

多重奏を対象とした音源同定：混合音テンプレートを用いた音の重なりに頑健な特徴量への重み付け及び音楽的文脈の利用

北原 鉄朗^{†a)} 後藤 真孝^{††} 駒谷 和範[†] 尾形 哲也[†]
奥乃 博[†]

Instrument Identification in Polyphonic Music: Feature Weighting Based on Mixed-Sound Template and Use of Musical Context

Tetsuro KITAHARA^{†a)}, Masataka GOTO^{††}, Kazunori KOMATANI[†], Tetsuya OGATA[†], and Hiroshi G. OKUNO[†]

あらまし 本論文では、多重奏に対する音源同定において不可避な課題である「音の重なりによる特徴変動」について新たな解決法を提案する。多重奏では複数の楽器が同時に発音するため、各々の周波数成分が重なって干渉し、音響の特徴が変動する。本研究では、混合音から抽出した学習データに対して、各特徴量のクラス内分散・クラス間分散比を求めることで、周波数成分の重なりの影響の大きさを定量的に評価する。そして、線形判別分析を用いることで、これを最小化するように特徴量を重み付けした新たな特徴量軸を生成する。これにより、周波数成分の重なりの影響をできるだけ小さくした特徴空間が得られる。更に、音楽的文脈を利用することで音源同定の更なる高精度化を図る。実楽器音データベースから作成した二重奏～四重奏の音響信号を用いた実験により、二重奏では 50.9%から 84.1%へ、三重奏では 46.1%から 77.6%へ、四重奏では 43.1%から 72.3%へ認識率の改善を得、本手法の有効性を確認した。

キーワード 音源同定, 楽器音, 混合音テンプレート, 音楽情景分析, MPEG-7

1. ま え が き

デジタル音楽配信や大容量携帯音楽プレーヤーの普及により、個人が大量の音楽音響信号を入手・利用できるようになったのに対し、目的の楽曲を簡単に探し出す手段は十分ではない。タイトルやアーティスト名などの書誌情報である程度絞り込むことはできても、音楽は内容を一覧することが困難なため、最終的には一曲一曲試聴して探すことが強いられている。この現状を打破するには、計算機を用いた音楽検索（音楽情報検索）の実現が不可欠である。中でも、計算機を用いて音楽の内容を統一的な枠組みで記述する技術は、効果的な音楽情報検索システム実現のための鍵技術と

期待されている。実際、音楽を含むマルチメディアコンテンツを記述する統一的枠組みとして MPEG-7 [1] が制定され、研究開発が進められている。

我々は、記述すべき音楽的内容として楽器名が重要であると考えている。どの楽器で演奏されたかという情報は、「ピアノソナタ」「弦楽四重奏」などの分類があるように、特にクラシック音楽では楽曲を特徴づける重要な要素である。そのため、例えば「弦楽四重奏の曲が聴きたい」といった楽器に基づく検索には一定の需要があると考えられる。また、楽器名は、聴取者の感性や主観に依存しないという観点からも記述すべき音楽的内容としてふさわしい。更に、「フルートが弾き始めるところから聴く」といった楽器名をキーとした頭出しに利用することもできる。

本論文では、上記で述べた音楽音響信号に対する楽器名情報の記述で中心的な処理となる、混合音からの楽器名の同定（音源同定）を扱う。音源同定は、パターン認識の一問題と考えることができ、音声情報処理における話者同定に対応する問題設定である。しかし、

[†] 京都大学大学院情報学研究所, 京都市 Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto-shi, 606-8501 Japan

^{††} 産業技術総合研究所, つくば市 National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba-shi, 305-8568 Japan

a) E-mail: kitahara@kuis.kyoto-u.ac.jp

通常複数の楽器が同時に演奏されるという点で難しく、従来研究において、単一音を対象としたもの [2] ~ [11] は 10 ~ 30 種類程度の楽器を扱っているのに対し、多重奏を対象としたもの [12] ~ [16] は 3 ~ 5 種類程度の楽器による二重奏 ~ 三重奏を扱うにとどまっている。多重奏に対する音源同定が難しい主たる原因は、周波数成分が重複することにより、各楽器の音響的特徴を正確に抽出できないことにある。もしも、音源分離技術により、混合音から各楽器音が完全に分離できるのであれば、混合音に対する音源同定は、単一音の音源同定の問題に帰着する。しかし、実際には、周波数成分の重複が頻繁に発生するため、混合音をひずみなく分離するのは困難である。

本論文では、この周波数成分重複の問題を、同定に用いる特徴量の重み付けにより解決する。特徴量が周波数成分重複の影響を多分に受けていれば低い重みを与え、逆にあまり影響を受けていなければ高い重みを与える。このような特徴量の重み付けが実現できれば、音の重なりに対する頑健性を改善することができる。本研究では、各特徴量がどの程度周波数成分重複の影響を受けているかを、その特徴量のクラス内分散・クラス間分散比として定量的に評価することで、上記の重み付けを線形判別分析による次元圧縮に帰着させる。線形判別分析は、クラス内分散・クラス間分散比を最小化するように特徴量を加重混合した新たな特徴量軸を作り出す手法で、混合音から抽出した学習データに対して線形判別分析を適用することで、周波数成分重複の影響を最小化した部分空間が得られる。本論文では、この方法を DAMS 法 (discriminant analysis with mixed sounds) と呼ぶ。従来、同問題に対して適応型混合テンプレート [13]、周波数成分の重なり適応 [15]、ミッシングフィーチャー理論 [16] などが提案されてきたが、音の重なりに対する頑健度に基づいて特徴量に重み付けをするという試みはなかった。

更に本論文では、音楽的文脈 (前後関係) に基づいて音源同定の更なる性能向上を図る。旋律の連続性を考慮して事後確率を計算することで、フルートによる一連の旋律の中でクラリネットが一音だけ出現するといった、音楽的に不自然な誤りを削減する。

2. 問題設定

本研究では、音源同定を単音^(注1)ごとに行うものとする。今、与えられた音楽音響信号に K 個の単音 n_1, \dots, n_K が含まれているとし、対象

楽器を $\omega_1, \dots, \omega_M$ とする。このとき解くべき問題は、 n_1, \dots, n_K の各々 (n_k と表記) に対して $\operatorname{argmax}_{\omega_i \in \{\omega_1, \dots, \omega_M\}} p(\omega_i | n_k)$ を求めることである。ここで、 $p(\omega_i | n_k)$ は単音 n_k が楽器 ω_i による演奏である確率を表す。実際には、単音 n_k から複数の音響的特徴を抽出し、この音響的特徴からなる多次元ベクトル x_k を考え、 $\operatorname{argmax}_{\omega_i} p(\omega_i | x_k)$ を求める。特徴量の詳細は 5. で述べるが、周波数重心 (各周波数成分のパワー値を重みとした周波数の重み付き重心) やパワー包絡 (各周波数成分のパワーの累積により計算) の近似直線の傾きなど、我々が独自に設計した 43 個を用いる。 $p(\omega_i | x_k)$ は事後確率と呼ばれており、ベイズの定理により以下のように展開できる：

$$p(\omega_i | x_k) = \frac{p(x_k | \omega_i) p(\omega_i)}{\sum_{j=1}^M p(x_k | \omega_j) p(\omega_j)}$$

ここで、 $p(x_k | \omega_i)$ は楽器 ω_i の確率密度関数、 $p(\omega_i)$ は事前確率である。確率密度関数は、あらかじめ用意された多数の学習用特徴ベクトルに基づき、これらが正規分布などの確率分布に従って生起すると仮定して計算する。この学習用特徴ベクトルは、各楽器の様々な音響信号から、楽器同定時と同様の手順により特徴抽出して得られたもので、楽器名ラベルが付与されている。このような楽器名ラベルが付与された学習用の特徴ベクトルを数多く集めたデータベースを、本研究では特徴量テンプレート [15] と呼ぶ^(注2)。

3. 音の重なり頑健な音源同定

本章では、音の重なりによる特徴変動に頑健な音源同定手法の設計について論ずる。まず、個々の楽器音の同定の際に、同時に発音する他の音の影響を抑えるため、調波構造を抽出することが有用であることを述べる。次に、調波構造抽出だけでは不十分であることを述べ、その解決策として、音の重なりに対する頑健度に応じて特徴量に重み付けをする方法を提案する。

3.1 調波構造モデルの利用

音声認識や話者認識などの研究では、メル周波数ケ

(注1): 本論文では、「単音」と「単一音」を異なる意味で用いる。前者は、処理の単位となる音で、楽譜上の一音符に相当する。通常、一つの単音は調波構造を一つだけ持つ。一方、後者は、単音が同時に一つしか鳴っていない音を指す。

(注2): 本研究では、テンプレートマッチングのような事例ベースのアプローチで用いられる個々の事例ではなく、統計的に処理される多数の特徴ベクトルから構成される集合全体をテンプレートと呼ぶ。

プストラム係数 (MFCC) をはじめとする、スペクトル包絡に関する特徴量がよく用いられる。これは、観測されたスペクトルの大まかな形をとらえるものであり、複数の単音が同時に鳴ったときのスペクトルから個々の単音に起因する部分に着目することは困難である。そこで、打楽器を除く多くの楽器音が調波構造をもつことを利用^(注3)し、音楽音響信号から個々の単音の調波構造を抽出して、そこから特徴量を抽出する方法が広く用いられてきた [12], [15], [19], [20]。

本研究においても、各単音の調波構造を抽出し、そこから特徴量を抽出する。ここでは、単音 n_k の調波構造モデル $\mathcal{H}(n_k)$ を以下のように表す。

$$\mathcal{H}(n_k) = \{(F_i(t), A_i(t)) \mid i=1,2,\dots,h, 0 \leq t \leq T\}$$

$F_i(t)$, $A_i(t)$ は時刻 t における i 次倍音の周波数と振幅を表し、組 $(F_i(t), A_i(t))$ を周波数成分と呼ぶ。 $F_i(t)$ は $F_1(t)$ のおよそ i 倍であるが、実際には誤差を生じる。時刻 t は発音開始 (オンセット) 時刻を基準 ($t=0$) とし、周波数は $F_1(t)$ の時間方向の中央値が 1 となるように除算によって正規化した相対周波数で表す。 h は最大倍音次数、 T は音長である。このように楽器音を調波構造としてモデル化すると、複数の楽器音が同時に発音したときの影響は周波数成分の重なりだけに限定して考えることができる。実際の楽器音は非調波成分を含み、それが楽器音を特徴づける要素となっている [7], [21]。しかし、混合音からの非調波成分の抽出は困難であり、信頼性に欠けるため、本研究では非調波成分は扱わないものとする。

3.2 音の重なりに関する頑健な特徴量への重み付け

前節で述べたように、同時発音する他の音の影響は、周波数成分の重なりだけに限定して考えることができる。仮に、単音 n_k の周波数成分が単音 n_j の周波数成分と一つも重なりがないとすると、同時発音による互いの影響は無視できるほどに小さくなる。しかし、実際には周波数成分の重なりは頻繁に発生する。例えば、音高が C4 (約 262 Hz) の単音と G4 (約 394 Hz) の単音が同時に発音するとき、C4 の $3m$ 次倍音と G4 の $2m$ 次倍音 (m は任意の自然数) は重なりを起こす。一般的に、協和する音程 (音高差) の 2 単音は多くの周波数成分が重なりを起こすため、重なって変化した周波数成分から特徴抽出することによる特徴変動の問題は、多重奏の音源同定においては深刻な問題である。

この特徴変動の問題を解決する上で有力となるアプローチは、特徴変動の度合いに基づいた特徴量の重み付

けである。変動の大きな特徴量には小さな重みを、変動の小さな特徴量には大きな重みを与えることができれば、頑健な音源同定を実現できるはずである。以下、この特徴量の重み付けによる頑健性の向上について論ずる。

3.2.1 関連研究

従来研究において、上記と類似したアイデアとして Missing Feature Theory [16]、周波数成分重なり適応処理 [15] が検討されてきた。

- Missing Feature Theory [16] は信頼できない特徴量をマスクすることで、事後確率の計算でこのような特徴量を使わないようにする方法である。特徴量をマスクすることはその特徴量に重み 0 を与えることに相当するため、上述した特徴量の重み付けの一実現法といえる。この方法は、マスクすべき特徴量が分かっていたら極めて効果的であることが音声認識の研究などで確かめられている [22]。しかし、その自動推定法は発展途上であり、現状ではローカルスペクトルなどの限られた種類の特徴量しか自動推定に利用できない。

- 周波数成分重なり適応処理 [15] は、各特徴量を、周波数成分の重なりによる変動の仕方でも「加算特徴量」「優先特徴量」「崩壊特徴量」に分類し、周波数成分の重なりがあったときに、この分類に従って特徴量を再計算したり無効化したりするものである。特徴量を無効化することは上記と同様に特徴量の重み付けの一実現法といえるが、特徴量の分類を手で行うため、多くの特徴量を利用するのは困難である。特徴量の重要度を計算する処理も導入されているが、単一音に基づいて計算されており、音の重なりによる特徴変動の程度に応じた重み設定という観点には至っていない。

- これらの他に多重奏の音源同定を扱った研究として適応型混合テンプレート法 [13] がある。ただし、特徴抽出を行わずに波形テンプレートのマッチングを行う手法であり、音の重なりによる特徴変動の程度に応じた特徴量の重み付けという立場からの研究ではない。

3.2.2 音の重なりに対する頑健度の定量化

上記で述べた特徴量の重み付けを実現する上での課題は、各特徴量が周波数成分の重なりによって受ける影響の大きさの定量化である。従来の研究では、い

(注3): 本研究では、打楽器は対象外とする。これは、ピアノやバイオリンなどの音 (楽音) が明確な音高や調波構造をもつものに対し、打楽器音 (非楽音) は音高や調波構造が明確に現れないために、打楽器音を楽音と同一のシステムで扱うのが困難だからである。打楽器に関しては文献 [17], [18] などにおいて一定の成果が上げられており、将来的にはこれらとの統合を進める予定である。

れも学習データを単一音から作成していたため、周波数成分の重なりの影響を学習データから評価することは不可能であった。

本論文で提案する DAMS 法では、混合音から抽出した特徴ベクトルを学習に用いることで、周波数成分の重なりの影響度をクラス内分散・クラス間分散比として定量化する。混合音から抽出した特徴ベクトルを学習データとした場合、学習データは既に音の重なりによる影響を受けている。そのため、音の重なりによって変化しやすい特徴量であれば、各楽器の学習データにおけるその特徴量の分布は大きな分散をもつはずであるし、音の重なりによる変化が大きい特徴量であれば、分布の分散も大きくなるはずである。すなわち、クラス内分散・クラス間分散比を最小化する重みを見つけることにより、音の重なり影響度をできるだけ小さくする特徴空間を得ることができる。クラス内分散・クラス間分散を最小化する方法は、線形判別分析として広く知られているので、目的の特徴量の重み付けは、混合音から得られた学習データに対して線形判別分析を行えばよい。

3.2.3 既存楽曲の音程混合頻度をを用いた混合音テンプレート作成

混合音から特徴ベクトルを抽出する処理の流れを図 1 に示す。学習用の混合音には、そこに含まれる各単音の楽器名・音高・発音時刻がラベルづけられている。そのラベルに基づき、各単音の調波構造を抽出する。次に、その調波構造に対して特徴抽出を行う。これをあらかじめ用意した様々な混合音に対して行い蓄積して特徴量テンプレート（学習用特徴ベクトルの集合）を作成する。本研究では、このようにして混合音から作成した特徴量テンプレートを混合音テンプレートと呼ぶ。

混合音テンプレート作成において重要となる課題は、学習用の混合音としてどのようなものを用意するかである。楽器は様々な音高や音の強さで発音されるため、たとえ楽器の種類を限定したとしても、複数楽器が同時に発音するときの音の組合せ方（混合パターンと呼ぶ）はばく大となり、すべてのパターンを網羅的に収集するのは現実的には不可能である^(注4)。そのため、楽器の音色を学習するのに必要十分な量の混合パターンを設計しなければならない。

この課題を解決する上でかぎとなる音楽的性質は、実際の楽曲で出現する混合パターンには、その音高差（音程）に一定の傾向があることである。後でも述べ

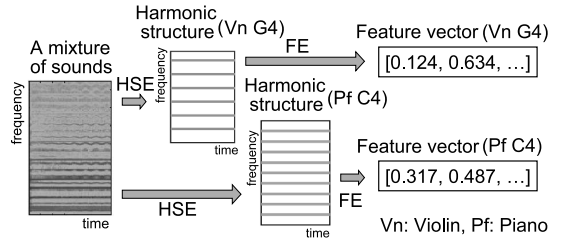


図 1 混合音からの特徴ベクトルの抽出（混合音テンプレート作成）HSE：調波構造抽出，FE：特徴抽出
Fig. 1 Construction of a mixed-sound template.

るように、短 2 度など一部の音程は不協和音を生むため、長 3 度や完全 5 度などに比べて使用頻度は低く、この傾向は楽曲が違って大きく異なるものではない。そのため、既存の楽曲の楽譜から混合音を作成することで、実際の楽曲における使用頻度を反映した混合パターンを得ることができる。そこでこれを混合音テンプレート作成に用いる。我々は、次の二つの理由により、同定対象とテンプレート作成用楽曲が異なっても、性能向上に貢献できると考えている。

● 混合パターンの音程の偏り

実際の楽曲で用いられる混合パターンには音程に偏りがあり、例えば C4, C#4, D4 の 3 音の同時発音は極度の不協和音を生むため、特別な効果をねらう場合以外はほとんど用いられない。このような音程の利用頻度は、人間の音に対する快・不快の印象と密接に関連するため、楽曲ごとに大きくは変わらない。

● 調波構造抽出による同時発音の影響の限定化

本研究では単音ごとに調波構造を抽出するので、複数楽器の同時発音の影響は、「どの倍音成分にどの程度他の単音の成分が重なってくるか」に限定される。概して「どの倍音成分に」の部分は当該単音と他の単音との音程で決まり、「どの程度」の部分は他の単音が何の楽器かに依存して決まるが、後者は楽器が変わっても大幅に変わるわけではない。そのため、混合パターンの音程が実際の音楽を反映していれば、楽器などの他の組合せが多少欠けていても十分利用できると考えられる。

(注4)：例えば、本論文の実験で用いる楽器音データベースには 5 楽器 2651 音が収録されているので、同時発音数を 3 に限定しても、その混合音は $2651C_3 = \text{約 } 31 \text{ 億通り}$ である。これは、1 秒 1 個の速さで学習しても約 98 年かかる量である。

4. 音楽的文脈の利用

本章では、音楽的文脈の利用法について述べる。これは、例えばフルートによる一連の旋律の中でクラリネットが一音だけ出現するといった、音楽的に不自然な誤りの削減を目的とする。本研究における音楽的文脈利用法のキーアイデアは、単音 n_k の事後確率を計算する際に、事前確率に前後の単音の事後確率を反映させることである（図 2）。この処理は、「 n_k の前後の単音が楽器 ω_i であれば、 n_k も ω_i である可能性が高い」という考え方に基づいている。この処理を実現するには、次の課題を解決する必要がある。

[課題] 前後の単音から n_k と同じ楽器の単音の抽出

図 2 の処理を実現するには、文脈として用いる単音が同定対象音 n_k と同じ楽器である必要がある。しか

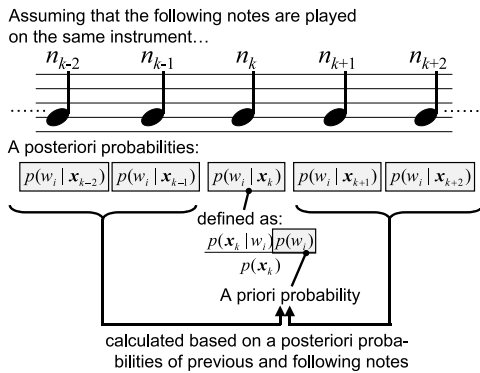


図 2 音楽的文脈利用のキーアイデア
(単音 n_k の事後確率を計算するために、 n_k の前後の単音の事後確率を利用する)

Fig. 2 Basic idea for using musical context.

し、実際の楽曲にはさまざまな楽器の音が存在するため、楽器 ω_i の単音を抽出しなければならない。

我々はこの課題を次のようにして解決する。

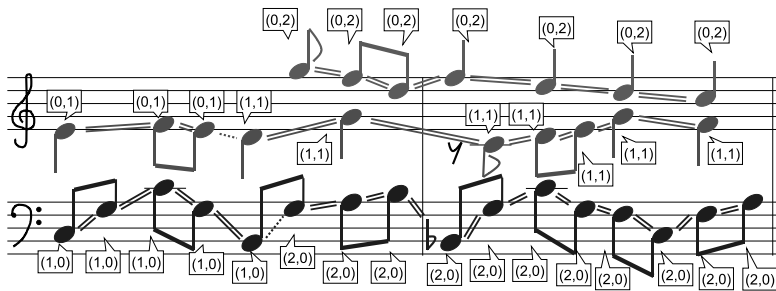
[解決策] 複数パート知覚における並行移動性の利用
複数の旋律が同時に進行する演奏を聴いたとき、人間は、旋律が交差せずに進んでいくように知覚することが知られている [23]。例えば周波数が下降する旋律と上昇する旋律がある時刻で交差するとき、周波数が接近して交差せずに再び離れていくように聞こえる。そのため、作曲家は、特別な効果をねらう場合以外ではできるだけ交差しないように旋律を設計すると考えられる。そこで、2 単音が同楽器によるもの（同パート）かの判断を次のように行う。単音 n_k の発音中に、 n_k よりも高音域で同時に発音する単音数の最大値を $s_h(n_k)$ 、低音域で同時に発音する単音数の最大値を $s_l(n_k)$ とすると、 $s_h(n_k) = s_h(n_j)$ かつ $s_l(n_k) = s_l(n_j)$ のときに 2 単音 n_k と n_j は同パートと判断する（図 3）。柏野ら [14] は単音連鎖ネットワークを形成するために役割同一性の概念を導入し、彼らは「最高音」と「最低音」の二つの音楽的役割がそれぞれ主旋律とベースラインに対応するとした。我々の方法は、この音楽的役割の一拡張法と考えることもできる。

以下、処理の詳細を述べる。

[第 1 パス] 文脈を考慮しない事後確率の仮計算

n_1, \dots, n_K の各々に対して $p(\omega_i | x_k)$ を計算する。前後の単音の事後確率から計算される事前確率 $p(\omega_i)$ は現段階では定まらないので、定数とみなして計算する。

[第 2 パス] 文脈を考慮した事後確率の再計算



— A pair of notes that are correctly judged to be played on the same instrument
..... A pair of notes that are not judged to be played on the same instrument although they actually are

図 3 隣り合う単音が同じ楽器によるかどうかの判定の例
(組“(a, b)”は $s_h(n_k) = a$ and $s_l(n_k) = b$ を表す)

Fig. 3 An example of judgment of whether adjacent notes are played on the same instrument or not.

n_1, \dots, n_K の各々 (n_k と表記) に対して, 時刻の早い方から (n_1, n_2, \dots の順で) 次の処理を行う.

(1) 前後の単音から n_k と同じ楽器の単音の抽出
 単音 n_k の前後の単音から $\{n_j \mid s_h(n_k) = s_h(n_j) \cap s_1(n_k) = s_1(n_j)\}$ を満たす単音を抽出する. この処理は単音 n_k に時間的に近いものから順に行い, c 個単音が抽出された時点で終了する. 以下, 抽出された単音の集合を \mathcal{N} で表す.

(2) 事前確率の計算

単音 n_k の事前確率をステップ (1) で抽出された単音の事後確率 (音楽的文脈) に基づいて計算する. 今, 音楽的文脈から計算される事前確率を $p_1(\omega_i)$, 他の手掛りから計算される事前確率を $p_2(\omega_i)$ とする. このとき, 求めるべき事前確率 $p(\omega_i)$ を

$$p(\omega_i) = \lambda p_1(\omega_i) + (1 - \lambda) p_2(\omega_i)$$

と定義する. ここで, λ は音楽的文脈の信頼度であり, ステップ (1) で抽出された単音がすべて楽器 ω_i であるときに単音 n_k も楽器 ω_i である確率として統計的分析によって得ることもできるが, ここでは簡単化のため定数とし, $1 - (1/2)^c$ を用いた. これは, 文脈に用いた単音が多いほどそこから得た情報は信頼できるというヒューリスティクスに基づくものである. 音楽的文脈に基づく事前確率 $p_1(\omega_i)$ は, ステップ (1) で求めた単音の, 第 1 パスで求めた事後確率の積を正規化したものとし,

$$p_1(\omega_i) = \frac{1}{\alpha} \prod_{n_j \in \mathcal{N}} p(\omega_i | x_j)$$

と定義する. ここで, x_j は単音 n_j から抽出された特徴ベクトル, α は各楽器の事前確率の総和を 1 とするための正規化係数で, $\alpha = \sum_{\omega_i} \prod_{n_j} p(\omega_i | x_j)$ である. また, 他の手掛りに基づく事前確率 $p_2(\omega_i)$ は, 単に $p_2(\omega_i) = 1/M$ とした.

(3) 事後確率の更新

ステップ (2) で求めた事前確率を用いて, 事後確率を再計算する.

5. 音源同定の処理の詳細

上記で議論した手法を用いた音源同定の処理の詳細について述べる. 全体の処理の流れを図 4 に示す. まず, 入力された音楽音響信号からスペクトログラムを求め, ピーク抽出を行う. 次に, あらかじめ既存手法 ([12], [20], [24] など) を使って推定した各単音の音高・

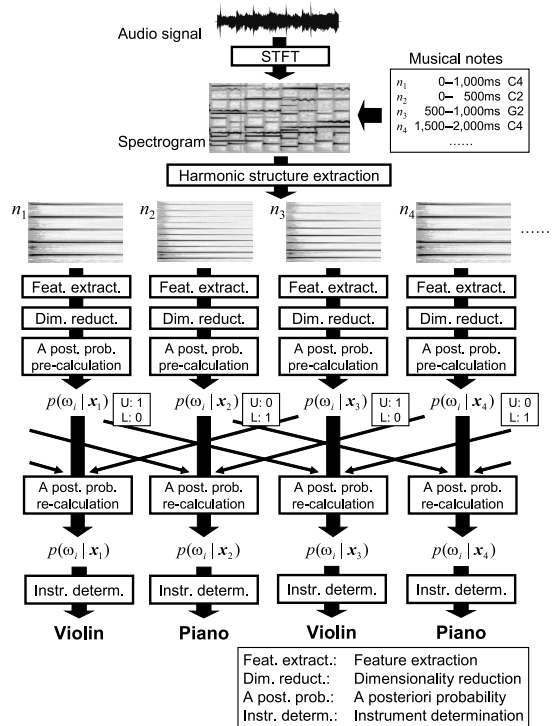


図 4 音源同定手法の処理の流れ (n_1, \dots, n_4 は単音を表し, 図右上の表は, 付与される各単音の発音時刻・発音終了時刻・音高を表す. “Harmonic structure extraction” の下の図は各単音の調波構造を表す)
 Fig. 4 Flow of our instrument identification method.

発音時刻・音長に基づき, 各単音の調波構造を抽出する. ただし, 本論文の実験では, 音源同定部のみの性能を評価するため, 各単音の音高・発音時刻・音長は正解を与えるものとする. その後, 特徴抽出, 次元圧縮, 事後確率算出, 楽器名決定の順序で処理を進める.

5.1 周波数解析

入力された音楽音響信号 (本論文の実験では, サンプリング周波数 44.1 kHz, 16 ビットリニア量子化, モノラルの音響信号を用いる) に対して, 短時間フーリエ変換を用いてスペクトログラムを求める. 窓関数にはハミング窓を使用し, 窓幅は 8192 点, シフト長は 10 ms とする. その後, フレームごとにパワースペクトルのピークを抽出する.

5.2 調波構造の抽出

あらかじめ推定 (ただし本実験では正解を付与) した各単音の音高, 発音時刻, 音長に基づいて, 単音 (n_k で表す) に対応する調波構造 $\mathcal{H}(n_k)$ を抽出する. 音高に対応する周波数 (A4 なら 440 Hz, 平均律で算

出)の近傍 200 cent 以内に存在する最もパワーの大きいピークを基本周波数成分とみなし、その周波数の整数倍のピークを 10 次倍音まで抽出する。ただし、整数倍のピークの抽出においては 5%までの周波数の誤差を許容する。その後、2. で述べたように、基本周波数成分の周波数 $F_1(t)$ の時間軸方向の中央値が 1 となるように周波数を除算によって正規化する。

5.3 音長の調節

特徴量テンプレート作成時に用いた単音と同定対象の単音との間で音長を合わせることで、特徴量の音長依存性を回避する。例えば、発音直後にパワーが急激に大きくなった後に急激に小さくなり、その後徐々に小さくなってやがて消音する音に対して、パワー包絡の近似直線の傾きを計算する場合、短い音と長い音とでは計算結果が大きく異なることが予想される。ここでの目的は、このような音の長短によって特徴量が変化する現象を避けることにある。音長を合わせる最も単純な方法は、最も短い音長に合わせることであり、しかし、音長を短くするとスペクトルが安定しにくくなるため、音長が長いままの状態に比べて同定が難しくなる。そこで本論文では、いくつかの音長パターン (300 ms, 450 ms, 600 ms. 左からパターン I, II, III と名づける) に対して特徴量テンプレートを作成し、同定対象音より短い範囲で最長のパターンに合わせる。

5.4 特徴抽出

各単音 n_k の調波構造 $\mathcal{H}(n_k)$ から、楽器名の同定に効果的と考えられる特徴量を抽出する。特徴量は、我々が以前単一音用に提案したもの [7] から混合音からの抽出が困難と思われるものを除いた最大 43 個 (音長が 600 ms の場合、音長が 300 ms, 450 ms の場合各 31 個, 37 個) である。使用する特徴量を表 1 に列挙する。

5.5 DAMS 法 (次元圧縮)

DAMS 法を適用して、周波数成分の重なりの影響を最小化した部分空間を得る。線形判別分析の計算を頑健に行うには、特徴空間が無相関であることが望ましいので、DAMS 法の前に主成分分析を適用して無相関な部分空間を得る。主成分分析後の次元数は累積寄与率が 99% になるように決め、学習データによって多少変化するが 20 次元前後である。この部分空間に対して DAMS 法を適用して (楽器数 - 1) 次元の更なる部分空間を得る。

5.6 事後確率算出

各単音 n_k に対して、事後確率 $p(\omega_i|x_k)$ を求める。

表 1 使用する特徴量の一覧
Table 1 Features used.

(1) スペクトルの時間平均に関する特徴	
1	周波数重心 (各高調波成分のパワー値を重みとした周波数の重み付き重心)
2	全高調波成分のパワー値の合計に対する基音成分のパワー値の割合
3, 10	高調波成分のパワー値の合計に対する基音から i 次までの高調波成分のパワー値の割合 ($i = 2, 3, \dots, 9$)
11	奇数次の高調波成分 (基音含む) と偶数次の高調波成分とのパワー値の比
12, 20	持続時間が、最長の高調波成分のその、 $p\%$ 以上ある高調波成分の個数 (持続時間はパワーがあるしきい値を超えている時間にて定義) ($p = 10, 20, \dots, 90$)
(2) パワーの時間変化に関する特徴	
21	パワー包絡の線形最小二乗法による近似直線の傾き
22, 30	発音開始直後 t 秒間のパワー包絡の微分係数の中央値* ($t = 0.15, 0.20, 0.25, \dots, 0.55$ [s])
31, 39	最大パワー値と、発音開始から t 秒後のときのパワー値の比*
(3) 各種変調の振幅と振動数	
40, 41	振幅変調の振幅と振動数
42, 43	周波数変調の振幅と振動数

*音長パターン I, II では、 t が音長を超える場合の特徴量は省略される。

これは 4. で述べた計算法により、音楽的文脈を考慮しながら処理を進める。なお、事後確率の計算における確率密度関数 $p(x_k|\omega_i)$ は、我々が以前提案した F0 依存多次元正規分布 [7] を用いる。F0 依存多次元正規分布とは、楽器の音色が音高に従って変化する現象を陽に表現するモデルで、多次元正規分布のパラメータである平均ベクトルの各要素が、基本周波数の関数として定義されたものである。

5.7 楽器名決定

最後に、事後確率 $p(\omega_i|x_k)$ が最大となる楽器名 ω_i を同定結果として決定する。

6. 評価実験

提案手法の有効性を示すため、以下の実験を行った。

6.1 多重奏の音響信号の作成

本実験では、各単音の正確な音高・発音時刻・音長のラベルを利用できるようにするため、実演奏ではなく計算機上で楽器音データベース (実楽器を発音可能な音域全体にわたって半音ごとに個別に発音したものを収録したもの) の音響信号を切り貼りして作成した音響信号を用いる。

まず、スタンダード MIDI ファイル (SMF) として

表 2 主要 4 パートからの二重奏～四重奏の抜粋方法

Table 2 Extraction of parts from the main four ones.

	二重奏 A	二重奏 B	三重奏	四重奏
第 1 パート	✓	—	✓	✓
第 2 パート	—	✓	✓	✓
第 3 パート	—	—	—	✓
第 4 パート	✓	✓	✓	✓

表 3 使用した楽器音データの内訳

Table 3 Details of musical sound data used.

楽器番号	楽器名 (楽器記号)	音域	バリエーション	強さ	データ数*
01	ピアノ (PF)	A0-C8	1, 2, 3	強・中・弱	792
09	クラシックギター (CG)	E2-E5	1, 2, 3	強・中・弱	702
15	バイオリン (VN)	G3-E7	1, 2, 3	強・中・弱	576
31	クラリネット (CL)	D3-F6	1, 2, 3	強・中・弱	360
33	フルート (FL)	C4-C7	1, 2	強・中・弱	221

奏法はノーマル奏法(記号:NO)のみを使用

(ただし、クラシックギターはアポヤンド/指弾き奏法(AF))

* 無音検出による自動切出しによって切り出された単音の個数

「RWC 研究用音楽データベース:クラシック」(RWC-MDB-C-2001)[25] 収録の楽曲番号 13, 16, 17 のものを用意した。これらは四重奏～五重奏程度の小規模編成のクラシック曲である。本実験では二重奏～四重奏を扱うので、各曲の SMF (各冒頭 100 小節まで) から主要な 4 パートを抜粋した後、表 2 に従って更にパートを抜粋して二重奏～四重奏の SMF を得た。また、テンプレート作成用に単旋律のものも用意した。

次に、これらの SMF に従って単音ごとの音響信号を切り貼りするプログラムを作成し、これを用いて二重奏～四重奏の音響信号を得た。このプログラムでは、SMF に記述された単音の各々に対し、楽器音データベースから適切な音響信号を探し出し、SMF に記述された音長に合わせて音響信号を切った後貼り付けていく。SMF 中の音長がもともとなる音響信号よりも長い場合には、生成される音響信号は意図する (SMF に記述された) 音長よりも短い音となるが、我々が使用したデータベースでは十分な長さで収録されており (およそ 3 秒程度、ピアノやクラシックギターといった減衰系楽器では 1 秒未満のものもあるが自然減衰により十分にパワーが小さくなった後の打ち切りである)、このような事態はほとんど発生しなかった。もともとなる音響信号には、「RWC 研究用音楽データベース:楽器音」(RWC-MDB-I-2001)[25] から抜粋した表 3 のものを用いた。RWC-MDB-I-2001 は、上述の条件を満たす単一音で構成された楽器音データベースである。学習用データと同定用データに同じ音響信号

表 4 各パートの楽器の候補

Table 4 Instrument candidates for each part.

第 1 パート	PF, VN, FL
第 2 パート	PF, CG, VN, CL
第 3 パート	PF, CG
第 4 パート	PF, CG

表 5 混合音テンプレートにおける単音数。Leave-one-out 法。音長パターン I のもののみ。音長パターン II 及び III については紙面の制約から省略するが、おおむね II が I の 1/2, III が I の 1/3～1/4 程度である。

Table 5 Number of notes in mixed-sound templates.

	S+D	S+D+T	サブセット
No. 13	PF 31,334	83,491	24,784
同定用	CG 23,446	56,184	10,718
	VN 14,760	47,087	9,804
	CL 7,332	20,031	4,888
	FL 4,581	16,732	3,043
No. 16	PF 26,738	71,203	21,104
同定用	CG 19,760	46,924	8,893
	VN 12,342	39,461	8,230
	CL 5,916	16,043	3,944
	FL 3,970	14,287	2,632
No. 17	PF 23,836	63,932	18,880
同定用	CG 17,618	42,552	8,053
	VN 11,706	36,984	7,806
	CL 5,928	16,208	3,952
	FL 3,613	13,059	2,407

S+D: 単旋律+二重奏

S+D+T: 単旋律+二重奏+三重奏

サブセット: 実験 3 で用いる楽器の組合せを減らしたのも

が用いられることを防ぐため、表 3 の音響信号のうち、同定用データにバリエーション番号「1」、強度「中」のもの (011PFNOM, 091CGAFM, 151VNNOM, 311CLNOM, 331FLNOM) を、学習用データにはそれ以外のすべての音響信号を用いた。各パートの楽器の候補を音域の制約から表 4 のように制限し、この制限内で総当たりとした。その結果、例えば四重奏では 48 通りの組合せとなった。

6.2 実験 1: Leave-one-out 法

最初の実験は、楽曲単位での Leave-one-out 法により行った。すなわち、3 曲のうち 1 曲を除いたデータを学習に用いて除外された 1 曲のデータを同定するという処理を、3 曲すべてが同定に用いられるまで繰り返した。混合音テンプレートは、単旋律と二重奏の音響信号から作成したもの (「単旋律+二重奏」) とそれに三重奏を加えたもの (「単旋律+二重奏+三重奏」) を用い、比較のため、表 3 の音響信号をそのまま学習データとした場合 (「単一音のみ」) も行った。作成した混合音テンプレート中の単音数を表 5 に示す。この表から、例えば、楽曲番号 13 を同定するとき用いる

表 6 実験 1 の結果 (認識率) (Leave-one-out 法利用 . 太字は 75%以上を表す)
Table 6 Results of Experiment 1.

テンプレート 音高依存 文脈	単一音のみ				単旋律+二重奏				単旋律+二重奏+三重奏				
	×		○		×		○		×		○		
	×	○	×	○	×	○	×	○	×	○	×	○	
二重奏	PF	53.7%	63.0%	70.7%	84.7%	61.5%	63.8%	69.8%	78.9%	69.1%	70.8%	71.0%	82.7%
	CG	46.0%	44.6%	50.8%	42.8%	50.9%	67.5%	70.2%	85.1%	44.0%	57.7%	71.0%	82.9%
	VN	63.7%	81.3%	63.1%	75.6%	68.1%	85.5%	70.6%	87.7%	65.4%	84.2%	67.7%	88.1%
	CL	62.9%	70.3%	53.4%	56.1%	81.8%	92.1%	81.9%	89.9%	84.6%	95.1%	82.9%	92.6%
	FL	28.1%	33.5%	29.1%	38.7%	67.6%	84.9%	67.6%	78.8%	56.8%	70.5%	61.5%	74.3%
平均	50.9%	58.5%	53.4%	59.6%	66.0%	78.8%	72.0%	84.1%	64.0%	75.7%	70.8%	84.1%	
三重奏	PF	42.8%	49.3%	63.0%	75.4%	44.1%	43.8%	57.0%	61.4%	52.4%	53.6%	61.5%	68.3%
	CG	39.8%	39.1%	40.0%	31.7%	52.1%	66.8%	68.3%	82.0%	47.2%	62.8%	68.3%	82.8%
	VN	61.4%	76.8%	62.2%	72.5%	67.0%	81.8%	70.8%	83.5%	60.5%	80.6%	68.1%	82.5%
	CL	53.4%	55.7%	46.0%	43.9%	69.5%	77.1%	72.2%	78.3%	71.0%	82.8%	76.2%	82.8%
	FL	33.0%	42.6%	36.7%	46.5%	68.4%	77.9%	68.1%	76.9%	59.1%	69.3%	64.0%	71.5%
平均	46.1%	52.7%	49.6%	54.0%	60.2%	69.5%	67.3%	76.4%	58.0%	69.8%	67.6%	77.6%	
四重奏	PF	38.9%	46.0%	54.2%	64.9%	38.7%	38.6%	50.3%	53.1%	46.1%	46.6%	53.3%	57.2%
	CG	34.3%	33.2%	35.3%	29.1%	51.2%	62.7%	64.8%	75.3%	51.2%	64.5%	65.0%	79.1%
	VN	60.2%	74.3%	62.8%	73.1%	70.0%	81.2%	72.7%	82.3%	67.4%	79.2%	69.7%	79.9%
	CL	45.8%	44.8%	39.5%	35.8%	62.6%	66.8%	65.4%	69.3%	68.6%	74.4%	70.9%	74.5%
	FL	36.0%	50.8%	40.8%	52.0%	69.8%	76.1%	69.9%	76.2%	61.7%	69.4%	64.5%	70.9%
平均	43.1%	49.8%	46.5%	51.0%	58.5%	65.1%	64.6%	71.2%	59.0%	66.8%	64.7%	72.3%	

「単旋律+二重奏」からなる混合音テンプレートでは、31,334 個の (単音から抽出した) 特徴ベクトルを用いて一つの F0 依存多次元正規分布のパラメータ (平均ベクトルと共分散行列に相当するもの) を推定していることが分かる。また、F0 依存多次元正規分布を用いた場合と用いなかった場合、及び音楽的文脈を用いた場合と用いなかった場合についても実験を行い、結果を比較した。なお、本論文のすべての実験で、音源同定のみの性能を評価するため、各単音の音高・発音時刻・音長は正解、すなわち、前節の音響信号作成時に使用したラベルをそのまま与えるものとした。これは、これらの推定を 100%の精度で行えた場合の評価である。

実験結果を表 6 に示す。表中の認識率は 3 曲の平均値である。「DAMS 法」「F0 依存多次元正規分布」「音楽的文脈」により、二重奏は 50.9%から 84.1%へ、三重奏は 46.1%から 77.6%へ、四重奏は 43.1%から 72.3%へ認識率を改善することができた。また、実験結果より以下のことが分かる。

- 「単旋律+二重奏」のテンプレート使用時と「単旋律+二重奏+三重奏」のテンプレート使用時とは結果に大きな差はなかった。特に四重奏に対しても、「単旋律+二重奏」のテンプレートだけで単一音のテンプレートに比べて高い認識率を示した。これは、必ずしも認識対象と同程度の複雑さの楽曲でなくても、同時発音する他の音からの影響を受けているデータであ

ば、認識率の改善に貢献できることを示している。

- 5 楽器のうち PF と CG は、F0 依存多次元正規分布の効果が大きかった。これは、これらの楽器の音域が他の楽器よりも広いために、音高による音色変化が顕著に現れたからだと考えられる。

- 音楽的文脈を利用することで、認識率が 10%前後改善された。これは、我々が用いた 3 曲にパート間の音高交差がなく、我々が用いた音楽のヒューリスティクスとよくマッチしていたからである。

- 単一音のみによるテンプレートをを用いた場合など、もともとの認識率が低いとき (30~40%台) に音楽的文脈を利用すると、認識率が低下することがあった。本手法のような、前後の単音の事後確率を事前確率に反映させる方法では、前後の単音の事後確率 (仮計算による確率) が一定の精度で計算される必要があり、その精度に満たなかったためと考えられる。混合音テンプレートをを用いた場合はこの現象は見られないため、音楽的文脈を用いる場合には、混合音テンプレートを併用するなどしてももとの性能を高めることが重要である。

- 5 楽器のうち PF はあまり認識率が高くない場合が多かった。これは、CG と音色が似ているからと考えられる。実際、両者の調波構造を抽出し再合成したものを試聴したところ、区別の難しいものが多かった。

DAMS 法によって得られた特徴量の重み値の主なものを表 7 に示す。比較のため、単一音データに対し

表 7 DAMS 法によって得られた主な特徴量の重み値
Table 7 Weights of features obtained from the DAMS method.

D	第 1 軸	15 (0.31), 16 (0.31), 30 (0.32), 40 (0.40)
A	第 2 軸	11 (-0.56), 20 (-0.58)
M	第 3 軸	4 (0.39), 5 (0.53), 42 (-0.43)
S	第 4 軸	3 (-0.39), 5 (0.48), 8 (0.32), 12 (-0.39)
単	第 1 軸	20 (0.33), 30 (-0.41), 40 (-0.42)
一	第 2 軸	11 (-0.55), 20 (-0.30)
音	第 3 軸	19 (-0.31), 20 (-0.42), 42 (-0.36)
	第 4 軸	12 (-0.34), 40 (0.60)

て主成分分析・線形判別分析を行って得られたものも同表に示す．単一音データに対して行った場合に比べると，DAMS 法では「スペクトルの時間平均に関する特徴」の重みが相対的に大きかった．これは「スペクトルの時間平均に関する特徴」に比べて，パワーを全高調波成分のパワーの累積として求めるために「パワーの時間変化に関する特徴」は，どこかの高調波成分にパワーの大きい他の音の高調波成分が重なったときに壊れやすいことが関係していると考えられる．また，両者の場合において，11 (奇数次と偶数次の高調波成分のパワー比)，20 (持続時間が最長の高調波成分の 90%ある高調波成分の個数)，30 (発音開始直後 0.55 秒間のパワー包絡の微分係数の中央値)，40 (振幅変調の振幅)，42 (周波数変調の振幅)の重みが大きかった．

6.3 実験 2 : 1 曲のみからのテンプレート作成

次に，混合音テンプレートを 1 曲のみから作成した場合と 2 曲用いて作成した場合 (Leave-one-out 法)での認識率を比較するため，混合音テンプレートを 1 曲のみから作成した場合の実験を行った．実験結果を表 8 に示す．1 曲のみからの作成でも CG, VN, CL は比較的高い認識率であった．FL は作成用楽曲によっては 30%の認識率であったが，2 曲用いて作成することで，高い認識率が得られた．これらは，クローズド実験 (ただし四重奏は学習データに含まれていない)の結果に匹敵する認識率となっている．このことから，複数楽器の同時発音の影響の多様性に対して十分な量の学習データがただか 2 曲から得られているといえる．ただし，学習に用いる曲も同定に用いる曲もともに小編成のクラシック曲となっており，異なるジャンルの楽曲を含む多様な音楽データベースに対する性能は，今後の更なる検証が必要である．

表 8 混合音テンプレートを 1 曲のみから作成した場合 (実験 2) の結果 (認識率, 四重奏) [単位: %]
Table 8 Results of Experiment 2.

	単旋律+二重奏				単旋律+二重奏+三重奏			
	13	16	17	*	13	16	17	*
PF	(57.8)	32.3	38.4	36.6	(67.2)	33.2	45.1	39.7
No. CG	(73.3)	78.1	76.2	76.7	(76.8)	84.3	80.3	82.1
13 VN	(89.5)	59.4	87.5	86.2	(87.2)	58.0	85.2	83.1
CL	(68.5)	70.8	62.2	73.8	(72.3)	72.3	68.6	75.9
FL	(85.5)	40.2	74.9	82.7	(86.0)	38.9	68.8	80.8
PF	74.1	(64.8)	61.1	71.2	79.6	(67.1)	73.0	78.3
No. CG	79.2	(77.9)	78.9	74.3	70.4	(82.6)	74.0	75.2
16 VN	89.2	(85.5)	87.0	87.0	86.0	(83.5)	84.7	85.0
CL	68.1	(78.9)	68.9	76.1	72.4	(82.8)	76.3	82.1
FL	82.0	(75.9)	72.5	77.3	77.9	(72.3)	35.7	69.2
PF	53.0	39.4	(51.2)	51.6	52.2	40.6	(55.7)	53.7
No. CG	73.7	69.0	(75.8)	75.0	76.0	74.3	(78.4)	80.0
17 VN	79.5	61.2	(78.3)	73.6	77.4	58.0	(78.7)	71.7
CL	51.3	60.5	(57.1)	57.9	61.1	62.6	(66.9)	65.4
FL	65.0	35.0	(73.1)	68.7	58.6	34.7	(70.9)	62.6

*Leave-one-out 法の場合 (参考)．括弧付は closed 実験．「単旋律+二重奏 13」は，楽曲 No.13 の SMF から抜粋した単旋律と二重奏のみを用いたテンプレート作成の意．

表 9 実験 3 において用いた楽器の組合せ
Table 9 Instrument combinations in Experiment 3.

単旋律	PF, CG, VN, CL, FL
二重奏	PF-PF, CG-CG, VN-PF, CL-PF, FL-PF
三・四重奏	使用せず

6.4 実験 3 : 楽器の組合せが不完全な場合

混合音テンプレート中の楽器の組合せが網羅されていないときに認識率がどの程度変化するか確認するため，テンプレート中の楽器の組合せを大幅に減らした状態で実験を行った．楽器の組合せは，楽器数 n に対して $O(n)$ のオーダになるように，表 9 のように定めた．混合音テンプレート作成は，同定対象曲以外の 2 曲から作成した単旋律と二重奏を用いた (Leave-one-out 法)．実験結果を表 10 に示す．楽器の組合せを減らした場合と減らさなかった場合とで認識率に大きな差はなかった．すべての楽器の組合せを網羅したテンプレートを作成するには，楽器数を n として m 重奏を扱う場合に $O(n^m)$ のオーダのデータ量が必要となり (ただし，音域的に発音不可能な組合せを除くと多少少なくなる)，将来より多くの楽器を扱う際に困難が予想される．しかし，楽器の組合せを網羅しないテンプレートでも性能向上を確認できたことから，組合せ爆発の問題は深刻にはならないと期待できる．

6.5 実験 4 : 線形判別分析の効果の評価

最後に，DAMS 法における線形判別分析の効果の評価するため，主成分分析のみで次元圧縮をした場合

表 10 楽器の組合せを減らした場合（サブセット）と減らさなかった場合（フルセット）の比較（実験 3）
Table 10 Comparison of the full set and the subset of instrument combinations.

		サブセット	フルセット
二重奏	PF	85.4%	78.9%
	CG	70.8%	85.1%
	VN	88.2%	87.7%
	CL	90.4%	89.9%
	FL	79.7%	78.8%
	平均	82.9%	84.1%
三重奏	PF	73.9%	61.4%
	CG	62.0%	82.0%
	VN	85.7%	83.5%
	CL	79.7%	78.3%
	FL	76.5%	76.9%
	平均	75.6%	76.4%
四重奏	PF	68.9%	53.1%
	CG	52.4%	75.3%
	VN	85.0%	82.3%
	CL	71.1%	69.3%
	FL	74.5%	76.2%
	平均	70.4%	71.2%

の実験を行った。実験方法は、実験 1 や実験 3 と同様に Leave-one-out 法とした。実験結果を図 5 に示す。主成分分析のみで次元圧縮をした場合、単一音を学習データとしたとき (S) と混合音を学習データとしたとき (S+D, S+D+T) との認識率の差が 6~14%程度だったのに対し、線形判別分析を用いた場合では 20~24%程度になった。主成分分析はクラス内分散・クラス間分散比を考慮しないため、必ずしも識別に有効な特徴量の重みが高くないのに対し、線形判別分析はクラス内分散・クラス間分散比を最小化するため、識別に有効な特徴量（すなわち混合音の影響がより少なかった特徴量）の重みが高くなったと考えられる。また、すべての条件において線形判別分析を用いた方が用いないよりも認識率は高かった。

6.6 考 察

我々は五つの楽器から選ばれた二重奏~四重奏について実験を行い、二重奏では平均 84.1%、三重奏で 77.6%、四重奏で 72.3%の認識率を得た。本節では、これらの結果について関連研究と比較しながら議論する。

音源同定の難易度を決める要因として、対象楽器数 (TI) と同時発音数 (SI) がある。柏野ら [13], [14] は、3 楽器を対象に三重奏 (TI, SI ともに 3) を扱い、88%の認識率を得た。木下ら [15] は、3 楽器によるランダムノートパターン (TI, SI ともに 3) を扱い、72~81%の認識率を得た。Eggink ら [16] は、5 楽器

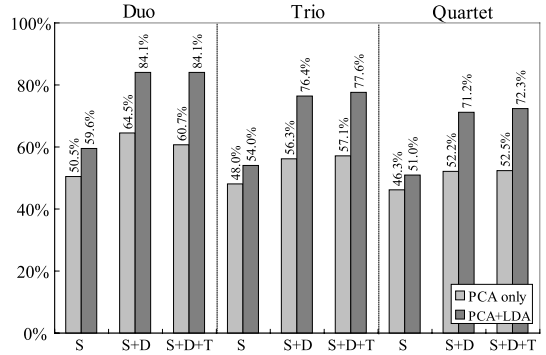


図 5 主成分分析 (PCA) のみを用いて次元圧縮した場合と PCA と線形判別分析 (LDA) とを併用した場合との比較 (実験 4)。グラフ上の「Duo」「Trio」「Quartet」は認識対象楽曲を表し。グラフ下の「S」「S+D」「S+D+T」はそれぞれ「単一音」「単旋律+二重奏」「単旋律+二重奏+三重奏」の特徴量テンプレートをを用いたことを表す。グラフから、LDA を用いることで混合音テンプレートがより効果的に認識率を改善していることが分かる。

Fig.5 Comparison of using only PCA and using PCA and LDA together.

から選ばれた二重奏 (TI: 5, SI: 2) に対して約 50%の認識率を得た。我々の問題設定は TI が 5, SI が 2~4 で、柏野ら、木下らよりも TI が多く、Eggink らと比べて SI が同じか多くなっている。Eggink らと本実験の二重奏は TI, SI ともに等しくなっているが、我々は Eggink らと比べて高い認識率 (84.1%) を得ている。ただし、実験データは異なるものを使用しているため、正確な比較はできない。

音源同定の難易度を決めるもう一つの要因として、各単音の発音時刻や音高の正解を与えるかどうかがある。本論文の実験も含め、これまでの多くの研究では、これらは正解を与えて実験をしていた [13], [14], [16]。実際、上で述べた認識率はいずれも音高などの正解を与えた場合である。これは、混合音中の発音時刻検出や音高推定はまだまだ困難な課題であり、音源同定部のみの性能を評価するためと考えられる。各単音の発音時刻や音高を自動推定して音源同定をした場合、性能が落ちることが予想される。例えば、木下ら [15] は、各単音の発音時刻や音高の正解を与えた場合の認識率が 72~81%だったのに対し、これらを自動推定した場合 66~75%であったと報告している。この問題に対処するには、発音時刻検出や音高推定の精度を上げるだけでなく、音源同定部においても、これらに誤りがあることを前提とした処理が求められる。このような処理の検討は、重要な今後の課題の一つである。

7. む す び

本論文では、多重奏に対する音源同定における最大の課題である「音の重なりによる特徴変動」に対する解決策として、特徴変動の程度に応じた特徴量の重み付けを提案した。この重み付けは、混合音から学習データを作成して線形判別分析を適用することで実現される。混合音から学習データを作成するというアプローチは、シンプルにもかかわらずこれまでの研究では試されてこなかった。その理由の一つに、あらゆる混合音を網羅した学習データを作るにはばく大なデータが必要となるということが考えられる。しかし、我々の実験により、あらゆる混合音を網羅していなくても、実楽曲の楽譜から混合音を作成することで、十分に性能向上に有効なデータを得られることが分かった。更に、音楽的文脈を考慮する手法として、事前確率を前後の単音の事後確率から計算する方法を提案した。これにより、メロディの時間的連続性を考慮した楽器同定が可能になった。

今後は、本実験で手動で与えた各単音の発音時刻や音高を自動で推定し、音楽音響信号に対する自動楽器インデキシングを実現する。更に、この技術を活用した音楽検索システムの構築を進めていく予定である。

謝辞 本研究の一部は、日本学術振興会科学研究費補助金(基盤研究(A), 特定領域「情報学」, 特定領域「情報爆発」, 特別研究員奨励費), 21世紀COEプログラム「知識社会基盤構築のための情報学拠点形成」, 科学技術振興機構CREST「時系列メディアのデザイン転写技術の開発」による。また、「RWC研究用音楽データベース」を使用した。

文 献

- [1] B.S. Manjunath, P. Salembier, and T. Sikora, Introduction of MPEG-7, John Wiley & Sons, 2002.
- [2] K.D. Martin, Sound-Source Recognition: A Theory and Computational Model, PhD thesis, MIT, 1999.
- [3] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," Proc. ICASSP, pp.735-756, 2000.
- [4] A. Fraser and I. Fujinaga, "Toward real-time recognition of acoustic musical instruments," Proc. ICMC, pp.175-177, 1999.
- [5] I. Fujinaga and K. MacMillan, "Realtime recognition of orchestral instruments," Proc. ICMC, pp.141-143, 2000.
- [6] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," EURASIP J. Applied Signal Process., vol.2003, no.1, pp.5-14, 2003.
- [7] 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楽器音の音源同定: F0 依存多次元正規分布に基づく識別手法," 情処学論, vol.44, no.10, pp.2448-2458, 2003.
- [8] J. Marques and P.J. Moreno, "A study of musical instrument classification using Gaussian mixture models and support vector machines," CRL Technical Report Series, CRL/4, 1999.
- [9] J.C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," J. Acoust. Soc. Am., vol.103, no.3, pp.1933-1941, 1999.
- [10] J.C. Brown, "Feature dependence in the automatic identification of musical woodwind instruments," J. Acoust. Soc. Am., vol.109, no.3, pp.1064-1072, 2001.
- [11] A.G. Krishna and T.V. Sreenivas, "Music instrument recognition: From isolated notes to solo phrases," Proc. ICASSP, vol.IV, pp.265-268, 2004.
- [12] 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," 信学論 (D-II), vol.J79-D-II, no.11, pp.1751-1761, Nov. 1996.
- [13] 柏野邦夫, 村瀬 洋, "適応型混合テンプレートをを用いた音源同定," 信学論 (D-II), vol.J81-D-II, no.7, pp.1510-1517, July 1998.
- [14] 柏野邦夫, 村瀬 洋, "単音連鎖確率ネットワークに基づく音楽演奏の音源同定," 人工知能誌, vol.13, no.6, pp.962-970, 1998.
- [15] 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," 信学論 (D-II), vol.J83-D-II, no.4, pp.1073-1081, April 2000.
- [16] J. Eggink and G.J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," Proc. ISMIR, 2003.
- [17] 後藤真孝, 村岡洋一, "打楽器音を対象にした音源分離システム," 信学論 (D-II), vol.J77-D-II, no.5, pp.901-911, May 1994.
- [18] 吉井和佳, 後藤真孝, 奥乃 博, "テンプレート適応を利用した実世界の音楽音響信号に対するドラムスの音源同定," 情処学研報, 2003-MUS-53, pp.55-60, 2003.
- [19] 中臺一博, 柏野邦夫, 田中英彦, "音楽音響信号を対象とする音源分離システム," 情処学研報, 93-MUS-1, pp.1-8, 1993.
- [20] 桜庭洋平, 奥乃 博, "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," 情処学研報, 2003-MUS-51, pp.35-42, 2003.
- [21] 安藤由典, 楽器の音響学, 音楽之友社, 1996.
- [22] 山本俊一, 中臺一博, 辻野広司, 奥乃 博, "ミッシングフィーチャー理論を利用した音源分離と音声認識のインターフェースと複数ロボットへの適用," 日本ロボット学会誌, vol.23, no.6, pp.743-751, 2005.
- [23] A.S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, 1990.
- [24] 亀岡弘和, 西本卓也, 嵯峨山茂樹, "調波時間構造化クラ

スタリング (HTC) による音楽音響特徴量の同時推定 ; 情処学研報, 2005-MUS-61, pp.71-78, 2005.

- [25] 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：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース” 情処学論, vol.45, no.3, pp.728-738, 2004.

(平成 17 年 11 月 17 日受付, 18 年 7 月 10 日再受付)



北原 鉄朗 (学生員)

2002 東京理科大・理工・情報科学卒。2004 京都大学大学院情報学研究科知能情報学専攻修士課程了。現在, 同大学院博士後期課程在学中。2005 より日本学術振興会特別研究員 (DC2)。音楽情報処理に興味をもつ。電気通信普及財団第 19 回テレコムシステム技術学生賞, 情報処理学会第 67 回全国大会大会奨励賞等受賞。情報処理学会, 人工知能学会, 日本音響学会, 日本音楽知覚認知学会, IEEE 各学生会員。



後藤 真孝 (正員)

1993 早大・理工・電子通信卒。1998 同大学院理工学研究科博士後期課程了。同年, 電子技術総合研究所 (2001 に独立行政法人産業技術総合研究所に改組) に入所し, 現在に至る。2000 から 2003 まで科学技術振興事業団さきがけ研究 21 「情報と知」領域研究員, 2005 から筑波大学大学院システム情報工学研究科助教授 (連携大学院) を兼任。博士 (工学)。音楽情報処理, 音声言語情報処理等に興味をもつ。2001 日本音響学会粟屋潔学術奨励賞・ポスター賞, 2003 インタラクシオン 2003 ベストペーパー賞, 2005 情報処理学会論文賞等 18 件受賞。情報処理学会, 日本音響学会, 日本音楽知覚認知学会各会員。



駒谷 和範 (正員)

1998 京大・工・情報工学卒。2000 同大学院情報学研究科知能情報学専攻修士課程了。2002 同大学院博士後期課程了。同年より京都大学情報学研究科助手。京都大学博士 (情報学)。情報処理学会平 16 年度山下記念研究賞, FIT2002 ヤングリサーチャー賞受賞。情報処理学会, 言語処理学会, 人工知能学会, ACL 各会員。



尾形 哲也

1993 早大・理工・機械卒。日本学術振興会特別研究員, 早稲田大学理工学部助手, 理化学研究所脳科学総合研究センター研究員, 京都大学大学院情報学研究科講師を経て, 2005 より同助教授。博士 (工学)。早稲田大学ヒューマノイド研究所客員助教授, 理化学研究所脳科学総合研究センター客員研究員を兼務。人間とロボットのインタラクシオンと協調, 神経回路モデルなどの研究に従事。2000 年度日本機械学会論文賞, IEA/AIE-2005 最優秀論文賞などを受賞。情報処理学会, 日本ロボット学会, 日本機械学会, 人工知能学会, IEEE 等各会員。



奥乃 博

1972 東大・教養・基礎科学卒。日本電信電話公社, NTT, JST 北野プロジェクト, 東京理科大学を経て, 2001 年 4 月より京都大学大学院情報学研究科知能情報学専攻教授。博士 (工学)。この間, スタンフォード大学客員研究員, 東京大学工学部客員助教授。人工知能, 音環境理解, 音楽情報処理, ロボット聴覚の研究に従事。1990 年度人工知能学会論文賞, IEA/AIE-2001, 2005 最優秀論文賞, 平 14 年度船井情報科学振興賞等受賞。情報処理学会, 人工知能学会, 日本ロボット学会, ACM, AAAI, IEEE 等会員。『インターネット活用術』(岩波書店), 『Computational Auditory Scene Analysis』(共編, LEA), 『Advanced Lisp Technology』(共編, Taylor & Francis) ほか。