音 楽 情 報 科 学 46－2
（２００２．７．７）

定位情報と音色情報を用いた複数楽器音の認識

櫻庭 洋平　　河原 達也　　奥乃 博

京都大学大学院情報学研究科知能情報学専攻知能メディア講座

sakuraba@kuis.kyoto-u.ac.jp　kawahara@kuis.kyoto-u.ac.jp　okuno@i.kyoto-u.ac.jp

あらまし 複数楽器音の自動採譜には音源分離同定処理が必須である．しかし，重奏から楽器ごとの情報を抽出する試みはそれほど多くなく，まだ十分な精度も得られていない．音源分離同定処理は周波数成分から単音を形成する同時的グルーピングと単音の流れを形成する継時的グルーピングの 2 つのグルーピングからなる．本稿では定位情報と音色情報を用いることで 2 つのグルーピングの曖昧性を解消することを試みる．同時的グルーピングでは，位相差の変動に着目して各周波数成分の重なりを判定し，重なり情報を利用して単音を形成する．継時的グルーピングでは，得られた単音の定位情報と音色情報を手がかりとしてパートごとの流れを形成する．本手法を実装・実験した結果，提案手法の有効性を確かめることができた．

Sound Source Identification of Multiple Musical Instruments by Combining Localization and Timbre

Yohei Sakuraba　　Tatsuya Kawahara　　Hiroshi G. Okuno

Dept. of Intelligence Science and Technology,
Graduate School of Informatics, Kyoto University

Abstract Some experimental systems developed so far have done sound source identification in homophony music, but only a few in polyphony music, in particular, of multiple musical instruments. The accuracy of automatic musical transcription is not sufficient. Sound source separation consists of two processes; simultaneous grouping and sequential grouping. In this report, we propose localization (the direction of a sound source) to improve the performance of groupings. In simultaneous groupings, overlaps of each frequency component are judged using the change of IPD, and a note is generated by using overlaps of frequency components. In sequential groupings, a temporal sequence of notes are generated by using timbre and localization of a note. Experimental results showed that the performance of sound source identification is improved by the proposed method.

1. はじめに

近年，計算機の処理能力の向上に伴い音楽信号を自動採譜する研究が行われている．複数楽器による演奏を楽器ごとに自動採譜するには，音響信号を楽器ごとに分離する音源分離と，分離された信号の音源名を同定する音源同定が必須である．しかし，複数楽器による演奏を音源分離同定することは，周波数成分が干渉し合い複雑なスペクトルになるため非常に困難であり，そのような試みはそれほど多くない．

音源分離同定処理は Bregman によると，周波数成分から単音を形成する同時的グルーピングと，何らかの一貫性に従って単音の流れを形成する継時的グルーピングの 2 種類のグルーピングから構成される[1)]．複数楽器による演奏をターゲットとする場合，各グルーピングが困難である原因はいくつかあり，特に（1）同時的グルーピングにおけるオクターブの関係の認識の問題，(2）継時的グルーピングにおける特徴量の問題，という 2 つの問題があげられる．

（1）は同時に発音する単音が同一または，整数倍の関係にある基本周波数を持つ場合には，周波数成分の大部分が重なってしまうため，調波構造のみを手がかりに正しく単音形成をすることは難しい問題である，という問題である．(2）は継時的グルーピングにおいて各単音の流れを形成するための一貫性としてどのような特徴量が有効なのか，という問題である．各単音の音色情報は継時的グルーピングの大きな手がかりとなる．しかし，重奏をターゲットとすると，周波数成分の重なりにより干渉し合うため，正確な音源同定は困難である．

音源分離同定の先行研究には柏野らによるOPTIMAがある[2)3)]．OPTIMAでは周波数成分・単音・和音の3つの抽象度の階層を持つベイジアンネットワークを備えている．ボトムアップ処理，トップダウン処理，に加え，和音の遷移確率をベイジアンネットワークで情報統合することで同時的グルーピングの曖昧性の解消を図っている．パート（各楽器が担当する単音全体）ごとに分類する特徴としては音色情報のみを用いている．情報統合を行うことで「オクターブの関係」にある単音形成精度は約17%向上し，60.1%になっているものの，精度向上の余地は十分にある．

木下らはOPTIMAにおけるパート抽出の精度の向上を音色類似度，音域類似度，旋律類似度を用いて試みている[4)]．3パートの楽曲に対して再現率82%でパート抽出に成功しているものの，さらなる高精度化のためには他の手がかりも利用する必要があることを指摘している．

我々は，2種類のグルーピングの精度向上には複数の手がかりを用いる必要があると考える．その第一段階として定位情報を利用することを提案する．なぜなら，定位情報は和音遷移確率や単音遷移確率と異なり，音楽のジャンルに依存しないと考えられるからである．同時的グルーピングでは，調波構造だけでなく，各周波数成分の定位情報を利用することで精度向上を試みる．継時的グルーピングでは，各単音ごとに音源同定を行うのではなく，各単音を定位情報を用いて複数の集合に分類し，得られたパートごとに音源同定を行う．単音ごとの音源同定では，周波数成分の重なり等の影響により同定に失敗することが多かったのに対し，本手法ではパートごとに音源同定をするため，同定の対象となる単音数が多くなり，精度の向上が期待できると考えられる．

定位を用いて継時的グルーピングの精度を向上しようという試みは三輪らも行っている[5)]．ステレオ音響信号を入力とし，左右の音量比のヒストグラムを作り，クラスタリングすることでパートごとの採譜を行っている．この手法では，入力音響信号が三重奏までに限定され，その三つの楽器の配置も一つは中央，残り二つは左右に一つずつに限られていた．しかし，楽器の数が増えた場合にも対応していくためにも，より詳細な定位が必要となる．

我々は，両耳間強度差（IID）・両耳間位相差（IPD）を用いることで左・中央・右のような大きなレベルでの定位だけではなく，左30度のようにより詳細な定位を求め，2種類のグルーピングの手がかりとして利用する．

以下，**2**．で定位を用いた同時的グルーピングの曖昧性解消について説明する．**3**．で定位情報と音色情報を用いた継時的グルーピングの曖昧性解消について説明する．**4**．で本研究のために作成したシステムについて説明する．**5**．で重奏の音源分離同定実験を行い，考察する．**6**．で結論と今後の課題を述べる．

2. 同時的グルーピングの曖昧性と定位の利用

複数音からなる音響信号を時間周波数解析した結果得られた周波数成分をどの単音にグルーピングするかを決定するのは難しい問題である．柏野らは倍音構造を手がかりにグルーピングを行っている[2)3)]．倍音構造だけを手がかりにすると，オクターブの関係のように，ある単音の基本周波数が他の単音の倍音と重なった場合は，周波数成分の大部分が重なり合うため，すべての単音を正しく抽出するのは難しい．

我々は柏野らによるOPTIMAにおいて，正しく同時的グルーピングを行うことが困難であったオクターブの関係にも有効な，定位情報を用いたグルーピング法を提案する．定位情報を用いたグルーピング法は（1）定位の変化による周波数成分の重なり判定，(2）定位を用いた単音形成という2段構成になっている．

各周波数成分はいくつかのピークから成り立っている．ピークごとにIPDを用いて定位を求める（詳細は4.2）．重なりのない周波数成分の定位は全ピークを通じて安定した値を取るのに対し，重なりのある周波数成分の定位は安定した値を取らない．これにより各周波数成分に重なりがあるかないかを判定することができる．変動が閾値以内なら安定と判断し，安定している周波数成分の定位は全ピークの平均値とした．

次に得られた各周波数成分の定位および重なり判定と調波構造を用いることで単音を形成する．単音形成処理では，次の3つを仮定している．

(1) 一つの単音に含まれるすべての周波数成分はその単音の基本周波数に対し，ほぼ高調波関係にある．

(2) 一つの単音に含まれるすべての周波数成分の立ち上がり時刻はほぼ同時である．

(3) 一つの単音に含まれるすべての周波数成分の定位はほぼ等しい．

重なりがあると判定された周波数成分は調波構造（仮定（1）(2)）のみによりグルーピングを行い，すべての重なりを満足される組み合わせの単音の組み合わせが出力されると単音形成処理は終了する．

3. 継時的グルーピングの曖昧性と定位の利用

同時的グルーピングの結果，各時点において同時に発音している単音の組み合わせが出力される．その結果得られた単音は，継時的グルーピングにより楽器ごとに分類される．音色情報は継時的グルーピングの大きな手がかりである．しかし，重奏をターゲットとす

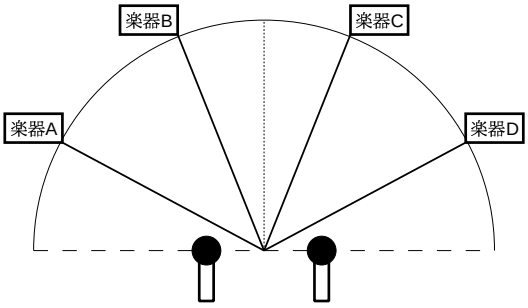

図 1　楽器の配置例

ると，周波数成分の重なりにより干渉し合い，音源同定の精度は単音に比べ低下する．

我々は，定位情報を用いてパートを形成し，パートごとに音源同定を行う．パートに属する全ての単音の音源同定の結果，多数決で最大数となる楽器をそのパートの楽器とすることで，音源同定の間違いを修正することができると考えられる．

定位を扱うために以下の 3 点を前提としている．

- 強度差と時間差をもつステレオ音響信号を入力とする
- 同じ定位には一つの楽器しかない
- 楽器の定位は移動しない

図 1 は前提を満たす楽器の配置の一例である．

定位情報を用いた継時的グルーピングは以下の手順で行う．

(1) ある単音の定位といずれかのパートの定位との誤差が閾値以内であれば，そのパートに属する単音とする．
(2) どのパートにも属さなければ新たなパートを生成する．
(3) 以上の処理を全ての単音に対して繰り返す．

以上により定位を用いてパートを形成する．

4. システムの構成

システムは図 2 で示すように，周波数解析部，定位抽出部，単音形成部，特徴抽出部，音源同定部，結果結合部の 6 モジュールから構成されている．また，システムは楽器の特徴量テンプレート（音源名と特徴ベクトルの集合）も持っている．本システムは 48kHz，16bit のステレオ音響信号の入力を前提としている．

入力のステレオ音響信号は，**周波数解析部**で左右それぞれ時間周波数解析し，ピーク抽出を行い時間方向に接続することで周波数成分を形成する．**定位抽出部**では得られた周波数成分ごとに定位を求める．**単音形成部**では調波構造および定位を用いて周波数成分をグルーピングし，単音を形成する．**特徴抽出部**は単音ごとにエンベロープや倍音構造に関する 23 次元の特徴量を抽出する．**音源同定部**は得られた特徴量とテンプレートの類似度を計算する．**結果結合部**は定位情報と音色情報を用いてパートごとにピアノロール譜を作成する．各処理部の詳細を以下で示す．

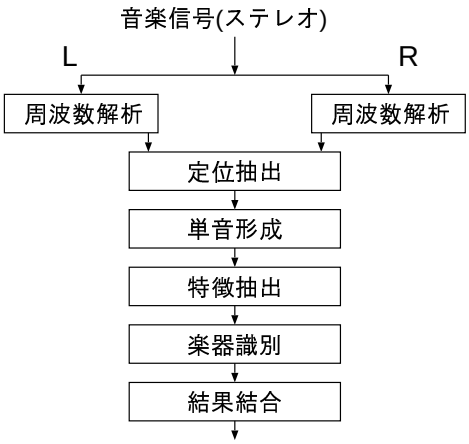

図 2　音源分離同定システム

4.1 周波数解析部

周波数解析部は入力信号に対し，短時間フーリエ変換（STFT）による時間周波数解析を行い，各フレームのピークを抽出する．FFT には FFTW[6] を用い，窓関数にはハミング窓を用いている．窓長は 4096 点（周波数分解能 11.8Hz）・シフト長は 1000 点である．時間周波数解析で得られたピークの中でパワーの大きい点を最大 60 点抽出する．

次に時間周波数解析で得られたピークを時間方向に接続する．ゆらぎによるピッチ変化を考慮し，50 セント☆までの変化を許容してピークを接続する．その結果，周波数成分が左右それぞれに形成される．

4.2 定位抽出部

4.1 の周波数解析処理により得られた周波数成分に対し，左右の対応を取る．左右の対応条件として

(1) 左右の周波数成分の音程の差が窓長 4096 における周波数分解能の 2 倍の 23Hz 以内である
(2) 左右の周波数成分が時間的に 0.1sec 以上の重なりを持つ

の 2 つの条件が成り立つ周波数成分を同一の周波数成分として対応づける．左右の対応が取れた周波数成分に対して，定位を求める．ここで言う定位とは水平方向の角度を指し，垂直方向の角度は考えない．

本システムでは強度差と時間差を利用して，定位を求める．図 3 のように音源が方向 θ で発音された場合，音源とマイクの距離がマイク間の距離 l に比べ十分大きいと考えると，左右の音波の到達距離の差は $d = l \sin\theta$ となる．そこで音源方向 θ を求めるために左右の対応

☆ 音高差を対数スケールで表現したもので，半音は 100 セントに相当する．

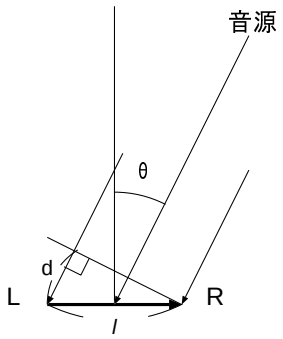

図 3　音源の方向と両耳間の時間差

が付けられた周波数成分に対し，STFT のフレームごとに以下の値を求める．$Sp(l)$ は左におけるスペクトルを表し，Re，Im でそれぞれ実部，虚部を表す．

- IID（両耳間強度差）
  各周波数成分に対し，パワースペクトルの比を求める．

$$IID = \frac{\sqrt{Re[Sp(l)]^2 + Im[Sp(l)]^2}}{\sqrt{Re[Sp(r)]^2 + Im[Sp(r)]^2}}$$

- IPD（両耳間位相差）
  1700Hz 未満の周波数成分に対し，位相差を求める．

$$\Delta\phi = \tan^{-1}(\frac{Im[Sp(l)]}{Re[Sp(l)]}) - \tan^{-1}(\frac{Im[Sp(r)]}{Re[Sp(r)]})$$

位相差（$\Delta\phi$）から，以下の式により時間差（Δt）が求められる．

$$\Delta t = \frac{1}{2\pi f}\Delta\phi$$

f は周波数成分のピッチを表す．次に

$$\theta = \sin^{-1}(\frac{c}{l}\Delta t)$$

により時間差（Δt）から音源方向（θ）を求める．c は音の速さ（340m/s），l はマイク間の距離（20cm）を表す．

位相差からは位相が進んでいるのか遅れているかはわからない．そこで，IID によって左右を定め，次に位相差から時間差を計算し方向を求めた．方向は各フレームごとの位相差の平均値から求める．

IPD は 1700Hz 未満の周波数成分に対して求める．1700Hz 以上のピッチをもつ周波数成分では位相が 2π 以上変化することもあるため，正しい時間差が求められないからである．(0.2/340 = 1/1700 より 20cm は 1700Hz の周波数が一周期の間に進む距離である．)

4.3　単音形成部

第 2 章で述べたように各周波数成分に対し重なり判定をし，定位情報と調波構造を用いて単音形成を行う．周波数成分の重なり判定の閾値は実験的に 6 度とした．同一単音としてグルーピングされる条件は次の 3 条件である．

(1)　一つの単音に含まれるすべての周波数成分はその単音の基本周波数に対し，整数倍の周波数と誤差 23Hz 以内．

(2)　一つの単音に含まれるすべての周波数成分はその単音の基本周波数に対し，時間的にその周波数成分の音長の半分以上が重なっている．

(3)　一つの単音に含まれるすべての周波数成分の定位は誤差 5 度以内．

閾値は実験的に定めた．ただし、1700Hz 以上の高次倍音については定位が一意に求められないので，(3) の条件は使用しない．

表 1　23 次元の特徴

特徴量一覧
周波数成分の最大パワーを 1 としたときのパワーの平均（基本波）
周波数重心を与える時間（基本波）
発音時からパワーが最大時までの時間（attack time）（基本波）
パワーが最大パワーの 5 割以上の時間（基本波）
パワーが最大パワーの 6.5 割以上の時間（基本波）
パワーが最大パワーの 8 割以上の時間（基本波）
最大パワーと中心時間のパワーの比（基本波）
attack 時のパワーと attack 時から 0.2sec までの最小パワーとの比
基本波のパワー値の時間変化
基本波のパワー包絡線の極値の個数（音の長さで正規化）
attack 時から音長の 75%までの基本波のパワー包絡線と近似直線の差の分散
各周波数成分のパワー値の時間変化の標準偏差の全倍音での平均
周波数重心（重みは各周波数成分の総パワー）
周波数重心の時間変化の標準偏差
基本波と第 2 倍音のパワー比
基本波と第 3 倍音のパワー比
基本波と第 4 倍音のパワー比
全パワーに対する 5 次倍音までのパワーの割合
偶数時倍音と奇数倍音の比（パワーの合計)
偶数時倍音と奇数倍音の比（attack 時のパワー)
周波数成分数
全持続時間の 7 割以上発音している高調波の個数
各周波数成分の総パワーの合計を基本波の総パワーで割った値

4.4　特徴抽出部

表 1 に示す 23 個の特徴量を抽出する．これらは，先行研究[7)8)9)10)]や楽器の特性を参考に決定した．また，音高により特徴量が変化する[10)]ことを考慮し，さらに基本周波数も特徴量として追加している．

4.5　音源同定部

各単音ごとに抽出した 24 次元の特徴量とテンプレートとの尤度を計算し一番尤度の高いクラスをその単音の楽器とする．識別器には多クラス対判別分析[11)]を用いる．

多クラス対判別分析は 2 段構成になっている．まず，群の対の組み合わせを設けて，2 群ごとにその平均間の距離を最大化するような変数を選択して判別分析を行い，次いで，各 2 群対から得られる対判別結果を min-

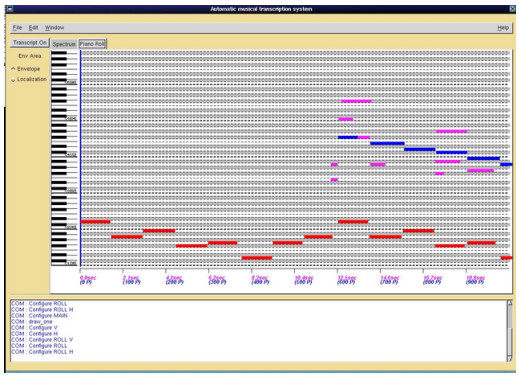

図 4　ピアノロール出力

imax 法により組み合わせて最終的な識別結果を決定する．以下に処理の概要を示す．

(1) すべての楽器対に対して対判別分析を行う．

(2) その結果得られた確率値から，あるサンプル x が i,j の 2 クラスの対判別分析でクラスΠ_iに属する確率の最小値を求める．

$$q_i = \min_j p_{i,j}(\Pi_i|x)$$

(3) q_iが最も大きいクラスを最終的な識別結果とする．

4.6　結果結合部

第 3 章では定位情報を用いてパートを形成し，パートごとに音源同定を行うことを述べた．しかし，周波数成分が重なり合うことで，安定した定位が求められず，定位を用いてパート形成をすることができない単音がある．

そこで，定位を求めることができた単音は，パート形成を行い，パート全体で音源同定を行う．定位を求めることができなかった単音はそれぞれ音源同定を行う．最後に，結果を結合することで，楽器ごとにグルーピングをすることができる．以上の処理により，ピアノロール形式（図 4）で結果が出力される．

4.7　音源同定予備実験

多クラス対判別分析の妥当性を検討するため，音源同定の予備実験を行う．実験には，音響信号単音データベース NTTMSA-P1 を用いた．NTTMSA-P1 のデータの内訳を表 2 に示す．本データベースは実楽器の単独発音を 48kHz，16bit，モノラルで収録したものである．表 2 のデータからランダムに 50%選び学習データとし，特徴量テンプレートを作成する．残りの 50%を実験の評価データとした．実験は 4 回繰り返し，その結果の合計を表 3 に示す．表 3 からわかるように単音では平均で 91.0%の音源同定精度である．

5.　システム評価実験

システムの評価のためのテスト曲として,「パッヘルベルのカノン」を作成した．NTTMSA-P1 のデータ

表 2　単音データベース NTTMSA-P1

楽器の種類	Piano（244 個），Violin（587 個），Trumpet（199 個），Flute（453 個），Clarinet（246 個）
音域	Piano:C0-C7，Violin:G2-C6，Trumpet:E2-C5，Flute:B2-D6，Clarinet:D2-G5
強さ	フォルテ，ノーマル，ピアノ
備考	通常の奏法（全楽器） ビブラート奏法（Violin,Flute） 各楽器に対して，2 種類の個体 （例 Piano:ヤマハ製，ベーゼンドルファー製）

表 3　実験結果（多クラス対判別分析）

	音源同定精度
Piano	96 %
Violin	95 %
Trumpet	86 %
Flute	88 %
Clarinet	90 %

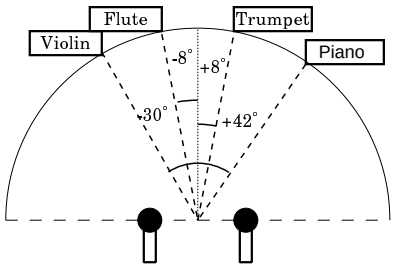

図 5　テスト曲の楽器配置パターン 1

を AKAI のサンプラー S6000 に格納し，パートごとに録音したものに時間差と強度差を与え，それを足し合わせて作成した．OPTIMA では「蛍の光」（Flute，Clarinet，Piano の 3 重奏）[2)3)]，三輪らはヴィヴァルディの四季「春」（Violin，Cello，Contrabass の 3 重奏）の第 1 楽章の最初の 4 小節[5)]を用いて実験を行っている.「パッヘルベルのカノン」は 4 声部からなる楽曲である．また，ある単音の基本波が他の単音の倍音構造と重なっている場合が大部分を占める．音源分離同定を行うには難易度の高い曲だと言える．テスト曲中に現れる 32 分音符は本システムでは対象外としている．楽器の配置図を図 5 に示す． 中央を 0 度とし，左は−右は+で表す．

5.1　同時的グルーピング実験

調波構造のみを用いて同時的グルーピングを行った場合と，調波構造と定位情報を用いて同時的グルーピングを行った場合の単音形成精度を比較する．楽器の定位情報や同時発音数などの事前知識は一切与えていない．

単音形成が成功した単音とは，音高が正しく，単音のオンセットが正解と誤差 0.2 秒以内である単音とした．評価として，単音形成の再現率，適合率を求める．再現率，適合率の求め方は以下の通りである．

$$再現率 = \frac{正解条件の通り単音形成された単音の数}{正解\ (楽譜)\ にある単音の数}$$

$$適合率 = \frac{正解条件の通り単音形成された単音の数}{本システムにより形成された単音の数}$$

5.1.1 ソロ演奏による評価

テスト曲をパートごとにソロ演奏で録音したデータに対して同時的グルーピングの精度を比較する．本実験で扱うテスト曲ではソロ演奏は単音旋律である．実験結果を表 4 に示す．

表 4　ソロ演奏単音形成結果

	単音形成再現率	単音形成適合率
調波構造のみを利用	98%	97%
調波構造と定位情報を利用	98%	97%

実験の結果，調波構造のみを利用した場合と，調波構造と定位情報を利用した場合では再現率，適合率の差はなかった．

5.1.2 重奏による評価

次にデュオ演奏，4 重奏を用いて同時的グルーピング精度を比較する．音長が短いと，単音形成，定位抽出の各精度に大きく影響する可能性がある．ここでは音長によって全体を 2 つのクラスに分類しそれぞれ評価することで，同時発音数の違い，音長の違いによる再現率，適合率の差が明確になるようにした．

(1) **クラス 1** 同時に発音する単音がすべて 8 分音符以上の長さである部分

(2) **クラス 2** 同時に発音する単音のうち少なくとも一つが 16 分音符である部分

まずはデュオ演奏に対して評価を行う．4 重奏の中の 2 つの楽器の組を複数（Violin-Piano，Flute-Piano，Trumpet-Piano，Violin-Flute の組み合わせ）選び実験を行った．各組の実験結果の平均を表 5，6 に示す．

重奏ではオクターブの関係が多く，調波構造のみによる同時的グルーピングではほとんど有効な処理結果を期待できない．それに対して，調波構造と定位情報を利用した場合では，単音形成の再現率はクラス 1 のパターンにおいてはソロ演奏から 2%しか低下しなかった．クラス 2 ではソロ演奏と比べ，15%程度の低下がみられるものの，調波構造のみを利用した場合と比べると約 40%高い．適合率はクラス 1，2 ともにソロと比べると大きく低下しているが，調波構造のみを利用した場合より約 20%高い．

次に 4 重奏に対して評価を行う．4 重奏では音の重なりが多くなり，調波構造のみによる同時的グルーピングは期待できなくなる．また，調波構造と定位情報を用いた同時的グルーピングの曖昧性の解消もいっそう複雑になる．

表 5　デュオ演奏単音形成結果（クラス 1）

	単音形成再現率	単音形成適合率
調波構造のみを利用	62%	53%
調波構造と定位情報を利用	96%	77%

表 6　デュオ演奏単音形成結果（クラス 2）

	単音形成再現率	単音形成適合率
調波構造のみを利用	43%	51%
調波構造と定位情報を利用	83%	71%

表 7　4 重奏単音形成結果（クラス 1）

	単音形成再現率	単音形成適合率
調波構造のみを利用	60%	62%
調波構造と定位情報を利用	90%	71%

表 8　4 重奏単音形成結果（クラス 2）

	単音形成再現率	単音形成適合率
調波構造のみを利用	34%	74%
調波構造と定位情報を利用	68%	62%

実験結果を表 7，8 に示す．デュオ演奏同様，調波構造のみによる同時的グルーピングはほとんど有効な処理結果を期待できない．それに対して，調波構造と定位情報を利用した場合では，基本周波数が整数倍の関係にある単音も形成することが可能であるため，調波構造のみを用いた場合に比べ再現率が高い．特にクラス 1 のパターンにおいては単音形成の再現率は 90%で高精度である．クラス 2 ではクラス 1 に比べが低下しているものの，調波構造のみを利用した場合より約 34%再現率が高い．

5.2 継時的グルーピング実験

継時的グルーピングにおける定位情報の有効性を検証する．実験は 4 重奏を用いて行う．単音形成は調波構造と定位情報を用いて行い，その結果得られた単音を音色情報のみを用いて継時的グルーピングを行った場合と，音色情報と定位情報を用いて継時的グルーピングを行った場合の音源分離同定精度を比較する．楽器は Piano，Violin，Trumpet，Flute，Clarinet のどれかであるとした．楽器の定位情報は与えていない．

実験は先の図 5 の楽器配置の他に，新たに図 6 の楽器配置でも実験を行う．図 6 は楽器を右側にすべて集めた場合であり，従来のパワー比のみを用いた継時的グルーピング[5] では難しい楽器配置である．

実験結果を表 9，10 に示す．最終的な音源分離同定結果なので，単音形成の再現率を越えることはない．つまり，クラス 1 では 90%，クラス 2 では 68%以上の音源分離同定精度はあり得ない．

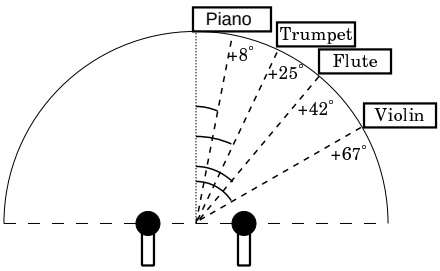

図 6　テスト曲の楽器配置パターン 2

表 9　テスト曲音源分離同定再現率（楽器配置パターン 1）

	クラス 1	クラス 2
音色情報のみを利用	62%	49%
音色情報と定位情報を利用	89%	62%

表 10　テスト曲音源分離同定再現率（楽器配置パターン 2）

	クラス 1	クラス 2
音色情報のみを利用	56%	47%
音色情報と定位情報を利用	84%	58%

音色情報のみを利用した場合では大きく再現率が低下しているのに対し，音色情報と定位情報を利用することで再現率低下を防ぐことができた．

5.3 考　　察

（1）同時的グルーピングにおける定位情報の有効性

表 5，6，7，8 からわかるように，調波構造のみによる同時的グルーピングは，オクターブの関係にある単音の組み合わせでは高い再現率は期待できない．それに対し，調波構造と定位情報を用いることで，周波数成分の重なりが認識され，重なり情報と各周波数成分の定位により，オクターブの関係にある単音も形成することが可能になる．4 重奏の平均で 32%再現率が向上したことから，同時的グルーピングにおいて調波構造と定位情報を用いることの有効性が示された．

（2）継時的グルーピングにおける定位情報の有効性

表 9，10 からわかるように，音色情報を用いて継時的グルーピングを行うと，正しい音源同定が行われていない．それに対し，定位情報を用いてパートを形成し，パート全体に対して音源同定を行うことで，継時的グルーピングの再現率は大きく向上している．音色情報のみを利用する場合に比べ，全体の音源分離同定精度が平均で約 20% 向上したことから，継時的グルーピングにおいて音色情報と定位情報を用いることの有効性が示された．

（3）位相差を用いることの有効性

図 6 のように 4 つの楽器を右側にすべて集めた場合は，パワー比のみを用いて継時的グルーピングが難しいと考えられる，位相差を用いることで，より詳細な定位を求めることができ，その結果，クラス 1 のパターンでは 84%で音源分離同定に成功した．

5.4 今後の課題

同時的グルーピング手法において本研究では 2 段階の処理を行っており，その 2 段階目の処理において重なりを満たす単音の組み合わせが出力された時点で終了している．しかし，重なりが正しく認識されているにもかかわらず，一意に単音の組み合わせを決定できない場合がある．例えば C3，C4，C5 が同時に発音している場合では，C3，C4 が出力された時点で終了してしまう．そのため 4 重奏において正しく重なりが認識されているにもかかわらず正しい単音の組み合わせが出力されない場合がある．また，一つの周波数成分の重なり判定を誤ると，連鎖的に複数の誤った単音が形成され，再現率が低下するという問題がある．

以上 2 つの問題に対処するためには，次の方法が有効だと考える．調波構造から考えられる単音の組み合わせの仮説を複数生成し，角度の変動により各周波数成分に重なりがある確率を求め，各仮説の尤度を計算する．その尤度が最も高い単音の組み合わせを出力する．この尤度計算は，単音の組み合わせの複数のパターンが確率で表されるため，今後他の情報との確率統合を行う場合にも有効であると考えられる．

6. おわりに

本論文では，自動採譜に必要な音源分離同定処理を同時的グルーピング・継時的グルーピングという 2 種類のグルーピング問題ととらえ，従来研究における各グルーピングの問題点を定位情報を用いることで解消を試みた．同時的グルーピングでは，調波構造と定位情報を用いることで，調波構造のみを用いるのと比べ，単音形成再現率が平均で 32%向上した．継時的グルーピングでは，音色情報と定位情報を用いることで，音色情報のみを用いるのと比べ，再現率が平均で 20%向上した．以上により，両グルーピングに定位情報を利用することの有効性が示された．

しかし，5.4 の今後の課題であげたように定位を用いた処理にはまだ性能向上の可能性が残されている．また，人間が音楽を聴く場合には定位情報・音色情報以外にも多くの手がかりを用いていると考えられる．今後はより多くの手がかりについても検討し，情報統合による高精度化を考えていくつもりである．

謝　　辞

本研究は，日本学術振興会から交付された科学研究費補助金および NTTCS 研から援助を受けた．また，音響信号データ NTTMSA-P1 の使用許可を下さった NTT コミュニケーション科学基礎研究所に感謝する．

参 考 文 献

1) A.S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.
2) 柏野邦夫, 中臺一博, 木下智義, 田中英彦. 音源情景分析の処理モデル optima における単音の認識. 電子情報通信学会論文誌, Vol. J79-DII, No. 11, pp. 1751–1761, 1996.
3) 柏野邦夫, 木下智義, 中臺一博, 田中英彦. 音源情景分析の処理モデル optima における和音の認識. 電子情報通信学会論文誌, Vol. J79-DII, No. 11, pp. 1762–1770, 1996.
4) 木下智義, 半田伊吹, 武藤誠, 坂井修一, 田中英彦. 自動採譜処理における知覚的階層に着目したパート分離処理. 電子情報通信学会論文誌, Vol. J85-DII, No. 3, pp. 373–381, 2002.
5) 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. 電子情報通信学会論文誌, Vol. J84-DII, No. 7, pp. 1251–1260, 2001.
6) *FFTW*. http://www.fftw.org/.
7) Antti Eronen and Anssi Klapuri. Musical instrument recognition using cepstral coefficients and temporal feature. In *Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.
8) Antti Eronen. Comparison of features for musical instrument recognition. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001.
9) 中臺一博, 田中英彦. 楽器演奏における単音の分離抽出とその音楽情景分析システムへの応用. Master's thesis, 東京大学, 1995.
10) 北原鉄朗, 後藤真孝, 奥乃博. '音高による音色変化に着目した音源同定手法. *SIGMUS*, Vol. 2001, No. 45, pp. 7–14, 2001.
11) T.Kawahara, T.Ogawa, S.Kitazawa, and S.Doshita. Phoneme recognition by combining Bayesian linear discriminations of selected pairs of classes. In *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, pp. 229–232, 1990.