バイナリマスク付き非負値行列因子分解に基づく発音時刻を用いた音源分離

Onset-informed Source Separation using Non-negative Matrix Factorization with Binary Masks

日下 湧太[1*]　糸山 克寿[1]　西田 健次[1]　中臺 一博[1, 2]
Yuta Kusaka[1], Katsutoshi Itoyama[1], Kenji Nishida[1], Kazuhiro Nakadai[1,2]

[1] 東京工業大学
[1] 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
[2] Honda Research Insitute Japan Co.,Ltd.

Abstract: 本稿では，バイナリマスク付き非負値行列因子分解に基づき，目的音源の発音時刻を補足情報として利用する新しい音源分離手法について説明する．複数の楽器音により構成される混合音から特定の楽器音のみを分離するタスクにおいて，多くの既存手法は作成に時間と手間がかかる補足情報を利用していた．提案法では，楽曲を聴取しながらデバイスをタッピングするだけで容易に作成が可能な補足情報として，発音時刻を利用して音源分離を行う．NMF ベースの音源分離パイプラインに発音時刻を組み込むために，楽器音のオンとオフを制御するバイナリマスクを導入する．バイナリマスクは楽器音の連続性に関する仮定に基づき，マルコフ連鎖を用いてモデル化する．発音時刻はバイナリマスクがオンからオフに変化する時刻として扱う．提案モデルはギブスサンプリングによって推定され，推定時に発音時刻を活用することで効率的に目的音源を推定できる．提案法を用いて音楽音響信号からメロディを演奏する楽器音を分離する実験において，分離音と残留ノイズ比による評価で，2 から 10 dB の改善が確認できた．さらに，入力発音時刻の一部が欠落した場合や，時間方向のずれを含む場合の分離精度を評価し，提案法の頑健性を検証した．

1　はじめに

複数の楽器を含む音楽音響信号から目的の楽器音のみを分離する音源分離技術は，重要なトピックとして長年研究されている．分離によって得られる楽器音は，楽器練習や楽曲のリミキシングに有用であり，楽曲編集 [1]，カラオケ音源作成 [2]，自動採譜 [3] や楽器判別 [4, 5] といった音楽情報処理システムの改善にも活用できる．さらに，楽曲から分離したメロディラインの信号は，音楽検索システム [6, 7] のようなシステムにも利用可能である．

音源分離には非負値行列因子分解（non-negative matrix factorization; NMF）[8, 9] や独立成分分析（independent component analysis; ICA）[10] が提案されており，そのなかでも NMF はモノラル音響信号に対して有効な音源分離手法として長年研究されている．音楽音響信号に NMF を適用すると，信号に含まれる楽器音に対応する複数の基底に分解することができる．NMF によって混合音から目的の楽器音を分離するには，分解された基底から目的楽器に対応する基底の集合を選択する必要がある．しかし，NMF によって得られた基底と楽器音は基本的に一対一対応しないため，大量の基底から目的楽器の対応する基底を全て選択する操作は現実的には難しい．

NMF のような音源分離手法に分離したい音源に関する補足情報を入力することで，分離を補助したり分離精度を向上させたりするアプローチを informed source separation (ISS) [11] と呼ぶ．ISS で利用される情報の例として，目的楽器の音色のようなスペクトル的情報や，楽譜のような時間的情報などが挙げられる．ISS は楽器音や歌声分離に対して強力なアプローチであるが，補足情報の入手可能性の問題などにより適用不可能な場面も多い．ユーザが作成可能な補足情報を利用して

*連絡先： 東京工業大学
152-8552 東京都目黒区大岡山 2-12-1
E-mail: kusaka@ra.sc.e.titech.ac.jp
本稿は DAFx2021 で採択された”ONSET-INFORMED SOURCE SEPARATION USING NON-NEGATIVE MATRIX FACTORIZATION WITH BINARY MASKS”を和訳したものである

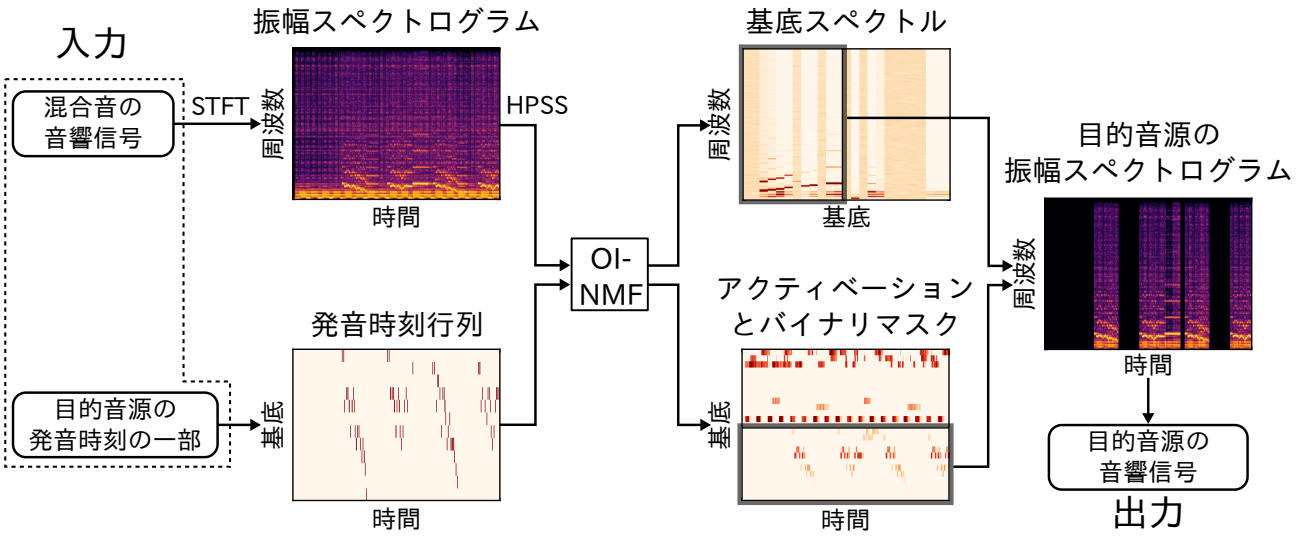


図 1: OI-NMF による目的音源分離の流れ．入力は混合音の音響信号と分離したい音源の発音時刻（の一部）である．入力は振幅スペクトログラムと発音時刻行列に変換され，OI-NMF によって基底スペクトル，アクティベーション，バイナリマスクに分解される．発音時刻を含めて推定を行った基底（図灰色の枠内の基底）が目的音源に対応しており，これらの基底を用いて信号を復元することで目的音源の音響信号を得ることができる．

音源分離を行うユーザガイドなアプローチ [12, 13] も提案されているが，情報の作成にはユーザの技量や手間が要求される．

そこで，本研究では分離したい楽器音の発音時刻を利用することでモノラルの音楽音響信号から目的の楽器音を分離可能な，新しい ISS 手法である *onset-informed NMF* (OI-NMF) を提案する．発音時刻は，ユーザが楽曲を聴取しながら分離したい楽器音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ピングすることで容易に作成可能である．OI-NMF は既存の NMF ベースの音源分離モデルを，発音時刻を補足情報として入力できるよう拡張したモデルである．OI-NMF の特徴として，発音時刻は全てのタイミングで与えられる必要はなく，一部が欠落したものでも分離に利用できる点が挙げられる．これにより，既存の ISS 手法で利用されていた準備が難しい情報に比べ，簡単に作成できる情報に基づいて目的楽器音を分離できる．本研究の貢献を以下に示す．

- 既存の NMF モデルを拡張し，発音時刻を補足情報として入力できる OI-NMF モデルを開発した．NMF の変数として楽器音のオン/オフを表現するバイナリマスクを導入し，発音時刻をマスクがオンからオフに変化する時刻として扱う．OI-NMF のモデルを確率モデルとして定義し，バイナリマスクと発音時刻を含めてベイズ則によりモデルを推論することで目的の楽器音を効率的に推定することができる．
- OI-NMF を実装し，実楽曲から目的の楽器音を分離する実験とその分離精度を評価した．発音時刻は楽曲データセットに含まれる F0 アノテーションから作成したものを用いた．提案法と発音時刻を利用しないベースライン手法を比較し，OI-NMF と発音時刻の有効性を確認した．さらに，発音時刻の一部が欠落していたり，時間方向にずれを含んだりする場合の分離の頑健性を検証する実験を行った．

2 関連研究

OI-NNF は分離したい楽器の発音時刻を補足情報として利用するため，ISS の一種に含められる．基本的な ISS は利用する補足情報の種類によって次のように大別することができる．

- 目的音源のスペクトル的情報を利用するアプローチ．目的音源が楽器の場合，その音色や調波構造などが利用される．教師あり NMF [14, 15] は，分離したい楽器音を表す基底スペクトルを事前に用意した音源から学習して分離に利用する．用意した音源と目的楽器音が完全に一致しない場合に分離精度が劣化する問題点があるが，スペクトルに関する制約を加えることで精度劣化を抑えている．
- 目的音源の時間的情報を利用するアプローチ．OI-NMF で利用する発音時刻もこちらに該当する．音楽音響信号に関する典型的な時間情報に楽譜が

挙げられる．楽譜は楽器音の発音時刻，消音時刻という時間情報に加え，楽音の音高というスペクトル情報も持ち，これを利用する score-informed NMF [16, 17] は高精度な分離を実現している．また，近年盛んに研究されている深層学習を用いた手法 [18, 19] も，目的音源のみを含むスペクトログラムを教師としてモデルを学習するため，このアプローチに含めることができる．

これらの補足情報は分離に有効であるが，クリーンな目的音源の信号や楽譜は準備に手間がかかることやそもそも存在しないことがある．また，深層学習ベースの手法も適切な学習データを大量に用意する必要がある．そのため，これらの手法を実際に適用することは難しい場面も多く存在する．

この問題を解決するため，ユーザが楽曲を聴取して作成できるような補足情報を利用して分離するアプローチも提案されている．例えば，分離したい音源を真似た鼻歌 [12] や，スペクトログラム上の目的音源に対応する領域につけたアノテーション [13] などを分離に利用する．これらの情報は楽譜等の情報の準備が難しい楽曲に対しても適用可能である一方，その作成にはユーザの技能や時間を要求する．OI-NMF で利用する発音時刻は，これらの情報に比べて簡単に作成可能である．

また，発音時刻と類似した時間的情報を利用する手法として，非負値テンソル因子分解に基づきユーザが作成した楽器音の存在区間アノテーション [20] を利用する音源分離手法も提案されている．OI-NMF はこの手法と比較すると，モノラル音響信号にも適用可能である点や，存在区間で必要とされる消音時間を利用しないため情報作成が簡単という点で優れている．

3 非負値行列因子分解

非負値行列因子分解（non-negative matrix factorization; NMF） [8, 9] はモノラル音響信号の分離に有効なアルゴリズムである．もとは画像処理分野で提唱された [21] 手法であるが，音源分離 [22, 23] や自動採譜 [24] といった音声分野への応用も研究されている．音源分離における NMF は，入力である混合音の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振幅スペクトログラムを，音響信号の低ランク性に基づき 2 つの非負行列に分解する．

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

ここで，$\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{F \times T}$ は振幅スペクトログラム（$\mathbb{R}_{\geq 0}$ は非負実数全体の集合），$f \in \{1, 2, \ldots, F\}$ は周波数ビン，$t \in \{1, 2, \ldots, T\}$ は時間フレームである．NMF の出力である $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{F \times K}$ は基底スペクトルと呼ばれ，振幅スペクトルに含まれる代表的なスペクトルパターンの基底 $k \in \{1, 2, \ldots, K\}$ から構成される行列である．もう一方の出力 $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{K \times T}$ はアクティベーションと呼ばれ，対応する基底の時間変化を表す行列である．K は基底数と呼ばれ，分解の基底の数を決めるハイパーパラメータである．

NMF によって得られた基底を用いて目的音源を分離するには，まず目的音源に対応する基底の組から選択して目的音源を表す基底スペクトル $\boldsymbol{W}_{\text{target}}$ とアクティベーション $\boldsymbol{H}_{\text{target}}$ を構成する．これらに対しウィナーフィルタを適用することで目的音源の振幅スペクトログラム $\boldsymbol{X}_{\text{target}}$ を得る．

$$\boldsymbol{X}_{\text{target}} = \frac{\boldsymbol{W}_{\text{target}}\boldsymbol{H}_{\text{target}}}{\boldsymbol{W}\boldsymbol{H}} \odot \boldsymbol{X} \tag{2}$$

ただし，$\odot$ は行列の要素ごとの積を表す．最後に，$\boldsymbol{X}_{\text{target}}$ 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音源の音響信号を復元できる．ここで利用する位相スペクトログラムは，入力信号に STFT を適用して得られるもので十分であり，$\boldsymbol{X}_{\text{target}}$ から推定されたもの [25] を用いることで分離精度を向上させることも可能である．

4 Onset-informed NMF

本節では，OI-NMF とこれを用いた目的音源の分離について説明する．図 1 に OI-NMF による音源分離の流れを示す．入力は混合音の音響信号に STFT を適用して得られた振幅スペクトログラムと，分離したい音源の発音時刻の一部，出力は目的音源の振幅スペクトログラムである．発音時刻は発音時刻行列という時間周波数領域の行列に変換される．発音時刻行列を含めて OI-NMF モデルの推論を行うことで，分離したい音源が入力発音時刻から続くように推定される．最後に，発音時刻を入力した基底から $\boldsymbol{W}_{\text{target}}$ と $\boldsymbol{H}_{\text{target}}$ を構成し，NMF と同様にウィナーフィルタ (2) によって得られた振幅スペクトログラムに ISTFT を適用することで目的音源の音響信号を復元できる．

OI-NMF の一番の特徴は，NMF による音源分離において，発音時刻を補足情報として扱えるように導入したバイナリマスク $\boldsymbol{S} \in \{0, 1\}^{K \times T}$ にある．バイナリマスクはアクティベーションと同サイズの 2 値行列であり，アクティベーションと要素積をとる形で導入される．バイナリマスクを導入した NMF の拡張モデルとして beta process sparse NMF (BP-NMF) [26, 27] が提案されており，これに従って OI-NMF を次のように定義する．

$$\boldsymbol{X} \sim \boldsymbol{W}(\boldsymbol{H} \odot \boldsymbol{S}) \tag{3}$$

ここで，$\boldsymbol{X} \in \mathbb{Z}_{\geq 0}^{F \times T}$ は入力振幅スペクトログラム（$\mathbb{Z}_{\geq 0}$ は非負整数全体の集合），$\boldsymbol{W}, \boldsymbol{H}$ は通常の NMF(1) と同様の基底スペクトルとアクティベーションである．バイナリマスクは，既存の NMF におけるアクティベーションの値，つまり対応する楽器音の音量変化を，その 1/0 の値でオン/オフする機能を持つ．バイナリマスクが 1 のフレームはアクティベーションの値で楽器が発音しており，0 のフレームはアクティベーションの値によらず楽器の音量は 0 になる．このバイナリマスクにおいて，発音時刻はマスクが 0 から 1 に変化するフレームとして扱う．

OI-NMF のモデル (3) は，BP-NMF と同様に階層ベイズモデルとして定義され，モデルの変数を確率変数として次のように事前分布を導入する．

$$X_{f,t} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S} \sim \mathrm{Poisson}\left(X_{f,t} \,\middle|\, \sum_{k=1}^{K} W_{f,k} H_{k,t} S_{k,t} \right), \tag{4}$$

$$W_{f,k} \sim \mathrm{Gamma}\left(W_{f,k} \mid \alpha^W, \beta^W\right), \tag{5}$$

$$H_{k,t} \sim \mathrm{Gamma}\left(H_{k,t} \mid \alpha^H, \beta^H\right), \tag{6}$$

ここで，$\alpha^W, \beta^W, \alpha^H, \beta^H$ はガンマ分布のハイパーパラメータである．α^W, α^H はガンマ分布の形状パラメータであり，基底スペクトルに関する α^W は楽器音の調波構造におけるスパース性を誘導するため 1 より小さい値に設定する．一方，アクティベーションは 0 になるとバイナリマスクが機能しなくなるため，α^H を 1 より少し大きい値に設定することで，一定の大きさを持った値を誘導する．

4.1 OI-NMF の構造

提案法の新規性は，新しく導入したバイナリマスクと発音時刻の組み合わせにある．本節では，バイナリマスクと発音時刻のモデリングおよびこれらの変数を含めた OI-NMF モデルの推論方法について説明する．

4.1.1 バイナリマスク

バイナリマスクの事前分布をモデリングする際に，楽器音はその種類によって一定時間持続するという仮定を考える．つまり，現在発音している楽器音は次の時間フレームでも発音している確率が高く，発音していない楽器音は次のフレームも発音していない確率が高い．この仮定に基づき，バイナリマスクの事前分布をマルコフ連鎖によってモデル化する．バイナリマスク $\boldsymbol{S}$ のある基底 $\boldsymbol{S}_k = \boldsymbol{S}_{k,:}$ が従うマルコフ連鎖による事前分布は以下のように表される．

$$p(\boldsymbol{S}_k) = p(S_{k,1}) \prod_{t=2}^{T} p\left(S_{k,t} \mid S_{k,t-1}\right) \tag{7}$$

第 1 項 $p(S_{k,1})$ は最初の時間フレームの要素が従う確率分布であり，初期確率 $a_0 \in (0,1)$ をパラメータとするベルヌーイ分布によって定義される．

$$p(S_{k,1}) = \mathrm{Bernoulli}\left(S_{k,1} \mid a_0\right) \tag{8}$$

$p\left(S_{k,t} \mid S_{k,t-1}\right)$ はバイナリマスクの t が 2 以上のインデックスの要素が従う確率分布であり，ベルヌーイ分布の積によって定義される．

$$\begin{aligned} p\left(S_{k,t} \mid S_{k,t-1}\right) = \mathrm{Bernoulli}\left(S_{k,t} \mid a_{1\to 1}\right)^{S_{k,t-1}} \\ \cdot \mathrm{Bernoulli}\left(S_{k,t} \mid a_{0\to 1}\right)^{1-S_{k,t-1}} \end{aligned} \tag{9}$$

ここで，$a_{1\to 1}, a_{0\to 1} \in (0,1)$ はバイナリマスクがオン状態からオン状態，およびオフ状態からオン状態に遷移する確率である．これらの値は，楽器音の連続性の仮定の基づき，$a_{1\to 1}$ は 1 に近い値，$a_{0\to 1}$ は 0 に近い値に設定する．(7) より，バイナリマスク $\boldsymbol{S}$ 全体の事前分布は以下のように表すことができる．

$$p(\boldsymbol{S}) = \prod_{k=1}^{K} p(\boldsymbol{S}_k) = \prod_{k=1}^{K} p(S_{k,1}) \prod_{t=2}^{T} p\left(S_{k,t} \mid S_{k,t-1}\right) \tag{10}$$

4.1.2 発音時刻行列

分離したい楽器音は J 個（ただし $J < K$）の音高を持っており，音高 $j \in \{1, 2, \ldots, J\}$ に対して発音時刻の系列 $\boldsymbol{\tau}_j = (\tau_{j,1}, \ldots, \tau_{j,n}, \ldots, \tau_{j,N_j})$ が与えられると仮定する．ここで，N_j は音高 j に対して与えられる発音時刻の個数であり，発音時刻 $\tau_{j,n}$ は時間周波数領域の時間フレーム単位で表される．この発音時刻は，後述するモデル推論の際に扱いやすくするため，バイナリマスクと同サイズの発音時刻行列 $\boldsymbol{O} \in \{0,1\}^{K \times T}$ の形で次のように定義する．

$$O_{k,t} = \begin{cases} 1, & \tau_{k,n} \leq t \leq \tau_{k,n} + T_{\mathrm{onset}} \\ & \quad (k = 1, 2, \ldots, J, n = 1, 2, \ldots, N_j) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

ここで，T_{onset} は発音時刻の許容幅を表す．許容幅を設けることで，発音時刻が目的音源より前のタイミングに入力された場合でも分離を行うことができる．T_{onset} が大きすぎると目的音源以外の音源も参照される可能性があるため，T_{onset} を 1/16 拍, 1/8 拍, 1/4 拍 ... と変化させ，推論がうまく動作する下限である 1/8 拍に経験的に設定した．

4.2 提案モデルの推論

OI-NMF の出力変数である基底スペクトル $\boldsymbol{W}$，アクティベーション $\boldsymbol{H}$ およびバイナリマスク $\boldsymbol{S}$ を推定するためには，ベイズ則によりこれらの事後分布を推論すればよい．しかし，事後分布を解析的に計算することは困難なため，ギブスサンプリングによって期待値で近似的に求める．ギブスサンプリングでは，他の変数が与えられた条件付き分布に従ってサンプル列を生成し，サンプル列の平均を取ることで近似を行う．OI-NMF の i 番目のサンプリング式は次のように表される．

$$\boldsymbol{W}^{(i)} \sim p\left(\boldsymbol{W} \,\middle|\, \boldsymbol{H}^{(i)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}\right) \tag{12}$$

$$\boldsymbol{H}^{(i)} \sim p\left(\boldsymbol{H} \,\middle|\, \boldsymbol{W}^{(i+1)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}\right) \tag{13}$$

$$\boldsymbol{S}^{(i)} \sim p\left(\boldsymbol{S} \,\middle|\, \boldsymbol{W}^{(i+1)}, \boldsymbol{H}^{(i+1)}, \boldsymbol{X}\right) \tag{14}$$

4.2.1 バイナリマスクのサンプリング

バイナリマスクの各要素は 0 か 1 の値をとるため，条件付き事後分布はベルヌーイ分布で表すことができる．

$$S_{k,t} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X} \sim \mathrm{Bernoulli}\left(S_{k,t} \,\middle|\, \frac{P_1}{P_1 + P_0}\right) \tag{15}$$

ここで，尤度 P_1, P_0 は $\boldsymbol{S}$ のインデックス k, t を除いた全ての要素 $S_{\neg k,t}$ を用いて次のように表される．

$$P_1 = p\left(S_{k,t} = 1 \mid S_{\neg k,t}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X}\right) \tag{16}$$

$$P_0 = p\left(S_{k,t} = 0 \mid S_{\neg k,t}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X}\right) \tag{17}$$

尤度 P_1(16) は以下のように書き下される．

$$P_1 \propto p(S_{k,t} = 1) p\left(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{k,t} = 1, S_{\neg k,t}\right) \tag{18}$$

(18) の第 1 項と第 2 項はそれぞれ次のように表すことができる．

$$p(S_{k,t} = 1) = \begin{cases} a_0, & t = 1 \\ a_{1\to1}^{S_{k,t-1}} a_{0\to1}^{1-S_{k,t-1}}, & t \geq 2 \end{cases} \tag{19}$$

$$p^1_{k,t} \triangleq p\left(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{k,t} = 1, S_{\neg k,t}\right) \tag{20}$$

$$\propto \prod_{f=1}^{F} (X^{\neg k}_{f,t} + W_{f,k} H_{k,t})^{X_{f,t}} \exp(-W_{f,k} H_{k,t}) \tag{21}$$

ここで，$X^{\neg k}_{f,t} = \sum_{l \neq k} W_{f,l} H_{l,t} S_{l,t}$ である．したがって，P_1 は次のように表すことができる．

$$P_1 = \begin{cases} a_0 p^1_{k,t}, & t = 1 \\ a_{1\to1}^{S_{k,t-1}} a_{0\to1}^{1-S_{k,t-1}} p^1_{k,t}, & t \geq 2 \end{cases} \tag{22}$$

同様に，P_0 は次のように表すことができる．

$$P_0 = \begin{cases} (1 - a_0) p^0_{k,t}, & t = 1 \\ (1 - a_{1\to1}^{S_{k,t-1}})(1 - a_{0\to1}^{1-S_{k,t-1}}) p^0_{k,t}, & t \geq 2 \end{cases} \tag{23}$$

ここで，$p^0_{k,t} \triangleq \prod_{f=1}^{F} (X^{\neg k}_{f,t})^{X_{f,t}}$ である．(22)，(23) および (15) に従って $t = 1$ から順にサンプリングすることで，バイナリマスク全体をサンプリングすることができる．

また，発音時刻が存在する部分は必ず楽器音はオン状態になっていると考え，発音時刻行列が $O_{k,t} = 1$ のインデックスの値をサンプリング結果にかかわらず $S_{k,t} = 1$ とする．それ以外のインデックスはサンプリング結果に従う．これにより，OI-NMF を推定する際に発音時刻を入力できる．

4.2.2 他の変数のサンプリング式

基底スペクトルとアクティベーションのサンプリング式は，BP-NMF のギブスサンプリング [27] と同様に以下のように導出できる．

$$W_{f,k} \mid \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{X} \sim \mathrm{Gamma}\left(\alpha^W + \sum_{t=1}^{T} X_{f,t} \phi_{f,t,k}\,,\ \beta^W + \sum_{t=1}^{T} H_{k,t} S_{k,t}\right) \tag{24}$$

$$H_{k,t} \mid \boldsymbol{W}, \boldsymbol{S}, \boldsymbol{X} \sim \mathrm{Gamma}\left(\alpha^H + \sum_{f=1}^{F} X_{f,t} \phi_{f,t,k},\ \beta^H + S_{k,t} \sum_{f=1}^{F} W_{f,k}\right) \tag{25}$$

4.2.3 OI-NMF のサンプリングアルゴリズム

アルゴリズム 1 に OI-NMF のギブスサンプリングアルゴリズムを示す．最初に各変数の初期化を行う．ギブスサンプリングで推論される確率分布は，初期値によらず定常分布に収束することが知られているが，どの音源がどの基底に推定されるかは初期値に大きく依存する．そのため，目的音源が発音時刻を与えた基底に出現するように誘導するため，確率変数は score-informed NMF [16, 17] を参考にして初期化する．

基底スペクトルは，ガンマ事前分布に従ってランダムに初期化する．アクティベーションは，発音時刻が与えられたフレームはガンマ分布で初期化する．また，発音時刻が与えられていない基底は全てのタイミング

Algorithm 1 OI-NMF のギブスサンプリング

1: $\boldsymbol{W}$, $\boldsymbol{H}$ および $\boldsymbol{S}$ を初期化
2: **for** $i = 1, 2, \cdots$ **do**
3: $\quad \phi_{f,t,k} = \frac{W_{f,k}H_{k,t}S_{k,t}}{\sum_l W_{f,l}H_{l,t}S_{l,t}}$ を計算
4: $\quad$ 式 (24) から $\boldsymbol{W}$ をサンプリング
5: $\quad$ 式 (25) から $\boldsymbol{H}$ をサンプリング
6: $\quad$ 式 (15), (22) および (23) から $\boldsymbol{S}$ をサンプリング
7: **end for**
8: サンプル列から $\boldsymbol{W}$, $\boldsymbol{H}$ および $\boldsymbol{S}$ の期待値を計算

で伴奏楽器が存在しうるため，同様にガンマ分布で初期化する．それ以外のフレームは 0 で初期化する．

$$H_{k,t} = \begin{cases} 0, & O_{k,t} \neq 1 \ (k = 1, 2, \ldots, L) \\ \frac{\alpha^H}{\beta^H}, & \text{otherwise} \end{cases} \tag{26}$$

バイナリマスクも同様のルールに従って初期化する．

$$S_{k,t} = \begin{cases} 0, & O_{k,t} \neq 1 \ (k = 1, 2, \ldots, L) \\ 1, & \text{otherwise} \end{cases} \tag{27}$$

その後，各変数のサンプリング式に従ってサンプル列を生成する．出力変数の値は，バーンイン後のサンプル列に対して平均をとったものとなる．ここで，バーンインとはサンプルが定常分布に達していないため破棄される期間を意味する．

4.3 目的音源の復元

ギブスサンプリングによって得られた出力変数を用いて，通常の NMF と同様に目的音源の音響信号を復元する．4.1.2 で述べたように，バイナリマスクに入力した発音時刻によって，基底 $k = 1, 2, \ldots, J$ に対応する音源が推定される．そのため，$\boldsymbol{W}_{\text{target}}$ と $\boldsymbol{H}_{\text{target}}$ は発音時刻を与えた基底を用いて次のように構成する．

$$\boldsymbol{W}_{\text{target}} = \boldsymbol{W}_{:,1:J}, \tag{28}$$

$$\boldsymbol{H}_{\text{target}} = \boldsymbol{H}_{1:J,:} \odot \boldsymbol{S}_{1:J,:}. \tag{29}$$

これに対してウィナーフィルタ (2) を適用し，得られた目的音源の振幅スペクトログラムに対して ISTFT を行うことで目的音源の音響信号を得ることができる．

5 評価実験

本節では，OI-NMF が発音時刻を利用して目的音源を分離できるか検証するために行ったメロディ分離実験について説明する．また，既存手法との比較による OI-NMF の有効性検証および頑健性評価についても説明する．

5.1 実験設定

入力楽曲には，音源分離実験用の実楽曲データセットである MedleyDB [28] から，ヴォーカルを含まずメロディのアノテーションが存在する楽曲として選択した，アーティストが MusicDelta であるジャズ楽曲 8 曲 (BebopJazz, CoolJazz, FreeJazz, FunkJazz, FusionJazz, LatinJazz, ModalJazz, SwingJazz) を利用した．これらの楽曲の wav ファイルから冒頭 20 秒を切り出し，22,050 [Hz] にダウンサンプリングした信号に対して，FFT サイズ 512 サンプル，オーバーラップ 50%，窓関数がハミング窓の STFT を適用することで得られた振幅スペクトログラムを OI-NMF の入力とした．なお，ドラムのような打楽器成分は，発音時刻が他の楽器と重複しやすいため，残存していると目的楽器音の分離が失敗しやすくなる．そのため，打楽器成分は予め調波・打楽器音分離 [29] によって除去した．

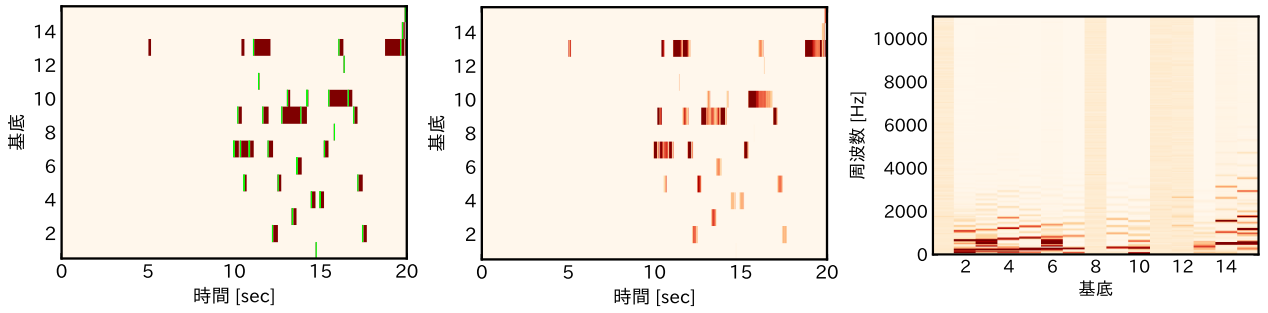
今回の実験では，メロディを演奏するアノテーションが付与された楽器を目的音源に設定し，これを分離する実験を行った．OI-NMF に入力するメロディ楽器の発音時刻は，MedleyDB データセットに含まれる F0 アノテーションから生成したものを用いた．F0 の値を MIDI ノート番号に変換し，ノート番号が変化する時刻を発音時刻とした．なお，ビブラートなどによる F0 の変化は発音時刻には含めない．

OI-NMF の基底数は十分大きい値として $K = 25$ とし，他のハイパーパラメータは $\alpha^W = 0.5$, $\beta^W = 1.0$, $\alpha^H = 1.1$, $\beta^H = 1.0$, $a_0 = 0.5$, $a_{1\to1} = 0.99$, $a_{0\to1} = 0.1$ に設定した．ギブスサンプリングは 200 回行い，得られたサンプルのうち開始から 100 サンプルはバーンインとして破棄して期待値を計算した．

5.2 分離精度評価指標

分離精度評価の指標には，signal-to-distortion ratio (SI-SDR), signal-to-interference ratio (SI-SIR), signal-to-artifacts ratio (SI-SAR) [30] を採用した．SI-SDR は推定された分離音と残差ノイズの比によって定義され，その値が大きいほど分離精度がよいことを表す．さらに，残差ノイズは目的音源以外の音源由来の干渉ノイズとアルゴリズム由来のノイズに分けられ，分離音とこれらのノイズの比によって SI-SIR と SI-SAR がそれぞれ定義される．SI-SIR と SI-SAR は互いにトレードオフの関係にあり，比較することでどちらのノイズ成分が支配的か調べることができる．

一般に，ブラインド音源分離の精度を評価するためには，上記の評価指標のスケール可変版である SDR, SIR および SAR [31] が用いられることが多い．しかし，OI-NMF のように分離音に無音区間が含まれると分離精度を正しく評価できなくなる．今回の実験設定



(a) $\boldsymbol{O}$ と $\boldsymbol{S}$　(b) $\boldsymbol{H} \odot \boldsymbol{S}$　(c) $\boldsymbol{W}$

図 2: OI-NMF による推定例（SwingJazz）．(a) 発音時刻行列（緑）とバイナリマスク（赤）．(b) アクティベーションとバイナリマスクの要素積．(c) 基底スペクトル．基底数 $K = 25$ のうち，発音時刻を与えた $J = 15$ 個の基底を表示している．

ではこれらの指標に代わり，スケール不変版を利用することでより正確な分離精度比較が可能となる．

5.3　分離例

まずはじめに，OI-NMF によるメロディ楽器分離の例を示す．ここでは，SwingJazz からクラリネットを分離した例を挙げる．発音時刻はすべてのタイミングで欠落なく与えられたものを利用した．

図 2 に推定された OI-NMF の各変数のヒートマップを示す．図 2(a) に入力発音時刻と推定されたバイナリマスクを示す．バイナリマスクが発音時刻から続いてピアノロールのように推定されている．基底 $k = 1, 8, 11, 12$ にはクラリネット以外の音源が推定されているが，発音時刻が存在しないフレームのマスクは 0 になっている．また，基底 $k = 13$ にはクラリネットとそれ以外の楽器音が同時に推定されてしまっている．図 2(b) に推定されたバイナリマスクとアクティベーションの要素積を示す．図 2(c) に推定された基底スペクトルを示す．クラリネットの調波構造が現れていることが確認できる．基底 $k = 1, 8, 11, 12$ にはクラリネットではない非調波成分が推定されているが，対応するアクティベーションと共に小さな値をとっているので信号復元時には打ち消される．この例の SI-SDR 改善率は約 4dB であった．ほかの楽曲の分離例と音源は次のリポジトリで確認できる[1]．

5.4　発音時刻の有効性検証

OI-NMF における発音時刻の貢献を示すため，OI-NMF と発音時刻を利用しない分離手法の精度を比較する実験を行った．比較対象には，発音時刻を入力しない OI-NMF と Bayesian NMF (BNMF) [32] を採用した．BNMF は通常の NMF を確率モデルとして推定を行う手法である．これらの発音時刻を利用しない手法により推定された基底は，目的楽器音とそれ以外の楽器音の基底で分類されていない．そのため，OI − NMF のように基底 $k = 1, 2, \ldots, J$ を利用して復元した信号は OI-NMF の分離精度の下限を与える．この下限と OI-NMF の分離精度を比較することで，発音時刻の有効性を確認することができる．

表 1: 発音時刻を利用する手法（発音時刻あり OI-NMF）と利用しない手法（発音時刻なし OI-NMF と B NMF）の SI-SDR 改善率 [dB] の平均．カッコ内の値は標準偏差を表す．

	OI-NMF		BNMF
	発音時刻あり	発音時刻なし	
Bebop	**5.62** (2.46)	−2.75 (3.32)	−4.42 (5.80)
Cool	**4.84** (2.75)	−0.56 (3.10)	−0.83 (5.38)
Free	**4.49** (1.63)	−3.37 (5.24)	−8.44 (14.7)
Funk	**9.86** (0.97)	−3.69 (3.99)	−4.84 (7.32)
Fusion	**7.09** (1.21)	−1.54 (3.33)	0.33 (3.22)
Latin	**5.77** (0.37)	−4.58 (11.9)	−6.67 (10.3)
Modal	**4.52** (1.92)	−5.60 (4.05)	−1.77 (5.52)
Swing	**4.29** (1.09)	−2.38 (2.36)	−6.08 (1.82)

各楽曲に対して 10 回分離を行ったときの，SI-SDR 改善率の平均と標準偏差を表 1 に示す．全ての楽曲において，発音時刻を入力した OI-NMF では平均が 0 以上であり，分離精度が改善していることが確認できる．一方，発音時刻を利用しない手法では平均は 0 未満であり，目的音源の分離ができていないことを示している．この結果より，OI-NMF は目的音源の分離に発音時刻を活用していることが確認できる．

[1] https://github.com/YutaKusaka/onset-informed-NMF-example

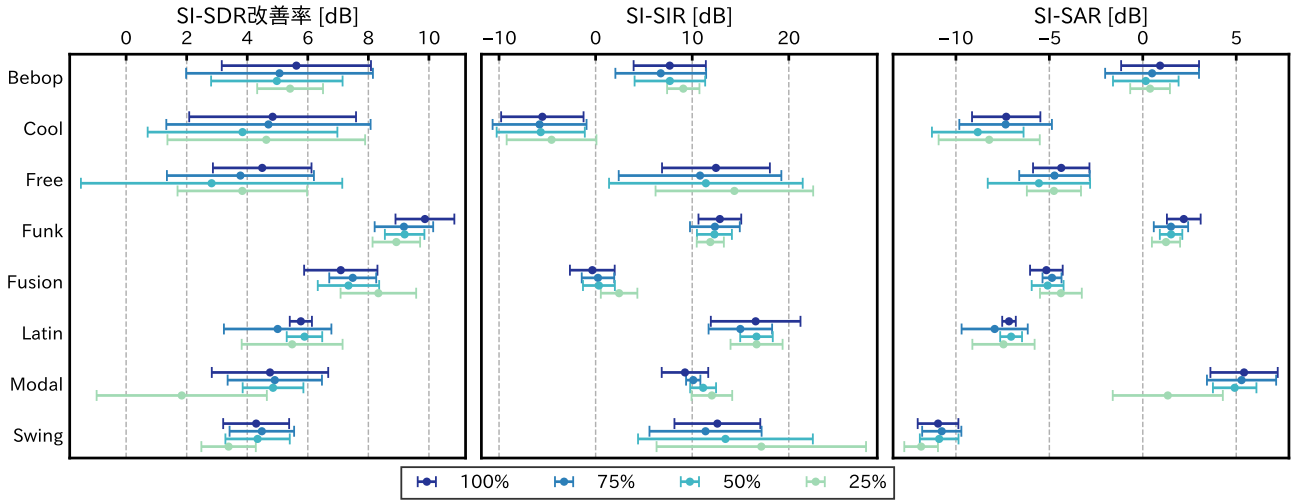


図 3: 発音時刻の存在割合を変化させたときの分離精度変化．ドットは 10 回分離を行った平均，エラーバーは標準偏差を表す．凡例は入力発音時刻の割合を示す．

5.5 発音時刻の欠落に対する頑健性評価

ユーザが楽曲を聴取しながら発音時刻を入力する際には，聞き逃しなどによる発音時刻の一部欠落が予測される．この一部が欠落した発音時刻が入力された場合でも OI-NMF による目的音源の分離は可能か検証する実験を行った．．この実験では，欠落がない発音時刻（100%）に対して存在割合が 75%, 50%, 25%になるよう各基底からランダムに欠落させた発音時刻を作成し，これらを入力した場合の分離精度を比較した．他の実験設定は 5.4 節と同様である．評価には SI-SDR 改善率に加え，どの種類のノイズが支配的か調べるために SI-SIR，SI-SAR も指標として利用した．

図 3 に評価結果を示す．全ての楽曲，発音時刻割合で SI-SDR の改善率の平均が 0 以上となり，発音時刻の割合が減少するにつれて，平均は減少し，標準偏差は増加する傾向にあることが確認できる．各試行で分解された基底を確認すると，発音時刻の割合が減少するにつれて発音時刻に対応する音の推定に失敗する基底が増加しており，このような傾向が現れていると考えられる．また，SI-SIR と SI-SAR に関しても SI-SDR と同様の変化傾向がみられ，これらを比較すると，OI-NMF においてはアルゴリズム由来のノイズのほうが支配的であることが確認できた．

また，発音時刻割合が 50%や 25%のように少ないとき，いくつかの試行で SI-SDR の改善率が 0 以下になり，分離に失敗していることが確認できた．さらなる考察を行うため，OI-NMF に入力される発音時刻が最悪の場合を想定し，各基底に対して 1 つだけ発音時刻を与えて分離する実験を各楽曲 10 回行った．その結果，多くの楽曲で分離に失敗している試行がみられ，FreeJazz では 4 回，LatinJazz では 6 回，SwingJazz では 8 回失敗していることが確認できた．これらの結果より，入力発音時刻が少なすぎる場合は分離に失敗すると予想される．一方で，50%以上の発音時刻が与えられる場合は分離精度の劣化は小さく抑えられており，発音時刻の一部の欠落を許容して分離ができると期待される．

5.6 発音時刻のずれに対する頑健性評価

ユーザが入力した発音時刻には，欠落だけでなく真の位置からのずれも含むことが予測される．この時間方向にずれを含む発音事項が入力された場合の，OI-NMF の分離の安定性を検証する実験を行った．発音時刻が含むずれは，実際にユーザに発音時刻する操作を行った研究で報告された統計値に基づき，平均が真の位置から 10 ms 後ろ [33], 標準偏差が 100 ms [34] の正規分布でモデル化した．つまり，ずれを含まない発音時刻を τ とすると，ずれを含む発音時刻 $\tilde{\tau}$ は次のように表すことができる．

$$\tilde{\tau} = \tau + \epsilon \tag{30}$$

$$\epsilon \sim \mathcal{N}(0.01, 0.1^2) \tag{31}$$

欠落のない 100%の発音時刻と，これに (30), (31) を適用して作成したずれを含む発音時刻の 2 種類を入力した際のそれぞれの分離精度を比較した．この実験も，5.4 節や 5.5 節の実験と同様のパラメータで，各楽曲に 10 回分離を行った．

図 4 に評価結果を示す．CoolJazz と FreeJazz を除く全ての楽曲で，ずれによって分離精度が劣化していることが確認できる．BebopJazz と ModalJazz の SI-SDR 改善率は，ずれがない場合に比べて大きく劣化しているが，他の楽曲では劣化幅が小さく抑えられてい

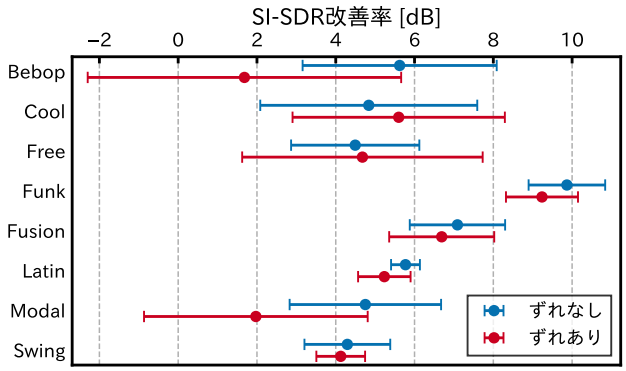


図 4: 発音時刻にずれを含む場合の分離精度の比較．ドットは 10 回分離を行った平均，エラーバーは標準偏差を表す．

る．BebopJazz と ModalJazz の大きな分離精度劣化の原因として，ノイズ性を持つ伴奏楽器が楽曲全体に存在しており，これを目的音源と誤って推定していることが考えられる．また，CoolJazz と FreeJazz においては，ずれを含まない場合よりも高い分離精度を示している．これは，データセットに含まれる F0 アノテーションから作成した発音時刻よりも，ずれを含んだ発音時刻のほうがより適しているためと考えられる．そのため，今後は入力発音時刻の作成方法による分離精度の変化なども考慮する必要がある．

以上の実験により，OI-NMF は発音時刻を利用して分離したい楽器音を効率的に推定できることが示された．さらに，人間が作成する際に想定されるレベルでの発音時刻の欠落やずれといった外乱をある程度許容した分離ができることを確認した．

6 おわりに

本稿では，分離したい音源の発音時刻を補足情報として利用する，新しい音源分離手法である onset-informed NMF を提案した．発音時刻を NMF ベースの音源分離フレームワークへ組み込むために，NMF の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し，マスク上で発音時刻を扱った．さらに，バイナリマスクと発音時刻も含めてモデルを推論するアルゴリズムを導出した．分離精度を検証する実験で，発音時刻の一部が欠落したり，時間方向にずれを含むような現実的な設定においても，安定した分離を実現することが期待できる結果を示した．

現在の問題として，目的音源と伴奏音源が同時に発音しているような場合，NMF の性質上，OI-NMF ではこれらを分離することは難しいと考えられる．そのため，基底スペクトルに対して目的音源の調波構造に関する制約を取り入れるなどして，目的音源推定の精度を高めることを考えている．さらに，目的音源以外に発音時刻が与えられてしまった場合も，分離精度が劣化すると予想されるため，これに対する検証実験も行う予定である．

さらに，現在は入力発音時刻は楽器の音高ごとにグルーピングされて与えられる仮定をおいている．この仮定は，実際にユーザが発音時刻を入力する際には手間がかかる操作になると予想される．そのため，発音時刻を音高に依存しない単一の時系列としてモデルに入力できるように拡張することも考えている．

謝辞

本研究は JSPS 科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた．

参考文献

[1] Kazuyoshi Yoshii et al. Drumix: An audio player with real-time drum-part rearrangement functions for active music listening. *Information and Media Technologies*, 2(2):601–611, 2007.

[2] A. J. Simpson et al. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *LVA/ICA*, pages 429–436, 2015.

[3] E. Benetos et al. Automatic music transcription: challenges and future directions. *J. Intell. Inf. Syst.*, 41(3):407–434, 2013.

[4] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *ICASSP*, volume 2, pages II753–II756, 2000.

[5] B. Kostek. Musical instrument classification and duet analysis employing music information retrieval techniques. *Proc. IEEE*, 92(4):712–729, 2004.

[6] M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *IEEE Trans. Multimedia.*, 10(8):1617–1625, 2008.

[7] S. S. Shwartz et al. Robust temporal and spectral modeling for query by melody. In *SIGIR*, pages 331–338, 2002.

[8] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. In *NIPS*, pages 556–562, 2001.

[9] C. Févotte et al. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Comput.*, 21(3):793–830, 2009.

[10] A. Hyvräinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4):411–430, 2000.

[11] A. Liutkus et al. An overview of informed audio source separation. In *WIAMIS*, pages 1–4, 2013.

[12] P. Smaragdis and G. J. Mysore. Separation by “humming”: User-guided sound extraction from monophonic mixtures. In *WASPAA*, pages 69–72.

[13] A. Lefèvre et al. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *ISMIR*, pages 1–6, 2012.

[14] D. Kitamura et al. Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing. In *ISSPIT*, pages 392–397, 2013.

[15] D. Kitamura et al. Music signal separation by supervised nonnegative matrix factorization with basis deformation. In *DSP*, pages 1–6, 2013.

[16] S. Ewert and M. Muller. Using score-informed constraints for NMF-based source separation. In *ICASSP*, pages 129–132, 2012.

[17] J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *ICASSP*, pages 888–891, 2013.

[18] S. Uhlich et al. Deep neural network based instrument extraction from music. In *ICASSP*, pages 2135–2139, 2015.

[19] P. Chandna et al. Monoaural audio source separation using deep convolutional neural networks. In *LVA/ICA*, volume 10169, pages 258–266, 2017.

[20] A. Ozerov et al. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *ICASSP*, pages 257–260, 2011.

[21] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[22] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *DMRN*, pages 1–5, 2005.

[23] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech Lang. Process*, 15(3):1066–1074, 2007.

[24] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *WASPAA*, volume 3, pages 177–180, 2003.

[25] D. W. Griffin and J. S. Lim. Signal Estimation From Modified Short-Time Fourier Transform. In *ICASSP*, volume 2, pages 804–807, 1983.

[26] D. Liang et al. Beta Process Sparse Nonnegative Matrix Factorization for Music. In *ISMIR*, pages 375–380, 2013.

[27] D. Liang and M. D. Hoffman. Beta Process Non-negative Matrix Factorization with Stochastic Structured Mean-Field Variational Inference. In *arXiv:1411.1804*, pages 1–6, 2014.

[28] R. Bittner et al. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *ISMIR*, pages 155–160, 2014.

[29] D. FitzGerald. Harmonic/Percussive Separation Using Median Filtering. In *DAFx*, pages 1–4, 2010.

[30] J. L. Roux et al. SDR - half-baked or well done? In *ICASSP*, pages 626–630, 2019.

[31] E. Vincent et al. Performance Measurement in Blind Audio Source Separation. *IEEE Trans. Audio Speech Lang. Process*, 14(4):1462–1469, 2006.

[32] A. T. Cemgil. Bayesian Inference for Nonnegative Matrix Factorisation Models. *Comput. Intell. Neurosci.*, 2009:1–17, 2009.

[33] P. Leveau et al. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *ISMIR*, pages 1–4, 2004.

[34] M. Cartwright et al. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proc. ACM Hum. Comput. Interact.*, 1:1–21, 2017.