音声言語情報処理　9－6
(1995. 12. 14)

音響ストリーム分離の音声認識からの評価

奥乃 博 ・ 中谷 智広 ・ 川端 豪

NTT基礎研究所
〒243-01 神奈川県厚木市森の里若宮 3-1
Email: *okuno@nuesun.brl.ntt.jp, nakatani@horn.brl.ntt.jp, kaw@idea.brl.ntt.jp*

あらまし

本稿では, 音響ストリーム分離を一般環境下での音声認識システムの前処理として使用するための問題点を明らかにするために行った予備実験について報告する. 音響ストリーム分離の結果, 入力音がスペクトル変形を受ける. その原因は, 調波構造抽出, 頭部伝達関数, およびグルーピングである. 離散型単一ゴードブック型 HMM-LR を対象として, これらのスペクトル変形の影響を調べ, 調波構造抽出については音声認識ほとんど影響がないこと, 頭部伝達関数とグルーピングによる影響ついては, 頭部伝達関数をかけた学習データで HMM-LR のパラメータの再学習が有効であることが判明した.

キーワード　音環境理解, 音響ストリーム分離, 音声認識, スペクトル変形, 調波構造分離, 頭部伝達関数

Evaluation of Sound Stream Segregation from the viewpoint of Speech Recognition

Hiroshi G. Okuno, Tomohiro Nakatani, and Takeshi Kawabata

NTT Basic Research Laboratories
3-1 Morinosato Wakamiya, Atsugi-city, Kanagawa 243-01 Japan
URL: *http://www.brl.ntt.jp/people/{okuno, kaw}/*

Abstract　This paper reports the results of the preliminary experiments for interfacing sound stream segregation systems to speech understanding system. Its main issue is spectrum distortion of input sound caused by three processing of sound stream segregation — harmonic structure extraction, head transfer function (HRTF), and grouping. The experiments with a discrete single codebook HMM-LR shows that spectrum distortion caused by harmonic structure extraction is negligible. Spectrum distortion caused by HRTF and grouping are considerable large and can be overcome by re-learning of the parameters of HMM-LR by training data applied by HRTF.

key words　Computational Auditory Scene Analysis, Sound Stream Segregation, Speech Understanding, spectrum distortion

1 はじめに

最近, 複数の音が存在する実環境で音一般を分析し, 理解しようという『音環境理解』 (Computational Auditory Scene Analysis) の研究が盛んになってきた [20]. 音環境理解の研究は, 音声や楽音などの個別の音だけを扱い, また, 多くの場合「研究室環境」という理想的な音場を想定していた従来の音響研究の反省に立ったものと言えよう*. 研究室環境での音声認識というレベルは, 人間にたとえると, 聴覚に障害のある人のレベルであり, 少しでも雑音が入ると適切な処理ができない. 一方, 健康な聴力を持つ人は, 入力音に含まれるさまざまな音の中からどれかの音に注目して聞くことができる. それでも人間は同時には1つのことしか聞けない.

我々は, 計算機のもつべき聴覚機能の1つとして同時に複数のことが聞ける『聖徳太子効果』 [5, 16] を設定し, 研究を進めてきた. 混合音からの音響ストリームの分離を基本機能として, 聖徳太子効果やカクテルパーティ効果のモデル化を行なった [16]. また, 音響ストリーム分離のためにマルチエージェントによる構成を提案し [10, 11], さらに, 音響ストリーム断片の抽出やそのグルーピングのための汎用なアーキテクチャとして残差駆動型アーキテクチャを提案してきた [13]. また, 調波構造に基づいた音響ストリーム分離 [12] や調波構造と方向同定に基づいた音響ストリーム分離 [17, 14] を開発してきた.

音響ストリーム分離は, どのような目的で使用するかによって, その評価基準が変わる. 本稿では, 音響ストリーム分離システムを音声認識システムの前処理 (front-end) として使用するという観点からこれまでに開発してきた音響ストリーム分離システムの予備評価を行なう. さらに, この評価を通して, 音響ストリーム分離システムと音声認識システムとの結合する上での問題点を解明し, それらの解決策を提案する.

2 音響ストリーム分離

2.1 音響ストリームとは

一般の音を扱うためには音の表現が不可欠である. 我々は音の表現として, ある一貫した特徴を持つ音のまとまりである**『音響ストリーム』**を用いる†. 音響ストリームを使用して, 聖徳太子効果やカクテルパーティ効果のモデル化を説明しよう. 聖徳太子効果のモデル化は, 入力音から複数の音響ストリームが分離される音響ストリーム分離と, 分離された音響ストリームからの音声ストリームの選別, および, 音声ストリームを入力として受けとる別々の音声理解システムによって行える (図1参照). このように人間には難しい聖徳太子効果が, 計算機では自然にモデル化が行える. また, カクテルパーティ効果は, 音響ストリーム分離と, 得られた複数の音響ストリームに対する動的な注意切替機構とでモデル化できる.

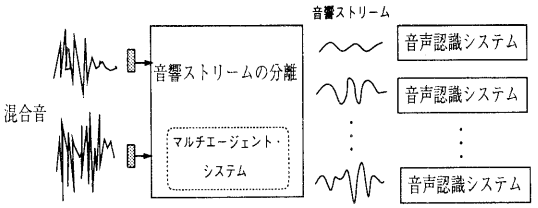

図 1: 聖徳太子効果のモデル化

2.2 音響ストリーム分離の処理

音響ストリーム分離は, 2段階から構成される.

(1) **音響ストリーム断片抽出** — 入力音から何らかの特徴を一貫して持った音のまとまりを抽出する.

(2) **グルーピング** — ストリーム断片を何らかの特徴でまとめ, 音響ストリームを作成する.

各段階で使用可能な特徴や情報は, さまざまなものがある. 我々は, 音響ストリーム分離の問題点を明確にするために, まず, 調波構造 [4, 18, 19] と言う低レベルな情報だけを用いることとした. これまでに調波構造に基づく音響ストリーム分離システム HBSS を開発してきた [10, 11]. HBSS は, 残差駆動型アーキテクチャ [13] に基づいて設計されており, 以下のような手順で, 音響ストリーム断片を抽出する (図2参照).

(1) 変化検出器が入力中に新しい音が入ってきたことを検知すると, 生成器に通知する.

(2) 生成器は, 新しい音に調波構造を発見したら, その基本周波数を求め, それを追跡する追跡器を生成する. 基本周波数が求まらなかったら, 雑音と見なし, 雑音追跡器を生成する. ただし, 雑音生成器がすでに生成されていると, 新たな雑音を雑音生成器に通知する.

(3) 追跡器は, 与えられた基本周波数を元に調波構造を抽出し, さらに, 次の時間フレームの信号を予測し, 変化検出器に送る. 雑音追跡器は, 背景雑音を抽出するとともに, その平均スペクトル強度を用いて, 次の時間フレームの雑音を予測し, 変化検出器に送る.

追跡器が動的に生成されるので, 入力中に含まれる音源の個数を予め与えたり, あるいは, 音源の個数が固定されている必要はない.

*我々のアプローチが, あくまで音声に留まり, 不特定話者を対象としたり, あるいは, 周囲雑音や残響に対してロバストにして, 何とか実用に供したいというアプローチとは全く発想を異にしていることに注意していただきたい.

†計算機科学の立場での『音の表現』は, もっと具体的なものである. たとえば, 『*sound-grep* ⟨ある音⟩ ⟨混合音⟩』を実現するためのデータ表現法を指そう. しかし, そのようなレベルでの表現はまだ提案されていない.

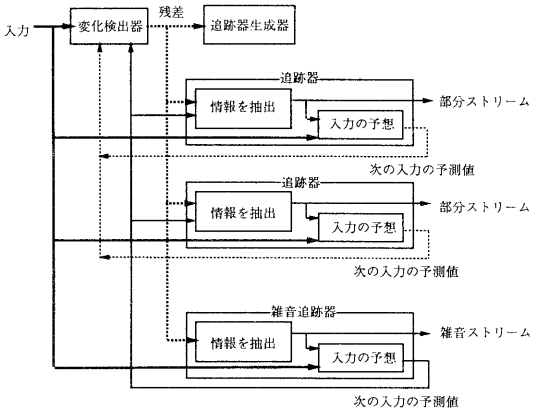

図 2: 残差駆動型アーキテクチャに基づく HBSS

表 1: 学習データの内訳

データ集合 (個数)		話者
0	(999)	MMS (男性)
1	(1,000)	MAU (男性)
2	(1,000)	MNM (男性)
3	(1,000)	MTK (男性)
4	(1,000)	MTT (男性)
5	(241)	MAU,MNM,MTT, 各 121 個, MMS, MTK 各 120 個
計	5,240	(総計 5,602)

表 2: 評価データ

データ集合	(個数)	話者
0	999	MAU (男性)

HBSS を拡張し, バイノーラル入力から方向情報 [7] を利用して, 音響ストリーム断片を抽出し, さらに, 調波構造と方向情報を用いてグルーピングする Bi-HBSS も開発した [14, 17]. バイノーラル入力は, 頭部伝達関数 (HRTF) のかかった音をマイク入力として受け取ることと等価である. Bi-HBSS でのポイントは, 両耳の調波構造だけを用いて音源方向を決定している点と, 抽出された音源方向を用いて調波構造の同定の洗練化を行っている点である.

グルーピングは 2 段階に分かれている. 第 1 段階は, 『**調波構造グルーピング**』により, 音響ストリーム断片がまとめられる. 調波構造グルーピングとして, 以下の 3 種類の方法が提供されている.

(1) **F-Grouping**: 調波構造の類似性 — 基本周波数の差が閾値以下なら同じグループと判定.

(2) **D-Grouping**: 方向情報の類似性 — 音源方位の差が閾値以下なら同じグループと判定.

(3) **B-Grouping**: 上記 2 つの組合せ — 両者の類似性に重みをかけて判定 [17].

HBSS では F-Grouping だけを使用する.

第 2 段階は, 『残差割り当て』である. これは, 無声子音のような調波構造を持たない音の分離のためのものである, 調波構造グルーピングでストリーム断片が存在しないが, 非調波構造の音が存在すると判定されたグループには, 残差が割り当てられる. その時間フレームの残差の割り当ては, 2 種類の方法が考えられる.

(1) **All-Residue**: 残差すべてを割り当てる.

(2) **Its-Residue**: 残差の内, 音源方位の成分だけを割り当てる.

(2) の方法では, 全帯域にわたって音源方向をチェックしなければならないので, その計算量が調波構造だけの場合と較べて膨大になる. これは, Bi-HBSS の利点を殺すことになるので, Bi-HBSS では採用していない. 本稿の実験では, 方向が分かっているものとして (1) の計算を行った.

2.3 音声認識システムとの結合上の課題

本稿で検討対象とした音声認識システムは, 離散型単一コードブック HMM-LR [6] である. 一般に, HMM による音声認識で用いられる特徴は主に次の 3 点である.

(1) **スペクトル包絡**

(2) **ピッチ**

(3) **ラベル** — 音のオンセットとオフセット

混合音から分離された音響ストリームは次の 3 つの過程で**スペクトル変形**を受けている.

(1) 調波構造抽出によるスペクトルの変形.

(2) 頭部伝達関数によるスペクトルの変形.

(3) グルーピングによるスペクトルの変形.

これらの影響により, スペクトル包絡が変化したことによる認識率の低下を測定する必要がある.

Bi-HBSS のピッチ抽出誤りは, 簡単なベンチマークで極めて小さいことが分かっている [12]. しかし, さまざまなベンチマークは行っていないので, 具体的な認識率への影響を測定する必要がある.

HBSS や Bi-HBSS ではオンセットの検出は, 調波構造の立ち上がりの検出で行っている. 子音で始まる発話では, グルーピング時に残差を割り当てる時に正しいオンセットとオフセットを計算しなければならない. しかし, 今回の実験では, 調波構造が始まる前に音がある場合には, 機械的にオンセットを一律 40ms 早め, 逆に調波構造が終ったあとも音がある場合には, オフセットを 40ms 遅す処理をした. また, 上記以外に, 調波構造以外の音が存在する時に, ストリーム断片抽出が安定化しないという問題点も予想される.

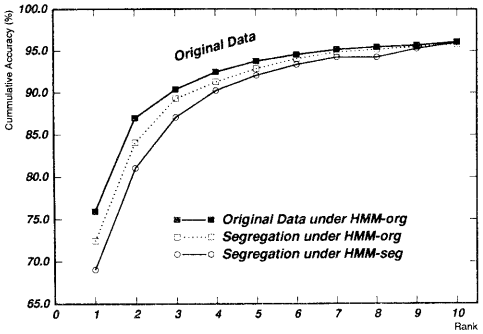

図 3: 調波構造再構成による認識率への影響と
再学習による認識率の改善

2.4 音声認識システムの諸元

本稿では, 離散型単一コードブック HMM-LR [6] を使用した. コードブックのサイズは 256 であり, 標準的な音韻データを元に作成をした. 学習には ATR で作成された 5 人の話者 (すべて男性) のデータを使用した (表 1). 文法は, スタート記号からターミナル記号に直接落ちるルールから構成される. システムの制約からルールは 1900 個である. 本稿では, 表 1のデータで学習させた音声認識システムを『**HMM-org**』と記する. また, 表 2に示したデータで評価した認識率を『**オリジナルデータ**』と略記する. 本稿で使用する認識率は, 認識結果の順位による累積認識率である.

3 音響ストリーム分離の認識率への基本的影響

本章では, 以下の実験によって音響ストリーム分離によって失われる情報が音声認識に与える影響を明らかにし, その改善策を検討する.

(1) 調波構造抽出による情報損失

(2) 頭部伝達関数による情報損失

(3) グルーピングによる情報損失

(1) に対して, 「HBSS による単音分離」実験を行い, (2) と (3) に対して, 「Bi-HBSS による単音分離」実験を行う.

3.1 調波構造抽出による影響

調波構造抽出が音声認識率に及ぼす影響を調べるために, HBSS から抽出された調波構造のストリーム断片, および, ストリーム断片が一切存在しない時間フレームには残差を割り当てて合成した音を用いて, 音声認識を行った. この操作で得られた音を『**調波構造再構成音**』と呼ぶ. 調波構造再構成音の認識率を図 3に示す. 凡例の最初が, 評価データの HMM-org による認識率 (オリジナルデータ) である. 2

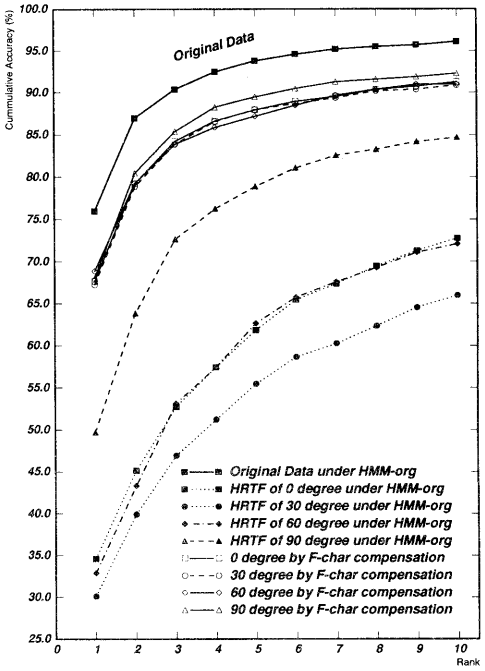

図 4: 頭部伝達関数による認識度への影響と
頭部伝達関数周波数補正による改善

番目 (□) の点線が, 評価データの調波構造再構成音の HMM-org による認識率である. 第 1 位では 3.5% 劣るが, 第 7 位でほぼ認識率が等しくなる. したがって, 調波構造抽出による影響はほとんどないということができる.

調波構造抽出によるスペクトルの変形による認識率劣化への対策として, 表 1の学習データの調波構造再構成音で音声認識システムのパラメータの再学習を行った. この音声認識システムを『**HMM-seg**』と記する. 評価データの調波構造再構成音の HMM-seg による認識率を同じ図 3の凡例の三番目 (○) の実線で示す. 図から, 再学習による認識率改善にはそれほど効果がないことが分かる. もちろん, 図には示していないが, 他のデータ集合では認識率が少し上がる場合もある. したがって, 調波構造抽出による影響は小さいと考えることができる.

3.2 頭部伝達関数による影響

バイノーラル入力は, モノラル入力に音源方向の頭部伝達関数がかかったものと等価である. 頭部伝達関数による影響は主として次の 3 点である.

(1) パワーの変化

(2) 周波数特性の変化

0° (真正面), 30°, 60°, 90° (真横) の 4 種類の頭部伝達関

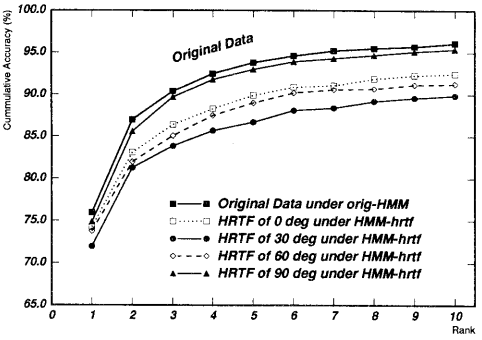

図 5: 頭部伝達関数による再学習後の認識度

数の認識率への影響を調べた. Bi-HBSS でストリーム断片を抽出し, 音源方位情報でグルーピング (D-Grouping) をし, さらに非調波構造のある時間フレームには全残差を割り当てて (All-Residue) 音響ストリームを分離する. さらに, 頭部伝達関数の影響を正確に調べるために, 分離された音のパワーの総和と入力パワーの総和との差が小さくなるように出力信号のパワーを調節し, かつ, 入力信号の無音状態が出力信号でも保存されるように調節した. そのようにして得られた信号に対する認識率を図 4の下 4 つの線で示す. 図から明らかなように, 方位による認識率低下の差が大きい. 90° 度の時が 20% と最も小さく, 他の角度では 40% 以上にも上る.

3.3 頭部伝達関数による認識率劣化への対策

頭部伝達関数による認識率劣化への対応策として次の 2 つの方法を検討した.

(1) 頭部伝達関数の周波数特性補正.

(2) 学習データに頭部伝達関数をかけて, 再学習.

頭部伝達関数をかけると高域が強調されることが分かっているので[‡], 周波数特性補正を行い, 頭部伝達関数の周波数特性を元に戻した音を HMM-org で認識させた. その認識率を図 4の真中 4 つの線で示す. 認識率は大幅に改善され, かつ, 音源方向による認識率のばらつきがなくなった.

表 1 の学習データすべてに 0°, 30°, 60°, 90° の頭部伝達関数をかけ, 上記のパワーと無音状態の調整を行ったデータを基に音声認識システムを再学習させた (学習データの個数は 4 倍に増える). この音声認識システムを『**HMM-hrtf**』と記する. 前節での頭部伝達関数をかけた評価データの HMM-hrtf による認識率を図 5に示す. 90° の場合の認識率は, オリジナルデータとほとんど遜色がない. 音源方向によって, 認識率に差が出るが, 頭部伝達関数の周波数特性補正よりも

[‡]90° の方位 (真横) から入力された『あじ』という音は, 高域が強調されて人間の耳には『あし』と聞こえる.

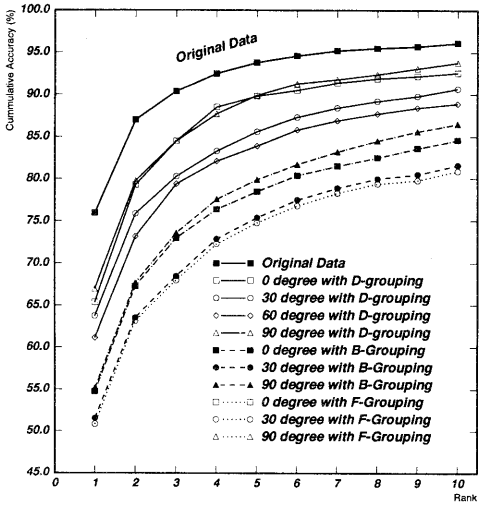

図 6: 調波構造グルーピングによる認識率への影響

認識率が改善されている.

頭部伝達関数によるスペクトル変形に対しては, 次の 2 つの理由により, 頭部伝達関数の周波数特性補正ではなく, HMM-hrtf を用いる方がよいと考える. 第 1 点は, 頭部伝達関数の周波数特性補正による認識率が劣ることである. 第 2 点は, 周波数特性補正のためには正確な音源方向が必要であるが, Bi-HBSS では, ±10° の精度でしか方向情報が分からないからである. また, 本稿で使用した頭部伝達関数のデータは 30° ごとにしか測定されていない.

3.4 グルーピングによる認識率への影響

調波構造グルーピングによる影響を調べるために音響ストリーム断片の 3 種類のグルーピング, F-, D-, B-Grouping を評価する. 4 つの方位, 0°, 30°, 60°, 90° について, Bi-HBSS で音響ストリーム断片を抽出した後で, それぞれのグルーピングを行い, パワーと無音状態の調整を行った後, HMM-hrtf で認識を行った. この結果を図 6 に示す. F-Grouping が一番悪いが, F-Grouping のアルゴリズムにある. 現在は, グルーピングの一貫性として使用しているのは, 基本周波数の類似性という極めて短時間の特徴である. しかし, 分離精度を向上させるためには, 基本周波数の変化動向などの中時間の特徴 [1], あるいは, 音声に特有の長時間の特徴などの活用を考慮しなければならない.

グルーピング後の残差割り当ての 2 つの方法, All-Residue と Its-Residue を評価するために, 上記と同様に, パワーと無音区間の調整を行った後で, HMM-hrtf によって, 認識を行った. その結果, 残差をすべて与えた方 (All-Residue) が認識率がよいことが分かった. また, HMM-hrtf 以外に,

HMM-org, HMM-seg でも認識実験を行ったが, HMM-hrtf 以外の認識率は極めて悪かった.

また, グルーピングによる認識率劣化への改善策として頭部伝達関数の周波数特性補正 (6KHz まで) を行って, HMM-org で認識させた. その結果, 頭部伝達関数をかけた後で, 調波構造に分離を行っているので, 図?? ほどには周波数特性補正が有効でないことが分かった.

4 考察と今後の課題

音響ストリーム分離システム HBSS, Bi-HBSS は, 調波構造抽出のために設計されている. そのような単純な音の特徴を用いて, 子音を含む音響ストリームの分離を行い, 音声認識システムを用いて分離された音の評価を行った.

(1) 調波構造抽出によるスペクトル変形はほぼ無視できる範囲にあることが分かった. 調波構造再構成音の 1 位での認識率が元のデータの認識率より約 3.5% 低下し, 7 位までの累積認識率はほぼ等しくなる.

(2) バイノーラル入力を Bi-HBSS では, 頭部伝達関数によるスペクトル変形が無視できず, 認識率が大きく低下する. その対策として, 頭部伝達関数をかけた学習データで HMM のパラメータを再学習すると, 認識率の低下が約 8% まで改善できることが分かった.

(3) Bi-HBSS へのグルーピングでは, 単純な調波構造の類似性だけを用いたのでは, 有効な判断基準にならないことが分かった.

(4) 調波構造抽出と全残差割り当てによる方法が子音を含む音の分離にも有効なことが分かった.

現在, 上記の知見を基に, 混合音での分離を行っている. 混合音からの分離の結果については, 別途報告する.

今後の課題としては, 通常のステレオ入力から音源方位を抽出する音響ストリーム分離 [8, ?, ?] の検討がまず挙げられる. というのは, ステレオ入力では頭部伝達関数の影響を受けないからである. また, グルーピングについては, 本文中で述べたようにピッチの動きを勘案したグルーピングなどを検討していく必要がある.

Bi-HBSS は, その設計思想から音声という属性を全く利用していない. 一旦分離している音が音声と分かれば, 音声の情報を活用した分離システム [15, 22] に制御を移すという枠組は, 従来から構想段階にあったが, 今後速やかに実現していかなければならない. その他に, 単一コードブック HMM に対して本稿で得られた知見の連続型 HMM への適用なども挙げられる.

5 まとめ

本稿では, 音響ストリーム分離を音声認識装置の前処理として使用するためには, 音響ストリーム分離によるスペクトル変形が課題であることを指摘した. 次に, 音響ストリーム分離の調波構造抽出, 頭部伝達関数, グルーピングという 3 つの段階でスペクトル変形の影響を調べ, その対策を提案した. 本稿では, 我々が提案してきた音響ストリーム分離を子音を含むさまざまな音に適用し, 評価を行った. その結果, システムのさまざまなチューニングの必要性が判明した.

最後に, バイノーラルシステムの設計と実装を共同で担当していただいた後藤真孝氏, 御討論いただいた柏野邦夫氏, 萩田紀博氏, 白木善尚氏, 計算環境の便宜をはかっていただいた誉田雅彰氏, 入野俊夫氏, 岡田美智男氏, 研究の機会を与えていただいた石井健一郎部長に感謝します.

参考文献

[1] Bodden: Modeling human sound-source localization and the cocktail-party-effect, *acta acustica* 1, ('93).

[2] Bregman: *Auditory Scene Analysis*, MIT Press ('90).

[3] Brown: Computational auditory scene analysis: A representational approach. PhD thesis, U. of Sheffield, '92.

[4] Cheveigne: Separation of concurrent harmonic sound: Fundamental frequency estimation and a time-domain cancellation model of auditory processing, *JASA*, 93 (6).

[5] Cooke 他: Computational Auditory Scene Analysis: listening to several things at once. *Endeavour*, 17(4), '93.

[6] 北, 川端, 斉藤: HMM 音韻認識と拡張 LR 構文解析法を用いた連続音声認識, 情処論, Vol. 31, No. 30 ('90).

[7] Lyon: A Computational Model of Binaural Localization and Separation, *ICASSP-83*.

[8] 黄, 大西, 杉江: 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 ('91)

[9] 森田: 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 ('90)

[10] 中谷, 奥乃, 川端: Auditory Stream Segregation in Auditory Scene Analysis with a MA System, *AAAI-94*.

[11] 中谷, 奥乃, 川端: 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, 10 (2), '95.

[12] 中谷, 奥乃, 川端,: A computational model of sound stream segregation with multi-agent paradigm, *ICASSP-95*.

[13] 中谷, 奥乃, 川端, : Residue-driven architecture for Computational Auditory Scene Analysis. *IJCAI-95*.

[14] 中谷, 後藤, 川端, 奥乃: 調波構造と方向同定に基づく音響ストリーム分離, 音響 7 年秋研講, '95.

[15] 長淵, 小林, 山本: 混合音声における音声強調・抑圧, 信学論, Vol.62-A, No.10 ('79).

[16] 奥乃, 中谷, 川端: Cocktail Party Effect with ComputationalAuditory Scene Analysis, *HCI-95*.

[17] 後藤, 中谷, 奥乃: カクテルパーティ効果実現のための音響ストリーム分離の検討 III. 両耳聴による音響ストリーム分離. 情処 51 全大, 2R-7, '95.

[18] Parsons: Separation of speech from interfering speech by means of harmonic selection, *JASA*, 60, 4 ('76).

[19] Ramalingam: Voiced-Speech Analysis Based on Residual Interfering Signal Canceler Algorithm, *ICASSP-94*.

[20] Rosenthal, 奥乃編: *Computational Auditory Scene Analysis — IJCAI Workshop —*, LEA Press, '96.

[21] Schmids: Multiple Emitter Location and Signal Parameter Estimation, *IEEE Trans. AP*, Vol.34, No.3 ('86)

[22] Waintraub: A Computational Model for Separating Two Simultaneous Talkers, *ICASSP-86*.