人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624

ＡＩチャレンジ研究会 (第２４回)

Proceedings of the 24th Meeting of Special Interest Group on AI Challenges

CONTENTS

【11 月 16 日】

【11 月 17 日】

日　時　2006 年 11 月 16 日～ 17 日　　場　所　京都, キャンパスプラザ京都
Campus Plaza Kyoto, Kyoto, Nov. 16–17, 2006

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

共催 社団法人 人工知能学会 言語・音声理解と対話処理研究会
JSAI SIG on Speech & Language Understanding and Dialogue Processing

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-0624-1 (11/17)

MFTを用いたロボットの動作音に頑健な音声認識手法の提案

Noise Robust Automatic Speech Recognition for a Humanoid with Motor Noise

***西村 義隆, **中臺 一博, **中野 幹生, **辻野 広司, *石塚 満**

Yoshitaka NISHIMURA, Kazuhiro NAKADAI, Mikio NAKANO, Hiroshi TSUJINO and Mitsuru ISHIZUKA

*東京大学大学院情報理工学系研究科

**(株) ホンダ・リサーチ・インスティチュート・ジャパン

Graduate School of Information Science and Technology, The University of Tokyo

Honda Research Institute Japan Co., Ltd.

nisshi@mi.ci.i.u-tokyo.ac.jp, ishizuka@i.u-tokyo.ac.jp, {nakadai, nakano, tsujino}@jp.honda-ri.com

Abstract

Automatic speech recognition (ASR) is essential for human-humanoid communication. One of the main problems with ASR is that a humanoid inevitably generates motor noises. However, it is possible to estimate these noises by using information on humanoid's motions and gestures. This paper proposes a method to improve ASR for a humanoid with motor noises by utilizing its motion/gesture information. The method consists of noise suppression and missing-feature-theory-based ASR (MFT-ASR). The proposed noise suppression technique is based on spectral subtraction, and white noise is added to blur distortion of suppression. MFT-ASR improves ASR by masking unreliable acoustic features in the input sound. The motion/gesture information is used for obtaining the unreliable acoustic features. We evaluated the proposed method through recognition of recorded by using Honda ASIMO in anechoic room. The experimental results show that the proposed method outperforms the conventional multi-condition training technique.

1 はじめに

近年，さまざまなロボットが開発されている．その中でも特にヒューマノイドロボットはコミュニケーションを通して，人と同じようにさまざまな仕事をこなすことが期待されている．人同士のコミュニケーションでは音声が最も一般的に用いられているため，人とロボットも音声でコミュニケーションを行うことが理想的であろう．しかし，ロボットが音声認識を行う上では多くの問題がある．実環境における音声認識では種々の雑音が混入する．特にロボットは自身の発する雑音が，定常時でもモータ音やファン音などの雑音，動作中には，手足の動作に伴うモータ音が発せられる．さらに，位置の変化等によりマイクに混入するモータ音やファン音も変化する．人・ロボットコミュニケーションの研究では，高雑音下での音声認識を避けるため，ロボット自身のマイクを用いずに接話マイクによる音声認識が行われている [1]．しかし，常に接話マイクを用いることは利用者にとって煩わしく，ロボット自身のマイクで音声認識を行うことは重要であると考えられる．

これまで，音声認識の先行研究においては数々の雑音への頑健性向上に対する手法が提案されている．マルチコンディション学習による音響モデルの学習は最も有効な手法の一つである．この手法は，あらかじめ雑音を含んだ音声を音響モデルの学習に用いるため，その雑音が既知である場合には強力である．しかし，雑音が大きい環境では，無音区間か発話区間かの区別すらできなくなる．また，定常的な雑音については効果的な学習が期待できるが，非定常な雑音に対しては難しい．このため，高雑音下ではこの手法には限界があると考えられる．

MLLR(Maximum Likelihood Linear Regression)[2] は，アフィン変換を用いて音響モデルを雑音に適応するアプローチである．これにより，音響モデルは学習時とは異なる認識環境の雑音や話者に適応される．MLLR も有効な手法であるが，雑音が非常に大きい環境や非定常雑音においては効果が薄いと考えられる．

このように従来の音声認識では，音響モデルを雑音へ適応するための研究が多く行われてきた．これは，入力信号から雑音を取り除くというアプローチをとると，音声の歪みが大きくなり，結果的に音響モデルの雑音への適応を行った方が性能が出やすいという側面を有するからと考えられる．しかし，ロボットにおける音声認識では，従来の音声認識が想定していた雑音よりも雑音の大きな環境 (SNR 0dB 以下である場合もある) での認識が必要

となる．このような環境では音響モデルを雑音へ適応化しても，もはや元の信号の情報はほとんど残っておらず，音声認識を行うことは困難である．したがって，雑音を除去する仕組みが必要となる．

ロボットにおける音声認識では，その前処理に用いるため，マイクロホンアレーを用いた音源分離が数多く行われている．ビームフォーミング (BF:Beam Forming)[3]，独立成分分析 (ICA:Independent Component Analysis)[4] あるいは幾何学的音源分離 (GSS:Geometric Source Separation)[5] による手法が提案されている．BF は一般的な音源分離手法であるが，音源分離による音声信号の歪みが生じる．歪みの少ない適応 BF も提案されているが，計算量が膨大であるという欠点がある．ICA は音源の独立性を仮定するだけで分離を行うことができる有効な手法であるが，実環境においてはしばしばこの仮定が成立しないことがあり，各周波数での分離信号が同じ音源に対応するように分離信号を並べ変えなければならないという permutation 問題も生じる．BF と ICA の中間的な手法として，GSS が挙げられる．GSS では音源位置とマイク位置及び音源の相関に基づいて音源分離を行うが，実環境では位置の正確な抽出が難しく，分離性能に影響を与える．

ロボットの音声認識で問題となる雑音には，動作音の他，環境雑音などがある．環境雑音は非定常であり，音源位置や音源数の情報もないため，雑音の推定にはマイクロホンアレーを用いた手法が必要となる．しかし，本研究で対象とする動作音はロボット自身が発するものであり，ロボットは自己の動作情報を取得可能なため，動作音の推定が可能である．よって，マイクロホンアレーのような多くの情報を用いて雑音への頑健性を向上させなくとも，もっと少ない情報で効率的に適応ができると考えられる．

本研究と同様に動作音を対象とし，マイク 1 本で雑音への適応を行うアプローチとして，SS(Spectrum Subtraction) を用いた手法がある [6]．従来の SS[7] は無音区間などを用いて定常雑音の推定を行い，スペクトル領域において推定雑音成分を減算することにより音声信号の抽出を行うものである．伊藤らは SS を AIBO の動作音の軽減に用いた [6]．具体的には関節角度や位置を入力としたニューラルネットワークで推定雑音の学習をさせ，これを用いて SS の減算に用いる雑音信号の推定を行い，シミュレーション上での認識性能を報告している．しかし，実環境でのパフォーマンスについて言及されていないため反響音のある環境や，マルチコンディション学習による音響モデルを用いた手法と比べ，有効性があるのかどうかは不明である．また，SS は定常雑音に対しては有効であると考えられているが，非定常雑音に対しては歪みが生じることがあるため有効な手法とは言い難い．

非定常雑音に対しても有効な手法として，MFT(Missing Feature Theory)[8] を用いた手法がある．MFT は音声信号のうち雑音や歪みのない部分の情報のみを用いて音声認識を行うアプローチである．信頼性の低い部分はマスクされることにより音声認識には用いられない．MFT はマスクするかしないかの二者択一とする狭義の MFT と，信頼性の大きさに応じてマスクを連続的な値とする広義の MFT があり，本稿では広義の MFT の意で用いる．関連する研究として重みづけを用いたマルチバンド音声認識 [9],[10] がある．重みづけを用いたマルチバンド音声認識では，信頼性の低い周波数帯域は重みを小さく，信頼性の高い周波数帯域は重みを大きくすることによりその重みを尤度に反映させて音声認識を行う．MFT を用いた方法では，信頼性の推定を正確に行うことができれば，認識性能は他の雑音適応手法と比較して大きく向上する．信頼性の推定を正確に行うためには雑音の推定が必要であるが，ブラインドで雑音推定を行うこと自体が音声認識と同レベルの難しさを有するという問題がある．従来の音声認識では，この信頼性推定が非常に困難であるため，MFT が有効な手法として用いられることが少なかった．しかし，本研究で対象とするロボットの動作音はその雑音推定が容易であるため，MFT が有効に利用できると考えられる．

本研究ではまず入力信号から雑音除去処理を行う．動作音の混入した環境では SNR が低いため，雑音除去処理は必要である．次に，雑音除去処理での雑音の引き残し成分を平坦化するため，白色雑音の重畳を行う．SNR が高い環境では雑音除去処理による音声信号の歪みは小さいと考えられるが，SNR が低い環境ではその歪みは大きく，雑音除去処理を行うことでかえって認識性能が劣化するという事態も考えうる．雑音の除去により，モータ音などの定常的な雑音の多くは取り除くことができるが，動作による非定常成分の雑音への適応は不十分であると考えられる．これらに適応するため，MFT を利用した音声認識を行う．MFT のマスク生成には推定動作雑音を用い，雑音の多く重畳した箇所は信頼性が低く，音声認識への関与を低くするようにする．次節において提案手法の詳細を説明する．

2 MFT を用いた動作音への雑音適応化手法

図 1 に提案する雑音適応化手法のブロック図を示す．以下，それぞれの処理について示す．

2.1 雑音除去処理

入力信号の SNR は低い (0dB 以下である場合もある) ため，このような環境で音声認識に有効な音声特徴量を抽出することは難しい．そこで，入力信号の SNR を改善するため雑音除去処理を行う．雑音除去処理には式 (1) に示

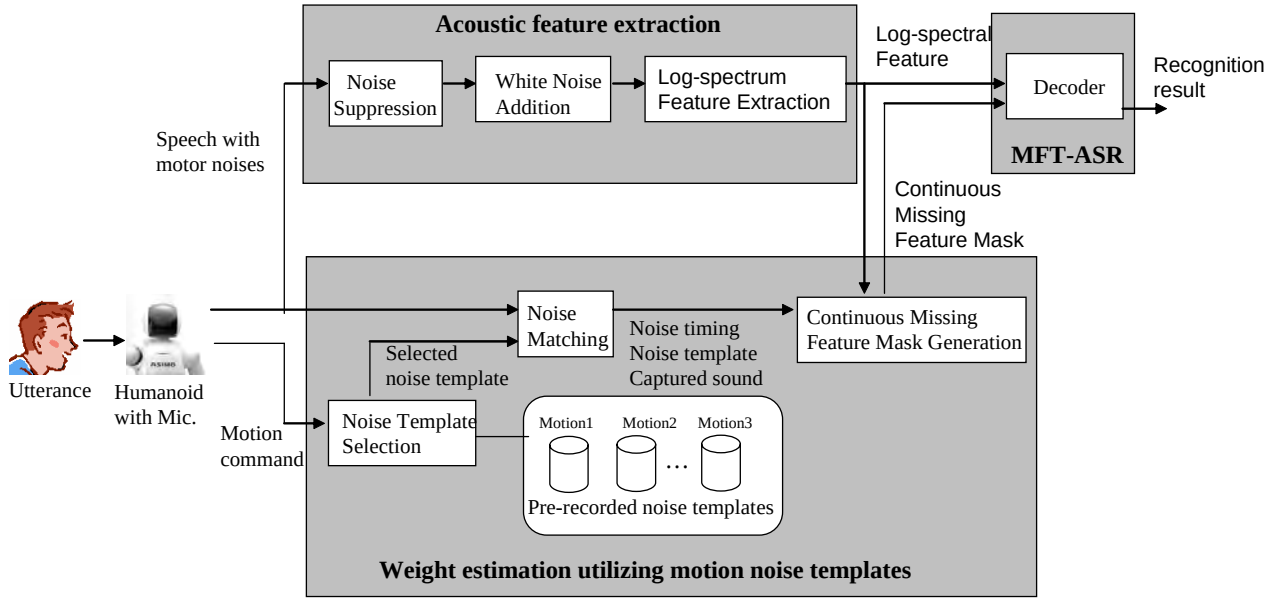


Figure 1: *Block diagram of the proposed method*

される SS を用いた．

$$|X(f)| = \max\{|X(f)| - \sqrt{\alpha}|\bar{N}|, \sqrt{\beta}|\bar{N}|\} \qquad (1)$$

$X(f)$ は入力信号のスペクトルを示し，$\bar{N}$ は入力信号に重畳している雑音信号の平均スペクトルを示す．α, β は SS を行う際のパラメータであるが，本稿では一般的によく用いられている値 $(\alpha = 1, \beta = 0.1)$ を用いた．

2.2 白色雑音重畳

雑音除去処理は SNR を向上させるが，同時にスペクトルの歪みを生み出す．このスペクトル歪みが認識性能に悪影響を及ぼす．雑音除去手法に関わらず，背景雑音の状況によっては大きな歪みを生じることがあり，音声認識ではスペクトル歪みに対する処理が必要である．特に本稿の対象とするロボットの動作雑音では，雑音パワーが大きく，歪みも大きいことが予測される．そこで，本稿ではこのスペクトル歪みを軽減するため，雑音除去処理の後に薄く白色雑音を重畳させることとした．同様の方法は山出 [11] らの報告にも述べられており，定常雑音を加えることで，雑音の引き残し成分を平坦化し，認識性能を高めることが期待される．

白色雑音の重畳には，入力信号のある程度のレベルの白色雑音を加えることが歪みを抑制するのに役立つと考え以下のような式を用いた．

$$y'(t) = y(t) + \frac{2p}{T}\sum_{t=1}^{T}|y(t)| \cdot \mathrm{random}(1) \qquad (2)$$

$y(t)$ は雑音除去処理後の信号であり，random(1) は -1 から 1 までの任意の実数値をランダムに返す関数である．本稿では $p = 0.1$ とした．すなわち，平均して入力信号の 1 割程度の大きさの白色雑音が加わることとなる．

2.3 対数スペクトル特徴量の抽出

白色雑音を重畳した後に音声特徴量を抽出する．音声特徴量には音声認識に一般的に用いられる MFCC(Mel Frequency Cepstrum Coefficients) ではなく，対数スペクトル特徴量 [12, 13] を用いた．動作音などの雑音は，スペクトル領域において加算される．しかし，従来用いられている MFCC はスペクトルをさらに DCT(Discrete Cosine Transform) した領域であるため，ある周波数帯域に加算された雑音は全ての特徴量に影響を与えてしまう．MFT を用いた音声認識では，雑音に埋もれた信頼性の低い周波数帯域を抽出することが必要であるため，ケプストラム領域の音声特徴量よりもスペクトル領域の音声特徴量の方が都合がよい．MFCC ではケプストラム領域に変換された後，C_0 項の除去，リフタリング，CMS (Cepstrum Mean Subtraction) の 3 つの正規化処理が行われる．これらの正規化処理は音声認識性能を向上させる上で重要であることが知られているため，使用した対数スペクトル特徴量においても，対数スペクトル領域において同様の正規化処理を施している．

以下，正規化処理について概説する．初めに C_0 項の除去であるが，これは対数スペクトルの平均を引くことによって同等の処理を行う．フレーム f における i 次元目のスペクトルを S_{fi} とすると，C_0 正規化後の対数スペクトルは以下のように示される．

$$S^1_{fi} = S_{fi} - \frac{1}{N}\sum_{j=1}^{N} S_{tj} \qquad (3)$$

次に，リフタリング処理ではスペクトル構造の山と谷が強調されるため，対数スペクトル特徴量の正規化では以

下のリフターにかける．

$$H(z) = 1 - 0.9z^{-1} \tag{4}$$

$H(z)$ のインパルス応答を h_x とし，

$$S^2_{fx} = h_x * S^1_{fx} \tag{5}$$

として処理を行う．$*$ は畳み込み演算を表す．最後に CMS であるが，以下のように対数スペクトルの時間方向の平均を減算することで同じ処理が行われる．

$$S^3_{fi} = S^2_{fi} - \frac{1}{F}\sum_{j=1}^{F} S^2_{ji} \tag{6}$$

2.4 MFT マスクの生成

MFT マスクはフレームごと，周波数帯域ごと (音声特徴量の次元ごと) に生成される．自動的なマスクの生成は Raj らの報告 [14] がある．しかし，完全に理想的なマスクを生成することは現実的には困難である．本研究では，ロボット自身の動作情報は動作前に取得できるため，これに基づいて動作音の推定を行う．動作音の推定については，あらかじめ収録したテンプレート雑音と現在入力されている動作音との時間的なマッチングにより行う．そして，入力された信号と推定された動作音に基づいてマスクの生成を行う．詳細については次に示す．

2.5 テンプレート雑音の選択

あらかじめ収録した動作音をテンプレート雑音としてデータベース化する．本研究では，34 種類の動作音を用意した．ロボットが動作を行う際には，データベースから動作種類に応じたテンプレート雑音を選択する．現在発せられている動作音はこのテンプレート雑音と同じであると仮定し，テンプレート雑音を用いた雑音推定を行う．

2.6 雑音マッチング

テンプレート雑音の選択が行われても，その雑音と現在発せられている雑音が時間的にはマッチしていない．そこで，時間的に雑音をマッチングさせる必要が生じる．マッチングは以下の方法により行われる．$T_d(f)$ をテンプレート雑音のスペクトル系列，$I_d(f)$ を入力信号のスペクトル系列とする．f はフレームとし，d は周波数軸方向のスペクトルの次元とする．D を 1 フレームの窓長 (サンプル数) とすると，$1 \leq d \leq D$ である．また，テンプレート雑音における各次元のスペクトルの最大値を M_d とする．

ここで，入力信号 $I_d(f)$ について，M_d を超えるものは音声信号が重畳しており，ミスマッチの要因となると考え，そのようなスペクトル系列の値を 0 とする．

$$I'_d(f) = \begin{cases} I_d(f) & if \quad I_d(f) \leq M_d \\ 0 & if \quad otherwise. \end{cases} \tag{7}$$

マッチングは I'_d と T_d の相互相関をとることにより行った．最も相関が高いフレーム s_d は

$$s_d = \underset{\tau}{\operatorname{argmax}} \sum_{f=0}^{N-1} I'_d(f) T_d(f-\tau) \tag{8}$$

である．得られた s_d の $1 \leq d \leq D$ のうち，最も s_d の値の数が大きいものを s_{match} としてマッチングに用いる．

マッチング後の推定雑音 $E_d(f)$ は，

$$E_d(f) = T_d(f - s_{match}) \tag{9}$$

により得られる．

2.7 マスクの生成

まず，マッチングされたテンプレート雑音 $T(s_{match})$ は対数スペクトルに変換される．変換された対数スペクトルの雑音を $n(k,f)$ とする．k は次元 (周波数軸方向) を示し，f はフレーム (時間軸方向) を示す．同様に，入力された雑音を含む対数スペクトルを $y(k,f)$，雑音除去処理後，白色雑音を重畳した対数スペクトルを $p(k,f)$ とする．推定された音声信号は以下のように表される．

$$c'(k,f) = y(k,f) - n(k,f). \tag{10}$$

マスク $m(k,f)$ は以下のように計算される．

$$m(k,f) = \frac{|C'(k,f) - \mathrm{median}_k(C'(k,f))|}{P(k,f) - C'(k,f)} \tag{11}$$

$\mathrm{median}_k(a(k))$ は $a(k)$ の中央値を得る関数である．$P(k,f)$ および $C'(k,f)$ は対数スペクトル $p(k,f)$ および $c'(k,f)$ に正規化処理を施したものである．$m(k,f)$ がとても大きな値になることを防ぐため，閾値 t_{th} を設けた．したがって，$m(k,f)$ のとる範囲は 0 から t_{th} である．t_{th} は実験的に 5.0 とした．

さらに，MFT マスクの正規化を行う．この正規化は，MFT を用いた音声認識を行うことで挿入ペナルティなどの最適パラメータの変化を抑えるために行う．正規化後の MFT マスクを $w(k,f)$ とし，1 フレームにおける $w(k,f)$ の合計が音声特徴量の次元数 K と同じになるように正規化を施す．

$$\begin{aligned} w(k,f) &= \frac{m'(k,f)}{\sum_{k=1}^{K} m'(k,f)} \\ m'(k,f) &= \begin{cases} m(k,f) & if \quad m(k,f) < t_{th}, \\ t_{th} & if \quad otherwise. \end{cases} \end{aligned} \tag{12}$$

2.8 MFT に基づく尤度の計算方法

MFT は非定常な雑音に対しても効果がある．雑音除去処理や白色雑音の重畳によって SNR は改善されるが，MFT

Table 1: *Experimental Conditions*

Condition	**A**	**B**	**C**	**D**	**E**	**F**
Multi-condition		✓	✓			
Noise Suppression (SS)			✓	✓	✓	✓
Adaptation for SS				✓	✓	✓
White Noise Addition				✓	✓	✓
MFT					✓	
MFT (a priori mask)						✓
Acoustic Model	AM-1	AM-2	AM-2	AM-3	AM-3	AM-3

を用いることでさらに非定常な雑音成分に対しても効果があると期待できる．しかし，テンプレート雑音と実際に生じた雑音に大きな差がある場合には効果は薄い．

MFT では信頼性の高い特徴成分に対しては大きな重みを，信頼性の低い特徴成分に対しては小さな重みを用いて尤度の計算を行う．MFT を用いない従来の音声認識では，音素モデル q_l，音声特徴量 $\mathbf{s_f}$ の尤度は以下の式によって与えられる．

$$L(\mathbf{s_f}|q_l) = \sum_{i=1}^{M} L(\mathbf{s_f}(i)|q_l). \tag{13}$$

MFT を用いた尤度計算は，マスクを $\omega(k,f)$ として以下のように定義される．

$$L(\mathbf{s_f}|q_l) = \sum_{i=1}^{M} \omega(i,f) L(\mathbf{s_f}(i)|q_l). \tag{14}$$

3 評価実験

3.1 実験条件

Honda ASIMO を用いて評価実験を行った．ASIMO の左マイクを用いて音声の収録を行い，孤立単語認識による評価を行った．評価用データには ATR 音素バランス単語を用いた．音素バランス単語には男性 12 話者，女性 13 話者の合計 25 話者の音声データが含まれ，1 話者あたりの発話数は 216 である．各発話は “いきおい”, “いよいよ” などの単語発声である．

音響モデルの構築には男性 9 話者女性 10 話者の合計 19 話者の音声データ (学習セット A_1) を用いた．このデータは無響室において 100 cm の距離から収録を行い，音圧の変化にも柔軟に対応できるようにするため，SNR のレベルを変化させて (+5 dB, +10 dB, +15 dB) 学習を行った．

テスト用のデータは男性 3 話者女性 3 話者の合計 6 話者の音声データ (テストセット R_1) を用いた．このデータは音響モデルの学習とは異なる話者から構成されている．収録は 7 m (W) × 4 m (D) × 3 m (H) の部屋において行った．実用的な環境においても性能を発揮するか検証するため，家庭のリビングを想定した大きさの部屋で，反響音のある環境で収録を行った．話者とロボットのマイクの距離は 50 cm, 100 cm, 150 cm, 200 cm の 4 距離である．

ロボットの動作雑音については，32 種類の動作を用いて認識実験を行った．この動作音は ASIMO の電源を投入し，動作を全く行っていない定常雑音 1 種類と「バイバイ」や「お辞儀」などの上半身の動作を主とするジェスチャ雑音 25 種類および「直進」や「回転」など足を用いた動きを主とする歩行雑音 8 種類より構成される．テストセット R_1 に動作音を重畳したものをテストセット R_2 とする．

提案手法と従来の有効な手法であるマルチコンディション学習による音響モデルを用いた手法の比較を行うため，マルチコンディション学習用のデータを用意した．マルチコンディション学習では A_1 のデータに加え，ASIMO の電源を投入したときのモータ音やファン音などの定常雑音を重畳したデータを用いた．これを学習セット A_2 とする．認識実験においては，以下の 4 つの音響モデルを用意した．

AM-1 学習セット A_1 を用いたモデル (クリーンモデル)

AM-2 学習セット A_1 と A_2 を用いたモデル (マルチコンディション学習モデル)

AM-3 学習セット A_1 と A_2 に雑音除去処理を施した後に白色雑音を重畳した A_3 を用いたモデル

評価は表 1 に示す 6 つの手法を比較することにより行った．手法 **A** はクリーンモデルを用いた一般的な音声認識である．手法 **B** は雑音に頑健な手法として一般的によく用いられているマルチコンディション学習による音響モデルを用いた音声認識である．手法 **C** は雑音除去処理を行い，マルチコンディション学習による音響モデルを用いて認識を行ったものである．手法 **D** はスペクトルの歪みを抑えるために雑音除去処理の後白色雑音を重畳したものである．手法 **E**，手法 **F** はともに MFT を用いた音声認識で，手法 **E** は音声と雑音の混入した入力信号から雑音マッチングを用いてマスクの推定を行った提案手法，手法 **F** は入力信号の雑音が完全に既知であると仮定してマスクの生成を行ったものである．

Table 2: Isolated Word Recognition (% Word Correct)

	50 cm							100 cm						
Condition	**A**	**B**	**C**	**D**	**E** (p-value)		**F**	**A**	**B**	**C**	**D**	**E** (P-value)		**F**
Motor noise	59.19	83.34	67.60	*84.42*	**85.27**	0.04	83.41	40.44	74.23	60.11	*78.94*	**79.48**	0.00	80.79
Right hand (1)	50.93	81.25	64.74	*82.87*	**83.18**	0.06	82.72	30.71	67.67	57.87	*73.38*	**75.78**	0.00	77.86
Right hand (2)	45.45	*77.63*	58.10	75.54	**80.87**	0.00	81.72	26.00	*64.97*	47.84	62.66	**72.54**	0.00	74.23
Right hand (3)	51.32	79.63	63.89	*80.79*	**81.87**	0.01	81.64	32.56	67.75	55.02	*71.45*	**76.54**	0.00	76.47
Right hand (4)	52.70	82.64	64.51	**83.64**	*83.41*	0.42	82.80	32.95	69.14	55.40	*72.92*	**75.23**	0.00	76.93
Right hand (5)	50.70	81.02	63.35	*81.33*	**83.80**	0.00	82.41	31.10	68.52	54.63	*72.84*	**75.85**	0.00	76.70
Left hand (1)	50.31	78.32	62.04	*81.18*	**81.95**	0.00	81.95	30.48	65.05	49.54	*69.52*	**73.23**	0.00	76.62
Left hand (2)	44.60	74.69	56.02	*75.62*	**78.55**	0.00	79.48	24.31	59.88	45.53	*60.26*	**68.29**	0.00	71.99
Left hand (3)	51.55	80.25	64.74	*82.33*	**83.72**	0.00	82.41	32.72	68.21	55.40	*70.99*	**74.62**	0.00	76.93
Left hand (4)	47.45	*74.62*	50.93	68.52	**78.78**	0.00	79.48	29.32	*60.42*	36.04	54.87	**68.75**	0.00	71.22
Left hand (5)	52.01	79.40	62.04	**82.72**	*82.03*	0.01	82.41	32.72	65.43	52.47	*70.37*	**74.00**	0.00	75.54
Both hands (1)	45.07	76.70	59.11	*78.47*	**80.79**	0.00	80.33	25.31	61.19	46.53	*66.52*	**71.84**	0.00	73.77
Both hands (2)	44.29	74.54	57.57	*77.93*	**78.94**	0.00	80.33	25.47	59.18	48.15	*63.27*	**69.14**	0.00	71.61
Both hands (3)	43.60	73.61	56.10	*76.39*	**78.40**	0.00	80.40	25.62	58.10	45.76	*62.12*	**67.67**	0.00	70.76
Both hands (4)	43.37	74.08	58.88	*77.62*	**78.78**	0.00	79.79	24.62	59.88	47.30	*63.12*	**69.83**	0.00	71.30
Both hands (5)	45.91	76.93	60.19	*78.16*	**81.41**	0.00	80.33	26.01	61.58	50.16	*65.82*	**71.22**	0.00	73.46
Head (1)	45.84	*63.89*	40.97	62.50	**70.76**	0.00	74.15	23.69	*50.23*	30.40	48.15	**58.57**	0.00	63.27
Head (2)	53.25	73.23	52.86	*73.84*	**75.16**	0.09	78.01	34.57	59.80	44.99	*61.34*	**65.97**	0.00	72.77
Head (3)	50.39	74.31	50.70	*74.54*	**75.93**	0.21	78.48	31.79	*62.04*	40.20	*62.04*	**67.21**	0.00	73.07
Head (4)	29.94	*44.91*	26.39	36.88	**51.39**	0.00	56.87	17.83	*33.88*	20.06	27.24	**41.67**	0.00	47.76
Head (5)	30.79	*46.22*	29.25	41.36	**50.23**	0.00	57.41	17.90	*34.88*	24.00	31.72	**42.83**	0.00	48.23
Head and Hands (1)	51.24	74.38	51.62	*75.62*	**77.32**	0.02	80.63	33.57	61.81	39.12	*62.96*	**67.59**	0.00	73.84
Head and Hands (2)	51.31	66.59	41.98	*67.52*	**71.22**	0.00	77.47	31.18	*55.79*	31.33	53.32	**58.49**	0.03	68.91
Head and Hands (3)	37.50	63.81	41.59	*64.28*	**70.61**	0.00	75.24	19.53	*48.85*	29.17	44.76	**56.33**	0.00	61.89
Head and Hands (4)	46.38	77.70	61.81	*80.17*	**81.56**	0.00	80.87	27.24	63.04	49.54	*66.52*	**70.76**	0.00	73.31
Head and Hands (5)	47.38	*73.30*	45.30	62.58	**73.61**	0.76	77.86	27.47	*60.27*	34.19	46.92	**62.43**	0.00	69.14
Walking Motion (1)	50.00	*74.00*	49.46	67.67	**76.39**	0.05	76.78	29.94	*62.19*	41.67	49.85	**64.43**	0.10	64.90
Walking Motion (2)	46.91	*71.30*	43.21	63.04	**73.61**	0.05	75.08	29.40	*56.95*	32.80	44.22	**60.81**	0.00	64.74
Walking Motion (3)	49.38	74.08	50.62	*74.85*	**77.62**	0.00	79.32	30.32	*64.05*	42.75	58.72	**68.36**	0.00	69.14
Walking Motion (4)	50.78	*75.85*	53.32	73.92	**79.17**	0.01	78.24	32.33	*65.51*	45.37	59.26	**69.06**	0.00	70.29
Walking Motion (5)	52.01	76.08	52.62	*77.78*	**81.10**	0.00	80.94	34.11	*67.59*	43.21	65.13	**71.69**	0.00	75.47
Walking Motion (6)	46.99	*74.85*	50.39	71.99	**76.93**	0.04	78.17	29.63	*61.81*	41.90	55.10	**67.13**	0.00	67.75
Walking Motion (7)	49.85	*75.24*	54.17	74.85	**79.17**	0.00	79.56	32.41	*66.51*	46.76	61.35	**70.22**	0.00	72.30
Walking Motion (8)	50.85	*76.62*	54.09	76.47	**79.94**	0.00	80.17	33.41	*67.75*	44.76	61.88	**72.15**	0.00	73.77
	150 cm							200 cm						
Condition	**A**	**B**	**C**	**D**	**E** (P-value)		**F**	**A**	**B**	**C**	**D**	**E** (P-value)		**F**
Motor noise	27.86	64.20	54.02	*70.60*	**73.69**	0.00	73.85	19.83	54.01	49.93	*59.72*	**66.75**	0.00	68.44
Right hand (1)	19.68	56.02	48.69	*61.81*	**67.75**	0.00	70.14	12.27	45.60	43.36	*49.00*	**59.73**	0.00	61.89
Right hand (2)	15.36	*50.70*	38.97	48.23	**62.96**	0.00	64.12	9.57	*41.44*	31.79	36.11	**52.01**	0.00	57.49
Right hand (3)	21.37	56.56	47.30	*58.80*	**66.75**	0.00	66.51	12.89	*46.53*	41.82	*46.53*	**59.11**	0.00	60.80
Right hand (4)	20.99	56.02	45.37	*58.64*	**67.67**	0.00	68.60	12.50	46.30	41.05	*46.46*	**60.03**	0.00	62.35
Right hand (5)	18.52	55.48	47.46	*58.34*	**66.51**	0.00	68.52	11.27	46.61	42.59	*47.15*	**57.80**	0.00	61.65
Left hand (1)	18.83	54.55	41.20	*56.10*	**62.89**	0.00	68.06	10.73	42.29	35.19	*44.75*	**54.86**	0.00	59.65
Left hand (2)	13.81	*47.38*	35.80	46.45	**56.72**	0.00	59.42	6.71	*36.65*	30.56	35.88	**47.23**	0.00	50.16
Left hand (3)	21.45	55.25	46.53	*58.72*	**65.05**	0.00	67.83	12.74	44.76	40.05	*46.68*	**58.95**	0.00	61.88
Left hand (4)	16.82	*50.00*	27.63	41.82	**59.73**	0.00	61.88	10.34	*39.51*	21.07	30.79	**50.85**	0.00	55.79
Left hand (5)	19.91	54.01	41.82	*56.49*	**65.97**	0.00	68.13	11.19	43.75	36.50	*44.99*	**55.71**	0.00	59.57
Both hands (1)	15.20	*51.01*	37.50	50.31	**59.80**	0.00	64.66	7.79	39.05	31.72	*40.67*	**51.70**	0.00	55.48
Both hands (2)	15.44	*49.00*	39.82	48.15	**59.65**	0.00	62.50	8.57	*38.82*	31.87	37.43	**50.54**	0.00	54.94
Both hands (3)	16.52	45.99	36.50	*48.54*	**57.18**	0.00	60.73	8.57	35.42	30.56	*36.65*	**48.38**	0.00	52.94
Both hands (4)	15.13	46.61	37.27	*49.31*	**56.41**	0.00	60.57	7.56	36.35	31.71	*37.35*	**49.16**	0.00	52.47
Both hands (5)	15.59	49.00	40.82	*52.01*	**60.57**	0.00	63.58	8.41	39.35	34.80	*41.83*	**52.24**	0.00	55.56
Head (1)	13.74	*36.89*	22.61	32.80	**45.99**	0.00	52.24	7.02	*26.78*	16.59	21.76	**37.35**	0.00	43.98
Head (2)	21.68	*51.62*	37.58	*51.62*	**57.64**	0.00	64.20	14.81	*42.05*	32.57	*42.05*	**50.39**	0.00	57.72
Head (3)	18.45	51.01	33.26	*51.93*	**56.56**	0.00	63.43	11.81	*41.13*	26.93	40.43	**51.24**	0.00	57.64
Head (4)	11.81	*24.62*	14.82	20.45	**32.56**	0.00	40.28	7.26	*19.83*	12.58	15.90	**27.16**	0.00	34.42
Head (5)	10.57	*28.47*	18.91	24.77	**36.88**	0.00	41.20	6.56	*21.22*	16.36	20.06	**30.94**	0.00	36.27
Head and Hands (1)	20.84	*52.55*	33.03	50.16	**58.65**	0.00	64.82	14.20	*41.67*	27.93	39.90	**50.00**	0.00	58.95
Head and Hands (2)	19.91	*44.14*	23.54	42.29	**49.16**	0.00	59.19	13.66	*35.88*	18.91	34.11	**42.44**	0.00	53.17
Head and Hands (3)	9.11	*38.04*	20.84	32.64	**44.37**	0.00	51.08	4.94	*27.39*	15.97	21.38	**36.04**	0.00	43.13
Head and Hands (4)	15.21	50.23	41.98	*51.62*	**62.04**	0.00	63.66	8.49	38.97	35.42	*40.05*	**53.32**	0.00	56.64
Head and Hands (5)	16.67	*47.77*	25.54	34.57	**52.70**	0.00	59.41	9.11	*37.27*	21.22	24.70	**45.06**	0.00	53.01
Walking Motion (1)	19.29	*52.09*	37.11	37.12	**53.63**	0.10	53.55	12.12	*41.13*	30.87	27.47	**45.14**	0.00	45.91
Walking Motion (2)	18.68	*45.68*	27.78	31.56	**49.23**	0.01	52.39	11.58	*36.96*	21.92	21.07	**40.36**	0.00	44.53
Walking Motion (3)	19.29	*51.47*	34.11	43.52	**56.87**	0.00	57.95	12.19	*42.21*	30.25	32.87	**48.31**	0.00	50.70
Walking Motion (4)	19.91	*55.10*	38.20	43.98	**56.87**	0.03	57.87	12.35	*45.06*	32.64	30.02	**47.30**	0.02	48.77
Walking Motion (5)	20.91	*55.09*	37.66	51.70	**62.89**	0.00	64.89	12.42	*45.99*	32.33	41.20	**53.70**	0.00	57.57
Walking Motion (6)	17.83	*51.55*	33.42	41.21	**54.87**	0.00	58.26	10.50	*40.90*	28.32	28.78	**47.53**	0.00	48.54
Walking Motion (7)	19.91	*54.40*	40.28	45.84	**58.72**	0.00	60.96	12.97	*44.37*	34.11	35.50	**50.77**	0.00	52.32
Walking Motion (8)	20.45	*55.71*	36.03	49.08	**61.50**	0.00	65.05	12.89	*43.37*	28.86	36.27	**52.47**	0.00	55.48

3.2 実験結果

表 2 に実験結果を示す．**F** は雑音が既知であり，このような環境は実用的には有り得ないので参考として示している．**A** から **E** までの中で最も認識性能のよかった手法をボールド体で示し，二番目に性能のよかったものをイタリック体で示す．提案手法の有意性を示すため，p 値 [15] を合わせて示している．p 値は手法 **B** をベースラインとし，提案手法である **E** の危険率を示す．

手法 **F** はマスクの生成過程において雑音が既知であるため，認識性能が一番よい．しかし，この手法を除くと提案手法である **E** が最もよく，手法 **D** と手法 **B** が次に続く．全ての距離，全ての動作雑音に対し提案手法 **E** はベースライン **B** よりも高い性能を示した．さらに，34 雑音× 4 距離のうちで，p 値が 5% を超えるものはわずか 7 のみであった．

4 考察

提案手法 **E** は従来の手法 **B** よりも全ての環境で高い性能を示している．特に SNR の低い環境である 200 cm においてはその有効性が大きい．MFT を用いない **D** と従来手法 **B** を比較するとその有効性が大きいとはいえないが，MFT を用いることでロボットの動作音に頑健な音声認識を行うことが可能となる．

提案手法 **E** において，**D** よりも性能が高くなっていることから，雑音のマッチングもうまく動作していることが確認できる．本稿の手法では，音声と雑音の重畳した入力信号を用いてテンプレート雑音とのマッチングを行うが，データベースにおける雑音よりもパワーの大きな箇所は音声信号も有すると仮定し，マッチングの際に考慮に入れないようにすることで，音声信号が含まれていたとしても雑音推定が可能であることが確認できる．

E は音声と雑音の重畳された入力信号からテンプレート雑音との雑音マッチングを行って推定雑音を求めるが，**F** は雑音が既知として MFT のマスク生成を行う．**F** は **E** と比べると理想的な環境であるため，認識性能も向上している．しかし，**E** も **F** と比較してよい性能を示しており，雑音マッチングが音声と雑音が重畳していても上手くできることが確認できる．50 cm の環境では，**F** の方が **E** よりも性能が低くなっているものも見られる．条件 **F** は雑音が既知であるが，これを用いた MFT マスクは正解を導くのに a priori なマスクということはできない．本稿で用いているマスク生成手法は，音声認識において重要と考えられるスペクトルの山と谷に重みを置き，さらに雑音の小さな箇所に大きな重みとなるようにするものである．しかし，音響モデルは完全にクリーンな音声のみで学習されたモデルではないため，このマスク生成手法がどのような入力信号に対しても最もよいマスクを生成するとは限らない．したがって，雑音が既知であっても性能が最高ではない環境が現れたと考えられる．しかし，全体的に見ると MFT による効果は明らかであるため，このマスク生成手法は多くの場合に有効な手法であると捉えることができる．

今回比較した実験では，定常雑音を音響モデルの学習に組み込んだ．すなわち，定常雑音を含んだ音声で学習したマルチコンディションによる音響モデルを用いた手法 **B** と，定常雑音から雑音除去処理，白色雑音の重畳を行った音響特徴量を学習した音響モデルを用いた提案手法 **E** の比較を行った．これは，音響モデルの学習では定常雑音の学習が有効であると考えたからである．しかし，定常雑音のみではなく，動作時の非定常雑音の学習も行うことで，認識性能向上の可能性があると考えている．

5 まとめ

ロボットの動作雑音除去を目的とした雑音適応手法の提案を行った．提案手法では，雑音除去処理と，歪みを補正するための白色雑音の重畳，MFT を用いることによる非定常雑音への適応を行う．雑音除去処理は SNR の高い環境では有効であるが，雑音が大きな環境においては歪みが大きく生じる．本研究ではこの歪みを抑えるために，白色雑音を重畳した．白色雑音を重畳させることは SNR を低下させ，一見，認識性能を下げるようにも思えるが，雑音の引き残し成分を平坦化し歪みを補正することで認識性能は大きく改善した．また，雑音除去処理は定常雑音には有効であるが，非定常雑音に対しては適応しきれない部分が存在する．この点を補完するため，MFT を用い，信頼性の低い部分が音声認識に関与する割合を低くすることにより認識性能の向上を図った．提案手法を用いることにより，従来から用いられている有効な方法であるマルチコンディション学習による音響モデルを用いた手法よりも高い認識性能を達成できた．

6 今後の課題

今後の課題としては，白色雑音の重畳の割合をいかにするとよい性能が出るか検討を行う必要があると考えている．また，今回は収録した音声および雑音を用いて提案手法の有効性の検証を行ったが，実際に ASIMO の音声認識システムの中に提案手法を組み込み，リアルタイムで処理できることを確認していきたいと考えている．

謝辞

MFT を用いるにあたり貴重なアドバイスを頂いた東京工業大学教授古井貞熙氏，および岩野公司氏に感謝する．また，本実験を行うにあたり貴重なアドバイスを頂いた HRI-JP の船越孝太郎氏および雑音の収録にあたりお手伝いいただいた京都大学山本俊一氏に感謝する．

参考文献

[1] C. Breazeal, *Designing Sociable Robots*, MIT press, 2002.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoo, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proc. of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2004)*. 2004, pp. 2404–2410, IEEE.

[4] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[5] S. Yamamoto, K. Nakadai, J. M. Valin, J. Rouat, F. Michaud, T. Ogata, H. Komatani, and H. G. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2005)*, IEEE, Ed., 2005, pp. 897–892.

[6] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal noise suppression for speech recognition by small robots," in *Proc. of European Conference on Speech Communication and Technology (Eurospeech-2005)*, 2005, pp. 2685–2688.

[7] Boll and S. F., "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*. 1979, pp. 200–203, IEEE.

[8] J. Barker, M. Cooke, and P. Green, "Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. of 7th European Conference on Speech Communication Technology (Eurospeech-2001)*. 2001, vol. 1, pp. 213–216, ESCA.

[9] A. Hagen and A. Morris, "Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR," in *Proc. of International Conference on Spoken Language Processing (ICSLP-2000)*, 2000, vol. 1, pp. 345–348.

[10] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. of International Conference on Spoken Language Processing (ICSLP-1996)*, 1996, vol. 1, pp. 426–429.

[11] 山出慎吾, 馬場朗, 芳澤伸一, 李晃伸, 猿渡洋, 鹿野清宏, "実環境における頑健な音声認識のための音韻モデルの教師なし話者適応," **電子情報通信学会論文誌**, vol. J87-D-II, no. 4, pp. 933–941, 2004.

[12] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," in *Proc. of 148th Acoustical Society of America Meetings*, ASA, Ed., 2004, p. 1aSC7.

[13] 西村義隆, 篠崎隆宏, 岩野公司, 古井貞熙, "周波数帯域ごとの重みつき尤度を用いた雑音に頑健な音声認識," **信学技報**, *SP2003-116*, 2003, pp. 19–24.

[14] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[15] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89)*, IEEE, Ed., 1989, pp. 532–535.

ICA よる音源分離と MFT に基づく音声認識の同時発話認識による評価

Evaluation of the ICA BSS and Missing Feature Theory Based-ASR with Simultaneous Speech Recognition

○武田 龍, 山本 俊一, 駒谷 和範, 尾形 哲也, 奥乃 博

Ryu TAKEDA, Shun'ichi YAMAMOTO, Kazunori KOMATANI, Tetsuya OGATA, and Hiroshi G. OKUNO

京都大学大学院情報学研究科 知能情報学専攻

Graduate School of Informatics, Kyoto University

{rtakeda, shunichi, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

Robot audition systems require capabilities for sound source separation and the recognition of separated sounds. We report a robot audition system with a pair of omni-directional microphones embedded in a humanoid that recognizes two simultaneous talkers. It first separates the sound sources by Independent Component Analysis (ICA). Then, spectral distortion in the separated sounds is then estimated to generate missing feature masks (MFM). Finally, the separated sounds are recognized by missing-feature theory (MFT) for Automatic Speech Recognition (ASR). We estimate of spectral distortion in the temporal-frequency domain in terms of feature vectors and generate MFM. The resulting system outperformed the baseline robot audition system by 13 % with isolated word recognition, and 6 % with continuous speech recognition.

1 はじめに

将来, 様々な面で人間をサポートするようなヒューマノイドロボットは人間と同等の認識能力を有する必要がある. 音声は人間同士のコミュニケーションにおいて重要な位置を占めており, 実環境における音声認識は基本的なロボット聴覚機能といえる. 音声を認識するには目的音以外を除去する必要がある. 特に複数の話者が同時に話している時には雑音除去だけでなく, それぞれの音声を聞き分ける機能は不可欠である.

今後ロボットは様々な環境で動作することを考えると, 音源分離・混合音認識といったロボット聴覚機能を実現する上で必要不可欠な条件は, できるだけ特定の環境に特化しない処理を実現することである. これまでに提案されている分離手法の多くは, ロボットに装着されたマイクロホンの位置や, 目的話者の位置などの情報を必要とする[1]. しかし, 実環境において分離に十分な精度の情報を動的に取得することは難しい. また, 混合音認識においても, 例えば, ノイズなどを学習データに加えることで, 環境の変化に対応したマルチコンディション学習が有効である[2]. ところが, 環境毎に学習データを準備しなければならず, 汎用的に利用できない欠点がある.

本稿では, (a) 音声の独立性のみを仮定する ICA (独立成分分析) による音源分離を行う. さらにクリーン音声での学習のみで, ICA の分離による歪みに対応できる (b) ミッシングフィーチャ理論 (MFT) を応用した音声認識を用いる. これにより必要な事前情報を必要最小限に抑えることができる. ここで課題となるのは, (1) 特徴量における歪みの検出, (2) 特徴量の信頼度設定, (3) 分離出力からのミッシングフィーチャマスク (MFM) の自動生成, である. これらの問題に対し, ノイズによる特徴量の変化量に着目し, 分離出力から擬似的にノイズ成分を変動させることで特徴量の歪みを検出する. 歪みの大きい部分を適切な閾値処理を行い, MFM を自動生成する.

ICA による分離では近年リアルタイムで動作が可能であるため [3], 処理速度による影響はほとんどないといえる. また, 同様のアプローチとして Kolossa ら [4] による手法があるが, 特徴量として MFCC (mel frequnecy cepstral coefficient) を用いており, スペクトル特徴量での検討は行われてはいない.

2 基本手法

我々がロボット聴覚システムを構築する上で用いる音源分離手法, 及びミッシングフィーチャ理論を応用した音声認識システムを説明する. これらを用いたシステムの全体像は 図 1 のようになる.

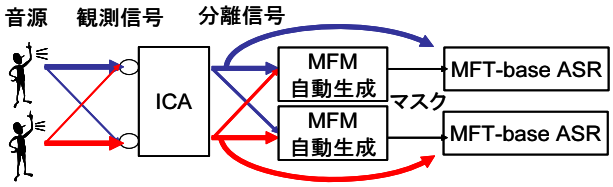


図 1: 処理の概要

2.1 ICA による音源分離

ICA は音源の独立性のみを仮定する分離手法であり, Blind Source Separation の一つである. ICA は時間領域, 周波数領域のどちらでも適用することが可能であるが, ここでは収束の早い周波数領域で ICA を適用する.

2.1.1 音声の混合過程

一般に, 複数の音源信号が線形不変な伝達系を経て混合された場合, その観測信号は次式で表される.

$$\boldsymbol{x}(t) = \sum_{n=0}^{N-1} \boldsymbol{a}(n)\boldsymbol{s}(t-n) \tag{1}$$

ここで, $\boldsymbol{s}(t) = [s_1(t), ..., s_I(t)]^T$ は音源信号ベクトル, $\boldsymbol{x}(t) = [x_1(t), ..., x_J(t)]^T$ はマイクロホンアレイにおける観測信号ベクトル, $\boldsymbol{a}(n) = [a_{ji}(n)]_{ji}$ は伝達系のインパルス応答を表す J 行 I 列の混合行列である. また, $[\cdot]_{ji}$ は j 行 i 列要素が $\cdot$ である行列を表す. 本研究では音源数 I とマイクロホンの数 J は等しく 2 であると仮定する. 周波数領域で ICA を適用すると, そのモデルは瞬時混合モデルとなる.

2.1.2 周波数領域 ICA

周波数領域で ICA を適用するため, 短時間分析を用いてフレーム毎に離散フーリエ変換された信号を入力とする. これより観測信号ベクトルは $\boldsymbol{X}(\omega, t) = [X_1(\omega, t), ..., X_J(\omega, t)]$ と表現できる. 次に, 分離行列 $\boldsymbol{W}$ を用いて, 分離信号 $\boldsymbol{Y}(\omega, t) = [Y_1(\omega, t), ..., Y_I(\omega, t)]$ を周波数毎に独立に以下の式で求める.

$$\boldsymbol{Y}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, t) \tag{2}$$

式 (2) を解くため, KL 情報量最小化に基づいて分離行列を推定する. ここでは以下の反復学習則を用いる[5], [6].

$$\boldsymbol{W}^{j+1}(\omega) = \boldsymbol{W}^{j}(\omega) - \alpha\{\text{off-diag}\langle\boldsymbol{\phi}(\boldsymbol{Y})\boldsymbol{Y}^{h}\rangle\}\boldsymbol{W}^{j}(\omega) \tag{3}$$

ここで, α は学習係数, $[j]$ は更新回数, $\langle\cdot\rangle$ は平均である. また, off-diag $(\boldsymbol{X})$ は対角要素を零に置き換える演算であり, 非線形関数ベクトル $\boldsymbol{\phi}(\boldsymbol{y})$ は $\phi(y_i) = \tanh(|y_i|)e^{j\theta(y_i)}$ である. スケーリング問題は Projection Back [7] によって解決した.

2.2 ミッシングフィーチャ理論に基づく音声認識

ミッシングフィーチャ理論を適用した音声認識を利用する場合, 次の 2 点が音声認識の核を成す部分であるため非常に重要である.

1. 音声認識特徴量
2. 信頼度を取り入れた出力確率の算出

この章ではこれら 2 点について検討する.

2.2.1 音声認識特徴量

MFT ベースの音声認識システムでは音声認識の特徴量として, MFCC ではなく, スペクトル特徴量 (mel scale log spectrum: MSLS) を用いる. MFCC は入力音声の歪みが少ない場合は有効であるが, 入力スペクトルに歪みがあると, 特徴量全体に影響を与えてしまい, 認識性能が低下する [1]. 一方, MSLS はスペクトル領域の特徴量であるため, ノイズは加法的であり, 歪みの検出が比較的容易である [8].

本研究ではスペクトル特徴量として MFCC の計算過程のケプストラム平均除去後, 逆コサイン変換を行いスペクトル領域に戻した 24 次元と, 1 次のデルタ特徴量 24 次元と合わせた計 48 次元を用いる.

2.2.2 信頼度を取り入れた出力確率の算出

MFT に基づく音声認識システムでは, 信頼度を考慮した出力確率計算を行う. この信頼度付きの出力確率の計算として, 西村ら [8] の周波数毎に重みをつける手法を採用する. この手法では計算量が比較的少なく, 高速に動作するという利点がある. なお, このように尤度計算に歪みを考慮する手法は一般的に marginalization approach (周辺化法) と呼ばれる [9].

特徴ベクトル x, 状態 S_j の時の正規分布の確率密度関数を $f(x|S_j)$, L を混合正規分布の混合数, $P(l|S_j)$ を混合係数とする. この時, 通常の連続分布型 HMM では出力確率は以下のように定義される.

$$b_j(x) = f(x|S_j) = \sum_{l=1}^{L} P(l|S_j) f(x|l, S_j) \tag{4}$$

ここで, MFT に基づく音声認識では, 出力確率 $b_j(x)$ は信頼できる特徴量ほど出力確率に大きく貢献し, 信頼できないほど貢献しないように設計する. 特徴量の各成分 i に対する信頼度を表す MFM ベクトル $M(i)$, 及び特徴量の次元数 N を用いて出力確率を次のように定義する.

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left(\sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j)\right), \tag{5}$$

信頼度が 0 である時, その特徴量に対する尤度は (5) 式よりすべてのクラスに対して等しくなる. 信頼度が全て 1 である時, 従来の音声認識デコーダと同じ動作をする.

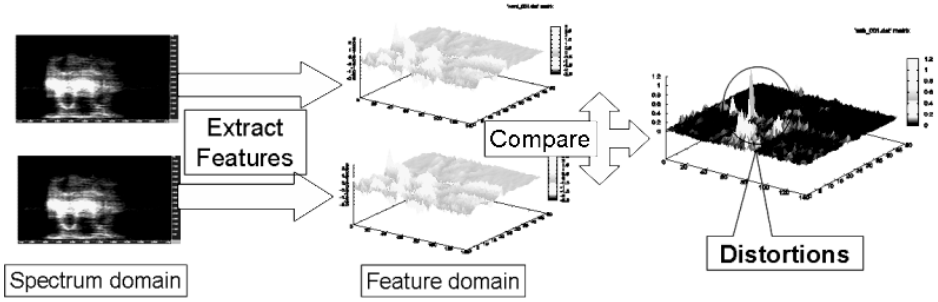


図 2: MSLS における歪みの検出

3 ミッシングフィーチャマスクの自動生成

MFM は特徴量領域において歪んだ特徴量を検出して作成される. MFM の生成には次の 3 ステップが必要である.

1. 特徴量領域での歪みの検出
2. 歪みに基づいた信頼度の設定
3. 分離出力からのマスク生成

特に特徴量領域での歪みの検出では, その検出手法は使用する特徴量に依存する.

3.1 特徴量歪みの検出

3.1.1 特徴量歪み

特徴量は本来定数倍に関して無視できるものである. この処理を行わない場合, 音量によって認識率が変化するからである. 特徴量への写像 $\boldsymbol{F}$ が有限の不連続点を除き, 連続微分可能であると仮定する.

時間周波数領域 $(t = 0, \ldots, T, w = 0, \ldots, W)$ の信号 $\boldsymbol{x} = (x_{0,0}, \ldots, x_{0,T}, x_{1,T}, \ldots, x_{W,T})$ において, 各周波数毎 $(w = 0, \ldots, W)$ の定数倍 $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_0, \alpha_1, \ldots, \alpha_W)$ に対して,

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{\alpha} .* \boldsymbol{x}) \tag{6}$$

が成り立つ. ここで, $.*$ はベクトルの要素同士の積 $\boldsymbol{x} \cdot \boldsymbol{y} = (x_0 y_0, x_1 y_1, \ldots, x_n y_n)$ を表す演算子である. ここで, 混合信号 $\boldsymbol{\alpha x} + \boldsymbol{\beta y}$ の特徴量を目的信号 $\boldsymbol{s}$ に着目し,

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\alpha x} + \boldsymbol{\beta y}) &= \boldsymbol{F}(\boldsymbol{x} + (\boldsymbol{\beta} ./ \boldsymbol{\alpha})\boldsymbol{y}) & (7) \\ &= \boldsymbol{F}(\boldsymbol{x} + \boldsymbol{\theta} .* \boldsymbol{y}) & (8) \end{aligned}$$

のように表現することにする.

ここで, 信号 $\boldsymbol{s}$ に雑音源 $\boldsymbol{n}$ が加わった $\boldsymbol{s} + \boldsymbol{n}$ の特徴量歪み $\boldsymbol{D}$ を以下のように定義する.

$$\boldsymbol{D} = \boldsymbol{F}(\boldsymbol{s} + \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s}) \tag{9}$$

つまり, 特徴量歪みをノイズ込みの特徴量と目的音源の特徴量との差であるとする.

3.1.2 歪みの検出

歪みの検出は一般的に利用する特徴量に依存する. ここでは, スペクトル特徴量で歪みを検出する方法を説明する.

スペクトル特徴量において, ノイズに対して特徴量における歪みは単調に変化すると仮定する. 特に, ノイズに対する変化がほぼ線形的であるとみなせる時, 特徴量における歪みはその変化量に比例することになる.

$$\begin{aligned} \boldsymbol{D} &= \boldsymbol{F}(\boldsymbol{s} + \theta\boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s}) & (10) \\ &\simeq \theta \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}(\boldsymbol{s})\boldsymbol{n} & (11) \end{aligned}$$

今, 2 つの特徴量 $\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{\alpha n})$, $\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{\beta n})$ が得られているとする. 特にこの係数 $\boldsymbol{\alpha}, \boldsymbol{\beta}$ の比が γ に近いとき, 上式にしたがって特徴量歪みは定数倍の曖昧性を除き,

$$\begin{aligned} \boldsymbol{D} &\simeq \boldsymbol{F}(\boldsymbol{s} + \boldsymbol{\alpha} .* \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s} + \boldsymbol{\beta} .* \boldsymbol{n}) & (12) \\ &\simeq \boldsymbol{F}(\boldsymbol{s} + \boldsymbol{\alpha} .* \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s} + \gamma\boldsymbol{\alpha} .* \boldsymbol{n}) & (13) \\ &\simeq \gamma \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}(\boldsymbol{s})\boldsymbol{\alpha} .* \boldsymbol{n} & (14) \end{aligned}$$

によって検出することができる. また, 係数の比が一定でなくとも, ノイズによる影響が大きい特徴量はその変化も大きいと考えられるため, 歪みを検出することできる.

3.2 信頼度の設定

次に, 検出した歪みに基づいて信頼度の設定を行う. 信頼度には reliable / unrelible の 2 値を用いる. 理想的なマスク *A priori* マスクの生成法と, 検出した歪みに基づくマスクの自動生成を説明する.

3.2.1 *A priori* マスク

A priori マスクは真の特徴量からの差, つまり上記で定義した歪み $\boldsymbol{D}$ を一定の閾値で切ることで生成される. 生成すべきマスク $\boldsymbol{AM}$ を閾値 $\boldsymbol{T}$ を用いて以下のように定義する.

$$\boldsymbol{AM} = \begin{cases} 1 & |\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s})| < \boldsymbol{T} \\ 0 & otherwise \end{cases} \tag{15}$$

ここで, この絶対値はベクトルの各要素ごとに掛かるものとする. この閾値 $\boldsymbol{T}$ は学習した音響モデルに合わせて, それぞれの要素ごとに最適な値を設定するべきであるが, 今回は n 次デルタ特徴量には特定の閾値 T_n を設ける.

3.2.2 自動生成マスク

式（15）によって検出された歪み $\boldsymbol{D}$ に対してマスクを生成する．

$$\boldsymbol{D} \simeq |\boldsymbol{F}(\boldsymbol{s}+\boldsymbol{\alpha}.*\boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s}+\boldsymbol{\beta}.*\boldsymbol{n})| \quad (16)$$

この歪みに対して，閾値 T_n によりマスク M を以下のように作ればよい．

$$\boldsymbol{M} = \begin{cases} 1 & |\boldsymbol{F}(\boldsymbol{s}+\boldsymbol{\alpha}.*\boldsymbol{n}) - \boldsymbol{F}(\boldsymbol{s}+\boldsymbol{\beta}.*\boldsymbol{n})| < T_n \\ 0 & otherwise \end{cases} \quad (17)$$

ここで，T_n は n 次デルタ特徴量に対する閾値である．

3.2.3 ICA の出力を用いたマスク生成

次に ICA の出力からミッシングフィーチャマスクの自動生成を行う方法を提案する．ICA で推定された分離フィルタを $\boldsymbol{W}$，真の混合行列を $\boldsymbol{M}$ とし，分離フィルタの誤差を $\boldsymbol{E} = \boldsymbol{M}^{-1} - \boldsymbol{W}$ で表す．また，元信号を $\boldsymbol{s}$，観測信号を $\boldsymbol{x}$ としたとき，ICA で分離された信号 $\boldsymbol{y}$ は周波数領域で次のように表現される．

$$\begin{aligned} \boldsymbol{y}(\omega,t) &= \boldsymbol{W}(\omega)\boldsymbol{x}(\omega,t) \\ &= \boldsymbol{W}(\omega)\boldsymbol{M}(\omega,t)\boldsymbol{s}(\omega,t) \\ &= \left(\boldsymbol{M}^{-1}(\omega) - \boldsymbol{E}(\omega)\right)\boldsymbol{M}(\omega)\boldsymbol{s}(\omega,t) \\ &= \boldsymbol{s}(\omega,t) - \boldsymbol{E}(\omega)\boldsymbol{M}(\omega)\boldsymbol{s}(\omega,t) \end{aligned} \quad (18)$$

このように，分離された信号における誤差は混合行列 M と誤差行列 E に依存する．ここで，2 音源の場合 $\boldsymbol{y}$ 中のある信号 y_1 に着目した時，その信号はスケーリング w_1 を合わせることで，

$$w_1 y_1(\omega,t) = w_1\left(1+e_1\right)s_1(\omega,t) - \hat{e}_1 w_1 s_2(\omega,t) \quad (19)$$

と表現できる．ただし $e_1, \hat{e}_1$ は適当な係数である．y_2 に対しても同様に得られる．このとき，

$$\hat{y}_1(\omega,t) = w_1 y_1(\omega,t) - \gamma w_2 y_2(\omega,t) \quad (20)$$

によって，$w_1 y_1, \hat{y}_1$ の 2 つの信号を得ることができ，式 (14) によって歪みを検出できる．なお，γ を適切に定めることができれば，$w_1 y_1$ 中の y_2 成分の影響を最小限にすることができるが，それを現実的に行うことは非常に困難である．実際に検出した歪みを図 2 に示す．

4 実験

本システムの評価を行うためヒューマノイド SIG2 (図 4) の外耳道モデル (図 3) に埋め込まれた 2 本の無指向性マイクロホンで 2 話者同時認識実験を行った．評価は次の 5 点についておこなう．

1. 孤立単語認識実験
   (a) パラメータ関係，γ と T の評価
   (b) ICA 出力と MFM の効果
   (c) MSLS と MFCC との特徴量比較
2. 連続音声認識実験
   (a) ICA 出力と MFM の効果
   (b) MSLS と MFCC との特徴量比較

図 3: SIG2 に設置された耳

図 4: Humanoid SIG2

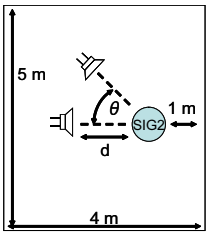


図 5: 配置 1: 半対称

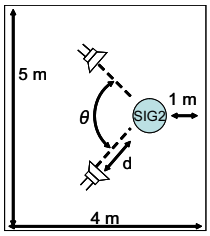


図 6: 配置 2: 対称

孤立単語認識実験の目的は，実用上の観点から文法ベースでの評価を行うため，連続音声認識実験の目的は，一般的な条件下での評価を行うためである．これらの実験によって，(1) 本手法で用いられる 2 つのパラメータの関係，(2) ICA による分離効果と MFM の効果，(3) MSLS と MFCC の特徴量に関して検証する．また，観測信号自体をそのまま用いてマスクを生成したものと比較し，ICA による分離によってマスク効果に変化があるかを検証した．

4.1 録音条件

録音には上述した SIG2 に設置されたマイクロホンを利用した．録音を行った部屋は 4m × 5m × 2.5m の広さで，残響時間 (RT20) が約 0.25 秒であった．このような条件で，実験 (1) では，2 話者同時発話 200 組を録音し，発話データは ATR 音素バランス単語を実験 (2) では，2 話者同時発話 100 組を録音し，発話データは ASJ-JNAS の評価用データセットを用いて録音を行った．

データセットは，(男性, 女性), の組み合わせで，マイクとスピーカの距離は 約 1.0m, スピーカーの配置は 1 つが正面固定・もう一つが右側に 30 度, 60 度, 90 度間隔で配置したもの (図 5), 正面に女性話者，右側に男性話者のものと，左右対称に配置し，その間隔が 30 度, 60 度, 90 度の組合わせとした (図 6). 録音したデータは 48kHz でサンプリングし，16kHz にダウンサンプリングを行った．

4.2 実験条件

音声認識エンジンにはマルチバンド版 Julian[8]を MFT に基づく音声認識システムとして使用した．音響モデルは実験 (1) では，トライフォン (3 状態 4 混合の HMM), ク

	孤立単語認識	連続音声認識
テストデータ	ATR 音素バランス単語: 男女各 200 語	ASJ-JNAS 新聞記事読み上げ: 男女各 100 文
学習データ	ATR 音素バランス単語: 男性 10 人, 女性 12 人, 各 216 語	新聞記事読み上げ + 音素バランス文 男性 100 人, 女性 100 人, 各 150 文
音響モデル	トライフォン: 3 状態 4 混合 HMM	トライフォン: 3 状態 8 混合 HMM
言語モデル	有限状態文法	統計的モデル: 毎日新聞記事 2 万語

表 1: 実験設定

リーン音声 22 話者 (男性 10 人, 女性 12 人) 分の ATR 音素バランス単語 216 語で学習, 実験 (2) では, PTM トライフォン (3 状態 8 混合の HMM), クリーン音声不特定話者約 200 人 (男性 100 人, 女性 100 人) 分の新聞記事読み上げ文と音素バランス文の計 150 文で学習した. また, 評価用データは学習に使用されていない. 連続音声認識での言語モデルは毎日新聞記事 2 万語の統計的言語モデルである. 表 1 にこれらをまとめた.

ICA のパラメータは, 録音データ 16 kHz サンプリングに対し, 窓幅 2048 点, シフト幅 512 点とした. 分離行列の初期値はランダム値である. マスクの閾値パラメータ T_0 は実験的に定め, MSLS の場合 m 孤立単語認識で 0.005, 連続音声認識で 0.02, MFCC では孤立単語認識で 0.01, 連続音声認識で 0.51 とした. スケーリングの値 γ は MSLS の場合 0.02, MFCC の場合 0.2 とした. デルタ特徴量に関してはマスクを行っていない. また, オフライン処理であるため, 孤立単語の分離においては 3〜5 秒程度単語を連結し, ある程度の分離精度は確保している.

5 実験結果

5.1 孤立単語認識実験

パラメータ関係 パラメータ関係を図 7 に示す. この図は $\gamma = 0.01, 0.1, 1.0$ の場合 (MFCC に関しては $\gamma = 1.0$ ではなく $\gamma = 0.5$) と *A priori* マスクの両方の閾値による認識率の変化を示している.

どのグラフでも, 閾値がある程度小さくなった場合と大きくなった場合では, 認識率がある値に収束している. *A priori* マスクにも見られるように, 閾値の変化に対して認識率は, 両端が一定で, ピークがある山形のような曲線を描いている. これは MFCC でも, MSLS でも同様の傾向がある.

さらに MSLS では, γ と T に相関があるとみなせる. 認識率の曲線が γ の値に対してほぼ線形な変化があると考えられ, スペクトル特徴量の歪みがノイズに対してほぼ線形に加わっているとみなせる.

ICA と MFM の効果 図 8 と図 9 に ICA による分離, 及びミッシングフィーチャマスクの効果をそれぞれの特徴量に関して示す.

MSLS, MFCC のいづれの特徴量でも認識率の向上が確認できる. MSLS では ICA の分離で平均 24 %, MFM で平均 13.3 % の認識率向上が見られる. MFCC においても 分離で平均 24 %, MFM で平均 6.2 % の向上がある.

また, 分離なしでマスクを生成した場合, MSLS では約 8.2 %, MFCC では約 2.1 % の向上であり, ICA 適用後のマスクよりもその効果は下回っている.

正解の特徴量に基づき作成した *A priori* マスクでは, ほぼ 90 % 以上の認識率を達成している.

MSLS と MFCC の比較 スペクトル特徴量単体であると, MFCC よりも認識率が低下している場合がある. MFM を用いることで, いづれの場合も MFCC 並の認識率を確保できている. しかし, 元の認識率が MFCC の方がいいため, MFCC + マスクの方が認識率が向上している.

5.2 連続音声認識実験

ICA と MFM の効果 2 つの特徴量による結果を図 10, 11 と, 図 12, 13 に示す.

分離なしの混合音声を認識させた場合では, 単語正解精度・単語正解率ともに認識率が悪い. 特に MSLS では, 単語正解精度がマイナスの値になっていることが分かる.

ICA による分離で, 単語正解率で MSLS が平均 7.22 %, MFCC では平均 11.07 % の改善, MFM でさらに, MSLS で 5.11 %, MFCC で 1.79 % の認識率の向上がある. 単語正解率では, ICA により, MSLS で平均 11.0 %, MFCC で 17.23 % の向上, MFM ではそれぞれ 8.68 %, 1.89 % となっている.

A priori マスクでは, 大幅な改善率を誇っているが, 上限値である単一発話者での認識率 MSLS 約 70%, MFCC 約 80 % に達していない. また, 自動生成マスクとも性能の差が大きいことがわかる.

MSLS と MFCC の比較 今回の設定では, MSLS よりも MFCC 特徴量の方が全体的な認識率は良い. ICA による分離では MFCC の方が改善率が高く, MFM では MSLS の方が高い傾向にある.

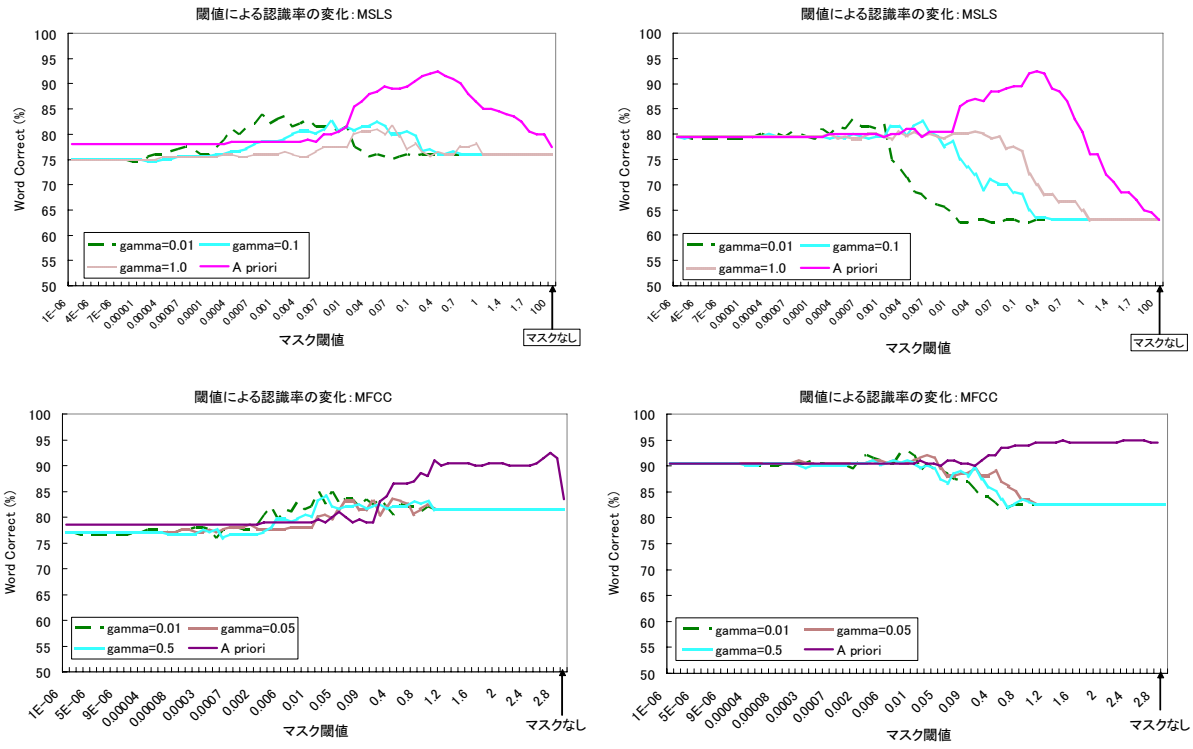


図 7: 閾値による認識率の変化：非対称 60 度間隔配置, 左:男性話者 右:女性話者, 上: MSLS, 下: MFCC

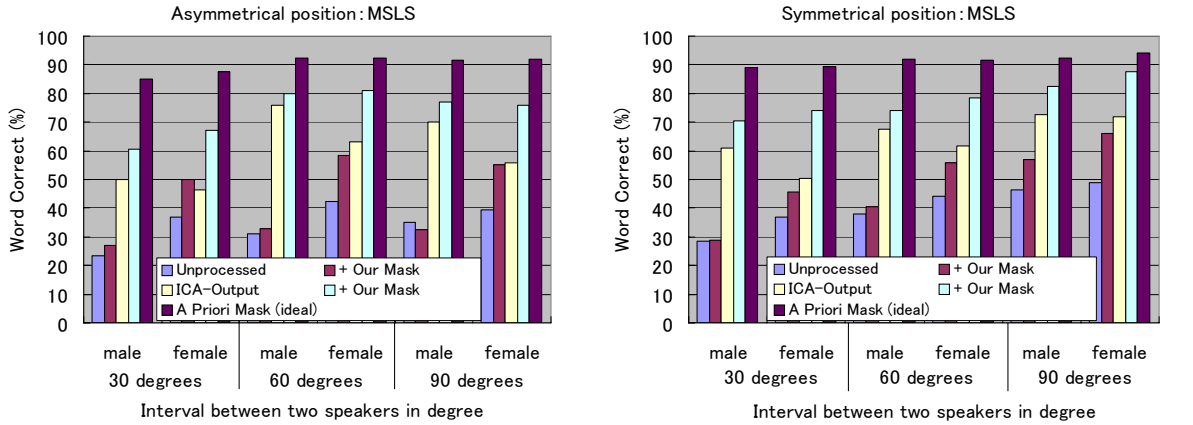


図 8: 孤立単語認識: 単語正解率：MSLS

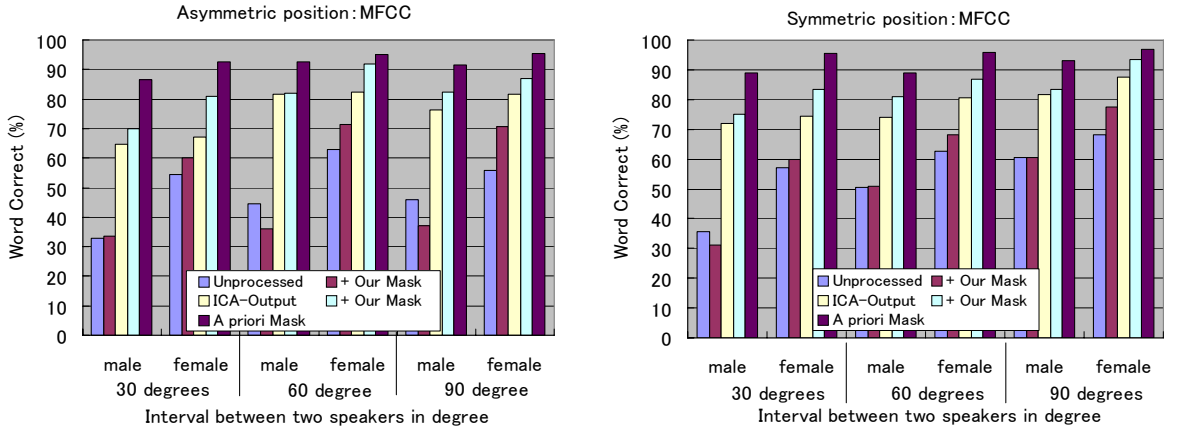


図 9: 孤立単語認識: 単語正解率 MFCC

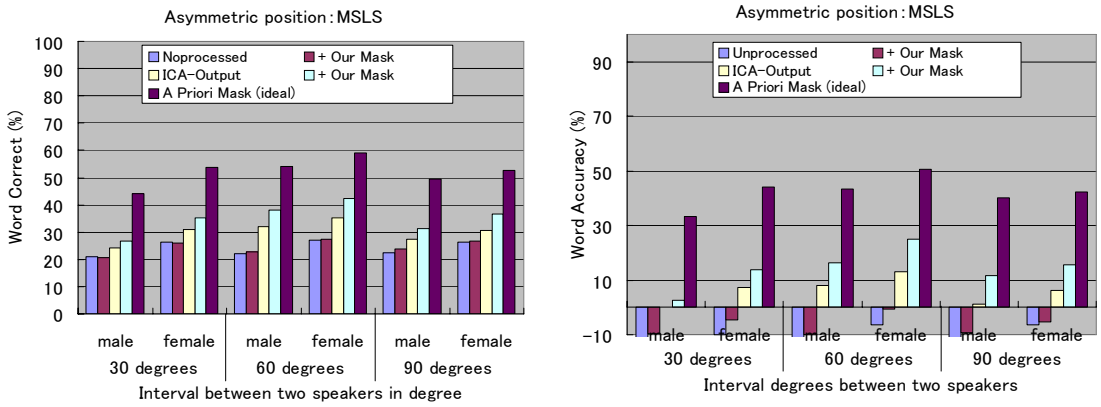


図 10: 非対称配置における単語正解率・正解精度: MSLS

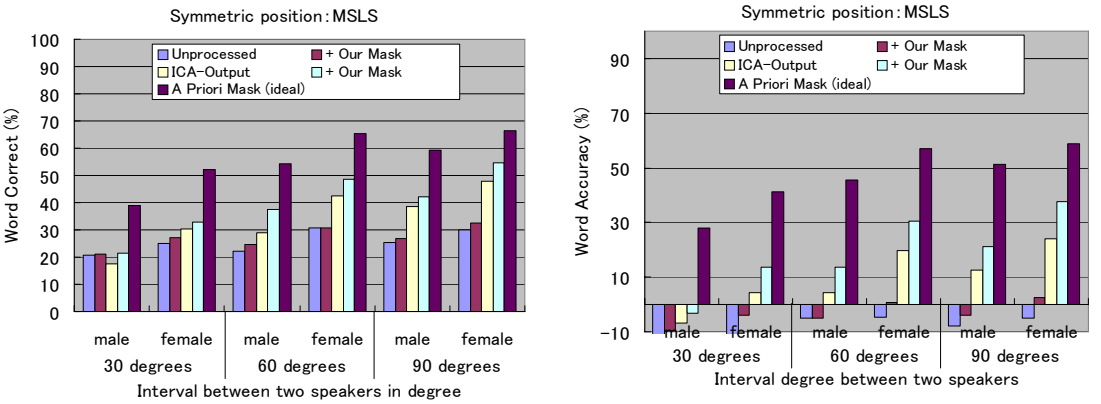


図 11: 対称配置における単語正解率・正解精度: MSLS

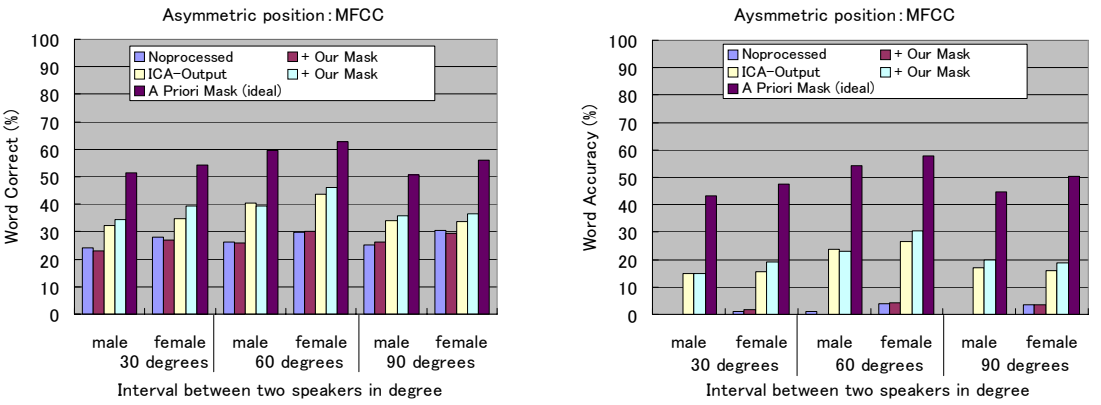


図 12: 非対称配置における単語正解率・正解精度: MFCC

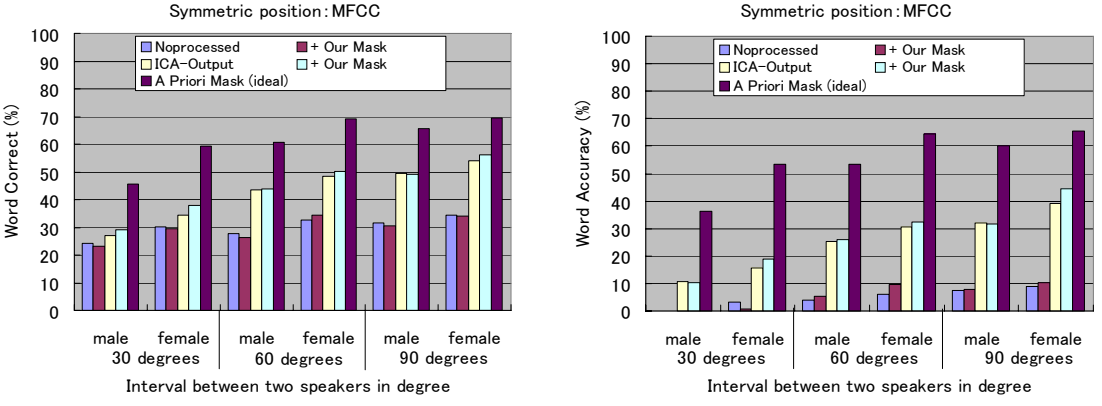


図 13: 対称配置における単語正解率・正解精度: MFCC

6 考察

6.1 特徴量歪みの検出法

今回の歪みの検出では, スペクトル特徴量を仮定し, ノイズの変化に対する特徴量の変化量で歪みを検出した. スペクトル特徴量はノイズが加法的に加わるため, 特徴量の変化量によって歪みやすい特徴量を検出できる. 一方, ケプストラル特徴量である MFCC では, ノイズが非線形に加わるため, ある点の変化量だけを見て歪みやすい部分を特定することは難しい.

孤立単語正解率では, 語彙数が少ないために MFCC マスクでも認識率が向上している. しかし, 連続音声認識では探索空間が巨大になり, より正確なマスク生成が要求されるため, 認識率の変化はほとんどない. 実際, MSLS でも MFCC でもマスクによる改善率が大きく落ちている. 分離を行うことでマスクによる効果は向上していることから, 他の手法との統合による歪み推定などが必要だろう.

閾値による認識率の変化をみると, 自動生成マスクでも閾値が小さくなるにつれて *A priori* マスクと同様にある一定の値に収束している. 閾値が極めて小さな値であるとき, 今回の方法だと確実に信頼できる部分しか残さないことを意味する. このため, 自動生成されたマスクがある程度の精度を保っていることが言える.

信頼度の設定では, 同一の歪みを検出しても音響モデルによって性能に差がでることも予想できる. そのため, 閾値の決定にはノイズレベルと同様に音響モデルに対しても適応することが望ましい.

6.2 音声認識特徴量

実験結果を見ると, MSLS よりも MFCC の方が絶対的な認識率が良い. これは, MFCC が音声特徴を良く捉えているからだと考えられる. ノイズや反響の無い条件では, MSLS でも次元数などを変えることで, MFCC と同様の性能を確保することが可能であるが, ノイズ環境化で同様の性能を出すことは難しい.

信頼度の推定では, ノイズに対して線形的に変化する特徴量の方が歪みを推定しやすい. 一方, ノイズに対する耐性では, 非線形的に変化する特徴量が有利である. MFT による音声認識を利用する場合では, 両方に有利な特徴量の追求が必要である.

7 おわりに

汎用性のあるロボット聴覚機能を実現するため, 事前情報の少ない音源分離と分離音声を認識可能にすることを目指した. 音源分離に ICA を用い, その後段処理として MFM の自動生成を行い, MFT を応用した音声認識器を利用した.

ノイズによる特徴量での変化量に基づき MFM の自動生成を行い, 孤立単語及び連続音声の 2 話者同時発話認識における認識率の改善を達成した. 今回の実験では, MFM の効果は孤立単語認識で約 13%, 連続音声認識で 約 5 % であることを確認した.

MFT による音声認識を用いる場合, 特徴量・信頼度付き尤度計算・信頼度設定・音響モデルなどが密接に関わっているため, それらの親和性が高い方法を採用する必要があるといえる.

今後の課題として挙げられるのは, 閾値の自動設定, 信頼度の設定方法の改善, 実時間での動作などを含め, より効果的な混合音認識手法の検討などがある.

参考文献

[1] 山本 他: “ミッシングフィーチャ理論を適用した同時発話認識システムの同時発話文による評価”, AI チャレンジ研究会, 22, pp.101-106, 2005.

[2] 中臺 他: “ロボットを対象とした散乱理論による三話者同時発話の定位・分離・認識の向上”, *JSAI Technical Report SIG-Challenge-0318-6*, pp.33-38, 2003.

[3] Saruwatari 他: “Two-Stage Blind Source Separation Based on ICA and Binary Masking for Real-Time Robot Audition System”, *Proc. of IROS 2005*, pp.209–214, 2005.

[4] Kolossa 他: “Separation and Robust Recognition of Noisy, Convolutive Speech Mixtures Using Time-Frequency Masking And Missing Data Techniques”, *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.82–85, 2005.

[5] Choi 他: “Natural Gradient Learning with a Nonholonomic Constraint for Bline Deconvolution of Multiple Channels”, *Proc. of International Workshop on ICA and BBS*, pp.371-376, 1999.

[6] Sawada 他: “Polar Coordinate based Nonlinear Function for Frequency-Domain Blind Source Separation”, *Proc. of IEICE Trans. Fundamentals*, 3, E86-A, pp.505-510, 2003.

[7] Murata 他: “An approach to blind source separation based on temporal structure of speech signals”, *Neurocomputing*, pp.1–24, 2001.

[8] 西村 他: “周波数毎の重みつき尤度を用いた音声認識の検討”, 音響学会 2004 年春講論, vol. 22, pp.117-118, 2004.

[9] Raj 他: “A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition”, *Speech Communication*, pp.379–393, 2004

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624-3 (11/17)

空間的サブトラクションアレーにおける 雑音推定処理の独立成分分析による高精度化

Improvement of Accuracy of Noise Estimation Based on Independent Component Analysis in Spatial Subtraction Array

高橋 祐，高谷 智哉，猿渡 洋，鹿野 清宏

Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, and Kiyohiro Shikano

奈良先端科学技術大学院大学

Nara Institute of Science and Technology

yuu-t@is.naist.jp

Abstract

In this paper, we propose a new spatial subtraction array (SSA) structure which includes independent component analysis (ICA)-based noise estimator. Recently, SSA has been proposed to realize noise-robust hands-free speech recognition. In SSA, noise reduction is achieved by subtracting the estimated noise power spectrum from the noisy speech power spectrum. The conventional SSA uses null beamformer (NBF) as a noise estimator, but NBF suffers from the adverse effect of microphone-element errors and room reverberations in real environments. To improve the problem, we newly replace NBF with ICA which can adapt its own separation filters to the element error and the reverberation. The affections by the element error and the reverberation can be mitigated in the proposed ICA-based noise estimator. Experimental results reveal that the accuracy of noise estimation by ICA outperforms that of NBF, and speech recognition performance of the proposed method overtakes that of the conventional SSA.

1 Introduction

A hands-free speech recognition system is essential for realizing an intuitive and stress-free human-machine interface. However, the quality of the distant-talking speech is always inferior to that of using close-talking microphone, and this leads to degradations of speech recognition. One approach for establishing a noise-robust speech recognition system is to enhance the speech signals by introducing microphone array signal processing. In delay-and-Sum (DS) array, we compensates the time delay for each element to reinforce the target signal arriving from the look direction. On the other hand, null beamformer (NBF) [1] provides more efficient noise reduction in which we steer the directional null to the direction of the noise signal. Moreover, Griffith-Jim adaptive array (GJ) [2] can achieve a superior performance relative to others. However, GJ requires a huge amount of calculations for learning adaptive multichannel FIR filters of, e.g., thousands or millions taps in total.

Spatial subtraction array (SSA) [3] is a successful candidate for hands-free speech recognition, and SSA is specifically designed for a speech recognition application. In SSA, noise reduction is achieved by subtracting the estimated noise power spectrum by NBF from the power spectrum of noisy observations in mel-scale filter bank domain. Since a common speech recognizer is not so sensitive to phase information, SSA which is performing subtraction processing only in the power spectrum domain is more applicable to the speech recognition, and it is reported that the speech recognition performance of SSA outperforms those of DS and GJ [3]. In SSA, noise estimation is performed by NBF which has decent performance under ideal conditions. However, NBF sustains the negative affection by microphone-element error and room reverberations. Therefore, in the real environment where the element error and the reverberation are always included, the performance of SSA significantly decreases because the noise-estimation accuracy by NBF decreases.

In this paper, we propose a new SSA structure which replaces NBF-based noise estimator with independent component analysis (ICA)[4]-based noise estimator. ICA is a technique for source separation based on indepen-

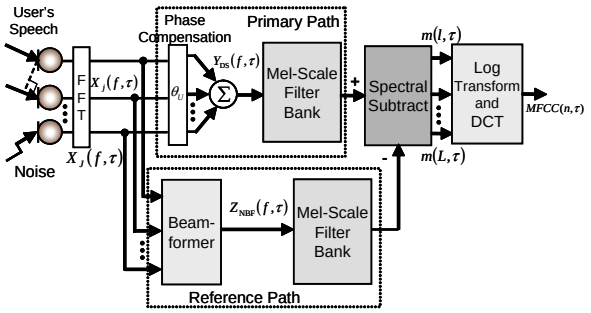


Figure 1: Block diagram of conventional SSA.

dence among multiple source signals. In acoustic source separation scenarios, ICA can also extract each source signal only using observed signals at the microphone array, and ICA does not require characteristics about sensor elements and the reverberation. Therefore, it is well expected that ICA can adapt its own separation filters to the element error and the reverberation. Accordingly the adverse effect by the element error and the reverberation can be mitigated in the proposed ICA-based noise estimator. Real-recording-based simulations are conducted, and we can indicate that the proposed method outperforms the conventional SSA on the basis of speech recognition performances.

2 Conventional Spatial Subtraction Array

2.1 Overview

The conventional SSA [3] consists of a DS-based primary path and a reference path via the NBF-based noise estimation (see Fig. 1). The estimated noise component by NBF is efficiently subtracted from the primary path in the power spectrum domain without phase information. In SSA, we assume that the target speech direction and speech break interval are known in advance. Detailed signal processing is shown below.

2.2 Partial speech enhancement in primary path

First, the short-time analysis of observed signals is conducted by a frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider these values as a time series. Hereafter, we designate the time series as

$$\boldsymbol{X}(f,\tau) = [X_1(f,\tau), \ldots, X_J(f,\tau)]^{\mathrm{T}}, \tag{1}$$

where J is the number of microphones, f is the frequency bin and τ is the frame number. Also, $\boldsymbol{X}(f,\tau)$ can be rewritten as

$$\boldsymbol{X}(f,\tau) = \boldsymbol{A}(f)\left(\boldsymbol{S}(f,\tau) + \boldsymbol{N}(f,\tau)\right), \tag{2}$$

$$\boldsymbol{S}(f,\tau) = [\underbrace{0,\ldots,0}_{U-1}, S_U(f,\tau), \underbrace{0,\ldots,0}_{K-U}]^{\mathrm{T}}, \tag{3}$$

$$\begin{aligned}\boldsymbol{N}(f,\tau) = [&N_1(f,\tau),\ldots,N_{U-1}(f,\tau),0,\\ &N_{U+1}(f,\tau),\ldots,N_K(f,\tau)]^{\mathrm{T}},\end{aligned} \tag{4}$$

where $\boldsymbol{A}(f)$ is a mixing matrix, $\boldsymbol{S}(f,\tau)$ is a target speech signal vector, $\boldsymbol{N}(f,\tau)$ is a noise signal vector, U expresses the target speech number, and K is the number of sound sources.

Next, the target speech signal is partly enhanced in advance by DS. This procedure can be given as

$$\begin{aligned}Y_{\mathrm{DS}}(f,\tau) &= \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{X}(f,\tau)\\ &= \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{S}(f,\tau)\\ &\quad + \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau),\end{aligned} \tag{5}$$

$$\boldsymbol{W}_{\mathrm{DS}}(f) = [W_1^{(\mathrm{DS})}(f),\ldots,W_J^{(\mathrm{DS})}(f)]^{\mathrm{T}}, \tag{6}$$

$$W_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_s d_j \sin\theta_U/c\right), \tag{7}$$

where $Y_{\mathrm{DS}(\mathrm{f},\tau)}$ is a primary-path output which slightly enhances the target speech, $\boldsymbol{W}_{\mathrm{DS}}(f)$ is a filter coefficient vector of DS, M is the DFT size, f_s is sampling frequency, d_j is a microphone position, and c is sound velocity. Besides, θ_U is a known direction-of-arrival (DOA) of the target speech. In Eq. (5), the second term in the right-hand side expresses the remaining noise in the output of the primary path.

2.3 Noise estimation in reference path

In the reference path, we estimate the noise signal by using NBF. This procedure is given as

$$Z_{\mathrm{NBF}}(f,\tau) = \boldsymbol{W}_{\mathrm{NBF}}^{\mathrm{T}}(f)\boldsymbol{X}(f,\tau), \tag{8}$$

$$\boldsymbol{W}_{\mathrm{NBF}}(f) = \left\{[1,0]\cdot[\boldsymbol{a}(f,\theta_O),\boldsymbol{a}(f,\theta_U)]^{+}\right\}^{\mathrm{T}}, \tag{9}$$

$$\boldsymbol{a}(f,\theta) = [a_1(f,\theta),\ldots,a_J(f,\theta)]^{\mathrm{T}}, \tag{10}$$

$$a_j(f,\theta) = \exp\left(i2\pi(f/M)f_s d_j \sin\theta/c\right), \tag{11}$$

where $Z_{\mathrm{NBF}}(f,\tau)$ is the estimated noise by NBF, $\boldsymbol{W}_{\mathrm{NBF}}(f)$ is a NBF-filter coefficient vector which steers the directional null in the direction of the DOA of the target speech, θ_U, and steers unit gain in the arbitrary direction $\theta_O(\neq\theta_U)$. $\boldsymbol{a}(f,\theta)$ is a steering vector which expresses phase information of the sound source arriving from the direction θ. Besides, $\boldsymbol{M}^{+}$ denotes Moore-Penrose pseudo inverse matrix of $\boldsymbol{M}$. This processing can suppress the target speech arriving from θ_U, which is equal to an extraction of noises from sound mixtures if we take into account affections of sensor errors and

reverberations. Thus we can estimate the noise signals by NBF under ideal conditions. Note that $Z_{\mathrm{NBF}}(f,\tau)$ is the function of the frame number τ, unlike the constant noise prototype estimated in the traditional spectral subtraction method [5]. Therefore, SSA can deal with a *non-stationary* noise.

2.4 Mel-scale filter bank analysis

SSA includes mel-scale filter bank analysis, and outputs mel-frequency cepstrum coefficient (MFCC) [6]. The triangular window $W_{\mathrm{mel}}(k;l)$ $(l = 1,\cdots,L)$ to perform mel-scale filter bank analysis is designated as follows:

$$W_{\mathrm{mel}}(f,l) = \begin{cases} \dfrac{f - f_{\mathrm{lo}}(l)}{f_{\mathrm{c}}(l) - f_{\mathrm{lo}}(l)} & \big(f_{\mathrm{lo}}(l) \le f \le f_{\mathrm{c}}(l)\big), \\ \dfrac{f_{\mathrm{hi}}(l) - f}{f_{\mathrm{hi}}(l) - f_{\mathrm{c}}(l)} & \big(f_{\mathrm{c}}(l) \le f \le f_{\mathrm{hi}}(l)\big), \end{cases} \tag{12}$$

where $f_{\mathrm{lo}}(l)$, $f_{\mathrm{c}}(l)$, and $f_{\mathrm{hi}}(l)$ are the lower, center, and higher frequency bins of each triangle window, respectively. They satisfy the relation among adjacent windows as

$$f_{\mathrm{c}}(l) = f_{\mathrm{hi}}(l-1) = f_{\mathrm{lo}}(l+1). \tag{13}$$

Moreover, $f_{\mathrm{c}}(l)$ is arranged in regular intervals on mel-frequency domain. Mel-scale frequency $Mel_{f_{\mathrm{c}}(l)}$ for $f_{\mathrm{c}}(l)$ is calculated as

$$Mel_{f_{\mathrm{c}}(l)} = 2595 \log_{10}\left\{1 + \frac{f_{\mathrm{c}}(l) f_s}{700 \cdot M}\right\}. \tag{14}$$

2.5 Noise reduction processing

In SSA, noise reduction is carried out by subtracting the estimated noise power spectrum from the partly enhanced target speech power spectrum in the mel-scale filter bank domain as

$$m(l,\tau) = \begin{cases} \displaystyle\sum_{f=f_{lo}(l)}^{f_{hi}(l)} W_{\mathrm{mel}}(f;l)\{|Y_{\mathrm{DS}}(f,\tau)|^2 - \alpha(l)\cdot\beta\cdot|Z_{\mathrm{NBF}}(f,\tau)|^2\}^{\frac{1}{2}} \\ \qquad (\text{ if } \ |Y_{\mathrm{DS}}(f,\tau)|^2 - \alpha(l)\cdot\beta\cdot|Z_{\mathrm{NBF}}(f,\tau)|^2 \ge 0\ ), \\ \displaystyle\sum_{f=f_{lo}(l)}^{f_{hi}(l)} W_{\mathrm{mel}}(f;l)\{\gamma\cdot|Y_{\mathrm{DS}}(f,\tau)|\} \ (\text{otherwise}), \end{cases} \tag{15}$$

where $m(l,\tau)$ is the output from the mel-scale filter bank. The system switches in two equations depending on the conditions in Eq. (15). $m(l,\tau)$ is a function of the over-subtraction parameter β and the parameter $\alpha(l)$ which is determined during a speech break so that the resultant output $m(l,\tau)$ is zero. On the other hand, if the power spectrum takes a negative value, $m(l,\tau)$ is obtained by using flooring processing, where γ is the flooring coefficient.

Since a common speech recognition is not so sensitive to phase information, SSA which is performing subtraction processing in the power domain is more applicable to the speech recognition. Moreover, in general, the order of the filter bank l is set to 24, and consequently SSA optimizes only 24 parameters. On the other hand, GJ requires the adaptive learning of FIR-filters of thousands or millions of taps. Finally, we perform mel-scale filter bank analysis, log transform and discrete cosine transform to obtain MFCC for speech recognizer.

3 Proposed Method

3.1 Error robustness analysis for noise estimation by NBF

In this section, we discuss the problem of the conventional SSA. The NBF-based noise estimator is used in the conventional SSA, but NBF suffers from the adverse effect of the microphone element error and the room reverberation. NBF is a technique to suppress an interference source signal by generating a null against the direction of the interference source signal. If the interference source signal arrives from the same direction as the null, we can suppress the interference source signal perfectly. In a reverberant environment, however, the interference source signal arrives from not only the null's direction but also outside of the direction. Therefore, in the reverberant room, we cannot suppress the interference source signal sufficiently. In addition, a microphone element usually involves gain and phase errors. NBF is designed under the ideal assumption that all elements have the same characteristics. In the real environment, however, the characteristics of each element are different. From the above-mentioned fact, the directivity pattern shaped by NBF in the ideal environment is apart from that of in the real environment.

Figure 2 illustrates directivity patterns which are shaped by two-element NBF in the ideal (solid line) and the real (dotted line) environment where the reverberation time is 200 ms. In this figure, the null direction is set to zero degree. We can see that the depth of the null in the real environment which contains the element error and the reverberation shallows. Therefore, we cannot suppress the interference source signal completely in the real environment by using NBF. Indeed, in SSA, we perform noise estimation via NBF which steers null against the target speech signal, but we cannot suppress the target speech signal sufficiently. In fact, NBF cannot

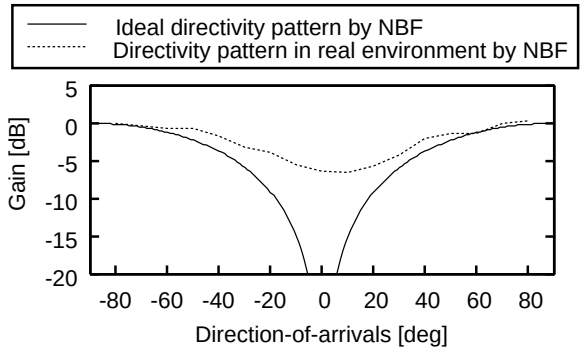


Figure 2: Directivity patterns shaped by NBF in ideal environment and real environment which contains element error and reverberation.

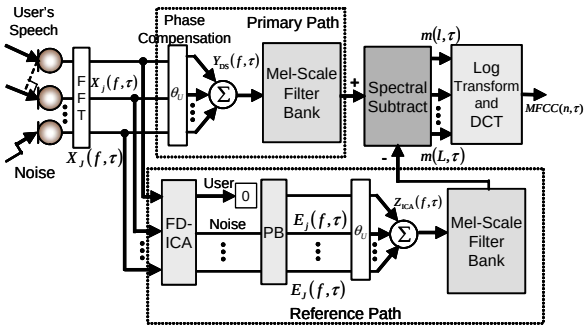


Figure 3: Block diagram of proposed method.

estimate noise signal completely. Thus the improvement of robustness in the noise estimator part is a problem demanding prompt attention.

3.2 Strategy of proposed method

We propose an improved SSA which includes ICA-based noise estimator instead of NBF-based noise estimator to address the problems which are discussed in the previous section. In the proposed method, the primary path and noise reduction processing are the same as the conventional SSA. As for the reference path, we newly introduce ICA as a robust noise estimator for adapting the filters to the element error and the reverberation (see Fig. 3). In ICA, an unmixing matrix is optimized so that output signals become mutually independent only using observed signals, and a priori information about the sensors and the room acoustics is not required. Therefore the proposed method can reduce these adverse effects because ICA can estimate noise signals which involve whole characteristics of the microphone elements and the reverberation. Detailed signal processing is shown below.

3.3 ICA-based noise estimation in reference path

The proposed method includes ICA-based noise estimation. In ICA part, we perform signal separation using the complex valued unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{O}(f,\tau) = [O_1(f,\tau),\ldots,O_J(f,\tau)]^{\mathrm{T}}$ become mutually independent; this procedure can be represented by

$$\boldsymbol{O}(f,\tau) = \boldsymbol{W}(f)\boldsymbol{X}(f,\tau), \tag{16}$$

$$\boldsymbol{W}(f) = \boldsymbol{P}(f)\boldsymbol{W}_{\mathrm{ICA}}(f), \tag{17}$$

where $\boldsymbol{P}(f)$ is a permutation matrix and $\boldsymbol{W}(f)$ is a new unmixing matrix which resolves the permutation problem. The permutation matrix $\boldsymbol{P}(f)$ is determined by looking at null directions in the directivity pattern which is shaped by $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [1], so that the U-th output $O_U(f,\tau)$ is set to the target speech signal. The optimal $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is obtained by the following iterative updating equation [7]:

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu\left[\boldsymbol{I} - \langle\boldsymbol{\Phi}\left(\boldsymbol{O}(f,\tau)\right)\boldsymbol{O}^{\mathrm{H}}(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \tag{18}$$

where μ is the step-size parameter, $[p]$ is used to express the value of the p-th step in the iterations, and $\boldsymbol{I}$ is an identity matrix. Besides, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes conjugate transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function [1]. In the reference path, the target signal is not required because we want to estimate only the noise component. Accordingly we remove the separated speech component $O_U(f,\tau)$ from ICA outputs $\boldsymbol{O}(f,\tau)$, and construct the following "noise-only vector, " $\boldsymbol{Q}(f,\tau)$;

$$\boldsymbol{Q}(f,\tau) = [O_1(f,\tau),...,O_{U-1}(f,\tau)\ ,0, O_{U+1}(f,\tau),...,O_J(f,\tau)]^{\mathrm{T}}. \tag{19}$$

Next, we apply the projection back (PB) [8] method to remove the ambiguity of amplitude. This procedure can be written as

$$\boldsymbol{E}(f,\tau) = \boldsymbol{W}^{+}(f)\boldsymbol{Q}(f,\tau). \tag{20}$$

Here, $\boldsymbol{Q}(f,\tau)$ is composed of only noise components. Therefore, $\boldsymbol{E}(f,\tau)$ is a good estimation of the received noise signals at the microphone positions;

$$\boldsymbol{E}(f,\tau) \simeq \boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \tag{21}$$

Finally, we obtain the estimated noise signal $Z_{\mathrm{ICA}}(f,\tau)$ by performing DS as follows:

$$Z_{\mathrm{ICA}}(f,\tau) = \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{E}(f,\tau) \simeq \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \tag{22}$$

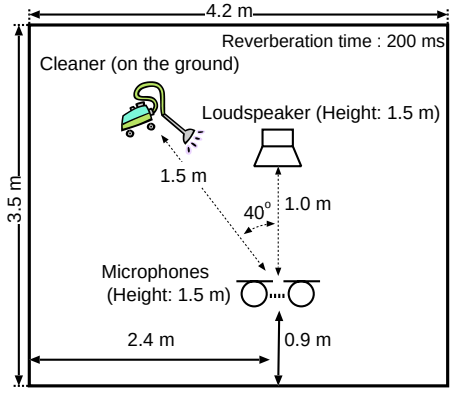


Figure 4: Layout of reverberant room used in our experiment.

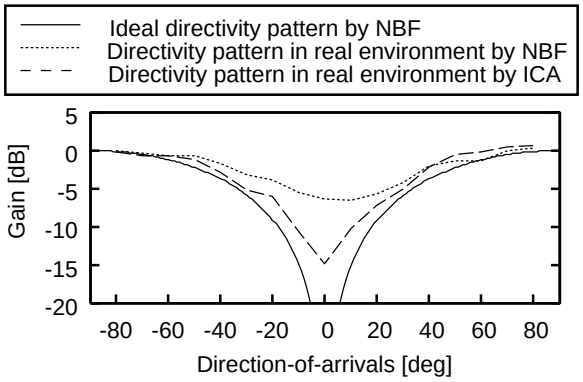


Figure 5: Directivity patterns shaped by NBF and ICA in ideal environment and real environment which contains element error and reverberation.

Equation (22) is expected to be equal to the noise term of Eq. (5) in the primary path. Of course, Eq. (22) contains estimation errors to some extent. Even though the level of the noise estimation error is not negligible, we can still enhance the target speech via over-subtraction [5] in the power spectrum domain.

4 Experiments And Results

4.1 Experimental setup

Figure 4 shows a layout of the reverberant room used in our experiments. We use the following 16 kHz sampled signals as test data; the original speech convoluted with the impulse responses recorded in the real environment, and added with a cleaner noise which was recored in the real environment. The cleaner noise is not a point source but consists of several non-stationary noises emitted from, e.g., a motor, air duct and nozzle. Moreover the cleaner noise includes background noise. The input signal-to-noise ratio (SNR) is set to 5, 10, or 15 dB at the array. A four-element array with the interelement spacing of 2 cm is used, and DFT size is 512. Over-subtraction parameter β is 1.4 and flooring coefficient γ is 0.2.

4.2 Accuracy of estimated noise signal

First, we analyze the directivity pattern shaped by ICA in the real environment. Figure 5 depicts the directivity pattern of ICA (broken line) in the real environment. From this result, we can confirm that the null shaped by ICA becomes deep compared with that of the NBF-based conventional SSA. Therefore, it is expected that the target speech suppression performance of ICA (equals the accuracy of the noise estimation) outperforms that of NBF. Next, we compare the conventional SSA and the proposed method in the accuracy of the estimated noise

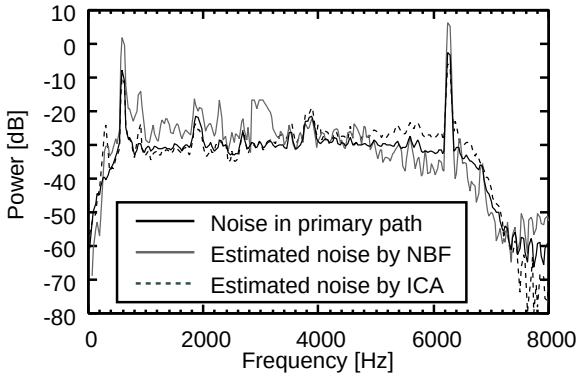


Figure 6: Accuracy of estimated noise signal by NBF and ICA.

signal. Figure 6 shows the long-term-averaged power spectra of the estimated noise signals by NBF and ICA. The black solid line indicates the power spectrum of the noise signal in the primary path, and this power spectrum is needed to be estimated. The gray solid line represents the power spectrum of the estimated noise signal by NBF, and the dotted line shows the power spectrum of the estimated noise signal by ICA. We can see that the power spectrum of the estimated noise signal by NBF is not accurate. This is due to that the target speech component still remains in the output of NBF because the null shaped by NBF is shallow. On the other hand, we can see that the power spectrum of the estimated noise signal by ICA is a good estimation because the depth of the null shaped by ICA is enough for suppressing the target speech. This result points out that ICA-based noise estimator is a more accurate noise estimator than NBF-based one. This gives propriety in which we use ICA as a noise estimator.

Table 1: Conditions for speech recognition

Database	JNAS [9], 306 speakers (150 sentences / 1 speaker)
Task	20 k newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [9], clean model
Number of training speakers for acoustic model	260 speakers (150 sentences / 1 speaker)
Decoder	JULIUS [9] ver 3.5.1

4.3 Results of speech recognition performance

We compare DS, the conventional SSA, and the proposed method on the basis of word accuracy scores. Table 1 describes the conditions for speech recognition, and we use 46 speakers (200 sentences) as original speech. Figure 7 shows the word accuracy in each method. Here, "Unprocessed" refers to the result without any noise reduction processing. From this result, we can see that the word accuracy of the proposed method is obviously superior to those of the conventional methods. This is a promising evidence that the proposed method has an applicability to noise-robust speech recognition rather than the conventional SSA.

5 Conclusions

In this paper, we proposed a new SSA which involves ICA-based noise estimation to realize a robust hands-free speech recognition in noisy environments. First, we pointed out NBF suffers from the adverse effect of the element error and the reverberation in the real environment. Secondly, based on the above-mentioned fact, we proposed a new SSA structure which replaces NBF-based noise estimator in the conventional SSA with ICA-based noise estimator. Finally, it was confirmed that the word accuracy of the proposed method overtook that of the conventional SSA in the experiment.

Acknowledgement

The work was partly supported by MEXT e-Society leading project.

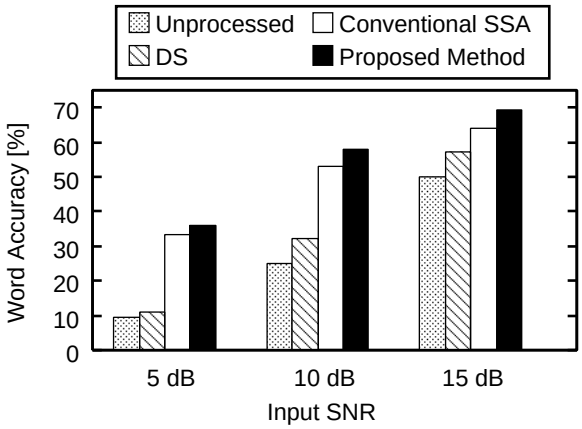


Figure 7: Results of word accuracy in each method.

References

[1] H. Saruwatari, et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Proc.*, vol.2003, no.11, pp.1135–1146, 2003.

[2] L. J. Griffith, and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol.30, no.1, pp.27–34, 1982.

[3] Y. Ohashi, et al., "Noise robust speech recognition based on spatial subtraction array," *Proc. NSIP*, pp.324–327, 2005.

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

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Proc*, vol.ASSP-27, no.2, pp.113–120, 1979.

[6] S. B. Davis, et al., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol.ASSP-28, no.4, pp.357–366, 1982.

[7] P. Smaragdis, "Blind separation of convoluted mixtures in the frequency domain," *Neurocomputing*, vol.22, pp.21–34, 1998.

[8] S. Ikeda and N. Murata, "A method of ICA in the frequency domain," *Proc. International Workshop on ICA and BSS*, pp.365–371, 1999.

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624-4 (11/17)

コミュニケーションロボットにおける音声認識システムの実環境での評価

Evaluation in real environments of a speech recognition system for communication robots

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
松田茂樹（ＮＩＣＴ／ＡＴＲ音声言語コミュニケーション研究所）
神田崇行（ＡＴＲ知能ロボティクス研究所）
實廣貴敏（ＡＴＲ知識科学研究所）
石黒浩（ＡＴＲ知能ロボティクス研究所）
中村哲（ＮＩＣＴ／ＡＴＲ音声言語コミュニケーション研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI (IRC Labs., ATR), Shigeki MATSUDA (NICT / SLC Labs., ATR), Takayuki KANDA (IRC Labs., ATR), Takatoshi JITSUHIRO (KSL Labs., ATR), Hiroshi ISHIGURO (IRC Labs., ATR), Satoshi NAKAMURA (NICT / SLC Labs., ATR), Norihiro HAGITA (IRC Labs., ATR)

carlos@atr.jp, shigeki.matsuda@atr.jp, kanda@atr.jp, takatoshi.jitsuhiro@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, satoshi.nakamura@atr.jp, hagita@atr.jp

Abstract - コミュニケーションロボットにおける重要な要素となる音声認識システムを実環境で評価した。音声認識システムは、フロント・エンドと認識エンジンに大きく分けられる。フロント・エンドでは、12 チャンネルのマイクロホンアレイを用い、RGSC処理により周辺の雑音を抑え、MMSE基準による特徴空間の雑音除去処理により音声区間が強調される。音声区間切り出しには、GMMによる自動切り出しを用いている。認識エンジンでは、大人と子供の音響モデルを作成し、パラレルデコーディングを行う。最終仮説は事後確率により選択する。認識結果は、GWPPにより信頼性の高いものに絞られる。評価実験では、短い文を発声した大人と子供の発話音声と、食堂の雑音を用い、各モジュールの性能を確かめた。70 dBAの雑音レベルにおいて、8割以上の単語正解率が得られた。

1 Introduction

Our research aims to develop “communication robots” that can naturally interact with humans and support everyday activities. Since the target audience of a communication robot is the general public who does not have specialized computing and engineering knowledge, a conversational interface using both verbal and non-verbal expressions is becoming more important. Previous studies in robotics have emphasized the merit of robot embodiments, showing the effectiveness of non-verbal information like facial expression [1], eye-gazing [2], and gestures [3].

Recently, several practical robots have been developed, such as therapy tools [4], museum orientation tool [5], and entertainments [6]. Moreover, robots are enlarging their working field in our daily lives. In one of our previous work, we tested a child-size interactive humanoid robot at an elementary school for several weeks [7]. The robot interacted with children by using speech and gestures in a free play situation. A similar project was run in a science museum where a humanoid robot interacted with visitors in a free-play situation and also conducted a museum tour, which contributed to visitors to grow interests in science and technologies [8].

One criticism to these two field trials was that these robots lacked speech recognition capability. The robots interacted with humans by speaking and making gestures, which are important elements for creating a sense of reality in humanoid robots. Language-based communication is indispensable, in order to fully utilize their human-like presence. However, one of the difficulties concerned speech recognition in noisy environments. Current technology has a good performance in recognizing formal utterances in noiseless environments, but the performance drastically degrades in noisy environments.

Several researchers are recently endeavoring to solve such problem so called “robot audition” [9]-[13]. Most of these works makes use of microphone array technology, for realizing sound source localization and separation, prior to speech recognition. However, the evaluation is usually done by controlling the direction of the noise or the interference.

Further, although most works evaluating speech recognition by robots have focused only on adult speech, these field trials (in both elementary school and science museum), indicated that children are important robot users, as well as adults. Thus, such communication robots should be able to deal with speech recognition of both adults’ and children’s speech. However, the performance of speech recognition also degrades due to differences on speaker age.

In this paper, we describe our ASR (automatic speech recognition) system, which accounts for these two

problems (caused by noisy environments and differences on speaker age), and evaluate it in a real noisy environment. Although we are conscious that a full communication could be reached by considering both linguistic and paralinguistic information included in the speech signal [14], in this paper, we focus only on the evaluation of the linguistic information processing.

The rest of the paper is organized as follows. In Section 2, we introduce our ASR system, and describe the techniques used in each module. In Section 3, we present the recognition performance results for several system structures, and for several noise conditions. We offer our conclusions in Section 4.

2 System Description

Accounting for the two problems (caused by noisy environments and differences on speaker age) described in Section I, we developed an ASR system to be robust to both background noise and speakers of different ages. Fig. 1 shows the overall structure of our ASR system. It consists of two major blocks.

The first block is a front-end processing. It contains a twelve-element microphone array, as depicted in Fig. 2. The real-time multichannel system for suppressing interference and noise and for attenuating reverberation consists of an outlier-robust generalized sidelobe canceller (RGSC) and a feature-space noise suppression (MMSE). The MMSE noise suppression is applied after RGSC to reduce the residual noise at the RGSC output. After that, the speech activity period detected by the GMM-based end-point detection (GMM-EPD) is transferred to the second block.

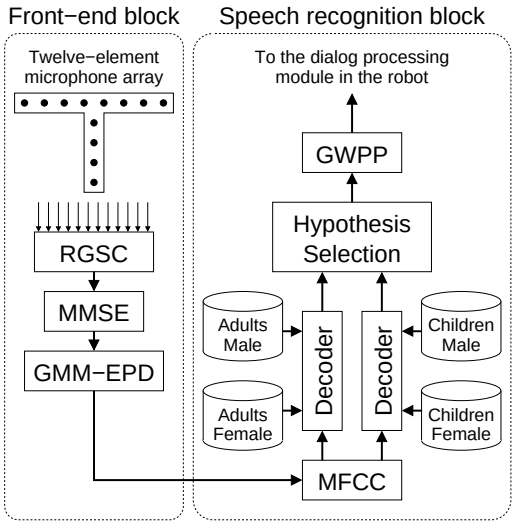


Fig. 1. The structure of the ASR system robust to noise and speakers of different age.

In the second block, there are two decoders depending on the age of the speaker (adult or child); each decoder works using gender-dependent acoustic models. The noise-suppressed speech at the first block is recognized using these two decoders, and one hypothesis is selected based on posterior probability. Finally, the hypothesis is measured using a generalized word posterior probability (GWPP)-based confidence measure. The hypothesis with confidence score higher than a threshold can then be transferred to a subsequent dialog processing module. The following sub-sections describe each module of our ASR system.

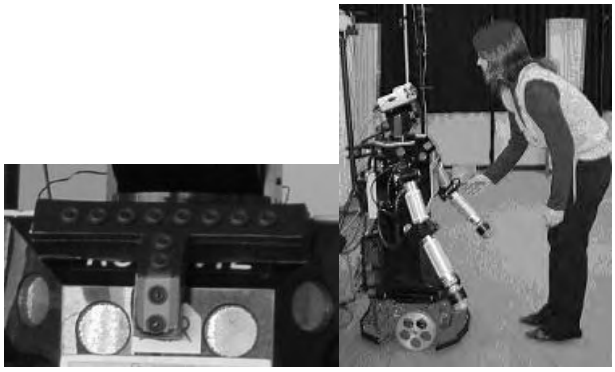


Fig. 2. Twelve-element microphone array in the Robovie's chest. Robovie wears the microphone array on its chest.

2.1 Twelve-element microphone array

In our system, we use a twelve-element microphone array for capturing speech. Omni-directional condenser microphones of type DPA 4060 are used for high-quality sound capture of distant-talking speech. The microphones are arranged in a T-shape with eight microphones on the horizontal axis and four microphones along vertical axis, with a spacing of 2-cm, as shown in Fig. 2.

We decided to arrange the microphone array in the robot chest, instead of the ear position or the head of the robot, for two reasons. One is the geometric limitations of the robot head, which would constraint the effective frequency range of the array processing. The other reason is that our robot makes rapid head movements, which would make difficult to set a target direction for the array processing.

Although the use of more microphones would provide a larger signal-to-noise ratio, we decided to limit to twelve, by considering hardware and real-time processing limitations.

2.2 Outlier-robust generalized sidelobe canceller (RGSC)

Many sound source separation algorithms have been proposed in order to reduce background noise coming from different directions. Here, we use an outlier-robust generalized sidelobe canceller (RGSC), proposed in [15]. The RGSC is applied to the audio signals captured using the twelve-element microphone array. The RGSC system is composed by a fixed beamformer, an adaptive blocking matrix, and an adaptive interference canceller.

The fixed beamformer steers the sensor array to the direction of the desired source and enhances the desired signal relative to the surrounding interference and noise. A simple uniformly weighted delay & sum beamformer is used. The fixed beamformer forms the reference path of the GSC. The blocking matrix is an adaptive spatial

filter which suppresses the desired signal and which passes interference and noise, such that the output of the blocking matrix is a reference for interference and noise. The adaptive interference canceller is realized by a multichannel adaptive filter between the output of the blocking matrix and the output of the fixed beamformer. The estimative of interference and noise is subtracted from the reference path at the output of the fixed beamformer so that the suppression of interference and noise is maximized.

The blocking matrix should be adapted when the signal-to-noise ratio (SNR) is high, while the interference canceller should be adapted when the SNR is low to prevent instability of the adaptive filters. A DFT bin-wise classifier for 'desired signal only', 'interference only' and 'double-talk' between the desired signal and interference or noise is then used for optimally tracking the desired signal and the interference. An "outlier-robust" adaptive filtering in the DFT domain for bin-wise adaptation control derived from [16] is then used to maximize the robustness against errors in the DFT bin-wise classifier.

2.3 Feature-space noise suppression using clean speech GMMs

The feature-space single-channel noise suppression is applied after the RGSC to reduce the residual noise at the RGSC output.

A GMM (Gaussian Mixture Model)-based MMSE (Minimum Mean Square Error) estimator is used to estimate a Wiener filter for suppressing background noise. The feature space is constituted by log Mel-spectral energy coefficients. For each frame i, one Wiener filter is obtained as a linear interpolation of multiple sub-Wiener filters, which are calculated using individual mixture component k of a clean speech GMM ($\boldsymbol{\mu}_{s,k}$, $\boldsymbol{\Sigma}_{s,k}$), and the observed noise $\mathbf{n}(i)$. The weights of the multiple sub-Wiener filters are optimized, based on the MMSE criteria, by maximizing the likelihood between te clean speech GMM and the input speech noise-suppressed by the Wiener filter. The filtered signal $\mathbf{g}(\mathbf{s}(i),\mathbf{n}(i))$ obtained in the log Mel-frequency domain is transformed back to the time domain for obtaining the impulse response $g(t)$. The clean speech is then estimated by convoluting the input noisy speech $y(t)$ with the impulse response $g(t)$. More details about the evaluation of the present noise suppression module can be found in [15].

2.4 GMM-Based End-Point Detection

An End-Point Detection (EPD) module is necessary for communication robots to properly interact with the user. In our ASR system, a GMM-based end-point detection (GMM-EPD) is used for detecting speech activity periods. This type of EPD architecture is widely used as a noise-robust EPD. First, we estimate the GMMs of noisy speech and noise in advance using a sufficient amount of training data.

During the detection of speech activity periods, we calculate the likelihood between each GMM and the input noisy speech. If the likelihood of the noisy speech GMM is higher than the noise GMM, the current frame is labeled as speech.

2.5 Hypothesis selection

For the speech recognition decoder engine, we use a hypothesis selection technique based on posterior probability [17] for improving robustness to speakers of different ages. One advantage of such approach is that it does not need to previously recognize the speaker age. Instead, the hypothesis with the highest score is selected from multiple hypotheses as follows:

$$\hat{k} = \arg\max_{k=1}^{K} H_k \tag{1}$$

$$H = \log P(\boldsymbol{X}|\boldsymbol{W}) + \lambda \log P(\boldsymbol{W}) \tag{2}$$

where H_k is the score of the hypothesis obtained from the k-th decoder and K denotes the number of decoders. The hypothesis obtained from the $\hat{k}$-th decoder has the highest score, which is defined as the sum of the log acoustic model likelihood $log\ P(\boldsymbol{X}|\boldsymbol{W})$ and the log language model probability $\log P(\boldsymbol{W})$ of a hypothesis. $\boldsymbol{X}$, $\boldsymbol{W}$, and H are the observed feature vector sequence, the hypothesis represented by a word sequence, and the score for the hypothesis, respectively. λ denotes a language model weight used for the hypothesis selection.

2.6 GWPP-based word confidence and rejection

So far, several techniques were described for improving the robustness of the ASR system to noise and to speakers of different ages. Nevertheless, the performance of an ASR system may degrade due to a mismatch between the training and testing channels, interference from environmental noise, etc. If the recognition results contain some fatal errors, this will adversely affect or prevent natural interaction between the robot and a human. Further, the system has to be able to reject utterances which are not included in the language model, in order to reduce insertion errors. To measure the reliability of the recognition results of the ASR system, we use a generalized word posterior probability (GWPP)-based confidence scoring [18].

In this method, the joint confidence of all component words in a recognition result is used to measure the confidence of a recognized utterance. The GWPP of a word is a measure of its correctness, or the probability of a binomial distributed "word correct" event. Thus, the probability of a "sentence correct" event is a product of all probabilities of component word correct events, assuming that all word events are statistically independent. A hypothesis with a probability of "sentence correct" event that is higher than a threshold is transferred as the final recognition result to the dialog processing module in the robot.

3 Experiments

3.1 Preparation of the modules

The acoustic models for adults were trained by using five hours dialogue speech from the ATR travel arrangement task database and 25 hours read speech of phonetically balanced sentences [19]. The training data was contaminated with eight types of noises listed in Table I at three types of SNR (20, 15, and 10dB), and the MMSE noise suppression was applied to the whole training data. A state-tying structure with 2,089 states was generated by using the MDL-SSS (Maximum Description Length Successive State Splitting) algorithm [20]. Each HMM state has five Gaussian distributions.

The feature vector consists of 12 MFCCs, Delta-pow, 12 Delta-MFCCs calculated with a 10-ms frame period and a 20-ms frame length. Cepstrum Mean Subtraction (CMS) was applied to the MFCC features, to reduce the effects of channel distortion.

The acoustic models for children were constructed by adaptation with an MLLR (Maximum Likelihood Linear Regression) algorithm using 12,000 words uttered by 238 child speakers in the CIAIR-VCV (Database of children's speech while playing video games) [21], which is provided from Nagoya university. The child training data was also contaminated with the same eight types of noises at 20, 15 and 10dB SNR.

For the MMSE noise suppression, we prepared a clean speech GMM with 512 Gaussian distributions using 24 log Mel-spectral energy coefficients. This clean speech GMM was trained with clean training data for adult speech only.

For the EPD module, 128-mixture GMMs were prepared using noisy speech and noise data. The feature vector is constituted by 12 MFCCs, Delta-pow and 12 Delta-MFCCs.

The language model is based on FSA (Finite State Automaton). The language model was constructed in order to recognize short Japanese sentences. This language model consists of 46 nodes and 205 links, and the lexicon size is 115.

The ATRASR speech recognition system developed by ATR Spoken Language Communication Laboratories was used in all experiments.

TABLE I
NOISE TYPES USED FOR TRAINING

Street	High-speed railway	Rice paddies
Airport lobby	Underground mall	Forest
Boiler room	Driving car	

3.2 Experimental conditions

To evaluate our ASR system's robustness to noise and to speakers of different ages, we tested it using a child and adult multichannel speech database that was recorded using the Robovie with the microphone array and a SANKEN CS-1, which is a directional microphone. This database consists of 1,464 short Japanese sentences uttered by 12 children and 12 adults. Each speaker uttered 61 sentences, such as "*Hello Robovie*", "*Come here*" and "*What can you do?*", in front of the Robovie. The children's ages ranged from 6 to 12. The distance between the speaker and the Robovie was 1 m.

We also recorded cafeteria noise using the same microphone array; this noise data was recorded at lunch time, therefore, it contains many types of noises such as talking voice, and clattery from dishes. The noise level was about 70 dBA. The cafeteria noise was recorded using the same volume level (amp gain) as when recording the speech database.

The speech data was contaminated by cafeteria noise at 65 dBA, 70 dBA, and 75 dBA.

3.3 Experimental results

To investigate the performance of the individual techniques described in Section II, we evaluated several ASR systems with different structures, listed in Table II.

3.2.1 *Evaluation of robustness to noise*

Fig. 6 shows the word accuracies of several ASR systems for adults' speech that was contaminated by cafeteria noise at 65 to 75 dBA. The performance of system D, which uses the twelve-element microphone array, the RGSC, and the MMSE noise suppression, was better than that of system A, B and C. And, system E, which has acoustic models trained with noise-contaminated adults' speech, achieved the best performance. System E reduced the errors by 85.5 %, 84.3 %, 80.5 %, and 48.3 %, in comparison to system A, B, C and D respectively. Clearly, performance is widely improved by applying all individual techniques described in Section II.

3.2.2 *Evaluation of robustness to speakers of different ages*

Regarding the evaluation of robustness to speakers of different ages, we evaluated the use of hypothesis selection using AMs of both adults' and children's speech (System G). It contained two decoders with acoustic models depending on the age of the speaker

TABLE II
EVALUATED STRUCTURES OF ASR SYSTEMS

System	A	B	C	D	E	F	G	H	I
Microphone type	1ch DPA	CS-1	array	array	array	array	array	array	array
Outlier-robust GSC	no	no	yes	yes	yes	yes	yes	yes	yes
MMSE noise suppression	no	no	no	yes	yes	yes	yes	yes	yes
Training data	clean	clean	clean	clean	noisy	noisy	noisy	noisy	noisy
AMs dependent on	adult	adult	adult	adult	adult	child	both	both	both
Segmentation	hand	hand	hand	hand	hand	hand	hand	auto	auto
GWPP-based rejection	no	no	no	no	no	no	no	no	yes

(adult or child). For comparison, we evaluated the recognition performance of a system which contains acoustic models for adults' speech only (System E) and for children's speech only (System F).

Fig. 7 shows the word accuracies for adults' and children's speech. We can see that a model depending on the age of speaker which is matched to that of input speech achieved the best performance. The system using hypothesis selection (System G) performed almost equally to the systems in matched case. A slight degradation for adults' speech occurred when using adults AM only, as shown in the left part of Fig. 7. However, the right part of Fig. 7 shows that the improvement for children speech is much more relevant.

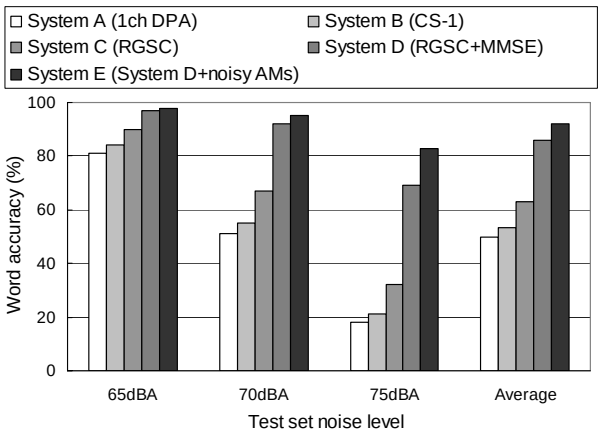


Fig. 6. Performance of system A to E for noise-contaminated adults' speech.

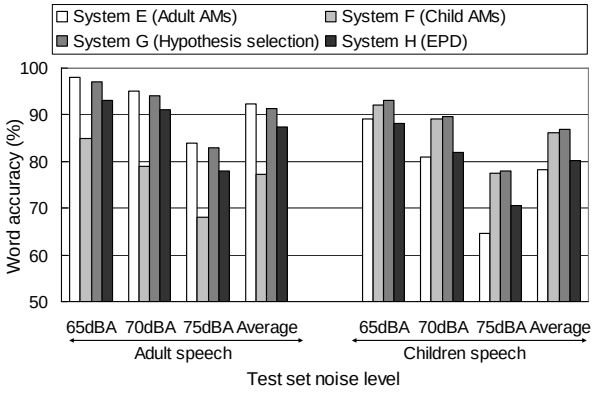


Fig. 7. Performance of system E to G for adults' and children's speech.

3.2.3 *Evaluation of the overall ASR system*

We tested the recognition performance of System H, which includes the GMM-based EPD module. As evident from Fig. 7 the GMM-based EPD module introduced some errors because of misdetections of speech activity periods. Table III shows the word accuracies calculated only for the speech detected by the EPD module. Values in brackets are the word rejection rates by the EPD module. As can be observed from these results, the word accuracies with the EPD module (System H) is almost the same as when manual segmentation is used (System G).

To improve the reliability of our ASR system, a GWPP-based confidence scoring module was implemented (System I). Table III shows the word accuracies obtained with the rejection module (EPD+GWPP). From these results, it is clear that word accuracies above 90 % were achieved at 70 dBA cafeteria noise. However a high rejection rate is also observed for high noise levels, indicating a tradeoff between word confidence and word rejection.

3.4 Evaluation of the overall ASR system in a real noisy environment

So far, the evaluation of the recognition system was realized by mixing clean speech and cafeteria noise, which were recorded separately. The purpose was to allow the control of different noise levels for evaluating the robustness of the system. In this section, we provide a more realistic evaluation, by recording speech in a real noisy environment.

Eight adult speakers (four males and four females) uttered the same 61 short Japanese sentences of the previous experiments, resulting in a database of 488 utterances.

The robot was placed in the cafeteria, in the same conditions (location and lunch time) used to record noise data in the previous experiment. Also, the distance between the speaker and the Robovie was about 1 m.

Recognition results indicated an average of 73 % word accuracy for all subjects. Word accuracies were between 70 to 84 % for seven subjects and 53 % for one of the subjects. Further analysis indicated that the SNR was about 10 dB for the seven subjects with higher scores and about 5 dB for the subject with the lowest score. The overall rejection rate (of correctly recognized words) was 8 %, while the overall rejection rate of insertions and incorrectly recognized words was 13 %.

A detailed analysis on the recognition errors revealed that some of the sentences, e.g. "utatteyo" ("sing!") and "tookudayo" ("it is far!"), showed low accuracies (less than 25 %). However it was observed that these errors were caused most due to rejection, rather than deletion or substitution. Also, monosyllabic sentences, like "hai" ("yes"), were found to be easier to be deleted, or misrecognized. In general, sentences composed by long lexicons were more reliably recognized.

3.5 Real-time issues

Finally, although most of the evaluations in this paper were realized in off-line, the system was verified to run in real-time as well, by using a remote PC with a Core 2 Duo Intel Xeon CPU at 3GHz and 1GB RAM. The audio data from the twelve-element microphone array was sent from the Robovie to the remote PC by TCP/IP network transmission.

The microphone array processing (RGSC module) is implemented by using Intel IPP (Integrated Performace Primitives). The use of Intel IPP allows real-time processing for the twelve-channel microphone array. The recognition engine (decoder) is the other critical part in terms of the processing time. However it is more

TABLE III
WORD ACCURACIES (%) OF SYSTEM H AND I. VALUES IN PARENTHESIS ARE WORD REJECTION RATES (%) BY THE EPD AND THE GWPP MODULES.

	Adult speech			Children speech		
	65 dBA	70 dBA	75 dBA	65 dBA	70 dBA	75 dBA
System H (EPD)	95.84 (2.28)	94.87 (4.06)	89.52 (14.19)	90.42 (3.52)	86.96 (6.63)	80.78 (15.91)
System I (EPD+GWPP)	96.33 (3.14)	96.58 (6.49)	92.05 (19.12)	91.42 (5.70)	91.04 (13.46)	88.57 (28.63)

difficult to guarantee real-time processing of the decoder module, since it depends on several factors, like the lexicon size, the language model complexity, the length of the detected speech segment, and the degree of noise (SNR). In our experiments, we observed that most of utterances were recognized within one second after the subject finishes uttering. However, sometimes the recognition results came after three to five seconds from the utterances. In these cases, a high noise level was observed, and the EPD module usually failed resulting in long segments containing long noise portions besides the real speech portion. The recognition results and processing time are thought to be improved by the EPD module.

4 Conclusions

In this paper, we described a robust ASR system for communication robots, and evaluated its robustness to real noisy environments and to speakers of different ages. In our ASR system, a twelve-element microphone array arranged in the robot chest, an RGSC-based microphone array processing, an MMSE-based feature-space noise suppression, and multi-conditionally trained acoustic models were used to improve robustness. Moreover, to improve the robustness to speakers of different ages, we used two decoders for children's and adults' speech respectively. Finally, the recognition results were scored using GWPP-based confidence measure, for reducing insertion errors. Experimental results in several noise level conditions indicated that our ASR system could achieve word accuracies of more than 80 % with 70 dBA of background cafeteria noise, for both adult and children speech. Further evaluation in a real noisy environment resulted in 73 % word accuracy for adult speech.

These recognition rates can still be increased by improving EPD (end-point detection) module. This topic is left for a future work.

Also as next steps of our work, a dialogue module will be developed for evaluating human-robot interaction in a real environment. We also intend to include a paralinguistic information extraction module [14], for also allowing a non-verbal communication between the robot and a human.

Acknowledgements

This work was partly supported by the Ministry of Internal Affairs and Communications.

References

1) C. Breazeal and B. Scassellati, A context-dependent attention system for a social robot, *Int. Joint Conf. on Artificial Intelligence (IJCAI'99)*, 1146-1151, 1999.
2) K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano: 'Real-Time Auditory and Visual Multiple-Object Tracking for Robots,' Proc. *IJCAI 2001*, 1425-1432, 2001.
3) O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Three-layered Draw-Attention Model for Humanoid Robots with Gestures and Verbal Cues," *IROS2005*, 2140-2145, 2005.
4) T. Shibata, "An overview of human interactive robots for psychological enrichment", Proceedings of *the IEEE*, Vol.92, No.11, 2004.
5) W. Burgard, et al., The interactive museum tour-guide robot, *National Conference on Artificial Intelligence*, 11-18, 1998.
6) M. Fujita, AIBO; towards the era of digital creatures, *Int. J. of Robotics Research*, Vol. 20, No. 10, 781-794, 2001.
7) T. Kanda, et al., Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial, *Human Computer Interaction*, Vol. 19, No. 1-2, 61-84, 2004.
8) M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, Interactive Humanoid Robots for a Science Museum, 1st Annual Conference on Human-Robot Interaction (*HRI2006*), 2006.
9) K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, Applying Scattering Theory to Robot Audition System, *IROS2003*, 1147-1152, 2003.
10) Asoh, H., Hayamizu, S., Hara, I., Motomura, Y., Akaho, S., and Matsui, T. "Socially Embedded Learning of the Office-Conversant Mobile Robot Jijo-2," *IJCAI'97*, 1997.
11) T. Takatani, S. Ukai, T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind sound scene decomposition for robot audition using SIMO-model-based ICA," *IROS2005*, 215-220, 2005.
12) Y. Ohashi, et al., "Noise-robust hands-free speech recognition based on spatial subtraction array and known noise superimposition," *IROS2005*, 533-537, 2005.
13) S. Yamamoto, et al, "Making a robot recognize three simultaneous sentences in real-time," *IROS2005*, 897-902, 2005.
14) C. T. Ishi, H. Ishiguro, N. Hagita: "Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information," *IROS2006*, 2006.
15) W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, "Hands-free speech recognition and communication on PDAs using microphone array technology," Proc. *ASRU2005*, 302-307, 2005.
16) W. Herbordt, et al., "Application of a double-talk resilient DFT-domain adaptive filter for bin-wise stepsize controls to adaptive beamforming," Proc.*IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 175.181, 2005.
17) S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," *IEICE Trans. Inf. & Syst.*, vol. E89-D, No. 3, 989--997, 2006.
18) F.K. Soong, W.K. Lo, and S. Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words," Proc. *SWIM2004*, 2004.
19) T. Takezawa, T. Morimoto, and Y. Sagisaka, "Speech and language databases for speech translation research in ATR," In Proc. *the 1st International Workshop on East-Asian Language Resources and Evaluation (EALREW 98)*, 148--155, 1998.
20) T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic Generation of Non-uniform HMM Topologies Based on the MDL Criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, 2121--2129, 2004.
21) Center for Integrated Acoustic Information Research, http: //db.ciair.coe.nagoya-u.ac.jp/eng/dbciair/dbciair2/kodomo.htm

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624-5 (11/17)

音声相互模倣過程を収束に導くマグネット効果

三浦勝司 [1)2)], 吉川雄一郎 [1)] and 浅田稔 [1)2)]

Katsushi MIURA[1)2)], Yuichiro Yoshikawa[1)] and Minoru ASADA[1)2)]

[1)] 科学技術振興機構 ERATO 浅田共創知能システムプロジェクト

[2)] 大阪大学大学院　工学研究科

[1)]**Asada Synergistic Intelligence Project, ERATO, JST**

[2)]**Graduate School of Eng., Osaka Univ., 2-1 Yamada-oka, Suita, Osaka 565-0871 Japan**

{miura,yoshikawa,asada}@jeap.org

Abstract

Human-robot communication is expected to be realized by the usual means for human-human communication. However, it is difficult for the robot to directly copy the human's means due to the difference of bodies. As argued in the issue of imitation with dissimilar bodies, what kinds of representation could correspond between human and robots is one of fundamental issues. The previous work [1] has hypothesized that mutual imitation of voice between the robot and the caregiver leads robot vowels to be more natural but the underlying mechanism has not been deeply argued. This paper focuses on two types of the magnet effects, the perceptual magnet effect and what we call *the articulatory magnet effect* as the underlying mechanism of mutual imitation. Toward the design principle of the robot behavior through mutual-imitation, we examine these magnet effects in the experiment of imitation of the vocalizing robot with human subjects.

1 はじめに

ヒューマノイドロボットは音声などのように，人が人とコミュニケーションをする際に用いる様式を用いて，人とのコミュニケーションを実現することが期待される．しかし，ロボットと人の身体構造や運動能力は異なるため，人のコミュニケーション行動をロボットがそのまま再現することは困難である．従ってロボットの行動は，それが人にどのように解釈されるかを考慮して構成されるべきである．一方，人の乳児は未発達のため，親の発声をそっくりそのままコピーすることはできないにもかかわらず，親とのインタラクションを通じて音韻様の発声を獲得することができるが，その発達メカニズムは明らかでない．従って，この乳児の発達メカニズムをモデル化 [2] することは，母音を発声できるロボットの実現だけでなく，人の乳児の言語獲得に至る認知発達過程の理解にも関連した非常に興味深い課題といえる．

インタラクションを通じた母音の獲得の従来研究として，複数の発話エージェントが知覚の自己組織化によって共通の母音を獲得するモデルが提案されている [3, 4]．しかし，これらの研究ではエージェント同士が同じ身体構造を持つことが仮定されており，乳児と母親のような身体構造が異なるエージェント同士がどのように母音を共有するかについては扱われていない．

身体構造の異なるエージェント同士が母音を共有する問題を扱った従来研究に，母親の模倣が乳児の発声を促し [5]，乳児の母音様の発声が母親の模倣を促す [6] という 2 つの知見に基づく母子間インタラクションモデルがある [7]．この研究では，発話ロボットの発声を教示者が母音でオウム返しすることで，ロボットが母音を獲得可能であることを示した．さらに，Miura et al. [1] は人が発声困難な音を模倣するとき，実際に模倣で返すべき音よりも無意識のうちに自身の母音よりの発声を返すとの仮説を基に，ロボットと人が互いの発声を模倣し合うことでロボットの発声を明瞭な母音に導くことができることを示した．しかし，人がなぜ無意識のうちに自身の母音よりの発声を返すのかについて十分に議論されていなかった．

人は自身の知覚する音を実際の音よりも自身の言語環境における特徴的な音素である母音や子音に似た音として知覚することが知られている．この現象は知覚のマグネット効果と呼ばれている．本論文では，相互模倣において人が無意識的に自身の母音よりの発声を返してしまう原因として，この知覚のマグネット効果に加え，我々が構音のマグネット効果と呼ぶ現象に注目する．これは人の発声が実際に出そうとした発声よりも自身の構音機構の制御や運動能力によって母音や子音に近い発声になる現

象である．

この構音のマグネット効果を示すための実験として，ロボットの発声を聞いたときに，人はどのように知覚し，模倣するかを検証する 2 種類の実験を行う．一つはロボットの発声を被験者が模倣する実験であり，もう一つはロボットの発声を日本語 5 母音のどれに聞こえたか被験者が判定する実験である．次節では相互模倣におけるマグネット効果についての仮説を説明する．そして，本研究で使用する発話ロボットについて紹介した後，実験および実験結果の解析について述べる．

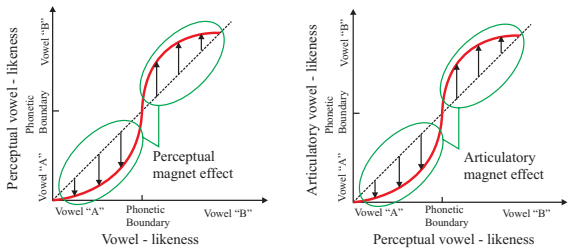
2　相互模倣におけるマグネット効果の仮説

発話の相互模倣を扱った先行研究 [1] では，人が模倣困難なロボットの発声を模倣しようとすると，無意識のうちに実際に発声すべき音よりも自身の母音に似た発声をしてしまうとの仮説が立てられている．そして，教示者と発話ロボットとの相互模倣を通じて発話ロボットに母音を獲得させることにより，仮説どおり教示者が母音よりの模倣をすることを示した．しかし，なぜ教示者が母音様の音声で模倣してしまうのかについての議論は尽くされていない．本研究では，この無意識的な母音様の模倣二間して知覚のマグネット効果と我々の提案する構音のマグネット効果の観点から考える．

知覚のマグネット効果とは，人の知覚する音が実際の音よりも自身の言語環境における特徴的な音素である母音や子音に似た音として知覚される現象である [8] (Figure 1 (a) 参照)．人はこの知覚のマグネット効果によって，他者の母音様の発声を実際よりも母音に近い音として知覚するため，模倣音声も実際より母音に近い音になる (Figure 1 (b) 参照)．一方，我々が構音のマグネット効果と呼ぶもうひとつのマグネット効果とは，人の発声する音が実際に知覚した音よりも自身の構音機構の制御や運動能力の拘束によって無意識のうちに自身の母音や子音に近い音になる現象である．従って，人は他者の発声を模倣するときに知覚と構音の 2 つのマグネット効果によって無意識的に自身の母音や子音に似た音を発声することになるため，相互模倣によってロボットの発声を人の母音に導くことが可能であると考えられる．

3　発話ロボット

本研究で用いる発話ロボット (Figure 2) は先行研究[7] [9] に習い，ソースフィルタ理論[10] に基づいて設計されている．この発話ロボットはエアコンプレッサ，人工声帯，声道，唇を備えており，母子間インタラクションのモデル化を目的として作成されている (Figure 3 参照)．コンプレッサから供給された空気はチューブを通して人工声帯を振動させ，音源になる．生成された音源は中空のシリコン製声道内を通過することで，声道形状に応じた共鳴周波数



(a) Perperception　　(b) Articulation

Figure 1: The shift of perceptual/articulatory vowel-likeness by the perceptual/articulatory magnet effect

を持つ音声として産出される．

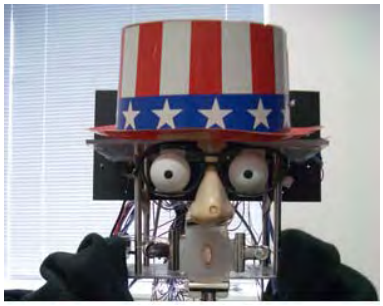


Figure 2: The vocalizing robot

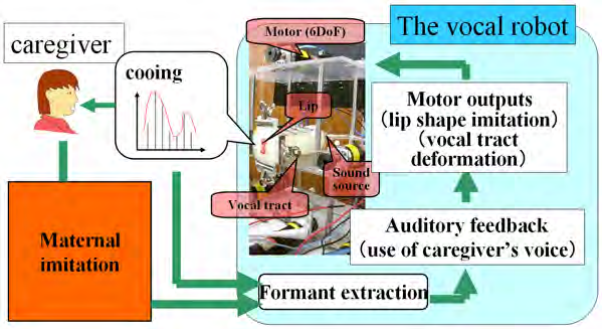


Figure 3: Vocalizing robot to model mother-infant interaction

ロボットは声道形状と唇形状をそれぞれ 4 つのモータを用いて変形させることにより，生成される音声の共鳴周波数を制御可能である．空気の流れはバルブの開閉でコントロールされており，モータとバルブのコントローラはホストコンピュータからの指令によって制御されている．ホストコンピュータはマイクから信号を受け取り，母音認識の特徴量でよく知られるフォルマント [11] を抽出する．

3.1 構音能力

声道形状の変形に用いる 4 つのモータの出力をそれぞれ 0.0(無変形) から 1.0(変形量最大) までの 5 段階に量子化し，唇の形状を人の母音 /a/，/u/，/e/ の発声時の唇の形状に似せた 3 種類に設定した．従って，ロボットが取りうる声道部と口唇部の形状は全部で 1875 通り (3×5^4) である．ここで /i/，/o/ の口唇形状を除外したのは，それぞれ/e/，/u/ の口唇形状との間にフォルマントの分布の差がなかったためである．

Figure 4 は横軸を第 1 フォルマント，縦軸を第 2 フォルマントとするフォルマント空間上にロボットが 1 秒間発声したフォルマントの平均値をプロットした結果である．以降フォルマント空間上の位置ベクトルをフォルマントベクトルと呼ぶ．また，人とロボットのフォルマントの分布を比較するため，日本人の男女 7 人が発声した日本語母音 /a/, /i/, /u/, /e/, /o/ のフォルマントの平均も Figure4 に示す．Figure 4 から，ロボットのフォルマントと人のフォルマントの分布は重なりあっていないことがわかる．すなわち，ロボットも人も互いの発声のフォルマントベクトルを再現することはできないことを示しているといえる．

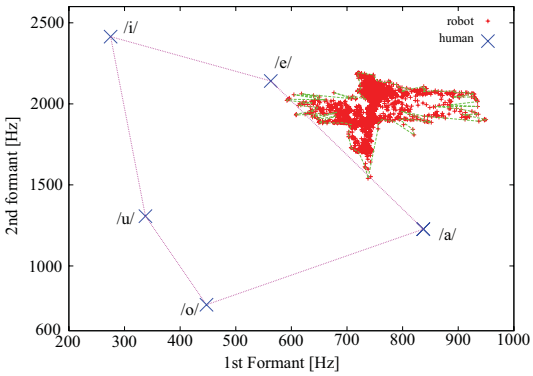
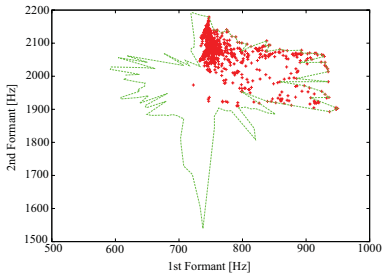
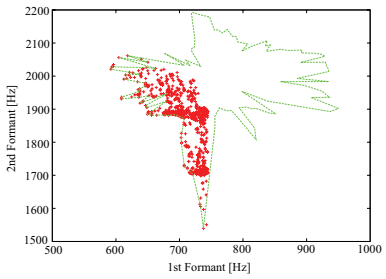


Figure 4: The distributions of the 1st and the 2nd formants of utterances by a human and the robot.

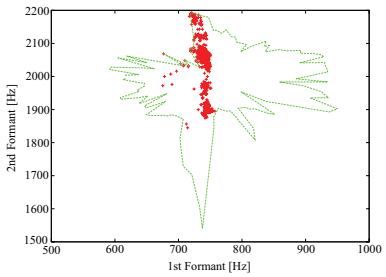
Figure 5 は母音 /a/, /u/, /e/ を発声するときの人の唇の形状 (Figure 5 (d),(e),(f)) とその母音に対応するロボットの唇の形状 (Figure 5 (g),(h),(i)) を示してある．Table 1 はそのときの唇のモータ出力である．さらに，その対応する唇の形状でロボットが発声したときのフォルマントの分布 (Figure 5 (a),(b),(c)) が示されているおり，口唇形状の違いによって，ロボットが発声するフォルマントの分布領域は異なることがわかる．この分布位置は人の母音のフォルマントの相対的位置関係 (Figure4 参照) に似ており，ロボットに人の口唇形状を模倣させることによって，口唇形状に対応した母音をロボットに発声させやすくなると考えられる．



(a) Formant distribution for the lip shape /a/



(b) Formant distribution for the lip shape /u/



(c) Formant distribution for the lip shape /e/

(d) /a/ (e) /u/ (f) /e/

(g) /a/ (h) /u/ (i) /e/

Figure 5: Formant distribution of the utterances, the lip shape of the human model and the mapped shape onto the robot lip for three vowels, /a/, /u/, /e/

Table 1: Relation between robot's lip shape and motor outputs

Deformation	/a/	/u/	/e/
vertical direction	1.0	0.0	0.0
horizontal direction	1.0	0.0	1.0

4 実験

マグネット効果が人の知覚や模倣にどのような影響を与えるかを検証するために 2 種類の実験を行った．一つ目の実験では，目隠しをした被験者にロボットの発声を模倣させることで，知覚と構音の 2 つマグネット効果の影響について調べる．もうひとつの実験では，目隠しをした被験者にロボットの発声が日本語の 5 母音のうちのどれに聞こえたか判定させ，そのときのロボットの発声の母音らしさを評価させることで，知覚のマグネット効果の影響について調べる．実験の被験者には大学院在学中の 10 人の被験者を選び，実験の試行順は被験者ごとに入れ替えた．この 2 つの実験の結果を比較することで，相互模倣インタラクションにおけるマグネット効果の影響について議論する．

4.1 刺激

２章で議論したマグネット効果から，模倣者が普段発声する母音にどの程度近い音声で模倣するかは呈示される音声の母音らしさの程度のシグモイド関数で近似可能であることが予想される (Figure 1 参照)．そこで，母音らしさの程度が異なる様々なロボットの発声を被験者に模倣させ，これらの関係を観察することを考える．

ただし，ロボットが発声可能な音の中で，どのようなフォルマントベクトルを持つ音が最も母音らしいかは不明であるため，最も母音らしい音に対応するフォルマントベクトル $\boldsymbol{r}^{/v/}$ を以下のように操作的に定義する．すなわち，日本人の男女 7 人が発声した日本語母音のフォルマントの平均 $\boldsymbol{h}^{/v/}$ とその日本語の 5 母音のフォルマント空間上での重心 $\boldsymbol{h}_c$ を用いて $\boldsymbol{r}^{/v/} = (\boldsymbol{h}^{/v/} - \boldsymbol{h}_c) + \boldsymbol{r}_c$ のように与える (Figure 6 参照)．ここで $\boldsymbol{r}_c$ はロボットが構音可能なフォルマントの分布 (Figure 4 参照) の重心である．

そして，各母音についてロボットが発声する母音らしさを変えた 5 つのフォルマントベクトル $\boldsymbol{r}_i^{/v/} (i = 1, \cdots, 5)$ をロボットが構音可能なフォルマントの分布の重心 $\boldsymbol{r}_c$ と最も母音らしい発声 $\boldsymbol{r}^{/v/}$ を用いて次式のように定めた．

$$\boldsymbol{r}_i^{/v/} = \boldsymbol{r}_c + \frac{i}{5}\alpha(\boldsymbol{r}^{/v/} - \boldsymbol{r}_c), \quad (i = 1, \cdots, 5), \qquad (1)$$

ここで α は $\boldsymbol{r}_i^{/v/}$ がロボットの構音可能な領域内に収まるようにするためのスケーリング係数である．これにより，実験では，各母音につき 5 通り，合計 15 通りのロボットの発声が模倣の対象として被験者に提示される．

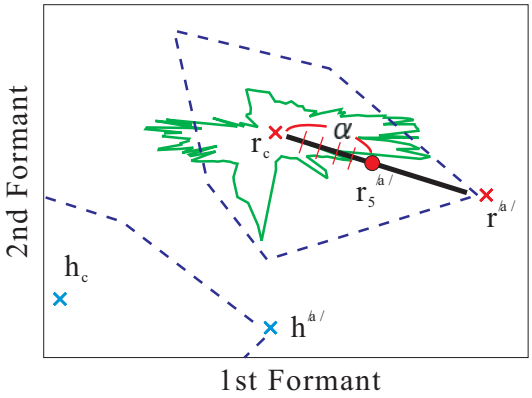


Figure 6: The formant vectors of the most "vowel-like" sound and the test sounds in the case of a vowel /a/. Note that this figure is schematic.

ロボットには実験の各試行ごとに 2 つの音を選択させる．ただし，一つ目の音に $\boldsymbol{r}_i^{/v/}$ が選ばれた場合，連続して発声される次の音は $\boldsymbol{r}_i^{/v'/}, (/v'/ \neq /v/)$ となるように選択させる．従って，発声される連続音の組み合わせの場合の数は，母音の選び方 ${}_3P_2$ 通りに母音らしさの選び方 ${}_5P_1$ 通りで合計 30 通りとなる．それぞれの音の組み合わせはランダムで 1 度だけ被験者に聞かされる．

発声する音が決まった後，ロボットはあらかじめ準備しておいたフォルマントと唇や声道形状を決定するモータ出力とのマッピングを利用して発声する．

4.2 実験手順

音声模倣実験　被験者に"ロボットの発声に対して第 3 者が同じ音だと知覚できるように模倣してください"と説明した後，被験者にロボットの 2 つの母音の連続発声を聞かせ，それぞれの音を模倣させた．これを 1 試行とし全部で 30 試行繰り返した．また，模倣実験の開始前に被験者に日本語の 5 母音を発声させ，フォルマントベクトルを抽出した．この抽出した日本語の 5 母音のフォルマントと被験者が模倣発声したフォルマントベクトルを比較することで，模倣時の知覚のマグネット効果と構音のマグネット効果を調べた．

母音判別実験　被験者に目隠しした状態でロボットの 2 つの母音の連続発声を聞かせ，日本語の 5 母音のうちのどれに聞こえたかと，そのときの自身の判断に対する自信の度合いを 5 段階で評価させた．そのときの評価値は '1' (適当に選んだ), '3' (なんとなくそう聞こえた), '5' (確信を持ってその母音だといえる) であり，'2'，'4' はそれぞれの中間の値である．これを 1 試行とし 30 試行繰り返した．そして，ロボットが発声した母音と被験者が答えた母音とが一致したときの評価値によって母音判別時の知覚のマグネット効果を調べた．

4.3 結果

まず初めに，我々が設定したロボットの母音らしさが適切であったかを検討する．設定した母音らしさが適切であるならば，ロボットの母音らしさが増すに連れ，被験者の模倣音声が母音に近付いていかねばならない．

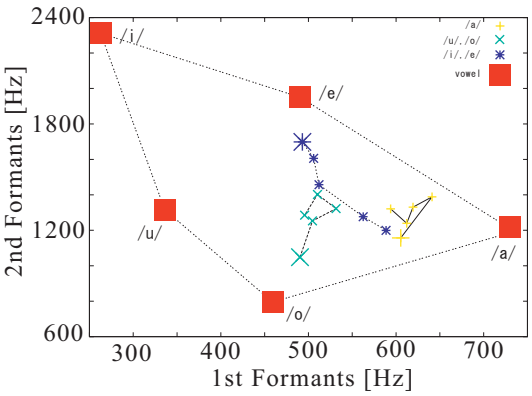


Figure 7: The average of formant vectors of subjects imitation and the average formant vectors.

Figure 7 はロボットの発声 $\boldsymbol{r}_i^{/v/}$ (/v/ = {/a/, /u/, /e/}，$i = 1, \cdots, 5$) に対して被験者が模倣したときのフォルマントである．ただし，＋ は /v/ = /a/，× は /v/ = /u/，* は /v/ = /e/ を示しており，$i = 5$ のみ大きなプロットとして $i = 1, \cdots, 5$ までを線で結んである．また，■を頂点とする五角形は被験者 10 人の母音のフォルマントベクトルの平均である．Figure 7 より，ロボットの母音らしさ i の増加に合わせて被験者の模倣がロボットの発声と同じ母音に近付いているのは，ロボットが /e/ を発声したときのみであり，ロボットの母音 /a/，/u/ に設定した母音らしさは不適切であったと考えられる．従って，ロボットの母音らしさの適切であると考えられる /e/ の結果のみについて考察する．

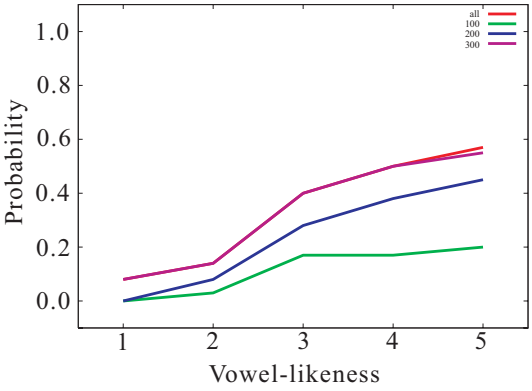


Figure 8: Probabilities in which the subjects replied with their vowels of which lip shapes were corresponded to the robot's one. Note that vowel-lbikeness corresponds to i in equation (1).

本実験で使用した発話ロボットでは構音可能領域が狭いため，スケーリング係数 α により $\boldsymbol{r}_i^{/i/}$，$\boldsymbol{r}_i^{/o/}$ がそれぞれ $\boldsymbol{r}_i^{/e/}$，$\boldsymbol{r}_i^{/u/}$ とほぼ同じ値となった．このようにフォルマントベクトルがほぼ同じである場合，人は母音の違いを判別できないと考えられるため，ロボットの発声 /u/，/e/ に対応する被験者の発声はそれぞれ /u/，/o/ と /i/，/e/ であるとする．

Figure 8 はロボットの発声に対して被験者が /i/ または/e/ の母音で模倣した確率がロボットの発声の母音らしさによってどのように変化するかを示している．この確率は人が感じる音の差を低周波，高周波に限らず一定の間隔であらわせるように周波数を変換した値である MEL を閾値とし，各被験者の平均値で表される．Figure 8 は，模倣音声と母音との差が MEL 空間上で 100 以内，200 以内，300 以内，制限なしの 4 つを閾値として用いた結果である．

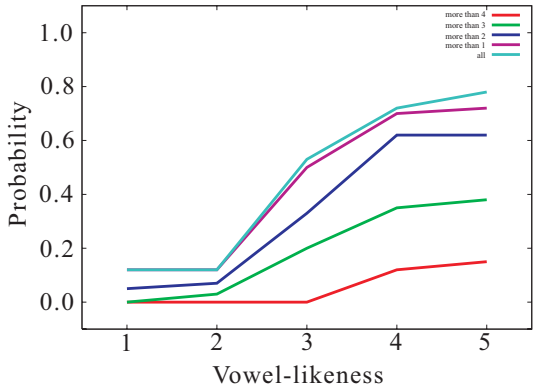


Figure 9: Probabilities in which the subjects confidentially categorized the heard sound of which lip shapes were corresponded to the robot's one. Note that vowel-likeness corresponds to i in equation (1).

一方，Figure 9 はロボットの発声に対して被験者が /i/ または/e/の母音であると答えた確率がロボットの発声の母音らしさによってどのように変化するかを示している．この確率は被験者の評価点を閾値とし，各被験者の平均値で表される．Figure 9 は評価値が 5 点以上，4 点以上，3 点以上，2 点以上，1 点以上の 5 種類の結果であ

ある母音らしさ i でのロボットの発声 4 回に対し，被験者が母音 /i/ または /e/ と答えたときの回数を被験者 10 人で平均化したものである．また，各線は評価値による閾値が 5 点以上，4 点以上，3 点以上，2 点以上，1 点以上の 5 つを閾値として用いた結果である．

発声の必要がない母音の判別では知覚のマグネット効果のみが働くと考えられる．一方，模倣する場合にはロボットの発声を聞くときに知覚のマグネット効果が，さらにその知覚した音を発声する際に構音のマグネット効果が働くと考えられる．そこで，Figure 8，9 を比較することで構音のマグネット効果が模倣時にどのような働きを

しているか考察する．Figure 8，9 はどちらもシグモイド関数に似たデータの変化を示していることがわかる．これは，ロボットの発声が母音様に近付くと，マグネット効果によって急激に母音であると知覚，模倣しやすくなることをあらわしている．

ここで，いつマグネット効果が現れ始めたのかを調べるため，各データの縦軸の最大値のと最小値の中間となる地点を通過するときの母音らしさを計算することでマグネット効果の現れるタイミングをあらわした．ただし，各母音らしさは線形で補間し，被験者による差や個人の応答のばらきをの影響を防ぐため，それぞれの実験の各条件における平均値で計算した．結果，母音らしさが模倣時は各条件での平均値で 2.65，判別のときは各条件での平均値で 3.04 のときに中間点を通過した．これは模倣時のほうがマグネット効果が早くあらわれると考えられることを示している．つまり，人はロボットの発声を模倣するときに，知覚のマグネット効果だけでなく構音のマグネット効果が加わることで，より母音に近い音を発声すると考えられる．

5 結論

本論文では，前研究で仮定した相互模倣が母音に収束するメカニズムの原因となる知覚のマグネット効果と，構音のマグネット効果の 2 つのマグネット効果について議論した．被験者がロボットの発声を模倣，またはどの母音であるか判別する実験より，知覚のマグネット効果だけではなく構音のマグネット効果によって被験者が母音様の発声で模倣することを示した．

参考文献

[1] Katsushi Miura, Minoru Asada, Koh Hosoda, and Yuichiro Yoshikawa. Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction. In *The 5th International Conference on Development and Learning (ICDL'06)*, 2006.

[2] Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro, and Yasuo Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous System*, 37:185–193, 2001.

[3] B. de Boer. Self-organization in vowel systems. *Journal of Phonetics*, 28:441–465, 2000.

[4] P.-Y. Oudeyer. Phonemic coding might result from sensory-motor coupling dynamics. In *Proceedings of the 7th international conference on simulation of adaptive behavior (SAB02)*, pages 406–416, 2002.

[5] M. Peláez-Nogueras, J. L. Gewirtz, and M. M. Markham. Infant vocalizations are confitioned both by maternal imitation and motherese speech. *Infant behavior and development*, 19:670, 1996.

[6] N. Masataka and K. Bloom. Accoustic properties that determine adult's preference for 3-month-old infant vocalization. *Infant Behavior and Development*, 17:461–464, 1994.

[7] Yuichiro Yoshikawa, Minoru Asada, Koh Hosoda, and Junpei Koga. A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, 15(4):245–258, Dec 2003.

[8] Patricia K. Kuhl. *Plasticity of development*, chapter 5 Perception, cognition, and the ontogenetic and phylogenetic emergence of human speech., pages 73–106. MIT Press, 1991.

[9] T. Higashimoto and H. Sawada. Speech production by a mechanical model construction of a vocal tract and its control by neural network. In *Proc. of the 2002 IEEE Intl. Conf. on Robotics & Automation*, pages 3858–3863, 2002.

[10] Philip Rubin and Eric Vatikiotis-Bateson. *Animal Acoustic Communication*, chapter 8 Measuring and modeling speech production. Springer-Verlag, 1998.

[11] R. K. Potter and J. C. Steinberg. Toward the specification of speech. *Journal of the Acoustical Society of America*, 22:807–820, 1950.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624-6 (11/17)

音声の構造的表象を通して考察する幼児の音声模倣と言語獲得

Consideration on infants' speech mimicking and their language acquisition based on the structural representation of speech

峯松信明 †，西村多寿子 ‡，櫻庭京子 *
Nobuaki Minematsu†, Tazuko Nishimura‡, Kyoko Sakuraba*
† 東京大学大学院新領域創成科学研究科 / Graduate School of Frontier Sciences, The University of Tokyo
‡ 東京大学大学院医学系研究科 / Graduate School of Medicine, The University of Tokyo
* 清瀬市障害者福祉センター / Kiyose-shi Welfare Center for the Handicapped
mine@gavo.t.u-tokyo.ac.jp, nt-tazuko@ams.odn.ne.jp, sakuraba@mtd.biglobe.ne.jp

Abstract

In speech communication, acoustic distortions are inevitably involved by speakers, channels, and hearers. However, infants acquire a spoken language mainly with speech samples of their mothers and fathers. They can solve the variability problem only with a remarkably biased speech corpus. Why and how is it possible? To answer this hard question, we already proposed a speaker-invariant structural representation of speech. In this report, the proposed representation is mathematically shown to be invariant also with non-linear transformations. Based on this representation, the speech recognition processes of dyslexics and autistics, often viewed as paradox, could be taken for granted. Finally, we discuss that speech communication should be based on relative sense of sounds.

1 はじめに

音声コミュニケーションには，話者・環境・聴取者に起因する，多様な音響歪みが不可避的に混入する。その一方で幼児は，大部分が「母親と父親の音声」という音声資料の提示を通して音声言語を獲得する。これは，音響的に非常に偏った話者性の音声資料の提示を通して，多様な音響歪みに関する対処法を獲得することを意味する。偏った音声資料の提示は，その後一生続く。何故ならば，人の聞く声の半分は自分の声だからである。人は偏った音声提示しか受けられないのである。何故，このように音響的に偏った音声提示環境の下で，人（幼児を含む）は多様な音響歪みに対処できるのだろうか？音響音声学／音声工学では，この多様性問題を直接的に解くことはせず，個々の音素の音響モデルを，数千・数万という話者の音声を集め，分布としてモデル化することでその解決を図ってきた。それでも多様性問題は解けず，音響モデル適応／特徴量正規化などの技術を編み出して来た。全く異なる戦略を示す両者の，本質的差異はどこにあるのだろうか？

話者 A の音声を書き起こす。話者 B の音声を書き起こす。この時，話者 A によって発せられたある音響事象を「あ」という記号で表記し，話者 B によって発せられたある音響事象も同様に「あ」という記号で表記する。当然，両音響事象間に物理的等価性は保証されない。物理的に異なる音響事象群を，同一の表記を用いて書き起こす訳である。なぜ話者毎に「あ」という記号の変種を用意することなく,「あ」と表記できるのだろうか？

提示された曲を，階名を用いて「ドレミ」として書き起こす。曲が階名として聞こえてくる聴取者は，その曲を移調しても，書き起こされる「ドレミ」列は変わらない。相対音感者である[1]。移調によっていくら「ド」の音高が変わろうとも，彼らは「ド」として表記する。第一著者は絶対音感を持っており，この階名での書き起こしが全くもって理解できない一人である。異なる音高に同一の音ラベルを振ることなど，全く理解不能である。

異なる話者間で「あ」の同一性が感覚できない人がいるのだろうか？音の絶対特性に執着し，両者の同一性が感覚できない人がいるのだろうか？感覚できない「機械」が，限られた話者の音声から構築された（特定話者）音声認識器である。そして，感覚できない「人」として，一部の自閉症者がいる[1]。優れた絶対音感を持つ率が，健常者と比較して遥かに高い自閉症者の中には,「ハ」の音（固定ドとしてのドレミの場合は「ド」の音）で始まらない「カエルの歌」を，それと認めない者もいる[2]。

相対音感者による階名による書き起こしは，音階の構造（全全半全全全半という音高遷移の枠組み）をメロディーの中に感覚し，例えば長調の場合，主音をドとして認識し，同様にして，上主音，中音，下属音を，レミファ，として認識する。即ち，音列の流れを通して，全体的なメロディー構造（音楽学では，これを「横の構造」と言う）

[1] なお，階名の書き起こしが出来ない（ハミングしかできない）相対音感者もいる。この場合，彼らは「言語化が困難な」相対音感者である。

の認知が先に起こり，それに基づいて個々の要素音の（他音群との関係によって定まる）機能的・相対的価値を認識する訳である。その結果，要素音の絶対的物理特性とは全く独立に，個々の要素音が同定されることになる。物理的には全く異なる二音が同一の機能的価値を有した時，両者が同一音として「聞こえる」ことになる[3, 4]。

幼児は，極端に限られた音声の多様性に接することで，広範囲に渡る音声の多様性に対処できるようになる。発達心理学によれば，「幼児の音声言語獲得は，分節音の獲得の前に語全体の音形・語ゲシュタルトの獲得から始まる」とされている[5, 6]。個々の音韻意識が定着するのは小学校入学以降であり，それまでは「しりとり」に難を示す児童もいる[7]。即ち幼児の音声コミュニケーション（例えば音声生成）は，個々の音韻（モーラ）を一つ一つ音に変換する形では（少なくとも意識の上では）行なうことは困難である。日本語には母音が 5 つあり，それが/あ/い/う/え/お/であることを知る以前に，幼児は両親と音声コミュニケーションを行ない，自己主張までする。

全体的なメロディー構造を通して，移調された曲同士の同一性を感覚し，更には，個々の要素音の（階名としての）同一性を感覚する。その結果，物理的に全く異なる二音を同一であると感覚する。幼児の言語獲得も同様に，語全体の音形の獲得から開始される，とした場合に，この語全体の音響表象が，音楽同様，音声の移調（非言語要因による不可避的な音響変動）による多様性問題を解く鍵になるのだろうか?従来より筆者らは，非言語的要因による音響変動に対して簡素な数学モデルを考え，この問題を解決してきた[8, 9, 10, 11, 12, 13]。本稿ではこれを一般化し，非常に広範囲な変換（非線形変換を含む）においても，移調不変な構造的表象が普遍的に存在することを数学的に示す。そして幼児の音声模倣，更には，自閉症・失読症者の音声認知を，この構造的表象を通して考察する。最後に，自閉症者のビヘービアと音声認識システムのそれとが非常に類似していることを指摘すると共に，人間の音声情報処理と機械の上に実装された情報処理との本質的・根源的な差異について，音情報処理の生物進化に伴う変遷を踏まえ，筆者らの意見を述べる。

2 非言語的音響変動不変の音声の構造的表象

2.1 2 つの空間における頑健な不変量

図 1 に示す様な，二つの空間 A と B を考える。両者には一対一の対応関係があり，空間 A のある点は空間 B の対応点へ写像され，逆もまた成立する。但し，その写像関数は明示的には与えられていないとする。以下，一般性を失わない範囲で 2 次元空間を用いて説明する。空間 A，B の対応する二点を (x, y)，(u, v) とし，両空間の対応付け（変数変換）を一般的に下記の様に示す。

$$x = x(u, v), \quad y = y(u, v) \tag{1}$$

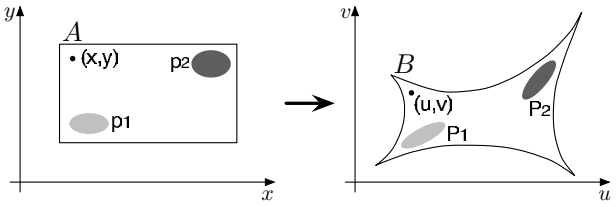


Figure 1: 一対一対応関係を有する二つの空間 A と B

空間 A における事象を考える。但し，全ての事象は空間内の点ではなく，確率密度分布関数として存在するものとする。即ち事象 p は次式を満たす。

$$1.0 = \iint p(x, y)dxdy \tag{2}$$

空間 A における積分演算は，変数変換によって空間 B における演算へと変換可能である。

$$\iint f(x, y)dxdy = \iint f(x(u, v), y(u, v))|J(u, v)|dudv \tag{3}$$

$$= \iint g(u, v)|J(u, v)|dudv \tag{4}$$

$g(u, v) \equiv f(x(u, v), y(u, v))$ であり，$J(u, v)$ はヤコビアンである。分布関数も同様に空間 A から B へ写像される。

$$1.0 = \iint p(x, y)dxdy \tag{5}$$

$$= \iint p(x(u, v), y(u, v))|J(u, v)|dudv \tag{6}$$

$$= \iint q(u, v)|J(u, v)|dudv \tag{7}$$

$$= \iint P(u, v)dudv \qquad p() \text{ in A} \rightarrow P() \text{ in B} \tag{8}$$

$q(u, v) \equiv p(x(u, v), y(u, v))$，$P(u, v) \equiv q(u, v)|J(u, v)|$ であり，変数変換後にヤコビアンを掛けることで写像される。

以上の道具を用いて，空間 A と B の間に存在する不変量について考察する。空間 A における二つの分布，p_1 と p_2，を考える。これらを空間 B へ写像して得られる分布を P_1，P_2 とすると，当然 p_i と P_i の絶対的特性は異なる。p_1 と p_2 に対するバタチャリヤ距離は下記式で表記される。

$$BD(p_1, p_2) = -\ln \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy \tag{9}$$

これは，下記の様に空間 B における積分演算へ変換される。

$$BD(p_1, p_2) \tag{10}$$

$$= -\ln \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy \tag{11}$$

$$= -\ln \iint \sqrt{q_1(u, v)q_2(u, v)}|J(u, v)|dudv \tag{12}$$

$$= -\ln \iint \sqrt{q_1(u, v)|J|}\sqrt{q_2(u, v)|J|}dudv \tag{13}$$

$$= -\ln \iint \sqrt{P_1(u, v)P_2(u, v)}dudv \tag{14}$$

$$= BD(P_1, P_2) \tag{15}$$

即ち，空間 A におけるバタチャリヤ距離は，空間 B における対応する二分布間のバタチャリヤ距離と等しくなる。この性質は，**式 (1)** の空間 A，B の対応付けに対して，強い制約を求めない。ヤコビアンによる変数変換が可能であれば，上記性質は満たされるため，一対一対応空間に対して付加的に要求される制約は，1)$x(u,v)$，$y(u,v)$ が偏微分可能で，導関数が連続，2) 空間 B の積分領域においてヤコビアン J が非零，のみとなる。結局，これらの条件を満たす，非線形変換を含む，広い変換群に対して，バタチャリヤ距離は不変となる。この変換不変性は，カルバックライブラ距離，ヘリンジャ距離などでも成立する一般的性質である。以上，各事象が分布として存在し，かつ，その推定が正確に行なわれれば，二分布間距離が非常に頑健な変換不変量として存在することを示した。この時，両空間の写像関数やヤコビアンを求める必要は無い。

2.2 不変事象間距離から普遍的に存在する不変構造へ

三辺の長さを規定すれば，三角形の形状は一意に定まる。同様に，ユークリッド空間に存在する n 点からなる幾何学構造は，${}_nC_2$ 個だけ存在する二点間距離を全て求めれば，(鏡像の曖昧性を無視すれば) その構造を一意に規定することになる。即ち，距離行列は幾何学構造を規定することになる。距離行列による構造定義は，タンパク質の構造解析など，広く用いられている方法である。距離行列と幾何学構造を等価であると考えるならば，空間に存在する N 個の分布群によって張られる距離行列，即ち，幾何学構造は，前節で数学的に導出した様に，一切変換不変となる。そして空間 A，B を二人の話者の音声音響空間 A，B とすれば，両者の間において不変構造が存在することになる。これはどの二話者でも成立するため，結局，話者非依存の普遍性を持つ（音響的普遍構造）。

この数学的性質は，非常に強力な枠組みとなると考えられる。従来筆者らは，英語学習者における英語母音群構造に対して，構造解析を行なってきた[12, 13]。話者／マイクなどの不可避的な音響歪みを除去し，外国語訛のみを構造歪みとして抽出することが目的であった。しかし，本来の構造不変性は，幅広い変換群で成立するため，学習者空間と教師空間との間に一対一対応があれば，外国語訛を超えて，構造の不変性・同一性を約束することになる。しかしこの場合，空間 A でのガウス性分布が空間 B では非ガウス性の分布として変換されるなど，分布形状の極端な歪みを生じることが予想される。例えば，変換後もガウス分布となるという制約の下で本数学的性質を使う等，積極的かつ妥当な制約導入によって，本性質は有効利用されると考える。逆に言えば，正確な分布の推定が可能であれば，それほど頑健な不変構造が，数学的には普遍的に存在する，ということである。本稿ではこの普遍的な不変構造の存在を基に，音声認知に関する種々の考察を行なう。

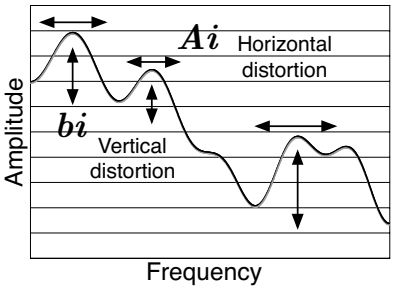


Figure 2: スペクトルの水平・垂直歪みと一次変換

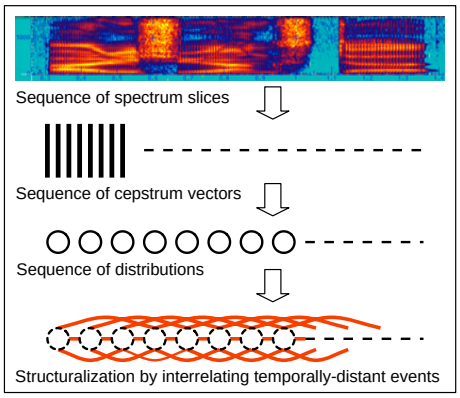


Figure 3: 事象間差のみを抽出して構成される不変構造

3 音声の構造的表象に対する実験的検討

従来より筆者らは，非言語的要因による音響変動をケプストラムの一次変換としてモデル化して議論してきた[8, 9]。これは，特にスペクトルの水平方向歪みが $c' = Ac$ として，垂直方向の歪みが $c' = c + b$ として記述できることによる（**図 2** 参照）。この場合，ガウス分布はガウス分布へと変換されることになる。これに対して**図 3** に示す様に，絶対項を全て捨象し，音事象間差異のみを求めることで，話者／マイク不変の構造的表象が得られる。実際に，孤立発声された 5 母音系列[2]をタスクとした音声認識では（語彙数 120 の孤立単語認識に相当），LPF などの前処理が必要ではあったが，一人の話者の音声で不特定話者音声認識が可能であることを示した[10, 11]。この実験では，4,130 人の話者の音声から構築された HMM よりも高い頑健性を示した。話者性を消去する，という方法論は，発音学習支援にも応用されている。特定の学習者と特定の教師の発音を，体格／性別／年齢といった違いを無視した形で，直接的・構造的に比較することが可能となっており，種々の興味深い実験結果が得られている[12, 13]。発音ポートフォーリオの提案，効率的学習のための教示生成，更には学習者分類などについて検討している。

図 3 に示す構造化による不変項の導出は「音声の非言語的特徴は時不変である」という仮定の上で成立する。即ち，話者性が時変であれば，不変項は導出されない。HMM 音声合成技術を用いて，時間的に話者性が変化する合成音

[2]連続発声を対象とした分布列推定方法がまだ確立できていないため，孤立発声母音系列という，人工的なタスクを用いた。

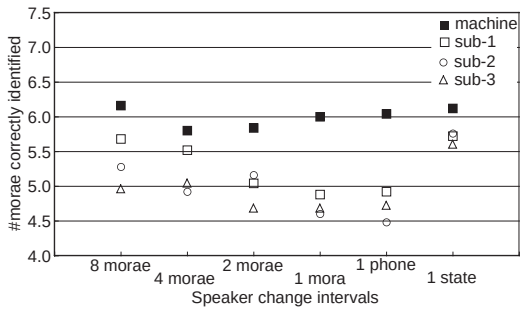


Figure 4: 非音声研究従事者を対象としたモーラ同定率

声を聴取させると，話者性変化頻度の向上によって無意味モーラ列音声のモーラ同定率が低下する様子が観測されている（**図 4**）[14]。なお，話者変化頻度を極端に上げた場合，HMM 音声合成の内部処理である時間方向のスペクトル平滑化によって話者性変化は消失されるとの予測が成立するが，実験結果も，その予測の妥当性を示している。

4　幼児は親の声の何を模倣しているのか？

幼児は親の声の模倣を通して音声言語を獲得すると言われている（音声模倣）[6]。しかし幼児は，親の声そのものを模倣しようとはしていない。太くて低い声を出そうと努力している幼児はいない。音韻意識が未熟である彼らは，個々のモーラ（話者非依存の音シンボル）を一つずつ生成する，という術は少なくとも意識的には不可能である。となると彼らは親の声の何を模倣しているのだろうか？九官鳥による音声模倣では，話者性までも真似ることが知られている[15]。優秀な九官鳥は，その音声模倣を聞いただけで飼い主が分かるが，どんなに優秀な幼児の音声模倣を聞いても，飼い主（親）の同定は不可能である。音響的な音声模倣と言語的な音声模倣は何が違うのだろうか？

「幼児の音声言語獲得は，分節音の獲得の前に語全体の音形・語ゲシュタルトの獲得から始まる」との主張が正しければ[3]，この語ゲシュタルトの音響的実体には，非言語的情報は含まれないはずである。もし含まれていれば，幼児は父親の声が出せるよう，日々努力するはずである。筆者らはこの語ゲシュタルトの音響的定義について，多くの発達心理学・言語獲得研究者に問いかけてみたが[16]，残念なことに，明確な答えは得られなかった。

近年の脳科学の進歩により，聴覚音声学の議論は，蝸牛から，聴覚皮質のモデリングに移行しつつある[17]。脳科学における多くの知見は偶然によって齎されている[18, 19]。交通事故や，一部の医師の不適切な処置が原因で不幸にも脳損傷を負った患者を通して多くの知見が得られている。動物実験でも同様，偶然的な刺激提示によって重要な知見が得られている。前頭葉，海馬，扁桃体の機能，更には，ミラーニューロンなどは良い例である。脳は研究者の机上の議論を超えた処理を行なっている，と解釈することもできる[19]。さて，聴覚皮質モデリングであるが，視覚皮質のような定説が存在する状況には無いが，幾つか興味深い主張がある。まず「音声の言語的情報と非言語的情報（話者の情報）は分離されて処理されている」との主張である[17, 20, 21]。特に [21] では，音楽と音声とを対比し，音声の言語情報は，音声の動きの情報（speech motions）によって伝搬されると主張している。音楽で言えばメロディーである。一方「話者の同定は音楽で言う楽器の同定に相当し，それは時不変の情報として処理される」と主張している。**図 3** に示した音声の構造的表象は，音声を「音の運動」と考え，その運動（コントラスト）成分のみを抽出する形となっている。即ち音声から「音であること」を一切捨て去った物理表象である。何が動いているのかは不明である。「動きだけを抽出した時に，話者／年齢／性別を超えた頑健な不変表現が数学的に入手できる。それこそ言語である」と主張するのが音響的普遍構造である。

近代言語学の祖ソシュールが一世紀以上も前に興味深い主張をしている[22]。The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from all others. 即ち，音ではなく，音的差異の重要性を説いている。差異を捉えることで単語が同定できる，との主張である。彼はまた Language is a system of conceptual differences and phonic differences. と主張している。「言語＝差異・動きのシステム」である。分布間差異を集めたものが頑健な不変構造を成し，それを用いた語同定が可能である。この不変構造こそ語ゲシュタルトではないだろうか。

音声が伝搬する情報は，言語／パラ言語／非言語情報と分類される。各情報を担う音響量に着眼すると，言語及び非言語情報は声道情報となるため，スペクトル包絡に相当し，パラ言語情報は音源情報となるため，F_0，パワー，継続長に相当する。即ちソース・フィルタの分離である。幼児の音声模倣は，親の声からまず非言語情報を分離すると考えられる。音声＝［言語＋パラ言語］＋［非言語］という枠組みである。しかし，音声科学・工学が構築した枠組みは，音声＝［言語＋非言語］＋［パラ言語］という枠組みである。調音音声学の価値観に基づけば，声道と音源を分離する，自然かつ妥当な枠組みである。しかし，音声コミュニケーションの観点から考えると，この枠組みでは幼児の音声模倣問題は解けない。非言語情報を頑健に分離する術が無いからである。人と音声との遭遇は聴取であって，生成ではない。しかし，科学は音声と生成を通して遭遇した，というのは歴史的事実である[23]。音声科学が実験科学である以上，それは時代の技術的制約の下で議論を重ねなければならない。聴覚音声学は観測技術の未熟さから，調音・音響音声学と比較して，その進展

[3] なお，日本人乳児が [r] と [l] を弁別できることが広く知られているが，これは 2 音の弁別ができるのであって，[r] を /r/ として同定している訳では無い。同定能力の獲得の前に，まず，弁別・区別，即ち差異の知覚が可能になることは重要である。

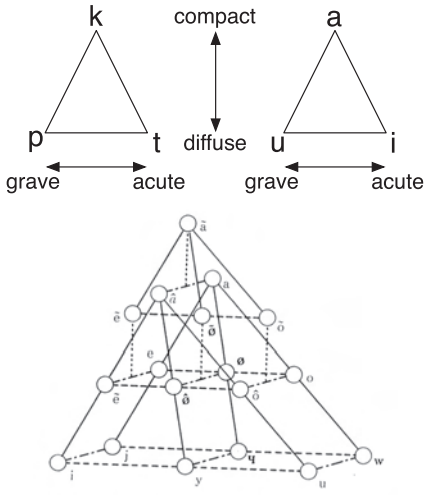


Figure 5: 母音・子音三角形[24] と仏語母音群構造[25]

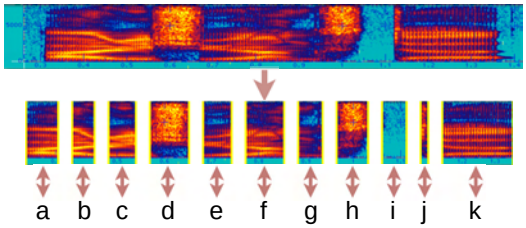


Figure 6: 音声ストリームの細分化と絶対音感的要素同定

が遅れざるを得なかった。脳科学の進展によって「話されたもの」としてではなく「耳に届くもの」として音声に焦点が当たった時に，従来の枠組みでは想像できない情報処理を脳が行なっていたとしても何ら不思議ではない。

ソシュールの phonic differences という言葉はやがて，ヤコブソンの弁別素性，即ち構造音韻論へと引き継がれる。音そのものではなく，音/x/と音/y/はどう違うのか，その違いを定性的に表現するために弁別素性が使われた（**図 5** 参照）。つまり，音素群が成す外部構造の議論である[25]。やがて，弁別素性はそれが束となって音素を表象する，即ち音素の内部構造の議論に使われるようになる[24]。この素性の束としての音素は音楽の「和音」をメタファーとして生まれた[26]。和音＝音素，音符＝素性，である。音楽学では，音楽には横の構造（メロディー構造）と縦の構造（ハーモニー構造）があると説く。和音は縦の構造であり，ヤコブソンの素性による内部構造の議論は音楽の縦構造を発端としている。一方，筆者らが提唱する音声の構造的表象は当然のことながら，音楽の横構造に相当する。縦と横の構造，どちらが音楽にとってより本質的かと言えば，当然，横構造である。和音の無い音楽はあれど，メロディーの無い音楽は存在しない。

筆者らの知る限り，音声工学において横構造の議論が皆無であり，縦構造の議論[27, 28, 29] が多い理由は，音声の表記方法に起因すると考える。音声を（話者非依存の）音シンボル列として表記し，音声をシンボルに対応させて区切れば，その時点で横構造は消失する。例えば Bloomfield は，

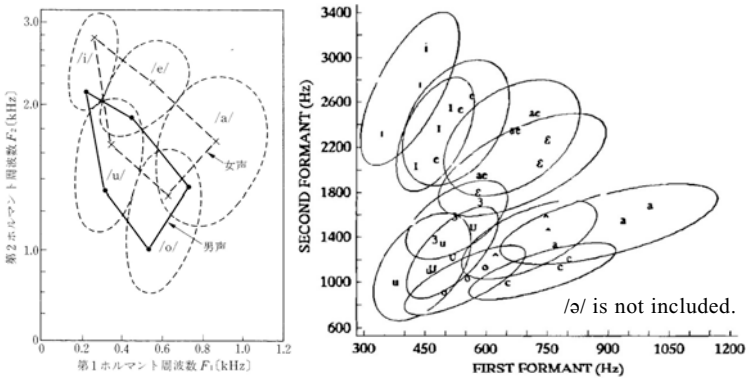


Figure 7: 日本語 5 母音と米語 12 母音の F_1/F_2 図[31, 32]

Writing is not language, but merely a way of recording language by visible marks. と述べているが，文字言語は本来音声言語の副産物であり，5 万年以上の音声活動を通して人類が造り出した「音声言語の視覚化技術」でしかない[4]。即ち「結果」であって「原因」ではない。しかし「結果」を「前提・原因」として捉え，**図 6** に示す様に，音声の物理現象を切り刻んで来た，という歴史は否めない。この一次元的音声視覚化技術は物理的に正しいのだろうか？そもそも科学は要素還元主義の上に構築され，その枠組みの限界が指摘されたのはごく最近である[30]。

5　音韻の意識は音声言語運用に必要なのか？

提案している音声表象は音声ストリームをメロディとして捉え，その横構造を頑健な不変項として導出している。孤立発声母音の系列という人工的タスクではあるが，個々の音事象の絶対的物理特性は一切用いずに，個々の音事象の同定は一切行なわずに，単語の同定が可能であることを示した[10]。音楽の場合，横構造を通して各音事象の機能的・相対的価値を感覚し，「ドレミ」が聞こえてくる[5]。音声の音韻知覚も同様の枠組みとして捉えられないだろうか？全体を通して要素同定する，音声の相対音感である。

音楽の場合，如何なる鍵盤（即ち物理特性）も「ド」になれる。逆にある鍵盤は「ドレミ…」のいずれにもなれる。しかし音声の場合，例えば F_1/F_2 図において任意の点の音を「あ」と知覚できるか，と考えれば，それは困難である。**図 7** に示す様に日本語の場合，男女を考慮しても 5 母音は凡そ分離している。音韻とその物理実体との対応が凡そ一対一，即ち，絶対音感的である。しかしフォルマント周波数が発声者の声道長に依存していることを考えれば，例えば目玉親父やウルトラマンが日本語母音を発声したとすると，これらの分布群は大きな重なりを呈することになる。一対一対応の崩壊である。このような場合，音の絶対量に基盤を置く処理系は機能せず，音と音の相対量に基盤を置いてこそ，頑健な処理系が期待できる。

[4] 文字起源は象形文字であるため，本来文字は意味の視覚化技術であり，音の視覚化技術ではない。表音文字は単なる借り物技術でしかない。
[5] 人が沢山いるそうである，としか第一著者は言えない。

アニメの世界では音声は相対音感的でなければならない。

アニメの世界を想像しなくても，相対音感の世界を創成することは容易である。母音の数を増やせばよい。**図 7** には米語 12 母音の F_1/F_2 図についても示している。成人男性・女性・子供（10〜12 歳）139 名の/h V d/から得られた結果である[32]。なお，音質が容易に変動する/ə/はこの図には含まれていない。これだけの重なりは，複数話者のデータを同時に表示するから生じるのであり，話者別に示せば当然重ならない。この事実を顧みずに，母音毎に，複数話者データに対して物理的な絶対量を統計的にモデル化しても，母音認識は困難となる。日本語は絶対音感的，米語は相対音感的なのだろうか？何れの場合も，個々の音はシンボル化される。音楽の場合でもドレミは階名としても（移動ド），音名としても（固定ド）使われている。

相対音感者の多くは，言語化できない相対音感者である。メロディーの記述を，音名／階名で行なうのではなく，「ラ〜ラ」即ちハミングで行なう相対音感者である。彼らも主音は認知しており[3]，音楽の横構造を認識しているが，主音に対して「ド」を対応させることが困難である。そもそも音高（基本周波数）と「ドレミ」という声（スペクトル包絡）とは無関係であり，これを恣意的に結びつけたのが階名である[6]。音声に対して「言語化できない相対音感者」とはどのような存在になるのだろうか？他者が歌った歌を「ラ〜ラ」として再生する際に頻繁に移調されることを考えれば，ある話者の発声を移調して再生することは，「繰り返し発声」に相当する。一方，曲を「ドレミ」に落とす作業はどうなるであろうか？スペクトル特性とは全く関係の無い，「声」に対して恣意的に関連付けられた「モノ」を考えれば明らかなように，それは「（表音）文字」に落とす作業となる。以上の考察から得られる帰結は「相対音感的な音認知が不可避的に要求される言語の場合，文字の読み書きに困難を覚える人が多い」となるが，こんな考察，意味があるのだろうか？

第一著者は，このような無意味かもしれない考察の最中に失読症（dyslxeia）を知った。「頭が良いのに，何故か本が読めない」方々である[33, 34, 35]。具体的な症状は様々であるが，共通項として存在する症状が音韻意識が希薄，即ち，単語音声に対して，それを個々の音に分割したり，個々の音が連結して単語音声になる，ということを感覚することが困難な方々である。**図 6** の枠組みを理解することが困難な方々である。幼児の音声認知をそのまま引きずっており[33]，個々の音をカテゴリとして知覚するのが困難である一方で，異音の区別は健常者よりも成績が良い[35]。これは [r] を/r/というカテゴリとして同定できないが，[r] と [l] が区別できることに相当する。米国では程度の差こそあれ，約 20%の人が失読症である[33]。政治家，作家，起業家，学者にも失読症者はおり，グラハム・ベルもその一人である。彼に音声認識・合成器を作らせても，音シンボル列と音声間の変換技術など作ら（れ）なかったはずである。「そんなモノの上に言語は出来ていない」と主張したであろう。幼児の言語獲得は，彼らの認知能力の未熟さが，個々の音韻を意識させないのだろうか？無意識下では音韻を操作しているのだろうか？或いは，個々の音韻意識は音声言語運用に不要なのだろうか？**図 6** の「音声 ↔ 音シンボル列変換」に難を示す多数の音声ユーザの存在を，音声科学・工学者はどう考えるべきなのだろうか？

言語化できる相対音感者が時として犯す勘違いとして，次のようなものがある。全ての長調の曲が「ハ長調」として聞こえる，というものである。凡そ全ての曲は主音で終了する。即ち，階名で「ドレミ」が聞こえて来る相対音感者は，曲の終わりは全て「ドー」と聞こえる[7]。常に「ドー」と聞こえるから，その箇所では同一の鍵盤を押している，即ち，同一の物理音が出ている，と解釈した訳である。機能的・相対的等価性が物理的等価性を上書きし，異なる音群に対して同一物理音を認知させた訳である。そのような相対音感者に対しては，絶対音感者が「それは物理的には錯覚，勘違いの一種である」と説明する。機能的・相対的等価性を物理的等価性と入れ違えた結末であると説明する。音声科学・工学では，話者 A の音声中のある音が音韻「あ」と感覚され，話者 B のある音も音韻「あ」と感覚された場合，**図 6** に示した音声ストリームの細分化を行ない，両話者の該当区間の物理現象に何らかの絶対的同一性を期待する。その二音の物理的相違は明らかであるにも拘らず，数千，数万人の話者から「あ」と感覚される音声区間を集め，統計的にモデル化する。音韻とは心的表象である[36]。心的表象とは物理実体が存在しないことを意味する。よってその心的表象は，物理的にはある種の「錯覚，勘違い」の産物ということになる[37]。音響音声学，音声工学が大前提とする**図 6** の枠組みは，物理的に妥当なのだろうか？音楽の相対音感者の勘違いは音楽の絶対音感者が是正してくれた様に，**図 6** の枠組みが研究者の単なる勘違いであるとするならば，音声の絶対音感者が，彼らを是正してくれるのだろうか？

6　究極の音声絶対音感者と音声言語

極端な絶対音感を持つ奏者は，オーケストラ／ホールが変わる度に十分な耳慣らしが必要となる。基準音がオーケストラ／ホールによって，数 Hz 異なるからである。参照パターンとして絶対項を持ってしまうと（例えば，基準音＝ 440Hz），環境の変化に対して柔軟に対応できなくなる。音声の極端な絶対音感者は話者 A の「おはよう」と話者 B の「おはよう」の同一性の認知が困難になると考えられるが，自閉症者の一部に，特定話者の音声のみ言語メッセージになる者がいる[1]。自閉症は端的に「関係の

[6]「ドレミ」という命名は僧侶の名前の第一音節から来ている。

[7]らしい。くどいようであるが，第一著者には皆目見当がつかない。

病」と言われるように[38]，入力される情報の整理整頓が困難であり，個々の要素的事象を丹念に記憶する。日付，曜日，電話番号，住所など互いに無関係なものを膨大に記憶する一方で，物事の因果関係や複数の刺激群が成すパターンの抽出，事象の抽象化に困難を示す。そのため，目の錯覚などが起き難い。マガーク効果が起き難い。顔の要素的特徴を覚える一方で，顔を見て表情や話者を同定することが苦手である。優れた音感を持ち，絶対音感者が多い。一言で言えば，ゲシュタルト知覚が困難である[39]。

第一著者にとって「ドレミ」とは音名であるため「曲が階名として聞こえる」という事実は想像を絶する。「ソ」が「ド」と聞こえる，というのは「え」が「あ」と聞こえる，というのに等しい。勘違いか錯覚の類いではないか，とさえ考えることもある[8]。極端な音声の絶対音感を持つと考えられる自閉症者にとって，物理的に異なる特性を持つ話者 A の音と話者 B の音を「同一音」として認知する健常者の感覚こそ，想像を絶するものであると推測する。彼らが「勘違いか錯覚の類いではないか」と主張しても不思議ではない。異なる二音を「あ」と感覚できる健常者の認知能力が，音の絶対項に基づくものなのか，あるいは，音間の相対項に基づくものなのか，彼らこそ，その回答をもたらしてくれるもの，と期待されるが，残念なことに，彼らの多くは口を開かない。何故なら，極端な絶対音感を持つ自閉症者は，音声言語を持たないからである。二話者の「おはよう」の同一性が認知できなければ，音声言語が破綻するのは自明である。音声言語は，ある種の錯覚・勘違いの上に成立する，と考察することもできる。音声言語を持たない自閉症者の中には，ごく稀に，文字言語を通して言語コミュニケーションを開始する場合がある[1, 40]。音は全て聞こえているにも拘らず，聞こえ過ぎるが故に，文字（視覚図形）言語が第一言語となる。自閉症者は，常に変化する環境を頑健に対処する術を持ち合わせていないと言われる。文字は変わらない。しかし，音声はいつも変わる。だから図形言語が第一言語となる。確かに，人，場所，時，あらゆる要因が音声の絶対項を変える。しかし着目する時間長において，その要因が時不変であれば，構造は一切不変である。変わることが許されない。

音響空間を [音素数]^3 の部分空間に分け，各々の独立性を仮定して各空間における観測量を絶対的にモデル化し，保持するのが現在の音響モデリング技術の常套手段である triphone である。その結果，環境が変わる度に耳慣らし（音響モデル適応／特徴量正規化）が不可欠となる。似ていないだろうか？筆者らには，音声認識における音響モデリングは，自閉症者の音感そのものであるように思える。問題の本質は，**図 6** に示した「音声 ↔ 音シンボル列変換」を物理的前提として音声の物理現象を解析することにあると考える。音声ストリームに対して，聴取者が感覚する音韻列を並べ，各音韻に対応する音声区間を切り出す。音韻は話者不変であるが，一方の物理現象は，人，場所，時，あらゆる要因がこれを変え，多様性問題に直面することになる。従来の音声科学・工学は「集めること」でこの問題を回避しようとしたが，本稿は，これを直接的に解く方法を提供している。筆者らは，**図 6** に示す一次元的音声視覚化技術は，物理的には，バグのある技術であると主張する。このバグのために「音声言語の正規ユーザ」が悩んでいても，何ら不思議ではない（失読症）。このバグのために，音シンボルの物理的対応物の不変性を信じて，その物理的対応物を絶対的に記憶する方々が音声言語運用に悩んでいても，何ら不思議ではない（自閉症）。

工学システムと自閉症者との類似性の議論は，古くはロボット工学に見られる。フレーム問題に端を発してロボットと自閉症児との類似性が議論されており，現在でも続いている[41, 42]。自閉症者は環境の些細な変化に非常に弱い側面を見せる。花瓶の位置が変わっただけでパニックに陥る場合もある。同様に，指定された部屋の情報を全てインプットされたロボットが，猫の来訪など，予期せぬ出来事にパニックに陥る。多様に変化する環境を頑健に対処できない両者に，工学者が，自閉症セラピストが，互いの類似性を認め合った経緯を持つ。環境の多様性を生き抜く術を与えるべく，工学者・セラピストが協力している。

言語発達に遅れの無い自閉症をアスペルガー症候群と言うが，彼らの音声言語活動は，やはり健常者とは異なる側面を示す[43, 44]。音声をまず文字化し，テキストを通して理解しようとする。そのため言語の論理面（文字面）だけの解釈となり，パラ言語的情報など文字化で消失する情報の処理が困難である。その音声が発せられた場・文脈を通して発言を解釈しようとせず，表層文だけに基づいて解釈を試みるため，場に合った対応ができず，多義性を解決する，行間・真意を読むなどの処理が苦手である。元々音声は苦手であり，電話音声などは特に困難である[43]。これらは，現状の音声対話システムに対しても，広く当てはまる性質である。アスペルガー症候群を患う者を家族に有する者は「計算機に音声コマンド入力するようなもの」と，彼らとの音声対話を記述している[43]。彼らの多くは自らを「地球生まれの異星人」と呼ぶ。感覚系・知覚系が健常者とは大きく異なるからである。「音声認識技術は，人間シミュレータを目指す必要は無い」という議論は古くからある。システムの入出力さえ模擬できれば，内部処理の実装まで模擬する必要は無い，という議論である。しかし，実際に構築したシステムは，そのビヘービアのみならず，内部処理の実装に至るまで非常に類似している「現実の対象物」が存在している。残念ながらそれは人間ではなく，自称異星人である，というのが筆者らの意見である。ヒューマノイドという名称で呼ばれる機械が巷に溢れているが，この「異星人」の存在を知る筆者らには，

[8] 実際には「ド」の意味が異なるので，勘違いでも錯覚でも無い。

少なくともその機械の音声処理系に関して「ヒューマン」という名称を使うことに強い抵抗を感じざるを得ない[9]。ロボット工学同様，音声の多様性を生き抜く術を両者に与えるべく，議論を重ねる必要があると考える。

7　生物進化と音の情報処理　〜絶対と相対〜

多くの動物は刺激間の相対的特性よりも，対象とする刺激の絶対的特性に基づいた処理を行なう傾向にあることが知られている。これは，相対的特性に基づく処理系の方が，より高度な認知能力を要求するからである，と考えられている[46]。音高に関しては，ラットやオオカミは絶対音感であることが報告されている。アカゲザルも基本的には絶対音感であるが，絶対性に基づく処理が失敗すると，相対性に基づく判断も行なう[47]。またニホンザルも同様，絶対音感としての処理が基本となっており，局所的な手がかりに着目する様子が報告されている[48]。このように生物進化の過程の中で音高処理が，絶対的な属性から相対的な属性へと遷移してきた様子が論じられている[49]。

本稿で論じてきた音声の構造的表象は，音高ではなく，スペクトル包絡という形で物理的に観測される音質に対する相対的な処理を対象としている。この音質に関する相対処理というのは，ヒト以外の動物では考察が困難である。そもそも，ヒトがこれだけ多様な母音を生成できるのは，二足歩行による喉頭の下落により，口腔に十分な空間を有するようになったからである。調音器官を制御して口腔を変形させることで，様々な共鳴パターンを生じさせ，これが様々な母音の生成を可能とした。当然口腔のサイズ／形状は話者依存であるため，音質の多様性は拡大する。本稿では，口腔のサイズ／形状に起因する静的な音響歪みを頑健に消失させる方法論として，音声の構造的表象を提案し，様々な観点から本手法を考察した。

8　まとめ

筆者らが提唱する音響的普遍構造が頑健な変換不変性を有することを数学的に示し，相対音感としての音声認知を通して，言語獲得，失読症，自閉症を考察した。本表象では音声の多様性問題が何ら問題になり得ず，また，パラドックスとも言われる，失読症や自閉症の音声認知についても，凡そ自然な考察で説明可能であることを示した。しかし，本考察がこれら障害の全容を網羅している訳ではなく，例えば失読症と自閉症の合併症が存在するのも事実である[50]。また，幾つかの凡そ典型的と考えられる症状について示したが，これらの障害は非常に多様な症状を呈しており，記述した各項目が常に観測される訳では無いことを断っておく。しかし，自閉症と音声認識技術の類似性について考察したように，自らを異星人を呼ぶ彼らのビヘービア及び認知特性は，より人間らしい機械を構築することを目指す工学者にとって，非常に有益な情報を提供していると筆者らは考える。**図 6** に示す音声の分節化及び要素の絶対的同定は，問題の要素還元に基づく方法論である。要素間の独立性を仮定した方法論である。その仮定に本質的な不備がある場合，要素分割とは異なる枠組みが必要となる。本稿はその一提案である。

[9]改名すべきモジュールが音声処理系だけであるかどうかは，言及しない。しかし，アスペルガー症候群の方々の身体の運動制御が，健常者のそれとは，やはり異なっていることを指摘しておく[45]。

参考文献

[1] 東田他, この地球にすんでいる僕の仲間たちへ, エスコアール出版社 (2005)
[2] 奥平, 自閉症の息子ダダくん 11 の不思議, 小学館 (2006)
[3] 谷口, 音は心の中で音楽になる, 北大路書房 (2003)
[4] 東川, 読譜力ー「移動ド」教育システムに学ぶ, 春秋社 (2005)
[5] 加藤, コミュニケーション障害学, 20, 2, pp.84–85 (2003)
[6] 早川, 月刊言語, 35, 9, pp.62–67 (2006)
[7] 原, コミュニケーション障害学, 20, 2, pp.98–102 (2003)
[8] 峯松他, 信学技報, SP2005-12, pp.1-8 (2005)
[9] 峯松他, 信学技報, SP2005-131, pp.121-126 (2005)
[10] 村上他, 信学技報, SP2005-14, pp.13-18 (2005)
[11] 村上他, 信学技報, SP2005-130, pp.115-120 (2005)
[12] 朝川他, 信学技報, SP2005-24, pp.25-30 (2005)
[13] 朝川他, 信学技報, SP2005-156, pp.37-42 (2006)
[14] 峯松他, 信学技報, SP2004-27, pp.47-52 (2004)
[15] 宮本, 音を作る・音を見る, 森北出版 (1995)
[16] N. Minematsu, *et al.*, "Universal and invariant representation of speech," Proc. Int. Conf. Infant Study (2006)
[17] 柏野, 月刊言語, 33, 9, pp.102–107 (2004)
[18] M. Spitzer, 脳・回路網の中の精神, 新曜社 (2001)
[19] 茂木, 心を生みだす脳のシステム, 日本放送出版協会 (2001)
[20] K. S. Scott *et al.*, Trends in Neurosci., 26, pp,100–107 (1003)
[21] P. Belin *et al.*, Nature Neurosci., 3, 10, pp.965–966 (2000)
[22] F. D. Saussure, Course in general linguistics, McGraw-Hilll Humanities/Social Sciences/Langua (1965)
[23] 前川, 音声研究, 8, 3, pp.35–40 (2004)
[24] R. Jakobson *et al.*, Preliminaries to speech analysis, MIT Press, Cambridge, MA (1952)
[25] R. Jakobson *et al.*, Notes on the French phonemic pattern, Hunter, N.Y. (1949)
[26] S. E. Blache, The acquisition of distincitve features, Univ. Park Press (1978)
[27] K. N. Stevens, J. Phonetics, 17, p.3–45 (1989)
[28] L. Deng *et al.*, Speech Comm., 33, 2–3, pp.93–111 (1997)
[29] M. Ostendorf, Proc. ASRU, pp.79–84 (1999)
[30] M. M. Waldrop, 複雑系, 新潮社 (2000)
[31] R. K. Potter *et al.*, JASA, 22, 6, pp.807–820 (1950)
[32] J. Hillenbrand *et al.*, JASA, 97, 5, pp.3099–3111 (1995)
[33] S. Shaywitz, 読み書き障害（ディスレクシア）のすべて〜頭はいいのに本が読めない〜, PHP 研究所 (2006)
[34] 石井, 科学技術政策研究所・科学技術動向 45, pp.13–24 (2004)
[35] W. Serniclaes *et al.*, Cognition, 98, pp.B35–B44 (2005)
[36] H. A. Gleason, An introduction of descriptive linguistics, Holt, Rinehart & Winston (1961)
[37] A. J. Lotto *et al.*, Chicago University Society, 35, pp.191–204 (2000)
[38] 酒木, 自閉症の子どもたち, PHP 研究所 (2001)
[39] U. Frith, 自閉症の謎を解き明かす, 東京書籍 (1991)
[40] R. Martin, 自閉症児イアンの物語, 草思社 (2001)
[41] 渡部, 鉄腕アトムと晋平君, ミネルヴァ書房 (1998)
[42] J. Nade, "The developing child with autism," Tutorial Session of IEEE Int. Conf. Development and Learning (2005)
[43] 泉, 僕の妻はエイリアン, 新潮社 (2005)
[44] 榊原, アスペルガー症候群と学習障害, 講談社 (2002)
[45] ニキリンコ, 自閉っ子, こういう風にできてます！, 花風社 (2004)
[46] D. J. Levitin *et al.*, Trends in Cognitive Sciences, 9, 1, pp.26–33 (2005)
[47] A. A. Wright *et al.*, Journal of Experimental Psychology, General, 129, pp.291–307 (2000)
[48] A. Izumi, Journal of Comparative Psychology, 115, pp.127–131 (2001)
[49] M. D. Hauser *et al.*, Nature Neurosciences, 6, pp.663–668 (2003)
[50] 月文他, 自閉症者からの紹介状, 明石書店 (2006)

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624-7 (11/17)

複数マイクロホンアレイのパーティクルフィルタ統合による実時間音源追跡

Real-Time Multiple Sound Source Tracking by Particle-Filter-based Integration of Heterogeneous Microphone Arrays

中臺 一博† **中島 弘史**† **村瀬 昌満**‡ **奥乃 博**‡ **長谷川 雄二**† **辻野 広司**†

Kazuhiro Nakadai†, Hirofumi Nakajima†, Masamitsu Murase‡,
Hiroshi G. Okuno‡, 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

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

Abstract

Real-time and robust sound source tracking is an important function for a robot operating in a daily environment, because the robot should recognize where a sound event such as speech, music and other environmental sounds originate from. This paper addresses real-time sound source tracking by spatial integration of an in-room microphone array (IRMA) and a robot-embedded microphone array (REMA). The IRMA system consists of 64 ch microphones attached to the walls. It localizes multiple sound sources based on weighted delay-and-sum beamforming on a 2D plane. The REMA system localizes multiple sound sources in azimuth using eight microphones attached to a robot's head on a rotational table. A particle filter integrates their localization results to track multiple sound sources. The experimental results show that particle filter based integration improved accuracy and robustness of sound source tracking even when the robot's head was in rotation.

1 はじめに

知覚のロバスト性向上において，様々な情報を統合することは本質的である．例えば，人間の知覚では，視聴覚の時間的な統合 [16] や, 音声認識における McGurk 効果 [10]，音源定位における視聴覚統合[11] などが報告されている．また，音源定位では，両耳間位相差 (interaural phase difference)，両耳間強度差 (interaural intensity difference) という二つの異なる情報を統合することによって，広周波数域にわたってロバストに音源を定位することを可能にしている[8]. こうした知見に基づき，実環境を扱うことを目的とした視聴覚統合システムも複数実装され，その有効性が報告されている[14, 6]. これは，統合が実環境で動作を行うロボットにとっても知覚を向上させるために本質的であることを物語っていよう．実際，これまでに，三話者が同時に発話した場合でも，視聴覚統合によって音源定位，分離，分離音認識が行えるロボット聴覚システムを報告し，人・ロボットコミュニケーションに有効であることを示してきた[12]. 報告したシステムはロボットの耳部に搭載された２本のマイクのみを利用していた．しかし，工学的なシステムとして捉えた場合，ロボット搭載型のマイクだけではなく周囲の環境に埋め込まれたマイクロホンも利用することはパフォーマンス向上のためには有効であろう．本稿では，ロバスト性・精度といったパフォーマンス向上を目的として複数のマイクロホンアレイを統合する空間統合 (*spatial integration*) を提案する．

1.1 二種類のマイクロホンアレイとその統合

空間統合用のマイクロホンアレイとして，ロボット搭載型マイクロホンアレイ (*Robot-Embedded Microphone Array, REMA*)，および室内設置型マイクロホンアレイ (*In-Room Microphone Array, IRMA*) という二種類の異種マイクロホンアレイの使用を検討する．REMA はロボット搭載マイクロホンを用いてロボット聴覚を向上させる直接的なアプローチであり，ロボット近傍で高い解像度が得られるという特徴を持っている．実際，8 チャネルの REMA を用いて，両耳聴システムよりも音源定位・分離で優れた性能をもったシステムが報告されている[5, 18]．しかし，このアプローチはロボット動作中に動作と収音音響信号の正確な同期を取ることが難しいこと，動作によって動的に変化する音響環境への対応が難しいこと，距離が離れた音源の情報を正確に抽出することが難しいことといった本質的な欠点を抱えている．一方で，IRMA は，相当数のマイクロホンを必要とし，比較的処理の解像度は低いものの，設置型のアレイであるため，常に静止しており，動作に起因する問題を扱う必要がない．また，マイクロホンを部屋中にちりばめることによって，人とロボット間の距離に影響されず，部屋内の位置とは無関係に音源情報を抽出することが可能である．実際に，音源定位や分離を目的とした大規模マイクロホンアレイも複数報告されている[1, 15, 23]．このように，REMA と IRMA は，お互いの欠点と利点が相補的関係にあるため，両者を統合することにより，お互いの曖昧性を解消することが可能であると考えられる．本稿では，REMA, IRMA という二種類のマイクロホンアレイを統合するため，パーティクルフィルタに基づいた手法を提案する．また，実際に 8 チャネルの REMA と 64 チャネルの IRMA をパーティクルフィルタで統合した

空間統合システムを構築し，その効果を複数音源追跡を通じて示す．

以後，2 章では，各マイクロホンアレイで用いた定位のアルゴリズムについて述べ，3 章では，マイクロホン統合に用いたパーティクルフィルタについて解説する．4 章では，空間統合による音源追跡システムの実装について述べ，5 章でシステムの評価を行う．最後に，6,7 章でまとめ，今後の課題について議論する．

2 定位のアルゴリズム

2.1 ロボット搭載型マイクロホンアレイ (REMA)

REMA を用いた実時間音源定位に関しては，これまで 2 本のマイクによる両耳聴システム[12]，および遅延和型ビームフォーマーに基づくマイクロホンアレイシステム[24]を報告した．本稿では，より音響環境の変化にロバストな手法として，適応ビームフォーマの一種である *Multiple Signal Classification (MUSIC)* [3] を採用した．MUSIC は環境の変化に逐次的に適応することにより音源定位のロバスト性を向上させることが出来る．また，事前計測のインパルス応答を用いて伝達関数を生成しておくことにより，実時間動作も可能である．なお，MUSIC は，産総研の実装[5]を利用した. これは，実環境で動作するヒューマノイドロボットを対象に開発された実装であり，実時間でロバストに動作することが特徴である．アルゴリズムの詳細は文献[3]を参照されたい．

2.2 室内設置型マイクロホンアレイ (IRMA)

IRMA については，重み付き遅延和法 (*weighted delay-and-sum beamforming (WDS-BF)*) [13] を用いた． 一般に, 典型的なビームフォーミングでは，システム出力 $Y_{\boldsymbol{p}}(\omega)$ は，下記の式で表される．

$$Y_{\boldsymbol{p}}(\omega) = \sum_{n=1}^{N} G_{n,\boldsymbol{p}}(\omega) X_n(\omega) \tag{1}$$

$$X_n(\omega) = H_{\boldsymbol{p},n}(\omega) X(\omega) \tag{2}$$

ここで $X(\omega)$ は，座標 $\boldsymbol{p}$ に置かれた音源 S のスペクトルである．$H_{\boldsymbol{p},n}(\omega)$ は，S から n 番目のマイクへの伝達関数を表す． $X_n(\omega)$ は，n 番目のマイクによって収音された信号のスペクトルである．$G_{n,\boldsymbol{p}}(\omega)$ は，n 番目のマイクへの入力信号のスペクトルから $\boldsymbol{p}$ におけるスペクトルを推定するためのフィルタ関数を示す．WDS-BF では，測定結果や計算的に導出されたものなど様々なタイプの伝達関数を統一的に扱えるよう一般化を行っている．また，伝達関数 $H_{\boldsymbol{p},n}$ の動的変化や入力信号 $X_n(\omega)$ の歪みなどにロバストになるように，$G_{n,\boldsymbol{p}}(\omega)$ のノルムを最小化している[25]．しかし，マイクロホン数が多くなると，計算量の増大により，実時間動作は困難となってしまう．そこで，本稿では，計算量を削減するためにマイクロホンアレイの部分集合のみを利用するサブアレイ法を導入する．サブアレイ法におけるマイクの選択方法は音源と各マイクロホンの距離によって決める．具体的には，n 番目のマイクロホンと音源との距離 r_n が r_{th} 以下の場合, そのマイクロホンを選択し，そうでない場合はそのマイクは選択せず，伝達関数の値を 0 に設定する．

また，WDS-BF は，式 (1), (2) における $\boldsymbol{p}$ を $\boldsymbol{p}' = (\boldsymbol{p}, \theta)$ と置き換えることにより指向特性推定に適用可能である．指向特性推定を応用すれば，音源が向いている方向を推定したり，実際に人間が話しているのか，スピーカから出力された音声なのかを判断したりすることが可能である．詳細は，[13] を参照されたい．

3 マイクロホンアレイの統合

二種類のマイクロホンアレイを統合するためパーティクルフィルタ[2] 用いた. パーティクルフィルタは物体の視覚的な追跡や *Simultaneous Localization And Mapping (SLAM)* [17] を効率的に解くために用いられる手法であり，パーティクルを用いて状態をサンプリングし，サンプリングによって得られたパーティクルを遷移モデル，および，観測モデルを用いて更新していくことにより，観測結果から内部状態を推定する．パーティクルフィルタは，線形な遷移しか扱えない Kalman フィルタなどと異なり，非線形の動きを追跡する枠組みを備えていること，ガウシアン以外の分布を扱えること，動作速度がパーティクルの数で制御可能なため実時間動作が可能なことといった特長を持っている．また，パーティクルフィルタは，遷移モデル，および，観測モデルを用意すれば，データの種類を問わず利用できるため，音源追跡への適用も試みられいる[22, 21, 20]．例えば，Valin[19] らは，物体追跡で利用されている方法[22, 9, 7]を適用し，複数音源に対応したパーティクルフィルタを報告している．麻生らは，パーティクルフィルタを拡張し，視聴覚統合による音源追跡を報告している[4]．しかし，異なるタイプのマイクロホンアレイの統合を考えた際，得られる定位結果の座標系の違いに由来する問題のため，こうした手法をそのまま適用することは難しい．

3.1 マイクロホンアレイ統合の課題

異種マイクロホンアレイを統合ためには，ロボット座標系 vs. 絶対座標系，および極座標系 vs. x-y 座標系に由来する二つの課題を考慮する必要がある．

REMA は移動可能であるため，定位結果は常にロボット座標系で観測される．一方，IRMA は静止しているため，絶対座標系で観測される．つまり，統合のためには，REMA の座標系を絶対座標系へ変換する必要がある．このためには，音響処理とロボット動作の高精度な時間同期が必要である．この同期問題を解決するために二つのアプローチが考えられる．一つはパーティクルフィルタに時間方向の分布を表すファクターを導入して，時間同期の曖昧性を解いてしまうソフトウェアベースのアプローチである．もう一つは，動作と音響信号取得を高精度に同期できるハードウェアを導入するハードウェアベースのアプローチである．前者は，研究課題として興味深いが，確率的表現を用いている以上，少なからず同期にエラーが混入してしまう．そこで，本稿では，後者のアプローチを取ることにした (4.1 節参照).

2 つめの課題は定位の次元数に由来する．IRMA は 2 次元の定位を行うことに対し，REMA では 1 次元 (水平角) の定位を行う．かつ，これらは x-y 座標系と極座標系という異なる座標系で観測される．本稿では，それぞれの座標系で尤度関数を用意し，定位結果に対して，座標系と独立な値として尤度を算出し，尤度レベルで統合を行う手法を提案する．

3.2 パーティクルフィルタ

パーティクルフィルタでは, 内部状態 $\boldsymbol{x}(t)$ の遷移モデル $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1))$ と観測モデル $p(\boldsymbol{y}(t)|\boldsymbol{x}(t))$ を確率的な表現として定義する．なお，$\boldsymbol{y}(t)$ は観測ベクトルを表す. i 番目のパーティクルは内部状態 $\boldsymbol{x}_i(t)$ とそのパーティクルが真値（ここでは，音源追跡結果）にどの程度貢献するかを示す重要度 $w_i(t)$ を持っている．重要度は一般に尤度として定義される．本稿では，観測ベクトル $\boldsymbol{y}(t)$ として，REMA から $\boldsymbol{Y}_{\mathrm{REMA}}(t)$，IRMA から $\boldsymbol{Y}_{\mathrm{IRMA}}(t)$ という 2 種類が時刻 t で得られるとする．

$$\boldsymbol{Y}_{\mathrm{REMA}}(t) = \{\boldsymbol{y}_{a_1}(t), \cdots, \boldsymbol{y}_{a_l}(t), \cdots, \boldsymbol{y}_{a_{L_t}}(t)\}, \quad (3)$$

$$\boldsymbol{Y}_{\mathrm{IRMA}}(t) = \{\boldsymbol{y}_{b_1}(t), \cdots, \boldsymbol{y}_{b_m}(t), \cdots, \boldsymbol{y}_{b_{M_t}}(t)\} \quad (4)$$

L_t，M_t は，それぞれ REMA，IRMA から得られる時刻 t における観測の数である．$\boldsymbol{y}_{a_l}$ と $\boldsymbol{y}_{b_m}$ は下記のように定義する．

$$\boldsymbol{y}_{a_l} = \{a_{\theta_l}, a_{p_l}\}, \quad (5)$$

$$\boldsymbol{y}_{b_m} = \{b_{x_m}, b_{y_m}, b_{o_m}, b_{p_m}\}, \quad (6)$$

a_{θ_l} は，絶対座標系での水平角を示し，b_{x_m}，b_{y_m} は，絶対座標系での位置を示す．また，b_{o_m} は音源の進行方向, a_{p_l}，b_{p_m} は，推定されたパワーを示している．

このように複数の異なる定位が得られる場合に，複数の音源が存在していても音源追跡が実現できるよう一般的なパーティクルフィルタに対して改良を行った．改良は，主に次の 3 点について行った．

1. 複数音源が扱えるよう複数のパーティクルグループを許容し，観測状況により，グループ数を動的に変化させる機構の実装
2. 移動音源はある程度の速度があれば急激に進行方向を変えないという仮定の下，ランダムウォークと運動方程式を音源速度に応じて使い分ける非線形な遷移モデルの採用
3. REMA，IRMA から得られる次元数の異なる定位情報を透過的に統合するため，尤度レベルで定位情報を統合する機構の実装

実際に構築した改良パーティクルフィルタは，初期化, 音源生成・消滅チェック, 重要度サンプリング *(Importance sampling)*, 選択, and 出力 の５つのステップからなっている. 以下に，詳細を記述する．

Step 1 – 初期化

パーティクルの初期化を行う．i 番目のパーティクルの内部状態 $\boldsymbol{x}_i$ を $(x_i(t), y_i(t), v_i(t), o_i(t))$ とする．$(x_i(t), y_i(t))$ は音源の位置，$v_i(t)$ は音源の速度，$o_i(t)$ は音源の進行方向である．初期化では，すべてのパーティクルを一様にかつランダムに分布させる．また，複数音源を扱うために, パーティクルグループを導入して，重要度を下記のように定義した．

$$\sum_{i \in P_k} w_i = 1, \quad \sum_{k=1}^{S} N_k = N, \quad (7)$$

ここで，N_k は k 番目のパーティクルグループ P_k のパーティクルの数，S は音源の数. N は全パーティクルの総数である．

Step 2 – 音源生成・消滅チェック

この Step は，複数音源を扱うために新規に追加した．パーティクルグループ P_k の内部状態は下記のように定義される．

$$\hat{\boldsymbol{x}}_k(t) = \sum_{i \in P_k} \boldsymbol{x}_i(t) \cdot w_i(t) \quad (8)$$

時刻 t における REMA もしくは IRMA からの j 番目の観測 $\boldsymbol{y}_{a_j}(t)$ もしくは $\boldsymbol{y}_{b_j}(t)$ をまとめて $\boldsymbol{y}_j(t)$ と表すことにすると $||\hat{\boldsymbol{x}}_k(t) - \boldsymbol{y}_j(t)|| < D_{th}$ を満たす場合, $\boldsymbol{y}_j(t)$ は P_k にアソシエートされる．ここで，$||\cdot||$ は，ユークリッド距離を表している．また，$\boldsymbol{y}_m(t)$ に対応するパーティクルグループが見つからない場合，新規にパーティクルグループを生成する．パーティクルグループ P_k にアソシエートする観測が一定時間 T_{th} 以上得られなかった場合，P_k は，消滅する．いずれの場合も，パーティクルは，式 (7) が満たされるように再配置される．

Step 3 – 重要度サンプリング

この step では，まず，遷移モデル $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1))$ を用いて，$\boldsymbol{x}_i(t-1)$ から $\boldsymbol{x}_i(t)$ を推定する．次に，$w_i(t)$ を式 (20) を用いて更新する．最後に, 式 (7) に従って，$w_i(t)$ を正規化する．

遷移モデルに対しては，前述のようにランダムウォークと運動方程式を音源の速度に応じて使い分けるような非線形のモデル化を行った．具体的には，音源速度が v_{th} 以下の場合, システムはランダムウォークモデルを採用する（本稿では，v_{th} の値は $2\,m/s$ とした）．この場合の遷移モデルは下記の式で定義される．

$$x_i(t) = x_i(t-1) + r_x, \quad (9)$$

$$y_i(t) = y_i(t-1) + r_y, \quad (10)$$

$$v_i(t) = v_i(t-1) + r_v, \quad (11)$$

$$o_i(t) = o_i(t-1) + r_o, \quad (12)$$

なお，r_* は白色雑音を示している．なお，分散は実験的に求めた値を使用した．

音源速度が v_{th} より大きい場合は, システムは下記に定義される運動方程式に基づくモデルを用いる．

$$\begin{aligned} x_i(t) &= x_i(t-1) \quad (13)\\ &+ v_i(t-1) \cdot \cos(o_i(t-1)) + r_x, \\ y_i(t) &= y_i(t-1) \quad (14)\\ &+ v_i(t-1) \cdot \sin(o_i(t-1)) + r_y, \\ v_i(t) &= \alpha \cdot v_i(t-1) \quad (15)\\ &+ (1-\alpha) \cdot \sqrt{\Delta x_i(t)^2 + \Delta y_i(t)^2} + r_v, \\ o_i(t) &= \alpha o_i(t-1) \quad (16)\\ &+ (1-\alpha) \cdot \tan^{-1}\left(\frac{\Delta y_i(t)}{\Delta x_i(t)}\right) + r_\theta, \end{aligned}$$

ここで α は重みパラメータであり，実験的に求めた．$\Delta x_i(t)$ と $\Delta y_i(t)$ は下記式で定義される．

$$\Delta x_i(t) = x_i(t) - x_i(t-1), \quad \Delta y_i(t) = y_i(t) - y_i(t-1).$$

尤度の定義は下記に示すとおりである．

$$l_{\mathrm{REMA}}(t) = exp\left(-\frac{(\angle(\boldsymbol{x}_i(t) - \boldsymbol{P}_{\mathrm{REMA}}(t)) - \theta_l)^2}{2R_{\mathrm{REMA}}}\right) \quad (17)$$

$$l_{\mathrm{IRMA}}(t) = exp\left(-\frac{||\boldsymbol{x}_i(t) - \boldsymbol{y}_{b_m}(t)||^2}{2R_{\mathrm{IRMA}}}\right) \quad (18)$$

ここで $\angle(\boldsymbol{x})$ は，ベクトル $\boldsymbol{x}$ が x 軸となす角度を示す．R_{REMA} および R_{IRMA} は REMA と IRMA の音源定位結果の分散である．$\boldsymbol{P}_{\mathrm{REMA}}$ は，ロボットの位置を示す．$l_{\mathrm{REMA}}(t)$ および $l_{\mathrm{IRMA}}(t)$ を下記の式で統合し，$l_I(t)$ を導出することによりマイクロホンアレイの尤度レベルの統合を行っている．

$$l_I(t) = \alpha_l \cdot l_{\mathrm{REMA}}(t) + (1 - \alpha_l) \cdot l_{\mathrm{IRMA}}(t) \quad (19)$$

ここで α_l は統合用重みパラメータである．最後に w_i を下記式で更新する．

$$w_i(t) = l_I(t) \cdot w_i(t-1). \quad (20)$$

Step 4 – 選択

重要度 w_i に応じて, パーティクルの更新を行う．$i \in P_k$ を満たす i に対するパーティクル数は下記式で更新される．

$$N_{k_i} = round(N_k \cdot w_i). \quad (21)$$

この場合，R_k 個のパーティクルが更新されないままとなっている．

$$R_k = N_k - \sum_{i \in P_k} N_{k_i}. \quad (22)$$

これらのパーティクルも，残差重みパラメータ R_{w_i} に従って分配される．

$$R_{w_i} = w_i - N_{k_i} / \sum_{i \in P_k} N_{k_i} \quad (23)$$

この際, 一般的な *Sampling Importance Resampling (SIR)* アルゴリズム [2] を用いている．

Step 5 – 出力

更新後のパーティクルの密度から事後確率 $p(\boldsymbol{x}(t)|\boldsymbol{y}_m(t))$ を推定する. 音源 k に対するパーティクルグループの内部状態は式 (8) によって推定される. Steps 2 – 5 が追跡が終了するまで繰り返される．

4 システム実装

図 1 に空間統合システムの構成図を示す．REMA 搭載ロボット, IRMA システム, マイクロホンアレイ統合器, 音源ビューワの 4 つのコンポーネントからなる．

4.1 REMA 搭載ロボット

REMA 搭載ロボットを図 2a) に示す．ロボットは 8 チャネルの REMA を搭載した Honda ASIMO の頭部，回転台，全方向移動台車から構成されている．REMA は，ゴム製のヘアバンドに 8 本の無指向性マイクを等間隔で配置したものであり，これを ASIMO の頭部に設置した．回転台，全方向移動台車は PC から制御可能になっている．ただし，

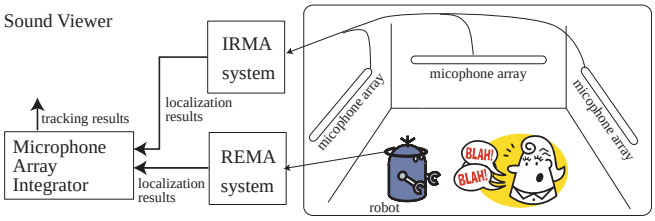
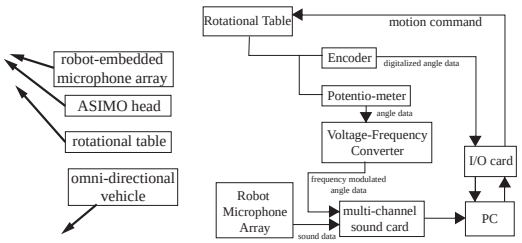


Figure 1: Spatial Integration System



a) REMA on Robot　　b) Architecture

Figure 2: REMA System

今回は回転台のみを用い，全方向移動台車は静止させて使用した．

回転台の動作情報と REMA の音響信号処理をハードウェア的に高精度に時間同期させ，音源定位結果を絶対座標系に変換するため，図 2b) に示すアーキテクチャを構築した．

回転台の回転角はエンコーダを用いて 0.0015° の解像度で精度よく計測できるが，情報出力までに遅延が生じる．同期を行う際には，この遅延を考慮する必要がある．一方，ポテンショメータは，アナログ出力であるため，解像度は，0.95° 程度であるが，時間遅延はほぼ 0 で角度情報を計測することができる．そこで，ポテンショメータも同時に回転台に設置し，両センサからの情報を比較することにより，遅延の計測を試みた．なお，ポテンショメータは電圧-周波数コンバータを介してサウンドカードに接続するようにしている．これは，ポテンショメータから得られる回転情報がサウンドカードで DC 成分としてフィルタリングされないようにするためである．比較した結果，エンコーダには平均 32.9 ms の時間遅延があることがわかった．そこで，絶対座標変換の際にはこの値を考慮して変換を行うことにより高精度な同期を実現した．

4.2 IRMA システム

64 ch の IRMA システムを構築した. 構築システムは 4 台の JEOL 製 RASP II を用いて，同期して 64 チャネルの信号を 16 kHz サンプリングで収音することが出来る．図 3 は IRMA システムが実装されている 4.0 m × 7.0 m の部屋を示している. 三方の壁は吸音材で覆われており，残り一方の壁はガラス製である．また，室内にはキッチン台が置かれているなど，反響が一様でない部屋となっている．壁に設置されたマイクの高さは 1.2 m である．マイク配置は，なるべく部屋全体をカバーできるような配置となっている．IRMA 用のビームフォーマを設計するため，まず，25 cm メッシュを用いて，室内の離散化を行った．離散化した領域は，X 軸方向が 1.0 m – 5.0 m，Y 軸方向が 0.5 m – 3.5 m まで，Z 軸（高さ）方向は 1.2 m で固定した．従って，離散化によって 221 点の音源定位用の評価点をサンプルした．次に，反響が一様でない環境や任意のマイクレ

Table 1: The effect of a sub-array on computational cost (simulation)

r_{th}	computational cost (%)	# of ch to use	
		Max	Min
7	100	64	64
6	99.9	64	63
5	97.4	64	41
4	82.4	64	33
3.5	68.8	62	22
3	53.0	56	19
2.5	37.5	39	12
2	23.2	29	0

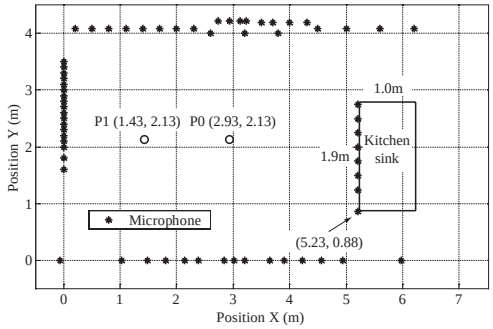


Figure 3: Layout of Microphones

イアウトに対応するため，すべての音源定位用評価点でスピーカを P_0 に置かれたロボットの方向に向けた状態でインパルス応答の計測を行い，伝達関数を計測した．この伝達関数によって得られるビームフォーマを “M-BF” と呼ぶものとする．また，M-BF に対してサブアレイ法を適用したビームフォーマ（以後，“MS-BF” とする) を設計した．この際，2.2 節で述べた距離の閾値 r_{th} は，3.5 m に設定した．この場合，表 1 に示したように，平均 30% の計算量削減を見込むことが出来ることをシミュレーションによって確認した．

4.3 マイクロホンアレイ統合器とサウンドビューワ

マイクロホンアレイ統合器では，3.2 節で説明したパーティクルフィルタによって，REMA と IRMA の定位結果を統合し，音源追跡を行う．追跡結果はサウンドビューワに送られる．サウンドビューワは，Java3D で実装されており，実時間 3D 表示機能を有している．

5 評価

構築システムを用いて，音源定位，および音源追跡のパフォーマンスの評価を行った．

音源定位に関しては，まず，その基本性能を知るため，単一音源の定位を IRMA および REMA を用いて行い，定位の誤差平均とその標準偏差を計測した．音源には，図 3 に示す P1 に配置したアクティブスピーカ GENELEC 1029A を用いて再生した録音音声を用いた．スピーカの方向は 0° とした．なお，(1,0) ベクトルの方向を 0 度とし，＋方向は反時計回りの方向とした．

IRMA 用のビームフォーマには，“M-BF”, “MS-BF”, “Sim-BF”, “RSim-BF” の 4 種類を用いた．“M-BF”, “MS-BF” は 4.2 節で説明したものである．“Sim-BF” は単に室内が自由空間であることを仮定して，“RSim-BF” は壁の反響を考慮して，シミュレーション計算により設計したものである．

音源追跡に関しては，下記の 5 つの状況を設定し，複数音源の同時追跡行った．

Ex.2A: 録音音声を出力するスピーカを (2.93 m, 0.63 m) から (2.93 m, 3.63 m) まで，P_0 を中心とした半径 1.5 m の弧上を反時計回りに動かした．ロボットは P_0 に設置し，向きは 180° の方向に固定した．

Ex.2B: スピーカを P_1 に，ロボットを P_0 に固定した. ただしロボットは 90° から 270° まで回転させた. 他の条件は **Ex.2A** と同じである．

Ex.2C: スピーカを **Ex.2A** と同様に移動させた. ロボットは P_0 に設置したが，その向きはスピーカの動きに追従するように 90° から 270° まで移動させた．

Ex.2D: 2 人の男性被験者 (A 氏と B 氏) に発話しながら，中心 P_0，半径 1.5 m の円に沿って移動するように依頼した．発話は，日本語の文章であり，常にロボットの方向を向いてしゃべってもらった．A 氏は (2.93 m, 0.63 m)，（つまりロボット座標系で水平角が 90°）を起点に時計方向に 0° まで移動してもらい，そこから折り返して，270° まで反時計回りに移動してもらった. B 氏は (2.93 m, 3.63 m) を起点として，A 氏と対称になるように移動してもらった．つまり，まず，反時計回りに 0° まで移動し, 次に時計回りに 90° まで移動した．2 人は 0° で近づいてから離れ，180° では近づいてそのまま交差するという状況になっており，音源追跡ではこの曖昧性を解決する必要があるような状況である．ロボットは P_0 に固定し，向きも 180° に固定した．

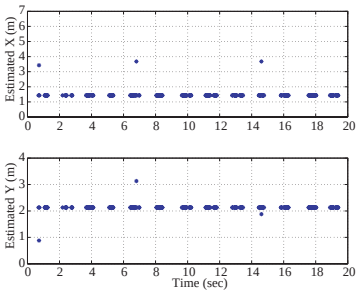
Ex.2E: 被験者の動作は **Ex.2D** と同じである. ロボットの位置は P_0 固定であるが，その向きは，常に A 氏の方向を向くように回転させた．

リファレンスデータを取得するために，超音波 3D タグシステム (U3D-TS) を用いた．このシステムは，超音波 3D タグを，数センチの誤差で定位することが可能である．[13]，今回は，被験者やスピーカにタグを設置して同時にリファレンスデータを取得した．また，実験では, IRMA 用のビームフォーマとして MS-BF を用いた．

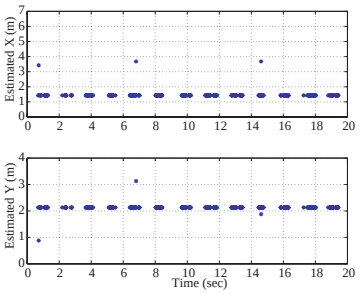
5.1 結果

図 4a)-d) は IRMA による単音源の定位結果を示している．横軸は時刻，縦軸は推定した X, Y の値をメータで示したものである．図 4e) は REMA による定位結果を示している．横軸は時刻，縦軸は推定した音源の極座標系での水平角となっている．図 4f) は定位の平均誤差と標準偏差を示している．

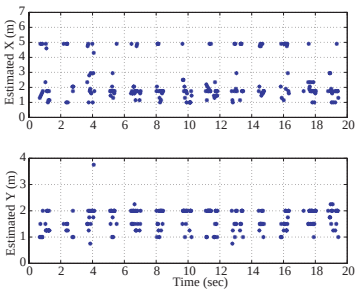
図 5 は音源追跡実験の結果を示している．図 5 の各列は上からそれぞれ実験 **Ex.2A** – **Ex.2E** に対応している. 左の行は REMA による定位結果を示している. 横軸は時間であり，縦軸は推定した水平角である．青いアスタリスクは，ロボット座標系での定位結果を，赤線はエンコーダから得られたロボットの動作情報を絶対極座標系で表している．赤いプラスマークは，絶対極座標系に変換した後の定位結果を表している．中央の行は IRMA による定位結果を表している. 青いアスタリスクは絶対 x-y 座標系



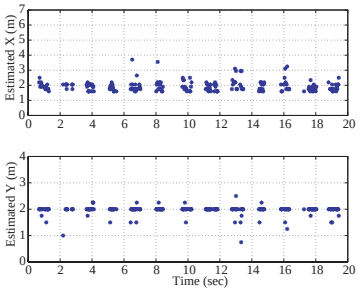
a) M-BF(IRMA)



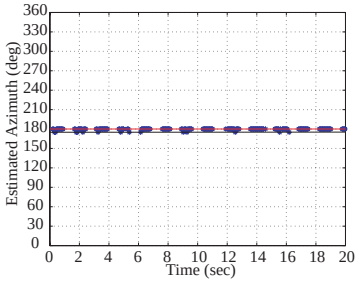
b) MS-BF(IRMA)



c) Sim-BF(IRMA)



d) RSim-BF(IRMA)



e) MUSIC(REMA)

Beamformer	Error	
	Avg.	S.D.
M-BF (IRMA)	0.016 m	0.039 m
MS-BF (IRMA)	0.019 m	0.041 m
Sim-BF (IRMA)	0.95 m	1.19 m
RSim-BF (IRMA)	0.50 m	0.52 m
MUSIC (REMA)	4.56°	1.41°

f) Error of Localization

Figure 4: Sound Source Localization Results

Table 2: Localization Error with REMA and IRMA

	REMA		IRMA	
	Avg.(deg)	S.D.(deg)	Avg.(m)	S.D.(m)
Ex.2A	4.01	16.18	0.217	0.157
Ex.2B	3.25	7.61	0.082	0.249
Ex.2C	5.96	3.16	0.190	0.303
Ex.2D	6.14	10.66	0.194	0.173
Ex.2E	7.46	7.83	0.234	0.200

Table 3: Tracking Error with Particle Filter

	IRMA Only		Integration of IRMA and REMA	
	Avg.(m)	S.D.(m)	Avg.(m)	S.D.(m)
Ex.2A	0.12	0.062	0.10	0.040
Ex.2B	0.06	0.012	0.06	0.012
Ex.2C	0.11	0.075	0.10	0.071
Ex.2D	0.16	0.084	0.16	0.083
Ex.2E	0.18	0.133	0.17	0.123

での定位結果を表している．右の行は，パーティクルフィルタを用いた音源追跡結果である．赤線は IRMA から得られた定位結果のみを用いた場合の音源追跡結果である．青線は，REMA と IRMA 両方の定位結果を用いた場合の音源追跡結果である．また，各図の黒線と緑線は U3D-TS によって得られた音源方向のリファレンスデータである．表 2 は REMA および IRMA における音源定位誤差の平均および標準偏差を表している．また，表 3 は音源追跡の誤差を表している．

5.2 考察

音源定位実験からは，M-BF および MS-BF の精度がよいことがわかる．定位誤差は 15 – 20 cm 程度であり，メッシュサイズが　25 cm であることを考えれば，小さい値であるといえよう．これらのビームフォーマは測定した伝達関数に基づいて設計されたものであり，測定環境における反響などのノイズ成分にロバストである．処理速度も考えた場合，表 1 に示したとおり, 定位精度を保ったまま 30% の計算量削減を達成した．実際 IRMA システムは，サブアレイ法の導入によって，約 16 fps の処理をリアルタイムで可能にした．

事前計測した離散点では伝達関数が利用可能であるが，その他の点をサポートするためには，何らかのインターポレーションが必要である．そのような場合には，RSim-BF を利用することも可能であると考えられる．REMA で MUSIC を用いた場合，約 4.5° の定位誤差となった．これは，ロボットから 1.5 m 離れた場所では，12 cm のエラーに相当する．つまり，ほぼ IRMA の定位精度と同等であるといえる．定位の解像度はより近い音源では精度が高くなり，遠くなるに従い悪くなる．今回の音源追跡実験はロボットと音源の距離は約 1.5 m であったので，統合用の重みパラメータ α_l は 0.5 とした．

音源追跡実験における，REMA と IRMA の定位結果については，U3D-TS の追跡結果と比較すると，一部に定位結果の飛びが見受けられる．また，表 2 では，定位誤差は音源数が増加したり，音源が移動したりすると誤差が増大することを示している．しかし，その増加は数センチ，もしくは数度程度の範囲に収まっており，REMA, IRMA ともに 2 つの移動音源の精度よい音源定位が行われていることがわかる．また，REMA の座標系変換に関しては，変換結果が U3D-TS から得られた定位結果にフィットしており，精度の高い時間同期が達成できたといえる．

しかし，音源定位だけでは，定位結果と対応する音源とのアソシエーション問題は解決されていない．これは，ICA による音源分離ではパーミュテーション問題と呼ばれ，複数音源を扱う際には本質的な問題である．図 5 の右図のようにパーティクルフィルタにより，この問題が解決されることがわかる．特に，前述した 10 秒付近と 20 秒付近に生じている曖昧性が解決され，正しい追跡がなされていることがわかる．これは，被験者が，自然に，近づいて離れる際には速度を落とし，交差する際には速度を落とさないという状況を非線形な遷移モデルがうまく扱っていることを示している．加えて，図 3 から平均定位誤差が 2 cm – 9 cm 程度，標準偏差が平均 10 cm 程度低

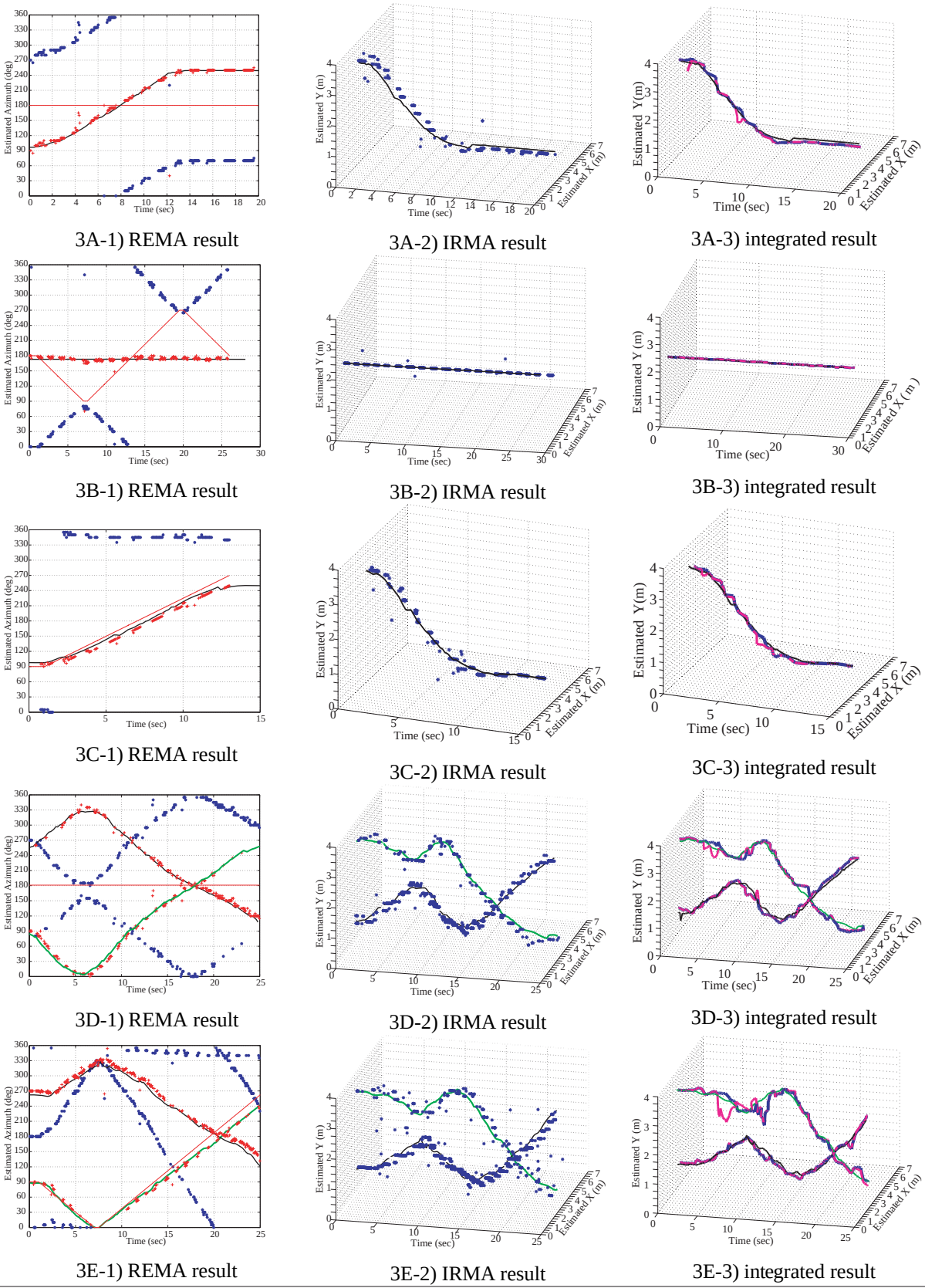


Figure 5: Tracking Results

減されており，パーティクルフィルタは音源追跡の精度，ロバスト性を向上させることがわかる．実際に，IRMA の結果のみを使った追跡結果（赤線）は，図 5 の 5 秒から 10 秒付近で大きな誤差が生じているが，IRMA, REMA を統合した追跡結果 (青線) ではこのエラーが低減されていることがわかる．

6 結論

ロボット聴覚を向上させるために異なる 2 つのタイプのマイクロホンアレイをパーティクルフィルタによって統合する空間統合システムを報告した．IRMA については，重み付遅延和アレイを実時間で動作させるためにサブアレイ法を導入し，その効果を示した．また，空間統合のため，新規に複数音源に対応したパーティクルフィルタを提案した．実際に，64 チャネルの IRMA，および 8 チャネルの REMA を用いて実時間空間統合システムを構築した．6 種類の状況設定を行って音源追跡実験をした結果，提案手法が，精度，および，ロバスト性を向上させることを示した．

7 今後の課題

実際にはパーティクルフィルタを利用する際にいくつかのパラメータを設定する必要がある．現状では，実験的にこれらの値を設定しているが，自動的に設定出来るようにする必要がある．また，音源数に関しては，現状，最大 2 つまでという制約を置いているが，この制約も緩和する必要があろう．また，IRMA については，複数の小さなアレイを利用してパフォーマンスを落とさずにマイク数を削減できる可能性がある．これは今後の課題である．本稿では，音源追跡を報告したが，音源分離や音声認識に関しても今後報告していきたい．多数のマイクロホンが室内に配置されている状況は，多数のセンサを用いるユビキタス社会を念頭に置けば，決して，飛躍した考え方ではないと考えている．

謝辞

本研究を進めるにあたり，サポートや貴重な意見を頂いた京都大学の海尻聡氏，山本俊一氏，産総研の浅野太氏，麻生秀樹氏に感謝する．

参考文献

[1] P. Aarabi and S. Zaky. Robust sound localization using multi-source audiovisual information fusion. *Information Fusion*, 2(3):209–223, 2001.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[3] Futoshi Asano, Masataka Goto, Katunobu Itou, and Hideki Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. In ISCA, editor, *Proc. of European Conforence on Speech Processing (Eurospeech 2001)*, pages 1013–1016, 2001.

[4] H. Asoh, F. Asano, K. Yamamoto, T. Yoshimura, Y. Motomura, N. Ichimura, I. Hara, and J. Ogata. An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion. In *International Conference on Information Fusion*, pages 805–812, 2004.

[5] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoo. Robust speech interface based on audio and video information fusion for humanoid HRP-2. In *Proc. of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2004)*, pages 2404–2410. IEEE, 2004.

[6] J. Hershey, H. Ishiguro, and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Neural Information Processing Systems*, volume 12, pages 813 – 819. MIT Press, 2000.

[7] C. Hue, J.-P. L. Cadre, and P. Perez. A particle filter to track multiple objects. In IEEE, editor, *IEEE Workshop on Multi-Object Tracking*, pages 61–68, 2001.

[8] L.A. Jeffress. A place theory of sound localization. *Journal of Comparative Physiology and Psychology*, **41**:35–39, 1948.

[9] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, 39(1):57–71, 2000.

[10] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, **264**:746–748, 1976.

[11] D. H. Mershon, D. H. Desaulniers, S. A. Kiefer, T. L. Amerson, Jr., and J. T. Mills. Perceived loudness and visually-determined auditory distance. *Perception*, **10**:531–543, 1981.

[12] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, 44:97–112, 2004.

[13] K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura, and H. Tsujino. Sound source tracking with directivity pattern estimation using a 64 ch microphone array. In *Proc. of the IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS 2005)*, pages 196–202, 2005.

[14] G. Potamianos and C. Neti. Stream confidence estimation for audio-visual speech recognition. In *Proceeding of the International Conference on Spoken Language Processing (ICSLP 2000)*, pages 746–749. ISCA, 2000.

[15] H.F. Silverman, W.R. Patterson, and J.L. Flanagan. The huge microphone array. Technical report, LEMS, Brown University, 1996.

[16] Y. Sugita and Y. Suzuki. Audiovisual perception: Implicit estimation of sound-arrival time. *Nature*, **421**:911, 2003.

[17] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, 2005.

[18] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In IEEE, editor, *Proc. IEEE International Conference on Robotics and Automation (ICRA 2004)*, 2004.

[19] Jean-Marc Valin. *Auditory System for Robot*. PhD thesis, Universitè de Sherbrooke, 2005.

[20] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In IEEE, editor, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3021–3024, 2001.

[21] D. B. Ward, E. A. Lehmann, and R. C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11(6):826 – 836, 2003.

[22] D. B. Ward and R. C. Williamson. Particle filtering beamforming for acoustic source localization in a reverberant environment. In IEEE, editor, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume II, pages 1777–1780, 2002.

[23] E. Weinstein, K. Steele, A. Agarwal, and J. Glass. Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces. MIT/LCS Technical Memo MIT-LCS-TM-642, MIT, 2004.

[24] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno. Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory. In IEEE, editor, *Proc. of IEEE-RAS International Conference on Robots and Automation (ICRA-2004)*, pages 1517–1523, 2004.

[25] 中島 弘史. 不定項を利用した平均サイドローブエネルギー最小ビームフォーミングの実現. **日本音響学会誌**, 62(10):726–737, 2006.

視聴覚情報統合及び EM アルゴリズムを用いた人物追跡システム実現

Real-Time Auditory and Visual Talker tracking through an EM algorithm

*金 鉉燉, 駒谷 和範, 尾形 哲也, 奥乃 博
*Hyun-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno
京都大学　大学院情報学研究科　知能情報学専攻
{hyundon, komatani, ogata, and okuno}@kuis.kyoto-u.ac.jp

Abstract— This paper presents techniques that enable a talker tracking for effective human-robot interaction. We propose a way to use an EM algorithm to select an appropriate path for tracking a talker. The proposed algorithm is simple because it contains relatively few conditional statements. Moreover, the proposed way can easily adapt new kinds of information for tracking talker to our system. This is because our system estimates the position of a desired talker through means, variances, and weights calculated from EM training regardless of the number and kinds of information. In addition, to enhance a robot's ability to track a talker in real-world environments, we applied a particle filter to the talker tracking after performing EM algorithm. Besides, we have integrated a variety of auditory and visual information regarding sound localization, face localization, and lip movement detection. Notably, we have applied a sound classification function that allows our system to distinguish between voice, music, or noise. Also, we developed a vision module that can locate moving objects.

1. INTRODUCITION

In the near future, we expect the participation of intelligent robots in human society to grow rapidly. Therefore, since effective interaction between robots and the average person will be essential, robots should identify people in social and domestic environments, pay attention to the voices of people and look at speakers to identify them visually and associate voice and visual images so as to robustly realize interaction between the robot and the desired person [1-4]. To cope with the rapidly changing circumstances and technology related to robots, robots should be able to easily adapt themselves to new environments and technologies. For example, people have recently taken interest in the possibilities offered by remote control and information exchange between robots or robots and various electronic appliances; that is, what is called *ubiquitous* environment. For such applications, robots receiving information regarding new circumstances will need to apply themselves to these circumstances without the assistance of robot experts or developers. That is, the installed software of robots should be flexible enough that programmers do not have to modify the program or the algorithm whenever robots encounter new conditions.

The objective of this research has been to develop techniques that enable a talker tracking for effective human-robot interaction. Recently, Nakadai et al. developed real-time auditory and visual multiple-talker tracking technology [1, 2]. However, the program of this system has many conditional statements to enable multiple-talker tracking. Specifically, this system has auditory, vision, and motor modules and generates a stream through events extracted by each module. And streams can be associated in a pair of auditory and visual streams to create a higher level stream called an associated stream. Unfortunately, this algorithm needs many conditional statements to create associated streams because the system has to compare for every stream that differs from the others. Moreover, if an event of an entirely new kind is applied to the system, the program structure has to modify all parts of the algorithm related to streams and associations. For this reason, the program is complex and difficult to modify for changed conditions. We propose a way to select an appreciate path for tracking a talker from various events through an expectation maximization (EM) algorithm [5]. Our method is simple because many conditional statements are not needed to create associated streams and is flexible because newly added streams can be easily applied to the system without modification of an entire algorithm. Moreover, to obtain the reliable tracking path, we added a particle filter [12] to a tracking process after performing EM algorithm. That can help the robot to track a designated talker continuously.

Besides, our auditory system includes a sound classification module that can distinguish between voice, music and noise to enable reliable talker tracking in real environments. To realize this, we used a Gaussian mixture model (GMM). To make up for the fact that a face detection module cannot detect a face which is turned away or tilted, we also developed a module to locate moving objects. The vision system can detect lip movement to identifying a talker.

2. DESIGN OF SYSTEM

2.1 Robot Hardware

As a test of real-time talker tracking, we use a humanoid robot called SIG2 (Figure 1). SIG2 has two omni-directional microphones inside humanoid ears at the left and right ear position, its head and body respectively have three degree of freedom (DOF) and one DOF, each of which is enabled by a DC motor controlled by an encoder sensor. SIG2 is equipped with a pair of CCD cameras, but the current vision module uses only one camera.

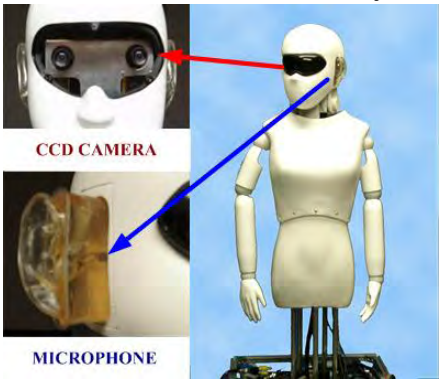


Fig. 1. SIG2

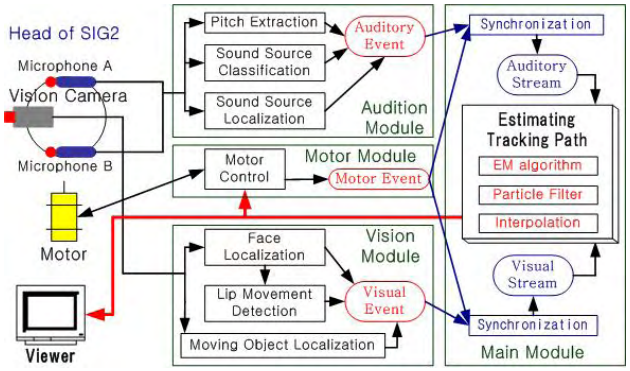


Fig. 2. System Overview

2.2 Design of Subsystems

Figure 2 shows the structure of the system based on a client/server model. Our system consists of four client modules (auditory, vision, motor, and viewer) and a server module (main). Each client controls the following modules:

1) Auditory: Generates auditory events through pitch extraction, sound source localization, and sound source classification. In particular, an auditory module can discriminate among three classes (voice, music, and noise). The sampling frequency is 16 kHz, the processing time for each event generation is 32 ms, and 4.5 frames (one frame consists of 1024 samples) must be calculated for sound source classification.

2) Vision: Generates vision events through face localization, lip movement detection, and moving object localization. The processing time for each event generation is up to 250 ms.

3) Motor: Generates motor events and controls motors for a talker tacking. The time needed for each event generation is 100 ms.

4) Viewer: Displays various streams, result data, and the tracking status.

2.3 Design of Main System

The main (server) module can currently create four streams (sound, face, moving object, and motor) using events extracted by the subsystems. Beyond that, to track the desired talker among a group of people in a noisy environment, the main module estimates an appropriate tracking path through the EM and particle filter.

1) Stream Formation: The server firstly synchronizes the events provided by other modules. A motor event is used in synchronization between the current motor position and the horizontal angle of localization extracted from auditory and vision events over time. This process is important for tracking a talker because it can use an absolute coordinate regardless of which way the robot's head is turned. After that, an auditory event is connected to the nearest auditory stream within $\pm 15°$ with a common pitch. Each auditory stream can be classified according to three sound classes (voice, music, and noise). A visual event is connected to the nearest visual stream within $\pm 5°$. For a visual stream, there is a face stream, which includes the status of lip movement detection, and an object moving stream. If any appropriate stream is found, such an event becomes a new stream. If no event is connected to an existing stream within 1 sec, the stream terminates.

2) Estimating A Tracking Path: In a conventional system, a pair of auditory and visual streams can be associated to enable robust tracking of multiple objects. However, the stream association depends on many conditional statements of the algorithm. To avoid problem, we use EM algorithm that allows a robot to classify the range for a tracking path among a lot of events or streams. Then, a particle filter helps it to estimate the reliable path from the classified range in order to track a designated talker continuously. Finally, for generating a smooth motion when turning a head's motor, we applied an interpolation method using Bezier curve to the talker tracking.

3. AUDITORY SYSTEM

3.1 Sound Source Localization

We use a CSP method for sound source localization. For the purpose of multiple sound localizations, after we calculate CSP at every 0.5 frame (one frame consists of

1024 samples) for 4.5 frames, we can estimate the multiple directions of sounds. At that time, we assume that more than two sounds will not simultaneously enter the microphones with the same magnitude because CSP cannot detect sounds from multiple directions at the same time.

1) Cross-Power Spectrum Phase: The direction of the sound source can be obtained by estimating the time delay of arrival (TDOA) between two microphones [6]. When there is a single sound source, the TDOA can be estimated by finding the maximum value of the cross-power spectrum phase (CSP) coefficients [7], as derived from

$$csp_{ij}(k) = IFFT\left[\frac{FFT[s_i(n)]FFT[s_j(n)]^*}{\left|FFT[s_i(n)]\right|\left|FFT[s_j(n)]\right|}\right] \quad (1)$$

$$\tau = \arg\max\left(CSP_{ij}(k)\right) \quad (2)$$

where k and n are time delays, FFT (or IFFT) is the fast Fourier transform (or inverse FFT), * is the complex conjugate, and τ is the estimated TDOA. The sound source direction is derived from

$$\theta = \cos^{-1}\left(\frac{v\cdot\tau}{d_{\max}\cdot F_s}\right) \quad (3)$$

where θ is the sound direction, v is the sound propagation speed, F_s is the sampling frequency, and d_{max} is the distance with a maximum time delay between two microphones.

3.2 Audio Feature Analysis

Our system can classify three types of sounds (voice, music and noise) for talker tracking in a real environment. It uses four auditory features (pitch, SF, MFCC, and sound localization), and needs to calculate a period of 4.5 frames for sound classification.

1) Pitch Extraction: "Cepstrum" means the signals made by inverse Fourier transform of the logarithm of Fourier transform of sampled signals. One of the most important features of the cepstrum is that if a signal is periodic, the cepstrum will present peaks at intervals for each period [8]. Therefore, the cepstrum can reliably extract the pitch of a speech signal. Given a signal $x(\omega)$, the equation of the cepstrum is denoted as

$$c_c(\tau) = IFFT\left\{\log\left|x(\omega)\right|\right\} \quad (4)$$

In the sequence to extract pitch signals, we first apply a Hamming window to the sampled signals to minimize frequency leakage effects. Then, after performing fast Fourier transform (FFT), it performs inverse fast Fourier transform (IFFT) of the logarithm of these signals. Finally, when the number of samples between two peak signals is found, the pitch can be detected by:

$$Pitch = \frac{\text{Sampling Frequency}}{\text{Number of samples between the two peaks}} \quad (5)$$

*2) Spectrum Flux:*Spectrum flux (SF) is the average variation value of the spectrum between two adjacent frames [9]. SF is denoted as

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{n=1}^{K-1}\left[\log\left(A(n,k)\right)-\log\left(A(n-1,k)\right)\right]^2 \quad (6)$$

where *A(n,k)* is the discrete Fourier transform of the n-*th* frame of the input signal, *N* is the total number of frames, and *K* is the order of FFT. In our experiments, we found that, in general, the SF values of voice are higher than those of music or noise. Therefore, SF is a good feature for classifying speech signals. This feature is used to discriminate between speech and non-speech.

3) Mel Frequency Cepstral Coefficients: There are two dominant types of acoustic measurement of a speech signal for the feature extraction of speech. One is the parametric approach, which was developed to match closely the resonant structure of the human vocal tract that produces the corresponding speech sound. It is mainly derived from linear predictive analysis, such as LPC-based cepstrum (LPCC). The other is the non-parametric method which models the human auditory perception system. Mel frequency cepstral coefficients (MFCCs) are used for this purpose [10]. Here, we used the 0 to 12th MFCCs. MFCC provides good information useful for discriminating between speech and non-speech. Usually, the 0 to 12th MFCCs of speech signals have different patterns for speech, music, or noise respectively.

3.3 Sound Source Classification by GMM

Figure 3 shows the processing flow for classifying sound signals by auditory features. First, we apply each feature data extracted from the cepstrum, MFCC, SF, or CSP to mean defined as (7) and covariance defined as (8)

$$\mu = \frac{1}{M}\sum_{m=1}^{M} X_m \quad (7)$$

$$\sigma^2 = \frac{1}{M}\sum_{m=1}^{M}\left(X_m-\mu\right)^2 \quad (8)$$

where X_m is data, μ is the mean, σ is the variance, and M is the number of data. Then, for the cepstrum, SF and CSP, if the values calculated from the mean and the covariance are within the boundary of those of speech signals, each final feature value, $f_t^{2\sim4}$, will have the re-

sulting value as a speech signal.

When using MFCCs, we apply the 0 to 12th MFCCs to Gaussian Mixture Model (GMM) defined by (9) and the weight as denoted by (10). The GMM is a powerful statistical method widely used for speech classification [11].

$$P_{mixture}\left(X_{0\sim12}|\theta_{0\sim12}\right)=\sum_{L=0}^{12}P_L\left(X_L|\theta_L\right)w(L) \tag{9}$$

$$\sum_{L=1}^{L+1}w(L)=1,\quad 0\le w(L)\le 1 \tag{10}$$

where P is the component density function, L is the number of the MFCC order, X is the 0 to 12th MFCC data, and θ is the parameter vector concerning each MFCC. Moreover, to classify speech signals robustly, we designed two GMM models for speech and noise derived as

$$f_t^1=\log\left(P_s\left(x_s(t)|\Theta_s\right)\right)-\log\left(P_n\left(x_n(t)|\Theta_n\right)\right) \tag{11}$$

where P_s is the GMM related to speech, and $X_s(t)$ is the speech feature data set at the t-*th* frame belonging to the speech parameters, Θ_s. On the other hand, P_n is the GMM related to noise and $X_n(t)$ is the noise feature data set at the t-*th* frame belonging to the noise parameters, Θ_n. Finally, all final feature values, $f_t^{1\sim4}$, that have appropriate weights, $w^{1\sim4}$, are combined to judge whether the frame is voice, music or noise. To train the GMM parameter, we used 30 speech datum (15 male and 15 female), 15 noise datum (white, brown, pink, and clapping), and 15 music datum (normal pop music excluding vocals).

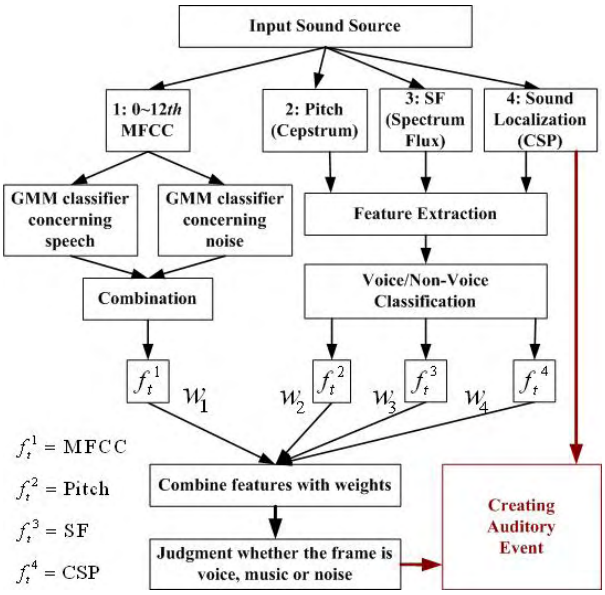


Fig. 3. Sound source classification by auditory features

4. VISION SYSTEM

4.1 Face Localization by OpenCV

For the purpose of detecting human faces, we used open computer vision (OpenCV), the open source vision library created by the Intel Company. This library supplies functions for detecting human faces. Therefore, we can get the number and the coordinates of the detected faces through OpenCV [4]. Our system used 320 x 240 images and can calculate about four images per second.

4.2 Lip Movement Detection

We achieve lip movement detection using an OpticalFlow function in OpenCV. This function can detect a variation between a former picture and a present one. Therefore, our system can accurately distinguish when a speaker is talking.

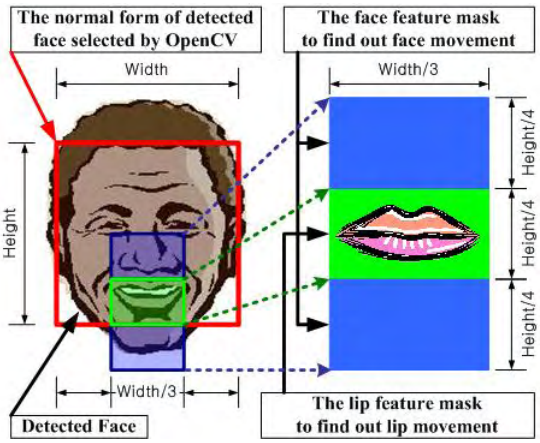


Fig. 4. Feature mask for detecting lip movement

Figure 4 shows feature masks applied to the area of detected faces to detect lip movement. If the amount of variation detected by the lip feature mask is large, the system will infer a person is talking. However, if the amount of a variation detected by the face feature mask exceeds that detected by the lip feature mask, the system will regard the person, who is not talking, because the amount of a variation detected by the feature mask is also increased when a person swings his face. This prevents misdetection for detecting lip movement when a face is moving.

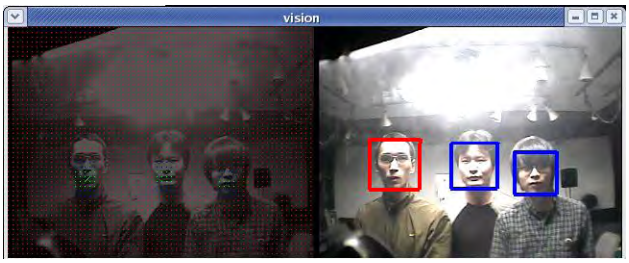


Fig. 5. Results of face detection and finding the talker

Figure 5 shows the results of face and lip movement detection among three people. In this picture, blue boxes indicate detected faces and a left person, whose box has turned red, is selected as the talker because lip movement is detected among the detected faces. The left part of Figure 5 shows the applied feature masks.

4.3 Moving Object Localization

OpenCV has some limitations. First, it cannot determine a face over 2m away by using 320 x 240 images. Second, it cannot detect a face which is turned away or tilted. Consequently, a person must be looking straight at the camera. To overcome these shortcomings, we developed a function for moving object localization by using OpticalFlow. It can infer a moving object's position from the position where the value calculated by OpticalFlow is high. Therefore, if faces are not detected although people are in front of the camera, it can obtain the positions of people when they are moving. The right side of Figure 6 shows an image captured by SIG's camera, and the red lines in the left part of Figure 6 show how different objects are moved between a former image and a present one.

Fig. 6. Moving object localization by OpticalFlow

5. TALKER TRACKING SYSTEM

For the purpose of tracking a desired talker, we should first estimate the appropriate tracking path. Therefore, we applied an EM algorithm in this process [5], which allows us to easily obtain the range of direction for tracking from among various streams. However, if a robot gets several sound streams that have the same condition or weight, it will be difficult to maintain the designated path which a robot has tracked. Therefore, we also proposed the way to add the particle filter to the tracking process after performing EM algorithm. The applied particle filter helps the robot to track the designated path continuously even if the condition and weight of streams or events is the same and the distance between those is also close. Moreover, we applied interpolation method using Bezier curve to the final process of a tracking algorithm so that the robot has the smooth motion when motor is rotating.

Figure 7 shows the process to select an appropriate tracking path by the EM and particle filter algorithm. In Figure 7, for simplicity there are just two kinds of stream (sound stream and vision stream) and four Gaussian mixture components for EM training. However, our system actually has three kinds of streams (sound, face, and moving object localization) and uses eight Gaussian mixture components for training. Section 5.1 to 5.4 describe this processing in detail.

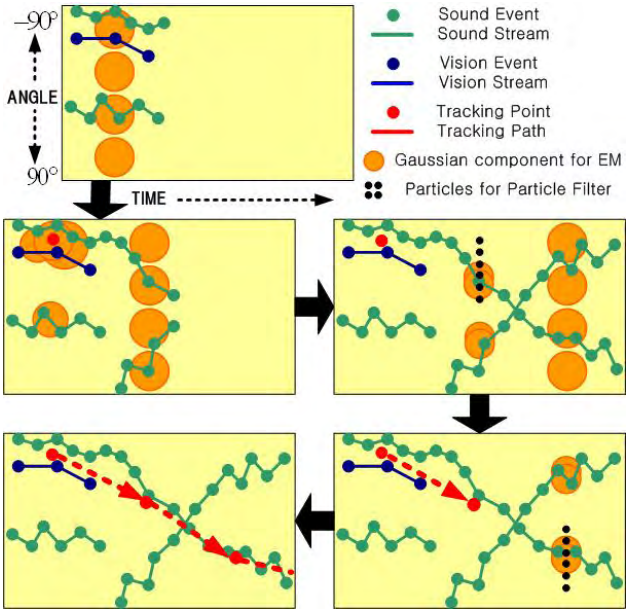


Fig. 7. Process to select the path for talker tracking

5.1 Arranging the Gaussian mixture components

First, according to the conditions of the streams, the system increases the number of events. For example, for sound events extracted from a voice or face events when lip movement is detected, it increases the actual number of events by 3 to 4 times so that the Gaussian components for EM training are gathered near the area of the stream that has the highest priority. After that, the set of increased datum, X_m, are substituted for a one-dimensional Gaussian mixture which is denoted as

$$P\left(X_m \middle| \mu_k, \sigma_k\right) = P\left(X_m \middle| \theta_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(X_m-\mu_k)^2}{2\sigma_k^2}} \quad (12)$$

where μ_k is the mean, σ_k^2 is the variance, θ_k is a parameter vector, and k is the number of mixture components. The objective is to find the parameter vector θ_k describing each component density $P\left(X_m \middle| \theta_k\right)$.

Second, for EM training (iteration), eight Gaussian components are located between -90° and 90° at 1 sec intervals and the interval to run EM algorithm also shifts every 100 ms. At that time, if the coordinates of the robot's head change, the position to locate components will also change corresponding to the coordinates of the motor. This step is shown in the top of Figure 7.

5.2 Performing the EM algorithm

After locating the Gaussian components, the system runs the E-step and M-step for less 10 iterations. This EM step in detail is as follows.

1) E-step: The expectation step essentially computes the expected values of the indicators $P(\theta_k|X_m)$ that each data point X_m was generated by component k, given N is the number of mixture component, the current parameter estimates θ_k and weight w_k, using Bayes' Rule derived as

$$P(\theta_k|X_m)=\frac{P(X_m|\theta_k)\cdot w_k}{\sum_{k=1}^{N}P(X_m|\theta_k)\cdot w_k} \tag{13}$$

2) M-step: At the maximization step, we can compute the cluster parameters that maximize the likelihood of the data assuming that the current data distribution is correct. Accordingly, we obtain the recomputed mean using (14), the recomputed variance using (15), and the recomputed mixture proportions (weight) using (16).

$$\mu_k=\frac{\sum_{m=1}^{M}P(\theta_k|X_m)\cdot X_m}{\sum_{m=1}^{M}P(\theta_k|X_m)} \tag{14}$$

$$\sigma_k^2=\frac{\sum_{m=1}^{M}P(\theta_k|X_m)\cdot(X_m-\mu_k)^2}{\sum_{m=1}^{M}P(\theta_k|X_m)} \tag{15}$$

$$w_k=\frac{1}{N}\sum_{m=1}^{M}P(\theta_k|X_m) \tag{16}$$

where M is the total number of data.

Finally, we can obtain the estimated mean, variance and weight corresponding to the current data distribution if the E and M steps are iterated an adequate number of times.

Consequently, Gaussian mixture components are relocated around the streams and the components are mainly concentrated where the density of streams is high. In addition, the mean, variance, and weight of components are decided according to, respectively, the location value, the distributional range, and the priority of streams. Also, the start point for a talker tracking is determined where the weight calculated by EM is the highest. This step is shown in the middle-left part of Figure 7.

5.3 Particle Filter Implementation

For the purpose of obtaining the reliable tracking path, we added the particle filter [12] to the tracking process after performing EM algorithm. Therefore, particle filter can help a robot to maintain the designated tracking path which the robot has tracked regardless of changing the location of talkers. The detail process of applied particle filter is as follows. First, we can estimate a present tracking position by former tracking positions. This model is defined as

$$x_{t+1}^i=x_t^i+\dot{x}_t^i\cdot T_s+\ddot{x}_t^i\frac{T_s^2}{2} \tag{17}$$

where i is the number of particle, T_s is a sample period, x^i_{t+1} is a estimated tracking position, x^i_t is the former tracking position, $\dot{x}^i_t$ is the differential value between x^i_t and x^i_{t-1}, and the differential value between $\dot{x}^i_t$ and $\dot{x}^i_{t-1}$ is $\ddot{x}^i_t$. Then, the particle filter spreads particles in the range of $\pm 15°$ of the estimated position. This step is shown in the middle-right part of Figure 7.

Second, it calculates the equation (18) by using result values (mean, variance, and weight) calculated by EM algorithm and then it should iterate the update routine until the condition of resample is satisfied. The equation (19) defines the condition of resample. Then, tracking points can be determined as you see the bottom-right part of Figure 7. The iteration routine of this particle filter in detail is as follows.

1) Measurement update: Update the weights by the likelihood:

$$\omega_t^i=\omega_{t-1}^i P(\theta_t^i|x_t^i)=\omega_{t-1}^i\frac{P(x_t^i|\theta_t^i)\cdot w_t^i}{\sum_{i=1}^{N}P(x_t^i|\theta_t^i)\cdot w_t^i} \tag{18}$$

$i=1,2,\mathrm{K},N$ and normalize to $\omega_t^i:=\omega_t^i\left[\sum_{i=1}^{N}\omega_t^i\right]^{-1}$

As an approximation to, take $x_t\approx\sum_{i=1}^{N}\omega_t^i x_t^i$

2) Re-sampling:

(a) Bayesian bootstrap: Take N samples with replacement from the set $\{x_t^i\}_{i=0}^N$ where the probability to take sample i is ω_t^i. Let $\omega_t^i=1/N$. This step is also called Sampling Importance Re-sampling (SIR).

(b) Importance sampling: Only resample as above when the effective number of samples is less than a threshold N_{th},

$$N_{eff}=\frac{1}{\sum_{i=0}^{N}(\omega_t^i)^2}<N_{th} \tag{19}$$

Here $1 \le N_{eff} \le N$, where the upper bound is attained when all particles have the same weight, and the lower bound when all probability mass is at one particle. The threshold can be chosen as $N_{th} = 2N/3$. Let $t:=t+1$ and iterate to measurement update.

5.4 Estimating the Tracking Path

Finally, the desired tracking path can be determined by iterating EM and particle filter according to time. Therefore, as you see the bottom-left part of Figure 7, although the classified areas of streams have the same condition or the paths of streams are even crossed each other, it is able to estimate a reliable tracking path. Moreover, it is necessary to produce a smooth path from estimated tracking points in order to turn motor smoothly. For this step, we used the interpolation method through Bezier curve. The Bezier parametric curve is given by $B(u)$ as follows:

$$B(u) = \sum_{k=0}^{N} P_k \cdot \frac{N!}{k!(N-k)!} \cdot u^k \cdot (1-u)^{N-k} \qquad 0 \le u \le 1 \qquad (20)$$

where the number of data is N and control points P_k with $k=0$ to N.

Consequently, the robot can have the tracking path without an oscillation. Figure 8 shows that the real path of a motor is converted to the smooth path generated by interpolation through Bezier curve.

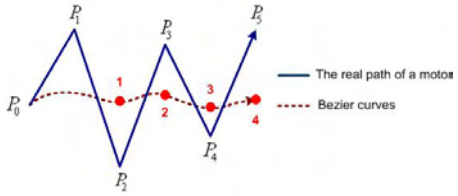


Fig. 8. Interpolation through Bezier curve.

6. EXPERIMENTS AND EVALUATION

As you see Figure 9, a viewer module displays various streams, the current position of a motor, and the results and status of a talker tracking received from main module. The red rings indicate the path selected for talker tracking by the EM algorithm and the particle filter. To realize the reliable talker tracking in a real environment, the proposed system was designed with the following points in mind.

1) The sound stream created from noise is rejected for tracking. Besides, when there are only moving object streams, they are also rejected for tracking. However, the sound stream created from voice and the face stream create from a face localization are accepted.
2) If two sound streams created from voice occur at the same time, the system will select the sound stream where vision information (face and moving object) exists nearby. Also, when a sound stream created from music or noise occurs with a face stream, the system selects the face stream.

Figure 9 shows that the robot is actually turning its head towards the direction of a path selected by the EM and paticle filter algorithm. The pink area indicates the visibility range of the vision camera and the center of the area indicates the position of the rotation motor of a head. In (A) of Figure 9, we can see that a designated tracking path, that was first started compared to another path, is continuously selected by our proposed way even if sound streams exist at the same time. However, if vision stream appears on another area which did not belong to the tracking path, the tracking path will be changed. This is because vision information is usually more reliable than audio information. In (B) of Figure 9, we can see that although the paths of streams are crossed each other, it is able to maintain the tracking path which was first started. (C) of Figure 9 shows the talker tracking with all kinds of stream. Needless to say, the area including all kinds of stream has top priority when tracking the talker. Therefore, the area is always determined as the tracking path.

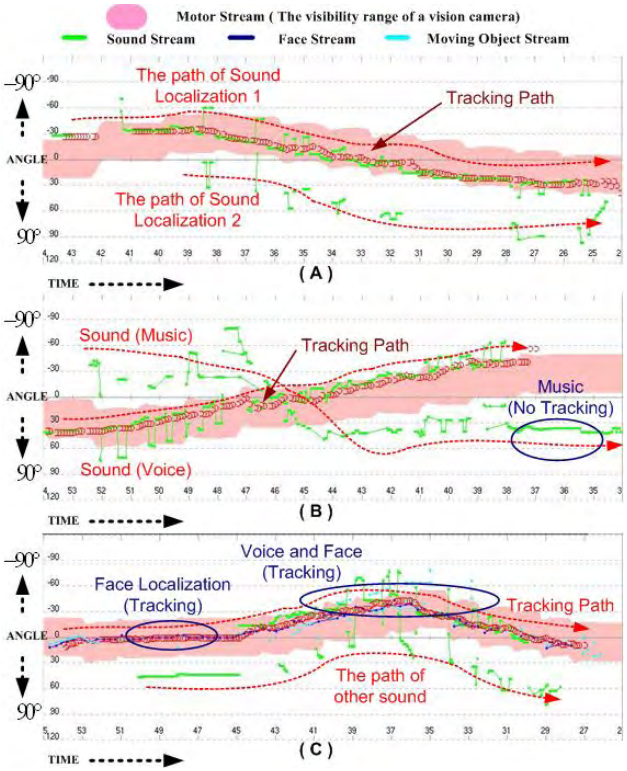


Fig. 9. Results of talker-tracking experiments

7. CONCLUSION AND FUTURE WORK

We have described a way to use an EM and particle filter algorithm to select an appropriate tracking path for

the purpose of tracking a talker. Our system based on this approach has some principal merits. First, the proposed algorithm is simple because it contains relatively few conditional statements. It is also not necessary to associate streams, unlike the conventional system, because our system can easily infer the distributional range of streams from the calculated variance by EM. Second, although developers do not modify the entire algorithm, the proposed system can easily adapt to new kinds of events or streams to the tracking system. Since this system estimates the position of a desired talker through means, variances, and weights calculated from EM training regardless of the number and kinds of event and stream, they only determine the initial condition according to the priority of the new event or stream. Finally, to produce the reliable tracking path, we added the particle filter to the tracking process after performing EM algorithm. Particle filter can help it to maintain the designated tracking path which the robot has tracked regardless of changing the condition of streams and the position of a tracking path.

To realize real-time auditory and visual talker tracking in practical environments, though, we need to refine our system. First, we plan to develop a system that can track a group of talkers in practical environments. Therefore, sound identification and face recognition will be necessary. And reliable multiple sound source localization will be also necessary. In this respect, we are considering a way to integrate the good points concerning several methods for sound source localization. Second, to realize a practical active auditory system, we need to add speech recognition and voice synthesis function to our system so that it will be able to talk with humans. In addition, we will add sound source separation to our system so that it can dealing with various sound signals.

ACKNOWLEDGMENT

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research, and COE program of MEXT, Japan.

REFERENCES

[1] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi Mizoguchi, Hiroshi G. Okuno, and Hiroaki Kitano, “Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids,” in Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI-01), pp. 1425-1432, Seattle, Aug. 2001.

[2] Hiroshi G. Okuno, Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi Mizoguchi, and Hiroaki Kitano, “Human-Robot Interaction Through Real-Time Auditory and Visual Multiple-Talker Tracking,” in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001), pp. 1402-1409, Oct. 2001.

[3] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration," in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, Washington DC USA, May 2002, pp. 1043-1049.

[4] H. D. Kim, J. S. Choi, and M. S. Kim, "Speaker localization among multi-faces in noisy environment by audio-visual integration", in Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA2006), pp. 1305-1310, May, 2006.

[5] T. K. Moon. “The Expectation-Maximization algorithm,” IEEE Signal Processing Magazine, 13(6) pp. 47-60, Nov. 1996.

[6] H. D. Kim, J. S. Choi, C. H. Lee, and M. S. Kim, “Reliable Detection of Sound’s Direction for Human Robot Interaction,” in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, Sendai Japan, Sep. 2004, pp.2411-2416.

[7] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, “Localization of multiple sound sources based on a CSP analysis with a microphone array,” IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, June, 2000, pp 1053-1056.

[8] H. Kobayashi, and T. Shimamura, "A Modified Cepstrum Method for Pitch Extraction,” IEEE/APCCAS Int. Conf. Circuits and Systems, Nov. 1988, pp. 299-302.

[9] L. Lu, H. J. Zhang, and H. Jiang, “Content Analysis for Audio Classification and Segmentation,” IEEE Trans. on Speech and Audio Processing, vol. 10, no 7, pp. 504-516, 2002.

[10] J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantormo, “Robust Voiced/Unvoiced classification using novel feature and Gaussian Mixture Model,” IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, Montreal, Canada, May, 2004.

[11] M. Bahoura and C. Pelletier, “Respiratory Sound Classification using Cepstral Analysis and Gaussian Mixture Models,” IEEE/EMBS Int. Conf., San Francisco, USA, Sep. 1-5, 2004.

[12] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P. Nordlund, “Particle Filters for Positioning, Navigation and Tracking,” IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 50, no 2, pp. 425-437, Feb. 2002.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0624-9 (11/17)

逐次的な位相差補正処理を特徴とする音源定位方式:SPIRE

A sound source localization method named SPIRE (Stepwise Phase dIfference REstoration)

戸上真人, 住吉貴志, 神田直之, 天野明雄

Masahito TOGAMI, Takahshi SUMIYOSHI, Naoyuki KANDA, and Akio AMANO

（株）日立製作所 中央研究所

Central Research Laboratory, Hitachi Ltd.

{masahito.togami.fe, takashi.sumiyoshi.bf, naoyuki.kanda.kn, akio.amano.qb}@hitachi.com

Abstract

We propose a new methodology of sound source localization named **SPIRE** (Stepwise Phase dIfference REstoration) that is able to localize sources even if they are neighboring in a reverberant environment. The major feature of our proposed method is restoration of a microphone pair's phase difference (M1) by using the phase difference of another microphone pair (M2) under the condition that the distance between M1's microphones is longer than the distance between M2's microphones. This restoration process makes it possible to reduce the variance of an estimated sound source direction and to solve the spatial aliasing problem that occurs with the M2's phase difference. The experimental results in a reverberant environment (reverberation time = about 300ms) indicate that our proposed method can localize sources even if they are neighboring (even if the difference in the sources' directions equals 10 degree).

1 はじめに

音源定位技術とは、複数のマイクロホン素子を有するマイクロホンアレイを用いて、マイク素子間の位相差や振幅差から音の到来方向を推定する技術である。

話者方向を推定して、その方向に振り向いたり、所望方向の発話内容を聞き分けたりするロボット[1][2]や、話者方向に自動的にカメラを向けるテレビ会議システムなどで、高性能な音源定位技術が必要とされている。また雑音環境での遠隔音声認識のための前処理としても音源定位技術が使われることが多い。我々は、愛知万博でデモを行った人間共生ロボット EMIEW(Excellent Mobility and Interactive Existence as Workmate) に、音源定位機能や雑音下音声認識機能を搭載した[2]。

従来の音源定位法として、マイク間の相互相関値から音源方向を推定する加算型の音源定位法である遅延和アレイ法[3]が提案されている。遅延和アレイ法は、位相の遅延及び遅延信号の加算のみの単純な構成であるが、少数のマイクロホンでは複数音源が存在する場合の定位性能が悪いため、多数のマイクロホンを必要とする。またアレイ長が大きくなるという問題がある。そこで、最小分散ビームフォーマの分離フィルタを定位に応用した手法[3]や、入力相関行列の雑音に由来する固有ベクトルと各音源のステアリングベクトルとが直交することを利用した MUSIC 法[7]などの少数のマイクロホン素子で複数音源を定位することを目的とした高精度音源定位法が提案されている。中でも MUSIC 法は特に精度が高いが、音源数を予め知っているか別途推定する必要があり、誤った音源数を設定した場合、MUSIC 法の性能は大きく劣化する。高負荷な固有値計算を伴うため計算量が多いという問題がある。また仮想方向のステアリングベクトルを参照し、雑音に由来する固有ベクトルと比較して音源方向を求めるため、音源方向の探索分解能に処理量が依存する。

近年、スパース性と呼ばれる音声の性質を利用した音源定位法が検討されている[2][4][5][6]。音声は、短時間で見ると、少数の周波数からなるスパースな信号であり、複数の音源が同じ時間-周波数成分に混合することは稀である[4]。これらの手法はこの音声のスパース性に基づき、各時間-周波数毎に一つの音源のみ存在すると仮定し、その音源方向を推定する。そして推定した音源方向に該当する時間-周波数成分を振り分けて、ヒストグラムを生成し、ヒストグラムのピークを音源方向として検出する。マイク間位相差や振幅差から直接音源方向を推定する手法[4][5][6]では、音源方向の探索分解能に処理量が依存することは無いが、空間エイリアシング[8]の問題から、用いるマイ

クペアのマイク間隔全てが信号の最大周波数から決まる間隔（エイリアシング距離）以下である必要がある。しかし短いマイク間隔では位相差がばらつく場合の定位性能が劣化する問題があり、特に残響環境のように位相差のばらつきが大きい環境では性能劣化は大きくなる。修正遅延和アレイ法[2]は、スパース性の仮定の下、遅延和アレイ法を修正した手法である。修正遅延和アレイ法では、一つのマイク間隔がエイリアシング距離以下であればよく、全てのマイク間隔がエイリアシング距離以下である必要は無いため、マイク間隔に対する制約は少ない。また MUSIC 法と比較して固有値計算などの高負荷の計算を伴わない低処理で高精度な音源定位法である。しかし MUSIC 法と同様に音源方向の探索分解能に処理量が依存するため、高分解能な探索は難しいという問題があった。

本稿ではマイク間隔が狭いマイクペアの位相差を用いて、空間エイリアシングが生じるマイク間隔がマイクペアの位相差を補正することを特徴とする、スパース性に基づく音源定位法:SPIRE (Stepwise Phase dIfference REstoration)　法を提案する。SPIRE では、用いるマイクペアのうち少なくとも 1 つが空間エイリアシングを満たしていれば良く、その他のマイク間隔を広くとることができるため、残響環境などの位相差がばらつく環境での性能を向上させることができる。補正した位相差から直接音源方向を算出可能であるため、音源方向の探索分解能に処理量が依存しない。また固有値計算などの高負荷な処理を必要としないため、低処理量である。提案手法の有効性を残響時間 300ms の実環境での近接複数話者音源定位実験にて示す。

2　音源定位の問題設定

音源からマイクロホンアレイまでの伝達・混合過程モデル及びマイク間位相差と音源方向の関係について述べる。

2.1　座標系定義

マイクロホン及び音源の座標系を (Figure 1) に示す。

2.2　伝達・混合過程モデル

音源数 N, マイク素子数 M とする。j 番目の音源の原信号を $s_j(t)$ とする。i 番目のマイクの入力信号を $x_i(t)$ とする。j 番目の音源から i 番目のマイクまでの直接音のインパルス応答を $h_{j,i}(t)$, インパルス応答長を Imp とする。$n_i(t)$ を無指向性の背景雑音だと仮定する。$n_i(t)$ はマイク間で独立とする。i 番目のマイクの入力信号 $x_i(t)$ は、

$$x_i(t) = \sum_{j=0}^{N} \sum_{\tau=0}^{Imp} h_{j,i}(\tau) s_j(t-\tau) + n_i(t) \qquad (1)$$

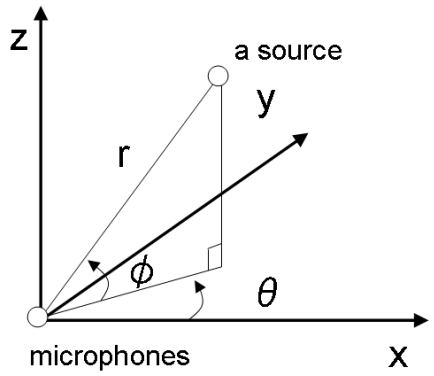


Figure 1: 座標系: r はマイクロホンアレイから音源までの距離, θ は方位角, and ϕ は仰角

となる。時間領域の畳み込み混合で表される伝達・混合過程モデルを短時間フーリエ変換で以下のように近似する。

$$x_i(f,\tau) \approx \sum_{j=0}^{N} h_{j,i}(f) s_j(f,\tau) + n_i(f,\tau) \qquad (2)$$

ここで、τ は短時間フーリエ変換のフレームインデックスであり、f は周波数ビン・インデックスである。

2.3　位相差と音源方向の関係

マイク間位相差から音源方向を推定可能であることを示す。$N = 1$ かつ無残響・無背景雑音、音源仰角 $\phi = 0$ とし、音源とマイクロホンアレイの距離が十分長く、到来波が平面波とみなせるとする。マイク間隔 d のマイクロホンペアが y 軸上にあるとすると、このマイクロホンペアの周波数 f の位相差 σ と音源方位角 θ は以下の関係にある。

$$\sin\theta = \frac{\sigma}{2\pi f dc^{-1}}, \qquad (3)$$

ここで c は音速である。マイクロホンペアの位相差 σ から音源方位の推定値 $\hat{\theta}$ を

$$\hat{\theta} = arcsin \frac{\sigma}{2\pi f dc^{-1}} \qquad (4)$$

で求めることができる。

3　従来手法

従来のスパース性に基づく音源定位法のアルゴリズムを示し、その推定精度とマイク間隔の関係について述べる。スパース性に基づく音源定位法では、マイク間隔が広いほど、音源定位性能が高まるが、空間的エイリアシングの問題からマイク間隔の上限値が存在するため、音源定位性能にも上限があることを示す。

3.1　音源定位

音声は、短時間で見ると、少数の周波数からなるスパースな信号であり、複数の音源が同一の時間-周波数で混合

することは稀であることが知られている[4]。この性質に基づき、時間-周波数毎に存在する音源は1つであると仮定し、その音源方向を式 (4) を用いてマイクロホン位相差から推定することができる。そして、時間-周波数毎に求めた音源方向のヒストグラムを取り、そのピークを求めることで、音源方向を得ることができる。$\Delta = \frac{2}{L}$ をヒストグラムの分割幅として、L を分割数とする。k はヒストグラムのインデックスである。ヒストグラム $P(k)$ を $k = 0$ から $L-1$ までピークサーチする。そしてピーク k_{peak} より音源方向は

$$\hat{\theta} = arcsin\Big(-1 + k_{peak}\Delta\Big) \tag{5}$$

と推定される。

3.2 音源方向の推定誤差とマイク間隔の関係

ここでは、マイク間隔を広げるほど、音源方向の推定誤差が小さくなることを示す。無指向性の背景雑音 $n_i(f,\tau)$ が無視できないとする。スパース性に基づき、時間-周波数毎に一つの音源のみ存在し、その音源の原信号を $s(f,\tau)$ として、i 番目のマイクロホンまでの直接音のインパルス応答を $h_i(f)$ とする。反響や残響については、厳密には無指向性ではないが、一般的に、直接音と比較して相対的に指向性が弱いため、$n_i(f,\tau)$ に含めるものとする。以下、周波数及びフレームインデックスに関する (f,τ) は省略する。

i 番目のマイクロホンと j 番目のマイクロホンの位相差 $arg\frac{x_i}{x_j}$ は、

$$arg\frac{x_i}{x_j} = arg\frac{s\cdot h_i + n_i}{s\cdot h_j + n_j} \tag{6}$$

$$= arg\frac{h_i}{h_j} + arg(1 + \frac{n_i}{s\cdot h_i}) - arg(1 + \frac{n_j}{s\cdot h_j}) \tag{7}$$

となる。式 (7) の第一項は、

$$arg\frac{h_i}{h_j} = 2\pi f d \sin\theta c^{-1}, \tag{8}$$

となる。d は i 番目のマイクロホンと j 番目のマイクロホンのマイク間隔である。式 (7) の第二項 $arg(1+\frac{n_i}{s\cdot h_i})$ は、音源と無指向性雑音の SNR にのみ依存する項でありマイク間隔や音源方向に依存しない。また n_i と n_j が独立であるから、第三項 $arg(1+\frac{n_j}{s\cdot h_j})$ と第二項は無相関となる。従って、マイク間位相差 $arg\frac{x_i}{x_j}$ の分散はマイク間隔に依存しないことになり、逆に位相差から推定する $\sin\theta$ の分散は、$\frac{1}{d^2}$ に比例することとなる。したがって、マイク間隔 d を広げるほど、$\sin\theta$ の推定値の推定誤差を小さくすることができる。

3.3 空間的エイリアシング

音源方向の推定誤差を小さくするためには、広いマイク間隔が必要となるが、空間的エイリアシングの問題から、音源の最大周波数で決まる距離以上にマイク間隔を広げることができない。f_{max} を音源の最大周波数とする。あるマイクペアのマイク間隔 d が $\frac{c}{2f_{max}}$（エイリアシング距離）を超えた場合、そのマイクペアの位相差のレンジが 2π を上回る。しかし、入力信号から位相差を算出する場合、arg のレンジが 2π であるため、位相差に 2π の整数倍の不定性が存在し、入力信号の位相差から音源方向を推定することができないという問題が生じる[8]。

4 提案する音源定位法

音源方向の推定誤差を抑えるためには、できる限りマイク間隔を広くとる必要がある。しかし音源の最大周波数で決まるエイリアシング距離以上にマイク間隔を設定すると、空間的エイリアシングの問題が生じるため、音源方向を正確に推定することが不可能となる。結果として、音源方向の推定誤差をある程度以上小さくすることができないことになる。特に残響環境などのようにマイク間位相差にばらつきが生じやすい環境では、音源方向の推定性能が十分な性能とならないという問題がある。この問題に対して、提案する **SPIRE** (Stepwise Phase dIfference REstoration) では、エイリアシング距離以上のマイクペアの位相差をエイリアシング距離以下のマイクペアの位相差で補正することで、音源方向の推定誤差を抑えるとともに、エイリアシング距離以上のマイクペアで生じる空間的エイリアシングの問題を解消する。(Figure 2) に提案手法のイメージを示す。提案手法は、複数のマイクペアを狭いマイク間隔から順に用いて、少しずつ音源方向を絞り込んでいく方式と考えることができる。

提案手法は**位相差算出プロセス**と**位相差補正プロセス**の2ステップから成る。位相差算出プロセスでは、各マイクペアの位相差を算出する。この際位相差のレンジは 2π である。位相差補正プロセスは位相差算出プロセスで求めた各マイクペア位相差の 2π の整数倍の不定項を算出し、位相差を補正する。不定項の算出はよりマイク間隔が狭いマイクペアから順に行う。最初に用いるマイクペアのマイク間隔はエイリアシング距離以下に設定する必要がある。その後、一つ前のマイクペアの補正後の位相差を用いて、逐次的に位相差の不定項を求めていく。以降、位相差補正法について述べた後、SPIRE のアルゴリズムを、マイクロホンアレイのマイク配置が直線配置の場合と、非直線配置の場合に分けて説明する。

4.1 直線配置への適用

マイク配置が直線配置の場合の SPIRE による音源定位アルゴリズムを示す。

使用するマイクロホンペアの数を L とする。L 個のマイクペアは、各マイクペアのマイク間隔の昇順でソートされているものとする。$\sigma\hat{}_1$ は 0 と、d_{-1} は 1　とする。

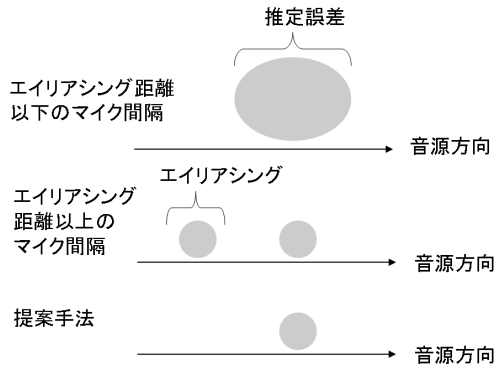


Figure 2: 提案手法のイメージ：上段は、エイリアシング距離より狭いマイク間隔を用いて音源方向を推定した時の推定誤差のイメージである。エイリアシングは生じないが、推定方向の誤差が大きくなる。中段は、エイリアシング距離以上のマイク間隔を用いて音源方向を推定したときのイメージである。マイク間隔を広げることで推定誤差は小さくなるが、エイリアシングが生じてしまい、音源方向を推定不能となる。下段は、提案手法のイメージである。提案手法ではエイリアシング距離より狭いマイク間隔で大まかに音源方向を特定する。そして音源方向がその範囲にあるという制約のもと、エイリアシング距離以上のマイク間隔を用いて、より緻密に音源方向を特定する。

位相差算出プロセス:

i 番目のマイクペアの入力信号をそれぞれ $x_{i,1}$,$x_{i,2}$ と書く。i 番目のマイクペアの位相差 σ_i は以下のように計算される。

$$\sigma_i = arg(\frac{x_{i,1}}{x_{i,2}}). \tag{9}$$

位相差補正プロセス:

ここでは、マイクは (Figure 1) の y 軸上に直線上に並んでおり、音源の仰角 ϕ は 0 とする。i 番目のマイクロホンペアのマイク間隔がエイリアシング距離を上回っている場合、空間的エイリアシングの問題が生じるため、マイク間位相差から音源方向を一つに定めることができない。この場合、未知の不定項 n_i を伴った $\sigma_i + 2n_i\pi$ が真の位相差となる。n_i は i 番目のマイクペアの情報からは求めることができない。そこで、よりマイク間隔の狭い $i-1$ 番目のマイクペアの情報から n_i を求め、i 番目の位相差を補正する。i 番目のマイクペアの補正後位相差 $\hat{\sigma_{i-1}}$ を以下のように求める。

$$\frac{\hat{\sigma_{i-1}} d_i}{d_{i-1}} - \pi \leq \sigma_i + 2n_i\pi \leq \frac{\hat{\sigma_{i-1}} d_i}{d_{i-1}} + \pi. \tag{10}$$

$$\hat{\sigma_i} = \sigma_i + 2n_i\pi. \tag{11}$$

ここで、$V(x)$ を x の分散と定義する。仮に、不定項 n_i が正確に求まった場合、$V(\hat{\sigma_i})$ は $V(\hat{\sigma_{i-1}}\frac{d_i}{d_{i-1}})$ の $\frac{d_{i-1}^2}{d_i^2}$ 倍となり、マイク間隔の 2 乗に反比例して分散が小さくなる。

位相差推定処理と位相差補正処理は、$i=0$ から $i=L-1$ まで実行され、最終的に最も広いマイク間隔の位相差を補正した $\hat{\sigma_{L-1}}$ が得られる。そして $\hat{\sigma_{L-1}}$ より時間 τ、周波数 f 毎の音源方向を以下の式で求める。

$$\hat{\theta} = arcsin\frac{\sigma_{L-1}}{2\pi f d_{L-1} c^{-1}}. \tag{12}$$

全時間-周波数に渡り求めた $\sin\hat{\theta}$ のヒストグラムをピークサーチすることで、音源方向を得ることができる。

4.2 非直線配置への適用

全方向音源定位への応用を考え、非直線配置へ適用できるようアルゴリズムを拡張する。

直線配置では複数のマイクペアを用いるが、非直線配置の場合、複数のサブマイクロホンアレイを用いる。サブマイクロホンアレイの数を U とする。各サブマイクロホンアレイは複数のマイクペアを保持する。サブマイクロホンアレイは各々の最大マイク間隔の昇順でソートされているものとする。L_l を l 番目のサブマイクロホンアレイのマイクペア数とする。位相差算出処理及び位相差補正処理はサブマイクロホンアレイ毎に行う。

位相差算出処理:

$\boldsymbol{p}_i = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ を i 番目のマイクロホンの (Figure 1) で定義される xyz 座標系における位置ベクトルとする。j_1 と j_2 を l 番目のサブマイクロホンアレイの j 番目のマイクペアのマイクロホン番号とする。また $\boldsymbol{d}_j = \boldsymbol{p}_{j_1} - \boldsymbol{p}_{j_2}$ とする。$\boldsymbol{D} = [\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{l-1}]$ とする。各マイクペアごとの位相差を要素に持つベクトルを $\boldsymbol{r} = [arg_0, \ldots, arg_{l-1}]^T$ と定義する。そして $\boldsymbol{D}^+$ を $\boldsymbol{D}$ のムーア・ペンローズ型一般化逆行列とする。$\boldsymbol{q}$ を (Figure 1) で定義される極座標系における音源の位置ベクトルとする。音源距離については正規化し、$|\boldsymbol{q}| = 1$ とする。$\boldsymbol{q} = \begin{bmatrix} \cos\theta\cos\phi \\ \sin\theta\cos\phi \\ \sin\phi \end{bmatrix}$ となる。$\boldsymbol{q}$ の推定値 $\hat{\boldsymbol{q}}$ は次式で求めることができる[5]。

$$\hat{\boldsymbol{q}} = \boldsymbol{D}^+\boldsymbol{r}(2\pi f c^{-1})^{-1} \tag{13}$$

位相差補正処理:

位相差算出処理において、l 番目のサブマイクロホンアレイのマイク間隔のうち少なくとも 1 つ以上のマイク間隔がエイリアシング距離を越えている場合、空間的エイリアシング問題が生じるため、$\hat{\boldsymbol{q}}$ は音源位置ベクトル $\boldsymbol{q}$ の良い推定値とはならない。位相差に 2π の整数倍の不定性が生じるため、不定性を解消した $\boldsymbol{D}^+(\boldsymbol{r} + 2\pi\boldsymbol{n})(2\pi f c^{-1})^{-1}$

が $\boldsymbol{q}$ の良い推定値となる。ここで、n_i を整数値として、$\boldsymbol{n} = [n_0, \ldots, n_{l-1}]$ である。提案手法は、直線配置の場合と同様に $\boldsymbol{n}$ を $l-1$ 番目のサブマイクロホンアレイの情報を利用して算出する。

$\boldsymbol{n}_{-1} = \boldsymbol{0}$、$\hat{\boldsymbol{r}}_{l-1} = \boldsymbol{0}$ とする。ここで $\boldsymbol{n}_l$ は、次式を満たすベクトルである。

$$\boldsymbol{D}_l \boldsymbol{D}_{l-1}^{+} \hat{\boldsymbol{r}}_{l-1} - \pi \boldsymbol{1} \leq_{each} \boldsymbol{r}_l + 2\pi \boldsymbol{n}_l \leq_{each} \boldsymbol{D}_l \boldsymbol{D}_{l-1}^{+} \hat{\boldsymbol{r}}_{l-1} + \pi \boldsymbol{1} \tag{14}$$

ここで $\boldsymbol{x} \leq_{each} \boldsymbol{y}$ は、$\boldsymbol{y}$ の全ての要素が対応する $\boldsymbol{x}$ の各々の要素以上であることを意味する。そして、$\boldsymbol{1}$ は全ての値が 1 をとるベクトルである。

l 番目のサブマイクロホンアレイの位相差ベクトル $\boldsymbol{r}_l$ は次式で求められる $\hat{\boldsymbol{r}}_l$ で補正される。

$$\hat{\boldsymbol{r}}_l = \boldsymbol{r}_l + 2\pi \boldsymbol{n}_l. \tag{15}$$

位相差算出処理及び補正処理は、$l = 0$ から $l = L-1$ までのサブマイクロホンアレイについて実行され、最後に最もサイズの大きいサブマイクロホンアレイの位相差ベクトルの補正値 $\hat{\boldsymbol{r}}_{L-1}$ が求められる。

時間 τ 及び周波数 f での音源方向は、

$$\hat{\boldsymbol{q}}_{L-1} = \boldsymbol{D}_{L-1}^{+} \hat{\boldsymbol{r}}_{L-1} (2\pi f c^{-1})^{-1} \tag{16}$$

となる。 $\hat{\boldsymbol{q}}_{L-1}$ のヒストグラムをピークサーチすることで、音源方向を求めることができる。

4.3 音源方向推定に失敗した成分の棄却

複数の音源がある時間-周波数成分に混合した場合や背景雑音または残響・反響の影響で音源方向の推定に失敗した時間-周波数成分を特定し、その時間-周波数成分を推定に用いないように制御する。直線配置では、$|\frac{\sigma_{L-1}}{2\pi f d_{L-1} c^{-1}}| > 1$ となる場合、棄却する。非直線配置では、$|\hat{\boldsymbol{q}}_{L-1}| > \alpha$ または $|\hat{\boldsymbol{q}}_{L-1}| < \beta$ の場合、棄却する。本稿の実験では、$\alpha = 1.1$、$\beta = 0.9$ に設定した。

5 実験

提案する音源定位方式 SPIRE の評価を残響環境での音源定位実験にて示す。直線配置及び非直線配置の 2 種類のマイクロホンアレイについて実験した。(Figure 3) に直線配置のマイク配置を示す。非直線配置としては、同心円の大きさの異なる 3 つの正三角形サブマイクロホンアレイ（1 辺がそれぞれ 1 cm、3 cm、9 cm）を有する配置を用いた。実験は残響時間約 300ms のオフィスルームで行った。音源とマイクロホンアレイの距離は 1 m に設定した。音源は $\phi = 0$ の平面に配置した。方位角のみを推定対象とした。サンプリングレートは 32kHz とした。短時間フーリエ変換のフレームサイズは 2048 ポイントとして、フレームシフトは 512 ポイントとした。原信号として、日本語男性音声 3 発話を用いた。信号長は約 2 秒とした。音源定位で用いるヒストグラムの分割数は 200 として、等分割とした。ヒストグラムを正規化した $P'(k) = \frac{P(k)}{\sum_i P(i)}$ を評価に用いる。

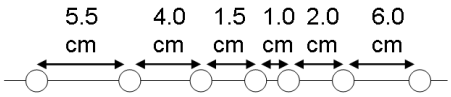


Figure 3: 直線配置のマイク配置: 用いたマイクペアのマイク間隔は 1.0 cm, 1.5 cm, 2.0 cm, 2.5 cm, 3.0 cm, 4.5 cm, 5.5 cm, 6.5 cm, 8.5 cm, 10.5 cm, 14.5 cm, 20.0 cm とした。

5.1 直線配置での音源定位結果

直線配置での音源定位結果を示す。1 つのマイクペアの位相差しか用いない従来の DUET 法[4]と比較する。2 音源を 85°,95° に近接して配置する。音源探索範囲は 180 °とする。直線配置における正規化したヒストグラムを (Figure 4) に示す。マイクペアを 1 個のみ使った場合 (従来の DUET 法に相当)、2 つの音源方向を別々のピークとして検出することができなかった。それに対して、マイクペアを 7 個または 12 個用いて提案する SPIRE を適用した場合では、近接する 2 音源を別々のピークとして検出することができた。マイクペアが 7 個の場合と 12 個の場合を比較すると、12 個の場合のほうがピークが鋭くなっており、使用するマイクペアを増やすことで性能が向上することが分かった。

5.2 非直線配置での音源定位結果

非直線配置での音源定位結果を示す。周波数毎に空間的エイリアシングを生じないという条件でもっともサイズの大きいサブマイクロホンアレイを選択して従来のスパース性に基づく音源定位法で定位する方法 (サブマイクロホンアレイ選択型) と比較する。

3 音源を −125°, −65°, 95° に配置し、非直線配置のマイクロホンアレイを用いて、360 °の音源定位実験を行った結果を (Figure 5) に示す。結果より空間的エイリアシングが生じておらず、各音源の定位に成功していることが分かる。サブマイクロホンアレイ選択型では、各音源方向に立っているピークの鋭さが提案手法と比較して鈍く、推定誤差が大きい時間-周波数成分が多いことが分かる。

3 音源を 75°, 85°, 95° に近接して配置した場合の音源定位結果を、(Figure 6) に示す。サブマイクロホンアレイ選択型では、3 つの音源方向を別々のピークとして検出することができなかったが、3 つのサブマイクロホンアレイを用いた SPIRE では方向差 10 °の近接する音源の方向を別々のピークとして検出することができた。周波数

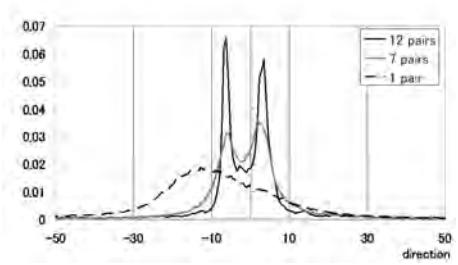


Figure 4: 直線配置マイクロホンアレイによる正規化ヒストグラム: 85°,95° に音源を配置。“1pair” はマイク間隔 1.0 cm のマイクロホンペアを用いたときの結果 (DUET 法に相当)。 “7 pairs” はマイク間隔 1.0 cm から 5.5 cm の 7 つのマイクロホンペアを用いて、SPIRE を適用した時の結果。“12 pairs” はマイク間隔 1.0 cm から 20.0 cm の 12 つのマイクロホンペアを用いて、SPIRE を適用した時の結果。

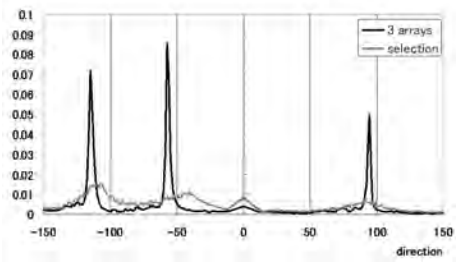


Figure 5: 非直線配置における音源定位結果: −125°,−65°, 95° に 3 音源を配置。“3 arrays” は 3 つのサブマイクロホンアレイ (1cm, 3cm, 9cm) を用いて SPIRE を適応した結果である．“selection” は　周波数毎に全てのマイク間隔がエイリアシング距離以下であるという条件の下、最大のサブマイクロホンアレイを選択して、そのサブマイクロホンアレイを用いて音源定位した場合の結果である。

毎にサブマイクロホンアレイを選択するよりも、逐次的に位相差ベクトルを補正する方式のほうが有効であることがわかった。

6　まとめ

本稿では、マイク間隔の異なる複数のマイクペアまたはサブマイクロホンアレイを用いて、マイク間隔の狭いマイクペアから順に用いて、空間的エイリアシングに伴う位相差の不定項を逐次的に推定することを特徴とした音源定位方式 **SPIRE** (Stepwise Phase dIfference REstoration) を提案した。SPIRE は直線配置及び非直線配置の双方に適用でき、360 °音源定位についても可能である。残響時間約 300ms の残響環境での評価の結果、方向差が 10 °の近接した音源を従来法では定位することができなかったが、SPIRE では定位することが可能であることを示した。

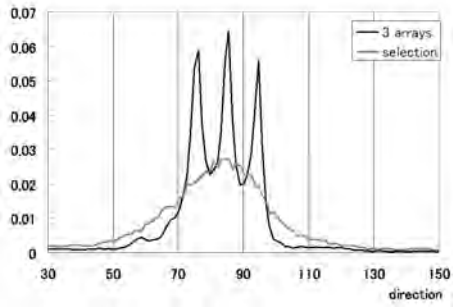


Figure 6: 非直線配置における音源定位結果: 75°,85°, 95° に 3 音源を配置。“3 arrays” は 3 つのサブマイクロホンアレイ (1cm, 3cm, 9cm) を用いて SPIRE を適応した結果である．“selection” は　周波数毎に全てのマイク間隔がエイリアシング距離以下であるという条件の下、最大のサブマイクロホンアレイを選択して、そのサブマイクロホンアレイを用いて音源定位した場合の結果である。

参考文献

[1] 鈴木薫, 古賀敏之, 廣川潤子, 小川秀樹, 松日楽信人, “ハフ変換を用いた音源音のクラスタリングとロボット用聴覚への応用,” 第 22 回 AI チャレンジ研究会, pp. 53-58, 2005.

[2] 戸上真人, 天野明雄, 新庄広, 鴨志田亮太, 玉本淳一, 柄川索, “人間共生ロボット “EMIEW” の聴覚機能,” 第 22 回 AI チャレンジ研究会, pp. 59-64, 2005.

[3] 大賀寿郎, 山崎芳男, 金田豊, “音響システムとディジタル処理,”　電子情報通信学会,1995.

[4] Ö. Yılmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE* Trans.SP, vol.52, no.7, pp. 1830-1847, 2004.

[5] S. Araki, H. Sawada, R. Mukai, S. Makino, “DOA Estimation for multiple sparse sources with normalized observation vector clustering,” *Proc. ICASSP2006*, vol.V, pp.33-36, 2006.

[6] M. Matsuo, Y. Hioka, N. Hamada, “Estimating DOA of multiple speech signals by improved histogram mapping method,” in *Proc. IWAENC2005*, pp.129-132, 2005.

[7] R. O. Schmidt, “Multiple Emitter Location and Signal Parameter Estimation,” *IEEE* Trans. Antennas and Propagation, vol.34, no.3, pp.276-280, 1986.

[8] D. H. Johnson and D. E. Dudgeon, “Array Signal Processing- Concepts and Techniques,” PTR Prentice Hall, New Jersey, USA, 1993.

別の部屋から呼ばれて赴くロボット

– 天井設置型および搭載型マイクアレイによる実現 –

"Calling from the other room"

– A robot task achieved by ceiling and onbody microphone array –

加賀美 聡 [1,2,3] **佐々木 洋子** [1,2] **Simon Thompson** [1]
西田 佳史 [1,3] **溝口 博** [2,1] **榎本 格士** [4]

1 産業技術総合研究所デジタルヒューマン研究センター
2 東京理科大学理工学部機械工学科
3 科学技術振興機構
4 関西電力（株）

Abstract

This paper describes a mobile robot for home service purpose together with ubiquitous ceiling ultrasonic locator and microphone array. User call can be detected by those ubiquitous devices and then mobile robot navigates toward given location. After reaching to the given location, the robot tries to find out user location by using stereo camera, laser and onbody microphone array. We implemented those system at our experimental house "Rokko Holone". System components and experimental results are shown.

1 はじめに

ロボットの移動と対人インタラクションの機能は、サービスロボットにとって本質的に重要である。本論文ではこの例として、別の部屋から呼ばれてそこまで赴くロボットタスクを考える。このような機能はロボット単体で実現するのは技術的にも物理的にも困難である。そこで部屋の天井に設置したユービキタスなマイクアレイとロボットと組み合わせたシステムにより、このタスクを実現することを考える。

システムのコンポーネントは以下の３つである。

1. 天井設置型の超音波位置センサとマイクアレイ、
2. 対人インタラクション用にステレオカメラ、マイクアレイ、レーザーレンジセンサを搭載した移動ロボット、
3. 家屋の外に光ファイバーで接続する、ユービキタスセンサとロボットの処理を行うためのサーバー計算機。

以下では第２節で天井設置型の超音波位置センサとマイクアレイについて、第３節でロボットシステムについて述べた後、第４節でこれらのシステムを実験住宅に設置し、本論文のタスクである別の部屋から呼ばれて赴く行動実験について述べる。

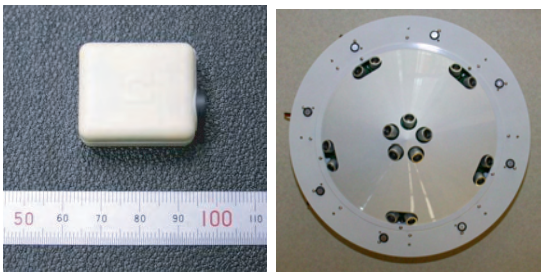


Figure 1: Ultrasonic tag (left) and 8ch ceiling microphone array (outer ring) together with 15 ultrasonic receivers (inner cone) (right)

2 天井設置型超音波タグ受信機およびマイクアレイ

2.1 超音波タグシステム

超音波タグシステムはタグの三次元位置を計測するために開発されたもので、以下の３つの要素で構成されている。1) 対象物に付けられたタグ（送信機）(Fig.1 左)、2) 天井に設置された受信機 (Fig.1 右)、3) タグと受信機の同期を無線で取る同期ユニット。

同期ユニットはタグの ID を 315MHz の搬送波で送信し、タグは自分の固有の ID を受け取ると 40khz の超音波を発信する。受信機はタグが送信した超音波が到達するまでの時間を計測し、その時間を RS485 バスによりサーバー PC に転送する。サーバー PC は各受信ユニットへの到達時間から統計的にタグの三次元位置を推定する。

i 番目の受信機の位置を (x_i, y_i, z_i) とし、タグ（送信機）までの距離を L_i とすると、i 番目の受信機を中心とした

Figure 2: Experimental House “Rokko Holone”

Figure 3: Views of the house with Ceiling Ultrasonic Locator and Microphone Array

球の式が置ける [Nishida, 2003]。

$$(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2 = L_i^2, \quad (0 < i < n) \quad (1)$$

従って原理的には直線上にない３つの受信機から、タグの位置は計算できることになる。誤差を ε_i とすると、ε_i は下記のように表せる。

$$\varepsilon_i = \left| l_i - \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} \right| \quad (2)$$

ε_i を最小化する (x, y, z) を求めることにより、タグ $(\hat{x}, \hat{y}, \hat{z})$ の位置を推定することができる。

$$(\hat{x}, \hat{y}, \hat{z}) \quad = \quad \min_{(x,y,z)} \sum_{i}^{n} \varepsilon_i \quad (3)$$

式 3 は非線形の最適化問題であり解析的には解けない。そこで RANSAC(Random Sampling Consensus) 法を用いて、限られた数のサンプリングから、この問題を精度良く解いている。

2.2 天井用マイクアレイ

マイク数や配置はマイクアレイの特性を決定する主要な要素である。ここではユービキタスセンサとして多数配置することを目的として、個々のユニット辺りマイク数を８としてシミュレーションを行い、直径 30cm のリングを選択した。

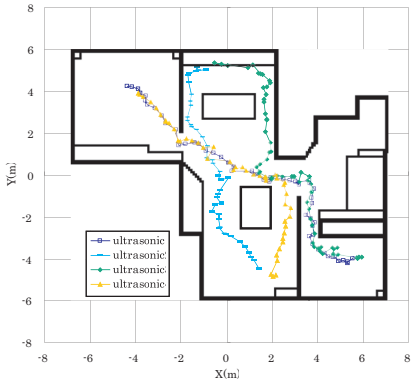


Figure 4: Ultrasonic Locator Results for the Same Four Runs

システムを接続するために PCI 128ch 44Khz 16bit 同時サンプリング AD ボードを開発した。このボードは全データを取得から 5us 以内に DMA 転送を用いて、サーバー PC のメインメモリに転送することができる。実時間を保証するためにシステムは ART-Linux を用いて周期タスクを実行している。

サーバー PC は Delayed-sum beam forming 法 (DSBF) と周波数帯域選択法 (FSB) を用いて、音源位置を 100msec 周期で探索し、見つけた音を分離している。分離された音は Julian[Lee, 2001] を用いて認識している。本論文では家の中の場所の名前、ロボットの動作など約２０個程度の単語のみの辞書を作成し、認識を行った。

2.3 実験ハウス “Holone”

天井設置センサは前述の 15ch の超音波受信機を中心部に、8ch マイクアレイを外周部に持つものである (Fig.1)。このユニットを実験ハウス “Holone” (Fig.2) のうちの４部屋（リビング、キッチン、玄関、書斎の計約 $106m^2$）に合計１６個設置した。Fig.3 はユニット設置後の各部屋の様子である。

2.4 超音波タグシステムによる位置計測

超音波タグをロボットに搭載して移動させ、システムの出力と、ロボットの軌跡を床にマークして手動で計測した結果を Fig.4 に示す。４回の走行で合計約 60.5m を走行し、誤差は平均 35cm であった。

2.5 マイクアレイによる音源定位

Fig.5 は音源定位の様子を示す。各ユニットからの仰角（４分割）と方位角（８分割）に応じて音の強度を示している。左下には一つのユニットからの出力を拡大して表示

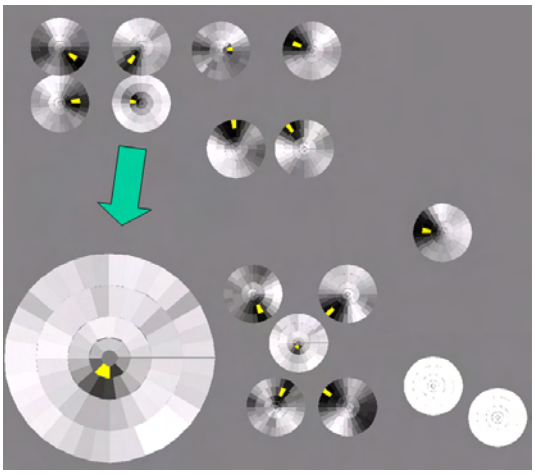


Figure 5: Sound Localization Results by Ceiling Microphone Array

している。黄色い領域はそのユニットでもっとも音の強い領域である。

このマイクアレイで得られる S/N 比は 1khz の sin 波で約 3dB であるため、大きなノイズ源があるときに、対象を認識可能な精度で分離することは難しかった。

3 車輪移動ロボット Penguin2

Penguin2(Fig.6) は、前述のシステムを実装するためにアールラボ社により設計・製作された、駆動輪二輪を前輪に、キャスター二輪を後輪にもつ小型車輪移動ロボットである。

屋内で使用することを前提に、駆動輪を中心に直径 50cm の空間でぶつからずにその場旋回が可能な配置としてあり、最大 7 度のスロープの登降、2cm の段差を乗り越えが可能となるように設計されている。走行の安定性のために、駆動輪、キャスター輪共にバネ・ダンパーによるサスペンションを搭載すると共に、キャスター輪はベルトを用いた半円形キャスターにより見かけ上の輪径を大径化している。また高速走行が可能となるように重心位置を極力下げている。

緒元は全長 40cm、全幅 45cm、全高 32cm、重量約 15kg、最高時速約 2m/s、重心高さ 16cm（駆動輪より 7cm 後ろ)、駆動輪中心間距離 35cm、ホイールベース（駆動輪とキャスター軸間）約 25cm である。駆動輪用に Maxon RE30(60W) DC モーターを 2 個搭載している。電源は Li-ION バッテリー (26V7.5Ah) を 2 個搭載し、モーターと計算機をそれぞれ独立に駆動している。連続走行中の計算機側のバッテリーの保持時間は約 4.5 時間である。

計算機は PentiumM-2.1GHz (FSB400MHz, 1GB メモリ) ハーフサイズボードを PCI バックプレーンに接続し、RT-Linux により 1ms のサーボループを実現している。また無線 LAN、IEEE1394、IO(エンコーダカウンタ、AD,DA) ボードをそれぞれ PCI バスに接続し、RS422 インターフェースを USB バスに接続している。また計算機システムはモジュール化され、簡便に取り出してメンテナンスや交換することが可能となっている。

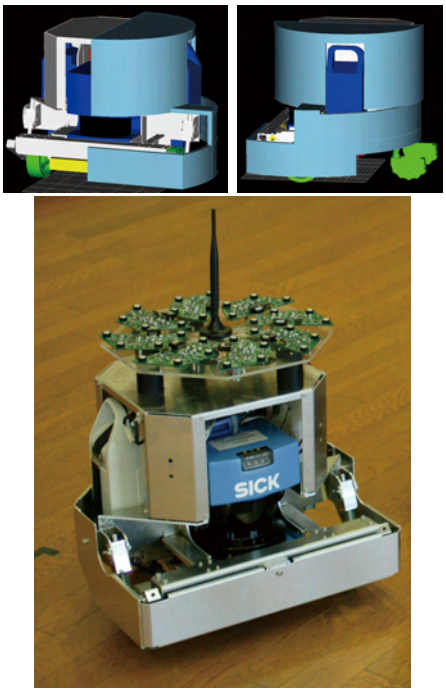


Figure 6: Mobile Robot “Penguin2”

外部センサーは Sick LMS-200 レーザー距離センサ (RS422 接続)、自作の 32ch 同時サンプリングマイクアレイボード（IEEE1394 接続）[佐々木, 2006]、Videre Design STH-DCSG-STOC ステレオカメラ (IEEE1394 接続）を搭載している。

また非常停止、外部からの電源供給と内部のバッテリー使用の切り替えリレー、LED による状態表示の制御回路が計算機とは独立に機能している。

3.1 Particle Fileter による自己位置同定

Particle Fileter（パーティクルフィルタ）は大きな状態空間に対して事後確率分布をオンラインで効率よく表現・計算する方法である (モンテカルロフィルタ (Monte Carlo filter)[Kitagawa, 1996]，Condensation アルゴリズム [Isard, 1998] などとも呼ばれている)．本手法の基本的なアイディアは，状態空間に値を持つ多数の粒子の状態空間中の分布によって確率分布を近似表現することである．カルマンフィルタと異なり、ガウス分布でない任意の状態遷移確率モデルを扱える点が、実世界の事象を扱うのに向いているために、近年盛んに研究されている。

時刻 t におけるロボットの状態を x_t, センサ入力を z_t, 制御出力を u_t とあらわすとする。ある時刻 t における状態は、初期状態、観測確率と出力の遷移からマルコフ仮定

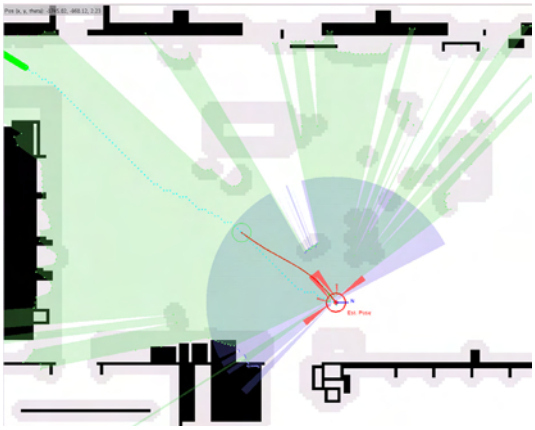


Figure 7: Particle Filter based localization, mapping and path planning result

を置くことにより以下のように表される。

$$p(x_t|x_{0:t-1}, z_{1:t-1}, u_{1:t}) = p(x_t|x_{t-1}, u_t) \quad (4)$$

$$p(z_t|x_{0:t-1}, z_{1:t-1}, u_{1:t}) = p(z_t|x_t) \quad (5)$$

$p(x_t|x_{t-1}, u_t)$ は状態遷移確率であり、$p(z_t|x_t)$ は観測確率である。Particle Filter を用い、あらかじめ与えた地図 m_t をセンサ入力 z_t とマッチングすることにより x_t の事後分布確率 $p(x_t|z_{1:t-1}, u_{1:t})$ を推定することができる。

つまり，事後確率 $p(x_t|z_{1:t-1}, u_{1:t})$ は，初期状態と状態遷移確率および観測確率の 3 つの分布により求めることができる．また，実際に Particle Filter を用いるときには、結果はロボットの初期位置に依存し、時間的に不変なロボットのセンサとモーションに関する確率遷移モデリングが必要となる．

筆者らはこの Particle Filter を用いて、レーザー距離センサ、ステレオ距離画像、超音波センサからの入力に対応したセンサモデルと、2.5 次元環境地図表現による自己位置同定システムの構築を行なった [Thompson, 2004a]. Fig.7 に地図と位置認識，経路計画の結果を示す．図中で赤い円がロボットの位置を表し、そこを中心に緑および青で示されたエリアがレーザー距離センサの入力である。ゴールは緑の棒で示され、そこにいたる経路がグリッドベースの最短経路が水色で、後述する 4 次多項式による滑らかな経路が赤色で示されている。

3.2 Rao-Blackwelllized Particle Filter による地図作成

Rao-Blackwellized Particle Filter は、ロボットの位置推定だけでなく、地図を含めた状態確率を最大化することが出来る。地図 m を導入した事後確率は $p(x_1 : t, m|z_{1:t}, u_{1:t})$ となる。この場合ロボットの移動してきた経路のすべての状態 $x_{1:t}$ に対する事後確率を計算するために計算量と必要なメモリは非常に多い。

Fig.8 に Fig.8 に示した環境を 3 周回りながら、Rao-Blackwellized Particle Filter によりマッピングし、さらに局所的な状態遷移誤差をループを検出することにより緩和した結果を示す。

Figure 8: Mapping result for Fig.1 area

3.3 ステレオ視による障害物発見

再帰相関演算手法 [Faugeras, 1993] と一貫性評価法 [Bolles, 1993] による高速なステレオ視により距離画像をリアルタイムに生成している．次に得られた距離画像を 2.5 次元の確率表現に変換し，その各セルの存在確率を時間方向に積分して行くことで，環境の 2.5 次元マップが作成される（Fig.9 ）．この 2.5 次元マップから障害物を検出することにより，レーザーレンジセンサで観測している 2 次元の世界では検出不能な障害物が発見可能である [Thompson, 2004b, Thompson and Kagami, 2004].

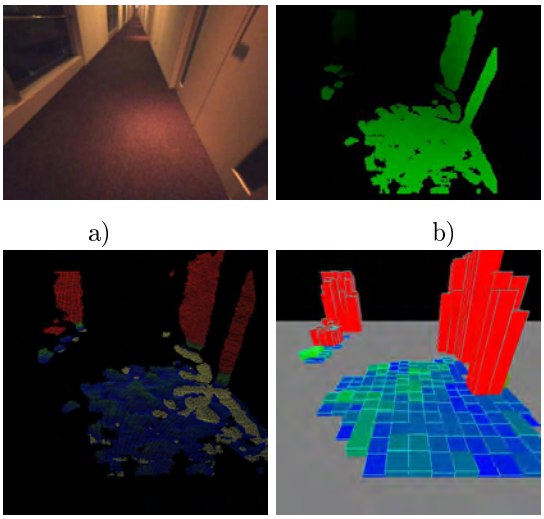


Figure 9: ステレオ画像処理， a) カメラ画像， b) 視差画像， c) グローバル座標に投影した各点， d) 各セルの確率

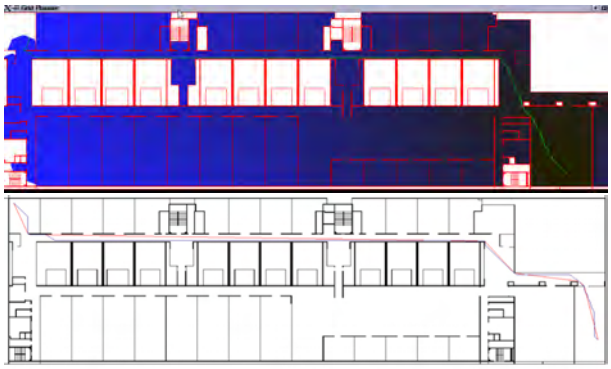


Figure 10: Grid Optimized A* for AIST Waterfront Building (3F) and path smmothing result

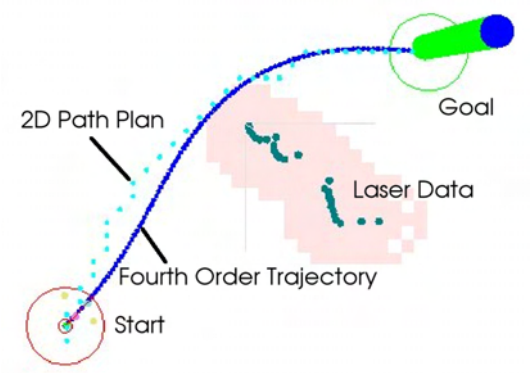


Figure 11: Generated trajectory from path planning results

3.4 経路計画

与えられたゴールまでの経路計画は、ビットマップのグリッドとして表現された地図を，A*アルゴリズムをグリッド探索用に最適化した手法 [Kuffner, 2004] を用いて探索し、最短経路を得ている。通常は 5 または 10cm のグリッドを用いており、例えば Fig.10 上の場合には 140 × 40m の研究室の平面図を 1400 × 400 のグリッドマップにして表現しており、対角の二点間の探索には約 1.5s 程度かかる。

また得られた経路は、滑らかではない上に、グリッドの８近傍を移動するという制約の元では最短ではあるが、グリッド上を通らなくてよければ最短ではないために、ランダムに２点を選びながら経路を滑らかに縮めてゆくパススムージング手法を取り入れている [金原, 2006]。Fig.10 下に、グリッド最適版 A*（青線）とパススムージングを施した結果（赤線）を示す。経路長は数%減少すると共に、経路の単位長さに対する平均進路変化は 1/20 になり、そのために要する計算時間はグリッド最適版 A*の約 10 %だった。

3.5 経路制御

経路計画により目的とするゴールまで計算したパスは、そのままでは滑らかでないために実現できない。そこで４次の多項式により障害物を避けた滑らかな経路を探索的に計算する。ロボットの状態を $(x, y, \theta, \kappa, L)$　で表すとする。ここで κ は曲率、L は障害物からの距離の逆数で定義される項である。

$$x(s) = x_0 + \int_0^s cos(\theta(s)), \quad y(s) = y_0 + \int_0^s sin(\theta(s)) \tag{6}$$

$$\theta(s) = \theta_0 + \int_0^s \kappa(s) = \theta_0 + \kappa_0 s + \frac{as^2}{2} + \frac{bs^3}{3} + \frac{cs^4}{4} + \frac{ds^5}{5} \tag{7}$$

L は次の式で表される。

$$L(s) = \int_0^s \left(\frac{\lambda}{v_0} + \ldots + \frac{\lambda}{v_{N-1}} \right) \tag{8}$$

λ は障害物の進入により定義される定数であり、v_i は i 番目の障害物とロボットの位置との距離である。初期状態を $\underline{x}_0 = (x_0, y_0, \theta_0, \kappa_0, L_0)$、ゴールを $\underline{x}_T = (x_T, y_T, \theta_T, \kappa_T, \mathrm{L}_T)$ とする。ロボットの状態ベクトルは４次多項式のパラメータの関数として表した非線形方程式を、その逆ヤコビアンを収束計算によって解き、解を得ることができる。実際のセンサ入力から計算を行った例を Fig.11 に示す。

3.6 搭載用 32ch マイクアレイ

ロボットに搭載するマイクアレイは定位と分離の精度を上げるために、サイドローブを低減化した 32ch のシステムを開発した。アレイのサイズは “Pen2” の外形から 33cm に限られている。シミュレーションを繰り返しながら実験的にサイドローブが小さくなるマイク配置を検討した。Fig.12(左) が開発したシステムのマイク配置である。 構成要素を分割可能とするための配置を検討し、等辺台形に４つのマイクを搭載したものを８つ合わせて、合計 32ch のシステムを設計した。Fig.12(右) がシステムの DSBF によりシミュレーションした感度分布である。Fig.13 に 1, 1.4, 2Khz 時のシミュレーション結果を示す。各周波数でサイドローブが低減化されていることがわかる。

Fig.14 が開発したシステムである。AD ボードは 16bit 解像度で、16khz 同時サンプリングを行い、IEEE1394 バス経由で PC にデータを転送する。Table 1 に AD ボードの仕様を示す。

4 実験結果

4.1 別の部屋から呼ばれて赴くタスク

第 2,3 節で述べたシステムにより、別の部屋から呼ばれてアプローチするタスクを実現した。天井に設置したマイ

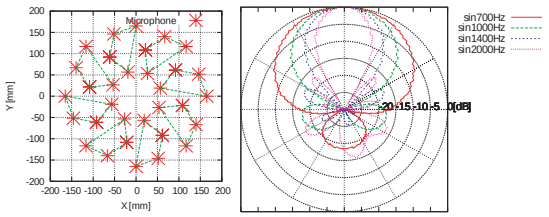


Figure 12: 32ch Microphone Arrangement(left) and Microphone Directional Pattern in DSBF(right)

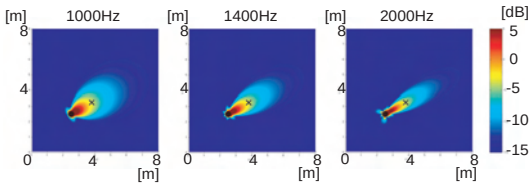
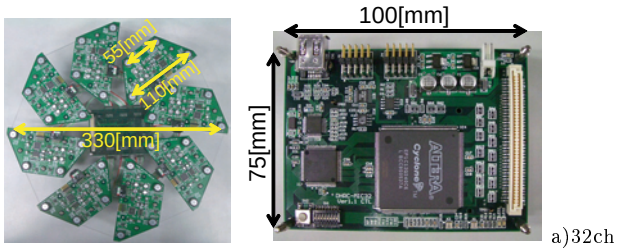


Figure 13: Beam Forming Simulation Result



a)32ch Microphone Array and b) DHRC Microphone AD Board

Figure 14: Microphone Array and AD board

Table 1: Spec. of Microphone AD Board

board size	w=75, d=100, h=30 [mm]
input channel	32 channel
interface	IEEE-1394
data transfer	isochronous transfer
sampling speed	16[kHz]
resolution	16bit, programable gain amp.
power supply	+5 [v]

クアレイにより、システムは建物内のどこから呼ばれたかをロボットのサーバーに伝え、ロボットはそこまで移動する経路を計画し、ローカルな障害物を避けながらそこまでナビゲートする。指定した部屋に入ったら、ロボットは搭載したセンサ（本論文の場合には搭載型マイクアレイ）とレーザー距離センサの入力と対応させ、対象物に近づく。Fig.15 に GUI 画面のスクリーンショットを示す。またビデオをキャプチャしたものを Fig.16 に示す。ユーザーが書斎からロボットを呼び (Fig.16 b)、ロボットがその部屋に向けて動き出す (Fig.16 c)。リビングルームを横切り (Fig.16 d,e)、次に玄関を横切り (Fig.16 f)、書斎に入りユーザーにアプローチする (Fig.16 g)。この実験における移動距離は 18.4m であり、ユーザーが呼びかけてから、ロボットが到達するまでの時間は 33 秒、平均移動速度は 0.7m/s であった。

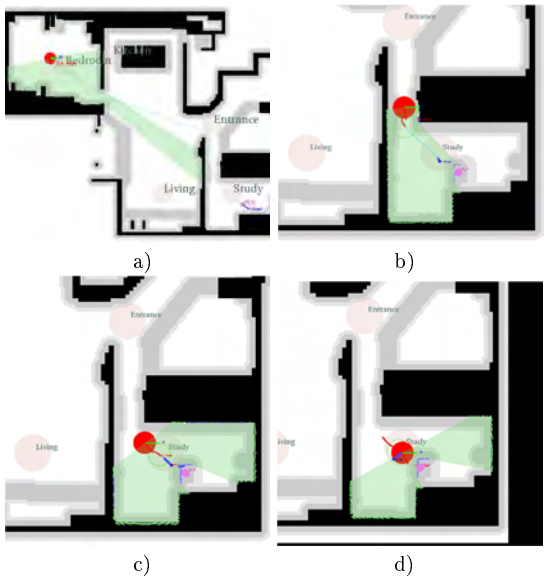


Figure 15: An example of approaching a sound source within the ubiquitous house. Figure a) shows the planned path to estimated tag location, while b), c) and d) show the robot approaching a local goal identified by fusing location with locally detected objects.

4.2 指示された場所に赴くタスク

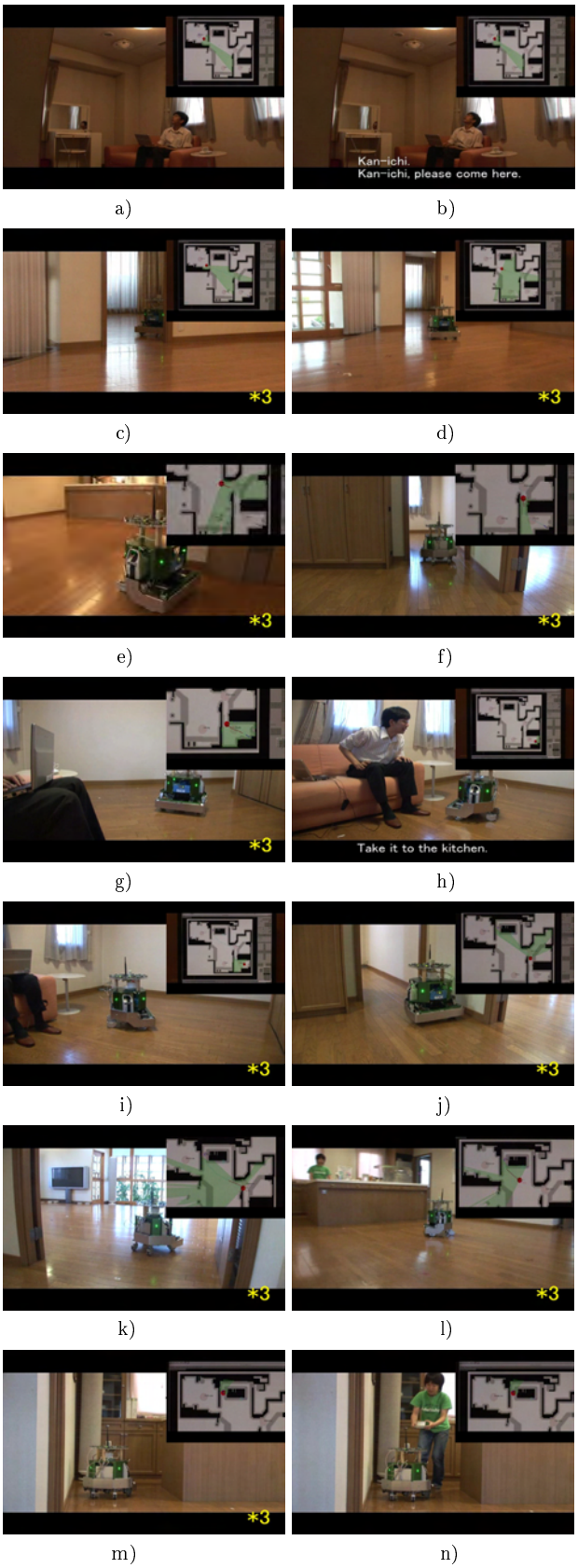
次にユーザーはコーヒーカップをロボットに載せ、キッチンまで持っていくように指令した。この指令は搭載のマイクアレイから Julian により認識されて、ロボットの行動に変換される。認識のための辞書は 1) 行け, 2) 来い、3) 挨拶、に大別されている。また場所の名前を約１０有している。

ロボットが場所を指定されると、その場所へのナビゲーションを開始する (Fig.16 i)。玄関を横ぎり (Fig.16 j)、リビングを横切り (Fig.16 k, l)、キッチンに到着する (Fig.16 m)。最後に別のユーザーがカップを受け取り、ロボットを解放する (Fig.16 n)。

5 おわりに

本論文では屋内にユービキタスに設置された超音波位置センサ、マイクアレイと、ロボットの統合によるロボットサービスのコンセプトを提案した。この提案を確認するために、1) 天井設置型の超音波位置センサ、マイクアレイ、2) 搭載型マイクアレイと移動ロボットシステム、3) 自律移動のための位置認識、地図作成、経路計画、経路制御システムを開発した。

別の部屋から呼ばれて赴くには３つのフェーズがある。1) 別の部屋からユーザーが呼びかけ、天井設置のマイク



a) b) c) d) e) f) g) h) i) j) k) l) m) n)

Figure 16: Snapshots on Experiment

アレイ（または超音波タグシステム）により位置を同定する、2) ロボットが自己位置を同定し、障害物を避けながらナビゲーションする、3) 同じ部屋に入ったらユーザーを検出するために搭載型マイクアレイ、音源定位、画像処理、レーザーセンサの結果を統合して同定し、接近する。

参考文献

[Lee, 2001] A.Lee, T.Kawahara and K.Shikano. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, pp. 1691–1694, 2001.

[Bolles, 1993] R. Bolles and J. Woodfill. Spatiotemporal Consistency Checking of Passive Renge Data. In T. Kanade and R. Paul, editors, *Robotics Research: The Sixth International Symposium*, pp. 165–183. International Foundation for Robotics Research, 1993.

[Faugeras, 1993] O. Faugeras, B. Hots, H. Mathieu, T. Viéville, Z. Zhang, P. Fua, E. Théron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy. Real Time Correlation-Based Stereo: Algorithm, Implementations and Applications. Technical Report N°2013, INRIA, 1993.

[Kitagawa, 1996] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

[Isard, 1998] M. Isard and A. Blake. “Condensation-conditional density propagation for visual tracking”. *International Journal of Computer Vision*, 28:5–28, 1998.

[Kuffner, 2004] James J. Kuffner. Efficient optimal search of Euclidean-cost grids and lattices. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

[Thompson, 2004a] Simon Thompson and Satoshi Kagami. Stereo Vision and Sonar Sensor Based View Registration for 2.5 Dimensional Map Generation. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS04)*, pp. 3444–3449, Oct. 2004.

[Thompson, 2004b] Simon Thompson and Satoshi Kagami. Stereo Vision Terrain Modeling for Non-Planar Mobile Robot Mapping and Nabigation. In *Proceedings of 2004 IEEE International Conference on Systems, Man & Cybernetics(SMC04)*, pp. 5392–5397, Oct. 2004.

[Thompson and Kagami, 2004] Simon Thompson and Satoshi Kagami. Revising stereo vision maps in particle filter based slam using localisation confidence and sample history. In *Proceedings of 2nd International Conference on Autonomous Robots and Agents(ICARA2004)*, pp. 218–223, New Zealand, Dec. 2004.

[Nishida, 2003] Y. Nishida, H. Aizawa, T. Hori, N. Hoffman, T. Kanade and M. Kakikura. 3D Ultrasonic Tagging System for Observing Human Activity. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 785–791, 2003.

[金原, 2006] 金原 正朋, 加賀美 聡, ジェームズ・カフナー, 溝口 博. 車輪型ロボットの滑らかな経路探索のための経路平滑化手法と Field A*との比較. ロボティクス・メカトロニクス講演会'06 講演論文集, pp. 2P2–E09, May 2006.

[佐々木, 2006] 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性能評価. 第２４回ロボット学会学術講演会講演論文集, pp. 985–986, 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-0624-11 (11/17)

ことばの前/下のインタラクション——ヒトの場合・ロボットの場合

Interaction before/under speech — in humans and in robots

小嶋 秀樹
Hideki KOZIMA
情報通信研究機構*
National Institute of Information and Communications Technology
xkozima@nict.go.jp

Abstract

This paper explores the epigenetic and robotogenetic origins of human communication. We built interactive robots, Infanoid and Keepon, with which we observed children's spontaneous communicative behavior; from this experiment, we learned that attentional and emotional exchange would play an indispensable role in the emergence of human communications including verbal one. We then conducted longitudinal field observations of a group of children with developmental disorders interacting with Keepon; from this field study, we learned that the children, including those with autism, spontaneously engaged not only in dyadic interaction with the robot, but also in triadic interaction among children and carers, where the robot functioned as a pivot of the interpersonal communication. Based on these findings, we further extend our idea of the origin of human communication into that children's motivation to share meanings of environmental events with others would probably be the most fundamental prerequisite for the genesis.

1 はじめに

コトバの使用は，ヒトを他の動物から区別する特徴といわれる．たしかに，二重分節された音声や，表現と意味の恣意的な対応づけは，ヒトのコミュニケーションだけに見られる特徴だろう．コトバで思考し，コトバで他者とコミュニケーションすることが，ヒトのもつ知性の大きな柱であることに疑いはない．しかし，コトバを使用する能力によって，はじめて豊かなコミュニケーションが可能になったのかというと，必ずしもそうではない．とくに個体発達という観点からコミュニケーションを見てみると，コトバの出現（およそ１歳から１歳半）のはるか前から，子どもは豊かなコミュニケーション——たとえば情動の共有・注意の共有・行為の共有など——を実践している．この前言語的コミュニケーションによってコトバの獲得，そしてコトバの後のコミュニケーションが可能になる．さらに，前言語的コミュニケーションは，コトバの獲得以降も作動しつづけ，言語による命題情報のやりとりの下で，意図や感情といった心理情報を交流させていく．ある意味，やりとりされる言語情報は〈搬送波〉であり，そこに変復調される心理情報こそが伝えたい・知りたい〈情報〉なのだろう．

この論文では,「ことばの前/下のインタラクション」とは何か，それは何によって可能になるのかを議論することで，ヒトのコミュニケーションの成り立ちを再考してみたい．第２節では，子どもとロボットのインタラクション観察を紹介し，注意（指さし・視線など）や情動（表情・韻律など）のやりとりが，言語的および非言語的コミュニケーションの底流をなしていることを説明する．第３節では，ロボットを使った発達障害児の療育支援を紹介し，注意や情動をやりとりする能力でさえも，より根源的なコミュニケーションへの動機づけと周囲からの働きかけによって，段階的に構成されていくことを，いくつかの事例をとおして説明する．第４節では，この知見を今後の人工知能研究にどのように反映させるべきかを考察したい．

2 注意と情動のインタラクション

子どものコミュニケーション，とくに母子間（主たる養育者と子どものあいだ）のインタラクションは，アイコンタクト（見つめあい）やそれに伴う表情や声のやりとりに始まり，やがて指さしや共同注意（同じ対象を見ること）

* 情報通信研究機構 知識創成コミュニケーション研究センター
〒 619-0289 京都府相楽郡精華町光台 3-5

Figure 1: Infanoid engaging in eye-contact (left) and joint attention (right) with a human interactant.

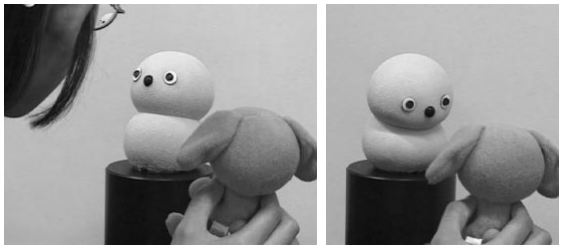


Figure 2: Keepon performaing eye-contact (left) and joint attention (right) with a human interactant.

による注意のやりとりへと，コミュニケーションを深めていく [Kozima and Ito, 2003]．注意対象を同じように知覚し，その対象に向けられた情動を参照しあうことで，子どもと養育者はたがいの存在や対象との関わりを共同化していく．このような営みのなかで，共感的コミュニケーションが育まれていく [小嶋・高田, 2001]．

2.1 ロボット

このようなコミュニケーション発達を観察・モデル化するために，いくつかのインタラクティブ ロボットを開発してきた．その開発コンセプトは『子どもから自発的なコミュニケーション行為を引きだす〈身体〉』である．現在までに,〈子ども型ロボット Infanoid〉と〈ぬいぐるみロボット Keepon〉などをデザイン・製作した．

Infanoid（図１）は，３〜４歳児とほぼ同じ大きさ（座高 480mm）の上半身ヒューマノイド（人間型ロボット）である [Kozima, 2002]．おもに幼児期から学童期の子どもたちとのインタラクションを想定してデザインした．唇や眉などによる情動の表出，視線や指さしなどによる注意の表出，そして何かに手を伸ばす・何かを手でつかむなどによる意図の表出が可能である．子どもたちが，自発的に情動・注意・意図などを Infanoid に帰属させ,〈心をもったエージェント〉としての Infanoid と遊んでもらうことをねらっている．

Keepon（図２）は，おもに乳児期から幼児期の子どもたちと，安全なインタラクションができるようにデザインされている [仲川ほか, 2004]．高さ 120mm・直径 80mm のシリコンゴムでできたダンゴ型の身体にできることは，(1) 注意の表出：顔（つまり視線）を人物や対象物に向けることと，(2) 情動の表出：身体を左右あるいは上下に揺すり，楽しさや興奮といった心の状態を表現することだけである．Keepon も何かを見たり喜んだりできることを，子どもたちに直観的に感じとってもらうことをねらっている．

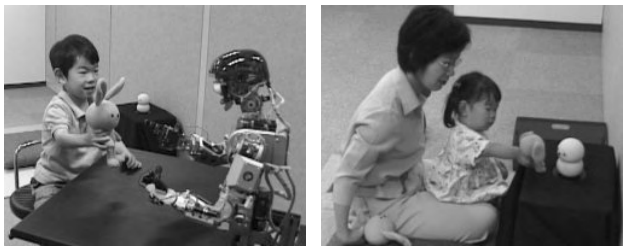


Figure 3: Children interacting socially with the robots.

2.2 子どもとのインタラクション

コミュニケーションは双方向的な行為のやりとりである．ロボットにアイコンタクトや共同注意の能力やその発達プロセスを実装しただけでは片手落ちで，そのロボットに人間（とくに子ども）がどのように関わろうとするのかを観察することも不可欠である．この観察から，(1) ロボットに欠けていた（あるいは過剰な）機能や形態が明らかになるし，また (2) 子どもの発達そのものをより詳細に調べることができる．

約 40 名の乳幼児を対象としたインタラクション実験（図３）から明らかになったのは，子どもたちがロボットとの関係を（時間経過とともに・発達年齢とともに）ダイナミックに変化させていくプロセスである．最初，子どもたちは〈動くモノ〉としてロボットを捉えているが，ロボットの視線・表情・身ぶりなどから，子どもたちは自律的な主体（＝生命性）をロボットのなかに発見し，知覚し応答する〈システム〉としてロボットを理解するようになる．やがて子どもたちは，ロボットの視線・表情・身ぶりなどが，子ども自身の行為に随伴している（時間的・空間的な関連がある）ことに気づいていき，心をもった〈エージェント〉としてロボットの行為を解釈しようする．子どもたちはロボットとの注意や情動のつながりを深めながら，共感的コミュニケーションへと入っていく [小嶋, 2003; Kozima et al.,2004]．

3 療育支援から見えてきたこと

前節で取り上げたインタラクション実験は，実験室でのその場かぎりの観察であり，子どものコミュニケーション能力の成り立ちを十分に見ることができなかった．そこで，より実践的なコミュニケーション発達の〈現場〉[小嶋, 2004] として，自閉症などの発達障害をもつ子どもたちの療育施設を長期訪問し，日常的な療育活動のなかで子どもたちとロボットのやりとりを縦断的に観察するこ

Figure 4: Keepon in the playroom at a day-care center.

とを開始した．

この療育施設では，子どもたち（おもに２～４歳）と母親たち，そして療育士たちが，さまざまな自由遊びやグループ遊びをくりひろげる．この多様でダイナミックな，それでいて限りなく日常的な実践のなかで，子どもたちの行為はゆっくりと意味づけられていく．

3.1　プレイルームでの Keepon

この療育施設のプレイルームに，ぬいぐるみロボット Keepon を置かせていただいた（図４）．約３時間の療育セッションのあいだ，子どもたちは好きなときに Keepon で遊ぶことができる．自由遊びのあいだは，さまざまなオモチャのひとつとして，いつでも Keepon で遊べる．またグループ遊びのあいだ，Keepon は邪魔にならない場所（プレイルームの隅など）に移されるが，グループ活動に飽きや疲れをみせた子どもはいつでも Keepon のところに来ることができる．

プレイルームでの Keepon は，高さ約 25cm のプラスチック製のカバーに入っている．その中に電池や無線装置などを格納することで，別室にいる操作者が Keepon を手動運転モードで遠隔操作できるようにした．操作者は，Keepon が子どもの顔やオモチャを注視するように，また子どもから何らかの働きかけ（アイコンタクトやタッチなど）があったときは，ポンポンポンと音を出しながら身体を数回伸縮させるといったポジティブな情動表出を行なうようにした．

3.2　Keepon からみた子どもたち

このプレイルームで Keepon と子どもたちのインタラクションを 2003 年 10 月以来観察してきた．現在（2006 年 10 月）までに，約 100 セッション（700 人回以上）のインタラクション観察を実施した．子どもとロボットのインタラクションを，これだけ長期縦断的に観察した例はほかにないと思われる．

この観察をとおして，子どもたちと Keepon のインタラクションを，Keepon 自身の眼から捉えることができた．Keepon という第１人称的な視点（パースペクティブ），つまり〈私〉の視点から，子どもたちと関わり，子どもたちの表情・しぐさ・声やコトバなどを記録・分析することができた．この〈私〉とは，実際には Keepon の〈操作者〉の主観になるのだが，Keepon というシンプルな身体をとおした子どもたちへの関わり（ロボットの動作）はすべて記録され，再現可能になっている．つまり Keepon は，子どもとやりとりする〈私〉という現象学的な〈主観性〉と，それを誰でも追体験できる〈客観性〉，これら２つをあわせもった〈メディア〉であるといえる．

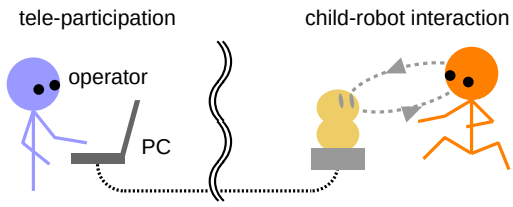


Figure 5: Teleparticipation in the child world.

Keepon からみた子どもたちは，Keepon への関わりをさまざまな形で見せてくれた．ときには，他人に（母親にも）あまり見せたことのない表情や，Keepon に帽子をかぶせてあげる・食べ物をたべさせる（フリをする）といった援助的な行為を，子どもたちは見せてくれた [仲川, 2005]．全体としては以下の点が示唆される．

- ヒトでもオモチャでもない Keepon だからこそ，対人コミュニケーションを苦手とする子どもたちが，安心感と好奇心をもって Keepon にアプローチすることができた．
- 子どもから Keepon への直接的な関わりだけでなく，そこで得られた楽しさ・驚きなどを他者（母親・療育士・ほかの子ども）と共有しようとするような，対人的な関わりへの発展も多くみられた．
- Keepon への関わり方とその変化は十人十色であり，たんなる障害名（「PDD」「自閉症」「ダウン症」など）を越えた，その子らしさ・その子の発達の道すじを物語っている．

現在，Keepon からみた子どもたちひとり一人の〈物語〉を，療育施設でのサービスや家庭での子育てに役立ててもらうために，保護者や療育士にフィードバックすることを進めている．

3.3　あるエピソードから

療育施設でのフィールド観察から得られた知見は，子どもたちのコミュニケーション能力――たとえば共同注意の能力――が実践をとおして形成されていくことである．つまり，注意や情動のやりとりが人間のコミュニケーションの出発点なのではなく，より根元的な能力あるいは動機づけをもって，原初的なコミュニケーション実践を適切

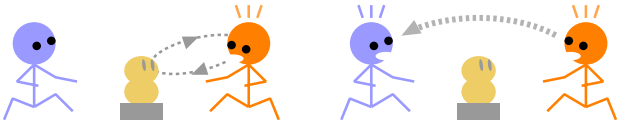


Figure 6: Relating my wonder to his/her wonder.

な養育環境で経験することで，注意や情動をやりとりする能力が形成されていくというものである．このことは，つぎにあげる自閉症児Nのエピソードに例示されている．

> 最初の 15 セッションの間，Nの Keepon への自発的な行為は，ちらっと見ることくらいだった．母親や療育士に抱かれて，半ば強制的に Keepon と対面させられても，インタラクションには発展しなかった．S10（第 10 セッション）で初めて Keepon に触ったが，モノとして Keepon の感触を確かめたようだった．
>
> S16（S15 から 3 カ月のブランクの直後）のおやつの時間，Nが Keepon の前に来て，指で Keepon の鼻を押した．Keepon がポンポンポンと身体を上下に伸縮させて応答すると，Nは少し驚いたような笑顔を見せた．それを見守っていた周囲の母親たちや療育士たちはどっと笑い出した．Nは Keepon の反応を引き出そうと同じ行為を繰り返し，Keepon が応答するたびに横にいる担当療育士や自分の母親に顔を向けて微笑んだ．あたかも自分と同じ驚きや楽しさを確かめ共有しようとするように．
>
> この S16 以降，Nが母親といっしょに Keepon と遊ぶ場面が多く見られるようになった．

S16 での出来事は，既に獲得していた〈社会的参照〉——注意を共有している他者の情動を参照する行為——の能力をNがたまたま行使したというものではない．むしろ，自分が Keepon に見つけた驚きや楽しさと，周囲の大人たちが（共感的に）同時表出した驚きや楽しさのつながりに気づき，それを確かめるように Keepon への関わりと母親や療育士への関わり（参照視＋微笑み）を繰り返していたのだろう（図６）．

4　意味の共有への動機づけ

前節で見たように，発達に必要なものは，個体に内在する根元的な動機づけと，その発現を待ち・読み取り・積極的に応答していく養育環境（養育者）である．これらによって，自発的に自分をとりまく物理的・社会的環境を探索し，そこで発見した驚き・楽しさなどを近しい他者と共有していることを確かめたり，あるいは積極的に伝えたりする行為がはじまる [Trevarthen, 2001]．

ここで大切なことは，あらかじめゴールやタスク (あるいはその達成度を表わす評価関数) が与えられているのではないことである．たとえば，ロボットにこのようなコミュニケーション行為を獲得させる場合，設定されたゴールやタスクをロボットに探索・学習させ，そのアルゴリズムや効率を問うだけでは，ロボットを〈発達〉させているとは言えない．〈発達〉の本質は，何かを求めて外に向かおうとする力が，未知環境とぶつかり，絶え間なく自分自身を変えていくプロセスにある．そもそも環境や身体は不変なものではなく，自然現象や社会活動，あるいは個体の成長や疾病によって，つねに変化していく．したがって〈発達〉にアプリオリなゴールを与えることは意味をなさない．高等動物 (とくにヒト) の〈発達〉は「何かを求めて外に向かおうとする力 (＝自発性)」によって駆動された open-ended なプロセスとしてモデル化されるべきであり，ロボットの〈発達〉もそうあるべきだ．そして，その自発性が養育環境（共感的な解釈のフィードバック）と出会うとき，社会的なコミュニケーションへの発達の道が拓けるだろう．

このような open-ended な〈発達〉の原動力となる「何かを求めて外に向かおうとする力」とは何なのか．この素朴な疑問への一般解が「内発的動機づけ (intrinsic motivation)」である．内発的動機づけとは，それ自体が内的報酬となるような活動への動機づけであり，内から外に向かう原動力といえる．たとえば「新奇性 (novelty)」や「学習の進み (learning progress)」への方向づけは内発的動機づけの好例であり [Barto, 2006]，これらは大脳基底核でのドーパミン系の働きとアナロジーがとれると言われる [Kaplan, 2006]．ゆえに，強化学習との相性もよいようだ．(ちなみに「外発的動機づけ (extrinsic motivation)」とは，外的報酬 (例：ボーナス) を得るためや罰 (例：罰金) をさけるための動機づけであり，身体内部のホメオスタシスによる動機づけ (例：摂食・睡眠) も含まれる.) 内発的動機づけをもったロボットは，未知環境のなかで自発的に活動し，自分自身を適応的に変化させていくことで，open-ended な〈発達〉を実現できるだろう．

謝辞

本研究に協力していただいた仲川こころさん（情報通信研究機構）・矢野博之さん（情報通信研究機構）・安田有里子さん（近江八幡市 心身障害児通園センター）・長谷川郁子さん（八王子保育園）・Jordan Zlatev さん（Lund 大学）・高田 明さん（京都大学）・Marek Michalowski さん（CMU）に感謝の意を表します．

参考文献

[Barto, 2006] Andrew Barto: Intrinsic motivation, cumulative learning, and computational reinforcement

learning, *The Sixth International Confernece on Epigenetic Robotics* (Paris, France), 2006.

[Kaplan, 2004] Frédéric Kaplan and Pierre-Yves Oudeyer: Neuromodulation and open-ended development. *The Third International Conference on Development and Learning* (La Jolla, CA.), 2004.

[小嶋・高田, 2001] 小嶋 秀樹・高田 明：社会的相互行為への発達的アプローチ：社会のなかで発達するロボットの可能性，人工知能学会誌，Vol. 16, pp. 812–818, 2001.

[Kozima, 2002] Hideki Kozima: Infanoid: A babybot that explores the social environment. K. Dautenhahn *et al.* (eds), *Socially intelligent agent*, Kluwer Academic Publishers, pp. 157–164, 2002.

[Kozima and Ito, 2003] Hideki Kozima and Akira Ito: From joint attention to language acquisition, J. Leather and J. van Dam (eds.), *Ecology of Language Acquisition*, Amsterdam: Kluwer Academic Publishers, pp. 65-81, 2003.

[小嶋, 2003] 小嶋 秀樹：赤ちゃんロボットからみたコミュニケーションのなりたち，発達，Vol. 24, No. 95, pp. 52-60, 2003.

[Kozima et al.,2004] Hideki Kozima, Cocoro Nakagawa, and Hiroyuki Yano: Can a robot empathize with people?, *International Journal of Artificial Life and Robotics*, Vol. 8, pp. 83–88, 2004.

[小嶋, 2004] 小嶋 秀樹： ロボットは障害児教育に何ができるか, 渡部 信一 (編著)「21 世紀テクノロジーと障害児教育」, 学苑社, pp. 105-113, 2004.

[仲川ほか, 2004] 仲川こころ・小杉 大輔・安田有里子・小嶋 秀樹：Keepon：子どもからの自発的な関わりを引き出すぬいぐるみロボット, 人工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A401-02, pp. 7-14, 2004.

[仲川, 2005] 仲川こころ： 人との関係に問題をもつ子どもたち——キーポンと療育教室の子どもたち，発達, Vol. 26, No. 104, pp. 89-96, 2005.

[Tomasello, 1999] Michael Tomasello: *The Cultural Origins of Human Cognition*, Harvard University Press, 1999.

[Tomasello et al., 2004] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll: Understanding and sharing intentions: The origins of cultural cognition, *Behavioral and Brain Sciences*, (in press: http://www.bbsonline.org/Preprints/Tomasello-01192004/Referees/)

[Trevarthen, 2001] Trevarthen, C.: Intrinsic motives for companionship in understanding: Their origin, development, and significance for infant mental health. *Infant Mental Health Journal*, Vol. 22, pp. 95–131, 2001.

〒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/