音環境理解：混合音の認識をめざして

奥乃　博
東京理科大学理工学部情報科学科／
科学技術振興事業団ERATO北野共生システムプロジェクト
okuno@nue.org

我々は音声，楽音といった特定の音ではなく，音一般を通じて環境を認識する「音環境理解（CASA，Computational Auditory Scene Analysis）」の研究を進めている．CASAでは，混合音からそれぞれの音を自動的に聞き分ける機能が不可欠である．本稿では，このような機能をマルチエージェントシステムで構成するためのモデルとして，残差駆動型アーキテクチャを紹介する．さらに，音を分離するために使用する情報が増えるにしたがって，混合音分離がどのように改善されるかを報告する．ここでは，モノラル音からの調波構造（基本周波数とその整数倍の倍音から構成される音），ステレオ音からの調波構造と音源方向，さらに，音源方向としてより正確な画像情報を使用した場合，を取り上げる．具体的な応用として開発した音声ストリーム分離システムを概説し，3名の発話を同時に理解するという「聖徳太子コンピュータ」の予備実験結果をモーダリティの観点から報告する．また，このようなモーダリティを増やすには，メディアオントロジーが重要な役割を果たすことを論ずる．最後に，関連研究を概観し，今後の研究課題を指摘する．

なぜ混合音の認識が必要か

音は，マルチモーダルコミュニケーションを実現するための重要な情報として注目されているが，入力の手段である『入力メディア』として文字や画像ほどにはまだ活用されていない．正常な聴力を有した健聴者の場合には，混み合ったパーティー会場でもある会話に入っていたかと思うと，別の会話にすぐに移れるという「カクテルパーティー効果」を有していることはよく知られた事実である．現在の音声認識システムは，口元にマイクロフォンが置かれており，入力音は音声であるということを仮定して構築されている．このような仮定の下でそれなりの性能が達成され，今や音声認識は実用段階に入っている．しかし，マイクロフォンを口元から離すと，音声認識率は一気に低下する．お手元のPCに備わった音声認識システムで簡単に実験ができるので，この事実を読者ご自身で確認していただきたい．

多くの実用的なシステムを構築する場合には，マイクロフォンや特に電話の受話器が話者の口元にあるように設計するのは妥当であろう．しかし，現実にはそのような設計ができない場合も数多くある．たとえば，音源位置が同定できない場合には，想定される音源の位置のそばにマイクロフォンを設置するということは，マイクロフォンの数が非常に多くなったり，あるいは，どのマイクロフォンに焦点を合わせるのかといった新たな問題が生じ，非現実的となる．また，モバイルロボットなどの移動型システムに組み込まれた音声認識システムでは，話者の位置を検出してそこまで近付いて聞くというような制御の問題が生じたり，あるいは，逆に話者がそのシステムの間近まで寄って話すということが必要になり，使い勝手が悪くなる．

音源のそばにマイクロフォンがないと，なぜ音声認識がうまく機能しないのか．音声認識の研究開発者は，雑音抑制などのSN比の向上の技術を用いて音声強調（speech enhancement）を進めているが，このような研究手法は，あくまで単音源からの音が「信号」であり，他の音は雑音であるというモデルに基づいたアプローチであるように思える．やはり，私たちが通常聞く音は混合音であるというアプローチをとる必要があるのではないだろうか．

音環境理解とは

混合音を認識するためには，人はどのような音を聞いているのかに対する我々の立場を明確にしておく必要があろう．**図-1**に示すように，さまざまな音源から発生した音響ストリームが伝達空間を通って我々の耳に入る．耳に同時に到着する音響ストリームは1つでは

図-1　私たちは音をどのように聞くのか

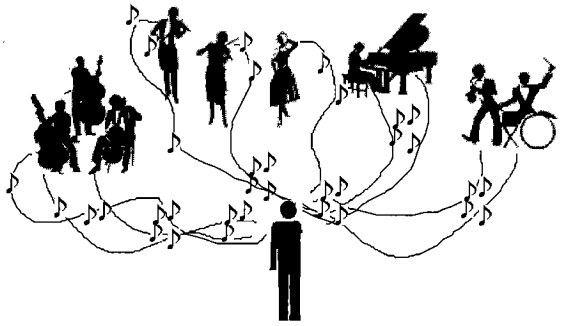

図-2　音響ストリームの階層性

ないので，人は複数の音から構成される混合音を聞くことになる．

混合音を処理したり，理解するためのアプローチは次のように大きく2つに分類することができよう．

(1) 特定の音に的を絞り，他の音を雑音として取り扱う．

(2) 混合音に含まれる（主要な）音を取り出したり，理解したりする．

前者のアプローチは従来の音響処理でとられているものである．音声認識だけでなく，伴奏つきの歌から歌だけを抽出したり，あるいは，ボールベアリングの音を聞いて不良ロットを検出するといった実用に供されるシステムもある．

後者のアプローチは，特定の音に依存せず，あらゆる音をできるだけ同等に扱うことによって，新たな音響処理を追求しようという立場である．つまり，音響処理でのfirst-class citizenを「音声」や「楽音」だけでなく，他の音へも広げるための枠組みを研究しようというわけである．「音一般」の理解をめざすこのような研究は，**音環境理解**☆1（Computational Auditory Scene Analysis, CASA）と呼ばれている．混合音に含まれる個々の音を認識するためのアプローチとして，音が混合したままで認識するという考え方と，個々の音を分離してから認識するという考え方がある．このうち著者らは後者の立場で，何らかの共通の特徴を持つ音のまとまりである「**音響ストリーム**」を分離することが重要であると考えている．

音響ストリームの階層性

音響ストリームを規定する共通の特徴の取り方には階層性がある．たとえば，**図-2**には，さまざまな楽器が演奏され，歌手が楽器演奏を伴奏にして歌っている状況が示されている．このような状況で聞こえる入力音から抽出される音響ストリームには，調波構造の場合もあれば，弦楽器，ピアノ，管楽器，打楽器の個々の楽器音や歌手の歌声などもあろう．また，楽器音にしても，2つのバイオリンが別々に分離される場合もあれば，2つのバイオリンがあたかも1つのバイオリンストリームとして分離される場合もあろう．楽器音が完全に調和してメロディーを奏でているときには，全体で1つの伴奏ストリームとして分離される場合もあろう．このように，音響ストリーム分離のための共通の特徴は，知識表現での階層的な構造をとることになる．音の共通的な表現を「**音オントロジー**」(sound ontology) [6] と呼ぼう．すると，混合音の理解は音オントロジーにしたがって，音響ストリーム分離を行うことということができよう．

逆にいうと，音響信号処理的なアプローチであるマイクロフォンアレイや櫛型フィルタでは，混合音中から音響ストリーム分離を実現しているが，階層的な表現を含めた音響ストリーム分離は考慮されてはいない．第1ステップとしては使えるものの，より上位の音響ストリーム分離には不十分である．

人間の聴覚による音響ストリーム分離についての研究は，「**聴覚による情景分析**」(Auditory Scene Analysis) [2] と呼ばれており，少なくとも1950年代から研究が行われている．しかし，その研究で得られた知見には工学的なモデル化という視点が欠如しており，そのため，1990年代になり，アルゴリズム的な観点から「音環境理解」(CASA) の研究が始まった．特に，混合音に含まれるそれぞれの音源の分離，複数の音を同時に聴く「聖徳太子コンピュータ」，特定の音に注目して耳を傾ける☆2「カクテルパーティーコンピュータ」など

☆1 あるいは聴覚的情景分析とも呼ばれている．

☆2「注視」に対して「傾聴」という用語が一般的であるが，後者に対応するのは「凝視」であり，「注聴」というような用語が必要と思われる．人間の場合には，同時にはたかだか2つの音しか聞き分けられないという心理実験の報告があるので，「注聴」はあり得ないのだが．

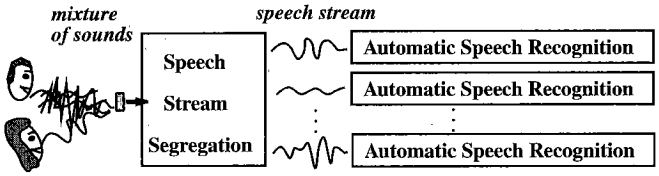

図-3　聖徳太子効果のモデル化

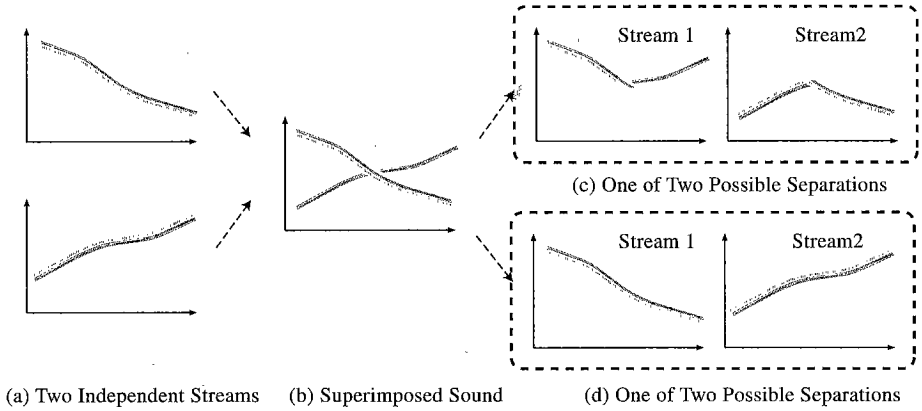

図-4　音響ストリーム分離における曖昧性

の研究が行われている[8]．

音響ストリーム分離での研究課題

複数の音声を同時に聴き，認識するような**聖徳太子コンピュータ**を実現するには，**図-3**に示す混合音から音響ストリームを分離し，その中の音声ストリームをそれぞれ音声認識すれば，そのような実現の第1近似とすることができる．もちろん，この第1近似は，人間のカクテルパーティー効果と比べると単純化されすぎているが，難聴者が補聴器を使用するという実際の状況をシミュレートしていると見なすことができる．音響ストリーム分離が使えるという立場からは，どれか1つの音響ストリームに注意を向けるというカクテルパーティーコンピュータよりも，聖徳太子コンピュータの方が，モデル化が簡単なことが分かる．

音響ストリームを分離するために使用できる音の特徴としては，調波構造（基本周波数とその整数倍の倍音から構成される音），音の立ち上がり（onset），音の立ち下がり（offset），パワー，共通するAM変調あるいはFM変調，音色，ホルマント，音源方向などがあろう．前者では，どの特徴が有効なのか，あるいは，特徴間の優位差はあるのか，といった課題は，現在，音響心理学だけでなく，音環境理解でも重要な研究課題となっている．また，逆に，音声や楽音といった音源のモデル，クラリネットや電話のベルといった音源の種類，あるいは，音源の個数，4拍子というリズムの指定，など何らかの音源情報が仮定できる場合もあろう．このような音オントロジーの上位の情報を用いて，下位の特徴を抽出するときの制御に使うのは，システム構成上重要な課題である．

音響ストリーム分離は，実世界では多くの場合不良問題の逆問題となるので，完全に解くことはできない．たとえば，**図-4**（a）に示された2つの音響ストリームを考えよう．この2つの音響ストリームが**図-4**（b）に示すように混合音となって聞こえたときに，低レベルの特徴だけを使用した音響ストリーム分離では，2つの音響ストリームが**図-4**(c)のように実際に交わっているのか，あるいは，**図-4**（d）のように，両者が接近し，再び離れていったのか，確実に決定することができず，曖昧性が残る場合がある．このような曖昧性を解消するには，使用する音の特徴だけでなく，入力としてモノラル音，ステレオ音，あるいは，画像など入力のモーダリティを増したときの効果も重要な研究課題である．

モーダリティから見た音響ストリーム分離

音オントロジーの音源情報として調波構造と音源方向という2つのレベルだけに注目し，入力としてマイクロフォン1本，2本，さらに画像情報が使えるときに，音響ストリーム分離の性能がどう改善するかを著者ら

の実験を通じて紹介しよう．

■モノラル音からの調波構造を用いた音響ストリーム分離

混合音（モノラル音）から次のような手順でまず，調波構造ストリーム断片を抽出する．

（1）今，追跡エージェントが追跡している調波構造ストリームから次の時間フレームの信号を予測し，次の時間フレームの入力からスペクトルで引き算をし，残差を求める．また，追跡している調波構造ストリームを出力する．

（2）次の時間フレームに追跡している調波構造が見つからなければ，追跡エージェントを終了する．

（3）残差の中に，調波構造が含まれていたら，新たな調波構造が出現したものとして，その調波構造を追跡するエージェントを生成する．

次に，抽出した調波構造ストリーム断片を何らかの類似度を使用してグループ化する．今の場合には，調波構造の基本周波数の近さを尺度とする．

このように漸次的に分離した調波構造ストリームは，基本周波数が重なっていないかぎり，うまく分離をすることができる．また，上記の手法は，音源数を想定していないので，動的な音源数の変化に対応できる．しかし，基本周波数が部分的に重なっている場合には，図-4に示した曖昧性を解消することはできない．

■ステレオ音からの調波構造と音源方向情報を用いた音響ストリーム分離

モノラル入力では周波数領域と時間領域の情報を使用しているだけであったので，音響ストリーム分離での曖昧性解消が難しかった．そこで，空間的な情報を利用するために2本のマイクロフォンを使用する．ただし，通常のステレオマイクロフォンではなく，ダミーヘッドに埋め込まれた1対のマイクロフォンから得られるバイノーラル音を使う．もちろん，2本などというケチなことをいわず多くのマイクロフォンを使ったマイクロフォンアレイなどの使用も考えられよう．

左右のマイクロフォンから調波構造ストリーム断片を上述した方法で抽出する．次に，基本周波数が一致する調波構造ストリーム断片のペアを見つけ，それから，左右の調波構造ストリーム断片のマイクロフォンに到達する時間の差，「両耳間時間差」（ITD，Interaural Time Difference）という，および，左右の調波構造ストリーム断片の強度差，「両耳間強度差」（IID，Interaural Intensity Difference）という，を求める．バイノーラル音を入力とすると，ダミーヘッドの頭の形が入力音に影響を与え，強度差が強調されることになる．このような影響を頭部伝達関数（HRTF，Head-Related Transfer Function）という．

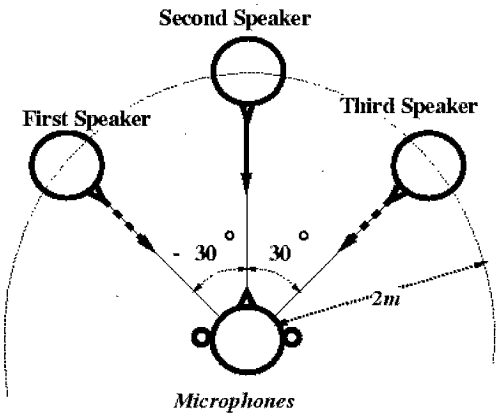

図-5　3話者同時発話認識の実験環境

両耳間の時間差と強度差を求めることは，ステレオ画像におけるステレオマッチングに相当するが，上記の方法では調波構造を音の“ステレオマッチング”のキーとして利用している．したがって，調波構造のスタートが違っており，基本周波数が重なっていても空間的な位置が違っているかぎり，音源がいくつあっても分離が原理的には可能となる．

このように分離された調波構造ストリーム断片は，基本周波数だけでなく音源方向も有している．したがって，グループ化に使用する特徴としては，基本周波数の近さ，音源方向の近さ，あるいは，両者の組合せなどが考えられる．

■音声ストリーム分離とその音声認識

調波構造ストリーム分離の応用として，音声ストリーム分離ができる．音声ストリーム分離は，有声音（母音と有声子音）と無声音（無声子音）に分けて分離をする．有声音は，上述した方法で調波構造ストリーム断片を抽出し，音源方向を用いたグループ化により分離する．

無声子音の分離は信号処理ではきわめて難しい問題であり，現在まで有効な手法は知られていない．そこで，我々は無声子音の分離を行う代わりに，上記の分離過程で生成される残差を用いることとした．というのは，残差には，理想的には調波構造はほとんど含まれていないからである．無声子音の抽出の代わりに，有声音間のギャップを残差でそのまま補完する．このようにして音声ストリームを得る．

次に，このような音声ストリーム分離が音声認識システムのフロントエンドとして使用できるかどうかを検討した．そのために，3話者による同時単語発話の認識実験を行った．**図-5**に実験条件を示す．ダミーヘッ

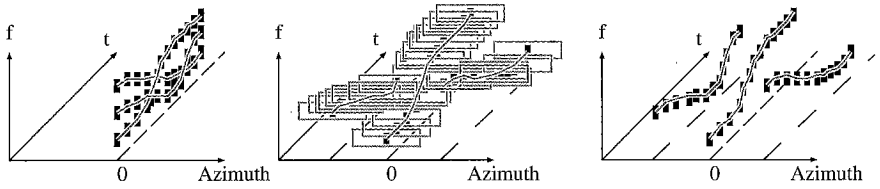

図-6　モダリティから見た音響ストリーム分離の曖昧性

ドマイクロフォンから半径2mの円周上に，中央と左右30度の位置に3人の話者がいる．ただし，ベンチマークデータは，3話者が同時に発話した実データではなく，無響室で採録したデータを頭部伝達関数を用いて解析的に合成したものである．200組の単語発話の混合音に対して，音声ストリーム分離を行った後に，隠れマルコフモデル（HMM）に基づいた音声認識システムを使用し，認識実験を行った．

単音で認識したときと比較して，混合音によって低下する認識率を認識エラーとすると，音声ストリーム分離によってそれがどの程度回復するかという「エラー削減率」で音声ストリーム分離の効率を測定した．第10位までに正解が含まれる認識率のエラー削減率は3話者の平均で，モノラル音の場合には17.25％であり，バイノーラル音の場合には52.2％であった．

なお，2本のマイクロフォンによる3話者同時発話認識実験は，「n本のマイクロフォンを使うと，n-1個の死角を構成できる」という音響信号理論，あるいは，「音源が相互に独立であると，n本のマイクロフォンでn個の音源が分離できる」という独立成分解析などの手法では，難しい課題である．この実験のポイントは，限られた本数のマイクロフォンでどれだけの音源が分離できるかということにある．

■画像からの音源方向情報を活用したモダリティの評価

今，音響ストリーム分離の曖昧性を解消するために，方向情報を導入した．上記の方法による方向情報の精度はたかだか±10°でしかない．ちなみに，健聴者が知覚できる方向情報の精度は±8°といわれているので，±10°という精度自体はそれほど悪いというわけではない．一方，距離や方向を正確に測定するには通常音ではなく，光（画像）が用いられる．この事実に注目し，次のような手法を考案した[4]．

（1）画像情報から方向情報を抽出する．

（2）画像処理は，音源の発見のためにバイノーラル音から得られる粗い方向情報を用いる．

（3）頭部伝達関数を用いて，特定の方向だけを通過させる方向通過型フィルターを構成する[5]．

この手法の特徴は，音響処理と画像処理とを信号レベルではなく，音源の記号表現レベル，上記の方法では，音源方向，で統合していることである．この手法で利用している方向通過型フィルタでは，10Hzごとにサブバンドを両チャンネルで比較し，両時間時間差と強度差を求め，それが指定された方向の両時間時間差と強度差を満足するかどうかを判定し，満足しないサブバンドは通過させず，破棄する．

前節と同様の3話者同時発話の認識実験を行い，得られたエラー削減率は，88.0％である．話者が静止しているので，画像からは正確な方向情報が得られるので，こういう高性能が達成された．第1位での平均エラー削減率は64.5％である．

■モダリティに関する考察

時間的，空間的に重なった音信号を分離する音響ストリーム分離では，分離の曖昧性をどのように解消するかが課題であると述べた．モダリティが分離の曖昧性を解消する理由を示したのが，**図-6**である．Monauralはモノラル音の入力を空間的時間的な情報空間を示し，空間軸の情報が縮退していることを示す．Binauralは空間次元の精度に誤差があるので，各時間での情報に曖昧性が残ることを示す．Binaural with Visionは空間次元の精度が高いので，分離がより正確に行えることを示している．

このようにモダリティが増すと，音響ストリーム分離の曖昧性が減少しているが，必ずしもこれが成立するとは限らない．独立な次元が増えたり，あるいは，ある次元での情報がより正確にならないと，音響ストリーム分離の曖昧性は解消しない．言い換えると，音響ストリーム分離で使用できる鍵となる特徴量が階層的な構造となっており，そのような階層的構造をオントロジーとして表現し，オントロジーの上位への方向にモダリティが増すように設計しなければならない．我々のこれまでの知見から，調波構造が音オントロジ

ーの最も基本的な特徴であり，方向情報がより上位にあるものの，方向情報だけでも分離が可能であることが分かった．

今後の研究課題

音響ストリーム分離が聖徳太子コンピュータを構築する上で重要な役割を果たすことを述べてきた．しかし，単一のメディアに固執していては音響分離の性能に限界があり，その限界を乗り越えるためには，他のメディアからの情報，特に画像情報の統合が不可欠である．そのためにはメディア間で共通の階層的な表現であるオントロジーを介して，階層的な音環境理解を実現していく計画である．

将来の展望として，以下のことが挙げられよう．

(1) オントロジーの上位レベルの情報を用いた分離．音源が分かれば，音源に特化した音響ストリーム分離ができよう．特に，話者が同定できれば，その話者のピッチや話し方など，話者適応を利用した音声ストリーム分離ができよう．また，楽器の種類が画像などから同定できれば，楽器特有の音の特徴を使用して楽器ストリームの分離ができよう．

(2) さらに，音声の場合には，話されている内容を活用して，音声認識に予測と類似した方法で音声ストリームを分離することも検討課題である．

(3) 音響ストリームが，聖徳太子をつくる技術に常に必要か．話している内容が分かれば，音響分離をしなくてもよい場合があろう．たとえば，怒られているときに，聞き流すという機能．

(4) 本文中でも触れたが，他のメディアとの情報統合は，きわめて重要である．本稿で紹介した方法以外に，信号レベルでの統合から知識レベルでの統合までいろいろなアプローチが考えられる．オントロジーのような階層的な表現が，統合に拡張性を与えるためには不可欠である．

(5) 本小特集「聖徳太子をつくる技術」では，独立成分解析に基づく手法と，音環境理解に基づく手法が取り上げられている．これら2つの手法の違いは端的にいえば，音源情報の特徴を積極的に使用するかしないかということである．しかし，独立成分解析においても，独立成分が求まった後の置換の処理では，何らかの音源の情報を使用している．したがって，もっと積極的に音源情報を使用すると，つまり，音環境理解の手法で得られるような特徴量を活用すれば，独立成分解析の精度をさらに向上させることが可能であろう．

(6) 人間の代替となる，Computer Human Interfaceとして音環境理解の成果を使用できるだけではない．本小特集で取り上げられた「聖徳太子をつくる技術」は，ロボット，特に，昨年来人気を呼んでいるエンタテイメントロボットの聴覚に不可欠な機能と考えられる．

実際，著者らは「ヒューマノイドオーディション」という研究分野を現在提案しており，そこで音環境理解の概念や技術が有効であると考えられる[3]．

(7) 多数のセンサから得られるマルチメディア情報を積極的に活用し，それらがお互いに相互作用をし，その過程を通じて，知能が発生するという『共生系知能』研究において，「聖徳太子をつくる技術」が重要な役割を果たそう．

(8) 特定の方向情報だけを通過させる方向通過型フィルタの設計には，ダミーヘッドの頭部伝達関数を利用している．しかし，一般に，頭部伝達関数は音源方向によって変化し，さらに，音の伝達系，たとえば，部屋の中の配置などにも影響を受ける．頭部伝達関数を利用しない方向通過型フィルタの設計も可能であり，ロボットオーディションに適用している[3]．

(9) 「聖徳太子をつくる技術」はさまざまな応用が考えられるが，その中で特に重要なのは，聴覚障害者への支援技術であろう．人間の聴覚機能は20歳を過ぎると低下し始めるといわれている．言葉は人間同士の重要なコミュニケーションの手段であるので，聴覚障害者への補助に活用していくことは重要と考える．

本稿が，新しいマルチメディア認識研究，ロボット研究，コンピュータヒューマンインタフェース研究などに寄与できれば幸いである．最後に，貴重なコメントをいただいた閲読者，かつての共同研究者であった中谷智広氏，川端豪氏，柏野邦夫氏，現在の共同研究者である中臺一博氏，中川友紀子氏，研究の機会を与えていただいた北野宏明氏に感謝する．

参考文献

1) 甘利俊一：聖徳太子かカクテルパーティか，科学 Vol.67, No.9, pp.648-652 (1997).
2) Bregman, A.S.: Auditory Scene Analysis — the Perceptual Organization of Sound, MIT Press (1990).
3) Nakadai, K., Lourens, T., Okuno, H.G. and Kitano, H.: Active Audition for Humanoid, In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pp.832-839, AAAI, Austin (July 2000).
4) Nakagawa, Y., Okuno, H.G. and Kitano, H.: Using Vision to Improve Sound Source Separation, In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), pp.768-775, AAAI, Orlando (July 1999).
5) Nakatani, T. and Okuno, H.G.: Harmonic Sound Stream Segregation Using Localization and Its Application to Speech Stream Segregation, Speech Communication, Elsevier, Vol.27, Nos.3-4, pp.209-222 (Apr. 1999).
6) 中谷智広, 奥乃 博: 音オントロジーに基づいた音環境理解システムの統合, 人工知能学会誌, Vol.14, No.6, pp.1072-1079 (Nov. 1999).
7) Okuno, H.G., Nakatani, T. and Kawabata, T.: Listening to Two Simultaneous Speeches, Speech Communication, Elsevier, Vol.27, Nos.3-4, pp.299-310 (Apr. 1999).
8) Rosenthal, D. and Okuno, H.G. (Eds.): Computational Auditory Scene Analysis, Lawrence Erlbaum Associates (1998).

(平成12年8月1日受付)