7T-01

マルコフ連鎖に基づくマスク付き NMF を用いた特定音源の分離

日下湧太 [1], 糸山克寿 [1], 西田健次 [1], 中臺一博 [1,2]
1 東京工業大学工学院システム制御系
2（株）ホンダ・リサーチ・インスティチュート・ジャパン

1 はじめに

本稿では，マルコフ連鎖に基づくバイナリマスクを導入した非負値行列因子分解（NMF）により，音楽音響信号から特定の楽器音のみを分離する手法を提案する．一般に，NMF [1] による音源分離では，必ずしも基底と楽器とが一対一に対応しない．これを解決する手法として，楽器の教師音により基底を事前学習する NMF [2] が提案されているものの，教師音を準備する手間が大きいという問題がある．提案法では，楽器音の立上り（オンセット）情報の一部を指定したうえで，新たに導入したバイナリマスクを自動推定することにより，教師音なしでの特定楽器音分離を行う．予備的な実験を行い，オンセットを事前情報として与えることで，特定の楽器音のみが分離できることを確認した．

2 提案法

提案法では，ベータ過程 NMF (BP-NMF) [3] と同様にアクティベーションの ON/OFF を表現するバイナリマスクを導入する．振幅スペクトログラムを $\boldsymbol{X} \in \mathbb{N}_+^{F\times T}$ としたとき，BP-NMF は次のように表される．

$$\boldsymbol{X} \simeq \boldsymbol{W}(\boldsymbol{H} \odot \boldsymbol{S}) \tag{1}$$

ここで，$\boldsymbol{W} \in \mathbb{R}_+^{F\times K}$ は基底スペクトル，$\boldsymbol{H} \in \mathbb{R}_+^{K\times T}$ はアクティベーション，$\boldsymbol{S} \in \{0,1\}^{K\times T}$ はバイナリマスクである．また，周波数ビンを $f = 1,\cdots,F$，時間フレームを $t = 1,\cdots,T$，基底を $k = 1,\cdots,K$ で表す．

BP-NMF は次のような事前分布を導入することによりベイズモデルとして解釈することができる．

$$X_{fk} \sim \mathrm{Poisson}\left(\sum_{k=1}^{K} W_{fk}H_{kt}S_{kt}\right) \tag{2}$$

$$W_{fk} \sim \mathrm{Gamma}(a,b),\ H_{kt} \sim \mathrm{Gamma}(c,d) \tag{3}$$

ここで，a,b,c,d は基底スペクトル $\boldsymbol{W}$ とアクティベーション $\boldsymbol{H}$ の事前分布であるガンマ分布のハイパーパラメータである．

このように表される BP-NMF に対して，提案法ではバイナリマスク $\boldsymbol{S}$ の事前分布に変更を加え，新たにオンセットを表現する行列 $\boldsymbol{I}$ を導入する（Fig. 1）．

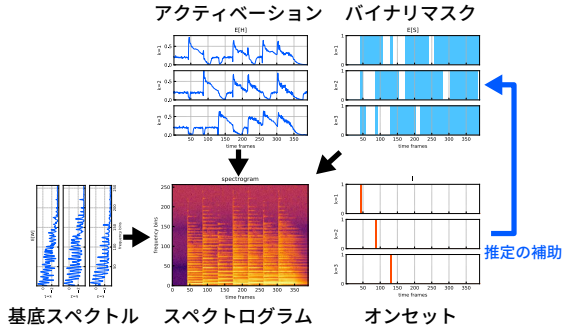


Fig. 1: 提案モデルの概略図

2.1 バイナリマスクの定式化

BP-NMF では，バイナリマスク $\boldsymbol{S}$ をベータ過程を用いてモデル化をしているが，提案法では楽器音は楽器の種類に応じたある程度の時間鳴り続けるという仮定のもと，各基底に対してアクティベーションの ON を表す状態 1 と OFF を表す状態 0 を遷移するマルコフ連鎖によってモデル化を行う．初期確率を ϕ，状態 1 から状態 1 への遷移確率を A_0，状態 0 から状態 1 への遷移確率を A_1 としたとき，$\boldsymbol{S}$ の同時確率は次のようにベルヌーイ分布の積の形で表される．

$$p(\boldsymbol{S}) = \prod_{k=1}^{K} p(S_{k1}) \prod_{t=2}^{T} p(S_{kt}|S_{kt-1}) \tag{4}$$

$$p(S_{k1}) = \mathrm{Bernoulli}(\phi) \tag{5}$$

$$p(S_{kt}|S_{kt-1}) = \mathrm{Bernoulli}(A_1)^{S_{kt-1}} \cdot \mathrm{Bernoulli}(A_0)^{1-S_{kt-1}} \tag{6}$$

2.2 オンセットの定式化

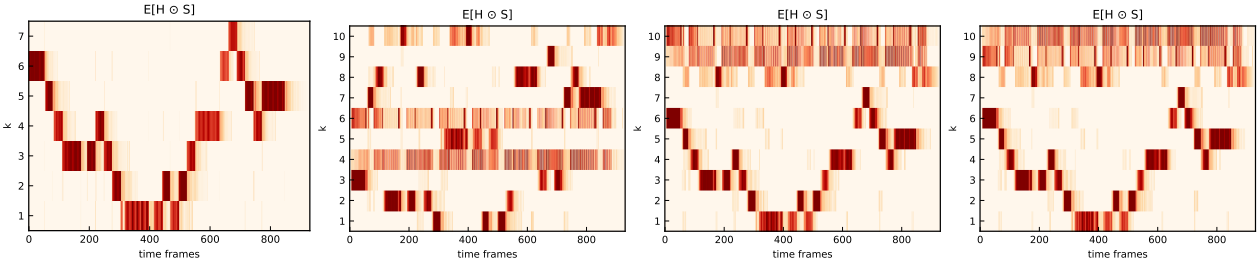
オンセット $\boldsymbol{I} \in \{0,1\}^{K\times T}$ は NMF のベイズモデルに含めず，後述のサンプリングの際に間接的に推定の補助として利用する．楽器音は 1 フレームで終了せず，一定フレームは持続するという仮定に基づき，指定された時間フレームを開始フレームとし，一定フレームの間 1 の値を取り続けるようにオンセット行列を設定する．

2.3 サンプリング

事後分布 $p(\boldsymbol{W},\boldsymbol{H},\boldsymbol{S}\,|\,\boldsymbol{X})$ を推定するために，ギブスサンプリングを用いる．最初に $\boldsymbol{W},\boldsymbol{H}$ を適当に初期化する．次に $\boldsymbol{S}$ の初期化を行うが，オンセットが与えられた基底はオンセット行列 $\boldsymbol{I}$ と対応するインデックスを 1，それ以外を 0 とし，オンセットが与えられていない基底は全て 0 と初期化することで，オンセッ

Sound Source Separation using Masked NMF based on Markov Chain Model

Yuta Kusaka[1], Katsutoshi Itoyama[1], Kenji Nishida[1], Kazuhiro Nakadai[1,2]
1 Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
2 Honda Research Institute Japan Co., Ltd.



(a) 事前学習　(b) オンセットなし　(c) オンセットあり（全て）　(d) オンセットあり（先頭のみ）

Fig. 2: サンプリング結果

トに対応する基底が誘起されるように誘導する．その後，$p(\boldsymbol{W} \mid \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{X})$, $p(\boldsymbol{H} \mid \boldsymbol{S}, \boldsymbol{W}, \boldsymbol{X})$, $p(\boldsymbol{S} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X})$ を BP-NMF のサンプリングを参考に順にサンプリングを繰り返すことで，事後分布を近似することができる．ただし，$\boldsymbol{S}$ をサンプリングするときは，オンセット行列 $\boldsymbol{I}$ の 1 に対応するインデックスの値は 1 で固定する．

3　評価実験

提案法によりオンセットを与えた基底が正しく分離されているかを確認するために，Dictionary Quality Evaluation (DQE) [4] を行った．DQE は，混合音から提案手法を用いて分離した基底と，事前に用意した分離済みの音源から得た基底の 2 種類のアクティベーションの相関係数を計算し，目的の音源が正しく分離できているかどうかを評価する方法である．RWC Music Database:Popular Music [5] の No.47 から約 10 秒間を切り出した音楽データ（サンプリングレート 22,050Hz に対して，フレーム長 512 サンプル，シフト幅 256 サンプル，窓関数をハニング窓とした STFT を行い，振幅スペクトルを得た後，定数倍と整数値への丸め込みを行った．これに対して調波打楽器分離 [6] を行ったものを入力として用いた．

用意した音楽データはメロディ，ギター，ピアノ，ベースの 4 種類の楽器から構成されており，メロディのオンセットを与えてメロディのみの分離を試みた．ハイパーパラメータを $a = b = 2, c = d = 1, \phi = 0.01, A_1 = 0.99, A_0 = 0.01$ とし，基底数 K はメロディの音高数 7 と，他の構成楽器数 3 の和である 10 とした．

以上のパラメータに対し，オンセットを与えない，オンセットを全て与える，オンセットを先頭のみ与える，の 3 つの場合でサンプリングを行う．それぞれの場合で得たメロディのアクティベーションとバイナリマスクの積 $\boldsymbol{H}_k \cdot \boldsymbol{S}_k$ $(k = 1, 2, \ldots, 7)$ (Fig. 2(b), 2(c), 2(d)) と，事前にメロディのみの音楽データから提案モデルにオンセットを与えずに学習しておいたアクティベーションとバイナリマスクの要素積 (Fig. 2(a)) の相関係数をプロットしたグラフを Fig. 3 に示す．

Fig. 3 は，基底 $k = 1, \cdots, 7$ における DQE の相関係数を箱ひげ図を用いてプロットしたものである．各基底の相関係数の値は，目的の基底が分離できている場合は 1 に近い値を取るため，オンセットを与えることにより目的の基底が分離できていることが確認できる．また，全てのオンセットを与える場合と比べてより厳しい条件であるオンセットを先頭のみ与えた場合でも，同等の分解性能が得られることが分かった．すなわち，複雑な楽曲に対して全てのオンセットを与えることが難しいような場合でも，提案法は頑健に動作することが期待される．

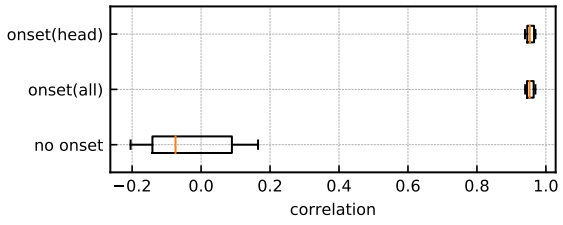


Fig. 3: 上から，オンセットを最初だけ与えた場合，全て与えた場合，与えなかった場合の相関係数

4　まとめ

本稿では，NMF にバイナリマスクを導入することにより，ユーザがオンセットを事前情報として与えることができるようになる手法を提案した．実際の音楽データに対して相関係数による評価を行うことで，特定の音源分離が可能であることが示された．

今後はマルコフ連鎖の状態に音楽的な構造を反映できるような拡張を行うことで，分離精度の向上を目指す．また，オンセットを事前に用意して提案モデルに入力したが，実際にはユーザが入力することになるため，ユーザが簡単にオンセットを作成できるインターフェースを設計する予定である．

謝辞　科研費 16H02884, 16K00294, 17K00365 および，JST ImPACT タフロボティクスチャレンジの支援を受けた．

参考文献

[1] Daniel D. Lee, et al. Algorithms for Non-negative Matrix Factorization. *Advances in neural information processing system*, pp. 556–562, 2001.
[2] Smaragdis, et al. Supervised and semi-supervised separation of sounds from single-channel mixtures. *International Conference on Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.
[3] Dawen Liang, et al. Beta process non-negative matrix factorization with stochastic structured mean-field variational inference. *arXiv*, Vol. 1411.1804, 2014.
[4] Dawen Liang, et al. Beta process sparse nonnegative matrix factorization for music. *ISMIR*, 2013.
[5] 後藤真孝ほか. 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. 情報処理学会論文誌, 第 45 巻, pp. 728–738, 2004.
[6] Derry Fitzgerald. Harmonic/percussive separation using median filtering. *International Conference on Digital Audio Effects (DAFx)*, pp. 1–4, 2010.