2003－MUS－51　(7)
2003／8／5

自動採譜におけるパート形成処理のための特徴量の検討

櫻庭 洋平　　奥乃 博

京都大学大学院 情報学研究科 知能情報学専攻

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

あらまし　複数楽器による多重奏の自動採譜には，入力の音楽音響信号から得られた単音列をパートごとに分類するパート形成処理が必要である．パート形成処理の重要な手がかりは楽器の音色である．通常，音色を周波数成分のパワー変化などの特徴量で表現することが多いので，同時に複数の単音が発音している場合，周波数成分の重なりにより，音色を正確に抽出することは困難である．本稿では，パート形成に有効だと思われる手がかりとして音色類似度，定位近接度，音高遷移確率，音高高低関係維持度の 4 つを検討する．それらから求まるパート形成度を用いてパート形成を行う．実験の結果，定位近接度が最も有効な手がかりであった．4 つの手がかりを統合することで，79.00%正しくパートを形成することができた．

Comparing Features for Forming Music Stream in Automatic Music Transcription

Yohei Sakuraba　　Hiroshi G. Okuno

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

Abstract　This paper describes music stream formation in automatic music transcription. Although timbre of notes are in general the main clues for music stream formation, it may blurred however, when two or more notes are simultaneously played, and thus precisely extraction of features is difficult. The four keys, that is timbre similarity, direction proximity, note transition probability and pitch relation are exploited . Music streams are formed by integrating the four key. The result of experiments showed that the performance of music stream formation is around 79.00%.

1. はじめに

自動採譜は，作曲家や編曲家の支援としてだけではなく，音楽検索，音楽を理解するロボットなど，応用の面から見ても重要な技術であり，実現が望まれている．しかし，これまでの自動採譜は対象が単旋律に限られた場合が多く，多重奏を対象とした研究が行われ始めたのは 1990 年代に入ってからである[1)2)3)4)5)6)]．

多重奏を対象として，パートごとに音高☆を推定するためには，入力音楽音響信号から同時に発音する複数の単音☆☆の音高を推定する処理と，単音をパートごとに分類する処理の 2 つが必須となる．本研究では，前者を**同時的グルーピング**（simultaneous grouping），後者を**継時的グルーピング**（sequential grouping）と呼ぶ[7)]．

一般に，継時的グルーピングの重要な手がかりは楽器の音色である．通常，音色を周波数成分のパワー変化などの特徴量で表現することが多いので，複数音が発音すると，周波数成分の重なりや干渉により単音の音色を正確に抽出することが困難である．ところで，人間が音楽を聴く場合には，各時点に存在する音のみではなく，音のつながり，つまりパート全体を継続的に認識している．そのため，演奏の一部分だけを切り出して聞いた場合には，正しく単音を認識することは難しいものである．本研究で提案する手法では，パート追跡に音源同定処理を用いていない．従来の処理では，音源同定処理により同じ楽器と判断された単音をまとめてパートとしていた[1)8)9)]．本手法では，木下らと同様の立場に立ち，各単音を分類して複数のパートに分類することを考える[10)]．このことは，人間が音楽を聴くときに，音源名がわからなくてもパート追跡が可能であることからも，妥当な処理であると言える．また，三輪らは，音源の定位が近接した単音は同じパー

☆ C4，C5 などオクターブ情報を含む音名のことを指す．
☆☆ 本論文では，楽器演奏中の個々の音を単音と定義する．楽譜における一音符に相当する．

トに属するという仮定のもとで，入力ステレオ音楽音響信号から左右の強度差のヒストグラムを作り，クラスタリングをすることでパートごとの採譜を行っている[11]．この手法では，入力音楽音響信号は三重奏までに限定され，その三つの楽器も一つは中央，残り二つは左右に一つずつに限られていた．しかし，楽器数が増えた場合にも対応するためにはより詳細な定位が必要となる．

本研究では，継時的グルーピングの手がかりとして，音色類似度，定位近接度，音高遷移確率，音高高低関係維持度という4つを用いる．それらから求まるパート形成度が最大となるようにパートを形成する．

2. 自動採譜におけるパート形成処理

本研究では，複数楽器の多重奏の音楽音響信号から，同時に発音する複数の音高を推定し，それをパートごとに分類する処理を検討する．ここでは，音符のような楽譜表現ではなく，音高の時間変化として各パートを表現する．

パート形成処理において，音色は大きな手がかりである．従来は，単音ごとに音源同定を行い，同じ楽器と考えられる単音の集合をパートとしていた[1]．しかし，複数楽器の多重奏では，同時に発音する単音同士が干渉しあい，それが原因で正確な音色を得ることができなかった．高性能なパート形成処理を目指し，複数楽器の多重奏からより正確な音色を得る研究[9)12)]や，他の手がかりを用いて情報統合により解決する研究[9)10)]が行われた．前者では，性能が向上したものの，結果としては正確な音色を得ることができたとは言いがたい．また，音色が同じ楽器が複数発音した場合にはパート形成をすることができない．後者では，入力の単音がすべて正解であるものとし，どの単音をどのパートに分類するかという問題のみを扱っていた．実際の自動採譜では，すべての単音の音高を正しく推定できるわけではなく，発音していないのに発音していると判断した単音や，発音しているのに発音していないと判断した単音が存在する．これらを考慮したパート形成処理が必要となる．

本研究ではパート形成を行うために，次の4つを仮定する．

(1) 一つの旋律は，類似した音色の系列を持つ．
(2) 一つの旋律は，近接した定位の系列を持つ．
(3) 一つの旋律に現れる音高の遷移には，傾向がある．
(4) 旋律同士の音高の高低関係は維持する傾向がある．

以上の仮定は，多くの場合（特にクラシックなど）にあてはまり，妥当であると考える．これら4つの手がかりを本研究では次のように呼ぶことにする．

(1) 音色類似度
(2) 定位近接度
(3) 音高遷移確率
(4) 音高高低関係維持度

特に，定位近接度と音高高低関係維持度は我々が新たに提案した手がかりである．音色類似度は，2単音の楽器名が等しいかどうかを学習した音色類似度モデルを利用して求める．音高遷移確率は，クラシック音楽50曲を分析して学習したトライグラムモデルを利用する．以下，4つの手がかりを説明する．

2.1 音色類似度

音色類似度は，ある2つの単音の音色がどれだけ類似しているかを示す値である．各単音は，パワーエンベロープの立ち上がりの強さや周波数重心などに関する23次元の音色特徴ベクトルで表現されている．音色特徴量は，従来研究を参考に抽出した[13)14)]．

2単音が同じ楽器である群，2単音が異なる楽器である群をそれぞれΠ_0，Π_1とする．2単音 $note_j$，$note_k$の音色特徴ベクトルの差を x_{jk}，音色類似度を $P_t(note_j, note_k)$ とする．x_{jk}が群Π_0に属する事後確率 $p(\Pi_0|x_{jk})$ を確率密度 $p(x|\Pi_i)$ と群Π_iの正規する事前確率 $p(\Pi_i)$ からベイズ規則により求め，それを音色類似度 $P_t(note_i, note_j)$ とする．ここでは，すべての群の事前確率を等しくしている．

$$P_t(note_j, note_k) = p(\Pi_0|x_{jk}) = \frac{p(x_{jk}|\Pi_0)}{p(x_{jk}|\Pi_0) + p(x_{jk}|\Pi_1)}$$

また，標本 x の確率密度関数 $p(x|\Pi_i)$ は

$$p(x|\Pi_i) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} exp\{-\frac{D_i^2(x;\mu_i)}{2}\}$$

で与えられる．ここで，dは各群の正規分布の次元数，$D_i^2(x;\mu_i)$ は x と群Π_iの平均μ_iとのマハラノビス距離であり，その定義は

$$D_i^2(x;\mu_i) = (x-\mu_i)^t \cdot \Sigma^{-1} \cdot (x-\mu_i)$$

である．ただし，各群の共分散行列を等しい（$\Sigma = \Sigma_i$）と仮定している．

ある時点において，パート $part$ が c 個の単音から成り立っているとする．$part$ と単音 $note$ の音色類似度 $P_t(part, note)$ は，$part$ に含まれるすべての単音 $note_i (1 \leq i \leq c)$ と $note$ の音色類似度の平均とする．

$$P_t(part, note) = \frac{1}{c}\sum_i^c P_t(note_i, note)$$

音色類似度モデルは，NTTMSA-P1（5楽器，1353サンプル）を用いて学習した．

2.2 定位近接度

定位近接度は，パートと追跡しようとする単音の定位がどれだけ近接しているかを示す値である．単音の

定位は，ステレオ音響信号の左右の位相差 IPD を利用して求める[15)16)]．

単音の定位は次のように求める．

（１）単音のオンセットからオフセットまでの全フレーム（短時間フーリエ変換の一区間）に対し，その単音の基本周波数から 20 次倍音までを抽出する．各周波数成分は，単音の音高に対応する周波数（平均律で算出）の近傍（50 セント以内）に存在するピークの周波数とする．

（２）ピークごとに左右のスペクトルから IPD を求める．

$$IPD = \tan^{-1}\left(\frac{Im[Sp(l)]}{Re[Sp(l)]}\right) - \tan^{-1}\left(\frac{Im[Sp(r)]}{Re[Sp(r)]}\right)$$

ただし，各ピークの左右のスペクトルを $Sp(l)$，$Sp(r)$ とし，Re，Im でそれぞれ実部，虚部を表す．IPD から次の式に基づいて定位θを求める．

$$\theta = sin^{-1}\left(\frac{c}{2\pi f l}\left(IPD \pm 2n\pi\right)\right) \quad (n = 1, 2, \cdots)$$

ただし，f はピークの周波数，l はマイク間の距離（$0.2m$），c は音速（$340m/s$）を表す．

（３）測定された各ピークの定位から，級間隔 1 度の定位ヒストグラムを得る．

（４）ヒストグラムの度数が最大となる階級値をその単音の定位 $D(note)$ とする．

パートの定位 $D(part)$ は，そのパートに属するすべての単音の定位ヒストグラムを足し合わせたものである．このとき，パート $part$ と単音 $note$ の定位近接度 $P_d(part, note)$ を次の式で定義する．

$$P_d(part, note) = 1 - \frac{|D(part) - D(note)|}{2\ T_d\left(D(part)\right)}$$

ここで，$T_d(x)$ は定位の誤差であり，次式で定義する．中央（0 度）で T_c度，外側（90 度）で T_o度，間は線形補間された値となるように設計した．

$$T_d(x) = T_c + (T_o - T_c) \cdot \frac{|x|}{90}$$

2.3 音高遷移確率

旋律中に現れる音高の遷移には，現れやすい遷移とそうでない遷移がある．音高の遷移はランダムに行われているわけではない．そこで，統計的に得た音高遷移確率をパート形成の手がかりとして利用する．

統計的なモデルとして音高の生起確率が直前の 2 個の音高のみに依存するとしたトライグラムモデルを用いる．このとき，パート $part$ と，ある単音 $note$ との音高遷移確率 $P_n(part, note)$ は，

$$P_n(part, note) = p(note|note_{c-1}, note_c)$$

と表せる．ただし，c は $part$ 内の単音の個数，$note_{c-1}$ は $c-1$ 番目の単音，$note_c$は c 番目の単音を表す．

トライグラムの学習には，RWC 研究用音楽データベース：クラシック音楽（RWC-MDB-C-2001）[17)] の全曲（50 曲）を用いた．データベース付属の MIDI データから単旋律のトラック（またはその一部）を抽出し，音高遷移を得た（総音符数：167179 個）．音高遷移の出現頻度は，その曲の調性に大きく依存するため，全体でそのまま統計をとることには意味がない．そこで，あらかじめ調性によって正規化した上で統計をとった．また，長調と単調でそれぞれ統計をとった．調は，RWC 研究用音楽データベースの楽譜から得た．トライグラムモデルの作成には，CMU-Cambridge ツールキットを用いた[18)]．

表 1　音高遷移確率のパープレキシティ

	長調	短調
音高数	84	82
パープレキシティ	6.54	7.76

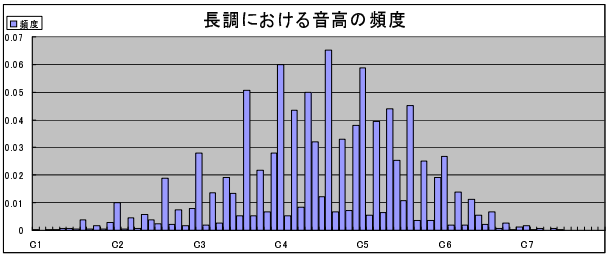

図 1　長調の音高出現頻度

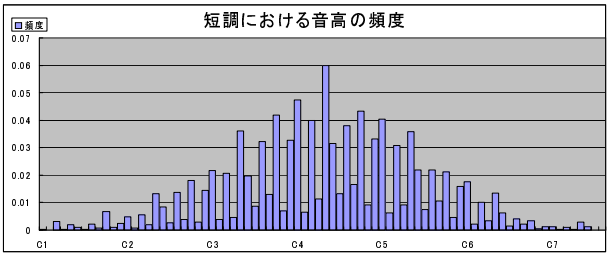

図 2　短調の音高出現頻度

求めたトライグラムのパープレキシティを表 1 に示す．パープレキシティは，情報理論的な意味での平均分岐数に相当する．

長調，短調における音高の出現頻度（ユニグラムに相当）を図 1，2 に示す．ただし，正規化後の中央のドが C4 となるように，表現している．音高，音域ごとに出現頻度にばらつきがあることがわかる．

2.4 音高高低関係維持度

アンサンブル演奏などの楽曲において，各パートは主旋律やベースなどといった，音楽的な役割を持っている．そのため，主旋律を担当している楽器とベースを担当している楽器が入れ替わることはまれである．また，弦楽四重奏（第一バイオリン，第二バイオリン，ビオラ，チェロ）などにおいても，大部分はチェロが

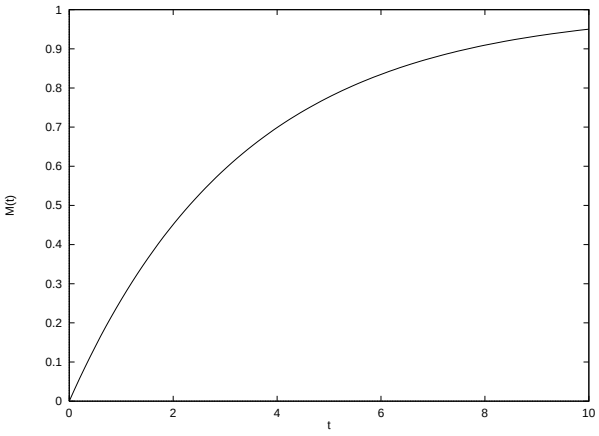

図 3　関数 $M(t)$ のグラフ（C=0.3 のとき）

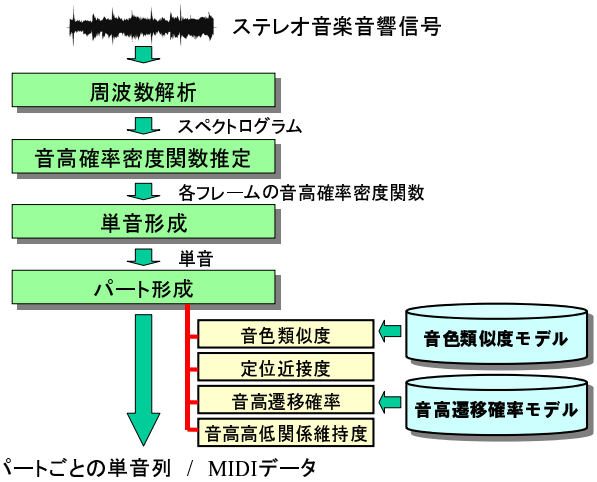

図 4　本研究で提案する自動採譜システムの流れ

最低音を担当し，第一バイオリンが最高音を担当する．

このような，パート間における音高の高低関係が維持される傾向をパート追跡に導入する．このとき，音高の高低が頻繁に入れ替わるパートには，悪影響がないように設計する必要がある．

ある時点でパート $part$ が単音 $note$ とパート形成をすることで，他のパート $part_i$ $(1 \leq i \leq N)$ より，音高が高く（低く）なるとする．$part$ と $part_i$が同時に発音している時間に対し，$part$ の音高が高い（低い）時間の割合を q_iとする．このとき，音高高低関係維持度 $P_r(part, note)$ を次のように定義する．

$$P_r(part, note) = \frac{1}{N}\sum_{i}^{N}\left(\frac{1}{2} + (q_i - \frac{1}{2}) \cdot M(time_i)\right)$$

$time_i$は $part$ と $part_i$が同時に発音している時間を 1 小節を 1 として正規化した値である．$M(t)$ は，

$$M(t) = 1 - exp(-C \cdot t)$$

であり，図 3 のように，同時に発音している時間が長くなるに連れて 1 に近づく関数である．音高高低関係維持度 $P_r(part, note)$ は，最初は$\frac{1}{2}$であるが，時間が経つに連れて$\frac{1}{N}\sum_{i}^{N} q_i$に近づく．

3. 本研究で提案する自動採譜システムの流れ

本章では，提案手法の流れを述べる．まず，入力の音楽音響信号に対し，周波数解析を行う．次に，EM アルゴリズムを用いて，フレームごとの音高確率密度関数を推定する．得られた確率密度関数を継時的に追跡し，単音を形成する．さらに，マルチエージェントモデルを用いて，単音からパートを形成する．パート形成のための手がかりには，音色類似度，定位類似度，音高遷移確率，音高高低関係維持度から求まるパート形成度を利用する．

3.1　周波数解析

入力音楽音響信号に対し，短時間フーリエ変換を用い，スペクトログラムを作成する（窓関数：ハミング窓，窓幅：4096 点，シフト幅：1000 点）．フレームごとに，パワースペクトルの極大値（ピーク）を抽出する．

3.2　音高確率密度関数推定

各フレームのパワースペクトルに対し，各高調波構造が相対的にどれだけ優勢かを表す音高の確率密度関数を求める[19]．本手法では，あるフレーム t の周波数成分の確率密度関数 $p^{(t)}(x)$ が，高調波構造をもつ音をモデル化した確率分布（音モデル）の混合分布モデルから生成されたと考える．音高（C4，G5 など）が F の音モデルの確率密度関数を $p(x|F)$ とすると，その混合分布モデル $p(x;\theta^{(t)})$ は次式で定義できる．

$$p(x;\theta^{(t)}) = \int_{Fl}^{Fh} w^{(t)}(F)p(x|F)dF$$

$$\theta^{(t)} = \{w^{(t)}(F)|Fl \leq F \leq < Fh\}$$

Fl，Fh は許容される音高の上限と下限であり，$w^{(t)}(F)$ は次式を満たす音モデル $p(x|F)$ の重みである．

$$\int_{Fl}^{Fh} w^{(t)}(F)dF = 1$$

もし，観測された周波数成分の確率密度関数 $p^{(t)}(x)$ が混合分布モデル $p(x;\theta^{(t)})$ から生成されたかのようにモデルパラメータ$\theta^{(t)}$を推定できれば，$p^{(t)}(x)$ が個々の音モデルへと分解されたとみなすことができる．その重み $w^{(t)}(F)$ を，音高の確率密度関数 $p^{(t)}_{pitch}(F)$ と解釈できる．

$$p^{(t)}_{pitch}(F) = w^{(t)}(F) \quad (Fl \leq F \leq Fh)$$

以上から，確率密度関数 $p^{(t)}(x)$ を観測したときに，そのモデル $p(x;\theta^{(t)})$ のパラメータ$\theta^{(t)}$を推定する問題を解けばよい．$\theta^{(t)}$の最尤推定量は，次式で定義される平均対数尤度を最大化することで得られる．

$$\int_{-\infty}^{\infty} p^{(t)}(x) \log p(x;\theta^{(t)}) dx$$

ここで，音モデルの確率密度関数 $p(x|F)$ を次のように仮定する．

$$p(x|F) = \alpha \sum_{h=1}^{N} c(h) G(x; F0(F) \times h, W)$$

$$G(x;m,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-m)^2}{2\sigma^2}}$$

ここで，αは正規化定数，N は考慮する周波数成分の数，$F0(F)$ は音高 F の基本周波数（平均率で算出），W はガウス分布 $G(x;m,\sigma)$ の分散を表す．$c(h)$ は h 次高調波成分の振幅を決める関数で，

$c(h) = G(h;1,H)$

とする（H は定数）．

3.3 単音形成

音高の確率密度関数の時間変化において，複数のピークの軌跡を継時的に追跡し，優勢で安定した音高の軌跡を単音とする．つまり，各フレームにおける音高の確率密度関数を入力として，単音（音高，オンセット，オフセットの組）を出力することを考える．

その際，動的に生成・消滅するピークの軌跡を，並行して追跡する必要がある．後藤は，マルチエージェントモデルを用いてこの処理を実現している[19]．その手法では，各エージェントは F0 を手がかりに，F0 確率密度関数の目立つピークを追跡している．

本研究では後藤のマルチエージェントモデルをもとにしている．しかし，追跡の手がかりに F0 ではなく音高を用いている点や，相互作用がないという点で大きく異なる．そのため，マルチエージェントモデルとは言わないが，便宜上音高の確率密度関数を追跡して単音を形成するものをエージェントと呼ぶ．

本研究で用いるモデルは，一つの特徴検出器と複数のエージェントで構成される．各エージェントは，追跡中の音高 M の他に，追跡中の軌跡の信頼度CM_{Note}と累積ペナルティPEN_{Note}を保持し，各時刻において以下のステップによってこれらを更新する．

(1) 検出器は，音高の確率密度関数 $p^{(t)}_{pitch}(F)$ の中で目立つピークを複数検出する．

(2) 目立つピークは，同じ音高のエージェントに割り当てられる．目立つピークが割り当てられなかった場合は，そのピークを追跡する新たなエージェントを生成する．

(3) 目立つピークが割り当てられたエージェントの累積ペナルティはリセットされる．割り当てられなかった場合は，一定のペナルティを受け，音高の確率密度関数 $p^{(t)}_{pitch}(F)$ の中から自分の追跡するピークを直接見つけようとする．それも見つからない時には，さらにペナルティを受ける．累積ペナルティが一定の閾値を超えると，そのエージェントは消滅する．

(4) 各エージェントは，割り当てられたピークに応じて，以下のように信頼度CM_{Note}を増減する．

$$CM_{Note} = \frac{1}{N} \sum_{t}^{N} p^{(t)}_{pitch}(M)$$

ピークの確率密度関数値の平均値である．ただし，N は追跡したピークの個数を表す．

(5) CM_{Note}が閾値 T_{CM}を超えるエージェントを単音として出力する．

3.4 パート形成

次に，得られた単音をパートごとに分類する．その際，動的に生成・消滅する複数のパートを，相互の干渉を考慮しながら並行して追跡することが不可欠である．そこで本研究では，動的な追跡処理を柔軟に制御することを可能にするために，マルチエージェントモデルを導入する．

本研究で提案するマルチエージェントモデルは検出器，複数のエージェント，音色類似度モデル，音高遷移確率モデル（音高の遷移のトライグラムモデル）で構成される．各エージェントはパートであり，単音を追跡し，パートを形成する．

本稿では，パート形成の手がかりとして **2** 章で述べた音色類似度，定位近接度，音高遷移確率，音高高低関係維持度の 4 つを用いる．4 つの積であるパート形成度が最大となるようにパートを形成する．

従来は，入力の単音がすべて正解であるものとし，どの単音をどのパートに分類するかという問題のみを扱っていた[9][10]．実際の自動採譜では，すべての単音の音高を正しく推定できるわけではなく，発音していないのに発音していると判断した単音や，発音しているのに発音していないと判断した単音が存在する．これらを考慮したパート形成処理が必要となる．

本研究では，楽器名が異なれば異なるパートに属し，定位が離れていれば異なるパートに属する，と仮定する．これらの仮定は多くの場合にあてはまり，妥当であると言える．そこで，パートと次の単音候補の音色類似度と定位近接度の平均が 0.5 未満であれば，その単音を候補から取り除く．

各エージェントは，追跡中の軌跡の信頼度 CM_{Part} と累積ペナルティPEN_{Part}を保持し，処理単位（32 分音符に相当する時間：1block）ごとに以下のステップ（最初の 3 ステップは図 5 に対応）によってこれらを更新する．

(1) 検出器は，処理単位内にオンセットを持つ，信頼度の高い単音（信頼度が T_{CM}を超える単音）

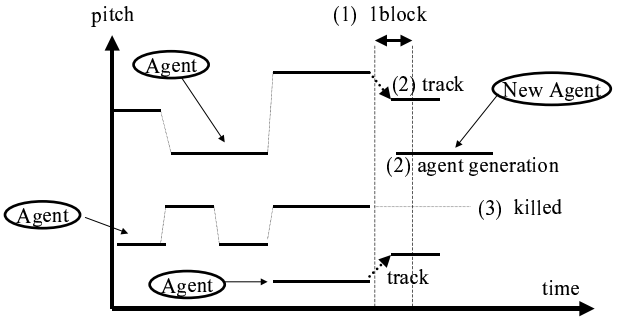

図 5 マルチエージェントモデルによるパート形成

を複数検出する．

(2) 検出された信頼度の高い単音は，音色類似度 P_t，定位近接度 P_d，音高遷移確率 P_n，音高高低関係維持度 P_r の積であるパート形成度が最大のエージェントに割り当てる．ただし，音色類似度 P_t と定位近接度 P_d の平均が 0.5 未満であれば，割り当て候補から取り除く．信頼度の高い単音がどのパートにも割り当てられなかった場合は，その単音を追跡する新たなエージェントを生成する．

(3) 信頼度の高い単音を割り当てられたエージェントの累積ペナルティは，リセットされる．割り当てられなかった場合は，一定のペナルティを受ける．累積ペナルティが閾値を越えるとエージェントは消滅する．

(4) 各エージェントの信頼度 CM_{Part} は，割り当てられた単音の信頼度 CM_{Note} に応じて増減する．割り当てられた単音 $Note_i(1 \leq i \leq N)$ の信頼度の平均とする．

$$CM_{Part} = \frac{1}{N}\sum_{i}^{N} CM_{Note_i}$$

4. システム評価実験

4.1 実験内容

ステレオ音楽音響信号を入力とし，推定した音程とパートを出力する提案手法に基づいたシステムを評価する（パラメータの値を表 2 に示す）．ベンチマークデータには無響室で録音した「パッヘルベルのカノン」の四重奏（演奏時間：6 分 30 秒）を用いた．四重奏を AKAI のサンプラー S6000（楽器音データベース NTTMSA-P1 を保存したもの）で 4 台スピーカーを用いて各パートを再生し，2 本のマイク（マイク間距離：$0.2m$，マイクの中心とスピーカーの距離：$1m$）で録音した．楽器の配置は，左から，violin，flute，trumpet，piano で固定とした．楽器の間隔は 20 度間隔，40 度間隔，60 度間隔の 3 パターンとした．

表 2 パラメータの値

$W = 0.015$	$H = 4.0$
Fl = D2	Fh = F#6
$T_c = 10$	$T_o = 20$
$C = 3.0$	$T_{CM} = 0.06$

表 3 パート形成結果

音色類似度	定位近接度	単音遷移確率	音高高低関係維持度	精度
○	−	−	−	62.60%
−	○	−	−	78.26%
−	−	○	−	53.18%
−	−	−	○	49.15%
○	○	−	−	76.64%
○	−	○	−	62.42%
○	−	−	○	63.04%
−	○	○	−	80.42%
−	○	−	○	78.64%
−	−	○	○	53.26%
○	○	○	−	79.00%
○	○	−	○	76.90%
○	−	○	○	62.47%
−	○	○	○	80.42%
○	○	○	○	79.00%

単音形成処理では，システムが出力した単音を楽譜と比較し，音高が正しく，発音時刻の誤差が 32 分音符以内であれば正解とする．再現率と適合率を次のように定義する．

$$\text{再現率} = \frac{\text{正しく形成された単音の個数}}{\text{楽譜上の実際の単音の個数}}.$$
$$\text{適合率} = \frac{\text{正しく形成された単音の個数}}{\text{システムが出力した単音の個数}}.$$

このとき，再現率，適合率はそれぞれ，66.4%，76.0%であった．

パート形成処理では，音高，発音時刻に加え，出力のパートに含まれる単音が入力の楽譜で同一音源に由来するものを正解とした．パート形成処理の精度 R は，

$$R = \frac{\text{正しくパート形成された単音の個数}}{\text{単音形成処理で出力された正解単音の個数}}$$

とした．

4 つの手がかりのうち，1 つの手がかりだけを用いた場合から 4 つの手がかり全てを用いた場合までの合計 15 通りでパート形成処理の精度を比較した．

4.2 実験結果と考察

結果を表 3 に示す．音色類似度のみでは，62.60%であるが，4 つの類似度を統合することで，79.00%まで性能が向上した．しかし，性能が最大となる手がかりの組み合わせは，音色類似度以外の 3 つを用いる場合であった．また，音色類似度を用いない場合に 80.42%であることは，全て同じ楽器の四重奏においても，8 割程度のパート形成が可能であることを意味している．

4.2.1 音色類似度の考察

4つの手がかりをすべて統合した場合よりも，音色類似度以外の3つの手がかりを利用した場合が性能が高かった．この理由は，音色類似度は単音で学習したモデルを用いており，調波構造の重なりによる特徴変動を考慮できていないためであると考えられる．今後は，調波構造の重なりを考慮した音色特徴を検討し，混合音にも有効な音色類似度モデルを用いる必要がある．

4.2.2 定位近接度の考察

単一の手がかりごとに比較した場合，定位近接度の効果が他の3つの手がかりに比べて大きい結果となった．定位近接度のみで78.26%の性能であり，統合した場合に大きな効果が得られている．位相差が含まれている音楽音響信号に対しては効果的であることを示している．今後は，強度差のみからなる音楽音響信号や，残響が含まれている音楽音響信号に対しても，定位近接度の有効性を確かめる必要がある．

4.2.3 音高遷移確率の考察

音高遷移確率以外の3つの手がかりを統合した場合の性能は76.90%であるが，音高遷移確率を統合することで79.00%まで性能が向上している．本研究で学習した音高遷移確率ではオクターブ以内の遷移が約97%あった．3オクターブのような大きな遷移（今回の場合は間違った遷移）を防ぐ意味で効果があったと考えられる．この手がかりは，型破りの演奏でなければ効果が得られると考えられる．今回はクラシック音楽のみを用いて統計をとったが，ジャズやポピュラーなどに対しても統計をとることで，より多くのジャンルへの対応が可能となると考えられる．

4.2.4 音高高低関係維持度の考察

音高高低維持度は，統合することで性能が上がっているものの，定位近接度に比べると効果は高くない．今回評価に用いた曲では，中音域の3パートの音高が頻繁に交差しており，音高高低関係維持度に差が出なかったためであると考えられる．しかし，この手がかりは，統合することで性能が下がることはなかった．また，パート間の音域が離れた曲に対しては有効に働くと考えられる．より多くの曲で実験し，評価する必要があると考えられる．

5. おわりに

本研究では，自動採譜におけるパート形成処理の手がかりとして，音色類似度，定位近接度，単音遷移確率，音高高低関係維持度の4つを提案した．4つの手がかりの有無によるパート形成性能を四重奏を用いて評価した結果，音色類似度のみでは62.60%の性能であるのに対し，4つの手がかりすべてを統合した場合に79.00%まで性能が向上した．しかし，このパート形成性能はまだ十分な値とは言えない．今後は，混合音にも有効な音色類似度の設計など，各手がかりについても洗練する必要がある．また今回は，パートと次の単音候補の音色類似度と定位類似度の平均が0.5未満であれば，単音候補から除去しているが，この除去処理についても評価する必要がある．

提案手法により，4つの手がかりを用いてある程度パート形成が可能であることがわかった．そこで今後は，パート形成処理により単音形成の誤りを補正する処理も実現していく予定である．

謝　辞

本研究は，日本学術振興会科学研究費補助金基盤研究 (A) 第15200015号，(B) 第12480090号およびサウンド技術振興財団研究助成による．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，無響室を貸してくださった国際電気通信基礎技術研究所に感謝する．また，有益なご助言をくださった片寄 晴弘氏（関西学院大学），柏野 邦夫氏（NTTコミュニケーション科学基礎研究所），中臺 一博氏（株式会社ホンダ・リサーチ・インスティチュート・ジャパン）に感謝する．

参考文献

1) 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.
2) 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.
3) A. Klapuri, T. Virtanen, A. Eronen, and J. Seppanen. Automatic transcription of musical recordings. In *Consistent & Reliable Acoustic Cues Workshop (CRAC)*, 2001.
4) K. D. Martin. *Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing.* M.I.T. Media Lab Perceptual Computing Technical Report #399, 1996.
5) R. Mani and S. H. Nawab. Integration of dsp algorithms and musical constraints for the separation of partials in polyphonic music. In *Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1741–1744, 1998.
6) 亀岡弘和, 西本卓也, 篠田浩一, 嵯峨山茂樹. ハーモニック・クラスタリングによる多重奏の基本周波数推定アルゴリズム. 情報処理学会研究報告 2002-MUS-50, pp. 27–32, 2003.
7) A.S. Bregman. *Auditory Scene Analysis.* MIT Press, 1990.
8) 柏野邦夫, 村瀬洋. 適応型混合テンプレートを用いた音源同定. 電子情報通信学会論文誌, Vol. J81-DII, No. 7, pp. 1510–1517, 1998.
9) 柏野邦夫, 村瀬洋. 単音連鎖ネットワークに基づく音楽演奏の音源同定. 人工知能学会誌, Vol. 13, No. 6, pp. 962–970, 1997.

10) 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.
11) 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.
12) 木下智義, 坂井修一, 田中英彦. 周波数成分の重なり適応処理を用いた複数楽器の音源同定処理. 電子情報通信学会論文誌, Vol. J83-DII, No. 4, pp. 1073–1081, 2000.
13) 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.
14) T. Kitahara, M. Goto, and H. G. Okuno. Musical instrument identification based on f0-dependent multivariate normal distribution. In *Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.
15) 境久雄. 日本音響学会編 音響工学講座 6 聴覚と音響心理. コロナ社, 1978.
16) K. Nakadai, H. G. Okuno, and H. Kitano. Real-time sound source localization and separation for robot audition. In *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, pp. 193–196, 2002.
17) 後藤真孝, 橋口博樹, 西村拓一, 岡隆一. Rwc 研究用音楽データベース: クラシック音楽データベースとジャズ音楽データベース. 情報処理学会研究報告 2002-MUS-44, pp. 25–32, 2002.
18) *CMU-Cambridge SLM ToolKit.* http://www-svr.eng.cam.ac.uk/~ prc14/toolkit.html.
19) 後藤真孝. 音楽音響信号を対象としたメロディとベースの音高推定. 電子情報通信学会論文誌, Vol. J84-DII, No. 1, pp. 12–22, 2001.