特徴量に注目した複数楽器の演奏における音源同定処理

木下 智義　坂井 修一　田中 英彦

東京大学大学院 工学系研究科
〒 113-8656 文京区本郷 7-3-1

{kino,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

筆者らはこれまでに音楽情景分析の処理モデル OPTIMA を提案し、その実験システムを構築した。しかしながら、その認識精度は実用上十分とは言えず、その改善が課題となっている。本稿では、従来の処理の問題点である周波数成分の重なりに対する脆弱性を改善するための新たな処理を提案する。本稿では、周波数成分が重なった時の特徴に合わせて特徴量を分類し、それに応じて重なりのある周波数成分の特徴量を適応的に変化させ、音源同定処理を行う。また、各特徴量の音源同定の手掛かりとしての重要度を計算し、同定処理に導入した。評価実験の結果、処理精度の向上が確認され、提案する処理の有効性が明らかになった。

音楽情景分析, 音源同定, 聴覚的情景分析, 周波数成分特徴量

Musical source identification based on frequency component features

Tomoyoshi Kinoshita　Shuichi Sakai　Hidehiko Tanaka

University of Tokyo
7-3-1 Bunkyo-ku, Tokyo, 113-8656

{kino,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

We have previously proposed a processing model OPTIMA for music scene analysis and implemented its experimental system. However, the system was not robust to signals with overlapped frequency components. In this paper, we present a new method that improves this problem by using overlap pattern of frequency components. Weighted template-matching method is applied to identify sound sources repeatedly to each frequency component cluster. The weight is evaluated according to the significance of each feature of the signal. When multiple components are overlapped, our system adaptively modifies features of an input signal to a combination of overlapped components. The experimental results show that the system can identify sound sources of 60 to 69% of musical notes. It also showed improvement the accuracy about 5%, compared to the result without the proposed mechanism.

Music scene analysis, Sound source identification, Auditory scene analysis, Frequency component feature

1　はじめに

筆者らは既に音楽情景分析の処理モデル OPTIMA を提案し [2, 3]、その実験システムを構築した。しかしながら、その処理精度は実用上十分であるとは言えず、改善が課題となっている。

OPTIMA における音源同定処理では、各周波数成分の物理的な特徴量を抽出した上で、主成分分析、判別分析といった統計的な処理が用いられている。ここでは複数の単音に由来する周波数成分が重なった場合においても同様の処理を行っている。しかしながら、周波数成分の重なりが起こった場合、それぞれの成分が干渉し合い、そこから得られる特徴量は大きく変動する。この変動は、ほぼ和になるという単純なものの他に、強い値が優先的に残る場合やあるいは値そのものに意味がなくなるという性質のものなどの種類がある。

そこで本稿では、特徴量を 3 つの種類に分類した後、周波数成分の重なりが存在した場合にこの分類に従って特徴量の再計算を行い、音源同定における誤りを軽減する新たな処理機構を提案する。

本稿の他にも、音源同定を扱った研究がある。最近のものとして、波形レベルでのテンプレートマッチングを用いた手法 [4] では、楽器の個体差を吸収するためにテンプレートのフィルタリングや位相トラッキングの処理を行なう。また、パワーの時間変化のみから音源の推定を試みた例もある [1]。

2　処理の構成

本稿で提案する処理は、図 1に示すように、7 つの処理ブロックと 1 つの知識ベースからなる。本稿ではこの知識ベースを特徴量テンプレートと呼ぶ。

入力音響信号は最初に前処理 (Preprocess) 部において時間-周波数解析され、次いでその結果から周波数成分が形成される。単音形成 (Sound Formation) 部では、前処理部において得られた周波数成分に対してクラスタリングを行う。ここでは、各クラスタが単音に相当する。ここでは単音形成と同時に周波数成分の重なりパターンが抽出される。続く特徴抽出 (Feature Extraction) 部では、それぞれの周波数成分から特徴量が得られる。この特徴量は、単音形成部で得られた重なりパターンに従い、特徴量適応 (Feature Adaptation) 部にて変形される。マッチング (Matching) 部では、変形された特徴量と特徴量テンプレートに格納されている特徴量の間の比較を行い、類似度が計算される。この類似度は仮説生成 (Hypotheses Creation) 部へと送られ、単音仮説が形成される。全てのクラスタについての音源同定が完了するまで特徴量適応部から仮説生成部の間の処理は繰り返し行われることになる。最後に後処理 (Postprocesses) 部にて単音データや楽譜が作成される。各処理部の動作に関する詳細は後述する。

特徴量テンプレートは、それぞれが 1 つの単音に相当するレコードからなる。各レコードには音源名と、特徴量の値のリストが含まれる。図 1にその例を示す。

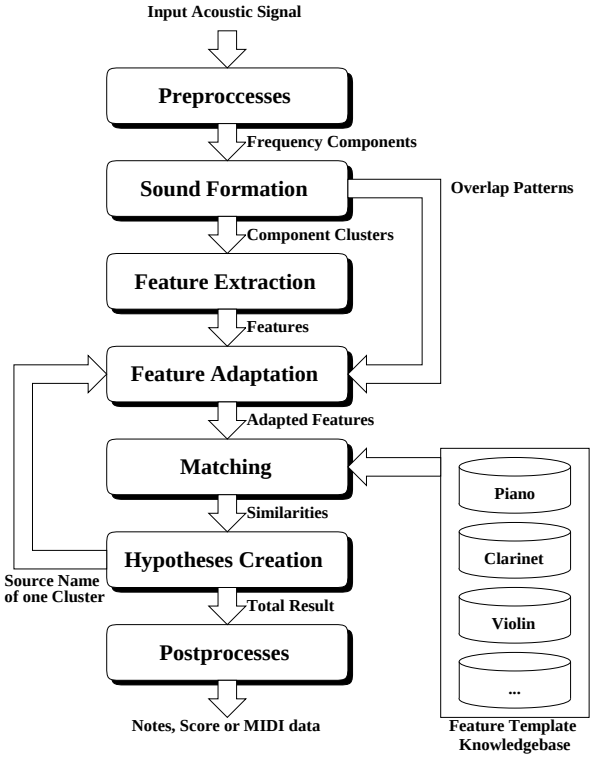


図 1: 処理の構成図

音源名	特徴量 #1	特徴量 #2	特徴量 #3	···
ピアノ	0.724	23.48	5.901	···
ピアノ	0.271	18.22	3.725	···
⋮	⋮	⋮	⋮	···
クラリネット	0.513	49.11	7.224	···
⋮	⋮	⋮	⋮	⋱

表 1: 特徴量テンプレートの例

2.1　前処理および単音形成

前処理部においては、最初に入力信号に対して時間周波数解析を行い、続いてここから周波数成分を抽出する。これらの処理においては、IIR フィルタバンクと挟平面回帰法を用いた [2]。また、特徴量テンプレートを作成する際にも、モデルとなる単音データに対して同様の処理を施す。

続いて、単音形成クラスタリングを行う。この処理において、周波数成分は単音ごとにクラスタリングされる。ここでは柏野の手法 [2] を用いた。この手法では、調和性や複数の周波数成分の間の立上り時刻のずれを抽出し、この結果に応じて同一の単音によると考えられる周波数成分どうしを集めてクラスタリングを行う。

本稿ではこれに加えて周波数成分の重なりパターンを

抽出した。ここで、重なりパターンとは、複数のクラスタに属する周波数成分と、この成分が属するクラスタの組の集合として定義される (図 2)。

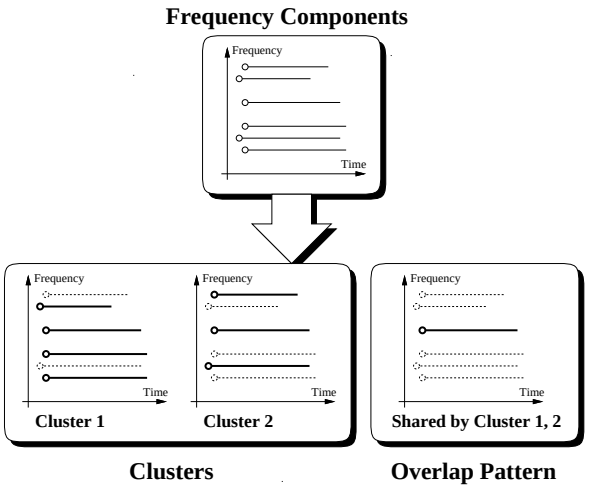


図 2: 単音形成クラスタリングと重なりパターンの抽出

2.2 特徴量抽出

次に、前段階で得られた周波数成分から特徴量を抽出する。本稿では特徴量はパワー包絡線の形状や、立上りの強さ、あるいは各周波数成分のパワーの比といった物理的な量としている。表 2に特徴量の例を示す。

一部の高調波について、そのパワー値が小さい場合にはこの周波数成分から得られた特徴量には意味がないものとして無効化し、後述するマッチング部において類似度計算の対象から外すものとする。

- 各周波数成分のパワー値
- 中心周波数 (各成分の周波数値のパワーを重みとした加重平均)
- 周波数成分のパワー包絡線をパワー値の分布とみた時の、歪度と尖度

表 2: 周波数成分特徴量 (抜粋)

2.3 特徴量の適応処理

実音楽の場合、通常は同時に複数の単音が存在し、また、周波数成分のうちのいくつかは重なりあうことが多い。これにより、それぞれの周波数成分は変形され、特徴量は変化してしまう。そのため、重なりのある周波数成分については、その特徴量をそのまま用いて音源同定すると誤認識の原因となる。

そこで本稿では、周波数成分の重なりパターンと、各周波数成分の性質により特徴量の変形を行うものとする (図 3)。

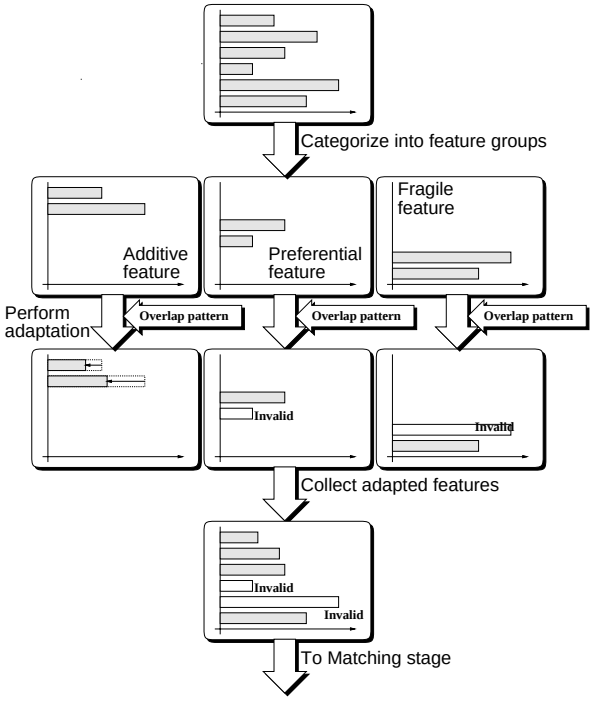


図 3: 周波数成分特徴量の適応機構

2.3.1 特徴量の分類

まず、事前に特徴量をその特質によって 3 種類に分類した。以下にその分類を示す。

1. **加算特徴量**

   周波数成分が重なった時に、その周波数成分から計算される特徴量も概ねそれぞれ単独の場合の和になるもの。
   (例: 周波数成分のパワー値)

2. **優先特徴量**

   重なった周波数成分の特徴量のうち、最大もしくは最小の値が全体の特徴量として得られるもの
   (例: 立上りの強さ)

3. **崩壊特徴量**

   周波数成分が重なった場合、得られた特徴量が意味をなさなくなるもの
   (例: パワー包絡線形状の対称性)

2.3.2 適応処理

周波数成分がただ 1 つのクラスタに属する時には前段で計算された特徴量がそのまま用いられる。一方、複数の周

波数成分が重なった場合には、前項による分類に従って、特徴量の再計算を行う。再計算は以下のように行われる。

1. **加算特徴量**

   適応処理は以下のアルゴリズムに従って行われる

   If 周波数成分が属するクラスタのうち 1 つについて、既に音源名が決定されている

   Then

   既に決定されている音源の特徴量テンプレートから特徴量を得、入力信号から計算された特徴量から引く

   Else

   特徴量の再計算は行われない

2. **優先特徴量**

   以下に述べるようなアルゴリズムに従って適応処理をする

   If 周波数成分が属するクラスタのうち 1 つについて、既に音源名が決定されている

   Then

   既に決定されている音源の特徴量テンプレートから特徴量を得る

   If 入力信号から計算された特徴量と、テンプレートから得られた特徴量が十分近い値となっている

   Then

   入力からの特徴量は、既に決定している音源によるものと判断し、特徴量を無効にしてマッチング部での類似度計算の対象から外す

   Else

   特徴量の再計算は行われない

   Else

   特徴量の再計算は行われない

3. **崩壊特徴量**

   既に特徴量が意味をなさないものとなっていると判断されるため、特徴量を無効にし、マッチング部での類似度計算の対象から外す

2.4 マッチング

音源同定は、適応処理を施した特徴量とテンプレート中の特徴量との類似度を計算することで行う。

2.4.1 重み値の計算

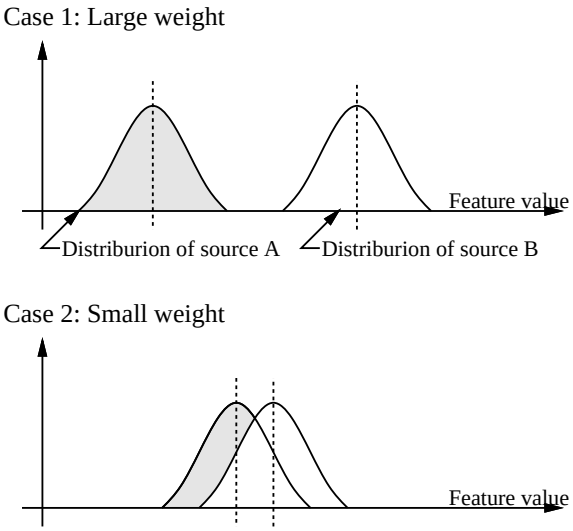
周波数成分の特徴量が、音源同定の手掛かりとなることは示されているものの [2]、全ての特徴量が、手掛かりとして同程度の意味があるとは限らない。実際、クラリネットは、偶数次高調波のパワーが非常に小さいという特徴があり、この点はクラリネットの同定において他の特徴量と較べて大きな手掛かりとなる。そこで、本稿では事前に各音源ごとに、特徴量の重要度を計算した。

まず、事前に特徴量テンプレートに格納されている各特徴量について、音源ごとに平均と標準偏差を計算する。続いて、以下の式にしたがって各特徴量の重み値を計算する。本稿では、上付文字で特徴量の種類を、下付文字で音源の種類を表すものとする。

$$W_s^i = \sqrt{\sum_{t \in S, t \neq s} \left\{ 2 \left(\Phi \left(\frac{|\mu_t^i - \mu_s^i|}{\sigma_s^i} \right) - \frac{1}{2} \right) \right\}^2} \quad (1)$$

ここで、Sは音源名の集合を表す $(=\{Piano,\ Clarinet,\ \cdots\})$。また、$s$ は個々の音源を表すものとする $(s \in S)$。μ、σはそれぞれ特徴量の平均値と標準偏差で、Φは、累積正規確率分布関数とする $(\Phi(z) = \int_{-\infty}^{z} (1/\sqrt{2}) \exp(-x^2/2) dx)$。

今、D_s^iを音源 s の i 番目の特徴量の分布とする。上述の式により、W_s^iは D_s^iが他の音源の分布と離れている場合に大きな値となり、逆に近い時に小さな値となる。例えば、D_s^iが他の音源の D_s^iと十分離れていれば W_s^iは 1 となり、全ての音源について D_s^iが同じ位置に分布する時には W_s^iは 0 となる。こうして、大きな W_s^iを持つ特徴量 i は音源同定の手掛かりとして重視される (図 4)。こうして得られた W_s^iは次段の類似度計算にて用いられる。



Case1: 2 つの特徴量の分布が十分に離れている場合、この特徴は音源同定の手掛かりとして重要であると判断し、大きな重み値を与える。

Case2: 分布が近い場合には、音源同定の手掛かりにはならないと判断する。

図 4: 特徴量の分布に応じた重み値の計算

2.4.2 類似度計算

次いで、入力信号から得られた特徴量とテンプレートから得られる特徴量の間の類似度を計算する。この類似度が音源同定の根拠として用いられる。

まず、類似度そのものの計算に先だち、各特徴量ごとに入力信号の特徴量とテンプレートの特徴量の間の距離を計算する。距離は以下のように計算される。

$$d_s^i = 2\left\{1 - \Phi\left(\frac{|f^i - \mu_s^i|}{\sigma_s^i}\right)\right\} \tag{2}$$

ここで、d_s^i、f^iはそれぞれ i 番目の特徴量に関する入力とテンプレートの間の距離、入力信号から得られた i 番目の特徴量の値を意味する。Φは、前項と同様、累積正規確率分布関数である。

最後に、類似度を以下の式に従って計算する。

$$S_s = \exp\left(\sum_i W_s^i \log d_s^i \Big/ \sum_i W_s^i\right) \tag{3}$$

ここで、W_s^iと d_s^iはこれまでの処理で得られた値である。

この式において、i は各特徴量を表すが、特徴量抽出や適応処理の段階でこの特徴量が無効であると判断されている場合には、$\sum$の計算から除外される。

2.5 仮説生成

マッチング処理の後、最も低い基本周波数を持つクラスタについて、その音源名を確定させる。この一部の音源名が確定したデータは適応処理部へフィードバックされる。適応処理部では、確定した音源の情報を用いて再度適応処理を行うこととなる。実際には音源名を一意には確定させず、複数の候補を作成してそれぞれについてフィードバックを行うことになる。

全ての音源名が確定した時点で、単音仮説を生成する。各単音仮説は複数の単音を含み、それぞれの単音は開始時刻、継続時間、音高、音源名の情報を持つ。

2.6 後処理

OPTIMA の枠組において、仮説生成部で得られた単音仮説は他の処理モジュールから得られる確率情報と統合される [2]。この統合処理により出力された単音仮説に含まれる誤りが訂正されることが期待される。

3 評価実験

本稿で提案した処理を検証するために、2 種類の評価実験を行った。まず、特徴量の重み値の計算を行い、その結果の妥当性を確認した。続いてベンチマークデータに対する音源同定処理の精度を計算し、適応処理の有無による差を評価した。

3.1 重み値計算の評価

まず、表 3にマッチング部における W_s^iの計算結果を示す。

音源名	特徴量
ピアノ	基本波のパワー値の時間方向対称性
	2 倍音のパワー値の時間方向対称性
	基本波のパワー値分布の歪度
クラリネット	基本波のパワー値分布の尖度
	3 倍音のパワー値の時間方向対称性
	2 倍音のパワー値
ヴァイオリン	基本波の立上りの強さ
	2 倍音の立上りの強さ
	基本波のパワー値分布の尖度

表 3: 各音源における重み値の大きい特徴量上位 3 傑

この結果は、直感に合ったものとなっている。ピアノの周波数成分は鋭く立上り、緩やかに減衰する。またそのパワーは立上り付近に主に分布し、減衰域でのパワー値は小さなものとなっている。上表での時間方向対称性と歪度はいずれもピアノのパワー包絡線形状の非対称性を反映したものとなっている。クラリネットは、偶数次の高調波のパワーが非常に小さく、またそれぞれの成分は台形状のパワー包絡線を持つ。従って、パワー分布の尖度や 2 倍音のパワー値の小ささがその特徴として現れている。ヴァイオリンの立上りは本稿で用いた音源の中では比較的緩やかであった。それを反映して、立上りの強さの重み値が大きなものとして得られている。(図 5)

3.2 音源同定処理の評価

ここでは、ベンチマークデータとして用意したランダムノートパターンを対象に、音源同定処理の認識精度の評価を行った。

ランダムノートパターンは、同時に立ち上がる 3 つの単音の組の集合である。各単音の音高と音源名はランダムに決定されている。また、周波数成分の重なりに応じ以下のように分類をした。

クラス 1: 1 つの単音の 2 倍音の成分が、別の単音の基本波と重なるような単音の組。この場合、1 つの単音の周波数成分の全てが別の単音の周波数成分と重なることになる。本稿で分類した 3 つのクラスの中では最も認識が困難になる。

クラス 2: 1 つの単音の 3 倍音の成分が、別の単音の 2 倍音成分と重なるような単音の組。これら 2 つの単音は基本周波数の比が 2 : 3 で、完全 5 度の関係になっている。

クラス 3: クラス 1 にもクラス 2 にも分類されないもの。この場合、周波数成分の重なりは比較的少ない。

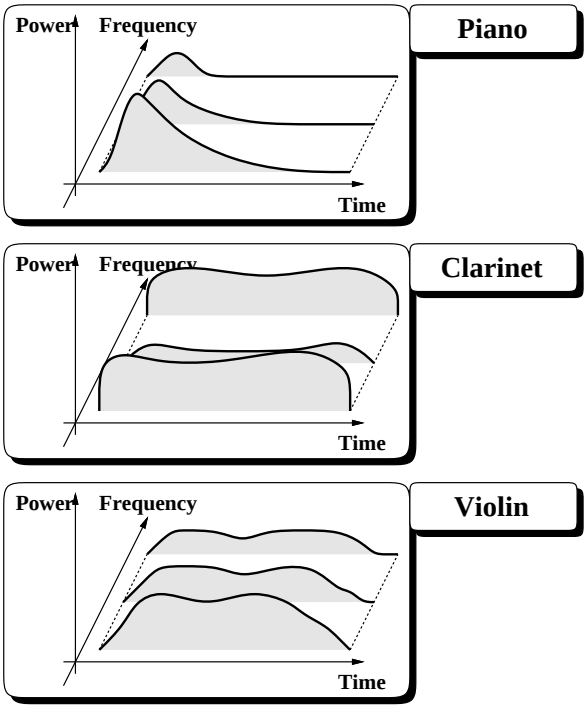


図 5: 各音源の周波数成分の標準的な形

本実験では、各クラス 100 パターンずつ用意し、合計 300 パターンに対して 3 音の加算合成を行ったものを入力として処理を行った。入力信号とテンプレートを作成するデータとして音響信号データ NTTMSA-P1 の単音データを用いた。この信号データは、単一の実楽器による演奏を録音したものである。また、処理対象となる音響信号と、特徴量テンプレートの作成に用いた音響信号は、同一の音源名の場合でも別の楽器個体からのものを用いた。

図 6 に認識結果の精度を示す。各クラス左から順に、単音形成精度、音源同定精度、適応処理を行わなかった場合の精度、類似度計算で重み値を用いなかった場合の精度、を表す。単音形成精度は、音高のみを用いて正誤判定を行ったもので、本稿における処理では音源同定処理の前に単音形成を行うため、音源同定処理がこの数値を上回ることはない。

また、認識精度として、本稿では再現率と適合率の平均を用いた。すなわち、

$$\left(\frac{\text{出力中の正解数}}{\text{出力された単音数}}+\frac{\text{出力中の正解数}}{\text{入力に含まれる単音数}}\right)\times\frac{1}{2}$$

である。

4 おわりに

本稿では、音源同定処理に、周波数成分の重なりに応じた特徴量の適応処理を導入することで、精度を向上させることに成功した。しかしながら、精度の向上はわずかなものにとどまり、また精度自体は改善の余地を多く残している。

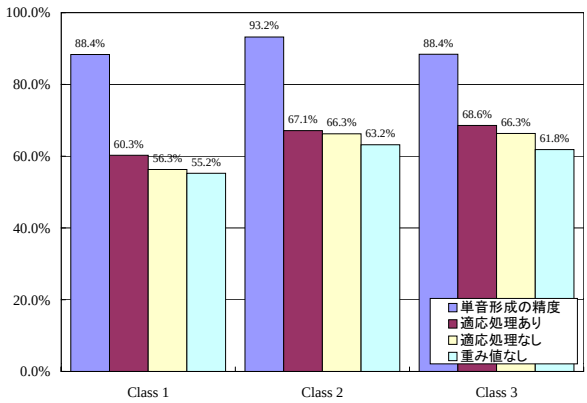


図 6: ベンチマークデータに対する処理結果

これは、本稿では適応処理のうち一部で「何もしない」あるいは「特徴量を無効にする」といった簡単な処理を用いたことが一因として挙げられる。今後は同定の対象となる楽器の種類等をも考慮にいれて適応処理を改善して行く必要があるだろう。

全体の処理精度の改善のためには、演奏された個々の単音の間の特徴量の差を考慮したより柔軟な類似度計算の方法の検討などが今後の課題となるだろう。

謝辞

本稿は、文部省科学研究費補助金 (課題番号 09-07628) による研究成果の一部である。

また、音響信号データ NTTMSA-P1 の使用許可をいただきました NTT コミュニケーション科学基礎研究所に感謝いたします。

参考文献

[1] 三輪, 田所, 斎藤. くし形フィルタを利用した採譜のための異楽器音中のピッ チ推定. 電子情報通信学会論文誌, J81-DII(9):1965–1974, 9 1998.

[2] 柏野, 中臺, 木下, 田中. 音楽情景分析の処理モデル OPTIMA における単音の認識. 電子情報通信学会論文誌, J79-DII(11):1751–1761, 11 1996.

[3] 柏野, 木下, 中臺, 田中. 音楽情景分析の処理モデル OPTIMA における和音の認識. 電子情報通信学会論文誌, J79-DII(11):1762–1770, 11 1996.

[4] 柏野, 村瀬. 適応型混合テンプレートを用いた音源同定 — 音楽演奏 への応用 —. 電子情報通信学会論文誌, J81-DII(7):1510–1517, 7 1998.