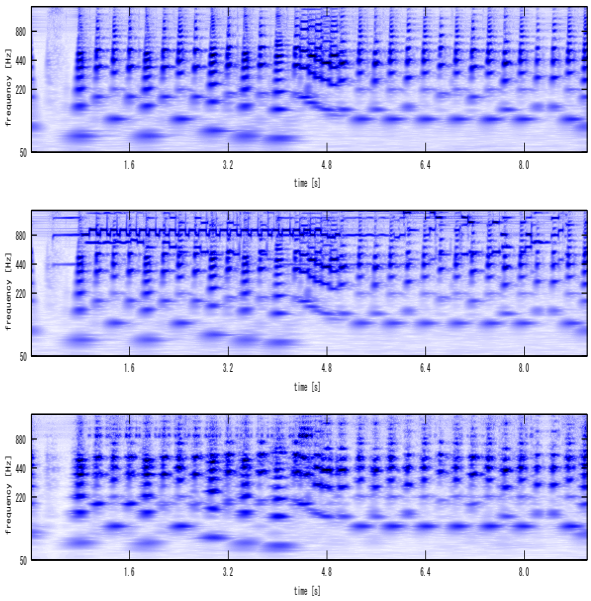

Fig. 2 元信号（上段）と 3 種類の信号加算による混合信号（中段）と楽器音分離結果（下段）のスペクトログラム

Table 1 分離性能（SNR [dB]）

	signal 1	signal 2	signal 3
mixture	1.81	-8.72	-2.47
separated	3.04	-4.69	3.16

$$\mathrm{SNR[dB]} = 10 \log_{10} \frac{\sum_{f,t} |Y_{f,t}|^2}{\sum_{f,t} |Y_{f,t} - \hat{Y}_{f,t}|^2} \tag{7}$$

によって定義した．提案手法による音源分離によって，いずれの音源信号に対しても混合信号からの SNR の改善がみられた．

5 まとめ

本稿では，基底の類似性と単旋律性に基づいた楽器音クラスタリングを，NMF の後処理として提案し，モノラル混合音楽音響信号に対しての楽器音分離を検討した．また本稿で後処理として行ったクラスタリングを NMF の内部処理として行うことによって，基底分解の精度を向上させることを今後の課題として検討している．

謝辞 本研究の一部は、文部科学省科学研究費補助金基盤研究 (A) (課題番号 00303321) と科学技術振興機構 CrestMuse プロジェクトの支援を受けて行われた．

参考文献

[1] D. D. Lee and H. S. Seung, Advances in Neural Information Processing Systems, vol. 13, pp. 556–562, 2000.
[2] T. Virtanen, IEEE Trans. Audio Speech Language Process., vol. 15, no. 3, pp. 1066–1074, 2007.
[3] 元田, 津本, 山口, 沼尾, “データマイニングの基礎,” オーム社, 2006.
[4] Z. Duan, Y. Zhang, C. Zhang and Z. Shi, IEEE Trans. Audio Speech Language Process., vol. 16, pp. 766–778, 2008.