社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－MUS－61（13）
2005／8／5

HMM と MFCC を用いた楽器音の音源同定の検討

馬場 貴之†　山田 武志†　北脇 信彦†

† 筑波大学大学院システム情報工学研究科

〒305-8573 茨城県つくば市天王台 1-1-1

E-mail: †i011214@coins.tsukuba.ac.jp

あらまし　自動採譜や音楽検索において,楽器音の音源同定は重要な要素技術である.本稿では,単一楽器の孤立発音を対象とした，HMM と MFCC を用いた楽器音の音源同定について述べる．まず，適切な HMM のパラメータを決定するために，RWC の楽器音データベースを用いて同定実験を行った．その結果，HMM の状態数と混合分布数が各々9，12 のときに 88.65%の同定率が得られることが分かった．次に，学習データの量が同定率に及ぼす影響を調査し，学習データの量と同定率の関係を明らかにした．

Examination of Musical Instrument Identification Using HMM and MFCC

Takayuki BABA†　Takeshi YAMADA†　Nobuhiko KITAWAKI†

† Graduate School of Systems and Information Engineering,
University of Tsukuba

1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan

E-mail: †i011214@coins.tsukuba.ac.jp

Abstract　Musical instrument identification is one of key technologies for automatic music transcription and music information retrieval. In this paper, we describe musical instrument identification using HMM and MFCC for monophonic sound. First, we conduct an experiment on identification using HMM and MFCC to investigate the proper parameter of HMM. From the result, we obtained the identification rate of 88.65% when a number of states and Gaussian mixtures are 9 and 12, respectively. Second, we examined the effect of an amount of training data on the identification rate and showed the relationship by performing an experiment.

1 まえがき

近年，音楽圧縮技術の向上によるディジタル音楽配信の普及に伴い，膨大な数の音楽音響信号がインターネット上に蓄積されつつある．一般のユーザがこれらの音楽音響信号の中から目的の音楽を的確かつ素早く検索するためには，個々の音楽音響信号に楽曲名やアーティスト名，メロディーといったメタデータを付与することが有効である．ところが，インターネット上に既に存在する音楽音響信号に対して人手でメタデータを付与することは,コストの点から現実的ではない.よって,音楽音響信号のみから自動的にこれらの情報を検出する技術，とりわけ自動採譜の技術が必要となる．自動採譜は音楽音響信号の複雑さ，多様さのためにまだ実用には至っておらず，様々な課題が残されている[1].

本稿では，自動採譜のための要素技術の一つである楽器音の音源同定を扱う．従来，単一楽器の孤立発音，複数楽器の同時発音を対象とし，様々な音源同定手法が提案されている[2-7]．前者については，適応型混合テンプレートを用いて楽器の個体差や音高の揺らぎを吸収する手法[2]や，音高による音色変化をモデル化する手法[3]などが提案されており

表 1 実験に用いたデータ

楽器名	楽器記号	カテゴリー	音域	強さ	奏法	データ数
ピアノ	PF	ピアノ	A0-C8	それぞれ強･中･弱の3種類	通常の奏法のみ	790
クラシックギター	CG	ギター	E2-E5			2807
ウクレレ	UK		F3-A5			468
アコースティックギター	AG		E2-E5			2808
バイオリン	VN	弦楽器	G3-E7			576
ビオラ	VL		C3-F6			540
チェロ	VC		C2-F5			565
トランペット	TR	金管楽器	E3-A#6			302
トロンボーン	TB		A#1-F#5			286
ソプラノサックス	SS	サックス	G#3-E6			297
アルトサックス	AS		C#3-A5			297
テナーサックス	TS		G#2-E5			294
バリトンサックス	BA		C2-A4			298
オーボエ	OB	複こう楽器	A#3-G6			288
ファゴット	FG		A#1-D#5			360
クラリネット	CL	クラリネット	D3-F6			360
ピッコロ	PC	無こう楽器	D5-C8			290
フルート	FL		C4-C7			332
リコーダー	RC		C4-B6			225

10～30 種類の孤立発音に対して，70～80%程度の同定率が得られたことが報告されている[5]．後者については，egginkらが，5種類の楽器の2重奏に対して，49%の同定率が得られたことを報告している[6]．このように単一楽器の孤立発音でさえも十分な同定率が得られていないのが現状である．

これまでに，楽器音の音源同定に隠れマルコフモデル（HMM）を適用することが検討されている[7]．HMMは非定常信号源を定常信号源の連鎖として表現するモデルであり，音声信号と同様に音楽音響信号に対しても有効であると考えられている．

一般に，HMMの学習のためには大量のデータが必要であることが知られている．そこで，本稿では，単一楽器の孤立発音を対象とし，楽器音の同定のためにはどの程度の学習データが必要となるのかを調査する．また，従来は独自に収集したデータを用いて性能評価が行われていたために，HMMの有効性が必ずしも明らかではなかった．しかし，最近になって，様々な楽器の孤立発音や同時発音を収録したデータベース[8]がリリースされ，同じ条件下での性能比較ができるようになった．そこで，本稿では，このデータベースを用いて行われた先行研究[3]との比較によりHMMの有効性を評価する．なお，本稿では，特徴量としては音声認識で一般に用いられているメルケプストラム（MFCC）を用いることとし，今後はこれをベースラインとして特徴量の検討を進めていくことを考えている．

以下，2章では本稿で用いたデータについて述べる．3章ではHMMの状態数と混合分布数を実験的に決定し，4章では学習に必要なデータの量を調査する．最後に，5章でまとめと今後の課題を述べる．

2 実験に用いたデータ

実楽器の単音データベースとしてRWC研究用データベースの楽器音データベースRWC-MDB-I-2001[8]を用いた．楽器の種類については，先行研究[3]に準拠し，オーケストラで一般的に使用される楽器のうち，打楽器等を除いた19種類とした．本実験で用いたデータの詳細を表1に示す．各楽器には楽器個体，もしくは奏者が3つずつ用意されており，同じ楽器の同じ音高のファイルが3つ存在している．今回はそれらをすべて用いた．強さについては強，中，弱の3つ，奏法については通常の奏法のみを用いている．また，音域に

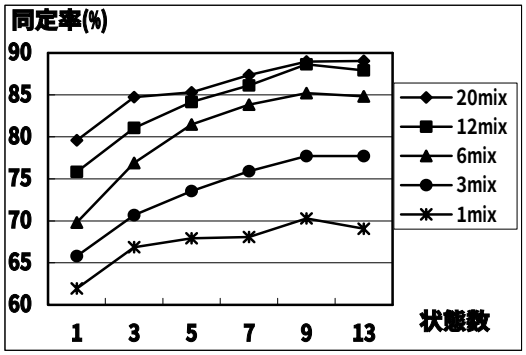

図 1 HMM のパラメータと同定率の関係

ついては表 1 に示す通りである．さらに，カテゴリーについては先行研究[3]と同様とした．これは，発音構造や楽器特有のエンベロープによって，よく似た音色を持つ楽器ごとにカテゴライズしたものである．楽器音データベースに収録されている 1 ファイルには，複数の単発音が含まれているため，無音部検出によって自動的に分割した．具体的には，振幅が 0 の部分が一定数続いた部分は単発音間の区切り区間とみなし，その部分を取り除くことによって分割した．よって分割後のファイルは,「収録時のノイズ＋楽器音＋収録時のノイズ」という形になっている．

3 HMM のパラメータの検討

HMM を用いて楽器音の音源同定を行う際には，特徴量と HMM のパラメータ（状態数や混合分布数）を適切に決める必要がある．そこで，本章では，特徴量としては音声認識で一般に用いられている MFCC を用いることとし，その条件の下で HMM の状態数と分布数を実験的に決定する．

3.1 実験条件

3.1.1 MFCC のパラメータ

MFCC の計算の際には，表 2 に示すようなパラメータを設定する必要がある．表 2 には，音声信号（8kHz サンプリング）に対する設定値の典型例と，音楽音響信号（44.1kHz サンプリング）に対して本実験で設定した値を示している．ここで，後者の場合のメルフィルタバンクは，音声帯域においては音声信号と同一，それ以上の帯域についてはその拡張となっている．

表 2 MFCC の諸元

	音声信号の設定値	音楽音響信号の設定値
メルフィルタバンク数	23	42
MFCC 次数	12	22
リフタ長	22	40
フレーム長	25msec	25msec
フレーム周期	10msec	10msec
高域強調	on	off

表 3 HMM の諸元

状態数	1,3,5,7,9,13
混合分布数	1,3,6,12,20
モデル	各楽器(PF,CG,…) 無音(sil)

最終的な特徴量は，MFCC 22 次元，パワー1 次元，及びそれらの一次微分と二次微分を合わせた計 69 次元である．

3.1.2 HMM のパラメータ

HMM の条件を表 3 に示す. HMM の状態数として 1，3，5，7，9，13 の 6 種類，混合分布数として 1，3，6，12，20 の 5 種類を考え，その全ての組み合わせについてモデルを学習する．モデルは楽器名(pf,cg,…)を単位として用意し，別途，音楽音響信号の前後の無音部を表すモデル(sil)を学習した．

学習には表 1 に示したデータのうちの 9 割を用いた．学習データの量が同定性能に及ぼす影響については 4 章で議論する．

3.2 実験結果と考察

学習に用いなかった残りのデータ(全体の 1 割）を用いて，状態数，混合分布数が異なるすべての条件において同定実験を行った．実験結果を図 1 に示す.

ここで，図 1 の縦軸は楽器毎の同定率の平均，横軸は状態数である．また，記号の違いは，混合分布数が異なることを表す．

図 1 より，状態数と混合分布数が増えるにつれて同定率が大きく向上していることが分かる．一方，同定率の伸びは，状態数が 9，混合分布数が 12 のときに頭打ちとなっていることが見て取れる．そのときの同定率は 88.65%であり，先行研究[3]の 79.73%に比べ

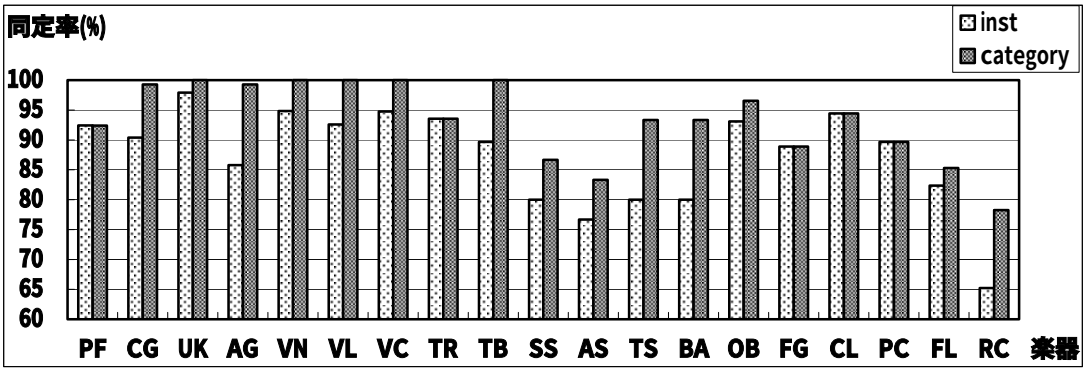

図 2 各楽器の楽器レベルとカテゴリーレベルにおける同定率

て 9%ほど高いことが分かった．ただし，本実験で用いたデータは先行研究[3]に準拠しているものの，完全に同一ではないので，参考としての比較である．

図 2 に各楽器の楽器毎の同定率とカテゴリー毎の同定率を示す．カテゴリー毎の同定では，同カテゴリーに属する楽器に誤って同定した場合には，それを正解とみなす．例えば，クラシックギターを同定した結果がアコースティックギターとなっていた場合，楽器レベルでは不正解となる．しかし，同じギターカテゴリーへの誤りであるため，カテゴリーレベルでは正解とみなす．

まず、楽器毎の同定率について見ると，楽器によってかなりのばらつきがあることが分かる．同定率が低下している原因を詳しく調べたところ，同じカテゴリーに属する別の楽器に誤って同定していることが分かった．例えば，サックス系の楽器（SS，AS，TS）は人間が同定することさえも困難であることが知られている．今回用いた MFCC ではこれらの楽器の違いを適切に表現できていないと考えられるので，適切な特徴量を導入する必要がある．

次に，状態数 9，混合分布数 12 のカテゴリー毎の同定率の平均は 93.38%であり，楽器毎の同定率に比べて当然ながら良い結果となっている．特に、ギター系の楽器においてはほぼ 100%の同定率が得られていることが分かる．サックス系の楽器についても同定率に向上が見られた．

最後に，HMM の状態数について考察する．図 3 は楽器音のエンベロープを示している．これは，楽器音が鳴り始めてから鳴り終わるまでの音量・音色・音程などの時間的な変化を表したもので，図 3 の場合，縦軸がパワー，横軸が時間を表している．特に音量の時間的変化は特徴的で，減衰系（ピアノなど），持続系（ストリングスなど），パーカッション系といったものでそれぞれ大きく異なる．この知見によって，状態数を 4 程度にするという考え方がある[7]．一方，本実験ではそれ以上の状態数が必要であるとの結果になった．適切な状態数は，どのような特徴量を用いるのかによって変化することが想定されることから，今後特徴量の検討を進める際には，HMM のパラメータの変化についても十分に考慮する必要がある．

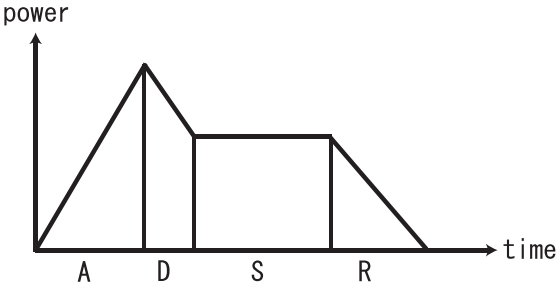

図 3 楽器音のエンベロープ

4 学習データ量の検討

前章では，HMM を用いることにより 88.65%という高い同定率が得られることを示した．しかし，HMM の学習には大量のデータが必要であり，学習データの量が少ない場合には同定性能が低下してしまう．そこで，本章では，学習データの量と同定率の関係を調査する．

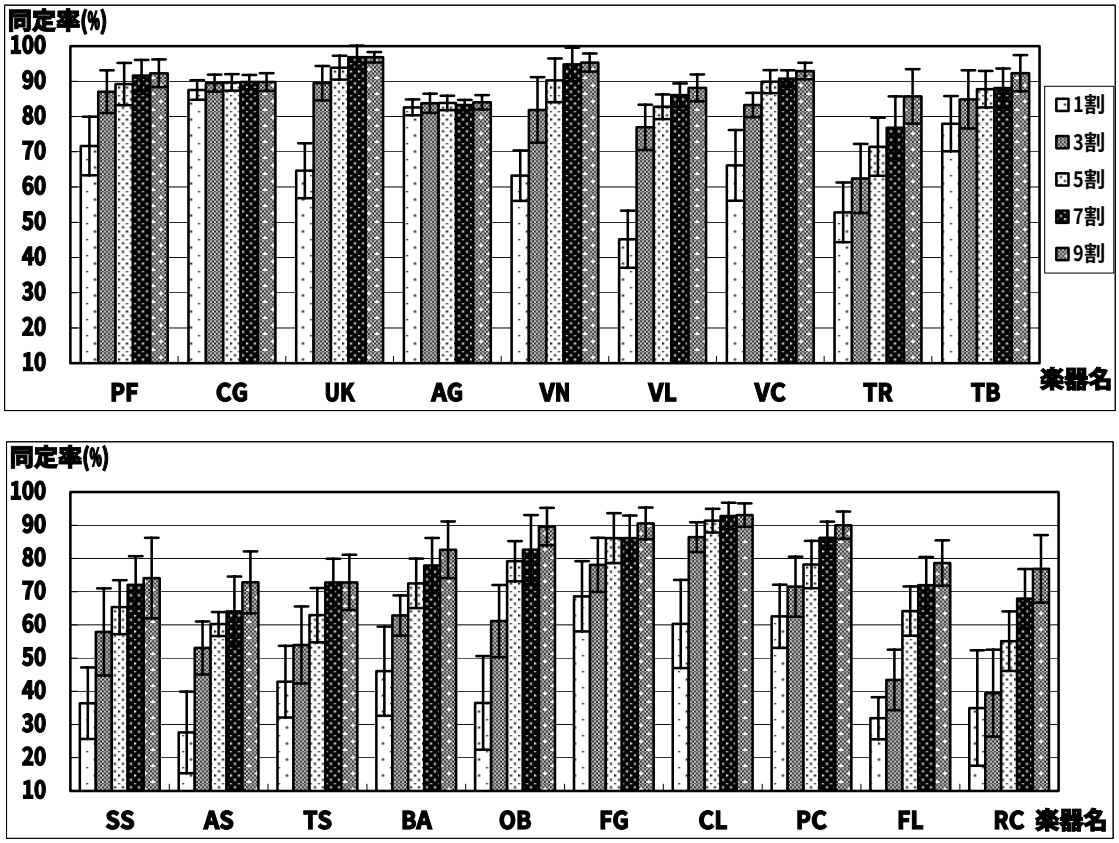

図 4 学習データの量を変化させた時の楽器ごとの同定率と標準偏差

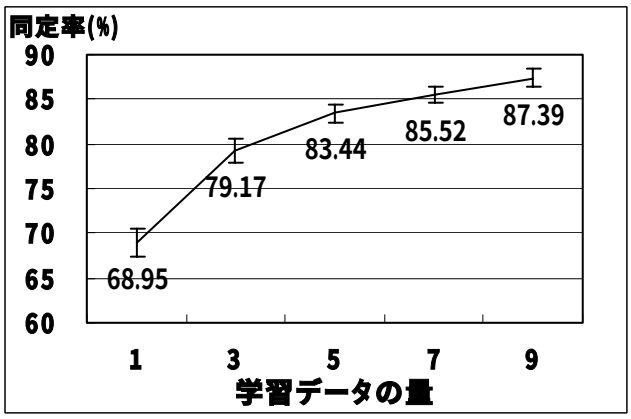

図 5 学習データの量を変化させた時の楽器ごとの同定率と標準偏差の平均

4.1 実験条件

表 1 に示したデータを 10 等分し, 10 個のデータセットを作成した. そのうちの 1 セット, 3 セット, 5 セット, 7 セット, 9 セットを学習データ, 学習に用いなかったデータセットの 1 つを評価データとして用いた. 学習データの違いによる同定率の偏りを防ぐため, 10 個のデータセットの各々を評価用データとしてクロスバリデーションを行った.

特徴量は 3.1.1 節で述べたものと同一である. また HMM のパラメータは 3.2 節の結果から状態数を 9, 混合分布数を 12 とした.

4.2 実験結果と考察

学習データの量を変化させたときの楽器毎の同定率を図 4 に示す. また, 楽器毎の同定率の平均を図 5 に示す. ここで, 同定率は,

クロスバリデーションにおける評価データセット毎の同定率の平均であり，その標準偏差を併記している．

図 4 と図 5 より，学習データの量が少なくなるにつれて，同定率が低下し，また標準偏差が大きくなっていくことが分かる．図 5 より，一つの目安として 80%程度の同定率を得るためには，5 割の学習データが必要であることが見て取れる．このときの同定率は，学習データとしてデータ全体の 9 割を用いた場合と比べて 3 ポイント程度の低下である．ただし，本実験では，音高や音色の異なる学習データがバランス良く配分されていたことに注意が必要である．

また，図 4 より，楽器毎に見ると，学習データの量に対する同定率の低下の度合いが楽器によって異なることが分かる．これは，各楽器音の学習データの絶対量にばらつきがあることによると考えられる．

5 あとがき

本稿では，単一楽器の孤立発音を対象とした，HMM と MFCC を用いた楽器音の音源同定について述べた．まず，適切な HMM のパラメータを決定するために，RWC の楽器音データベースを用いて同定実験を行った．その結果，HMM の状態数と混合分布数が各々9, 12 のときに 88.65%の同定率が得られることが分かった．次に，学習データの量が同定率に及ぼす影響を調査し，学習データの量と同定率の関係を明らかにした．今後は，MFCC では同定が困難な楽器音が存在することから，別の特徴量を追加することにより基本性能の改善を図り，また複数音源の同時発音を対象としていく予定である．

謝辞

本研究の一部は，総務省戦略的情報通信研究開発推進制度，及び NTT サービスインテグレーション基盤研究所の研究委託による．

また、本研究の実験において、文献[8]の「RWC 研究用音楽データベース：楽器音」（RWC-MDB-I-2001）を使用した。

参考文献

[1] 後藤真孝，平田圭二，“音楽情報処理の最近の研究，”日本音響学会誌 60 巻 11 号，pp. 675-681, 2004.

[2] 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.

[3] 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, Vol. 44, No. 10, Nov. 2003.

[4] Martin,K.D.,“Sound-Source Recognition:A Theory and Computational Model,” Ph.D. Thesis, MIT (1999).

[5] 北原鉄朗，後藤真孝，奥乃博，“混合音テンプレートを用いた多重奏の音源同定，”情報処理学会研究報告, 2004-MUS-56-9, Vol. 2001, No. 84, pp. 57-64, Aug. 2004.

[6] J.Eggink,G.J.Brown,“A Missing Feature Approach to Instrument Identification in Polyphonic Music, Proc. ICASSP, Vol V, pp. 553-556, 2003.

[7] 大下隼人，甲籐二郎，“HMM を用いた音源同定アルゴリズムに関する一検討，”情報科学技術フォーラム FIT2003，F-002，pp. 207-208，Sept. 2003.

[8] 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，”情報処理学会研究報告，2002-MUS-45，pp. 19-26, 2002.