論　文

Technical Papers

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

Residue-Driven Architecture and Its Application to Sound Stream Segregation

中谷　智広[*1]　後藤　真孝[*2]　川端　豪[*1]　奥乃　博[*1]
Tomohiro Nakatani　Masataka Goto　Takeshi Kawabata　Hiroshi G. Okuno

*1 NTT 基礎研究所
Basic Research Laboratories, Nippon Telegraph and Telephone Corporation, Atsugi 243-01, Japan.
*2 早稲田大学理工学部
School of Science and Engineering, Waseda University, Tokyo 169, Japan.

1995年9月25日　受理

Keywords: multi-agent system, sound stream segregation, computational auditory scene analysis, harmonics, localization.

Summary

This paper presents the *Residue-Driven Architecture* (*RDA*) as a general computational framework for sound stream segregation based on a multi-agent paradigm. Sound stream segregation is an important primary processing for computationally understanding sounds (*Computational Auditory Scene Analysis*) in the real-world. Since RDA is designed without assuming any specific sound attributes, it can be applied to various kinds of sound stream segregation problems. The RDA consists of three kinds of agents: an event-detector, a tracer-generator, and tracers. The event-detector calculates a residue by subtracting the predicted input from the actual input. When a residue exceeds a threshold value, tracer-generator generates a tracer that extracts a sound stream from the residue and returns a predicted input of the next time frame to the event-detector. The RDA is applied to the design of two subsystems: A monaural subsystem segregates sound streams under background noise using harmonic structure; a binaural subsystem refines the sound streams segregated by the monaural system using the direction of the sound source. These subsystems can be concisely designed and simply implemented based on the RDA; therefore, the effectiveness of the RDA is proven. In addition, experimental results show that the capability of the sound stream segregation system is improved by combining these subsystems.

1. はじめに

これまで，音声や楽音などの特定の音が持つ情報を抽出するために，多くの計算技術が研究されてきた．しかし，これらの技術の多くは，特定の音だけが，制限つきの条件下で入力されるという理想的な「研究室環境」を前提とした，個別の技術であった．しかも，通常の環境では，多様な音が同時に発生しており，単純に個々の技術を高度化するだけでは，混合音中の音の干渉や，入力音の多様性を取り扱うのは困難である．これは，人工知能での「スケールアップ問題」[北野 92]に対応する問題といえる．このような問題に対処するためには，多様な音の混合音入力を前提とした，一般的な音の理解の枠組みが必要である．

一般の音を対象とした人間の聴覚機構を解明するために，聴覚心理学では，聴覚的情景分析(Auditory scene analysis)の枠組みが提案され，さまざまな知見が得られている[Bregman 90]．しかし，これらの知見は，計算機上で一般の音を理解するためのモデル化や実装方法については何ら言及していない．これに対し，計算機上でのモデル化や実装のために計算論的な

アプローチを重視した音環境理解(Computational auditory scene analysis)*1が提案されている[Brown 92, Cooke 93, Green 95, Nakatani 94a, Rosenthal 95]. 音環境理解によって,「カクテルパーティ効果」のような騒音下での音の理解機構を実現できると期待される[Okuno 95a]. 実環境下で音環境理解を実現するための課題の一つは,混合音から何らかの一貫した特徴を持った音の集まりを分離することである. このような一貫した特徴を持つ音を**音響ストリーム**(sound stream)と呼ぶ. 音響ストリームは,音響事象の最も単純な表現の一つとみなすことができる.

音響ストリームを分離するには,何らかの音の特徴を手掛りにして,同じ特徴を持つ音を追跡しなければならない. そのような特徴は,音源の性質に基づくものと空間伝達の性質に基づくものに大別することができる. 前者には,低レベルの特徴として,音の立上り,立下り,調波構造,共通FM特性,共通AM特性などがあり,高レベルの特徴として,音声,ピアノの音,電話のベルの音などがある. 一方,後者には,低レベルの特徴として,両耳間時間差,両耳間強度差から求まる音源の方向などがあり,高レベルの特徴としては,部屋の残響特性,視覚により求まる音源位置などがあげられる. 混合音の構造は複雑であるので,どの特徴を使用するかを慎重に選択しないとシステムの構造が極めて複雑になってしまう. 本論文では,低レベルの特徴だけで音響ストリーム分離がどれだけ可能か,また,その限界はどこにあるのかを探るというアプローチをとる.

これまで,低レベルの特徴を用いる工学的な分離法の多くは,それぞれ,別々の技術として研究されてきた. そのため,例えば,調波構造を用いる方法[Cheveigne 93, 長渕 79, Parsons 76, Waintraub 86]は,モノラル音に適用できるが,基本周波数が近接した複数音の区別が困難であったり,無声子音のような複雑な音が扱えなかった. また,音源の位置情報を用いる方法[Bodden 93, Lyon 83, 黄 91, 森田 90, Schmids 86]は,音源自体の性質に関係なく利用できるが,多数のマイクとそれぞれへの空間の伝達特性が必要であったり,扱える周波数帯域に制限があったり,調波構造のような有効な手掛りも利用できなかった. また,多くの分離法では,入力中の音源の数を既知としており,一般の混合音を扱うには制限が多かった.

これらの問題に対処するためには,システムアーキテクチャの観点から考えて,個別の特徴には依存しない,汎用なシステム構成法を構築することが重要である. 我々は,まず,マルチエージェントに基づく構成法を提案し*2,調波構造を使って分離を行うシステムを開発した[中谷 95a]. 本論文では,これをHBSS(Harmonic-Based Stream Segregation)と呼ぶ[Nakatani 94a]. HBSSは,音源の数を仮定せず,調波構造で予測した入力波形と実際の入力波形の減算を行い,その残差を用いて音源数の動的な変化に対応する. しかし,このシステムは,調波構造を前提としたシステムであり,汎用の構成法を与えたものではない*3.

本論文では,まず,HBSSの要点である残差の考え方に着目し,これを,一般の分離問題に適用できるように抽象化したシステムの構成法として残差駆動型アーキテクチャを提案する. 本アーキテクチャから見ると,HBSSは一つの応用システムである. さらに,本論文では,① 雑音に対処した調波構造による分離,② 調波構造と音源方向を組み合わせた分離,③ 音響ストリームの時系列グルーピング,の各処理を行うシステムを設計することを通じて,その実用性を示す.

以下,**2章**で,残差駆動型アーキテクチャについて説明する. **3章**で,残差駆動型アーキテクチャを適用して調波構造による音響ストリーム分離システムを設計し,実装したシステムを**4章**で評価する. **5章**で,バイノーラル入力を用いた音響ストリーム分離システムを残差駆動型アーキテクチャを用いて設計する. この設計でのポイントは,調波構造と方向情報との統合方法である. **6章**で,その実装したシステムを評価し,**7章**でまとめをする.

2. 残差駆動型アーキテクチャ

音響ストリーム分離は,二つの処理に分けることができる. 最初の処理は,ストリーム断片の抽出である. ストリーム断片は,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る. 次の処理は,ストリーム断片を時間的につないでストリームを形成するグルーピングの処理である. **残差駆動型アーキテクチャ**(Residue-Driven Architecture)は,これら二つの処理に共通したアーキテクチャである.

各処理に共通する課題は,何らかの音の特徴に基づ

---

*1 「音環境理解」という用語を用いたのは,聴覚心理学からのアプローチとの違いを明確にし,さらに,研究の進むべき方向性を出すためである.

*2 ここでのエージェントとはごく限られた能力しか持たないMinsky流のエージェントである[Minsky 86].

*3 本論文で提案する残差駆動型アーキテクチャとHBSSの関係は,EMYCINとMYCINの関係と類似しており,前者が後者を一般化したアーキテクチャとなっている.

いて，① 分離対象となる音の発見，② 音の終了の検出，③ 各音の時間的な追跡，④ 各音への混合音の排他的な分配，⑤ 雑音*4 の除去，を行うことである．残差駆動型アーキテクチャでは，これらの課題に対し，入力音中の各構成音を分散管理し，各時刻に次の入力を予測して，実際の入力から予測信号を減算して得られる残差を用いて対処する．すなわち，残差を用いて，各音どうしの影響を抑制し，個々の音の処理を個別に扱えるようにすることによって，各課題に対する処理を単独音を扱うように単純なものにする．

これまで，残差を用いる分離法は，これ以外にも，いくつか研究されているが[Cheveigne 93, Ramalingam 94]，単純なサイン波や調波構造に注目してつくられたものであり，雑音下での処理や，他の特徴を使った分離問題までは，考慮されていない．

2・1 残差駆動型アーキテクチャの構成

残差駆動型アーキテクチャは，**変化検出エージェント，生成エージェント，追跡エージェント**という3種類のエージェントで構成される(図1)．追跡エージェントは通常のものと雑音追跡エージェントに分かれる．変化検出エージェントと生成エージェントはつねに一つだけ存在するのに対して，追跡エージェントは入力音の変化に伴って動的に生成されたり，消滅したりする．また，残差駆動型アーキテクチャ自体は特定の特徴に依存しないが，具体的な実現では何らかの特徴を手掛りとして使うことになる．本論文では，そのような手掛りを**一貫性要因**と呼ぶ．一貫性要因に基づく，各エージェントの働きと全体の動作原理は以下のようにまとめられる．

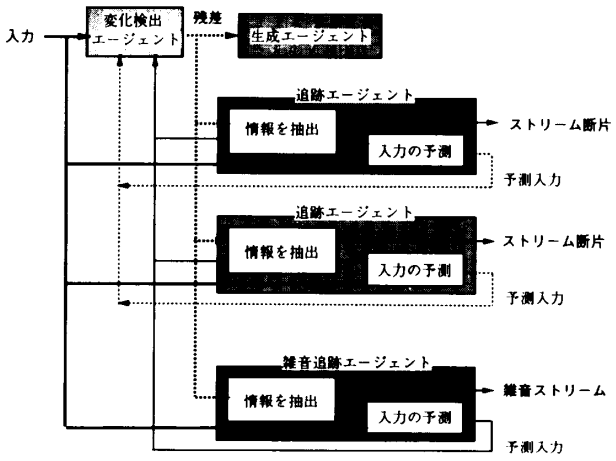


図 1 残差駆動型アーキテクチャ

- 変化検出エージェントは，追跡エージェントからフィードバックされた予測入力を実入力から除き，残差を求める．
- 生成エージェントは，残差がしきい値よりも大きいときに，入力中に新たな情報が入ってきたものとみなし，残差中の新たな情報の属性を一貫性要因に基づいて求める．一貫性要因を満たす成分が見つかれば，それに対する追跡エージェントを生成し，その一貫性要因に基づいてストリームを分離するように指示する．さもなければ，新たな情報を「雑音」とみなし，雑音として一貫性要因を抽出し，雑音追跡エージェントを生成する．ただし，すでに雑音追跡エージェントが生成されている場合には，その一貫性要因を，雑音追跡エージェントに与える．
- 各追跡エージェントは一貫性要因を用いてストリームを追跡し，ストリーム断片を分離するとともに，さらに，次の入力を予測する．雑音追跡エージェントは，雑音についての一貫性要因をもとに雑音ストリームを分離する．

残差駆動型アーキテクチャで構成されるシステムは，初期状態では追跡エージェントは存在しない．新たな情報が入ってくると，追跡エージェントが生成され，その情報の追跡を開始する．新たな情報が追加されない場合には，残差は0となるので，システムの状態は安定する．さらに新たな情報が入ってくると，再び，追跡エージェントが作成され，システムの状態は安定する．ある情報が入力中からなくなると，それを追跡していた追跡エージェントは消滅する．

残差を用いて音を排他的に分配する方法では，予測誤差などの原因で，分配が適切に行えないときがある．これは，残差の誤差が大きくなる原因となり，ひいては，追跡エージェントが多数生成されてしまい，その結果，短いストリーム断片が数多く抽出されるという悪循環に陥る．これを防止するために，残差駆動型アーキテクチャでは，音の感度を動的に変化させて，誤差の拡大を抑え，排他的分配を適正化する機構を有する．

以下では，残差駆動型アーキテクチャを使って，具体的な音響ストリーム分離システムを設計する方法を述べる．音響入力は，12 kHz 標本化，16 bit 量子化，30 ms ハミング窓，7.5 ms フレーム周期で処理されているものとする．

＊4 本論文では，ある特徴で音響ストリームを分離するときに，その特徴を持たない音をすべて「雑音」として取り扱う．

3. 背景雑音を考慮した調波構造による音響ストリーム分離の設計

調波構造に基づく音響ストリーム分離システムHBSSを残差駆動型アーキテクチャを用いて再構成する[Nakatani 95c]．すなわち，一貫性要因として**調波構造**を用いて，さらに，HBSSでは扱えなかった調波構造を持たない音を**雑音追跡エージェント**で処理する．以下では，この新しいシステムを，①ストリーム断片の抽出，②ストリーム断片のグルーピングの各処理を行う二つのサブシステムで構成し，それぞれ，残差駆動型アーキテクチャに基づいて個別に設計する．なお，本章で扱う入力音は，モノラルである．

3・1 ストリーム断片の抽出

背景雑音下で，調波構造を一貫性要因として音響ストリーム分離をするとき，他の音との干渉のために，適切な属性値が取り出せない倍音がある場合がある．HBSSでは，そのような倍音を正しく扱えていなかったために，基本周波数の同定がうまくいかず，精度の高い分離ができなかった．これに対処するために，新システムでは，調波構造の追跡エージェント(以下，**調波構造追跡エージェント**と呼ぶ)は，属性を適切に取り出せる倍音(以下，**有効倍音**と呼ぶ)だけを用いて基本周波数を決定するとともに，雑音追跡エージェントを追加する[Nakatani 95b]．

有効倍音の判定条件を与える前に，いくつかの用語を定義しておく．各入力フレームに対して，調波構造追跡エージェントが抽出する調波構造の属性値には，基本周波数 ω，k 次倍音の強度 A_k，その位相 ϕ_k，がある．これによって，k 次倍音 $H_k(t)$，および調波構造ストリームの波形 $W(t)$ は，次式で表現される．

$$H_k(t)=A_k\cdot\sin(k\omega t+\phi_k) \tag{1}$$

$$W(t)=\sum_k H_k(t). \tag{2}$$

このとき，『有効倍音』は，次の2条件で判定される．

[**条件1**] 強度 $A_k \geq \theta(k\omega)$

[**条件2**] 強度 $\tilde{A}_k$ の時間変化が連続的

ここで，$\tilde{A}_k$ は，混合音のスペクトルグラム上での，調波構造の第 k 倍音に相当する周波数成分の強度を指す．

条件1は，強度の大きい音がある場合に，その周波数近傍の強度のしきい値 θ を上げて，感度を下げるための条件である．現在の実装では，入力の離散フーリエ変換 $\mathrm{DFT}_t(k\omega)$ により，次式で動的に決定される．

$$\theta(k\omega)=c\cdot\|\mathrm{DFT}_t(k\omega)\| \tag{3}$$

(実験的に $c=0.15$ と設定した)．

条件2は，倍音の強度変化の連続性を，周波数成分(混合音スペクトルグラム上の周波数ピークを時間方向に追跡したもの)に関して調べる条件である．ここでは，この強度の時間変化(約50～100 ms)を直線近似し，その2乗誤差の割合が一定値以下かどうかで判定している(詳細は，[Nakatani 95b]を参照のこと)．

調波構造ストリームの発見・追跡・終了検知は，有効倍音をもとにして処理される．一つのストリームは，最低一つの有効倍音を持つ調波構造があるときに発見される．以後，追跡エージェントが，各時刻で有効倍音を判定しながら追跡を行う．まず，各時刻に，残差信号を受け取り，一時刻前に自身が合成した予測信号を加算して入力を得る．この入力に対し，一時刻前の基本周波数の整数倍に相当するすべての周波数の強度 A_k を計算し，どの倍音が有効倍音であるか判定する．次に，現在時刻の基本周波数を，有効倍音の強度の総和を極大にする周波数 ω として，入力から求める．求められた基本周波数をもとに，入力に対する各倍音の位相 ϕ_k と強度 A_k を求め，式(1)，(2)を使ってストリームを波形で合成する．同時に，位相 ϕ_k を次入力に適合するように進めて波形合成を行い，予測信号を合成する．ストリームの終了は，有効倍音が一つもなくなったときに検知される(各属性値の抽出や信号予測の詳細は，[中谷 94b，Nakatani 95b]を参照のこと)．

非調波構造をすべて扱うような雑音追跡エージェントの設計は難しいので，本論文では定常的な背景雑音だけを扱う．雑音追跡エージェントは，調波構造追跡エージェントが生成されていないときに，入力の平均スペクトル強度を計算し，これをもとに，背景雑音の推定を行う．すなわち，予測入力として，雑音を平均スペクトル強度で予測し，実入力から予測された雑音をスペクトル減算する[Boll 79，Graf 93]．これにより，不適切な追跡エージェントの生成が抑制されるだけでなく，調波構造追跡エージェントのストリーム追跡が背景雑音の影響を受けにくくなることを目指す．

3・2 ストリーム断片のグルーピング

前節で抽出された調波構造ストリーム断片をグルーピングする処理では，①新しいグループの生成，②各グループの追跡および消滅検知，③グループ間の競合の解消，という三つの処理が必要である．このグ

ルーピング処理を残差駆動型アーキテクチャを用いて設計する．ただし，ここでは，基本周波数の近さが一貫性要因となる．

グルーピングへの入力は，各時刻ごとに抽出されたストリーム断片の集合である．グルーピング変化検出エージェントは，各ストリーム断片の基本周波数を調べ，追跡エージェントからフィードバックされた基本周波数に近接するストリーム断片を入力から除いて，残差を求める．もし，残差があれば，グルーピング生成エージェントは，グルーピング追跡エージェントを生成する．入力されるストリーム断片によって，追跡エージェントの個数が変化するのは，残差駆動型アーキテクチャの本質である．

追跡エージェントは，「同一グループ条件」を用いてグルーピングと次の入力の予測を行う．すなわち，「基本周波数の差 Δf（単位 cent[*5]）がしきい値 ν 以下である」という条件が成立すれば，二つのストリーム断片は同じストリームに属すると判断する．実験により，ν の値は，グループ内に追跡中の他のストリーム断片がある場合には 350 cent，ない場合には 600 cent と動的に変化させるように設定した．なお，継続時間が短い(約 100 ms 以下)ストリーム断片は，雑音と判断して，雑音追跡エージェントが受け取っている．

4. 調波構造による音響ストリーム分離の評価

前章で設計した調波構造による音響ストリーム分離システムの動作確認を次の 2 種類のベンチマークで行う[*6]．これらのベンチマークは，反響の少ない通常の部屋で録音した単独音を，計算機上で混合して作成した．なお，背景雑音下での分離音の品質，ならびに，雑音の強度に依存した品質の変化に関するより詳細な評価は，[Nakatani 95b]を参照のこと．

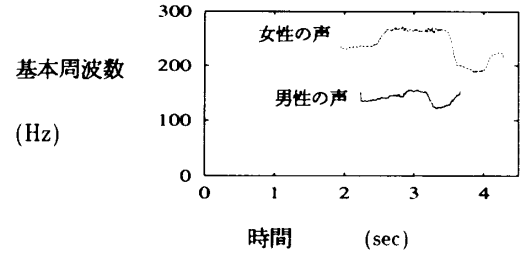
ベンチマーク B1　背景雑音下での 2 話者の声の混合音

白色雑音下で，男性と女性が「あいうえお」と発声している．男性の声の SN 比は，白色雑音に対して 0 dB，女性の声に対して −1.6 dB であり，男性と女性の声の開始時刻の差は 270 ms である．

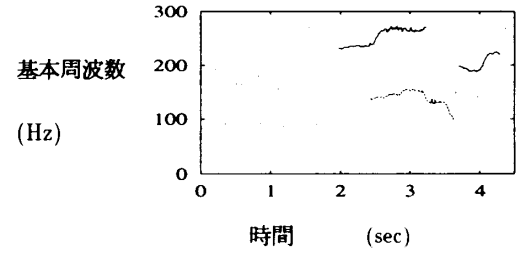
ベンチマーク B2　同じ女性の声の開始時間をずらした混合音

二つの音の基本周波数や音色が非常に接近した状況をつくるために人工的に作成したベンチマークである．女性の声は，B1 と同じであり，二つの声の開始時刻の差は 1.0 s である．

図 2 に，雑音追跡エージェントがある場合とない場合に，ベンチマーク B1 から分離された音響ストリームの基本周波数のパターンを示す．図より，雑音追跡エージェントがある場合のほうが，安定して各声を追

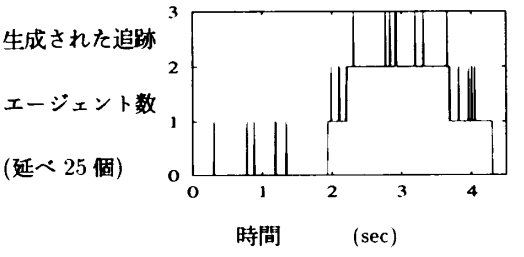


（a） 雑音追跡エージェントありの場合

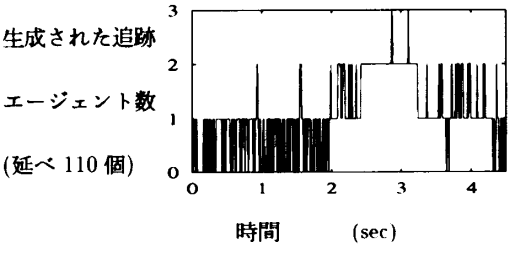


（b） 雑音追跡エージェントなしの場合

図 2　ベンチマーク B1 からの分離された音響ストリームの基本周波数パターン



（a） 雑音追跡エージェントありの場合(延べ 25 個)

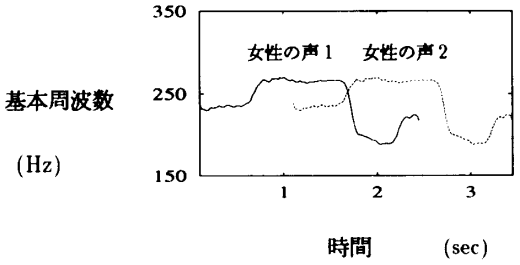


（b） 雑音追跡エージェントなしの場合(延べ 110 個)

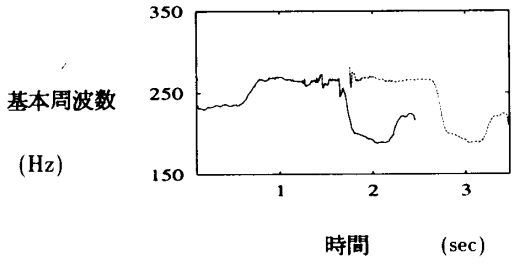
図 3　ベンチマーク B1 を入力したときに生成された，追跡エージェントの数の遷移

＊5　cent は音程の対数表記であり，1 オクターブの音程の周波数差は 1 200 cent となる．

＊6　ここで用いた 2 種類のベンチマークでは，各ストリームは単独のストリーム断片で構成されているので，グルーピングは必要ではなかった．



（a） ベンチマークB2の基本周波数パターン



（b） 分離された音響ストリームの基本周波数パターン

図 4 調波構造による音響ストリーム分離システムの分離結果

跡できていることがわかる．生成された追跡エージェントの個数も図3に示すように大幅に減少しており，雑音追跡エージェントの効果が確認できる．

また，図4に，ベンチマークB2から分離された音響ストリームの基本周波数のパターンを示す．入力音とそれとを比較することにより，二つの音の基本周波数が接近すると(時間軸で1.2～1.8)，分離の性能が低下していることがわかる．これに対処するには，調波構造以外の手掛りが必要である．このため，**次章**において方向情報を用いる方法を検討し，精度の向上を図る．

5. バイノーラルによる音響ストリーム分離の設計

本章では，残差駆動型アーキテクチャを用いて音源の方向情報に基づく分離法を設計し，**前章**で設計した分離システムの性能が向上することを示す．入力は，バイノーラルマイクロフォンの左右2チャネル入力を用い，そこから，方向情報を抽出して分離を行う．

バイノーラル音響ストリーム分離システムを設計するときの主要な課題は，方向情報の抽出，および，調波構造と方向構造の二つの一貫性要因の情報統合である．これら二つの課題を，**前章**で設計したモノラルによる音響ストリーム分離システムを部品として活用し，さらに，残差駆動型アーキテクチャによってモデル化し，実装することによって解決する．

残差駆動型アーキテクチャの各エージェントは，モノラルシステムの対応する左右1対のエージェントを

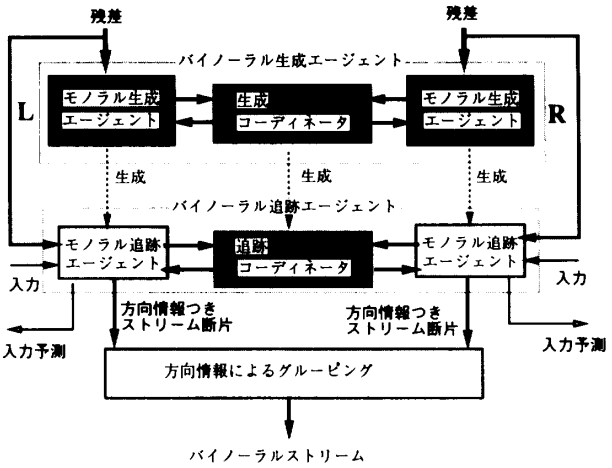


図 5 バイノーラルシステムの構成

用いて構成する(図5)．変化検出エージェントは，モノラルの場合とまったく同じであり，左右の情報交換を行わないので，図中には省略してある．生成エージェントは，生成コーディネータを介して左右の情報交換を行う．生成エージェントが追跡エージェントを生成するアルゴリズムにおいて，モノラルの場合と違うのは，どちらか片方のモノラルシステムの生成エージェントが残差中に新たな音を検出したら，生成コーディネータが，双方で同じ調波構造を追跡する追跡エージェントの生成を指示するとともに，追跡コーディネータを生成する点である．また，左右で同時に異なる調波構造を検知した場合には，生成コーディネータは，強度の強いほうだけについて，追跡エージェントの生成を指示する．このとき，追跡を開始されなかったもう一つの調波構造は，次の入力フレーム中において同様に検知されるので，結果的に，追跡エージェントが生成される．なお，ストリームの発見においては，方向情報は抽出も使用もされない．

追跡エージェントは，二つのモノラルシステムの追跡エージェントが，左右の調波構造の抽出を担当し，それをもとに，追跡コーディネータが，① 音源の方向情報の抽出，② 方向情報による調波構造の洗練化，を担当して，ストリーム分離を行う．

5・1 方向情報の抽出

追跡コーディネータは，左右で分離された調波構造について，両耳間の強度差と位相差を用いて方向同定を行う．この方法では，分離された音を単独音のように扱うことができるので，方向同定の曖昧性が少なく，かつ安定している．

具体的な方向情報の抽出法を述べる．モノラルの場合と同様に，調波構造を一貫性要因として，左右の調波構造が分離され，各倍音の位相と強度が求まる．こ

れらの値をもとに，音源の方向 d を以下のようにして計算する．ただし，方向 d は，その方向から来る音が両耳に到達する時間差(ms)で表し，真左を$-90°$ (-0.75 ms)，正面を$0°$ (0 ms)，真右を$90°$ (0.75 ms)とし，その間を19段階(段階の幅 λ は，$\lambda=0.083$ ms，約$10°$)に離散化して扱う．

(1) 各倍音の方向を左右の強度差と位相差から求める．

(2) 各方向について，その方向と判定された倍音の強度を加算したヒストグラムを計算する．

(3) ヒストグラムで，最大値の方向を音源方向とする．

(1)の倍音ごとの方向同定は以下のように行う．左右の入力ごとに求められる倍音(周波数が ω_k)の強度の対数値と位相の両耳間差を，それぞれ，Δp, $\Delta\phi$ とする．そのとき，倍音の方向が d[ms] であると判定されるのは，位相差の条件式(4)を満たす整数 n が存在し，強度差の条件式(5)が成立するときである．

$$2\pi\omega_k(d-\lambda)\leq\Delta\phi+2n\pi\leq2\pi\omega_k(d+\lambda) \tag{4}$$

$$\begin{cases}\Delta p>0, \text{ if } d>2\lambda,\\ \Delta p<0, \text{ if } d<-2\lambda,\\ |\Delta p|<\varepsilon, \text{ otherwise,}\end{cases} \tag{5}$$

(実験的に，$\varepsilon=0.4$ と設定した．)

(2)において強度の重みつきヒストグラムを用いる理由は，雑音の影響を低減するためである．

さらに，混合音中でも安定して正しい方向が得られるようにするために，同一の方向が継続的に(75 ms以上)得られるときのみ，正しく方向同定されたと判断する．しかし，いったん，正しく方向同定された後でも，瞬間的に音源方向の抽出が不安定になることがある．これに対処するため，音源方向が急激に変化する場合は，その時刻以前に，最後に安定的に取り出された方向を，その時刻の音源方向とする．この安定性は，緩やかな音源の移動(約$130°$/s以下)にも追従できるように，移動速度のしきい値で判定している．

5・2 方向情報による調波構造の洗練化

一つの音源の方向がいったん求まると，それを利用し，その音の調波構造抽出の洗練化に利用する．残差駆動型アーキテクチャでは個々の音の処理が個別化されているので，このような処理の流れが実現しやすい．

調波構造の分離では，基本周波数が近接する場合，各倍音がどの音響ストリームに属するかの判定が難しい．その結果，有効倍音の判定が適切に行えなくなり，分離性能が悪化する．これに対処するために，有効倍音判定の二つの条件に，新たに，音源方向に関する次の条件を加える．

[**条件3**] 倍音と音源の一時刻前の方向差が一定値(0.083 ms$=\lambda$)以下である．

新たな有効倍音判定法により，基本周波数が近接する場合でも，音源方向が異なれば，異なる音の間で，倍音のいくつかは区別できるようになり，より正確な追跡が行えるようになる．

各時刻の分離の処理の流れは，(1)基本周波数抽出，(2)各倍音の位相・強度抽出，(3)方向同定，(4)次入力予測，である．このうち，各倍音の位相・強度抽出，および，次入力予測の処理は，モノラルシステムとほぼ同じ方法で，各モノラル追跡エージェントが行う．これに対し，追跡コーディネータは，方向に関する有効倍音の判定条件を調べながら，両耳の有効倍音の強度の総和が極大になるように，基本周波数を決定し，各時刻において左右で同じ基本周波数を追跡するように指示する．さらに，音源方向を，左右で得られる倍音の属性から同定する．また，ストリーム断片の終了を，両方のチャネルで同時に終了が検知されたときに判定する．

ただし，追跡エージェントが生成された直後は音源方向が求まっているとは限らない．したがって，そのときは，調波構造だけを用いてストリーム追跡を行い，方向情報が正しく求まってから初めて，分離の手掛りとして利用を始める．

5・3 方向情報によるストリーム断片のグルーピング

ストリーム断片のグルーピングを，モノラルの場合と同様に，残差駆動型アーキテクチャで設計する．一貫性要因としては，基本周波数の近さだけでなく，音源方向の近さも使用する．すなわち，モノラルの場合の「同一グループ条件」を，「音源方向の両耳間時間差 Δd がしきい値 ρ 以下であり，かつ，$\Delta f<\nu$ である」という条件に変える．なお，実験により ρ の値は0.167(ms)($=2\lambda$)に設定した．あるストリーム断片が二つ以上のグループに属すると判断された場合には，基本周波数の近さと音源方向の近さの両方の尺度を加味して，より適切なグループだけにそのストリーム断片を割り当てる．このため，次式のグループとストリーム断片の間の距離尺度を導入し，その値が最小となるグループに，そのストリーム断片を割り当てる．

$$K=\alpha\frac{|\Delta f|}{F}+(1-\alpha)\frac{|\Delta d|}{D}, \tag{6}$$

ここで，$F=600$ cent，$D=0.167$ ms は，それぞれのグルーピング判定しきい値，$\alpha=0.47$ は正規化のための係数である．

6. バイノーラルによる音響ストリーム分離の評価

前章で設計したバイノーラルの音響ストリーム分離システムの動作確認のために二つのベンチマークを使う．これらのベンチマークは，無響室で，スピーカーから鳴らした単独音をダミーヘッドで録音したものを，計算機上で混合して作成した．なお，より詳細な性能評価は，音声認識から評価した文献「奥乃 95 b]，スペクトルひずみで評価した文献[中谷 96a]を参照のこと．

ベンチマーク B3　同じ女性の声を，右 45° と左 45° から発声させ，開始時間をずらした混合音
B2 と同じ組合せの音声に，音源位置の情報を追加したベンチマークである．

ベンチマーク B4　B3 に加えて，真正面から三角波が断続的に聞こえる混合音
三つの音が同時に存在するため，片方の手掛りだけでは分離が困難なベンチマークである．

図 6 に，ベンチマーク B3 から分離された音響ストリームの基本周波数のパターンを示す．図 4 と比べると，基本周波数が近接している区間 1.2〜1.8 s についても分離が改善され，ほぼ完全に分離できている．また，分離音のピッチ推定エラーやスペクトルひずみが大幅に改善されることは，別の文献[中谷 95d]で，すでに示している．

図 7 に，ベンチマーク B4 を入力として用いた場合の音響ストリーム分離結果を示す．各ストリーム断片の基本周波数はほぼ正しく分離されているのがわかる．通常，マイクロフォンアレイを用いる分離法[黄 91，森田 90，Schmids 86]では，二つのマイクロフォンで三つ以上の音源を分離することは原理上困難であるが，本実験では，調波構造の手掛りと組み合わせることにより，同時に三つの音が存在する場合でも，適切に分離できている．また，グルーピングに関しても，三つのストリームは，それぞれ個別のストリームとして適切にグルーピングされている．

以上により，残差駆動型アーキテクチャに基づいて設計した調波構造と音源方向を用いる分離システムにより，ストリーム断片の分離，グルーピングの性能が向上することが確認された．

7. お わ り に

本論文では，HBSS の特長である，残差を用いて

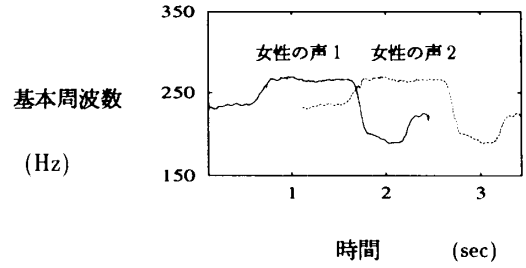
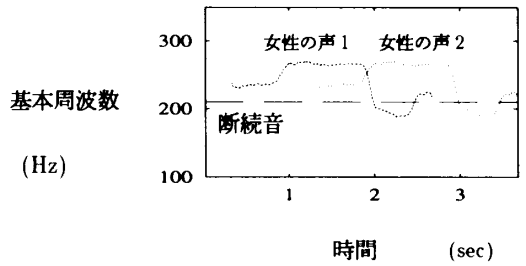
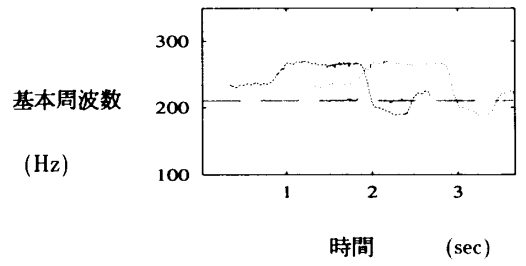


図 6　ベンチマーク B3 のバイノーラルによる分離結果



（a）　ベンチマーク B4 の基本周波数パターン



（b）　分離された音響ストリームの基本周波数パターン

図 7　ベンチマーク B4 のバイノーラルによる音響ストリーム分離の結果

任意個の特徴を追跡する方式を一般化し，さらに，HBSS の弱点である雑音処理を補強したシステムアーキテクチャとして，マルチエージェントに基づく残差駆動型アーキテクチャを提案した．提案した方式の有効性を確認するために，残差駆動型アーキテクチャに基づき，HBSS の再構成と拡張を行い，SN 比 0 dB，白色雑音下でもモノラル音の分離が行えることを示した．また，このモノラル分離システムを部品として用い，バイノーラル入力による音響ストリーム分離システムを，残差駆動型アーキテクチャを用いて設計した．このとき，調波構造と方向情報を組み合わせて分離を行う方法を考案し，モノラル入力では分離の困難な基本周波数が近接した混合音もうまく分離できるようになることを示した．これらの結果から，人が雑音下で音を選択的に聞くという「カクテルパーティ効果」を計算機上でモデル化し，実現するための糸口がつかめたと考えられる．今後の課題としては，アーキテクチャに関しては，トップダウン知識と組み合わせた分離方式の構築があげられ，分離手法に関して

は，子音のような複雑な音を分離する方法の構築があげられる．

最後に，バイノーラル音響デバイスを提供していただいた平原達也氏，入野俊夫氏，御討論いただいたIJCAI-95「音環境理解」ワークショップ参加者各位，竹内郁雄氏，柏野邦夫氏，柏野牧夫氏，河原英紀氏，萩田紀博氏，石井健一郎部長，貴重なコメントをいただいた査読者に感謝します．

◇参 考 文 献◇

[Boddden 93] Bodden, M.: Modeling human sound-source localization and the cocktail-party-effect, *acta acustica*, Vol. 1, pp. 43-55 (1993).

[Boll 79] Boll, S. F.: A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech, *ICASSP79*, pp. 200-203 (1979).

[Bregman 90] Bregman, A. S.: *Auditory Scene Analysis—the perceptual organization of sound*, MIT Press (1990).

[Brown 92] Brown, G.: Computational auditory scene analysis: A representational approach, PhD thesis, Dept. of Computer Science, University of Sheffield (1992).

[Cheveigne 93] Cheveigne: Separation of concurrent harmonic sound: Fundamental frequency estimation and a time-domain cancellation model of auditory processing, *J. Acoust. Soc. Am.*, Vol. 93, No. 6 (1993).

[Cooke 93] 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).

[Green 95] Green, P. D., Coock, M. P. and Crowford, M. D.: Auditory Scene Analysis and Hidden Markov Model Recognition of Speech in Noise, *Proc. ICASSP-95*, Vol. 1, pp. 401-404, May (1995).

[北野 92] 北野宏明：超並列人工知能，人工知能学会誌，Vol. 7, No. 2, pp. 244-262 (1992).

[Lyon 83] Lyon, R. F.: A Computational Model of Binaural Localization and Separation, *Proc. ICASSP83* (1983).

[黄 91] 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)

[Minsky 86] Minsky, M.: *Society of Minds*, Simon & Schuster, Inc. (1986).
邦訳：安西 訳：心の社会，産業図書 (1990).

[森田 90] 森田龍彌：音響パラメータ探索法による複数話者の音声分離，信学論，Vol. J73-A, No. 10, pp. 1551-1557 (1990).

[Nakatani 94a] Nakatani, T., Okuno, H. G. and Kawabata, T.: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System, *Proc. AAAI94*, pp. 100-107 (1994).

[中谷 94b] 中谷智広，奥乃　博，川端　豪：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—ストリーム分離の排他性の向上—，情処全国大会，Vol. 6, 6 213 (1994).

[中谷 95a] 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, pp. 232-241 (1995).

[Nakatani 95b] Nakatani, T., Kawabata, T. and Okuno, H. G.: A computational model of sound stream segregation with multi-agent paradigm, *Proc. ICASSP 95*, Vol. 4, pp. 2671-2674 (1995).

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

[中谷 95d] 中谷智広，後藤真孝，川端　豪，奥乃　博：調波構造と方向同定に基づく音響ストリーム分離，日本音響学会平成7年秋期研究発表会 (1995).

[中谷 96] 中谷智広，川端　豪，奥乃　博：調波構造分離と子音補完による音声ストリーム分離，日本音響学会平成8年春季研究発表会 (1996).

[長渕 79] 長渕裕実，小林　勉，山本　啓：混合音声における音声強調・抑圧，信学論，Vol. 62-A, No. 10 (1979).

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

[奥乃 95b] 奥乃智広，中谷智広，川端　豪：音響ストリーム分離の音声認識からの評価，信学技報，NLC95-51, SP95-86 (1995).

[Parsons 76] Parsons, T. W.: Separation of speech from interfering speech by means of harmonic selection, *J. Acoust. Soc. Am.*, Vol. 60, No. 4, pp. 911-918 (1976).

[Graf 93] Graf, J. and Hubing, N.: Dynamic time warping comb filter for the enhancement of speech degraded by white Gaussian noise, *Proc. ICASSP-93*, pp. 339-342 (1993).

[Ramalingam 94] Ramalingam, C. S. and Kumaresan, R.: Voiced-Speech Analysis Based on the Residual Interfering Signal Canceler (RISC) Algorithm, *Proc. ICASSP-94* (1994).

[Rosenthal 95] 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).

[Schmids 86] Schmids, R. O.: Multiple Emitter Location and Signal Parameter Estimation, *IEEE Trans. on Antennas and Propagation*, Vol. AP-34, No. 3, pp. 276-280 (1986)

[Waintraub 86] Waintraub, M.: A Computational Model for Separating Two Simultaneous Talkers, *Proc. ICASSP86* (1986).

〔査読者：阿曽弘具，中村貞吾〕

著者紹介

中谷　智広(正会員)

1989年京都大学工学部精密工学科卒業．1991年同大学院工学研究科応用システム科学専攻修士課程修了．同年，日本電信電話(株)入社．NTT基礎研究所に勤務．ヒューマンインタフェースに興味を持ち，人工知能を用いた音環境理解の研究に従事．情報処理学会，日本音響学会各会員．

後藤　真孝(学生会員)

1993年早稲田大学理工学部電子通信学科卒業．現在同大学院博士後期課程在学中．日本学術振興会特別研究員．音楽情報処理，音環境理解などに興味を持つ．1992年jus設立10周年記念UNIX国際シンポジウム論文賞受賞．1993年NICOGRAPH '93 CG教育シンポジウム最優秀賞受賞．情報処理学会，電子情報通信学会，日本ソフトウェア科学会，日本音楽知覚認知学会，日本神経回路学会，ICMA各会員．

川端　豪

1978年東北大学工学部電子工学科卒業．1983年同大学院工学研究科電気および通信工学専攻博士課程修了．工学博士．同年，日本電信電話公社入社．1986～89年，自動翻訳電話研究所に勤務．1990年，NTT基礎研究所に復帰，現在に至る．音声自動認識に関する研究に従事．電子情報通信学会，日本音響学会，IEEE各会員．

奥乃　博(正会員)

1950年生まれ(HALと同じ誕生日)．1972年東京大学教養学部基礎科学科卒業．同年電電公社(現NTT)入社．1986～88年スタンフォード大学知識システム研究所客員研究員．1992～93年東京大学工学部電子工学科知能工学寄付講座客員助教授．現在，NTT基礎研究所勤務．主幹研究員．博士(工学)．音環境理解，インターネット情報収集の研究に従事．1990年度人工知能学会論文賞受賞．IJCAI 97広報委員長．情報処理学会，日本ソフトウェア科学会，日本認知科学会，ACM，AAAI各会員．著編書「インターネット活用術」(岩波書店)，「*Computational Auditory Scene Analysis*」(共編，LEA，近刊)，他．