音オントロジーを用いた音楽情報処理の研究

課題番号 12480090

平成12年度～平成14年度

科学研究費補助金 (基盤研究(B)(2))

研究成果報告書

平成１５年３月

研究代表者　奥乃　博

(京都大学大学院 情報学研究科 教授)

音オントロジーを用いた音楽情報処理の研究

課題番号 12480090

平成12年度～平成14年度

科学研究費補助金 (基盤研究(B)(2))

研究成果報告書

平成15年3月

研究代表者　奥乃　博

(京都大学大学院 情報学研究科 教授)

はしがき

研究目的

これまでの音の認識及び生成の研究では, 音の表現がシステム毎に定められているので, 拡張性に欠けたり, 複数システムの統合が難しいという問題点があった. このような問題を解決するために, 人工知能の工学的応用の可能化技術 (enabling technologies) の一つとして注目を浴びているオントロジーの概念を音に適用することことが考えられる. さらに, インターネット上のさまざまなコンテンツを効率良く検索し, 流通していくという Semantic Web においても, Multimedia Content Description Interface として, MPEG-7 などの規格が定められており, その意味的なアノテーションに音オントロジーを含めた各種オントロジーの開発が必要とされている. このように, ソフトウェア構築上の課題を解決し, さらに, マルチメディアコンテンツ流通させ, 次世代インターネットとでもいうべき Semantic Web を実現していくためにも音オントロジーへの期待が高まっている.

本研究では, 音楽情報処理分野のための音オントロジーを確立するために, 楽器音同定手法や階層的なクラスタリング手法について研究・開発を行う. 具体的な研究目標は以下の通りである：

(1) 非打楽器音の自動認識手法,

(2) 打楽器音の自動認識手法,

(3) 複数楽器演奏からの混合音分離手法および認識手法,

(4) 音響的特徴から楽器音オントロジーのシステマティックな構築手法,

(5) 音オントロジーによる音楽コーパスのタグ付け手法.

研究組織

研究代表者　奥乃 博　(東京理科大学・理工学部・教授, 平成 12 年度)
(京都大学・大学院情報学研究科・教授, 平成 13 年度～)

研究分担者　後藤 真孝　(電子技術総合研究所・知能情報部・研究員, 平成 12 年度)
(産業技術総合研究所・情報処理研究部門・研究員, 平成 13 年度～)

河原 達也　(京都大学・大学院情報学研究科・助教授, 平成 13 年度～)

研究協力者　中臺 一博　(科学技術振興事業団・北野共生システムプロジェクト・研究員)

北原 鉄郎　(京都大学・大学院情報学研究科・修士課程)

桜庭 洋平　(京都大学・大学院情報学研究科・修士課程, 平成 13 年度～)

研究経費 (配分額)

	直接経費	間接経費	合 計
平成 12 年度	5,300 千円	0	5,300 千円
平成 13 年度	1,900 千円	0	1,900 千円
平成 14 年度	2,300 千円	0	2,300 千円
総 計	9,500 千円	0	9,500 千円

研究発表

【論文誌】

[1] 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. 情報処理学会論文誌, 条件つき採録 (Jan. 2003) 情報処理学会.

【国際会議予稿】

[1] Kitahara, T., Goto, M., and Okuno, H.G.: Musical Instrument Identification based on F0-dependent Multivariate Normal Distribution. in *Proceedings of 2003 International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003)*, IEEE, Vol.5, pp.421–424, Hong Kong, Apr. 2003.

[2] Kitahara, T., Goto, M., and Okuno, H.G.: Pitch-dependent Musical Instrument Identification and Its Application to Musical Sound Ontology. In Hinde, C. and Ali, M. (Eds.) Proceedings of Sixteenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2003), Loughborough, UK, Jun. 2003, Lecture Notes in Artificial Intelligence, , *to appear*, Springer-Verlag.

[3] Kitahara, T., Goto, M., and Okuno, H.G.: Musical Instrument Identification based on F0-dependent Multivariate Normal Distribution. in *Proceedings of 2003 International Conference on Multimedia and Expo (ICME 2003)*, IEEE, *to appear*, Baltimore, MD, Jul. 2003.

【解説論文】

[1] 奥乃 博, 中臺一博: ロボットの耳は二つで十分か. 日本音響学会誌, Vol.58, No.3 (Mar. 2002) pp.205–210.

【国内会議発表】

[1] 北原鉄朗, 後藤真孝, 奥乃博: 音響的類似性に基づく楽器音の階層的クラスタリング, 情報処理学会第 65 回全国大会, 1P-1, Mar. 2003.

[2] 櫻庭洋平, 奥乃博: 音色類似度と定位類似度の統合による自動採譜, 情報処理学会第 65 回全国大会, 1P-2, Mar. 2003.

[3] 吉井和佳, 北原鉄朗, 櫻庭洋平, 奥乃博: 教師なしクラスタリングと認識誤りパターンを利用した打楽器音の音源同定, 情報処理学会第 65 回全国大会, 1P-3, Mar. 2003.

[4] 北原鉄朗, 後藤真孝, 奥乃 博: 音高の音色変化に着目した音源同定手法. 音響学会秋季研究発表会講演論文集, 1-1-4, pp.643-644, Sep. 2002.

[5] 櫻庭洋平, 河原達也, 奥乃 博: 定位情報と音色情報を用いた複数楽器音の認識, 音楽情報処理研究会, SIG-MUS-46-1, 情処研報, Vol.2002, No.63, pp.1–8, 情報処理学会, Jul. 2002.

[6] 北原鉄朗, 後藤真孝, 奥乃 博: 音高の音色変化に着目した音源同定手法. 音楽情報処理研究会, SIG-MUS-46-2, 情処研報, Vol.2002, No.63, pp.9–16, 情報処理学会, Jul. 2002.

[7] 北原 鉄朗, 後藤 真孝, 奥乃 博: 音高の音色変化に着目した音源同定手法. 音楽情報処理研究会, SIG-MUS, 情報処理学会, May 2001.

[8] 北原 鉄朗, 後藤 真孝, 奥乃 博: 楽器音オントロジー作成のための楽器音特徴抽出. 情報処理学会第 62 回全国大会, 4M-5, Mar. 2001.

成果の概要

本報告書は, 以下の 4 章から構成される.

- 第 1 章では, 非打楽器音の単音の音源同定に取り組み, 楽器音同定のための音響的特徴量の抽出手法を開発すると共に, それらの特徴量の楽器音同定における有意さを検討する.
- 第 2 章では, 非打楽器音オントロジーのシステマティックな構築法に取り組み, 第 1 章で提案した特徴量を用いて, 決定木学習により楽器音階層を構築するとともに, 構築された階層的な表現の妥当性について検討する.
- 第 3 章では, 単音ではなく混合音のパート譜自動認識に取り組み, オクターブ関係にある音の分離で不可避な曖昧性を, 音源定位情報と音色情報により解消する手法を開発する.
- 第 4 章では, 打楽器演奏の音源同定に取り組み, ドラム類とシンバル類とを帯域フィルタにより分離した後, 音響的にばらつきの大きいドラムには教師なし学習で音源同定を行う手法を開発すると共に, シンバル類については認識誤りパターンを活用した誤り補正を行う手法を開発し, 提案する手法の有効性を検討する.

また, 付録には, 文献 [国際会議論文 -1] のコピーを掲載する.

【第 1 章】音高による音色変化に着目した楽器音の音源同定手法

本章では, 音高による音色変化を考慮した音源同定手法である F0 依存多次元正規分布に基づく識別手法を提案する. 楽器音のの音色が音高によって変化すること, あるいは, 楽器の音色が演奏法だけでなく個体差によっても大きく変化することは従来から広く知られていた. それにもかかわらず, これを適切に扱える音源同定手法については, ほとんど研究されてこなかった. 本研究では, 音高による音色変化を適切に扱うため, 平均が基本周波数によって変化する多次元正規分布を用いる. すなわち, 音色空間 (楽器音の特徴空間) 上で各楽器音データがこの分布に従うと仮定し, この分布のための識別関数をベイズ決定規則から定式化する. 提案手法を実装・実験した結果, 音高による音色変化を考慮しない多次元正規分布を用いた場合の誤認識全体のうち, 個々の楽器レベルでは 16.48% の誤認識を削減することができた. さらに、個別の楽器ではなく, 楽器グループで構成される階層的なカテゴリーレベルでは 20.67% の誤認識を削減することができた. 階層的なカテゴリーレベルでの認識により, システムが学習していない未知楽器に対して, システムが最も適切なレベル迄認識をすることが可能となった. 今後, 混合音の扱い方が一番重要な課題となる.

【第 2 章】音オントロジーのシステマティック構築手法

音オントロジーのシステマティックな構成法として, F0 依存多次元正規分布で表現された特徴量から, 決定木学習を繰り返すことにより, 楽器音の階層的な表現を自動的に構築することが可能となった. このような手法で構築された楽器音階層は, 従来教科書などに記述されている楽器音階層木とは一部異なっている. 例えば, クラリネットとリコーダは偶数次倍音が小さいという特徴から, 他の間楽器との類似性が低い, A2 (110Hz) ではバリトンサックスとファゴットはテナーサックスやトロンボーンよりもバイオリンとの類似度の方が高い, 楽器音 1 サンプルだけ使用して得られた階層構造は, 楽器の個体差に大きく影響を受ける, などの知見が得られた. また, クラリネットとサックスは共に単簧楽器 (single reed) であり, オーボエとファゴットは共に複簧楽器 (multiple reed) であるので, 通常統べてリード楽器として分類されている. しかし, リードだけではなく円筒管 (クラリネット) か円錐管 (サックス) という構造上の違いがスペクトルに影響を与えるので, 本階層的表現では類似度が小さくなっている. 今後, 音オントロジー構築のために, システマティックに構築した階層的表現がどの程度人間の知覚とあっているのかを調査し, その妥当性を明らかにする.

【第 3 章】定位情報と音色情報を用いた複数楽器音の認識手法

オクターブ関係にある音が同時に演奏されると，ピッチ (F0) の同定や音源定位が曖昧になるという問題を，演奏者の位置情報を使用して，解消する手法を開発した．ステレオ入力から両耳間位相差と両時間強度差を用いて，各楽器の定位を求める．この過程で，調波構造構造と定位情報とを統合し，両者の曖昧性を解消する．各周波数成分の重なりは位相差の変動に着目して判定する．重なり情報を利用して単音を形成する同時的グルーピングと，得られた単音の音色類似度と定位情報を手がかりとしてパートごとの流れを形成する継時的グルーピングの 2 つの処理から構成される複数楽器音の音源同定分離システムを構築した．無響室で録音した四重奏に対して，精度と再現率の統合指標 F-Measure がそれぞれ 10% 以上向上をし，提案手法の有効性を確認した．今後，第 1 章で提案した特徴量を使用した音源同定システムとの統合を図り，混合音分離・認識をよりロバストなものにする．

【第 4 章】教師なしクラスタリングと認識誤り補正による複数打楽器音の認識手法

ドラムやシンバルを含む複数の打楽器による演奏を対象とした音源同定手法を開発することで，非打楽器音も含めた一般の複数楽器演奏から音源分離や認識，さらには自動採譜への基礎技術を確立する．従来研究で得られている知見は，多くの場合打楽器単音を対象としているため，複数打楽器による演奏には簡単には適用できない．本研究では，ドラムのような膜鳴楽器類とシンバルのような体鳴楽器類を，まず，前者には低域通過フィルタ処理し，後者には高域通過フィルタ処理した音響信号だけに絞り込むことを通じて，各楽器類の音源同定を別処理により行う．膜鳴楽器は演奏法やチューニングにより音響特性が大きく変わり，また，非打楽器音と比較して学習用データベースが大きくないので，教師なしクラスタリング手法による音源同定手法を開発した．体鳴楽器の音源同定には，音の重なりによる特徴量変動と認識誤りに一定のパターンがあることに着目し，k-NN 識別後，認識誤り補正を行う手法を開発した．この手法により，識別すべき体鳴楽器が Crash Cymbal の残響下にある場合，Snare Drum と同時発音する場合など，音の重なりにより誤認識が生じやすい問題に対処できるようになった．ベンチマークにより，膜鳴楽器の音源同定率は，音源によらず 9 割程度を達成し，体鳴楽器の音源同定については，MIDI 音源で作成した評価用データに対し 50%，市販 CD に対し 10% の認識誤り削減率が得られた．今後，音源同定をよりロバストにすると共に，一般の楽音の音源同定と統合していく必要がある．

目 次

1 F0 依存多次元正規分布に基づく識別手法 1

1.1 はじめに . . . 1

1.2 F0 依存多次元正規分布 . . . 2

1.2.1 代表値関数 . . . 2

1.2.2 F0 正規化共分散行列 . . . 2

1.2.3 ベイズ決定規則による識別 . . . 3

1.3 処理の流れ . . . 3

1.3.1 調波構造の推定 . . . 4

1.3.2 特徴抽出 . . . 4

1.3.3 主成分分析・線形判別分析による次元圧縮 . . . 5

1.3.4 識別 . . . 5

1.4 評価実験 . . . 5

1.4.1 実験方法 . . . 5

1.4.2 実験結果 . . . 7

1.4.3 主成分分析に関する考察 . . . 7

1.4.4 線形判別分析に関する考察 . . . 9

1.4.5 実験結果に関する考察 . . . 9

1.5 k-NN 法との比較 . . . 10

1.6 おわりに . . . 11

2 楽器音オントロジーのシステマティックな構成法 17

2.1 Introduction . . . 17

2.2 F0-dependent Multivariate Normal Distribution . . . 18

2.2.1 Pitch and Non-pitch Dependencies . . . 18

2.2.2 Features for Musical Instrument Identification . . . 19

2.3 A Discriminant Function based on the Bayes Decision Rule . . . 19

2.4 Musical Instrument Identification . . . 20

2.4.1 Experimental Conditions . . . 20

2.4.2 Results of Musical Instrument Identification . . . 21

2.5 Evaluation of the Bayes Decision Rule . . . 22

2.6 Musical Instrument Ontology . . . 23

2.6.1 Musical instrument ontology by C5.0 . . . 23

2.6.2 Pitch-dependency in musical sound ontology by agglomerative hierarchical clustering . . 24
2.7 Conclusion . . . 25

3 定位情報と音色情報を用いた複数楽器音の認識 31
3.1 はじめに . . . 31
3.2 同時的グルーピングの曖昧性と定位の利用 . . . 32
3.3 継時的グルーピングの曖昧性と定位の利用 . . . 33
3.4 システムの構成 . . . 33
3.4.1 周波数解析部 . . . 34
3.4.2 定位抽出部 . . . 34
3.4.3 単音形成部 . . . 35
3.4.4 特徴抽出部 . . . 36
3.4.5 音源同定部 . . . 36
3.4.6 結果結合部 . . . 36
3.4.7 音源同定予備実験 . . . 37
3.5 システム評価実験 . . . 37
3.5.1 同時的グルーピング実験 . . . 37
3.5.2 継時的グルーピング実験 . . . 38
3.5.3 考察 . . . 39
3.5.4 今後の課題 . . . 39
3.6 おわりに . . . 39

4 教師なしクラスタリングと認識誤り補正による打楽器演奏の音源同定 43
4.1 はじめに . . . 44
4.2 打楽器音を対象とした音源同定の従来研究 . . . 45
4.2.1 打楽器音の識別手法の比較検討に関する従来研究 . . . 45
4.2.2 パワー分布に基づくパターンマッチング手法 . . . 46
4.2.3 本研究でのアプローチ . . . 46
4.3 打楽器演奏を対象とした音源同定手法 . . . 47
4.3.1 音源同定の対象とする問題設定 . . . 47
4.3.2 問題の所在と解決手法 . . . 48
4.3.3 打楽器音同定システムの構成 . . . 48
4.3.4 発音時刻検出 . . . 49
4.3.5 膜鳴楽器の教師なしクラスタリング . . . 50
4.3.6 体鳴楽器の音源同定と認識誤り補正 . . . 51
4.4 実験と考察 . . . 54
4.4.1 実験条件 . . . 54
4.4.2 発音時刻検出実験 . . . 54
4.4.3 膜鳴楽器の音源同定実験及び考察 . . . 56
4.4.4 体鳴楽器の音源同定実験及び考察 . . . 58
4.5 おわりに . . . 61

参考文献 62

ICASSP'2003 掲載論文 67

第 1 章

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

本章では, 音高による音色変化を考慮する音源同定手法を提案する. 楽器音の音色が音高によって変化することは, 従来から広く知られているにも関わらず, これを適切に扱える音源同定手法については, ほとんど研究されてこなかった. これに対して, 我々は音高による音色変化を適切に扱うため, 平均が基本周波数によって変化する多次元正規分布を提案する. そして, 音色空間 (楽器音の特徴空間) 上で各楽器音データがこの分布に従うと仮定し, この分布のための識別関数をベイズ決定規則から定式化する. 提案手法を実装・実験した結果, 音高による音色変化を考慮しない多次元正規分布を用いた場合の誤認識全体のうち, 個々の楽器レベルでは 16.48%, カテゴリーレベルでは 20.67% の誤認識を削減することができた.

1.1 はじめに

楽器音の同定がパターン認識の研究対象として広く扱われるようになったのは, 音声や文字などより遅く, 1990 年代に入ってからのことである [6, 4, 23, 26, 29, 35, 13, 9]. そのため, 音声認識や文字認識に比べて得られている知見は少ない. また, 楽器音響学の分野では古くからさまざまな分析が行われてきた [47, 19, 39, 1] が, 音源同定の工学的モデルの実現には至っていない.

我々は, 音源同定を**音楽情景分析** [23] (音楽音響信号を対象とした計算機による聴覚的情景分析 [3]) の重要な要素技術の 1 つと位置づけ,

(1) 単音の音源同定,

(2) 混合音の音源同定,

という 2 段階のアプローチをとって研究を進めている. このように 2 段階のアプローチをとるのは, 音源同定は単音であっても難しい問題だからである. たとえば, 文献 [35] では, 音楽経験者を対象とした聴取実験 (14 楽器の単音を聞いて, その楽器名を 27 個の楽器名が書かれたリストから選ぶという実験) で, 45.9% の正解率が報告されている.

楽器音の音源同定の精度を高くするのが難しい原因は, 楽器の音色が, 楽器の個体差や音高などのさまざまな要因により変化するからである. しかし, 従来の音源同定研究の多く [6, 4, 23, 29, 35, 13, 9] は, こうした音色変化の問題を明示的に扱っていなかった. それに対し, 柏野らの適応型混合テンプレート法 [26] は, テンプレートフィルタリングにより楽器の個体差を, 位相トラッキングにより音高の揺らぎを吸収することで, 音色変化の問題に対処していた. ただし, テンプレートを各楽器で半音毎に用意しているものの, 音高が変化すると音色がどのように変化するかを効率的にモデル化する手法については扱われていなかった. 一般に楽器の音域は広く, たとえば, ピアノの音域は 7 オクター

ブにも渡る．そのため，高い音域と低い音域では音色が大きく異なり，音高を考慮しなければ，各楽器の全音域で適切に音源同定するのが困難となる．しかも楽器の音色は，通常，低域から高域にいくにしたがって連続的に変わることを，音源同定手法は考慮しなければならない．

本論文では，音高による音色変化を考慮する音源同定手法として，**F0 依存多次元正規分布**に基づく識別手法を提案する．これは，音高が物理量 (基本周波数) とほぼ 1 対 1 で対応することに着目して，音高による音色変化を基本周波数の関数として表現するものである．具体的には，音色空間 (楽器音の特徴空間) 上で，各楽器音データは，基本周波数によって平均が変化する多次元正規分布 (F0 依存多次元正規分布) に従うと仮定し，この分布のパラメータ推定法を提案する．そして，この分布を仮定した場合の識別関数をベイズ決定規則から定式化する．このように，特徴変動をその要因となる物理量の関数としてとらえるアプローチは，従来の研究 [6, 4, 23, 26, 29, 35, 13, 9] ではあまり議論されてこなかった．

以下，まず **2.** で F0 依存多次元正規分布を提案し，この分布を仮定した場合の識別関数をベイズ決定規則に基づいて定式化する．次に，**3.** で提案手法の処理の流れを述べ，**4.** で評価実験について述べる．さらに，**5.** でベイズ決定規則と k-NN 法を比較し，最後に **6.** でまとめをする．

1.2 F0 依存多次元正規分布

本論文の音源同定方式では，各楽器名がラベルづけられた楽器音の特徴ベクトルのデータベース (個々の特徴ベクトルを学習パターンと呼ぶ) に基づいて音源同定を行う．各楽器の学習パターンが多次元正規分布に従うと仮定し，多次元正規分布のパラメータを推定して各楽器の事後確率を計算する．そして，事後確率の最も高い楽器名を同定結果として出力する．ただし，学習パターンは，以下の理由により音高に依存する：

(1) 音高が低くなれば，発音体は大きくなる．発音体の質量が大きくなると慣性も大きくなり，発音の立ち上がりや減衰に，より多くの時間を要する [19].

(2) 音高が高くなると振動損失が大きくなるため，高次の高調波は発生されにくくなる [19].

(3) 一部の楽器では音高により発音体が異なり，各発音体は異なる材質からできている．

この問題に対する 1 つの解決法は，各楽器の学習パターンが音高ごとに異なる多次元正規分布に従うと仮定し，入力 (同定対象) と同じ音高の学習パターンのみを使って分布のパラメータを推定することである．しかし，分布のパラメータ推定には多くの学習パターンを必要とし，音高ごとに多くの学習パターンを用意するのは非現実的である．

本論文では，各楽器の学習パターンは，平均が基本周波数によって連続的に変化し，共分散行列が基本周波数に依存しない多次元正規分布に従うと仮定する．このように仮定することで，音高による音色変化を考慮しながら，全音域の学習パターンを 1 つの分布で表現することができる．本論文では，このように拡張された多次元正規分布を **F0 依存多次元正規分布**と呼ぶ．以下，F0 依存多次元正規分布のパラメータ (平均と共分散) の推定法を述べる．

1.2.1 代表値関数

基本周波数によって変化する分布の平均を，最小二乗法による関数近似で推定する (図 1.1)．この近似曲線を**代表値関数**と呼び，$\mu_i(f)$ と書く (i：楽器名)．ここでは，近似関数として 3 次関数を用いる．これは，複雑な音色変化を表現できることと，少ないデータからでも推定できることを両立できる次数を，予備実験により求めたものである．

1.2.2 F0 正規化共分散行列

F0 依存多次元正規分布における共分散行列の算出法を述べる．F0 依存多次元正規分布の共分散行列は，代表値関数からのちらばりの程度を表す．代表値関数は音高による音色変化を表すので，共分散行列は，音高以外の要因による

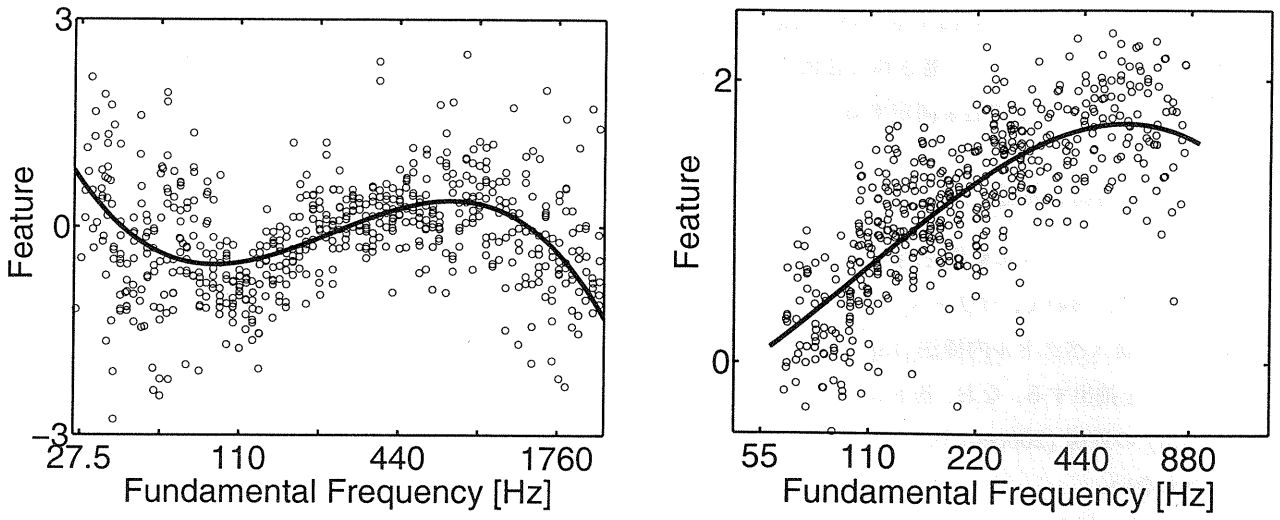


Figure 1.1: 代表値関数 (太字) の例. 左の図は線形近似では精度が不十分な例 (ピアノの第 4 軸) で, 右の図は音高による音色変化が特に顕著な例 (チェロの第 1 軸) である.

音色変化を表していると考えることができる. そこで, 音色空間を代表値関数で正規化することで音高による音色変化を除去してから, 共分散行列を求める. 本論文では, この F0 依存多次元正規分布における共分散行列を **F0 正規化共分散行列**と呼び, Σ_i と書く.

1.2.3 ベイズ決定規則による識別

ベイズ決定規則に基づいて識別関数を定式化する. 各楽器 ω_i の学習パターンが, F0 依存多次元正規分布に従うと仮定し, この分布の確率密度関数 $p(\boldsymbol{x}|\omega_i; f)$ を使って, パターン $\boldsymbol{x}$ が入力されたときの識別関数 $g_i(\boldsymbol{x}; f)$ を次式で定義する:

$$g_i(\boldsymbol{x}; f) = \log p(\boldsymbol{x}|\omega_i; f). \tag{1.1}$$

ここで, パラメータ f は入力パターン $\boldsymbol{x}$ の基本周波数で, 本研究で新たに導入されたものである. また, F0 依存多次元正規分布の確率密度関数は

$$p(\boldsymbol{x}|\omega_i; f) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f))\right\}$$

で与えられる. ここで, d は音色空間の次元数, D はマハラノビス距離であり, その定義は

$$D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) = (\boldsymbol{x} - \boldsymbol{\mu}_i(f))'\Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i(f))$$

で与えられる ($'$ は転置). この式を式 (1) に代入すると次の識別関数 $g_i(\boldsymbol{x}; f)$ が得られる:

$$g_i(\boldsymbol{x}; f) = -\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) - \frac{1}{2}\log|\Sigma_i| - \frac{d}{2}\log 2\pi.$$

基本周波数 f を発音できる楽器のみを対象として, この $g_i(\boldsymbol{x}; f)$ を最大にする楽器名, すなわち $k = \mathrm{argmax}_i\, g_i(\boldsymbol{x}; f)$ とおいたときの ω_k を同定結果として出力する.

1.3 処理の流れ

本章では, 提案手法の処理の流れを述べる. まず, 前処理としてスペクトログラムを作成し, 調波構造を推定する. 次に, 特徴抽出を行う. 抽出する特徴量は, 後で特徴空間の変形 (次元圧縮) をすることを前提に, 識別に有効と期待で

きるものをできるだけ多く抽出するという設計方針をもとに, 129 個定めた. その後, 主成分分析・線形判別分析により次元圧縮を行う. そして, 圧縮された音色空間上で各楽器のパターンが F0 依存多次元正規分布に従うと仮定し, ベイズ識別規則を用いて楽器名を同定する.

1.3.1 調波構造の推定

まず, 短時間フーリエ変換を用いてスペクトログラムを作成し (ハニング窓使用, 窓幅: 4096 点, 時間分解能: 10ms), 各時刻において, パワースペクトログラムの周波数方向の導関数の零交差からピーク抽出を行う. ピーク位置推定には, 複素スペクトル内挿法 [18] をハニング窓用に拡張したもの [14] を使用し, 抽出されたピークから調波構造 (30 次まで) を抽出する. なお, 基本周波数に関しては, 音高 (C4 など) を人手で与え, その音高に対応する周波数 (平均律で算出) の近傍 (200cent 以内) に存在するピークの周波数とする. また, 周波数とパワーはともに対数で表し, 正規化は行わない.

1.3.2 特徴抽出

次に示す 129 個の特徴量を抽出する. これらは, 先行研究 [23, 35] や楽器音響学・楽器物理学などの知見 [47, 19, 39, 1] を参考にしながら決定した.

(1) スペクトルに関する特徴

ここでは, 主に音の甲高さなどスペクトルの定常的な特徴を抽出する. そのため, 各高調波成分の周波数やパワー値は, その時間方向の中央値を用いる. 具体的には次に示す 40 個の特徴量を抽出する:

1 周波数重心 (各高調波成分のパワー値を重みとする周波数の重みつき平均),

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,

3~30 全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合計の割合 ($i = 2, 3, \cdots, 29$),

31 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合計の比,

32~40 音が鳴り続けている時間 (周波数成分全体のパワーがしきい値を越えている時間) に対して, その高調波成分の鳴り続けている時間 (パワー値が同じしきい値を越えている時間) が p% である高調波成分の個数 ($p = 10, 20, \cdots, 90$).

(2) パワーの時間変化に関する特徴

ここでは, パワーの時間変化に関する特徴を抽出する. 以下の特徴量 41 で, 大局的な音量変化 (通常, 音が減衰するか持続するか) を表し, 特徴量 42~75 で, より細かな音量変化を表す.

41 パワー包絡線の線形最小二乗法による近似直線の傾き,

42~58 発音開始直後 t 秒間のパワー包絡線の微分係数の中央値 ($t = 0.15, 0.20, \cdots, 0.95$),

59~75 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 ($t = 0.15, 0.20, \cdots, 0.95$).

(3) 各種変調の振幅と振動数

以下の変調の振幅と振動数を抽出する. ここで, 各種変動の振動数は導関数の零交差点数から, 振幅は, 十分に平滑化された信号と元の信号との差に対する四分位幅 (上位 25% と下位 25% の値を無視したときの最大値と最小値の差) からそれぞれ算出する. 平滑化には, Savitzky と Golay の 2 次多項式適合による平滑化法 [46] を使用する.

76 振幅変調の振幅,

77 振幅変調の振動数,

78 周波数変調の振幅,

79 周波数変調の振動数,

80 周波数重心の時間変化の振幅,

81 周波数重心の時間変化の振動数,

82~94 k 次のメル周波数ケプストラム係数 (MFCC) の時間変化の振幅 ($k = 1, 2, \cdots, 13$),

95~107 k 次の MFCC の時間変化の振動数 ($k = 1, 2, \cdots, 13$).

(4) 発音開始直後のピーク尖度に関する特徴

発音開始直後 150ms 間において, 各高調波成分のピーク周辺にどの程度非調波成分があるかを, 各高調波成分のピークの尖度から抽出する. まず, 発音開始時刻から 150ms までの各時刻のパワースペクトルから, 基音から 11 次倍音までの各高調波成分のピーク付近 (ピークの周波数を f[Hz] とすると, $0.75f$[Hz] から $1.5f$[Hz] までの範囲) の部分を切り出す. そして, 切り出された各ピーク付近がどの程度とんがっているかを 4 次モーメントから算出する. このとき, 非高調波成分が多く含まれていれば, 高調波成分のピークが非高調波成分に埋もれる形となるため, ピークの尖度は低くなり, 逆に, 非高調波成分があまり含まれていなければ, ピークの尖度は高くなる. そこで, 各高調波成分に対する各時刻のピーク尖度の時間方向の平均値 (特徴量番号108~118) と, 時間変化の振幅 (特徴量番号119~129) をそれぞれ抽出する.

1.3.3 主成分分析・線形判別分析による次元圧縮

上記の特徴量を平均が 0, 分散が 1 になるように正規化し, 主成分分析により次元を圧縮する. 累積寄与率 99% で, 129 次元から 79 次元に圧縮される.

次に, 線形判別分析によりさらに次元を圧縮する. 本論文では 19 種類の楽器を扱うので, 特徴空間は 18 次元に圧縮される. 線形判別分析は, クラス内分散・クラス間分散比を最大にする部分空間を求める手法で, 識別を考慮した次元圧縮法である. そのため, 主成分分析のみで同次元に圧縮するのに比べて高性能になると予測される. このことは, 後述の実験で確認する.

1.3.4 識別

2. で述べたように, 主成分分析・線形判別分析によって圧縮された 18 次元の特徴空間上で, 各クラスのパターンが F0 依存多次元正規分布に従うと仮定し, ベイズ決定規則を用いて楽器名を同定する.

1.4 評価実験

提案手法の有効性を確認するため, 評価実験を行う.

1.4.1 実験方法

実楽器の単音データベースとして, RWC 研究用音楽データベースの楽器音データベース RWC-MDB-I-2001[15] を使用する. これは, 50 種類の実楽器の単独発音を半音ごとに収録 (サンプリング周波数: 44.1kHz, 16 ビットリニ

Table 1.1: 使用した楽器音データベースの内訳

楽器番号	楽器名 (楽器記号)	楽器個体	音域	強さ	奏法	データ数 *
01	ピアノ (PF)	3	A0–C8			508
09	クラシックギター (CG)	3	E2–E5			696
10	ウクレレ (UK)	3	F3–A5			295
11	アコースティックギター (AG)	3	E2–E5	それ		666
15	バイオリン (VN)	3	G3–E7	れ		528
16	ビオラ (VL)	3	C3–F6	ぞ		472
17	チェロ (VC)	3	C2–F5	れ	通	558
21	トランペット (TR)	2	E3–A♯6	強	常	151
22	トロンボーン (TB)	3	A♯1–F♯5	・	の	262
25	ソプラノサックス (SS)	3	G♯3–E6	中	奏	169
26	アルトサックス (AS)	3	C♯3–A5	・	法	282
27	テナーサックス (TS)	3	G♯2–E5	弱	の	153
28	バリトンサックス (BS)	3	C2–A4	の	み	215
29	オーボエ (OB)	2	A♯3–G6	3		151
30	ファゴット (FG)	3	A♯1–D♯5	種		312
31	クラリネット (CL)	3	D3–F6	類		263
32	ピッコロ (PC)	3	D5–C8			245
33	フルート (FL)	2	C4–C7			134
34	リコーダー (RC)	3	C4–B6			160

* 無音検出による自動切り出しによって切り出された単音の個数.

ア量子化, モノラル) したもので, 各楽器音には, 原則 3 種類の楽器個体, 3 種類の音の強さ, 複数の奏法が含まれている.

このデータベースのうち, オーケストラで一般的に使用される楽器から, 打楽器, 収録時のノイズが大きいものなどを除いた 19 種類の楽器を使用する. 使用したデータ (総数 : 6247 個) の内訳を表 1.1に示す. 表 1.1のデータ全体を無作為に 10 等分し, クロスバリデーションを行って認識率を求める. すなわち, 10 個のグループそれぞれに対して, そのグループ以外のデータで学習してそのグループのデータで評価するという実験を繰り返して, 認識率の平均を求める.

楽器音を扱う場合, 個々の楽器の認識率だけでなく, 弦楽器, 金管楽器などのカテゴリーレベルの認識率も重要である. なぜなら, 実際の応用においてカテゴリーレベルの情報が分かるだけで有用な場面が多いからである. たとえば, ピアノソロ曲を検索する場面では, 音楽音響信号に擦弦楽器や管楽器などが含まれていることが分かれば, それだけで検索対象からはずすことができる. また, フルートとピアノのアンサンブル曲を自動採譜する場面で, 個々の楽器名を正しく同定できなくても, カテゴリーレベルで両者を区別することはできる.

本論文では, カテゴリーレベルの認識率を, 表 1.2に示す 8 つのカテゴリーを用いて算出する. この分類は, 楽器の

Table 1.2: 19 楽器の分類

カテゴリー	属する楽器
ピアノ	ピアノ
ギター	クラシックギター, ウクレレ, アコースティックギター
弦楽器	バイオリン, ビオラ, チェロ
金管楽器	トランペット, トロンボーン
サックス	ソプラノサックス, アルトサックス, テナーサックス, バリトンサックス
複簧楽器	オーボエ, ファゴット
クラリネット	クラリネット
無簧楽器	ピッコロ, フルート, リコーダ

発音機構や従来研究 [35, 9] に基づいて本研究で定義したものである (ただし, Eronen は, 複簧楽器とクラリネットを 1 つのカテゴリーに, 金管楽器とサックスを 1 つのカテゴリーにまとめた 6 カテゴリーを用いている [9]). まず, 同定対象を個々の楽器レベルで (すなわち楽器名を) 同定し, 同定結果と正解とがカテゴリーレベルで一致しているかどうかを表 1.2の分類に基づいて決定する. そして, 同定結果と正解とがカテゴリーレベルで一致した音響信号の個数から, カテゴリーレベルの認識率を算出する.

1.4.2 実験結果

通常の多次元正規分布を仮定して識別した場合 (音高による音色変化を考慮しない場合) と, F0 依存多次元正規分布を仮定して識別した場合 (提案手法) の両方の認識率を表 1.3に示す. 本論文で提案する F0 依存の処理を導入することで, 個々の楽器レベルで, 平均の認識率が 75.73% から 79.73% と 4.00% 改善された. これは, 音高による音色変化を考慮しない場合の誤認識全体を 1 とすると, その 16.48% を削減できたことを意味する. また, カテゴリーレベルでは, 平均の認識率は 88.20% から 90.65% と 2.45% 改善された. これは, 音高による音色変化を考慮しない場合の認識率全体を 1 とすると, その 20.67% を削減できたことを意味する.

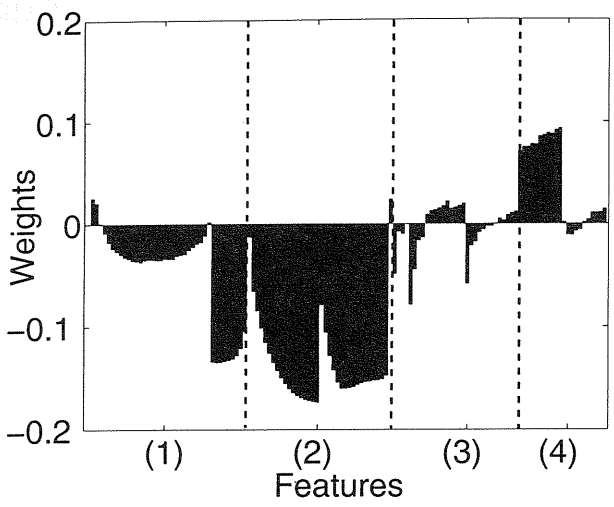
これらの結果が有意であることを t 検定 (片側検定) で示す. 各楽器における両手法の認識率の差を d_i $(i = 1, \cdots, n)$ とすると, 検定統計量 t_0 は,

$$t_0 = \frac{|\bar{d}|}{\sqrt{\sum_i (d_i - \bar{d})^2 \Big/ n(n-1)}}$$

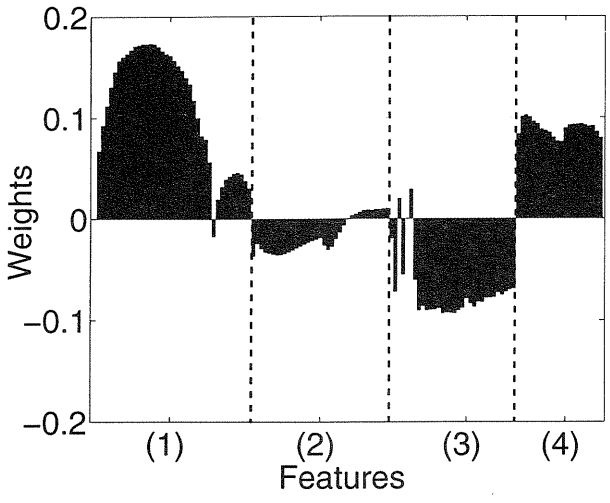
で与えられる. ここで, $\bar{d}$ は $d_1, \cdots, d_n$ の平均値である. この統計検定量は, 個々の楽器レベルとカテゴリーレベルでそれぞれ 5.4781, 3.9482 で, ともに有意水準 0.05% (棄却域 : $(3.9217, \infty)$) で有意である.

1.4.3 主成分分析に関する考察

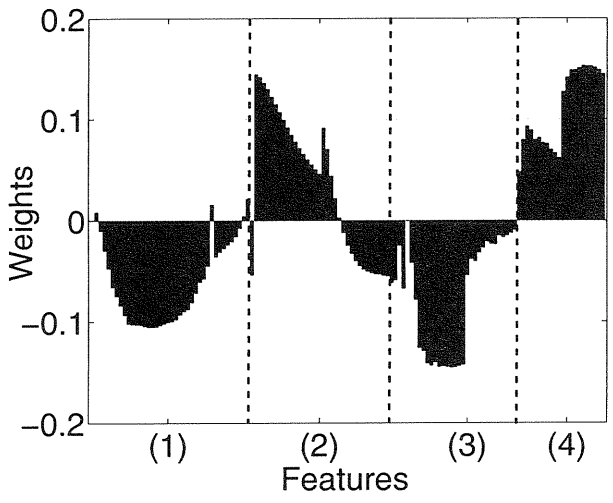
各主成分の重みの一部を図 1.2に示す. 図 1.2(a), (b) から, 第 1 主成分は高調波成分の個数 (32~40) とパワーの時間変化 (41~75) を, 第 2 主成分は高調波成分に関する定常的特徴 (2~30) を総合的に表していると考えられる. 従来より, 音色を規定する要因としてスペクトルに関する定常的特徴とパワーの時間変化に関する特徴が重要である



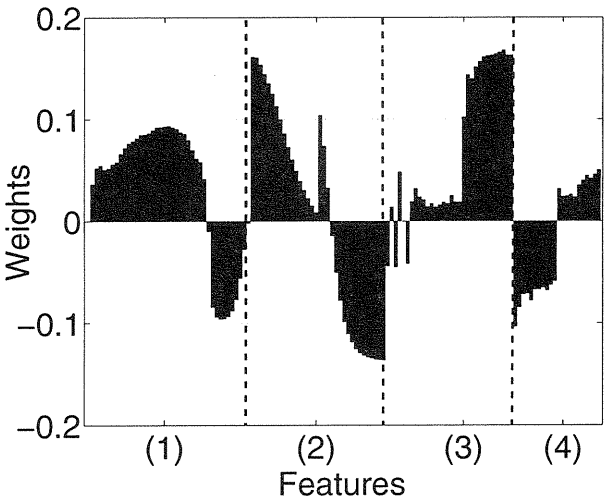
(a) 第 1 主成分



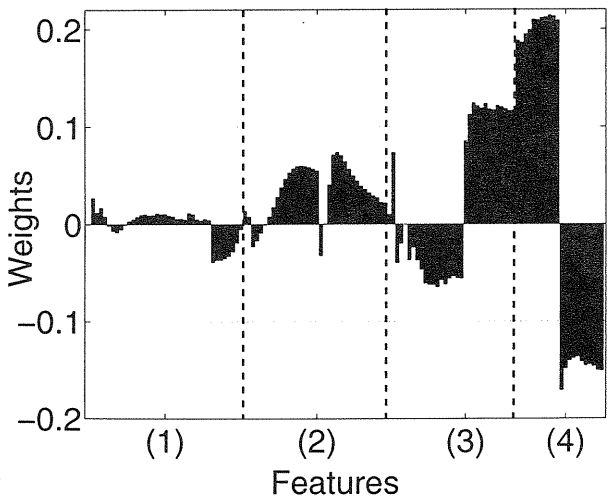
(b) 第 2 主成分



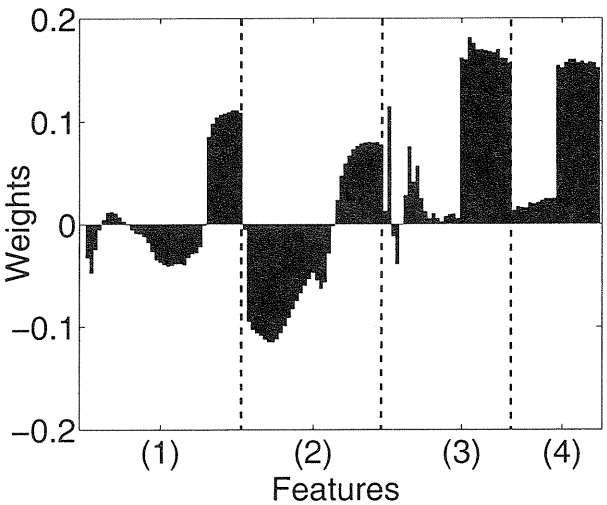
(c) 第 3 主成分



(d) 第 4 主成分



(e) 第 5 主成分



(f) 第 6 主成分

注 図の横軸は, **3.2** の 129 個の特徴量に対応し, 図中の (1) ~ (4) は, **3.2** の特徴量の説明における (1) ~ (4) に対応する.

Figure 1.2: 主成分分析による各主成分の重み値

と言われており[1], このことを裏付ける結果になったといえる. 第3主成分はMFCCや発音開始直後のピーク尖度の時間変化の振幅 (82~94, 119~129), 第4主成分はMFCCの時間変化の振動数 (95~107), 第5主成分はMFCCと発音開始直後のピーク尖度 (95~129), 第6主成分はMFCCの時間変化の振動数 (95~107) と発音開始直後のピーク尖度の時間変化の振動数 (119~129) などを総合的に表していることが分かる.

ここで, 発音開始直後のピーク尖度に関する特徴量 (95~129) が, 第3主成分, 第5主成分, 第6主成分と多くの主成分に現れている. この特徴量は, 高調波成分周辺の非調波成分の多さをモデル化したものである. 楽器音の非調波成分のモデル化は, 従来からその必要性が認識されながらも[2][1], ほとんど考慮されてこなかった. これらの特徴量が多くの主成分に現れたことは, 音楽情景分析において非調波成分を適切に扱う必要があることを示唆している.

1.4.4 線形判別分析に関する考察

主成分分析と線形判別分析によって特徴空間変形を行った際の重み (基底ベクトル) の一部を表1.4に示す. 表から以下の考察が得られる:

(1) スペクトルに関する特徴量について

第9軸に2 (全高調波成分のパワー値の合計に対する基本成分のパワー値の割合) の重みが0.3586と高く現れた他は, 32~40 (音が鳴り続けている時間に対して, その高調波成分の鳴り続けている時間がp%以上である高調波成分の個数) の重みが, 第5軸, 第7軸, 第9軸, 第10 ~ 13軸で高かった (絶対値で0.2721 ~ 0.4363). 音色を規定する要因としてスペクトルに関する特徴量が重要であることは以前から知られており [41, 42], 上記の結果は, これを裏付ける結果になったといえる.

(2) パワーの時間変化・各種変調について

41 (パワー包絡線の線形最小二乗法による近似直線の傾き) の重みが, 第3軸, 第4軸で高く (それぞれ0.5977, −0.2578), 42 (発音開始直後150ms間のパワー包絡線の微分係数の中央値) の重みが第10軸で−0.3200, 73~75 (最大パワー値と, 発音開始から$t(= 0.85, 0.90, 0.95)$秒後のときのパワー値の比) の重みが第1軸で0.2701 ~ 0.3926であった. また, 76~79 (振幅変調, 周波数変調の振幅／振動数) は, 第1軸, 第2軸, 第4 ~ 6軸, 第8軸, 第12軸, 第14軸と, 多くの軸で大きな重みが現れた (絶対値で0.2755 ~ 0.4425). これらから, パワーの時間変化や各種変調などの動的特徴が識別に効果的であるといえる. 実際, 人間の音色知覚においても, このような動的特徴が重要であることが知られており [1, 41, 42], 楽器音の音響信号を逆転再生すると音源同定能力が低下するという実験結果 [2] などもこのことを裏付けている.

(3) 発音開始直後のピーク尖度に関する特徴について

108, 109 (発音開始直後の基音成分／2次高調波成分のピーク尖度) が, 第3 ~ 8軸, 第10軸, 第11軸と, 多くの軸で大きな重みが現れた (絶対値で0.2607 ~ 0.5400). これは, **4.3** でも述べたように, 音楽情景分析において非調波成分を適切に扱う必要があることを示唆している.

1.4.5 実験結果に関する考察

実験結果について考察する.

---

[1]たとえば, 古典的な音合成方式では, 周波数エンベロープオシレータにより定常的なスペクトルを制御し, パワーエンベロープオシレータによりパワーの時間変化を作り出している [43].

[2]たとえば, 安藤は, 非調波成分を「雑音的成分の混在」と称し, 「楽器音の それらしさ を構成する重要な因子」と述べている [1].

(1) ピアノの性能改善が顕著 (74.21% から 83.27%, 9.06% の改善) である. これは, ピアノの音域が広く, 音高による音色変化が顕著に現れるからと考えられる. 楽器音響の分野では, ピアノの音色は,

- 低音ほど倍音が豊富である,
- 低音ほど弦が太く, 弦の質量が大きくなるため, 振幅の時間変化が緩やかになる,
- 低音では 1 本, 中音では 2 本, 高音では 3 本の弦が 1 つの鍵盤に対して使われており, 中音・高音では調律の微妙なずれにより うなり が発生する,

ということが知られている [1, 19]. 実際, 主成分分析・線形判別分析で得られた特徴空間において, 第 2 軸, 第 4 軸, 第 14 軸に音高による特徴変動が顕著にみられた. これらの軸は, 41 パワー包絡線の線形最小自乗法による近似直線の傾き, 76 ~ 79 振幅変調・周波数変調の振幅／振動数, などの重みが大きく, 上記と一致する部分がみられる.

(2) クラシックギター, ウクレレ, リコーダーでは, 提案手法の有効性を確認することはできなかった. これは, 元々の認識率が 90% を越えており, 改善の余地が小さかったからと考えられる.

(3) ギター, 弦楽器のカテゴリーレベルの認識率が, 他の楽器に比べ高かった (94.92% ~ 99.05%). これは, 管楽器は種類が多く, いくつかのカテゴリーにまたがって存在するのに対し, ギターや弦楽器は, 他のカテゴリーに発音機構の似た楽器が存在しないためと考えられる.

(4) サックスは, カテゴリーレベルの認識率 (77.66% ~ 92.16%) に比べ, 個々の楽器レベルの認識率 (47.87% ~ 73.95%) が低かった. これは, サックス内の個々の楽器の音色が非常に似ているためと考えられる. 実際, これらは人間でも識別するのが難しく, 文献 [35] によれば, 被験者が聴いた音の楽器名を 27 個の楽器名が書かれたリストから選ぶ, という実験で正しく認識できた人 (音楽経験者) は, ソプラノサックスで 7.1%, アルトサックスで 28.6% と少なかった. ただし, この聴取実験では 10 秒程度の旋律の抜粋を用いており, 本実験の結果と直接比較することはできない.

(5) リコーダーの認識率が, 個々の楽器レベルで 0.63%, カテゴリーレベルで 1.25% 下がり, オーボエのカテゴリーレベルの認識率が 0.67% 下がった. しかし, リコーダーやオーボエはデータ数が少なく (それぞれ 160 個, 151 個), リコーダーの個々の楽器レベルで 1 個, カテゴリーレベルで 2 個, オーボエのカテゴリーレベルで 1 個, 誤認識が増えたに過ぎない.

1.5 k-NN 法との比較

本章では, F0 依存多次元正規分布を仮定してベイズ決定規則を用いた場合 (提案手法) と他の手法 (ノン・パラメトリックな手法) を用いた場合, および, 線形判別分析を用いた場合と用いなかった場合とで, 認識率を比較する. なお, ノン・パラメトリックな手法としては, k-NN 法 ($k = 3$) を取り上げた.

実験方法は **4.** と同じく, 表 1.1のデータ (総数： 6247 個) を使ってクロスバリデーションを行う. 実験結果 (表 1.5) から以下の知見が得られる：

(1) 平均の認識率で, 主成分分析・線形判別分析で次元圧縮した後に F0 依存多次元正規分布を仮定してベイズ決定規則を用いた場合 (提案手法) が最も高かった (79.73%). また, この方法は楽器毎の性能の偏りも最も小さかった.

(2) トランペット, ソプラノサックス, テナーサックス, オーボエ, フルートについて, 主成分分析で 79 次元に圧縮してベイズ決定規則を用いた場合の認識率がいずれも低い (30.07% ~ 48.52%) のに対して, 主成分分析による

次元圧縮を 18 次元にすると，認識率に大幅な改善が見られた (66.84% ～ 84.33%)．これは，79 次元正規分布のパラメータを推定するのに十分な数の学習データがなかったため (いずれも 170 個未満) と考えられる．しかし，全体では 62.11% から 66.50% に改善されたにすぎない．これは，主成分分析が識別を考慮した次元圧縮ではないため，識別に効果的な特徴が落とされる可能性があるからと考えられる．それに対し，線形判別分析はクラス内分散・クラス間分散比最大化に基づく識別を考慮した次元圧縮法で，実際に認識率は 79.73% と大幅に改善された．

(3) 本論文では，線形判別分析のみを用いて次元圧縮した場合については実験しなかった．これは，線形判別分析で用いる逆行列は，特徴空間に相関性の高い軸が含まれていると誤差が大きくなるため，線形判別分析による部分空間が，正常に算出されないためである．主成分分析は，特徴空間の次元を圧縮するだけでなく，各軸が直交するように空間を変形する．そのため，線形判別分析を適用する前に，主成分分析を用いて各軸を直交化することが有効である．

1.6 おわりに

本論文では，音高による音色変化を考慮する音源同定手法として，F0 依存多次元正規分布に基づく識別手法を提案した．この手法は，各楽器音データが，基本周波数によって平均が変化する多次元正規分布に従うと仮定し，基本周波数によって変化する平均を関数近似により求めるものである．実験の結果，音高による音色変化を考慮しない場合の誤認識全体のうち，個々の楽器レベルで 16.48%，カテゴリーレベルで 20.67% の誤認識を削減することができた．

本論文で提案した F0 依存多次元正規分布は，ベイズ決定規則への応用のみに限定されるものではない．今後は，この枠組みを応用して，より高性能な識別手法の設計に取り組むとともに，より多くの楽器に対応できるよう，他の特徴量の導入も検討する．さらに，混合音への適用などにも取り組んでいく予定である．

Table 1.3: 実験結果 (通常の多次元正規分布の場合の認識率と F0 依存多次元正規分布の場合の認識率)

楽器	個々の楽器レベル			カテゴリーレベル		
記号	Normal	F0-dpt	差	Normal	F0-dpt	差
PF	74.21%	83.27%	+9.06%	74.21%	83.27%	+9.06%
CG	90.23%	90.23%	±0.00%	97.27%	97.13%	−0.14%
UK	97.97%	97.97%	±0.00%	97.97%	98.31%	+0.34%
AG	81.23%	83.93%	+2.70%	94.89%	95.65%	+0.76%
VN	69.70%	73.67%	+3.97%	98.86%	99.05%	+0.19%
VL	73.94%	76.27%	+2.33%	93.22%	94.92%	+1.70%
VC	73.48%	78.67%	+5.19%	95.16%	96.24%	+1.08%
TR	73.51%	82.12%	+8.61%	76.82%	85.43%	+8.61%
TB	76.72%	84.35%	+7.63%	85.50%	89.69%	+4.19%
SS	56.80%	65.89%	+9.09%	73.96%	80.47%	+6.51%
AS	41.49%	47.87%	+6.38%	73.76%	77.66%	+3.90%
TS	64.71%	66.01%	+1.30%	90.20%	92.16%	+1.96%
BS	66.05%	73.95%	+7.90%	81.40%	86.05%	+4.65%
OB	71.52%	72.19%	+0.67%	75.50%	74.83%	−0.67%
FG	59.61%	68.59%	+8.98%	64.74%	71.15%	+6.41%
CL	90.69%	92.07%	+1.38%	90.69%	92.07%	+1.38%
PC	77.56%	81.63%	+4.07%	89.39%	90.20%	+0.81%
FL	81.34%	85.07%	+3.73%	82.09%	85.82%	+3.73%
RC	91.88%	91.25%	−0.63%	92.50%	91.25%	−1.25%
平均	75.73%	79.73%	+4.00%	88.20%	90.65%	+2.45%

Normal: 通常の多次元正規分布を仮定した場合

F0-dpt: F0 依存多次元正規分布を仮定した場合 (提案手法)

Table 1.4: 特徴空間変形における基底ベクトルの一部

	特徴量と重み値
第 1 軸	73 (0.2701), 74 (0.3220), 75 (0.3926), 79 (−0.3204), 81 (0.2559)
第 2 軸	40 (−0.2721), 76 (0.4425), 78 (0.3554), 82 (−0.2771),
第 3 軸	41 (0.5977), 109 (0.2607)
第 4 軸	41 (−0.2578), 79 (−0.2917), 109 (0.2944)
第 5 軸	40 (0.4286), 78 (0.3219), 108 (0.5400)
第 6 軸	76 (−0.2755), 108 (−0.4529)
第 7 軸	40 (0.3974), 108 (−0.4576)
第 8 軸	76 (0.3378), 85 (0.2614), 108 (−0.4541)
第 9 軸	2 (0.3586), 40 (−0.2783), 84 (0.4525)
第 10 軸	40 (0.2887), 42 (−0.3200), 108 (−0.3292), 109 (0.4508)
第 11 軸	32 (0.4363), 36 (−0.2837), 109 (−0.2732)
第 12 軸	39 (0.2794), 78 (0.3174), 81 (0.2704)
第 13 軸	40 (0.3521), 120 (−0.2522)
第 14 軸	76 (−0.3484), 77 (0.4201)

Table 1.5: **5.** の実験結果 (k-NN 法とベイズ決定規則との認識率の比較；個々の楽器レベルの認識率のみ)

楽器	k-NN 法 ($k=3$)			ベイズ決定規則		
記号	PCA1	PCA2	LDA	PCA1	PCA2	LDA
PF	53.94%	46.46%	63.39%	55.91%	59.06%	83.27%
CG	79.74%	77.16%	75.72%	98.28%	97.27%	90.23%
UK	94.58%	92.54%	97.63%	67.12%	80.00%	97.97%
AG	95.05%	92.79%	97.00%	19.97%	44.14%	83.93%
VN	47.73%	46.02%	45.83%	89.58%	84.47%	73.67%
VL	55.93%	54.24%	61.86%	71.19%	79.24%	76.27%
VC	86.20%	85.84%	84.23%	45.16%	30.82%	78.67%
TR	36.42%	38.41%	47.02%	41.72%	72.85%	82.12%
TB	70.99%	54.58%	77.86%	75.19%	78.24%	84.35%
SS	23.08%	14.20%	24.85%	48.52%	66.86%	65.89%
AS	37.59%	29.79%	40.43%	72.70%	41.84%	47.84%
TS	62.09%	66.01%	68.63%	30.07%	61.44%	66.01%
BS	68.84%	67.91%	66.98%	55.35%	54.42%	73.95%
OB	47.68%	48.34%	49.01%	43.71%	81.46%	72.19%
FG	64.10%	65.06%	74.36%	40.38%	30.12%	68.59%
CL	93.45%	87.93%	93.10%	95.51%	93.45%	92.07%
PC	84.08%	84.90%	84.08%	63.27%	58.37%	81.63%
FL	88.06%	72.39%	94.03%	35.82%	84.33%	85.07%
RC	97.50%	93.75%	97.50%	85.00%	96.25%	91.25%
平均	70.27%	66.98%	72.53%	62.11%	66.50%	79.73%

PCA1：主成分分析のみを用いて 79 次元に圧縮した場合.

PCA2：主成分分析のみを用いて 18 次元に圧縮した場合.

LDA：主成分分析を用いて 79 次元に圧縮し，さらに線形判別分析で 18 次元に圧縮した場合.

Table 1.6: **3.** の実験結果

	Indivisual-level						Category-level					
	k-NN ($k=3$)			Mahalanobis distance			k-NN ($k=3$)			Mahalanobis distance		
	PCA1	PCA2	LDA	PCA1	PCA2	LDA	PCA1	PCA2	LDA	PCA1	PCA2	LDA
PF	53.94%	46.46%	63.39%	96.26%	93.70%	74.21%	53.94%	46.46%	63.39%	96.26%	93.70%	74.21%
CG	79.74%	77.16%	75.72%	68.53%	56.18%	90.23%	99.14%	98.28%	99.86%	74.71%	61.21%	97.27%
UK	94.58%	92.54%	97.63%	84.75%	88.47%	97.97%	99.32%	99.66%	100.00%	86.78%	89.15%	97.97%
AG	95.05%	92.79%	97.00%	72.97%	54.65%	81.23%	96.10%	94.74%	98.80%	76.28%	57.96%	94.89%
VN	47.73%	46.02%	45.83%	86.36%	71.59%	69.70%	96.02%	94.70%	97.92%	97.35%	92.61%	98.86%
VL	55.93%	54.24%	61.86%	37.92%	35.59%	73.94%	84.32%	86.02%	90.04%	88.35%	73.73%	93.22%
VC	86.20%	85.84%	84.23%	77.60%	79.21%	73.48%	92.83%	92.83%	93.91%	93.73%	88.35%	95.16%
TR	36.42%	38.41%	47.02%	2.65%	50.99%	73.51%	47.68%	45.03%	64.90%	3.97%	51.66%	76.82%
TB	70.99%	54.58%	77.86%	53.82%	68.32%	76.72%	70.99%	54.58%	78.63%	53.82%	70.23%	85.50%
SS	23.08%	14.20%	24.85%	2.96%	24.26%	56.80%	51.48%	50.30%	63.31%	61.54%	71.60%	73.96%
AS	37.59%	29.79%	40.43%	71.63%	56.38%	41.49%	64.89%	63.12%	71.28%	78.37%	76.95%	73.76%
TS	62.09%	66.01%	68.63%	15.69%	42.48%	64.71%	75.82%	82.35%	88.24%	86.93%	91.50%	90.20%
BS	68.84%	67.91%	66.98%	53.95%	72.56%	66.05%	71.16%	71.63%	73.02%	68.37%	83.72%	81.40%
OB	47.68%	48.34%	49.01%	2.65%	35.10%	71.52%	58.94%	60.93%	67.55%	2.65%	75.50%	75.50%
FG	64.10%	65.06%	74.36%	91.35%	89.74%	59.61%	65.38%	66.99%	75.96%	91.35%	89.74%	64.74%
CL	93.45%	87.93%	93.10%	51.38%	53.10%	90.69%	93.45%	87.93%	93.10%	51.38%	53.10%	90.69%
PC	84.08%	84.90%	84.08%	90.61%	86.12%	77.56%	95.10%	93.88%	94.29%	90.61%	86.12%	89.39%
FL	88.06%	72.39%	94.03%	0.74%	9.70%	81.34%	94.78%	88.06%	97.01%	33.58%	27.61%	82.09%
RC	97.50%	93.75%	97.50%	5.63%	17.50%	91.88%	99.38%	97.50%	99.38%	6.25%	20.00%	92.50%
Ave.	70.27%	66.98%	72.53%	62.94%	62.37%	75.73%	79.88%	81.37%	87.55%	68.74%	74.10%	88.20%

第 2 章

Systematic Generation of Musical Sound Ontology by Pitch-dependent Musical Instrument Identification

Musical sound ontologies are essential in annotation of musical instruments in musical sound archives and their retrieval by specifying musical instruments, because perceptual categorization of musical instruments is not based on sound features.

Musical instrument ontologies are essential for annotation and retrieval of musical sound archives by specifying musical instruments. However systematic classification of musical instrument sounds is not easy, because taxonomic classification may differ from perceptual classification at some frequency regions. Therefore, ontologies for musical sounds are required to resolve ambiguities caused by such differences. In this paper, we present systematic construction of musical instrument ontology by musical instrument identification that exploits the *pitch dependency* of timbre, which has not been fully exploited so far. This dependency is represented by the *F0-dependent multivariate normal distribution*, whose mean is represented by a cubic polynomial of fundamental frequency (F0). The F0-dependent mean function represents the pitch dependency, while the F0-normalized covariance represents the non-pitch dependency. Musical sounds are first analyzed by the F0-dependent multivariate normal distribution, and then identified by the Bayes decision rule. Experimental results of identifying 6,247 solo tones of 19 musical instruments by 10-fold cross validation showed that the proposed method improved the recognition rate at individual-instrument level from 75.73% to 79.73%, and the recognition rate at category level from 88.20% to 90.65%. Based on these results, musical sound ontology is constructed by using the C5.0 decision tree program.

2.1 Introduction

Musical instrument ontologies are essential for annotation and retrieval of musical sound archives or multimedia database by specifying musical instruments. The approaches to musical instrument classification can be classified into two ways [16]. One is *perceptual classification*, which may be obtained by psychoacoustical studies of perceptual similarity of human subjects [22, 48]. The other is *taxonomic classification*, which may be obtained by judgments of sounding mechanisms and types of musical instruments done by skilled people. The problem of systematic classification is that perceptual and taxonomic classification may differ at frequency regions. To bridge this kind of difference, ontologies for musical instrument sounds are required.

The way of constructing musical instrument ontologies should be consistent with musical instrument iden-

tification, because this identification is an important subtask for many applications including computational auditory scene analysis [44] and multimedia retrieval as well as for reducing ambiguities in systematic music transcription. Musical sound ontology has been usually given in advance [35, 38], and its systematic construction has not been reported yet.

Although there are considerable works on musical instrument identification [20], its difficulties reside in the fact that some features are affected by individual instruments and pitch. In particular, timbres of musical instruments are obviously affected by the pitch due to their wide range of pitch, e.g., seven octaves covered by piano. In addition, database from which musical instrument ontology is constructed should include more than one individual for each instrument, because each individual instrument, for example, for pianos, Steinway & Sons, Bechstein, Bosendörfer, Yamaha, and Kawai sounds quite different.

To attain high performance of musical instrument identification, it is indispensable to cope with this *pitch dependency* of timbre. Most studies on musical instrument identification, however, have not dealt with the pitch dependency [5, 11, 13, 26, 35]. Martin used 31 features including spectral and temporal features with hierarchical classification and attained about 70% of identification by the benchmark of 1,023 solo tones of 14 instruments. He pointed out the importance of the pitch dependency, but left it as future work [35]. Eronen *et al.* used spectral and temporal features as well as cepstral coefficients used by Brown [5] and attained about 80% of identification by the benchmark of 1,498 solo tones of 30 instruments [11]. They treated the pitch as one element of feature vectors, but did not cope with the pitch dependency. Kashino *et al.* also treated the pitch similarly in the automatic music transcription system [26]. They also coped with the difference of individual instruments, but did not deal with the pitch dependency [27].

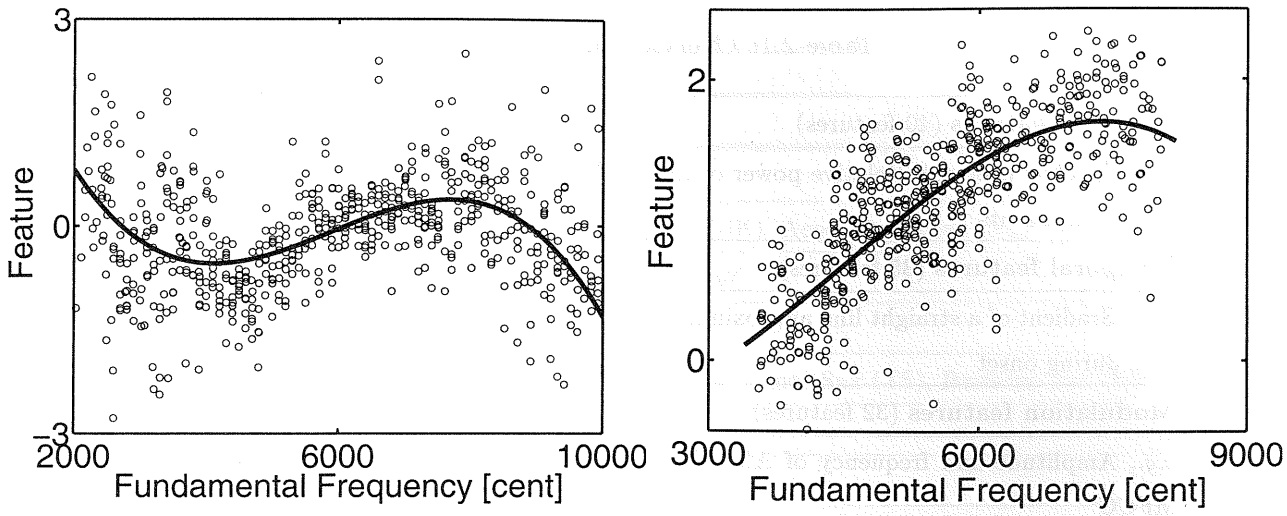
To capture the pitch dependency of timbre in musical instrument identification, a basic vector of features is represented by an *F0-dependent multivariate normal distribution*, the mean of which is represented by a function of fundamental frequency (F0) [33]. This *F0-dependent mean function* represents the pitch dependency of each feature, while the *F0-normalized covariance* represents the non-pitch dependency.

In this paper, the F0-dependent musical instrument identification and musical instrument ontology constructed by the C5.0 decision tree program [40] are reported. The rest of this paper is organized as follows: Section 2 proposes the F0-dependent multivariate normal distribution, and Section 3 describes the algorithm of musical instrument identification. Sections 4 and 5 report the experimental results. Section 6 applies the results to construct musical sound ontology by using the C5.0 decision tree program, and finally Section 7 concludes this paper.

2.2 F0-dependent Multivariate Normal Distribution

2.2.1 Pitch and Non-pitch Dependencies

The distribution of tone features in the feature space is represented by an *F0-dependent multivariate normal distribution* with two parameters: the *F0-dependent mean function* and *F0-normalized covariance*. The reason why the mean of the distribution is approximated as a function of F0, that is, an F0-dependent mean function is that tone features at different pitches have different positions (means) of distributions in the feature space. In this paper, the F0-dependent mean function for each musical instrument ω_i, $\mu_i(f)$, is approximated as a cubic polynomial by using the least squares method. For example, piano's fourth basic vector of features and cello's first basic vector are depicted in Figure 2.1 (a) and (b), respectively.



(a) Piano's 4th basic vector of features.

(b) Cello's first basic vector of features.

Figure 2.1: Examples of F0-dependent mean functions.

On the other hand, the non-pitch dependency of each feature is represented by the *F0-normalized covariance*. Since the F0-dependent mean function represents the mean of features, the covariance obtained by subtracting the mean from each feature eliminates the pitch dependency of features. For each musical instrument ω_i, the F0-normalized covariance Σ_i is defined as follows: $\Sigma_i = \frac{1}{n_i}\sum_{\boldsymbol{x}\in\chi_i}(\boldsymbol{x}-\boldsymbol{\mu}_i(f_{\boldsymbol{x}}))(\boldsymbol{x}-\boldsymbol{\mu}_i(f_{\boldsymbol{x}}))'$, where $'$ is the transposition operator, χ_i and n_i are the set of the training data of the instrument ω_i and its total number, respectively. $f_{\boldsymbol{x}}$ denotes the F0 of the data $\boldsymbol{x}$.

2.2.2 Features for Musical Instrument Identification

We used spectral, temporal, and modulation features as well as non-harmonic component feature resulting in 129 features in total listed in Table 2.1. The features except the non-harmonic component features are determined by consulting the literatures [35, 11, 25]. The non-harmonic component features are original and have not been used in the literature. We incorporated features as many as possible, since the feature space is transformed to a lower-dimensional space.

Each musical instrument sound sampled by 44.1 kHz with 16 bits are first analyzed by STFT (short time Fourier transform) with Hanning windows (4096 points) for every 10 ms, and spectral peaks are extracted from the power spectrum. Then, the F0 and the harmonic structure are obtained from these peaks.

The number of dimensions of the feature space is reduced by principal component analysis (PCA): the 129-dimensional space is reduced to a 79-dimensional space with the proportion value of 99%. It is further reduced to the minimum dimension by linear discriminant analysis (LDA). In this paper, the space is reduced to an 18-dimensional space, since we deal with 19 instruments.

2.3 A Discriminant Function based on the Bayes Decision Rule

Once pitch and non-pitch dependencies of feature vectors are represented, the Bayes decision rule is applied to identify the musical instrument or category of instruments. The discriminant function $g_i(\boldsymbol{x}; f)$ for the musical instrument ω_i is defined by

Table 2.1: Overview of 129 features.

(1)	**Spectral features** (40 features)
	e.g., Spectral centroid, Relative power of the fundamental component, Relative power in odd and even components
(2)	**Temporal features** (35 features)
	e.g., Gradient of a straight line approximating power envelope, Average differential of power envelope during onset
(3)	**Modulation features** (32 features)
	e.g., Amplitude and frequency of AM, FM, modulation of spectral centroid and modulation of MFCC
(4)	**Non-harmonic component features** (22 features)
	e.g., Temporal mean of kurtosis of spectral peaks of each harmonic component (Their values become lower as sounds contain more non-harmonic components.)

$$g_i(\boldsymbol{x}; f) = \log p(\boldsymbol{x}|\omega_i; f) + \log p(\omega_i; f), \tag{2.1}$$

where $\boldsymbol{x}$ is an input data, $p(\boldsymbol{x}|\omega_i; f)$ is a probability density function (PDF) of this distribution and $p(\omega_i; f)$ is a priori probability of the instrument ω_i.

The PDF of this distribution is defined by

$$p(\boldsymbol{x}|\omega_i; f) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f))\right\}, \tag{2.2}$$

where d is the number of dimensions of the feature space and D^2 is the squared Mahalanobis distance defined by

$$D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) = (\boldsymbol{x} - \boldsymbol{\mu}_i(f))'\Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i(f)).$$

Substituting equation (2) into equation (1), thus, generates the discriminant function $g_i(\boldsymbol{x}; f)$ as follows:

$$g_i(\boldsymbol{x}; f) = -\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) - \frac{1}{2}\log|\Sigma_i| - \frac{d}{2}\log 2\pi + \log p(\omega_i; f).$$

The name of the instrument that maximizes this function, that is ω_k satisfying $k = \mathrm{argmax}_i\, g_i(\boldsymbol{x}; f)$, is determined as the result of musical instrument identification.

The a priori probability $p(\omega_i; f)$ represents whether the pitch range of the instrument ω_i includes f, that is,

$$p(\omega_i; f) = \begin{cases} 1/c & (\text{if } f \in R_i) \\ 0 & (\text{if } f \notin R_i) \end{cases}$$

where R_i is the pitch range of the instrument ω_i, and c is the normalizing factor to satisfy $\sum_i p(\omega_i; f) = 1$.

2.4 Musical Instrument Identification

2.4.1 Experimental Conditions

Musical instrument identification is performed not only at individual instrument level but also at category level to evaluate the improvement of recognition rates by the proposed method based on the F0-dependent

Table 2.2: Categorization of 19 instruments.

CATEGORY	Instruments (abbreviation)
PIANO	Piano (PF)
GUITARS	Classical Guitar (CG), Ukulele (UK), Acoustic Guitar (AG)
STRINGS	Violin (VN), Viola (VL), Cello (VC)
BRASS	Trumpet (TR), Trombone (TB)
SAX	Soprano Sax (SS), Alto Sax (AS), Tenor Sax (TS), Baritone Sax (BS)
DOUBLE REEDS	Oboe (OB), Faggoto (FG)
CLARINET	Clarinet (CL)
AIR REEDS	Piccolo (PC), Flute (FL), Recorder (RC)

multivariate normal distribution. The recognition rate was obtained by 10-fold cross validation. We compared the results by the method using usual multivariate normal distribution (called *baseline*) with those by the method using the proposed F0-dependent multivariate normal distribution (called *proposed*).

The benchmark used for evaluation is a subset of the large musical instrument sound database RWC-MDB-I-2001 developed by the former RWC [15]. This subset summarized in Table 2.3 was selected by the quality of the recorded sounds and consists of 6,247 solo tones of 19 orchestral instruments. All data are sampled by 44.1 kHz with 16 bits.

The categories of musical instruments summarized in Table 2.2 are determined based on our musical instrument ontology obtained by C5.0, which is described later. This categorization is quite similar to that of existing studies [35, 11].

The category of instruments is useful for some applications including music retrieval. For example, when a user tries to find a piece of piano solo on a music retrieval system, the system can reject pieces containing instruments of different categories, which can be judged without identifying individual instrument names.

2.4.2 Results of Musical Instrument Identification

Table 2.4 summarizes the recognition rates by both the *baseline* and *proposed* methods. The proposed F0-dependent method improved the recognition rates at individual-instrument level from 75.73% to 79.73% and at category level from 88.20% to 90.65% in average. It also reduced recognition errors by 16.48% and 20.67% in average at individual-instrument and category levels, respectively.

The observation of these experimental results is summarized below:

Improvement by the pitch dependency

The recognition rates of six instruments (Piano (PF), Trumpet (TR), Trombone (TB), Soprano Sax (SS), Basitone Sax (BS), and Faggoto (FG)) were improved by more than 7%. In particular, the recognition rate for pianos was improved by 9.06%, and its recognition errors were reduced by 35.13%. This big improvement was attained, since their pitch dependency is salient due to their wide range of pitch.

Improvement by the category level identification

Table 2.3: Contents of the database used in this paper.

Instrument name	Abbrev.	pitch range	# of tones	# of individuals	Intensity	Articulation
Piano	PF	A0–C8	508			
Classical Guitar	CG	E2–E5	696			
Ukulele	UK	F3–A5	295			
Acoustic Guitar	AG	E2–E5	666	3		
Violin	VN	G3–E7	528		Forte,	
Viola	VL	C3–F6	472			
Cello	VC	C2–F5	558			
Trumpet	TR	E3–A♯6	151	2		normal
Trombone	TB	A♯1–F♯5	262		normal,	
Soprano Sax	SS	G♯3–E6	169			
Alto Sax	AS	C♯3–A5	282	3		only
Tenor Sax	TS	G♯2–E5	153		&	
Baritone Sax	BS	C2–A4	215			
Oboe	OB	A♯3–G6	151	2		
Faggoto	FG	A♯1–D♯5	312		piano	
Clarinet	CL	D3–F6	263	3		
Piccolo	PC	D5–C8	245			
Flute	FL	C4–C7	134	2		
Recorder	RC	C4–B6	160	3		

The recognition rates of the four types of saxophones at individual-instrument level (47–73%) were lower than those at category level (77–92%). This is because sounds of those saxophones were quite similar. In fact, Martin reported that sounds of various saxophones are very difficult for the human to discriminate [35].

Effectiveness of the flat categorization

Since we adopt the flat (non-hierarchical) categrization, the the recognition rates at category level depend on the category. The recognition rates of GUITARS and STRINGS at category level were more than 94%, while those ofBRASS, SAX, DOUBLE REEDS, CLARINET and AIR REEDS were about 70–90%. A conventional categorization has a hierarchy of musical instruments; categories such as brass and sax are sub-categories of "wind instruments." Our preliminary studies showed that the indentification at category level did not improve the recognition rates at individual-instrument level, which is simlar to Erone's results [11].

2.5 Evaluation of the Bayes Decision Rule

The effect of the Bayes decision rule in musical instrument identificaton was evaluated by comparing with the k-NN rule (k-nearest neighbor rule; $k = 3$ in this paper) with/without LDA. Three variations of the

dimension reduction are examined;

(a) Reduction to 79 dimension by PCA,

(b) reduction to 18 dimension by PCA, and

(c) reduction to 18 dimension by PCA and LDA.

The last one is adopted in the proposed system.

The experimental results listed in Table 2.5 showed that the proposed Bayes decision rule performed better in average than the 3-NN rule . Some observations are as follows:

(1) The Bayes decision rule with 79-dimension showed poor performance for Acoustic Guitar (AG), Trumpet (TR), Soprano Sax (SS), Tenor Sax (TS), Oboe (OB), and Flute (FL), since the training data is not enough for estimating parameters of a 79-dimensional normal distribution. For small traning sets with 79-dimension, k-NN is superior to the Bayes decision rule.

(2) LDA with the Bayes decision rule improved the accuracy of musical instrument identification from 66.50% to 79.73% in average. Although it seemed that PCA with 79-dimension performed better than LDA for Classical Guitar (CG), Violin (VN), and Alto Sax (AS), the cumulative performance of LDA for the categories of strings and sax is better than that of PCA.

2.6 Musical Instrument Ontology

Usually, musical sound ontology are borrowed from the classification of musical instruments specified in musical literature, and has been constructed manually [35, 38]. In this section, we use the pitch-dependent features to construct musical instrument ontologies by using the C5.0 decision tree program, a successor of C4.5 program [40] and by using agglomerative hierarchical clustering.

2.6.1 Musical instrument ontology by C5.0

From the naive decision tree obtained by applying the C5.0 to all the notes listed in Table 1.1, the hierarchy of Figure 2.2 is formed. The top level category of musical sound ontology consists of **Decayed instruments** and **Sustained instruments** and the latter consists of **Strings** and **Wind instruments**. This categorization is reasonable because it matches that of musical instrument classification. However, but the lower level categories are not the case. For example, the classification of **Winds** consists of 9 subcategories based on the features of the gradient of a straight line approximating power envelop by LSM ([41]) and amplitude of Amplitude Modulation (AM) ([78]).

The observations of the systematic construction of musical sound ontology by the C5.0 decision tree program with the proposed F0-dependent features are summarized below:

(a) **Wind instruments** can be classified **Recorder** and **Non-Recorder**. This discrimination matches the known fact that **Recorder** is different from other winds instrument.

(b) The categorization of **Wind instruments** except **Recorder** differs from that of musical literature. In particular, **Brass**, **Sax**, **Faggoto** are classified in the same category due to the pitch range, although their sounding mechanisms are different.

The top level category of musical sound ontology consists of **Decayed instruments** and **Sustained instruments** and the latter consists of **Strings** and **Wind instruments**. This categorization is reasonable

because it matches that of musical instrument classification. However, but the lower level catagories are not the case. For example, the classification of **Winds** consists of 9 subcategories based on the features of thegradient of a straight line approximating power envelop by LSM ([41]) and amplitude of Amplitude Modulation (AM) ([78], which is shown in Table2.6

The observations of the systematic generation of musical sound ontology by the C5.0 decision tree program with the proposed F0-dependent features are summarized below:

(a) **Wind instruments** can be classified **Recorder** and **Non-Recorder**. This descrimination matches the known fact that **Recorder** is different from other winds instrument.

(b) The categorization of **Wind instruments** except **Recorder** differs from that of musical literature. In particular, **Brass**, **Sax**, **Faggoto** are classified in the same category due to the pitch range, although their sounding mechanisms are different.

(c) Musical sound ontology needs plural aspects of sound features, in particular, sounding mechanism and pitch.

2.6.2 Pitch-dependency in musical sound ontology by agglomerative hierarchical clustering

In order to investigate the pitch dependency of musical instrument ontologies, one variant of agglomerative hierarchical clustering, group average agglomerative clustering [17] is used to obtain a dendrogram as an ontology. We focus on A2 (110Hz, the pitch mainly used for bases), A3 (220Hz, the pitch mainly used for accompaniment), and A4 (440Hz, the pitch mainly used for melody). Two kinds of sample data are used; One includes all the individuals of each instrument and the Mahalanobis distance based on the F0-dependent multivariate normal distribution is used. The other includes the data of one individual of each instrument and Euclidean distance is used. The resulting dendrograms obtained by group average agglomerative clustering are shown in Figures 2.4 and 2.5.

Observations concerning pitch-dependency of musical instrument ontologies are summarized below:

(a) The top level of dendrograms in Figure 2.4 classifies **Decayed instruments** and **Sustained instruments**, which is consistent with musical instrument ontology obtained by C5.0. This result is also consistent with judgments of human subjectives [34].

(b) The top level of dendrogram in Figure 2.5 (c) is consistent with the above result, but not in (a) and (b).

(c) Clarinet (CL) is isolated from **Wind instruments** in Figure 2.4 (b), partially because the fact that the powers of overtones of even order, in particular, 2nd and 4th, are small compared with other overtones in clarinet affects features at lower pitches.

(d) From A4 to A2, **Wind instruments** group is gradually breaking; at A3 or (b), clarinet (CL) is out of the group, and at A2 or (a), tenar sax (TS) and trombone (TB) are out of the group.

We think that the F0-dependent multivariate normal distribution succeeds in capturing the continuous change of timbre according to pitch. This observation is also consistent with the one that human subjects may not notice the change of timbre between one octave.

2.7 Conclusion

In this paper, we presented a method for musical instrument identification using the *F0-dependent multivariate normal distribution* which takes into consideration the pitch dependency of timbre. The method improved the recognition rates at individual-instrument level from 75.73% to 79.73%, and at category level from 88.20% to 90.65% in average, respectively. Based on the F0-dependent multivariate normal distribution, musical instrument ontology is constructed by using C5.0. It showed that top level categorization matches the conventional hierarchy of musical instruments. However, the categorization of wind instruments differs much from the conventional one. Musical instrument ontology based on the F0-dependent multivariate normal distribution is also validated by comparing with dendrograms obtained from sample data consisting of only one individual of each instrument.

Future works include evaluation of the method with different styles of playing, evaluation of the robustness of each feature against mixture of sounds, application of musical instrument ontology as annotation for MPEG-7, and automatic music transcription.

Table 2.4: Accuracy by usual distribution (baseline) and F0-dependent distribution (proposed).

	Indiv.-instr. level (%)			Category level (%)		
	Basel.	*Prop.*	Improv.	*Basel.*	*Prop.*	Improv.
PF	74.21	83.27	+9.06	74.21	83.27	+9.06
CG	90.23	90.23	±0.00	97.27	97.13	−0.14
UK	97.97	97.97	±0.00	97.97	98.31	+0.34
AG	81.23	83.93	+2.70	94.89	95.65	+0.76
VN	69.70	73.67	+3.97	98.86	99.05	+0.19
VL	73.94	76.27	+2.33	93.22	94.92	+1.70
VC	73.48	78.67	+5.19	95.16	96.24	+1.08
TR	73.51	82.12	+8.61	76.82	85.43	+8.61
TB	76.72	84.35	+7.63	85.50	89.69	+4.19
SS	56.80	65.89	+9.09	73.96	80.47	+6.51
AS	41.49	47.87	+6.38	73.76	77.66	+3.90
TS	64.71	66.01	+1.30	90.20	92.16	+1.96
BS	66.05	73.95	+7.90	81.40	86.05	+4.65
OB	71.52	72.19	+0.67	75.50	74.83	−0.67
FG	59.61	68.59	+8.98	64.74	71.15	+6.41
CL	90.69	92.07	+1.38	90.69	92.07	+1.38
PC	77.56	81.63	+4.07	89.39	90.20	+0.81
FL	81.34	85.07	+3.73	82.09	85.82	+3.73
RC	91.88	91.25	−0.63	92.50	92.50	0.00
Ave	75.73	79.73	+4.00	88.20	90.65	+2.45

Baseline: Usual (F0-independent) distribution

Proposed: F0-dependent distribution

Table 2.5: Accuracy by k-NN rule and the Bayes decision rule.

	k-NN rule ($k = 3$)			Bayes decision rule		
	(a)	(b)	(c)	(a)	(b)	(c)
	79-Dim.	18-Dim.		79-Dim.	18-Dim.	
	PCA		PCA&LDA	PCA		PCA&LDA
PF	53.94%	46.46%	63.39%	55.91%	59.06%	83.27%
CG	79.74%	77.16%	75.72%	98.28%	97.27%	90.23%
UK	94.58%	92.54%	97.63%	67.12%	80.00%	97.97%
AG	95.05%	92.79%	97.00%	19.97%	44.14%	83.93%
VN	47.73%	46.02%	45.83%	89.58%	84.47%	73.67%
VL	55.93%	54.24%	61.86%	71.19%	79.24%	76.27%
VC	86.20%	85.84%	84.23%	45.16%	30.82%	78.67%
TR	36.42%	38.41%	47.02%	41.72%	72.85%	82.12%
TB	70.99%	54.58%	77.86%	75.19%	78.24%	84.35%
SS	23.08%	14.20%	24.85%	48.52%	66.86%	65.89%
AS	37.59%	29.79%	40.43%	72.70%	41.84%	47.84%
TS	62.09%	66.01%	68.63%	30.07%	61.44%	66.01%
BS	68.84%	67.91%	66.98%	55.35%	54.42%	73.95%
OB	47.68%	48.34%	49.01%	43.71%	81.46%	72.19%
FG	64.10%	65.06%	74.36%	40.38%	30.12%	68.59%
CL	93.45%	87.93%	93.10%	95.51%	93.45%	92.07%
PC	84.08%	84.90%	84.08%	63.27%	58.37%	81.63%
FL	88.06%	72.39%	94.03%	35.82%	84.33%	85.07%
RC	97.50%	93.75%	97.50%	85.00%	96.25%	91.25%
Average	70.27%	66.98%	72.53%	62.11%	66.50%	79.73%

(a) Dimensionality reduction to 79 dim. using PCA only

(b) Dimensionality reduction to 18 dim. using PCA only

(c) Dimensionality reduction to 18 dim. using both PCA and LDA

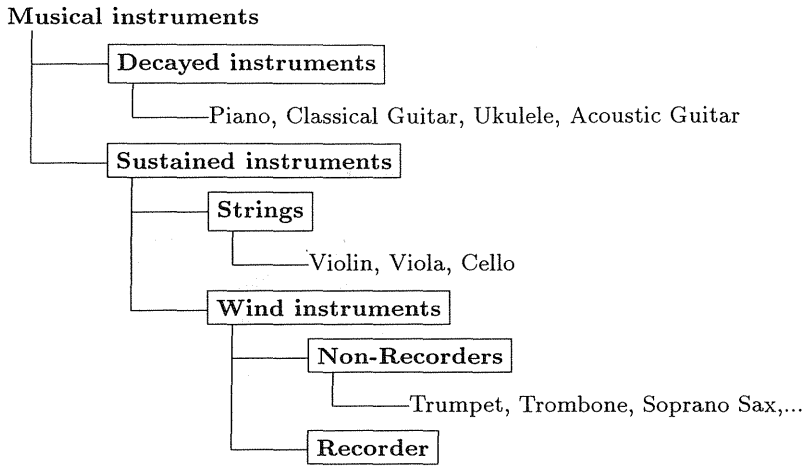


Figure 2.2: Top level category of musical instrument ontology obtained by C5.0

```
.....[74] > 5.621357
:    :...[61] > -4.37589
:    :    :...-1.35449 < [41] <= -0.666931 and [78] > 0.746687
:    :    :    => Decayed(0)  Strings(150)  Winds(48)
:    :    :...else
:    :         => Decayed(46)  Strings(157)  Winds(2131)
:    :...[61] <= -4.37589
:         :...[82] > 1.521393
:         :    => Decayed(0)  Strings(2)  Winds(59)
:         :...[82] <= 1.521393
:              => Decayed(10)  Strings(1156)  Winds(139)
:...[74] <= -5.621357
     :...[41] <= -0.625842 and [78] > 0.936492
     :    => Decayed(2)  Strings(55)  Winds(5)
     :...[41] <= -0.625842 and [78] <= 0.936492 and [82] > 0.91987
     :    => Decayed(0)  Strings(3)  Winds(26)
     :...else
          => Decayed(1913)  Strings(141)  Winds(53)
```

Where the features of sounds specified by a pair of bracket are summarized below:

[41]	Gradient of a straight line approximating power envelop by LSM
[61]	Ratio of the maximum power and power of 0.20 sec after onset.
[74]	Ratio of the maximum power and power of 0.90 sec after onset.
[78]	Amplitude of AM (Amplitude Modulation)
[82]	Amplitude of the first coefficient of MFCC modulation

Figure 2.3: Top level category of musical sound ontology

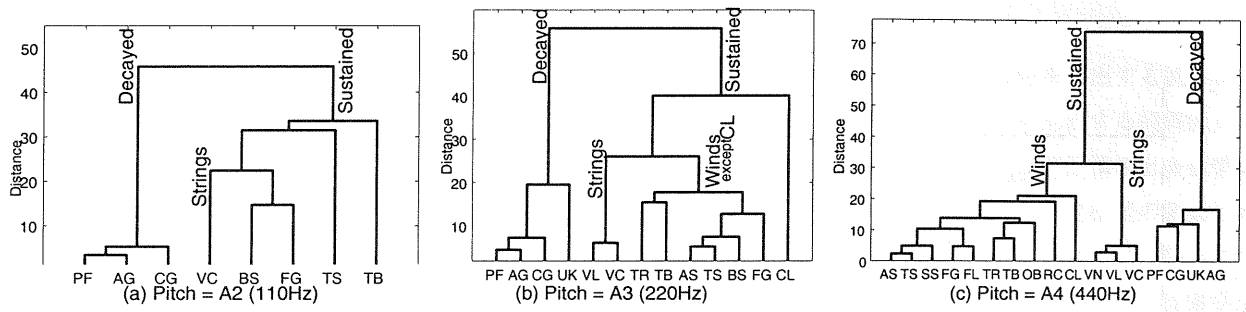


Figure 2.4: Musical instrument ontology constructed based on F0-dependent multivariate normal distribution. Instruments out-of-pitch-range are removed. The y-axis is Mahalanobis distance.

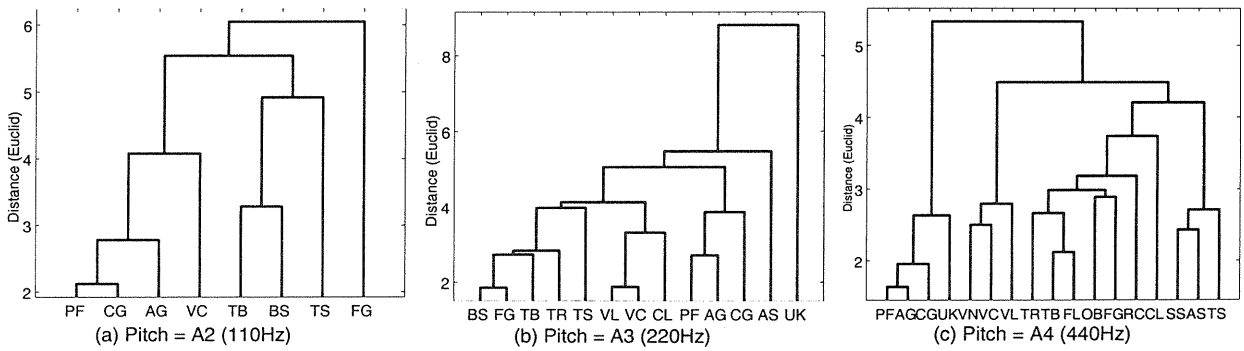
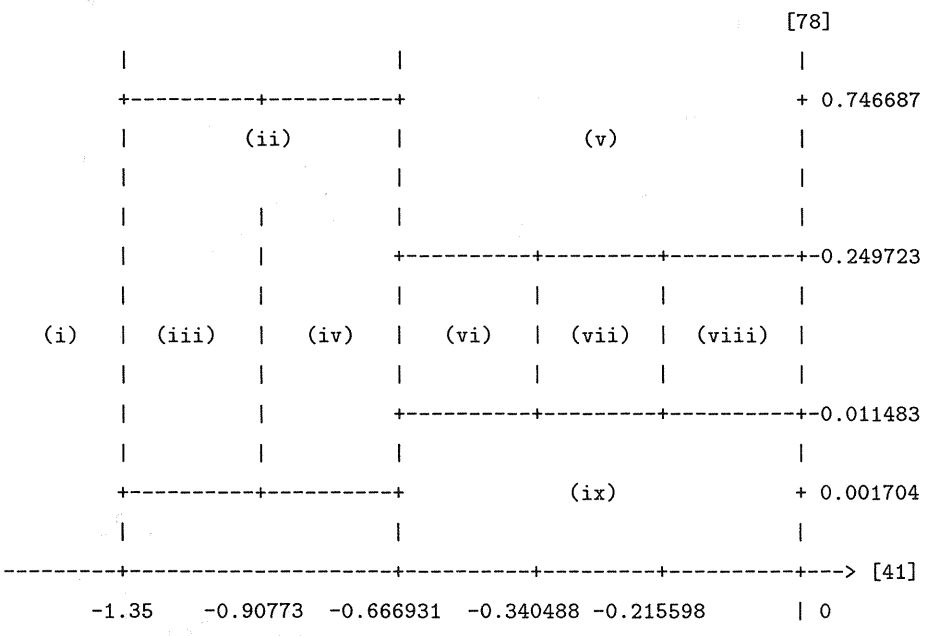


Figure 2.5: Musical instrument ontology constructed with only one individual of each instrument. Instruments out-of-pitch-range are removed. The y-axis is Euclidean distance.



Nine categories is defined by each corresponding rule. For example, rule (i) specifies **Sax**, while rule (ii) checks whether the instrument is **Oboe** or not.

Table 2.6: Details of classification of **Wind instruments** except **Recorder**

第 3 章

定位情報と音色情報を用いた複数楽器音の認識

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．

3.1 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 [3]．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 [23, 24]．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

[30]. 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 [36]. 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, 第 2 章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. 第 3 章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. 第 4 章で本研究のために作成したシステムについて説明する. 第 5 章で重奏の音源分離同定実験を行い, 考察する. 第 6 章で結論と今後の課題を述べる.

3.2 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 [23, 24]. 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 (詳細は第 4.2 節). 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.

(a) 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.

(b) 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.

(c) 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.

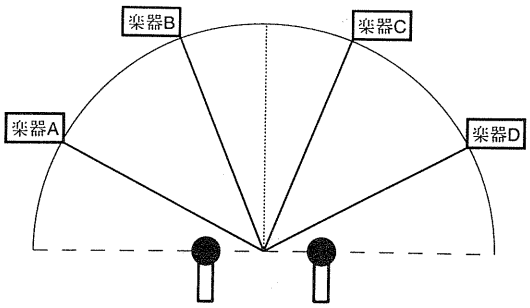


Figure 3.1: 楽器の配置例

3.3 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 3.1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.

(a) 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.

(b) どのパートにも属さなければ新たなパートを生成する.

(c) 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.

3.4 システムの構成

システムは図 3.2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

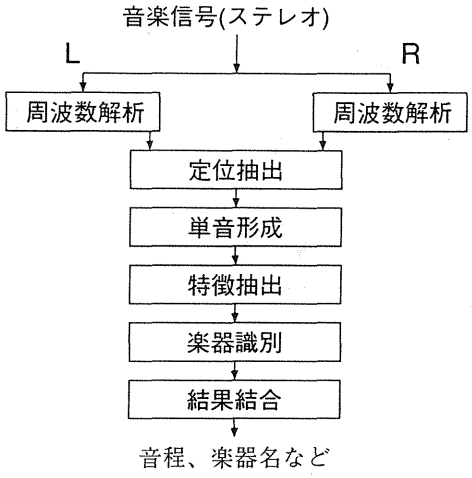


Figure 3.2: 音源分離同定システム

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．

3.4.1 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[12] 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[1]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．

3.4.2 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

(a) 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

(b) 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 3.3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．

[1]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．

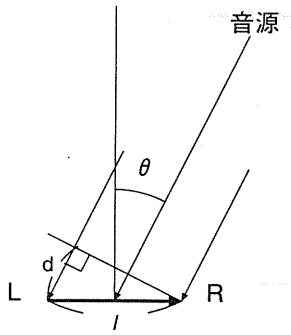


Figure 3.3: 音源の方向と両耳間の時間差

- 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.）

3.4.3 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.

(a) 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.

(b) 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.

(c) 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.

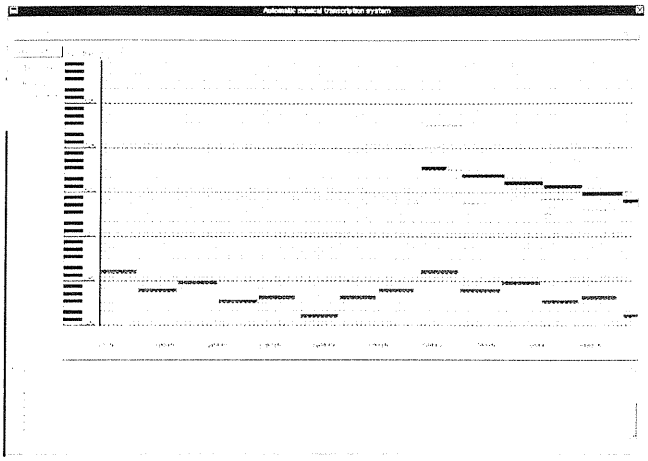


Figure 3.4: ピアノロール出力

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.

3.4.4 特徴抽出部

表 3.1に示す 23 個の特徴量を抽出する. これらは, 先行研究 [8][10][37][32] 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 [31, 32] 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.

3.4.5 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 [28] 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 minimax 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.

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

(3) q_i が最も大きいクラスを最終的な識別結果とする.

3.4.6 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 3.4) で結果が出力される.

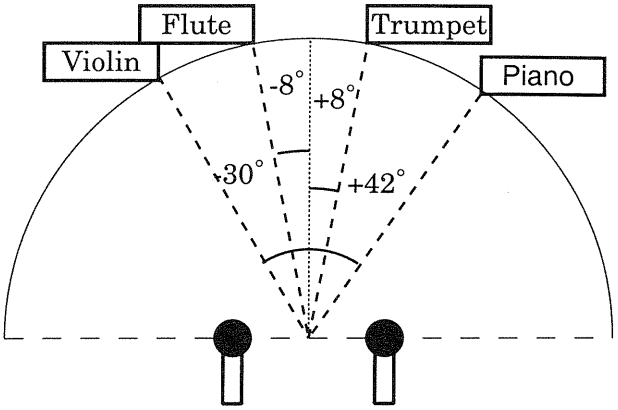


Figure 3.5: テスト曲の楽器配置パターン 1

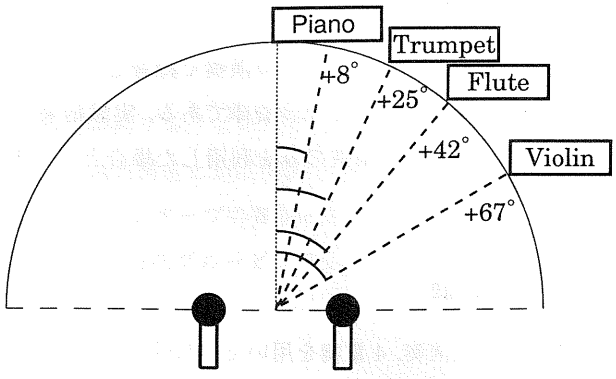


Figure 3.6: テスト曲の楽器配置パターン 2

3.4.7 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 3.2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 3.2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 3.3に示す．表 3.3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．

3.5 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）[23, 24], 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 [36] 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 3.5に示す．中央を 0 度とし，左は − 右は + で表す．

3.5.1 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．

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

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

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 3.4に示す．

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．

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．

(a) **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

(b) **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 3.5, 3.6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．

実験結果を表 3.7, 3.8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．

3.5.2 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 3.5の楽器配置の他に，新たに図 3.6の楽器配置でも実験を行う．図 3.6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 [36] では難しい楽器配置である．

実験結果を表 3.9, 3.10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．

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．

3.5.3 考察

(a) 同時的グルーピングにおける定位情報の有効性

表 3.5, 3.6, 3.7, 3.8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.

(b) 継時的グルーピングにおける定位情報の有効性

表 3.9, 3.10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.

(c) 位相差を用いることの有効性

図 3.6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.

3.5.4 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.

3.6 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, 第 5.4 節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.

Table 3.1: 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

Table 3.2: 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)

Table 3.3: 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 %

Table 3.4: 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%

Table 3.5: デュオ演奏単音形成結果 (クラス 1)

利用情報 \ 単音形成	再現率	適合率
調波構造のみ	62%	53%
調波構造と定位情報	96%	77%

Table 3.6: デュオ演奏単音形成結果 (クラス 2)

利用情報 \ 単音形成	再現率	適合率
調波構造のみ	43%	51%
調波構造と定位情報	83%	71%

Table 3.7: 4 重奏単音形成結果 (クラス 1)

利用情報 \ 単音形成	再現率	適合率
調波構造のみ	60%	62%
調波構造と定位情報	90%	71%

Table 3.8: 4 重奏単音形成結果 (クラス 2)

利用情報 \ 単音形成	再現率	適合率
調波構造のみ	34%	74%
調波構造と定位情報	68%	62%

Table 3.9: テスト曲音源分離同定再現率 (楽器配置パターン 1)

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

Table 3.10: テスト曲音源分離同定再現率 (楽器配置パターン 2)

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

第 4 章

教師なしクラスタリングと認識誤り補正による打楽器演奏の音源同定

音楽情報処理に関する研究は 1970 年以降広く行われるようになり, 近年, 様々な分野で, その重要性が認識されてきている. 特に, マルチメディアデータを扱う上では音楽情報処理技術は欠かせない, 例えば, 蓄積された音楽データに対して, 自動タグ付けを行い, 検索が簡単に行えるようにするなどの応用がある. しかし, 打楽器音を対象とした研究は, 楽音 (調波構造を持ち音高が明確な音) を対象とした研究よりもずっと少なく, 余り取り組まれてこなかった. 打楽器の発音機構は他の楽器とは大きく異なるので, まず, 楽音と打楽器音とで別々に音源同定する手法を開発する必要がある.

本研究では, ドラムやシンバルを含む複数の打楽器による演奏を対象とする. 音響信号が入力として与えられると, どのタイミングでどの楽器が発音したかが出力される. 従来研究の多くは, 打楽器単音の音源同定を扱っている. このような研究で得られた知見を, 複数打楽器による演奏には簡単には適用できない.

本研究では, ドラムのような膜鳴楽器類とシンバルのような体鳴楽器類を, まず, フィルタ処理により分離し, 各楽器類の音源同定を別処理により行う. 膜鳴楽器類の音源同定は, 低域通過フィルタ処理した音響信号に対して行われ, 体鳴楽器類の音源同定は, 高域通過フィルタ処理した音響信号に対して行わる, これらのフィルタにより, 互いのスペクトルの干渉を大きく減少させることができる.

膜鳴楽器の音源同定は, 教師なしクラスタリングに基づく. 従来通りの統計的な手法やテンプレートマッチングを採用した場合, 事前学習だけでは十分に対応できないと予想される. しかし, 同一曲内では音色変化が小さいので, クラスタリング手法は有効であると考えられる. また, 打楽器音は楽音に比べてデータベースが不十分であるので, 教師なしであることにより, 学習データを必要としない点は教師あり学習よりも有利である.

体鳴楽器の音源同定には, 音の重なりによる特徴量変動と認識誤りに一定のパターンがあることに着目し, k-NN 識別後, 認識誤り補正を行う手法を開発した. この手法により, 識別すべき体鳴楽器が Crash Cymbal の残響下にある場合, Snare Drum と同時発音する場合など, 音の重なりにより誤認識が生じやすい問題に対処できる. 各認識誤りパターンに対する補正法の学習は事前に自動で行った. また, 抽出した特徴量に対する信頼度を定義し, 信頼度に関するヒューリスティクスを与えることにより, どの認識誤りパターンに対する補正法を k-NN 識別結果のどこに適用するかの自動判定も行っている.

複数のデータを用いて音源同定実験を行い, 手法の有効性の検証を行った. 本研究で識別対象となる楽器は, 膜鳴楽器類として Bass Drum, Snare Drum, Low Tom, Middle Tom,High Tom の 5 種類, 体鳴楽器類として Crash Cymbal, Hihat Open, Hihat Close の 3 種類である. 体鳴楽器の学習用データは MIDI 音源 1 種類で合計 80 サンプル作成した. 膜鳴楽器の音源同定には学習用データは必要ない. 評価用データは学習用データ作成に用いた MIDI 音源と他の異なった MIDI 音源でそれぞれ 2 種類作成した. これは, 標準的な 8 ビートのドラムパターンである. また,

市販 CD に録音された打楽器演奏も評価対象とした．実験の結果，膜鳴楽器の音源同定率は，音源によらず 9 割程度を達成した．評価用データ中に数の少なかったタム類の識別に誤りは生じなかった．体鳴楽器の音源同定については，認識誤り補正法の導入により，MIDI 音源で作成した評価用データに対し 50%，市販 CD に対し 10% の認識誤り削減率が得られた．これらの結果から本研究の提案する教師なしクラスタリングと認識誤り補正法が打楽器演奏の音源同定に有効であることが示された．

4.1 はじめに

AV 機器の発達やデジタル信号処理技術の発展，計算機性能の向上により，楽器音や音楽演奏を対象とした研究は 1970 年以降広く行われるようになった．近年，音楽情報処理という研究領域が確立し，活発な研究が行われている．主要なテーマとして，音響信号に対する音源同定問題がある．音源同定とは，「入力として与えられた単音あるいは混合音から，どのタイミングでどの楽器が発音したかを出力する処理」のことを指す．入力音響信号は単音であったり，混合音であったり様々である．音源同定では，何らかの音響的な特徴あるいは他のメディア情報を活用して，演奏されているそれぞれの楽器を同定することになる．

健聴者は音源同定のような聞き分ける能力を生まれながらに備えており，楽音 (調波構造を持ち音高が明確な音) や打楽器音が複雑に混じりあった音響信号を入力として扱える．音源同定を計算機上で実現することは，人間の認知メカニズムの解明にもつながると考えられ，音源同定問題は非常に興味深い研究テーマとなっている．

打楽器を含む音楽演奏を入力信号として音源同定処理を行うためには，打楽器の発音機構は他の楽器とは大きく異なるので，まず，打楽器だけの演奏を対象とした音源同定技術を開発する必要がある．楽音を対象とした研究例として，柏野らは，音楽情景分析の処理モデル OPTIMA を開発している [23, 24]．OPTIMA では，音楽演奏の音響信号から，単音や和音などの音楽情報を記号表現として抽出する．ベイジアンネットワークによる情報統合の機構を組み込み，音源同定のみならず和音認識にも取り組んでいる．和音を扱う場合には，ボトムアップ処理だけでなく，和音を構成する単音に関する統計情報を統合することで和音認識率の向上が得られている．また，4 重奏を入力音響信号とした研究では，音の重なりによる曖昧性を解消するために，定位や音色類似度を用いる手法が提案されている [45]．しかし，これらの研究では，楽音だけしか扱われてこなかった．打楽器音を対象とした研究は，楽音を対象とした研究と比較して，研究事例は多くなく，音楽演奏を入力音響信号として扱った研究はその中でもさらに少ない．

本研究が扱う打楽器演奏の音源同定においては，同じ楽器であっても個体差により，また混合音を扱う上での音の重なりにより特徴量が大きく変動することを考慮にいれなければならない．打楽器音を対象とした従来研究の多くは，打楽器の単独発音から特徴量を抽出し，k-NN 法や PCA などの識別，次元圧縮の手法を利用することで音源同定処理を実現している．これは楽音を対象とした音源同定でも共通した手法である．しかし，本研究にこのような従来法を適用しただけでは，十分な音源同定精度は得られない．また，特にドラム類やタム類にあてはまるが，楽器サイズやミューティングの具合などの個体差による特徴量変動は大きく，事前学習を用いた識別では不十分である．

本研究では，打楽器による演奏の自動採譜処理の実現を主眼におき，前述のような問題を解決し，打楽器音の音源同定を行う手法を開発する．ドラム類のような膜鳴楽器の音源同定には教師なしクラスタリングを用い，楽器個体差に対処する．シンバル類のような体鳴楽器の音源同定には一度従来通りの識別手法を適用した後，認識誤り補正を行うことで認識誤りを削減する．また，演奏を扱う上では，適宜，演奏に関する事前知識を有効に利用していくことができる．例えば，一般的なドラムスの演奏においては，バスドラムとスネアドラムはタム類よりも多く叩かれるといった知識である．こういった知識の利用は，打楽器単音を対象とした音源同定にはない特性である．本研究では，知識を積極的に利用することで，入力音響信号への制約，音源同定処理の実現を行っていく．

以下，第 2 章では従来の打楽器音の音源手法について述べ，第 3 章では本研究の提案する手法を具体的に説明する．第 4 章では提案手法を用いた実験を行うことで．手法の有効性の評価を行う．第 5 章で本研究の結論を述べる．

Table 4.1: 実験に用いられたデータとその分類

Super-category	Basic-level	Sub-category
Membranes (380)	Bass (115)	Bass (115)
	Snare (150)	Snare (150)
	Tom (115)	Low (42)
		Medium (44)
		High (29)
Plates (263)	Hihat (142)	Open (70)
		Closed (72)
	Cymbal (121)	Ride (46)
		Crash(75)

* 括弧内はサンプル数

Table 4.2: 実験結果

	音源同定率	識別手法	特徴量選択手法
Super-category	99.3%	CDA	CDA
Basic-level	97.4%	K*	ReliefF
Sub-category	90.7%	K*	CDA

4.2 打楽器音を対象とした音源同定の従来研究

4.2.1 打楽器音の識別手法の比較検討に関する従来研究

Herrera らは, 標準的なドラムセットを構成する 9 種類の打楽器を対象として, 単独発音に対する音源同定問題を扱う際の識別手法について比較検討している [21]. 実験に用いられたデータとその分類を表 4.1に示す. カテゴリは一般的な楽器分類に基づき設計している.

識別手法として, k-NN, K*, C4.5, PART (PARtial decision Trees), CDA (Canonical discriminant analysis) の 4 つを取り上げている. k-NN, K* は学習サンプルの分布によらないアルゴリズムである. K* は距離尺度にユークリッド距離ではなくエントロピーを用いて探索を行う. C4.5 は決定木学習アルゴリズムであり, PART は C4.5 を改良したバリエーションの 1 つである. CDA は統計的な手法であり, クラス間距離ができるだけ大きく, クラス内分散ができるだけ小さくなるような分割を求めるアルゴリズムである.

また, 特徴量選択には, CFS (Correlation-based Feature Selection), ReliefF, CDA を利用している. CFS は, 特徴量の可能な部分集合全てに対して, その特徴量集合を用いた場合の識別率と特徴量集合の冗長性, 相関性を求めるアルゴリズムである. ReliefF は各特徴量の重要性を評価するアルゴリズムである.

音源同定実験に取り組み, 10 クロスバリデージョンを評価に用い, 実験データのうち 90% を学習用データ, 残りの 10% を評価用データとしている. これを 10 回繰り返して, 音源同定率の平均値を求めるものである. 各カテゴリレベルで, 最良の音源同定率が得られた手法の組み合わせとその音源同定率を表 4.2に示す. 単音を対象とした場合の各

打楽器名レベルでの音源同定率は 90.7% である.

4.2.2 パワー分布に基づくパターンマッチング手法

後藤らは, 複数の打楽器音が混合したモノラルの音響信号を入力とし, 打楽器の種類とその発音時刻・強度を表すシンボルを標準 MIDI ファイルの形式で出力する音源同定システムを開発している [14]. 識別対象となる楽器は, 標準的なドラムセットで, 9 種の打楽器で構成される. サンプラーで作成した典型的な 8 ビートのドラムパターンの演奏を入力信号とする. また, 予め使用する楽器のパワー分布をテンプレートパターンとして登録しておく必要がある.

まず, 発音時刻を検出し, そこを開始点として一定区間のパワー分布を取り出す. テンプレートパターンと入力パターンとの距離をテンプレートマッチングにより求める. ここで, 以下のような問題が生じる.

(a) テンプレートパターンと入力パターンの音量が異なると, 両者が同じ音色でも距離が大きくなる.

(b) 複数の打楽器が同時になると距離が大きくなる.

(c) 認識するのに関係のない周波数成分が異なっても距離が大きくなる.

上記のような問題を解決するために, それぞれの問題に対し, 次のような方法により解決を図るとしている.

(a) の解決策 : 音量補正法

図 4.1のように, 入力パターン中にテンプレートパターンが含まれていると仮定して, 入力パターン中のテンプレートパターンに相当する部分とテンプレートパターンとを比較し, 音量補正値を決定する.

(b) の解決策 : 音源分離を実現する距離尺度

複数の打楽器が発音した場合に対応するために, 入力パターン中にテンプレートパターンが含まれているかいないかを判断する距離尺度を用いる.

(c) の解決策 : 選択的注意の機構

距離を求める際に, すべての周波数成分を等しく評価すると, 認識に関係のない周波数成分が異なっても距離が大きくなるので, 各楽器ごとに重み関数を用意し, 周波数時間平面上でその楽器固有の領域に重みつける.

このシステムは, 前述の入力信号に関する制約内で高精度な音源分離を達成している. しかし, 同じ楽器だとしても, 実際には様々な音色のバリエーションが存在するので, 全てに対応するためには大量のテンプレートパターンの用意が必要となる問題が存在する. このような大量のテンプレートパターンを用意することは現実的ではなく, システムに登録されていないような音色を持つ個体による演奏に対しても有効に機能する音源同定システムの開発を目指す必要がある.

4.2.3 本研究でのアプローチ

本研究では, 打楽器による音楽演奏を対象とした音源同定を行う. 音楽演奏から特徴量を抽出する場合, 音の重なりにより特徴量が大きく変動する. これが原因で, 従来通りの識別手法を適用しただけでは音源同定率が低下することが大きな問題である. しかし従来研究では, 使用楽器の特徴は事前に登録されているなどの仮定をおいて本問題を回避しており, この問題に正面から取り組んでは来なかった. 打楽器単音を対象とした音源同定で識別に有効であった特徴量が, 混合音を対象とした場合に信頼できなくなる場合が多い. 本研究では, 抽出した特徴量に対する信頼度を導入し, 信頼度に関するヒューリスティクスにより, 音源同定の誤りを補正する手法を組み込むことにで, この問題の解決を目指す.

本研究では, 膜鳴楽器の音源同定に, 教師なしクラスタリングを用いる. 音楽情報処理分野において広く用いられる統計に基づく識別手法は採用しない. というのは, 様々な音色を持つ膜鳴楽器の識別には, 統計に基づく手法より,

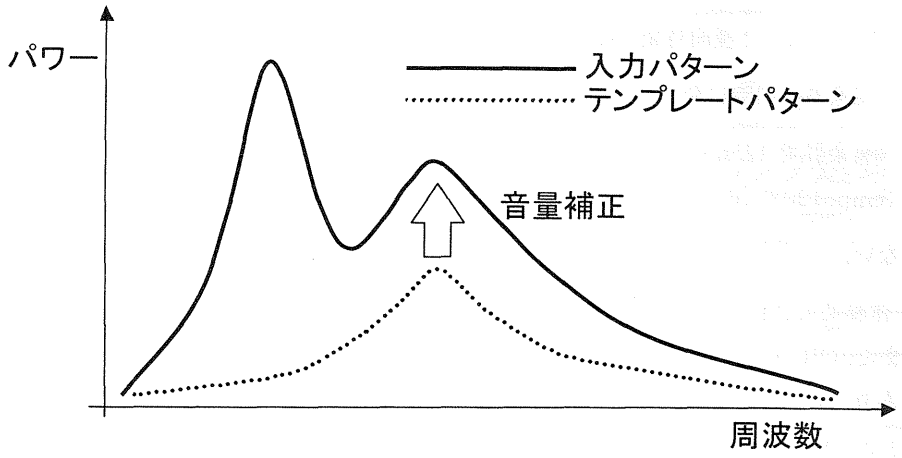


Figure 4.1: 音量補正の模式図

Table 4.3: 本稿で扱う打楽器群

膜鳴楽器	Bass Drum (BD), Snare Drum (SD), Low Tom (LT), Middle Tom (MT), High Tom (HT)
体鳴楽器	Crash Cymbal (CR), Hihat Close (HC), Hihat Open (HO)

同一楽器では曲内の音色変化が小さいことに着目した教師なしクラスタリングの方が有効であると考えられるからである．パワー分布によるテンプレートマッチングは，統計的手法ではないが，個体差が大きい膜鳴楽器の音源同定には適切ではない．教師なしクラスタリングは，個体差による特徴量変動に対しよりロバストである．

本研究では識別すべき対象として，Ride Cymbal は除外する．現段階で，Ride Cymbal を扱うほど十分なデータが集まっていない．しかし，本研究で提案する手法は，Ride Cymbal を含めて 9 種楽器の音源同定を行う問題を設定したときも有効性を失わないように設計していく．

4.3 打楽器演奏を対象とした音源同定手法

4.3.1 音源同定の対象とする問題設定

本研究においては，前述したように「入力として与えられた音響信号から，どのタイミングでどの楽器が発音したかを出力する処理」を音源同定と呼ぶ．例えば，CR が 5.4 秒後，8.8 秒後に発音したという情報が出力できればよい．また，各音符が何拍目であるかや，小節の区切りや演奏速度 (Tempo) の推定などは行わない．

本研究では，入力音響信号は，標準的な構成のドラムセットによる Tempo120 より遅い 8 ビートのドラム演奏であるとする．ドラムセットは，表 4.3の 8 種類の打楽器で構成されている．ここで，膜鳴楽器と体鳴楽器という分類は，Erich von Hornbostel と Curt Sachs の体系的楽器分類に基づく [7]．前者は 1kHz 以下の中低域の周波数成分が多く，後者は 4kHz 以上の高周波数域に広くスペクトルが分布している．使用楽器および演奏法について，以下の仮定をおく：

(1) 体鳴楽器が同時に 2 種類発音されることはない.
また, 膜鳴楽器も同様に 2 種同時発音はない.

(2) 体鳴楽器と膜鳴楽器は同時に発音されてもよい.

(3) 発音間隔は膜鳴楽器で 125ms 以上, 体鳴楽器で 250ms 以上とする.
(125ms は Tempo120で 16 分音符に相当する.)

(4) 未知楽器はない.

(1) に関して, 体鳴楽器 2 種, あるいは膜鳴楽器 2 種を同時に発音させることは, 実際の演奏においてはそのような場合は少ない. また, CR (Crash Cymbal) と HC (Hihat Close) が同時発音したとしても, この認識は人間にとっても非常に困難であり, 本研究ではこのような場合は対象としない. (2) に関して, 例えば BD (Bass Drum) と CR, SD (Snare Drum) と HC などの同時発音は普通に行われる. (3) に関して, Tempo120 は 8 ビートのドラム演奏上標準的な速度であり, 膜鳴楽器の音符の最小単位は多くの場合 16 分音符である. 同様に, 体鳴楽器の音符の最小単位は 8 分音符であるとできる. (4) に関して, 本研究では標準的な構成のドラムセットを対象としているので, どのような楽器が実際に使用されるかは所与のものとする. このような理由から, (1) ~ (4) の仮定は妥当であると考える.

4.3.2 問題の所在と解決手法

前述のような問題設定のもとでの, 打楽器音の音源同定の主な問題の所在は以下の 2 つである.

(a) 個体差の大きい膜鳴楽器は十分な事前学習ができない. 例えば BD を考えてみると, 胴のサイズやミュートの具合などによって, 特徴量は大きく変動する. これらを網羅的にデータ収集することは難しい.

(b) 演奏を扱う上で, 音の重なりによる認識誤りは避けられない. 特に, 体鳴楽器による残響の影響は大きい. また, SD は裏面に金属の紐が張ってあり, 膜鳴楽器でありながら金属振動も生じるため, 高周波数域までスペクトルが分布している. そのため, 体鳴楽器と SD の同時発音時については特別な処理が不可欠である.

上記問題に対し, 帯域フィルタ処理後にそれぞれの楽器類用に設計された信号処理手法を適用するというアプローチをとる. 具体的には, 以下に提案する音源同定処理を行うための前処理として, 膜鳴楽器の認識に用いる音響信号には低域通過フィルタで, 体鳴楽器の認識用に用いる音響信号には高域通過フィルタで意図的な周波数帯域制限をかけることで, 互いのスペクトル混合による互いの影響を大きく抑制する. この処理の後, 以下の 2 つの手法により問題の解決を図る.

(a) に対しては, 同一曲内では同じ楽器を使う以上, 音色変化が少ないことに着目し, 膜鳴器認識に教師なしクラスタリングを利用する. 事前学習を必要としないので, 楽器個体差による特徴量変動の影響は教師あり学習よりも小さいと考えられる. さらに, 打楽器音は楽音に比べてデータベースが不十分であるので, 教師なしクラスタリングは教師ありの識別手法よりも有利と考えられる.

(b) に対しては, 体鳴楽器認識時に, 体鳴楽器の残響や SD との同時発音などの音の重なりによる特徴量変動と認識誤りに一定のパターンがあることに着目し, k-NN 識別後, 認識誤り補正を行う手法を開発する.

4.3.3 打楽器音同定システムの構成

打楽器の音源同定処理は, 前述したように膜鳴楽器識別と体鳴楽器識別で別々に行う. 図 3.1 に処理の流れを示す. 膜鳴楽器の認識では, 入力音響信号に図 3.2 左の低域通過フィルタを適用した後, (1) 発音時刻検出, (2) 教師なしクラスタリングの順に処理を行い, 膜鳴楽器名を出力する. 体鳴楽器の認識では, 入力音響信号に図 3.2 右の高域通過フィルタを適用した後, (1) 発音時刻検出, (2) k-NN 法による識別を行った後, その結果をもとに (3) 認識誤りパターンを利用した補正を行し, 体鳴楽器名を出力する.

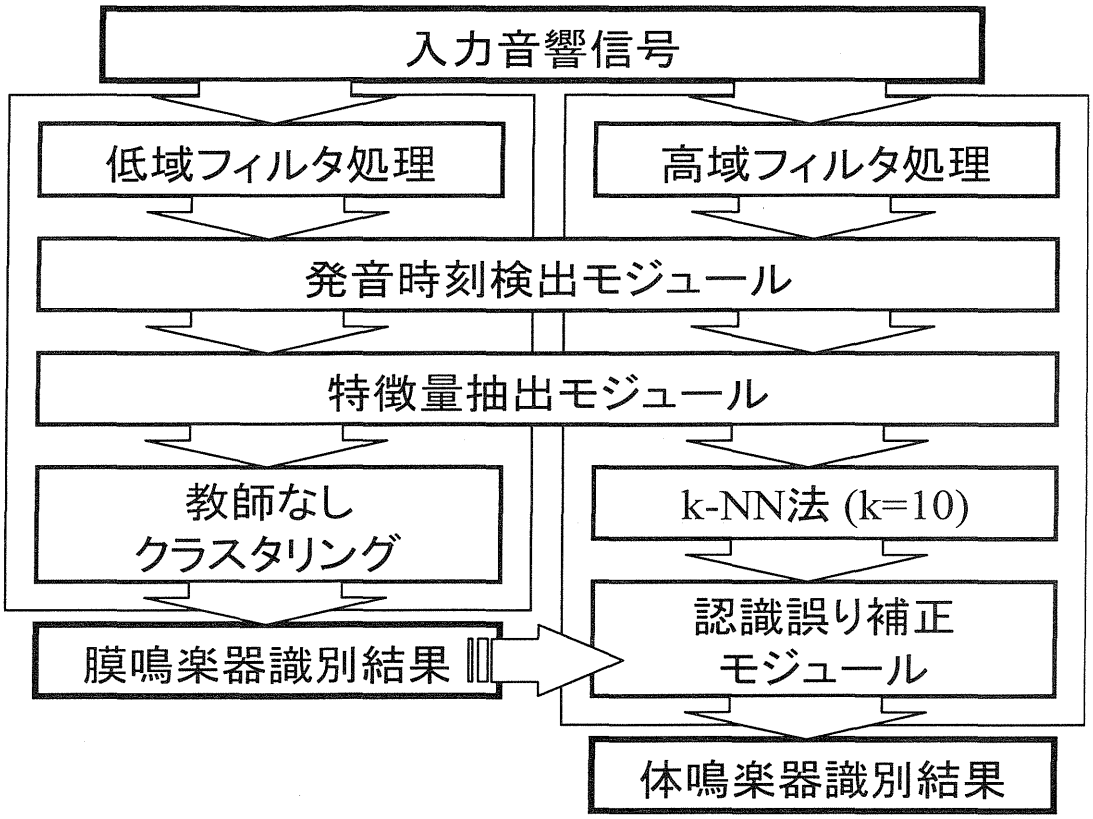


Figure 4.2: 打楽器音同定システムの構成

4.3.4 発音時刻検出

発音時刻検出には, 後藤らの手法を用いる [14]. ただし, パワー分布形状を表現する周波数軸のスケールは, 2 種の対象楽器類で異なったものを使用する. 膜鳴楽器の発音時刻検出には, 中低域の分布をよく表すように周波数軸を対数尺度の Cent で表し, 体鳴楽器の場合には高域の分布をよく表すように周波数軸を Hz で表す. 一定の周波数幅 f_c で周波数軸を区切り, 各区間内の最大パワーをその区間の代表値とする. (f_c = 80[Cent], 130[Hz] とした) 計算されたパワー分布形状の時刻 t, 周波数 f におけるパワーをそれぞれ $P_k(t,f)$ とする. (k = Cent, Hz) こうして得られたパワー分布形状から, 発音時刻を検出する. これは, パワー分布形状の時間方向の 1 次微分値が大きい値をとる時刻を発音時刻とするものである. 立ち上がりの度合い $Q_k(t,f)$ 算出と, 発音時刻検出の手順を以下に示す.

(1) $t = l-1$, l, $l+1$ の各時刻において連続して

$$\frac{\partial P_k(t,f)}{\partial t} > 0$$

を満たすとき, $t = l$ における $\partial P_k(t,f)/\partial t$ を $Q_k(t,f)$ とする. 満たさないときは $Q_k(t,f) = 0$ とする.

(2) 各時刻 t ごとに $Q_k(t,f)$ の重み付き合計値 $S_k(t)$ を次式により定める.

$$S_k(t) = \sum_f F_k(f) Q_k(t,f)$$

ただし, $F_k(f)$ は図 4.4のような楽器の特性に応じた重み関数である.

(3) $S_k(t)$ に対し, Savitzky と Golay の方法による平滑化と微分 [46] を用い, 極大値を与える時刻を検出する. $S_{Cent}(t)$ から求めた時刻を膜鳴楽器の発音時刻とし, $S_{Hz}(t)$ から求めた時刻を体鳴楽器の発音時刻とする.

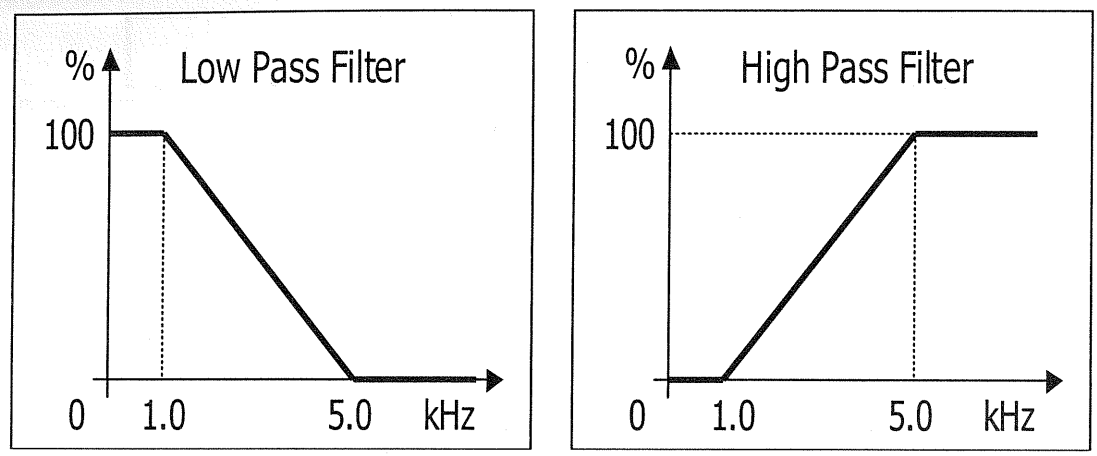


Figure 4.3: 低域通過フィルタと高域通過フィルタ

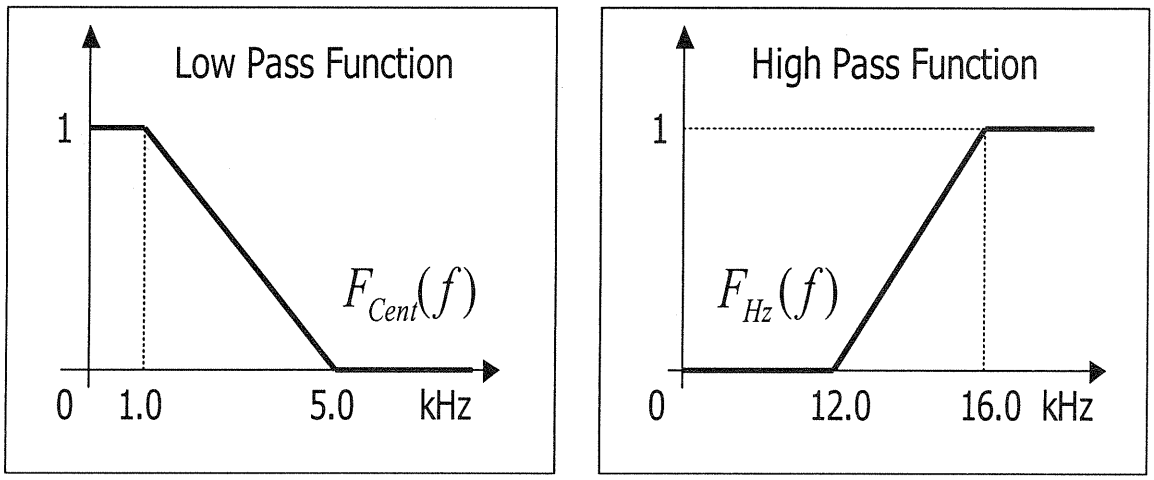


Figure 4.4: 重み関数：$F_{Cent}(f)$ と $F_{Hz}(f)$

また, 各音符の発音時刻を代表する代表パワー分布形状 $V_k(f)$ を以下のように定義する.

(1) 発音時刻以降, 100ms 以内の最大パワーフレームを求める.

(2) 最大パワーフレーム後 100ms 間の $P_k(t,f)$ の時間方向の平均値を $V_k(f)$ とする.

従来研究から, スペクトルのパワー分布は音源同定に有用な特徴量であることが分かっているので, 代表パワー分布形状を $V_k(f)$ を膜鳴楽器識別, 体鳴楽器識別の両方に利用する.

4.3.5 膜鳴楽器の教師なしクラスタリング

代表パワー分布形状 $V_{Cent}(f)$ に対し, k-means 法を利用して教師なしクラスタリングを行う. 膜鳴楽器識別に用いられる全ての音響信号は低域通過フィルタ処理されているため, 低域のみのパワー分布を考えるだけでよい. ここでは, $f = 0, \cdots, 49$ ($\times f_c$[Cent]) とし, 50 次元のベクトルをクラスタリングする. 本研究では, 打楽器を対象とした音源同定問題におけるクラスタリング手法の有効性を確かめるために, クラス数 (演奏中での実際の使用楽器数) は, 既知

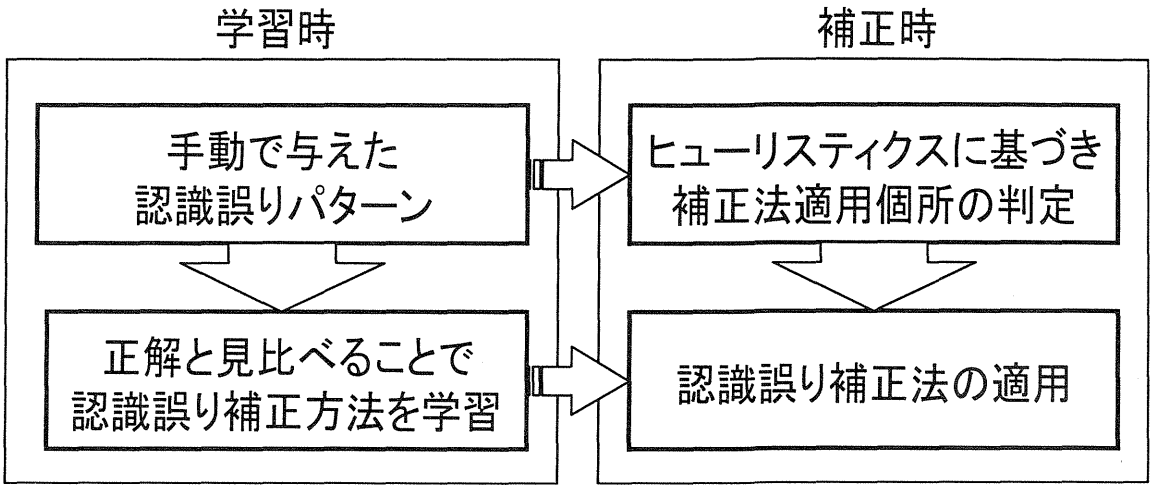


Figure 4.5: 認識誤り補正法の学習と適用の流れ

とする. 例えば, BD, SD, LT, HT (MT 不使用) ならば4クラス分類問題となる.

クラスタリング後の楽器名同定処理には, 以下の2つの事前知識を用いる.

- 一般的なドラム演奏においては, BD と SD がタム類よりも多く叩かれる.
- タム類は LT, MT, HT の順に音が高くなるように知覚される. これは, 周波数重心が順に高くなることが原因と考えられる.

これらの知識を用いることで, 以下の処理により楽器名同定を行う:

(1) 教師なしクラスタリングにより得られたクラスタのうち, 要素数が最大のクラスタと2番目のクラスタを選ぶ.

(2) 各クラスタの周波数重心の平均値を計算し, 小さい値のクラスタを BD, 大きい値のクラスタを SD とする.

(3) 上位2位以外のクラスタに対しても同様に周波数重心の平均値を計算し, 小さいものから順に LT, MT, HT とする. タム類3種類すべてが使用されない音響信号の場合, 例えば LT, HT のみ使用される場合, 周波数重心の小さいものから LT, HT とする.

4.3.6 体鳴楽器の音源同定と認識誤り補正

代表パワー分布形状 $V_{Hz}(f)$ を特徴量として k-NN 法 (k=10) により得られる識別結果を, 本章で述べるアルゴリズムで補正する. 詳しくは4.3.6章以降で説明するが, 認識誤りパターンを手動で与えることで, その誤り補正法を自動学習し, k-NN 識別結果のどこに適用すればいいかを自動判別する. 処理の流れを図4.5に示す.

認識誤りパターンと補正法の学習

打楽器演奏においては, 同時発音の影響や過去の発音の残響成分にスペクトルが埋もれることにより認識誤りが起きやすい個所が存在し, 特徴量変動の要因と認識誤りには一定のパターンがあると推測される. 本研究では, これを認識誤りパターンと呼ぶ. こでまでの予備実験から得られた以下の4パターンを検出した.

[I] HC, SD の同時発音時に, HC→CR と誤認識する

[II] CR の残響影響下において, HC→CR と誤認識する

Table 4.4: 打楽器の音色を表す 35 個の特徴量の概要

(1)	**スペクトルの定常的特徴** (3 個 : FT1 - FT3)
	周波数重心, 最大パワー周波数, 最大から NTH 番目までのパワーを持つ周波数の時間方向の平均
(2)	**2-4 次モーメントに関する特徴** (3 個 : FT4 - FT6)
	パワーの分散, 歪度, 尖度の時間方向の平均
(3)	**アタック区間に関する特徴** (8 個 : FT7 - FT14)
	発音から最大パワーフレームまでの時間とその対数, パワーの時間方向の平均値, パワー包絡の面積とその割合, ゼロクロスの割合, 時間方向の重心とアタック時間に対する割合
(4)	**ディケイ区間に関する特徴** (6 個 : FT15 - FT20)
	ゼロクロスの割合, 周波数重心の時間方向の平均値と分散, パワーの分散, 歪度, 尖度などの時間方向の平均
(6)	**残響成分に関する特徴** (2 個 : FT21 - FT22)
	最大パワーフレーム後 Y(ms) 後までの残響度合い (Y = 100, 200) (エンベロープの面積 / 最大パワー * Y ms)
(7)	**MFCC に関する特徴** (13 個 : FT23 - FT35)
	特徴量抽出区間内の 13 次元 MFCC の時間平均

* アタックとは最大パワーフレームまでの区間, ディケイとはそれ以降の区間を指す.

[III] 残響系の特徴を持つため, HO→CR と誤認識する

[IV] パワー分布が似ているため, HO→HC と誤認識する

[I], [II] は音の重なりに起因した誤り, [III], [IV] は音の特徴の類似性に起因した誤りである. 特に, 前者の誤りは, 単音で学習して演奏中の混合音を認識する上で避けられない認識誤りであり, この補正を行うことは極めて重要である.

上記の各認識誤りパターンの補正を行う識別機械として, 決定木 T_i $(i = I, \cdots, IV)$ を以下の手順により構築する.

(1) 認識誤りパターンに該当する演奏データを作成する. 例えば, [I] の認識誤りに対する補正法を学習させるためには, SD, HC が同時発音するような演奏データを作成し. 学習用データとして利用する.

(2) 作成した演奏データから特徴量を抽出し, 特徴量集合 S とする. 特徴量は従来研究 [21, 32] を参考に定めた 35 個である (表 4.4).

(3) $A \rightarrow B$ と誤認識する認識誤りパターンの補正法として, クラス B に属する特徴量と S を識別する決定木を決定木学習法 C5.0 により構成する. 例えば, [I] の認識誤りに対する補正法学習の場合には, A=HC, B=CR であり, S とは SD と HC との混合音から抽出した特徴量の集合である.

上記の手順により, [I] から [IV] に対し, それぞれ個別に補正用の識別機械が求まる. この補正法を k-NN 識別結果に対し, どこに適用するかについては, 4.3.6章以降で説明する. ここで注意しなければならないのは, 補正法を適用したからといって, 補正前と補正後で, 常に識別結果が変わるわけではないということである. 認識誤り補正法の目的は, 特別に設計した識別機械により再識別を行うことによって, 最終的な音源同定結果の信頼性を高めることにある.

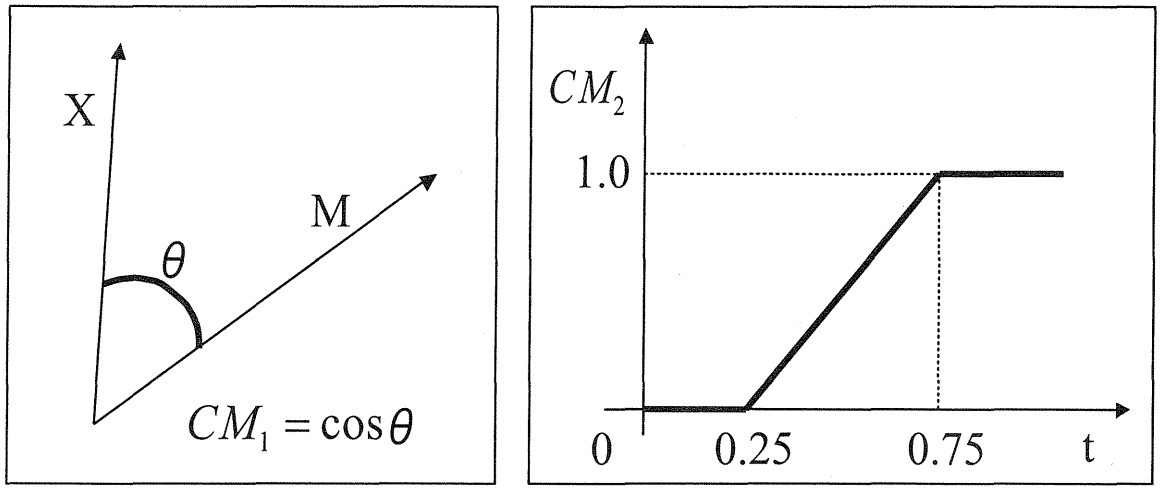


Figure 4.6: 信頼度の算出：CM_1 と CM_2

信頼度の導入

どの時点でどの認識誤り補正法を適用するかを判定するために，抽出した特徴量ベクトルに対する信頼度 (CM) を導入する．CM が低いものは，認識誤りを起こしている可能性が高い．よって，CM を用いた適当なヒューリィスティクスを与えることで，補正法を適用すべき個所を判定可能である．まず，抽出した特徴量ベクトルに対する CM を次の 2 つの値の積として定義する．

- CM_1：抽出した特徴量ベクトル X と各識別対象クラスの特徴量ベクトルの平均 M との類似度を示す．これは次式で計算され，2 つのベクトルの間の角のコサイン値を表す (図 4.6左)．

$$CM_1 = \frac{(X, M)}{|X||M|}$$

- CM_2：$0.5 < CM$ の CR 検出後，特徴量の信頼性は時間経過で上昇すると仮定し，モデル化した値である (図 4.6右)．残響の非常に大きい CR の後では認識誤りが起きやすく，特別に考慮する必要がある．

信頼度 CM は，

$$CM = CM_1 * CM_2$$

で計算し，k-NN 法による識別結果クラスに対してだけでなく，全クラスに対して求め，CM_{Class} (Class = CR, HO, HC) とする．

ヒューリスティクスの利用と補正法の適用

4.3.6の認識誤りパターンに対する補正法を適用するためには，認識誤りを起こしている個所の同定と適用すべき補正法を決定する必要がある．そのために，CM を利用したヒューリスティクスを用いる．以下に，本研究で利用するヒューリスティクスを図 4.7に示す．

ヒューリスティクスの設計方針は以下の通りである．前述したように，認識誤りパターン [I], [II] は混合音を扱う上で避けられない誤りであるので，[III], [IV] に対する誤り補正よりも優先して考慮しなければならない．そのため，[I], [II] に当てはまるかどうかを [III], [IV] に当てはまるかどうかよりも先に判定する．

if CR と識別 && $CM_{CR} < \theta_1$
if SD が同時発音 **then** *apply*(決定木 T_I)
elseif CR 検出後 0.5s 以内 **then** *apply*(決定木 T_{II})
else *apply*(決定木 T_{III}) **fi**
elseif HC と識別 && $CM_{HC} < \theta_2 < CM_{HO}$
then *apply*(決定木 T_{IV}) **fi**

Figure 4.7: 認識誤り箇所の同定と適用すべき補正法決定のためのヒューリスティクス

Table 4.5: 学習用データの内訳

楽器名	楽器個体	強弱	総数
CR	10	4 種類	40 音
HO, HC	5	4 種類	20 音

また, CR 残響下での認識誤りパターン [I] に対する補正が必要かどうか判定する際には, 判定すべき個所の前に実際に CR が発音されているかどうかが重要である. 誤って CR と識別されたとすると, そのあとの補正に悪影響が考えられる. そこで, 閾値 θ_1 により, 信頼性の高い CR の検出時のあとにだけ補正法を適用することにする. [IV] の補正法適用時にも, 閾値 θ_2 を設定する. HO と HC の特徴は似ており, 演奏中にはもともと HC が多いことから, 閾値を設定することにより, 必要以上に補正法を適用してしまう可能性を排除する.

4.4 実験と考察

4.4.1 実験条件

体鳴楽器識別のための学習サンプルは, Roland 社製 MIDI 音源 SC-88VL を用いて作成した合計 80 音である. 詳細を表 4.5に示す. 膜鳴楽器識別には教師なしクラスタリングを利用するため, 事前学習のためのサンプルを必要としない.

評価用のデータは SC-88VL 及び YAMAHA 社製 MIDI 音源 MU-2000 で作成した. これは, 4.3.1章で述べたような条件に適合する典型的な 8 ビートのドラム演奏であり, 4 小節の演奏を 2 セット作成した (以降, 本稿では評価用データをそれぞれ SC1, SC2, MU1, MU2 と呼ぶことにする). この評価用データのスコアを図 4.8に示す. また, ドラムの実演奏を収録した市販 CD も評価の対象とした. この CD 内の演奏にはタム類, CR の発音はなかった. スコアを図 4.9に示す. つまり, 膜鳴楽器識別では 2 クラス分類問題を解くことになる. これらの音響信号は, すべて 16bit, 44.1kHz のモノラル信号である.

4.4.2 発音時刻検出実験

4.3.4節の手法を評価するために発音時刻検出実験を行った. 周波数解析には, 窓幅 2048 点, 窓シフト 882 点として短時間フーリエ変換 (STFT) を用いた. 窓幅を長くとりすぎると, 周波数分解能は向上するが, 時間分解能の低下により, 音符の速い変化についていけなくなる問題が生じる. そのため, 発音時刻検出部では時間分解能を重視し, 後の音源同定処理部では周波数分解能を重視するというようにパラメータの設定を変えた. 検出された時刻と正解の時

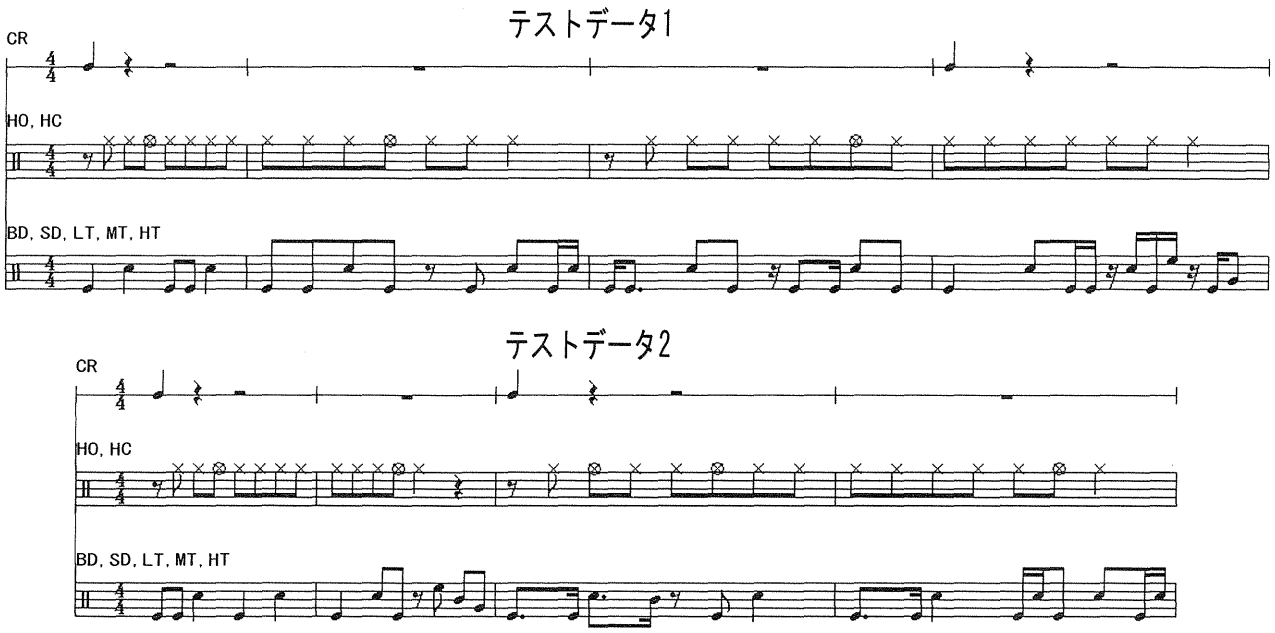


Figure 4.8: 評価用データのスコア (MIDI 音源で作成)

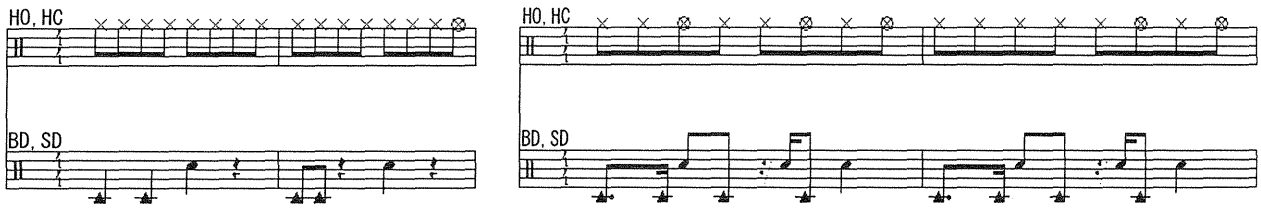


Figure 4.9: 評価用データのスコア (市販 CD)

刻とのずれが, 膜鳴楽器の場合 62.5ms 以内 (Tempo120 で 32 分音符の長さのずれまで許容), 体鳴楽器の場合 125ms 以内 (Tempo120 で 16 分音符の長さのずれまで許容) のとき, 発音が正しく検出できたとする. 実験結果を表 4.6に示す. 膜鳴楽器の発音時刻検出率は, 再現率, 適合率ともに 100% であった. 例として, SC1 に低域通過フィルタ処理後の信号のパワーエンベロープを図 4.10に, 各時刻におけるパワーの立ち上がり度 $S_{Cent}(t)$ を図 4.11に示す.

パワーエンベロープの裾野の広がりによらず, パワーの急激な立ち上がり時刻だけに急峻なピークを持つグラフが得られた. このピーク以外での値は小さく, 容易にピークの抽出ができるので, 本手法はパワーエンベロープより直接ピークを求める手法よりも有利であることが分かる.

実験結果から, 膜鳴楽器の発音時刻検出は非常に高精度に行えることが示された. 膜鳴楽器のスペクトル分布は, 例えば BD ならば 60Hz 付近, SD ならば 200Hz 付近の低周波数域に非常に大きなパワーのピークを持ち, パワーの集中が見られる. 低域通過フィルタで体鳴楽器のスペクトル成分をカットしきれなくても, 1kHz 以下の低周波数域では, 膜鳴楽器に由来するパワーの方が体鳴楽器に由来するものよりもずっと優勢であるので, 膜鳴楽器の発音時刻検出にはほとんど影響を与えない.

Table 4.6: 体鳴楽器の発音時刻検出率

	テストデータ 1	テストデータ 2	合計
SC-88VL	100.0% / 100.0%	100.0% / 96.6%	100.0% / 98.3%
MU-2000	100.0% / 100.0%	100.0% / 100.0%	100.0% / 100.0%
市販 CD	100.0% / 93.5%		

* 再現率 / 適合率 で表記している.

逆に, 体鳴楽器の発音時刻検出率の適合率は低い. 体鳴楽器のスペクトルは 5kHz から 20kHz まで広範囲な周波数域に分布し, ある周波数域へのパワー集中は見られない. また, 楽器の特性として, 体鳴楽器はアタックや残響が長いのに対し, 膜鳴楽器の場合には比較的短い. これらのことから, 体鳴楽器の各時刻のパワー変化は, 膜鳴楽器ほど大きくはない. 4.3.4節の発音時刻検出法は, 各時刻におけるパワー変化を求め, 立ち上がりが急激な時刻を発音時刻とするものである. 体鳴楽器検出用の立ち上がり度 $S_{Hz}(t)$ を求めたときに, 膜鳴楽器ほど急峻なピークは得られず, 体鳴楽器の発音時刻検出は難しい. 適合率が低い理由は, SD や高い周波数重心をもつタム類のスペクトルの影響を高域通過フィルタによって完全にカットできなかったことによる. SD では, 裏面に張った金属の紐の振動により, 高周波域までスペクトルが分布している. そのため, 実際に体鳴楽器が発音されていなくとも, SD の発音が体鳴楽器の発音として誤検出されやすく, 適合率が低下する原因となる.

今回の実験内においては, 膜鳴楽器の発音時刻検出率は再現率, 適合率ともに 100% であったが, 体鳴楽器の適合率に関しては, 100% を達成できていない. 再現率が 100% でない, つまり, 発音が再現できなかった個所については, 後続の音源同定部に情報が伝わらないので音源同定することはできない. しかし, 適合率が 100% でない, つまり, 実際に発音されていないのに誤って検出されたものについては, 音源同定処理部でなんらかの楽器名に同定されることになり誤った結果が出てくる. 今後, こういったものについて, 発音誤検出されたものであると判定する仕組みを検討していく必要がある.

4.4.3 膜鳴楽器の音源同定実験及び考察

4.3.5節の手法を評価するために膜鳴楽器の音源同定実験を行った. 周波数解析には, 周波数分解能を重視して, 窓幅 4096 点, 窓シフト 220 点として STFT を用いた. 評価用データ内の膜鳴楽器の内訳を表 4.7に, 実験結果を表 4.8に示す. 従来研究において, ドラム単音に対しての音源同定率が 9 割程度であることを考慮すると, 混合音で 88% 以上というのは非常に精度の良い結果であると言える. このことから, 教師なしクラスタリングが膜鳴楽器の音源同定に有効であることが示された.

認識誤りが生じた個所について考察する. 評価用の演奏データには, タム類の数が少なく, SD, BD の識別を重視すれば, 全体の音源同定率は向上する. 極端に言えば, 全て SD か BD に認識するシステムを構築すれば, 全体の音源同定率の向上につながる可能性がある. しかし, このようなシステムは自動採譜の面から意味がない. タム類が正しく認識できることは重要である. 本実験において, タム類に関する間違いは生じていない. 認識誤りは, 例えば BD, SD が 16 分音符で連続して発音されるといった個所で生じている. 教師なしクラスタリング対象とした代表パワー分布形状 $V_{Cent}(f)$ は, 4.3.4節で定義した. $V_{Cent}(f)$ は $P_{Cent}(t, f)$ の 100ms 間の時間方向の平均値と定義しているので, 膜鳴楽器の発音が密集していると, 周辺のスペクトルの影響を大きく受け, 単独発音から求めた $V_{Cent}(f)$ とは違ったものになる. BD, SD が 16 分音符で連続している個所では $V_{Cent}(f)$ がそれらの間の中間的なものになるので, クラスタリングの曖昧性が排除できない.

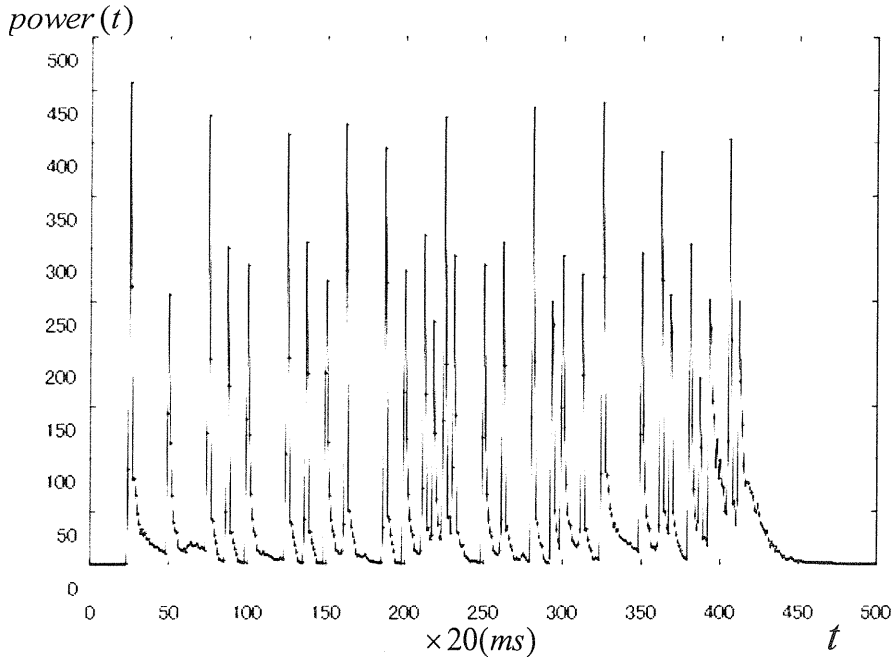


Figure 4.10: 低域通過フィルタ処理された後のパワーエンベロープ

このような発音の密集によるスペクトルの混合の問題は, 本研究で採用したパワー分布形状に基づく手法, あるいは従来研究のようなスペクトル上の特徴量に基づく手法のいずれにも, 混合音を扱う上では必然的に起こる. 教師なしクラスタリングは, 楽器の個体差の問題に対して非常に有効な手法であるものの, スペクトルの混合の問題を扱うのに十分ではない. 今後, 音符の遷移確率を導入することで, この問題に対処していきたい.

次に, 教師なしクラスタリング時に与えるクラス数について検討する. 本研究では, 使用楽器数としてクラス数は既知であるとした. つまり, テストデータ1に対しては4クラス分類問題を, テストデータ2に対しては5クラス分類問題を解くことによって, 音源同定を実現している. 実際には, クラス数が未知の場合もある. この問題に対処するために, まず2クラスへ大きく大別してから, 表4.4のような特徴量に基づき, 各打楽器名の同定を行う手法を検討する. このクラスタリングで, BD とそれ以外の2クラスか, BD, SD とそれ以外の2クラスに分かれることが予想される. 実験結果を表4.9に示す.

同じ音源を用いて作成したSC1とSC2のクラスタリング結果を比較する. 大まかに言えば, SC1では, BD, LT とそれ以外, SC2では, BD, LT, MT とそれ以外に分類されている. BD と LT は, 100Hz 以下の特に低い周波数域にパワーピークを持ち, $V_{Cent}(f)$ はよく似たものになることから同じクラスタに分類された. MT は HT と BD とを比べた場合, $V_{Cent}(f)$ は BD のものにより近いという結果になった. MU2でも同様のことがいえる. しかし, MU1だけは他とは様子が異なり, HT とそれ以外で2クラス分類された. HT だけスペクトル上のパワーピークが, 他のものと大きく離れたところにあり, パワーピークの位置が比較的近い HT 以外のの楽器全部で1つのクラスタにまとめられた. このように, 常に予想した通りの2クラス分類結果は得られるとは限らず, 予めどのようなクラスタが得られるかを定めておくのは困難である. この対処法は今後の検討課題とする.

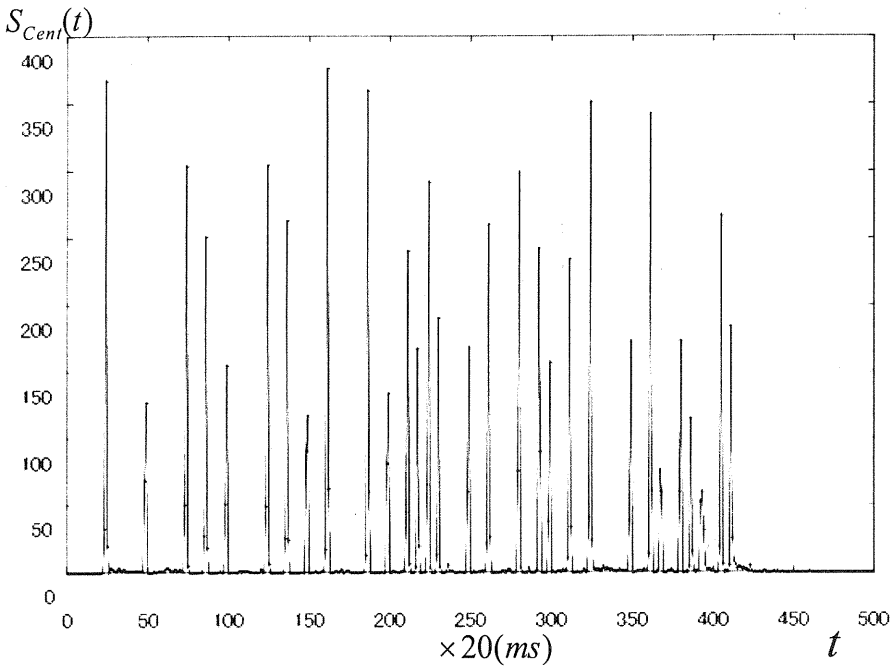


Figure 4.11: 各時刻におけるパワーの立ち上がり度

4.4.4 体鳴楽器の音源同定実験及び考察

4.3.6節の手法を評価するために体鳴楽器の音源同定実験を行った．周波数解析には，窓幅 4096 点，窓シフト 220 点として STFT を用いた．認識誤り補正時の適用個所の判定で用いる閾値は $\theta_1 = 0.5$, $\theta_1 = 0.4$ とした．評価用データ内の膜鳴楽器の内訳を表 4.10に，実験結果を表 4.11に示す．認識誤り補正法を適用する前 (k-NN 識別のみ) と補正後での識別結果をそれぞれ示してある．検出された発音時刻が，再現かつ適合したものに対して，音源同定率を求めている．

本研究で提案する認識誤り補正法の適用により，すべての音源に対して，音源同定精度の向上が見られ，本手法の有効性を示す結果となった．音源同定精度の改善の度合いは，SC-88VL, MU-2000 を音源として用いた場合に大きく，どちらも 50% 程度の認識誤り削減率を示した．また，市販 CD に対しては改善の度合いは大きくないように見える．この理由については後述する．

各評価データに対し，k-NN 識別で認識誤りが生じた個所に対して，認識誤り補正法がどのように適用されたかについて具体的に考察する．認識誤り個所の発見率を表 4.12に，認識誤り補正法を適用したもののうち，正しい結果に補正ができた個数と不必要な補正で誤りを増加させた個数を表 4.13に示す．また，適用された認識誤り補正法の内訳を各誤りパターンごとに集計したものを表 4.14に示す．学習により得られた認識誤り補正用識別機械の判定条件と補正の方法を表 4.15に示す (具体的な特徴量の説明は表 4.4参照)．

評価用データが SC1, SC2 の場合について考察する．認識誤り補正法の適用により正解数は増加し，補正前に正解であるところを不必要に補正し，逆に誤りを増加させる個所は見られなかった．認識誤り補正法の学習に用いたデータも SC-88VL で作成したものであるため，close 実験であり，妥当な結果が得られたと言える．また，認識誤りパターンは [IV], [II] が多く検出され，このパターンに対する認識誤り補正法が有効に働いていることが分かる．

次に，評価用データが MU1, MU2 の場合について考察する．全体として見ると音源同定率は大きく向上している．これは，認識誤りパターン [I], [III] が多く検出され，このパターンに対する認識誤り補正法が有効に働いているこ

Table 4.7: 評価用データ中に含まれる膜鳴楽器の内訳

	BD	SD	LT	MT	HT	合計
テストデータ 1	19	9	1	0	1	30
テストデータ 2	13	9	1	2	1	26
市販 CD	16	10	0	0	0	26

Table 4.8: 膜鳴楽器の音源同定結果

	音源同定率	各楽器の正解数				
		BD	SD	LT	MT	HT
SC1	90.0% (27/30)	17	8	1	0	1
SC2	92.3% (24/26)	11	9	1	2	1
MU1	90.0% (27/30)	17	8	1	0	1
MU2	88.5% (23/26)	12	7	1	2	1
市販 CD	100.0% (26/26)	16	10	0	0	0

とによる．しかし，認識誤り補正法の適用により，MU1, MU2 ともに 2 個所で誤りを増加させている．この誤り増加は，認識誤りパターン [IV] の補正が正しく機能せず，k-NN で HC と識別されたものに対し，不必要な HO への補正を行ったためである．表 4.15より，認識誤りパターン [IV] の補正の閾値は 0.265 である．認識誤り補正法の習用データと評価用データの音源が異なることで，認識誤り補正が最も有効に働く閾値が変化し，閾値をわずかに越えたものに対して，誤った補正が行われた．実際，MU1 の 2 個所から抽出した特徴量 FT21 の値は 0.267, 0.290 であり，閾値から非常に近いところにあった．

市販 CD に対しては，SD との重なりによる認識誤りパターン [I] の補正法は有効に働いた．しかし，音源同定率の改善の度合いは大きくないように見える．市販 CD の演奏データには HC が多く，k-NN 識別で多くが正しく HC と出力された．よって，もともと補正するべき箇所が少なく，間違って HO への不必要な補正を行う場合があり，正しい

Table 4.9: 2 クラスへ教師なしクラスタリングした実験結果

	SC1	SC2
クラス 1	BD 18, LT 1	BD 11, SD 1, LT 1, MT 2
クラス 2	BD 2, SD 8, HT 1	BD 2, SD 8, HT 1

	MU1	MU2
クラス 1	BD 20, SD 8, LT 1	BD 8, SD 1, LT 1, MT 2
クラス 2	HT 1	BD 4, SD 9, HT 1

Table 4.10: 評価用データ中に含まれる体鳴楽器の内訳

	CR	HO	HC	合計
テストデータ 1	2	3	24	29
テストデータ 2	2	5	21	28
市販 CD	0	6	23	29

Table 4.11: 体鳴楽器の音源同定結果

	補正前	補正後	認識誤り削減率
SC1	79.3% (23/29)	86.2% (25/29)	33.3% (2/7)
SC2	78.6% (22/28)	96.4% (27/28)	83.3% (5/6)
MU1	55.2% (16/29)	79.3% (23/29)	53.8% (7/13)
MU2	50.0% (14/28)	75.0% (21/28)	50.0% (7/14)
市販 CD	65.5% (19/29)	69.0% (20/29)	10.0% (1/10)

Table 4.12: 認識誤り個所の発見率

	再現率	適合率
SC1	83.3% (5/6)	71.4% (5/7)
SC2	100.0% (5/5)	100.0% (5/5)
MU1	100.0% (13/13)	66.7% (13/18)
MU2	100.0% (14/14)	77.8% (14/18)
市販 CD	80.0% (8/10)	80.0% (8/10)

Table 4.13: 認識誤り補正法の効果

	正しい補正による正解増加数	不必要な補正による誤り増加数
SC1	2 / 7	0 / 7
SC2	5 / 7	0 / 7
MU1	9 / 18	2 / 18
MU2	9 / 18	2 / 18
市販 CD	5 / 10	4 / 10

補正による認識誤り削減効果が減少してしまったからである．これは，MIDI 音源 1 種類での補正法の学習が十分でないためだと考えられる．今後，学習サンプル数を増やすことで対処できる．

これまで見てきたように，音源が異なれば誤りパターンの現れやすさも異なる．各音源ごとに現れやすい認識誤りパターンの傾向を表 4.16に示す．これより，ヒューリスティクスの改善により識別率向上が望めると考えられる．現れやすい認識誤りパターンの補正法のみを適用するようにヒューリスティクスを改良した場合の識別率を表 4.17に示す．大幅な改善が得られており，ヒューリスティクスを評価データに適応させていく手法も検討していく必要性がある．

また，今回は認識誤り補正法の学習に C5.0 を用い，表 4.15に示すように，1 種類の特徴量により判定する決定木が得られた．学習データのサンプル数も多くなく，補正するかしないかの 2 クラス識別であったためである．特徴量 1 つしか見ずに補正するかどうかを判定するため，MU1, MU2 のように，特徴量の値が閾値から非常に近いところにある場合，誤った補正判定を行う場合が考えられる．今後，学習サンプル数が増えれば，2 クラス判定が精度よく行えると言われる SVM (Support Vector Machine) やその他の統計的識別手法を採用することができる．特徴量ベクトル全

Table 4.14: 適用された認識誤り補正法の内訳

	[I]	[II]	[III]	[IV]	合計
SC1	0/0 (0/0)	1/0 (1/0)	0/2 (0/0)	1/3 (1/0)	2/5 (2/0)
SC2	0/0 (0/0)	1/0 (1/0)	0/0 (0/0)	4/0 (4/0)	5/0 (5/0)
MU1	6/0 (6/0)	0/3 (0/0)	3/0 (3/0)	2/4 (0/2)	11/7 (9/2)
MU2	6/0 (5/0)	0/2 (0/0)	6/0 (4/0)	2/2 (0/2)	14/4 (9/2)
市販 CD	2/0 (2/0)	1/0 (0/0)	3/0 (3/0)	4/0 (0/4)	10/0 (5/4)

* 結果が変わる個数 / 結果が変わらない個数 (正解増加数 / 誤り増加数) で示した.

Table 4.15: 認識誤り補正法の学習で得られた識別機械

	判定条件	補正の仕方
[I]	0.499 > FT21	CR→HC
[II]	1826892 < FT4	CR→HC
[III]	5299 > FT17	CR→HO
[IV]	0.265 < FT21	HC→HO

体を考慮にいれることができ, 認識誤り補正法の音源の違いに対する汎用性が増し, 補正の精度が向上すると期待できる.

統計的手法を採用することができれば, 本研究の k-NN 識別部を置き換えることができる. 体鳴楽器は膜鳴楽器ほど個体差による音色の違いはなく, 特徴量分布は正規分布に従うと予想される. そのため, 統計的手法を用いれば, 特徴量の分布を仮定しない k-NN 識別よりも音源同定率が向上する可能性がある. しかし, この場合でも音の重なりによる特徴量変動は避けられないので, 認識誤り補正法の導入により, さらに音源同定率の向上が見込める.

4.5 おわりに

本研究では, 打楽器の演奏を対象とした音源同定問題を扱った. 打楽器は, バスドラムのように膜の振動により発音する膜鳴楽器とシンバルのように金属振動により発音する体鳴楽器とに大別される. スペクトル分布の違いから, 帯域フィルタ処理により互いのスペクトル混合の影響を抑制し, その後, 各楽器類ごとに別々に設計された音源同定手法

Table 4.16: 音源による各認識誤りパターンの現れやすさ

音源	現れやすい認識誤りパターン
SC-88VL	[II], [IV]
MU-2000	[I], [III]
市販 CD	[I]

Table 4.17: ヒューリスティクスの改良前と改良後の音源同定率

	補正前	補正後	
		改良前	改良後
MU1	55.2% (16/29)	79.3% (23/29)	86.2% (25/29)
MU2	50.0% (14/28)	75.0% (21/28)	82.1% (23/28)
市販 CD	65.5% (19/29)	69.0% (20/29)	82.8% (24/29)

を適用するという音源同定手法を開発した.

音色のバリエーションに富み, 個体差が大きな膜鳴楽器の音源同定には, 教師なしクラスタリングを用いた. 大量のテンプレートが必要となるテンプレートマッチング法や, 従来通りの統計的手法は有効に働かないと予想される. 教師なしクラスタリングの採用により, 音源によらず, 90% 程度の音源同定率を達成でき, クラスタリングにより膜鳴楽器の音源同定ができることが示された. 本研究では, クラス数は既知としたが, 今後, クラス数未知でのクラスタリング手法や, 2 から 5 のクラス数でそれぞれクラスタリングを行い, 音楽知識を用いて, 最もありえそうなものを選択するといった手法について検討を進めていく.

また, 残響や SD との同時発音による特徴量変動が原因で, 認識誤りが生じやすい体鳴楽器の音源同定には, 認識誤り補正法を導入し, 認識率の向上を図った. 補正法の学習アルゴリズムとして決定木学習法 C5.0 を採用し, 自動学習が行えるようにした. また, 抽出した特徴量に対する信頼度を定義し, 信頼度に関するヒューリスティクスを与えることで, 補正法の適用個所の同定と適用すべき認識誤りパターンの判定も自動で行えるようにした. 補正法の導入により, 認識誤りが MIDI 音源に対しては 50% 程度, 市販 CD に対しては 10% 削減できた. 補正法の学習用サンプルの数が少なく, 補正法の汎用性が高くなかったため, 市販 CD に対しては効果は大きくない. これは, 今後学習サンプル数を増やすことで対応可能と考える.

本研究の提案する教師なしクラスタリングと認識誤り補正法が有効に機能することは実験により確認した. 今後, 学習サンプルを増やし, 識別や補正に有効な特徴量を考慮することで, さらに精度を上げることができると考える. 特徴量選択や次元圧縮の手法を検討していく予定である.

参考文献

[1] 安藤 由典: 楽器の音響学, 音楽之友社 (1996).

[2] K. W. Berger: Some Factors in the Recognition of Timbre, *J. Acoust. Soc. Am.*, Vol.36, No.10, pp.1888–1891 (1964).

[3] A. S. Bregman: *Auditory Scene Analysis*, MIT Press (1990).

[4] G. J. Brown and M. Cooke: Perceptual Grouping of Musical Sounds: A Computational Model, *J. New Music Research*, Vol.23, pp.107–132 (1994).

[5] Brown, J. C. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *Journal of Acoustic Society of America 103*, 3 (1999), 1933–1941.

[6] P. Cosi, G. D. Poli and G. Lauzzana: Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification, *J. New Music Research*, Vol.23, pp.71–98 (1994).

[7] ダイヤグラムグループ編, 皆川達夫監修: 『楽器』 (1992).

[8] 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)* (2000), IEEE, pp. 753–756.

[9] A. Eronen: Automatic Musical Instrument Recognition, M.Sc. Thesis, Tampere Univ. of Tech. (2001).

[10] 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.

[11] Eronen, A. and Klapuri, A.: Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of 2000 International Conference on Acoustics, Speech and Signal Processing (ICASSP-2000)*, pages 753–756. IEEE, 2000.

[12] *FFTW*. http://www.fftw.org/.

[13] I. Fujinaga and K. MacMillan: Realtime Recognition of Orchestral Instruments, In *Proceedings of International Computer Music Conference (ICMC)* (2000).

[14] 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, Vol.J77-D-II, No.5, pp.901–911 (1994).

[15] 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: 音楽ジャンルデータベースと楽器音データベース, 情処研報, 2002-MUS-45, pp.19–26 (2002).

[16] Guoyon, F. and Herrera, P.: Exploration of techniques for automatic labeling of audio drum tracks' instruments. In *Proceedings of Workshop on Current Directions in Computer Music*. MOSART, 2001.

[17] El-Hamdouchi, A. and Willet, P.: Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220–227, 1989.

[18] 原 祐一郎, 井口 征士: 複素スペクトルを用いた周波数同定, 計測論, Vol.19, No.9, pp.718–723 (1983).

[19] 早坂 寿雄: 『楽器の科学』, 電子情報通信学会 (1992).

[20] Herrera, P., Yeterian, A., and Gouyon, F.: Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques. In *Second International Confernece on Music and AI (ICMAI 2002)*, Lecture Notes in Artificial Intelligence 2445, pages 69–80. Springer Verlag, 2002.

[21] Perfecto Herrera, Alexandre Yeterian and Fabien Gouyon: Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques, *ICMAI*, LNAI2445, pp. 69–80 (2002).

[22] Iverson, P. and Krumhansl, C.L.: Isolating the dynamic attributes of musical timbre. *Journal of Acoustic Society of America*, 94(5):2595–2603, 1993.

[23] 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, Vol.J79-D-II, No.11, pp.1751–1761 (1996).

[24] 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.

[25] KASHINO, K., NAKADAI, K., KINOSHITA, T., AND TANAKA, H. Application of the bayesian probability network to music scene analysis. In *Computational Auditory Scene Analysis* (1998), D. Rosenthal and H. G. Okuno, Eds., Lawrence Erlbaum Associates, pp. 115–137.

[26] 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, 信学論, Vol.J81-D-II, No.7, pp.1510–1517 (1998).

[27] KASHINO, K., AND MURASE, H. A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication 27*, 3-4 (1999), 337–349.

[28] 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.

[29] 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, Vol.J83-D-II, No.4, pp.1073–1081 (2000).

[30] 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.

[31] 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.

[32] 北原鉄朗, 後藤真孝, 奥乃博: 楽器音を対象とした音源同定：音高による音色変化を考慮する識別手法の検討, *MUS-46-1*, 63, Vol. 2002, pp. 1–8 (2002).

[33] Kitahara, T., Goto, M., and Okuno, H.G:. Musical instrument identification based on f0-dependent multivariate normal distribution. In *Proceedings of 2003 International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003)*, IEEE, Vol.5, pp.421–424, Hong Kong, Apr. 2003.

[34] Marozeau, J.P., de Cheveigne, A., McAdams, S., and Winsberg, S.: The perceptual interaction between the pitch and timbre of musical sound. *Journal of Acoustic Society of America*, 109(5):2288, 2001.

[35] K. D. Martin: Sound-Source Recognition: A Theory and Computional Model, PhD Thesis, MIT (1999).

[36] 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.

[37] 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.

[38] NAKATANI, T., AND OKUNO, H. G. Sound ontology for computational auditory scene analysis. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)* (1998), AAAI, pp. 1004–1010.

[39] H. F. Olson (平岡 正徳訳) : 音楽工学, 誠文堂新光社 (1969).

[40] QUINLAN, J. R. *C4.5 Programs for Machine Learning.* Morgan Kaufmann, 1993.

[41] R. A. Rasch and R. Plomp (宮坂 栄一訳) : 楽音の知覚, 『音楽の心理学 (上) 』第 1 章, 西村書店 (1987).

[42] J. C. Risset and D. L. Wessel (宮坂 栄一訳) : 分析と合成による音色の探求, 『音楽の心理学 (上)』第 2 章, 西村書店 (1987).

[43] C. Roads (青柳 龍也 他訳) : 『コンピュータ音楽 — 歴史・テクノロジー・アート —』, 東京電機大学出版局 (2001).

[44] ROSENTHAL, D., AND OKUNO, H. G., Eds. *Computational Auditory Scene Analysis.* Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[45] 櫻庭洋平, 河原達也, 奥乃博: 音色情報と定位情報を用いた複数楽器音の認識, *MUS-46-1*, 情報処理学会, Vol. 2002, pp.1–8 (2002).

[46] Savitzky A. and Golay M.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Anal. Chem.*, 8, Vol. 36, pp. 1627–1630 (1964).

[47] 山口 公典, 安藤 繁雄: 短時間スペクトル分析法の自然楽器音への適用, 音響誌, Vol.33, No.6, pp.291–300 (1977).

[48] Weldin, L. and Goude. G.: Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psycology*, 13:228–240, 1972.

MUSICAL INSTRUMENT IDENTIFICATION BASED ON F0-DEPENDENT MULTIVARIATE NORMAL DISTRIBUTION

Tetsuro Kitahara,[†] *Masataka Goto,*[‡] and *Hiroshi G. Okuno*[†]

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

[‡]"Information and Human Activity", PRESTO,
JST / National Institute of Advanced
Industrial Science and Technology

kitahara@kuis.kyoto-u.ac.jp m.goto@aist.go.jp okuno@i.kyoto-u.ac.jp

ABSTRACT

The *pitch dependency* of timbres has not been fully exploited in musical instrument identification. In this paper, we present a method using an *F0-dependent multivariate normal distribution* of which mean is represented by a function of fundamental frequency (F0). This F0-dependent mean function represents the pitch dependency of each feature, while the F0-normalized covariance represents the non-pitch dependency. Musical instrument sounds are first analyzed by the F0-dependent multivariate normal distribution, and then identified by using the discriminant function based on the Bayes decision rule. Experimental results of identifying 6,247 solo tones of 19 musical instruments by 10-fold cross validation showed that the proposed method improved the recognition rate at individual-instrument level from 75.73% to 79.73%, and the recognition rate at category level from 88.20% to 90.65%.

1. INTRODUCTION

Musical instrument identification is an important subtask for many applications including auditory scene analysis and multimedia retrieval as well as for reducing ambiguities in automatic music transcription. The difficulties in musical instrument identification reside in the fact that some features depend on pitch and individual instruments. In particular, timbres of musical instruments are obviously affected by the pitch due to their wide range of pitch. For example, the pitch range of the piano covers over seven octaves.

To attain high performance of musical instrument identification, it is indispensable to cope with this *pitch dependency* of timbre. Most studies on musical instrument identification, however, have not dealt with the pitch dependency [1]–[6]. Martin used 31 features including spectral and temporal features with hierarchical classification and attained about 70% of identification by the benchmark of 1,023 solo tones of 14 instruments. He pointed out the importance of the pitch dependency, but left it as future work [1]. Eronen *et al.* used spectral and temporal features as well as cepstral coefficients used by Brown [2] and attained about 80% of identification by the benchmark of 1,498 solo tones of 30 instruments [3]. They treated the pitch as one element of feature vectors, but did not cope with the pitch dependency. Kashino *et al.* also treated the pitch similarly in their automatic music transcription system [4]. They also coped with the difference of individual instruments, but did not deal with the pitch dependency [5].

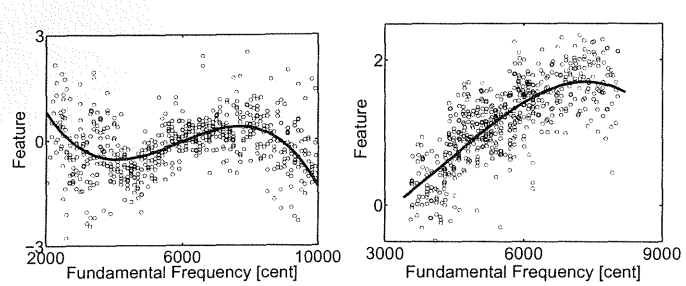
In this paper, to take into consideration the pitch dependency of timbre in musical instrument identification, each feature or basic vector of features is represented by an *F0-dependent multivariate normal distribution* of which mean is represented by a function of fundamental frequency (F0). This *F0-dependent mean function* represents the pitch dependency of each feature, while the *F0-normalized covariance* represents the non-pitch dependency. Musical instrument identification is performed both at individual-instrument level and at non-tree category level by a discriminant function based on the Bayes decision rule.

The rest of this paper is organized as follows: Section 2 proposes the F0-dependent multivariate normal distribution, and Section 3 describes the features and the discriminant function used in this paper. Sections 4 and 5 report the experimental results, and finally Section 6 concludes this paper.

2. F0-DEPENDENT MULTIVARIATE NORMAL DISTRIBUTION

The distribution of tone features in the feature space is represented by an *F0-dependent multivariate normal distribution* with two parameters: the *F0-dependent mean function* and *F0-normalized covariance*. The reason why the mean of the distribution is approximated as a function of F0, that is an *F0-dependent mean function*, is that tone features at different pitches have different positions (means) of distributions in the feature space. In this paper, the F0-dependent

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research (B), No.12480090, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan)



(a) Piano's 4th basic vector of features. (b) Cello's first basic vector of features.

Fig. 1. Examples of F0-dependent mean functions.

mean function for each musical instrument ω_i, $\boldsymbol{\mu}_i(f)$, is approximated as a cubic polynomial by using the least squares method. For example, piano's fourth basic vector of features and cello's first basic vector are depicted in **Fig. 1** (a) and (b), respectively.

On the other hand, the non-pitch dependency of each feature is represented by the *F0-normalized covariance*. Since the F0-dependent mean function represents the mean of features, the covariance obtained by subtracting the mean from each feature eliminates the pitch dependency of features. For each musical instrument ω_i, the F0-normalized covariance Σ_i is defined as follows:

$$\Sigma_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \chi_i} (\boldsymbol{x} - \boldsymbol{\mu}_i(f_{\boldsymbol{x}}))(\boldsymbol{x} - \boldsymbol{\mu}_i(f_{\boldsymbol{x}}))',$$

where $'$ is the transposition operator, χ_i and n_i are the set of the training data of the instrument ω_i and its total number, respectively. $f_{\boldsymbol{x}}$ denotes the F0 of the data $\boldsymbol{x}$.

3. FEATURES AND A DISCRIMINANT FUNCTION

3.1. Features for Musical Instrument Identification

We used spectral, temporal, and modulation features as well as non-harmonic component features resulting in 129 features in total listed in **Table 1**. The features except the non-harmonic component features are determined by consulting the literatures [1, 3, 4]. The non-harmonic component features are original and have not been used in the literature. We incorporated features as many as possible, since the feature space is transformed to a lower-dimensional space.

Each musical instrument sound sampled by 44.1 kHz with 16 bits are first analyzed by STFT (short time Fourier transform) with Hanning windows (4096 points) for every 10 ms, and spectral peaks are extracted from the power spectrum. Then, the F0 and the harmonic structure is obtained from these peaks.

The number of dimensions of the feature space is reduced by principal component analysis (PCA): the 129-dimensional space is reduced to a 79-dimensional space with the proportion value of 99%. It is further reduced to the minimum dimension by linear discriminant analysis (LDA).

Table 1. Overview of 129 features.

(1)	**Spectral features** (40 features)
	e.g., Spectral centroid, Relative power of the fundamental component, Relative power in odd and even components
(2)	**Temporal features** (35 features)
	e.g., Gradient of a straight line approximating power envelope, Average differential of power envelope during onset
(3)	**Modulation features** (32 features)
	e.g., Amplitude and frequency of AM, FM, modulation of spectral centroid and modulation of MFCC
(4)	**Non-harmonic component features** (22 features)
	e.g., Temporal mean of kurtosis of spectral peaks of each harmonic component (Their values become lower as sounds contain more non-harmonic components.)

In this paper, the space is reduced to an 18-dimensional space, since we deal with 19 instruments.

3.2. A Discriminant Function for the F0-dependent Multivariate Normal Distribution

Once parameters of the F0-dependent multivariate normal distribution are estimated, the Bayes decision rule is applied to identify the musical instrument or category of instruments. The discriminant function $g_i(\boldsymbol{x}; f)$ for the musical instrument ω_i is defined by

$$g_i(\boldsymbol{x}; f) = \log p(\boldsymbol{x}|\omega_i; f) + \log p(\omega_i; f), \qquad (1)$$

where $\boldsymbol{x}$ is an input data, $p(\boldsymbol{x}|\omega_i; f)$ is a probability density function (PDF) of this distribution and $p(\omega_i; f)$ is a priori probability of the instrument ω_i.

The PDF of this distribution is defined by

$$p(\boldsymbol{x}|\omega_i; f) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2} D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f))\right\}, \qquad (2)$$

where d is the number of dimensions of the feature space and D^2 is the squared Mahalanobis distance defined by

$$D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) = (\boldsymbol{x} - \boldsymbol{\mu}_i(f))' \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i(f)).$$

Substituting equation (2) into equation (1), thus, generates the discriminant function $g_i(\boldsymbol{x}; f)$ as follows:

$$g_i(\boldsymbol{x}; f) = -\frac{1}{2} D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) - \frac{1}{2} \log |\Sigma_i| - \frac{d}{2} \log 2\pi + \log p(\omega_i; f).$$

The name of the instrument that maximizes this function, that is ω_k satisfying $k = \mathrm{argmax}_i\, g_i(\boldsymbol{x}; f)$, is determined as the result of musical instrument identification.

The a priori probability $p(\omega_i; f)$ represents whether the pitch range of the instrument ω_i includes f, that is,

$$p(\omega_i; f) = \begin{cases} 1/c & (\text{if } f \in R_i) \\ 0 & (\text{if } f \notin R_i) \end{cases}$$

where R_i is the pitch range of the instrument ω_i, and c is the normalizing factor to satisfy $\sum_i p(\omega_i; f) = 1$.

Table 2. Contents of the database used in this paper.

Instrument names	Piano (PF), Classical Guitar (CG), Ukulele (UK), Acoustic Guitar (AG), Violin (VN)!$ Viola (VL), Cello (VC), Trumpet (TR), Trombone (TB), Soprano Sax (SS), Alto Sax (AS), Tenor Sax (TS), Baritone Sax (BS)!$ Oboe (OB), Fagotto (FG), Clarinet (CL), Piccolo (PC), Flute (FL), Recorder (RC)
Individuals	3 individuals except TR, OB, FL. TR, OB, FL: 2 individuals.
Intensity	Forte, normal, piano.
Articulation	Normal articulation style only.
Number of tones	PF: 508, CG: 696, UK: 295, AG: 666, VN: 528, VC: 558, TR: 151, TB: 262, SS: 169, AS: 282, TS: 153, BS: 215, OB: 151, FG: 312, CL: 263, PC: 245, FL: 134, RC: 160.

Table 3. Categorization of 19 instruments.

Categories	Instruments
Piano	Piano
Guitars	Classical Guitar, Ukulele, Acoustic Guitar
Strings	Violin, Viola, Cello
Brasses	Trumpet, Trombone
Saxophones	Soprano Sax, Alto Sax, Tenor Sax, Baritone Sax
Double Reeds	Oboe, Faggoto
Clarinet	Clarinet
Air Reeds	Piccolo, Flute, Recorder

4. EXPERIMENTS AND RESULTS

4.1. Experimental Conditions

Musical instrument identification is performed not only at individual-instrument level but also at category level to evaluate the improvement of recognition rates by the proposed method based on the F0-dependent multivariate normal distribution. The recognition rate was obtained by 10-fold cross validation. We compared the results by the method using usual multivariate normal distribution (called *baseline*) with those by the method using the proposed F0-dependent multivariate normal distribution (called *proposed*).

The benchmark used for evaluation is a subset of the large musical instrument sound database RWC-MDB-I-2001 developed by Goto *et al.* [7, 8]. This subset summarized in **Table 2** was selected by the quality of recorded sounds and consists of 6,247 solo tones of 19 orchestral instruments. All data are sampled by 44.1 kHz with 16 bits.

The categories of musical instruments summarized in **Table 3** are determined based on the sounding mechanism of instruments and existing studies [1, 3]. The category of instruments is useful for some applications including music retrieval. For example, when a user wants to find a piece of piano solo on a music retrieval system, the system can reject pieces containing instruments of different categories, which can be judged without identifying individual instrument names.

4.2. Results of Musical Instrument Identification

Table 4 summarizes the recognition rates by both the *baseline* and *proposed* methods. The proposed F0-dependent method improved the recognition rate at individual-instrument level from 75.73% to 79.73% and reduced recognition errors by 16.48% in average. At category level, the proposed method improved the recognition rate from 88.20% to 90.65% and reduced recognition errors by 20.67%. The observation of these experimental results is summarized below:

Improvement by the pitch dependency

The recognition rates of six instruments (PF, TR, TB, SS, BS, and FG) were improved by more than 7%. In particular, the recognition rate for pianos was improved by 9.06%, and its recognition errors were reduced by 35.13%. This big improvement was attained, since their pitch dependency is salient due to their wide range of pitch.

Difference between accuracy at two levels

The recognition rates of the four types of saxophones at individual-instrument level (47–73%) were lower than those at category level (77–92%). This is because sounds of those saxophones were quite similar. In fact, Martin reported that sounds of various saxophones are very difficult for the human to discriminate [1].

Instrument-dependent difficulty of identification

Since we adopt the flat (non-hierarchical) categorization, the recognition rates at category level depend on the category. The recognition rates of guitars and strings at category level were more than 94%, while those of brasses, saxophones, double reeds, clarinet and air reeds were about 70–90%. This is because instruments of these categories have similar sounding mechanism: these categories are subcategories of "wind instruments" in conventional hierarchical categorization.

5. EVALUATION OF THE BAYES DECISION RULE

The effect of the Bayes decision rule in musical instrument identification was evaluated by comparing with the 3-NN rule (3-nearest neighbor rule) with/without LDA. Three variations of the dimension reduction are examined:

(a) reduction to 79 dimension by PCA,
(b) reduction to 18 dimension by PCA, and
(c) reduction to 18 dimension by PCA and LDA.

The last one is adopted in the proposed method.

The experimental results listed in **Table 5** showed that the Bayes decision rule performed better in average than the 3-NN rule. Some observation are as follows:

Table 4. Accuracy by usual distribution (baseline) and F0-dependent distribution (proposed).

	Individual-instrument level			Category level		
	Usual	F0-dpt	diff.	Usual	F0-dpt	diff.
PF	74.21%	83.27%	+9.06%	74.21%	83.27%	+9.06%
CG	90.23%	90.23%	±0.00%	97.27%	97.13%	−0.14%
UK	97.97%	97.97%	±0.00%	97.97%	98.31%	+0.34%
AG	81.23%	83.93%	+2.70%	94.89%	95.65%	+0.76%
VN	69.70%	73.67%	+3.97%	98.86%	99.05%	+0.19%
VL	73.94%	76.27%	+2.33%	93.22%	94.92%	+1.70%
VC	73.48%	78.67%	+5.19%	95.16%	96.24%	+1.08%
TR	73.51%	82.12%	+8.61%	76.82%	85.43%	+8.61%
TB	76.72%	84.35%	+7.63%	85.50%	89.69%	+4.19%
SS	56.80%	65.89%	+9.09%	73.96%	80.47%	+6.51%
AS	41.49%	47.87%	+6.38%	73.76%	77.66%	+3.90%
TS	64.71%	66.01%	+1.30%	90.20%	92.16%	+1.96%
BS	66.05%	73.95%	+7.90%	81.40%	86.05%	+4.65%
OB	71.52%	72.19%	+0.67%	75.50%	74.83%	−0.67%
FG	59.61%	68.59%	+8.98%	64.74%	71.15%	+6.41%
CL	90.69%	92.07%	+1.38%	90.69%	92.07%	+1.38%
PC	77.56%	81.63%	+4.07%	89.39%	90.20%	+0.81%
FL	81.34%	85.07%	+3.73%	82.09%	85.82%	+3.73%
RC	91.88%	91.25%	−0.63%	92.50%	91.25%	−1.25%
Ave.	75.73%	79.73%	+4.00%	88.20%	90.65%	+2.45%

Usual: Usual (F0-independent) distribution (baseline)
F0-dpt: F0-dependent distribution (proposed)

Table 5. Accuracy by 3-NN rule and the Bayes decision rule.

	3-NN rule			Bayes decision rule		
	(a)	(b)	(c)	(a)	(b)	(c)
PF	53.94%	46.46%	63.39%	55.91%	59.06%	83.27%
CG	79.74%	77.16%	75.72%	98.28%	97.27%	90.23%
UK	94.58%	92.54%	97.63%	67.12%	80.00%	97.97%
AG	95.05%	92.79%	97.00%	19.97%	44.14%	83.93%
VN	47.73%	46.02%	45.83%	89.58%	84.47%	73.67%
VL	55.93%	54.24%	61.86%	71.19%	79.24%	76.27%
VC	86.20%	85.84%	84.23%	45.16%	30.82%	78.67%
TR	36.42%	38.41%	47.02%	41.72%	72.85%	82.12%
TB	70.99%	54.58%	77.86%	75.19%	78.24%	84.35%
SS	23.08%	14.20%	24.85%	48.52%	66.86%	65.89%
AS	37.59%	29.79%	40.43%	72.70%	41.84%	47.84%
TS	62.09%	66.01%	68.63%	30.07%	61.44%	66.01%
BS	68.84%	67.91%	66.98%	55.35%	54.42%	73.95%
OB	47.68%	48.34%	49.01%	43.71%	81.46%	72.19%
FG	64.10%	65.06%	74.36%	40.38%	30.12%	68.59%
CL	93.45%	87.93%	93.10%	95.51%	93.45%	92.07%
PC	84.08%	84.90%	84.08%	63.27%	58.37%	81.63%
FL	88.06%	72.39%	94.03%	35.82%	84.33%	85.07%
RC	97.50%	93.75%	97.50%	85.00%	96.25%	91.25%
Ave.	70.27%	66.98%	72.53%	62.11%	66.50%	79.73%

(a) Dimensionality reduction to 79 dim. using PCA only
(b) Dimensionality reduction to 18 dim. using PCA only
(c) Dimensionality reduction to 18 dim. using both PCA and LDA

(1) The Bayes decision rule with 79-dimension showed poor performance for AG, TR, SS, TS, OB and FL, since the number of their training data is not enough for estimating parameters of a 79-dimensional normal distribution. For such small training sets with 79-dimension, 3-NN is superior to the Bayes decision rule.

(2) LDA with the Bayes decision rule improved the accuracy of musical instrument identification from 66.50% to 79.73% in average. Although it seemed that PCA with 79-dimension performed better than LDA for CG, VN and AS, the cumulative performance of LDA for the categories of strings and saxophones is better than that of PCA.

6. CONCLUSIONS

In this paper, we presented a method for musical instrument identification using the *F0-dependent multivariate normal distribution* which takes into consideration the pitch dependency of timbre. The method improved the recognition rates at individual-instrument level from 75.73% to 79.73%, and at category level from 88.20% to 90.65% in average, respectively. The Bayes decision rule with dimension reduction by PCA and LDA also performed better than the 3-NN method.

Future works include evaluation of the method with different styles of playing, evaluation of the robustness of each feature against mixture of sounds, and automatic music transcription.

Acknowledgments: We thank everyone who has contributed to building and distributing the RWC Music Database (Musical Instrument Sound: RWC-MDB-I-2001) [7, 8]. We also thank Kazuhiro Nakadai and Hideki Asoh for their valuable comments.

7. REFERENCES

[1] K. D. Martin, "Sound-Source Recognition: A Theory and Computational Model," PhD Thesis, MIT, 1999.

[2] J. C. Brown, "Computer Identification of Musical Instruments Using Pattern Recognition with Cepstral Coefficients as Features," *J. Acoust. Soc. Am.*, **103**, 3, pp.1933–1941, 1999.

[3] A. Eronen and A. Klapuri, "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features," *Proc. of ICASSP*, pp.753–756, 2000.

[4] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, "Application of the Bayesian Probability Network to Music Scene Analysis," *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno (eds.), Lawrence Erlbaum Associates, pp.115–137, 1998.

[5] K. Kashino and H. Murase, "A Sound Source Identification System for Ensemble Music Based on Template Adaptation and Music Stream Extraction," *Speech Communication*, **27**, pp.337–349, 1999.

[6] I. Fujinaga and K. MacMillan, "Realtime Recognition of Orchestral Instruments," *Proc. of ICMC*, 2000.

[7] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," *IPSJ SIG Notes*, 2002-MUS-45, pp.19–26, 2002. *(in Japanese)*

[8] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," *Proc. of ISMIR 2002*, pp.287–288, 2002.