音楽情報科学 19−12
(1997. 2. 21)

背景音楽つき音声に対する音響ストリームの分離

中谷 智広　　柏野 邦夫　　奥乃 博

NTT 基礎研究所

nakatani@horn.brl.ntt.co.jp　kunio@ca-sun1.brl.ntt.co.jp　okuno@nue.org

〒243-01 神奈川県厚木市森の里若宮 3-1

あらまし 音楽つきナレーションの様に，音声と音楽が混在する入力に対して，各構成音（ストリーム）を分離する方式を検討する．これまで，音声と音楽を同時に扱えるような方式は提案されていなかった．我々は，二つの音源の特徴にあった処理方式をうまく使い分けることで異なる音源グループに対応する方式をとる．このように複数の対象を扱うためには，それらの関係を体系的に記述した知識表現は重要である．本稿では，オントロジーを用いた音の知識表現と，それに基づいて分離を実現する階層的マルチエージェントアーキテクチャを提案する．また，このアーキテクチャに基づいて音声と音楽を分離するシステムの設計を行なう．

キーワード 音環境理解，音響ストリーム分離，オントロジー，音声，音楽，マルチエージェントシステム

Sound Stream Segregation from a Mixture of Speech and Background Music

Tomohiro Nakatani　　Kunio Kashino　　Hiroshi G. Okuno

NTT Basic Research Laboratories

nakatani@horn.brl.ntt.co.jp　kunio@ca-sun1.brl.ntt.co.jp　okuno@nue.org

3-1, Morinosato-Wakamiya, Atsugi, Kanagawa 243-01 JAPAN

Abstract　A new methodology is proposed to segregate individual sound streams from a mixture of different kinds of sounds, such as speech and background music. No appropriate methodologies are so far presented to cope with such a sound mixture. The proposed methodology adaptively utilizes verieties of processing methods to deal with such different kinds of sounds. For this purpose, formulation of the sound knowledge representation is required. Therefore, a sound knowledge representation method based on ontology is presented, and then a hierarchical multi-agent architecture is constructed for segregating sound streams. A prototype system for segregating speech and background music is designed.

key words　auditory scene analysis, sound stream segregation, ontology, speech, music, multi-agent system

1 はじめに

本稿では，背景音楽つきナレーションのように，音声と音楽が混在した入力から各構成音を分離するための方式を提案する．混合音の取り扱いは，実環境での音響入力や，マルチメディア信号データベースを扱うために重要である．これに対し，近年，音環境理解 (Computational auditory scene analysis) の分野 [1, 19, 23] では，混合音から何らかの一貫した特徴を有する構成音 (音響ストリーム) を抽出して，個別に処理できるようにする研究が行なわれている．本稿では，特に，異なる性質を持つ音が混在する入力に対し，各音源の性質の違いを表現する知識表現法を与えるとともに，それを用いた分離システムのアーキテクチャの検討を行なう．

音声や楽音 (音楽) のどちらかだけが混在する入力に関して，我々は，これまでも，音響ストリームを分離・認識するシステムの研究を行なってきた．中谷らは，複数の音声が混在する入力から，調波構造をもとに有声音部分を抽出し，入力中に分離されずに残った残差を無声部分として割り当てることによって，音声の分離を実現している [14, 18]．また，柏野らは，個々の楽器の音の特徴，音楽の和声・リズムの構造を用いて，音楽を分離認識するシステムを構築している [6, 7]．音声と音楽が同時に混在する入力に対しても，これらのシステムで用いられている処理技術を適用できれば，効率的にシステム構築が行なえるであろう．また，上記以外にも，単独音の処理技術は様々に開発されてきており，これを分離に活用できれば，非常に有効である．

一方，様々な音響処理技術を活用するためには，処理ルーティンごとに概念や用語，関数，あるいはパラメータが異なっていることがしばしば問題になる．例えば，上記の音声分離と音楽分離システムでは，調波構造の分離処理は類似しているが，子音の補完は音声分離だけの処理である．このような違いから，同じような処理を複数のルーティンで行うという無駄が生じたり，パラメータの違いからある処理での計算結果を他の処理では活用できないといった不都合が生じる．本稿では，このような問題点を解決するために音の概念表現 (オントロジー) を用いる [16]．具体的には，それぞれの処理での音の表現はその処理に依存したものが多いので，一旦より概念的な表現 (オントロジー) に変換し，その後で他の処理に依存した表現に変換することによって，ある処理を他の処理で活用できるようにする．つまり，個々の表現をオントロジーレベルに抽象化することで，様々な処理法を統合する手法を構築する．

本稿では，まず，音の概念構造を階層的なネットワークで表現した“音のオントロジー”を構築し，入力音の状態を表現できるようにする．次に，この表現を用いて分離を行なう計算モデルとして，階層的マルチエージェントアーキテクチャを提案する．このモデルは，個々の音を個別に管理するエージェントのネットワークを動的に構築することで適応的な処理を実現する．また，音声と音楽を分離するプロトタイプシステムを設計する．

2 関連研究

音声や楽器音などの特定の種類の音に関して，認識，音色変換，話速変換，ビートトラッキングなど，多くの信号処理技術が開発されてきている [4, 5, 21]．また，音響ストリーム分離の手法としては，調波構造 [8, 10, 15]，音源方向 [22]，共通 AM などに基づく手法が研究されている．オントロジーに基づく音響ストリーム分離では，これらの処理技術は，利用可能な部品として扱う．

分離のアーキテクチャとして提案されたものとしては，まず，Cooke らによる黒板アーキテクチャがある [3]．この計算モデルでは，入力混合音の特徴を表現する複数の特徴マップを作成し，黒板を用いてそれらを統合する．しかし，個別の音源の性質を明示的に扱う枠組が提供されておらず，異なる種類の音源を同時に扱うのは困難である．中谷らの提案した残差駆動型アーキテクチャ [12, 14] は，音源の特徴を個別に管理するエージェントが相互作用して分離を行なうものである．このモデル自身は任意の音源の特徴を個別に扱えるが，異なる種類の音源に対して，どのように処理を行なうかについては検討されていなかった．また，柏野らが提案している OPTIMA [6, 7] は，音楽の楽器音の分離に関して，仮説ネットワークを用いてボトムアップ情報と個々の楽器の音色や音楽の規則性との情報統合を行なう．しかし，単一の音楽を扱うように設計されているため，ナレーションのような別の音源を同時に扱うことができない．

音響ストリーム分離ではないが，種類の異なる音源を特定するアーキテクチャとして，Lesser らの IPUS がある [17]．IPUS は，ヘヤドライヤ，足音，電話音，火災警報，滝の音などの様々な知識源を追加することができ，実際の音を特定することができる．しかし，基本的に単一音を特定するモデルである上，各知識源間の関係を扱っていないため，複数音源を特定する方策は示されていない．

3 音声と音楽の知識表現

3.1 音の概念構造

女性の声は人の声の一種であったり，複数の楽器音が集まって一つの音楽を形成したりするように，一般に音には階層構造がある．我々は，この音の階層構造を，次の二種類に分けて考える．

- 音の抽象度のレベル
- 音の包含関係

前者は，女性の声と人の声の関係であり，人の声は女性の声より抽象度の高い記述であるという意味で階層の上位にくる．後者は，音楽と楽器音の関係であり，音のグループがその構成音に対して上位にくる．音響ストリーム分離にとって，これらの階層構造は重要な意味を持つ．すなわち，分離が進行していく過程で，例えば，人の声が女性の声であると分かることによって，声の基本周波数の高さを制限できたり，ある音が音楽の構成音であることが分か

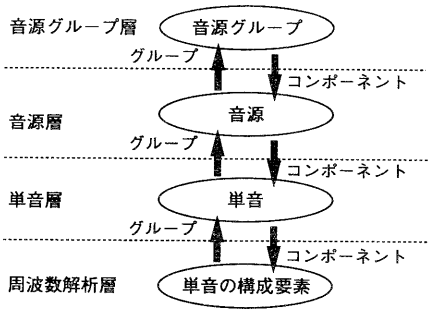


図 1: 音クラスの part-of 階層

ることによって, 和声やリズムの制約を用いてより正確な分離を行なえるようになるからである. したがって, 音の知識表現にとって階層構造は重要である.

このような階層構造を持つ対象の知識表現は, オントロジーで行なうことができる [16]. オントロジーは, 人工知能の研究分野, とくに, 知識共有 (Knowledge sharing) や知識再利用で 10 年ほど前から使われるようになった手法である. オントロジーは, 語彙の形式的な仕様と考えればよい. オントロジーは, さまざまな AI ソフトウェア間での知識共有を容易にする手段として期待されている. 共通のオントロジーを使用することにより, 処理 (問い合わせや言明) の一貫性を保証することができる.

オントロジーでは, 知識はクラスとそのクラス間の関係によって表現される. 通常, クラスはノードで表現し, クラス間の関係をノード間のリンクで表現したネットワークを形成する. また, クラスには以下の 2 つの階層構造がある.

1. クラス階層
2. part-of 階層

クラス階層は抽象度のレベルの階層に相当し, 上位のクラスをスーパークラス, 下位をサブクラス呼ぶ. また, part-of 階層は包含関係の階層に相当し, 上位をグループクラス, 下位をコンポーネントクラスと呼ぶ. また, 各クラスは固有の属性値を有する. クラスの階層において, 下位クラスの属性は, 特別に定義されない限り, 上位クラスの属性をそのまま継承する.

音の場合のクラスとは "音声" のように, 何らかの意味のある音のまとまりを指す抽象的なコンセプトである. また, 実際に発話された個々の音声は音声クラスのインスタンスとして, クラスとは区別される. 特に, 音響ストリーム分離においては, クラスのインスタンスは分離された音響ストリームに相当する. また, 音の階層構造はクラスの階層構造で表現される. 例えば, 人間の声は女性の声のスーパークラスであり, また, 音楽は楽器音に対してグループクラス, 逆はコンポーネントクラスである. 音のクラスの属性としては, 例えば, 音楽クラスのリズムや和声の属性, 音声クラスの基本周波数の属性などを考える.

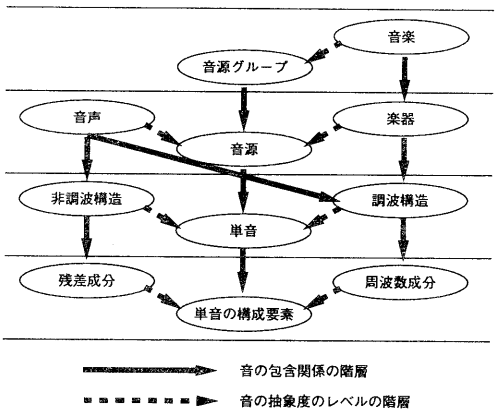


図 2: 音声と音楽のオントロジー

3.2 音クラスの階層

音クラスの part-of 階層

音声や音楽に共通の音クラスの基本的な part-of 階層として, 図 1に示す 4 つの階層を考える. 図中の, 音源とは, 一つの音源から出た音の時系列を指し, 単一楽器から出た音や一人の人の音声などに相当する. 音源グループとは, 複数の音源が, ある規則性を持って集まって出来る音のグループを指し, 楽器のアンサンブルや会話などに相当する. また, 単音とは, 調波構造のような低レベルの一貫した特徴を持つ時間的に切れ目のない音のまとまりであり, 楽器の一音や音声の一続きの有声音などを指す. 単音の構成要素とは, 調波構造の各周波数成分のように, 単音を構成する音のエネルギーである. それぞれの階層で, ある規則性を持って上位クラスが下位クラスを構成要素として所有している. 例えば, 調波構造を持った単音は高調波関係を持った周波数成分を所有し, 楽器のアンサンブルは和声やリズムの規則にしたがった楽器音を所有する. 本稿では, 単音層以上のクラスを音響ストリームのクラスと呼び, そのインスタンスを音響ストリームと呼ぶ.

音クラスのクラス階層

音クラスのクラス階層には, 様々なものをあげることができる. 例えば, 音声に関していえば, もっとも抽象度が高い音源から, 音声, 性別の声, 特定の人の声, その人の風邪声というように, 任意の抽象度のレベル分けでクラス階層を構築することができる. どの程度のレベル分けが必要であるかは, 対象とする音響ストリーム分離システムの問題によって決まる. いずれにせよ, すべてのクラス階層でもっとも上位に位置するクラスは, 各 part-of 階層において, 図 1 に示した 4 つのクラスである.

本稿の以下の節において, 音声と音楽の分離システムを構築する際に利用する, 音声と音楽のオントロジーを図 2 に示す. 図では, 例えば, 音声が非調波構造部と調波構造部からなることや, 複数の楽器が集まって音楽を形成する

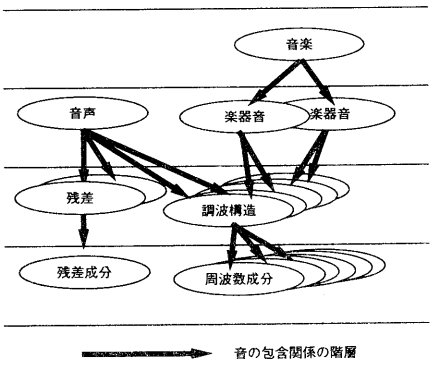


図 3: 入力音の表現例: 音声と 2 楽器のアンサンブル

ことなどが表現されている.

クラス間の関係クラスとその階層

クラス間の関係もクラスとして考える. この関係クラスには次の 2 つを用意する.

1. part-of 関係クラス
2. 層内関係クラス

前者は, part-of 階層で上下に位置するクラス間の関係であり, 後者は, 同一の part-of 階層に位置する任意の二つのクラス間の関係のことである. 関係クラスは音クラスの 2 項関係で表現でき, 例えば, 音声クラスと楽器クラスの関係クラスは, [音声, 楽器] のように表記される. 関係クラスにも, 音のクラスと同じ 2 種類の階層構造が存在する. これは, 2 つの関係クラスの間で各項ごとの音クラスの上下階層が一致する場合に, その階層を関係クラスの階層とする. 例えば, [音声, 楽器] クラスは, [音源, 音源] クラスに対しクラス階層で上位に位置し, [音声, 音楽] クラスに対し, part-of 階層で下位に位置する. この関係クラスは, 音響ストリーム分離において, 音源間の関係を表現するために利用される.

3.3 入力音の表現 − インスタンスの生成

オントロジーを利用すると, 実際の入力の状態は各クラスのインスタンスを生成することによって表現される. 例えば, 2 つの楽器が音楽を演奏し, 一つの音声が発話している状態では, 図 3のようなインスタンスが生成される. この時, 音響ストリームの分離とは, 1) 入力中に含まれる個々の音を発見し, 2) その音が属するクラスのインスタンスを生成し, 3) その属性を求めることに相当する.

4 階層的マルチエージェントアーキテクチャ

4.1 分離のアーキテクチャの概要

前節で述べたように, オントロジーに基づく音響ストリームの分離は, クラスのインスタンス生成と, その属性抽出によって実現される. したがって, 音響ストリーム分離のアーキテクチャの課題は, オントロジーに基づいて下記の処理方法を規定することであるといえる.

1. 入力中の音響ストリームの発見
2. 音響ストリームのクラスの同定
3. 各ストリームのクラスに合わせた属性抽出
4. ストリームどうしの相互の影響の抽出・分離へ還元

まず, あらかじめ入力中にどのようなクラスの音が含まれているかはわかっていないので, 通常, 発見される音響ストリームは抽象度の高いクラスのインスタンスである. その後で, 音響ストリームの分離が進むにつれて, より詳細なストリームの記述が得られてはじめて, より抽象度の低いクラスへとストリームの同定を進めることができる.

また, 音響ストリームの属性抽出に関して, 同定されたクラス固有の処理方法を定義しておけば, 入力の状態に応じて適切な処理を使い分けることができる. さらに, 次の 2 つの音響ストリームどうしの関係は, 属性抽出処理にとって重要な情報となる.

1. part-of 関係クラス間の相互の影響
2. 層内関係クラス間の相互の影響

例えば, 楽器音の属性抽出に音楽のリズムや和音が有効であるように [6], part-of 階層において, ある音響ストリームのコンポーネントの属性抽出にそのグループの属性を利用することは有効である場合がある. また, 同じ層内に存在するストリームどうしの関係も区別することは重要である. これは, 例えば, 同じ調波構造を持った音どうしでも, 後述するように, 楽器音どうしの関係の扱いと楽器音と音声の関係の扱いは違っているからである.

以上のような要件を備えたアーキテクチャには, 次のような特徴があると期待できる.

1. 入力音の状態に応じた分離処理の適応性
2. 時間が経過して情報が増えるにつれて, 入力の構造が明確になる漸時性
3. 拡張性 (モジュール性) の高さ
4. 他のメディアとの統合の容易さ

4.2 マルチエージェントによるモデル化

4.1節で述べた処理の流れを階層的マルチエージェントアーキテクチャを用いてモデル化する. 階層的マルチエージェントアーキテクチャは, 以下のような構成を持つ (図 4) .

- 音のオントロジーの part-of 階層に相当する階層構造をもつ.
- 最下位の階層は周波数解析層であり, 様々な手法に基づく解析結果を出力する.
- それより上の階層は, 新しいストリームを発見する一つの生成エージェントと, 発見されたストリームの状態変化を追跡する追跡エージェントとからなる. 追跡

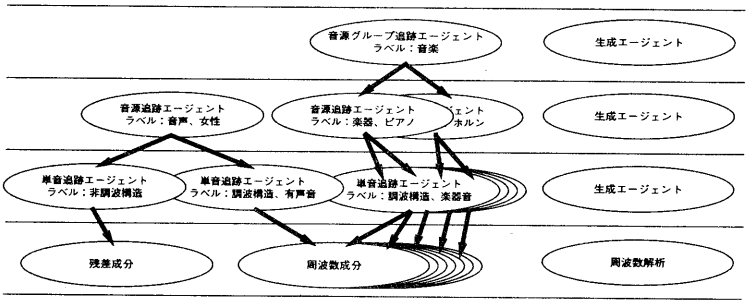


図 4: 階層的マルチエージェントアーキテクチャの構成

エージェントは，生成エージェントが音を発見するたびに動的に生成される．

- 関係クラスに対応する追跡エージェントどうしの通信と，各生成エージェントとすべての追跡エージェントとの通信が存在する．

ここで，各階層内の生成エージェントと追跡エージェント間の相互作用は，中谷らが提案した残差駆動型アーキテクチャと同一の考え方にしたがって構成される [12, 14]．いいかえると，階層的アーキテクチャは，残差駆動型アーキテクチャに階層構造を持たせることで拡張したものである．

各エージェントの動作は次のように規定される．

- 追跡エージェントは，自分が追跡しているストリームのクラスに応じて，次の 4 つの処理を行なう．
  1. ストリームの属性抽出
  2. ストリームのラベリング (抽象度の階層の下位にクラスへの同定)
  3. 他の追跡エージェントと関係クラスに応じた相互作用
  4. ストリームの終了の検出

  これら 4 つの処理法は，各音のクラスおよび関係クラスごとに定義される．その結果，入力音の状態に応じた適応的な処理を行なうことができるようになる．
- 生成エージェントは，すでに生成されている追跡エージェントが追跡していない音のコンポーネントを検出し，新しいストリームを発見する．

例をあげて処理の流れを説明しよう．初期状態においては，追跡エージェントは一つも生成されていない．新たに音が入力されてくると，まず，単音層の生成エージェントが単音を発見し，単音追跡エージェントを生成する．ここで，追跡エージェントの生成・消滅の流れは，残差駆動型アーキテクチャと同様である．生成された単音追跡エージェントは，時間をおって単音の状態変化を追跡し，追跡している音の種類を同定する．ここで，仮に楽器音であるとラベルづけされると，以後，楽器音の特性に応じた処理法で音の変化を追跡する．また，これにともなって，音源層，音源グループ層では，それぞれ楽器追跡エージェント，音楽追跡エージェントが生成される．以後，音楽追跡エージェントはリズムや和声に関する情報を抽出する．また，このリズムの情報は，part-of 階層の下位のエージェントに影響を与えて，例えば，単音層の追跡エージェントは，リズムにあう音の開始時刻や和声にしたがった基本周波数の追跡を行なうことができるようになる．次に，音楽とは別のナレーションが後から入ってきた場合，音楽の場合と同様に，音声の追跡エージェントが各階層で生成される．この時の音声と音楽は，音源層以上の階層では相互作用がほとんど生じないが，単音層以下では，周波数成分が重なったり基本周波数が交差するなどで相互に影響し合う．このような干渉は，層内関係クラスに応じて決まる追跡エージェント間の相互作用を通じて解消される．

以上のような処理を全体としてみると，階層的マルチエージェントシステムによる音響ストリーム分離は，オントロジーとして表現した音楽や音声の知識の体系を制約条件として，動的に生成される追跡エージェントが個々のストリームの記述を漸時的に詳細化していく過程としてとらえることができる．

5 音声と音楽の分離システムの設計

5.1 分離処理モジュール

オントロジーを用いた音声と音楽を分離するシステムの設計では，各処理はクラスごとに個別化されているので，既存の分離のための処理モジュールを再利用できる．したがって，本稿では，まず，すでに構築されている分離システムの処理モジュールを整理して，利用できるものは利用し，不足するモジュールを新たに追加するという方針で，システムを設計する．なお，以下では，1 つの音声+複数の楽器のアンサンブルからなる混合音を対象として，音声と音楽に分離するためのシステムを構成する．

まず，中谷ら [14, 19, 20] と柏野ら [6, 7, 9] が構築したシステムの処理モジュールを列挙する．

音声分離の処理モジュール

1. 有声音分離

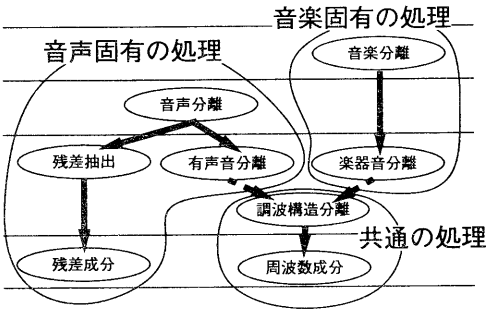


図 5: 音声と音楽の分離システムの構成

2. 無声子音 (入力から有声音を除去した残差で代用) の抽出
3. 有声音の時系列グルーピング
4. 無声子音の補間

音楽分離の処理モジュール

1. 単音抽出
2. 楽器種の同定
3. リズム抽出
4. コード抽出
5. 音楽情報の統計処理

このうち, 有声音抽出および単音抽出の処理は, どちらも, 調波構造を抽出する音声・音楽に共通の処理である. しかし, 二つのシステムでは, 異なる単音間で周波数成分が重複する場合の処理法が全く違っている. したがって, まず, 一般的な調波構造抽出および重複成分の分離処理を設計し, それが音声であるか楽器音であるかが判別された時に, それぞれに適した処理が起動されるようにすることが望ましい. また, 音声の抽出に関して, 有声音の時系列グルーピング, 無声子音の補間は, 音声に固有の処理であり, 音声追跡エージェントの処理として利用できる. 次に, 楽器種の同定, リズム抽出, コード抽出は, 音楽に固有の処理として, 楽器追跡エージェント, 音楽追跡エージェントで利用できる. また, 本稿の話題とは外れるが, 将来的に, 音楽情報の統計処理を組み合わせれば, 統計的な最適化の処理も追加することができるであろう. 以上の処理を組み合わせて, 設計できるシステムの構成図を図 5に示す. 個々の具体的な処理については, 5.2 節で述べる.

生成エージェント

各層において追跡エージェントを動的に生成するために, 各一つの生成エージェントが必要である. 各層における生成エージェントは, 基本的な仕組みとして, 同一層のどの追跡エージェントのコンポーネントにもならない音が一つ下位の層に存在する時, 新しいストリームを検出して, 新しい追跡エージェントを生成する. 特に, 単音層の生成エージェントが処理する調波構造追跡エージェントの生成は, 中谷らが設計した HBSS [10, 11] の生成エージェントとほぼ同一の機能であり, また, 音声追跡エージェントの生成は, 中谷らのグループ生成エージェントとほぼ同一の機能で実現できる. 音楽については, 最初に楽器音が見つかった時に自動的に音楽追跡エージェントを生成するものとする.

追跡エージェントの相互作用

音源間の相互作用としては, 本稿では単音層のものだけを考える. 単音層内の音源の関係は, 層内関係クラスで規定される. この関係クラスには, 調波構造, 有声音, 楽器音の 3 つから 2 つを取り出す組合せの数 (6 個) だけ存在するが, 本稿では, [調波構造, 調波構造] と [楽器音, 楽器音] の二つの関係についての相互作用のみを規定する. それ以外の関係については, 関係クラスの抽象度の階層によって, 相互作用の方法が規定できる. 例えば, [楽器音, 調波構造] は, [調波構造, 調波構造] のサブクラスなので, [調波構造, 調波構造] と同じ相互作用をする.

また, part-of 関係クラス間の相互作用の代表的なものとして, [音楽, 楽器音] などがある. この相互作用では, 音楽追跡エージェントは, 楽器音追跡エージェントが分離した音の属性 (音の開始時刻, 音の高さ) をもとに, 音楽のリズムやコードを抽出し, 一旦, これらの情報の抽出に成功すると, 逆に, 楽器音追跡エージェントがより正確に音の開始時刻, 音の高さを求めるために利用する. ただし, part-of 関係クラス間の相互作用については, 以下のプロトタイプシステムの設計では扱わないものとする.

5.2 プロトタイプシステムの設計例

音声と音楽の分離を行なうためのプロトタイプシステムの設計例を以下に示す. 入力音としては, 蛍の光とナレーションの混合音のようなものを対象とする. ただし, 紙面の都合により各処理において特徴的なもののみを列挙する. また, 和音・リズムの抽出に関しては本稿では扱わないこととする.

共通の処理

最も基本となる処理として調波構造を持つ音を分離するシステムを設計する.

- 調波構造の発見・追跡の基本的な設計は, HBSS と同じにする [10]. すなわち,
  1. 生成エージェントは, 閾値以上のパワーを持つ未発見の調波構造を見つけると新たに追跡エージェントを生成する.
  2. 各時刻において, 過去 n フレームから基本周波数・各周波数成分のパワーの現在値を直線外挿して予測し, 予測値と最も近い基本周波数・音色が得られる調波構造を抽出することで, ストリームを追跡す

る．また，基本周波数の予測値を $\hat{f}_0$ とすると，その探索範囲は，$\hat{f}_0 \pm 10\%$ 程度とする．

3. 調波構造を持つ音のパワーが閾値以下になると音の終了を検出する．

- 異なる2つの音響ストリーム間で周波数成分が重なる場合は，予測値とより近いパワーを持つ方にその成分を排他的に渡し，他方には予測値で代替する．

音声固有の処理

音声の大きな特徴は，有声音のあいだに，調波構造を持たない無声音が挿入されていることである．この音声の分離は以下のように設計できる．

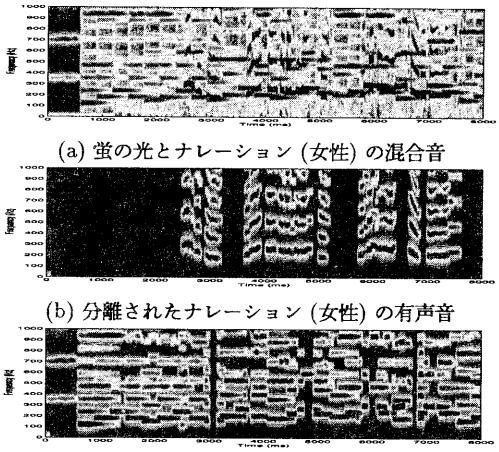
- 追跡している調波構造が音声の有声音部であると判断されると，音声だとラベルづけされる．音声か楽器音かを判別する方法には，1) 音記憶と照合させる方法 [9]，2) 部分空間法を用いる方法 [5] などがあげられる．
- 有声音ストリームが発見された時，音源層に音声追跡エージェントが生成されていなければ新たに生成し，すでに存在する時はそのエージェントに，音声の一部として渡される．
- 有声音の追跡，および周波数成分が重なる2音間での相互作用は，共通の処理と同じとする．
- 音声追跡エージェントは，分離された有声音に子音を補完して音声として分離する．子音の補完は，調波構造を分離した後の残差を割り当てるものとする．[19]．

音楽固有の処理

楽器音の特徴には，和声やリズムを持つこと以外に，各単音の音色・基本周波数の変化が比較的少なく，ある程度以上の継続時間を持ち，また，異なる楽器間で周波数成分がしばしば持続的に重複することなどである．これに基づいて，音声追跡エージェントを以下のように設計する．

- 分離中の調波構造ストリームが楽器音だと判定されると，楽器音とラベルづけする．楽器音の基本周波数遷移パターンや継続時間は特徴的であり，音の発見時に判別に用いることができる．
- 基本周波数の予測値としては，過去 n フレームの平均 $\hat{f}_0$ をとる．また，基本周波数の探索範囲は，音の高さの $\hat{f}_0 \pm 2 \sim 3\%$ と，共通処理よりは狭いものを使用する．
- 周波数成分が重なる楽器音どうしは，周波数成分をどれかに排他的に割り当てるのではなく，その成分を共有しながら追跡する．また，別の音の高調波を基本周波数として持つ音が見つかった時は，その高調波関係が継続していて，次にオンセット／オフセットが見つかるまで，音の長さを時間方向に延長する．

5.3 実装と動作例



(a) 蛍の光とナレーション (女性) の混合音

(b) 分離されたナレーション (女性) の有声音

(c) 分離された蛍の光

図 6: 混合音と分離音のスペクトログラム

5.2 節の設計に基づいて現在実装中のシステムを用いて，システムの動作例を示す．音源の種別判定には基本周波数の変動範囲の大きさを用いた．音声の子音の補完，音楽のリズム・和声の情報の利用，および音源間の相互作用は未実装である．入力は，蛍の光とナレーションの混合音を (12 kHz 標本化, 16 bit 量子化) 用いた．蛍の光の演奏は，2つの楽器 (フルートとピアノ) の音をサンプラで合成し (同時に最大2音が発音)，音声は，女性の声で「本日は御来店いただきましてありがとうございました.」と発話したものを収録した．混合音は，計算機上で音声と音楽を加算して合成した．図 5.3 に，混合音入力と分離された各音のスペクトログラムを示す．図より，概ね二つの音源に適切に分けられていることがわかる．

本稿で設計したシステムの性能評価については，今後，実装を進めて，稿をかえて報告する．

6 今後の課題

1. リズムやコード抽出までを含めた本格的なシステムの実装を行ない，どの程度の性能が得られるかを評価する必要がある．
2. ネットワーク構造を持つシステムでは，個別の緩い制約を全体として統合することができるが，逆に，時として強力なトップダウン制御を用いる方が有効な場合がある．これをネットワーク化したシステムに実現する手段としては，包摂アーキテクチャ [2] を導入するなどの方法が考えられる．
3. 入力の状態に応じて処理を切替えるシステムでは，性能向上のために複数のとりうる状態の可能性 (マルチワールド) を同時に追跡しながら，随時，良い解を選択するという方法が考えられる．これにより，状態の同定誤りによるシステムの性能低下をある程度防ぐことができる．ただし，単純に，マルチワールドに拡張する

だけでは，状態の場合わけがすぐに組合せ爆発を起こすため，効率的に計算量を減らす工夫が必要である．

4. 工学的に実用的なシステムでは，何らかの意味での最適性基準を導入できる枠組が必要な場合がある．本稿で提案したようなネットワーク化したシステムの最適化の方法は，興味ある研究課題である．

7 まとめ

音声と音楽のように，異なる性質を持った音が混在した入力音から，各音響ストリームを分離するための方式を検討した．音響ストリームを分離するためには，様々な処理方法を活用する必要がある．この時，性質の異なる音を扱うためには，対象ごとに用いられている概念や用語の一貫性を保証して，共通処理と個別処理の区別を行なう必要がある．このため，本稿では，人工知能の分野で知識共有 (Knowledge sharing) や知識再利用のために利用されているオントロジーを用いて，音の概念的な知識表現を行なった．この知識表現では，音は，クラス階層，part-of階層の２つの階層構造を持ったクラスのネットワークで表現される．さらに，この音の知識表現を用いて音響ストリームを分離するシステムを設計するために，階層的マルチエージェントアーキテクチャを構築した．このアーキテクチャでは，体系的に表現された音楽や音声の知識を制約条件として，動的に生成される追跡エージェントが個々のストリームの記述を漸時的に詳細化していくことによって音響ストリームを分離する．また，具体例として，複数の楽器アンサンブルとナレーションからなる混合音を分離するプロトタイプシステムについての設計指針を解説した．

謝辞

御討論いただいた，川端豪氏，柏野牧夫氏，後藤真孝氏，河原英紀氏，石井健一郎部長，および，NTT 基礎研究所の研究者の方々に感謝します．

参考文献

[1] Bregman, A.S.: *Auditory Scene Analysis - the perceptual organization of sound,* MIT Press (1990).

[2] Brooks, R.A.: A Robust Layered Control System for a Mobile Robot, *IEEE J. RA-2* '86.

[3] Cooke, M., Brown, G.J., Crawford, M., and Green, P.: Computational Auditory Scene Analysis: listening to several things at once. *Endeavour,* Vol. 17, No. 4 (1993).

[4] Goto, M. and Muraoka, Y.: Beat Tracking based on Multiple-agent Architecture — A Real-time Beat Tracking System for Audio Signals —, Proc. of ICMAS-96, pp.103-110, Dec. (1996).

[5] 木之下 秀二, 有木 康雄: 部分空間法を用いた音声・音楽・環境音の識別, 音講論集 2-5-4, pp.65–66, Mar, (1996).

[6] 柏野 邦夫, 中臺 一博, 木下 智義, 田中 英彦: 音楽情景分析の処理モデル OPTIMA における単音の認識. 信学論 D-II, **J79-DII**, 11, pages 1751–1761, 1996.

[7] 柏野 邦夫, 木下 智義, 中臺 一博, 田中 英彦: 音楽情景分析の処理モデル OPTIMA における和音の認識. 信学論 D-II, **J79-DII**, 11, pages 1762–1770, 1996.

[8] 長渕 裕実, 小林 勉, 山本 啓: 混合音声における音声強調・抑圧, 電子通信学会論文誌, Vol.62-A, No.10 (1979).

[9] 中臺 一博, 柏野 邦夫, 田中 英彦: 音楽音響信号を対象とする音源分離システム －音モデルに基づくアプローチ－, 情処学会 音情研 1-1, Apr. (1993).

[10] Nakatani, T., Okuno, H.G., and Kawabata, T.: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System, *Proc. of AAAI 94*, pp.100-107, Aug. (1994).

[11] 中谷 智広, 奥乃 博, 川端 豪: 音環境理解のためのマルチエージェントによる調波構造ストリームの分離. 人工知能学会誌, Vol.10, No.2 (Mar. 1995) pp.232–241.

[12] Nakatani, T., Okuno, H.G. and Kawabata, T.: Residue-driven architecture for Computational Auditory Scene Analysis. *Proc. of IJCAI-95*, pp.165–172, Aug. (1995).

[13] Nakatani, T., Goto, M., Hiroshi G. Okuno: Localization by harmonic structure and its application to harmonic sound stream segregation, *Proc. of ICASSP-96*, Vol.2, pp.653–656, May. (1995).

[14] 中谷 智広, 後藤 真孝, 川端 豪, 奥乃 博: 残差駆動型アーキテクチャの提案と音響ストリーム分離への応用. 人工知能学会誌, Vol.12, No.1, Jan. (1997).

[15] Nishi, K., Ando, S., and Aida, S.: Optimum Harmonics Tracking Filter for Auditory Scene Analysis, *Proc. of ICASSP-96*, May. (1996).

[16] 西田 豊明: 協調型アーキテクチャによる知識の共有と再利用, 人工知能学会誌 特集「知識の共有と再利用」, Vol.9, No.1, pp.23–28, Jan. (1994).

[17] Nawab, S.H., and Lesser, V.: Integrated Processing and Understanding of Signals, 251–285. in Oppenheim, A.V. and Nawab, S.H. eds. *Symbolic and Knowledge-Based Signal Processing*, Prentice-Hall, (1992).

[18] Okuno, H. G., Nakatani, T., and Kawabata, T.: Cocktail Party Effect with ComputationalAuditory Scene Analysis — Preliminary Report —, Anzai, Y et al. (Eds): *Symbiosis of Human and Artifact, Proc. of HCI International '95*, Vol.2, pp.503–508, Elsevier, Jul. (1995).

[19] Okuno, H. G., Nakatani, T., and Kawabata, T.: Interfacing Sound Stream Segregation to Speech Recognition Systems — Preliminary Results of Listening to Several Things at the Same Time, Proc. of AAAI-96, Vol.2., pp.1082–1089, Aug. (1996).

[20] 奥乃 博, 中谷 智広, 川端 豪: 音声ストリーム分離法の提案と複数音声の同時認識の予備実験. 情報処理学会論文誌, Vol.38, No. 3 (Mar. 1997) 掲載予定.

[21] 小坂 直敏: Sinusoidal model による音色補間, 情処学会 音情研 13-9, 7 年度 12 月研究会予稿 pp. 45-50, Dec. (1995).

[22] 黄 捷, 大西 昇, 杉江 昇: 音源の方位情報を用いた複数音源の分離, 日本ロボット学会誌, Vol.9, No.4, pp.409-414, (1991)

[23] Rosenthal, D., and Okuno, H.G. (eds.): *Working Notes of IJCAI-95 Workshop on Computational Auditory Scene Analysis, (to be published from Lawrence Erlbaum Associates),* (1995).