人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702

ＡＩチャレンジ研究会 (第２６回)

Proceedings of the 26th Meeting of Special Interest Group on AI Challenges

CONTENTS

◇ 【基調講演】話し言葉音声の認識精度向上のために 1
古井 貞熙, 中村 匡伸, ポール ディクソン, 大西 翼, 岩野 公司 (東京工業大学)

◇ 聴覚系を模倣した音源定位システムによる複数音源の 3 次元定位 8
中島 弘道 (理化学研究所), 河本 満 (産業技術総合研究所), 伊藤 雅紀 (名古屋大学), 向井 利春 (理化学研究所)

◇ メインローブモデルを用いた複数音源定位手法の動的環境下での性能評価 15
佐々木 洋子 (東京理科大学 / 産業技術総合研究所), 加賀美 聡 (産業技術総合研究所 / 東京理科大学), 溝口 博 (東京理科大学 / 産業技術総合研究所)

◇ Tracking A Varying Number of Speakers Using Particle Filtering 21
Angela Quinlan, Mitsuru Kawamoto, Futoshi Asano (AIST)

◇ 適応ステップサイズ BSS による音源分離のロボット聴覚への適用 29
中島 弘史, 中臺 一博, 長谷川 雄二, 辻野 広司 (HRI-JP)

◇ Semi-blind Cancellation of Robot Internal Noise for Hands-Free Speech Recognition
Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano (NAIST) 35

◇ Real-time Dereverberation Using Late Components of Impulse Response for Hands-Free Speech Recognition 40
Randy Gomez, Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano (NAIST)

◇ 発話音声に関わる頭部動作の分析及びアンドロイドロボットの頭部制御 46
石井カルロス寿憲 (ATR), 石黒 浩 (大阪大学 /ATR), 萩田 紀博 (ATR)

◇ 単旋律楽曲における機械学習を用いた楽器音の音源同定 52
井原 瑞希, 前田 新一 (奈良先端科学技術大学院大学), 石井 信 (京都大学)

◇ 口じゃんけん判定ロボットの開発 ～ロボット聴覚システムの応用に向けて～ 59
中臺 一博 (HRI-JP/ 東京工業大学), 山本 俊一, 奥乃 博 (京都大学), 中島 弘史, 長谷川 雄二, 辻野 広司 (HRI-JP)

◇ 小型音声対話モジュールによる耐雑音音声認識 65
佐藤 幹, 岩沢 透, 杉山 昭彦 (NEC)

日　時　2007 年 11 月 22 日　　場　所　東京, 東京工業大学
Tokyo Institute of Technology, Tokyo, Nov. 22, 2007

社団法人 人工知能学会
Japanese Society for Artificial Intelligence

共催 京都大学グローバル COE プログラム
「知識循環社会のための情報学教育研究拠点」
Kyoto University Global COE Program:
Informatics Education and Research Center for Knowledge-Circulating Society

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-1

話し言葉音声の認識精度向上のために

Toward increasing spontaneous speech recognition accuracy

○古井 貞熙, 中村 匡伸, ポール ディクソン, 大西 翼, 岩野 公司
東京工業大学大学院情報理工学研究科計算工学専攻
* Sadaoki Furui, Masanobu Nakamura, Paul Dixon, Tasuku Oonishi, Koji Iwano
Tokyo Institute of Technology, Department of Computer Science
{furui, masa, dixonp, oonishi, iwano}@furui.cs.titech.ac.jp

Abstract – This paper reports progress on two major fronts in our research on spontaneous speech recognition. First, by comparing various features of spontaneous speech and read speech, we have found that spectral space reduction and linguistic complexity are two major sources of decreases in spontaneous speech recognition accuracy. Second, by introducing on-the-fly composition, disk-based search and acoustic likelihood calculation using GPU hardware, we have constructed a high-flexibility and high-performance WFST-based decoder.

１．はじめに

大語彙の連続音声認識でも、テキストを読み上げた音声であれば、かなり高い精度で音声認識できるようになったが、普通の話し言葉、すなわち考えながら（あるいは考えるよりも前に）話している自発性の高い音声では、認識性能が大幅に下がってしまう。これは、読み上げ音声と話し言葉音声は、音響的にも言語的にも大きく違うことを示している。音声の現象を調べ、音声のモデルを作り、音声認識システムを構築するためには、大規模の音声コーパスを作成することが不可欠である。ところが、大規模の話し言葉音声コーパスを作成するには、実際の話し言葉音声を大量に録音し、人手で書き起こし、さらに形態素解析を行って形態素（単語）に区切ったり、実際の発音を与えたりしなければならないため、多大な人手とコストがかかる。このため、話し言葉音声の研究は、世界的にも、10 年くらい前まではほとんど行われず、そのために話し言葉と読み上げ音声の違いも強く意識されず、タスクを限定した対話音声システムを除くと、大語彙連続音声認識では、もっぱら読み上げ音声を用いた研究が行われていた。対話音声システムでも、ユーザは通常、コンピュータの能力を意識した発声をするため、その音声は、ある意味できちんとした読み上げ音声に近く、自然な話し言葉音声とは異なる。

しかし、ニュース音声を用いた研究[1]が進展するに伴って、アナウンサーがテキストを読み上げている音声から、現場からの中継や、スポーツの実況中継などの音声を認識するようになって、自然な話し言葉音声の認識の難しさが強く意識されるようになった。米国では、DARPA によって、Switchboard と呼ばれる、電話での会話音声が大量に録音され、音声コーパスが作成されるようになったが[2]、わが国では、その動きが遅れていた。そこで、国語研、通総研（当時）と東工大で、1998 年秋に、科学技術振興調整費の開放的融合研究推進制度によるプロジェクトに、「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」事業を提案し、多数の方々のご支援を得て、1999 年から 5 年間、プロジェクトが推進された[3]。このプロジェクトには、上記３機関の研究者だけでなく、京都大学など種々の大学や研究機関からの研究者が参加し、極めて密度の高い研究開発が行われた。そのプロジェクトでは、「日本語話し言葉コーパス（CSJ: corpus of Spontaneous Japanese）」を構築したが、その中では、共通の話者による話し言葉音声と読み上げ音声の違いが、解析できるようになっている[4]。

最近、中国語に関しても、多数話者による話し言葉音声と読み上げ音声を収録したコーパスが作られている。世界的には、会議、講演、講義、国会の討論など、多様な話し言葉を対象とした研究が活発に行われつつある[5]。以下に、日本語と中国語の話し言葉コーパスを用いた、話し言葉音声と読み上げ音声の違

いに関する分析結果について述べる。

話し言葉音声の認識のためには、種々の知識を容易に導入でき、スケーラビリティがよく、性能のよいデコーダを開発することが必要である。我々は、このような観点から、WFST（重みつき有限状態トランスデューサ, Weighted Finite State Transducer）に着目し、デコーダを開発してきた[6]。世界的にも、WFSTによるデコーダが主流になりつつある。我々のデコーダの開発現状についても、以下に述べる。

２．日本語話し言葉コーパスを用いた話し言葉の音響的・言語的特徴の分析

2.1 日本語話し言葉コーパス

上記の「日本語話し言葉コーパス（CSJ）」は、タスクを限定しない独話、特に講演（学会講演と模擬講演）を主たる対象として、のべ約 650 時間、単語（形態素）にして約 700 万語規模を有し、世界最大かつ高品質の話し言葉コーパスである[4]。独話、特に講演を対象としたのは、準備された講演音声は、音響的および言語的に、読み上げ音声と対話音声の中間に位置していると考えられ、話し言葉の特性を有しながら、対話音声よりもモデル化しやすいと考えられるためと、講演音声をコンテンツ化したり、自動的に字幕をつけたいというニーズは、極めて大きいためである。全体のコーパスの大きさは、音声認識のための統計的言語モデルを構築するために、最低限必要と思われるデータ量に基づいて決めた。研究の幅を広げるため、コーパスの一部に、インタビューなども含めた。

コーパス全体について、セグメンテーション、書き起こし（正書法による基本形と発音形）、形態素解析などを行ったが、全体の約 8%（50 万語）の「コア」については、人手による形態素解析結果、構文構造、要約の他、韻律情報などパラ言語情報まで含めたコーパスとした。コアの形態素解析結果を用いて、コーパス全体を自動的に形態素解析するためのツール（解析プログラム）の構築も行った。

2.2 音響的特徴の分析

CSJ には、同じ話者が学会講演（AP）、模擬講演（EP）、対話音声（学会講演の内容に関するインタビューと自由対話）（D）、および読み上げ音声（自分の学会講演の書き起こしや、対話形式エッセーを読み上げた音声）（R）がデータベース化されているため、この中の男性・女性話者各 5 名による音声を用いて、話し言葉と読み上げ音声の音響的特徴の違いに関する分析を行った[7]。異なる発声スタイルで、話者が共通しているため、話者による声質の違いの影響が除去できる。

話し言葉音声においては、読み上げ音声に比べて、いわゆる発声のなまけの効果で、スペクトル空間が縮小する傾向がある。これを確認するため、各話者の各音素の 3 状態 1 混合モノフォン HMM を作成し、音素ごとにその中心状態のケプストラムの平均ベクトルの、全音素の分布の中心（平均値）からの距離の縮小率を求めた。発話単位（400ms 以上の無音区間で区切られた約 10 秒長程度の区間）ごとに CMS 処理を行っている。実験の結果、ほとんどすべての音素において、読み上げ音声に比べて、話し言葉音声の平均ベクトルの中心からの距離が減少する傾向が見られ、対話音声において特に顕著になることが確認された。図 1 に、話し言葉音声の場合の、全音素の分布の中心からの各音素までの距離を、読み上げ音声の場合の値で割った縮小率を示す。発話スタイルごとに、母音・子音別に、音素と話者に関して平均して示した。

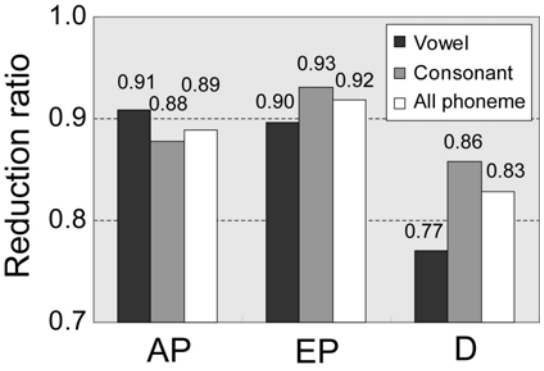


Fig. 1 – Mean reduction ratio of vowels, consonants and all phonemes, respectively, for each speaking style.

音声認識性能に直接関係するのは、異なる音素間の距離と考えられるので、次に、各音素モデル間のマハラノビス距離の分布を調べ

た。AP、EP、D に加えて、新聞読み上げ音声（R）について分析した。図２に、各音素間のマハラノビス距離の相対累積度数を、発話スタイルごとに示す。朗読（R)、自然独話（AP, EP)、対話（D）の順に、つまり自発性が高くなるに従い、音素間のマハラノビス距離が小さくなる傾向がある。

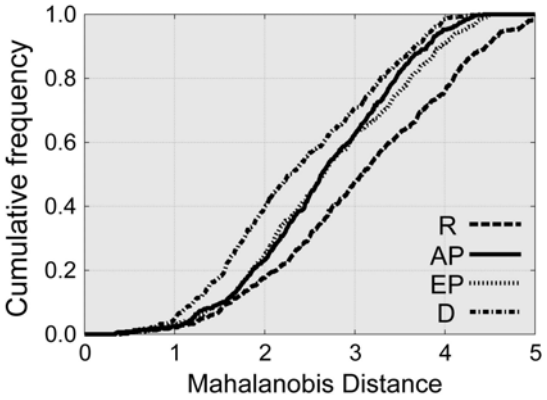


Fig. 2 – Distribution of Mahalanobis distances between phonemes for each speaking style.

次に、音素間のマハラノビス距離の平均値と音素認識正解精度の関係を調べた。音素モデルは、CSJ の学会講演と模擬講演の男女計 100 名の音声データを用いて作成し、各発話スタイルの音声を認識した。結果は図３に示す通りで、音素間のマハラノビス距離の平均値と音素正解精度の間には、高い相関があることがわかった。このことは、音素間のマハラノビス距離を調べれば、音素正解精度が予測できることを示している。

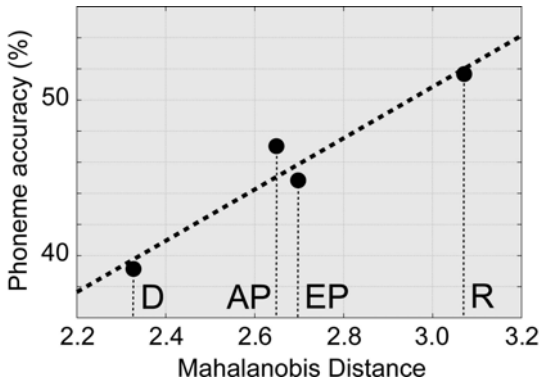


Fig. 3 – Relationship between Mahalanobis phoneme distance and phoneme recognition accuracy.

2.3 言語的特徴の分析

CSJ 中の種々の発話スタイルのコーパスに加えて、書き言葉およびそれに近いタスクとして、毎日新聞記事（NP）および放送のニュース解説（NC）のコーパスを用いて、それらの言語モデルの比較を行った[8]。各コーパスについて、トライグラムを作成し、コーパス相互間のテストセットパープレキシティと未知語率を調べた。コーパスによって語彙数や未知語率が異なるため、厳密な比較はできないが、テストセットと同じコーパスから言語モデルを作成した場合でも、書き言葉である新聞記事に比べて、講演、対話などの話し言葉では、パープレキシティが約 5 倍に大きくなることがわかった。各コーパス間のパープレキシティ行列を、距離行列化し、多次元尺度構成法により、言語モデル間の距離の可視化を行った結果を、図４に示す。第１軸（横軸）が、ほぼ自発性の度合いに対応している。

音素モデルを共通にし、これらの言語モデルを用いて各タスクの音声認識実験を行った結果、パープレキシティと単語正解精度の間に、高い相関があることが確認されている。

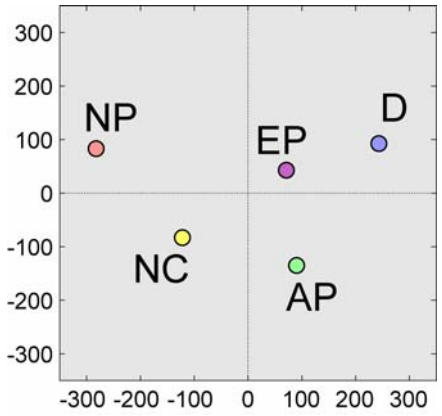


Fig. 4 – Relationship between the language models derived from the perplexity matrix.

３．中国語音声を用いた実験

3.1 中国語音声データベースと実験方法

話し言葉音声としては、電話での対話を想定して、通常のマイクロホンで収録した対話音声を用い、読み上げ音声としては、同一話者が発声したニュース記事読み上げ音声、および本や新聞を朗読した音声を用いた[9]。分

析対象の音響単位としては、トーン（四声および軽声）を区別した全 184 種類の中国語 Initial-Final モデル（以下、簡単のため、音素と呼ぶ）を用いた。

上記の日本語の音素の分析の場合は、簡単のために各音素を、単一ガウス分布を用いた 3 状態のモノフォン HMM でモデル化し、その中心の状態を用いたが、中国語音素の分析では、さらに精密化を図るため、話し言葉と読み上げ音声それぞれの各音素を、混合ガウス分布を用いたモノフォン HMM でモデル化した。日本語の場合と同様に、発話単位（無音区間で区切られた約 10 秒長程度の区間）ごとに CMS 処理を行っている。HMM の学習には、460 名（男女各 230 名）の話者の音声を用い、認識実験の評価話者には、30 名（男性 16 名，女性 14 名）の話者を用いた。

単一ガウス分布の場合は、マハラノビス距離を用いることによって、分布間の距離を測ることができるが、混合ガウス分布の場合はそれができないので、Kulback-Leibler 擬距離（KLD）を用いた。KLD を定義どおりに計算することは困難なので、ここでは unscented transform に基づく、次のような近似式を用いた[10]。

$$D(s\|\tilde{s}) \approx \frac{1}{2N}\sum_{m=1}^{M}\omega_m \sum_{k=1}^{2N}\log\frac{p(o_{m,k}|\,s)}{p(o_{m,k}|\,\tilde{s}\,)}$$

ただし、N は音響特徴量（ケプストラム、対数パワー、およびそれらの動的特徴量）の次元数（N=39）、M は混合数、ω_m は GMM における m番目のガウス分布の混合重み、$o_{m,k}$（$1 \leqq k \leqq 2N$）は、m 番目のガウス分布の k 番目の sigma point（標準偏差の位置）である。

3.2 全音素間の KLD の分布

話し言葉音声と読み上げ音声において、音素モデルの混合数を、1, 2, 4, 8, 16, 32, 64 の 7 段階に変化させた場合の、全音素相互間の KLD の変化を比較した。各音素に対して、10 個の近傍音素を選び、その KLD を求めた。音素間の KLD の値の相対累積度数を、図 5 に示す。音素モデルの混合数が増えるにしたがって、モデルがより正確になってくるので、全音素間の KLD が大きくなるのは当然であるが、読み上げ音声に比べて、話し言葉音声の場合に、混合数の増加による KLD の増加率が小さいことがわかる。

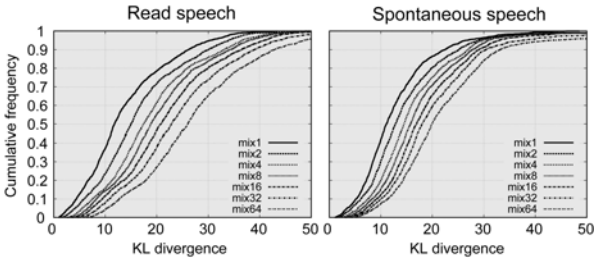


Fig. 5 – Comparison of distributions of KLD distances between phonemes for read and spontaneous speech as a function of the number of mixtures.

3.3 全音素間の KLD の中位値と、音素認識精度の関係

HMM の各状態の混合数を 1 から 64 まで増やしていったときの、音素（Initial-Final モデル）ベースのネットワークを用いた音素認識実験を行った。挿入ペナルティは、発話スタイルごとに最適化した。混合数が 1 の場合に対する音素認識誤りの削減率と、音素間の KLD の中位値との関係を、図 6 に示す。この結果から、KLD と音素認識誤りの削減率の間には、強い相関があることがわかる。ただし、同じ KLD の値でも、読み上げ音声と話し言葉音声では、音素正解精度の絶対値には顕著な違いがある。この原因を明らかにすることは、話し言葉音声の認識精度を向上させるための示唆を与える可能性がある。

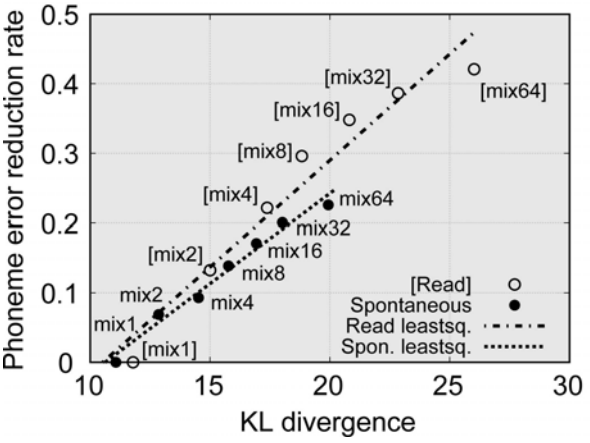


Fig. 6 – Relationship between median of KLD phoneme distances and phoneme recognition error reduction rate.

４．話し言葉音声の分析に関する考察

話し言葉音声の認識はどうして難しいのか？　その疑問への回答と、認識精度の向上を目指して、日本語と中国語の音声を用いて分析を行ってきた。これまでに、言語に依存せずスペクトル空間が縮小することなど、いくつかの定量的な事実が明らかになってきたが、まだ不明なことが多い。これまでの分析結果は、残念ながらまだ認識性能の向上には結びついていない。話し言葉音声の認識は、音声認識応用の展開において、最も重要なテーマの一つであるので、データベースの整備を含め、言語横断的な協力によって、分析を強力に進めていく必要がある。

５．話し言葉音声認識のためのデコーダ

5.1 WFST デコーダ

我々は、経済産業省「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発・音声認識基盤技術」プロジェクトの一環として、WFST (Weighted Finite State Transducer)を利用したデコーダを構築している。WFST に基づくデコーダは、高速で高性能、かつフレキシビリティの高い方法として注目されており、これまでに種々のデコーダが実際に構築されて、有効性が確認されている[11][12][13]。

WFST に基づく音声認識では、探索に先だち、音響モデルや言語モデル、単語発音辞書などの構成要素を合成してまとめあげ、一つの巨大な WFST 形式のネットワークを構築する。認識はこのネットワークを探索することで進められる。従来の認識手法に比べて、探索ネットワークを保持するためのメモリ量を多く必要とする反面、モデルの融合が事前に行われることから、探索時に動的なモデル融合を行う必要がなく、高速なデコーディングが可能となる。また、様々な形式のモデルや辞書を扱う必要が生じても、最終的に WFST の形式でネットワークに変換できれば、モデルに応じてデコーダ自体を変更するといった必要がないため、柔軟なデコーダが実現できる。我々は、スケーラビリティの向上を狙い、省メモリ化や高速化などのための、様々な機能の実装と検討を行っている。

5.2 WFST デコーダの構成

WFST は、与えられた入力記号列に対して状態遷移を繰り返し、それに対応した出力記号列と重みを出力する有限状態オートマトンの一種である。WFST を利用した音声認識では、まず音素モデルや言語モデル、単語発音辞書などをそれぞれ個別に WFST の形式に変換する。次に、基本演算の一つである合成(composition)演算を施して WFST 同士をまとめ、複数のモデルを組み込んだ一つの WFST を生成する。合成に際して、最小化(minimization)や決定化(determinization)などの演算を施すことにより、すべてのモデルを考慮したネットワーク全体に対して最適化が行われ、効率的な探索ネットワークを生成することができる。これにより、高速で高精度な音声認識を実現することができる。

入力音声は、フロントエンドを通して特徴ベクトルに変換され、デコーディングに利用される。本デコーダでは、Sphinx[14]で利用されている、多段フィルタによるフロントエンド設計を採用している。例えば、入力音声は、「窓掛け」や「FFT」などの個別の処理フィルタに順次通されることで、MFCC などの特徴ベクトルへ変換される。ユーザは、利用目的に応じて設計した処理フィルタを、容易に取り入れることができる。例えば、動画から画像特徴量への変換フィルタを作成することで、デコーダを動画認識やマルチモーダル音声認識に利用することが容易にできる。

探索は、フレーム同期型の１パス探索であり、第１位仮説からの尤度差と保持仮説数の上限値を用いた枝刈りを行っている。認識結果は 1-best や単語ラティス形式で出力する。認識結果の出力方式には、探索の終了時に一度に出力を行う「バッチ型」と、探索途中で確定した単語列を順次出力する「逐次型」を選択することができる。

音声への字幕付与などのアプリケーションでは、発話から単語列確定までの遅れ時間の短縮が非常に重要になる。逐次デコーディングにより、全ての発話が終了した後に最尤となる単語列を確定するのではなく、発話途中で早期に単語列を確定することができる。具体的手法として、保持している複数の仮説の

単語履歴に対し、履歴中の先頭からの部分単語列が共通になった段階で、その単語列を出力する手法[15]、過去の仮説単語履歴と比較する手法[16]、推定された無音区間毎に単語列を出力する手法[17]などがある。本デコーダでは[15]の手法を用いている。

本デコーダは音声検索、リスコアリング処理を利用したアプリケーションなどとの親和性を高めるため、ラティス形式での出力を行うことを可能にしている。ラティスは WFST 形式であり、仮説展開の際に同時に生成される。ラティスを構成する要素単位は、事前に合成されたネットワークに依存しており、HMM の状態を構成要素とした単語ラティスや、(文脈依存) 音素を構成要素とした単語ラティスが得られる。

5.3 省メモリ化

WFST による音声認識では、肥大化した探索ネットワークの読み込みに伴う、メモリ使用量の増大がしばしば問題となる。その対策として、1) 事前のネットワーク構築の段階ですべての WFST を合成せず、一部の WFST については、探索中に動的に合成するようにして、読み込む探索ネットワークの肥大化を防ぐ手法 (on-the-fly 合成[18][19][20])、2) 認識時に探索ネットワーク全体をメモリ上に読み込むのではなく、ディスク上に展開しておき、必要分だけを随時メモリ領域に読み込んで利用する方法 (disk-based search[21]) の２つについて検討を行った。On-the-fly 合成に関しては、事前の探索ネットワークの合成に利用する WFST を換えて、様々な方式について検討した。

5.4 高速化

混合ガウス分布を音響モデルとして利用する音声認識では、ガウス分布の混合数の増加に伴い、音響尤度計算に多くの時間を要する。このため音響尤度計算を効率的に行うことは、高速な音声認識において非常に重要である。

近年、グラフィックスカードに搭載された GPU(Graphics Processing Unit)の浮動小数点速度が、 CPU のそれと比較して飛躍的に向上しており、将来的には、汎用的な計算プロセッサとして GPU が広く利用されることが予想される。我々は高速に音響尤度を計算する一つのアプローチとして、GPU を利用した音響尤度計算手法を提案し、その実装を行った。このアプローチでは、高速な演算ユニットを利用して正確に音響尤度計算を行うため、近似的な計算により計算量を削減するアプローチと違い、認識率の劣化なしに高速に音響尤度を計算することができる。

5.5 評価結果

On-the-fly 合成において、オーバヘッドの増大を防ぐため、HMM としてスキップなしの left-to-right 型を扱うこととした。その結果、最大 60%以上のメモリ消費量の削減が実現できた。また disk-based search を行うことで、最大で 60%以上のメモリ消費量の削減を確認することができた。さらに、それらを組み合わせることで、全ての WFST を事前に合成した場合と比べて、80%以上のメモリ消費量の削減を確認することができた。

また GPU の利用により、最大で 25%程度の認識時間が削減できることが確認できた。

６．むすび

話し言葉音声の認識性能の向上を目指して進めている研究の中から、二つの最近の研究内容を紹介した。一つ目は、話し言葉音声と読み上げ音声の違いに関する分析で、話し言葉音声では、読み上げ音声に比べて、顕著にスペクトル（ケプストラム）空間が縮小するとともに、言語モデルの複雑さが増して、それらが音声認識誤りの増大の原因になっていることを明らかにした。

二つ目は、話し言葉音声認識のための、フレキシビリティの高い、高性能のデコーダの開発で、WFST をベースとして、種々の機能を持たせるとともに、省メモリ化、高速化を実現している。

人が話し言葉音声を理解する際には、文脈などを含む、極めて多様な知識を組み合わせて用いている。フレキシビリティの高いデコーダをベースに、如何にこれらの知識を組み込んだ音声認識の枠組みを構築していくかが、今後の進展の鍵になるであろう。

謝辞

本研究は、21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」および、経済産業省「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発・音声認識基盤技術」プロジェクトの支援を得て行われている。

参考文献

[1] J.-L. Gauvain et al., “Structuring broadcast audio for information access,” EURASIP Journ. on Applied Signal Process.,” vol.2003, no.2, pp. 140-150, 2003.

[2] J. J. Godfrey et al., “Switchboard: Telephone speech corpus for research and development,” Proc. ICASSP, pp. I-517-520, 1992.

[3] S. Furui, “Recent progress in corpus-based spontaneous speech recognition,” IEICE Trans. Inf. & Syst., vol.E88-D, no.1, pp. 1-11, 2005.

[4] 前川, “『日本語話し言葉コーパス』公開版の仕様,” 第３回話し言葉の科学と工学ワークショップ講演予稿集, pp. 7-14, 2004.

[5] 古井, “音声認識の動向[1]-話し言葉音声認識,” 電子情報通信学会誌, vol.89, no.8, pp. 746-751, 2006.

[6] 大西他, “WFST 音声認識デコーダの開発とその性能評価,” 情報処理学会研究報告, vol.2007, no.68, 2007.

[7] M. Nakamura et al., “Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances,” Proc. Interspeech, pp. 3381-3384, 2005.

[8] 中村他, “読み上げ音声に対する話し言葉音声の言語的特徴の分析,” 日本音響学会秋季研究発表会講演論文集, 3-6-10, 2005.

[9] 中村他, “KLD を用いた中国語における読み上げ音声と話し言葉音声の違いの分析,” 日本音響学会秋季研究発表会講演論文集, 3-6-10, 2007.

[10] J. Du et al., “Minimum divergence based discriminative training,” Proc. Interspeech, pp. 2410-2413, 2006.

[11] M. Mohri et al., “Weighted finite-state transducers in speech recognition, ” Computer Speech and Language, vol.16, no.1, pp.69-88, 2002.

[12] D. Moore et al., “Juicer: A weighted finite-state transducer speech decoder,” Proc. MLMI, 2006.

[13] T. Hori, “NTT Speech recognizer with OutLook On the Next generation; SOLON,” Proc. Communication Scene Analysis, 2004.

[14] P. Lamere et al., “Design of the CMU Sphinx-4 decoder,” Proc. ICSLP, pp.1181-1184, 2003.

[15] P. F. Brown et al., “Partial traceback and dynamic programming,” Proc. ICASSP, pp.1629-1632, 1982.

[16] 今井他, “最ゆう単語列逐次比較による音声認識結果の早期確定,” 電子情報通信学会論文誌 D-II, vol.J84-D-II, no.9, pp.1942-1949, 2001.

[17] 河原他, “話し言葉音声認識のための言語モデルとデコーダの改善,” 情報処理学会研究報告, vol.2001, no.55, pp.15-22, 2001.

[18] H. J. G. A. Dolfing et al., “Incremental language models for speech recognition using finite-state transducers,” Proc. ASRU, 2001.

[19] T. Hori et al., “Generalized fast on-the-fly composition algorithm for WFST-based speech recognition,” Proc. Interspeech, pp. 847-850, 2005.

[20] D. A. Caseiro et al., “A specialized on-the-fly algorithm for lexicon and language model composition,” IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.4, pp. 1281-1291, 2006.

[21] D. Willett et al., “Time and memory efficient Viterbi decoding for LVCSR using a precompiled search network,” Proc. Eurospeech, pp.847-850, 2001.

聴覚系を模倣した音源定位システムによる複数音源の3次元定位

A three-dimensional localization of multiple sound sources using the sound source localization system imitating a auditory system

○中島 弘道 (理化学研究所)
河本 満 (産業技術総合研究所)
伊藤 雅紀 (名古屋大学大学院)
向井 利春 (理化学研究所)

***Hiromichi NAKASHIMA(RIKEN), Mitsuru KAWAMOTO(AIST), Masanori ITO(Nagoya University) and Toshiharu MUKAI(RIKEN)**

nakas@bmc.riken.jp, m.kawamoto@aist.go.jp, ito-m@nagoya-u.jp, tosh@bmc.riken.jp

Abstract

Human beings and living thing have the capability of identifying the direction of two or more sound by a certain amount of correctness with only two ears. However it is difficult to achieve this capability by robots. Almost robots which have been proposed until now, have three or more microphones in order to localize sound sources. In this paper, we propose a technique of estimating two kind of directions, that is, vertical and horizontal directions, using a robot head consisted of two microphones in which the microphones have reflectors working like the pinna.

1 まえがき

人とロボットのコミュニケーションにおいて，人間からの呼びかけに反応して話しかけられた方向を向いて対応するような音源定位機能は重要な役割をはたす．ロボットが人と会話する場合，ロボットは話者の方を向いて会話をする方が望ましい．そのためには，話者の方向を検出する必要がある．また，実際に次世代ロボットが利用される生活環境においては，マイクで検出される音は複数の音源からの音が混ざり合った混合音となることが頻繁に起こると考えられる．そのような環境においてロボットを使用する場合には，ロボットはこのような混合音を聞き分け音源方向を見つける必要がある．人間や多くの生物は，2つの耳だけで複数の音の方向をある程度の正確さで識別する能力を持っているが，ロボットでこれを実現することは未だ困難であり，多くの音源定位機能を備えたロボットでは，3つ以上のマイクや聴覚以外のセンサ情報を用いることによって，様々な方向からの音源の定位に対応している．

安藤らによって開発されたSmartHead[1]は4本のマイクと2つのカメラを用いて低レベルな情報を抽出し，それらを統合することによって複数音源の定位を可能としている．また，中臺らは，アクティブな動作と視聴覚を統合することにより，2本のマイクで水平面状にある複数音源を定位・分離し，さらにこの分離フィルタをフロントエンド処理として利用した三話者同時発話の認識を実現している[2]．さらに東芝の聞き分けロボットは6つのマイクで全周囲から音声を取り込み，人から話しかけられた方向と内容を認識出来る．これにより，複数の人による全方向からの呼び掛けに対して，それぞれに回答することが出来る．

人間や多くの生物は複数の音源を2つの耳のみで分離し，その音源方向を見つけることが可能である．本研究では，このような処理を工学的に実現する手法を提案する．提案手法を用いることで，複数音源からの音を分離し，それぞれの音の上下左右の方向を推定することが可能となる．

2 提案する複数音源の定位法

通常，人の話す音声にはたくさんの無音部分が存在する．よって，複数の人間が同時に話していたとしても短い時間内であれば話をしている人間が一人だけとなる区間がいくらか存在すると考えられる．そこで，音声の中で1音源とみなせるような，一方の音が出ていない又は音が小さい区間を見つける．そこで，その区間のみを注目すれば1音源での音源定位の処理方法を用いる事が可能であると考えられる．具体的な処理の流れを以下に示す．

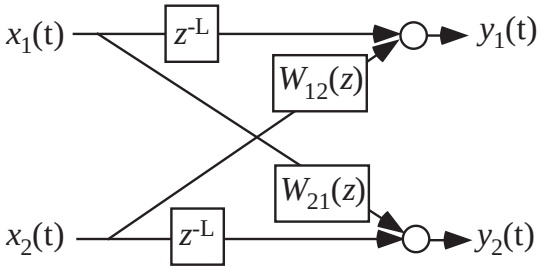


Figure 1: separation filter

1. 2 つのマイクで取得した音声の分離を行う．
2. 分離後の音声信号上で無音区間を検出する．
3. 音源が 1 つのみとなる区間を検出する．
4. 分離前の音声信号で音源が 1 つのみの部分を抽出する．
5. 抽出した部分の音声信号を用いて，音源定位を行う．

本提案手法は複数音源を対称とするものであるが，今回分離に使用した手法が 2 音源に対する手法であるため，これ以降は 2 音源の場合についてそれぞれの処理の詳細を説明する．

2.1 音源分離

音源分離には，ブラインド分離法を用いる[3]．つまり，分離フィルタ $W_{ij}(z)$=$\sum_{k=0}^{M} w_{ij}(k)z^{-k}$ (図 1) の係数 $w_{ij}(k)$ を以下の式で更新することによって，分離を実現する．

$$\Delta w_{ij}(k) \propto -\frac{y_i(t-L)y_j(t-k)}{\phi_i(t)}, \quad i,j = 1,2(i \neq j) \qquad (1)$$

ただし，係数 L は，0< L < M を満足する正の定数である．$y_i(t)$ $(i = 1, 2)$ は，分離フィルタの出力信号である．$\phi_i(t)$ は，$y_i(t-L)^2$ を時間平均で求めた値である．さらに，分離の精度を上げるために，分離した信号 $y_i(t)$ に対して，バイナリマスク[4]をかける．

2.2 単一音源部分の抽出法

音源の 2 つの信号をそれぞれ $s_1(t), s_2(t)$，マイクによる観測信号を $x_1(t), x_2(t)$，分離した結果得られた信号を $y_1(t), y_2(t)$ とする．ここで図 2 に示すような分離信号 $y_1(t), y_2(t)$ の 2 つの音声信号それぞれにおいて，一定時間 (T_m) 以上連続して閾値 (I_{m1}, I_{m2}) 以下となる部分を無音区間とする．図中で網掛けとなっている区間が無音区間と判断された領域である．ここで，無音部分とそれ以外の部分を表す関数 $b_1(t), b_2(t)$ は以下のように定義される．

$$b_i(t) = \begin{cases} 0, & \text{無音区間} \\ 1, & \text{それ以外} \end{cases} \quad i = 1,2 \qquad (2)$$

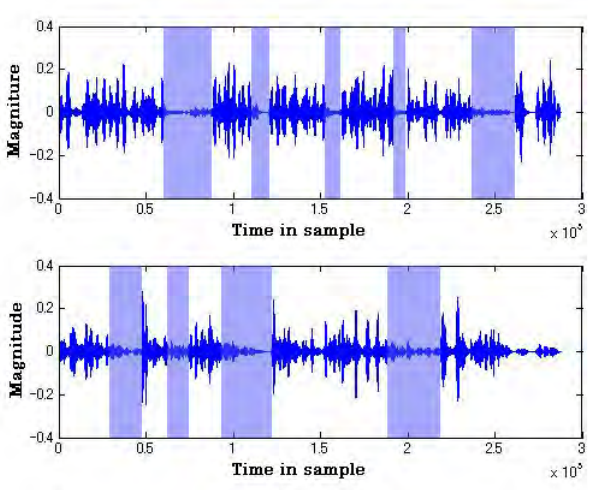


Figure 2: silent section of separation result

この無音部分から音源が 1 つのみとなる区間が図 3 のように得られる．図中，上の図が 1 つ目の音源 $s_1(t)$ の音のみが出ている区間 $m^1(t)$，下の図が 2 つ目の音源 $s_2(t)$ の音のみが出ている区間 $m^2(t)$ を表している．この処理を式で表すと，2 つの信号で片方の音のみが無音となる単一音声区間 $m^1(t), m^2(t)$ は

$$\begin{cases} m^1(t) = \bar{b}_1(t)b_2(t) \\ m^2(t) = b_1(t)\bar{b}_2(t) \end{cases} \qquad (3)$$

となる．ここで，$\bar{b}_i(t)$ は $b_i(t)$ の値が 0 のとき 1，1 のときに 0 に反転させる関数と定義する．すなわち，$\bar{b}_i(t)$ は音がある部分を表す．この関数 $m^i(t)$ をマスクとして用いると，観測信号 $x_j(t)$ に $m^i(t)$ のマスクを書けた時の信号を $x_j^i(t)$ とすると 1 つ目の音源部分 $x_1^1(t), x_2^1(t)$ および，2 つ目の音源部分 $x_1^2(t), x_2^2(t)$ は

$$\begin{cases} x_1^1(t) = m^1(t)x_1(t) \\ x_2^1(t) = m^1(t)x_2(t) \end{cases} \qquad (4)$$

$$\begin{cases} x_1^2(t) = m^2(t)x_1(t) \\ x_2^2(t) = m^2(t)x_2(t) \end{cases} \qquad (5)$$

となる．実際の抽出結果を図 4,5 に示す．この抽出された部分の信号を用いて音源定位の処理を行う．

2.3 単一音源の定位法

音源定位とは，音源が発する音からその音源物体の位置を判定することである．人間や多くの生物は，音の方向をある程度の正確さで識別することができる．この識別には，両耳間到達時間差（Inter-aural Time Difference，以下 ITD と略記），両耳間音圧差（Inter-aural Level Difference，以下 ILD と略記），耳介によるスペクトルパターンの変化等が手掛かりとして用いられることはよく知られている[5][6]．

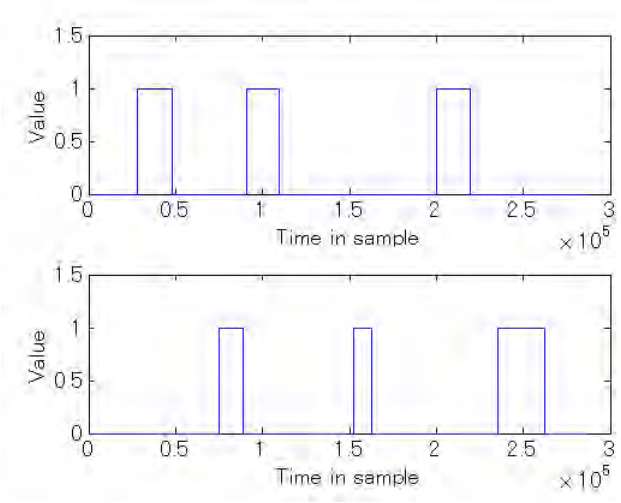


Figure 3: One sound source section

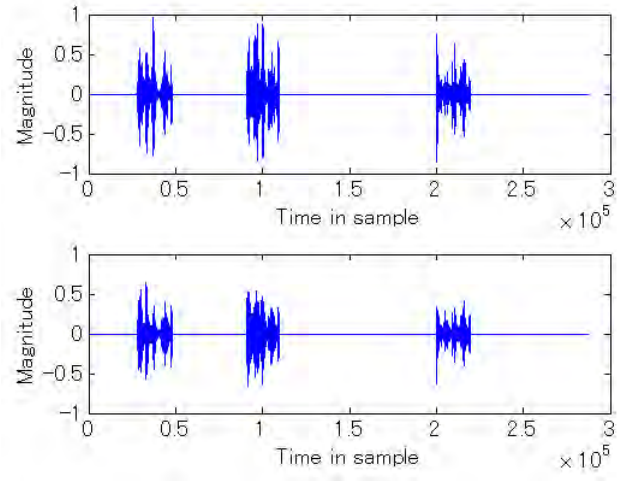


Figure 4: Extract only the sound of the 1st sound source($x_1^1(t), x_2^1(t)$)

2.3.1 水平方向の定位

心理実験から，人間は主に ITD と ILD を手掛かりとして，水平面内の音源方向を推定していると考えられる．つまり人間は，相対的に音が早く到達した方，又は相対的に音が大きい方に音源があるように認識する[7]．また，人間の場合，音源距離が 50cm 以上では，近似的に ITD の値は音源方向に対して，ほぼ一定となる[8]．人間では，時間差も音圧差も，音源の左右の位置を識別する手掛かりとして用いられているが，メンフクロウの場合には，音圧差は上下の音源を識別するために用いられている[9]．このようなメンフクロウの音源定位機構を模倣した，音源定位システムも提案されている[10]．

本研究では，水平方向を推定する手掛かりとして，ITD の情報のみを利用することにした．これは，ILD の情報

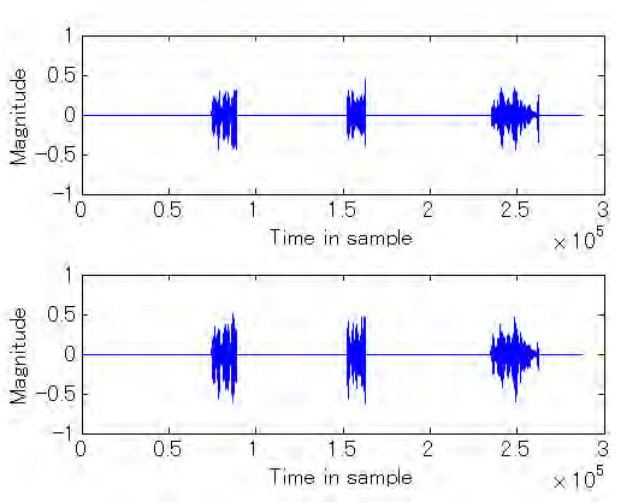


Figure 5: Extract only the sound of the 2nd sound source($x_1^2(t), x_2^2(t)$)

は雑音に弱く正確な方向推定に用いることは難しいのに比べ，ITD は検出が比較的容易であり，マイク間距離に比べて音源までの距離が十分大きい場合，原理的に ITD の値のみで音源方向が一意に決まり，ITD 情報のみで音源方向の推定が可能であるからである．

ITD の値は，左右の音データの相互相関係数値が最大値となる標本点の差 Δt とサンプリングレート f_s から求められ $\Delta t/f_s$ である．この ITD 値から音源方向の推定を行う．図 6 上の到達距離差 l は ITD 値と音速 c から次の式で求められる．

$$l = \frac{c\Delta t}{f_s} \tag{6}$$

この到達距離差 l から音源とマイク間の距離が十分遠い場合の音源方向 θ は，左右のマイク間距離を d とすると次式のように求められる．

$$\theta = \arccos(l/d) \tag{7}$$

2.3.2 垂直方向の定位

人間は，音の上下方向や前後方向の判断には，耳介などによる単耳手掛かりを使っている．耳介は，到達音に対して音源方向に依存したスペクトル上の変化を与える事によって，音源方向を推定する手掛かりを作っている．音源が広い範囲の周波数成分を多く含むほど，広い範囲に渡ってスペクトルの変化が起こるので，音源方向に関しての情報が多く得られ，音源方向の推定精度が高くなると考えられる．逆に，純音では情報がほとんど得られない為，定位することが困難となる．Hebrank らの心理実験によると，人間は 4k-15kHz の比較的高い周波数領域の音声情報を定位手掛かりとして用いていると考えられる[11]．また Shaw らは，人工外耳を用いた測定により，スペクトル

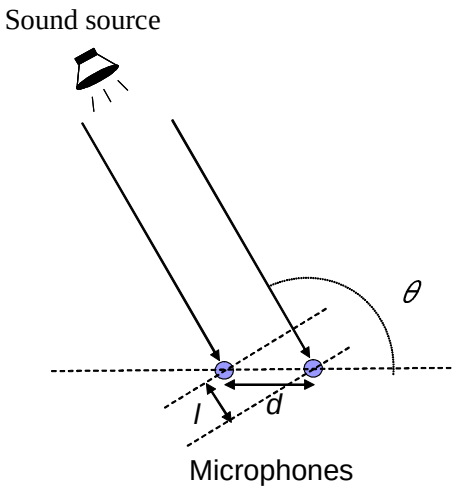


Figure 6: Estimate the direction of the sound source on the horizontal plane

に音源方向に依存した複数の窪みが存在することを示した[12]．さらに Hofman らは，ベイズ統計を用いたスペクトルを再構築することによって，スペクトルの特徴と音源方向の関連性を明らかにした[13]．小野らは，スペクトル形状が音源方向に依存して変化する特徴を利用し，反射板を用いた単耳音源定位システムを実現した[14]．このシステムでは，スペクトル波形上で極小値をとる周波数の対数が，音源方向に対して線形関係となるように反射板を設計することにより，広帯域雑音の音源方向を定位可能である．

我々は，上下方向（鉛直面方向）の音源定位動作を自己組織的に学習するロボットシステムの構築を行い，実験によってシステムが定位能力を学習可能である事を示した．[15] このシステムでの上下方向の定位手掛かりには，小野らの設計した反射板によって生じる周波数特性上の極小値を用いた．小野らのシステムでは，一部の周波数帯のみを使用し，キャリブレーションする事により正しい定位を可能としている．これに対し，本システムでは，より広い周波数帯を使用し，自己組織特徴マップによる学習によって定位能力を獲得している．より具体的には，小野らは１つの極小値周波数のみを含む周波数帯において，極小値周波数の対数と音源方向が線形になるように設計しているのに対し，本システムでは，複数の極小値周波数を含むような広い周波数帯を使用し，極小値周波数と音源方向の関係を学習によって獲得する．これによって，単一の情報のみを使用するのに比べて環境の変化に対してロバストな能力が獲得可能である．

本研究では，この我々の提案した手法を用いて上下方向の音源位置を推定することにした．

2.3.3 任意の音源への対応

この反射板を用いた上下定位の手法は音源に白色雑音を使用することを前提とした方法である為，音声等の場合には上下方向を正しく定位することが出来ない．これは音源が白色雑音等の周波数特性がフラットなもの以外の場合には，反射板の伝達特性を得る為に音源の周波数特性が必要となる為である．

我々はこの問題を解決するために，定位には使用しないもう一方のマイク情報を利用する手法を提案する．

音源の周波数特性を $S(\omega)$，音源から左右のマイクまでの伝達特性をそれぞれ $H_1(\omega), H_2(\omega)$，反射板の伝達特性を $R(\omega)$，マイクによる観測信号の周波数特性をそれぞれ $X_1(\omega), X_2(\omega)$ とする．音源に近い側のマイクに入る音 $(X_1(\omega))$ は反射板の影響を受けるが，音源から遠い側のマイクに入る音 $(X_2(\omega))$ は壁等の影響は受けるがその影響は反射板にくらべて十分小さいとし，これらの関係は次のように表されると仮定する．

$$X_1(\omega) = R(\omega)H_1(\omega)S(\omega) \quad (8)$$

$$X_2(\omega) = H_2(\omega)S(\omega) \quad (9)$$

この式から反射板の伝達特性 $R(\omega)$ は，

$$R(\omega) = \frac{H_2(\omega)X_1(\omega)}{H_1(\omega)X_2(\omega)} \quad (10)$$

となる．両耳への伝達特性の違いは無視出来ると仮定する．すなわち，音源-マイク間の伝達特性 $H_1(\omega), H_2(\omega)$ を $H_1(\omega) = H_2(\omega)$ とすると．

$$R(\omega) = \frac{X_1(\omega)}{X_2(\omega)} \quad (11)$$

となり，左右のマイクによる観測信号のみから反射板の伝達特性を $R(\omega)$ を計算することが可能となる．

この提案手法の検証の為に実験を行った．図 7 に示すような，周波数軸上に 3 つの窪みのある音源を用いて，左右のマイクで音を取り込んだ．図 8(a) は音源に近い側のマイクにおける周波数特性 $(X_1(\omega))$ ，図 8(b) は音源に遠い側のマイクにおける周波数特性 $(X_2(\omega))$ である．また図 8(c) が音源に白色雑音を用いた場合の周波数特性である．反響や伝搬による減衰等のない理想的な環境では，この周波数特性が反射板の周波数特性となる．そして，図 8(d) が計算により得られた周波数特性 $(R(\omega))$ となる．図 8(c) と図 8(d) を比較すると，形状はかなり異なるが，窪みの位置は近いことがわかる．白色雑音によって作成される特徴マップ[15]と比較する為に音源の高さを-60 度〜60 度まで変化させ特徴マップを作成した．その結果，図 9 に示すようにかなり似た特徴マップになっていることがわかる．このことから，この提案手法を用いることによって音源が白色雑音の場合とほぼ同等の定位が可能であると考えられる．よって，この手法を用いて音声の上下方向推定を行うこととした．

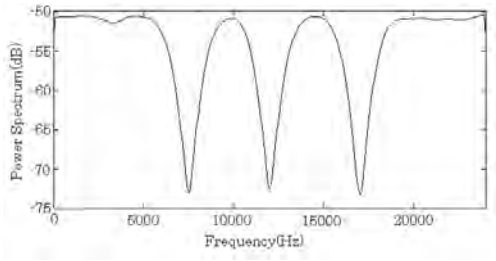
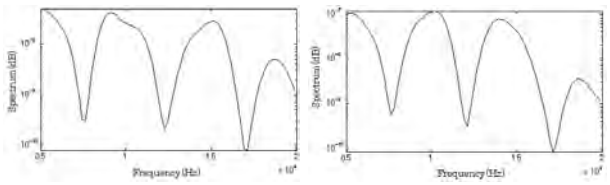
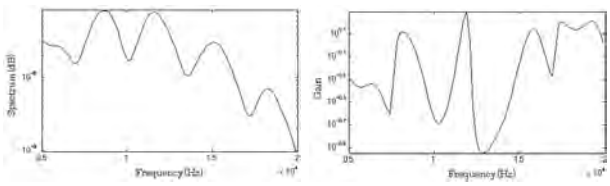


Figure 7: Frequency characteristic of the used sound



(a) Left microphone($X_1(\omega)$)　(b) Right microphone($X_2(\omega)$)



(c) Transfer function　(d) Result($R(\omega)$)

Figure 8: Frequency characteristic

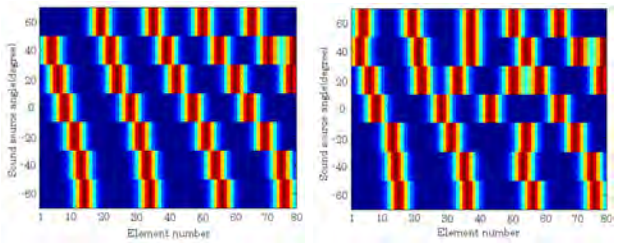
3　定位実験

3.1　実験環境

反射板に無指向性マイクを取り付けたものを左右の耳としてロボットに搭載した（図 10）．2 つのマイク間距離は 32cm とした．また，これらのマイクの床からの高さは 1.35m とした．マイクの音はサウンドボードを通して量子化ビット数 16 ビット，サンプリング周波数 48kHz の PCM データに変換される．スピーカは図 11 に示すようにマイクを搭載したロボットから 1.5m の距離の円周上に配置した．実験に使用する音声は，男性と女性の声のものを使用した．

3.2　実験方法

図 11 に示すように，1 つ目のスピーカ (Spk1) は水平方向の角度 30 度，高さ 1.35m で固定とし，2 つ目のスピーカ (Spk2) は水平方向の角度 45,60,80,120 度，高さ 1.05,1.35,1.65m の計 12 カ所を移動させた．音声は 1 つ目のスピーカから女性の声，2 つ目のスピーカから男性の声が出力される．2 つの混合音を 2 つのマイクによって約 17 秒間録音し，左右のマイクにより録音された量子化ビット数 16 ビット，サンプリング周波数 48kHz の PCM データを用いて音源定位処理を行う．無音部分を抽出する為の閾値 (I_{m1}, I_{m2}) は分離結果から適切に設定した．左右方向は式 (7) を用いて得られた角度を音源方向とする．また，上下方向は 1.05,1.35,1.65m の 3 つの高さの判別を行う．この判別処理は，予め白色雑音を用いて作成した各音源方向に対する特徴ベクトルのテンプレートとのマッチングにより行う．今回の実験において音源の高さを 30cm 刻みとしたのは，ロボットに搭載されているカメラの視野角がおよそ ±10degrees であることから，距離 1.5m の位置ではおよそ ±25cm の範囲を見ることが可能にする為である．



(a) White noise　(b) Test sound

Figure 9: Created feature vector

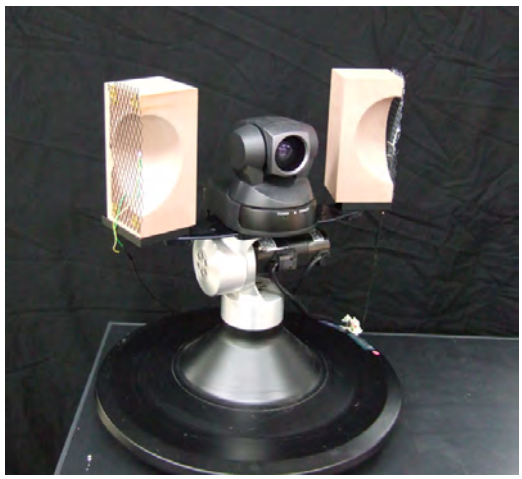


Figure 10: Head robot

3.3　実験結果

それぞれのスピーカ配置に対する水平方向の推定結果を表 1 に示す．表から 2 つの音源方向がほぼ正しく推定できていることがわかる．正面付近に比べ両端での精度が悪い．これはサンプリング周波数を 48kHz とした場合，正面付近では ITD の値が 1 違った場合に音源方向は 2 度程度しか変化しないが，両端付近では ITD の値が 1 違った場合に音源方向がおよそ 15 度も変化してしまう為である．また，垂直方向の推定結果を表 2 に示す．こちらも正しく推

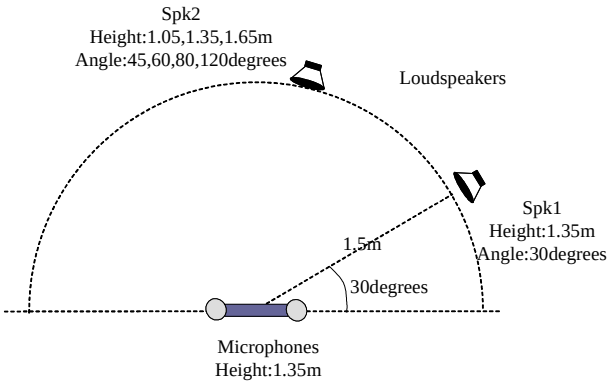


Figure 11: Arrangement of microphones and loudspeakers

Table 1: Estimation results on the horizontal plane

Height of speakers (cm)		Angle of speakers (degree)							
		Spk1	Spk2	Spk1	Spk2	Spk1	Spk2	Spk1	Spk2
Spk1	Spk2	30	45	30	60	30	80	30	120
135	105	35.5	50.7	35.5	64.6	35.5	79.6	35.5	111.2
135	135	35.5	49.0	35.5	64.6	35.5	76.9	35.5	118.4
135	165	35.5	43.6	35.5	60.2	35.5	79.6	35.5	118.4

定出来ていることがわかる．

スピーカの配置が 30 度と 60 度の場合の 1.05,1.35,1.65m の 3 つの高さに対する特徴ベクトルおよびテンプレートを図 12 に示す．図の上から 2 つ目の音源の位置が 1.05,1.35,1.65m の 3 つの高さに対するグラフである．図から 3 つの極小値の位置の内少なくとも 1 つはほぼ理想値と一致していることがわかる．このように複数の極小値周波数を用いることによって，誤認識を減らすことが出来ていると考えられる．

4 まとめ

本稿では，2 つの異なる位置から発せられる音声の音源方向を推定する手法を提案し，実験によりその有効性を示した．本手法を用いることにより，2 つのマイクのみで 2 つの音源の水平方向及び垂直方向の位置を推定することが可能となった．

今回提案手法を評価する為に行った実験では，スピーカを上下に 30cm 刻みで 3 カ所の高さに配置して音源方向の識別を行ったが，白色雑音を用いた場合にはより高い精度で識別が可能であるので，今後音声を用いた場合においても，精度を高めたい．また，本手法では反射板を経由せず直接マイクに到達する側の音を音源の音とみなし，音源の音の特徴を打ち消すことにより任意の音に対して上下方向の定位を可能にした．しかし，現在使用している反射板の配置では正面付近の音源に対しては正しく動作しない．これは，正面付近から来る音は左右共に反射版を経由しないかまたは左右共に反射を経由するかのどちらかになってしまう為である．今後この問題に対する解決方法を検討したい．さらに，実験において使用した分離信号に対する閾値は，現在分離結果から適切な値を設定しているが，これについても今後分離信号自身から生成する手法を検討したいと考えている．

Table 2: Estimation results on the vertical plane

Height of speakers (cm)		Angle of speakers (degree)							
		Spk1	Spk2	Spk1	Spk2	Spk1	Spk2	Spk1	Spk2
Spk1	Spk2	30	45	30	60	30	80	30	120
135	105	135	105	135	105	135	105	135	105
135	135	135	135	135	135	135	135	135	135
135	165	135	165	135	165	135	165	135	165

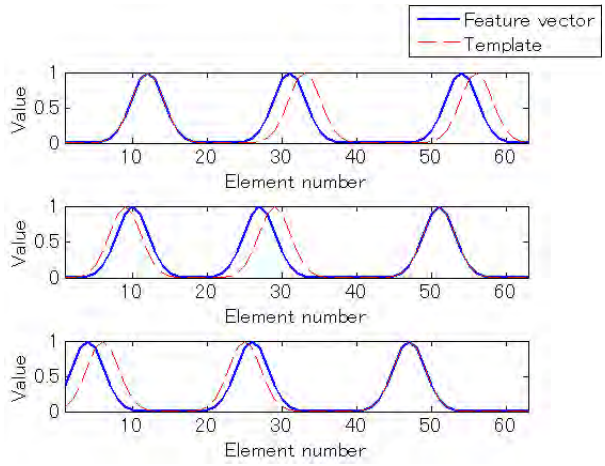


Figure 12: Feature vector and feature map to each sound source direction

参考文献

[1] 安藤繁, 篠田裕之, 小川勝也, 光山訓. 時空間勾配法に基づく 3 次元音 源定位センサシステム. 計測自動制御学会論文集, Vol. 29, No. 5, pp. 520–528, 1993.

[2] 中臺一博, 奥乃博, 北野宏明. アクティブオーディションによる複数音源の定位・分離・認識. 日本人工知能学会　第 16 回 AI チャレンジ研究会, pp. 25–32, 2002.

[3] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi. Real World Blind Separation of Convolved Non-Stationary Signals. In *ICA'99*, pp. 347–352, 1999.

[4] 松坂要佐, 小林哲則. 対面対話の収録と解析のための信号処理とパターン認識. 人工知能学会研究会資料 SIG-SLUD-A201, pp. 27–32, 2002.

[5] 大山正, 今井省吾, 和気典二（編）. 新編 感覚・知覚ハンドブック. 誠信書房, 1994.

[6] 吉田登美男, 亀田和夫. 新版 聴覚と音声, pp. 73–240. 電子情報通信学会, 1980.

[7] W.A. Yost. Lateral position of sinusoids presented with interaural intensive and temporal differences. *J. Acoust. Soc. Am.*, Vol. 70, No. 2, pp. 397–409, 1981.

[8] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *J. Acoust. Soc. Am.*, Vol. 106, No. 3, pp. 1465–1479, 1999.

[9] 小西正一. フクロウの音源定位の脳機構. 科学, Vol. 60, No. 1, pp. 18–28, 1990.

[10] L. Natale, G. Metta, and G. Sandini. Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. *Robotics and Autonomous Systems*, Vol. 39, pp. 87–106, 2002.

[11] J. Hebrank and D. Wright. Specral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.*, Vol. 56, No. 6, pp. 1829–1834, 1974.

[12] E. A. G. Shaw and R. Teranishi. Sound Pressure Generated in an External-Ear Replica and Real Human Ears by a Nearby Point Source. *J. Acoust. Soc. Am.*, Vol. 44, No. 1, pp. 240–249, 1968.

[13] P. M. Hofman and A. J. Van Opstal. Bayesian reconstruction for sound localization cues from responses to random spectra. *Biological Cybernetics*, Vol. 86, pp. 305–316, 2002.

[14] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi, and S. Ando. Biomimicry Sound Source Localization with Fishbone. *T.IEE*, Vol. 121-E, No. 6, pp. 313–319, 2001.

[15] 中島弘道, 向井利春, 大西昇. スペクトルの特徴マップを用いた上下方向音源定位学習システム. 電子情報通信学会論文誌, Vol. J87-DII, No. 11, pp. 2034–2044, 2004.

メインローブモデルを用いた複数音源定位手法の動的環境下での性能評価

Evaluation of Main-lobe Model Based Sound Localization on Mobile Robot

佐々木洋子 [1,2], **加賀美聡** [2,1], **溝口博** [1,2]

Yoko Sasaki, Satoshi Kagami, Hiroshi Mizoguchi

1. 東京理科大学 理工学研究科機械工学専攻

2. 産業技術総合研究所 デジタルヒューマン研究センター

Tokyo University of Science

National Institute of Advanced Industrial Science and Technology

{y-sasaki, s.kagami}@aist.go.jp, hm@rs.noda.tus.ac.jp

Abstract

The paper describes sound source localization for a mobile robot using microphone array. We use Main-Lobe Fitting method (MLF) for robust sound directional localization in varied real environment. MLF based sound localization is tested in different conditions, and the performance is compared with MUltiple SIgnal Classification(MUSIC) method. The result does not show big difference between two algorithms, but sound localization result by MLF shows beam forming approach can perform well as adaptive algorithm in varied conditions by selecting reliable peaks from observed signals.

1 はじめに

実環境下で周囲の音源を検出する機能は，音声認識や環境変化の初期知覚に役立つ．特に，直接見える範囲に限定される視覚情報と比較して，音はより広範囲を扱うことが可能であり，ロボットの環境知覚機能として有用である．

ロボットの音知覚機能としては，様々な環境でより広い範囲を確実に扱うことが重要である．しかし音は環境変化の影響を受けやすく，音源数の増加や，音源までの距離が長い場合，また反射や残響のある環境下では，音源定位，分離の性能は大きく低下する．これまでに，音声認識を目的としてロボットに装着したマイクによる音源定位，音源分離システムが多数提案されているが，広範囲を動く移動ロボットへ適用するためには，様々な環境下でより広範囲の音を扱うことが大きな課題である．

ロボットに搭載したマイクアレイによる動的環境下での音源定位については，パーティクルフィルタを用いた音源追跡 [1] や，カメラによる顔追跡を利用した話者定位 [2] などが報告されている．これらは，視覚情報を併用することで音情報のあいまい性を解消している．また Valin らは，8 チャンネルのマイクアレイを用いて移動ロボットによる複数音源の定位，移動音の追跡 [3] を実現している．

しかし，これらの研究ではより広範囲を扱うための，音源間の音圧差や音源数の増加については述べていない．この問題への対処法として，中臺ら [4] は，ロボット搭載型マイクアレイと環境に配置したマイクアレイを統合することで，音源までの距離増加に伴うあいまい性を解消している．また事前に環境内の静的な雑音源を観測することで，移動ロボットが対象音源を捉えやすい場所を選択する方法 [5] なども報告されている．

環境変化の初期知覚という観点からも，なるべく短区間のデータから複数の音源を確実に検出することが重要である．本稿では，ロボットに搭載したマイクアレイを用いた音源定位について述べる．移動中のロボットが，短区間のデータから複数の音源を検出するために，MLF 法を用いて信頼度の高い周波数成分を抽出することで，より頑健な音源定位を実現する．また音源数，音源までの距離変化に対する音源定位の性能を，MUSIC 法と比較して評価を行う．

2 音源定位手法

本節では，点音源に対して DSBF で観測される音圧分布を定義し，メインローブ部分をモデルとした音源定位手法 [6] について述べる．

2.1 遅延和ビームフォーミング

各チャンネルの信号出力が同位相となるように入力信号に遅延を与え加算することで目的方向の音を強調させる．マイクアレイの中心を点 O，点 O を中心とするアレイ直径より十分大きな円周上の点を C_ϕ とする．$\phi = 0$ の音源に対して，同位相化のために i 番目のマイクに与える遅延

時間 τ_i は，式 (1) で与えられる．

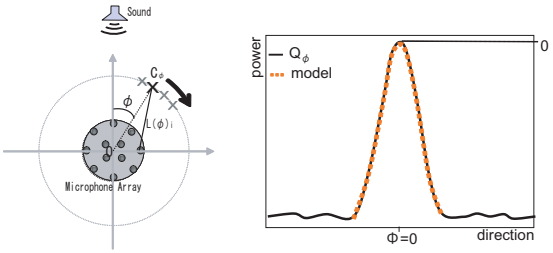
$$\tau_i = \frac{L_0(0) - L_i(0)}{V_s} \tag{1}$$

ただし，V_s は音速，$L_i(0)$ は C_0 から i 番目のマイク ($i = 1, 2, \cdots, M$) までの距離を表す．また $L_0(0)$ は C_0 からマイクアレイの中心までの距離とする．

Fig.1(a) のように，C_ϕ を $(-\pi < \phi \leq \pi)$ でスキャンさせることで，$\phi = 0$ の点音源に対して各方向で観測される音 $Q_\phi(\omega)$ が得られる．

$$Q_\phi(\omega) = \sum_{i=1}^{M} \sin\omega(t + \tau_i + \frac{L_i(\phi)}{V_s}) \quad (-\pi < \phi \leq \pi) \tag{2}$$

ただし，t は時刻，ω は音源の周波数を表す．



a) Scanning the focus　　b) Spatial spectrum

Figure 1: Spatial spectrum and main-lobe model

また音源定位のための空間スペクトルは次式で定義される．

$$Q'_\phi(\omega) = \frac{1}{N} \sum_{n=1}^{N} \sqrt{|\, Q_\phi(\omega, n)\,|^2} \tag{3}$$

観測されるピークの形状は，マイクアレイの指向特性により決まり，対象音源に対する離角 ϕ と，音源方向を基準とした音圧の減衰割合として，次式でメインローブモデルを定義する．

$$model(\omega, \phi) = \frac{|\, Q'_\phi(\omega)\,|}{|\, Q'_0(\omega)\,|} \qquad (-\varphi_m < \phi < \varphi_m) \tag{4}$$

2.2　メインローブモデル

観測された空間スペクトルのピークに対し，式 4 で求めたメインローブモデルを当てはめ，モデルと一致するピークを抽出することで，反射や他音源の干渉を受けた成分を除外する．

これ以降，簡単のために任意の周波数 ω について論じる．まず，n 番目 $(n = 1, 2, \cdots)$ の音源からの漏洩を $l_n(\omega)$，背景雑音を $BN(\omega)$ とすると，空間スペクトルの最大方向 θ_0 でのピーク値は，式 (5) で表される．

$$P_{\theta_0}(\omega) = S_0(\omega) + \{l_1(\omega) + l_2(\omega) + \cdots\} + BN(\omega) \tag{5}$$

ここで，$S_0(\omega)$ は θ_0 方向の音源の音圧を示す．また式 (4) のメインローブモデル $model(\omega, \phi)$ を用いて，空間スペクトル中の対象音源の推定スペクトルは，式 (6) で表される．

$$E(\omega, \theta) = \begin{cases} model(\omega, \theta) S_0(\omega) & \text{if } \theta_0 - \varphi_m \leq \theta \leq \theta_0 + \varphi_m \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

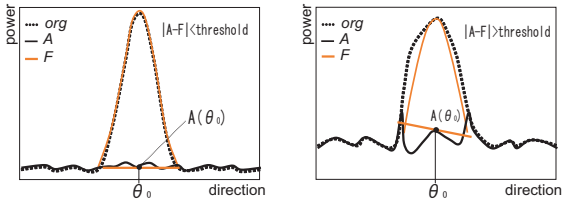
観測された空間スペクトルに対し，S_0 を変化させて推定スペクトル $E(\omega, \theta)$ を決定する．空間スペクトルから $E(\omega, \theta)$ を減算し，メインローブモデルとの適合判定を行う．θ_0 方向の音源を減算した空間スペクトルは，式 (7) となる．

$$A(\omega, \theta) = \{org(\omega, \theta) - E(\omega, \theta) + (P_{\theta_0}(\omega) - S_0(\omega))\} \tag{7}$$

減算後のスペクトルに対し，最小二乗法で $A(\theta_0)$ を通る直線を求め，メインローブ幅内でこの直線との残差 (絶対誤差の平均) をモデル適合の閾値に用いる．ここで，残差が閾値を上回りモデルと不適合と判断されたピークを除外し，以降，定位計算には用いない．モデルに適合した場合は，次のピークに対して同様にモデルの適合判定を行う．閾値を R_{th} として，モデル適合の判定式を次式に表す．

$$\frac{1}{2m} \sum_{\phi=-\varphi_m}^{\varphi_m} |\, A(\omega, \phi) - F(\omega, \phi)\,| < R_{th} \tag{8}$$

また Figure 2 に適合判定の模式図を示す．



a) Fit to the model　　b) Error fitting

Figure 2: Fit the main-lobe model to observed spectrum

観測された空間スペクトルに対し，最大ピークからメインローブモデルを当てはめ，適合した場合は空間スペクトルからモデル分を減算し，次のピークに対し，同様に適合判定を行う．周波数ごとにこの処理を繰り返し，複数のピークを検出する．

2.3　MLF を用いた音源定位

MLF により抽出された各周波数でのピークパワーを用いて音源方向を決定する．Figure 3 に音源定位の模式図を示す．左図は DSBF により得られる空間スペクトルの例であり，点線はモデルに適合するピークがない周波数成分，実線は適合するピークを持つ周波数成分を示している．右図に示した，モデルに適合したピークのピークパワー (左図赤点) の総和から音源定位を行う．信頼度の高いピークのみを抽出することで，近接した複数音源に対して，音源間の偽ピークの誤検出を減少させる．またピークが乱れた反射成分を除外することで，高反射，高残響下での音源定位を向上させる．

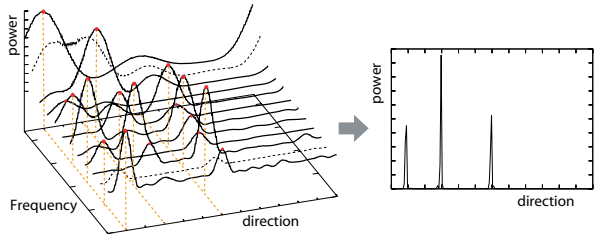


Figure 3: Directional localization process by MLF

2.4 MUSIC による音源定位

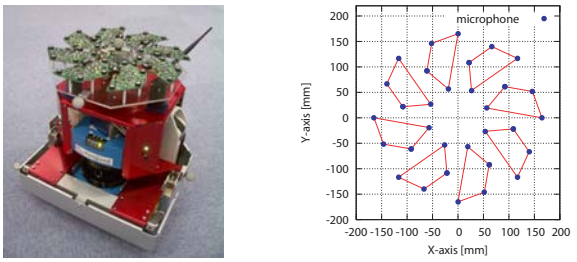
MLF 法による音源定位と比較するため，MUSIC(MUltiple SIgnal Classification) 法を用いて音源定位を行う．MUSIC はビームフォーミングと比較して，鋭いピークが得られるため，複数の音源を高精度に検出可能であることが知られている．

MUSIC には，浅野ら [7] の実装を用いる．定位計算には，音源数および伝達関数が必要であるが，本稿では音源数を既知とし，伝達関数にはマイク配置から計算したものを用い，部屋の残響等は含めないものとする．また空間スペクトルの計算には，周波数ごとの空間スペクトルの平均値を用いる．

3 音源定位実験

3.1 ハードウェア

32 チャンネルマイクアレイ [8] を搭載した車輪型ロボットを用いて，音源定位実験を行った．Figure 4 にマイクアレイを搭載した車輪型ロボットおよびアレイのマイク配置を示す．



a) Mobile robot with the array　b) Microphone arrangement

Figure 4: Mobile robot with microphone array

Figure 5 に，マイクアレイの指向特性を示す．(a) のビームフォーミングシミュレーションから求めた特性に対し，実測から求めた特性は-20(dB) 以上ではよく一致している．-20(dB) 以下での差は，ハードウェアの離散化によるもの，および残響等の影響であると考えられる．以下の実験では，シミュレーションから求めた，-20(dB) 以上のメインローブ部分をモデルとして採用する．

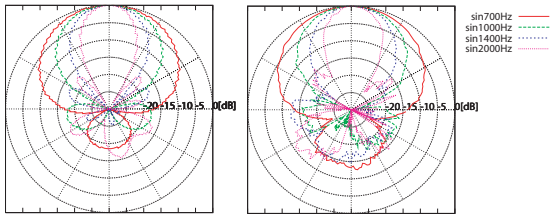
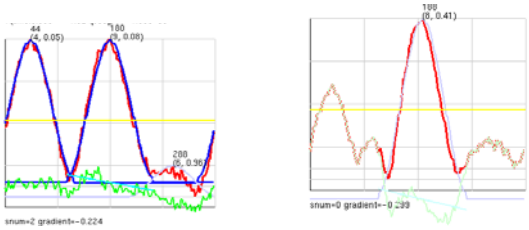


Figure 5: Directional pattern of simulation(left) and experimental measurement(right)

3.2 MLF 基礎実験

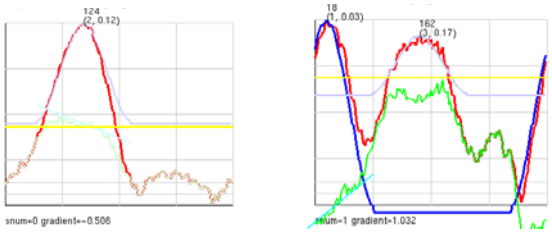
シミュレーションから作成したメインローブモデルを用いて，異なる環境でメインローブモデルの適合判定を行った．実験は，残響時間 (T_{60})450(msec)，騒音レベル (L_A)35(dB) の部屋 A，および残響時間 (T_{60})600(msec) の部屋 B の 2 箇所で行った．

Figure 6, 7 に 1kHz における空間スペクトルの例を示す．横軸は 0–359(deg) のロボットに対する水平角を示し，縦軸は最大値を 1 とした，各方向のパワーを表す．図中，赤線で示した観測スペクトルに対し，青線が推定されたメインローブモデル，緑線が観測スペクトルからモデルを減算したスペクトルを表している．



a) Detect 2 peaks　b) Fitting error

Figure 6: Example of MLF on 1kHz spatial spectrum

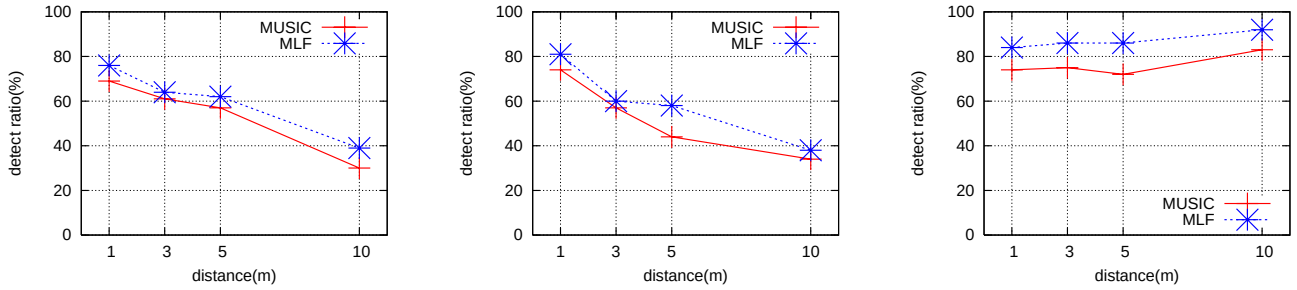


a) Fake peak by close 2 sounds　b) Reflection from wall

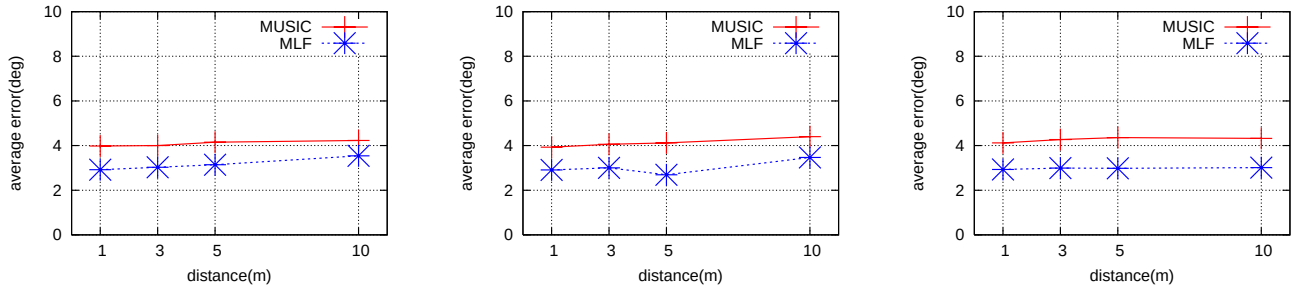
Figure 7: Examples of rejected peak by MLF

まず，1kHz のサイン波を音源として部屋 A で実験を行った．Figure 6 に，異なる時刻での空間スペクトルを示す．

音源方向は 45(deg) および 180(deg) で，ロボットからの距離はどちらも 2(m) である．Figure 6 a) は，44(deg) および 180(deg) に 2 つのピークを検出した例である．b) は，188(deg) にある第一ピークに対して，メインローブモデルに適合しなかった例を示している．



Detect ratio : expA, expB-female, expB-male from left side



Average error : expA, expB-female, expB-male from left side

Figure 8: Evaluation result of distant sound source localization

MLF による適合判定の結果，除外されたピークの例を Figure 7 に示す．実験は部屋 B にて行った．音源は，90(deg) および 135(deg) に配置した 1kHz のサイン波に対する空間スペクトルを Figure 7 a) に示す．ビームフォーミングではピーク幅が広いため，近接音源間に偽ピークが生じているが，124(deg) 方向に観測された第一ピークは，モデルには適合しない．Figure 7 b) は，反射により生じたピークの例である．音源は，20(deg) 方向，距離 2.5(m) の位置にあり，ロボットの後ろ (180(deg)) には，距離 1.5(m) の位置に壁がある．b) では，18(deg) に観測された第一ピークをモデルに適合するピークとして検出している一方で，162(deg) 方向に生じた第 2 ピークはモデルに適合せず除外されている．

Figure 6 b) や Figure 7 a) に示すように，2 つの音源が同一の周波数成分を持つ場合，干渉によって音源方向以外にピークが生じることがある．また Figure 7 b) のように，ビーム幅が広い DSBF では，反射音のような分布音源に対するピークは実際より大きく観測され得ることがわかる．

3.3 距離変化に対する精度評価

ロボット，音源共に静止した状態で，音源数および距離を変化させて音源定位の評価を行った．実験環境は，残響時間 (T_{60})450(msec)，騒音レベル (L_A)35(dB) の部屋である．音源には，あらかじめ接話マイクで録音した男声および女声の発話を用い，スピーカ (YAMAHA MS101II) から再生した．

まず，ロボットから音源までの距離を変化させて，音源定位を行った．男声音源を 135(deg) 方向，距離 1.5(m) に固定し，女声音源を 90(deg) 方向，距離 1, 3, 5, 10(m) の 4 箇所に変化させ，exp.A) 女声音源のみ，exp.B) 男声/女声の 2 音源，の 2 種類の実験を行った．それぞれ 15 秒間の録音に対し，検出率および平均誤差を求めた．なお，一回の定位に用いるデータ長さは 1024 点 (64msec) である．

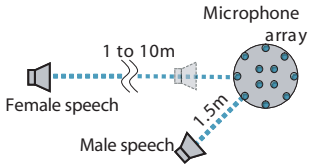


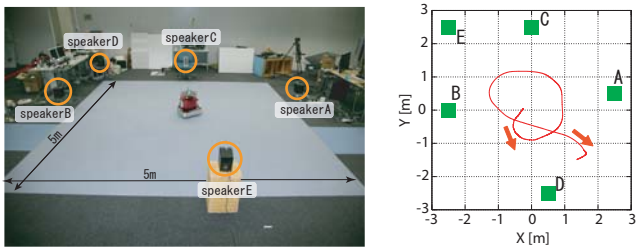
Figure 9: Speaker arrangement for distant sound source localization

Figure 8 にそれぞれの実験結果，Figure 9 に音源の配置図を示す．Figure 8 では上段に検出率，下段に平均誤差を示しており，それぞれ左から，exp.A 女声音源，exp.B 女声音源 (距離変化あり)，exp.B 男声音源 (固定) の結果を表す．検出率が距離の増加に伴って減少している一方，定位角の平均誤差には大きな違いはない．また MUSIC 法 (赤線) と MLF 法 (青線) を比較すると，MLF がやや良い結果を示しているものの，両者に大きな差はなく，距離の変化に対して同様の傾向を示していることが分かる．

3.4 移動中のロボットによる音源定位

移動中のロボットによる音源定位実験を行った．音源には，あらかじめ接話マイクで録音した男女 5 人の発話を用いる．再生には，冒頭に 100(msec) の 2.5(kHz) サイン波を挿入した，5ch DVD Audio (44kHz 64bit サンプリング) を作成し，スピーカから再生した．これは解析時に真値と同期を取り，発話と発話の間に生じる無音区間を検出率の計算から除くために用いる．ロボットの操作は JoyStick を用いた手動操作とした．また音源方向の真値を得るために，モーションキャプチャシステムを用いてロボットの軌跡を記録した．用いたシステムは MotionAnalysis の Eagle10 台で，$5 \times 5(m)$ の範囲を計測可能である．

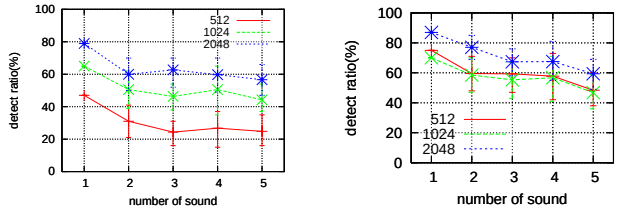
Figure 10 に実験環境の写真を示す．b) は 5 音源の実験におけるロボットの軌跡である．音源数を 1–5 とし，それぞれ 40(sec) 間，ロボットを走行させながら音源定位を行った．



a) Experimental setup　　b) Robot path

Figure 10: Robot path and speaker arrangement

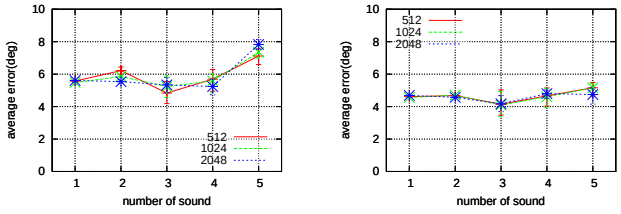
40(sec) 間の実験に対して，一回の定位に用いるデータ長さを 512, 1024, 2048 点 (それぞれ 32, 64, 128(msec)) として音源定位を行った．MUSIC 法，MLF 法それぞれについて，音源数の変化 (1–5) に対する検出率を Figure 11 に，平均誤差を Figure 12 にそれぞれ示す．なお，ここでの検出率は定位計算回数に対する音源の真値に対し ±15(deg) 以内で検出した回数の割合とし，計算には発話間に生じる無音区間 (一回の計算に用いるデータ全体が無音の場合) を除いてある．



a) MUSIC　　b) MLF

Figure 11: Detect ratio of sound source localization while moving

検出率については，MUSIC と比較して MLF が 5–10% 程度高い値を示している．MUSIC，MLF とも計算に用いるデータ長が短いほど，検出率が低くなっている一方，定位角の平均誤差はデータ長によらずほぼ一定で，MUSIC, MLF の間に大きな差はない．また MUSIC については 5 音源の場合に誤差が大きくなっている．



a) MUSIC　　b) MLF

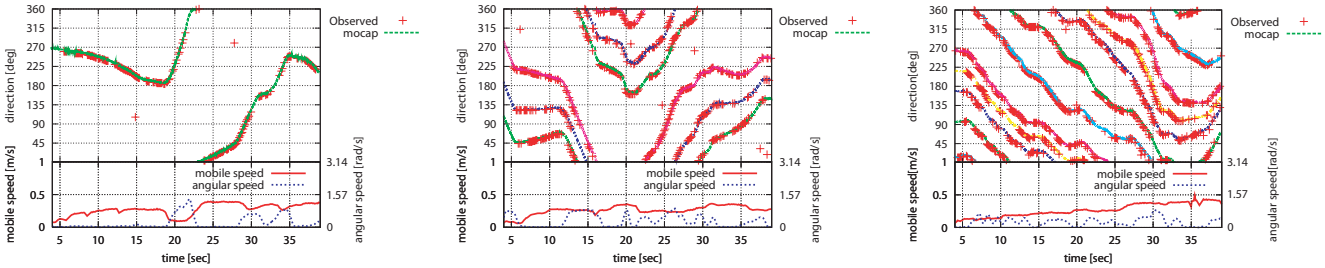
Figure 12: Average error of sound source localization while moving

1, 3, 5 音源の実験における，音源定位結果とロボットの速度を Figure 13 にまとめる．上段が MUSIC 法，下段が MLF 法による結果である．全体としては，MUSIC, MLF ともに似たような傾向を示しており，1 音源の 20–25(sec) 付近や 3 音源の 12–16(sec) 付近のように，ロボットの回転角速度が大きい場合には，MUSIC, MLF ともにうまく検出できていないことがわかる．また 5 音源の結果について，MUSIC と MLF を比較すると，10(sec) 前後で 120(deg) 方向の音源 (Figure 10 b) の音源 C) や，25(sec) 前後に 50(deg) 付近にある音源 (同図，音源 D) については，MUSIC の検出率がやや低くなっている．近くに他の対象音源があり，対象音源が離れている場合に，MUSIC の検出率が低くなる傾向があり，これが全体としての検出率の差の原因であると考えられる．

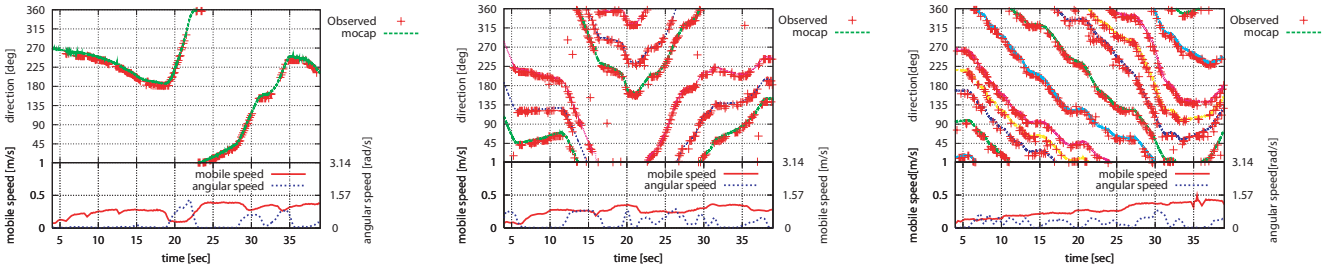
4 考察

音源数，音源までの距離等の条件を変化させて，MLF 法，MUSIC 法それぞれについて音源定位の性能評価を行った．全体的な結果としては，2 つの手法の間で大きな性能差はなく，環境条件の変化に対する性能の違いも似たような傾向を示した．定位精度に関しては，ビーム幅が広く高分解能を得にくい焦点型アルゴリズムを基本とした MLF 法であっても，信頼度の高いピークのみを抽出することで死角型アルゴリズムである MUSIC 法と同程度の精度を得られることがわかった．移動中のロボットによる音源定位の検出率については，MLF 法が 5-10%程度高い値となった．検出率の差は，近くに他の音源があり，対象音源までの距離が大きい場合に現れている．

ただし，今回の実験における MUSIC の実装に関しては，音源数を既知としており，この事前情報がない場合の評価は行っていない．伝達関数については，アレイのマイク配置から算出した音源からの距離差のみを用いたが，実装したアレイが平面配置であるため，インパルス応答の計測なしでも十分実用可能であることがわかった．また本稿の実験では，MUSIC の空間スペクトルは全周波数



MUSIC localization result (1, 3, 5 sound sources from left side)



MLF localization result (1, 3, 5 sound sources from left side)

Figure 13: Directional localization results while moving

の平均値としたが，計算に用いる周波数成分の選択や重み付けを行うことで，精度の向上が期待できる．

一方，ロボットの回転角速度が大きい場合には，MLF，MUSIC ともに大きく定位性能が劣化しており，DSBF であれば，マイクごとに与える遅延時間を一回の計算区間内で変化させる (時刻方向に可変にする) など，ロボットの動きを考慮した手法の改良が必要であるといえる．たとえば，角速度 180(deg/sec) では，512 点 (32(msec)) のデータの間で 5.8(deg) 姿勢が変化する．

また計算コストについては，本稿では厳密な検証をしていないが，MUSIC 法ではマイク数に比例するサイズの行列の固有値展開を多様しており，ビームフォーミングを基本とする MLF 法と比較すると計算コストは高い．

5 おわりに

本稿では，移動ロボットに搭載したマイクアレイを対象として，移動ロボットが複数の音源をより広い範囲で検出するための音源定位法を検討し，MLF 法と MUSIC 法の比較を行った．移動中のロボットによる音源定位結果から，どちらの手法でも 5(m) 以内の 5 つの音源を検出可能であるといえる．MLF 法による音源定位では，一般に高分解能を得にくいビームフォーミングを基本としているが，信頼度の高いピークを抽出することで，死角型アルゴリズムと同程度の性能を示すことがわかった．

MLF 法については，反射，回折等の影響を考慮したビームモデルを持つことで，ロボットから直接見えない音の検出も可能であると考えられ，点音源モデルで除外したピークの扱いが今後の課題である．

参考文献

[1] H. Asoh, I. Hara, and F. Asano. Tracking human speech events using a particle filter. In *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP 2005)*, pp. MSP–P2.6, Philadelphia, USA, 2005.

[2] H-D. Kim, J-S. Choi, and M. Kim. Speaker localization among multi-faces in noisy environment by audio-visual integration. In *Proceedings of IEEE-RAS International Conference on Robots and Automation (ICRA2006)*, pp. 1305–1310, Orlando, Florida, 2006.

[3] J-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems Journal*, Vol. 55, pp. 216–228, 2007.

[4] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, Y. Hasegawa, and H. Tsujino H G. Okuno. Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2006)*, pp. 852–859, Beijing, China, 2006.

[5] E. Martinson and A. Xchultz. Auditory evidence grids. In *Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robot and Systems(IROS2006)*, pp. 1140–1145, Beijing, China, 2006.

[6] 佐々木洋子, 加賀美聡, 溝口博. マイクアレイのメインローブモデルを用いた点音源検出手法. 第 25 回日本ロボット学会学術講演会講演論文集, p. 1N13, 千葉工業大学, 2007.

[7] F. Asano, M. Goto, K. Itou, and H. Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. In *Proceedings of European Conference on Speech Communication and Technology(Eurospeech2001)*, pp. 1013–1016, Aalborg, Denmark, 2001.

[8] 佐々木洋子, 加賀美聡, 溝口博. 移動ロボット搭載用 32ch マイクロホンアレイの設計と精度評価. 第 24 回日本ロボット学会学術講演会講演論文集, p. 1B19, 岡山大学, 2006.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-4

TRACKING A VARYING NUMBER OF SPEAKERS USING PARTICLE FILTERING

Angela Quinlan, Mitsuru Kawamoto and Futoshi Asano
AIST Information Technology Research Institute,
Tsukuba, Ibaraki 305-8568,
{angela-quinlan, m.kawamoto, f.asano}@aist.go.jp

Abstract

The extension of particle filtering techniques to the multiple speaker case is difficult as two distinct problems needs to be addressed. Firstly, the active speakers must be identified and their locations estimated, requiring the use of multi-dimensional likelihoods. Finally each speaker must be correctly associated with his corresponding location. In this paper we propose a multi-speaker tracking algorithm in which the number of active speakers is determined by estimating the profile of the noise-plus-reverberation covariance matrix eigenvalues. The multi-dimensional likelihoods are then decoupled using the Expectation Maximization (EM) algorithm. Using pause detection a continuously updated estimate of the noise-plus-reverberation covariance matrix is estimated. The results show the benefits of the proposed methods under difficult tracking situations.

1 Introduction

The ability to track the locations of a varying number of speakers in the presence of background noise and reverberation is of great interest due to the vast number of potential applications including security surveillance, pre-processing in speech recognition as well as many others. Kalman filters can be used for source tracking in situations where an assumption of a Gaussian uncertainty model and linear dynamics can be made. However, it is well known that in situations where these assumptions do not hold the Kalman filter is no longer accurate [1].

In these situations particle filtering techniques can instead be applied, and in recent years many authors have reported on the application of these techniques to tracking audio sources, e.g. [2, 3]. Using the particle filtering approach the location estimation is recursively updated using a two-step process of prediction and weighting. The prediction step uses prior information of the speaker's previous location together with a pre-defined motion model (usually random walk) to predict the current location of the speaker. This "prediction-likelihood" is then weighted by the received microphone signals - through the measurement likelihood, to give the posterior distribution from which the location estimate can be found.

While various particle filtering methods have been applied to the problem of tracking a single speaker, the extension of these techniques to the case of multiple speakers is not straightforward. This is because in the situation of multiple speakers two distinct problems now have to be solved, the estimation of the locations, involving multi-dimensional likelihoods, and also the association of the estimates with the correct source track. This data association problem is further complicated by the fact that one or more of the speakers may not be speaking making it necessary to estimate which speakers are "active" at a given time.

Multiple Hypothesis Testing (MHT) is an optimal way of performing state estimation by recursively building the data association hypotheses which assign the observations to either measurements or false alarms [4]. The probability that each hypothesis is correct is then found and the required source parameters are estimated using the most probable hypothesis. At each time step the hypothesis of a new source is tested resulting in an exponential increase in the computational complexity with time.

Various alternative approaches to the problem of multi-source tracking have been proposed, e.g. see [5][6] and the references therein. In [7] the estimation problem is cast as an incomplete data problem where the associations between targets and measurements are considered to be unobserved data. Using this framework the Expectation Maximization (EM) algorithm can be used to estimate the source locations and track associations. However, as the method proposed here uses the entire set of data in order to iteratively estimate the "unobserved data" an exponential growth in memory is required.

In [8] a similar approach is used, however in this case the problem is formulated in a recursive manner. The

computational complexity is then further reduced by modeling the association process as a 2nd order Markov Rand Field (MRF). While, in [5] the classical particle filter framework is extended to cover the multi-source situation using Gibbs sampler based estimation of the assignment probabilities. In this case the data association probabilities are also used for estimation of the number of targets present.

In this paper we are primarily concerned with tracking a varying number of speakers using particle filtering techniques. While many approaches for dealing with multi-source tracking using particle filters have been proposed, few of these techniques have been successfully applied to the speaker tracking problem. In [9] a method is proposed for tracking multiple speakers using fusion of both audio and video data. In this approach the speech activity state of each speaker is updated using a pre-defined transition matrix which takes into account conversation dynamics as the speakers are assumed to be conversing.

More recently a method for tracking multiple sources using audio signals only was proposed in [10]. In this case the computational complexity due to the multiple sources is reduced by exploiting the signal separation characteristics of the Expectation Maximization (EM) algorithm to estimate the particle filter weights. The activity state of the sources was assigned randomly, and the locations of two sources spaced a fixed distance apart were seen to be accurately tracked. This method was then extended on in [11] in order to avoid confusion of the source tracks in situations where one of the sources is inactive for a significant lengths of time.

In this paper we propose a method for estimating the number of active sources directly from the received data using an extension of the method proposed in [12]. This method is based on estimating the profile of the noise-plus-interference covariance matrix eigenvalues and comparing them with the eigenvalues of the observed covariance matrix.

The dimensionality problem due to the multiple sources is addressed by use of the Expectation Maximization (EM) algorithm. In the presence of M sources the inherent signal separation properties of the EM algorithm are used to decouple the M-dimensional likelihood into the M 1-dimensional likelihoods [10], thereby significantly reducing the computational complexity.

A pause detection step is used to select the relevant frequencies for calculating the measurement likelihoods in order to reduce errors due to the inclusion of inactive frequencies [11]. This step is also used to estimate the noise-plus-reverberation covariance matrix at each frequency. The proposed algorithm is validated using live recordings.

2 Problem Formulation

We consider the model of an array of M microphones located in a sound field generated by N_a active sources, which are assumed to be non-coherent.

Then, taking the short-term Fourier transform of the signals received by the microphones at time t, we obtain the following data model:

$$\mathbf{y}(\omega, \mathbf{t}) = \mathbf{A}(\omega, \mathbf{t})\,\mathbf{s}(\omega, \mathbf{t}) + \mathbf{n}(\omega, \mathbf{t}), \tag{1}$$

where ω is the frequency under consideration. In what follows we omit the frequency index for the sake of simplifying the notation. $\mathbf{A}(\omega, \mathbf{T})$ is the matrix of the L direct path transfer function vectors:

$$\mathbf{A} = [\mathbf{a}(\theta_\mathbf{1}), \ldots, \mathbf{a}(\theta_\mathbf{L})], \tag{2}$$

with
$\theta_l, l = 1, \ldots, N_a$ representing the 2D directions of the N_a sources. $\mathbf{s}(\mathbf{t}) = [\mathbf{S_1}(\mathbf{t}), \ldots, \mathbf{S_{N_a}}(\mathbf{t})]^\mathbf{T}$ is the source spectrum vector, and $\mathbf{n}(\mathbf{t}) = [\mathbf{N_1}(\mathbf{t}), \ldots, \mathbf{N_M}(\mathbf{t})]$ is the background noise spectrum vector. The signal and noise covariance matrices are defined respectively as:
$\mathbf{R_{ss}} = \mathbf{E}\left[\mathbf{s}(\mathbf{t})\,\mathbf{s^H}(\mathbf{t})\right], \quad \mathbf{R_{nn}} = \mathbf{E}\left[\mathbf{n}(\mathbf{t})\,\mathbf{n^H}(\mathbf{t})\right],$
where the superscript H denotes the conjugate transpose of the matrix.

2.1 Particle Filtering Algorithm

Using the framework of Bayesian hidden state sequence estimation, the particle filtering algorithm estimates the locations (θ_l) of moving sources by combining the information received in the observations with any available prior knowledge of the source transition model.

The hidden variable vector is defined as [13]:

$$\chi(\mathbf{t}) = \left[\mathbf{N_T}(\mathbf{t}), \chi_\mathbf{1}(\mathbf{t}), ..., \chi_{\mathbf{N_T(t)}}(\mathbf{t})\right], \tag{3}$$

where $N_T(t)$ is the total number of sources being tracked and the required parameters of the ith source are defined as $\chi_l(t) = (\theta_l, s_l(t))$. θ_l defines the location as in (2), and $s_l(t)$ is a Boolean variable which denotes the activity state of the ith source, i.e. whether the ith source is switched on/off. The observation variables, $\mathbf{Z}(\mathbf{t})$, are composed of the audio signals $z(t)$ received by the microphone array.

The posterior probability distribution $P\left(\chi_{\mathbf{1|T}}|\mathbf{Z_{1|T}}\right)$ specifies the likelihood of each possible $\chi_{\mathbf{1|T}}$ given the observations $\mathbf{Z}(\mathbf{t})$. The estimated hidden variable vector, $\widehat{\chi}_{\mathbf{1|T}}$, should then be selected so as to maximize this distribution.

Unfortunately, the distribution $P\left(\chi_{\mathbf{1|T}}|\mathbf{Z_{1|T}}\right)$ is not available. However, under certain non-restrictive assumptions, the required distribution can instead be approximated in accordance with Bayes' theorem using the measurement likelihood $P(\mathbf{Z}(\mathbf{t})\,|\chi(\mathbf{t}))$, and the state transition probability, $P(\chi(\mathbf{t}), \chi(\mathbf{t}-\mathbf{1}))$[13]:

$$P\left(\chi_{\mathbf{1|T}}|\mathbf{Z_{1|T}}\right) \propto \prod_{l=1}^{N_a} P(\mathbf{Z}(\mathbf{t})\,|\chi(\mathbf{t}))\,P(\chi(\mathbf{t}), \chi(\mathbf{t}-\mathbf{1})), \tag{4}$$

As no closed-form solution exists for (4) we approximate this distribution at a number of discrete points - or particles. Then, according to the central limit theorem as the number of samples increases toward infinity this approximation approaches the true posterior density.

The basic particle filtering framework can then be applied as a two-step *prediction* and *filtering* process. The

prediction step consists of propagating the particles according to the motion model and then in the filtering step the propagated particles are weighted according to the measurement likelihood corresponding to this particle location. The particles are then re-sampled according to these importance weights.

The final estimate of the source locations can then be found by taking the mean of the re-sampled particles.

$$\widehat{\chi} = \frac{1}{N_p} \sum_{i=1}^{N_p} \chi^{\mathbf{i}}, \tag{5}$$

where N_p is the total number of particles and $\chi^{\mathbf{i}}$ is the parameter vector associated with the *ith* particle.

3 Estimation of the Number of Active Sources

From equation (3) it can be seen that as the activity state of the speakers is unknown it is necessary to estimate N_a the number of active sources as well as their respective locations. There are two main approaches to this problem.

In the first case the particle filter can be applied to the joint problem of estimating and tracking the sources present. However this approach leads to high computational complexity. Therefore in this paper we instead use the alternative approach of firstly estimating the number of sources present and then using the particle filter to perform the tracking.

The MDL [14] and AIC [15] criteria are traditionally used for source number estimation. However, both these approaches are based on an assumption of white noise and are known to consistently over-estimate the number of sources present when reverberation is present [16]. In what follows we use the method proposed in [17] extended to cover reverberant environments as detailed in [12]. The spatial correlation matrix of the received signals $\mathbf{R_{yy}}(\omega)$ is defined as:

$$\mathbf{R_{yy}}(\omega) = \mathbf{E}\left[\mathbf{y}(\omega, \mathbf{T})\, \mathbf{y^H}(\omega, \mathbf{T})\right] \tag{6}$$

Then defining $\mathbf{R_{ss}}(\omega)$ as the spatial correlation of the source signals, $\mathbf{I}$ as the $M \times M$ identity matrix, and assuming the noise $\mathbf{n}(\omega, \mathbf{T})$ is spatially white and uncorrelated from the sources with noise power equal to σ^2, (6) can be re-expressed as:

$$\mathbf{R_{yy}}(\omega) = \mathbf{A}(\omega)\, \mathbf{R_{ss}}(\omega)\, \mathbf{A^H}(\omega) + \boldsymbol{\sigma^2}(\omega)\, \mathbf{I}, \tag{7}$$

The eigenvalues of $\mathbf{R_{yy}}(\omega)$ are therefore given by:

$$\lambda_1(\omega), ..., \lambda_M(\omega) = \gamma_1(\omega) + \sigma^2(\omega), ..., \tag{8}$$

$$\gamma_{N_a}(\omega) + \sigma^2(\omega), \sigma^2(\omega), ..., \sigma^2_M(\omega). \tag{9}$$

The number of eigenvalues corresponding to the signal subspace, the so-called signal eigenvalues, is equal to the number of active sources, and assuming that the source power is greater than that of the background noise, the number of sources present can now be easily determined as the number of eigenvalues not equal to σ^2.

In practice however, $\mathbf{R_{yy}}(\omega)$ is unknown and must instead be estimated using:

$$\widehat{\mathbf{R}}_{\mathbf{yy}}(\omega) = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{T=1}}^{\mathbf{N}} \mathbf{y}(\omega, \mathbf{T})\, \mathbf{y^H}(\omega, \mathbf{T}), \tag{10}$$

where N is the number of frames the spatial correlation is taken over.

In this case the active source number estimation problem still consists of distinguishing between the signal and noise eigenvalues. But, with the statistical fluctuations in $\mathbf{R_{yy}}(\omega)$, the noise eigenvalues are no longer all equal to σ^2. The separation between noise and signal eigenvalues is only clear now in the case of high Signal to Noise Ratio (SNR) and low reverberation, when a gap can be clearly observed.

In order to distinguish between the signal and noise eigenvalues we approximate the decreasing profile of the eigenvalues of the noise spatial correlation matrix, $\mathbf{R_{nn}}$, and compare this to the profile of the observed eigenvalues, where $\mathbf{R_{nn}}$ is defined as:

$$\mathbf{R_{nn}}(\omega) = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{T=1}}^{\mathbf{N}} \mathbf{n}(\omega, \mathbf{T})\, \mathbf{n^H}(\omega, \mathbf{T}). \tag{11}$$

As $\mathbf{R_{nn}}(\omega)$ has a Wishart distribution [18] it is extremely difficult to find the decreasing profile of its eigenvalues. However, this profile can be approximated using the first and second order moments of the eigenvalues together with an initial assumption of white noise [17].

The smallest observed eigenvalue is assumed to be a noise eigenvalue, corresponding to a noise subspace dimension of $P = 1$. Then letting $P = P + 1$ for each subsequent step until $P = M - 1$, the predicted profile of the noise only eigenvalues is found recursively using:

$$\widehat{\lambda}_{M-P}(\omega) = (P+1)\, J_{P+1} \widehat{\sigma}(\omega)^2, \tag{12}$$

where:

$$J_{P+1} = \frac{1 - r_{P+1,N}}{1 - (r_{P+1,N})^{P+1}}; \tag{13}$$

$$\widehat{\sigma}(\omega)^2 = \frac{1}{P+1} \sum_{i=0}^{P} \lambda_{M-i}(\omega); \tag{14}$$

$$r = e^{-2a}; \tag{15}$$

and:

$$a(M,N) = \sqrt{\frac{1}{2}\left\{ \frac{15}{M^2+2} - \sqrt{\frac{225}{(M^2+2)^2} - \frac{180M}{N(M^2-1)(M^2+2)}} \right\}}. \tag{16}$$

The relative differences between the predicted and observed eigenvalue profiles r_m are calculated using:

$$r_m(\omega) = \frac{\lambda_m(\omega) - \hat{\lambda}_m(\omega)}{\hat{\lambda}_m(\omega)}, \qquad m = 1, ..., M-1, \tag{17}$$

and r_m is then compared to a threshold value η_m in order to distinguish the signal eigenvalues. For a discussion on how to select this threshold value see [12].

The predicted noise eigenvalue profile $\left[\hat{\lambda}_1, ..., \hat{\lambda}_M\right]$ is based on the assumption that the background noise can be modeled as white noise. This approximation is valid in many practical situations when none of the speakers are active. Once some of the speakers are active though, reverberant tails arising due to the presence of speech violate this white noise assumption and lead to an increase in the noise eigenvalue profile.

In this case the noise eigenvalue profile predicted from equations (12)-(16) will be lower than that of the observed noise eigenvalues, resulting in frequent overestimation of the number of active sources. Therefore once it is known that at least one speaker is present, it is necessary to apply a correction factor to the predicted profile in order to account for the increase in the noise eigenvalues due to reverberation.

In order to calculate a suitable correction factor the eigenvalues of the estimated reverberation correlation matrix, $\lambda_1^{rev}(\omega), ..., \lambda_M^{rev}(\omega)$, are found. These values are then used to find the corresponding predicted noise eigenvalues $\hat{\lambda}_1^{rev}(\omega), ..., \hat{\lambda}_M^{rev}(\omega)$ as described in equations (12)-(16).

The difference between the predicted and observed profiles, relative to the largest observed eigenvalue, is then taken as a correction factor:

$$cf_m(\omega) = \frac{\lambda_m^{rev}(\omega) - \hat{\lambda}_m^{rev}(\omega)}{\lambda_1^{rev}(\omega)}, \qquad m = 2, ..., M. \quad (18)$$

In the presence of at least one active source the correction factor is then used to modify the originally predicted noise eigenvalue profile:

$$\hat{\lambda}_m^{mod}(\omega) = (1 + cf_m(\omega))\,\lambda_1(\omega). \quad (19)$$

Once again the predicted and observed profiles are compared by finding their relative difference:

$$r_m^{mod}(\omega) = \frac{\lambda_m(\omega) - \hat{\lambda}_m^{mod}(\omega)}{\hat{\lambda}_m^{mod}(\omega)}. \quad (20)$$

If $r_m > \eta_m$ then λ_m is a signal eigenvalue. The number of active sources is then equal to the number of signal eigenvalues.

4 Estimation of Measurement Likelihood Using Expectation Maximization

When tracking N_T sources the measurement likelihood distribution is an N_T-dimensional distribution and accordingly the calculational complexity grows exponentially as the number of sources increases.

A solution to this complexity problem proposed in [10] is the use of the Expectation Maximization (EM) algorithm. The main feature of the EM algorithm is that it decouples the N_T-dimensional likelihood distribution into N_T, 1-dimensional distributions which can be calculated in parallel.

This decoupling of the sources is achieved by decomposing the observed microphone signals into 'complete data' vectors which correspond to the signal due to each source:

$$\mathbf{y}(\mathbf{t}) = \sum_{\mathbf{l=1}}^{\mathbf{N_a}} \mathbf{x_l}(\mathbf{t}) = \mathbf{H}\mathbf{x}(\mathbf{t}), \quad (21)$$

where

$$\mathbf{x_l}(\mathbf{t}) = \mathbf{a}(\theta_\mathbf{l})\,\mathbf{S_l}(\mathbf{t}) + \mathbf{n_l}(\mathbf{t}),$$
$$\mathbf{x}(\mathbf{t}) = \left[\mathbf{x_1^T}(\mathbf{t}), \ldots, \mathbf{x_{Na}^T}(\mathbf{t})\right];$$
$$\mathbf{H} = [\mathbf{I}, \ldots, \mathbf{I}];$$

and the matrix $\mathbf{I}$ denotes the identity matrix. $\mathbf{n_l}(\mathbf{t})$ an arbitrary decomposition of the noise vector, which must satisfy $\mathbf{n}(\mathbf{t}) = \sum_{\mathbf{l=1}}^{\mathbf{N_a}} \mathbf{n_l}(\mathbf{t})$ and $\mathbf{R_{nl}} = \mathbf{E}[\mathbf{n_l}(\mathbf{t})\,\mathbf{n_l}(\mathbf{t})]$.

The likelihood of the complete data is then given by:

$$L_{xl}(\theta_l, \gamma_l | \mathbf{X_l}) = \Psi_{xl} \exp\left(-\frac{1}{2} tr\left[\mathbf{C_{xl}}\mathbf{K_{xl}^{-1}}\right]\right), \quad (22)$$

where:

$$\Psi_{xl} = (2\pi)^{-MN} \left[\det \mathbf{K_{xl}}\right]^{-N/2};$$

$$\mathbf{K_{xl}} = \gamma_\mathbf{l}\mathbf{a}(\theta_\mathbf{l})\,\mathbf{a^H}(\theta_\mathbf{l}) + \mathbf{R_{nl}}, \quad (23)$$

and the sample covariance matrix of the complete data $\mathbf{X_l}$ is given by:

$$\mathbf{C_{xl}} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{x_l}(\mathbf{n})\,\mathbf{x_l^H}(\mathbf{n}). \quad (24)$$

As the complete data is not known $\mathbf{C_{xl}}$ cannot be found directly and must instead be estimated using the following equations as in the Expectation step of the EM algorithm:

$$\mathbf{C_{xl}} = \mathbf{E}\left[\mathbf{C_{xl}} | \mathbf{C_y}; \widehat{\mathbf{K}}_\mathbf{y}\right]$$

$$= \widehat{\mathbf{K}}_\mathbf{xl} - \widehat{\mathbf{K}}_\mathbf{xl}\left(\widehat{\mathbf{K}}_\mathbf{y}\right)^{-1}\widehat{\mathbf{K}}_\mathbf{xl} + \widehat{\mathbf{K}}_\mathbf{xl}\left(\widehat{\mathbf{K}}_\mathbf{y}\right)^{-1}\mathbf{C_y}\left(\widehat{\mathbf{K}}_\mathbf{y}\right)^{-1}\widehat{\mathbf{K}}_\mathbf{xl}, \quad (25)$$

with:

$$\widehat{\mathbf{K}}_\mathbf{y} = \sum_{\mathbf{l=1}}^{\mathbf{N_a}} \widehat{\mathbf{K}}_\mathbf{xl} \quad (26)$$

$$\widehat{\mathbf{K}}_\mathbf{xl} = \widehat{\gamma}_\mathbf{l}\mathbf{a}\left(\widehat{\theta}_\mathbf{l}\right)\mathbf{a^H}\left(\widehat{\theta}_\mathbf{l}\right) + \widehat{\mathbf{R}}_\mathbf{nl}. \quad (27)$$

It can be seen that this expression requires an estimation of the decomposed noise covariance matrix $\widehat{\mathbf{R}}_{nl}$ and this step is discussed in detail in the following section.

Now, the importance weight for the particle filtering expression in (4) is calculated using $\mathbf{C_{xl}}$ as defined in (25) and $\widehat{\mathbf{K}}_\mathbf{xl}^{-1}$, where $\widehat{\mathbf{K}}_\mathbf{xl}$ is defined in (27).

$$L\left(\mathbf{y_{t|t+N}} | \chi(\mathbf{t})\right) = \exp\left(-\frac{1}{2} tr\left[\mathbf{C_{xl}^p}\widehat{\mathbf{K}}_\mathbf{xl}^{-1}\right]\right). \quad (28)$$

As this expression defines the likelihood at an individual frequency, the overall measurement likelihood is then given by:

$$P\left(\mathbf{z_{t|t+N}}|\chi\left(\mathbf{t}\right)\right)=\prod_{\omega}L\left(\mathbf{y_{t|t+N}}\left(\omega\right)|\chi\left(\mathbf{t}\right)\right). \quad (29)$$

5 Pause Detection Step

While the signals considered here are broadband in nature, not all frequency subbands will contain signal components. Inclusion of subbands containing only noise when calculating the measurement likelihood leads to severe degradation of the tracking results, and the spurious location results found can lead to incorrect association of the results and targets.

This becomes a serious problem when the number of sources changes during the tracking with some sources being inactive for significant periods. In such a situation the particle filtering algorithm proposed in [10] can no longer accurately distinguish between and track the individual targets.

In this paper we therefore propose the inclusion of a pause detection step at each frequency which determines which frequency subbands contain signal components and should therefore be included in the measurement likelihood calculation.

This step is based on the noise characterization method proposed in [19] in which a threshold is applied to each frequency subband in order to distinguish between frequencies containing noise only and those containing signal components.

The noise threshold η is calculated as:

$$\eta\left(\omega,k\right)=\beta E\left(\omega,k-1\right); \quad (30)$$

where k is the block index, (with N frames in a block). $E\left(\omega,k-1\right)$ is the energy of the previous noise at the given frequency ω, and β is a constant value lying between 1.5 and 2.5.

Then, if:

$$E\left(\omega,k\right)>\eta\left(\omega,k\right) \quad (31)$$

the frequency value ω is determined to contain signal components, and is included in (29) in order to find the measurement likelihood.

Meanwhile if $E\left(\omega,k\right)<\eta\left(\omega,k\right)$, the estimated noise spectrum estimate is updated as discussed in section 5.1.

5.1 Noise-plus-Reverberation Covariance Estimation

From equation (23) it can be seen that an estimate of the noise covariance matrix is needed in order to find the measurement likelihood. Previously this was found by making an assumption of white background noise of known power [10].

However, for most practical situations this assumption will not hold leading to degradation in the tracking results obtained, and it is therefore desirable to use a more accurate model of the background noise.

With this aim in mind a second original aspect of this paper is the estimation of the noise covariance for each frequency subband in order to find the EM pseudo-likelihood.

The noise covariance is estimated at each subband when that subband is determined to contain noise only as described in section 5. This covariance estimate is then smoothed over time:

$$\mathbf{R_{nn}}\left(\omega\right)=\frac{\mathbf{1}}{\mathbf{Q}}\sum_{\mathbf{q=1}}^{\mathbf{Q}}\mathbf{R_{nn}}\left(\mathbf{q},\omega\right) \quad (32)$$

where Q is the number of previous values used and depends on the statistics of the background noise.

This method allows for tracking the sources in situations where there is no prior knowledge of the background noise making it much more useful for practical tracking problems.

6 Data Association

The problem of data association - i.e. association of each location estimate with its corresponding source, arises when or more of the sources is inactive.

In this paper we make a decision on which source or sources are active by comparing the measurement likelihoods for each source.

$$P_l\left(\mathbf{z_{t|t+N}}|\chi\left(\mathbf{t}\right)\right)=\prod_{i=1}^{N_p}P_l\left(\mathbf{z_{t|t+N}}|\chi^{\mathbf{i}}\left(\mathbf{t}\right)\right) \quad (33)$$

The active sources are then estimated to be the sources resulting in the N_a largest values of $P_l\left(\mathbf{z_{t|t+N}}|\chi\left(\mathbf{t}\right)\right)$. These active sources are then updated using the measurement likelihood as discussed in section 2.1.

As the measurement likelihood is irrelevant for the inactive sources an average of the propagated particles is taken to estimate the location of these sources.

This method allows for tracking the sources in situations where there is no prior knowledge of the background noise making it much more useful for practical tracking problems, and the improvement in the tracking results due to the increased accuracy of the background noise model can be clearly seen in figure 4.

7 Experimental Setup

The proposed method was then tested using recordings taken in a middle sized meeting room with a reverberation time of $500ms$, using the experimental setup shown in figure 1. Throughout the recordings two sources were present, a human speaking Japanese and a loudspeaker playing music. The sources were placed so that the difference in the respective directions of arrival is 60^o.

The music was played continuously, however the human speaker spoke intermittently throughout the recording taking significant breaks between sentences. It is these pauses that cause difficulties in the tracking as the tracks of both sources can easily be confused when one source is silent.

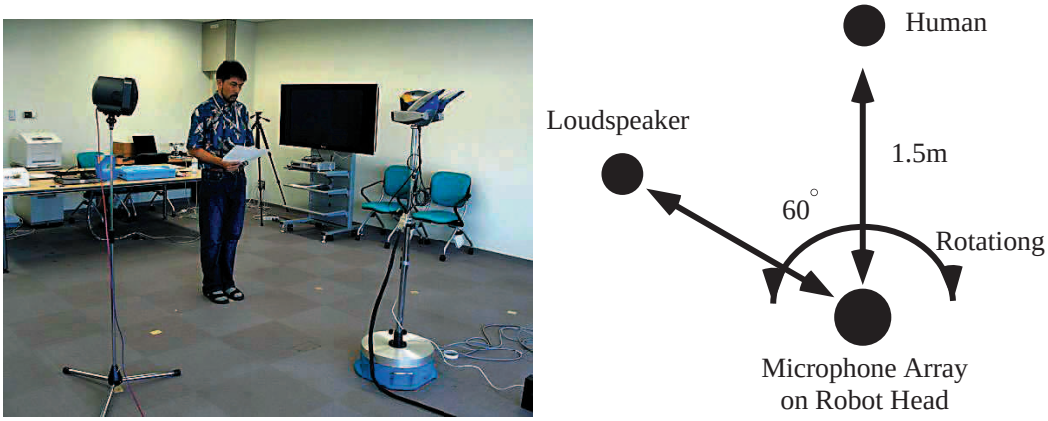


Figure 1: Experimental Setup

Sampling Frequency	16000Hz
FFT Length	512
FFT Shift	128
Frequency Range	800-1600Hz
Block Length	0.1s
Block Overlap	0.05s
Q	10s
β	1.5

Table 1: Experimental Parameters

A circular microphone array of 8 microphones was mounted on the head of a humanoid robot, the HRP-2 developed at AIST Japan, and mounted onto a computer controlled turntable. The turntable was then rotated from $0^o - 180^0$ at a constant speed while the sources remained at the same position. The distance between the sources and the microphones was $1.5m$ allowing for an assumption of far-field sources to be made.

The recorded signals were divided into frames of length $32ms$, with an averaging interval of $N = 9$, or approximately $0.1s$. The direction of the targets was quantized into 1-degree segments, and an estimate of location for both sources is found every $0.1s$. A full list of the experimental parameters used is given in table 1.

In order to evaluate the modifications proposed here the performance of the particle filtering algorithm with no pause detection step and using an assumption of white noise with known power (as proposed in [10]) is compared with the results achieved when the pause detection step is used to determine which frequencies contain speech. In both these cases the speech activity state of each speaker is randomly assigned.

We then compare the methods described in the previous paragraph with the proposed method. The results of the 3 different algorithms are shown in figure 2-4.

8 Results

In 2 it can be clearly seen that in the situation under consideration here the original particle filtering algorithm as proposed in [10] can no longer successfully track the sources. The absence of the human speaker leads to confusion of the separate tracks and there are a large number of spurious location estimates.

The addition of the frequency selection step results in a substantial improvement in the tracking results as seen in 3. The exclusion of those frequency components containing noise only dramatically reduces the number of spurious estimates, thereby allowing the tracks of the individual sources to remain identifiable.

Finally from figures 4 it can be seen that estimation of the noise-plus-reverberation covariance and the speaker activity states directly from the observed data leads to significantly better results than those achieved using the assumption of white noise. In this case the particle filter smoothly follows the true tracks of both sources.

In this case the particle filter can be seen to continue to estimate the location of the human speaker during the periods of silence, and while these results are based only on the propagation model and the previous location of the speaker, they are sufficiently accurate to allow for successful tracking once the speech begins again.

Future work will consider the performance of the algorithm for varying angles between the sources and for a larger number of speakers, with longer non-speech intervals for each speaker.

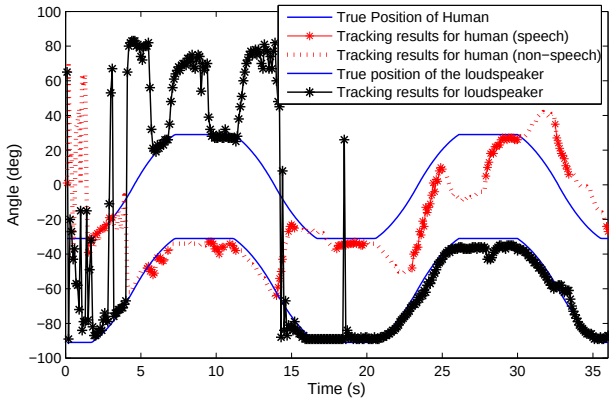


Figure 2: Tracking results using the PFA with no frequency selection.

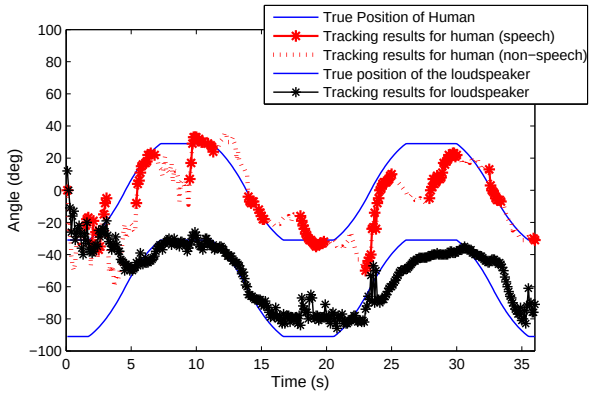


Figure 3: Tracking results using the PFA with frequency selection step.

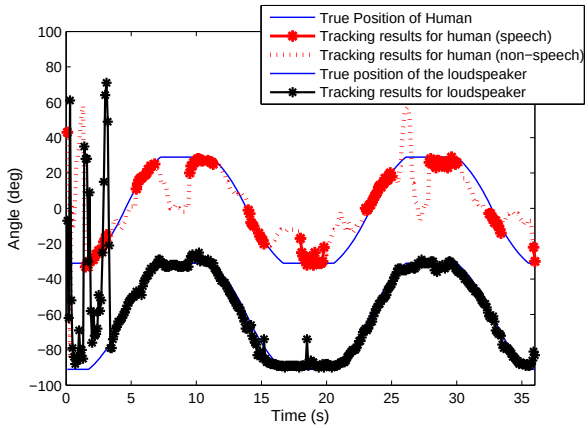


Figure 4: Tracking results using the PFA with frequency selection step and estimated noise.

9 Conclusion

This paper proposes a novel particle filtering scheme for tracking multiple speakers based on the approach proposed in [10]. This method was extended to include a pre-tracking active source number estimation step which is robust to the presence of reverberation. Furthermore a novel method of estimating the noise-plus-reverberation covariance matrix is proposed.

References

[1] B. Ristic, S. Arulampalam, and N. Gordon, in *Beyond the Kalman Filter; Particle Filters for Tracking Applications*. Artech House, 2004, ch. 12.

[2] J. Vermaak and A. Blake, "Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Salt Lake City, UT, 2001.

[3] D. Ward and R. Williamson, "Particle Filter Beamforming for Acoustic Source Localization in a Reverberant Environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Orlando, Fl, 2002.

[4] L. D. Stone, C. Barlow, and T. Corwin, in *Bayesian Multiple Target Tracking*. Artech House, 1999.

[5] C. Hue, J. L. Cadre, and P. Perez, "Sequential Monte Carlo Methods for Multiple Target Tracking and Data Fusion," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 309–325, 2002.

[6] J. Larocque, J. Reilly, and W. Ng, "Particle filters for tracking an unknown number of sources," *IEEE Trans. Signal Processing*, vol. 50, no. 12, pp. 2926–2937, 2002.

[7] D. Avitzour, "A Maximum Likelihood Approach to Data Association," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 2, pp. 560–566, 1992.

[8] K. Molnar and J. Modestino, "Application of the EM Algorithm for the Multitarget/Multisensor Tracking Problem," *IEEE Trans. Signal Processing*, vol. 46, no. 1, pp. 115–119, 1998.

[9] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Montreal, Quebec, 2004.

[10] M. Kawamoto, F. Asano, H. Asoh, and K. Yamamoto, "Particle Filtering Algorithms for Tracking Multiple Sound Sources Using Microphone Arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007.

[11] A. Quinlan, M. Kawamoto, F. Asano, H. Asoh, and K. Yamamoto, "Tracking a Varying Number of Sound Sources using Particle Filtering," in *IASTED Conference on Signal and Image Processing SIP 2007*, Honolulu, Hawaii, 2007.

[12] A. Quinlan and F. Asano, "Detection of Overlapping Speech in Meeting Recordings Using the Modified Exponential Fitting Test," in *Proc. 15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, 2007.

[13] A. Doucet, N. de freitas, and E. N. Gordon, in *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[14] J. Rissanen, "Modelling by Shortest Data Description Length," *Automatica*, vol. 14, pp. 465–471, 1978.

[15] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.

[16] A. Quinlan, F. Boland, J. Barbot, and P. Larzabal, "Determining the Number of Speakers with a Limited Number of Samples," in *European Signal Processing Conference EUSIPCO*, Florence, 2006.

[17] A. Quinlan, J.-P. Barbot, P. Larzabal, and M. Haardt, "Model Order Selection for Short Data: An Exponential Fitting Test (EFT)," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. Article ID 71953, 2007.

[18] J. Grouffaud, P. Larzabal, and H. Clergeot, "Some Properties of Ordered Eigenvalues of a Wishart Matrix: Application in Detection Test and Model Order Selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Atlanta, GA, 1996.

[19] H. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Detroit, MI, 1995.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-5

適応ステップサイズBSSによる音源分離のロボット聴覚への適用

Sound Source Separattion by Adaptive Step-size Contol and its Application to Robot Audition

中島 弘史, 中臺 一博, 長谷川 雄二, 辻野 広司

Hirofumi NAKAJIMA, Kazuhiro NAKADAI, Yuji HASEGAWA, Hiroshi TSUJINO

(株) ホンダ・リサーチ・インスティチュート・ジャパン

Honda Research Institute Japan Co., Ltd.

{nakajima, nakadai, yuji.hasegawa, tsujino}@jp.honda-ri.com

Abstract

This paper describes a method to adaptively control a step-size parameter which is used for updating a separation matrix to extract a target sound source accurately in blind source separation (BSS). The design of the step-size parameter is essential when we apply BSS to real-world applications such as robot audition systems. It is common to use a fixed step-size parameter which was obtained empirically. However, due to environmental changes and noises, the performance of BSS with the fixed step-size parameter deteriorates. We propose a general method that allows adaptive step-size control. The proposed method is applicable to any BSS algorithm. Actually, we applied it to six types of BSS algorithms for an 8 ch microphone array embedded in Honda ASIMO. Experimental results show that the proposed method improves the performance of these six BSS algorithms through experiments of separation and recognition for two simultaneous speeches.

1 はじめに

人と自然にインタラクションが可能なロボットを実現するためには，ロボットの聴覚機能[1]が必要である．実環境では，環境ノイズだけでなく，ロボット自身の動作音など，複数の音源が存在する．また自然なインタラクションでは，ある話者とロボットが話している最中に，別の話者が割り込んで話かける状況（バージイン）も起こるため，音源の数や位置は動的に変化する．このため，動的に変化する複数の音源信号から，各音源の信号を分離する音源分離処理は重要である．ブラインド音源分離 (*Blind Source Separation; BSS*)[2, 3]は，伝達系の情報が未知であっても高精度な分離が可能な分離手法として知られている．BSS では，適応時の収束性能と安定性を決めるステップサイズパラメータが重要である．従来の BSS の多くは，特定の静的な音響環境を仮定して実験的に定めた固定のステップサイズパラメータを利用しているため下記のような問題が起こる

- 雑音や音源数が多い環境に変化した時，ステップサイズが相対的に大きくなり，分離行列が発散する．
- 雑音や音源が少ない環境に変化した時，ステップサイズが相対的に小さくなり，収束速度が低下する．

またステップサイズパラメータは，本来周波数に依存して最適値が変化するが，従来法では周波数に関係なく一定値を用いているため，下記の問題も発生する．

- ステップサイズは発散しやすい周波数帯域においても安定動作を得るために小さな値を設定する必要があり，他の周波数帯域においては収束速度が低下する．
- 雑音や音源の周波数特性が変化した時，ステップサイズが相対的に大きくなり，分離行列が発散する．

本稿では，このような問題を解決するため，環境の変動に対応が可能で，どの周波数帯域においても最適なステップサイズを与えることが可能な適応ステップサイズ法を提案し，実際のロボット聴覚へ適応し提案法の有効性を確認する．

2 適応ステップサイズ法

2.1 BSS の定式化

図 1 に一般的な BSS の構成図を示す．M 個の音源と，N ($\leq M$) 個のマイクロホンがあるとする．周波数 ω での音源のスペクトル（ベクトル）を $\mathbf{s}(\omega)$, $[s_1(\omega)s_2(\omega)...s_M(\omega)]^T$，同様にマイクロホンの入力信号スペクトルを $\mathbf{x}(\omega)$,

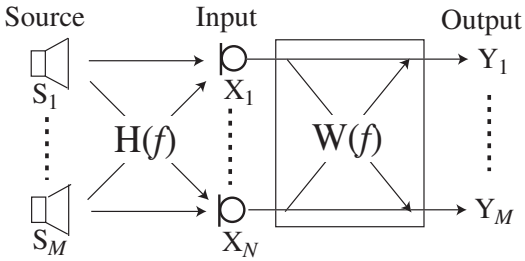


Figure 1: System Model for Blind Source Separation

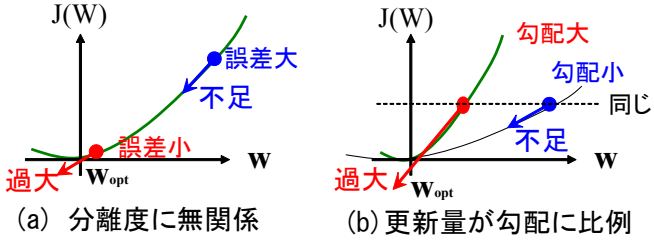


Figure 2: Problems of conventional methods

$[x_1(\omega)x_2(\omega)...x_N(\omega)]^T$ とすれば，$\mathbf{x}(\omega)$ は次式で計算できる．

$$\mathbf{x}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega), \tag{1}$$

ここで $\mathbf{H}(\omega)$ は伝達関数行列であり，$\mathbf{H}(\omega)$ の各要素 H_{ji} は，i 番目の音源から j 番目のマイクロホンまでの伝達関数を示している．音源分離処理は，次式で定式化できる．

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega), \tag{2}$$

ここで $\mathbf{W}(\omega)$ は分離行列である．音源分離処理は，出力信号 $\mathbf{y}(\omega)$ が音源信号 $\mathbf{s}(\omega)$ と一致するような $\mathbf{W}(\omega)$ を見つける問題となる．もし $\mathbf{H}(\omega)$ が事前に正確に得られていれば，$\mathbf{W}(\omega)$ は擬似逆行列 $\mathbf{H}^+(\omega)$ によって簡単に推定できる．しかし実際には，$\mathbf{H}(\omega)$ を正確に得ることは困難である．BSS はこの問題を解決する手法であり，$\mathbf{H}(\omega)$ が未知である場合や，$\mathbf{H}(\omega)$ の一部のみ（直接音成分など）が既知である場合も利用できる．BSS は，$\mathbf{y}$ の混合度を示すコスト関数 $J(\mathbf{y})$ を最小化する $\mathbf{W}_{opt}$ を推定する問題として，次式で書ける．

$$\mathbf{W}_{opt} = \underset{\mathbf{W}}{argmin}[J(\mathbf{y})] = \underset{\mathbf{W}}{argmin}[J(\mathbf{W}\mathbf{x})]. \tag{3}$$

多くの BSS では，次式を用いて適応的に $\mathbf{W}_{opt}$ を求める．

$$\mathbf{W}_{t+1} \quad = \quad \mathbf{W}_t - \mu\mathbf{J}'(\mathbf{W_t}). \tag{4}$$

ここで $\mathbf{W}_t$ は，時刻 t での $\mathbf{W}$ の値, $\mathbf{J}'(\mathbf{W})$ は，$\mathbf{W}$ の更新方向，μ はステップサイズパラメータを示す．従来の BSS では，このステップサイズとして，周波数に依存しない，固定値を利用する．しかし最適なステップサイズは，分離度や周波数によって異なるため，収束速度と分離精度を両立する事が困難である．速度と精度の両立が困難であるこ

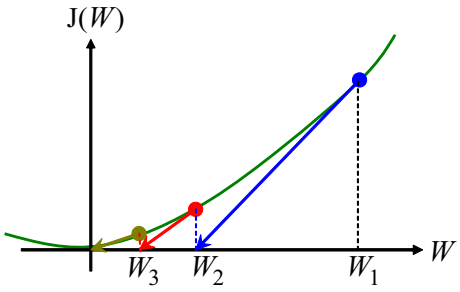


Figure 3: Newton's method

とを，式 (4) と図 2 を用いて説明する．式 (4) から，μ が固定値の場合，分離行列 $\mathbf{W}$ の更新量は，(a) コスト関数 $J(\mathbf{W})$ 自体には依存せず，(b) コスト関数の勾配 $\mathbf{J}'(\mathbf{W})$ のみに比例する．図 2(a) は，ステップサイズが $J(\mathbf{W})$ によらない事の問題点を示している．分離度が低く $J(\mathbf{W}_t)$ が大きい時は，最適な分離行列 $\mathbf{W}_{opt}$ と現在の分離行列 $\mathbf{W}_t$ の差は大きいため，早く収束させるためにはステップサイズを大きくとり，分離行列の更新量を大きくする必要がある．しかし逆に分離度が高く $J(\mathbf{W}_t)$ が小さい時は，最適値 $\mathbf{W}_{opt}$ と現在値 $\mathbf{W}_t$ の差は小さいため，更新量を小さくしなければ，$\mathbf{W}_{opt}$ へ精度良く $\mathbf{W}$ を収束させることができない．しかし従来法では，ステップサイズは一定値であるため，収束速度と収束精度を両立させることが困難である．図 2(b) は，ステップサイズを固定値とした場合，コスト関数が勾配 $\mathbf{J}'(\mathbf{W}_t)$ に比例する事による問題点を示している．勾配が急な時（$\mathbf{J}'(\mathbf{W}_t)$ が大きい時）は，僅かに分離行列を更新しただけでコスト関数が大きく変化する．そのため，ステップサイズを小さくして，$\mathbf{W}$ の更新量を押さえないと，$\mathbf{W}_{opt}$ へ精度良く $\mathbf{W}$ を収束させることができない．逆に勾配が緩やかな時（$\mathbf{J}'(\mathbf{W}_t)$ が小さい時）は，分離行列を大きく更新しないと，コスト関数が変化しない．そのため，早く収束させるためには，ステップサイズを大きくする必要がある．しかし (a) の場合と同様に，ステップサイズが一定値の場合，これらを両立することはできない．これら 2 点が，従来のステップサイズを固定した方法が原理的に収束速度と精度を両立することが難しい理由である．

2.2 適応ステップサイズ法の定式化

本節では，一般の BSS に対する適応ステップサイズ法の定式化を行う．ステップサイズを適応的に定める手法は，エコーキャンセラの分野では一般的[4]であるが，BSS への適用は難しく，その報告例は少ない．これは，N-LMS 法などの一般的なエコーキャンセラが単一チャンネルの実数信号を扱うのに対し，BSS は一般的に多チャンネルの複素信号であるため，最適値の算出が困難であるためである．本稿では，複素勾配[5]に基づく行列を用いた線形近似式と多次元のニュートン法に基づく最適値を利用す

ることにより，一般的な BSS で利用できる適応ステップサイズ法を提案する．ニュートン法は，図 3 に示す通り，現在の近似解による関数値とその勾配から，対象の関数を直線で近似し，近似した直線の解を，次の近似解とする方法である．ニュートン法による解の更新量は，誤差の大きさに比例し，誤差の勾配に半比例する．これらの性質から，前節の図 2 で述べた問題を原理的に解決できることがわかる．複素勾配理論[5]によれば，$J(\mathbf{W}_t)$ の周辺における $J(\mathbf{W})$ は，次式で近似できる．

$$J(\mathbf{W}) \approx \tilde{J}(\mathbf{W}) = J(\mathbf{W}_t) + 2\mathrm{MA}(\nabla_{w*}J(\mathbf{W}), \mathbf{W} - \mathbf{W}_t) \tag{5}$$

ここで $\mathrm{MA}(\mathbf{A}, \mathbf{B}) = Re[\sum_{i,j} a^*_{i,j} b_{i,j}]$ であり，行列 $\mathbf{A}^*$ と $\mathbf{B}$ の各要素の積の和の実部を示す，また ∇_{w*} は複素勾配演算[5]である．ニュートン法による最適なステップサイズ μ_{opt} は，$\tilde{J}(\mathbf{W}) = 0$ とすることで得られる．式 (4) と式 (5) より μ_{opt} は，

$$\mu_{\mathrm{opt}} = \frac{J(\mathbf{W}_t)}{2\mathrm{MA}(\nabla_{w*}J(\mathbf{W}_t), \mathbf{J}'(\mathbf{W}_t))} \tag{6}$$

である．式 (6) は，分子にコスト関数 $J(\mathbf{W}_t)$，分母にその勾配 $\nabla_{w*}J(\mathbf{W}_t)$ を含み，先ほど述べた性質を有することがわかる．式 (6) は一般的な適応処理におけるステップサイズの最適値を示しており，目的とする BSS のコスト関数 $J(\mathbf{W})$ に置き換えることで，様々な種類の BSS に対して利用できる．$\mathbf{J}'(\mathbf{W}) = \nabla_{w*}J(\mathbf{W})$ が成り立つ場合，μ_{opt} は更に次式で簡略化できる．

$$\mu_{opt} = \frac{J(\mathbf{W}_t)}{2\|\mathbf{J}'(\mathbf{W}_t)\|^2}, \tag{7}$$

ここで $\|\cdot\|^2$ は行列のフロベニウスノルムを示している．本手法を用いることで，収束の初期段階や変動が起こった直後など，混合度が高い時には大きい値のステップサイズとなり収束速度が速くなる．また，十分収束して混合度が低くなった時には，小さなステップサイズとなり，高精度で安定した分離が実現できる．

3 適応ステップサイズ BSS

本章では，適応ステップサイズ法を実際の BSS に適用した際の具体的な計算式について記述する．BSS 法は，無相関化に基づく分離 (*Decorrelation based Source Separation; DSS*)，独立成分分析 *Independent Component Analysis; ICA*)，幾何制約付きの分離 (*Geometric-constrained Source Separation; GSS*)，幾何制約付きの ICA(*Geometric-constrained ICA; GICA*)，高次の無相関化に基づく分離 (*High-order DSS; HDSS*)，幾何制約付きの HDSS(*Geometric-constrained HDSS; GHDSS*) の 6 種類とした．適応ステップサイズ BSS の基本式は，式 (2)〜(4) に示した通りであり，実際に各 BSS に適応した結果は，$J(\mathbf{W})$ と $\mathbf{J}'(\mathbf{W})$ の定義が異なるだけである．そこで，以降の各節では，式 (7) の中で定義される μ_{opt} がどのように計算されるかについて取り扱う．

3.1 無相関化に基づく分離 (DSS)

DSS のコスト関数は，次式で与えられる．

$$\begin{aligned} J_{DSS}(\mathbf{W}) &= \|E[\mathbf{E}]\|^2 \\ \mathbf{E} &= \mathbf{y}\mathbf{y}^H - \mathrm{diag}[\mathbf{y}\mathbf{y}^H], \end{aligned} \tag{8}$$

ここで $E[\cdot]$ は期待値演算を示す．更新方向 $\mathbf{J}'(\mathbf{W})$ は，$\nabla_{w*}J(\mathbf{W})$ を計算し，$E[\cdot]$ を無くすことにより得られ，

$$\mathbf{J'}_{DSS}(\mathbf{W}) = 2\mathbf{E}\mathbf{W}\mathbf{x}\mathbf{x}^H \tag{9}$$

となる．従って最適なステップサイズは，次式となる．

$$\mu_{opt_{DSS}} = \frac{\|\mathbf{E}\|^2}{2\|2\mathbf{E}\mathbf{W}_t\mathbf{x}\mathbf{x}^H\|^2} \tag{10}$$

3.2 独立成分分析 (ICA)

一般的に利用されている Kullback-Liebler 情報量および自然勾配を利用した ICA[2]では，コスト関数 $J(\mathbf{W})$ とその勾配 $\mathbf{J}'(\mathbf{W})$ として次式を利用する．

$$J_{ICA}(\mathbf{W}) = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{q(\mathbf{y})} d\mathbf{y}, \tag{11}$$

$$\mathbf{J'}_{ICA}(\mathbf{W}) = \mathbf{E}_\phi \mathbf{W}, \tag{12}$$

$$\mathbf{E}_\phi = \phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H],$$

ここで，$p(\mathbf{y})$ は $\mathbf{y}$ の同時確率密度分布，$q(\mathbf{y})$ は $\mathbf{y}$ の同時確率密度分布であり，$\prod_k p(y_k)$ で計算できる．$\phi(\mathbf{y})$ は，下記で定義される関数である．

$$\begin{aligned} \phi(\mathbf{y}) &= [\phi(y_1), \phi(y_2), \cdots, \phi(y_N)]^T \\ \phi(y_i) &= -\frac{\partial}{\partial y_i} \log p(y_i). \end{aligned} \tag{13}$$

一般的に利用される $\phi(y_i)$ は，いくつかの種類があるが，本報告では，ハイバボリックタンジェントを利用した関数[10]を用いた．

$$\phi(y_i) = \tanh(\eta|y_i|)e^{j\cdot\theta(y_i)}, \tag{14}$$

ここで η はスケーリングパラメータである．

J_{ICA} を実際に計算することは困難であるため，J_{ICA} の変わりに $\|\mathbf{E}_\phi\|^2$ を用て，最適なステップサイズを求めた．この結果，最適なステップサイズは，次式で計算できる．

$$\begin{aligned} \mu_{opt_{ICA}} &= \frac{\|\mathbf{E}_\phi\|^2}{2\mathrm{MA}(\mathbf{E}_\phi\mathbf{W}_t, 2\mathbf{E}\tilde{\phi}(\mathbf{y})\mathbf{x}^H)} \\ \tilde{\phi}(\mathbf{y}) &= [\tilde{\phi}(y_1), \tilde{\phi}(y_2), ..., \tilde{\phi}(y_N)]^T \\ \tilde{\phi}(y_i) &= \phi(y_i) + y_i \frac{\partial \phi(y_i)}{\partial y_i} \end{aligned} \tag{15}$$

3.3 幾何制約付き音源分離 (GSS)

GSS は，ICA で問題となる周波数帯域毎の音源の入れ替わり問題 (パーミュテーション) や，音源毎のゲインが一定とならない問題 (スケーリング) を，マイクロホンと音源の位置から計算される幾何的な制約を付加することで解決できる手法である．GSS は，実環境処理に適しており，ロボット聴覚で実際に利用されている[9]．GSS の $J(\mathbf{W})$ は，式 (8) のコスト関数 $J_{DSS}(\mathbf{W})$ と幾何制約のコスト関数 $J_{GC}(\mathbf{W})$ の合成として与えられる．

$$J_{GSS}(\mathbf{W}) \quad = \quad J_{DSS}(\mathbf{W}) + \lambda J_{GC}(\mathbf{W}) \qquad (16)$$

ここで λ は幾何制約の強さを示す重み係数である．$J_{GC}(\mathbf{W})$ はいくつかのバリエーションがあるが，遅延和ビームフォーミングに基づくコスト関数 (文献[7]では C1) では，次式で計算できる．

$$\begin{aligned} J_{GC}(\mathbf{W}) &= \|\mathbf{E}_{GC}\|^2 \qquad (17)\\ \mathbf{E}_{GC} &= diag[\mathbf{WD} - \mathbf{I}] \end{aligned}$$

ここで $\mathbf{D}$ は，音源からの直接音成分に対する伝達関数行列である．$\mathbf{J}'(\mathbf{W})$ は，

$$\begin{aligned} \mathbf{J'}_{GSS}(\mathbf{W}) &= \mathbf{J'}_{DSS}(\mathbf{W}) + \lambda \mathbf{J'}_{GC}(\mathbf{W}) \qquad (18)\\ \mathbf{J'}_{GC}(\mathbf{W}) &= \mathbf{E}_{GC}\mathbf{D}^H. \end{aligned}$$

となる．GSS の更新式は，

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu_{DSS}\mathbf{J'}_{DSS}(\mathbf{W_t}) - \mu_{GC}\mathbf{J'}_{GC}(\mathbf{W_t}). \qquad (19)$$

固定ステップサイズを利用した場合，μ_{GC} は次式で計算できる．

$$\mu_{GC} = \lambda \cdot \mu_{DSS}. \qquad (20)$$

提案法では，μ_{DSS} と μ_{GC} の両方のステップサイズを最適に定める．μ_{DSS} の最適値は，式 (10) に示した通りであり，μ_{GC} の最適値は，下記で計算できる．

$$\mu_{opt_{GC}} = \frac{\|\mathbf{E}_{GC}\|^2}{2\|2\mathbf{E}_{GC}\mathbf{D}^H\|^2} \qquad (21)$$

3.4 幾何制約付き ICA (GICA)

GICA は，ICA のアルゴリズムに幾何制約を加えた手法である．この手法は，GSS に対するコスト関数式 (16) の $J_{DSS}(\mathbf{W})$ を $J_{ICA}(\mathbf{W})$ に置き換えれば実現できる．

$$J_{GICA}(\mathbf{W}) \quad = \quad J_{ICA}(\mathbf{W}) + \lambda J_{GC}(\mathbf{W}) \qquad (22)$$

GICA の最適なステップサイズは，式 (15) および (21) で計算できる．GICA は文献[8]でも報告されているが，本手法とは幾何制約の誤差が許容される点で異なる．実際の応用では，幾何制約の元となる座標情報には誤差が含まれているため，提案法の方が実用的には適している．

Table 1: J, $\mathbf{J}'$ and μ_{opt} for each BSS algorithm

	J	$\mathbf{J}'$	μ_{opt}
DSS	$\|E[\mathbf{E}]\|^2$	$2\mathbf{E}\mathbf{W}_t\mathbf{x}\mathbf{x}^H$	$\frac{\|\mathbf{E}\|^2}{2\|2\mathbf{E}\mathbf{W}_t\mathbf{x}\mathbf{x}^H\|^2}$
ICA	Eq. 11	$\mathbf{E}_\phi\mathbf{W}_t$	$\frac{\|\mathbf{E}_\phi\|^2}{2\mathrm{MA}(\mathbf{E}_\phi\mathbf{W}_t, 2\mathbf{E}\tilde{\phi}(\mathbf{y})\mathbf{x}^H)}$
HDSS	$\|E[\mathbf{E}_\phi]\|^2$	$2\mathbf{E}_\phi\tilde{\phi}(\mathbf{y})\mathbf{x}^H$	$\frac{\|\mathbf{E}_\phi\|^2}{2\|2\mathbf{E}_\phi\tilde{\phi}(\mathbf{y})\mathbf{x}^H\|^2}$
G***	$\|\mathbf{E}_{GC}\|^2$	$\mathbf{E}_{GC}\mathbf{D}^H$	$\frac{\|\mathbf{E}_{GC}\|^2}{2\|2\mathbf{E}_{GC}\mathbf{D}^H\|^2}$

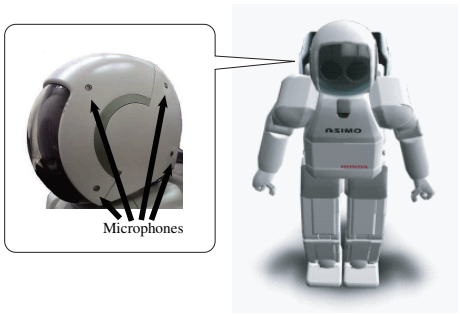


Figure 4: ASIMO with 8 microphones

3.5 高次無相関化 SS (HDSS)

高次の無相関化規範を利用した HDSS における $J(\mathbf{W})$ および $\mathbf{J}'(\mathbf{W})$ は，次式の通りである．

$$\begin{aligned} J_{HDSS}(\mathbf{W}) &= \|E[\mathbf{E}_\phi]\|^2 \qquad (23)\\ \mathbf{J'}_{HDSS}(\mathbf{W}) &= 2\mathbf{E}_\phi\tilde{\phi}(\mathbf{y})\mathbf{x}^H \qquad (24) \end{aligned}$$

また最適ステップサイズは，

$$\mu_{opt_{\mathrm{HDSS}}} = \frac{\|\mathbf{E}_\phi\|^2}{2\|2\mathbf{E}_\phi\tilde{\phi}(\mathbf{y})\mathbf{x}^H\|^2}. \qquad (25)$$

で計算できる．

3.6 幾何制約付き高次無相関化 SS (GHDSS)

GHDSS は，HDSS に幾何制約を加えた手法であり，そのコスト関数 $J_{GHDSS}(\mathbf{W})$ は

$$J_{GHDSS}(\mathbf{W}) \quad = \quad J_{HDSS}(\mathbf{W}) + \lambda J_{GC}(\mathbf{W}). \qquad (26)$$

で与えられる．最適ステップサイズは，式 (25) および (21) の通りである．

以上，各 BSS についてのコスト関数 J，更新方向 $\mathbf{J}'$ および最適ステップサイズ μ_{opt} についてまとめたものを表 1 に表記する．

4 評価

提案法の効果を示すため，6 種類の BSS アルゴリズムに対し，提案法を利用した場合としなかった場合での性能を比較した．図 4 に，評価で利用したマイクロホンアレイを

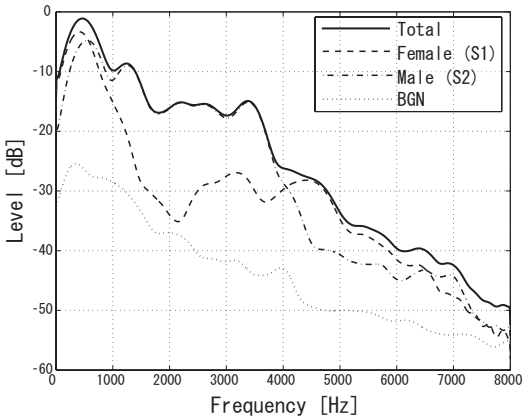


Figure 5: spectrum of each input signal

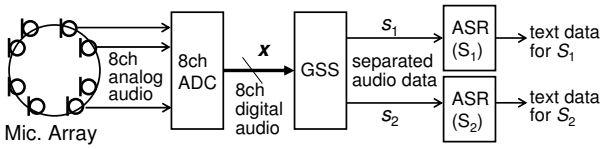


Figure 6: System con guration for WCR evaluation

示す．アレイは，Honda ASIMO の頭に左右対称に４つづつマイクロホンを埋め込んだ，計８素子のものである．始めにアレイを利用して，ロボットの動作ノイズを含む暗騒音および，スピーカ (GENELEC 1029A) を利用して音源からアレイまでの室内伝達関数のインパルス応答を測定した．部屋は，幅 4m, 奥行 7m, 高さ 7m の大きさであり，残響時間 (RT_{20}) は 0.3 から 0.4 秒である．音源は，アレイの中心から距離 1.5 mの正面 (S_1) および右 90°(S_2) の位置にあるものとした．音源信号は，S_1 は女声，S_2 は男声である．マイクロホンへの入力データは，音源信号とインパルス応答を畳み込み，暗騒音を付加して合成した．図 5 に，畳み込み後の各音源信号，暗騒音および最終的に合成した信号のスペクトルを示す．図より，各音源信号のレベルはほぼ同等であり，暗騒音は音声信号に対し 10〜20dB 低い．表 2 に，BSS の設定条件を示す．従来法の固定のステップサイズ値 μ は，0.1, 0.01, 0.001 の３種類とした．分離行列の初期値は，直接音成分の伝達関数行列の複素共役転置行列で与えた．

幾何制約の無い BSS で問題となる音源の入れ替わり問題は初期値によって解決されるものとして付加的な処理は行わず，周波数帯域毎の振幅の不定性問題は分離行列のノルムを正規化する事で解決した．また従来法で利用する幾何制約の重み係数 λ は，文献[9]に従って，$\|\mathbf{y}\mathbf{y}^H\|^{-2}$ と定めた．また 6 種類の BSS 以外に，ベースラインの性能を評価するために，マイクロホンの入力信号をそのまま利用した場合と，遅延和ビームフォーミングにより強調した信号を利用した場合もあわせて評価した．評価指標は，S/N 比 (SNR), 平均相関係数 (CC)，音声認識率 (WCR)

Table 2: BSS Setting

sampling frequency	16 kHz
window function	Hanning
window length	512 (32 ms)
shift length	256 (16 ms)
scaling parameter η	1

の３種類とした．全ての手法に対し，10 秒の音声を用いて SNR と CC の算出を行った．WCR については，SNR が最も良い GSS についてのみ実験を行った．この場合のシステム構成を図 6 に示す．SNR は次式で計算した．

$$SNR = 10\log_{10}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{|y|^2}{|\hat{n}|^2}\right], \tag{27}$$

ここで y は分離した出力信号，$\hat{n}$ は y に含まれるノイズ成分であり，$\hat{n} = y - \hat{s}$ で計算した．$\hat{s}$ は，単一の音声信号 S_i のみが存在する場合の出力信号である．CC は，時間周波数領域で，相関係数を平均する事で計算した．

$$\begin{aligned} CC \ \ [\mathrm{dB}] &= 10\log_{10} E_\omega[CC_\omega(\omega)], \\ CC_\omega(\omega) &= \frac{|E_t[|y_1^*(\omega,t)y_2(\omega,t)|]}{\sqrt{E_t[|y_1(\omega,t)|^2]}\cdot\sqrt{E_t[|y_2(\omega,t)|^2]}} \end{aligned} \tag{28}$$

ここで，$E_\omega[\cdot]$ および $E_t[\cdot]$ はそれぞれ，周波数および時間領域での平均を示す．また $y_i(\omega,t)$ は，時刻 t，周波数 ω における i 番目の出力信号である．CC は，2 つの音源の相関に基づく分離度を示しており，完全に分離された場合，$-\infty$ dB となる指標である．WCR は，音声認識エンジン Julian[11]を利用した孤立単語認識の正解率から算出した．言語モデルはネットワーク文法，音響モデルは残響や雑音のないクリーンな音声を用いて学習したもの用いた．単語セットは，216 語の ATR 音素バランス単語セットを用いた．

4.1 結果

図 7〜9 に，分離した音声に対する SNR, CC および WCR を示す．図 7 から，提案法 (AS) は GSS, DSS, HDSS および GHDSS の４つの手法に対して，SNR が向上することがわかる．しかし ICA および GICA では，SNR の向上は見られなかった．分離音を試聴した所，分離が困難な低周波の雑音が強調されており，これが SNR 悪化の原因と推定される．しかし音声は分離されており，音声帯域では正常に動作することがわかる．図 8 は，平均相関係数 (CC) での評価値を示している．図より，ICA および GICA を含め，AS が全ての手法に対して相関係数が低く，高精度に分離が可能であることがわかる．図 9 は，音声認識率 (WCR) を示している．提案法は，音声認識率が最も高く，実際にロボット聴覚へ応用する際の有効性が確認できる．

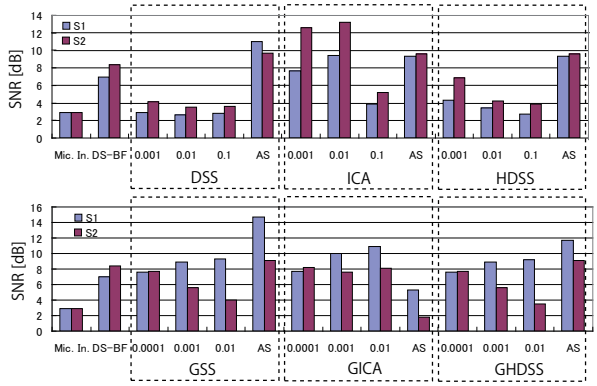


Figure 7: Improvement in SNR for two simultaneous speeches

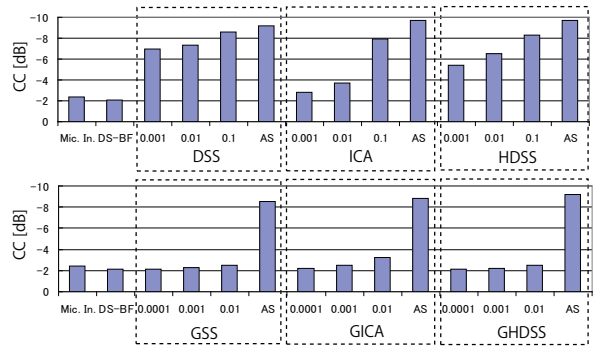


Figure 8: Correlation Coe cient (CC)

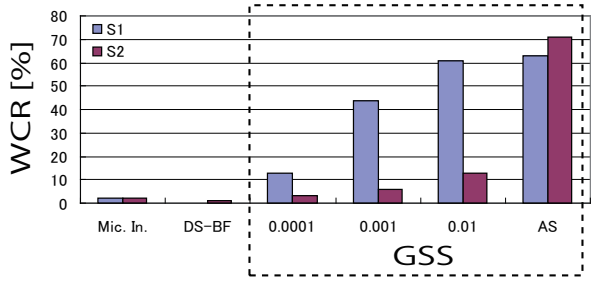


Figure 9: Improvement in WCR of separated speech

5 まとめと課題

本稿では，実環境での音源分離性能の向上を目指して適応ステップサイズ法を提案した．提案法は，ステップサイズを多次元のニュートン法に基づき最適値に適応制御するもので，さまざまな種類の BSS へ利用できる．本稿では，提案法を用いた 6 種類の BSS について定式化した．ロボットの頭部に埋め込んだ 8ch のマイクロホンアレイを用いた 2 音源の分離実験では，全ての BSS において相関値が低下し分離度が高い事が確認できた．また多くの手法で SNR の向上し，特に GSS でその効果が顕著であった．GSS については，音声認識システムによる評価も行い，音声認識率も向上することが確認できた．動的な実環境での評価，リアルタイムのロボット聴覚システムへの実装などが今後の課題である．

参考文献

[1] K. Nakadai, *et. al.*, “Active audition for humanoid,” in *17th National Conf. on Artificial Intelligence (AAAI2000)*. AAAI, 2000, pp. 832–839.

[2] S. Ikeda and N. Murata, “A method of ica in time-frequency domain,” *Workshop Indep. Compom. Anal. Signal.*, pp. 365–370, 1999.

[3] H. Saruwatari, *et. al.*, “Blind source separation based on a fast-convergence algorithm combining ica and beamforming,” *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[4] S. Yamamoto and S. Kitayama, “An adaptive echo canceller with variable step gain method,” *Trans. of the IECE of Japan*, vol. E 65, no. 1, pp. 1–8, 1982.

[5] B.A D.H. Brandwood, “A complex gradient operator and its application in adaptive array theory,” *IEE Proc.*, vol. 130, no. 1, pp. 251–276, 1983.

[6] S. Amari, “Natural gradient works e eciently in learning,” *Neural Compt.*, vol. 10, pp. 251–276, 1998.

[7] L. Parra and C. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[8] M. Knaak, *et. al.*, “Geometrically constrained independent component analysis,” *IEEE Trans. on Speech and Audio Processing*, vol. 15, no. 2, pp. 715–726, 2007.

[9] J. Valin, *et. al.*, “Enhanced robot audition based on microphone array source separation with post- lter,” in *2004 International Conference on Intelligent Robots and Systems (IROS2004)*. IEEE/RSJ, 2004, pp. 2123–2128.

[10] H. Sawada, *et. al.*, “Polar coordinate based nonlinear function for frequency-domain blind source separation,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2002, pp. 1001–1004.

[11] A. Lee, *et. al.*, “Julius - an open source real-time large vocabulary recognition engine,” in *7th European Conf. on Speech Communication and Technology*, 2001, vol. 3, pp. 1691–1694.

Semi-blind cancellation of robot internal noise for hands-free speech recognition

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate school of information science
Nara Institute of science and technology
Ikoma, Nara, Japan
even@is.naist.jp

Abstract

Hands-free speech recognition is the most natural interface for human/robot interaction. Recently, several methods based on frequency domain blind signal separation were proposed in order to recover the user's voice. However, one specific problem that arises in the human/robot hands-free speech recognition is that the robot itself creates some noises. In this paper, to resolve this problem, we propose a new frequency domain semi-blind source separation approach that exploits some additional sensors placed inside the robot. Some experimental results shows that the proposed method is able to incorporate the additional information efficiently and that the performance is improved in term of SNR and word accuracy in a speech recognition task.

1 Introduction

Nowadays, communicating with machines is usually not natural and requires some adaptation or training. In order to improve the usability of these machines and reduce the burden for the users it is important to recreate the natural human communication interface: Speech. The most difficult task being to give machines the ability to listen. Speech recognition is working well if we use a microphone close to the user's mouth but this is not a natural interface and not a convenient one in many situations. For these reasons, the focus is now on hands-free speech recognition. In hands-free speech recognition, the user's voice is picked at distance by a microphone array making a more natural interface with the machine. However, the cost is that noise and reverberation deteriorate the received speech quality. Hence it is necessary to improve the quality of the received speech before speech recognition is performed.

In order to deal with the noise, blind signal separation (BSS) based techniques is a strong candidate for processing the multidimensional observation given by microphone arrays (see review paper [1]). The goal of BSS is to separate the observed signals in its different components. Ideally, receiving the user's speech contaminated with noise, we would recover the speech and the noise separately. The frequency domain approach, referred to as FD-BSS, is especially of great interest since the convolutive mixture modeling the reverberant environment can be efficiently processed in the frequency domain. However, this is still a challenging task in a real environment where the number of interfering noise signals is large and the amount of data is limited.

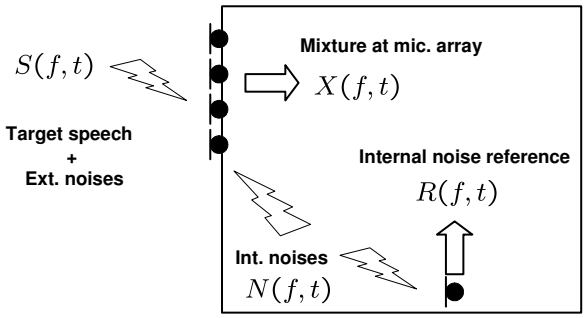


Figure 1: General situation.

In this paper, we consider the special case of the human/robot interaction. The specificity of this case is that the robot is assumed to have some moving parts that create noises. In the remainder, we referred to the noises created by the robot itself as *internal noises* and to the noises comming from outside the robot as *external noises*. The main idea is that it is possible to obtain some information on the internal noises by placing additional sensors inside the robot. To exploit this additional information, we replace the FD-BSS approach by a semi-blind signal separation method that operates in the frequency domain (FD-SBSS). Figure 1 illustrates the situation. The hands-free speech recognition system uses a microphone array that picks the user's speech and noises. The noises are composed of the external noises and the internal noises. The additional sensor inside the robot gives a reference of the internal noises. In the remainder of the paper, after briefly presenting coven-

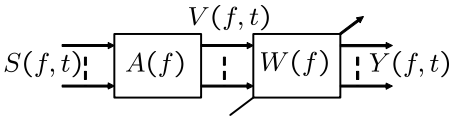


Figure 2: Mixture and blind separation at frequency bin f.

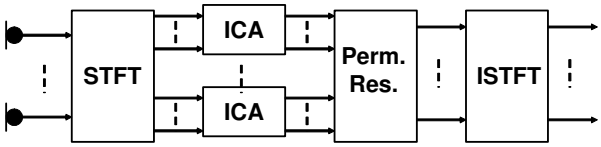


Figure 3: Frequency domain blind signal separation overview.

tional FDBSS and the proposed FD-SBSS method, their performance is compared in a realistic environment.

2 Frequency domain blind signal separation

In acoustic, the observed signals received by a microphone array in a reverberant environment are convolutive mixtures of some signals emitted from different locations. The goal of BSS is to recover the emitted signals knowing only the observed mixtures. In the frequency domain approach to BSS, a short time Fourier transform STFT is applied to the observed signals to get the frequency domain observations. $V(f,t) = [v_1(f,t),\ldots,v_n(f,t)]^T$ denotes the observation at the fth frequency bin where t is the frame index. The observed signal at the fth frequency bin is

$$V(f,t) = A(f)S(f,t) \tag{1}$$

where the $n \times n$ matrix $A(f)$ represents the mixture, $S(f,t) = [s_1(f,t),\ldots,s_n(f,t)]^T$ is the emitted signal. Consequently when using a F points analysis frame for the STFT the convolutive mixture is replaced by F instantaneous mixtures and the goal is to estimate the components of the emitted signals $S(f,t)$ in each frequency bin.

In the fth frequency bin, the estimate is obtained by applying an unmixing matrices $W(f)$ to the observed signals (see Fig. 2)

$$Y(f,t) = W(f)V(f,t) = W(f)A(f)S(f,t). \tag{2}$$

A usual assumption in BSS is that the components of $S(f,t)$ are statistically independent in each frequency bin. Then the component of $Y(f,t)$ are statistically independent if and only if $W(f)$ is such that

$$Y(f,t) \quad = \quad P(f)\Lambda(f)S(f,t)$$

where $P(f)$ is a $n \times n$ permutation matrix and $\Lambda(f)$ is a diagonal $n \times n$ matrix [2].

As a consequence, in each frequency bin, it is possible to recover the components of $S(f,t)$ up to scale and permutation indeterminacy by finding the unmixing matrix $W(f)$ that gives an estimate with statistically independent components. This problem is often referred to as independent component analysis (ICA). To complete the separation, it is necessary to match the components belonging to the same signal across all the frequency bins before applying the inverse STFT otherwise the time domain signals are still mixtures of the desired signals.

Figure 3 gives an overview of the FD-BSS approach. The resolution of the permutation indeterminacy and scale factor is done in the permutation resolution block (perm. res.).

Since our proposed semi-blind method is derived from the iterative INFOMAX method [3], we briefly present this method (see review [1] for reference to other methods). In the frequency bin f at the kth iteration, the separation equation is

$$Y^{(k)}(f,t) = W^{(k)}(f)V(f,t) \tag{3}$$

The mutual information of $Y^{(k)}(f,t)$ is iteratively minimized by updating the matrix $W^{(k)}(f)$ with the following rule (the frequency and frame indexes were dropped due to space limitation)

$$W^{(k+1)} = W^{(k)} + \mu(I - < \Phi(Y^{(k)})Y^{(k)H} >_t)W^{(k)} \tag{4}$$

where $< \cdot >_t$ denotes frame averaging and $\Phi(\cdot)$ denotes the vector of score functions. For $Y = [y_1,\ldots,y_p]^T$ the score function is defined by

$$\begin{aligned}\Phi(Y) &= [-\frac{\partial}{\partial y_1}\log P_{y_1}(y_1),\ldots,-\frac{\partial}{\partial y_1}\log P_{y_1}(y_1)]^T \\ &= [\phi_1(y_1),\ldots,\phi_p(y_p)]^T\end{aligned}$$

where $P_{y_i}(y_i)$ is the probability density function of y_i. In practice the score functions are unknown and should be estimated from the data or prior knowledge on the signal densities is available.

3 PROPOSED METHOD

3.1 Block structure

The goal of the proposed semi-blind approach is also to recover some unknown signals when only some mixtures of these signals are available. However, contrary to the fully blind separation case, we are also given an additional information about the observed mixtures. We know that the mixing process has the following block structure

$$\begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix} = \begin{bmatrix} A(f) & B(f) \\ 0 & C(f) \end{bmatrix} \begin{bmatrix} S(f,t) \\ N(f,t) \end{bmatrix}. \tag{5}$$

The observed signals and the sources are both partitioned in two vectors. The first components of the observation $X(f,t)$, of size $(p \times T)$ with T the number of frame, is a mixture of both of the source vectors $S(f,t)$ $(p \times T)$ and $N(f,t)$ $(q \times T)$ whereas the second component of the observation $R(f,t)$ $(q \times T)$ is only a function of $N(f,t)$. This structure corresponds to the situation described in Fig. 1. Figure 4 gives an overview of the

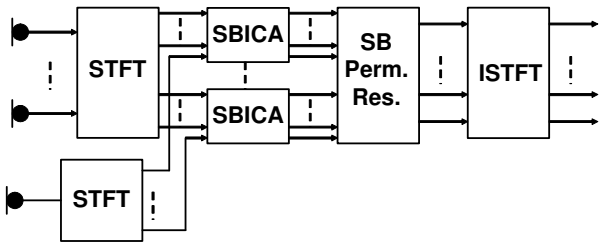


Figure 4: Frequency domain semi-blind signal separation overview.

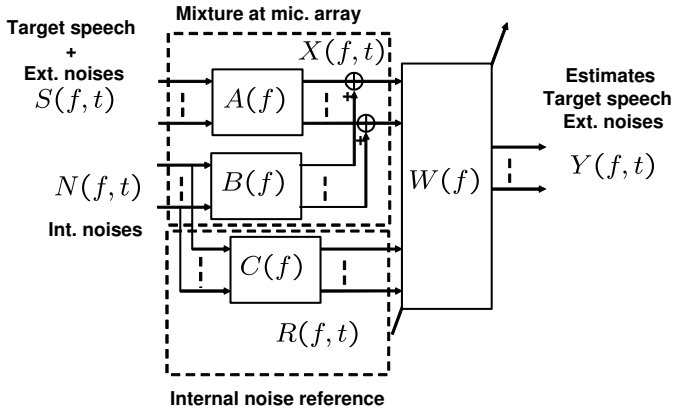


Figure 5: Block structure of the mixture.

semi-blind approach (in the case of one additional sensor). If compared to Fig. 3, we can see that the ICA and permutation resolution blocks are replaced by their semi-blind counterparts.

In the following we use the terms reference for $R(f,t)$ and observation for $X(f,t)$. A diagram of the mixing is given in Fig. 5. The proposed demixer has a block structure of compatible dimensions with the matrices $A(f)$, $B(f)$ and $C(f)$.

$$\begin{bmatrix} Y(f,t) \\ Q(f,t) \end{bmatrix} = \begin{bmatrix} W_1(f) & W_2(f) \\ 0 & W_3(f) \end{bmatrix} \begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix}.$$

Compared to the blind problem of same dimension, the number of coefficients to update is reduced.

Using the results in [2] presented in Sect.2, the components of $Y(f,t)$ and $Q(f,t)$ are all statistically independent if and only if the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$ are such that

$$\begin{bmatrix} W_1(f) & W_2(f) \\ 0 & W_3(f) \end{bmatrix} \begin{bmatrix} A(f) & B(f) \\ 0 & C(f) \end{bmatrix} = \begin{bmatrix} P_1(f)\Lambda_1(f) & 0 \\ 0 & P_2(f)\Lambda_2(f) \end{bmatrix}$$

where $P_1(f)(p \times p)$ and $P_2(f)$ $(q \times q)$ are permutation matrices and $\Lambda_1(f)(p \times p)$ and $\Lambda_2(f)$ $(q \times q)$ are diagonal matrices. Consequently it is possible to estimate the components of $S(f,t)$ and $N(f,t)$ by updating $W_1(f)$, $W_2(f)$ and $W_3(f)$ until the components of $Y(f,t)$ and $Q(f,t)$ are all statistically independent.

3.2 Proposed algorithm

The proposed method semi-blind separation method uses the mutual information of $Y(f,t)$ and $Q(f,t)$ to measure the statistical independence of their components. The criterion is optimized by an iterative gradient descent on the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$. At iteration k, we have the following unmixing system

$$\begin{bmatrix} Y^{(k)}(f,t) \\ Q^{(k)}(f,t) \end{bmatrix} = \begin{bmatrix} W_1^{(k)}(f) & W_2^{(k)}(f) \\ 0 & W_3^{(k)}(f) \end{bmatrix} \begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix}.$$

To obtain the update rules for these matrices we rewrite the update rule in the blind case eq.(4) with the proposed demixer structure

$$\begin{aligned} \begin{bmatrix} W_1^{(k+1)}(f) & W_2^{(k+1)}(f) \\ 0 & W_3^{(k+1)}(f) \end{bmatrix} = & \begin{bmatrix} W_1^{(k+1)}(f) & W_2^{(k+1)}(f) \\ 0 & W_3^{(k+1)}(f) \end{bmatrix} \\ -\mu \left(I_{p+q} - < \begin{bmatrix} \Phi(Y^{(k)}(f,t)) \\ \Phi(Q^{(k)}(f,t)) \end{bmatrix} \begin{bmatrix} Y^{(k)}(f,t) \\ Q^{(k)}(f,t) \end{bmatrix}^H >_t \right) & \\ \times \begin{bmatrix} W_1^{(k+1)}(f) & W_2^{(k+1)}(f) \\ 0 & W_3^{(k+1)}(f) \end{bmatrix}. & \end{aligned}$$

Then the update rule for the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$ are extracted (A semi-blind method for instantaneous mixtures in the time domain uses the same approach to get the update rule in [4]). The update rules for the matrices have the following form

$$W_j^{(k+1)}(f) = W_j^{(k)}(f) + \mu \Delta W_j^{(k)}(f)$$

where (dropping the frequency and frame indexes for $Y(f,t)$ and $Q(f,t)$)

$$\begin{aligned} \Delta W_1^{(k)}(f) &= \left(I - < \Phi(Y^{(k)}) Y^{(k)H} >_t \right) W_1^{(k)}(f) \\ \Delta W_2^{(k)}(f) &= \left(I - < \Phi(Y^{(k)}) Y^{(k)H} >_t \right) W_2^{(k)}(f) \\ &\quad - \left(< \Phi(Y^{(k)}) Q^{(k)H} >_t \right) W_3^{(k)}(f) \\ \Delta W_3^{(k)}(f) &= \left(I - < \Phi(Q^{(k)}) Q^{(k)H} >_t \right) W_3^{(k)}(f). \end{aligned}$$

The frequency domain signals are approximately circular because they were obtained by a STFT. For a circular random variable $y = |y| \mathbf{e}^{\mathbf{j} \arg(\mathbf{y})}$ we have (see [6] for details)

$$\phi(y) = \phi(|y|) \mathbf{e}^{\mathbf{j} \arg(\mathbf{y})}.$$

Here, we use a kernel based approach to estimate the score function of the modulus [7, 8].

After the semi-blind separation is performed in all the frequency bins, the permutation resolution is also simplified because of the block structure.

4 EXPERIMENTAL RESULTS

To demonstrate the importance of the internal noise reference we performed some experiments mixing the noise

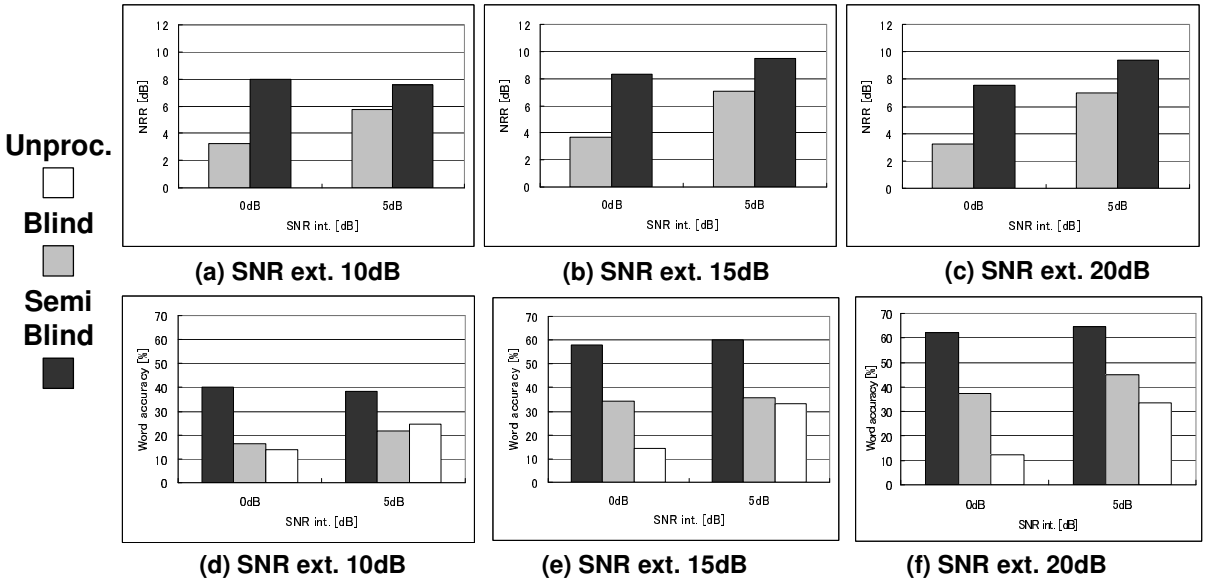


Figure 6: Noise rate reduction (NRR) and word accuracy for different SNRs.

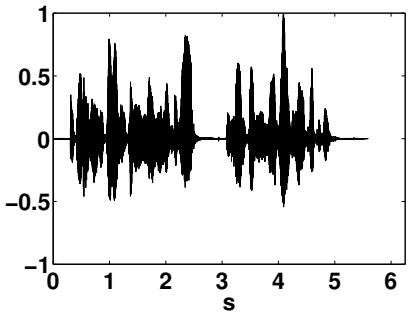


Figure 7: Speech signal at first microphone (SNR ext. 20dB SNR int. 0dB).

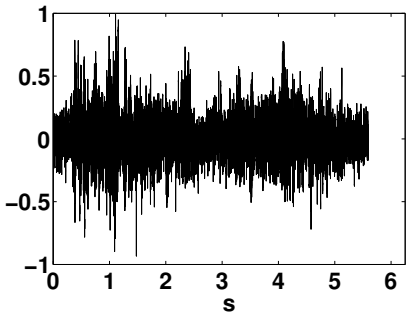


Figure 8: Observed signal at first microphone (SNR ext. 20dB SNR int. 0dB).

recorded in a train station as external noise and a synthetic non stationary noise as internal noise. The impulse response of the train station hall was also measured for a speaker at 50cm in front of a four microphone array (inter mic. spacing is 2.15cm), see Fig. 9. 200 Japanese sentences of different length were used as speech signals (duration 1s to 14s at 16kHz from the JNAS database). The observed signal is obtained in two

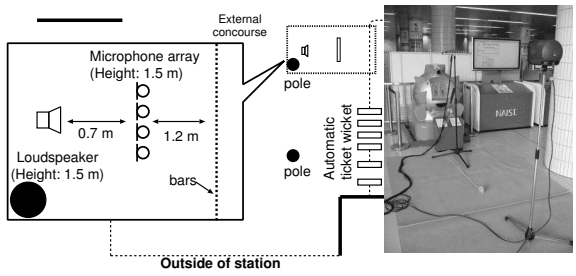


Figure 9: Recording conditions.

steps. First a speech signal convoluted by the impulse response is mixed with the recorded noise (the external moise). The SNR in this mixture is called SNR ext. Then the mixed speech and external noise is mixed with the internal noise that is filtered by a low pass filter. The SNR for this second mixture is SNR int. We also filter the internal noise to obtain the reference. The low pass filter and the filter used for the reference represent the propagation of the internal noises from their sources to the microphone array and to the additional sensor.

Figure 7 shows the speech signal as it is received at the first microphone before any noise is added. The corresponding observed signal obtained after addition of external (SNR ext. 20dB) and internal (SNR int. 0dB) noises is showed in Fig. 8.

In all experiments we compared the INFOMAX approach (blind) to the proposed approach (semi-blind). The STFT is performed with a 512 points hanning window with 256 points overlap. The separation matrices are initialized to identity in all frequency bins then 200 iterations are performed with an adaptation step $\mu = 0.1$. The speech signal is selected out of the separated components in all the frequency bins using the same method for both approaches.

The INFOMAX method consider the reference signal as a fifth observation. Then both algorithms have the

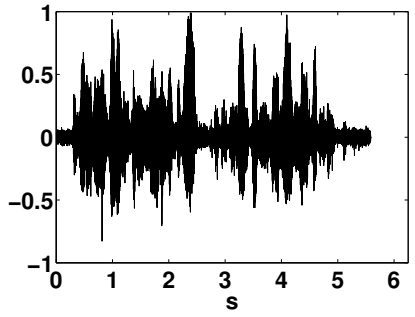


Figure 10: Recovered speech at first microphone with blind approach (SNR ext. 20dB SNR int. 0dB).

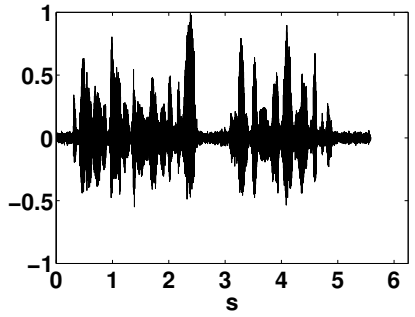


Figure 11: Recovered speech at first microphone with proposed semi-blind approach (SNR ext. 20dB SNR int. 0dB).

same amount of statistical information. Thus the only difference is that the semi-blind approach knows that the mixture has the block structure showed in Fig. 5.

The estimation quality is measure in term of noise reduction rate (NRR) defined as the difference of the SNR for the speech estimates (after processing) and the SNR for the observations (before processing). Consequently, a positive NRR means that the speech estimate quality is improved. Figures 6(a), (b) and (c) show the NRR for different mixtures (averaged on the 200 test signals). The second measure of performance is the word accuracy for a continuous speech recognition task. The speech recognition conditions are in given in table 1 and the results in Figs. 6(d), (e) and (f).

The blind method is able to improve the speech signal but using the block structure gives the advantage to the semi-blind method when a limited number of iteration is done. The performance of the blind method would increase if the number of iteration is larger but in a real situation computation time is limited (The performance difference is also larger for the shorter sentences).

Figures 10 and 11 show the recovered speech corresponding to the observation in Fig. 8. We can see that the proposed method results in cleaner speech as it is closer to the signal in Fig. 7.

Table 1: Condition for speech recognition

Task	20k word newspaper dictation
Acoustic model	phonetic tied mixture, clean model [5]
Acoustic model training	260 speakers (150 sentences/speaker)
Decoder	JULIUS ver 3.2 [5]

5 conclusion

In this paper we proposed a semi-blind separation approach that operates in the frequency domain. The method easily incorporates the information given by additional sensors to the FD-BSS based approach. Experiments showed that this is can be very beneficial in a hands-free speech recognition scenario for treating internal noises generated by a moving robot. Our future work is to apply the proposed method on real data recorded inside a robot (i.e. replacing the synthetic data used for the internal noises by real measurements).

References

[1] M.S. Pedersen et al., "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Communication*, 2007.

[2] P. Comon, "Independent component analysis, a new concept ?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[3] A. J. Bell and T. J. Sejnowski, "An information maximimization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[4] M. Joho, H. Mathis, and G. S. Moschitz, "Combined blind/nonblind source separation based on the natural gradient," *IEEE Signal Processing Letters*, vol. 8, no. 8, pp. 236–238, 2001.

[5] A. Lee et al., "Julius - an open source real-time large vocabulary recognition engine," *EUROSPEECH,*, pp. 1691–1694, 2001.

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino. Polar coordinate based nonlinear function for frequency-domain blind source separation. *IEICE Trans. Fundamentals*, E86-A(3):590–596, 2003.

[7] N. Vlassis and Y. Motomura. Efficient source adaptivity in independent component analysis. *IEEE Trans. Neural Networks*, 12(3):559–566, 2001.

[8] B.W. Silverman. Kernel density estimation using the fast fourier transform. *J. Roy. Statist. Soc. Ser. C: Appl. Statist.*, 31(1):93–99, 1982.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-7

Real-time Dereverberation Using Late Components of Impulse Response for Hands-Free Speech Recognition

Randy Gomez, Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science
Nara Institute of Science and Technology , JAPAN
randy-g@is.naist.jp

Abstract

We proposed a robust dereverberation technique based on multi-band spectral subtraction through the e ective use of the late component of the the room impulse response. The approximate late reverberant power spectrum used in the multi-band spectral subtraction is derived by ltering the observed reverberant signal by the late component of the impluse response. To compensate for the approximation error, multi-band coe cients are computed offline based on the minimum squared error criterion and introduced in the actual spectral subtraction. In this work, we perform recognition experiments using realistic impulse response containing utter echo and succesfully improved the recognition performance of the system. Moreover, the proposed method is robust to di erent microphone distances and recognition experiments are performed in a realistic reverberant environment condition.

1 INTRODUCTION

Hands-free speech recognition is common to robot applications where robots can move freely without the constraint of microphone wiring. However, the observed speech signal at the microphone is smeared by a phenomenon known as reverberation. Recognition performance for reverberated speech signi cantly decreases compared to the performance obtained for clean speech. The room impulse response gives a good insight of the reverberation and is often used to experimentally create a reverberated speech. The early part of the impulse response contains the direct speech sound and the early re ections and these vary in proportion to the distance between the microphone and the speaker. The late part however, tends to vary less with the distance since it can be viewed as a collective overlapping of re ections from di erent directions. Model-based speech recognition system like the Hidden Markov Models (HMM) approach can handle the e ects of the early reverberation part through adaptation [1]. There are several techniques being implemented to address the problem caused by reverberation. Although microphone array processing is a popular candidate [2][3], the method presented here is a single channel approach. But the generalization to multiple channels is straightforward. A single channel framework was proposed in [4]. The authors argued that the late reverberation part of the observed reverberant signal should be suppressed through signal processing whereas the early reverberant part, more likely to vary with talker-microphone distance, can be handled by the HMM based recognition system.

A novel approach based on this framework is proposed in [5]. This approach employs a numerical criterion based on minimum squared error through multi-step Linear Prediction Coe cients (LPC) to e ectively estimate the late reverberant component and makes use of spectral subtraction to remove it. Although [5] works well in estimating the late reverberant component for spectral subtraction, this approach requires the complete reverberant utterance for processing. Thus, realtime speech recognition is di cult to realize.

In our proposed method, the primary goal is to perform real-time speech recognition in realistic reverberant environment conditions. As in [5], we devise an approach to estimate the late reverberant component and suppress its e ect through spectral subtraction. But, in our case, there is no need to wait for the whole reverberant utterance before performing spectral subtraction. The spectral subtraction is conducted on a frame-wise basis. In order to achieve this purpose, some assumptions are done to simplify the online computation load. The main assumption is that the late reverberant components can be estimated by ltering the observed reverberant signal in a frame-wise manner using the late component of the impulse response. But this assumption results in a non-negligible estimation error. Thus we propose to correct this error in the spectral subtraction step by using a multi-band spectral subtraction with appropriate band coe cients. Moreover, the multi-band coe cients used to correct the approximation error are estimated offline

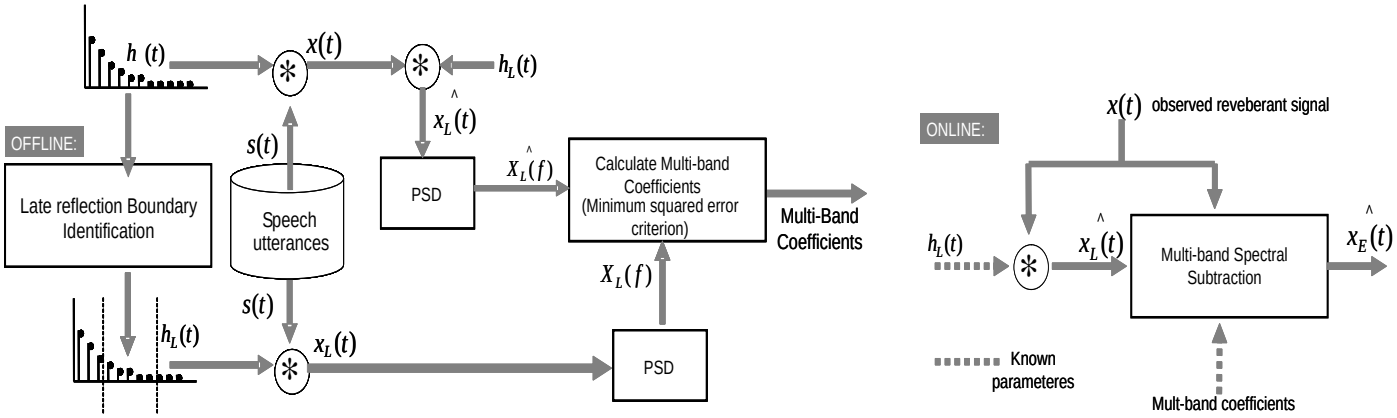


Figure 1: Overall Block Diagram of the Proposed Dereverberation Method.

by training on a small speech database. Then, in the online processing, all we need is just a simple ltering for estimating the late reverberant part and the multi-band spectral subtraction handles the correction of the estimate resulting in a fast dereverberation method.

In section 2, we will brie y discuss the basic concept of the proposed method with the detailed explanation of the multi-band spectral subtraction. Experimental evaluation is discussed in Section 3 and in Section 4, we conclude this paper. Throughout the paper, t denotes discrete time. The short time Fourier transform of a signal $x(t)$ is denoted $X(f, \tau)$ where f is the frequency channel index and τ the frame index. The actual observed reverberant signal and the synthetically reverberated signal using the impulse response are both referred to as $x(t)$.

2 Proposed Method

In [5], multi-step LPC is used to estimate the late reverberant components of the reverberant signal. Then the power spectrum of this component is suppressed by spectral subtraction to obtain the dereverberated signal. Although the proposed method works in the same principle, the approach di ers in the process of estimating the late reverberant component of the signal. In order to achieve real-time processing, the proposed estimate of the late reverberant component is based on the two following assumptions:

(a) Filtering the observed reverberated signal with the late impulse response gives an estimate of the late reverberant component.

(b) The late part of the impulse response used in assumption (a) should be robust with the variation of mic-speaker distance.

Although the prior knowledge of the impulse response is an advantage, without access to the clean utterance we cannot use this knowledge directly to estimate the late reverberant component. But the estimate of the late reverberant component can be obtained in real-time by ltering with late impulse response component using assumption (a). In addition, we propose to replace the single channel spectral subtraction in [5] by a speci - cally trained multi-band spectral subtraction (see Section 2.3). The purpose of the multi-band spectral subtraction is to correct the estimation error to render assumption (a) realistic. Thus it is necessary to measure the impulse response and determine the late part (see Section 2.1).

Assumption (b) is based on the fact that the late part of the impulse response does not vary as much as the earlier part for di erent speaker-microphone distances. Namely, if we measure the impulse response from a far speaker-microphone distance then we can use it to obtain the late reverberant component for closer distances. This is actually an issue of the robustness of the proposed approach given the same environment with varying distances. Later in Section 3 we will show that our approach is indeed robust.

The use of the prior information of the impulse response should not pose any concern because the real issue is whether, in the same reverberant environment, the late part of the impulse response varies signi cantly with the change in the speaker-microphone distance, as this would preclude the use of only a single impulse response. It should be noted that it is usually possible to measure the impulse response of the room beforehand. Moreover, it is relatively common to use such measures for simulating reverberant condition during development of a system. In this sense, the proposed dereverberation implementation just exploits the available prior information (i.e. impulse response). The proposed dereverberation scheme shown in Figure 1 is primarily composed of the following steps:

offline:

(1) The identi cation of the late components of the room impulse response $h_L(t)$, (2) The calculation of multi-band coe cients based on the minimum square error criterion using the power spectral densities (PSD) which will be discussed further in Section 2.3 used to correct the approximation error in assumption (a). The results from this stage are xed and used in the online portion.

online:

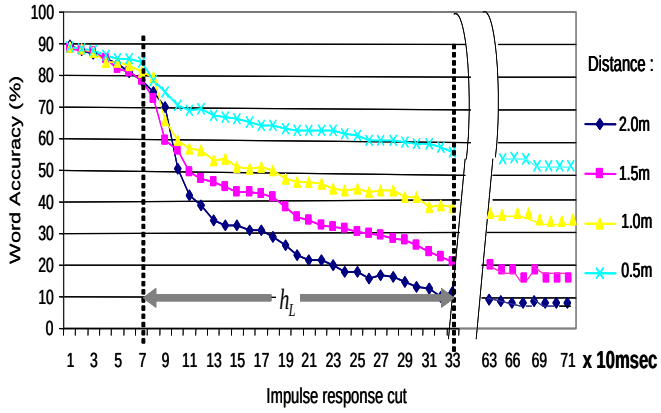


Figure 2: Late Re ection Boundary Identi cation.

(1) Realtime dereverberation is achieved by ltering the observed signal x with the late component of the impulse response h_L to approximate the late reverberant $\hat{x}_L$ signal of the observed reverberant signal x, (2) spectral subtraction is conducted using the pre-calculated and xed multi-band coe cients and the power spectrum of both x and $\hat{x}_L$ to recover $\hat{x}_E$ (reverberant signal minus the e ect of late reverberation) and (3) speech recognition is performed.

2.1 Identifying the Coefficients of the Late Components of The Room Impulse Response

For the proposed method to work, it is imperative to identify the boundary between the early components of the room impulse response and the late component, where the latter contributes to the late reverberation (see the rst block of the offline part of Figure 1). To determine this boundary, we control the length of impulse response and perform recognition using a clean model. In Figure 2, the horizontal axis is the length of impulse response and the vertical axis shows the recognition performance. It is obvious that the steep decrease in the performance starts at 70 ms which suggests the beginning of the e ect of the late reverberation. Consequently, the proposed dereverberation method has to suppress the reverberation e ect of the impulse response part after 70ms because the HMM based speech recognition system cannot handle the e ect of reverberation from this point onwards. The end-boundary of the late-part of the impulse response can be set to the taps of the lter where performance saturates as shown in the gure.

2.2 Estimating The Power Spectrum Attributed to the Late Reverberant Component

In a real-time speech recognition task, the estimate of the late reverberant component must be available on a frame by frame basis. Namely, contrary to the method in [5], it is not possible to estimate it from the whole utterance. Moreover, the necessary processing should be reduced to the minimum. To achieve this goal, the proposed method relies on several assumptions that reduce the online computation. Moreover we use the assumption that $h_E(\tau)$ and $h_L(\tau)$ can be treated independently as stressed in [4][5].

Considering an impulse response $h(\tau)$ measured for a speaker far from the microphone, the early and late parts of the impulse response, respectively $h_E(\tau)$ and $h_L(\tau)$, are determined as in Section 2.1. For a given utterance $s(t)$, the microphone receives the reverberant signal $x(t) = h*s(t)$. The assumption (a) is that $h_L*x(t)$ is a crude estimate of the late reverberant component $h_L * s(t)$ (Section 2.3 shows how to correct this error). Thus the estimate of the late reverberant component is simply obtained by ltering the received signal $x(t)$ by the measured $h_L(t)$, (see online part in Figure 1). Then its power spectrum is computed on a frame by frame basis by using short time Fourier transform. A few frames can be averaged in order to correct the estimation error of the power spectrum without a noticeable delay.

2.3 Multi-band Spectral Subtraction

In order to attenuate the e ect of the estimation error due to Section 2.2, multi-band spectral subtraction similar to that in [6] is used instead of conventional spectral subtraction [7]. Contrary to the work in [6], the coef- cients of the multi-band spectral subtraction are optimized for minimizing the estimation error.

In multi-band spectral subtraction, the frequency is partitioned in K bands $B_1, \ldots, B_K$ and the spectral subtraction is conducted in each bands. The additional ex- ibility over conventional spectral subtraction is the possibility to modulate the amount of subtracted signal in each band by the band coe cients $\delta_1, \ldots, \delta_K$. With the previous notation, we have

$$|\hat{X}_E(f,\tau)| = \begin{cases} |X(f,\tau)|^\gamma - \delta_k|\hat{X}_L(f,\tau)|^\gamma \\ \text{if } |X(f,\tau)|^\gamma - \delta_k|\hat{X}_L(f,\tau)|^\gamma > 0 \\ \beta|\hat{X}_L(f,\tau)|^\gamma \\ \text{else} \end{cases} \quad (1)$$

for $f \in B_k$ with β the ooring coe cient and γ the power exponent as in conventional spectral subtraction.

In [6], some values for the bands $B_1, \ldots, B_K$ and the coe cients $\delta_1, \ldots, \delta_K$ are empirically determined. But we use the multi-band spectral subtraction to correct the estimation error for the late reverberation component. Consequently, the bands and the coe cients are to be determined in order to minimize this estimation error.

This was achieved by training with a speech database, see the offline portion of Figure 1. For each clean signal $s(t)$ of the database, the real late reverberant component $x_L(t) = h_L * s(t)$ and the estimated late reverberant component $\hat{x}_L(t) = h_L * h * s(t)$ are computed using the late part of the impulse response determined in Section 2.1. Then the power spectral densities (PSD) $X_L(f)$ and $\hat{X}_L(f)$ of both signals are estimated using Welch's method. The window type, overlap and length of the frame are the same as the one used in the multi-band spectral subtraction. Figure 3 shows the PSDs of both signals. Then for a given set of bands $B_1, \ldots, B_K$, the

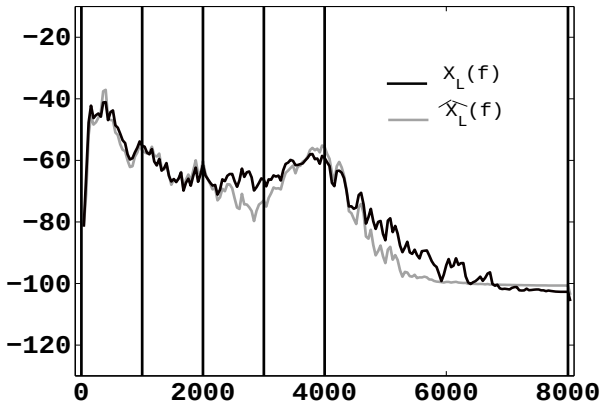


Figure 3: Power spectral densities of the real late reverberant component $X_L(f)$ and estimated late reverberant component $\hat{X}_L(f)$.

coe cients $\delta_1, \ldots, \delta_K$ are determined by minimizing the squared error

$$E_k = \sum_{f \in B_k} |X_L(f) - \delta_k \hat{X}_L(f)|^2 \qquad (2)$$

in each band.

The frequency range $[0, 8]$ kHz was rst partitioned in 8 bands of 1 kHz then the upper bands were merged resulting in the partition represented by the vertical lines in Figure 3. The merging of the upper bands was performed because partitioning the frequency range $[4, 8]$ kHz did not change the performance signi cantly. The bands coe cients δ_k are $\{0.68, 0.91, 1.17, 0.81, 0.71\}$.

2.4 Advantage of The Proposed Method

Unlike that of [5], the proposed method does not need to wait for the whole observed reverberant utterance in order to estimate the late reverberant component. Instead, we make use of the room impulse response (i.e. the $h_L(t)$ as discussed in Section 2.1) and the observed frames of the reverberant utterance to approximate the late reverberant power spectrum. This paves the way to a fast dereverberation for real-time speech recognition application.

3 EXPERIMENTAL RESULTS

3.1 Experimental Conditions

We will discuss the conditions of the experimental set-up. First, we elaborate on the reverberant environment conditions. This includes the room impulse response measurement and synthetically create reverberant database. Moreover, we will show the parameters used in the recognizer.

3.1.1 Reverberant Environment

To simulate a reverberant condition, we use the Time Stretched Pulse (TSP) method to obtain the actual room impulse response [8]. This method re ects e ectively the actual impulse response of the room. In this experiment we use a single channel directional microphone with distances from source of 0.5m, 1.0m, 1.5m, and 2.0m. Figure 4 shows the actual impulse response obtained using TSP with 550ms and 400ms reverberation time and with microphone distance of 2.0m and 1.0m respectively. In the right-most part of this gure, we can see the impulse response having some signi cant utter echos while in the left-most part of the gure we see a smooth impulse response with a less signi cant utter echo. The presence of the utter echo is a characteristic of a realistic reverberant environment often discarded in arti cially generated impulse response. By using movable panels and changed some settings depicted in Figure 5 we were able to generate these two sets of impulse responses (400ms and 550ms) using the TSP method. Reverberant signals are obtained using 6000-tap lter (corresponding to the end-boundary in Section 2.1).

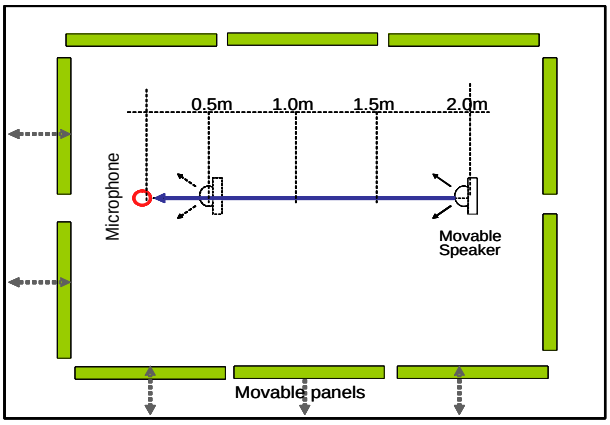


Figure 5: An Illustration of Acquiring the Room Impulse Response using TSP

Table 1: System speci cations

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior by JNAS
Test data	Adult and Senior by JNAS

3.1.2 Speech Recognition

In the recognition experiment we use JULIUS [9] as the recognizer using phonetically tied mixture (PTM) [10] model with 20K-word Japanese dictation task. The database comes from JNAS [11] with a combined 561 speakers (male and female). The open test set constitutes 44 (male and female) speakers with a combined 200 utterances. Summary of the conditions used in recognition is given in Table 1.

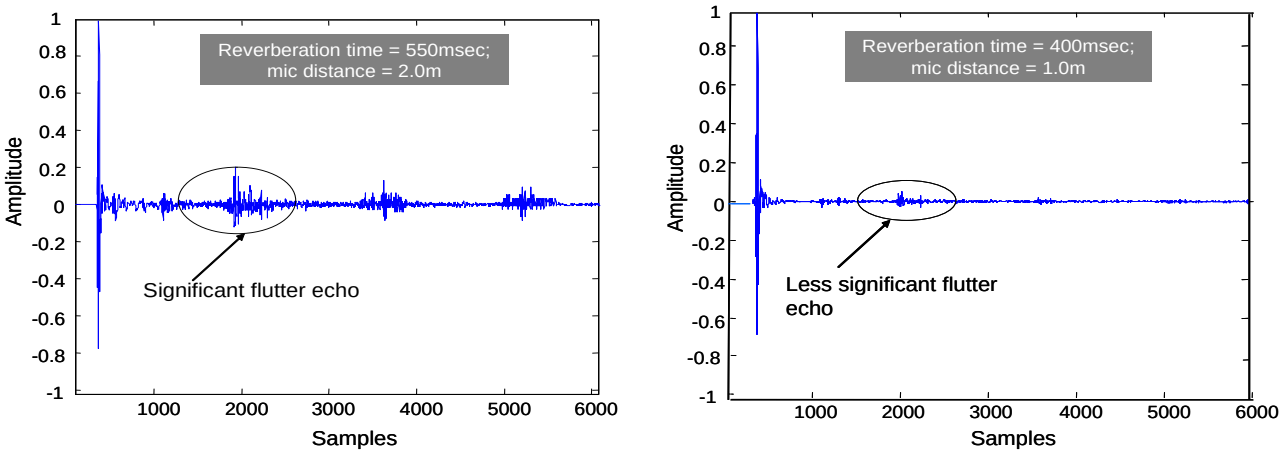


Figure 4: Room Impulse Response Obtained by Using TSP with 400ms and 550ms.

3.2 Recognition Performance

In this section, we will show the recognition experiment results for the proposed dereverberation method and compare to baseline results. There are two baseline results in our experiment and we use each of these to compare wherever appropriate. The rst baseline (1) is when using a clean model without any dereverberation and using a reverberated test data without dereverberation. The second one, baseline (2) is when using a reverberated model and a reverberated test data without any dereverberation at all (reverberated matched condition). Moreover, as stated in Section 3.1.1 we perform recognition experiments in a realistic environment that includes the e ects of utter echo. This explains the lower recognition performance baseline as opposed to arti cially generated room impulse response.

3.2.1 Basic Results

In Figure 6 we show the basic recognition results of the proposed method as opposed to the baseline. At the bottom of the gure we show the conditions of the experiment. "Clean" refers to the non reverberated model/test, while "Reverb" refers to reverberated model/test but without dereverberation, "multi-LPC" refers to a reverberated model/test processed (dereverberated) with multi-LPC, while "Proposed" refers to a reverberated model/test processed (dereverberated) with the proposed method. In this gure, it is apparent that the proposed method performs far better than when using the baseline (1) "clean model" without dereverberation and the baseline (2) "reverberated matched condition" without dereverberation. Moreover, the proposed method has an increased in performance compared to the multi-LPC approach and this is attributed to the fact that we have prior information of the impulse response and make use of it. These results a rm the validity of our proposed method of estimating $\hat{x}_L$, the late reverberant component which is used in spectral subtraction. It also means that our assumption (a) is realistic.

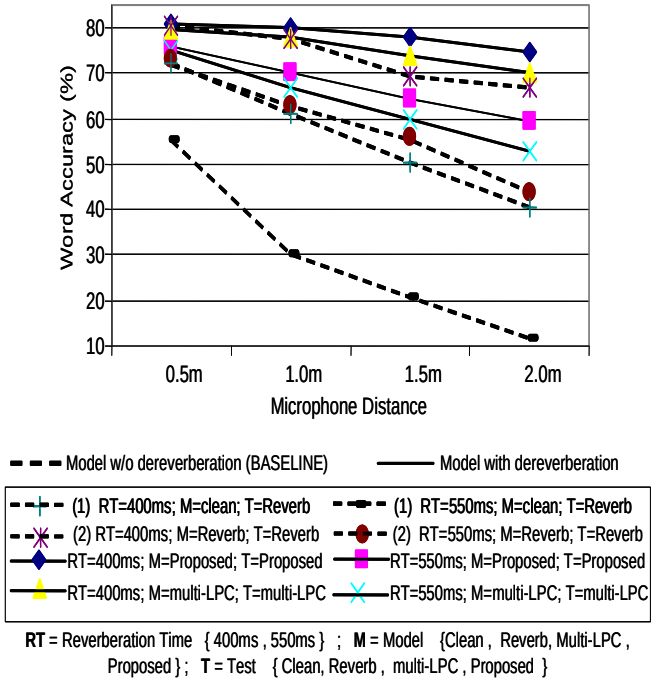


Figure 6: Basic Recognition Performance.

3.2.2 Robustness to Distance

We conducted a recognition experiment that tests the robustness of the approach. By cross validating the performance of the individually created models with a corresponding distance to the test data of di erent distances. Thus, we have two variables in this experiment, the mic distance in training the model and the mic distance in creating the test data. "Mismatched Variables" would mean mismatch distance between the model and the test distances and "Matched Variables" would simply mean that the mic distance used in training the model is the same mic distance in creating the test data. In Figure 7 we show the average recognition performance at each distances. In this gure, it can be shown that the proposed method is very robust in 400ms reverberation as

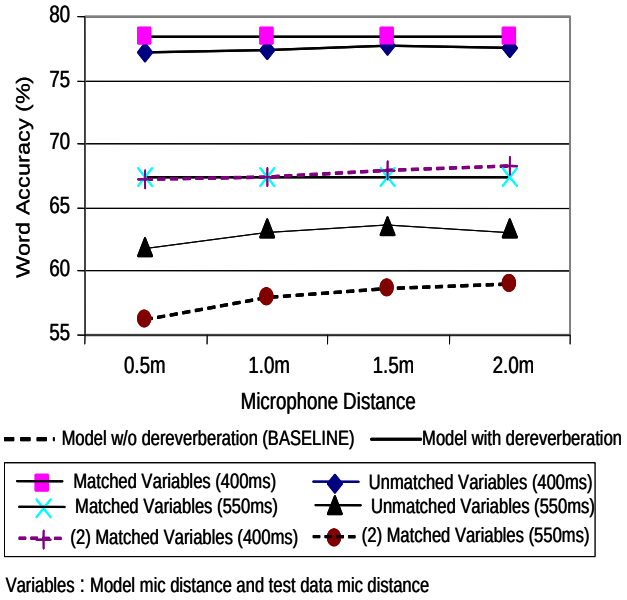


Figure 7: Robustness of the Proposed Method.

compared to the same reverberation time with matched condition. With 550ms reverberation time however, it can be observed that it is not as robust as in the 400ms case but still the performance is far better than using the baseline (1) "clean models" previously shown in Figure 6. If we consider the result of baseline (2) "reverberated matched model" without any dereverberation processing as shown in dashed lines, still the proposed method performs far better. Moreover, in this gure we can claim that, recognition performance improves when the reverberant signal is processed as opposed to without dereverberation at all. In general, Figure 7 shows that the proposed method is robust to distance from speaker to the microphone, thus assumption (b) is valid.

4 Conclusion

The method proposed by using multi-step LPC [5] is novel in the sense that it can adaptively identify the late reverberant components of the reverberant utterance without any information of the impulse response and this cannot be done with the proposed method. However, in practical realtime speech recognition applications it is imperative that we estimate the late reverberant power spectrum to be used in spectral subtraction in a real-time manner. This is the motivation for the assumptions in the proposed method. The result shows that the method is robust to the variation of the microphone distance and does not require to change the impulse response at various distances. The good recognition performance justi es the validity of our assumptions. In our future work we will test for robustness in di erent environment conditions (reverberation time) and consider adaptation approach for the spectral subtraction for robustness

5 Acknowledgment

This work is supported by the Japanese MEXT e-Society project.

References

[1] C.J.Leggeter and Woodland "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models" *In Proceedings of Computer Speech and Language*, vol.9,pp.171-185, 1995

[2] J.L. Flanagan, J.Johnston, R.Zahn, and G. Elko, "Computer-steered Microphone Arrays For Sound Transduction in Large Rooms" *In Proceedings of Acoustics Society of America*,Vol.78,pp.1508-1518, 1985

[3] S. Nakamura, and K. Shikano "Room Acoustics and Reverberation: Impact on Hands Free Speech Recognition" *Invited Paper for EUROSPEECH 1997*, pp.2419-2422, 1997

[4] K. Kinoshita, T. Nakatani, and M. Miyoshi "Ef- cient Dereverberation Framework For Automatic Speech Recognition" *In Proceedings of ICSLP* , Vol 1, pp 92-95, 2005

[5] K. Kinoshita, T. Nakatani, and M. Miyoshi "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *In Proceedings of ICASSP*, 2006

[6] S. Kamath, and P. Loizou "A Multi-Band Spectral Subtraction Method for enhancing Speech corrupted by colored Noise" *In Proceedings of ICASSP*, 2002

[7] S.F Boll "Suppression of Acoustic Noise in Speech using Spectral Subtraction" *IEEE Trans. on ASSP*, vol. 27(2), pp. 113–120, 1979

[8] Y. Suzuki, F. Asano, H.-Y. Kim, and Toshio Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" *J. Acoust. Soc. Am. Vol.97(2), pp.-1119-1123, 1995*

[9] "Julius, an Open-Source Large Vocabulary CSR Engine - http://julius.sourceforge.jp"

[10] A. Lee, T. Kawahara, K. Takeda and K. Shikano, "A New Phonetic Tied-Mixture Model For E cient Decoding" *In Proceedings of ICASSP* , pp. 1269-1272 2000.

[11] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research" *The Journal of Acoustical Society of Japan*, vol. 20, pp. 199-206, 1999

発話音声に関わる頭部動作の分析及びアンドロイドロボットの頭部制御

Analysis of head motions during speech utterances, and head motion control in an android

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
石黒浩（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）
* Carlos Toshinori ISHI, Hiroshi ISHIGURO, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract - With the aim of automatically generating head motions during speech utterances, analyses are conducted for verifying the relations between head motions and linguistic and paralinguistic information carried by speech utterances. Motion captured data are recorded during natural dialogue, and the rotation angles are estimated from the head marker data. Analysis results showed that nods frequently occur during speech utterances, not only for expressing specific dialog acts such as agreement and affirmation, but also as indicative of syntactic or semantic units, appearing at the last syllable of the phrases, in strong phrase boundaries. Analyses are also conducted on the dependence on linguistic, prosodic and voice quality information of other head motions, like shakes and tilts, and discuss about the potentiality for their use in automatic generation of head motions. The paper also proposes a method for controlling the head actuators of an android based on the rotation angles, and evaluates the mapping from the human head motions.

1　はじめに

人が発話をする際、自然に頭部動作が伴う。頭部動作はなんらかの意図を示すため意識的に行う場合がある。例えば頷くことにより、同意や肯定を示し、首を横に振ることにより、否定を示す。その一方、多くの場合は、発話に伴い、無意識に頭部動作が生じる。しかし、無意識に生じる頭部動作の場合も、発話となんらかの関連が存在すると考えられる。

我々の研究背景としては、人型ロボット（アンドロイドなど）の遠隔操作において、音声に伴う頭部動作を音声信号から自動的に生成することを目的としている。そこで本研究では、音声に含まれる言語及びパラ言語情報と頭部動作との関連を調べた。

具体的には、モーション・キャプチャーにより測定された頭部動作（頷き、首かしげ、首振り）と、それに伴う発話音声が伝達する談話機能（質問、相槌、発話権（ターン）の保持・譲渡など）や、音声に含まれる形態素、韻律、声質などの言語およびパラ言語情報との関連を分析した。

頭部動作と音声の関連に関しては、頭部動作と基本周波数（F0）およびパワーなどの韻律特徴を関連付ける研究が多い[1]-[6]。

例えば、Yehiaら[1]は、英語と日本語の読み上げ文に対し、頭部動作からのF0の推定は平均70%以上推定可能だが、F0からの頭部動作の推定においては英語の場合50%、日本語においては30%以下と報告している。Grafら[4]は、英語の文発話に対し、頭部動作と音声の関連を調べ、文内の単語の強調は頷きが頻繁に伴い、頭部を上げる動作は声を上げることに対応すると報告している。彼らはこれらの動作を「視覚的韻律」(“visual prosody”)と呼んでいる。

Beskowら[5]は、スウェーデン語の読み上げ文に対し、強調する単語を変えながら、肯定、質問、迷い、楽しさ、怒りなどのさまざまな表現を変えた場合の発話と、頭部動作と表情を含んだ顔のパラメータとの関連を調べている。結果として、すべての表現において、強調された単語において、強調されていない単語よりも、顔のパラメータの分散が大きかったと報告している。また、岩野ら[7]は、視覚情報を利用して対話理解を向上させることを目的とし、日本語の対話音声における頭部動作の役割を分析している。発話権や発話意図が考慮され、肯定・同意・応答・相槌では縦方向の動作、相手に応答を求める場合は顔を上げる動作が頻繁に生じることを報告している。

我々は人型ロボットの発話に伴う自然な頭部動作を生成する問題に対し、３つのステップを考慮する。第１ステップは発話音声と頭部動作の関連を調べることである。第２ステップは、人型ロボットが自然な動作を再現できるかを調べることである。最後の第３ステップは、発話音声から頭部動作を生成することである。本報告では最初の第１と第２ステップについて述べる。

本稿は以下のように構成される。次ぐ第２節ではデータ収集及びラベル付与（頭部動作と言語・パラ言語情報など）について述べる。第３節では頭部動作と言語・パラ言語情報の関連を分析し、音声からの頭部動作の予測について議論する。第４節ではア

ンドロイドロボットの頭部を制御する方法を説明し、第５節で結論を述べる。

2　音声、頭部動作データ及びラベリング

2.1 データ収集

分析用データとして、母語話者同士（大学院生男女各 1 名）の 30 分程度の自由対話を用いた。分析対象となる女性話者には、頭部と上半身に反射マーカーを貼り付け、音声とビデオと同時に、モーションキャプチャシステム(Hawk system of Motion Analysis)による頭部のマーカーデータを収集した。図１に示すように、頭部、顔面、上半身に付けた計 38 個のマーカーより３次元空間の位置データを測定した。多くのマーカーは唇の動きや表情などの情報を測定するためのものであり、本研究では、表情や唇の動きとは無関係となる頭部と鼻と耳に付けた計６つのマーカーを用いて頭部動作を表現する。

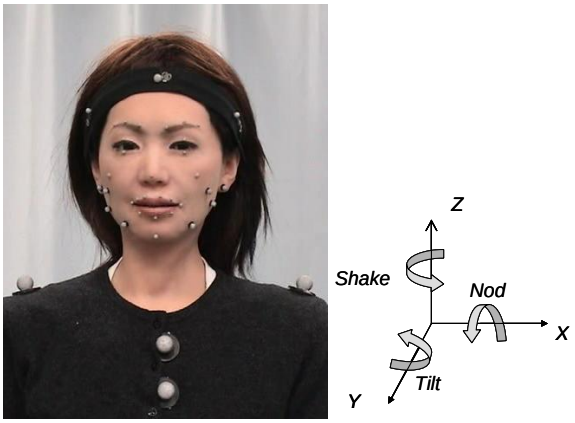


Fig. 1.　Markers and angles used to describe head motions.

図１に示した頭部動作を表現した３次元の回転角度をマーカーデータより SVD 法(Singular Value Decomposition method)に基づいて、以下の数式を用いて推定した[8]。

$$[U,D,V^T] = svd\ (\ reference * target\ ), \quad (1)$$

ここで *reference* と *target* はそれぞれニュートラルと対象となる３次元空間のマーカー位置を示す。また、マーカーの centroid が原点となる新しい座標にすべてのマーカー位置を移動させた。回転行列は以下の式により求められる。

$$R = V*U^T. \quad (2)$$

回転角度は､回転行列Rの要素にarccosineをかけることにより求める。図１の座標に示すように、頷き（nod）、首かしげ（tilt）、首振り（shake）の動作は、それぞれX軸、Y軸、Z軸の周りの回転により表現される。

2.2 頭部動作のタグ

回転角度の時系列とビデオ情報を基に、以下に示す**頭部動作タグ**を提案した。

- fu (face up)：顔を上げる動作
- fd (face down)：顔を下げる動作
- nd (nod)：頷く動作
- sh (shake)：首振り
- ti (tilt)：首をかしげる動作
- no：頭部動作無し

頷きは完全な上下方向の動きとして実現される訳ではない。緩い首かしげに伴って起きることも観察された。この場合は、度合いが強い方の動きを捉えることにした。

2.3 言語情報（形態素）

予備的な分析により、「ね」「で」「から」などの助詞や、「うん」「ええ」などの感動詞に頷きが多く生じることが観察された。これらの形態素は句の境界に通常現れるため、本研究では発話音声を**アクセント句単位**に区切り、アクセント句単位で頭部動作を調べた。

日本語母語話者１名が、発話音声をアクセント句単位に区切り、**句末の形態素**を書き起こした。535 個のアクセント句が揃った。

2.4 談話機能のタグ

本研究では、肯定的・否定的応答、発話権を考慮した談話機能と頭部動作との関連を考慮し、[9]で提案された以下に示す**談話機能**のタグを用いた。

- k (keep)：発話権の保持（強い句境界）
- k2 (keep)：弱い句境界（発話権の保持）
- k3 (keep)：発話末を伸ばし、発話の途中であることを表現（発話権の保持）
- f (filler)：「えっとー」「あのー」など、考え中であることを表現する感動詞
- f2 (conjunctions)：「じゃ」などの接続詞（短いフィラーとして捉えられる）
- g (give)：対話相手への発話権の譲渡
- q (question)：発話権の譲渡（対話相手に応答を求める場合）
- bc (backchannels)：「うん」「はい」などの相槌を表現する感動詞
- su (surprise/admiration)：「えー！」「へー」など、驚きや感心などの感情を表現する感動詞
- dn (denial)：「いいえ」「ううん」などの否定を表現する感動詞

発話行為タグは被験者１名により付与され、その後、別の被験者により確認・訂正された。

2.5 韻律・声質のタグ

本研究では談話機能に関連する句末音調の特徴に注目することにした。韻律・声質特徴の自動抽出法に関しては、[10][11]で提案されたものが存在するが、本研究で得られた音声データには雑音や障害音が多く混合していたため、手動でタグを付与することにした。また、句末では声質の変化も多く観察され、F0 が存在しない場合もあるため、声質のタグも付与した。本研究では、[9]の句末音調のラベルセットを基に、以下に示すタグセットを用いた。

- rs (rise)：上昇調
- fa (fall)：下降調
- fr (fall-rise)：下降上昇調
- hi (high)：高ピッチ
- mi (mid-height)：中ピッチ
- lo (low)：低ピッチ
- cr (creaky)：フライ発声（声帯振動が不規則であるため F0 の計測が難しい）
- wh (whisper)：ささやき声（F0 が無い）

韻律・声質タグは、著者がスペクトログラム及び F0 軌道を参考に、音声を聴取しながらラベル付与を行った。

3　音声と頭部動作の関連

3.1 頭部動作と形態素

まず、頭部動作と形態素の種類との関連に関しては、助詞や感動詞に伴って頻繁に頷きが生じることを確認した。ただし、これらの品詞には限らず、むしろ、強い句境界を伴う句末で生じやすいことがわかった。ここで「強い句境界」とは、句境界にポーズまたは明らかなピッチの立て直しがある場合を差す。予備実験で助詞や感動詞に頷きが多く観られたのは、これらの品詞は通常句末に出現するからである。また、頷き動作は品詞や形態素の種類よりも、談話機能に強い依存性を持つ傾向が観られた。この結果については、次節で詳しく述べる。

首振りに関しては、データが少なかったが、「ううん」や「。。。ない」を含む、否定を表現した発話に伴って出現した。

3.2 頭部動作と談話機能

表 1 は頭部動作と談話機能との関連を示す。頷き(nd)が最も頻繁に出現した頭部動作である。まず、予想されていた結果として、相槌(bc)のほとんどに頷きが起きることが分かる。また、3.1 節でも述べたように、品詞の種類に関わらず、強い句境界(k, g, q)でも頷きが頻繁に起きる結果が得られた。

発話末尾が上昇調になる質問系(q)の発話でも、顔を上げる動作(fu)よりも、頷き(nd)の方が多かった。これはピッチと頭部動作の相関を低くする一つの原因と考えられる。

話者が発話権を保持した場合(k)、頷き(nd)が多く出現する一方、顔を上げる動作(fu)はほとんど出現しなかった。

発話中の弱い句境界(k2)、及び話者が考え中もしくは発話がまだ終わっていない場合(k3, f, f2)には、頭部動作が伴わない(no)傾向が観られた。

驚きや感心を示す発話(su)は小数であったが、顔を上げる動作(fu)もしくは首をかしげる動作(ti)が伴う傾向が観られた。これらの感情を表現する発話に関してはデータ数を増やして検証する必要がある。

句頭においても頷きが出現する場合が（表 1 から省いた）10 個の発話に観られた。これは対話相手から発話権を交替する合図とも考えられる。顔を上げる動作も同様な機能で句頭に出現したものとみられる。

頭部動作の形状に関しては、頷きの場合、下降上昇動作に先立ち、小さい上昇動作が生じる場合が多く観られた。同意を表す際に、頷きが発話全体に渡って連続に起きる例もあり、特に「うんうんうんうん」のように相槌が連続して発声される場合、頷きも連続して起き、最初の頷きが後続のものよりも大きい傾向が観られた。

首振り(sh)は、本データに数少なく出現したが、否定を表現し、「ううん」や「。。。ない」など否定を表現する発話に伴うことが観られ、発話内容との依存性が強いと考えられる。

首かしげ(ti)は、相槌(bc)、質問(q)、否定(dn)以外の談話機能で出現した。ただし、首かしげは頷きと首振りに比べて比較的長い持続時間に渡って生じ、複数のアクセント句を含む場合が多かった。首かしげは話者が考え中、または自信がない発話で、言いよどみまたは言い詰まりの直後に出現する傾向がみられた。首かしげの左右における違いは観られなかった。

TABLE 1
DISTRIBUTION OF DIALOG ACTS (ROWS) AND HEAD MOTIONS (COLUMNS).

		nd	**fd**	**fu**	**ti**	**sh**	**no**
	total	***145***	24	48	33	4	189
k	61	***35***	***5***	1	3	0	17
k2	137	2	2	12	11	0	***106***
k3	28	4	1	3	4	0	***16***
f1	15	3	1	1	2	0	***8***
f2	22	2	0	3	5	0	***12***
g	79	***30***	***11***	***15***	5	2	14
q	25	**9**	3	5	0	0	7
bc	71	***60***	1	2	0	0	7
su	12	0	0	***6***	3	0	1
dn	4	0	0	1	0	***2***	1

3.3 頭部動作と韻律・声質特徴

ピッチと頭部動作の間に、なんらかの関連は存在するものの、その相関はあまり強くない[1]。日本語はピッチアクセント言語であり、単語のアクセント核ではピッチの動きが起きる。しかし、本実験では、

ピッチの動きが大きいアクセント核よりも句境界で頭部動作が生じやすい傾向が観られた。

更に、句境界音調と頭部動作との間には１対１の関係が成り立たない結果が得られた。表２は頭部動作と句末音調の分布を示す。

TABLE 2
DISTRIBUTION OF HEAD MOTIONS (ROWS) AND PHRASE FINAL TONES (COLUMNS).

	rs	fa	fr	hi	mi	lo	cr	wh
total	25	108	3	26	205	14	58	13
nd	8	***83***	1	0	29	4	***20***	4
fd	2	***4***	0	1	5	2	***7***	2
fu	8	3	0	5	22	3	7	2
ti	1	4	0	2	20	1	4	1
sh	0	0	2	0	1	0	1	0
no	6	14	0	***18***	***126***	3	18	4

表２よるおおまかな結果として、下降調(fa)とフライ発声(cr)では頷き(nd)が伴い、ピッチが高(hi)と中(mi)の場合は頭部動作が伴わない(no)ことが多い。

音調と頭部動作の間には強い関連が得られなかった一方、音調と談話機能との間には、表３に示すように、より強い関連が観られた。上昇調(rs)は質問系(q)、下降調(fa)は相槌(k)または発話権の保持(k)、高ピッチ(hi)及び中ピッチ(mi)は弱い句境界(k2)、低ピッチ(lo)、フライ発声(cr)、ささやき声(w)は、発話権の譲渡(g)に多く出現する結果が得られた。音調とともに形態素の情報も考慮すれば、発話行為とのより高い関連が得られると考えられる[9]。

TABLE 3
DISTRIBUTION OF DIALOG ACTS (ROWS) AND PHRASE FINAL TONES (COLUMNS).

	rs	fa	fr	hi	mi	lo	cr	wh
k	1	***45***	0	1	12	0	2	0
k2	0	1	0	***12***	***110***	2	11	1
k3	0	10	0	4	12	0	2	0
f1	0	3	0	0	11	0	1	0
f2	0	2	0	2	14	0	4	0
g	2	1	0	1	20	***9***	***36***	***10***
q	***19***	1	1	1	1	0	1	0
bc	0	***45***	0	0	19	3	1	2
su	2	0	0	5	5	0	0	0
dn	1	0	2	0	1	0	0	0

最後に、頭部動作と（上述の表に示されていない）他の声質に関して、頷きと首かしげにおいて、自信のない発話では柔らかい声質に伴い、頷きの強さが弱くなり、首かしげの出現または頭部動作が伴わない傾向が観られた。この結果に関しても更なる検証が必要である。

4　頭部動作のアンドロイドへの再現

前節では、音声に含まれるさまざまな言語やパラ言語的情報と頭部動作との関連を分析した。しかし、得られた結果を利用して、音声から頭部動作を生成する完全なシステムの構築に役立てるためには、音声信号から言語やパラ言語情報の自動抽出を実現する必要がある。現段階では自動抽出の問題は未解決であり、今後の課題として残される。

本節では、頭部動作を生成するシステムのハードウェアに関連する部分に対し、人間の頭部動作をアンドロイドで再現するためのマッピングについて紹介する。また、アンドロイドの機械的制限を考慮したうえで、人間の自然な頭部動作が再現できるか検証する。

4.1 アンドロイドの頭部アクチュエータ

本研究では、頭部動作の制御を評価するテストベッドとして、日本人女性をモデルとした大阪大学に存在するアンドロイドロボットRepliee Q2 [12]を用いた。このアンドロイドは図２に示すように頭部動作を制御するため３つのアクチュエータが備えられている。１４番と１５番アクチュエータは独立に制御した場合、それぞれ右上から左下と左上から右下の斜め方向に頭部を動作させる。これらの両アクチュエータを同時に同じ制御値で動作させることにより上下方向に頭部を動作させる。１６番アクチュエータは頭部を左右水平方向に動作させる。

アクチュエータの制御値は１４番と１５番では 0 から255 の範囲に、１６番では 50 から 205 の範囲に設定可能である。アクチュエータの中間制御値は127（ニュートラル位置）である。１６番のアクチュエータの制限は、首を左右方向に動かす際に、アンドロイドの首の周辺の皮膚が裂けることを避けるためである。

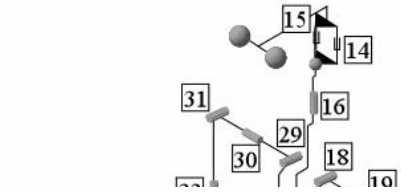


Fig. 2.　Android actuators (14-16 head actuators).

本アンドロイドの頭部動作の幅は、人間の頭部動作の幅に比べて制限されている。表４はアンドロイドRepliee Q2で測定された頭部の回転角度の幅を被験者のものと比較する。回転角度はニュートラルの位置に対するものである。*Angle_X,Y,Z* は（図１に示した座標の）X, Y, Z 軸の周りの回転角度を示す。

TABLE 4
RANGE OF THE HEAD ROTATION ANGLES (IN DEGREES) MEASURED IN THE HUMAN SUBJECT AND IN THE ANDROID RELATIVE TO THE NEUTRAL POSITIONS.

	human	android
Minimum *Angle_X* (downward direction)	-50	-12
Maximum *Angle_X* (upward direction)	30	12
Minimum *Angle_Y* (slantwise left direction)	-30	-12
Maximum *Angle_Y* (slantwise right direction)	30	12
Minimum *Angle_Z* (left direction)	-50	-20
Maximum *Angle_Z* (right direction)	40	20

回転角度の幅は、表４に示すように、人間の方がアンドロイドよりも２倍以上大きいが、自然対話の際には、図３に示すように、人間の頭部動作の動きの幅は比較的小さいことが分かる。

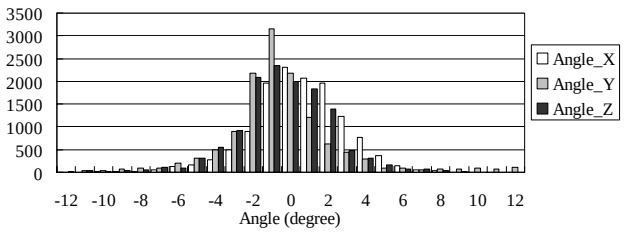


Fig. 3. Histograms of the measured head rotation angles for the subject during natural conversation for an interval of four minutes.

図３の分布では、回転角度が-12 度から 12 度の幅に収まることが観られ、本アンドロイドを用いても対話における頭部動作は再現可能といえる。

4.2 アンドロイドの頭部アクチュエータ

頭部アクチュエータにおいては、フレーム毎に計算した3次元の回転角度を使用する。

各アクチュエータにおける制御値は以下の数式により求められる。

$$act[14] = act\_neutral + (Angle\_X / MaxAngle\_X * 127) + (Angle\_Y / MaxAngle\_Y * 127) \quad (4)$$

$$act[15] = act\_neutral + (Angle\_X / MaxAngle\_X * 127) - (Angle\_Y / MaxAngle\_Y * 127) \quad (5)$$

$$act[16] = act\_neutral + (Angle\_Z / MaxAngle\_Z * 78) \quad (6)$$

ここでは、*act_neutral* はニュートラル状態におけるアクチュエータの値（すべてのアクチュエータで127）、*Angle_X,Y,Z* はターゲットとなる回転角度、そして*maxAngle_X,Y,Z* は制御値が最大の時にアンドロイドで測定された最大の回転角度を示す。また、推定された値がアクチュエータの範囲を超える場合は、最大値もしくは最低値に制限する。

本手法を評価するため、人間の測定された頭部動作をアンドロイドにマッピングした。入力としてはモーションキャプチャで測定された人間のマーカーデータを用い、2.1節で説明した手順に基づいて３次元の回転角度（*Angle_X,Y,Z*）を求める。アンドロイドの頭部のアクチュエータの制御値は、これらの回転角度の値を用いて、式 (3) ～ (5) により推定される。図４で、対話中の被験者のマーカーデータとそれに対応したアンドロイドのマーカーデータより推定した頭部の回転角度の軌道を比較している。いずれの回転軸においてもよいマッチングが得られたことが分かる。ただし、*Angle_X* と*Angle_Y* では幾箇所でターゲットの角度に達成することができなかったが、これはアクチュエータ14と15の相互依存による制限が原因と考えられる。被験者とアンドロイドの頭部の回転角度 *Angle_X, Angle_Y,* および *Angle_Z*の間に、それぞれ 0.74, 0.94 および 0.91 の相関値が得られた。

予備的な主観的評価も高い自然度と人間の動きとの視覚的対応が得られた。（デモビデオを参照。デモビデオでの唇と表情のマッピングは著者らの別の論文[13]の結果により再現されたものである。）

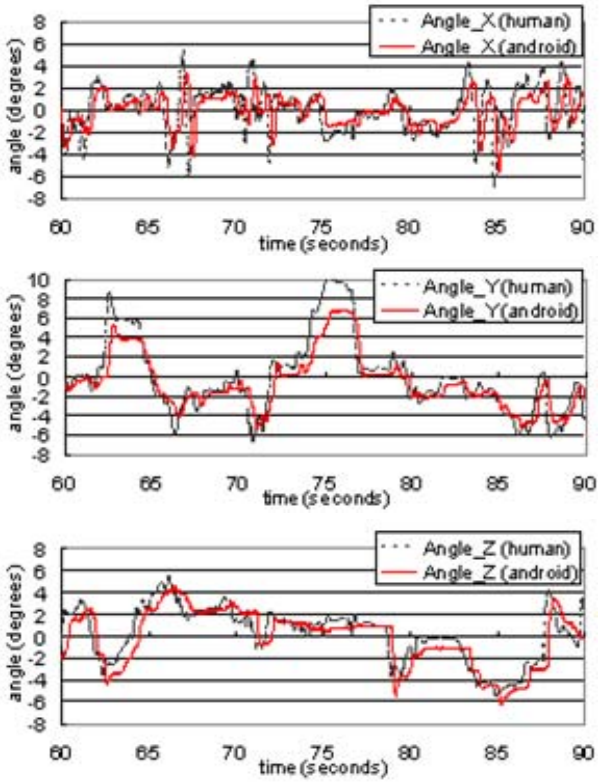


Fig. 4. Head rotation angles measured in the subject (human) during natural conversation and in the android after mapping from the subject motions.

5　頭部動作の自動生成に向けての今後の課題

3 節では音声と頭部動作との関連を分析し、4 節ではアンドロイドロボットの頭部動作の制御に関して検証した。しかし、音声信号から完全に自動的に頭部動作を生成するには、以下に示すように、いくつものステップをクリアする必要があり、今後の課題として残される。

- 発話意図や談話機能の自動検出
- 発話権の保持・譲渡の判別
- 強い句境界の自動検出
- それぞれの発話意図や談話機能の言語的キーワードの収集（例えば、否定を示す言語的表現）
- 言い詰まり・言い淀みの自動検出
- 韻律・声質特徴の頑健な自動抽出
- 感情表現の認識（感情表現のデータの増加、分析が必要；声質特徴が効果的である）

6　おわりに

音声信号から頭部動作を生成することを目的とし、音声に含まれる言語情報や韻律・声質のパラ言語情報及び音声によって伝達される談話機能と、頭部動作との関連を調べた。

さまざまな頭部動作のうち、頷きが最も頻繁に出現し、肯定や同意などの談話機能以外にも、強い句境界における句の最後の音節に高い頻度で出現する結果が得られた。首かしげは、話者が自信がない発話に出現し、フィラーや言い詰まりの直後に起きる傾向が観られた。さらにソフトで気息性を含んだ声質が伴う傾向も観られた。首振りは否定を示し、発話の内容により依存すると考えられる。

全般的な結果として、韻律・声質特徴と頭部動作を直接に対応付けるのは難しいが、言語情報との組み合わせにより、談話機能を介して頭部動作と関連付けられる可能性が示された。

回転角度を用いてアンドロイドロボットの頭部を制御する手法も提案した。アンドロイドの頭部動作には機械的な制限が示されたが、対話における頭部動作は再現可能と示した。アンドロイドに自然な頭部動作が再現され、人との頭部動作のマッピングにおいても良い対応が得られた。

今後の予定は、多人数のデータも分析し、頭部動作の個人性を検討する。また、句境界検出や言い詰まりの検出などを含んだ言語・パラ言語情報の自動抽出法を検討し、本研究の分析結果を利用して、人間型ロボットの自然な頭部動作を音声から自動的に生成することを試みる。

謝辞

本研究は総務省の研究委託により実施したものである。モーションキャプチャのデータ収集やアンドロイドの操作に貢献した大阪大学のJudith Haas氏、Freerk Wilbers氏、及び高野絵里氏に感謝する。データ解析に貢献した中西京子氏に感謝する。

参考文献

1) H.C. Yehia, T. Kuratate, E. Vatikiotis-Bateson, “Linking facial animation, head motion and speech acoustics,” *J. of Phonetics*, Vol. 30, pp. 555-568, 2002.
2) M.E. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, A.M. Tekalp, “Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis,” Proc *IEEE International Conference on Multimedia*, 2006.
3) K.G. Munhall, J.A. Jones, D.E. Callan, T. Kuratate, E. Vatikiotis-Bateson, “Visual prosody and speech intelligibility – Head movement improves auditory speech perception,” *Psychological Science*, Vol. 15, No. 2, pp. 133-137, 2004.
4) H.P. Graf., E. Cosatto, V. Strom, F.J. Huang, “Visual prosody: Facial movements accompanying speech,” Proc. IEEE *Int. Conf. on Automatic Face and Gesture Recognition* (*FGR’02*), 2002.
5) J. Beskow, B. Granstrom, D. House, “Visual correlates to prominence in several expressive modes,” Proc. *Interspeech 2006 – ICSLP*, pp. 1272-1275, 2006.
6) C. Busso, Z. Deng, M. Grimm, U. Neumann, S. Narayanan, “Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis,” *IEEE Trans. on Audio, Speech and Language Processing*, March 2007.
7) Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, K. Shirai, “Analysis of head movements and its role in spoken dialogue,” Proc. *ICSLP’96*, pp. 2167-2170, 1996.
8) M.B. Stegmann, D.D. Gomez, “A brief introduction to statistical shape analysis,” published online, 2002.
9) C.T. Ishi, H. Ishiguro, N. Hagita, “Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts,” Proc. *Interspeech’2006 - ICSLP*, pp. 2006-2009, 2006.
10) C.T. Ishi, “Perceptually-related F0 parameters for automatic classification of phrase final tones,” *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 3, pp. 481-488, 2005.
11) C.T. Ishi, H. Ishiguro, N. Hagita, “Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information,” Proc. *IROS 2006*, pp. 374-379, 2006.
12) T. Minato, M. Shimada, H. Ishiguro, S. Itakura. “Development of an android robot for studying human-robot interaction,” *Innovations in Applied Artificial Intelligence*, Springer Verlag, pp. 424-434, 2004.
13) F.P. Wilbers, C.T. Ishi, H. Ishiguro, “A blendshape model for mapping facial motions to an android,” Proc. *IROS 2007*, 2007.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-9

単旋律楽曲における機械学習を用いた楽器音の音源同定

Sound-source Identification on Monophonic Music Using Machine-Learning Approaches

井原瑞希, 前田新一
Mizuki Ihara, Shin-ichi Maeda
奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Nara
{mizuki-i ichi}@is.naist.jp

石井信
Shin Ishii
京都大学
Kyoto University, Kyoto
ishii@i.kyoto-u.ac.jp

Abstract

In this study, we propose an instrument feature extraction method applied for musical instrument identification. Although there are many approaches to extract instrument features, convensional feature extraction methods use heuristic assumptions on timbre; for example, the certain portion of spectrum such as spectrum envelopes represent the instrument characteristics. On the contrary, our proposed method does not assume any predetermined features. We first directly utilize the raw log-power spectrum as spectrum features. The problem occurs here is high complexity because of the high feature dimensionality. To solve that, two machine learning methods are employed for dimensionality reduction, which are principal component analysis (PCA) and linear Fisher discriminant analysis (LFDA). Instrument identification with parameters selected by PCA-LFDA resulted in 94.5% identification rate with only ten feature parameters.

1 はじめに

少ない特徴数で話者や楽器といった音源の同定ができれば，音楽検索，採譜，音合成のような様々な分野に応用が可能である．音源同定の課題の一つとして楽器推定が挙げられる．楽器推定に関する研究は単音，単旋律，多旋律を対象として行われており，本研究では単旋律楽曲における楽器推定を行う．

Essid は，スペクトル情報，時間変化情報，音量情報を含む 160 次元の音色特徴[1]を用意したあと，遺伝的アルゴリズム（*genetic algorithms*; GA）と特徴空間射影を用いた慣性比最大化（*inertia ratio maximization using feature space projection*; IRMFSP）を使用して楽器同定に重要と考えられる 70 個の特徴を選択し，その特徴を使って楽器識別を行った結果，10 種類の楽器において 87% の楽器判別率を達成した [1]．同様に，Livshin は，62 次元の時系列情報やハーモニック情報などの中から Essid とは異なる独自の特徴抽出法を用いて 20 次元の特徴を選択し，それを楽器固有の特徴とした．この実験では，次元削減をしてもほぼ同等の識別結果が得られている [2]．これらの研究では，音色特徴を発見的に仮定しているためにより優れた音色に対応するパラメータを見過ごしている危険性がある．

楽器の音色特徴を一種類に仮定する方法の例には，スペクトル情報もしくはケプストラム情報を用いる手法が成功を収めている．代表的な特徴抽出法として，メル周波数ケプストラム係数 (MFCCs) が挙げられる．Essid らの実験では，5 種類の木管楽器からそれぞれ 10 個の MFCCs を抽出し，楽器特徴として使用されている [3]．さらに，Marques は MFCCs，線形予測係数 (LPCs) とケプストラム係数による楽器判別率をそれぞれ単独で比較し，MFCCs が最も高い認識率を示したと報告している [4]．最近では，Chétry が線スペクトル周波数 (LSFs) のみで楽器同定を行い，6 種類の楽器において LPCs や MFCCs よりも高い，86% の判別率をあげている [5]．以上の実験は，すべて単旋律の楽曲において実験が行われており，単音や多旋律の実験は含まれない．

上で挙げた MFCCs，LPCs，LSFs はそれぞれスペクトル情報に基づいた表現であり，それらが楽器特徴を表現すると仮定している．これらの特徴はスペクトル情報のみから計算が可能なこと，また，これらの特徴による楽器同定が高い判別率となっていることから，スペクトル情報のみでも精度の高い楽器同定が出来ることを示唆している．一方で，これらの研究では，音色特徴をあらかじめ仮定しているが，その音色特徴の仮定が正しいとは言い切れない [6].

[1]本研究では，音量と音高には依存しない楽器固有の特徴を音色特徴とよぶ.

そこで本研究では，スペクトル情報全体でどれくらい楽器同定ができるかどうか確かめ，そのスペクトル情報の特徴次元圧縮に関して考察した．次元圧縮法として，機械学習法である主成分分析と局所フィッシャー判別分析の組み合わせを提案し，そのパラメータを楽器の音色特徴とした上での楽器識別の結果を従来手法やモデルを仮定しない主な機械学習の次元圧縮法と比較した結果から，提案手法が楽器特徴の抽出において優れていることを示す．

2 楽器の特徴抽出

本章ではまず，楽器推定や話者推定の従来研究の多くで採用されている，スペクトル情報に基づく特徴抽出方法について述べた後，提案手法を含む，機械学習に基づく次元削減法について説明する．

2.1 従来の特徴抽出方法

2.1.1 メル周波数ケプストラム係数 (MFCCs)

楽器の特徴や音声の個人性を表現するパラメータとしてしばしば用いられる係数として，メル周波数ケプストラム係数（*mel-frequency cepstral coefficients*; MFCCs）が挙げられる．音の対数スペクトル強度を逆フーリエ変換することで以下のようなケプストラム $c(\tau)$ が得られる．

$$c(\tau) = \mathcal{F}^{-1} \log |X(\omega)| \tag{1}$$
$$= \mathcal{F}^{-1} \log |G(\omega)| + \mathcal{F}^{-1} \log |H(\omega)| \tag{2}$$

ここでケプストラム (cepstrum) とは spectrum のつづりを逆にした造語であり，その変数 τ はケフレンシーとよばれる．$\mathcal{F}^{-1}$ はフーリエ逆変換，ω は周波数，$X(\cdot)$，$G(\cdot)$，$H(\cdot)$ はそれぞれ音信号，音源特性（高ケフレンシー），調音特性（低ケフレンシー）を表現していると考える．音声の分野では，周波数スペクトルの中の包絡に音色成分が含まれると仮定されることが多く [7, 8]，包絡を推定することができれば楽器の特徴が抽出できると考えられる．ケプストラム分析において，スペクトルの緩やかな変化はフーリエ変換後のケプストラム領域の低周波成分（低ケフレンシー）で表現できるはずである．そこで，低ケフレンシー成分 H がスペクトル包絡を表わすと仮定する．このケプストラム分析に加えて，MFCC を計算するには，対数スペクトルの周波数 $\mathbf{f}$Hz を

$$\mathbf{f}_{\mathrm{mel}} = 2595 \log_{10} \quad 1 + \frac{\mathbf{f}}{700\mathrm{Hz}} \quad . \tag{3}$$

のように，まずメル尺度に変換する．メル尺度は高周波数よりも低周波数が強調されるといった人間の聴覚特性を模擬している [9]．楽器の特徴を表現するパラメータとしては他の手法が MFCC よりも良い結果を表しているという研究結果もあるが [10]，係数の抽出が簡単であるという利点から，頻繁に楽器特徴として使われている [11, 12].

2.1.2 線形予測符合化 (LPC) 係数

線形予測分析（*linear predictive coding*; LPC）は音の周波数スペクトルの概形を予測する代表的な手法の一つである．時間領域での音信号の時刻の標本値を $s_t (t = 1, \cdots T)$ とする．LPC はこの標本値を p 個の過去の時刻の標本値の線形結合に時間ごとに独立な平均 0，分散 σ^2 のガウスノイズ ϵ_t が加えられたものとして以下のように推定する．

$$s_t = \sum_{i=1}^{p} \alpha_i s_{t-i} + \epsilon_t. \tag{4}$$

ここで，α_i $(i = 1, \cdots, p)$ は最尤推定で決定される LPC 係数である．z 変換を用いると，z 領域でのスペクトル包絡は

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-i}} \tag{5}$$

と書ける．従って周波数領域で書き直すと，Fs がサンプリング周波数，$\tilde{\omega}$ が正規化周波数 $\tilde{\omega} = \frac{\omega \mathrm{Fs}}{2\pi} (-\pi \leq \tilde{\omega} \leq \pi)$ のとき，

$$H(\tilde{\omega}) = \frac{\sigma^2}{2\pi} \frac{1}{A_0 + 2\sum_{i=1}^{p} A_i \cos(i\tilde{\omega})} \tag{6}$$

となる．ここでパラメータ A_i は，$\alpha_0 = 1$ として，LPC 係数の相関

$$A_i = \sum_{j=0}^{p-|i|} \alpha_j \alpha_{j+|i|} \tag{7}$$

で計算される [13, 14]．LPC はスペクトル包絡を少数のパラメータで推定可能であるが，LPC 係数の丸め誤差に対して鋭敏に変化するための安定性を失うおそれがある [15].

2.1.3 線スペクトル周波数 (LSF) パラメータ

LPC 係数の丸め誤差に対する鋭敏性を避けるため，LPC と 1 対 1 対応をもつ順方向と逆方向の予測誤差の相関（偏相関）係数を表す 偏自己相関（*Partial Auto Correlation*; PARCOR）係数が提案された [16]．このモデルの改良として，声道モデルにおいて声道内部を無損失と仮定した線スペクトル周波数（*line spectral frequencies*; LSF）が考案された．LSF は LPC と同様に，周波数スペクトルの包絡を推定することができる．LSF 係数は LPC 係数や PARCOR 係数に変換することができ，それらより高い復元率と安定性があることが示されており [17]，そのため，音声圧縮や音声合成など広い分野で現在も使用されている [18, 19]. LSF のスペクトル包絡は

$$H(\tilde{\omega}) = 2^{1-p} \left\{ \sin^2 \frac{\tilde{\omega}}{2} \prod_{n=2,4,\cdots,p} (\cos\tilde{\omega} - \cos b_n)^2 + \cos^2 \frac{\tilde{\omega}}{2} \prod_{n=1,3,\cdots,p-1} (\cos\tilde{\omega} - \cos b_n)^2 \right\}^{-2} \tag{8}$$

で表すことができる．この方法はすでに従来研究で試されており，単音データを使用した楽器判別において MFCCs と線形予測ケプストラム係数と比較して良好な結果を示している [10].

2.2 スペクトルの次元圧縮方法

前節では，スペクトル包絡が楽器特徴に対応していると仮定し，スペクトル包絡の抽出方法を楽器の特徴抽出として使用されている従来方法について説明した．この節では，楽器特徴とスペクトルの対応を前もって仮定しない機械学習手法について説明する．

2.2.1 主成分分析 (PCA)

機械学習の教師なし代表的な次元圧縮法として，主成分分析（*principal component analysis*; PCA）がある．PCA は入力空間 $\mathbf{x}$ から特徴空間 $\mathbf{y}$ への射影が線形変換

$$\mathbf{x} = \mathbf{W}^T \mathbf{y} \tag{9}$$

に基づいている．ここで T は転置行列を表しており，PCA では $\mathbf{x}$ の分散が最大となるような直交行列 $\mathbf{W}$ を求めている．PCA は多変量データの情報損失（平均自乗誤差）を最小限におさえて情報を圧縮する次元圧縮法である．第 1 主成分は分散が最大になるように選択される．第 2 主成分は，第 1 主成分と直交する部分空間上で分散が最大になるように選ばれ，第 3 主成分以降も同様に決定され，それによって PCA の変換行列が決められる．この手法は自乗誤差を最小にする線形変換という点で学習データ全体の特徴を抽出するための最適な変換を行っているが，教師なし学習を行うため，入力データに対応するラベルは考慮できない.

2.2.2 線形判別分析 (LDA)

線形判別分析（*linear discriminant analysis*; LDA）は，教師あり学習による次元圧縮法である．PCA と同様に，LDA の特徴空間 $\mathbf{x}$ から判別空間 $\mathbf{y}$ への射影は式 (9) のような線形変換に基づいている．LDA においては $\mathbf{W}$ が，クラス間分散を最大にし，クラス内分散を最小にするように決定される．LDA はクラスラベル情報を考慮してクラスごとに次元圧縮をおこなう．n 次元の入力空間を $d \times n$ の変換行列 W によって d 次元の判別空間に射影するとき，LDA の目的関数は

$$\max_{W} \mathrm{Tr} \quad W^T S_W W \ ^{-1} W^T S_B W \tag{10}$$

となり，ここで，S_W と S_B はそれぞれクラス内共分散とクラス間共分散行列を表わしている．すべての入力サンプルの平均値を μ，総クラス数を C とし，C_i，n_i，μ_i を i 番目のクラスのサンプル集合，サンプル数，クラス内の平均値としたとき，クラス内共分散行列 S_W とクラス間共分散行列 S_B は次のように定義される.

$$S_W = \sum_{i=1}^{C} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^{\mathrm{T}}, \tag{11}$$

$$S_B = \sum_{i=1}^{C} n_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}. \tag{12}$$

変換行列 W は

$$S_W^{-1} S_B \ \ W = W \ \ W^T S_W W \ ^{-1} \ \ W^T S_B W \tag{13}$$

を解くことで求めることができる [20]．しかし，LDA は判別空間の次元数がクラス数より小さい場合，クラス間共分散行列で固有値問題が起こる [21].

2.2.3 局所フィッシャー判別分析 (LFDA)

LDA はデータに多峰性がある場合や判別空間の次元数が総クラス数よりも小さい場合には好ましくない結果になることがある [22, 21]．その点を考慮した教師ありの線形変換に基づく次元圧縮法として，局所フィッシャー判別分析（*local Fisher discriminant analysis*; LFDA）が考案された．この方法は，LDA と教師なし次元圧縮法のひとつである局所性保存射影（*locality preserving projection*; LPP）を組み合わせた方法である．LPP は近くのサンプルを近くに射影する線形変換で，顔認識において PCA や LDA より良い結果を出している [23]．LFDA は，近傍のサンプルのみを考慮に入れることで，多峰性のあるデータへの対応も可能にしている [24].

2.2.4 提案する特徴抽出法 (PCA-LFDA)

クラスラベル情報が使えるという点で教師なし学習である PCA よりも教師あり学習の LDA の方が楽器同定に有効な特徴が抽出できると期待されるが，LDA はデータ数が少ない場合，オーバーフィッティングしやすいという問題点がある．この問題を解決する方法として，しばしば顔認識の研究では，入力データを PCA によって低次元特徴ベクトルに変換した後に LDA で次元圧縮を行う方法がとられている [25, 26]．初めに PCA を適用することでデータに含まれる冗長な次元を線形変換によって取り除いた後，LDA を用いてさらに低次元の判別空間に射影する．このように PCA によって LDA で学習するパラメータ数を減らすことにより，オーバーフィッティングを避けた楽器特徴を抽出することができる．本研究では PCA と LFDA を組み合わせた手法を提案し，この手法を PCA-LFDA と呼ぶ.

3 分類手法

統計的学習法に基づく分類手法の一つとして，サポートベクタマシン (*support vector machines*; SVM) がある．分類結果の良さから，単音，単旋律，多旋律に関わらず，これ

までの楽器識別研究では最も頻繁に使われている分類手法である．Marques らは，SVM と混合ガウス分布モデル（*Gaussian Mixture Model*; GMM）による単音分類実験を行い，GMM よりも SVM が単音分類に適していることを示した [4]．同様に Agostini も，複数の分類方法を比較し，SVM の良さを提示した [27]．このような理由から，本研究の実験では SVM を採用する．

SVM はマージン最大化とカーネルトリックという二つの重要な概念に基づいている．サンプルが線形分離可能である場合，学習データを正確に分類する超平面は数多く存在する．未知のテストデータも正確に分類するために，それぞれのクラスで境界に一番近いサンプルと境界の距離を最大化する．これをマージン最大化とよぶ．また，サンプルが線形分離不可能である場合，サンプルを高次元の空間に射影することで線形分離可能な空間を作ることができる．しかし，高次元空間への写像は計算量がかかるため，SVM ではカーネル関数

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{14}$$

を採用することでこの問題を解決している [28]．これをカーネルトリックという．ここで $\mathbf{x_i}$ と $\mathbf{x_j}$ はそれぞれ学習データ点ベクトルとテストデータ点ベクトル，$k(\cdot)$ はカーネル関数，$\phi(\cdot)$ は写像した特徴空間の要素を表わしており，$\phi(\cdot)$ が生じる空間が特徴空間にあたる．実験においては，libsvm パッケージ [29] を使用した．

4　実験設定と結果

4.1　実験用データ

単旋律の楽器推定実験は，共通のデータベースが存在しないという点で客観的評価や従来手法との比較が困難である．従来の単旋律楽曲における実験で市販の CD から学習データとテストデータを作っていることに倣って，この実験でも市販の CD（サンプリング周波数：44.1kHz）を使用する．市販の CD はそれぞれの楽器ごとに 5 枚ずつ用意し（単音データベース [30] を含む），その中から毎試行ランダムにテストデータと学習データがそれぞれ 449 サンプルずつ選ばれる．楽器は，弦楽器のバイオリン (Vn.)，チェロ (Vc.)，ギター (Gt.)，ピアノ (Pf.)，木管楽器のフルート (Fl.)，オーボエ (Ob.)，金管楽器のホルン (Hr.)，トランペット (Tp.) の 8 種類を含んでいる．つまり，テストデータが合計 3592 個，学習データも合計 3592 個用意する．それぞれのデータは 0.5 秒で，音量の小さいものや無音区間は含まない．0.5 秒というサンプルの長さは，多旋律に応用する際に，サンプル内で旋律の数が変動しないくらい短い長さで，かつ，ある程度の判別率を保つように選んだ．また，サンプルに含まれる基本周波数に対して，特に制限は設けなかった．

4.2　スペクトル分析

各 0.5 秒の音波形は，フーリエ変換によってスペクトル情報に変換される．一般的に，低周波数成分に音の識別に関わる情報が含まれると考えられているため，スペクトル情報の中でも，11020Hz より低い周波数のみを使う．さらに，10Hz ごとに平滑化することで，さらに 1102 次元のスペクトルベクトルで表現する．これの対数をとったものを実験では生対数スペクトル (Lspectrum) 特徴とする[2]．実験における包絡抽出 (LPC, LSF, MFCC) や次元圧縮 (PCA, LDA, LFDA, PCA-LFDA) はこの生対数スペクトルに対して行う．PCA-LFDA パラメータに関しては，まず PCA によって 1102 次元から 289 次元に圧縮した．この 289 次元の主成分は，学習データにおいて，用いる主成分の累積寄与率を 95% から 99.5% まで 0.5% 刻みに変化させ，それらの主成分をさらに LFDA によって 10 次元に変換した特徴を用いて SVM で識別を行った際に，判別率が最適であった値（97.5%）に定めた．（図 1）

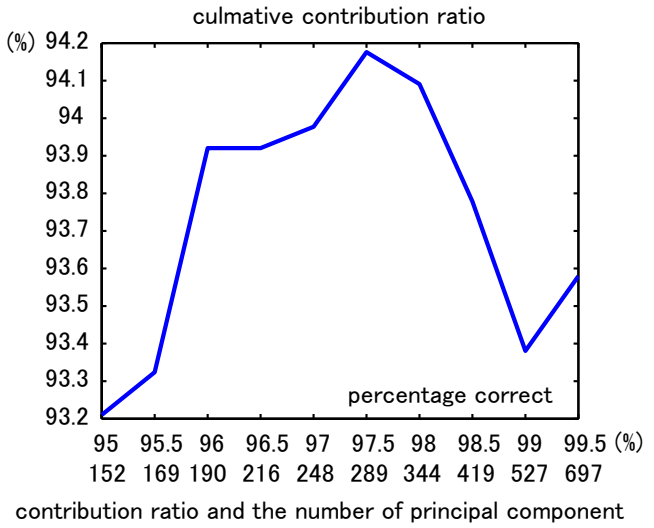


Figure 1: LFDA パラメータの次元を 10 に固定して主成分の累積寄与率を 95% から 99.5% に変化させたときのそれぞれの判別率の変化

特徴パラメータの次元数をどのように設定するかについては，計算量とスペクトル推定の正確さのトレードオフの問題になっている．10 個の LPC 係数で 8kHz でサンプリングされた標準的な音声波形が十分表現できるという従来研究 [31] を参考にして，それぞれの包絡推定法や次元圧縮法のパラメータの数を 10 個に設定する．実験データのサンプリング周波数は 44.1kHz であるが，楽音波形は音声波形よりも音源の特徴を抜き出しやすいと考えられるため，パラメータ数は 10 個で十分であると考えられる．

[2] 生対数スペクトルの生は，次元圧縮していない楽器特徴という意味である．

4.3 楽器判別実験

4.3.1 実験 1: カーネルの選択

最初の実験では，生対数スペクトルが，実際に楽器判別に使えるかどうかと，カーネルによる違いを検証する．前述の 1102 次元の対数スペクトルを用いて 2 種類のカーネルで楽器識別結果を比較した結果が表 1 である．RBF カーネルのパラメータである c と γ はそれぞれ 10-fold クロスバリデーションによって 2 と 1.7 に決めた．この結果

カーネルの種類	線形	RBF
判別精度 (%)	94.51	94.99

Table 1: 対数スペクトル情報のみによる線形と RBF カーネルを用いた SVM での楽器判別結果

から，どちらのカーネルもほとんどのテストサンプルが正しく分類されていることがわかる．このことから，対数スペクトル情報のみである程度の楽器分類が可能であるといえる．以後の実験においては，線形カーネルを使用する．

4.3.2 実験 2: 同じデータでの従来手法との比較

ここでは，過去の実験で楽音特徴として使用された特徴抽出法のいくつかを実装し，対数スペクトル (*Lspectrum*) と，Lspectrum を PCA-LFDA で次元削減したパラメータと同じデータセットを用いて比較している．比較対象として用いた特徴抽出法は 2 章で紹介した MFCCs [4, 12], LPC [4] と LSF [5] の 3 つの包絡を表現するパラメータである．(参考文献の中でこれらの特徴抽出法が楽器分類に使用されている.) 提案手法と比較した結果が表 2 である．表 2 は，全データをランダムに学習データとテストデータに分けて，学習データのみから識別器を学習し，その識別器の学習データ，テストデータそれぞれに対する判別率を調べることを 1 試行としたとき，30 試行の平均値を表示している．この表から，*Lspectrum* が一番良い識別結果となっており，ほぼ完璧に近い識別ができていることがわかる．また，LPC 係数と LSF 係数による楽器判別結果で LSF 係数の結果は約 20 パーセントも高かったが，LPC 係数と LSF 係数が同等のスペクトルを再構成できることを示すために，LPC と LSF で再構成したスペクトルで楽器判別を行った．それらの特徴はそれぞれ *LPC reconst.* と *LSF reconst.* で表現している．また，*Lspectrum* を 10 次元に圧縮した方法（*PCA-LFDA*，機械学習法，包絡圧縮法）の中では提案手法である PCA-LFDA が一番高い識別結果となっている．また，LPC と LSF のパラメータから包絡を再構成した 1102 次元の特徴 (*LPC reconst.* と *LSF reconst.*) で分類した結果は，包絡を圧縮した結果よりも良く，LPC 係数による再構成も LSF 係数による再構成もほぼ同じ結果となった．このことは LPC 係数は丸め誤差の影響で精

特徴	特徴数	テスト 判別率 (%)	学習データ 判別率 (%)
提案手法			
(Lspectrum)	1102	94.51	99.99
(PCA-LFDA)	10	94.07	97.01
機械学習法			
(PCA)	10	79.26	80.38
(LDA)	10	91.26	98.42
(LFDA)	10	77.48	90.17
包絡圧縮法			
MFCCs	10	71.67	72.67
LPCs	10	58.77	58.85
LSFs	10	79.88	80.62
包絡再構成			
LPC reconst.	1102	83.63	86.11
LSF reconst.	1102	83.00	85.63

Table 2: 8 楽器かつテストデータと学習データがそれぞれ異なる 449 サンプルの共通データを使用したときの，提案手法 (PCA-LFDA)，機械学習手法，従来手法で抽出した特徴パラメータによる楽器判別結果の平均値の比較

度が落ち，線スペクトル対によるスペクトル包絡圧縮の方が線形予測分析による圧縮に比べて優れていることと同じ情報を有していても LPC や LSF 係数で楽器特徴で表現するよりスペクトルそのものを楽器特徴とした方が，線形カーネル SVM においては分離しやすい情報になっていることを示している．

4.3.3 PCA-LFDA のパラメータの可視化

PCA と LFDA によって次元圧縮されたパラメータはどのように分布しているのかを見るため，PCA-LFDA パラメータのはじめの 2 次元を可視化したものが図 2 である．各楽器 449 サンプルあるが，サンプルの分布を理解しやすいように，449 個のテストサンプルから 100 個をランダムに選んで表示した．わずか 2 次元であっても，楽器ごとにクラスタができているだけでなく，それぞれ音色が似た楽器のサンプルは近くに分布しており，PCA-LFDA パラメータによるスペクトル情報の次元圧縮がある程度成功していることがわかる．

4.3.4 PCA-LFDA パラメータの混同行列

テストデータによる評価は，それぞれサンプルが異なる組み合わせで 30 回行った．表 3 は，その 30 試行のうちの 1 回のテストデータの *PCA-LFDA* 特徴の混同行列を表わしている．この結果から，多くの誤分類はギターとピアノ間，ホルンとフルートやオーボエ間など，音源特性が似ていると考えられる楽器，もしくは音色に影響を与えている可能性のある基本周波数の音域が近い楽器間で生

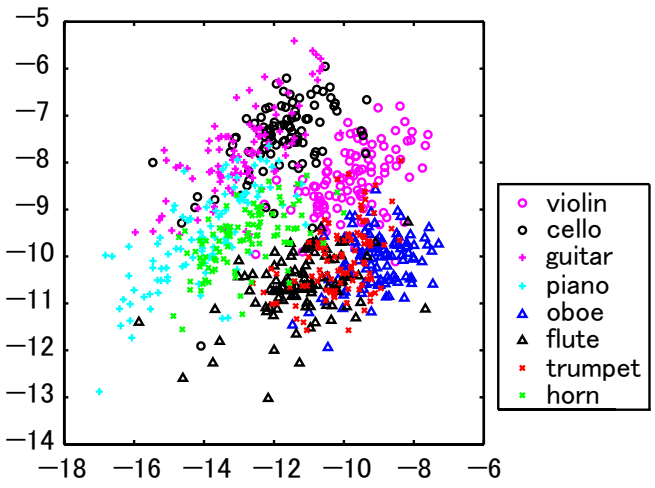


Figure 2: PCA-LFDA で次元削減されたパラメータの 2 次元可視化

	Vn.	Vc.	Gt.	Pf	Fl.	Ob.	Hr.	Tp.
Vn.	445	1	0	0	1	0	2	0
Vc.	0	440	4	0	0	2	3	0
Gt.	1	2	432	12	0	0	0	2
Pf.	0	0	11	428	0	3	0	7
Fl.	1	0	0	0	431	4	13	0
Ob.	1	0	1	1	11	424	7	4
Hr.	2	0	0	0	7	10	429	1
Tp.	6	1	2	3	1	7	1	428

Table 3: PCA-LFDA パラメータを楽器特徴として，8 種類の楽器，各楽器 449 のテストデータで楽器判別をした結果の混同行列（30 試行のうちの 1 試行）

じている．

5 まとめと今後の展開

本稿では，スペクトル情報と機械学習法を用いて単旋律の楽曲における楽器の特徴抽出方法を提案し，線形カーネル SVM によってその特徴を 8 種類の楽器判別に適用した．まずスペクトルの情報だけで楽器判別が可能であることを確認したあと，低次元かつ高精度の楽器特徴パラメータを求めるためにスペクトル情報の次元圧縮方法について考察した．従来研究ではスペクトルから発見的に求めた楽器特徴パラメータを使用していたが，提案手法ではそれらの仮定を取り除き，教師なし学習である主成分分析と教師あり学習である局所フィッシャー判別分析を組み合わせた方法で，スペクトルから直接，楽器判別に重要である部分を抜き出した．さらに，従来研究ではほとんど同じデータを使った実験結果の比較が行われていなかったことが問題であったため，本研究では共通のデータを使用して提案手法と従来手法のうちのいくつかを実装して比較した．今後は，より高精度の楽器判別が行えるように，音の時間変化情報を考慮した判別も行っていきたい．

参考文献

[1] S. Essid, G. Richard, and B. David, “Musical instrument recognition by pairwise classification strategies,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1401–1412, July 2006.

[2] A. Livshin and X. Rodet, “Musical instrument identification in continuous recordings,” in *Proc. of International Conference on Digital Audio Effects (DAFx)*, Oct. 2004.

[3] S. Essid, G. Richard, and B. David, “Musical instrument recognition based on class pairwise feature selection,” in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, Oct. 2004.

[4] J. Marques and P. Moreno, “A study of musical instrument classification using gaussian mixture models and support vector machines,” Compaq Computer Corporation, Tech. Rep., June 1999.

[5] N. Chétry and M. Sandler, “Linear predictive models for musical instrument identification,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, May 2006, pp. 5083–5086.

[6] S. Namba, “Definition of timbre,” *Journal of Acoustical Society of Japan*, vol. 49(11), pp. 823–831, 1993, (in Japanese).

[7] H. von Helmholtz, *On the Sensation of Tone - As a physiological basis for the theory of music*, 2nd ed. New York: Dover, 1954, originally written in 1877. Translated by A.J. Ellis from 4th German edition.

[8] K. Ito and S. Saito, “Effects of acoustical feature parameters of speech on perceptual identification of speaker,” *Transaction of IEICE of Japan*, vol. J65-A, no. 1, pp. 101–108, Jan. 1982, (in Japanese).

[9] L. Deng, *Speech Processing: A Dynamix and Optimization-Oriented Approach*. New York: Marcel Dekker, Inc., 2003.

[10] A. Krishna and T. Sreenivas, “Music instrument recognition: From isolated notes to solo phrases,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2004, pp. 265–268.

[11] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[12] S. Essid, G. Richard, and B. David, "Musical instrument recogntion on solo performance," in *Proc. of European Conference on Signal Processing (EUSIPCO)*, Vienna, Austria, Sept. 2004.

[13] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Transaction of IEICE of Japan*, vol. 53-A, pp. 36–43, 1970, (in Japanese).

[14] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, Apr. 1971.

[15] N. Sugamura and F. Itakura, "Speech data compression by LSP speech analysis-synthesis technique," *Transaction of IEICE of Japan*, vol. 64A, no. 8, pp. 599–606, Aug. 1981, (in Japanese).

[16] F. Itakura and S. Saito, "Analysis synthesis telephony based on the partial autocorrelation coefficient," *Acoustical Society of Japan Meeting*, 1969, (in Japanese).

[17] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of Acoustical Society of Japan*, vol. 57, p. S35, Apr. 1975, (in Japanese).

[18] N. Sugamura and F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signal and its statistical properties," *Transaction of IEICE of Japan*, vol. J64-A, no. 4, pp. 323–330, Apr. 1981, (in Japanese).

[19] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, Mar. 1984, pp. 37–40.

[20] C. Bishop, *Pattern Recognition and Machine learning*. New York, NY: Springer Science+Business Media, LLC, Feb. 2006.

[21] A. Martínez and A. Kak, "PCA versus LDA," *IEEE Transcations on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, Feb. 2001.

[22] K. Fukunaga, Ed., *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston: Academic Press. Inc., 1990.

[23] H. Xiaofei and N. Partha, "Locality preserving projections," University of Chicago Computer Science, Tech. Rep., Oct. 2002.

[24] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," Department of Computer Science, Tokyo Institute of Technology, Japan, Tech. Rep., 2006.

[25] W. Hwang, T. Kim, and S. Kee, "LDA with subgroup PCA method for facial image retrieval," in *International workshop on image analysis for multimedia interactive services (WIAMIS)*, Apr. 2004.

[26] P. Belhumeur, J. Hepanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transcations on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, July 1997.

[27] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," in *Proc. of European Conference on Signal Processing (EUSIPCO)*, vol. 1, 2003.

[28] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[29] C. Chang and C. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[30] The University of Iowa, "Electronic music studios: Musical instrument samples," http://theremin.music.uiowa.edu/MIS.html.

[31] P. Campbell and T. Tremain, "Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1986.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-10

口じゃんけん判定ロボットの開発 ～ロボット聴覚システムの応用に向けて～

Development of A Referee Robot for Rock-Paper-Scissors Sound Games

中臺 一博[†,*] **山本 俊一**[‡] **奥乃 博**[‡] **中島 弘史**[†] **長谷川 雄二**[†] **辻野 広司**[†]

Kazuhiro Nakadai[†], Shunichi Yamamoto[‡], Hiroshi G. Okuno[‡],
Hirofumi Nakajima[†], Yuji Hasegawa[†], Hiroshi Tsujino[†]

†(株) ホンダ・リサーチ・インスティチュート・ジャパン
‡ 京都大学大学院 情報学研究科 * 東京工業大学大学院 情報理工学研究科

†**Honda Research Institute Japan Co., Ltd.** ‡**Kyoto University** * **Tokyo Institute of Technology**
nakadai@jp.honda-ri.com

Abstract

This paper describes a referee robot for "rock-paper-scissors (RPS)" sound games; the robot decides the winner from a combination of rock, paper and scissors uttered by two or three players simultaneously without using any visual information. In this referee task, the robot has to cope with noisy speech due to a mixture of speeches, robot motor noises, and ambient noises. We constructed the referee robot by using a real-time robot audition subsystem and a dialog subsystem focusing on RPS sound games. The former can recognize simultaneous speeches by exploiting two key ideas; *preprocessing* to enchance target speeches with a microphone array, and *system integration* based on missing feature theory. The latter controls RPS sound games as a system-initiative dialog system for multiple players based on deterministic finite automata. The referee system is constructed for Honda ASIMO with an 8-ch microphone array. In the case with two players, we attained a 70% task completion rate for the games on average.

1 はじめに

実環境では，複数の音を同時に聴くことは本質的である．例えば，複数人が参加するゲーム，オークション，築地などの魚河岸では，少なからずこうした能力が求められる．ロボットが人と自然にコミュニケーションを行うためには，やはり，こうした状況を扱う必要があろう．また，日常的な対話でも，対話中に他者からの割り込み発話，所謂バージインが起きる場合もある．従って，ロボットも単一のユーザからの発話だけではなく，複数のユーザによる同時発話を認識する必要がある．

これまで，ロボットの対話処理に関しての研究については，いくつかの報告例がある．例えば，麻生らはオフィス環境で対話を通じて，認識や学習を行うロボット Jijo-2 を報告している [Asoh *et al.*, 1997]. 松坂らは，多人数の対話を想定して，音声や目の動きから話者交代を予測する ROBISUKE を開発した [Matsusaka *et al.*, 1999]. 中野らは，ドメイン推定を行うことによって複数ドメインのタスクが実行可能な対話システムを構築している [Nakano *et al.*, 2005]. Mavridis らは対話を通じた内部表象構築や記号接地問題を扱う状況接地モデル (grounded situation model) を提案し，アームロボットを用いてその有効性を示している [Mavridis and Roy, 2006]. こうした対話ロボットのほとんどは，音声認識性能を向上させるため，ヘッドセットを用いている．これにより，高い S/N 比 (signal-to-noise ratio) で目的音声を得ることができるだけでなく，バージインのような状況も避けるができる．しかし，常にユーザがヘッドセットを装着することは一般的とはいえず，ロボット自身のマイクを用いる方が自然である．従って，実環境を扱うためには同時発話の認識を扱う必要である．

こうした問題を解決するため，我々は，ロボット自身のマイクを用いて，実環境で音環境理解を実現する新しい研究分野として，**「ロボット聴覚」**を提案した [Nakadai *et al.*, 2000]. これまで，音源定位・分離といった信号処理技術の適用を中心に数多くの報告が行われている [Nakadai *et al.*, 2004, Valin *et al.*, 2004, Hara *et al.*, 2004, Yamamoto *et al.*, 2006]. しかし，分離音声認識や人・ロボットインタラクションに向けたロボット聴覚の応用という観点からの報告は少ない．例えば，浅野らは一般的なオフィス雑音環境下で、ＴＶやビデオの制御を行う簡単な対話機能を備えたロボット聴覚システムを開発した [Hara *et al.*, 2004]. しかし，認識対象を一人に絞り，同時複数ユーザへの対応は行っていない．我々は，実際に同時発話認識が可能なロボット聴覚システムを構築したものの，実際のインタラクションに適用する場合の有効性の評価は行っていない．

本稿では、同時発話への対応が不可欠なインタラクションの一例として，手を用いず，発声のみでじゃんけんを行う「口じゃんけん」ゲームに着目し，口じゃんけん判定ロボットを構築した．ロボットには 8 ch のマイクロホンアレイを搭載したホンダ ASIMO を用いた．孤立単語認識率，タスク達成率を通じて評価を行い，ロボット聴覚の応用という観点からその有効性を示す．

2 口じゃんけん判定ロボット

開発したロボット聴覚システムの構成を図 1 に示す．図 1 左に用いたロボット，および，ロボット左側面のマイク配

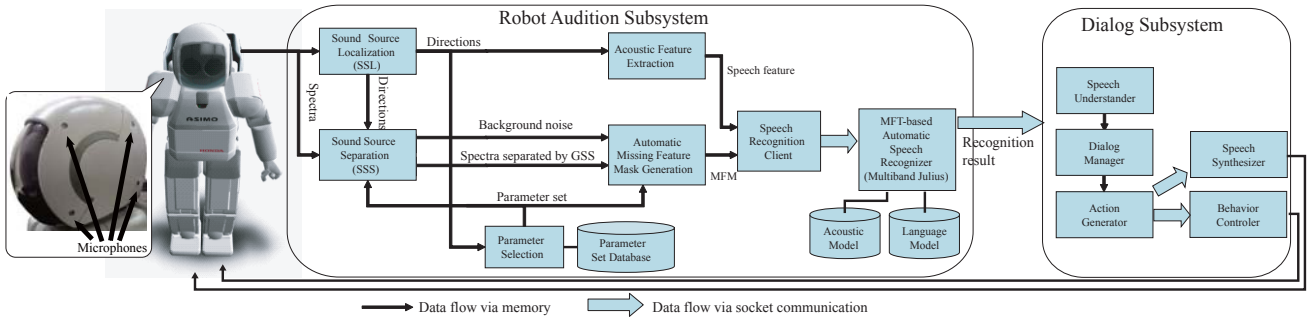


Figure 1: A Referee system for Rock-Paper-Scissors Sound Games

置 (4 ch 分) の拡大図を示した．搭載したマイクロホンアレイは 8 ch であり，左右対称にマイクロホンが配置されている．

システムは，大きく，ロボット聴覚システム (robot audition subsystem) と音声対話システム (dialog subsystem) から構成されている．ロボット聴覚システムは，下記の 7 つのモジュールから構成されている．

- **「音源定位」**(Sound Source Localization (SSL)),
- **「音源分離」**(Sound Source Separation (SSS)),
- **「音声特徴量抽出」**(Acoustic Feature Extraction),
- **「ミッシングフィーチャマスク (MFM) 自動生成」**(Automatic Missing Feature Mask Generation),
- **「パラメータ選択」**(Parameter Selection),
- **「音声認識クライアント」**(Speech Recognition Client),
- **「ミッシングフィーチャ(MFT) 音声認識」**(Missing Feature Theory based Automatic Speech Recognition).

「MFT 音声認識」を除く 6 モジュールは図 1 に示すように複雑で大量のデータ通信が生じる．データフローベースのミドルウェアである *FlowDesigner* [Côté *et al.*, 2004] を用いることによって，このような場合でも，柔軟で高速なモジュール間接続を実現した．**「MFT 音声認識」**については，認識時のＣＰＵ負荷が高いこと，通信量が他のモジュールと比較して少ないことから，別プロセスとして切出し，ネットワーク経由で通信を行うものとした．音声対話システムは，下記の 5 つのモジュールからなる．

- **「音声理解」**(Speech Understander),
- **「対話管理」**(Dialog Manager),
- **「行動生成」**(Action Generator),
- **「音声合成」**(Speech Synthesizer),
- **「行動制御」**(Behavior Controller)

実装上は，最初の 3 つのモジュールは，1 つのプログラムに統合されている．他の 2 つについては，それぞれ別個のプログラムとなっており，ネットワーク経由で通信を行う構成となっている．

2.1 ロボット聴覚システム

ロボット聴覚システムの各モジュールについて説明する．

「音源定位」は MUltiple SIgnal Classification (MUSIC) [Asano *et al.*, 1999]と呼ばれる周波数領域の適応ビームフォーマを用いている．MUSIC は，遅延和ビームフォーマなど他の音源定位手法と比較し，空間スペクトル上で音源方向に対して急峻なピークが得られるため，実環境で高精度な音源定位が可能である．また，8 ch 程度のマイクロホンアレイであれば，実時間処理が可能である．

「音源分離」は，Geometric Source Separation (GSS) とポストフィルタを組み合わせた手法 [Yamamoto *et al.*, 2006] を用いた．GSS は，各音源に相関がないことを仮定して，出力が無相関化されるような処理を行う．この点では，独立成分分析 (ICA) と呼ばれる音源同士の独立性を仮定して分離を行う手法に類似している．しかし，GSS は，ICA と異なり分離時に音源とマイクの位置関係を制約条件として利用する．このため，ICA では分離結果を時間方向に接続する際に生じるパーミュテーション，スケーリングといった問題を扱う必要がないという利点がある．ロボットでは，マイク同士の位置関係は図 1 のように固定であり，音源位置は，移動する場合であっても音源定位によって逐次的に得られるため，この制約を用いることは実際の使用上は特に問題とならない．ポストフィルタは GSS の分離結果に対して，適応的なスペクトルフィルタを用いて音声強調を行う手法である．具体的には，Ephraim & Malah の手法 [Ephraim and Malah, 1984] をマイクロホンアレイ用にマルチチャネルに拡張して利用している．また，ポストフィルタは非線形フィルタであり，音声品質は向上 (10 dB 程度) するが，スペクトルゲインが小さい部分がある場合などは音声認識に悪影響を与える．そこで，ポストフィルタをかけた分離音声に白色雑音を加えることにより，スムーズ化して認識劣化を防ぐ工夫 [Nishimura *et al.*, 2006] も行っている．この手法は，最初から複数の音声音源が同時に存在することを仮定しているため，同時発話のような場合に，より効果的な手法であるといえる．

「音声特徴量抽出」は，分離音声のスペクトルから音声認識用の特徴量を計算する．音声認識システムでは一般にメル周波数ケプストラム係数（MFCC）が特徴量として用いられることが多い [Kawahara and Lee, 2000]．しかし，分離音声に含まれる分離歪みや分離誤りがケプストラム領域では全ての係数に広がってしまい，最適な特徴量とはいえない．本システムでは，MFCC を逆離散コサイン変換したメルスケール対数スペクトル特徴量 [Yamamoto *et al.*, 2006]を利用した．この特徴量は，周波数領域の特徴量であるため，MFCC に比べ特定の周波数帯域に偏在する歪みが扱い安いという特徴がある．具体的には，スペクトル特徴量 24 次元とその一次回帰係数 24 次元で構成される 48 次元の特徴量ベクトルを用いた．

「MFM 自動生成」は，音声認識サブシステムで用いられるミッシングフィーチャマスク (MFM) を生成する．音声認識サブシステムに導入されている MFT [Raj and Stern, 2005] は，分離歪みなどに起因する信頼できない特徴量を MFM によりマスクすることによって認識精度を改善する手法である．MFM は，音声特徴量ベクトルに対応した 48 次元のベクトルとして表現される 2 値（信頼できる場合 1,

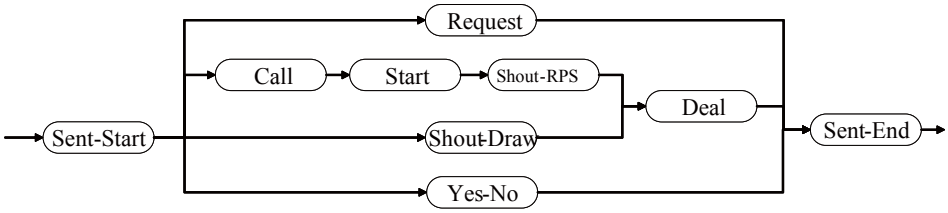


Figure 2: Network Grammar Model

信頼できない場合 0）のマスクである．MFM を正しく推定することが MFT 音声認識の課題となっている [Raj and Stern, 2005]が，我々は音源分離で用いるポストフィルタで推定される他チャネルからのリークエネルギー情報を利用した高精度な MFM 自動生成方法を開発した [Yamamoto *et al.*, 2006]．本稿でもこの手法を用いて自動生成を行っている．

「パラメータ選択」は，音源定位結果から現在の状態に最適なパラメータを選択する．パラメータデータベースは音源位置の組 $\boldsymbol{\theta}(i) = (\theta_1(i), \theta_2(i), \cdots, \theta_M(i))$ ごとに最適パラメータ値の集合が関連付けられている．この最適パラメータ値の集合を $P(\boldsymbol{\theta}(i))$ と表す．M は音源数を表す．パラメータ総数は 11 であり，互いに複雑な依存関係があり，手動での最適化は難しい．このため，遺伝的アルゴリズム (GA) を用いて，各方向ごとの最適パラメータの組を導出した．ϕ_m を音源 m の方位角とすると，時刻 t の音源定位結果が $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_M)$ のとき，以下の式を満たすパラメータセット $P(\boldsymbol{\theta}(i))$ が選択される．

$$\forall m \, |\phi_m - \theta_m(i)| < \theta_\delta \tag{1}$$

ここで，θ_δ は ϕ_m を $\theta_m(i)$ に割り当てるための閾値である．

「音声認識クライアント」は，**「MFT 音声認識」**に対してソケット通信で，音響特徴量と MFM を送信する．

「MFT 音声認識」は，入力音声の音響特徴量を MFM 情報に基づき，マスクする処理が追加されているものの，入力特徴量から，音響モデルや言語モデルを参照しながら，隠れマルコフモデル（HMM）を用いて，音素の列を推定するという点では一般的な音声認識と同様である．マスク情報を利用できるよう HMM から音響スコアを計算する部分の処理に変更が加えられている．音響スコアは遷移確率と出力確率に基づいて計算され，MFT に基づく音声認識では以下のように定義される．

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \tag{2}$$

ここで，$M(i)$ は i 次元目の特徴量に対する MFM を，$b_j(x)$ は出力確率を表す．また，$P(\cdot)$ は確率を，$x(i)$ は特徴量ベクトルを表す．N は特徴量ベクトルの次元数を，S_j は j 番目の状態を表す．$M(i) = 1$ とすれば，一般的な音声認識の尤度計算と同じになる．

MFT に基づく音声認識の実装として，Julian [Kawahara and Lee, 2000] をベースに MFT に基づく音声認識を行えるように改良したマルチバンド版 Julian[1] を利用した．音響モデルは，音源分離の出力として得られる音声データを用いて学習を行っている．これにより，音源分離の性質を考慮した音響モデルを用いることができ，MFT との併用で，分離音声の認識率を大きく向上することができた [Yamamoto *et al.*, 2006].

[1] http://www.furui.cs.titech.ac.jp/mband_julius/

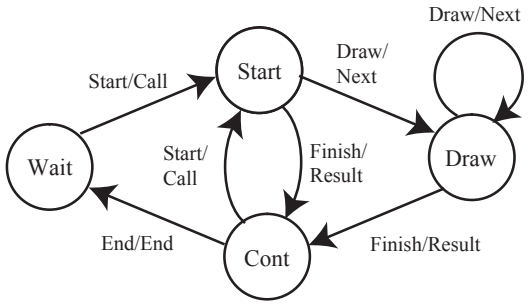


Figure 3: Dialog State Transition

Table 1: Request, State, and Output for Dialog Subsystem

a) Dialog Requests

Request	Description
Start	a request when a game-start command is detected
Draw	a request when a draw occurs
Finish	a request when the winner was found
End	a request when a game-end command is detected

b) Dialog States

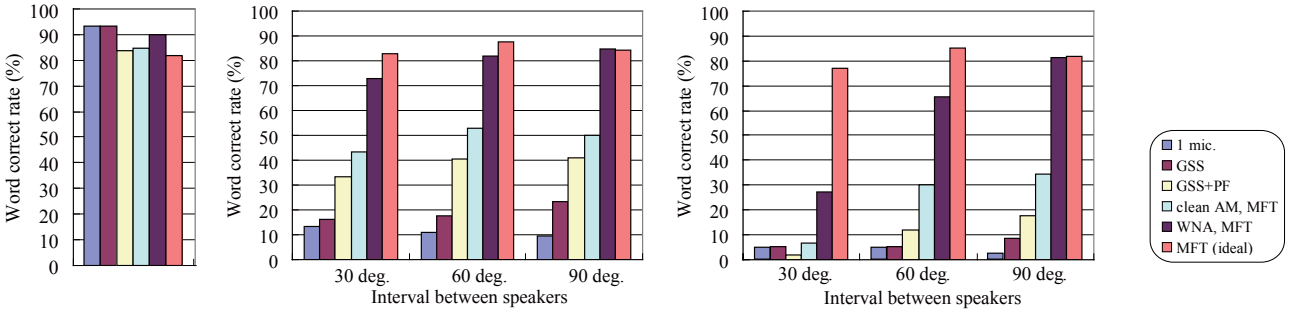
State	Description
Wait	waiting for a start trigger command from a player
Play	waiting for a speech mixture of rock, paper, and/or scissors
Draw	**Play** after a draw.
Cont	waiting for a reply to a question if the players want to start the next game

c) Dialog Outputs

Output	Description
Call	Urge each player to say one of three words
Next	**Call** after a drawn case
Result	Inform the players of the winner with gestures and speech. For gestures, ASIMO turns its face to and points its hand to the winner by using a winner direction obtained from sound source localization. For speech, ASIMO answers the winner's choice like "the player who said rock is the winner."
End	Just finish the game without any gesture and speech

3 音声対話システム

口じゃんけん用の音声対話システムについて述べる．口じゃんけんでは，各プレーヤは手を使う代わりに,「グー」,「チョキ」,「パー」の中から任意の一つを選択して，その言葉を発声する．判定ロボットは，同時発話音声から勝者を判定するため，同時発話認識機能が必要となる．システムは，対話の状態ごとに利用者の発話内容を限定して対話制御を行うシステム主導型の音声対話システムとして実装した．一般的な音声対話システムでは，利用者が一人であることが想定されているが，このシステムでは，



(a) Single Speaker(T1) (b) Two Simultaneous Speakers(T2) (c) Three Simultaneous Speakers(T3)

Figure 4: Word correct rates of three simultaneous speakers with our system

複数人のユーザが同時に発話することを許容する構成となっているのが特長である．

「MFT 音声認識」の言語モデルは，図 2 に示すようなネットワーク文法を用いている．図中の “Request” は,「じゃんけんをしましょう」といったゲーム開始契機の待ちうけに対する文法，“Call”, “Start”, “Shout-RPS”，“Deal” はゲーム中に用いられる「最初はグー、じゃんけんパー」といった発話に対応した文法を定義している．“Shout-Draw” は，あいことなった後の発話である「あいこでグー」などに相当する文法が定義されている．“Yes-No” は，再度ゲームを行うかどうかという問いに対する回答に対応した文法が定義されている．

「音声理解」 は，ロボット聴覚システムの分離音声認識結果から，表 1a) に示す 4 種類の対話要求を生成する．口じゃんけんのタスク内容は比較的単純であるため，一人のユーザの発話を理解するのは容易である．しかし，口じゃんけんでは複数のユーザの発話から発話内容を同時に理解する必要がある．この場合，送られてくる認識結果が，単一発話に対応した認識結果であるか，同時発話に対応した認識結果であるかを検出する必要がある．認識結果は発話が終了した順番で**「音声理解」**に送られるよう設計されているため，発話終了時刻に基づき単一発話・同時発話の検出を行った．まず，認識結果到着後は，他に同時発話である認識結果が到着する可能性があるので一定期間待機する．同時発話の認識結果が到着した場合は，待機時間を延長し，次の同時発話認識結果の到着を待つ．しかし，時間が経つにつれて同時発話である可能性は低くなるので，待機する時間を短くする．具体的な同時発話判定のアルゴリズムは下記の通りである．

1. 同時発話集合 S の初期状態を空集合とする．
2. 時刻 t に，認識結果が到着した場合，この認識結果を r_1 として，$S \leftarrow S \cup \{r_1\}$ とし，次の認識結果までの待ち受け時間 $\tau = T$ とする．
3. τ 時間以内に次の認識結果が到着し続ける間，以下の処理を繰り返す．
   - 認識結果を r_{i+1} とする (ただし，i は S の要素数)．
   - $S \leftarrow S \cup \{r_{i+1}\}$ と S の更新を行う．
   - $\tau = \frac{T}{i+1}$ と τ の更新を行う．
4. 最終的な集合 S の要素数が 1 であれば単一発話であり，複数であれば同時発話と判断する．
5. 1 に戻り，また次の同時発話を検出を開始する．

なお，T は実験的に 1 sec とした．

「対話管理」は，対話要求に応じて，対話の状態遷移を管理する．表 1b) に示す 4 種類の状態を定義した．状態遷移は決定性有限オートマトン（Deterministic Finite Automata, DFA）を利用して，図 3 のように定義した．**「対話管理」**が対話要求を受け取った場合，遷移した対話の状態に応じて，表 1c) に示す 4 種類の対話出力を行う．

「行動生成」は，対話出力を受けて，ロボットのジェスチャ生成コマンドと音声合成コマンドを発行する．**「行動制御」**では，ジェスチャ生成コマンドを受け，実際にロボットを制御してジェスチャを行う．**「音声合成」**では，音声合成コマンドを受け，音声を合成し出力する．これらのコマンドの通信はネットワーク経由で行われる．

4 評価

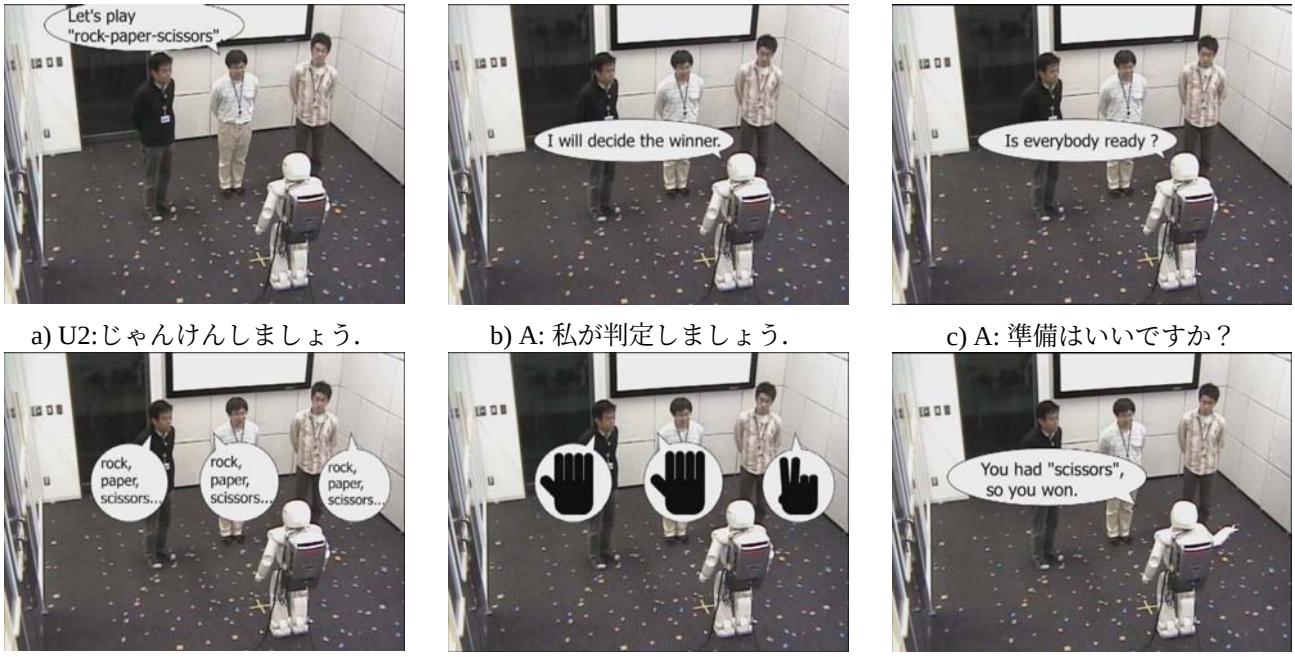
開発した口じゃんけん判定ロボットを評価するため，下記の 2 種類の実験を行った．

1. 同時発話認識実験 – 同時発話 (1,2,3 話者) に対する孤立単語正解率算出
2. 口じゃんけん判定実験 – 口じゃんけん判定タスクのじゃんけん判定率・タスク達成率算出

実験 1 は，ロボット聴覚システムの性能確認，実験 2 は，システムトータルの性能確認のための実験である．どちらも使用した環境は 4 m × 7 m の大きさであり，三方が吸音壁，一方がガラス壁という反響が一様でない部屋を用いた（残響時間 (RT_{20}) は 0.3 秒程度).

4.1 同時発話認識実験

同時発話者数が 1 – 3 の場合に正面話者に対する孤立単語正解率を測定した．同時発話者数に対応して T1 – T3 の 3 種類の認識用の音声データセットを用意した．T1 – T3 はそれぞれ，200 セットの同時発話からなっている．各同時発話セットの組合せは，男性、女性各 5 名ずつの ATR 音素バランス単語 216 語から任意に音声データを選択することによって決定した．この同時発話の組合せを，ASIMO の電源を ON にして，スピーカ (Genelec 1029A) から出力し，ロボットに搭載したマイクロホンアレイで実際に収録した．各スピーカとロボットの距離は 1.5 m とした．T1 では，スピーカはロボットの正面方向に設置した．T2 では，一方をロボット正面に固定し，もう一方をロボットの右 30, 60, 90 度に変更して 3 種類の収録を行った．T3 では，1 本のスピーカは正面方向に固定し，他の 2 本のスピーカは左右対称に ±30 度から ±90 度まで 30 度おきに変更して 3 種類の測定を行った．



a) U2:じゃんけんしましょう. b) A: 私が判定しましょう. c) A: 準備はいいですか？

d) U1-U3: 最初はグー，じゃんけん e) U1: パー U2: パー U3: チョキ f) A: チョキを出したあなたの勝ちです.

Figure 5: Snapshot of a RPS sound game (A: ASIMO，U1:left player，U2:center player，U3: right player)

収録した認識データに対して，以下の 6 つの条件で音声認識実験を行った.

(1) ASIMO 頭部の左前のマイクロホンの入力に対して，前処理を行わず，単純に音声認識実験を行った．クリーン音響モデルを用いた.
(2) 前処理として GSS を用いた音源分離を行い，分離音声を直接認識した．クリーン音響モデルを用いた.
(3) 前処理として GSS に加え Post-filter を用いた上で，認識を行った．クリーン音響モデルを用いた.
(4) 音声認識に，**「MFM 自動生成」**で推定された MFM を用いて MFT 音声認識を行った．その他の条件は (3) と同様とした.
(5) 音響モデルに白色雑音重畳したデータで学習を行った白色雑音重畳モデルを用いた．その他の条件は，(4) と同様とした.
(6) MFM の自動生成を行わず正解の MFM を与えた以外は，(5) と同様の条件で認識を行った．この条件では，正解マスクを用いるため，MFT 音声認識の性能の上限を知ることができる.

なお，クリーン音響モデルの学習には日本語新聞記事読み上げコーパス (JNAS) を用いた．従って，実験環境は話者・語彙ともにオープンである．白色雑音重畳モデルは，JNAS の各データにピークパワーの -40 dB 程度の白色雑音を重畳して，マルチコンディション学習を行い，作成した．音響モデルは，いずれも 3 状態 4 混合の triphone HMM を用いた.

T1 – T3 に対する認識実験結果を図 4 に示す. 自動生成した MFM を用いた MFT 音声認識が効果的に働いていることがわかる．また，白色雑音重畳音響モデルを用いた方が，正解率が高いことがわかる．白色雑音重畳音響モデルは環境や条件に関する先見的な知識を用いずに作成でき，性能も高くなることから，クリーン音響モデルを用いるよりもロボット聴覚システムに適しているといえる．特に T3 に関しては，スピーカ間の角度が狭くなると正面話者から見て妨害音源の影響が強くなるため，性能が劣化することがわかる．正解マスクを与えた場合は，さらに性能が高いことから，MFM の自動生成アルゴリズムの改良によりさらなる改善が期待できる.

4.2 口じゃんけん判定実験

プレーヤが 2 名の場合の口じゃんけん判定タスクの評価を行った．実験は，このシステムの熟練者である A 氏, B 氏，および，初めてシステムを利用した C 氏, D 氏の 2 組のペアで行った．各ペアには，じゃんけんの勝負が決するまでを 1 回と数えて，1 セット 10 回，計 11 セットの口じゃんけんを行ってもらった．前半の 6 セットは，A 氏, C 氏はロボットの正面方向に立ってもらった．B 氏, D 氏にはセットが変わるたびに −30 度から −180 度まで 30 度間隔で移動してもらった．後半 5 セットは，2 人のプレーヤには左右対称となるように ±30 度から ±150 度まで 30 度おきに移動してもらった．ASIMO は，部屋の中央に配置し，各プレーヤとロボットの距離は 1.5 m とした．図 5 プレーヤが 3 名の場合のスナップショットを示す．この例では，勝者が一意に決まっているが，決まらない場合は「あいこ」として扱われ，勝者が決するまでじゃんけんを行うシステムとなっている.

評価指標として，口じゃんけん判定成功率とタスク達成率を用いた．口じゃんけん判定性効率は，口じゃんけん判定機会の総数に対する判定成功数の比であり，判定は，「勝ち」「負け」「あいこ」のいずれかであるので，チャンスレートは 33%となる．これに対して，タスク達成率は，勝負が決するまでを一回と数えた場合の総回数に対する勝者判定成功数の比を示す．実際には，この 2 種類の指標は，あいこがまったくなければ同じ値となる.

表 2 に実験結果を示す．平均では，口じゃんけん判定成功率は 75%であり，タスク達成率は 70%であった．A 氏と B 氏はシステムの熟練者であるにもかかわらず，未経

Table 2: Result of Two-Speaker Task (Rock-Paper-Scissors Games)

speaker		judgment		task completion	
Mr.A (deg.)	Mr.B (deg.)	#success/ #judgments	success rate (%)	#success/ #tasks	completion rate(%)
0	-30	11/15	73.3	6/10	60
0	-60	9/11	81.8	8/10	80
0	-90	9/12	75.0	8/10	80
0	-120	11/13	84.6	8/10	80
0	-150	6/14	42.9	2/10	20
0	-180	8/11	72.7	7/10	70
30	-30	17/18	94.4	9/10	90
60	-60	7/12	58.3	5/10	50
90	-90	12/16	75.0	6/10	60
120	-120	15/17	88.2	8/10	80
150	-150	17/18	94.4	9/10	90
average		112/157	71.3	76/110	69.1

speaker		judgment		task completion	
Mr.C (deg.)	Mr.D (deg.)	#success/ #judgments	success rate (%)	#success/ #tasks	completion rate(%)
0	-30	12/16	75.0	6/10	60
0	-60	10/11	90.9	9/10	90
0	-90	9/13	69.2	6/10	60
0	-120	17/21	90.0	6/10	60
0	-150	10/14	71.4	6/10	60
0	-180	9/11	81.8	8/10	80
30	-30	14/19	73.7	5/10	50
60	-60	11/13	84.6	8/10	80
90	-90	8/11	72.7	7/10	70
120	-120	11/12	91.7	9/10	90
150	-150	10/11	90.9	9/10	90
average		121/152	79.6	79/110	71.8
total		233/309	75.4	155/220	70.5

験者であったC氏とD氏の方がよい結果となった．HMMベースの音声認識システムの性能には，個人差があることが知られているため，この例からだけでは，結論付けることは難しいが，この結果は，構築したシステムが話者に依らず有効であることをある程度示唆しているのではないかと考えている．また，プレーヤの立ち位置が変わっても大きな性能差は見受けられない．当初，ロボットの雑音が大きい，ロボット後方では性能が劣化すると考えていたが，この結果は，ロボット聴覚システムがロボットから発生するノイズもうまく扱えること支持していると考えられる．

5 おわりに

本稿では，ロボット聴覚システムの応用として口じゃんけん判定タスクを取り上げ，口じゃんけん判定ロボットの構築について説明した．また，実験を通じて，ロボット聴覚システムの有効性，および，口じゃんけんなど同時発話認識が必要なタスクへの適用が有効であることを示した．実際に，同時発話認識など複数の音源を同時に認識する機能は，ロボットが実環境でロバストに音環境を理解するうえで本質的な機能であるといえる．本稿では，同時発話認識に着目して，ロボット聴覚システム自体，およびその応用に向けた評価を行った．しかし，理論的には，ロボット聴覚システムは，音声以外の方向性雑音やある程度定常な非方向性雑音に対しても頑健に働くことが期待できる．今後は，様々な状況への適用を通じて，ロボット聴覚システムの応用の可能性を探っていく予定である．

謝辞

CSIRO の Jean-Marc Valin 氏に感謝する．

参考文献

[Asano *et al.*, 1999] F. Asano, H. Asoh, and T. Matsui. Sound source localization and signal separation for office robot “Jijo-2”. In *Proc. of IEEE Int’l Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI-99)*, pages 243–248, 1999.

[Asoh *et al.*, 1997] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui. Socially embedded learning of the office-conversant mobile robot *jijo-2*. In *Proc. of 15th Int’l Joint Conf. on Artificial Intelligence (IJCAI-97)*, volume 1, pages 880–885. AAAI, 1997.

[Côté *et al.*, 2004] C. Côté, D. Létourneau, F. Michaud, J.-M. Valin, Y. Brosseau, C. Raievsky, M. Lemay, and V. Tran. Code reusability tools for programming mobile robots. In *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2004)*, pages 1820–1825. IEEE, 2004.

[Ephraim and Malah, 1984] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-32(6):1109–1121, 1984.

[Hara *et al.*, 2004] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto. Robust speech interface based on audio and video information fusion for humanoid HRP-2. In *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2004)*, pages 2404–2410. IEEE, 2004.

[Kawahara and Lee, 2000] T. Kawahara and A. Lee. Free software toolkit for japanese large vocabulary continuous speech recognition. In *Int’l Conf. on Spoken Language Processing (ICSLP)*, volume 4, pages 476–479, 2000.

[Matsusaka *et al.*, 1999] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface — a robot who communicates with multi-user. In *Proc. of Eurospeech-1999*, pages 1723–1726, 1999.

[Mavridis and Roy, 2006] N. Mavridis and D. Roy. Grounded situation models for robots: Where words and percepts meet. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 4690–4697. IEEE, 2006.

[Nakadai *et al.*, 2000] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proc. of 17th National Conf. on Artificial Intelligence (AAAI-2000)*, pages 832–839. AAAI, 2000.

[Nakadai *et al.*, 2004] K. Nakadai, D. Matasuura, H. G. Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, 44(1-4):97–112, October 2004.

[Nakano *et al.*, 2005] M. Nakano, Y. Hasegawa, K. Nakadai, T. Nakamura, J. Takeuchi, T. Torii, H. Tsujino, N. Kanda, and H. G. Okuno. A two-layer model for behavior and dialogue planning in conv ersational service robots. In *Proc. of IEEE/RSJ International Conference on Intelligent Robot s and Systems (IROS-2005)*, pages 1542–1547, 2005.

[Nishimura *et al.*, 2006] Y. Nishimura, M. Ishizuka, K. Nakadai, M. Nakano, and H. Tsujino). Speech recognition for a humanoid with motor noise utilizing missing feature theory. In *Proc. of 6th IEEE-RAS Int’l Conf. on Humanoid Robots (Humanoids 2006)*, pages 26–33, 2006.

[Raj and Stern, 2005] Bhiksha Raj and Richard M. Stern. Missing-feature approaches in speech recognition. *Signal Processing Magazine*, 22(5):101–116, 2005.

[Valin *et al.*, 2004] J.-M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2004)*, pages 2123–2128. IEEE, 2004.

[Yamamoto *et al.*, 2006] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J.-M. Valin, K. Komatani, T. Ogata, and H. G. Okuno. Real-time robot audition system that recognizes simultaneous speech in the real world. In *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2006)*, pages 5333–5338, 2006.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A702-11

小型音声対話モジュールによる耐雑音音声認識
A Noise-Robust Speech Recognition on a Compact Speech Dialogue Module

○佐藤 幹　　岩沢 透　　杉山 昭彦

NEC 共通基盤ソフトウェア研究所

*Miki SATO,　Toru IWASAWA,　Akihiko SUGIYAMA

NEC Common Platform Software Research Laboratories

m-sato@dh.jp.nec.com　t-iwasawa@bp.jp.nec.com　aks@ak.jp.nec.com

Abstract—This paper presents implementation and evaluation of noise-robust speech recognition on a compact speech dialogue module. This module consists of direction of arrival (DOA) estimation, noise cancellation, speech recognition, and text-to-speech (TTS) conversion. DOA estimation is essential to adjust the microphone look-direction toward the speaker. Noise cancellation is useful to reduce undesirable influence by ambient noise and interference. Speech recognition helps understand the input speech and TTS conversion transforms text information into an audible form for the user. These functions, equipped with in a personal robot PaPeRo, are implemented on an application processor that was developed primarily for mobile phones. Evaluation results with PaPeRo-mini which is a scaled-down version of PaPeRo with this module, demonstrates 54% improvement in the speech recognition rate in noisy environment, and an 85% correct rate in DOA estimation without any negative effect of downsizing.

1. はじめに

近年、ロボットやカーナビ等の様々な機器のインターフェースとして、音声対話機能が注目されている。これらの機器は、通常、音声コマンドによって、離れた位置から制御される。実環境でスムーズな音声対話を実現するために、様々な雑音が存在する環境に対処できる耐雑音性能が求められる。耐雑音性能を向上させ、正確に音声コマンドを認識するために、背景雑音や妨害信号の影響を低減する指向性マイクロホンが広く使われている。このため、音声の到来する方向を推定し、推定方向にマイクロホンの指向性を一致させることが重要となる[1]。マイクロホンの指向性だけで抑圧できない雑音や妨害信号は、音声強調処理によって、その影響を軽減する。応用毎に異なる要求条件に応じて、1 つ又は多数のマイクロホンを用いた雑音及び妨害信号の抑圧が、広く行われている[2]。様々な機器に組込まれる音声対話機能においては、音声用と雑音用の 2 つのマイクロホンを用いて雑音の消去を行う適応ノイズキャンセラが、マイクロホン数、雑音除去性能、及び歪の観点から見て、良い妥協策である。これまでに、係数更新ステップサイズを音声対雑音比（SN 比）に応じて制御することで、高い雑音消去性能と小さな音声歪を両立することができるノイズキャンセラが提案され、ロボットにおける音声認識に適用されている[3]。

一方、これらの従来技術を実際の端末に適用する際には、小さな空間に複数の機能をコンパクトに搭載する必要があり、音声対話機能を有する端末の開発にとって障害となっている。そのため、実環境での音声対話に必要な機能を、小型・低消費電力・低コストの組込用モジュールとして提供することができれば、様々な端末における音声対話機能搭載を促進することができる。

本稿では、小型音声対話モジュールによる耐雑音音声認識について報告する。携帯電話用アプリケーションプロセッサを利用することによって小型・低消費電力・低コストを達成し、実環境においてパーソナルロボット PaPeRo[4]における音声対話と同等の機能を提供する。以下、2 節で音声対話モジュールの機能、3 節でハードウェアの構成、4 節で実装形態について説明する。5 節では、本モジュールを搭載した PaPeRo-mini について紹介し、6 節で PaPeRo-mini を用いた実環境における性能評価結果を示す。

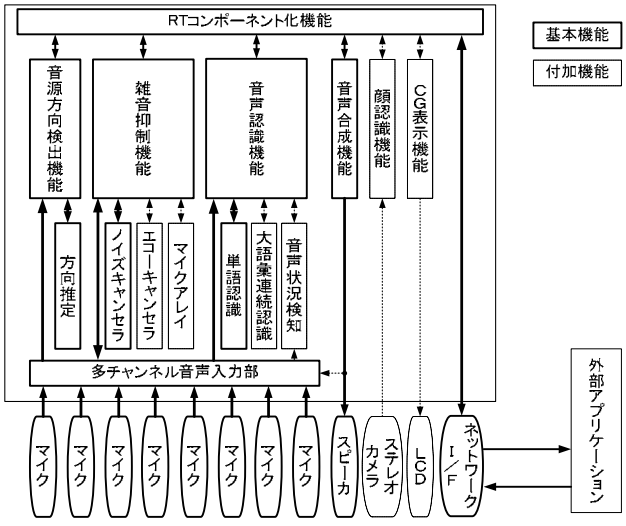


Figure 1. 音声対話モジュールの機能と構成

2. 音声対話モジュールの機能

音声対話モジュールの機能と構成を Fig.1 に示す。本モジュールは、実環境における音声対話の基本機能として、音声認識機能（単語認識機能）、雑音消去機能（ノイズキャンセラ機能）、音源方向検出機能、音声合成機能を有する。また、付加機能として、音声認識機能（大語彙連続認識機能、音声状況検知機能）、顔認識機能、ＣＧ表示機能を有する。これらの機能は、RT (Robot Technology) ミドルウェア[5]に対応した RT コンポーネントとして動作し、ネットワーク接続された外部アプリケーションから利用することが可能である。ここでは、本モジュールの基本機能について説明する。

2. 1. 音声認識機能

音声認識機能は、音声認識辞書に登録した単語を対象とした音声認識処理を行う。音声認識辞書を外部アプリケーションから供給することによって、様々な応用に対応できる。音声認識開始コマンドを受信することで動作を開始し、音声認識停止コマンドを受信するまで連続的に音声認識処理と認識結果の出力を行う。音声認識辞書は、W3C 勧告の記述仕様である SRGS (Speech Recognition Grammar Specification)[6]のサブセットに基づいて、XML 形式で記述する。音声認識辞書の一例を Fig.2 示す。

認識対象語彙は rule タグで表記された認識ルールを単位として記述する。音声認識辞書には複数の認識ルールを登録することができ、外部アプリケーションが音声認識開始時に認識ルールを選択できる。読みと表記に加えて、特定の意味の認識単語をまとめたグループを各認識単語に記載することによって、認識結果として取り出し、これを利用することができる。また、認識単語を連結した連続音声認識の記述も可能である。Fig.2 は、「ぜんしん/こうたい」と「して/してください」が連結された文、すなわち「ぜんしんして」「ぜんしんしてください」「ばっくして」「ばっくしてください」の 4 文を受理できることを示している。また、音声認識や音声検出の開始・終了などに代表される音声認識エンジン状態の通知機能や、不要雑音を認識単語と誤認識させないためのリジェクション機能[7]も提供する。

2. 2. 雑音消去機能（ノイズキャンセラ）

雑音消去機能は、音声用及び雑音用の 2 つのマイクロホンの入力信号を用いて、音声側入力信号に混入する雑音成分を推定して、消去する。外部アプリケーションから開始コマンドを受信することで動作を開始し、停止コマンドを受信するまで雑音消去処理を連続して行う。音声認識の前処理としてこの機能を使用することで、雑音下における音声入力の SN 比を向上させ、音声認識精度を向上させることができる。本機能は、雑音用マイクロホンに混入する音声成分の推定も行うことにより、低歪の雑音消去が可能である[3]。

```
<rule id="Move">
 <one-of>
  <item>ぜんしん<tag>grapheme="前進"group="前進"</tag>
  </item>
  <item>ばっく<tag>grapheme="バック"group=" 後退" </tag>
  </item>
 </one-of>
 <rulerefuri="#garbage">
</rule>
<rule id="garbage">
 <one-of>
  <item>して</item>
  <item>してください</item>
 </one-of>
</rule>
```

Figure 2. 音声認識辞書の記述例

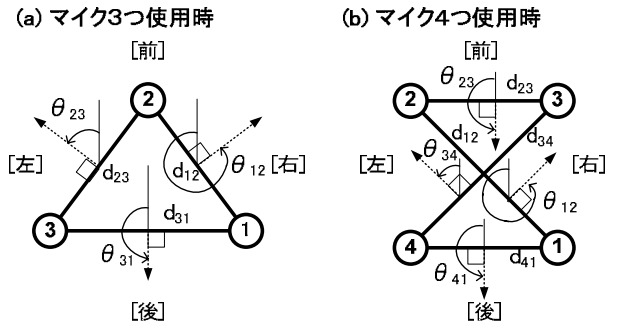


Figure 3. 音源方向検出のマイクロホン配置

2. 3. 音源方向検出機能

音源方向検出機能は、3 または 4 つのマイクロホンの入力信号を用いて、音の到来方向を推定する。外部アプリケーションから開始コマンドを受信することで動作を開始し、停止コマンドを受信するまで連続して音源方向検出処理を行う。検出結果として、0～360 度（反時計回り）の方向、および音量を出力する。使用するマイクロホン配置は、マイクロホン数、マイクロホンペア間の距離、及びマイクロホンペアを結ぶ直線の垂線方向により設定する。Fig.3 に示すように、マイクロホンを 3 つ使用する場合は、マイクロホンペア 3-1 を長辺とする二等辺三角形に配置し、マイクロホンペア 1-2、2-3、3-1 の距離 d12、d23、d31、垂線方向 θ12、θ23、θ31 を設定する。マイクロホンを 4 つ使用する場合は、マイクロホン 2-3、4-1 が平行になるように配置し、マイクロホンペア 1-2、2-3、3-4、4-1 の距離 d12、d23、d34、d41、垂線方向 θ12、θ23、θ34、θ41 を設定する。本機能は、近接音場を想定した方向推定を行うことにより、音源とマイクロホンが同一平面上になく、十分に距離が離れていない場合も、高精度な方向検出が可能である[8]。

2. 4. 音声合成機能

音声合成機能は、入力したテキストデータを音声信号に変換し、読み上げを行う。外部アプリケーションから、読み上げテキストを引数とした読み上げコマンドを受信することで動作し、合成された音声信号がスピーカから出力される。

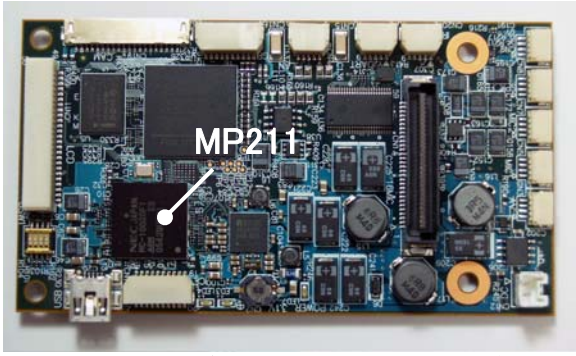


Figure 4. 音声対話モジュールの外観

Table. 1 音声対話モジュールの仕様

項目	仕様
CPU	ARM9(192MHz) ×3, DSP(SPXK6 192MHz)×1
メモリ	128MB, +128MB (拡張ボード使用時) フラッシュ 64MB
音声 I/F	マイクロホン入力 2ch ×2 系統 +16ch (AUDIO ボード使用時) スピーカ出力 2ch
画像 I/F	カメラ入力×2 系統、 ビデオ出力, LCD 出力
その他 I/F	USB, LAN, IrDA, GPIO CF Card (拡張ボード使用時)
サイズ	55mm × 100mm × 32mm (AUDIO ボード、拡張ボード使用時)

3. 音声対話モジュールのハードウェア

音声対話モジュールの外観を Fig.4 に、仕様を Table 1 に示す。本モジュールは、NEC エレクトロニクス社製携帯電話用アプリケーションプロセッサ MP211（ARM9×3, DSP×1 で構成）を搭載し、Linux OS によって動作する。

音声入力インターフェースとして、メインボード上に 2 チャンネル同期マイクロホン入力 2 系統を備え、拡張用 AUDIO ボードにより、16 チャンネル同期マイクロホン入力を増設することが可能である。その他に、スピーカ出力×2、カメラ入力×2、LCD 出力、USB、LAN などの周辺インターフェースを備え、拡張ボードにより、メモリカードの使用が可能である。

4. 基本機能の実装形態

音声対話モジュールの基本機能は、MP211 上の ARM9(192MHz) 1 個と DSP(192MHz)1 個により実現している。ARM-DSP 間の基本機能分割を Fig.5 に示す。音声認識機能、音声合成機能、RT コンポーネント化機能は ARM 上で動作する。雑音消去機能、音源方向検出機能は、ARM 上で動作する雑音消去制御部、音源方向検出制御部と、DSP 上で動作するノイズキャンセラコア部、方向推定コア部から構成される。マイクロホンに入力された信号は、11025Hz でサンプリングされ、ARM 用メモリに配置されたマルチリングバッファに書き込まれる。以下、ARM と DSP に分割された雑音消去制御部、ノイズキャンセラコア部、音源方向検出制御部、方向推定コア部について説明する。

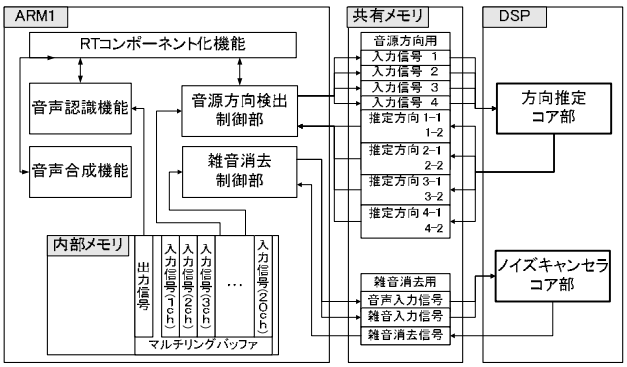


Figure 5. ARM-DSP 間の基本機能分割

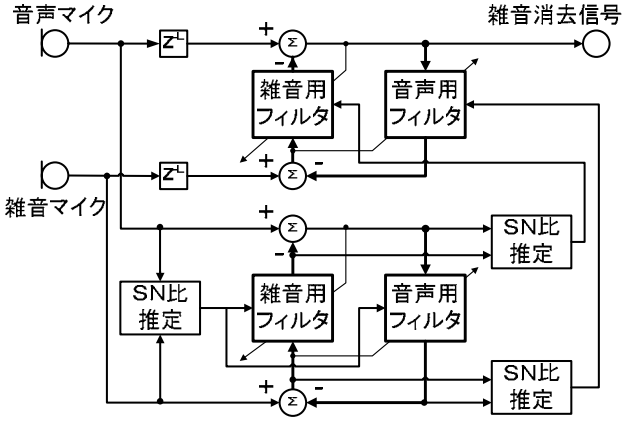


Figure 6. ノイズキャンセラコア部の構成

4. 1. 雑音消去制御部

雑音消去制御部は、外部から動作開始命令を受けると、音声用マイクロホンと雑音用マイクロホンの入力信号をマルチリングバッファから取り出し、ARM-DSP 間の共有メモリにコピーする。ノイズキャンセラコア部の処理が終了すると、共有メモリ経由で雑音消去信号を取得し、マルチリングバッファを介して音声認識機能に出力する。

4. 2. ノイズキャンセラコア部

ノイズキャンセラコア部の構成を Fig.6 に示す。ノイズキャンセラコア部は、入力信号を共有メモリから取得し、音声側入力に混入する雑音成分と、雑音側入力に混入する音声成分を、それぞれ個別の適応フィルタで推定し、推定結果を各信号から差し引くことにより雑音消去信号を生成し、共有メモリに書き込む。雑音成分用と音声成分用の適応フィルタは、それぞれに専用のサブ適応フィルタを用いて推定した SN 比でステップサイズを制御し、大きな消去量と小さな音声歪を両立する[3]。ノイズキャンセラコア部と共有メモリ間の入出力信号転送は 512 サンプル毎に行う。総演算量は 78MIPS である。

4. 3. 音源方向検出制御部

音源方向検出制御部は、外部から動作開始命令を受けると、マイクロホン数（3 または 4）に対応した入力信号をマルチリングバッファから取り出し、共有メモリを介して方向推定コア部に伝達する。方向推定コア部の処理が終了すると、共有メモリ経由でマイクロホンペア数に応じた複数の方向推定値を取得する。マイクロホンペアX-Yによる方向推定値は、Fig.7 に示すように、マイクロホンペアX-Yを結ぶ直線に関して対称な 2 つの値ϕ_{XY}、ϕ_{XY}として与えられる。音源方向検出制御部は、6 または８の方向推定値から、最終方向決定処理によって音源方向を決定する。最終方向決定処理は、以下の 2 つの性質に基づいて行う。

(1) 方向推定分解能は､マイクロホン間隔が大きいほど向上する
(2) 方向推定精度は､マイクロホンペアの垂線方向に近づくほど向上する

(A) マイクロホン数が３の場合

Fig.8(a)に各マイクロホンペアの推定範囲を示す。(1)の性質より、分解能が高いマイクロホンペア 3-1 の推定方向ϕ_{31}、ϕ_{31}を主推定方向とし、マイクロホンペア 1-2、2-3 の推定方向ϕ_{12}、ϕ_{12}、ϕ_{23}、ϕ_{23}を補助推定方向として用いる。(2)の性質より、主推定方向が左の時は､左前方の主推定方向ϕ_{31}と左前方の精度が高いϕ_{23}の差と、左後方の主推定方向ϕ_{31}と左前方の精度が高いϕ_{12}の差を比較する。前者の方が小さい場合はϕ_{31}を、後者の方が小さい場合はϕ_{31}を音源方向とする。同様に、主推定方向が右の時は、ϕ_{31}とϕ_{12}の差と、ϕ_{31}とϕ_{23}の差を比較し、前者の方が小さい場合はϕ_{31}を、後者の方が小さい場合はϕ_{31}を音源方向とする。

(B) マイクロホン数が４の場合

(A)と同様に､Fig.8(b)に示すマイクロホンペア 1-2、3-4 の推定方向ϕ_{12}、ϕ_{12}、ϕ_{34}、ϕ_{34}を主推定方向、マイクロホンペア 2-3、4-1 の推定方向ϕ_{23}、ϕ_{23}、ϕ_{41}、ϕ_{41}を補助推定方向とする。補助推定方向が左の時は､ϕ_{23}とϕ_{34}の差と、ϕ_{41}とϕ_{12}の差の比較を行い、前者の方が小さい場合はϕ_{34}を､後者の方が小さい場合はϕ_{12}を音源方向とする。補助推定方向が右の時は､ϕ_{23}とϕ_{12}の差と、ϕ_{41}とϕ_{34}の差の比較を行い、前者の方が小さい場合はϕ_{12}を､後者の方が小さい場合はϕ_{34}を音源方向とする。

4. 4. 方向推定コア部

方向推定コア部の構成を Fig.9 破線部に示す。方向推定コア部は、入力信号を共有メモリから取得し、マイクロホンペア毎の方向推定値を求める。各マイクロホンペアの方向推定値は、FFT 白色化、符号化を行った入力信号の相互相関から求める[8]。マイクロホン数３の時はマイクロホン 1-2、2-3、3-1 の３つのマイクロホンペア､4 の時はマイクロホン 1-2､2-3、3-4、4-1 の４つのマイクロホンペアの方向推定値を求める。マイクロホンペア数に応じた方向推定値は、

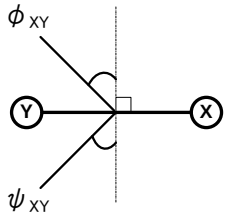


Figure 7.　マイクペアの推定方向

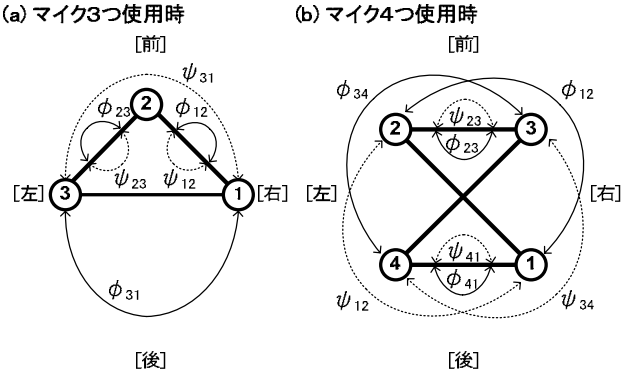


Figure 8.　マイクペアの推定範囲

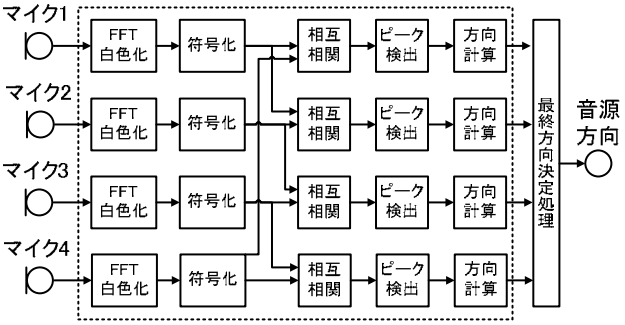


Figure 9.　方向推定コア部の構成

共有メモリを介して ARM に伝達される。方向推定コア部の方向推定値計算は 1024 サンプル毎に行い、総演算量は 35MIPS である。

5.　PaPeRo-mini

本モジュールの評価実証機として PaPeRo を小型化した PaPeRo-mini を開発した。PaPeRo-mini と PaPeRo の外観を Fig.10（左が PaPeRo-mini）に、主な仕様を Table 2 に示す。PaPeRo-mini は、音声対話モジュールをメインプロセッサとする､高さ 250mm 、幅 170mm、奥行 179mm、重量 2.5kg の自律移動型ロボットである。胴体上部に 8 個の無指向性マイクロホン、ステレオスピーカを搭載する。さらに、CCD カメラ、超音波センサ、赤外線センサ、タッチセンサ、焦電センサ、LCD を搭載し、４個のモータにより自律移動が可能である。

6.　性能評価

6. 1. 雑音下の音声認識性能

PaPeRo-mini を用いて、雑音消去機能の有無(ON/

Figure 10.　PaPeRo-mini と PaPeRo

Table 2. PaPeRo-mini と PaPeRo の仕様

	PaPeRo-mini	PaPeRo
CPU	MP211	Pentium-M 1.6GHz
OS	Linux	Windows XP
音声入力	無指向性マイク x8 個	無指向性マイク x7 個 指向性マイク x1 個
音声出力	ステレオスピーカ ライン出力 2ch	ステレオスピーカ ライン出力 2ch
画像入力	ステレオカメラ	ステレオカメラ
画像出力	コンポジットビデオ出力、LCD 表示	コンポジットビデオ出力、RGB 出力
その他 I/F	IrDA、USB	リモコン、USB
バッテリ	Li-ion 連続稼動 8 時間	Li-ion 連続稼動 2 時間
サイズ	250x170x179mm	385x248x245mm
質量	2.5 kg	5.0 kg

OFF)に対応した雑音下の音声認識性能を比較した。評価に用いた家庭環境を Fig.11 に示す。

PaPeRo-mini の正面方向(0 度方向)、距離 1.0m、高さ 1.0m に設置したスピーカから、男・女・子供合計 9 名、各 50 単語を音量 70dB で再生して認識対象音声、距離 1.0m、方向 90、135、180 の 3 方向から音量 60～65dB で再生したテレビの音を雑音とした。評価環境における平常時の騒音レベルは 40dB である。音声認識には、50 単語に雑音リジェクション単語を加えた辞書を用いた。得られた認識率を Fig.12 に示す。

比較のため、PaPeRo の性能評価結果[3]も示す。PaPeRo では、Windows XP で動作する Pentium M 1.6GHz により、同一の雑音消去処理が動作する。Fig.12 には、正面方向(0 度方向)、距離 1.0m 及び 1.5m、高さ 1.0m に設置したスピーカから、男・女・子供合計 30 名、各 50 単語の収録音声を音量 70dB で再生し、距離 1.0m、方向 90、135、180 の 3 方向から、音量 55～60dB 及び 65～70dB で再生したテレビの音を雑音とした時の平均認識率を示している。

結果を参照すると、PaPeRo-mini では、雑音が 90 度より後方にあるとき、40%以上認識率が改善し、最大改善率は 54%に達した。また、PaPeRo-mini と PaPeRo の結果を比較すると、雑音消去機能を使用しない場合は、PaPeRo の認識率が 15%以上上回った。

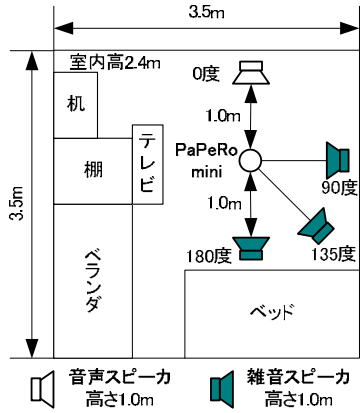


Figure 11.　音声認識の評価環境

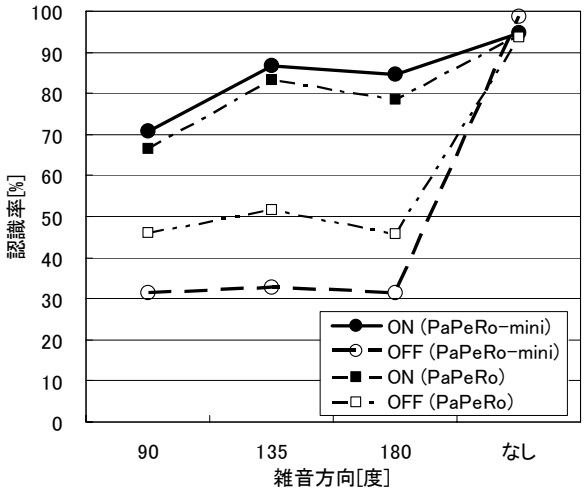


Figure 12.　雑音下の音声認識率

これは、指向性マイクロホンを音声用マイクロホンとして使用する PaPeRo と、無指向性マイクロホンを使用する PaPeRo-mini の違いであると考えられる。一方、雑音消去機能を使用した場合は、PaPeRo-mini の認識率が約 5%上回った。これは評価条件が完全に一致していないためであると考えられる。雑音なしで雑音消去機能を使用しない場合にも PaPeRo-mini の認識率が約 5%高く、音声認識率に基本的に約 5%の違いがあると理解することができる。これらの結果より、使用するマイクロホンの種類や小型化の影響を受けず、PaPeRo-mini で PaPeRo と同等の性能を得られることが確認された。

6. 2. 音源方向検出性能

PaPeRo-mini を用いて、音源方向検出性能を評価した。評価環境を Fig.13 に、マイクロホンの配置を Table 3 に示す。高さ 1.0m に設置したスピーカから男・女合計 5 名の収録音声(「おーい」)を音量 75dB で 10 回ずつ再生し、距離 1.5m に配置した PaPeRo-mini を 30 度ずつ 12 方向に回転させて、音源方向検出をおこなった。各方向の検出結果を Fig.14 に、検出率及び正解率を Fig.15 に示す。ただし、検出方向は真の方向±15 度までを正解とした。比較のため、Fig.15 に

Table 3. マイクロホン配置パラメータ

	PaPeRo-mini		PaPeRo
マイク数	4	マイク数	3
d12	0.13 m	d12	0.14 m
d23	0.12 m	d23	0.14 m
d34	0.13 m	d31	0.21 m
d41	0.12 m	–	–
θ12	337.5 度	θ12	320.0 度
θ23	180.0 度	θ23	40.0 度
θ34	22.5 度	θ31	180.0 度
θ41	180.0 度	–	–

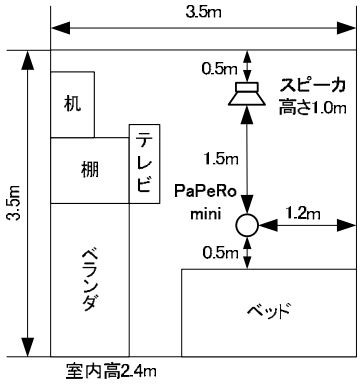


Figure 13. 音源方向検出の評価環境

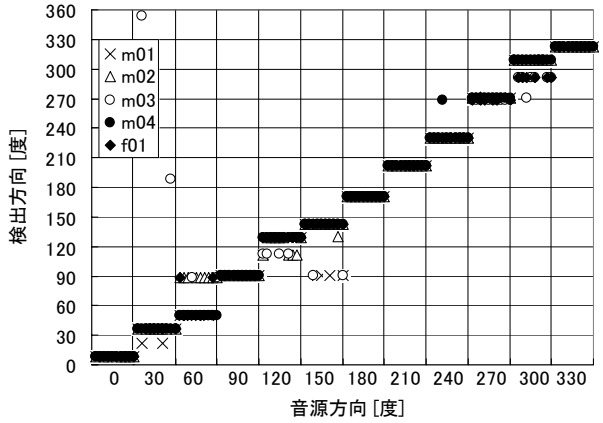


Figure 14. 音源方向検出の検出結果

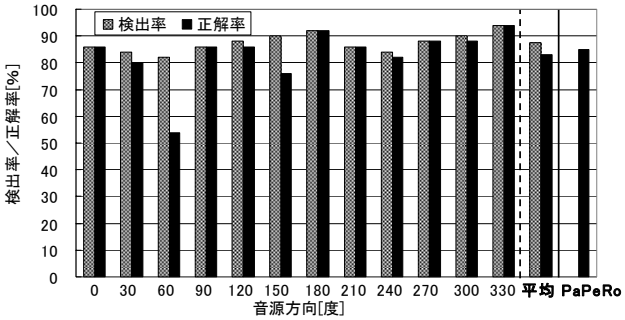


Figure 15. 音源方向検出の検出率及び正解率

PaPeRo による音源方向検出の正解率も示す。

Fig.14 と Fig.15 を参照すると、60 度方向の正解率が低く、60 度を 90 度として誤検出していることがわかる。これはマイクロホン間距離が小さいことによる方向推定分解能の低下に起因するものと考えられる。平均正解率は 83%に達し、PaPeRo と同等の性能を得られることが確認された。

7. 今後の課題

本モジュールは、現在、NEC を含む 5 機関のロボットに搭載され、機能検証中である。今後は、検証結果に基づき、基本機能の安定性とモジュールの汎用性を向上させることが課題である。また、実環境においてより高精度なインターフェースを実現するために、本稿では取り扱わなかったエコーキャンセラ、マイクロホンアレイ、画像認識などの付加機能についても評価を行い、これらの機能を統合することが必要となる。

8. おわりに

実環境で動作する音声対話機能を搭載した小型音声対話モジュールの機能、ハードウェア、及び実装形態について紹介し、本モジュールを搭載した PaPeRo-mini を用いた性能評価結果について報告した。本モジュールの雑音消去機能により、雑音環境での音声認識率が最大 54%改善し、音源方向検出機能の正解率が 83%に達することを示した。また、本モジュールの雑音消去機能、及び音源方向検出機能は、使用するマイクロホンの種類や配置の影響を受けることなく、PaPeRo と同等の性能が得られることを示した。

謝辞

本研究の一部は、NEDO の次世代ロボット共通基盤開発プロジェクトの一環として行っている。本研究をご支援いただいた関係各位に感謝する。

参考文献

1) 金田, "信号処理から見たロボット聴覚「音源の方向検出について」", 人工知能学会 AI チャレンジ研究会, Vol.22, pp.1-8, Oct. 2005.
2) M. Brandstein and D. Ward, "Microphone Arrays," Springer Verlag, Berlin, 2001.
3) M. Sato, A. Sugiyama, S. Ohnaka, "An Adaptive Noise Canceller with Low Signal-Distortion Based on Variable Stepsize Subfilters for Human-Robot Communication", IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2055-2061, Aug. 2005.
4) Y. Fujita, "Personal Robot PaPeRo," J. of Robotics and Mechatronics, Vol.14, No.1, pp.60-63, Jan. 2002.
5) 末廣, 北垣, 神徳, 尹, 安藤, "ＲＴ要素のモジュール化に関する検討―ＲＴミドルウェアの基本機能に関する研究開発(その１)", 日本ロボット学会学術講演会, Sep. 2003.
6) http://www.w3.org/TR/speech-grammar/
7) 岩沢, "パーソナルロボット PaPeRo の音声認識インターフェース", 人工知能学会 AI チャレンジ研究会, Vol.13, pp.17-23, Jun. 2001.
8) M. Sato, A. Sugiyama, O. Hoshuyama, N. Yamashita, Y. Fujita, "Near-Field Sound-Source Localization Based on a Signed Binary Code", IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2078-2086, Aug. 2005.

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW):
`http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/`