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

木下　智義[†,††]　半田　伊吹[†]　武藤　誠[†,†††]　坂井　修一[††††]
田中　英彦[††††]

Musical Part Separation Based on Perceptual Hierarchy

Tomoyoshi KINOSHITA[†,††], Ibuki HANDA[†], Makoto MUTO[†,†††], Shuichi SAKAI[††††], and Hidehiko TANAKA[††††]

あらまし　音響信号により外界の事象を理解する聴覚的情景分析に関して，従来多くの研究がなされてきた．特に対象を音楽に絞った場合，自動採譜等の実現を目指した研究例がいくつかある．しかしながら，従来の処理では各時点における局所的な処理に終始するものが多く，時間方向の処理を進めた例であっても，その対象は時間的に近接した範囲にとどまっていたため，処理性能に限界があった．本論文では，それを改善することを目的として，音響ストリームの知覚的な階層構造に着目し，より大局的な範囲での処理を用いて，楽曲からパートに相当する単音列を抽出する処理を提案する．本手法では，パート抽出に際してフレーズを中間的に形成した．フレーズ形成で局所的な手がかりを，パート形成で大局的な手がかりを用いる．これにより，計算量の爆発等の問題を招くことなくパートを抽出することに成功した．予備的な評価実験の結果，再現率 80%，適合率 85% 程度の精度でパート抽出に成功した．

キーワード　聴覚的情景分析，自動採譜，音響ストリーム，音源分離

1. はじめに

自動採譜処理に関する研究が行われるようになったのは，1970 年代のことである．この時期では主に FFT のみを用いた単純な手法であって，自動採譜よりはむしろ単音抽出というべきものであった．対象とする音響信号も，単一音源による単旋律に限定されていた．やがて 1990 年代に入り，数は少ないものの，複数音源による複数旋律の演奏を対象として自動採譜処理が試みられるようになった [2], [3], [6]〜[9]．しかしながら，いずれの例においても実用上十分な精度は得られてなかった．

† 東京大学大学院工学系研究科, 東京都
Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-8656
†† 現在，有限会社ネットコンパス, 東京都
NetCOMPASS Ltd., 5-17-8-207 Minami-Senju, Arakawa-ku, Tokyo, Japan, 116-0003
††† 現在，NTT サイバースペース研究所, 神奈川県
NTT Cyber Space Lab., 1-1 Hikari-no-Oka, Yokosuka-shi, Kanagawa, Japan, 239-0847
†††† 東京大学大学院情報理工学系研究科, 東京都
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-8656

従来処理において，十分な精度が得られなかった原因はいくつかあるが，その中心となるのは，処理システム上に準備されたテンプレートや知識と，入力信号との差を吸収しきれないことである．これは楽器個体間の特性の差や，同一個体でも音域の間の差などによって現れるが，他にも，同時刻に複数の音が存在することによる干渉等により受ける変形の影響が大きい．

我々は既にこの問題に対応するために，入力信号から得られる周波数成分の特徴量を音の重なりに応じて再計算する手法を提案したが，精度は改善されたものの，実用的な処理精度には届かなかった [3]．

一方，人間が音楽を聴く場合には，音の重なりによってある音が別の音源のものに聞こえたり，また楽器個体によって別の楽器に聞こえることは稀である．しかし，ある一つの音のみを聴いた場合，特にその継続時間が短い場合には，その音源が何であるかを判断するのは難しい．

このことは，人間が音楽を耳にするときには各時点に存在する音のみではなく，いくつかの音，あるいはパート全体を継続的に認識していることを意味する．実際，ある継続音中の一部分のみを雑音で隠蔽した場合にでも，正しく認識が行われるのはよく経験される

ことである．

このことから，人間が聴覚情景分析を行う際には，各時点の音を個別に弁別あるいは理解しているのではなく，継続的な音のつながりを理解の対象として扱っていることが分かる．この対象を Bregman は「音響ストリーム」と呼んだ [1]．実際に複数音が存在する環境下では，ある音が別の音によって隠蔽されることはしばしば起こる．そのため，計算機上で聴覚的情景分析を実現することを考えた場合にも，この点を考慮して，ストリームに相当する対象を抽出した上で，そこに含まれる時間的に継続した情報を元に重なりによって失われた情報を補完することが必要である．

しかしながら，従来研究ではこの点が十分に活用されてきたとは言いがたい．従来研究における時間処理の例を考えると，OPTIMA [6] では和音遷移に基づいた処理を行っているが，いずれも時間的に隣接する単音，和音間のみを考慮した処理がなされており，より長い時間的な要因を考慮した処理としては不十分なものである．また，Ipanema [8] においては,「単音連繋ネットワーク」を構成して，音源同定処理の誤りを補完する処理を用いている．しかしながら，このネットワークの構成には隣接単音間での遷移確率のみを用いており，局所的な処理に留まっている．

そこで，本論文では，実用的な処理精度を実現するために，時間的かつ大局的な情報を用いることで音源分離処理を精度よく行う手法を提案し，検討する．

2. 知覚的ストリーム

人間がある程度の長さをもつ音響信号を耳にするとき，断続的な信号や高さの連続しない音列を一まとまりに捉えることが多い．その一方で，個々の単音を個別に認識することも可能である．このことから，音響ストリームに関しても，階層的な知覚対象の構造を考えることができる．

特に対象を音楽に特化した場合，知覚の対象として以下に述べるものが考えられる．

1) 単音: 楽譜における音符に相当する音響信号．これはもっとも狭い意味で継続的な信号であると言える．

2) フレーズ: 楽譜においては 1 小節〜数小節程度に相当する．人間がひと続きのブロックとして認識が可能な，時間的に近接した複数単音の集合と言うことが可能である．本論文では，局所的な手がかりをもとに同一音源に由来すると判断される時間的に連続した複数単音の集合として定義する．

3) パート: 楽譜では各楽器の担当する譜面全体を指す．ピアノの演奏する部分，ヴァイオリンの演奏する部分，というように，一つの音源が一つのパートに相当するとも言える．

本論文においては，この単音，フレーズ，パートの 3 層構造を考慮し，音響ストリームの表現としてパートを抽出する手法を提案する．また，本論文では，単音からフレーズを形成する処理をフレーズ形成，フレーズからパートを形成する処理をパート形成，これら 2 つの処理をあわせてパート抽出と呼ぶ．

2.1 パート抽出の意義

本論文の手法で抽出するパートは，そのそれぞれが 1 つの音源に対応する．したがって，パートの抽出は音源分離処理に相当するものである．

一方で，提案する手法ではパート抽出に音源同定処理を用いていない．従来の処理では，音源同定処理の結果から入力単音を各音源に対応づけ，それをもって音源分離の結果としていた [6]．それに対し，本手法では各音源が何であるかの識別はさておき，各単音を分類して複数の集合に分けることを考える．このことは，音源同定処理の誤りの影響を受けないという点から有効な手法である．また，実際に人間が音楽を聴くときに，音源の種類を意識せずに音源分離が可能であるということからしても，自然な手法であると言える．

本手法のパート抽出処理を用いた場合，各単音がパートごとに分類された後に音源同定処理を行うことになる．従来の音源同定手法は，個々の単音に関してその特徴量等を用いて進められていたが，その場合は周波数成分の重なり等の影響で同定に失敗するケースが多かった．それに対し，本手法を用いた場合では各パートごとに音源同定を行うことで，同定の対象となる単音数が多くなり，そのため個々の単音に含まれる変形等の影響を軽減することが可能となる．

2.2 フレーズ形成の意義

従来の研究では，パートに相当する対象を抽出する手法は少なく，また抽出を行うものであっても，その手がかりとしては単音の遷移など局所的なものにとどまっていた [8]．これは，大局的な情報を与えようとする場合に，計算量が膨大になるからである．例えば，類似する旋律の繰り返しを利用する場合，その出現位置や繰り返し部分の長さなどの組合せは曲の長さの 2 乗のオーダーで増加し，それらすべてについて類似度の計算を行う必要が生じる．

本手法では，パート形成に先だってフレーズの形成

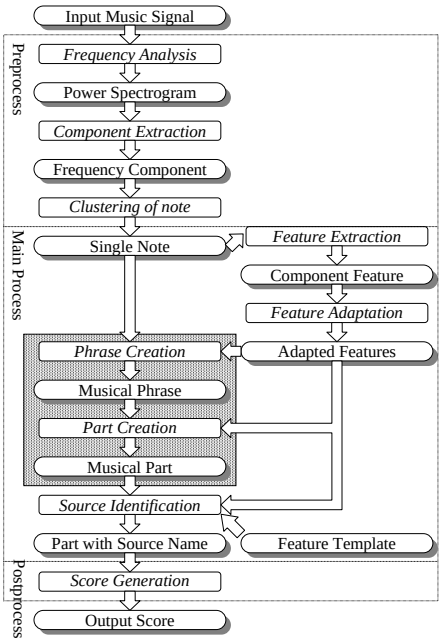


図 1　提案する処理システムの構成
Fig. 1　Block diagram of proposed process.

を行うことで，この問題を回避する．局所的な手がかりによってフレーズを形成し，その後大局的な手がかりを用いてパートを形成する．これにより，局所的および大局的な手がかりをともに利用でき，さらに，大局的な手がかりを与える対象として単音に比べ粒度の大きなフレーズを用いることで，計算量の増加を抑えることができる．

3. システム概要

処理システムの構成を図 1 に示す．本論文では，このうち網掛けを施した部分について議論する．

入力として，主に室内楽アンサンブル演奏がモノラル録音された音響信号を用いる．同時発音数は最大 3 程度を想定している．

音響信号は，フィルタバンクを用いた手法により時間周波数解析され，パワースペクトログラムとなる．続いてパワー値のピークを時間方向に追跡することにより，周波数成分が形成される．周波数成分は，高調波関係や立上りのずれに着目した処理によって単音に相当する集合へクラスタリングされる．これらの処理については，柏野らの手法 [6] を用いた．

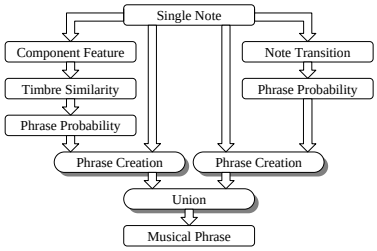


図 2　フレーズ形成処理
Fig. 2　Block diagram of phrase creation process.

単音を構成する周波数成分（周波数成分クラスタ）からは，特徴量が抽出され，複数の単音が同時に存在する場合に特徴量が変化することによる影響を軽減するため，特徴量の性質に応じて特徴量の値を再計算する処理が施される．この処理と，後述の音源同定処理は文献 [3] の手法に基づいて行った．

また，近接する単音が同一の音源に由来する確率を求め，フレーズを形成する（3.1 節）．フレーズ間について，大局的な手がかりを用いてパートを形成する（3.2 節）．

パートに含まれる単音を入力として，あらかじめ用意した特徴量テンプレートと比較することで音源同定処理を行い，各パートがどの音源に由来するものであるかを求め，最後に楽譜の形式で出力する．

本節では，図の主処理部の処理のうち，フレーズ形成（図 1: Phrase Creation）およびパート形成（図 1: Part Creation）およびについて詳述する．

3.1 フレーズ形成

前処理部による単音の抽出に続いて，フレーズの形成が行われる．処理の流れを図 2 に示す．

フレーズの形成では，時間的に隣接する単音間の局所的な手がかりを用い，それらが同一音源に属する確率（フレーズ形成確率）を計算する．

実際には，隣接する単音の音高と音色に着目して 2 通りの確率を計算してフレーズを形成し，それらを統合することで最終的なフレーズを得る．次項以降で，その具体的な手法について述べる．

3.1.1 音高遷移確率

まず，隣接する単音の音高から，それらが同一の音源に由来する確率を求める（図 2: Note Transition 〜 Phrase Probability）．この処理においては，あらかじめ統計的に求められた音高の遷移確率を用いる．

本論文では，遷移確率として，単音列データを解析することにより，楽曲における単音の遷移パターンに

関する以下の情報を抽出することにより求めた値を用いた [4]．このデータは，日本のポピュラー音楽を主とする 311 曲の MIDI データである．

a） 単音の遷移幅の出現頻度

まず，時間的に隣接する 2 つの単音について，その遷移における音高の差を遷移幅として抽出し，その幅ごとの出現頻度を調べた．

ここでいう単音の遷移幅とは，隣接する単音の音高の差を，半音を 1 として数えたもので，1 オクターブのずれが遷移幅 12 に相当する．

b） 単音の遷移パターンの出現頻度

単音の音高を調性によって正規化した上で，遷移の前後の音名の組を遷移パターンとして抽出し，パターンごとの出現頻度を調べた．

c） 確率値の統合

最後に，得られた上記 2 つの確率値を統合することで全体の遷移確率を得，音高遷移確率とする．本論文では上述の 2 つの値が遷移が同一フレーズに属する根拠を与えるものと解釈し，Dempster-Shafer の確率理論に基づいて統合を行う．

3.1.2 音色類似度

単音を形成する周波数成分からは，立上りの強さやパワー値の時間的変動などの物理的な特徴量が抽出されている．これを隣接する単音の間で比較することにより，それらの音色の類似度を計算する（図 2: Component Feature 〜 Phrase Probability）．計算される類似度は，文献 [3] によるものと同様である．各特徴量の分布をあらかじめ多量の単音データを元に準備し，比較する特徴量の差を分布内での距離に正規化した上で計算される．

3.1.3 フレーズ形成

算出された音高遷移確率と音色類似度を元に，それぞれをフレーズ形成確率と解釈する．なお，音色類似度については，類似度が 0 から 1 の範囲の値をとり，その値をそのままフレーズ形成確率とした．

続いて，以下の手順にしたがって単音の集合からフレーズを形成する．この手法は，文献 [8] において単音連繋確率ネットワークを構成するアルゴリズムと同様である．

（1） 処理は，開始時刻の早い順に単音を順次追加する形で進められる．

（2） ある単音が追加されると，近接する既存の単音との間のフレーズ形成確率が計算される．本論文では，時間的に近い単音を対象に処理を行った．閾値として，実験的に 2 秒を用いた．

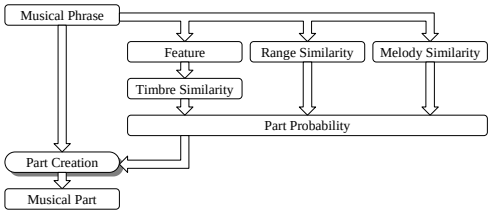


図 3 パート形成処理
Fig. 3 Block diagram of part creation process.

（3） 計算の結果，最も大きな値となった単音とフレーズを形成する．

（4） フレーズ形成の対象となった単音が既にそれより開始時刻の遅い別の単音とフレーズを形成していた場合には，古いほうのフレーズは切断される．

フレーズは単音遷移確率に基づくものと，音色類似度に基づくものの 2 通りを形成し，それらに共通して含まれる単音の組をパート抽出において用いるものとする．これは，フレーズ形成の段階で誤りが多く含まれると，後段のパート形成処理における処理で計算されるパート形成確率の値が不正確になるため，フレーズ形成の段階では結果の適合率を重視するためである．

また，同様の理由により，フレーズ形成確率に閾値を設定し，フレーズ形成確率がこれを越えたものに対してのみフレーズの形成を行う．閾値としては，実験的に 0.1 を用いた．

3.2 パート形成

形成されたフレーズを元に，それらを連結することでパートを形成する（図 3）．パートの形成においては，単音からフレーズを形成した時と同様，隣接したフレーズが同一のパートに属する確率を導出する．

この確率は複数の根拠に基づいて求められ，最終的に統合される．統合は，Dempster-Shafer の確率理論に基づいて進められる．ここでは，3 つの Dempster 確率: $m(C)$ = 隣接するフレーズが同一パートに属する確率，$m(\overline{C})$ = 隣接するフレーズが同一パートに属さない確率，$m(C,\overline{C})$ = 判断ができない確率，を用いた．

2 通りの根拠から 2 組の $m(C)$，$m(\overline{C})$，$m(C,\overline{C})$ が獲得されたとき，それらを Dempster の結合規則に基づいて統合することで新たな確率値を得る．

$$m(A_k) = \frac{\sum_{A_{1i}\cap A_{2j}=A_k} m_1(A_{1i})m_2(A_{2j})}{1-\sum_{A_{1i}\cap A_{2j}=\emptyset} m_1(A_{1i})m_2(A_{2j})}$$

ここで，A_k は $\{C\}$，$\{\overline{C}\}$，$\{C,\overline{C}\}$ のいずれかであり，2 組の $m(A_i)$ の値を m_1，m_2 と表した．

本手法では，時間的に隣接するフレーズについて後述する 3 つの根拠（音色類似度，音域類似度，旋律類似度）から確率値を計算し，それらを上記法則に基づいて統合する．

統合後の $m(C)$，$m(\overline{C})$，$m(C,\overline{C})$ から，対象となるフレーズが同一パートに属する事象の上限確率および，下限確率を計算し，それらの中間値を，最終的なパート形成確率として得る．

パートの形成は，以下のような手順で行われる．

（1） 各フレーズについて，近接するフレーズとの間のパート形成確率を計算する．本論文では，時間的に前のフレーズの終了時刻（最後の単音の終了時刻）と，時間的に後のフレーズの開始時刻（最初の単音の開始時刻）の差が 0.5 秒未満であるものを近接するフレーズと判断した．

（2） フレーズの組のうち，パート形成確率が最大の値となったものについて，それらを連結する．

（3） こうしてできたものを新しいフレーズと考え，その近接するフレーズとのパート形成確率を再計算する．

（4） これらの処理を，近接するフレーズが存在しなくなるまで繰り返す．

次項から，パート形成に用いる根拠についてその概要と確率値の導出方法を議論する．

3.2.1 音色類似度

隣接する 2 つのフレーズに対し，それらに含まれる単音の間の特徴量類似度を求める（図 3: Feature 〜 Timbre Similarity）．各フレーズには複数の単音が含まれるが，それらのうちいくつかは周波数成分の重なりにより特徴量が変形されている．そこで，周波数成分の重なりのある単音については類似度を計算する対象から除外する．なお，この周波数成分の重なりの有無は，単音抽出の段階で判定される [6]．

残った単音に関し，その全ての組合せについて類似度を計算する．類似度の計算方法は，フレーズ形成の際に用いたものと同様である（3.1 節）．

その平均値をフレーズ間の音色類似度とする．

3.2.2 音域類似度

同一のパートに含まれる単音は，ある一定の範囲の音域を推移することが多い．そこで，フレーズの音高分布を比較することでパートを形成する手がかりを求める．これを音域類似度と呼ぶことにする（図 3: Range Similarity）．

各フレーズに含まれる単音の音高に着目し，そのヒストグラムを作成する．音域類似度は，算出されたヒストグラムを比較することで得られる．以下にその導出の手順を示す．

まず，比較の対象となる 2 つのフレーズに対して，各フレーズを構成する単音の音高分布の平均と標準偏差を計算し，μ_1，μ_2 および σ_1，σ_2 とする．

この値を用いて，2 つのフレーズの音高分布を正規分布 $N(\mu_1,\sigma_1)$，$N(\mu_2,\sigma_2)$ で近似する．続いてこれらの分布の相関を求めるために，確率密度関数の積の積分 $\int_{-\infty}^{\infty} f_1(z)f_2(z)dz$ を計算し，$\mu_1=\mu_2$ となった時に 1 となるよう正規化した値

$$\exp\left(-\left(\mu_1-\mu_2\right)^2/2(\sigma_1^2+\sigma_2^2)\right)$$

を用いる．本論文では実験的にこの値を確率値 p とし，$m(C)=p$，$m(\overline{C})=1-p$，$m(C,\overline{C})=0$ とする．

また，フレーズを構成する単音の個数が少ない場合には，この値の信頼性が低いと考えられるため，ここで述べた確率の計算は行わず，$m(C)=0$，$m(\overline{C})=0$，$m(C,\overline{C})=1$ とする．なお，本論文では，この個数の下限として実験的に 3 を用いた．

3.2.3 旋律類似度

一般に，メロディに相当するパートと伴奏に相当するパートでは，旋律のパターンに大きな差がある．そこで，フレーズの旋律を比較することで，同一のパートに属するか否かを判断する根拠が得られる．これを旋律類似度と呼ぶことにする（図 3: Melody Similarity）．

比較の際には，同一パート内でも曲の推移に従って旋律に変動が加えられる可能性があることと，正確な音長の抽出が困難であることを考慮する．ここでは音程の推移を上行・下行等に，また音長の推移も単純化した上で，隣接するフレーズ間の相関を求める．

単純化は，文献 [5] のものを参考にし，隣接する 2 つの単音について，その音高の差から，U（上行），D（下行），E（同音）とする．また，音長からも，定数 r を用いて，L（後の音が前の音の r 倍以上の長さ），S（前の音が後の音の r 倍以上の長さ），E（それ以外）とする．なお，本論文では実験的に $r=1.25$ とした（図 4）．

隣接するフレーズのそれぞれから記号列を得た上で，その完全に一致する部分の最大長を求める．この長さの元の記号列の長さに対する比をもって，旋律類似度とする．分母となる記号列は，2 つの記号列のうち記

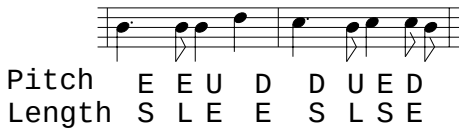


図 4 旋律の単純化
Fig. 4 Simplification of melody.

号列長が長くない方とした．

この値を p として，$m(C) = p$，$m(\overline{C}) = 0$，$m(C,\overline{C}) = 1 - p$ とする．ここで，$m(\overline{C}) = 1 - p$ とせずに $m(C,\overline{C}) = 1 - p$ としたのは，隣接するフレーズの旋律が異なることは，それらが異なるパートに属する根拠を与えるわけではないからである．

また，音域類似度と同様に，フレーズを構成する単音の個数が少ない場合にはここで述べた確率の計算は行わず，$m(C) = 0$，$m(\overline{C}) = 0$，$m(C,\overline{C}) = 1$ とする．なお，本論文では，この個数の下限として実験的に 10 を用いた．

4. 評 価 実 験

4.1 フレーズ形成評価

本論文にて提案したフレーズ形成処理について，その評価を行うための予備的な実験として，実験用のデータに対して処理を行い，処理結果の正当性を調べた．

実験では，実験曲「蛍の光」の室内楽アンサンブル演奏を対象に処理を進めた．この曲は，2 つの旋律パートと，1 つの伴奏パートからなる．用いた楽器は，ピアノ，フルート，ヴァイオリン，クラリネット，トランペットである．楽器の音域の関係から，伴奏パートにはピアノのみを用いた．3 つのパートの単音数は旋律パートが高音部から順にそれぞれ 59，61，伴奏パートが 122 である．

なお，実験においては，周波数成分抽出処理の誤差の影響を取り除くために，各単音の時刻と音高をあらかじめ人手で与えた．これにより，単音の抽出が完全に行われた場合のフレーズ形成の精度を評価する．

精度の計算は，出力のフレーズに含まれる単音の遷移（隣接する単音の組）について，それらの単音が入力の楽譜において同一音源に由来しているものをカウントすることで行った．評価は，適合率，および実験曲に含まれる遷移の総数を分母とした再現率によって行った．旋律パートの楽器を前述の楽器（ピアノを除く）から選んで組み合わせ，それらの結果の平均を求めて実験結果とした．

結果をパート数ごとに集計したものを図 5 に示す．

4.2 パート形成評価

本手法におけるパート形成の結果の正当性と，パート形成に用いた根拠の効果を見るために実験を行った．

パート形成処理における 3 つの根拠の効果を見るために，それぞれの根拠を用いた場合と用いなかった場合について処理を行い，その精度を比較した．実験で用いたデータは，フレーズ形成評価において全ての処理を適用した場合の出力を用い，処理精度の算出方法は，フレーズ形成精度のものと同様である．

結果をパート数ごとに集計したものを図 6 に示す．

5. 考　　察

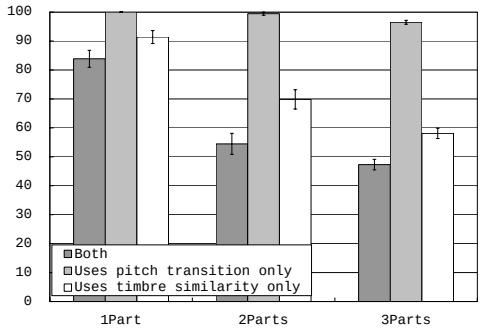
評価実験の結果について考察を加える．

5.1 フレーズ形成に関する考察

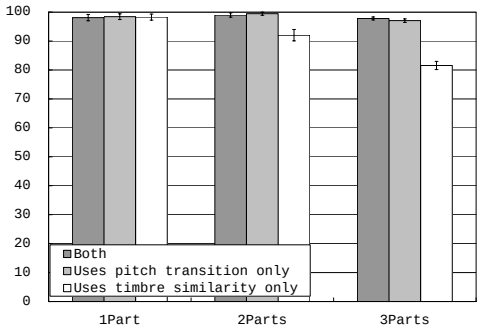
フレーズ形成精度の評価結果（図 5）から，3 パートの場合に約 98% の適合率でフレーズ形成が行えていることがわかる．一方，再現率は約 47% となっているが，これは本手法のフレーズ形成の目的が局所的な構造を抽出することであって，大局的なパートを得ることにはない点，および後段のパート形成処理により低下した再現率をカバーできる一方で，適合率の低下はカバーできない点から，フレーズ形成処理ではできるだけ適合率の高い結果を得る必要があることを考えれば妥当なものである．

フレーズ形成に用いた根拠別に見ると，音高遷移確率のみを用いてフレーズの形成を行った場合には，再現率が約 96% ，適合率が約 97% と，高い値が得られた．一方，音色類似度のみを用いた場合には特に 3 パートの場合で再現率で約 58%，適合率で約 82% と，音高遷移確率を用いた場合と比較して低い値となっている．パート数が 1 のものと 2 のもの，2 のものと 3 のものを比較した場合に，後者に較べて前者での再現率の低下が大きいことから，パートの数そのものよりも，複数になることによる影響が大きいと考えられる．また，音高遷移確率のみを用いた場合には，再現率と適合率の低下がともに少ないことから，多くの単音が存在すること自体よりも，複数音の重なりの発生の方が，音色類似度の値に与える影響が大きいことが分かる．ここでいう影響は，複数音の重なりによって，それぞれの音から得られる特徴量が変化することが主であると予想される．

本論文で用いた実験曲では，曲中 2 箇所において 2

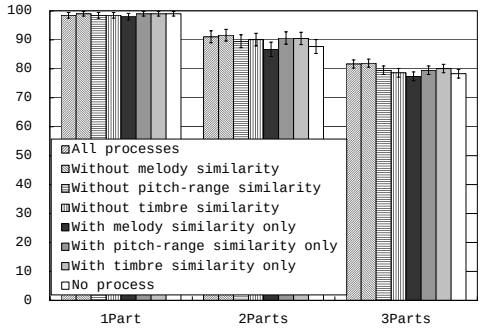


横軸がパート数，縦軸が再現率を表す（単位: %）．また，誤差範囲は 95% 信頼区間を表す

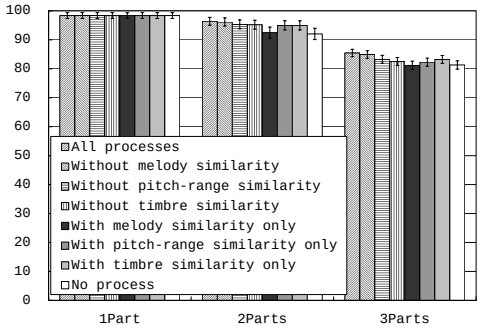


横軸がパート数，縦軸が適合率を表す（単位: %）．また，誤差範囲は 95% 信頼区間を表す

図 5　フレーズ形成の結果
Fig. 5　Result of phrase creation.



横軸がパート数，縦軸が再現率を表す（単位: %）．また，誤差範囲は 95% 信頼区間を表す



横軸がパート数，縦軸が適合率を表す（単位: %）．また，誤差範囲は 95% 信頼区間を表す

図 6　パート形成の結果
Fig. 6　Result of part creation.

つのパートが同一音高の単音を共有するか，音高の高低がパートの間で入れ替わっている．音高遷移のみに注目した場合，これらの箇所でフレーズ形成誤りが発生したが，音色類似度を導入することでこの箇所で形成されることはなくなり，結果として適合率が僅かではあるが向上した．このことは，音色類似度の利用の有効性を示唆するものであると言える．

一方で，今回用いた楽曲がこの 2 箇所を除いて，各パートの音域が比較的分離したものであったために，音高遷移処理のみで十分な精度が得られたと考えられる．今後，パートの音域がより近接した曲を用いて処理の検証を行いたい．

5.2　パート形成に関する考察

パート形成実験においては，3 つのパートからなる曲を処理の対象とした場合でも，最大で再現率で 90%，適合率で 94% 程度の再現率および適合率を得た（フルート，トランペット，ピアノの場合等）．このことから，本論文で提案した処理によって，音源の種類を同定することなくストリーム構造の抽出が可能であることが明らかになった．

その一方，今回実験に用いた曲は，前項でも述べた通り比較的パート間の音域が離れたものであった．そのため，音高のみでパート形成が可能となった可能性もある．今後，より多くの種類の曲を対象に実験を行う必要がある．

処理の構成による差をみると（図 6），有意な差があるとは言えないが，今後の検証の足がかりとして，結果から読み取れる傾向について考察を加えてみる．

各パート数の場合において，音域類似度，音色類似度の導入によって再現率，適合率ともに向上する傾向が見られる．一方，旋律類似度を用いた場合には，再現率，適合率ともに向上が見られなかった．この原因

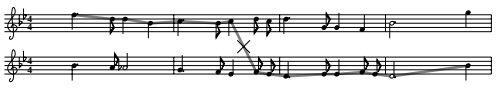


図 7　パート抽出ミスの例
Fig. 7　Example of part misextraction.

として，以下の点が考えられる．今回用いた手法では，旋律類似度を計算する対象のパートがほぼ一致している場合に根拠として意味をもち，共通部分があるだけの場合にはその寄与が少ない．また，パート形成の際にその一部分の単音を抽出できなかった場合にも類似度は低い値となる．今回用いた処理では音長，音高の変動，抽出誤りについては考慮したが，今後は単音抽出あるいはフレーズ形成の誤りに対応した処理が必要となるだろう．

また，旋律類似度の計算において，記号列パターンが完全一致している箇所のみを対象として類似度の計算を行った．そのため，パターンの途中で単音の抽出ミスがあった場合など，十分な効果が得られないケースが考えられる．この点に関して，今後より柔軟な類似度判定手法の検討を課題としたい．

パート形成の誤りの傾向をみると，図 7 のように複数のパートをまたがる形で形成されるものが多かった．この場合，音源同定の対象となる単音が，複数の音源に由来するものとなるため，音源同定処理に失敗する可能性が高くなり，また主要な音源の同定に成功した場合でも，その他の部分については音源同定に失敗することになる．

このような誤ったパートの形成は，図 7 の "×" で示した単音間の連結を防ぐことで回避することができる．十分長いフレーズを連結対象としてパート形成確率を計算する場合には，音色類似度等によりこれが可能であるが，そうでない場合には音色類似度の計算に用いる単音数が少なくなり，値に明確な差が出にくいことから，回避は困難である．

本論文で提案した処理においては，この問題は，パート形成においてパート形成確率の閾値を高く設定することで軽減できるが，一方で再現率の低下を招き，出力されるパートが短く切断されたものとなる．本質的な解決のためには，一旦パートが形成された後に各連結に関してその正当性を再検討するなどの処理を追加することが必要となる．この点に関しては，今後の検討課題とする．

また，本実験では，処理を用いた場合と用いなかった場合の間で結果に有意な差が出なかった．さらに別の面から評価実験を行い，処理の有効性について検討する必要がある．

6. おわりに

本論文では，音楽音響信号を対象とした自動採譜処理において，従来手法において問題となっていた時間方向に大局的な情報の利用の例として，知覚的階層構造に着目してパート抽出を行う手法を提案した．実験システムに対する評価実験の結果，3 パートの楽曲に対して再現率 82%，適合率 85% 程度でパートの抽出に成功した．

本論文では，複数の観点，あるいは特徴量に着目した上で処理を行った．しかしながら，人間が音楽を耳にする場合には今回用いた以外の手掛かりも利用していることは明らかである．例えば，各パートのテンポや音量等をも用いていると考えられる．今後はこの点を考慮した処理も検討したい．

一方，実験で用いた楽曲の数が少なく，部分的な処理の有無による処理精度の変化に有意な差が出ないなど，結果が予備的なものにとどまった．また，評価実験は，隣接する単音が同一パートに属するかを検証するにとどまった．今後は，各パートの音域等のような曲の種類や曲数そのものについてより多くを用い，また大局的な妥当性など複数の面からの検証を進めていく必要がある．

謝辞

本論文は，文部省科学研究費補助金（課題番号 09-07629）による研究成果の一部である．また，音響信号データ NTTMSA-P1 の使用許可をいただいた，NTT コミュニケーション科学基礎研究所に感謝する．

文　献

[1] A. S. Bregman. Auditory scene analysis. *MIT Press*, 1990.
[2] 三輪多恵子, 田所嘉昭, 斎藤努. くし形フィルタを利用した採譜のための異楽器音中のピッ チ推定. 電子情報通信学会論文誌, Vol. J81-DII, No. 9, pp. 1965–1974, 9 1998.
[3] 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, Vol. J83-DII, No. 4, pp. 1073–1081, 2000.
[4] 木下智義, 村岡秀哉, 田中英彦. 単音の遷移に注目した単音認識処理. 日本音響学会誌, Vol. 54, No. 2, pp. 190–198, March 1998.
[5] 園田智也, 後藤真孝, 村岡洋一. WWW 上での歌声による曲検索システム. 電子情報通信学会論文誌, Vol. J82-DII, No. 4, pp. 721–731, 1999.
[6] 柏野邦夫, 中臺一博, 木下智義, 田中英彦. 音楽情景分析

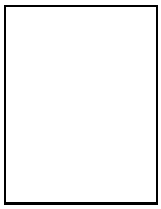
の処理モデル OPTIMA における単音の認識．電子情報通信学会論文誌, Vol. J79-DII, No. 11, pp. 1751–1761, 11 1996.

[7] 柏野邦夫, 木下智義, 中臺一博, 田中英彦．音楽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, 11 1996.

[8] 柏野邦夫, 村瀬洋．単音連繋確率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, 11 1998.

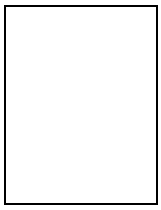
[9] 柏野邦夫, 村瀬洋．適応型混合テンプレートを用いた音源同定 — 音楽演奏 への応用 —．電子情報通信学会論文誌, Vol. J81-DII, No. 7, pp. 1510–1517, 7 1998.

（平成 xx 年 xx 月 xx 日受付）



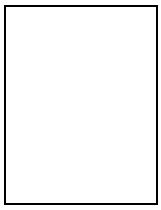
木下　智義　（正員）

1995 年東京大学工学部電子情報工学科卒業．2000 年同大大学院情報工学専攻博士課程修了．現在有限会社ネットコンパス勤務．情報処理学会，人工知能学会，各会員．



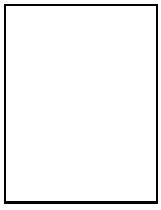
半田　伊吹

1997 年東京大学工学部電子工学科卒業．現在同大学大学院工学系研究科電気工学専攻博士課程在学中．音楽情報科学の研究に従事．情報処理学会，電気学会各会員．



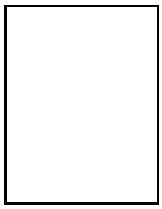
武藤　　誠

1999 年東京大学工学部電子情報工学科卒業．2001 年同大大学院工学系研究科電気工学専攻修士課程終了．現在 NTT サイバースペース研究所勤務．情報処理学会会員．



坂井　修一　（正員）

1958 年生．1981 年東京大学理学部情報科学科卒業．1986 年同大学院情報工学専門課程修了．工学博士．同年，電子技術総合研究所入所．1991 年 4 月より 1 年間米国 MIT 招聘研究員．1993 年 3 月より RWC 超並列アーキテクチャ研究室室長．1996 年 10 月筑波大学助教授（電子・情報工学系）．1998 年 4 月東京大学助教授（工学系研究科）．2001 年 4 月東京大学教授（情報理工学系研究科）．計算機システムとその応用の研究に従事．情報処理学会論文賞（1990 年度），日本 IBM 科学賞（1991 年），市村学術賞（1995 年），ICCD Outstanding Paper Award（1995 年）など受賞．情報処理学会，人工知能学会、IEEE，ACM，各会員．



田中　英彦　（正員）

1943 年生．1965 年東京大学工学部電子工学科卒業．1970 年同大学院博士課程修了．工学博士．同年東京大学工学部講師．1971 年助教授．1978 年〜1979 年ニューヨーク市立大学客員教授，1987 年教授，現在に至る．計算機アーキテクチャ，並列処理，人工知能，自然言語処理，分散処理，メディア処理等に興味を持っている．「非ノイマンコンピュータ」，「情報通信システム」著．「計算機アーキテクチャ」，「VLSI コンピュータ I，II」，「ソフトウェア指向アーキテクチャ」共著．情報処理学会，人工知能学会，日本ソフトウェア科学会，IEEE，ACM，各会員．