Blind Source Separation による 2 話者同時発話認識

Understanding Two Simultaneous Speeches by Blind Source Separation

奥乃 博
Hiroshi G. Okuno
科学技術振興事業団
ERATO 北野共生システムプロジェクト
Kitano Symbiotic Systems Project
ERATO, JST
okuno@symbio.jst.go.jp

池田 思朗
Shiro Ikeda
科学技術振興事業団 さきがけ研究 21/
理化学研究所 脳科学総合研究センター
"Information and Human Activity"
PRESTO, JST/RIKEN Brain Science Institute
Shiro.Ikeda@brain.riken.go.jp

Abstract

"*Price Shotoku Computer*", or computer audition system that can listen to several things at once, is studies by two separate groups. One group exploits Computational Auditory Scene Analysis (CASA) to segregate speech streams from a mixture of sounds, while the other exploits Blind Source Separation. The CASA group reported the performance of understanding two simultaneous speeches. This paper examines the performance of Blind Source Separation by using the same three benchmarks of 500 mixture of two speeches with/without an interfering sound recorded by a pair of microphones.

Blind Source Separation is performed by online algorithm proposed by N. Murata and S. Ikeda. First, the windowed-Fourier trnasformation is applied to observed signals by shifting Hamming window of 128 points, and on-line ICA (Independent Component Analysis) is applied to the frequency components of the non-symmetric 65 points. Since the output of ICA carries ambiguities in permutation of the frequent components, the permutation of components is determined on the basis of correlation between their envelops.

The performance of speech recognition is a little bit better than that of the CASA group for a mixture of two speeches, but worse for a mixture of two speeches and an interfering sound. This result follows the theoretical capability and limitations of Blind Source Separation. The paper also discusses integration of Blind Source Segregation and CASA from the view point of sound representations.

1 はじめに

同時に複数の音を理解する「聖徳太子コンピュータ」の研究が, 2 つのグループで独立に進められている [1, 10]. 甘利らのグループは, 音源情報を仮定しない Blind Source Separation の問題としてとらえ, ICA (Independent Component Analysis) を適用している. 一方, 奥乃らのグループは, 特定の音に限定せず音を使って環境を理解する音環境理解研究 (Computational Auditory Scene Analysis, CASA) [13] の一環として位置付けている.

従来の音響理解では, 入力音として特定の音を想定して研究が進められてきた. 例えば, 音声認識では, 音声が入力であることが仮定されており, また, 発話者の口元にマイクロフォンが置かれ, 入力が実質的に単一音であることが保証されていた. その結果, このような条件の下での音声認識は研究レベルから商品開発レベルへと移行し, 音声認識装置が市販のパーソナルコンピュータに搭載され始めている. 一方, 研究の方は, より現実的な環境, あるいは, 入力が単一音ではなく, 雑音, 反響, あるいは, 他の音声などが混じる混合音となる場合の「ロバスト」な音声認識へと課題が移っている.

音環境理解研究は, このような「音声」を立脚点とせずに, あらゆる音を同等の立場で扱うことによって, 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」だけでなく, 他の音へも広げるための枠組を研究しようというわけである.

音環境理解研究を促進する目的でマイルストーンとして, 奥乃らは「3 話者の同時発話認識」を第 15 回人工知能国際会議 IJCAI-97 で AI チャレンジとして提案している [12]. この AI チャレンジの最も簡単なアプローチは, 混合音から音声を抽出し, 抽出した音声を音声認識するという「音源分離と音声認識装置の組合せ」であろう. 音源から生成された音は時間的な連続性を持っているので「音響ストリーム」ととらえることができる[2]. 音響ストリームという概念を使用すると, 音源分離は音響ストリームの抽出ということができる.

人間の聴覚の音響ストリーム形成あるいは分離機構につ

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，カクテルパーティ効果の見つかった 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． つまり，CASA は，音の階層的な表現を想定し，その表現を基に処理機構を組み立てていくことを主目的としている．

音響処理のアプローチを検討する上で重要な課題は，信号処理だけで対応していくパターン処理，あるいは，音源情報を記号で表現し，記号処理の立場から信号処理をしていく記号処理という 2 つの両極端のアプローチをどう折衷するかにある． 例えば，音源分離の問題を考えてみると，音源情報を積極的に利用して分離をする CASA の立場と，それとは逆に，音源情報は互いに独立という条件だけを仮定して，音源情報から観測データを与える変換演算を観測データから推測し，その逆変換を求めることを通じて分離を行う Blind Source Separation [1, 7] があろう．

AI チャレンジの予備実験として，奥乃らは音響ストリーム分離による 2 話者同時発話の分離と音声認識を行っている [11]．本稿では，同じ問題を Blind Source Separation を用いて解くことにより，両者のアプローチの長所と短所を明らかにする．それを通じて，3 話者同時発話の分離と音声認識への糸口を検討する．

2 Blind Source Separation による分離

本稿では，blind source separation の手法として，村田と池田が提案した音声信号分離のためのオンラインアルゴリズム [7] を使用した．まず，128 点のハミング窓を 1 タップづつずらしながらかけ，その各 65 点 (対称性を除いて) の周波数成分に対しオンラインの ICA (Independent Component Analysis) をかける．次に，成分の置換については最初の 1 秒のデータを用い，envelope の相関に基づき解いた．

本アルゴリズムを簡単に紹介する

2.1 問題の定式化

今，n 個の成分 (音源) からなる音源信号ベクトル $\boldsymbol{s}(t)$ を (1) 式で，n 本のマイクロフォンで得られた観測信号ベクトル $\boldsymbol{x}(t)$ を (2) 式で表現されるとする：

$$\boldsymbol{s}(t) = (s_1(t), \cdots, s_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{1}$$

$$\boldsymbol{x}(t) = (x_1(t), \cdots, x_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{2}$$

ただし，各成分 (音源) は互いに独立であると仮定する．

さらに，観測データは音源信号の線形変換による混合音であると仮定する．つまり，線形変換オペレータ A によって，観測信号 $\boldsymbol{x}(t)$ は，$\boldsymbol{x}(t) = A\boldsymbol{s}(t)$ で与えられる，とする．

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) = \left(\sum_k a_{ik} * s_k(t) \right), \tag{3}$$

ただし，$a_{ik} * s_k(t) = \sum_{r=0}^{\tau_{max}} a_{ik}(\tau) * s_k(t - \tau)$

ここで，A は，瞬時的混合だけでなく，畳み込みによる混合を許すことに注意．瞬時的混合とは，観測される混合音には音源から発生した信号だけしか含まれないことを意味し，時間遅れなどは一切仮定しない．一方，畳み込みによる混合では，音源からの直接の信号以外に時間遅れの信号を含んでもよく，この結果，反響や残響に加えて，音源の移動などがモデル化できるので，実環境での混合音分離が扱い易くなる．

Blind source separation の目標は，観測信号 $\boldsymbol{x}(t)$ から $\boldsymbol{y}(t) = B\boldsymbol{x}(t)$ であるような相互に独立な $\boldsymbol{y}(t)$ を与える線形変換 B を見つけることである．ただし，線形変換 A や信号 $\boldsymbol{s}(t)$ の確率分布は未知とする．一般に B は A の逆変換とはならず，成分の振幅や組合せなどに曖昧性が残る．

2.2 オンラインアルゴリズムの概要

音声信号のための Blind Source Separation のアルゴリズムを設計するために，音声が 20 〜 30msec 以下の短時間では安定しているが，それ以上から 100msec 程度までは音声の周波数成分が変化し，安定しないという事実を使用する．

2.2.1 周波数の独立成分の抽出

まず，窓によるフーリエ変換を行い，スペクトルを求める．周波数を ω，離散型 FFT の離散点数を N，窓位置を t_s，窓関数を w，窓のシフト間隔を ΔT とすると，スペクログラムは次式で与えられる:

$$\widehat{\boldsymbol{x}}(\omega, t_s) = \textstyle\sum_t e^{-j\omega t} \boldsymbol{x}(t) w(t - t_s), \tag{4}$$

$$\omega = 0, \tfrac{1}{N}2\pi, \cdots, \tfrac{N-1}{N}2\pi, \quad t_s = 0, \Delta T, 2\Delta T, \cdots$$

ただし，w として ハミング窓を使用する．

次に適当な窓幅をを選んで，畳み込みの混合音を瞬時的混合音とみなす．つまり，ある固定周波数 ω に対して，A のフーリエ変換を $\widehat{A}(\omega)$ とすると，(4) 式を次式で近似する．

$$\widehat{\boldsymbol{x}}(\omega, t_s) = \widehat{A}(\omega)\widehat{\boldsymbol{s}}(\omega, t_s), \tag{5}$$

リカレントニューラルネットワークを使用して，各周波数チャネルの混合音信号から独立成分を抽出する． 次式で出力 $\boldsymbol{u}(\omega, t_s)$ を求める．

$$\widehat{\boldsymbol{u}}(\omega, t_s) = \widehat{\boldsymbol{x}}(\omega, t_s) - B(\omega, t_s)\widehat{\boldsymbol{u}}(\omega, t_s) \tag{6}$$

ここで，$B(\omega, t_s)$ の ij 要素は，出力 $\boldsymbol{u}(\omega, t_s)$ の j 成分から入力 $\boldsymbol{x}(\omega, t_s)$ の i 成分へのコネクションを示し，対角成分は 0 とする．

この式は

$$\widehat{\boldsymbol{u}}(\omega, t_s) = (B(\omega, t_s) + I)^{-1}\widehat{\boldsymbol{x}}(\omega, t_s) \tag{7}$$

と変形できるので，次のような学習アルゴリズムで解く．

$$B(\omega, t_s + \Delta T) =$$
$$B(\omega, t_s) - \eta(B(\omega, t_s) + I)(diag(\phi(\boldsymbol{z})\boldsymbol{z}^*) - \phi(\boldsymbol{z})\boldsymbol{z}^*),$$
$$\boldsymbol{z} = \widehat{\boldsymbol{u}}(\omega, t_s) \tag{8}$$

Table 1: Three Benchmark Sets of Mixed Sounds

ベンチマーク (位置)	第 1 話者 (左 30 度)	第 2 話者 (右 30 度)	スピーカ (中央)
Double	女性 1	女性 2	—
Triple	女性 1	女性 2	妨害音 (弱)
Triple'	女性 1	女性 2	妨害音 (強)

ここで, $diag(\cdot)$ は引数が対角要素となり, 他の要素はすべて 0 である対角行列を示し, $*$ は共役行列を示す. また,

$$\phi(z) = \tanh(\mathrm{Re}(z)) + i \cdot \tanh(\mathrm{Im}(z)) \qquad (9)$$

であり, 列ベクトルに対して要素毎に適用される.

このようにして

$$\widehat{\boldsymbol{v}}_\omega(t_s; i) = (B(\omega, t_s) + I)(0, \ldots, \widehat{u}_i(\omega, t_s), \ldots, 0)^T. \quad (10)$$

が得られる. もちろん, Blind Source Separation に固有の曖昧性から $\widehat{\boldsymbol{v}}_\omega(t_s; i)$ が他の周波数に対応する可能性は否定できない.

2.2.2 音声信号の復元

(10) 式で得られた成分から元の信号を復元するために, 同一音源から生成された信号は時間的な構造が共通であるという性質を使う. つまり, 振幅変調 (AM) が類似している周波数成分を同一音源からの信号と見なすわけである. 具体的には, M は正の定数とすると, 次式で表される envelop を計算する.

$$\mathcal{E}\widehat{\boldsymbol{v}}_\omega(t_s; i) = \frac{1}{M} \sum_{t'_s = t_s - M}^{t_s + M} |\widehat{\boldsymbol{v}}_w(t'_s; i)|, \qquad (11)$$

次に, $\mathcal{E}\widehat{\boldsymbol{v}}_\omega(t_s; \sigma_w(i))$ と

$$\mathcal{E}\widehat{\boldsymbol{y}}_\omega(t_s; i) = \mathcal{E} \sum_{\omega'} \widehat{\boldsymbol{v}}_{\omega'}(t_s; \sigma_{\omega'}(i))$$

との相関を最大にするような置換 $\sigma_w(i)$ を求め, 元の音源信号を復元する.

3 Blind Source Separation データに対する音声認識実験

Blind Source Separation と音声ストリーム分離との音声認識結果による比較を行うために, 音声ストリーム分離で使用されたデータ [11] の内, 2 名の女性によるデータを使用した. 表 1に使用した 3 種類の混合音を示す. いずれのベンチマークも 500 組の混合音から成り, その単語の組合せはすべてに共通である. 音響データは, 12kHz, 16 ビットサンプリングの (バイノーラル) ステレオ音である.

混合音は, ダミーヘッドに埋め込まれた 1 対のマイクロフォンで仮想的に録音した. つまり, 頭部音響伝達関数を測定し, それを使用して解析に合成をした. 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 150msec 後に右 30° に位置する話者が発声する. 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 50msec 休止) である. (図 3に話者とスピーカの位置を示す.) **Triple** と **Triple'** との違いは, 後者の方が妨害音の音量が約 2dB 大きいことである.

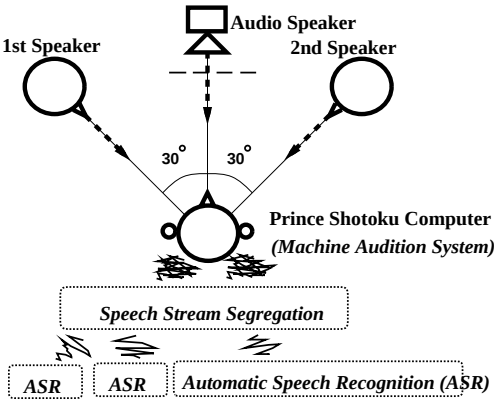

Figure 1: Allocation of Three Sound Sources for Experiments on Understanding Two Simultaneous Speeches.

3.1 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 [6] 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 12kHz で AD 変換後, フレーム周期 3msec 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 コード列に変換する. なお, コードブックのサイズは 256 である.
- 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, 女性の 5 名の話者のデータからそれを 30 度刻みの 4 種類の方向から発話したように解析的に加工したものを使用した. ただし, 混合音に使用した発話は HMM パラメータ学習用の発話データ集合には一切含まれていない.

3.2 分離音の音声認識結果

音声認識結果を上位 10 位までの累積正答率を図 2 に示す. また, 同じデータに対する 音声ストリーム分離システムによる分離音声の認識結果を図 3 に示す (文献[11]より引用).

両図から分かるように, 2 話者だけの混合音 **Double** では, 第 1 話者については, Blind Source Segpregation による分離音声の認識結果の方が, 第 1 位で 16%, 第 10 位までで 7% よいが, 第 2 話者については, 音声ストリーム分離の

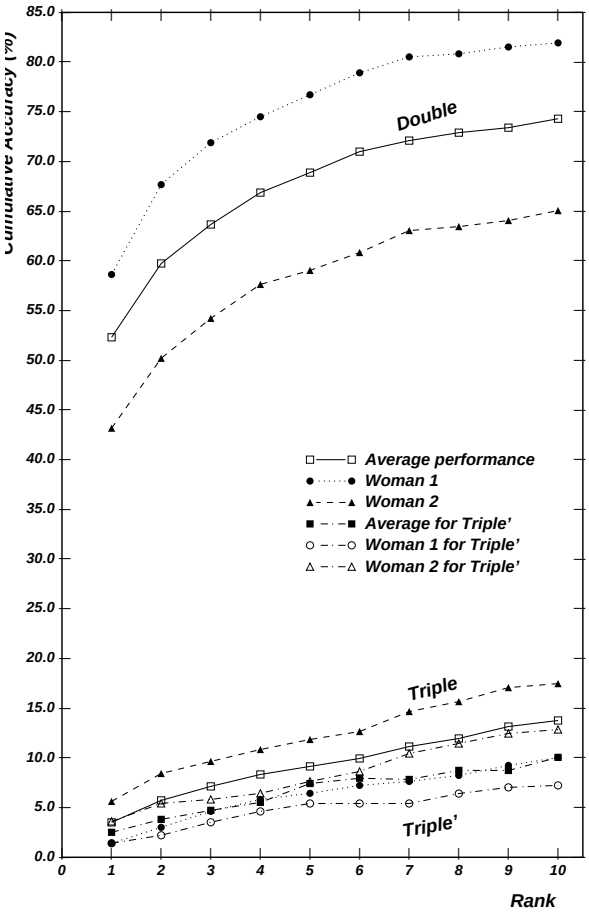

Figure 2: Cumulative Accuracy of Recognition up to the 10th rank by Blind Source Separation

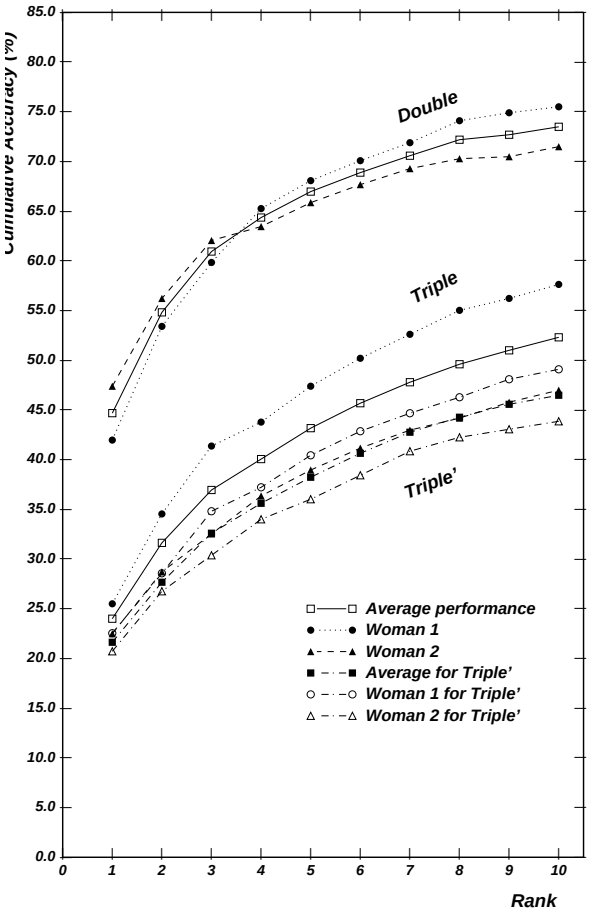

Figure 3: Cumulative Accuracy of Recognition up to the 10th rank by Speech Stream Segregation [11]

方がよい. また, 3 音混合の場合には, Blind Source Segpregation ではほとんど認識できていない.

3.3 認識結果の考察

Blind Source Separation による分離音の結果から得られた知見を以下にまとめる.

1. 2 音のベンチマーク **Double** の第 1 話者の音声の分離は, 音声ストリーム分離よりも認識率, とくに, 第 1 位の認識率が高い.

   これは, 第 1 話者が単独で発話していることが多いので [1], 抽出された独立成分から音声を復元するの精度が高くなったことが原因と考えられる.

2. 同じベンチマークでの第 2 話者の音声の分離は, 音声ストリーム分離よりもすべてのランクの認識率が 3 〜 6% 程度悪い.

   この原因として, 第 2 話者の発話は, ほとんどの場合, 第 1 話者の発話が同時に行われているので, SN 比が低下しており, この結果, 音声の復元精度が悪かったものと思われる.

   本稿で使用したオンラインアルゴリズムは, 文献[7]では文章の発話 (「Good morning」と「こんばんわ」) に対して適用してその有効性を確認している. 本稿でのベンチマークは, 単語の発話のために発話時間が短いものが少なからずあり, その結果 envelop の相関がうまく計算できず, 音声の復元に失敗した可能性がある.

3. 3 音のベンチマーク **Triple**, **Triple'** は, いずれの話者の分離された音声の認識率は, 音声ストリーム分離よりも悪い. 第 1 話者については, 混合音をそのまま認識した時よりも認識率が悪い.

   未確認であるが, 独立成分が得られた時点では, 3 つの音源がうまく分離できていたにもかかわらず, 音声の復元時に, 音源数がマイクロフォンの数以下であるという前提があったために思われる.

   というのは, 音声認識結果を調べてみると, 短い単語の場合, その音韻がすべて含まれているような単語として認識されている場合が少なからずあった. このようなより長い単語の部分列として認識される間違いは, 音声ストリーム分離の場合にも発生した.

[1] 第 1 話者の発話が第 2 話者のそれよりも 150msec だけ早いという混合音の合成を行ったが, 実際の発話では先頭にポーズが入っている場合もあるので, すべての混合音で第 1 話者が単独で発話しているとは限らない.

なお, 音声ストリーム分離では, 音源方向の利用が有効であった.

4. 本稿で使用した音声認識システムは left-to-right 型の探索をするので, 分離音の先頭だけが不正確である場合には認識結果が悪くなる. したがって, より信頼性の高い部分, あるいは, 音声ストリーム分離の場合には調波構造部から隠れマルコフモデルの探索を開始するような枠組が, 認識率を向上するためには不可欠である.

4 議論

Blind Source Separation には, (1) 音源数と同じだけの個数のマイクロフォンが必要であり, また, (2) 音源が相互に独立でないといけない, という 2 つの制約がある. このため, 3 音のベンチマークでは分離がうまくできなかった. 同様に, AI チャレンジの場合にも, 3 話者に対して 2 本のマイクロフォンしか使用できないとすると, Blind Source Separation だけでは音声認識は難しくなる.

また, 楽器の混合音から各楽器の音を分離する場合には, 和音やリズム等の情報を使用することが多い [3, 4, 5] ので, 音源の独立性が成り立たず, Blind Source Separation の適用は難しいと考えられる.

Blind Source Separation を活用するためには, 予想されたこととは言え, 上記の制約を回避するために他の手法との組合せを考えなければならない. ここでもう一度音源分離に音源や伝達路の情報が必要か, という根源的な問いを検討してみよう.

4.1 音源情報の必要性

本稿で比較の対象とした音源分離は, 「音声ストリーム分離」である. これは, 次のようの 3 ステップで構成される [11].

1. 混合音の中に含まれる調波構造と音源方向を手がかりとして, 調波構造ストリーム断片を抽出 [8],
2. 調波構造ストリーム断片を音源方向の連続性でグルーピングし, 音声の調波構造部分を再構築,
3. 調波構造間のギャップを入力音からすべての調波構造を除いた残差で補完し, 音声を「調波構造 + 非調波構造 + 調波構造」で再構築.

一方, 本稿で音声分離に使用した Blind Source Separation のオンラインアルゴリズムは, CASA の言葉で言い換えると, Blind Source Separation で独立成分を求めた後, 音声の時間的な特徴である共通 AM 変調を使用して, 周波数の独立成分の置換を行うということである. Blind Source Separation は, 「blind」という言葉から, 音源分離に音源や伝達路の情報を使用しないというニュアンスがあるが, それでは求まった独立成分のまとめ方に曖昧性が残るので, 実際には, 音源の何らかの特徴を利用して, 周波数の独立成分の置換を行い, 曖昧性を解消するわけである.

つまり, 研究課題は, どのような音源情報が有効なのか, あるいは, どのように音源情報を使用していくのかということであり, 音源情報が必要か不必要かということではない. このような問題を考える上で, 音表現が重要である. すでに中谷と奥乃は, 音響ストリーム分離における音表現の重要性を指摘し, その一案として「音オントロジー」を提案している.

4.2 音表現での課題

従来の音の認識システムは入力音を仮定していたために, 特定の音に対する表現があれば十分であったが, これでは次のような問題が生じる [9].

1. 複数の音ストリーム分離システムの統合 — 複数システムで同じ用語, 概念を使用していても, その定義や精度が異なっていて, 共用できない.
2. 音ストリーム分離システムと音声認識システムとのインターフェース — 使用する音声認識システムによって, 分離された音声ストリームの品質に対する要求が異なる.
3. ボトムアップ処理とトップダウン処理の制御 — 黒板モデルで両者の処理の制御を行う時に, アドホックなトークンを用いると, 拡張性の乏しくなる.

4.3 音オントロジーによる解決策

前節で指摘した表現の課題は, 複数のエキスパートシステムを統合するときに遭遇したものと全く同じである. その時の解法にならい, 音の共通的な表現として, 音オントロジーを考えると問題が扱いやすくなる. 音オントロジーは, クラス間の階層構造と, 各インスタンスの特徴を示す属性の階層構造とから構成される.

中谷らは, 音声と楽音に関するオントロジーを定義し, それを用いて混合音に含まれる音の調波構造の洗練化を行い, 調波構造の時間的な揺れから音声か楽音かを切り分けるシステムを施策し, その有効性を確認している [9].

本稿で使用した Blind Source Separation では, 得られた周波数の独立成分から音声信号を復元する時に共通 AM 変調を使用している. このような手がかりを本稿ではアプリオリに使用したが, 実環境の音理解を進める上では, 中谷らの方法のように自動的に調波構造から音声であることを同定したり, あるいは, 画像認識などから音源が人間であることを知って, 音声の特徴を使用していいく機能が必要となろう.

2 本のマイクロフォンで 3 話者同時発話を認識するのは, 複数の手法を効果的に組み合わせて解く必要がある. 音源情報や発話内容についての上位レベルの情報などを組合せていかなければならない. Blind Source Separation でそのような情報を使用するのが, 周波数の独立成分が求まった後の曖昧性解消のために置換を求めるときだけでよいのか, あるいは, ICA の中でも使っていった方がよいのかは, 今後の研究課題である.

5 おわりに

本稿では, 聖徳太子コンピュータ実現のために, 2 話者同時発話の認識を取り上げ, Blind Source Separation の可能性を検討した. 具体的には, 3 種類の 500 組からなる 2 音と 3 音の混合音を 2 本のマイクロフォンで収録したデータからオンラインアルゴリズムで音声を分離し, 得られた分離音声に対して音声認識をした. 2 音混合音に対しては, 従来得られていた音声ストリーム分離による認識結果と, 第 1 位の認識では優れ, 全体として少しよい認識結果が得られた.

また, 音情報という観点から, CASA の立場からの音声ストリーム分離と Blind Source Separation を組み合わせていくための枠組について検討を行った. 今後, AI チャレンジで提案されている 3 話者同時発話認識に得られた知見を基に取り組んでいく予定である.

最後に, 音源データの準備にご協力いただいた NTT マルチビジネス開発本部の中谷智広氏, 音声認識システムでご協力いただいた NTT 基礎研究所の川端豪氏, および, 御討論いただいた理化学研究所の甘利俊一グループディレクター, 村田昇研究員, 科学技術振興事業団, 北野宏明プロジェクトリーダ, NTT 基礎研究所の石井健一郎部長, 村瀬洋氏, 柏野邦夫氏に感謝する.

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 ('97).

[2] A.S. Bregman: *Auditory Scene Analysis – the Perceptual Organization of Sound*. MIT Press, Boston, 1990.

[3] M. Goto, and Y. Muraoka: 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).

[4] 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, pp.1751–1761 (1996).

[5] 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, pp.1762–1770 (1996).

[6] 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, 情処学会論文誌, Vol.31, No.3, pp.472–480 (1990).

[7] N. Murata, and S. Ikeda: An On-line Algorithm for Blind Source Separation on Speech Signals, *Proc. of 1998 International Symposium on Nonlinear Theory and its Applications (NOLTA'98)*, pp.923-927, September 1998.

[8] 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).

[9] T. Nakatani and H.G. Okuno: Sound Ontology for Computational Auditory Scene Analysis, *Proc. AAAI-98*, pp.1004–1010.

[10] H.G. Okuno, T. Nakatani and T. Kawabata: "Cocktail-Party Effect with Computational Auditory Scene Analysis — Preliminary Report —", in: Y. Anzai and Kato, eds., *Symbiosis of Human and Artifact — Proc. HCI Int'l '95*, Yokohama, July 1995, Elsevier, Vol.2, pp.503–508.

[11] 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. 情処学会論文誌, Vol.38, No. 3, pp.510-523 (1997).

[12] H.G. Okuno and T. Nakatani: Understanding Three Simultaneous Speeches, *Proc. IJCAI-97*, pp.30–35.

[13] D. Rosenthal and H.G. Okuno (eds.): *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, NJ, 1998.