Vol.2019-MUS-124 No.14
2019/8/28

バイナリマスク付き非負値行列因子分解による発音時刻を用いた音源分離手法とその評価

日下 湧太[1]　糸山 克寿[1]　西田 健次[1]　中臺 一博[1,2]

概要：本稿では，バイナリマスク付き非負値行列因子分解による発音時刻を事前情報として利用する音源分離手法の提案とその評価を行う．複数の楽器により構成されるモノラル音響信号から特定の楽器音のみを分離する処理には，目的の楽器の事前情報を利用する手法が主流である．しかし，これらの手法で扱われる事前情報はユーザが準備することが困難なことや作成するには手間がかかることが多い．この問題を緩和するため，事前情報として目的楽器の発音時刻を利用して分離を行う音源分離手法を提案する．提案法は音高や音の終了時刻は利用せず，発音時刻のみを事前情報として利用する．そのためユーザが楽曲を聴取し，目的楽器の発音に合わせてボタンを押下することで容易に作成できる．このときユーザがすべてのタイミングで発音時刻を与えずとも音源分離を行うことができる．非負値行列因子分解を音響モデルとし，アクティベーションにマルコフ連鎖に基づくバイナリマスクを導入することで発音時刻を扱うことができるように拡張を行う．提案法の分離精度を評価するため，実楽曲データベースを用いてメロディ分離実験を行った．この結果，発音時刻が各基底に対して 1 つしか与えられない場合でも十分な音源分離精度を実現できることを確認した．

1. はじめに

現在多くの人が音楽を手軽に楽しめる状況になっており，複数の楽器から構成される楽曲から特定の楽器音のみを聴きたい，楽器練習に役立てたい，また，特定の楽器音のみを除去して音楽鑑賞をしたいなどといった需要が出てきている．このような需要を実現する技術として音源分離が盛んに研究されている．音源分離は音楽編集 [1] やカラオケ音源作成 [2]，自動採譜 [3] など様々な応用が期待されている．分離された単独の楽器音の音源は楽器分類手法 [4], [5] に利用したり，メロディを演奏する音源を分離すれば query-by-melody[6], [7] のような音楽検索システムに活用したりすることができる．そのため，楽曲から特定の音源のみを分離することは音楽情報処理において重要なトピックとなっている．

音源分離を行う手法は多く提案されており，非負値行列因子分解（non-negative matrix factorization; NMF）や独立成分分析は複数音源から構成される入力音響信号を複数の基底に分解する．しかし，分解された基底には楽器音との対応情報は含まれておらず，基底と楽器音は一対一対応しないため，特定の楽器音のみを分離するためには対応する適切な基底の組を選択しなくてはならない．一般に，この基底選択の操作は目的楽器の音源サンプルや楽譜といった事前情報を用いて行われるが，これらの事前情報は準備に手間がかかることが多い．

本稿では，事前情報として目的音源の発音時刻の一部を利用し，目的音源のみを分離する新たな音源分離手法を提案する．発音時刻はユーザが楽曲を聴取し，分離したい音源の発音に合わせてキーボードやスマートフォンのようなデバイスをタップすることで作成できる．発音時刻は既存手法で利用する事前情報である音源サンプルや楽譜が存在しない場合でも利用可能であり，複雑な作業をユーザに要求せず直感的に作成できる．

提案法は NMF に基づき音源分離を行うため，発音時刻は目的音源を構成する基底ごとに与えられる．このとき，発音時刻は全ての発音タイミングで完全に与えられる必要はなく，いくつかのタイミングで欠落している場合でも十分な精度で分離ができる．また，発音時刻と基底の対応関係も入力時に考慮する必要がなく，各基底の区別さえできていればよい．

このように与えられる目的音源の発音時刻を事前情報として NMF に基づく音源分離モデルに入力するできるようにするため，既存の NMF のアクティベーションに音源の

[1] 東京工業大学
Tokyo Institute of Technology
[2] （株）ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan

発音の有無を表現するバイナリマスクを導入する．発音時刻をバイナリマスクの変化とみなすことで音源分離モデルの推論時に補助的に利用する．モデル推論によって得られた結果から，発音時刻を与えた基底のみを用いて信号を復元することで目的音源を分離することができる．

2. 関連研究

事前情報を用いて特定の音源を分離する手法は大きく次の 2 種類に分けられる．1 つ目は事前情報として音源のスペクトル情報を利用する教師あり NMF[8], [9] である．教師あり NMF は目的音源のサンプルである教師音を用いて基底スペクトルを学習し，この基底スペクトルを固定して再度 NMF を行うことで目的音源の分離を行う．2 つ目は事前情報として音源の時間情報を利用するものである．このアプローチでは目的音源の楽譜が利用されることが多い [10], [11]．楽譜には時間情報である楽器の発音時刻と終了時刻に加えて，スペクトル情報である音高が含まれているため，高い精度の音源分離が期待される．

2 つ目のアプローチは高い分離精度が期待されるが，楽譜が存在しない場合や入手性が悪い場合は適用が難しい．1 つ目のアプローチも教師音を用意しなければならないが，教師音は入力音響信号中の目的音源のみが発音している区間を切り出すことで代用できる．しかし，このような区間が存在しない場合はユーザが教師音を用意することとなり，大きな負担となる．

そこで，これらの事前情報が用意できない場合でもユーザが自ら作成できる事前情報を用いて音源分離を行うアプローチの音源分離も研究されている．ユーザが作成する事前情報の例として，目的音源を真似た鼻歌 [12] や，スペクトログラム上につけた目的音源が存在する領域を表すアノテーション [13] などが挙げられる．しかし，これらの事前情報も作成するユーザの習熟が必要であったり，作成に非常に時間を要するといった問題点があるため，できるだけ容易に作成できる事前情報を利用して目的音源の分離を行いたいというモチベーションがある．

3. 提案法

本稿では，我々が新たに提案する目的楽器の発音時刻を事前情報として利用する音源分離手法について説明する．提案法の概略図を図 1 に示す．提案法の入力は複数の楽器から構成される音楽音響信号に短時間フーリエ変換（short-time Fourier transform; STFT）を適用して得られるパワースペクトログラムと目的楽器の発音時刻の一部であり，出力は目的楽器のスペクトログラムである．出力された目的楽器のスペクトログラムと対応する位相スペクトログラムに逆短時間フーリエ変換（inverse STFT; ISTFT）を適用することで目的楽器の音響信号を復元することができる．位相スペクトログラムは入力信号に STFT を適用す

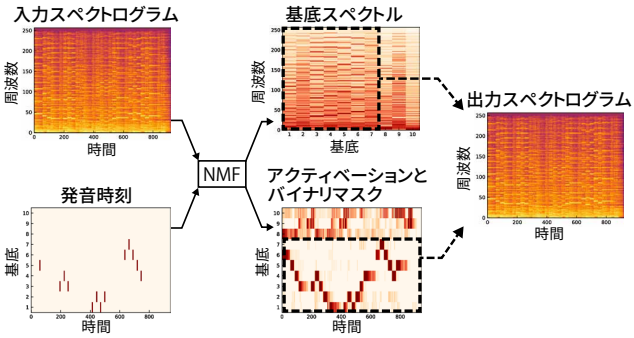


図 1: 提案法の概略図

ることで得られるものをそのまま用いてもよいし，出力された目的楽器のスペクトログラムから推定する手法 [14] を用いることで分離精度を改善することも可能である．

提案法は，モノラル信号の音源分離に広く利用されている手法である非負値行列因子分解（non-negative matrix factorization; NMF）[15], [16] に基づき分離を行う．NMF は全ての行列が非負である制約のもと，行列を 2 つの行列の積に分解するアルゴリズムである．NMF は画像処理の分野で提案された手法 [15] であるが，音楽情報処理の分野でも音楽音響信号に対する音源分離 [17] や自動採譜 [18] など様々なタスクに対して利用されている．

$\mathbb{R}_+$ を非負実数の集合とすると，周波数ビン $f = 1, 2, \ldots, F$，時間フレーム $t = 1, 2, \ldots, T$ のスペクトログラム $\boldsymbol{X} \in \mathbb{R}_+^{F \times T}$ が与えられたとき，NMF は次のようにスペクトログラムを分解する．

$$\boldsymbol{X} \simeq \boldsymbol{W}\boldsymbol{H} \tag{1}$$

ここで，$\boldsymbol{W} \in \mathbb{R}_+^{F \times K}$ はスペクトログラムに含まれるスペクトルパターンを表す基底スペクトル，$\boldsymbol{H} \in \mathbb{R}_+^{K \times T}$ は基底スペクトルの時間変化を表すアクティベーションである．$k = 1, 2, \ldots, K$ は基底スペクトルとアクティベーションの基底であり，基底数 K は予め与える必要がある．

提案法は NMF に目的音源の発音時刻を入力することで，特定の基底に目的音源を表す基底が出現するように誘導する．NMF においてアクティベーションが基底の時間情報を司るため，NMF に発音時刻を入力するためにはアクティベーションに発音時刻を扱える構造を導入すればよい．発音時刻は楽器の発音が発生する時間のみを持ち，発音の強さ（振幅）の情報は持たないためアクティベーションに直接与えることは困難である．そこで，提案法ではアクティベーションの存在，つまり音源が発音しているかしていないかを 1,0 の 2 値で表すバイナリマスクをアクティベーションと要素積をとる形で導入し，楽器の発音をバイナリマスクの 0 から 1 への変化と考えることで発音を時間のみの情報として扱うことができる．

アクティベーションにバイナリマスクを導入する NMF としてベータ過程 NMF（beta process sparse NMF; BP-

NMF）[19], [20] が提案されているため，BP-NMF と同じ形で提案法の定式化を行う．

$$\boldsymbol{X} \simeq \boldsymbol{W}(\boldsymbol{H} \odot \boldsymbol{S}) \tag{2}$$

ここで，$\boldsymbol{S} \in \{0,1\}^{K \times T}$ はバイナリマスクであり，アクティベーションと同サイズである．$\odot$ は行列の要素ごとの積を表す．BP-NMF は各変数に事前分布を与えることでモデルを階層ベイズモデルとして捉える．提案法では入力スペクトログラム $\boldsymbol{X}$，基底スペクトル $\boldsymbol{W}$，アクティベーション $\boldsymbol{H}$ には BP-NMF と同様の事前分布を与える．

$$X_{ft} = \sum_{k=1}^{K} Z_{ftk} \tag{3}$$

$$Z_{ftk} \sim \text{Poisson}(W_{fk} H_{kt} S_{kt}) \tag{4}$$

$$W_{fk} \sim \text{Gamma}(a, b) \tag{5}$$

$$H_{kt} \sim \text{Gamma}(c, d) \tag{6}$$

ここで，X_{ft}, W_{fk}, H_{kt} はそれぞれ $\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}$ の各要素，Z_{fkt} はモデル推論のために導入した補助変数 $\boldsymbol{Z} \in \mathbb{N}^{F \times T \times K}$ の各要素であり，a, b, c, d はそれぞれ事前分布のハイパーパラメータである．

提案法では BP-NMF と異なり，音楽的な仮定に基づきマルコフ連鎖による事前分布をバイナリマスクに設定する．また，発音時刻は行列の形で定義し，モデル推論の際に補助的に利用する．提案法は，入力スペクトログラムと発音時刻が観測されたときの基底スペクトル，アクティベーション，バイナリマスクの近似値をギブスサンプリングによって求める．得られた出力変数のうち，発音時刻を与えた基底には目的音源に対応する基底が出現しているため，これらの基底を用いることで目的音源が含まれるスペクトログラムを得ることができる．

3.1 バイナリマスク

バイナリマスクは，楽器音はその種類に基づき一定時間持続するという仮定に基づき，マルコフ連鎖を用いてモデル化する．楽器が発音しており，アクティベーションが大きな値をとるフレームではバイナリマスクの値が 1 となる（オン状態）．一方，楽器が発音しておらず，アクティベーションが 0 に近い値をとるフレームでは値が 0 となる（オフ状態）．$q_0, q_1 \in (0,1)$ をそれぞれオフ状態からオン状態，オン状態からオン状態に遷移する確率，$\phi \in (0,1)$ を初期確率とすると，バイナリマスクは図 2 のように基底 k の各時間フレーム S_{kt} が前のフレーム S_{kt-1} に依存して生成されるマルコフ連鎖として表すことができる．仮定より，ある時間フレームの状態は前の時間フレームと同じ状態になる確率が高いため，q_0 には 0 に近い値を，q_1 には 1 に近い値を設定する．また，最初の時間フレームでは楽器が発音していることは少ないという仮定に基づき，ϕ には 0 に近い値を設定する．

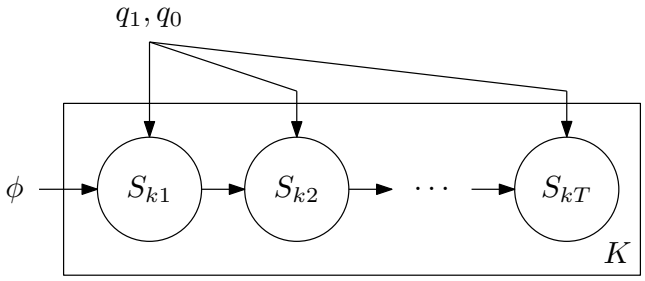


図 2: バイナリマスクの概略図

このように遷移確率と初期確率を設定すると，バイナリマスクの各基底 $\boldsymbol{S}_k$ の同時確率 $p(\boldsymbol{S}_k)$ は次のように表される．

$$p(\boldsymbol{S}_k) = p(S_{k1}) \prod_{t=2}^{T} p(S_{kt}|S_{kt-1}) \tag{7}$$

式（7）より，バイナリマスク全体の同時確率 $p(\boldsymbol{S})$ は次のように表される．

$$p(\boldsymbol{S}) = \prod_{k=1}^{K} p(\boldsymbol{S}_k) = \prod_{k=1}^{K} \left\{ p(S_{k1}) \prod_{t=2}^{T} p(S_{kt}|S_{kt-1}) \right\} \tag{8}$$

ここで，$p(S_{k1})$ はバイナリマスクの基底 k の最初の時間フレームが従う分布であり，バイナリマスクの各要素は 0 か 1 の 2 値をとるのでベルヌーイ分布とする．

$$p(S_{k1}) = \text{Bernoulli}(\phi) \tag{9}$$

同様に，$p(S_{kt}|Skt-1)$ はバイナリマスクの基底 k の時間フレーム t $(t \geq 2)$ が従う分布であり，ベルヌーイ分布の積の形で表すことができる．

$$p(S_{kt}|S_{kt-1}) = \text{Bernoulli}(q_1)^{S_{kt-1}} \cdot \text{Bernoulli}(q_0)^{1-S_{kt-1}} \tag{10}$$

3.2 発音時刻行列

与えられた発音時刻をモデル推論の際に扱いやすくするため，発音時刻行列 $\boldsymbol{O} \in \{0,1\}^{K \times T}$ を定義する．発音時刻行列のサイズはアクティベーション行列とバイナリマスクと同じであり，発音時刻行列の値が 1 であるフレームが発音の存在するフレームに対応する．発音時刻行列で値が 1 であるフレームと同じインデックスのバイナリマスクのフレームは，バイナリマスクのサンプリングを行う際にサンプリング結果にかかわらず 1 に固定することでバイナリマスクに発音時刻を与える．これにより，必ず発音が存在するフレームのバイナリマスクの値が 1 となるため，発音時刻を与えた音源が発音時刻を与えた基底に出現しやすくなる．

入力音響信号において，目的楽器が I 個の音高で発音していると仮定する．音高 $i = 1, 2, \ldots, I$ に対してユーザが作成した N_i 個の発音時刻の列 $\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,N_i}$ が与えら

れる．ここで，$\tau_{i,n}$ は音高 i の前から n 番目の発音時刻を表す．発音時刻が与えられた基底 $k = 1, 2, \ldots, I \leq K$ に対して，発音時刻 $\tau_{k,n}$ が存在するフレームを開始とし，N フレームの要素の値を 1 に，それ以外の要素の値を 0 に設定する．発音時刻を 1 フレームとすると，与えられた発音時刻にズレがある場合，誤ったフレームを参照し続けるため，誤った基底が分離されることがある．この問題を緩和するために発音時刻に幅をもたせている．今回 N は 1/8 拍の幅になるように設定した．また，発音時刻が与えられない基底 $k = I+1, I+2, \ldots, K$ に対しては，すべての要素の値を 0 に設定する．

3.3 モデル推論

提案モデルの出力変数（基底スペクトル，アクティベーション，バイナリマスク）を求めるため，これらの事後分布を計算する必要がある．事後分布はベイズの定理に基づき直接計算することは困難なため，ギブスサンプリングを用いて出力変数の期待値を近似的に求める．提案モデルのギブスサンプリングは BP-NMF のギブスサンプリング [20] と似た形で導出できる．提案法のサンプリングを行う条件付き分布は以下のとおりである．

$$\boldsymbol{Z}^{(i+1)} \sim p(\boldsymbol{Z} \mid \boldsymbol{W}^{(i)}, \boldsymbol{H}^{(i)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}) \tag{11}$$

$$\boldsymbol{W}^{(i+1)} \sim p(\boldsymbol{W} \mid \boldsymbol{Z}^{(i+1)}, \boldsymbol{H}^{(i)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}) \tag{12}$$

$$\boldsymbol{H}^{(i+1)} \sim p(\boldsymbol{H} \mid \boldsymbol{Z}^{(i+1)}, \boldsymbol{W}^{(i+1)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}) \tag{13}$$

$$\boldsymbol{S}^{(i+1)} \sim p(\boldsymbol{S} \mid \boldsymbol{Z}^{(i+1)}, \boldsymbol{W}^{(i+1)}, \boldsymbol{H}^{(i+1)}, \boldsymbol{X}) \tag{14}$$

3.3.1 $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{H}$ のサンプリング

補助変数 $\boldsymbol{Z}$ のサンプリングを行う条件付き分布は BP-NMF と同様に多項分布の形で表される．$\boldsymbol{Z}$ を構成するベクトル $\boldsymbol{Z}_{ft} = (Z_{ft1}, \ldots, Z_{ftK})$ は次の多項分布に従う．

$$\boldsymbol{Z}_{ft} \sim \mathrm{Mult}(X_{ft}, \phi_{ftk}) \tag{15}$$

ここで $\phi_{ftk} = \frac{W_{fk}H_{kt}S_{kt}}{\sum_l W_{fl}H_{lt}S_{lt}}$ である．サンプリングを行う際には $\boldsymbol{Z}_{ft}$ の条件付き期待値 $X_{ft}\phi_{ftk}$ をサンプリング結果として利用することができる．

基底スペクトル $\boldsymbol{W}$ とアクティベーション $\boldsymbol{H}$ のサンプリングを行う条件付き分布も BP-NMF と同様にガンマ分布の形で表される．

$$W_{fk} \sim \mathrm{Gamma}\left(a + \sum_{t=1}^{T} X_{ft}\phi_{ftk}, b + \sum_{t=1}^{T} H_{kt}S_{kt}\right) \tag{16}$$

$$H_{kt} \sim \mathrm{Gamma}\left(c + \sum_{f=1}^{F} X_{ft}\phi_{ftk}, d + S_{kt}\sum_{f=1}^{F} W_{kt}\right) \tag{17}$$

3.3.2 $\boldsymbol{S}$ のサンプリング

バイナリマスク $\boldsymbol{S}$ のサンプリングを行う条件付き分布は BP-NMF と似た形で導出できる．S_{kt} は 2 値をとるため，

Algorithm 1 提案モデルのギブスサンプリング

```
Initialize W, H
Initialize S
for i = 1, 2, ··· do
    Calculate φ_ftk = W_fk H_kt S_kt / Σ_l W_fl H_lt S_lt
    Sample W following Eq. 16
    Sample H following Eq. 17
    Sample S following Eq. 18, 24, 25
end for
```

ベルヌーイ分布に従う．

$$S_{kt} \sim \mathrm{Bernoulli}\left(\frac{P_1}{P_1 + P_0}\right) \tag{18}$$

ここで P_1, P_0 はそれぞれ以下のように表される．

$$P_1 = p(S_{kt} = 1 \mid S_{\neg kt}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{X}) \tag{19}$$

$$P_0 = p(S_{kt} = 0 \mid S_{\neg kt}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{X}) \tag{20}$$

式（19）を計算すると以下のようになる．

$$\begin{aligned} P_1 &= p(S_{kt} = 1 \mid S_{\neg kt}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{X}) \\ &\propto p(S_{kt} = 1)p(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{kt} = 1, S_{\neg kt}) \end{aligned} \tag{21}$$

ここで，第 1 項は次のように表される．

$$p(S_{kt} = 1) = \begin{cases} \phi & t = 1 \\ q_1^{S_{kt-1}} q_0^{1-S_{kt-1}} & t \geq 2 \end{cases} \tag{22}$$

また，第 2 項は次のように表される．

$$\begin{aligned} &p(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{kt} = 1, S_{\neg kt}) \\ &\propto \prod_{f=1}^{F} (X_{ft}^{\neg k} + W_{fk}H_{kt})^{X_{ft}} \exp\left(-W_{fk}H_{kt}\right) \triangleq P_{ftk}^{S=1} \end{aligned} \tag{23}$$

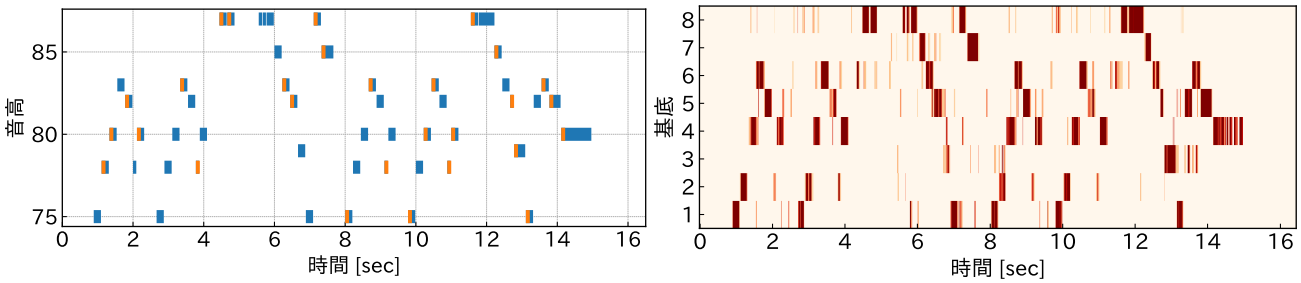
ここで $X_{ft}^{\neg k} = \sum_{l \neq k} W_{fl}H_{lt}$ である．以上をまとめると P_1 は次のように表される．

$$P_1 = \begin{cases} \phi P_{ftk}^{S=1} & t = 1 \\ q_1^{S_{kt-1}} q_0^{1-S_{kt-1}} P_{ftk}^{S=1} & t \geq 2 \end{cases} \tag{24}$$

同様に P_0 は次のように表される．

$$P_0 = \begin{cases} (1-\phi) P_{ftk}^{S=0} & t = 1 \\ (1 - q_1^{S_{kt-1}})(1 - q_0^{1-S_{kt-1}}) P_{ftk}^{S=0} & t \geq 2 \end{cases} \tag{25}$$

ここで $P_{ftk}^{S=0} \triangleq \prod_{f=1}^{F} (X_{ft}^{\neg k})^{X_{ft}}$ である．$t = 1$ から順に時間方向へ式（24），（25）を計算し，式（18）からサンプリングすることで $\boldsymbol{S}$ 全体をサンプリングすることができる．このとき，3.2 節で述べたように，発音時刻行列を考慮してサンプリングを行う．



(a) メロディの楽譜（青）と与えた発音時刻（橙）．
発音時刻は各音高に対して 50%与えられる

(b) アクティベーションとバイナリマスクの要素積．
発音時刻を与えた 8 つの基底のみを表示している．

図 3: 予備実験の入力と出力結果

3.3.3 提案モデルのギブスサンプリングアルゴリズム

Algorithm 1 に以上をまとめた提案モデルのギブスサンプリングアルゴリズムを示す．まず最初に各変数の初期化を行う．$\boldsymbol{W}$ と $\boldsymbol{H}$ は事前分布であるガンマ分布に従いランダムに初期化を行う．$\boldsymbol{S}$ については，発音時刻が与えられた基底は発音時刻行列 $\boldsymbol{O}$ の対応する基底と同じ形で初期化を行う．また，発音時刻が与えられない基底はすべて 1 の値で初期化を行う．このように初期化を行うことで，発音時刻が与えられていない基底に目的音源以外の音源が現れるように誘導することができる．初期化を行った後，それぞれの変数の条件付き分布に従って繰り返しサンプリングを行うことでサンプル列を生成し，サンプル列の期待値を計算することで出力とする．

4. 評価実験

最初に提案法が実際に音源分離を行うことができるかどうか確認するために予備実験を行った．動作確認を行った後に，複数の音源分離指標を用いて提案法の音源分離精度の評価を行った．

4.1 予備実験

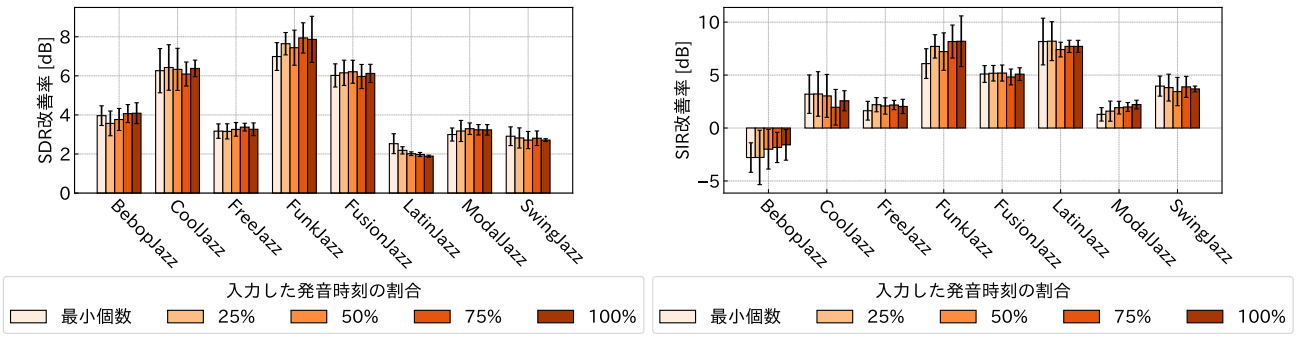
RWC Music Database: Popular Music [21] の No.1 を用いて提案法の動作確認を行った．この楽曲のサビに対応する 8 小節の MIDI ファイルを切り出し，WAV ファイルに変換したものを入力音響信号とした．サビはフルートによって演奏されるメロディと伴奏によって構成されており，メロディに対して発音時刻を与えてこれを分離する（図 3(a)）．WAV ファイルのサンプリング周波数は 22,050 Hz であり，予め調波・打楽器音分離 [22] を適用することで打楽器音を除去する．これに窓幅 512 サンプル，オーバーラップ 256 サンプル，窓関数をハン窓とする短時間フーリエ変換を適用することで得られた周波数ビン数 257，時間フレーム数 1417 のパワースペクトログラムを提案モデルに入力する．提案モデルのハイパーパラメータは $a = 0.5, b = 1.0, c = 3.0, d = 3.0, \phi = 0.01, q_0 = 0.01, q_1 = 0.99$ とし，基底数 K は十分大きい値として 30 に設定した．イテレーション 400 回のギブスサンプリングにより得られたサンプル列に対してバーンイン 200 サンプルを設けたサンプル列の期待値を計算することで各変数の出力結果とした．

図 3(b) に予備実験のよって得られたアクティベーションとバイナリマスクの要素積のヒートマップを示す．目的音源であるメロディの音高数は 8 であり，図 3(b) には発音時刻を与えた 8 つの基底のみを示している．ヒートマップの色の濃い部分が値が大きい，つまり対応する基底が大きな値をもつフレームを表しており，図 3(a) に示すメロディの楽譜と同じパターンが現れていることが確認できる．また，発音時刻を与えなかった基底 $k = 9, \ldots, 30$ にはメロディ以外の音源を表す基底が現れていることが確認できた．

4.2 音源分離精度評価実験

提案法で与える発音時刻の割合を変化させたときの音源分離精度の変化を調べる実験を行った．使用した音源は音源分離評価用の多重音トラックデータセットである MedleyDB[23] に含まれるアーティストが MusicDelta の楽曲のうち，ジャンルがジャズである 8 曲（BebopJazz, CoolJazz, FreeJazz, FunkJazz, FusionJazz, LatinJazz, ModalJazz, SwingJazz）の冒頭 30 秒を切り出したものである．MedleyDB にはメロディの F0 とこれに対応する楽器を表すアノテーションが含まれており，F0 から発音時刻を作成し提案法に与えることでメロディの分離を行う．入力する発音時刻の割合を各音高に対して 1 個，25%，50%，75%，100%と変化させ，出力された音源の分離精度を評価する．音源のサンプリングレートは 22050 Hz にダウンサンプリングし，4.1 節と同様の STFT を適用することでパワースペクトログラムを得る．ハイパーパラメータとギブスサンプリングのパラメータも 4.1 節の実験と同じに設定した．

音源分離精度の評価には source to distortion ratio (SDR) と source to inference ratio (SIR)[24] の改善率を使用する．SDR は分離信号と分離により発生する全てのノイズの比，SIR は分離信号と他の音源からの干渉によるノイズの比によって定義され，どちらの指標も値が大きければ分離精度が良いことを表す．SDR と SIR は `BSS Eval` toolbox[25]



(a) SDR　　**(b)** SIR

図 4: 入力する発音時刻の割合を変化させたときの音源分離精度の変化．エラーバーは標準偏差を表す．

を用いて計算した．

音源分離精度の評価には提案法の出力変数（基底スペクトル・アクティベーション・バイナリマスク）から目的音源のみが含まれる信号を復元する必要があるが，これはウィナーフィルタにより実現できる．発音時刻を与えた基底を取り出すことで得られる出力変数をそれぞれ $\boldsymbol{W}_{\text{target}}, \boldsymbol{H}_{\text{target}}, \boldsymbol{S}_{\text{target}}$ とすると，目的音源のパワースペクトログラム $\boldsymbol{X}_{\text{target}}$ はウィナーフィルタにより次のように求めることができる．

$$\boldsymbol{X}_{\text{target}} = \frac{\boldsymbol{W}_{\text{target}}(\boldsymbol{H}_{\text{target}} \odot \boldsymbol{S}_{\text{target}})}{\boldsymbol{W}(\boldsymbol{H} \odot \boldsymbol{S})} \odot \boldsymbol{X} \qquad (26)$$

その後，目的音源のパワースペクトログラムと元の位相スペクトログラムに対して ISTFT を適用することで目的音源の信号を得ることができる．

図 4 に入力する発音時刻の割合を変化させたときの評価結果を示す．グラフの各グループは左から順に発音時刻を最小個数，25%，50%，75%，100%与えたときの結果を示しており，基底スペクトル・アクティベーション・バイナリマスクの初期値を変えて試行を 10 回行ったときの平均と標準偏差を表す．図 4 (a) に示す SDR を見ると，全ての楽曲で入力発音時刻の割合によらず SDR が改善しており，複数の楽曲で入力発音時刻の割合が増加するにつれ標準偏差が減少する傾向があることが確認できる．これは，入力発音時刻の割合が増えるとサンプリング時に参照する目的音源が存在するフレームが増加し，目的音源を表す基底が現れやすくなるためと考えられる．一方，入力発音時刻の割合が増加すると，FunkJazz では標準偏差が増大し，LatinJazz や SwingJazz では平均値が減少していることが確認できる．これは，入力発音時刻の割合が増加すると，サンプリング時に参照されるフレームが増加し，過学習のような状態になっていると考えられる．

図 4 (b) に示す SIR を見ると，発音時刻の割合が増加すると標準偏差が減少する傾向があることが確認できる．しかし，BebopJazz において SIR が改善せず，劣化している．BebopJazz のメロディを演奏するトランペットの発音時刻の基底数は 13 であるため，入力発音時刻の割合を 100%で固定し，提案モデル全体の基底数を 14 から 50 の間で変化させて SIR の改善率の変化を調べたところ，基底数 20 付近で改善率が最大となり，その後基底数が増加するにつれ改善率が減少する傾向にあることが確認できた．基底数 50 のときのアクティベーションを確認したところ，発音時刻を与えていない基底 $k = 14, \cdots, 50$ のほとんどが非スパースになっていた．これは，NMF において基底数を増やすと，基底によるスペクトルパターンの表現の幅が増えるためである．これらの基底には目的音源を表現する基底も含まれているため，ウィナーフィルタを適用すると伴奏音源の方にもトランペットが出現してしまい，その結果分離精度が劣化すると考えられる．そのため，提案法においてよい分離結果を得るためには適切な基底数を選択する必要があるが，これは今後の課題とする．

5. おわりに

本稿では目的音源の発音時刻の一部を事前情報として利用することで目的音源のみを分離する新たな手法を提案した．マルコフ連鎖に基づくバイナリマスクをアクティベーションに導入することで，NMF による音源分離のフレームワークで発音時刻を利用できるように拡張した．また，新たに導入したバイナリマスクと発音時刻を含めた提案モデルの推論を行うためにギブスサンプリングアルゴリズムの改善を行った．提案法の音源分離精度を評価するための実験を行い，全ての発音時刻が与えられずとも，目的音源の分離が可能であることを示した．

提案法では発音時刻は音高別に与えられるという設定になっているため，実際にユーザが発音時刻を作成するとなると，全てのタイミングで与える必要はないという特徴があるとしても難しい作業となると思われる．この問題は，基底スペクトルとアクティベーションの時間方向シフトの畳み込みの形で分解を行う手法 [26] を利用し，楽器ごとのアクティベーションを求めるアプローチをとることで発音時刻の音高を考慮しなくてもよくなると考えられる．また，現在は発音時刻をモデル推論の補助として利用してい

るが，発音時刻にも事前分布を導入し確率モデルに含めて事後分布の推定を行うことで，発音時刻がずれて与えられた場合に正しい位置を推定し，分離精度の改善を行うことも考えている．

参考文献

[1] Yoshii, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Drumix: An audio player with real-time drum-part rearrangement functions for active music listening, *Information and Media Technologies*, Vol. 2, No. 2, pp. 601–611 (2007).

[2] Simpson, A. J., Roma, G. and Plumbley, M. D.: Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network, *LVA/ICA*, pp. 429–436 (2015).

[3] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. and Klapuri, A.: Automatic music transcription: challenges and future directions, *Journal of Intelligent Information Systems*, Vol. 41, No. 3, pp. 407–434 (2013).

[4] Eronen, A. and Klapuri, A.: Musical instrument recognition using cepstral coefficients and temporal features, *ICASSP*, Vol. 2, pp. II753–II756 (2000).

[5] Kostek, B.: Musical instrument classification and duet analysis employing music information retrieval techniques, *Proceedings of the IEEE*, Vol. 92, No. 4, pp. 712–729 (2004).

[6] Marolt, M.: A mid-level representation for melody-based retrieval in audio collections, *IEEE Transactions on Multimedia*, Vol. 10, No. 8, pp. 1617–1625 (2008).

[7] Shalev-Shwartz, S., Dubnov, S., Friedman, N. and Singer, Y.: Robust temporal and spectral modeling for query by melody, *SIGIR*, pp. 331–338 (2002).

[8] Kitamura, D., Saruwatari, H., Yagi, K., Shikano, K., Takahashi, Y. and Kondo, K.: Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing, *ISSPIT*, pp. 392–397 (2013).

[9] Kitamura, D., Saruwatari, H., Shikano, K., Kondo, K. and Takahashi, Y.: Music signal separation by supervised nonnegative matrix factorization with basis deformation, *DSP*, pp. 1–6 (2013).

[10] Ewert, S. and Müller, M.: Using score-informed constraints for NMF-based source separation, *ICASSP*, pp. 129–132 (2012).

[11] Fritsch, J. and Plumbley, M. D.: Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis, *ICASSP*, pp. 888–891 (2013).

[12] Smaragdis, P. and Mysore, G. J.: Separation by "humming": User-guided sound extraction from monophonic mixtures, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 69–72 (2009).

[13] Lefèvre, A., Bach, F. and Févotte, C.: Semi-supervised NMF with time-frequency annotations for single-channel source separation, *ISMIR* (2012).

[14] Griffin, D. and Lim, J.: Signal estimation from modified short-time Fourier transform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 2, pp. 236–243 (1984).

[15] Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, *Nature*, Vol. 401, No. 6755, pp. 788–791 (1999).

[16] Févotte, C., Bertin, N. and Durrieu, J.-L.: Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis, *Neural Computation*, Vol. 21, No. 3, pp. 793–830 (2009).

[17] Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 1066–1074 (2007).

[18] Smaragdis, P. and Brown, J. C.: Non-negative matrix factorization for polyphonic music transcription, Vol. 3, No. 3, pp. 177–180 (2003).

[19] Liang, D., Hoffman, M. D. and Ellis, D. P.: Beta Process Sparse Nonnegative Matrix Factorization for Music., *ISMIR*, pp. 375–380 (2013).

[20] Liang, D. and Hoffman, M. D.: Beta process nonnegative matrix factorization with stochastic structured mean-field variational inference, *arXiv:1411.1804* (2014).

[21] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases., *ISMIR*, Vol. 2, pp. 287–288 (2002).

[22] Fitzgerald, D.: Harmonic/Percussive Separation Using Median Filtering, *DAFx*, pp. 1–4 (2010).

[23] Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C. and Bello, J. P.: MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research, *ISMIR* (2014).

[24] Vincent, E., Gribonval, R. and Févotte, C.: Performance measurement in blind audio source separation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462–1469 (2006).

[25] Stöter, F.-R., Liutkus, A. and Ito, N.: The 2018 signal separation evaluation campaign, *LVA/ICA*, pp. 293–305 (2018).

[26] Schmidt, M. N. and Mørup, M.: Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation, *Independent Component Analysis and Blind Signal Separation*, pp. 700–707 (2006).