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

奥乃 博† 中谷智広† 川端 豪†

本稿では，一般環境下での音声認識のための前処理として音響ストリーム分離を使用するうえでの問題点について検討する．本稿の前半では，音声ストリーム分離の方法を提案する．提案する方法は，調波構造ストリーム断片の抽出とそのグルーピング，および，入力音からすべての調波構造を除いた残差での非調波構造の補完から構成される．本稿の後半では，分離した音声ストリームを離散型単一コードブック型HMM-LRで認識するうえでの問題点を解明し，その解決策を提示する．提案する音声ストリーム分離方法で方向情報抽出のために用いたバイノーラル入力がスペクトル変形を引き起こし，音声認識に影響を与えることが判明した．この対策として，4方向で頭部音響伝達関数をかけた学習データでHMM-LRのパラメータを再学習する方法を提案した．2人の話者の500組の子音を含んだ発話（SN比0～−3 dB）の音声認識実験を5種類行い，音声ストリーム分離により上位10候補累積認識率に対する混合音による認識誤りを最大77%削減することができた．

Speech Stream Segregation and Preliminary Results on Listening to Several Speeches Simultaneously

HIROSHI G. OKUNO,† TOMOHIRO NAKATANI† and TAKESHI KAWABATA†

This paper reports the preliminary results of experiments on listening to several sounds at once. Two issues are addressed: segregating speech streams from a mixture of sounds, and interfacing speech stream segregation with automatic speech recognition (*ASR*). Speech stream segregation (*SSS*) is designed as three processes: extracting harmonic fragments; grouping these extracted harmonic fragments according to their directions; and substituting the non-harmonic residue of harmonic fragments for non-harmonic parts of each group. The main problem in interfacing SSS with HMM-based ASR is how to reduce the recognition errors caused by spectral distortion of segregated sounds mainly due to binaural input. Our solution is to re-train the parameters of the HMM with training data binauralized for four directions. Experiments with five sets of 500 mixtures of two women's/men's utterances of a word (SNR is 0 dB to −3 dB) showed that the error of up to the 10th candidate of word recognition was reduced up to 77% by speech stream segregation.

1. はじめに

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．1953年にカクテルパーティ効果がCherryによって報告されて以来，音の分離の知覚機構について聴覚心理物理学の立場から研究が進められてきた．この研究分野は，Auditory Scene Analysis（ASA）と呼ばれ，膨大な知見が得られている[3]．しかし，これらの知見はあくまで人間の聴覚機構についてのものであり，直接工学的なモデルあるいは技術として使用できるようなものではない．1990年代になると，工学的な立場から複数の音が存在する実環境で音一般を分析し，理解することを目指すComputational Auditory Scene Analysis（CASA,『音環境理解』研究）が急速にさかんになってきた[4),5),32)]．音一般を扱おうとする音環境理解の研究の立場は，音声や楽音などの個別の音だけを扱ってきた従来の音響研究とは異なり，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研究手法の反省のうえに立ったものといえよう．本稿では音声認識を研究の対象としているが，「あくまで音声にとどまり，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する」というアプローチではなく，音一般を扱う音環境理解研究のアプローチから検討を行う．

† 日本電信電話株式会社基礎研究所
NTT Basic Research Laboratories, Nippon Telegraph and Telephone Corporation

研究室環境での音声認識というレベルは，人間に例えると，聴覚に障害のある人が静かな環境で会話をし理解できるというレベルに相当し，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．これが前述したカクテルパーティ効果の1つである．しかし，人間は同時には2つのことは聞けない．奥乃らは，計算機による聴覚機能（Computer Audition）の1つとして同時に複数のことを聞くことができるという『聖徳太子効果』[5),26)]を設定し，研究を進めている．混合音から複数の音を同時に聞くためには，音源分離の研究が重要である．実際，混合音からの音響ストリーム分離を基本機能とし，それを基に聖徳太子効果やカクテルパーティ効果のモデル化を行っている[26)]．

混合音に対する音声認識は，従来雑音抑制（noise reduction）として取り扱われてきた．雑音抑制で扱われる音声の信号雑音比（Signal-Noise-Ratio，SN比）は，10 dB以上であるのに対して，2話者が同時に同じ音量で話したとすると各々の音声の信号雑音比は0 dB以下である．言い換えると，従来の雑音抑制技法では現実的な問題が扱えているとは言い難い．したがって，音声ストリームの分離は，音声認識に対する雑音抑制，音声強調という面からも重要と考える．

混合音を分離する研究は，特定の音，特に音声に関しては古くから研究が行われており，最近では，音声・音響研究だけでなく，人工知能やロボットなどの分野でも報告が増えつつある．たとえば，中谷らは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[20),21)]，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いる[23)]．また，調波構造に基づいた音響ストリーム分離[22)]や調波構造と方向同定に基づいた音響ストリーム分離[8),24)]を開発している．しかし，これらの分離手法で分離可能な音は調波構造を持った音だけであり，調波構造を持たない無声子音を含むような一般の音声を分離する手法はまだ提案されていない．

英国シェフィールド大学のグループでは，混合音に含まれるさまざまな特徴— オンセット（音が立ち上がる時刻）やオフセット（音が消える時刻），調波構造（基本周波数の音とその整数倍の周波数を持つ倍音からなる音），共通FM変調（周波数の変化が一致しているかどうかという特徴）や共通AM変調（振幅の変化が一致しているかどうかという特徴）など — を『音響マップ』として表現し，複数の音を抽出する手がかりとしている[4)]．複数の特徴のどれを主として利用するのか，あるいは，特徴間の重要性が状況によって異なるので，注目すべき特徴と抽出法についてのプランを立てるのに，黒板モデルを用いて実現をしている[5)]．しかし，音の特徴を一括して音響マップとしてとらえるためにどの特徴が他の特徴よりも優位にあるのかといった音響ストリーム分離上の知見は，黒板モデルの知識源の中に埋め込まれており，詳細な報告は行われていない．

混合音の中から特定の音だけを分離する研究は，適応型櫛型フィルタとして実現されることが多い[2),6),30),31)]．たとえば，ドイツのグループは，『カクテルパーティ効果コンピュータ』を目指して精力的に研究を進めている[2)]．その手法は，混合音中の特定の音声を人間の頭の形をした疑似頭の耳の部分に仕込んだ一対のマイクロフォンによって録音したバイノーラル（両耳聴）入力からその音源方向を求め，適応型フィルタでその方向の音を抽出する．ただし，複数の音声や一般の音を抽出するということは考慮していないし，また，どの音に注目するのかという注視の機能も明確にはなっていない．調波構造に着目する研究[6),30)]では，子音も含むような音声の抽出についてはあまり考慮していない．また，バイノーラル入力では頭の形によってスペクトルが歪むという点についても考慮していない．なお，このようなスペクトルの変形を記述する関数を「頭部音響伝達関数」（Head-Related Transfer Function，HRTFとも略す）という．

文献31）は特定の正弦波を抽出するシステムであるが，複数の音声の抽出は考慮していない．MITメディアラボのグループは，人間の聴覚モデルに基づいたautocochleagramを用いて，音声の抽出を行っている[7)]．この手法は膨大な計算が必要であるので，実時間の処理が難しい．2話者の発話をピッチ追跡により理解するための計算モデル[36)]も提案されているが，数字列の理解にとどまっている．方向情報を用いた音源分離としては，上述のドイツのグループのようにバイノーラル入力を用いる研究[1),2)]以外に，マイクロフォンアレイを用いる方法[29),34)]などもある．

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，中谷らが開発してきた音響ストリーム分離システムを基にして，音声ストリーム分離システムを設計する．次に，音声ストリーム分離と音声認識システムとを統合するうえでの問題点を解明し，その解決策を提案する．さらに，得られたシステムを用いて同時に複数の音を聞く聖徳太子コンピュータの予備実験

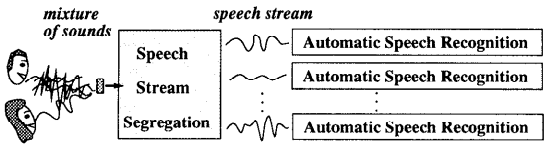


図1 聖徳太子効果のモデル化
Fig. 1 Modeling Prince Shotoku effect.

を行い，その結果を報告する．ここで『予備実験』といったのは，音声認識を目標とするならば人間のような豊富な音声に対する経験や知識を十二分に活用するのが本来であるが，本稿ではそのような情報をいっさい使わない場合に，どの程度の認識ができるかを追求することを目的としたからである．

以下，2章で，調波構造と方向を用いた音響ストリーム分離を基に音声ストリーム分離手法の提案を行う．3章で，提案する音声ストリーム分離手法の音声認識への影響を評価する．4章で，音声ストリーム分離手法と音声認識システムとの結合方法の提案を行う．5章で，複数の音声を同時に認識する実験を報告する．6章で，得られた実験結果の考察を行い，今後の研究課題を指摘する．7章で，まとめを行う．

2. 音響ストリーム分離手法

本章では，従来提案されてきた音響ストリーム分離手法を概観し，本稿の理解に必要な背景知識を説明する．次章では，それに基づいた音声ストリーム分離法を提案する．

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すると，聖徳太子効果やカクテルパーティ効果のモデル化を簡明に行うことができる．同時に10人の訴えを聴いたという聖徳太子[13]の聴覚機能は，計算機上では次のようにモデル化できよう：音響ストリーム分離により入力音から複数の音響ストリームを分離する．次に，分離した音響ストリームの中から音声ストリームだけを選別する．そして，個々の音声ストリームを別々の音声理解システムに入力をし，発話内容を理解させる（**図1** 参照）．このように人間には実現が難しい聖徳太子効果が，コンピュータによる聴覚では比較的容易に実現が行えると期待される．

☆ 計算機科学の立場での『音の表現』といえば，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意味しよう．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．

また，本稿では扱わないが，カクテルパーティ効果も同様に，音響ストリーム分離機構と，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え（focusing）機構とでモデル化できる．このようなモデル化を具体的に実現するために，このモデル化をブレークダウンして，個々の技術を確立することが本稿の目的である．

2.2 音響ストリーム分離の処理

音響ストリーム分離は，一般に2つの処理から構成されるととらえることができる[3]．

(1) **空間的な情報抽出：** 入力音から同一時間に発生した音を何らかの空間的な特徴でまとめた音響ストリーム断片抽出する．

(2) **時間的な情報抽出：** 入力音を何らかの時間的な特徴でまとめ音響ストリームを作成する．

ただし，その2つの処理がどのように構成されるのか——一方の処理に次いで，もう一方の処理が実行されるのか，あるいは，同時並行的に実行されるのか——については，まだ，十分な知見が得られていない[3]．

各処理段階で使用可能な特徴や情報にはさまざまなものがある．たとえば，低レベルの特徴としては，音のオンセットやオフセット，調波構造，共通FM変調や共通AM変調などがよく使われる．高レベルの特徴としては，音源のモデル（たとえば，音声，楽音），音源の種類（たとえば，ヘアドライヤー，電話のベル），音源の個数，リズム（たとえば，4拍子）などがある．

中谷らは，音響ストリーム分離の問題点を明確にするために，まず，調波構造という低レベルの特徴だけを用いた場合の分離性能を検討している．すなわち，現時点まで『一貫した特徴を持った音のまとまり』という曖昧な用語を用いてきたが，中谷らは『調波構造の基本周波数は連続的に変化する限り同一のストリームが継続する』という制約条件によって『調波構造ストリーム断片』を定義している．そして，調波構造に基づく音響ストリーム断片分離システムHBSSを開発している[20),21]．HBSSは，残差駆動型アーキテクチャ[23]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している（**図2** 参照）．

(1) 変化検出器（Event Detector）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（Tracer Generator）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（F_0）を求め，それを追跡する追跡器（Tracer）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（Noise-Tracer）に通知する．

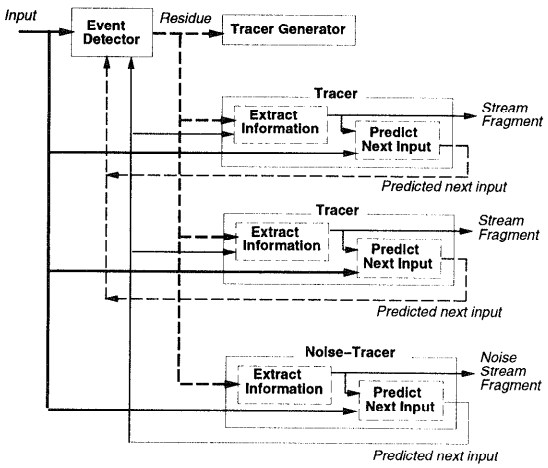


図2 残差駆動型アーキテクチャに基づくHBSS
Fig. 2 HBSS designed by residue-driven architecture.

(3) 追跡器は，与えられた基本周波数を基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あらかじめ与えたり，あるいは，音源の個数が固定されている必要はない．

HBSSでは，基本周波数が接近している複数の調波構造ストリームを分離することは難しい．たとえば，2つの調波構造ストリームの基本周波数が交差しているときに，2つのストリームが実際に交差したのか，それとも接近した後再び離れていくのかは調波構造だけでは分からない．

このような問題に対して，音源の方向情報を用いて解決する方法を中谷らは提案し，Bi-HBSSというシステムを開発している．Bi-HBSSは，モノラル入力から調波構造を分離するHBSSシステムを左右2個使用し，調波構造の基本周波数（F_0）の同定を行う．方向情報は，バイノーラル入力の左右の各チャンネルから各々のHBSSが分離した同じF_0を持つ調波構造ストリーム断片の強度差（IID, Interaural Intensity Difference）と到着時間差（位相差）（ITD, Interaural Time Difference）から求める．通常，方向情報を求める場合には全周波数帯域に対するIIDとITDを計算する[16]．それに対して，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．音源方向は，角度ではなくITDで表現している．また，調波構造ストリーム断片の分離の終了時に，そのストリーム断片と同じ方向に後続した無声子音があるかをチェックし，存在する場合には後続フラグを設定する．

IIBSSとBi-HBSSは，残差駆動型アーキテクチャ[23]で構成されており，調波構造を入力音から減算し，得られた残差を使用して，新しい入力音の検出と分離中の複数の音響ストリーム断片への信号の分配を行っている[20),21]．したがって，残差にはほとんど調波構造が含まれていない．

3. 音声ストリーム分離法の設計

中谷らの提案した音響ストリーム分離システムは，混合音から調波構造の音を分離することができている．しかし，音声には母音や有声子音などの調波構造の音だけでなく，無声子音のように調波構造を持たない音も含まれる．したがって，本稿の前半の課題は，調波構造と音源方向を用いた調波構造ストリーム分離システム，Bi-HBSS，を使用して，音声を抽出するシステムを設計し，実装すること，と言い換えることができる．

音声には『母音（V）・子音（C）・母音（V）』という構造が含まれることが多い．このうち，混合音から調波構造ストリーム断片（母音や有声子音）を分離するために，前述したBi-HBSS[23),25)]を用いることとした．Bi-HBSSで調波構造ストリーム断片を分離する際に入力音から追跡中のすべての調波構造を除いた残差が副作用として得られる．この残差を非調波構造部分の補完に使用することとした．

つまり，本稿で提案する音声ストリーム分離システムは，以下のような3段階から構成される．

(1) 調波構造ストリーム断片の抽出
(2) 調波構造ストリーム断片のグルーピング
(3) 非調波構造部分の残差による補完

各段階では，出力は逐次的に行われるので，全体の処理は，ある段階での処理がすべて終了した後，次の段階での処理が始まるのではなく，ある段階での出力がされると，後続の段階の処理が始まる．

3.1 入力音の録音条件と実験条件

調波構造ストリーム断片の抽出部分は，前述したように既発表のものを使用したので，説明は省略する．本節では，本研究での入力音の条件を説明する．

本研究で使用した単一音源のバイノーラル音と混合音はすべて合成したものである．このような合成音の使用は，以下に述べる簡単な実験により妥当と判断をした．

実験1： 実際に測定したバイノーラル音から合成した混合音から音響ストリーム分離の実験を行う．混合音の作成方法を下記に示す．

(1) 静かな部屋でマイクロフォンによりDATで「あいうえお」を録音した．この録音データを「モノラル・データ」と呼ぶ．

(2) 無響室でスピーカ1個からモノラル・データを再生し，ダミーヘッドマイクロフォンによりバイノーラルで録音した．ダミーヘッドに埋め込まれたマイクロフォンは無指向性であり，マイクロフォンの間隔は20 cmである．また，スピーカは，ダミーヘッドマイクロフォンを中心に半径2 mの円周上に，0°（真正面），30°，60°，90°（真横）に置いた．この4種類の録音データを「バイノーラル・データ1」と呼ぶ．

(3) 別々の角度のバイノーラル・データ1を時間を1秒ずらして重ね合わせた混合音を合成した．これを「実録音による混合音」と呼ぶ．

実録音による混合音をBi-HBSSで分離し，2つの調波構造ストリームが分離できることを確認した．

実験2： 音源方向によるスペクトル変形を記述した頭部音響伝達関数を使用してバイノーラル音を作成し，それらを基に合成した混合音から音響ストリーム分離の実験を行う．

使用した頭部音響伝達関数は，予備実験1と同じ条件で，音源（スピーカ）を水平方向で30°ごとに，垂直方向で10°ごとに半径2 mの球面上に置き，ダミーヘッドの左右1対のマイクロフォンで測定したインパルス応答の時系列信号である[12]．つまり，インパルス法[1]による測定である．モノラル信号とこのインパルス応答のたたみこみ演算を行うことによって，左右2チャンネルからなるバイノーラル信号を合成する．

混合音の作成方法を以下に示す．

(1) モノラル・データに頭部音響伝達関数をかけて，0°，30°，60°，90°の方向からのバイノーラル・データを合成する．このデータを「バイノーラル・データ2」と呼ぶ．

(2) 実データの場合と同様にバイノーラル・データ2から混合音を合成する．これを「合成による混合音」と呼ぶ．

合成による混合音をBi-HBSSで分離し，2つの調波構造ストリームが分離できることを確認した．

この2つの実験から，実録音による混合音および合成による混合音ともBi-HBSSで分離ができることを確認した．さらに，分離されたバイノーラル音の単一音源のバイノーラル音（元の音）に対する平均LPCスペクトル歪みが，両者の実験でほとんど差がないことも確認をした．したがって，合成による混合音を使用しても音響ストリーム分離から得られる結果が実録音による混合音を使用した場合とそれほど差がないと判断した．

3.2 調波構造ストリーム断片のグルーピング

調波構造ストリーム断片は，基本周波数と方向情報を持っているので，それらをグループ化するためのストリーム（断片）間の近接性として，基本周波数の近さ，音源方向の近さ，両者の組合せなどが考えられる．今，調波構造グループあるいは調波構造ストリーム断片 ϕ の時間フレーム t での基本周波数（F_0）を $f_{\phi,t}$ で表現することにする．本稿では調波構造グルーピングとして，以下の3種類の方法を検討した．

(1) **F-グルーピング**：調波構造の近接性に着目し，基本周波数の差が閾値以下なら同じグループと判定する．

具体的には，ストリーム断片 ϕ に対して，$|f_{\phi,t} - f_{\Psi,t-1}| < \delta$ を満たすようなすでに存在するグループ Ψ を探す．予備実験の結果から，δ の値として，ϕ と同時に他の新しいストリーム断片がない場合には300 cent☆とし，それ以外の場合には600 centとする[8]．複数の既存のグループが上記の条件を満足する場合には，ϕ に最も近いグループに統合し，1つのグループしか存在しない場合にはそのグループに統合する．もし，そのような既存のグループが見つからなかった場合には，新たなグループを生成する．

(2) **D-グルーピング**：方向情報の近接性に着目し，音源方向の差が閾値以下なら同じグループと判定する．閾値としては両耳間時間差（ITD）で0.167 msecを使用している．この時間差に対応する角度差は方向によって異なるが，おおむね20°である．

複数のグループが候補にあがったときには，F-グルーピングと同様の方法でグルーピングを行う．

(3) **B-グルーピング**：上記2つの組合せであり，両者の近接性に重みをかけて判定する[8]．

ストリーム断片 ϕ に対して，上記2つの条件を満足する既存のグループ Ψ を探す，もし，1つだけ見つかった場合には ϕ を Ψ に統合する．複数のグループが見つかった場合には，以下のような組合せ距離 K を定義し，最少の組合せ距離を持つグループ Ψ に統合する．

☆ centは音程の単位であり，1オクターブが1200 centになる．

$$K = \alpha\frac{|\triangle f|}{c_f} + (1-\alpha)\frac{|\triangle d|}{c_d}$$

ただし，c_f は300 centであり，c_d は0.167 msecである．また，正規化因子 α の値は0.47である．

なお，HBSSでは方向情報が得られないのでF-グルーピングだけが適用できる．

連続していないストリーム断片を同じグループに入れるかどうかは，不連続時間許容値で制御する．すなわち，あるグループに対して，ただちに後続するストリーム断片がなく，不連続時間許容値内に別のストリーム断片が発生した場合には「中抜け」という情報とともにそのストリーム断片が同じグループに統合される．現在の不連続時間許容値は，500 msecである．この値は，使用した単語発話データベースで継続するVとVとの間隔の最大値から求めた．

3.3 非調波構造の残差による補完

一般に無声子音の分離はきわめて困難である．人間の聴覚知覚では，『**聴覚的誘導**』(*auditory induction*)という現象がASA分野では報告されている[9),14),35)]．すなわち，ある音が別の音でマスキングされているにもかかわらず，マスキングされた元の音があたかも聞こえるような現象である．本稿では，子音の部分をそれに似た音で補完することにより音声認識システムに対して『聴覚的誘導』を生じさせ，実質的に子音の抽出に相当させることができないかを検討することとした．

調波構造グルーピングにおいて，調波構造ストリーム断片は存在しないが，最後の調波構造ストリーム断片の後続フラグが立っている場合には，その調波構造ストリーム断片が属するグループには，中抜けの部分に残差が割り当てられる．残差の割当て方法としては，次の2つの方法が考えられる．

(1) **全残差（All-Residue）**：残差すべてを割り当てる．

(2) **自己残差（Its-Residue）**：残差のうち，そのグループの音源方向の成分だけを割り当てる．

後者の方法では，音源方向が分かっていなければならないが，本手法で用いた音源方向同定の精度は約20°とかなり粗い．また，残差には主として非調波構造だけが含まれるので，全帯域にわたって音源方向による抽出をしなければならず，その計算量が調波構造だけの場合と比べて膨大になる．さらに，非調波構造の部分はパワーがもともと弱いので，特定の方向だけを取り出すとさらにパワーが弱くなり，有意な情報が得られなくなる可能性がある．この3つの理由から，本稿では前者を採用した．ただし，評価のために，本章の実験では，方向が分かっているものとして（2）の計算を行った．

3.4 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[15)]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) 頭部音響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

これらの影響をより詳しく調べるために，スペクトル包絡が変化したことによる累積認識率の低下を測定する．

Bi-HBSSでは，音源の方向情報を用いてオンセット，オフセットを求めている．それに対して，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ので，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とき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，本稿では，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40 msec早め，逆に調波構造が終わった後も音がある場合には，オフセットを40 msec遅らす処理をした．

3.5 音声認識システムの諸元

本稿で使用した音声認識システムは，ATRで開発された隠れマルコフモデルに基づいたHMM-LRシステム[15)]である．HMM-LRの諸元を以下に示す．

- 音響分析条件：サンプリング周波数12 kHzでAD変換後，フレーム周期3 msごとに256点ハミング窓で切り出し，12次LPC分析，16次PWLR距離尺度を用いてVQコード列（コードサイズ256）に変換する．
- 認識モデル：過渡的な音韻には4状態3ループ，定常的な音韻には2状態1ループのモデル構造を設定し，音韻ごとのleft-to-right型の離散型隠れマルコフモデルを使用する．

音声データは，ATRで作成されたものを使用し，標準的な音韻データを男女別に作成した．HMMパラメータの学習には，男女それぞれ5人の話者のデータを使用した（**表1**）．本稿では，『**単語認識**』を対象とするために，LR文法は，スタート記号からターミナ

表1 学習データの内訳
Table 1 Training data.

データ集合	個数	話者（男性）	話者（女性）
D_0	999	MMS	FAF
D_1	1,000	MAU	FKM
D_2	1,000	MNM	FKS
D_3	1,000	MTK	FSU
D_4	1,000	MTT	FYM
D_5	121	MAU, MNM, MTT	FAF, FKM, FKS
D_6	120	MMS, MTK	FSU, FYM
計	5,240	（総計 5,602）	（総計 5,602）

表2 評価データ
Table 2 Evaluating data.

データ集合	（個数）	話者（男性）	話者（女性）
D_0	999	MAU	FAF

ル記号に直接落ちるルールだけから構成される．

本稿では，累積認識率の尺度として，認識結果の順位による上位10候補累積認識率を使用する．上位10候補を考慮する理由は以下のとおりである．通常の音声認識システムは単独で使用されるのではなく，後続の言語処理と組み合わせて使用されることが多く，その場合音声認識部からの複数の候補が言語処理部によってさらに絞り込まれていくからである．

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上位10候補累積認識率を『**オリジナルデータ**』と略記する．

4. 提案する音声ストリーム分離法の音声認識への影響

提案する音声ストリーム分離法は，オリジナルの音からバイノーラル化された入力音をとり，調波構造と非調波構造に分解し，それからオリジナルの音を再構築する．このような変換によって，オリジナルの音から何らかの情報が失われている可能性がある．そのような原因としては，次の3点が考えられる．

(1) 調波構造抽出による情報損失
(2) 頭部音響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

本章では，上記3つの情報損失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．具体的には，(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．

なお，これらの評価には以下の理由により男性のデータを使用した．女性用のコードブックは，代表点が男性のと比べると近接しており，小さなノイズ等の影響を受けやすいからである．特に，無音モデルに認識結果が敏感に影響を受けるので，上記の3種類の実験のパラメータの制御が難しいと判断した．

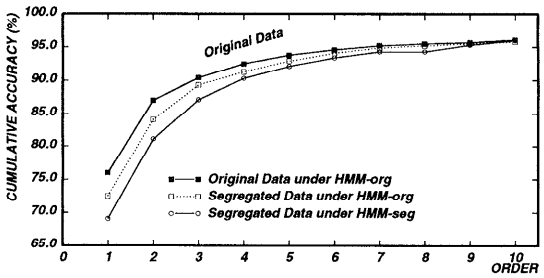


図3 調波構造再構成による累積認識率への影響と再学習による累積認識率の改善
Fig. 3 Effect on recognition performance by harmonic structure reorganization and recovery by retraining HMM parameters.

4.1 調波構造抽出による影響

調波構造抽出が上位10候補累積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，ストリーム断片がいっさい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上位10候補累積認識率を図**3**に示す．1番目（■）の実線が，評価データのHMM-orgによる上位10候補累積認識率（オリジナルデータ）である．2番目（□）の点線が，評価データの調波構造再構成音のHMM-orgによる上位10候補累積認識率である．第1位候補認識率では3.5%劣るが，上位7候補累積認識率でほぼ等しくなる．したがって，上位10候補累積認識率の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累積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による上位10候補累積認識率を同じ図3の3番目（○）の実線で示す．図から，再学習による上位10候補累積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，D_0以外のデータ集合では上位10候補累積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に対する特別の対策は必要がないと判断をした．

4.2 頭部音響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

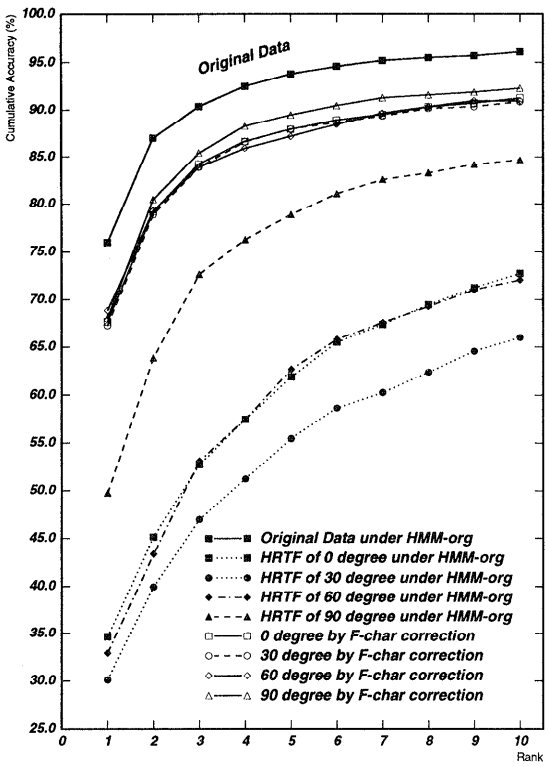


図4 頭部音響伝達関数による累積認識率への影響と頭部音響伝達関数周波数補正による改善

Fig. 4 Effect on recognition performance by HRTF and recovery by F-char correction.

部音響伝達関数がかかったものと等価である．頭部音響伝達関数による影響は主として次の2点である．

(1) パワーの変化

(2) 周波数特性の変化

0°，30°，60°，90° の4種類の頭部音響伝達関数の上位10候補累積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，音源方向情報でグルーピング（D-グルーピング）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，頭部音響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上位10候補累積認識率を**図4**の下4つの線で示す．図から明らかなように，方向による上位10候補累積認識率低下の差が大きい．90° のときが20%と最も小さく，他の角度では40%以上にも上る．

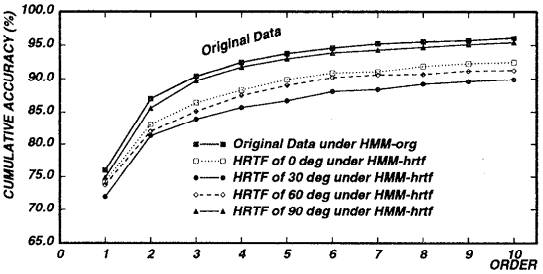


図5 バイノーラルデータによる再学習後の累積認識率の改善

Fig. 5 Recognition Performance by HMM-LR re-trained with binauralized training data.

4.3 頭部音響伝達関数による累積認識率劣化への対策

頭部音響伝達関数による上位10候補累積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) 頭部音響伝達関数の周波数特性補正（F-char correction と略す）

(2) 頭部音響伝達関数をかけた学習データによるHMMパラメータの再学習．

頭部音響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，頭部音響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．その上位10候補累積認識率を図4の真ん中4つの線で示す．上位10候補累積認識率は大幅に改善され，かつ，音源方向による上位10候補累積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° の頭部音響伝達関数をかけ，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のHMMパラメータを再学習させた（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．前節での頭部音響伝達関数をかけた評価データのHMM-hrtfによる上位10候補累積認識率を**図5**に示す．90° の場合の上位10候補累積認識率は，オリジナルデータと比べてほとんど遜色がない．音源方向によって，上位10候補累積認識率に差が出るが，頭部音響伝達関数の周波数特性補正よりも上位10候補累積認識率が改善されている．

頭部音響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，頭部音響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，頭部音響伝達関数の周波数特性補正による上位10候補累積認識率が劣ることである．第2

☆ 90° の方向（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．

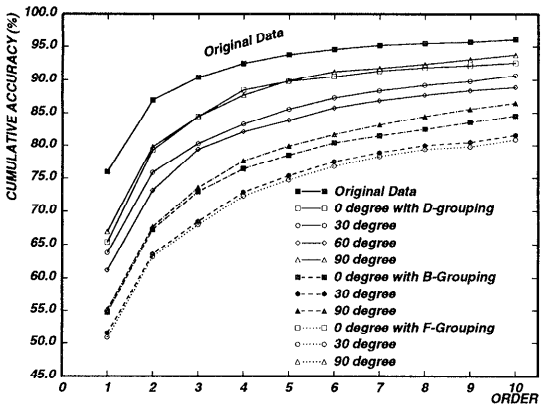


図6 調波構造グルーピングによる累積認識率への影響
Fig. 6 Effect on recognition performance by harmonic structure grouping.

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音響伝達関数のデータは 30° ごとにしか測定されていない[12]．

4.4 グルーピングによる上位 10 候補累積認識率への影響

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-グルーピングを評価する．4 つの方向，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-グルーピングが一番悪い．しかし，これは現在の F-グルーピングのアルゴリズムが単純すぎることに原因がある．というのは，F-グルーピングの一貫性として使用しているのは，1 つ前のフレーム（7 ms）での基本周波数の近接性というきわ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，基本周波数の時間的な変化傾向といったより長い時間の特徴[2]，あるいは，音声に特有な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）が上位 10 候補累積認識率が良いことが分かった．また，HMM-hrtf 以外に，HMM-seg や HMM-org（頭部音響伝達関数の周波数特性補正をかけた後で）でも認識実験を行ったが，HMM-hrtf と比べて上位 10 候補累積認識率はきわめて悪かった．

表 3 5 組のベンチマークの内訳
Table 3 Five set of benchmarks of mixed sounds.

No.	第 1 音	第 2 音	第 3 音
1	女性 1	女性 2	—
2	男性 1	女性 2	—
3	男性 1	男性 2	—
4	女性 1	女性 2	断続音
5	女性 1	女性 2	断続音（4 より強い）

1，2，3 での各音の SN 比は 0 dB，4，5 でのそれは 0 dB 以下．

5. 音声認識システムとの結合による複数音声の同時認識実験

前章での結果から，音声ストリーム分離には，Bi-HBSS を用いて調波構造ストリーム断片を抽出し，次に，抽出された調波構造ストリーム断片を D-グルーピングで調波構造部分を回復し，さらに，非調波構造部分を全残差によって補完する方法が，最も音声認識率が良いことが分かった．また，音声認識システム側の対処としては，さまざまな方向から入ってくるバイノーラル音により HMM パラメータを再学習する必要があることが分かった．

5.1 HMM パラメータの再学習

学習に用いた 5,240 個の単語発話データを 4 つの角度（0°，30°，60°，90°）でバイノーラル化し，それで再学習させた．前節で述べたこの再学習による上位 10 候補累積認識率低下の改善は，4 方向に対するデータを作成したことにより，量子化ベクトルがなまり，少々のスペクトル変形に対してロバスト性が増したからだ，と考えられる．なお，バイノーラル化は，頭部音響伝達関数を基に解析的に行った．また，無音モデルとして，学習データの無音部分のベクトルを利用した．

5.2 オープンテスト用ベンチマーク

2 音と 3 音の混合音 500 組を 5 種類用意した（**表 3**）．単語の 500 組の組合せはすべてに共通である．第 1 音はマイクから見て正面から左 30° に位置する話者が発声し，第 2 音はその 150 msec 後に右 30° に位置する話者が発声した．一方の音は，他方の音に対して雑音として働くので，2 音混合音中では，両者の平均信号雑音比（SN 比）は，各々0 dB である．混合音は既録音のデータを加えて作成し，混合音の最大振幅が 16 ビットに収まるように最大 2 dB まで減衰をさせた．また，混合音をそのままでも認識できるように第 1 話者と第 2 話者の発話をずらした．なお，HMM-org による男性 1（MTT），男性 2（MAU），女性 1（FSU），女性 2（FKM）の単一話者での上位 10 候補累積認識率（単

表4 混合音による認識誤り率
Table 4 Error rates caused by mixed sounds.

混合音	第1話者	第2話者
1	76.19%	95.50%
2	57.79%	95.70%
3	56.39%	94.70%
4	94.99%	95.70%
5	94.99%	95.90%

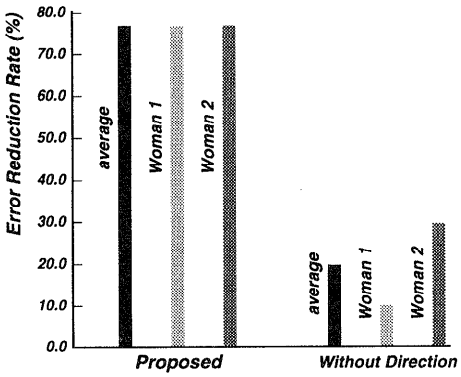


図7 ベンチマーク1の認識誤り削減率
Fig. 7 Error reduction rates for benchmark 1.

語認識）は，各々94.19%，95.10%，94.99%，96.10%である．

第3音は第1話者が話す前から真正面（0°）からずっと鳴っている F_0 が250 Hzの断続音（1秒継続後50 msec休止）である．ベンチマーク4と5はベンチマーク1に断続音が加わったものである．ベンチマーク4と5の断続音のパワーは，第1音と第2音の平均パワーに対して各々−1.7 dBと1.3 dBになるように調整した．また，2音の混合音の場合と同様に，混合音の最大振幅を16ビットに収まるように最大4 dBまで減衰させた．1つの音声から見ると，もう1方の音声と断続音は妨害音として働くので，SN比がベンチマーク1よりも一般には低下する．ただし，音が混合されることによって混合音のパワーが増える場合もあれば，逆に音がキャンセルされて混合音のパワーが減る場合もあり，各音の混合音中でのSN比は簡単には測定できない．

混合音にすることによって，元の音の累積認識率は大幅に低下する．本稿では，上位10候補累積認識率の低下を『**混合化による認識誤り率**』と定義し，以降単に『認識誤り率』と呼ぶ．モノラルの混合音を元の単音でのラベルを使用して，そのままHMM-orgで単語認識させたときの上位10候補累積認識率を $\mathcal{CA}_{mix}$ とする．また，単音をHMM-orgで認識させたオリジナルデータの上位10候補累積認識率を $\mathcal{CA}_{org}$ とする．このとき，認識誤り率 $\mathcal{E}$ を $\mathcal{E} = \mathcal{CA}_{org} - \mathcal{CA}_{mix}$ で定義する．**表4** に5種類の混合音でのそれぞれの音の認識誤り率をまとめておく．

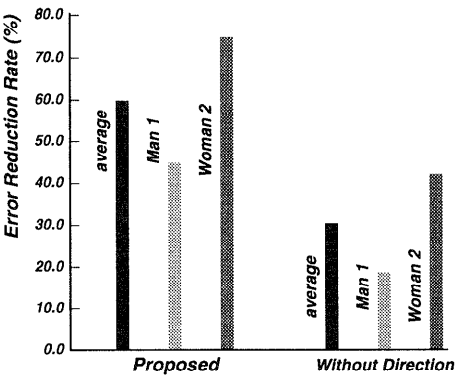


図8 ベンチマーク2の認識誤り削減率
Fig. 8 Error reduction rates for benchmark 2.

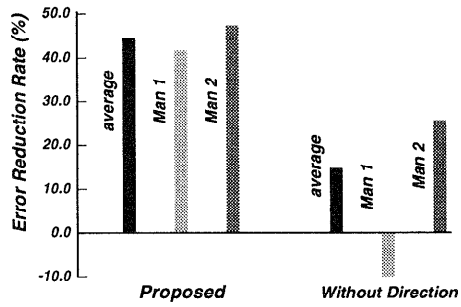


図9 ベンチマーク3の認識誤り削減率
Fig. 9 Error reduction rates for benchmark 3.

次に，音声ストリーム分離による『**認識誤り削減率**』（*Error Reduction Rate*）を以下のように定義する．音声ストリーム分離によって得られた音声ストリームに対する単語認識の上位10候補累積認識率を $\mathcal{CA}_{seg}$ とすると，**認識誤り削減率** $\mathcal{R}$ を次式で計算をする．

$$\mathcal{R} = \frac{\mathcal{CA}_{seg} - \mathcal{CA}_{mix}}{\mathcal{CA}_{org} - \mathcal{CA}_{mix}} \times 100.0$$

$$= \frac{\mathcal{CA}_{seg} - \mathcal{CA}_{mix}}{\mathcal{E}} \times 100.0$$

5.3 2音の混合音の音声認識の改善

ベンチマーク1〜3における音声ストリーム分離による上位10候補累積認識率の認識誤り削減率を**図7**〜**図9** に示す．『本手法』（図中の **Proposed**）の項が本稿で提案する音声ストリーム分離法によって分離した音声ストリームをHMM-hrtfで認識させたときの単語認識誤りの削減率である．『方向なし』（図中の

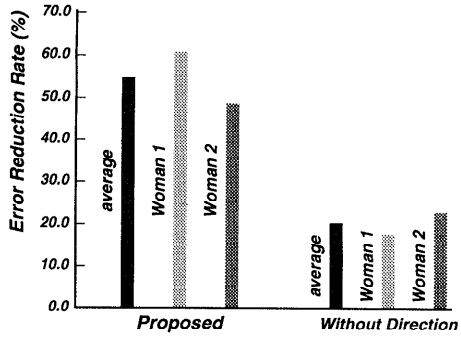


図10 ベンチマーク4の認識誤り削減率
Fig. 10 Error reduction rates for benchmark 4.

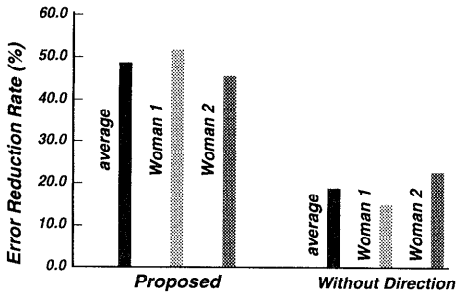


図11 ベンチマーク5の認識誤り削減率
Fig. 11 Error reduction rates for benchmark 5.

Without Direction）はモノラル入力をHBSSで調波構造ストリーム断片を分離し，F-グルーピングで調波構造ストリーム断片をグループ化し，さらに，全残差で補完した音をHMM-orgで認識させたときの認識誤り削減率である．

女性2名の場合には上位10候補累積認識率に対して77%強の認識誤り削減率が達成されている．男性の発話に対する認識誤り削減率は，女性の発話に対するものよりも劣っている．この理由は，一般に基本周波数 F_0 が低いほど，また，2つの音の F_0 が接近しているほど，調波構造の分離が難しくなるからである．男性の F_0 が女性よりも低く，男性2名の場合には F_0 の近接度が必然的に大きくなるので，分離の精度が悪く，認識誤りがそれほど削減されていない．特に，男性1の F_0 は100 Hz前後であるので，方向なしの場合には，逆に認識誤りが増している（図9の方向なしの男性1）．これは，ベンチマーク3で，混合音による認識誤り率が男性1と男性2とで大きく違うことに起因している（表4の混合音3）．すなわち，方向なしで分離された男性2の音声ストリームの先頭部分の分離精度が悪いためにHMMレベルで間違いが増幅されたのに対して，混合音をそのまま認識させた場合には男性1の音声ストリームの後半部が男性2の音声で破壊されているだけであるので，HMMレベルで間違いが回復できたため，認識誤り率がそれほど大きくはなかった．方向情報を用いた場合には，両方とも同じ割合で認識誤りが削減されているので，本手法の有効性が確認できたと考えられる．

5.4 3音の混合音の音声認識の改善

ベンチマーク4と5に対する音声ストリーム分離による単語認識（上位10候補）での認識誤り削減率を**図10**と**図11**に示す．断続音が入ることによって，音声のSN比が他の音が同時に存在するときに0 dBから，ベンチマーク4では1 dB程度低下し，ベンチマーク5ではベンチマーク4よりさらに1 dB程度低下する．その結果，認識誤り削減率がそれぞれ平均23%，28%低下した．従来の音声認識のための雑音抑制技術が音声のSN比として10 dB程度を想定していることを考慮すると，このように大幅にSN比が低下しているにもかかわらず，ベンチマーク4では平均55%の認識誤り削減率が達成できたととらえるべきであろう．さらに，ベンチマーク5では，断続音のパワーがベンチマーク4のそれよりも倍になっているにもかかわらず，認識誤り削減率は平均6%しか低下していない．これらの2つのベンチマークでは，第1音の前からずっと断続音が鳴っているので，表4に示したように分離をしない場合にはまったく認識できていない．

6. 考察と今後の課題

音響ストリーム断片分離システムHBSS，Bi-HBSSは，調波構造抽出のために設計されている．調波構造と音源方向という単純な音の特徴だけを用いて，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．さらに，実際に2つの音声の混合音とさらにもう1音混ざった混合音から音声ストリームを分離し，分離した音声ストリームを音声認識する実験も行った．得られた結果を以下にまとめる．

（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第10位候補認識率が元のデータのそれより約3.5%低下し，上位7候補累積認識率はほぼ等しくなる．

（2） バイノーラル入力をとる本手法では，頭部音響伝達関数によるスペクトル変形が無視できず，第10位候補累積認識率が大きく低下する．その対策として，頭部音響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，第10位候補累積認識率の

低下が約8%まで改善できることが分かった．

(3) Bi-HBSSで抽出された調波構造ストリーム断片のグルーピングでは，単純な調波構造の近接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．また，音源の方向情報の近接性を用いたグルーピングの方が有効であることが分かった．

(4) 非調波構造部分の抽出を音声認識の観点から見ると，入力音から追跡中の調波構造を除いた残差そのもので補完した方が残差の方向成分を取り出した場合よりも，累積認識率が良いことが分かった．この手法は，聴覚心理物理学で知られている『聴覚的誘導』という概念の工学的な実現となっている．

(5) 調波構造と音源の方向情報というきわめて単純な音の特徴しか使用していないにもかかわらず，2話者の混合音から音声ストリームを分離することによって，上位10候補単語認識の認識誤りを最高77%削減できた．ここで用いた音声のSN比が，0dBあるいはそれ以下という従来の雑音抑制法が扱えないような悪い領域での音声認識であり，実環境での音声認識への可能性が示されたのではないかと考える．

調波構造と音源方向という最低レベルの手がかりだけを使用した場合でも音声強調に有効であることが確かめられたと考える．しかし，本稿で得られた認識誤り削減率では，音声ストリーム分離を実用化するにはまだ程遠い．この原因としては，

(1) 調波構造ストリーム分離システムの性能が不十分であるために残差に調波構造が残っている，

(2) 音源方向抽出にバイーノラル音を使用したことにより頭部音響伝達関数によるスペクトル歪みが無視できないくらい大きい，

(3) 使用した音声認識システムが古く，その性能が最近のものと比べて劣っている．

などが考えられる．

今後の検討課題としては，上記の原因をさらに追求し，調波構造が有効な手段かを判断することが必要である．音声に関する限り，最も得やすい情報は調波構造（ピッチ）であるので，調波構造を中心に，さらに他の情報を二次的に使用することによって，音声ストリーム分離がどの程度の性能まで行くかを見極めることが重要であると考える．具体的な検討課題としては，以下のような項目を列挙することができる．

(1) 本稿で利用したバイノーラル入力では，頭部音響伝達関数の影響が大きいので，通常のステレオ入力，あるいはマイクロフォンアレイによる音源方向を抽出する手法[18),29),33)]を検討する必要がある．また，調波構造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手法などを検討していく必要がある．

(2) 本稿で得られた単一コードブック型HMMに対する知見を援用して，音声認識性能のより優れた連続型HMMを使用し，提案する音声ストリーム分離システムの性能を再検討する必要がある．

(3) 本稿で提案した手法は，調波構造と方向情報だけを使用して調波構造ストリーム断片を抽出しており，音声という属性は非調波構造部分の補完にしか利用していない．しかし，人間は豊富な経験と膨大な知識を用いて音声に対する聴覚処理を行っている[10)]．コンピュータ聴覚でも同じようなことが行える枠組みを追求する必要がある．たとえば，分離している音がいったん音声と分かれば，音声の情報を活用した分離システム（たとえば，文献19), 36)）に制御を移すという統合的な枠組みを今後検討していかなければならない．

(4) 音声認識では，単語全体の認識は失敗しているものの，部分的な認識には成功している場合も少なからずあるが，これらは現在の評価では失敗として扱われている．このことから，より上位の言語モデルと音声ストリーム分離とを結合するようなボトムアップ処理とトップダウン処理との統合方式を検討していく必要がある．

7. まとめ

本稿では，音声ストリーム分離を音声認識システムの前処理として使用するための課題について検討を行った．まず，中谷らが開発した調波構造ストリーム分離システムを用いた音声ストリーム分離システムを提案した．本稿で新たに検討した部分は，調波構造ストリーム断片のグルーピング方法と残差を用いた非調波構造部分の補完方法である．

次に，音声ストリーム分離システムと音声認識システムとの結合時の問題点として，音声ストリーム分離によるスペクトル変形を検討した．音響ストリーム分離の調波構造抽出，頭部音響伝達関数，グルーピングという3つの段階でスペクトル歪の影響を子音を含むさまざまな音で調べた．その結果，バイノーラル入力が受ける頭部音響伝達関数によるスペクトル変形が無視できないことが判明した．

このスペクトル変形による音声認識システムへの影響を削減するために，4方向からの音声データによってHMMパラメータの再学習を行い，音声認識システムのロバスト性を強化した．この結果，2人の話者の500組の子音を含んだ単語発話5種類の音声認識にお

いて，他の音が混合することによってもたらされる上位10候補単語認識における認識誤りを最大77%削減することができた．この性能はまだまだ実用化には程遠いが，混合音の音声認識の可能性が少し開けたと考える．

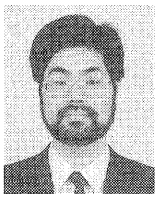
謝辞 最後に，バイノーラルシステムの設計と実装を共同で担当していただいた後藤真孝氏，ご討論いただいた柏野邦夫氏，柏野牧夫氏，萩田紀博氏，白木善尚氏，計算環境の便宜を図っていただいた誉田雅彰氏，頭部音響伝達関数のデータを提供いただいた入野俊夫氏，研究の機会を与えていただいた石井健一郎部長に感謝する．

参考文献

1) Blauert, J.: *Spatial Hearing*, MIT Press (1983). 邦訳：空間音響，鹿島出版会 (1986).
2) Bodden, M.: Modeling Human Sound-source Localization and the Cocktail-party-effect, *Acta Acustica*, Vol.1, pp.43–55 (1993).
3) Bregman, A.S.: *Auditory Scene Analysis – the Perceptual Organization of Sound*, MIT Press (1990).
4) Brown, G.J.: Computational Auditory Scene Analysis: A Representational Approach, Ph.D diss., Univ. of Sheffield (1992).
5) Cooke, M.P., Brown, G.J., Crawford, M. and Green, P.: Computational Auditory Scene Analysis: Listening to Several Things at Once, *Endeavour*, Vol.17, No.4, pp.186–190 (1993).
6) de Cheveigne, A.: Separation of Concurrent Harmonic Sound: Fundamental Frequency Estimation and a Time-domain Cancellation Model of Auditory Processing, *J. Acoust. Soc. Amer.*, Vol.93, No.6, pp.3271–3290 (1993).
7) Ellis, D.P.W. and Rosenthal, D.F.: Mid-level Representations for Computational Auditory Scene Analysis, *Workshop Notes of IJCAI-95 Workshop on Computational Auditory Scene Analysis* (Aug. 1995). Also in Ref. 32).
8) 後藤真孝，中谷智広，奥乃 博：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，第51回情報処理学会全国大会論文集，2R-7 (1995).
9) Green, P.D., Cooke, M.P. and Crawford, M.D.: Auditory Scene Analysis and Hidden Markov Model Recognition of Speech in Noise, *Proc. 1995 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-95)*, pp.401–404, IEEE (1995).
10) Handel, S.: *Listening: An Introduction to the Perception of Auditory Events*, MIT Press (1989).
11) Hansen, J.H.L., Mammone, R.J. and Young, S.: Editorial for the Special Issue on Robust Speech Processing. *Trans. Speech and Audio Proc.*, Vol.2, No.4, pp.549–550 (1994).
12) Irhino, T.: Modeling of the Head Related Transfer Function to Extract Features Usable in Sound Localization, Tech. Report: ISRL-93-7, NTT BRL (1993).
13) 推古天皇即位前記—元年四月『日本書紀』第二十二巻，日本古典文学全集，pp.172–173, 岩波書店.
14) 柏野牧夫：音の流れを聞き取る，科学，Vol.62, No.6 (1992).
15) 北 研二，川端 豪，斉藤博昭：HMM音韻認識と拡張LR構文解析法を用いた連続音声認識，情報処理学会論文誌，Vol.31, No.3, pp.472–480 (1990).
16) Lyon: A Computational Model of Binaural Localization and Separation, *Proc. 1983 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-83)*, IEEE (1983).
17) Minami, Y and Furui, S.: A Maximum Likelihood Procedure for a Universal Adaptation Method Based on HMM Composition, *Proc. 1995 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-95)*, Vol.1, pp.129–132, IEEE (1995).
18) 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, pp.1551–1557 (1990).
19) 長淵裕実，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, pp.627–634 (1979).
20) Nakatani, T., Okuno, H.G. and Kawabata, T.: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-agent System, *Proc. 12th Nat. Conf. on Artificial Intelligence (AAAI-94)*, pp.100–107 (1994).
21) 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).
22) Nakatani, T., Kawabata, T. and Okuno, H.G.: A Computational Model of Sound Stream Segregation with the Multi-agent Paradigm, *Proc. 1995 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-95)*, pp.2671–2674, IEEE (1995).
23) Nakatani, T., Okuno, H.G. and Kawabata, T.: Residue-driven architecture for Computational Auditory Scene Analysis, *Proc. Fourteenth Intl. Joint Conf. on Artificial Intelligence (IJCAI-95)*, pp.165–172 (1995).
24) Nakatani, T., Goto, M., Itoh, T. and Okuno, H.G.: Multi-agent Based Binaural

Sound Stream Segregation, *Workshop Notes of IJCAI-95 Workshop on Computational Auditory Scene Analysis* (Aug. 1995). Also in Ref. 32).
25) Nakatani, T., Goto, M. and Okuno, H.G.: Localization by Harmonic Structure and Its Application to Harmonic Sound Stream Segregation, *Proc. 1996 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-96)*, Vol.II, pp.653–656, IEEE (1996).
26) Okuno, H.G., Nakatani, T. and Kawabata, T.: Cocktail-party Effect with Computational Auditory Scene Analysis, Symbiosis of Human and Artifact, *Proc. 6th Intl. Conf. on Human Computer Interaction*, Anzai, Y., et al.(Eds.), Vol.2, pp.503–508, Elsevier (1995).
27) 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. 13th Natl. Conf. on Artificial Intelligence (AAAI-96)*, Vol.2, pp.1082–1089 (1996).
28) Okuno, H.G., Nakatani, T. and Kawabata, T.: A New Speech Enhancement: Speech Stream Segregation, *Proc. Int'l Conf. on Spoken Language Processing (ICSLP96)*, ASA, Vol.4, pp.2356–2359, 日本音響学会 (1996).
29) 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).
30) Parsons, T.W.: Separation of Speech from Interfering Speech by Means of Harmonic Selection, *J. Acousl. Soc. Amer.*, Vol.60, No.4, pp.911–918 (1976).
31) Ramalingam, C.S. and Kumaresan, R.: Voiced-speech Analysis based on the Residual Interfering Signal Canceler (RISC) Algorithm, *Proc. 1994 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-94)*, pp.473–476, IEEE (1994).
32) Rosenthal, D. and Okuno, H.G. (Eds.): Computational Auditory Scene Analysis, *Proc. IJCAI-95 Workshop*, Lawrence Erlbaum Associates, 近刊.
33) Schmids, R.O.: Multiple Emitter Location and Signal Parameter Estimation, *IEEE Trans. Antenna and Propagation*, Vol.AP-34, No.3, pp.276–280 (1986).
34) Stadler, R.W. and Rabinowitz, W.M.: On the Potential of Fixed Arrays for Hearing Aids, *J. Amer. Soc. Acoustics*, Vol.94, No.3, pp.1332–1342 (1993).
35) Warren, R.M.: Perceptual Restoration of Missing Speech Sounds, *Science*, No.167, pp.392–393 (1970).
36) Weintraub, M.: A Computational Model for Separating Two Simultaneous Talkers, *Proc. 1986 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-86)*, Vol.1, pp.81–84, IEEE (1986).

(平成8年7月19日受付)
(平成9年1月10日採録)



奥乃 博（正会員）

1950年生．1972年東京大学教養学部基礎科学科卒業．同年，日本電信電話公社（現NTT）入社．1986～1988年スタンフォード大学客員研究員．1992～1993年東京大学工学部客員助教授．現在，NTT基礎研究所勤務．主幹研究員．博士（工学）．推論機構，音環境理解の研究に従事．1990年度人工知能学会論文賞受賞．IJCAI-97広報委員長．本学会英文図書委員．日本ソフトウェア科学会理事．人工知能学会，日本認知科学会，日本ソフトウェア科学会，ACM，AAAI各会員．著編書：『インターネット活用術』（岩波書店，1996），『知的プログラミング』（共著，オーム社，1993），"Computational Auditory Scene Analysis"（共編，LEA，近刊）等．

中谷 智広（正会員）

1967年生．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基礎研究所勤務．研究主任．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．人工知能学会，日本音響学会各会員．

川端 豪

1955年生．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研究科博士課程修了．工学博士．同年，日本電信電話公社（現NTT）入社．1986～1990年，ATR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．主幹研究員．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．