社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－MUS－63 (5)
2005／12／23

周波数解析と独立成分分析に基づく ステレオ音楽音響信号の音源分離

南里保仁 守田 了

山口大学工学部

宇部市常盤台 2557

一般に独立成分分析により音源分離を行う場合, 観測信号と音源の数が一致することが必要である. そのため, ステレオ音楽音響信号を独立成分分析を用いて音源分離する場合は, 音源が2つの場合可能であるがそれ以上の場合困難である. 本論文では, 音源が2つより多い3重奏からステレオ音楽音響信号を用いて音源を分離する手法を提案する. ステレオ音楽音響信号の左右の音を各周波数ごと観測すると, たくさんの音源の音が混合していても, 周波数空間では単音であるか高々2つの音の混合であることが多いことに着目し, 各周波数ごとに音源を分離する. 各周波数に単音を含む場合はそのまま音を分離し, 2つの音を含んでいる場合は 2×2 の混合行列を推定する独立成分分析を用いて分離する. それ以外は単音の解析および2つの音の解析から推定される $m \times 2$ の混合行列 $(m > 2)$ を用いて, 音源を分離する. 各周波数に単音か2つの音か2つ以上の音が含まれているかの判断には, 左右のパワースペクトルの比のヒストグラムと, 推定された 2×2 の混合行列から得られる分離後の音源が含む観測信号の比のヒストグラムを用いる. 本手法は音源が3つより多くても音源分離が可能である. 実際にビバルディの四季" 春" の第一楽章の三重奏のステレオ音楽音響信号に対して音源分離を行い有効性を示す.

キーワード 独立成分分析, 周波数解析, ステレオ音楽音響信号, 音源分離,

Sound Source Separation of Stereo Music Sound Signals Based on Frequency Analysis and Independent Component Analysis

Yasuhito NANRI and Satoru MORITA

Faculty of Engineering,Yamaguchi University

2557 Tokiwadai, Ube, 755, Japan

It is necessary that the number of the observed signals equals to the number of source signals, if the independent component analysis is used to separate sound sources. It is difficult to separate the sound sources more than two sound sources, if the stereo music sound signals are used. We propose the method to separate the stereo music sound signals that the number of sound sources is more than two sound sources using the frequency analysis and the independent component analysis. We separate the stereo music sound signals to the sound sources in each frequency, because a sound source signal often exist and two sound source signals are often mixed in a frequency, even if many sound sources are mixed. We use the power ratio between right and left observed sounds in a frequency, if a sound signal exist in a frequency. The mixed matrix of 2×2 is calculated, if two sound signals are mixed in a frequency. The mixed matrix of $m \times 2 (m > 2)$ is used, if many sound signals are mixed in a frequency. In the case of finding the number of sound sources included in a frequency, we use the histograms generated from the power ratio between right and left observation signals and the ratio that a sound source is included in the right and left observation signal gotten from the 2×2 mixed matrix. We show the effectiveness by applying the method to the stereo music sound signals of the spring composed by Vivaldi played in trio.

keyword: Independent Component Analysis, Frequency Analysis, Stereo Music Sound Signal, Sound Source Separation

1 はじめに

音響信号を分離する音源分離にはさまざまな方法が提案されているが, その手法はおおきく2つに分けられる. その1つは多数のマイクを用いる音源分離であり [1][2][3][4][5][6], もう1つは高々2つのマイクを用いる音源分離である [7][8][9][10][11][12][13][14]. 我々がテレビや CD などで日常接している音楽はモノラルまたはステレオで録音されているため, 音楽を対象とした音源分離を行う場合は, 高々2つのマイクを用いる方が現実的である.

近年独立成分分析により音源分離を行う手法が提案されている [1]. 一般に独立成分分析により音源分離を行う場合, 観測信号と音源の数が一致することが必要である. そのため, ステレオ音楽音響信号を独立成分分析により音源分離する場合は, 音源が2つの場合は可能であるが, それ以上の場合は困難である.

他方で, ステレオ音楽音響信号を左右の2つの音楽音響信号のスペクトル比のヒストグラムを用いて, 音源分離を行う手法が提案されている [7][8]. ここでは各音源の楽譜を推定できるものの, 各音源の音の復元を目的としていない.

本論文では, 音源が2つより多い3重奏からステレオ音楽音響信号を用いて, 各音源の音の復元を目的とするために, 先に挙げた2つの手法の長所を組み合わせて, 左右の2つの音響信号の各周波数におけるスペクトル比と独立成分分析に基づく音源分離を提案する.

たくさんの音源の音が混合していても, 周波数空間では単音であるか2つの音の混合であることが多いことに着目する. まずステレオ音楽音響信号の周波数スペクトル比のヒストグラムから複数の極大値を抽出する. この極大値の近傍以外は複数の音が混在している可能性が高いので, それ以外の場合, 各周波数ごと独立成分分析を用いて 2×2 の混合行列を推定し, 2個の音源に分離する. 推定された 2×2 の混合行列から分離後の各音源を含む左右の観測信号の比のヒストグラムを作成し, 得られたヒストグラムから複数の極大値を抽出する. 分離結果の音源が異なる2つの極大値の近傍であれば, その周波数では2つの音源が混合している可能性が高い. それ以外の場合はヒストグラムから得られた複数の極大値を含む 2×2 の混合行列から $m \times 2$ の混合行列 $(m > 2)$ を推定する.

実際にビバルディの四季の第一楽章の三重奏のステレオ音楽音響信号に対して音源分離を行い有効性を示す.

2 ステレオ音楽音響信号の音源分離

2.1 左右のパワースペクトルの比による音源分離

ステレオ音楽音響信号の左右の音は 44.1kHz で 16bit でサンプルされている. 256 音が 16 分音符にあたるように 25 音ごとに 2^{16} 個の音をサンプルし, 2^{16} 個の FFT をかけ周波数分解する. 左右のパワースペクトルの比から時刻 t, 周波数 f における定位 $h(t,f)$ を以下のように定義する.

$$h(t,f) = \begin{cases} \frac{p_R(t,f)}{p_L(t,f)} & (p_R(t,f) <= p_L(t,f)) \\ 2 - \frac{p_R(t,f)}{p_L(t,f)} & (p_R(t,f) > p_L(t,f)) \end{cases} \quad (1)$$

但し $p_L(t,f)$ は時刻 $t(s)$, 周波数 $f(Hz)$ における左のパワースペクトルであり, $p_R(t,f)$ は時刻 $t(s)$, 周波数 $f(Hz)$ における右のパワースペクトルである. $h(t,f)$ は左に定位しているほど 0 に近い値になり, 中央に定位しているほど 1 に近い値になり, 右に定位しているほど 2 に近い値になる. 16 分音符にあたる 256 音に対して 0 から 2 までの値を 0.025 ごとの間隔で投票しヒストグラムを作成する. m 重奏の場合, ヒストグラムから m 個の極大値を抽出する. m 個の極大値を $pk_0(t), pk_1(t), \cdots, pk_{m-1}(t)$ とおく. もし $h(t,f)$ が $pk_i(t) - width$ より大きく, $pk_i(t) + width$ より小さいならその時刻および周波数において, i 番目の音源と判断する. 本実験では $width = 0.05$ を用いている. この時点で m 個の定位の音と判断されなかった音は2つ以上の音が混合している可能性が高いので, 次の処理により分離する.

2.2 2×2 の混合行列を推定する独立成分分析による音源分離

16 分音符にあたる 256 音に対して, 各周波数に複数の音が混合していると判断された音について, 独立成分分析により音源分離を行う. 独立成分分析を行う場合, 2個の観測信号に対して, 2つに音源分離できる.

本研究では最尤推定による FastICA アルゴリズムを用いた [1]. 各周波数の 256 音のデータを中心化して平均を 0 とする. 相関行列 $C = E\{xx^T\}$ を計算する. 乱数により分離行列 B を初期化する. ここでは, B の逆行列が混合行列にあたるので, 混合行列を乱数により初期化する. その後混合行列の逆行列を分離行列に用いる. まず, 無相関化と正規化を以下の計算により行う.

$$B \leftarrow (BCB^T)^{-\frac{1}{2}} B \quad (2)$$

以下の計算を行う.

$$y = Bx \tag{3}$$

$$\beta_i = -E\{y_i \cdot g(y_i)\}(i = 0, 1) \tag{4}$$

$$\alpha_i = \frac{-1}{\beta_i + E\{g'(y_i)\}}(i = 0, 1) \tag{5}$$

分離行列を以下により更新する. 但し $g(y) = tahn(y)$

$$B \leftarrow B + diag(\alpha_i)[diag(\beta_i) + E\{g(y)y^T\}]B \tag{6}$$

収束していなければ収束するまで繰り返す.

推定された分離行列の逆行列から混合行列 A が推定できる.

$$A = B^{-1} \tag{7}$$

観測信号 x, 音源 s と混合行列を A とすると, 以下が成立する.

$$x = As \tag{8}$$

$$A = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix} \tag{9}$$

分離結果における観測信号における単音における右の観測信号と左の観測信号に含まれる割合は以下の式により計算される. 分離は音源が 2 つと仮定して計算しているために, 分離後の 2 つの音源に対して右の観測信号と左の観測信号に含まれる割合が得られる. 得られた混合行列 $A(x,t)$ のうち, i 番目の分離された音源を右の観測信号に含む音は $a_{i0}(x,t)$ であり, 左の観測信号に含む音は $a_{i1}(x,t)$ である. 得られた混合行列から時刻 t, 周波数 f における i 番目の分離された音の定位 $h'_i(t,f)$ を以下のように定義する.

$$h'_i(t,f) = \begin{cases} \frac{a_{i0}(t,f)}{a_{i1}(t,f)} & (a_{i0}(t,f) <= a_{i1}(t,f)) \\ 2 - \frac{a_{i0}(t,f)}{a_{i1}(t,f)} & (a_{i0}(t,f) > a_{i1}(t,f)) \end{cases} \tag{10}$$

$$(i = 0, 1)$$

2 つの $h'_i(t,f)$ を 0 から 2 までの値を 0.025 ごとの間隔で投票しヒストグラムを作成する. m 重奏の場合, ヒストグラムから m 個の極大値を抽出する. m 個の極大値を $pk'_0(t)$,$pk'_1(t)$,$\cdots$, $pk'_{m-1}(t)$ とおく. i 番目の分離結果 $h'_i(t,f)$ と m 個の極大値との距離を比較し, 最も近い音源と判断する. この方法によって任意の個数の音源を分離できる.

さらに左右のパワースペクトルの比のヒストグラムによる方法と組み合わせることで性能の向上が期待できる. 2×2 の混合行列を推定する独立成分分析では, 2 つの音源の混合音を分離できる. もし $h'_i(t,f)$ が $pk'_j(t) - width$ より大きく, $pk'_j(t) + width$ より小さいならその時刻および周波数において, j 番目の音源と判断する. 本実験では $width = 0.1$ を用いている. もし $h'_i(t,f)$ において $i = 0, 1$ の 2 つの場合に異なる音源が得られている場合に, 2 つの音源の混合と判断される. それ以外の場合は 2 つ以上の音が混合している可能性が高いので, 次の処理により分離する. i 番目の音源が右の観測信号に含まれる割合 $a_{i0}(x,t)$ と左の観測信号に含まれる割合 $a_{i1}(x,t)$ の比のヒストグラムから得られる m 個の極大値を $pk_0(t)$,$pk_1(t)$,$\cdots$, $pk_{m-1}(t)$ における $a_{i0}(x,t)$ と $a_{i1}(x,t)(i = 0, 1)$ の平均を計算し, j 番目の定位について $a'_{j0}(x,t)$ と $a'_{j1}(x,t)(j = 0, 1, \cdots, m-1)$ を得る.

2.3 $m \times 2$ の混合行列 $(m > 2)$ による音源分離

2×2 の混合行列を推定する独立成分分析により得られた $h'_i(t,f)$ のヒストグラムから, $m \times 2$ の混合行列 $(m > 2)$ を推定する. 2×2 の混合行列から分離後の各音源ごと 1×2 の混合行列を抽出し, 平均することによって, $m \times 2$ の混合行列 $A'(x,t)$ が求まる. $A'(x,t)$ は $a'_{ij}(x,t), (i = 0, 1, \cdots, m-1), (j = 0, 1)$ である. 2 個以上音を含んでいる場合は, 各定位の混合行列が推定されているため, 2 個の観測信号から一般化逆行列により m 個の異なる音源に分離できる. すなわち y を観測信号, s を音源であるとし, A' は $m \times 2$ の行列であるので, A' が推定されていれば 2 つの観測信号から m 個の音源を求めることができる.

$$y = A's \tag{11}$$

$$A'^{-1} = A'^T(A'A'^T)^{-1} \tag{12}$$

$$s = A'^{-1}y \tag{13}$$

一般化逆行列のみを用いて音源分離する場合は, 観測信号に推定された A'^{-1} を掛け合わせることにより音源が推定できる.

先に提案した左右のパワースペクトルの比による方法と 2×2 の混合行列を推定する独立成分分析とを組み合わせることで, 性能の向上が期待できる. すなわち, 周波数分解後の各周波数で単音が鳴っていると判断される場合は左右のパワースペクトルの比による方法を用い, 2 つの音が鳴っていると判断される場合は 2×2 の混合行列を推定する独立成分分析による方法を用い, それ以外の場合は一般化逆行列による方法で m 個の音源に対する周波数分解後の値を推定し, フーリエ逆変換により m 個の音源が推定できる.

2.4 音源の復元

m 個の音源に対する周波数空間の実部と虚部の値を, 分離結果を用いて推定する. すべての周波数, すべての時刻について求められた m 個の音源に対する周波数空間の実部と虚部の値からフーリエ逆変換を用いて各音源の音を再構成する.

2.5 音源分離結果の評価

定位前の音と分離後の音を比較することで, 分離結果を評価する. 各楽器の分離結果の評価式を以下に示す. $44.1kHz$ で抽出されたすべての音に対して評価した. m 個の音源を用いた分離で音の総数を $tmax$ とし, j 番目の定位前の音源の音を $as_j(t_{ti})$ とし, 分離後の結果を $ak_j(t_{ti})$ とする.

$$E_j = \frac{\sum_{ti=0}^{tmax-1} \frac{as_j(t_{ti}) - ak_j(t_{ti})}{as_j(t_{ti})} * 100.0}{tmax} \quad (14)$$

全体の分離結果はすべての音で評価し平均をとることにより計算した.

$$E = \frac{\sum_{j=0}^{m-1} \sum_{ti=0}^{tmax-1} \frac{as_j(t_{ti}) - ak_j(t_{ti})}{as_j(t_{ti})} * 100.0}{tmax * m} \quad (15)$$

2.6 ステレオ音楽音響信号の音源分離のながれ

- ステレオ音楽音響信号の右と左の観測信号を 16 分音符から 256 音に対して 2^{16} の FFT を用いて周波数分解する.
- 右と左のスペクトル比 $h(f,t)$ を計算し, ヒストグラムを作成し, m 個の極大値 $pk_j(t)$ を求める. 右と左のスペクトル比 $h(f,t)$ が m 個の極大値 $pk_j(t)$ の近傍にあるものをその周波数 f, 時刻 t について単音であると判断する.
- FastICA を用いて音源を 2 つと仮定した音源分離を行う. 推定された 2×2 の混合行列 A から分離後の各音源を含む右と左の観測信号の比 $h'_i(f,t)(i=0,1)$ を計算し, ヒストグラムを作成する. m 個の極大値 $pk_j(t)$ を求める. 右と左の観測信号の比 $h'_i(t)$ が m 個の極大値 $pk_j(t)$ の近傍にあるものをその周波数 f, 時刻 t について 2 つの音を含んでいると判断する.
- 各時刻各周波数について, 単音もしくは 2 つの音の混合と判断されなかった場合, $m \times 2$ の混合行列 $A'(x,t),(m > 2)$ を推定し, 一般化逆行列を用いて逆変換する.
- 上の処理によって, 各時刻各周波数の m 個の音源について, 周波数空間の実部と虚部を推定する. 推定された音をフーリエ逆変換により m 個の音を再構成する.

音源分離のながれを図 1 に示す.

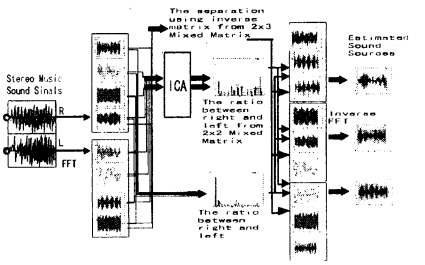

図 1: 音源分離のながれ

3 3重奏の音源分離

ビバルディの四季" 春" の第一楽章の最初の 4 小節である. 図 2 を左にバイオリン, 中央にチェロ, 右にコントラバスを定位させ, MIDI によってステレオ音楽音響信号を作成した. 図 3 にバイオリン, チェロ, コントラバスの配置を示した. この信号から本手法を用いて音源分離を行った. 定位前のバイオリン, チェロ, コントラバスの楽譜中の一小節目の最後の 16 符音符を図 6 に示す. さらに 16 符音符の中央の音を中心に 2^{16} をサンプルし, FFT をかけて得られた周波数スペクトルを図 7 に示す. 定位後のステレオ音楽音響信号の右と左の観測信号を図 8 に示す. さらに右と左の観測信号の周波数スペクトルを図 9 に示す. 図 4 と図 5 に $h(f,t)$ と $h'_i(f,t)$ のヒストグラムを示す. $pk_0(t), pk_1(t), pk_2(t)$ はそれぞれ $0.17, 1.0, 1, 83$ が得られている. 音源を分離し音を再構成後の結果を図 10 に示す. 音源を分離し音を再構成する前の周波数空間の値を図 11 に示す. 定位前の音 (図 6) と分離結果の音 (図 10) を比較すると, よく似ていることがわかる. 定位前の周波数スペクトル (図 9) と分離結果の周波数スペクトル (図 11) を比較しても, よく似ていることがわかる.

評価結果を表 1 に示す. 従来提案されている周波数解析とパワースペクトルの左右の比を用いた手法 (表 1 中 (I))[7][8] では 75%, 2 × 2 の混合行列を推定する独立成

分分析による手法 (表 1 中 (II)) では 72%程度の分離性能が得られた. $m \times 2$ の混合行列を推定後, 一般化逆行列を用いる手法 (表 1 中 (III)) では 65%の分離性能が得られた. 3 つの手法を組み合わせた手法 (表 1 中 (IV)) では 85% 程度の分離性能が得られた. 従来手法に比べて性能が向上していることがわかる. ビバルディの四季" 春" の第一楽章の最初の 2 小節に対する評価では, 周波数解析とパワースペクトルの左右の比を用いた手法 (表 2 中 (I)) を用いた場合, 全体で 71% 程度の分離性能が得られているのに対し, 3 つの手法を組み合わせた手法 (表 2 中 (IV)) を用いた場合, 全体で 77% 程度の分離性能が得られており, 性能が向上していることがわかる. 但し, 表 1(I) 表 2(I) では各楽器を分ける境界に $C1 = 0.675$, $C2 = 1.375$[7][8] を用いた.

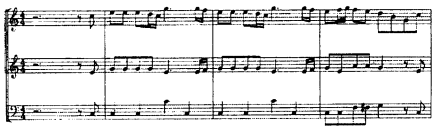

図 2: ビバルディの四季" 春" の第一楽章の最初の 4 小節

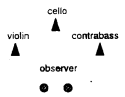

図 3: バイオリン, チェロ, コントラバスの配置

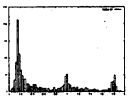

図 4: 音量比 $h(t,f)$ のヒストグラム

4 おわりに

音源が 2 つより多い 3 重奏からステレオ音楽音響信号を用いて音源を分離する手法を提案した. 左右のパワースペクトルの比のヒストグラムと独立成分分析から推定された 2×2 の混合行列から得られる分離後の音源を含む右と左の観測信号の比のヒストグラムを用いることにより, $m \times 2$ の混合行列 $(m > 2)$ を推定し, 各音源を精度よく分離できることを示した. 実際に分離した結果, 従来方法に比べて性能が向上していることを確認した.

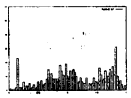

図 5: 推定された混合行列から得られた各音源の左右に含まれる観測信号の比 $h'_i(t,f)(i = 0,1)$ のヒストグラム

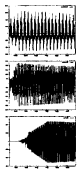

図 6: 定位前のバイオリン, チェロ, コントラバスの音源

本手法は音源が 3 つより多くても音源分離が可能な手法である.

謝辞

いつも研究に協力いただいている宇部市民オーケストラ, 山口大学管弦楽団の皆様に感謝いたします。

参考文献

[1] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis,2001

[2] O. M. E. Mitchell, C. A. Ross, G,H. Yatos,"Signal Processing for Cooktail Party Effect," J. Acoust. Soc. Am., Vol. 50, no. 2, pp. 656-660, 1971

[3] 永田仁史, 安部正人, 城戸健一, "多数センサによる音源波形の推定," 日本音響学会誌, 弟 47 巻 4 号,pp. 268-273, 1991

[4] 安部正人, "多数センサによる音源推定," 日本音響学会誌, 弟 51 巻 5 号,pp. 384-389, 1995

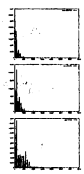

図 7: 定位前のバイオリン, チェロ, コントラバスの周波数スペクトル

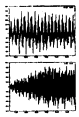

図 8: 定位後のステレオ音響信号の右と左の観測信号

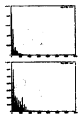

図 9: 定位後のステレオ音楽音響信号の観測周波数スペクトル

[5] 柳田益造, 角所収, 植田章彦, 野村康雄," 一般逆行列を用いたカクテルパーティー効果の知覚的検討, 信学技報, EA80-69, 1981

[6] C. Jutton, J. Herault,"Blind separation of sources, Part I:An adaptive algorithm based on neuromimetic architecture, Signal Processing, vol.24, no. 1, pp. 1-10, 1991

[7] 三輪明宏, 守田了, 能動環境におけるステレオ音楽音響信号を用いた 3 重奏の音源分離, 信学論, vol. J84-DII, no.1, pp.23-30,2001

[8] 三輪明宏, 守田了, ステレオ音楽音響信号を用いた三重奏に対する自動採譜, 信学論, vol. J84-DII, no.7, pp.1251-1260,2001

[9] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am., vol. 60, no. 4, pp. 911-918, 1976

[10] A. Nohorai, and B. Porat, "Adaptive Comb Filtering for Harmonic Signal Enhancement, IEEE Trans. on ASSP, vol. 34, no. 5, pp. 1124-1138, 1986

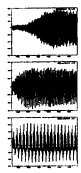

図 10: 音源分離によりステレオ音楽音響信号から推定された音源

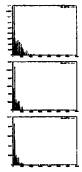

図 11: 音源分離によりステレオ音楽音響信号から推定された音源の周波数スペクトル

表 1: ビバルディの四季" 春" の一楽章の一小節目の最後の 16 符音符に対する音源分離性能の評価

	violin	cello	contrabass	average
I	60.8(%)	78.7(%)	85.5(%)	75.0(%)
II	59.4(%)	78.0(%)	77.1(%)	71.5(%)
III	67.9(%)	61.6(%)	67.3(%)	65.6(%)
IV	76.6(%)	88.6(%)	89.1(%)	84.8(%)

[11] M. Abe, S. Ando, "Application of Loudness/ Pitch/ Timbre Decomposition Operators to Auditory Scene Analysis, Proc. of ICASSP, pp. 1124-1138, 1986

[12] 中谷智広, 後藤真孝, 川端豪, 奥野博, "残渣駆動型アーキテクチャの提案と音響ストリームの分離への応用, 人工知能学会誌, vol.20, no.1, pp. 111-119, 1997

[13] 柏野邦夫, "音楽情景分析の処理モデル OPTIMA における単音の認識, 信学論, vol. J79-D-II, no. 11, pp. 1751-1761, 1996

[14] 柏野邦夫, "音楽情景分析の処理モデル OPTIMA における和音の認識, 信学論, vol. J79-D-II, no. 11, pp. 1762-1770, 1996

表 2: ビバルディの四季" 春" の第一楽章の最初の 2 小節に対する音源分離性能の評価

	violin	cello	contrabass	average
I	68.5(%)	70.8(%)	73.3(%)	70.9(%)
IV	73.3(%)	77.7(%)	78.7(%)	76.6(%)