スパース音場分解における時間周波数領域低ランクモデルの導入 *

☆村田直毅 [1]，小山翔一 [1]，亀岡弘和 [1,2]，高宗典玄 [1]，猿渡洋 [1]
[1] 東京大学　大学院情報理工学系研究科
[2] 日本電信電話株式会社　NTT コミュニケーション科学基礎研究所

1　はじめに

音場分解は，複数のマイクロフォンで得られた音圧を用いて，音場を波動方程式（またはヘルムホルツ方程式）の基本解の和として表現することを目的としており，音場の解析や可視化，再現の基礎となっている．一般的な手法として，音場を平面波に相当するフーリエ基底で展開する手法 [1] があり，その高い安定性と計算効率に有用性を持つ．一方で，音場を構成する要素（例えば点音源）の数は少ないという仮定から，音場分解の問題をスパース分解問題の枠組みで解く手法と，その応用が提案されている [2–4]．

音場のスパース分解の応用の一例として音場収音・再現がある．これは，収音領域で複数のマイクロフォンを用いて得た観測信号を用いて，再現領域で複数のスピーカを用いて収音領域の音場を再現することである．このとき，再現領域のスピーカの駆動信号は収音領域のマイクロフォンで得られる音圧の観測信号を用いて計算できる．マイクロフォンとスピーカの配置が平面状または直線状の場合，波面再構成フィルタ [5] により駆動信号への信号変換が実現可能である．しかしながら，この手法はマイクロフォンアレイの素子間隔により決まる空間ナイキスト周波数以上の帯域において，空間エイリアシングの誤差が生じ，音像の定位感が劣化してしまうという問題があった．文献 [4,6] は，この空間エイリアシングの誤差を軽減するため，音場のスパース分解に基づく信号変換手法を提案している．この手法は，音場をモノポール音源成分と平面波成分の重ね合わせでモデル化している．ここでマイクロフォン近傍の領域におけるモノポール音源成分の数は少数であると仮定できるため，観測信号を，グリーン関数で表現される基底関数上にスパースに展開することができる．また，このような分解が正確に実現できれば，空間ナイキスト周波数以上の帯域での再現精度を向上することが可能であることが報告されている．すなわち，高い再現精度を実現するためには，高精度な音場分解手法が必要となる．

しかしながら，例えば残響環境下において鏡像音源を含めた音場分解を行う場合，目的音のパワーが直接音に比べて極めて小さくなってしまうため，分解の精度が低下するという問題があった．これまでの手法は，時間周波数領域の事前情報を用いていなかったが，実用上は，分解すべき信号の時間周波数構造が限られており，それらがあらかじめ学習可能であるような状況は十分に考えられる．

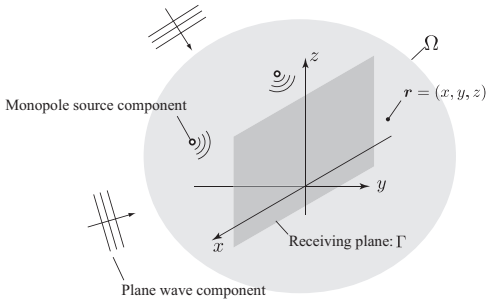


Fig. 1　Generative model of sound field

そこで，本稿では，時間周波数領域低ランクモデルを導入した音場のスパース分解を提案する．文献 [7] で提案されている複素 NMF (Non-negative Matrix Factorization) では，音響信号の時間周波数領域の情報を，低ランク行列 (特に rank1 行列) で表された振幅スペクトログラムと位相スペクトログラムの積の和でモデル化している．この手法は，音響信号が限られた音のパターンから構成されていることを仮定するもので，モノラル音源分離において有効な手法となっている．提案手法は，音場を構成するモノポール音源成分を，時間周波数領域上では [7] で提案されたように複素 NMF の形でモデル化できると仮定するもので，モデルに合致しない音響信号（雑音など）の干渉に頑健な分解が実現できると期待できる．

2　時間周波数領域低ランクモデルを導入した音場の生成モデル

2.1　音場の生成モデルとそのスパース分解

最初に，文献 [4,6] で提案された音場の生成モデルについて概観する．これは Fig. 1 のように音場をモノポール音源由来の成分と平面波由来の成分との和として表現するものである．まず，収音場をある閉曲面の内部と外部の領域に分け，内部領域を Ω とする．モノポール音源成分は，領域 Ω 内にのみ存在するとする．位置 $\mathbf{r}$ における周波数領域での音圧を $p(\mathbf{r})$ とすると，これは非斉次項 $p_{\mathrm{i}}(\mathbf{r})$ と斉次項 $p_{\mathrm{h}}(\mathbf{r})$ の和として以下のように書ける．

$$\begin{aligned} p(\mathbf{r}) &= p_{\mathrm{i}}(\mathbf{r}) + p_{\mathrm{h}}(\mathbf{r}) \\ &= \int_{\mathbf{r}'\in\Omega} Q(\mathbf{r}')G(\mathbf{r}|\mathbf{r}')d\mathbf{r}' + p_{\mathrm{h}}(\mathbf{r}) \end{aligned} \quad (1)$$

---

* Sparse sound field decomposition with low-rank modeling in time-frequency domain. by MURATA Naoki[1], KOYAMA Shoichi[1], KAMEOKA Hirokazu[1,2], TAKAMUNE Norihiro[1], and SARUWATARI Hiroshi[1]. [1]Graduate School of Information Science and Technology, The University of Tokyo, [2]NTT Communication Science Laboratories.

ここで，$G(\mathbf{r}|\mathbf{r}')$ は 3 次元自由空間のグリーン関数であり，$Q(\mathbf{r})$ は Ω 内でのモノポール音源成分の分布を表す．斉次項 $p_{\mathrm{h}}(\mathbf{r})$ は平面波の和として表現できる．また，Ω 内の平面 Γ において，音圧分布を観測することとする．

式 (1) を離散化することで以下に示すようにスパース分解の枠組みで扱うことができる．まず，領域 Ω を格子点の集合として離散化し，マイクロフォンは平面 Γ 上に離散的に配置するとする．ここで，格子点は領域 Ω 内に密に配置する必要があるため，格子点の数を N，マイクロフォンの数を M とすると，$N \gg M$ と仮定できる．このとき，式 (1) は以下のような離散形式で書ける．

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{z}, \tag{2}$$

この式で，$\mathbf{y} \in \mathbb{C}^M$ は各マイクロフォンの観測信号を，$\mathbf{x} \in \mathbb{C}^N$ は各格子点におけるモノポール音源成分の分布を表す．$\mathbf{D} \in \mathbb{C}^{M\times N}$ は各格子点と各マイクロフォンとの間のグリーン関数を要素に持つ辞書行列となる．ここでは，モノポール音源由来の成分である $\mathbf{D}\mathbf{x}$ が $\mathbf{y}$ の支配的な成分であることを仮定し，斉次項に対応する $\mathbf{z}$ は残差成分として扱う．ここで，領域 Ω 内でモノポール音源成分の数は格子点の数に比べて十分に少ないと仮定できるため，$\mathbf{x}$ は少数の要素のみに非ゼロの値を持つ，すなわちスパースな構造を持っていると考えられる．これにより，音場のスパース分解は，$\mathbf{x}$ のスパース性を誘導する罰則項を $J(\mathbf{x})$ としたときに，次のような最適化問題を解くことにより実現できる．

$$\min_{\mathbf{x}}\Big\{||\mathbf{y} - \mathbf{D}\mathbf{x}||_2^2 + \lambda J(\mathbf{x})\Big\} \tag{3}$$

ここで，λ は残差項と罰則項のバランスを決めるパラメータである．

式 (2) は，ある特定の時間周波数ビンのみにおける信号モデルを表現している．文献 [4,6] では，音場の物理的性質から誘導されるグループスパースモデルを導入することで複数の信号モデルを統一的に扱い，より正確な音場分解を実現している．例えば [4] では，複数の時間フレームの信号を同時に扱い，モノポール音源成分の位置が複数の時間フレームで変化しないことを仮定したとき，複数の時間フレームのモノポール音源成分の分布を表すベクトル（式 (2) における $\mathbf{x}$) のスパース構造が変化しないことを利用している．すなわち，ベクトル中の非ゼロの位置が全観測時刻で共通であると仮定している．このようなスパース分解は同時スパース分解と呼ばれる [8]．文献 [4] では，M-FOCUSS アルゴリズム [9] を用いて同時スパース分解を行っている．

2.2 時間周波数領域低ランクモデルの導入

ここでは，式 (2) のモノポール音源成分 $\mathbf{x}$ に時間周波数領域での低ランクモデルを導入することを考える．音響信号処理の分野においては，時間周波数領域の情報を用いて音源分離や多重音解析などがなされている [7,10]．文献 [7] で提案されている複素 NMF では，時間周波数領域の複素スペクトログラムを以下のようにモデル化している．

$$f_{k,t} \approx \sum_l h_{k,l} u_{l,t} e^{j\phi_{k,l,t}} \tag{4}$$

ここで，それぞれ k は周波数，t は時刻に対応するインデックスである．$h_{k,l}$ は時刻に依らず，グローバルに決定される基底スペクトルのパターンであり，l はそのインデックスを表す．$u_{l,t}$ は基底スペクトルのアクティベーションであり，$\phi_{k,l,t}$ は位相スペクトルである．これらは各時刻毎に自由度を持つ．これは，音響信号が限られた種類の基底スペクトルの信号だけで構成されており，それらのアクティビティと位相スペクトルが自由に時変して実現されるものという考えに基いている．

式 (2) のモノポール音源成分 $\mathbf{x}$ にもこのモデルを導入することが可能である．ここから，n は格子点のインデックスを表すとする．$x_{n,k,t}$ が，ある格子点 n の，時間周波数領域での複素スペクトログラムの成分を表すとすると，以下のようにモデル化できる．

$$x_{n,k,t} \approx \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}} \tag{5}$$

基底スペクトル $h_{k,l}$ は，事前に音源の情報として学習するとする．$\phi_{n,k,t}$ は本来ならば基底のインデックスを持つべきであるが，ここでは格子点 n 毎に位相スペクトログラムが対応するとする．このモデルを導入したもとで，マイクロフォンのインデックスを m として，各マイクロフォンで得られる観測信号 $y_{m,k,t}$ は，以下のようになる．

$$y_{m,k,t} \approx \sum_n d_{m,n,k} \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}} \tag{6}$$

ここでは，$d_{m,n,k}$ は周波数 k の，格子点 n とマイクロフォン m との間のグリーン関数を表す．また，文献 [4] に倣って，$u_{n,l,t}$ に対して次のような罰則項を加える事で空間的にスパースな解を求める．

$$J^{p,2}(\mathbf{U}) = \sum_n \Big(\sum_{l,t} u_{n,l,t}^2\Big)^{p/2} \tag{7}$$

今後，$\mathbf{U}$ と $\mathbf{\Phi}$ は，$u_{n,l,t}$ と $\phi_{n,k,t}$ を要素にもつテンソルを表すものとする．以上のモデル化のもとに，最適化問題の定式化を行う．ここで，各格子点と各マイクロフォンとの間のグリーン関数 $d_{m,n,k}$ は既知とし，モノポール音源の基底スペクトル $h_{k,l}$ は事前の学習により求まっているとする．式 (6) のモデル化誤差は

$$R(\mathbf{U},\mathbf{\Phi}) = \sum_{m,k,t} |y_{m,k,t} - \sum_n d_{m,n,k} \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2 \tag{8}$$

と表せるので，最適化問題は次のように書ける．

$$\min_{\mathbf{U},\mathbf{\Phi}}\big\{R(\mathbf{U},\mathbf{\Phi}) + 2\lambda J^{p,2}(\mathbf{U})\big\} \tag{9}$$

λ は，モデル化誤差と罰則項のバランスを決めるパラメータである．

3　補助関数法に基づく最適化アルゴリズムの導出

この最適化問題を解くアルゴリズムを補助関数法 [7] より導く．最初に，目的関数の補助関数を，以下の不等式（証明略）を用いて導く．

$$|y_{m,k,t} - \sum_n d_{m,n,k} \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2 \leq \sum_{n,l} \frac{|\overline{y}_{m,n,k,l,t} - d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2}{\beta_{m,n,k,l,t}} \quad (10)$$

$$2(\sum_{l,t} u_{n,l,t}^2)^{p/2} \leq p(\sum_{l,t} \overline{u}_{n,l,t}^2)^{p/2-1}(\sum_{l,t} u_{n,l,t}^2 - \overline{u}_{n,l,t}^2) + 2(\sum_{l,t} \overline{u}_{n,l,t}^2)^{p/2} \quad (11)$$

$\overline{y}_{m,n,k,l,t}$，$\overline{u}_{n,l,t}$ は補助変数であって，式 (10) は $\sum_{n,l} \overline{y}_{m,n,k,l,t} = y_{m,k,t}$ の条件のもとで，式 (11) は $0 < p < 2$ の条件のもとで成り立つ．また，$\beta_{m,n,k,l,t}$ は $0 < \beta_{m,n,k,l,t} < 1$，$\sum_{n,l} \beta_{m,n,k,l,t} = 1$ を満たす任意の定数である．よって，

$$f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}}) = \sum_{m,n,k,l,t} \frac{|\overline{y}_{m,n,k,l,t} - d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2}{\beta_{m,n,k,l,t}} + \lambda \sum_n \Big\{ p\Big(\sum_{l,t} \overline{u}_{n,l,t}^2\Big)^{p/2-1} \Big(\sum_{l,t} u_{n,l,t}^2 - \overline{u}_{n,l,t}^2\Big) + 2\Big(\sum_{l,t} \overline{u}_{n,l,t}^2\Big)^{p/2} \Big\} \quad (12)$$

とすれば，$f(\mathbf{U}, \mathbf{\Phi}) \leq f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}})$ が成り立ち，

$$\overline{y}_{m,n,k,l,t} = d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}} + \beta_{m,n,k,l,t}\Big(y_{m,k,t} - \sum_{n,l} d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}\Big) \quad (13)$$

$$\overline{u}_{n,l,t} = u_{n,l,t} \quad (14)$$

のとき $f(U, \phi) = f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}})$ が成り立つため，f^+ は補助関数の定義を満たす．ここで，$\overline{\mathbf{Y}}$，$\overline{\mathbf{U}}$ はそれぞれ $\overline{y}_{m,n,k,l,t}$，$\overline{u}_{n,l,t}$ を各要素に持つテンソルである．

以上により，$\mathbf{U}$ と $\mathbf{\Phi}$ の各要素の反復更新式を求めることができる．本来は $u_{n,l,t}$ の更新においては非負値制約を課す必要があるが，ここではここでは簡単のため無制約で $u_{n,l,t}$ を最適化する．まず $u_{n,l,t}$ について，$\partial f^+ / \partial u_{n,l,t} = 0$ を解いて，

$$u_{n,l,t} \leftarrow \frac{\sum_{m,k} \frac{\Re[\overline{y}^*_{m,n,k,l,t} d_{m,n,k} h_{k,l} e^{j\phi_{n,k,t}}]}{\beta_{m,n,k,l,t}}}{\sum_{m,k} \frac{|d_{m,n,k}|^2 h_{k,l}^2}{\beta_{m,n,k,l,t}} + \lambda p (\sum_{l,t} \overline{u}_{n,l,t}^2)^{p/2-1}} \quad (15)$$

を得る．ここで，$\Re[\cdot]$ は複素数の実部をとる演算子である．変数の肩の $*$ は，複素共役を表す．

次に $\phi_{n,k,t}$ の更新式を導く．$\sum_i |a_i - b_i e^{j\theta}|^2$，$a_i, b_i \in \mathbb{C}$，$\theta \in \mathbb{R}$ を最小化する θ が $e^{j\theta} = \sum_i a_i b_i^* / |\sum_i a_i b_i^*|$ を満たすことを利用すると，$\phi_{n,k,t}$ の更新式

$$e^{j\phi_{n,k,t}} \leftarrow \frac{\sum_{m,l} \overline{y}_{m,n,k,l,t} d^*_{m,n,k} h_{k,l} u_{n,l,t}}{|\sum_{m,l} \overline{y}_{m,n,k,l,t} d^*_{m,n,k} h_{k,l} u_{n,l,t}|} \quad (16)$$

を得る．

従って，提案アルゴリズムは，まず $\mathbf{U}$，$\mathbf{\Phi}$ の初期値を定め，式 (13)，式 (16)，式 (15) により $\overline{\mathbf{Y}}$，$\mathbf{\Phi}$，$\mathbf{U}$ を順に更新する．

ここでは，

$$\beta_{m,n,k,l,t} = \frac{|d_{m,n,k}|^2 h_{k,l}^2 u_{n,l,t}^2}{\sum_{n,l} |d_{m,n,k}|^2 h_{k,l}^2 u_{n,l,t}^2} \quad (17)$$

を用いて式 (15) を，

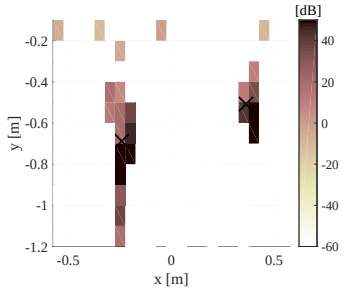
$$\beta_{m,n,k,l,t} = \frac{|d_{m,n,k}| h_{k,l} u_{n,l,t}}{\sum_{n,l} |d_{m,n,k}| h_{k,l} u_{n,l,t}} \quad (18)$$

を用いて式 (16) を更新する．

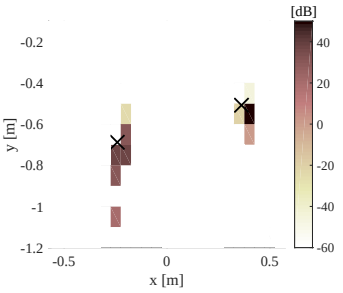
4　シミュレーション実験

音場分解に関して，2 次元自由空間内でのシミュレーション実験を行った．

収音場に，直線状の無指向性のマイクロフォンアレイが原点を中心として x 軸上に配置されているとし，マイクロフォンの数は 10，素子間隔は 0.12 m とした．よって，アレイ長は 1.2 m である．領域 Ω 内に 2 つの音源があるとし，それらの位置はそれぞれ (–0.24, –0.69, 0) m，(0.36, –0.51, 0) m とした．音源は実環境に近づけるため，それぞれ指向性を持つとした．音源はそれぞれ $-\pi/8$，$\pi/6$ の方向を向いた単一指向性を持つとした．音源信号は，MIDI 音源より作成したものを用いて，2 つの音源の音源信号は同一とする．楽器はオーボエである．サンプリング周波数は 16 kHz であり，時間周波数解析のフーリエ変換長は 32ms，シフト長はその半分とした．また，各マイクロフォンの観測信号に SN 比が 10 dB となるように，白色雑音を付加した．以上により得られた観測信号を，提案法 (Proposed) と，M-FOCUSS [4] によって分解した結果を比較した．格子点は，(0.0, –0.6, 0.0) m を中心とする x-y 平面の矩形領域 1.2×1,2 m^2 に，x 軸方向には 0.05 m，y 軸方向には 0.1 m の間隔で配置した．基底スペクトルの学習は，別に MIDI 音源から作成した観測信号中の音域を含む 25 個の音階の音源を用いて，音源の振幅スペクトログラムを非負値行列分解することにより行った．M-FOCUSS アルゴリズムにおいて，罰則項のパラメータはそれぞれ $p = 0.8$，$q = 2$ また残差項と罰則項を調整するパラメータは $\lambda = 0.5$ とした．M-FOCUSS の最大反復回数は 400 回であり，提案法の最大反復回数は 50 回と



(a) M-FOCUSS



(b) Proposed

Fig. 2　Sound field decomposition results

した．提案法において，式 (9) のパラメータ λ は，より安定な収束のため，最初の反復 5 回は 10^{-8} とした後，残りの 45 回は 10^{-1} とした．

信号分解性能を定量的に評価するため，F-measure(F_{msr}) を以下のように定義する．

$$F_{\mathrm{msr}} = 2\frac{|\mathrm{supp}(\mathbf{X}_{\mathrm{est}}) \cap \mathrm{supp}(\mathbf{X}_{\mathrm{ideal}})|}{|\mathrm{supp}(\mathbf{X}_{\mathrm{est}})| + |\mathrm{supp}(\mathbf{X}_{\mathrm{ideal}}|} \qquad (19)$$

ここで，$\mathrm{supp}(\mathbf{X}_{\mathrm{est}})$ は，ある閾値以上のパワーを持つ格子点のインデックス集合を表し，$\mathrm{supp}(\mathbf{X}_{\mathrm{ideal}})$ は，真の音源位置近傍の 3 つの格子点のインデックス集合を表す．ここで，閾値は，–20 dB とした．

Figure. 2 は，提案法，M-FOCUSS アルゴリズムにおける分解結果である．各グリッド点におけるパワーを，対数スケールで表したものである．真の音源の位置を “×” 印で表している．M-FOCUSS アルゴリズムおいては，真の音源位置以外にもパワーが分散して推定されているが，提案法においては，概ね真の音源位置周辺にパワーが集中して推定されている．また，上で定義した F_{msr} の値は，M-FOCUSS では 0.19 であるのに対し，提案法では 0.46 となった．これは，分解の段階で，雑音や指向性などのモデルから外れた誤差がある場合でも，予め学習した音源の情報を用いて，正しく分解がなされていることを示している．

5　おわりに

本稿では，音源に時間周波数領域低ランクモデルを導入したスパース音場分解手法を提案した．補助関数法を用いた最適化アルゴリズムの導出を行った．シミュレーション実験において，雑音が付加され，また音源がモデルとは離れた指向性を持っている場合においても，頑健な分解が可能であることを示した．

参考文献

[1] E. G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, New York, 1999.

[2] G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval. Near-field acoustic holography using sparsity and compressive sampling principles. *J. Acoust. Soc. Am.*, 132(3):1521–1534, 2012.

[3] A. Asaei, H. Bourlard, M. Taghizadeh, and V. Cevher. Model-based sparse component analysis for reverberant speech localization. In *Proc. IEEE ICASSP*, pages 1453–1457, Florence, May 2014.

[4] S. Koyama, S. Shimauchi, and H. Ohmuro. Sparse sound field representation in recording and reproduction for reducing spatial aliasing artifacts. In *Proc. IEEE ICASSP*, pages 4476–4480, Florence, May 2014.

[5] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda. Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(4):685–696, 2013.

[6] S. Koyama, N. Murata, and H. Saruwatari. Structured sparse signal models and decomposition algorithm for super-resolution in sound field recording and reproduction. In *Proc. IEEE ICASSP*, pages 619–623, Brisbane, Apr. 2015.

[7] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama. Complex NMF: A new sparse representation for acoustic signals. In *Proc. IEEE ICASSP*, pages 3437–3440, Taipei, Apr. 2009.

[8] A. Rakotomamonjy. Surveying and computing simultaneous sparse approximation (or group-lasso) algorithms. *Signal Process.*, 91:1505–1526, 2011.

[9] S. F. Cotter, D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Process.*, 53(7):2477–2488, 2005.

[10] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pages 177–180, New Paltz, Oct. 2003.