スパース性と連続性の制約付き非負値行列分解に基づく調波音・打楽器音分離*

◎北野佑, 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、モノラル音楽音響信号から調波音と打楽器音を分離する手法について議論する。この分離手法は、打楽器やノイズなどの非調波成分を含んだ多声音楽信号の楽音分析における前処理、打楽器パートの強調や打楽器パターン変更といった音楽加工など、多くの応用が期待される。

調波音・打楽器音分離の従来手法として、観測スペクトログラムに対して非負値行列分解 (NMF) を適用した際に発生する基底の Permutation 問題を、予め学習したサポートベクターマシーンの識別面により解くものがある [2]。この手法は分離だけでなく、打楽器音の種類 (バスドラムやスネアなど) ごとに分解し、さらにそれぞれの Transcription もできるというメリットがあるが、予め調波音と打楽器音のサンプルを用意し、学習しておく必要があるのが難点である。

一方、調波音と打楽器音のスペクトログラム上での滑らかさの異方性に着目し、事前学習なしで、調波的な音と打楽器的な音に分離することのできる HPSS (Harmonic/Percussive Source Separation) と呼ばれる手法も提案されている [3]。リアルタイムで分離することのできる手法であるが、前述の非負値行列分解を用いた手法とは違い、打楽器音の種類ごとに分解はできない。

これらを踏まえ、本研究では、事前学習なしに調波音・打楽器音分離を行い、さらに打楽器音の種類ごとに分解する手法を提案する。HPSS における調波音と打楽器音のモデルと非負値行列分解における基底分解モデルを組み合わせた定式化を以下で導入し、シミュレーション実験を行うことにより、提案法の有効性を検討する。

2　非負値行列分解

本稿において信号は全て短時間 Fourier 変換、もしくは wavelet 変換により、時間周波数領域へ変換されたものとして扱う。観測信号の振幅 (もしくはパワー) スペクトログラム $Y_{x,t}$ 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、

$$Y_{x,t} \simeq F_{x,t} = \sum_{k \in K} B_{x,k} G_{t,k} \tag{1}$$

となるように近似的に基底 $\mathbf{B} \equiv (B_{x,k})_{X \times K}$ とゲイン $\mathbf{G} \equiv (G_{t,k})_{T \times K}$ に非負制約下で分解するのが非負値行列分解である。ここで x, t はそれぞれ周波数、時間の index にあたる。また k は基底の index、K は基底の集合であり $F_{x,t}$ はスペクトログラムモデルに当たる。なお分解スケールの任意性を防ぐために、

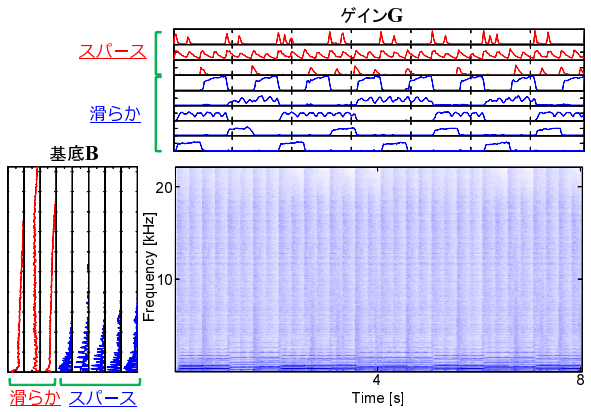

Fig. 1　調波音と打楽器音の混合信号に非負値行列分解を適用した例。調波音の基底・ゲインは青、打楽器音の基底・ゲインは赤でプロットした。これらを見てもわかる通り、それぞれの基底・ゲインは連続性とスパース性の性質を持っている。

$$\sum_{x} B_{x,k}^2 = 1 \tag{2}$$

とする。

一般的に非負値行列分解は観測とモデルの間の何らかの距離尺度を最小化する問題として解かれる。目的関数として Frobenius ノルムや Kullback-Leibler divergence などがよく用いられるが、いずれの場合も乗法更新アルゴリズム [1] と呼ばれる高効率なアルゴリズムにより、非負性の保証された解を得ることができる。調波音と打楽器音の混合信号に非負値行列分解を適用する場合、理想的には調波音と打楽器音を構成する基底系が得られ、これらの基底を調波音のクラス K_H と打楽器音のクラス K_P へ分類することにより、調波音・打楽器音分離が可能になると思われる。よってこのクラスタリングをどのように解くかが問題となってくる。

3　提案手法

非負値行列分解によるスペクトログラムの分解において、各基底を打楽器音や調波音に誘導するために、非負値行列分解の目的関数に制約項を加えることを考える。そこでスペクトルグラム上に現れる調波音と打楽器音の性質として、調波音と打楽器音のスペクトログラムの異方性に着目する。スペクトログラム上で調波音は周波数方向に急峻かつ時間方向に滑らかな成分を持ち、打楽器音は周波数方向に滑らかかつ時間方向に急峻な成分を持つ [3]。非負値行列分解において、基底は周波数方向の情報にあたり、ゲインは時間方向の情報にあたるため、先述の性質は、調波音の基底の周波数方向のスパース性とゲインの時間方向の連続性、打楽器音の基底の周波数方

*Harmonic and Percussive Source Separation based on Nonnegative Matrix Factorization with sparse and continuous constraints. by KITANO Yu, NAKANO Masahiro, ONO Nobutaka, SAGAYAMA Shigeki(The University of Tokyo)

向の連続性とゲインの時間方向のスパース性に対応していると考えられる (Fig. 1)。よってこれらの性質を制約として目的関数に加えることにより、基底クラスタリングを解きつつ、基底分解をすることが可能になると考えられる。今、$\theta \equiv \{\mathbf{B}, \mathbf{G}\}$ とし、スパース性を与える制約項には [5] に、連続性を与える制約項には [3] にならうことにすると、調波音らしさの制約項 $\mathcal{I}_H(\theta)$ と打楽器音らしさの制約項 $\mathcal{I}_P(\theta)$ は、それぞれ

$$\mathcal{I}_H(\theta) = \lambda_H \sum_{x,k \in K_H} |B_{x,k}|^{p_B} + \gamma_H \sum_{x,t,k \in K_H} B_{x,k} \left(\sqrt{G_{t,k}} - \sqrt{G_{t-1,k}} \right)^2 \quad (3)$$

$$\mathcal{I}_P(\theta) = \lambda_P \sum_{t,k \in K_P} |G_{t,k}|^{p_G} + \gamma_P \sum_{x,t,k \in K_P} G_{t,k} \left(\sqrt{B_{x,k}} - \sqrt{B_{x-1,k}} \right)^2 \quad (4)$$

と書ける。但し、$0 < p_B, p_G \leq 1$、$\gamma_H, \gamma_P, \lambda_H, \lambda_P \geq 0$ とする。これらの一項目は L^p ノルムで各々のパラメータをスパースにする効果があり、二項目は各軸方向の滑らかさのコストにあたる。今、非負値行列分解の距離尺度として、Kullback-Leibler divergence

$$\mathcal{J}(\theta) = \sum_{x,t} \left(Y_{x,t} \log \frac{Y_{x,t}}{F_{x,t}} - Y_{x,t} + F_{x,t} \right) \quad (5)$$

を考えると、解くべき問題は、(2) の条件下で観測 $Y_{x,t}$ から

$$\mathcal{F}(\theta) = \mathcal{J}(\theta) + \mathcal{I}_H(\theta) + \mathcal{I}_P(\theta) \quad (6)$$

を最小とする θ を推定する問題となる。この目的関数を直接最適化することは困難であるが、補助関数法を用いることにより反復的に目的関数を単調収束させるようなパラメータの更新則を導出することが可能である。アルゴリズム、更新式は紙面の都合上省略する。なお反復の度に (2) を満たすように規格化することとする。

4 シミュレーション実験

提案法の有効性を検証するために、シミュレーション実験を行った。実験に用いた観測信号には SiSEC[6] のデータベースから打楽器 (バスドラム、スネアドラム、ハイハット) と調波音 (ベース) により演奏された楽曲を用いた。時間周波数解析には短時間フーリエ変換 (サンプリング周波数 16kHz、フレーム長 64ms、フレームシフト 32ms、Hanning 窓) を用いた。総基底数を 9、打楽器音に割り当てる基底数は 4 とし、制約項への重みは $\lambda_H = \lambda_P = 0, \gamma_H = 0.005, \gamma_P = 1, p_B = p_G = 1$ として提案法を適用した。なおこれらのパラメータは実験的に決めた。反復回数を 200 回とし、$\mathbf{B}, \mathbf{G}$ の初期値は乱数によって与えた。提案法により基底分解した結果を Fig. 2 に示した。上の 4 つが打楽器音の割り当てに対応しており、打楽器の種類ごとに分解されていることが確認出来る。また, 調波音と打楽器音の分離性能として HPSS [3] と SNR を比較した結果を Table 1 に示した。なお調波音と打楽器音の分離は、推定された $\mathbf{B}, \mathbf{G}$ より各振幅スペクトルを求め、Wiener フィルタに基づく時間周波数マスクにより分離した。この結果から、HPSS に遜色のない性能であることが確認出来る。

Table 1　分離性能 (SNR[dB])

	Harmonic	Percussive	Avg.
mix	14.9	-14.0	0.4
HPSS [3]	14.3	-1.7	6.3
Proposed	16.0	-2.4	6.8

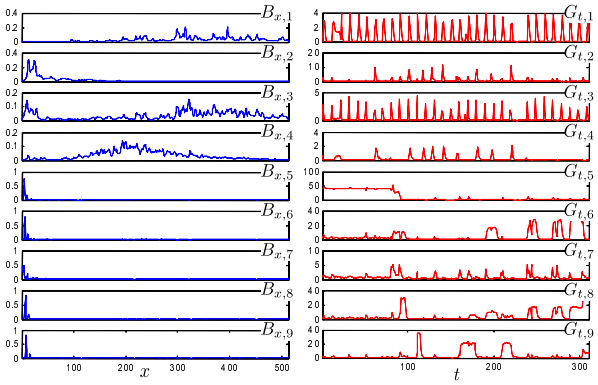

Fig. 2　学習された基底とゲイン。上の 4 つが打楽器音に対応している。

5 おわりに

本研究では、モノラル音楽音響信号から調波音と打楽器音を分離し、かつ基底分解する手法としてスペクトログラムの異方性に基づく制約付き非負値行列分解を提案した。今後は各基底が打楽器と調波音のいずれに当たるのかを隠れ状態として同時に推定する枠組みへの拡張を検討している。また本手法を [4] のように、基底を時間軸方向へ拡張したモデルも検討している。

参考文献

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization, "in *Proc. NIPS*, vol.13, pp.556–562, 2001.

[2] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," *Proc. EUSIPCO*, Sep. 2005.

[3] N. Ono *et. al.*, "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," *Proc. EUSIPCO*, Aug. 2008.

[4] P. Smaragdis, "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs, "In *Proc. ICA*, pp. 494–499, 2004.

[5] H. Kameoka *et. al.*, "Complex NMF: A New Sparse Representation for Acoustic Signals," In *Proc. ICASSP*, pp. 3437–3440, 2009.

[6] http://sisec.wiki.irisa.fr/tiki-index.php