重畳マルコフ連鎖スペクトルモデルに基づく半教師あり学習による楽器音分離*

☆田沼 巌[1], △中野 允裕[1], △藤田 卓[2], 亀岡弘和[1 3], 嵯峨山茂樹[1]
([1] 東大院・情報理工, [2] 東大・工, [3]NTT CS 研)

1 はじめに

音楽信号からの楽器音分離は音響信号処理の重要な課題の一つであり, 楽器音一音単位での分離が達成されればユーザが各楽器ごとの音量を調整しながら鑑賞することが出来るようになるなど, 音楽鑑賞・加工の基盤技術としてその確立が期待されている.

本研究は非負値行列因子分解 (Nonnegative matrix factorization: NMF) に基づく音源分離の近年の発展と密接に関わっている. NMF に基づく音楽信号解析は一般に, 音楽信号のスペクトログラムを非負値行列に見立て, それを低ランクの非負値行列の積に分解することによって行われる. これはスペクトログラムを少数の頻出のスペクトルパターンとその音量変化によって近似しようとするものであり, 音楽信号においては各楽器音は曲中で繰り返し演奏されることが多いため, 頻出のスペクトルパターンは自ずと各楽器音の平均的なスペクトルとして表出することが期待される. NMF によって音楽信号を分解する研究は盛んに研究され, 自動採譜や楽器音分離, 楽器音認識などの音楽信号処理の多くの問題に適用されてきた. しかし, 実際には多くの楽器音は時間的に変化する多様なスペクトルパターンを持っており, これらは NMF に基づく音楽信号解析の大きな障害となってきた. そこで我々は時間変化する楽器音スペクトルを表現するために, 重畳マルコフ遷移スペクトルモデルと呼ぶ音楽音響信号モデルに音源分離技術を提案してきた. これは各楽器音をマルコフ遷移するスペクトルの系列として表現し, 音楽信号はそれらを重畳したものとしてモデル化することによって, 観測信号からそれらのパラメータを推定する問題として楽器音分離を扱ったものである.

本稿ではこの重畳マルコフ遷移スペクトルモデルへの補助情報の利用を議論する. 重畳マルコフ遷移スペクトルモデルは多くの音楽信号に適用可能なモデルと期待されるが, その柔軟さ故に入力データに対して過学習を起こしてしまう危険が伴う. すなわち, 我々の目的は音楽信号を楽器音一音ごとに分解することであるにも関わらず, 例えば複数の楽器音を一連のマルコフ連鎖によって表現してしまう可能性が十分にある. そこで楽器音がどのようなものであるかの手掛かりとして, 解析したい入力信号以外に楽器音信号を付加することを検討する. これらの補助情報を与えることによって信号の分解を所望の結果に導くために, 楽器音分離を重畳マルコフ遷移スペクトルモデルの半教師つき学習問題として扱う枠組みを提案する.

2 重畳マルコフ遷移スペクトルモデルに基づく音楽信号分解

楽器音は多様な音色の変化によって音楽に彩りを与えており, 音楽信号から楽器音分離においてもこれらの時間的に変化するスペクトルパターンを抽出する技術が求められてきた. 例えばピアノの場合, 発音からアタック, ディケイ, サステイン, リリースといった音色変化を経て消音すると一般的に捉えられている. また, 歌声や弦楽器は, ヴィブラートのように基本周波数を変化させることで演奏に表情付けを行っている. 我々はこのような楽器音の非定常なスペクトル変化を表現するモデルとして重畳マルコフ遷移スペクトルモデルと呼ぶ音楽信号生成モデルを提案してきた.

今, D 個の楽器音によって構成された音楽信号 $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\,\Omega\times T}$ (ただし, $\omega = 1,\cdots,\Omega$ は周波数インデックス, $t = 1,\cdots,T$ は時間インデックスを表す) を考える. 以後, これらの D 個の構成要素は, NMF による音楽信号分解の通例に倣いコンポーネントと呼ぶことにする. ただし, 例えばピアノの C が楽曲中で複数回演奏された場合も, それらは一つのコンポーネントとして扱われることに注意して頂きたい. ここで, d 番目のコンポーネントは K 個の基底スペクトルパーツ $\boldsymbol{H} = \{(H^{(1)}_{\omega,d})_\Omega, \ldots, (H^{(K)}_{\omega,d})_\Omega\}$ を持っており, 各時刻のスペクトルはそのうちの一つのパーツから生成されたものと仮定する. このとき, 時刻 t に選ばれた基底スペクトルのインデックスを $Z_{d,t}$, 音量を $U_{d,t}$ とすると, d 番目のコンポーネントのスペクトログラム $(C_{\omega,t,d})_{\Omega\times T}$ は次のような生成モデルに基づいて生成されたものと見なすことが出来る:

$$C_{\omega,t,d} \sim \mathrm{Poisson}\Big(H^{(Z_{d,t})}_{\omega,d} U_{d,t}\Big) \quad (1)$$

$$H^{(k)}_{\omega,d} \sim \mathrm{Gamma}(a_H, b_H) \quad (2)$$

$$U_{d,t} \sim \mathrm{Gamma}(a_U, b_U). \quad (3)$$

例えばピアノの 1 コンポーネントを想像すると, 基底スペクトルパーツはアタック, ディケイ, サステイン, リリースの代表的なスペクトルパターンに対応し, 各時刻はそれらの一つがある音量で鳴ることでスペクトルが生成されたと見なしていることに相当している. ピアノの音色変化は上記 4 つの順に変化していくと言われるように, ある時刻でアタックの基底スペクトルが採用されているのであれば, 次の時刻は引き続きアタックであるか, もしくはディケイに移ることが予想される. またサステインが採用された次の時刻にはサステインを維持するかもしくはリリースに移ることが期待され, ディケイに戻るようなことは起こりにくいと期待される.

このように, 基底スペクトルの移り変わりはある種の頻出の規則があると考えられるため, これらを

*Semi-supervised approach to monaural music source separation based on Markov-chained spectrogram model. by Iwao TANUMA[1], Masahiro NAKANO[1], Suguru FUJITA[1], Hirokazu KAMEOKA[1 2], Shigeki SAGAYAMA[1], ([1]The University of Tokyo, [2]NTT Communication Science Laboratories)

Hidden Markov model(HMM) の状態遷移のように捉え, 状態遷移確率を表す $\boldsymbol{\pi} = (\pi_{k,k})_{K\times K}$ を用いて, 次のような 1 次のマルコフ連鎖として表現するの有効であると期待される:

$$Z_{d,t} \mid Z_{d,t-1}, (\pi_{d,k})_{k=1}^{K} \sim \boldsymbol{\pi}_{d,Z_{d,t-1}}. \tag{4}$$

以上より, 観測される音楽信号スペクトログラム $\boldsymbol{Y}$ はコンポーネントが重ねあわされたものとして

$$Y_{\omega,t} \sim \delta\left(Y_{\omega,t} - \sum_{d} C_{\omega,t,d}\right) \tag{5}$$

とみなし, 観測データから各パラメータを推定することで各コンポーネントへの分解が実現される. ただし, $\delta(A-B)$ は $A=B$ のとき 1 をとり, それ以外の場合は 0 をとるものとする. 本稿では出力分布として Poisson 分布を用い, 基底スペクトルと音量への事前分布として Gamma 分布を設定したが, 他にも様々な代替案が考えられる. 上記の生成モデルにおいて $K=1$ としたとき, I-divergence 規準の NMF のベイズ的な生成モデルとしての解釈と等価になっていることにも注意して頂きたい. 出力分布の選択や NMF との関係についての詳細は [1] を参照して頂きたい.

我々は上記のモデルを音楽信号に適用する場合におけるモデルの自由度の設計法についても検討を行ってきた. コンポーネントの数 D と各コンポーネントの持つ基底スペクトルパターンの個数 K の設定は信号に分解結果に重大な影響を与える. 例えば 10 種の楽器音で構成された曲に対し, $D=20$ と設定したモデルを適用した場合, パラメータは観測データを 20 個のコンポーネントで説明するように推定されるため, 本来一つのコンポーネントに割り当てられて欲しい一つの楽器音が複数のコンポーネントに分断されて割り当てられてしまう可能性が考えられる. 同様に基底スペクトルパターン数 K の設定も重要であり, 例えばピアノであれば 4 つ程度の状態で十分に表現しうることが期待されるが, ヴァイオリンのヴィブラートのような演奏表情はより多様な音色変化を伴っており, もっと多くの状態が必要であることが予想される. そこで, 我々はノンパラメトリックベイズと呼ばれる枠組みに基づいてこれらのモデル自由度を適切に扱うことを試みてきた. 詳細は [1] を参照して頂きたいが, 各コンポーネントに対し総音量 $\boldsymbol{\theta} = (\theta_d)_D$ を導入し, それらが Gamma process から生成されたと見なし, また各コンポーネントのスペクトル遷移を構成していた HMM に関して Hierarchical Dirichlet process HMM への拡張を行う. 生成モデルは $\boldsymbol{H}$ および $\boldsymbol{U}$ への事前分布に加え, 式 (1) に $\boldsymbol{\theta}$ を導入することで,

$$\begin{aligned}
C_{\omega,t,d} &\sim \mathrm{Poisson}(\theta_d H_{\omega,d} U_{d,t}) \\
\theta_d &\sim \mathrm{Gamma}(\eta/D, \eta\lambda) \\
\boldsymbol{\beta}_d &\sim \mathrm{GEM}(\gamma) \\
\boldsymbol{\pi}_{d,k} \mid \boldsymbol{\beta}_d &\sim \mathrm{DP}(\alpha, \boldsymbol{\beta}_d) \\
Z_{d,t} \mid Z_{d,t-1}, (\pi_{d,k})_{k=1}^{K} &\sim \boldsymbol{\pi}_{d,Z_{d,t-1}}.
\end{aligned}$$

と表すことが出来る. ただし, $\boldsymbol{\beta}_d \sim \mathrm{GEM}(\gamma)$ は

$$\beta'_k \sim \mathrm{Beta}(1,\gamma) \tag{6}$$

$$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l) \tag{7}$$

を表しており, D は K は無限大にすることで所望のモデルとなる. (η, λ, α, γ については [1] を参照して頂きたい) 本研究では後に述べる推論との兼ね合いで, 実用上は D や K を十分大きな値で打ち切ることによって近似したモデルを扱う. ここで重要なのは, 適切な D や K を見つける必要はなく, 十分に大きな D や K を設定しておけば良いことにある. 例えば 10 種の楽器音で構成された楽曲にこのモデルを適用した時, D=100 に設定した場合も D=10000 と設定した場合も主にアクティブになるコンポーネントは 10 程度となり他は抑圧することが期待される.

3 半教師あり学習による楽器音分離

重畳マルコフ遷移スペクトルモデルは多くの音楽信号に適用可能なモデルと期待されるが, その柔軟さ故に入力データに対して過学習を起こしてしまう危険が伴うため, 所望の分解結果を得るために補助情報を援用することが有効となることが期待される. 今, 解析したい音楽信号が与えられているとして, どのように補助情報を用いることが出来るかを考えてみたい. 例えば解析したい信号に, あるヴァイオリンが含まれているとする. もしその信号と同一環境下で録音された同一楽器の単音データベースが得られたとすると, これらは多重音から単音を抽出する際の大きな手掛かりとなると考えられる. 逆に手元にピアノの単音データベースが得られていたとしても, 解析したい信号中がピアノを含んでいない場合, これらは信号の分解において有効な手掛かりとしては機能しないと予想される. では, 解析したい信号に用いられたものではないが, 何らかのヴァイオリンのある環境下で録音されたデータベースが手元にあるとき, これらは有効に活用することは出来るだろうか. これは本当に活用出来るのかは定かではない. つまり, 手元に得られる補助情報の中から入力信号を解析する上で有効なものだけを選択的に用い活用するような仕組みが理想的には望まれる. 我々はこのような仕組みを実現する一つの方法として半教師あり学習として楽器音分離を扱う枠組みを提案する.

従来の楽器音分離 (特に近年の NMF に基づく音源分離) は大きく教師あり学習と教師なし学習と二つに分類することが出来る. 例として, ピアノの C, E, G の 3 種類の音で構成された楽曲を考える. この信号は, C の音が単音で鳴っている箇所や, E と G が重なってなっている箇所, 3 音全てが同時に発音している箇所がある. この信号に対し重畳マルコフ遷移スペクトルモデルを適用すると, モデルは主要なコンポーネントによって信号を説明しようと働く. このとき, 3 つのピアノの音は楽曲を構成する頻出のパーツであることから, 3 つの主要なコンポーネントとしてそれら楽器音一つ一つが表出してくることが期待される. このように, モデルは楽器音がどのようなものであるかの先見知識なしに, 入力信号だけを手がかりとして頻出のコンポーネントに分解することで楽器音分離が達成される. 統計モデルのパラメータ推定の観点から見た時, このような信号の分解は”教師なし学習問題” として捉えることが出来る.

では, もし入力信号において C と E が毎回同時に演奏されていた場合にこのモデルはどのように学習されるだろうか. この場合, モデルそれ自体には C と E を分解する手掛かり, すなわち楽器音とはどのよう

なものであるかの手掛かりはなく, おそらく 2 つをまとまった単音と見なし一つのコンポーネントして説明しようと働くことが予想される. そこで, あらかじめ各楽器音の単音データ (もしくはそれらに似た音色の単音データ) を用意して基底スペクトルパーツを事前に学習しておき, それらを用いて解析したい信号を分解する方法もしばしば用いられてきた. これは事前に “教師データ” を用いて基底スペクトルパーツを学習し, それらを用いて未知データを解析していることから, “教師あり学習問題” として捉えることが出来る.

本節の冒頭で述べた通り, “教師データ” の援用は楽器音分離への有用な手立てではあるが, 解析したい信号への手掛かりとして有効なデータのみを事前に用意出来るような状況は稀であり, 入力信号に応じて選択的に有効なデータのみを活用するような枠組みが求められている. 本研究では, 教師あり学習と教師なし学習両者の利点を生かしつつそれらの欠点を補うために半教師あり学習の枠組みを用いる.

本研究の目的は, 解析したい観測スペクトログラム $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\,\Omega\times T}$ をアクティブな D 個のスペクトログラム $[\boldsymbol{C}_1,\ldots,\boldsymbol{C}_d,\ldots,\boldsymbol{C}_D] = [(C_{\omega,t,1})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\,\Omega\times T},\ldots,(C_{\omega,t,d})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\,\Omega\times T},\ldots,(C_{\omega,t,D})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\,\Omega\times T}]$ に分解し, これら一つ一つを楽器音一音一音に対応させることである. このとき教師データとして D' 個の楽器音スペクトログラム $[\boldsymbol{Y'}_1,\ldots,\boldsymbol{Y'}_{d'},\ldots,\boldsymbol{Y'}_{D'}] = [(Y_{\omega,t,1})_{\Omega\times T_1} \in \mathbb{R}^{\geq 0,\,\Omega\times T_1},\ldots,(Y_{\omega,t,d'})_{\Omega\times T_{d'}} \in \mathbb{R}^{\geq 0,\,\Omega\times T_{d'}},\ldots,(Y_{\omega,t,D'})_{\Omega\times T_{D'}} \in \mathbb{R}^{\geq 0,\,\Omega\times T_{D'}}]$, そして各楽器音データのインデックスと分離されたスペクトログラムのインデックスの対応関係 $f : d' \in \{1,\ldots,D'\} \to d \in \{1,\ldots,D\}$ が入力として与えられるとし, このとき $\boldsymbol{C}_d$ には $f(d') = d$ となるような $\boldsymbol{Y'}_{d'}$ に類似する音が含まれることが期待される. ここで関数 f は単射とは限らない, すなわち 1 つの楽器音モデルを学習するにあたり複数の教師データを入力とすることが可能であることに注意されたい. 本稿では同一楽器かつ同一音高のものを共通の楽器音モデルで学習する. 教師データに適合しない要素は、教師なし学習と同様の枠組みで解析したい対象の信号のみを手掛かりとして自身を説明するパラメータを推定するよう働くことが期待される. 本稿では楽器音一音単位での分離が目的であるため, 上記のように各教師データのコンポーネントの割り当てを既知な情報として用いる場合のみを議論するが, 例えば [8] のように楽器単位で (ドラムとそれ以外のように) 分離を行うことが目的である場合は, 楽器名の情報だけを使う方法も考えられる.

D' 個のスペクトログラム $(Y'_{\omega,t,1})_{\Omega\times T_1} \in \mathbb{R}^{\geq 0,\,\Omega\times T_1},\ldots,(Y'_{\omega,t,D'})_{\Omega\times T_{D'}} \in \mathbb{R}^{\geq 0,\,\Omega\times T_{D'}}$ が教師データとして入力される楽器音信号とする. 本稿では教師データが 1 つの楽器音に割り当てられている状況を想定しており, 教師データは分離信号と楽器音モデルを共有することで,

$$Y'_{\omega,t,d'} \sim \mathrm{Poisson}\Big(\theta'_{d'} H^{(Z'_{d',t})}_{\omega,f(d')} U'_{d',t}\Big) \quad (8)$$

$$Z'_{d',t} \sim \boldsymbol{\pi}_{f(d'),Z'_{d',t-1}} \quad (9)$$

$$U'_{d',t} \sim \mathrm{Gamma}(a_U, b_U) \quad (10)$$

$$\theta'_{d'} \sim \mathrm{Gamma}(\eta, \eta\lambda^{l}_{d'}) \quad (11)$$

と表わされる.

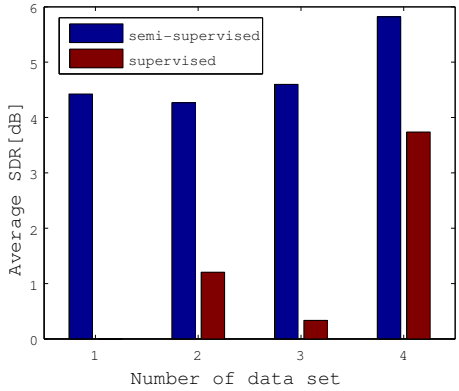

Fig. 1　各楽器音における SDR の平均値による予備実験の評価. 半教師あり学習により少ない学習データにおいても高い分離性能を保っていることが確認できる.

提案手法の推論に関しては, 主にマルコフ連鎖モンテカルロ法と用いる方法と変分ベイズに基づく基づく方法が考えられる. 紙面の都合上推論の詳細は省略するが, 以下の実験では変分ベイズ法 [9] により推論を行った結果を示す.

4　評価実験

4.1　予備実験

RWC 楽器音データベース [10] を用いて本手法において半教師あり学習の枠組みがどれほど寄与するかの評価を行うための実験を行った. バイオリン (RWC-MDB-I 2001 No.15) の 5 種類の奏法に関してそれぞれ C5, E5, G5 の 3 音ずつを用意し, そのうち 1 つを合成し, 観測信号とした. 残りの 4 種類のデータを教師データとし, 同一音高の教師データは同一の分離信号に対応付け, 分離実験を行った. 観測信号, 教師データ共に短時間フーリエ変換 (サンプリング周波数 16kHz, フレーム長 64ms, フレームシフト 32ms, Hanning 窓) により計算した振幅スペクトログラムを入力とした. 各パラメータについては, $a_H = b_H = 0.01$, $a_U = b_U = 1$, $\alpha = \gamma = 2$, $\eta = 0.1$, ガンマプロセスと棒折り過程における打ち切り数は 15 とした. 本手法と, その比較手法として, 提案手法における H と $n_{d,j,k}$ を教師データのみから更新するように変更した教師あり学習に基づく手法に対して実験を行った. 評価には C5, E5, G5 の平均 SDR(Signal to Distortion Ratio) を用いた [11].

バイオリン C5, E5 ,G5 各音の教師データが対応する分離スペクトログラムの SDR の平均を図 1 に示す. 教師ありの手法では, 顕著に性能の低下が確認された. これはほとんどが未知楽器に分類されてしまっていることに起因する. それに対して, 提案法の半教師あり学習に基づくものは, 概ね高い分離性能が得られた. また, いずれにおいても教師データを増やすことによって分離性能が向上するような傾向が確認された.

4.2　実際の楽曲に対する手法の適用

音楽音響信号の分離において, 本手法の有用性を示すため, バイオリンとピアノのパートによって構成さ

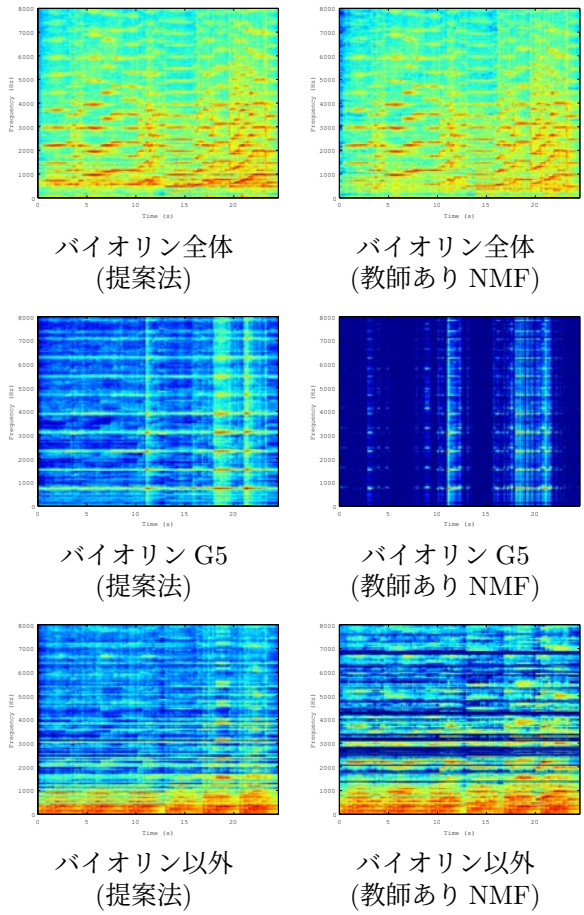

Fig. 2　左側が提案法, 右側が教師あり NMF による分離結果, 1 音高単位 (中段) で見ると教師あり NMF ではビブラートが取れておらず, また未知楽器成分 (下段) にバイオリン成分が多く残っている.

れた楽曲 (RWC-MDB-C-2001 No.40) を用いて従来手法 (教師あり NMF) との比較実験を行った [12]. 教師データとしてバイオリン (RWC-MDB-I 2001 No.13) の全音高分の楽器音 (46 音) を与えた. パラメータ等の条件はガンマプロセスの打ち切り数を 70 とした以外予備実験と同じ条件である. 従来手法の教師あり NMF は 1 音高につき 3 つの基底スペクトルで学習, すなわち 138 個の基底スペクトルを事前に学習し, 未知楽器用基底スペクトルとして 20 個, 合計 158 個の基底スペクトルを用いて分離を行った.

図 2 はそれぞれ提案法と教師あり NMF によって分離されたスペクトログラムにおける, バイオリン成分を全て足し合わせたもの (上段), バイオリン単音 (G5) のスペクトログラム (中段), そしてバイオリン成分以外を全て足し合わせたスペクトログラムである. バイオリン全体のスペクトログラムを見るとあまり差異は無いが, 1 音単位で見ると教師あり NMF ではビブラートが取れていないことが確認された. このような時間的に非定常な楽器音に対しては重畳マルコフ連鎖スペクトルが非常に有効であると考えられる. またピアノ成分のみが残ると期待されるバイオリン以外のスペクトログラムにおいても, バイオリンの音が多く残っている. これは教師データへ過剰な適応により, 教師データと観測信号におけるバイオリン成分との差が残ってしまったものだと考えられる. それに対して, 提案法ではバイオリンの成分は殆ど残っておらず, 半教師あり学習の枠組みによって, このような教師データと観測信号に含まれる楽器音における音響的なギャップを埋めたものと考えられる.

5　おわりに

本稿では, 音楽信号からの楽器音分離において, 信号を楽器音一音ごとに分解するために, 補助情報として楽器音の単音信号を援用し半教師あり学習問題として扱う枠組みを示した. 今後は楽器音認識や多重音解析を同時に実現するような方向へ拡張することを検討したい.

参考文献

[1] M. Nakano, *et al.*, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden markov model," in *Proc. of WASPAA*, pp. 325–328, IEEE, 2011.

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription, " *Proc. of WASPAA.*, pp. 177–180, IEEE, 2003.

[3] D. Lee, *et al.*, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[4] A. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, p. 17, 2009.

[5] M. Schmidt, *et al.*, "Bayesian non-negative matrix factorization," *Independent Component Analysis and Signal Separation*, pp. 540–547, 2009.

[6] P. Smaragdis, *et al.*, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[7] M. D. Hoffman, *et al.*, "Bayesian nonparametric matrix factorization for recorded music," *Proc. of ICML*, pp. 439–446, 2010.

[8] M. Kim, *et al.* , "Nonnegative Matrix Partial Co-Factorization for Spectral and Temporal Drum Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, vol.5, no.6, pp.1192-1204, Oct. 2011

[9] Y. Tehz, *et al.*, "Collapsed variational inference for hdp," *NIPS*, vol. 20, no. 20, pp. 1481–1488, 2008.

[10] M. Goto, *et al.*, "Rwc music database: Music genre database and musical instrument sound database," in *Proc. of ISMIR*, pp. 229–230, 2003.

[11] E. Vincent, *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. on ASLP* , vol. 14, no. 4, pp. 1462–1469, 2006.

[12] M. Goto, *et al.*, "RWC Music Database: Popular , Classical, and Jazz Music Databases," in *Proc. of ISMIR*, pp. 287–288, 2002.