Experimental evaluation of superresolution-based nonnegative matrix factorization for binaural recording

北村大地 (総合研究大学院大学), 猿渡洋 (東京大学), 中村哲 (奈良先端科学技術大学院大学)
高橋祐 (ヤマハ株式会社), 近藤多伸 (ヤマハ株式会社), 亀岡弘和 (東京大学)

1. 研究背景

- 近年，音楽信号分離技術が注目されている

応用例

- 自動採譜
- 音の拡張現実 (AR)　　etc.

■ 先行研究

- **Supervised NMF (SNMF) [1]**
  モノラル信号が対象
  各楽器音の調波構造を利用して特定楽器音のみを抽出
  分解したスペクトル基底を楽器ごとに分類することが困難
- **Multichannel NMF [2]**
  NMF をマルチチャネル信号用に拡張
  各楽器音の調波構造とチャネル間の位相情報を利用
  分解行列の初期値依存性が強く，頑健性に欠ける
- **超解像型SNMF及びハイブリッド法 [3]**
  マルチチャネル信号を対象
  方位分解とスペクトル分解の 2 ステップで目的音を抽出

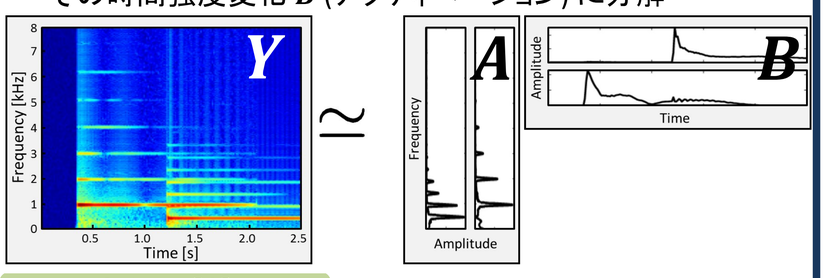
本研究の目的

ハイブリッド手法の分離精度向上を目指す
多重ダイバージェンスに基づく超解像型SNMFを提案し，実録音信号での実験結果を示す

2. NMFによる音源分離

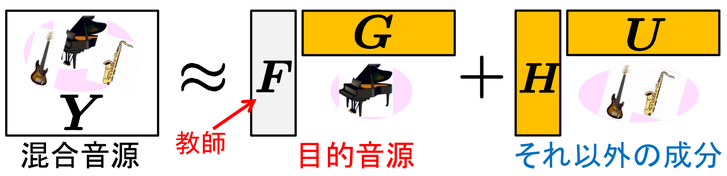
音響信号における NMF [4]

- スペクトログラム $\boldsymbol{Y}$ を有限個のスペクトルパーツ $\boldsymbol{A}$ (基底) とその時間強度変化 $\boldsymbol{B}$ (アクティベーション) に分解



SNMF [1]

- 教師スペクトルパーツ $\boldsymbol{F}$ を事前学習で作成しておく
- 「目的音源$\boldsymbol{FG}$」と「それ以外の成分$\boldsymbol{HU}$」に分離可能



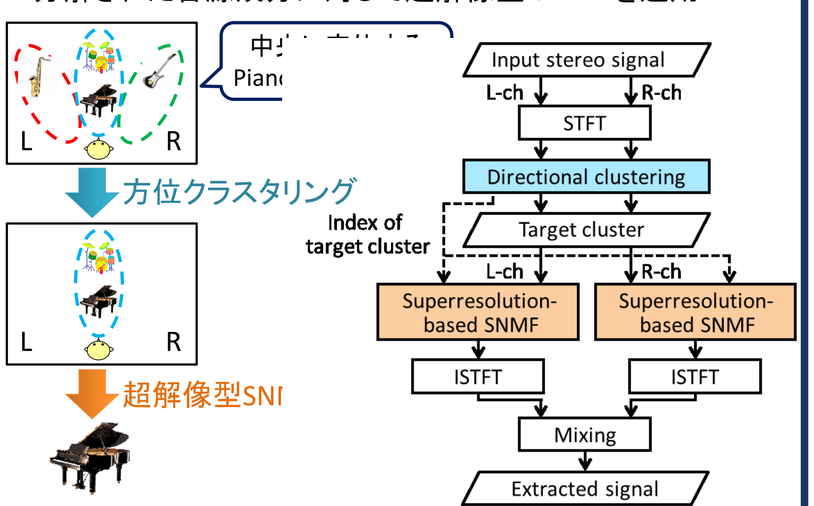
SNMF の問題点

- 分離対象信号に含まれる楽器数が多い場合（4 つ以上）は分離精度が極端に落ちる
- 楽器数が多い場合，楽器間で似たスペクトルパターンが多く現れることが原因

3. 超解像型 SNMF とハイブリッド法

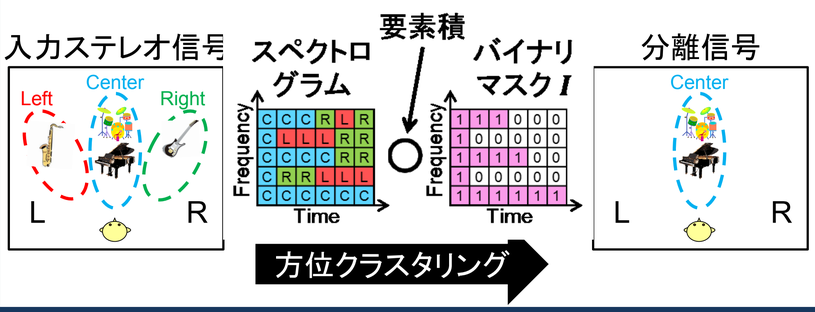
■ ハイブリッド音源分離法

- 方位クラスタリングと SNMF を組み合わせた2ステップのマルチチャネル信号分解手法
- 前段の方位クラスタリングにより，分離目的音源が含まれる方位を抽出
- 分解された音源成分に対して超解像型 SNMF を適用



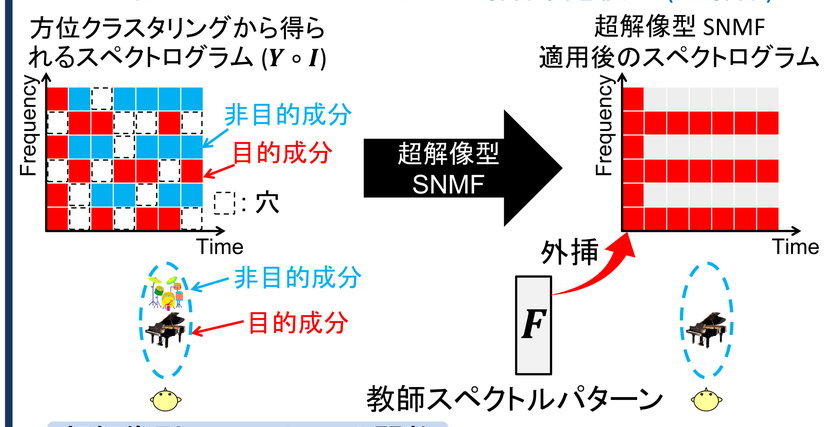
■ 方位クラスタリング (ステップ1)

- チャネルの振幅差を利用して k-means クラスタリング
- k-means の結果からバイナリマスク $\boldsymbol{I}$ が得られる
- 入力スペクトログラムに対するバイナリマスキングと等価



■ 超解像型 SNMF (ステップ2)

- 残留している非目的音源を抑圧
  ➡ 目的音源の分離
- 方位クラスタリング (バイナリマスキング) で失われた目的音源成分を教師スペクトルの外挿によって復元
  ➡ 低下したスペクトログラムの解像度を復元 (超解像)



超解像型 SNMF のコスト関数

$$\mathcal{J} = \sum_{\omega,t} i_{\omega,t} \mathcal{D}_{\beta_{\mathrm{NMF}}}\left(y_{\omega,t} \| \textstyle\sum_k f_{\omega,k} g_{k,t} + \sum_l h_{\omega,l} u_{l,t}\right) + \lambda \sum_{\omega,t} \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\mathrm{reg}}}\left(0 \| \textstyle\sum_k f_{\omega,k} g_{k,t}\right) + \mu \|\boldsymbol{F}^{\mathrm{T}} \boldsymbol{H}\|_{\mathrm{Fr}}^2$$

正則化項 (外挿誤りを防ぐ)　　罰則項 (分離性能を向上)

$\boldsymbol{I}$: バイナリマスク行列，$i_{\omega,t}, y_{\omega,t}, f_{\omega,k}, g_{k,t}, h_{\omega,l}, u_{l,t}$: それぞれの行列の要素
$\overline{\cdot}$: 論理反転，λ, μ: 重み係数，$\|\cdot\|_{\mathrm{Fr}}$: フロベニウスノルム，

- $\mathcal{D}_\beta(\cdot\|\cdot)$ は一般化距離関数: β-ダイバージェンス [5]
- ユークリッド距離 ($\beta = 2$) と KL-ダイバージェンス ($\beta = 1$) がよく用いられる

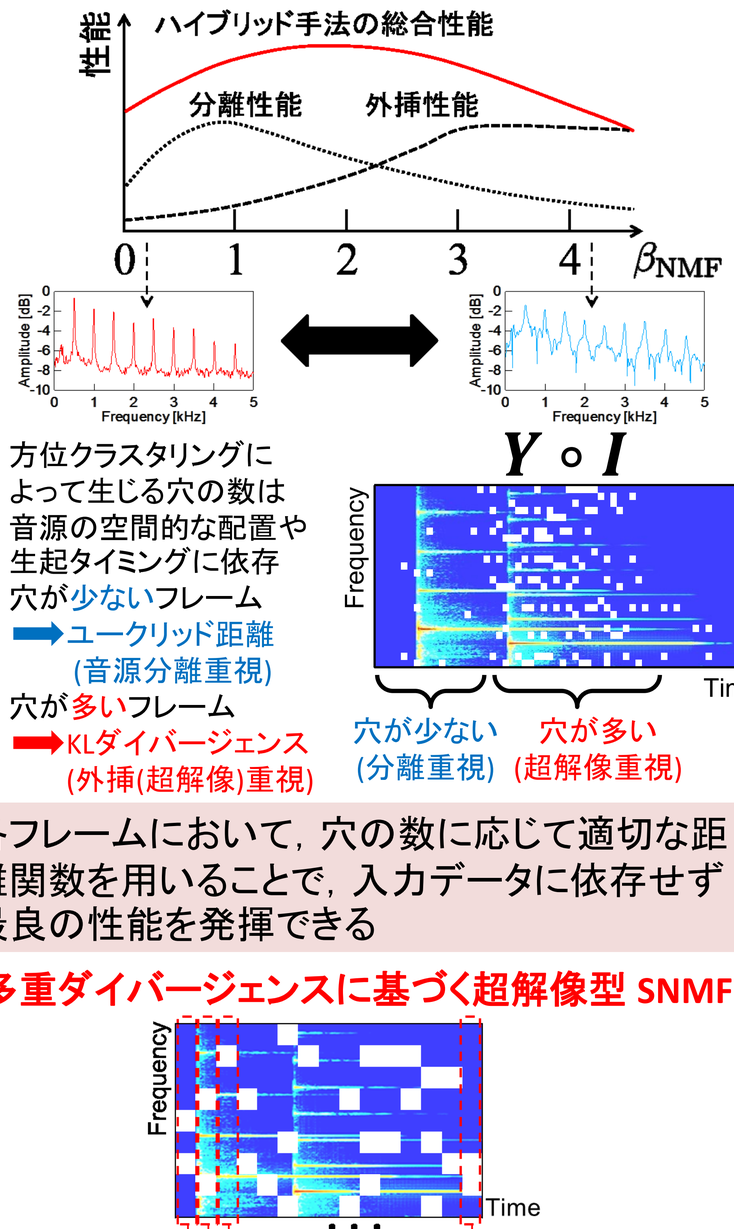
4. 多重ダイバージェンス

■ ダイバージェンスの選び方

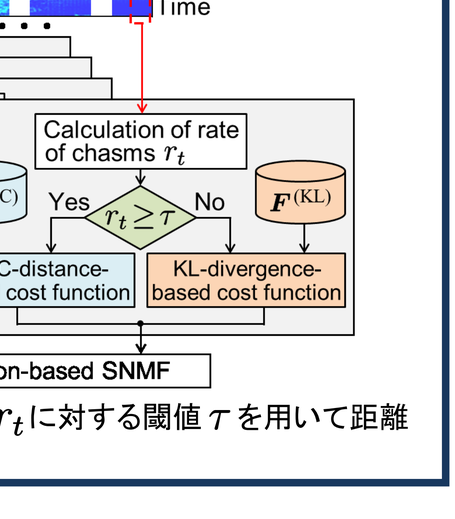
- 超解像型 SNMF には二つの異なるタスクがある

超解像型 SNMF ＝ 目的音源の分離 ＋ 基底の外挿（超解像）

- SNMFによる音源分離にはKL-ダイバージェンスが適切
- 教師基底の外挿 (超解像) にはユークリッド距離が適切
- 総合性能は分離と外挿のトレードオフによって決まる [6]



- 方位クラスタリングによって生じる穴の数は音源の空間的な配置や生起タイミングに依存
- 穴が少ないフレーム
  ➡ユークリッド距離 (音源分離重視)
- 穴が多いフレーム
  ➡KLダイバージェンス (外挿(超解像)重視)



各フレームにおいて，穴の数に応じて適切な距離関数を用いることで，入力データに依存せず最良の性能を発揮できる

多重ダイバージェンスに基づく超解像型 SNMF

- フレームごとに穴の密度 r_t に対する閾値 τ を用いて距離関数を決定

多重ダイバージェンスに基づく超解像型 SNMF のコスト関数

$$\mathcal{J}_{\mathrm{m}} = \textstyle\sum_t \mathcal{J}_t,$$

$$\mathcal{J}_t = \begin{cases} \sum_\omega i_{\omega,t} \mathcal{D}_{\beta=2}(y_{\omega,t} \| s_{\omega,t}^{(\mathrm{EUC})}) \\ \quad + \lambda^{(\mathrm{EUC})} \sum_\omega \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\mathrm{reg}}}(0 \| \sum_k f_{\omega,k}^{(\mathrm{EUC})} g_{k,t}) \\ \quad + \mu^{(\mathrm{EUC})} \|\boldsymbol{F}^{(\mathrm{EUC})\mathrm{T}} \boldsymbol{H}\|_{\mathrm{Fr}}^2 & (r_t \geq \tau) \\ \sum_\omega i_{\omega,t} \mathcal{D}_{\beta=1}(y_{\omega,t} \| s_{\omega,t}^{(\mathrm{KL})}) \\ \quad + \lambda^{(\mathrm{KL})} \sum_\omega \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\mathrm{reg}}}(0 \| \sum_k f_{\omega,k}^{(\mathrm{KL})} g_{k,t}) \\ \quad + \mu^{(\mathrm{KL})} \|\boldsymbol{F}^{(\mathrm{KL})\mathrm{T}} \boldsymbol{H}\|_{\mathrm{Fr}}^2 & (r_t < \tau) \end{cases}$$

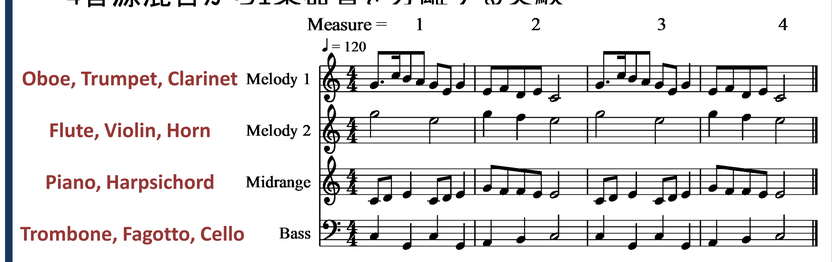
$$s_{\omega,t}^{(*)} = \textstyle\sum_k f_{\omega,k}^{(*)} g_{k,t} + \sum_n h_{\omega,n} u_{n,t}, \quad * = \{\mathrm{EUC}, \mathrm{KL}\},$$

$$r_t = \textstyle\sum_\omega \overline{i_{\omega,t}} / \Omega$$ ：各フレームにおける穴の密度

5. 人工及び実録音信号による実験

実験条件

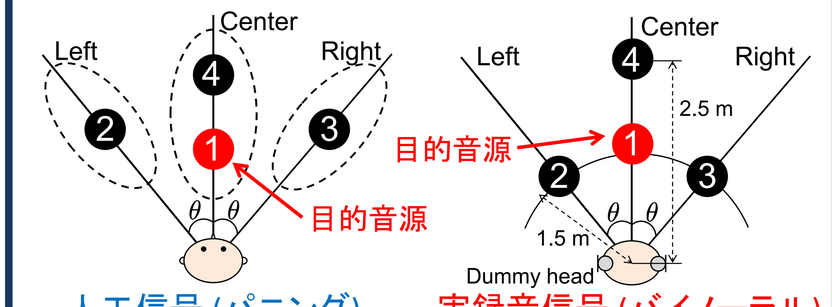
- 11種類の楽器音，4種のメロディのMIDIから混合音源を作成
- 4音源混合から1楽器音を分離する実験



- 各音源に対して，小節毎に左右音源の空間配置が異なる 4 種類の混合音源(SP1-SP4)を作成

Spatial condition	Measure 1	2	3	4
SP1	$\theta = 45°$	$\theta = 0°$	$\theta = 0°$	$\theta = 0°$
SP2	$\theta = 45°$	$\theta = 45°$	$\theta = 0°$	$\theta = 0°$
SP3	$\theta = 45°$	$\theta = 45°$	$\theta = 45°$	$\theta = 0°$
SP4	$\theta = 45°$	$\theta = 45°$	$\theta = 45°$	$\theta = 45°$

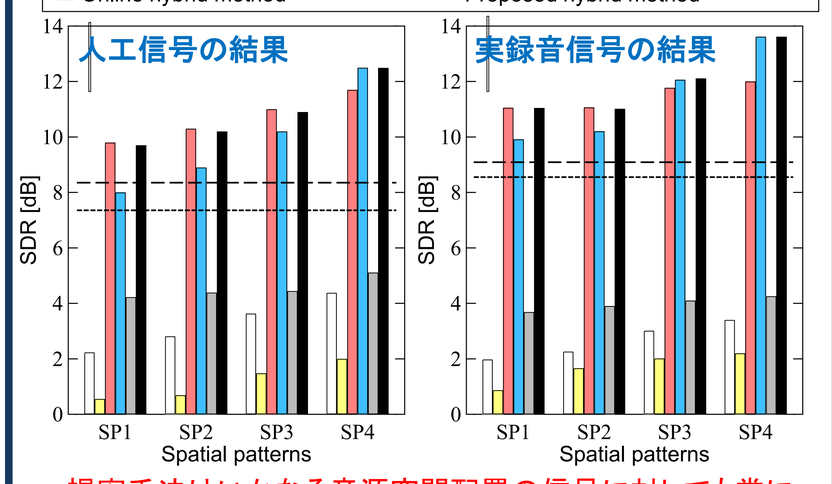
- 人工信号: 左右の振幅差 (パニング) で作成した混合信号
- 実録音信号: 残響時間 200 [ms] の環境でダミーヘッドを用いたバイノーラル録音で作成した混合信号



- 教師信号は24音(2オクターブ)が半音ずつ上昇する音源
- ダイバージェンス切り替えの閾値 τ は20%に設定
  ➡ 各フレームで全体の20%以上穴が開いていると ユークリッド距離で測る

実験結果

- SDR (分離度合と目的音の品質を含む総合的な分離精度) を手法毎に比較する (36 種の楽器パターンの平均評価値)
- オンライン法: 1フレーム毎に独立してハイブリッド法を適用した単純な例 (ただし距離関数は閾値で切り替える)



- 提案手法はいかなる音源空間配置の信号に対しても常に高いスコアをマークしている

参考文献

[1] P. Smaragdis, et al., “Supervised and semi-supervised separation of sounds from single-channel mixtures,” Proc. 7th International Conference on Independent Component Analysis and Signal Separation, pp.414–421, 2007.
[2] H. Sawada, et al., “Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization,” Proc. IEEE ICASSP, pp.261–264, 2012.
[3] D. Kitamura, et al., “Superresolution-based stereo signal separation via supervised nonnegative matrix factorization,” Proc. 18th DSP, T3C-2, 2013.
[4] D. D. Lee, et al., “Algorithms for non-negative matrix factorization,” Proc. Advances in Neural Information Processing Systems, vol.13, pp.556–562, 2001.
[5] S. Eguchi, et al., “Robustifying maximum likelihood estimation,” Technical Report of Institute of Statistical Mathematics, 2001.
[6] D. Kitamura, et al., "Divergence optimization in nonnegative matrix factorization with spectrogram restoration for multichannel signal separation," 4th HSCMA, 2014.