音色類似度と定位類似度の統合による自動採譜

櫻庭 洋平† 奥乃 博†
† 京都大学 情報学研究科 知能情報学専攻

1. はじめに

複数楽器による重奏の自動採譜には，周波数成分から単音を形成する『同時的グルーピング』と単音をパートごとに分類する『継時的グルーピング』が必要となる.『同時的グルーピング』には，オクターブの関係にある単音形成の曖昧性解消が必要であるが，従来は明示的に扱われていない[1) 2)].『継時的グルーピング』については，音色と定位を用いて実験をおこなった例はない.

調波構造は『同時的グルーピング』の大きな手がかりである．実際には，複数音が同時に発音している場合は調波構造に基づく『同時的グルーピング』には曖昧性が生じる．それは，単独の単音が発音している場合と複数の単音が発音している場合（ある単音の基本周波数が，別の単音の基本周波数の整数倍の関係にある場合）の調波構造が極めて類似するためである．この曖昧性を明示的に解消し，精度向上を試みた研究は筆者の知る限り報告されていない．本稿では，周波数成分の重なり確率を求める．それに基づき単音を形成する．

また，我々はこれまでに定位情報を用いた『継時的グルーピング』を研究してきた[4)]．サンプラーで合成した音響信号を入力とし，強度差，位相差から得られる定位情報が信頼できるものとして，定位が近接する単音を同一パートとする．しかし，実際に演奏された音楽音響信号から得られる定位は変動するため，定位が必ずしも信頼できるとは限らない.

本稿では，同一パートと判断する特徴として音色と定位を用いる．音色や定位が類似する単音は同じパートであると考える．それぞれの類似度を**音色類似度**，**定位類似度**と呼ぶ．2 つの類似度を Dempster-Shafer の確率理論を用いて統合し，より高性能な『継時的グルーピング』を目指す.

2. 同時的グルーピング

2.1 周波数成分の重なり確率

入力のステレオ音楽音響信号から，周波数成分を抽出する．周波数成分は短時間フーリエ変換で得られるパワーの極大値（ピーク）を時間方向に接続して得られる．各ピークは強度差，位相差から求まる定位を持っている[4)].

周波数成分内のピークの定位の変動を利用して，周波数成分の重なり確率を求める．重なりのない周波数成分は定位の変動が小さく，重なりのある周波数成分は定位の変動が大きくなる．定位の変動を，周波数成分内のピークの定位の標準偏差（σ）と最小二乗近似直線の傾き（g）で表現できると仮定し，それぞれから得られる重なり確率を Dempster-Shafer の確率理論を用いて統合し，最終的な重なり確率とする.

S，$\overline{S}$がそれぞれ周波数成分に重なりがあるという信念，ないという信念を表すとすると，3 つの基本確率は

$m(S)$: 周波数成分に重なりがある基本確率
$m(\overline{S})$: 周波数成分に重なりがない基本確率
$m(S,\overline{S})$: 判断できない基本確率

となる．σ，g それぞれから得られる基本確率を m_σ，m_g とし，次のように定義する.

$$m_x(S) = r \times \{1 - Sig(x, T_x)\} \quad x = \sigma, g$$
$$m_x(\overline{S}) = r \times Sig(x, T_x)$$
$$m_x(S,\overline{S}) = 1 - r$$

$Sig(x,T)$ は $Sig(T,T)$=0.5，$Sig(0,T)$=0.99 となる $Sigmoid$ 関数である．ここでは T_σ=5 度，T_g=9.6 度/sec とした．$r = \frac{\text{定位が求められたフレーム数}-1}{\text{周波数成分のフレーム数}-1}$である．$m_\sigma$，$m_g$を以下の Dempster の結合規則に基づいて統合し，新たな確率値を得る.

$$m(A_k) = \frac{\sum_{A_i \cap A_j = A_k} m_\sigma(A_i) m_g(A_j)}{1 - \sum_{A_i \cap A_j = \phi} m_\sigma(A_i) m_g(A_j)}$$

こうして得られた全体の $m(S)$，$m(\overline{S})$，$m(S,\overline{S})$ に対して，上界確率 $P^* = m(S) + m(S,\overline{S})$ と，下界確率 $P_* = m(S)$ を求め，その中間値を統合後の重なり確率とする.

2.2 単音仮説生成・評価

周波数成分から，調波構造に基づいて，考えられる単音の組み合わせ（単音仮説）を生成する．このとき，調波構造のほとんどが重なる関係にある単音仮説も生成される.

単音仮説の尤度 L は周波数成分 f_i（$0 \le i < N$）の，重なり確率 $P(f_i)$ や定位 $Pan(f_i)$（周波数成分内の全ピークの定位の平均）をどれだけ満たすかを表し，以下の 3 ステップで求める．N は周波数成分の個数である.

（1）f_iに重なりがある場合．$Score(f_i) = P(f_i)$
（2）f_iに重なりがない場合.
$Score(f_i) = Sig(\|Pan(f_i) - Pan(n)\|, T_\sigma)$
ただし，n は f_iの定位に最も近い定位を持つ単音である.
（3）$Score(f_i)$ の合計を L とする.

3. 継時的グルーピング

3.1 パート形成確率

P，$\overline{P}$がそれぞれ，二つの単音が同じパートであるという信念，異なるパートであるという信念を表すとすると，3 つの Dempster 確率は

$m(P)$:二つの単音が同じパートである基本確率
$m(\overline{P})$:二つの単音が異なるパートである基本確率
$m(P,\overline{P})$:判断できない確率

となる．音色類似度，定位類似度を表す基本確率をそれぞれ m_t，m_pとする．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などに関する 23 次元の特徴量[4)]で表現さ

Automatic musical transcription by integrating direction proximity and timbre similarity
by Yohei Sakuraba, and Hiroshi G. Okuno (KyotoUniv.)

れている．システムは知識として，2 つの単音の特徴量の差を Support Vector Machine を用いて学習した音色類似度モデルを持っている．学習には NTTMSA-P1 の 5 楽器，約 1100 サンプルを用いた．

このとき，m_t，m_pを以下の式で定義する．

$m_t(P) = \int_{-\infty}^{SVMscore} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$

$m_t(\overline{P}) = \int_{SVMscore}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$

$m_t(P, \overline{P}) = 0$

$m_p(P) = r_1 \times r_2 \times Sig(\|Pan(n_1) - Pan(n_2)\|, T_p)$

$m_p(\overline{P}) = r_1 \times r_2 \times \{1 - Sig(\|Pan(n_1) - Pan(n_2)\|, T_p)\}$

$m_p(P, \overline{P}) = 1 - r_1 \times r_2$

Pan は定位を表す．r_1，r_2はそれぞれ単音 n_1，n_2の定位の信頼度であり，単音の周波数成分の重なりがない確率の和である．T_pは次式で，中央で小さく，外側で大きくなるように設計されている．

$T_p = \frac{c}{l} sin^{-1}(\frac{l}{c} sin(Pan(n)) + \frac{4.0}{f_s})) - Pan(n);$

c は音速，l はマイク間の距離，f_sはサンプリングレートである．Dempster-Shafer の確率理論に基づき，上界確率，下界確率からパート形成確率を求める．

3.2 パート形成処理

パート形成処理は，単音を時間順に追加して行われる．ある時点のパート数を C，$Part_i$（$0 \leq i < C$）内の単音の個数を c_i，単音を n_{ij}（$0 \leq j < c_i$）とする．$Part_i$と入力単音 n とのパート形成確率 $P(Part_i, n)$ はパート内のすべての単音と入力単音とのパート形成確率の平均である．

パート形成処理の流れは以下の 4 ステップとなる．

（1）全パートと，入力単音 n との $P(Part_i, n)$ を求める．

（2）n は $P(Part_i, n)$ が最大となるパートに追加する．

（3）最大値が 0.5 未満であれば，新しいパートを形成する．

（4）以上の処理を全ての入力単音に対して繰り返す．

4. システム実装・実験・評価

音楽音響信号を入力し，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，構築した．入力データは無響室で録音した「パッヘルベルのカノン」の 4 重奏である．楽器の位置は，左から，violin，flute，trumpet，piano で固定とした．角度は 20 度間隔，40 度間隔，60 度間隔の 3 パターンとした．ただし，システムにはそのような情報は与えていない．システムは各楽器の位置が不変であることのみを知っている．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）で録音した．

評価は，再現率 R と適合率 P から求まる F 値$\frac{2 \times R \times P}{R+P}$で行う.『同時的グルーピング』では，音名が正しく，発音時刻の誤差が 32 分音符以内であれば正解とする.『継時的グルーピング』では，音名，発音時刻に加え，パートが正しく判定されれば正解とする．音長により性能に差が出ると考えられるので，8 分音符までしか表れない小節と 16 分音符を含む小節に分けて実験を行う．32 分音符を含む小節は実験対象外とした.『同時的グルーピング』,『継時的グルーピング』の結果をそれぞれ表 1・2，表 3・4 に示す．

『同時的グルーピング』では，本稿で提案する周波数成分の重なり確率を導入することで，すべての楽器配置パターンで F 値が向上した.『継時的グルーピング』では，音色類似度と定位類似度を統合することで，一方のみを用いる場合より F 値が向上した．適合率と再現率でみると,『同時的グルーピング』では，再現率が大きく向上しているが，適合率が低下している．同時発音数を与えていないため，数多くの単音からなる仮説も出力されることがあるためだと考えられる．また,『継時的グルーピング』のみの性能を表 5 に示す．両類似度を統合することで，8 分音符では 90%の性能である．

表 1　同時的グルーピングの結果（8 分音符）

間隔	調波構造のみ	調波構造 + 重なり確率
20 度	0.63（0.46, 0.98）	0.78（0.76, 0.79）
40 度	0.63（0.46, 1.00）	0.76（0.80, 0.72）
60 度	0.60（0.43, 0.97）	0.61（0.66, 0.57）

表 2　同時的グルーピングの結果（16 分音符）

間隔	調波構造のみ	調波構造 + 重なり確率
20 度	0.44（0.31, 0.79）	0.58（0.51, 0.66）
40 度	0.44（0.31, 0.79）	0.57（0.53, 0.62）
60 度	0.45（0.31, 0.82）	0.49（0.41, 0.61）

表 3　継時的グルーピングの結果（8 分音符）

間隔	音色類似度のみ	定位類似度のみ	両類似度を統合
20 度	0.65 (0.64, 0.65)	0.68 (0.68, 0.69)	0.71 (0.70, 0.71)
40 度	0.60 (0.63, 0.56)	0.66 (0.71, 0.63)	0.69 (0.73, 0.65)
60 度	0.52 (0.56, 0.48)	0.52 (0.57, 0.49)	0.56 (0.61, 0.53)

表 4　継時的グルーピングの結果（16 分音符）

間隔	音色類似度のみ	定位類似度のみ	両類似度を統合
20 度	0.40 (0.35, 0.46)	0.50 (0.45, 0.57)	0.51 (0.45, 0.58)
40 度	0.35 (0.32, 0.39)	0.49 (0.46, 0.53)	0.50 (0.46, 0.54)
60 度	0.34 (0.28, 0.43)	0.34 (0.28, 0.44)	0.38 (0.31, 0.50)

※表中の数値は F 値（再現率，適合率）である．

表 5　継時的グルーピングのみの結果

間隔	音色類似度のみ	定位類似度のみ	両類似度を統合
20 度	0.84，0.69	0.89，0.88	0.92，0.88
40 度	0.79，0.60	0.89，0.87	0.91，0.87
60 度	0.85，0.67	0.86，0.68	0.92，0.76

※左が 8 分音符の結果，右が 16 分音符の結果である．

5. おわりに

本稿では，自動採譜を『同時的グルーピング』,『継時的グルーピング』の二つにわけ，それぞれの曖昧性解消について述べた．無響室で収録した音楽音響信号を用いて実験し，性能の向上を確認した．今後は，音高遷移確率などを導入し，より高性能な自動採譜システムの実現をしていく．

謝辞　本研究は，学振科研費および，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，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．

参考文献

1) 柏野ほか：“音源情景分析の処理モデル OPTIMA における単音の認識”，信学論，J79-DII，11，pp.1751-1761，1996.
2) 柏野ほか：“音源情景分析の処理モデル OPTIMA における和音の認識”，信学論，J79-DII，11，pp.1762-1770，1996.
3) 三輪・守田：“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，J84-DII，7，pp.1251-1260，2001.
4) 櫻庭ほか：“定位情報と音色情報を用いた複数楽器音の認識”，情処研報，2002-MUS-46，pp.9-16，2002.