Vol.2021-MUS-131 No.10
Vol.2021-SLP-137 No.10
2021/6/18

基底共有型非負値行列因子分解を用いた楽器音の音色変換

北村 大地[1]　香西 海斗[1,†1]

概要：複数の楽器音間の音色の違いを解析する手法として，基底共有型非負値行列因子分解（basis-shared nonnegative matrix factorization: BSNMF）が提案されている．この手法では，複数の楽器音の振幅スペクトログラムを低ランクな非負行列で近似表現する際に，一部の基底ベクトルを共有することで，入力楽器音間の共通音色成分及び各入力楽器音の固有音色成分をそれぞれスペクトルパターンとして同時に推定する．本稿では，この BSNMF で得られる共通・固有スペクトルパターンを用いて，複数の入力楽器音の音色を入れ替える変換を実現する手法を提案する．この音色変換法の利点として，複数の楽器音のパラレルデータ（同一の旋律を異なる楽器で演奏したデータ）が無い場合でも，信号中に同じ音高や和音が含まれていれば音色変換ができる点が挙げられる．実験では，変換前の 2 つの楽器音と変換後の楽器音を被験者に聴取してもらう XAB 法に基づく主観評価を実施し，主観的にも認知できる品質で音色変換が成功することを示す．

Timbre Conversion of Musical Instruments Using Basis-Shared Nonnegative Matrix Factorization

Kitamura Daichi[1]　Kozai Kaito[1,†1]

Abstract: Basis-shared nonnegative matrix factorization (BSNMF) was proposed to analyze the difference of timbres between multiple musical instruments. This method approximates observed amplitude spectrograms of the multiple musical instrument signals by low-rank nonnegative matrices, where some basis vectors are shared within the matrices. As a result, common and individual spectral patterns between the musical instruments can be simultaneously estimated. In this paper, we propose BSNMF-based timbre conversion of the input musical instruments. The advantage of the proposed timbre conversion is that the parallel data of the musical instruments (signals of the same melody played with the different musical instruments) are not necessary. In the experiment, we conduct an XAB subjective test using two original instrumental sound and one timbre-converted sound, which validates quality of timbre conversion of the proposed method.

1. はじめに

一般的に，個人の演奏や楽器本体の芸術的価値は，一定の品質を超える範囲において，評価者の主観に基づいて評価される．例えば，「アマチュア奏者とプロフェッショナル奏者の演奏の差異」や「安価な楽器と高価な楽器の奏でる音の違い」が主観的に語られることは多い．しかしながら，これらの観点について主観を廃して定量的に議論する方法はあまり確立されておらず，とくに音楽演奏や楽器に対して精通していない者にとっては，芸術的価値を判断する材料がないため他者の主観的評価に頼らざるを得ない．

そこで著者らは，複数の楽器音信号間の差異を客観的かつ定量的に議論するための音響特徴量抽出手法の構築を目指し，新たな特徴量抽出手法を提案している [1]．本手法は, 非負値行列因子分解（nonnegative matrix factorization: NMF）[2], [3] と呼ばれる行列分解理論を用いて，複数の楽器音信号間の「共通する音響特徴量」及び「固有の音響特徴量」を同時に推定・抽出する新しいアルゴリズムである．NMF における基底行列を複数の楽器音信号のモデル間で共有していることから，基底共有型 NMF（basis-shared NMF: BSNMF）と呼ぶ．

[1] 香川高等専門学校
National Institute of Technology, Kagawa College
[†1] 現在，豊橋技術科学大学
Presently with Toyohashi University of Technology

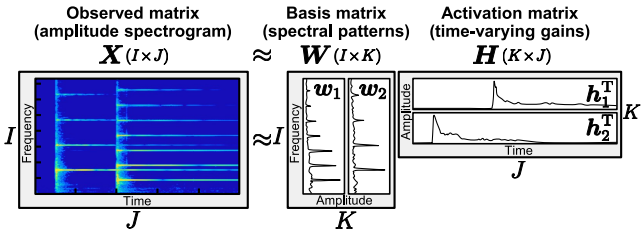

Fig. 1 NMF for audio signals, where $K = 2$.

音響信号の解析では，これまでにピッチ，スペクトル包絡，メル周波数ケプストラム係数 [4] 等様々な特徴量が考案され，広く利用されている．また，楽器音の物理現象を対象とした解析も歴史は古く，ピアノやヴァイオリン等，個々の楽器の物理音響的側面から発音機構が解析されており [5]，2000 年以降では，楽器同定の分野で楽器音の音響特徴量が各種検討されている [6], [7]．著者らの提案する BSNMF で得られる音響特徴量が上記の歴史的に有名な特徴量と異なるのは，音響信号としての絶対的な特徴量ではなく，入力された複数の楽器音信号間の相対的な差異を表すことを目的としている点が挙げられる．このような複数の楽器音信号間の相対的な特徴量は，アマチュア奏者がより良い演奏をするために必要な技術の提示や，より芸術的価値の高い楽器の設計製作等に役立てることができる他，楽器音信号の音色変換にも応用することが可能である．

本稿では，BSNMF で推定される共通及び固有音響特徴量が有用であることを確認するために，これらの特徴量を用いた楽器音の音色変換アルゴリズムを提案する．また，主観評価実験を実施し，音色変換の精度について調査する．

2. 既存手法

2.1 NMF の概要

NMF [2], [3] は，次式に示すように，非負の観測行列を別の二つの非負行列の行列積に近似的に分解する．

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

ここで，$\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{I \times J}$ は全要素が非負の観測行列であり，$\boldsymbol{W} = [\boldsymbol{w}_1\ \boldsymbol{w}_2\ \cdots\ \boldsymbol{w}_K] \in \mathbb{R}_{\geq 0}^{I \times K}$ 及び $\boldsymbol{H} = [\boldsymbol{h}_1\ \boldsymbol{h}_2\ \cdots\ \boldsymbol{h}_K]^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{K \times J}$ は NMF で推定すべき非負変数行列である．また，$\cdot^{\mathrm{T}}$ は転置を表す．$\boldsymbol{W}$ 及び $\boldsymbol{H}$ はそれぞれ基底行列及びアクティベーション行列と呼ばれる．$\boldsymbol{w}_k$ は基底ベクトルと呼ばれ，その本数 K は $K \ll \min(I, J)$ となるように設定される．ここで，$k = 1, 2, \cdots, K$ は基底ベクトルのインデクスを示す．従って，NMF は $\boldsymbol{X}$ を $\boldsymbol{W}\boldsymbol{H}$ で低ランク近似する行列分解であり，$\boldsymbol{X}$ 中に頻出する少数（K 個）の潜在的なパターンを基底ベクトルとして抽出できる [2]．

NMF を音響信号に適用する場合，短時間フーリエ変換（short-time Fourier transform: STFT）を経て得られる振幅（又はパワー）スペクトログラムを非負観測行列 $\boldsymbol{X}$ とするのが一般的である．この場合，Fig. 1 に示すように，音響信号中の頻出スペクトルが $\boldsymbol{w}_k$ として得られ，さらに各スペクトルの時間的強度変化が $\boldsymbol{h}_k$ となる．NMF は音響信号中のスペクトルパターンを抽出できるため，音楽信号解析 [8] や音源分離 [9], [10], [11], [12] 等に適用される．

2.2 NMF における変数行列の最適化

NMF は次式の最適化問題により各変数行列を推定する．

$$\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{D}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H})\ \text{ s.t. } w_{ik}, h_{kj} \geq 0\ \forall i, j, k \tag{2}$$

ここで，$\boldsymbol{X}$ は観測された音響信号に STFT を適用して得られる振幅スペクトログラム，w_{ik} 及び h_{kj} はそれぞれ $\boldsymbol{W}$ 及び $\boldsymbol{H}$ の要素，$i = 1, 2, \cdots, I$ 及び $j = 1, 2, \cdots, J$ はそれぞれ周波数ビン及び時間フレームのインデクスを示す．また，$\mathcal{D}(\cdot|\cdot)$ は 2 つの入力行列間の類似度を測る関数である．本稿では，次式で表される二乗 Euclid 距離を用いる．

$$\mathcal{D}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) = \sum_{i,j} \left(x_{ij} - \sum_k w_{ik} h_{kj} \right)^2 \tag{3}$$

このとき，変数行列 $\boldsymbol{W}$ 及び $\boldsymbol{H}$ は，いずれも非負の乱数で初期化したうえで，コスト関数値が収束するまで次の乗算型更新式を反復的に計算することで最適化できる [3]．

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\boldsymbol{X}\boldsymbol{H}^{\mathrm{T}}}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}} \tag{4}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}}{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H}} \tag{5}$$

ここで，$\odot$ 及び行列の分数はそれぞれ要素毎の積及び商を表す．式 (4) 及び (5) は補助関数法 [13] と呼ばれる最適化アルゴリズムで導出されたものであり，毎反復でのコスト関数の単調非増加性が理論的に保証されている．

2.3 BSNMF の分解モデル

BSNMF [1] では，複数の音響信号間の共通成分及び固有成分を推定する．いま，n 番目の観測信号の振幅スペクトログラムを $\boldsymbol{X}_n \in \mathbb{R}_{\geq 0}^{I \times J_n}$（$n = 1, 2, \cdots, N$ は観測信号のインデクス）と表すとき，次の連立低ランク近似を考える．

$$\begin{cases} \boldsymbol{X}_1 \approx \boldsymbol{W}\boldsymbol{H}_1 + \boldsymbol{F}_1\boldsymbol{H}_1 \\ \boldsymbol{X}_2 \approx \boldsymbol{W}\boldsymbol{H}_2 + \boldsymbol{F}_2\boldsymbol{H}_2 \\ \quad \vdots \\ \boldsymbol{X}_N \approx \boldsymbol{W}\boldsymbol{H}_N + \boldsymbol{F}_N\boldsymbol{H}_N \end{cases} \tag{6}$$

ここで，$\boldsymbol{W}$ は全観測信号のモデルで共有される基底行列であり，$\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ 間で共通のスペクトルパターンを K 個含む．従って，行列 $\boldsymbol{W}\boldsymbol{H}_n$ は共通スペクトルパターンで表された $\boldsymbol{X}_n$ 中の成分（共通スペクトル成分）となる．一方，基底行列 $\boldsymbol{F}_n = [\boldsymbol{f}_{n1}\ \boldsymbol{f}_{n2}\ \cdots\ \boldsymbol{f}_{nK}] \in \mathbb{R}_{\geq 0}^{I \times K}$ は n 番目の観測信号 $\boldsymbol{X}_n$ にのみ含まれる固有のスペクトル成分を K 個含んでおり，$\boldsymbol{W} + \boldsymbol{F}_k$ として共有基底行列 $\boldsymbol{W}$

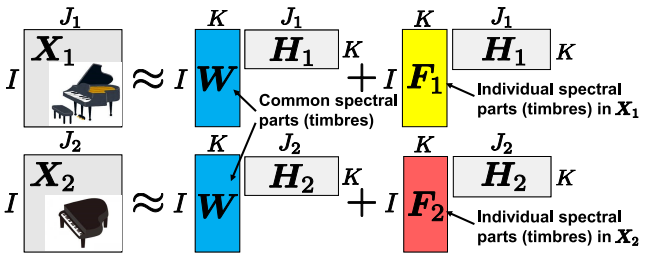

Fig. 2 Decomposition model in proposed basis-shared NMF, where $N = 2$.

と合わせることで, $\boldsymbol{X}_n$ 中のスペクトルパターンとなる．すなわち，固有基底行列 $\boldsymbol{F}_n$ とアクティベーション行列 $\boldsymbol{H}_n = [\boldsymbol{h}_{n1}\ \boldsymbol{h}_{n2}\ \cdots\ \boldsymbol{h}_{nK}]^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{K \times J_n}$ の行列積 $\boldsymbol{F}_n\boldsymbol{H}_n$ は，$\boldsymbol{X}_n$ にのみ含まれる成分（固有スペクトル成分）を表現する．このような分解から，N 個の観測信号中の共通及び固有成分がそれぞれ推定できる．

なお，式 (6) では，固有成分のアクティベーション行列 $\boldsymbol{H}_n$ を共有基底行列 $\boldsymbol{W}$ と固有基底行列 $\boldsymbol{F}_n$ の間で共有することで，K 本の基底ベクトルのそれぞれに対する共通・固有成分への分解（$\boldsymbol{w}_k$ 及び $\boldsymbol{f}_k$）を実現している．基底ベクトル単位での共通・固有成分への分解を必要としない場合は，式 (6) の各近似式を $\boldsymbol{X}_n \approx \boldsymbol{W}\boldsymbol{H}_n + \boldsymbol{T}_n\boldsymbol{G}_n$ に変更すればよい．ここで，$\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times K_n}$ 及び $\boldsymbol{G}_n \in \mathbb{R}_{\geq 0}^{K_n \times J_n}$ である．このモデルでは，$\boldsymbol{W}\boldsymbol{H}_n$ 及び $\boldsymbol{T}_n\boldsymbol{G}_n$ としてスペクトログラム単位での共通・固有成分の抽出が可能となる他，固有成分と共通成分で異なる基底数を設定できる利点がある．両分解モデルは，推定された共通・固有成分の用途に応じて選択すればよい．本稿で示す音色変換への活用では，その実現方法に起因して，式 (6) を用いる必要がある．

2.4 NMF に基づく既存の音色変換手法

NMF に基づく音色変換は，これまでに声質変換 [14], [15], [16] を目的とした手法が提案されている．この手法では，変換前の話者（ソース）と変換後の話者（ターゲット）の同一発話内容のサンプル信号（パラレルデータ）を用意し，dynamic time warping 等を適用してサンプル信号間の時間フレームの同期を取った 2 つのデータを NMF の基底行列として利用する．入力音声信号をソースの基底行列とアクティベーション行列でモデル化し，基底行列をターゲットの基底行列に差し替えることで声質変換が実現できる．同様のアプローチは楽器音の音色変換にも応用できると思われるが，パラレルデータを用意するコストが生じるほか，サンプル信号間の時間フレームの同期が正しく取れない場合は変換精度の劣化を招くことが予想される．

NMF を用いたより強力な音色変換として，オーディオモザイキングのための手法が提案されている [17], [18]．この手法では，ターゲットの音響信号の非負スペクトログラムを NMF の観測行列とし，さらにソースの音響サンプルの非負スペクトログラムそのものを基底行列としたうえでアクティベーション行列を推定しており，「音楽」と「蜂の羽音」等のように音色が全く異なる音響信号同士の音色変換さえ実現している[*1]．この手法は信号間で時間フレームの同期を取る必要が無いという利点があるが，アクティベーション行列の推定時に「要素音が時間的に連続しすぎない」・「同時に生起する要素音数が多くなりすぎない」・「要素音が時間的に途切れない」等の制約を与えて音質を担保せねばならず，各制約の強さの調整が必要となる．

3. 提案手法

3.1 BSNMF を用いた音色変換

本稿では，BSNMF で推定できる共通・固有スペクトル成分の活用例として，2 種の楽器音信号間の音色変換を実現するアルゴリズムを提案する．提案手法では，2 種の楽器音信号の振幅スペクトログラムをそれぞれ $\boldsymbol{X}_1$ 及び $\boldsymbol{X}_2$ とし，式 (6) や Fig. 2 のように BSNMF でモデル化する．変数行列 $\boldsymbol{W}$，$\boldsymbol{F}_1$，$\boldsymbol{F}_2$，$\boldsymbol{H}_1$，及び $\boldsymbol{H}_2$ を推定後は，固有基底行列 $\boldsymbol{F}_1$ 及び $\boldsymbol{F}_2$ のみを次式のように入れ替える．

$$\begin{cases} \boldsymbol{W}\boldsymbol{H}_1 + \boldsymbol{F}_2\boldsymbol{H}_1 = \boldsymbol{Y}_1 \\ \boldsymbol{W}\boldsymbol{H}_2 + \boldsymbol{F}_1\boldsymbol{H}_2 = \boldsymbol{Y}_2 \end{cases} \tag{7}$$

式 (7) では，共有基底行列及びアクティベーション行列を変えずに固有基底行列のみを交換しているため，$\boldsymbol{Y}_1 \in \mathbb{R}_{\geq 0}^{I \times J_1}$ は，理想的には「$\boldsymbol{X}_2$ の音色に変換された $\boldsymbol{X}_1$ のメロディを含む音響信号の振幅スペクトログラム」となる．同様に，$\boldsymbol{Y}_2 \in \mathbb{R}_{\geq 0}^{I \times J_2}$ は「$\boldsymbol{X}_1$ の音色に変換された $\boldsymbol{X}_2$ のメロディを含む音響信号の振幅スペクトログラム」として合成される．

3.2 二乗 Euclid 距離基準 BSNMF の反復更新式

式 (6) の変数行列 $\boldsymbol{W}$，$\boldsymbol{F}_n$，及び $\boldsymbol{H}_n$ を最適化する反復更新式は，通常の NMF と同様に補助関数法を用いて導出できる．文献 [1] では，一般化 Kullback–Leibler ダイバージェンスを類似度関数に用いた場合の反復更新式が導出されているが，音色変換の用途においては式 (3) の二乗 Euclid 距離を用いたほうが高音質となることを実験的に確認している．そこで本稿では，二乗 Euclid 距離基準 BSNMF の補助関数法に基づく反復更新式の導出を示す．

最適化問題のコスト関数は次式となる．

$$\begin{aligned} \mathcal{J} &= \sum_{n,i,j_n} (x_{ij_nn} - \hat{x}_{ij_nn})^2 \\ &= \sum_{n,i,j_n} \left[x_{ij_nn}^2 + \hat{x}_{ij_nn}^2 - 2x_{ij_nn}\hat{x}_{ij_nn}\right] \end{aligned} \tag{8}$$

$$\hat{x}_{ij_nn} = \sum_k w_{ik}h_{kj_nn} + \sum_k f_{ikn}h_{kj_nn} \tag{9}$$

ここで，$j_n = 1, 2, \cdots, J_n$ は $\boldsymbol{X}_n$ の時間フレームのインデ

[*1] 音色変換の音響サンプル：https://www.audiolabs-erlangen.de/resources/MIR/2015-ISMIR-LetItBee

クスを示す．また，x_{ij_nn}，f_{ikn}，及び h_{ij_n} はそれぞれ $\boldsymbol{X}_n$，$\boldsymbol{F}_n$，及び $\boldsymbol{H}_n$ の要素である．

反復更新式を導出するために，式 (8) の第 2 項に関して Jensen の不等式

$$\begin{aligned}\hat{x}_{ij_nn}^2 &= \left[\sum_k w_{ik}h_{kj_nn} + \sum_k f_{ikn}h_{kj_nn}\right]^2 \\ &= \left[\sum_k \alpha_{ij_nk}\frac{w_{ik}h_{kj_nn}}{\alpha_{ij_nk}} + \sum_k \beta_{ij_nk}\frac{f_{ikn}h_{kj_nn}}{\beta_{ij_nk}}\right]^2 \\ &\leq \sum_k \alpha_{ij_nk}\left(\frac{w_{ik}h_{kj_nn}}{\alpha_{ij_nk}}\right)^2 + \sum_k \beta_{ij_nk}\left(\frac{f_{ikn}h_{kj_nn}}{\beta_{ij_nk}}\right)^2 \\ &= \sum_k \left(\frac{w_{ik}^2h_{kj_nn}^2}{\alpha_{ij_nk}} + \frac{f_{ikn}^2h_{kj_nn}^2}{\beta_{ij_nk}}\right) \qquad (10)\end{aligned}$$

を適用し，$\mathcal{J} \leq \mathcal{J}^+$ なる補助関数 $\mathcal{J}^+$ を設計する．

$$\begin{aligned}\mathcal{J}^+ = \sum_{n,i,j_n}\Bigg[&x_{ij_nn}^2 + \sum_k\left(\frac{w_{ik}^2h_{kj_nn}^2}{\alpha_{ij_nk}} + \frac{f_{ikn}^2h_{kj_nn}^2}{\beta_{ij_nk}}\right) \\ &- 2x_{ij_nn}\left(\sum_k w_{ik}h_{kj_nn} + \sum_k f_{ikn}h_{kj_nn}\right)\Bigg] \qquad (11)\end{aligned}$$

ここで，$\alpha_{ij_nk}, \beta_{ij_nk} > 0$ は $\sum_k \alpha_{ij_nk} + \sum_k \beta_{ij_nk} = 1$ を満たす補助変数であり，式 (10) の等号成立条件は

$$\alpha_{ij_nk} = \frac{w_{ik}h_{kj_nn}}{\sum_{k'} w_{ik'}h_{k'j_nn} + \sum_{k'} f_{ik'n}h_{k'j_nn}} \qquad (12)$$

$$\beta_{ij_nk} = \frac{f_{ikn}h_{kj_nn}}{\sum_{k'} w_{ik'}h_{k'j_nn} + \sum_{k'} f_{ik'n}h_{k'j_nn}} \qquad (13)$$

となる．即ち，補助変数が式 (12) 及び (13) を満たすとき唯一，$\mathcal{J}^+$ は補助変数に関して最小化され $\mathcal{J}^+ = \mathcal{J}$ となる．

次に，補助関数 $\mathcal{J}^+$ を各変数で偏微分し反復更新式を求める．$\partial\mathcal{J}^+/\partial w_{ik} = 0$ より，

$$\sum_{n,j_n}\left(2\frac{w_{ik}h_{kj_nn}^2}{\alpha_{ij_nk}} - 2x_{ij_nn}h_{kj_nn}\right) = 0 \qquad (14)$$

が得られ，式 (14) を w_{ik} について解くと次式となる．

$$w_{ik} = \frac{\sum_{n,j_n} x_{ij_nn}h_{kj_nn}}{\sum_{n,j_n} h_{kj_nn}^2/\alpha_{ij_nk}} \qquad (15)$$

従って，式 (12) 及び (13) による補助変数の更新と式 (15) による w_{ik} の更新を反復することで，本来のコスト関数である $\mathcal{J}$ を間接的に最小化できる．式 (12) 及び (13) を式 (15) に代入することで，より簡潔な反復更新式が得られる．

$$w_{ik} \leftarrow w_{ik}\frac{\sum_{n,j_n} x_{ij_nn}h_{kj_nn}}{\sum_{n,j_n}\left(\sum_{k'} w_{ik'}h_{k'j_nn} + \sum_{k'} f_{ik'n}h_{k'j_nn}\right)h_{kj_nn}} \qquad (16)$$

式 (16) を行列形式で表すと次式となる．

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\sum_n \boldsymbol{X}_n\boldsymbol{H}_n^{\mathrm{T}}}{\sum_n (\boldsymbol{W}\boldsymbol{H}_n + \boldsymbol{F}_n\boldsymbol{H}_n)\,\boldsymbol{H}_n^{\mathrm{T}}} \qquad (17)$$

同様に $\partial\mathcal{J}^+/\partial f_{ikn} = 0$ 及び $\partial\mathcal{J}^+/\partial h_{kj_nn} = 0$ より，$\boldsymbol{F}_n$ 及び $\boldsymbol{H}_n$ の反復更新式は次のように得られる．

$$\boldsymbol{F}_n \leftarrow \boldsymbol{F}_n \odot \frac{\boldsymbol{X}_n\boldsymbol{H}_n^{\mathrm{T}}}{(\boldsymbol{W}\boldsymbol{H}_n + \boldsymbol{F}_n\boldsymbol{H}_n)\,\boldsymbol{H}_n^{\mathrm{T}}} \qquad (18)$$

$$\boldsymbol{H}_n \leftarrow \boldsymbol{H}_n \odot \frac{(\boldsymbol{W} + \boldsymbol{F}_n)^{\mathrm{T}}\,\boldsymbol{X}_n}{(\boldsymbol{W} + \boldsymbol{F}_n)^{\mathrm{T}}\,(\boldsymbol{W}\boldsymbol{H}_n + \boldsymbol{F}_n\boldsymbol{H}_n)} \qquad (19)$$

3.3 スケールフィッティング

式 (17)–(19) を反復計算することで，変数 $\boldsymbol{W}$，$\boldsymbol{F}_n$，及び $\boldsymbol{H}_n$ を最適化できる．その後，式 (7) のように固有基底行列 $\boldsymbol{F}_1$ 及び $\boldsymbol{F}_2$ を入れ替えてスペクトログラム $\boldsymbol{Y}_1$ 及び $\boldsymbol{Y}_2$ を得る．しかしながら，式 (7) において共有成分 $\boldsymbol{W}\boldsymbol{H}_n$ と固有成分 $\boldsymbol{F}_n\boldsymbol{H}_n$ がそれぞれどの程度のパワーを持っているかは楽器音信号 n によって異なる．この原因は，BSNMF の最適化問題が「$\boldsymbol{X}_n$ と $\boldsymbol{W}\boldsymbol{H}_n + \boldsymbol{F}_n\boldsymbol{H}_n$ の類似度の最小化」として定式化されているためである．即ち，楽器音信号 n 間で基底行列 $\boldsymbol{W}$ は共有されているが，$\boldsymbol{W}$ と $\boldsymbol{H}_n$ の間にはスケールの任意性（$a > 0$ に対して $\boldsymbol{W}\boldsymbol{H}_n = (a\boldsymbol{W})(\boldsymbol{H}_n/a)$）が存在するため，共有成分 $\boldsymbol{W}\boldsymbol{H}_n$ と固有成分 $\boldsymbol{F}_n\boldsymbol{H}_n$ のパワーバランスは楽器音信号 n によって異なる．この状態で固有基底行列 $\boldsymbol{F}_n$ を $\boldsymbol{F}_m$（ただし，$m \neq n$ は変換先の楽器音信号のインデクス）に入れ替えても，構成されるスペクトログラム $\boldsymbol{W}\boldsymbol{H}_n + \boldsymbol{F}_m\boldsymbol{H}_n$ は歪んだ信号となってしまう．

この問題の解決策として，固有基底行列 $\boldsymbol{F}_n$ を $\boldsymbol{F}_m$ に入れ替えた後に，$\boldsymbol{F}_m$ 中の各基底ベクトル $\boldsymbol{f}_{km}$ のスケールを観測信号 $\boldsymbol{X}_n$ のスケールにフィッティングさせる処理を施す．いま，楽器音信号 $\boldsymbol{X}_n$ を m 番目の楽器音信号の音色に変換する場合，次式のモデルを仮定する．

$$\boldsymbol{X}_n \approx \boldsymbol{W}\boldsymbol{H}_n + (\boldsymbol{F}_m\boldsymbol{D}_n)\boldsymbol{H}_n \qquad (20)$$

ここで，スケール行列 $\boldsymbol{D}_n \in \mathbb{R}_{\geq 0}^{K\times K}$ は各固有基底ベクトル $\boldsymbol{f}_{km}$ の大きさを定める係数 $d_{kn} \geq 0$ を対角要素に持つ対角行列である．式 (20) の分解モデルにおいて，左辺と右辺をできるだけ近似するスケール行列 $\boldsymbol{D}_n$ を求めることでスケールフィッティングができる．これは，次の最適化問題に帰着する．

$$\min_{\boldsymbol{D}_n} \mathcal{D}\left[\boldsymbol{X}_n | \boldsymbol{W}\boldsymbol{H}_n + (\boldsymbol{F}_m\boldsymbol{D}_n)\boldsymbol{H}_n\right] \quad \text{s.t. } d_{kn} \geq 0\ \forall k, n \qquad (21)$$

式 (21) の最適化問題は，通常の NMF と同様に補助関数法に基づく反復更新式として解くことができる．導出は 3.2 節と同様であるため割愛するが，二乗 Euclid 距離のコスト関数では次式の反復更新式が導かれる．

$$\boldsymbol{D}_n \leftarrow \boldsymbol{D}_n \odot \frac{\boldsymbol{F}_m^{\mathrm{T}}\boldsymbol{X}_n\boldsymbol{H}_n^{\mathrm{T}}}{\boldsymbol{F}_m^{\mathrm{T}}\left[\boldsymbol{W}\boldsymbol{H}_n + (\boldsymbol{F}_m\boldsymbol{D})\boldsymbol{H}_n\right]\boldsymbol{H}_n^{\mathrm{T}}} \qquad (22)$$

この反復更新式で $\boldsymbol{D}_n$ を求めることで，音色変換された楽器音信号は $\boldsymbol{Y}_n = \boldsymbol{W}\boldsymbol{H}_n + (\boldsymbol{F}_m\boldsymbol{D}_n)\boldsymbol{H}_n$ として得られる．

この振幅スペクトログラムに，$\boldsymbol{X}_n$ に対応する位相スペクトログラムを付与して逆 STFT を施すことで，歪みのない音色変換された楽器音信号が得られる．

4. 主観評価実験

4.1 実験条件

本実験では，MIDI 音源で作成した楽器音信号 $\boldsymbol{X}_1$ 及び $\boldsymbol{X}_2$ を用いた．各楽器音信号は，Fig. 3 の楽譜に基づき，ピアノ音源 Iowa Piano[*2]及び Sketch Upright Piano[*3]により電子的に生成した．ここで，Iowa Piano 及び Sketch Upright Piano はそれぞれグランドピアノ及びアップライトピアノのサンプリング音源である．STFT の窓長及びシフト長はそれぞれ 92.9 ms 及び 23.2 ms とし，窓関数は Hamming 窓を用いた．基底数 K の値は Fig. 3 の楽譜毎に変更し，音色変換後の音質が最良となるように調節した．式 (17)–(19) を 1000 回反復して各変数行列を推定した後，式 (20) を 1000 回反復し，音色変換後の信号を生成した．

Fig. 3 の楽譜は，Score 1 及び 2 が 1–3 音の和音，Score 3 及び 4 は 2–4 音の和音，Score 5 及び 6 は 2–5 音の和音でそれぞれ構成される．そして，Score 1 及び 2 は 3 個の基本音（C4 音，E4 音，及び G4 音）が存在し，その他の和音はこの 3 個の基本音の組み合わせである．同様に，Score 3–6 も 4 個の基本和音が存在し，その他の和音は基本和音の組み合わせである．

本実験において，前節の方法で高精度に音色が変換されたかどうかを評価するために，ABX 法に基づく主観評価実験を実施した．ABX 法とは，被験者に対して A と B をまず提示し，その後に続けて A と B のいずれかである X を提示する実験である．被験者は最後に提示された X が A と B のどちらと等しいかを回答する．ABX 法では「A と B には差がない（区別できない）」ことを帰無仮説とし，上記試問に対する回答を複数集めることで，帰無仮説が棄却できるか否か，即ち A と B が本当に区別可能であるか否かを有意水準に基づき検定する．今，得られた回答数（標本数）を N_{answer} とし，全試問中の正答数 N_{correct} とする．帰無仮説が正しいならば，提示された X に対して「X は A か B か」という問いには正当も誤答も期待確率 0.5 である．ABX 法では，被験者は 2 つの選択肢の内 1 つを選ぶと正当か誤答かが決するため，自由度 1 のカイ二乗分布での検定が可能である．このとき，χ^2 の値は次式となる．

$$\chi^2 = \frac{4}{N_{\text{answer}}}\left(N_{\text{correct}} - \frac{N_{\text{answer}}}{2}\right)^2 \qquad (23)$$

カイ二乗分布の値より，$\chi^2 > 3.84$ ならば有意水準 5 %で，$\chi^2 > 6.63$ ならば有意水準 1 %で帰無仮説が棄却される．

本節の主観評価では，2 つの異なる MIDI 音源で生成した同じ楽譜の楽器音信号を A と B として被験者に提示する．その後，X として，音色変換後の同じ楽譜の楽器音信号を提示し，A と B のどちらに近いかを回答させる ABX 法を実施した．このような提示は，X が A 又は B そのものではないため本来の ABX 法とは異なるが，次の 2 つの前提条件が成立するならば，「A と B には差がない」という帰無仮説を棄却できる場合に音色変換が高精度であることを示す．

- A と B の音色は明らかに差がある
- A と B の音色の違いは全被験者が完全に区別できる

1 つ目の前提条件は，異なる MIDI 音源を用いていることから成立する．2 つ目の前提条件は，2 年以上の楽器経験者を被験者として採用することで成り立つものと仮定する．実際に今回用いた Iowa Piano と Sketch Upright Piano は明確に音色差が聴感上感知できるものであるため，上記の 2 つの前提条件は成立しているとみなしてよい．

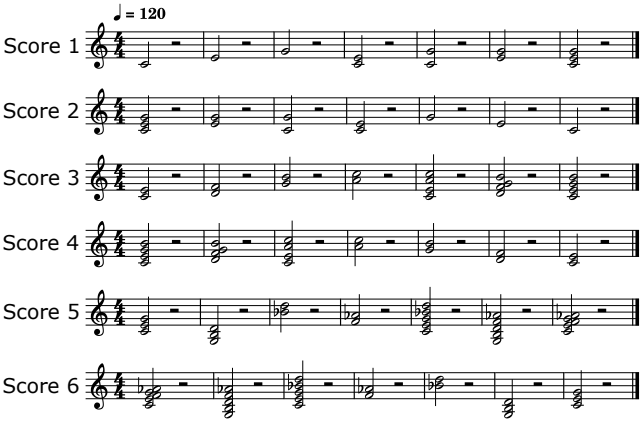

Fig. 3 Music scores used in experiment of timbre conversion.

前述のとおり，2 年以上の楽器経験者 14 名（男性 7 名及び女性 7 名）を対象に主観評価実験を行った．楽譜は Score 1 と Score 2，Score 3 と Score 4，Score 5 と Score 6 をそれぞれペアとして BSNMF に基づく音色変換を適用することで，各楽譜について Sketch Upright Piano を Iowa Piano に音色変換した合成音，Iowa Piano を Sketch Upright Piano に音色変換した合成音の合計 12 個の楽器音信号を用意し，これを提示音 X とした．この 12 個の X の音源それぞれに対して，「A を Iowa Piano，B を Sketch Upright Piano とした場合」及び「A を Sketch Upright Piano，B を Iowa Piano とした場合」の 2 パターンを考え，合計 24 個の ABX 音源を被験者に提示した．このとき，A と B と X は全て同じ Score の楽器音信号とした．また，A と B の間及び B と X の間にはそれぞれ 1 秒の無音区間を挿入した．なお，24 個の音源の提示順は被験者毎にランダムとし，各問の聞き直しは 2 回までとした．

4.2 実験結果

14 名の被験者の各回答の χ^2 値の箱ひげ図を Fig. 4 に示

[*2] https://vst4free.com/plugin/2294/
[*3] https://vis.versilstudios.com/upright-1.html

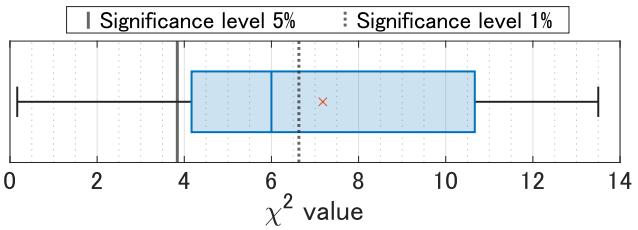

Fig. 4 Box plot of χ^2 values for chi-squared test. Central line of box plot indicates median, cross mark in box indicates average, and left and right edges of box indicate 25th and 75th percentiles, respectively.

す．また，各被験者の年代及び性別の情報と回答結果から算出した正答率及び χ^2 値を Table 1 に示す．Fig. 4 から，χ^2 値が多くの被験者において有意水準 5% よりも高い値に位置することがわかる．例えば，全被験者の結果での中央値は $\chi^2 = 6$ となった．また Table 1 より，正答率が 50 % を下回る被験者はいなかった．以上の結果より，有意水準 5 %で「A と B は差がない」という帰無仮説が棄却でき，前節で述べた通りこの結果は音色変換を高精度にできているという結果を表している．

実験後に被験者にヒヤリングを実施したところ，一部の被験者は 2 種類のピアノ音がグランドピアノとアップライトピアノの違いであることを認識できていた．当該被験者らの主張では，これらの違いは高周波帯域に明確に現れており，「低音の豊かな柔らかい音がグランドピアノ」及び「高音の豊かなきらびやかな音がアップライト」等と形容された．そして，「提示音 X がこれらの違いを十分反映しており，それを手がかりとして識別できた」という意見があり，これは音色変換した音が，変換先の種類のピアノとして十分認識できることを示している．従って，被験者が形容した「低音の豊かな柔らかい音」や「高音の豊かなきらびやかな音」は固有基底行列 $\boldsymbol{F}_1$ 及び $\boldsymbol{F}_2$ で表現されており，このような楽器経験者の主観的な形容を，$\boldsymbol{F}_1$ 及び $\boldsymbol{F}_2$ のスペクトルとして定量的に議論できる大きな可能性を示唆している．

5. まとめ

本稿では，複数楽器信号間の共通・固有成分を教師無し学習として抽出できる BSNMF を紹介し，推定された各成分の用途の一つである音色変換アルゴリズムを新たに提案した．提案手法では，固有成分を楽器音信号間で交換することで音色変換を実現しており，複数楽器音信号のパラレルデータを必要としない点に大きな利点がある．主観評価の結果より，人が認知可能な精度で音色変換が成功していることが確認された．今後の課題として，提案手法と同様にパラレルデータを必要としない音色変換手法 [17], [18] と精度を比較することが挙げられる．

謝辞 本研究の一部は，公益信託小野音響学研究助成基金の助成を受けた．

Table 1 Accuracy and χ^2 value of each subject

ID	Age	Gender	Accuracy [%]	χ^2 value
Subject no.1	20s	Female	70.83	4.17
Subject no.2	30s	Female	66.67	2.67
Subject no.3	30s	Female	75.00	6.00
Subject no.4	20s	Male	70.83	4.17
Subject no.5	20s	Male	87.50	13.50
Subject no.6	Teens	Male	83.33	10.67
Subject no.7	Teens	Male	83.33	10.67
Subject no.8	Teens	Male	70.83	4.17
Subject no.9	Teens	Female	75.00	6.00
Subject no.10	Teens	Female	54.17	0.17
Subject no.11	Teens	Male	79.17	8.17
Subject no.12	20s	Female	75.00	6.00
Subject no.13	20s	Male	87.50	13.50
Subject no.14	20s	Female	83.33	10.67

参考文献

[1] 香西海斗, 北村大地, “基底共有型非負値行列因子分解に基づく楽器音の共通・固有成分の分析,” 日本音響学会 *2021* 年春季研究発表会講演論文集, pp. 1109–1112, 2021.
[2] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[3] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization” *Proc. NIPS*, pp. 556–562, 2000.
[4] F. Zheng, G. Zhang, and Z. Song, “Comparison of different implementations of MFCC,” *J. Comput. Sci. and Tech.*, vol. 16, no. 5, pp. 582–589, 2001.
[5] N. H. Fletcher and T. D. Rossing, “The physics of musical instruments,” *Springer Sci. & Business Media*, 1991.
[6] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Musical instrument recognizer “instrogram” and its application to music retrieval based on instrumentation similarity,” *Proc. Int. Symp. Multimedia*, pp. 265–274, 2006.
[7] C. Joder, S. Essid, and G. Richard, “Temporal integration for audio classification with application to musical instrument classification,” *IEEE Trans. ASLP*, vol. 17, no. 1, pp. 174–186, 2009.
[8] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis,” *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.
[9] D. FitzGerald, M. Cranitch, and E. Coyle, “On the use of the beta divergence for musical source separation,” *Proc. Irish Signals Syst. Conf.*, 2009.
[10] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, “Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties,” *IEICE Trans. Fundamentals*, vol.E97-A, no.5, pp.1113–1118, 2014.
[11] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, “Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration,” *IEEE/ACM Trans. ASLP*, vol. 23, no. 4, pp. 654–669, 2015.
[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.
[13] D. R. Hunter and K. Lange, “A tutorial on MM algorithms,” *The American Statist.*, vol. 58, no. 1, pp. 30–37, 2004.
[14] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” *Proc. IEEE Spoken Lang. Tech. Workshop*, pp. 313–317, 2012.
[15] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” *Proc. ICASSP*, pp. 7944–7948, 2014.
[16] 李権俊, 相原龍, 滝口哲也, 有木康雄, “複素 NMF を用いた声質変換の検討,” 日本音響学会 *2016* 年秋季研究発表会講演論文集, pp. 277–280, 2016.
[17] J. Driedger, T. Prätzlich, and M. Müller, “Let It Bee – Towards NMF-inspired audio mosaicing,” *Proc. ISMIR*, pp. 350–356, 2015.
[18] 池田将也, 小坂直敏, “NMF を用いたサウンドコラージュの合成,” 情報処理学会研究報告, vol. 2020-MUS-126, no. 8, pp. 1–6, 2020.