5S-01

バイナリマスク付き非負値行列因子分解に基づく音源分離手法の発音時刻ずれに対する性能評価

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

本稿では目的音源の発音時刻を事前情報として利用するバイナリマスク付き非負値行列因子分解による音源分離手法 [1] において，発音時刻に時間のずれが含まれている場合の分離精度の変化について評価を行う．複数の楽器から構成されるモノラル音響信号から特定の音源のみを分離する処理には，目的音源の事前情報を利用する手法が主流となっており，我々は既存手法で用いられている事前情報よりユーザが容易に作成可能な事前情報として目的音源の発音時刻を利用する音源分離手法 [1] を提案した．これまでの報告においては，提案法に入力する発音時刻は入力楽曲に付帯する MIDI やアノテーションから作成された理想的な状況に限定されていた．本報告では，人間が音響信号を聴取し発音時刻を入力した際に発生する時間のずれをモデル化し，このずれが目的音源の分離精度に及ぼす影響を評価するための実験とその評価を行い，このようなずれに対して頑健な分離が期待できることを確認した．

2 分離手法の概要

ここでは，我々が提案した発音時刻を事前情報として利用する音源分離手法 [1] の概要について説明する．この分離手法は目的音源の発音時刻を利用して分離を行うため，ここでは OI-NMF (onset-informed NMF) と呼ぶことにする．入力変数は音響信号に短時間フーリエ変換（short-time Fourier transform; STFT）を適用することで得られる振幅スペクトログラムと，目的音源の音高別の発音時刻の系列である．OI-NMF は生成モデルとしての非負値行列因子分解（non-negative matrix factorization; NMF）に基づく音響モデルに対し，発音時刻を取り扱うためのマルコフ連鎖に基づくバイナリマスクを導入した階層ベイズモデルとなっている．入力変数が観測されたとき，モデルの未知の確率変数である基底スペクトル，

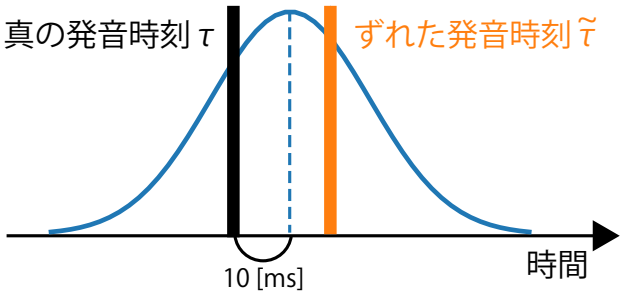

Fig. 1 発音時刻のずれの概略図．真の発音時刻から 10[ms] 後ろの時刻に平均をもつ正規分布に従ってずれた発音時刻が生成される．

アクティベーションおよびバイナリマスクの事後分布をギブスサンプリングにより推論する．その後，推定された変数のうち発音時刻を与えた基底を用いてスペクトログラムを再構成し，逆短時間フーリエ変換を適用することで目的音源の信号を分離する．

3 発音時刻ずれのモデル

以前の報告 [1] では，入力発音時刻は入力音響信号と同じデータベースに含まれる MIDI や F0 アノテーションから作成された発音時刻を利用していた．しかし，実際に人間が楽曲を聴取し，デバイスを目的音源の発音に合わせてタッピングして発音時刻を作成する場合，実際の発音時刻と入力発音時刻の間にはずれが発生することが予想される．

OI-NMF は入力発音時刻が与えられた時間フレームに目的音源が必ず存在すると仮定してモデル推論を行うため，目的音源が存在しない時間フレームに対して発音時刻が入力されると目的音源の基底を正しく推定できず分離精度が劣化する可能性がある．そこで本報告では発音時刻のずれをモデル化し，ずれを加えた発音時刻を OI-NMF に入力した場合の分離精度の評価を行うことで，OI-NMF の性能評価を行う．

発音時刻は自動採譜など様々な音楽情報処理において有用な事前情報として利用されており，人間が音響信号を聴取し作成した発音時刻と真値のずれの関係について報告されているため，これらの報告に基づいてモデル化を行う．楽器音の発音時刻を作成する実験ではずれの平均の絶対値が 10[ms] になること [2] が，音イベント区

Performance Evaluation of Separation Method based on Onset-informed Non-negative Matrix Factorization with Binary Mask for inaccurate Time Onset
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.

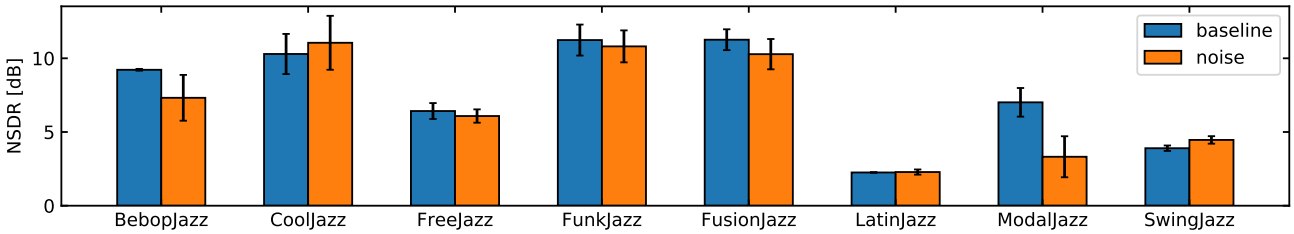

Fig. 2　入力発音時刻にずれが含まれていない（baseline）と含まれている（noise）場合の OI-NMF による分離結果の NSDR．エラーバーは標準偏差を表す．

間推定のための発音時刻ラベル作成実験ではずれが標準偏差が約 0.1[ms] の正規分布に似た分布で表されること [3] が報告されている．そのため，ずれを含む発音時刻 $\tilde{\tau}_{t,i}$ を以下のように定義する．

$$\tilde{\tau}_{t,i} = \tau_{t,i} + \epsilon \tag{1}$$

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

ここで，$\tau_{t,i}$ は時刻 t，音高 i の真の発音時刻である．Fig.1 に発音時刻のずれの様子を示す．

4　評価実験

アノテーションから作成した真の発音時刻と，これにずれを加えた発音時刻を OI-NMF に入力して目的音源の分離し，分離精度評価を行いずれの影響を評価する．

4.1　実験設定

実験に用いた音源は実楽曲データベース MedleyDB [4] に含まれるアーティストが MusicDelta の楽曲のうち，ジャンルがジャズの 8 曲である．これらの楽曲の冒頭から 20 秒を切り出し，22,050Hz にダウンサンプリングし，窓幅 512 サンプル，オーバーラップ 256 サンプル，窓関数をハミング窓とする STFT を適用することで得られる振幅スペクトログラムと，データベースに含まれるメロディの F0 アノテーションから作成した発音時刻を OI-NMF に与えることで混合音からメロディ楽器パートを分離する．なお，混合音からドラムを除去するため予め調波・打楽器音分離を行っている．NMF の基底数は十分大きい値として 25，モデルのハイパーパラメータは $a = 0.5, b = 1.0, c = 3.0, d = 3.0, \phi = 0.01, q_0 = 0.01, q_1 = 0.99$ に設定した．イテレーション 200 回，バーンイン 100 回のギブスサンプリングにより得られたサンプル列にから計算される期待値を各変数の出力とする．

音源分離精度の評価指標には source to distortion ratio (SDR) [5] の改善率（normalized SDR; NSDR）を用いた．SDR は `bss_eval` toolbox を用いて計算し，分離実験を 10 回行ったときの平均と標準誤差によって評価する．

4.2　実験結果

Fig. 2 に評価結果を示す．ずれが含まれている場合は，ずれが含まれていない場合に対して分離精度が劣化する傾向があるものの，大きな精度の差はないことが確認できる．また，CoolJazz に関しては，ずれが含まれている場合のほうが分離精度が改善されていることが確認できる．これにより，OI-NMF は入力発音時刻に多少のずれが含まれている場合も頑健な目的音源の分離が期待できることが分かる．

5　おわりに

本稿では，目的音源の発音時刻を事前情報として利用する分離手法に対して，入力発音時刻にずれが含まれている場合に分離精度がどのように変化するかについて評価を行い，OI-NMF はずれに対して頑健な分離を行うことができることを確認した．現段階では，OI-NMF の入力発音時刻は音高ごとに与えられることを仮定している．より簡単に発音時刻を入力するため，発音時刻を単一の系列として扱えるように拡張を行う予定である．

謝辞　本研究は JSPS 科研費 16H02884, 17K00365 および 19K12017 の助成を受けた．

参考文献

[1] 日下ほか. 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. 研究報告音楽情報科学（MUS）, Vol. 2019-MUS-124, No. 14, pp. 1–7, 2019.

[2] L. Daudet, et al. Methodology and Tools for the evaluation of automatic onset detection algorithms in music. In *ISMIR*, pp. 72–75, 2004.

[3] M. Cartwright, et al. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proc. of the ACM on Human-Computer Interaction*, Vol. 1, No. CSCW, 2017.

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

[5] E. Vincent, et al. Performance Measurement in Blind Audio Source Separation. *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462–1469, 2006.