人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0522

ＡＩチャレンジ研究会 (第２２回)

Proceedings of the 22nd Meeting of Special Interest Group on AI Challenges

CONTENTS

【10 月 14 日】

【裏へ続く】

日　時　2005 年 10 月 14 日～ 15 日　　場　所　伊豆, ラフォーレ修善寺
Laforet Shuzenji, Izu, Oct. 14–15, 2005

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

共催 社団法人日本ロボット学会 ロボット聴覚研究専門委員会
Robotics Society of Japan, Research Committee on Robot Audition

【10 月 14 日】 (続き)

【10 月 15 日】

信号処理から見たロボット聴覚:「音源の方向検出について」

Sound Source Localization;
Robot Audition System from the Signal Processing Point of View

金田 豊 (東京電機大学工学部)

* Yutaka KANEDA (Tokyo Denki Univ.)

kaneda@c.dendai.ac.jp

Abstract— This paper describes techniques for sound source localization. At first, commonly used techniques are explained. The methods are time delay estimation, cross-spectrum phase method, MUSIC, and so on. Then, characteristics of these methods are discussed. Problems encountered in practical application are also discussed. Finally, a possible solution for the localization error under reverberant room condition is introduced. Some experimental results are demonstrated.

1．はじめに

音源方向の検出は、ロボット技術において大変重要な役割をはたしている。例えば、ロボットが人間と会話をする時、ロボットは話者の方を向いて会話を行うことが基本的である。そのためには、話者（音源）の方向（または位置）を検出することがまず必要となる。

周囲に不要音が存在する音環境において、所望音声の方向を知ることができれば、指向性受音器（マイク）などを用いた選択収音を効果的に行うことができる。所望音声を高い SN 比で受音できれば、音声認識性能も向上し、人間－ロボットの対話を円滑化に寄与が大きい。また一方、話者方向を特定することにより、ロボットは発声者の画像を得ることができる。その結果、遠隔地へ話者の姿を伝えたり、話者の認識やセンサフュージョン的な処理へと進めることができる。

人間は音源の方向を判断する際には、聴覚だけではなく視覚も利用しているし、また、より高次の知能処理も利用している。その結果、聴覚では間違いの多い前後判断や、高騒音環境での方向検出も可能としている。ロボット聴覚においても、そのようなセンサフュージョン的処理や高度な知能処理を組み込むことは必須であると思われる。しかし本稿ではそれらの基本技術として、聴覚系のみの処理に限定して話を進めることにする。

波の到来方向 (DOA:Direction-of-Arrival) の検出は、レーダやソナーの基本技術として古くから研究がなされてきた[1][2]。可聴音波の到来方向検出を行うロボットへの応用では、レーダやソナーとは異なった点もあるが、基本的には同様な手法が適用されている。

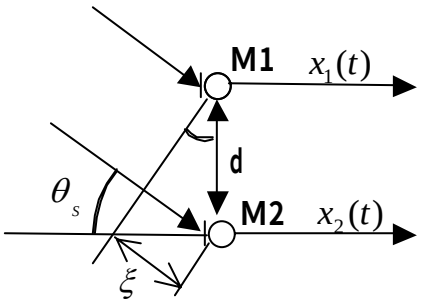


Fig.1 Sound wave and microphones.

本論文では、まず、従来の音源方向検出の代表的手法について概説する。続いて、これらの手法の定性的性質やアンテナ分野との相違について議論をする。そして、これらの手法を実環境で応用する際の問題点を議論する。最後に、問題点の一つである室内反射音の悪影響を回避する一つの方法を説明し、実験により、その効果を示す。

２．代表的な音源方向検出技術

音の空間的情報を得るには、複数のマイクロホン（マイクロホンアレー）を利用するのが通常である。音源検出手法は複数のマイクロホンによる受音信号の間の時間差に基づく方法、および複数の信号間の統計量（相関行列）に基づく方法に大別できる。

2.1 時間差に基づく方法

最初に単一音源を仮定して考える。図１は、音波を２つのマイクロホンで受音している様子を示している。音波は図の左上方向から平面波として到来している。到来方向は２つのマイクロホンを結んだ線に垂直な線を基準として θs で表している。このとき、音波はまずマイクロホン M1 で受音され、時間 τs 遅れて、マイクロホン M2 で受音される。よって、マイクロホン M1 で受音される信号 $x_1(t)$ を $s(t)$ と表すと、マイクロホン M2 で受音される信号 $x_2(t)$ は、

$$x_1(t) = s(t) \quad (1)$$

$$x_2(t) = s(t - \tau_s) \quad (2)$$

と表される。この時間差 τs は、音波が図の距離 ξ を

進む時間であり、

$$\tau_s = d \sin\theta_s / c \qquad (3)$$

d: マイク間距離、c: 音速

と表される。これより、時間差τsを求めることができれば、音源方向は、

$$\theta_s = \sin^{-1}(c\tau_s / d) \qquad (4)$$

と、求めることができる。

２つの信号に含まれる成分の時間差（遅れ時間）を推定する方法はTDE(Time Delay Estimation)[3]と呼ばれ、さまざまな方法が提案されているが、代表的な方法を以下に示す。

(a) 相互相関関数$\phi_{12}(\tau)$

次式で定義される相互相関関数を利用する。

$$\phi_{12}(\tau) = \int x_1(t)x_2(t+\tau)dt \qquad (5)$$

式(1)(2)を代入すると、

$$\phi_{12}(\tau) = \int s(t)s(t-\tau_s+\tau)dt = \phi_{ss}(\tau-\tau_s) \qquad (6)$$

ただし$\phi_{ss}(\tau)$は、次式で表される信号$s(t)$の自己相関関数である。

$$\phi_{ss}(\tau) = \int s(t)s(t+\tau)dt \qquad (7)$$

よく知られているように、$s(t)$が周期関数でなければ、自己相関関数$\phi_{ss}(\tau)$は$\tau=0$で単独の最大値をとる。よって相互相関関数$\phi_{12}(\tau)=\phi_{ss}(\tau-\tau_s)$は、$\tau=\tau_0$で最大値をとる。

以上のことより、相互相関関数$\phi_{12}(\tau)$を計算し、$\phi_{12}(\tau)$が最大値をとるτの値を求めれば、それが求めていた時間差τsとなる。

(b) 一般化相関関数 (generalized correlation) [4]

雑音や複数音源などの条件下で性能を確保するために、相関関数にさまざまな周波数重みをつけることが提案されている。これを一般的に表したものが一般化相関関数$R_{12}(\tau)$で、次式で表される。

$$R_{12}(\tau) = \int \psi(\omega)\Phi_{12}(\omega)e^{j\omega t}d\omega \qquad (8)$$

ただし、$\Phi_{12}(\omega)$は、相互相関関数$\phi_{12}(\tau)$のフーリエ変換（＝信号$x_1(t)$と$x_2(t)$のクロススペクトル）である。式(8)は、相互相関関数$\phi_{12}(\tau)$の周波数成分である$\Phi_{12}(\omega)$に$\psi(\omega)$で重み付けをした後、逆フーリエ変換をして時間軸に戻したものである。

この重み関数$\psi(\omega)$としては、いくつかの評価基準に基づいたものが提案されているが、定性的には、

- SN比の大きい帯域を強調し、雑音が優勢な帯域を抑制する。
- 周波数白色化を行って相関関数のピークを際立たせる

の２つの考えがある。ここでは、前者の代表としてSCOT(Smoothed Coherence Transform)とPHAT(Phase Transform)を説明する。（注：これらの方法は他の名称でも呼ばれているが、ここでは、文献[4]に従った）

SCOT

$$\psi_s(\omega) = 1/\sqrt{\Phi_{11}(\omega)\Phi_{22}(\omega)} \qquad (9)$$

ただし、$\Phi_{11}(\omega)\Phi_{22}(\omega)$はそれぞれ信号$x_1(t)$および$x_2(t)$のパワースペクトルを表す。これを式に代入すると、

$$R_{12}(\tau) = \int \frac{\Phi_{12}(\omega)}{\sqrt{\Phi_{11}(\omega)\Phi_{22}(\omega)}} e^{j\omega t}d\omega \qquad (10)$$

となる。この式の被積分項$\gamma_{12}(\omega)$

$$\gamma_{12}(\omega) = \frac{\Phi_{12}(\omega)}{\sqrt{\Phi_{11}(\omega)\Phi_{22}(\omega)}} \qquad (11)$$

は、コヒーレンス関数となっている。コヒーレンス関数は、該当する周波数成分のSN比が大きい場合には、その絶対値が１に近く、SN比が小さい場合には０に近い値をとる。（ただし、２つのチャンネルに含まれる雑音は無相関と仮定）よって、高いSN比で時間差情報を含んだ周波数帯域の強調を行うことになる。

PHAT

$$\psi_p(\omega) = 1/\left|\Phi_{12}(\omega)\right| \qquad (12)$$

これを式に代入すると、

$$R_{12}(\tau) = \int \frac{\Phi_{12}(\omega)}{\left|\Phi_{12}(\omega)\right|} e^{j\omega t}d\omega \qquad (13)$$

となり、被積分項は、クロススペクトルの振幅を平坦化し、位相項のみを表す。相関関数が白色化されるので相関関数のピークが先鋭化され、複数音源の分離性能が向上する。この方法は、白色化相互相関[5]、CSP[6]、などとも呼ばれている。

白色化の際にSN比の悪い帯域も持ち上げるため、SN 比の悪い帯域を含んでいる場合には雑音の影響を受けやすいことが予想されるが[4][7]、この方法を利用している報告は多い。またこの方法は、反射音の影響を受けにくいという報告[7]もあるが、定量的な検証が必要であると考えている。

(c) クロススペクトルの位相特性

時間差τs を持った信号$s(t)$と$s(t-\tau_s)$のクロススペクトル$\Phi_{12}(\omega)$ は、

$$\Phi_{12}(\omega) = \Phi_{ss}(\omega)e^{-j\omega\tau_s} \qquad (14)$$

と表される。$\Phi_{ss}(\omega)$は信号$s(t)$のパワースペクトルで、実数であるので、$\Phi_{12}(\omega)$の位相特性$\varphi(\omega)$は、

$$\varphi(\omega) = \omega\tau_s \qquad (15)$$

となる。この特性の傾きより時間差τsが得られる。

2.2 相関行列に基づく方法

M個のマイクロホンで受音する事を考える。θ方向にある音源から各マイクロホンまでの伝達関数を $G_i(\omega,\theta), i=1,2,\cdots M$ と表す。このとき、周波数ωを固定して考えて、方向ベクトル$\mathbf{d}(\theta)$を次式で定義する。

$$\mathbf{d}(\theta)=[\,G_1(\omega,\theta), G_2(\omega,\theta),\cdots, G_M(\omega,\theta)\,]^T \tag{16}$$

この$\mathbf{d}(\theta)$は、1つのマイクまでの伝達関数$G_i(\omega,\theta)$を基準とした相対的特性

$$\mathbf{d}(\theta) \;\rightarrow\; \mathbf{d}(\theta)/G_i(\omega,\theta) \tag{17}$$

と置き換えても良い。

ここで、簡単のため、各マイクロホンで受音される音の大きさは等しい、すなわち、

$|G_1(\omega,\theta)|=|G_2(\omega,\theta)|=\cdots=|G_M(\omega,\theta)|$ が成立すると仮定し、$\mathbf{d}(\theta)$の各項を$G_1(\omega,\theta)$で除したものを改めて$\mathbf{d}(\theta)$とすると、

$$\mathbf{d}(\theta)=[1, e^{-j\omega\tau_2(\theta)}, e^{-j\omega\tau_2(\theta)},\cdots, e^{-j\omega\tau_M(\theta)}\,]^T \tag{18}$$

と表される。このとき、$\omega\tau_i(\theta), i=2,3,\cdots M$ は第ｉ番目のマイクと１番目のマイクの位相差を、$\tau_i(\theta)$は時間差を表している。

各マイクロホンの受音信号を短時間周波数分析したものを $X_i(\omega,t), i=1,2,\cdots M$ と表す。入力信号ベクトル$\mathbf{x}$を

$$\mathbf{x}=[X_1(\omega,t), X_2(\omega,t),\cdots, X_M(\omega,t)]^T \tag{19}$$

と定義し、入力相関行列$\mathbf{R}_x$を次式で定義する。

$$\mathbf{R}_x=E[\mathbf{x}\mathbf{x}^T] \tag{20}$$

ただし、$E[\cdot]$は期待値を表すが、実際的には時間平均で計算する。音源方向検出はこの方向ベクトル$\mathbf{d}(\theta)$および入力相関行列$\mathbf{R}_x$を用いて行われる。以下にその中の代表的手法を示す。なお、スペースの関係で各手法導出の詳細は省略するので文献[8]などを参照されたい。

(a) DSA（Delay-and-sum-array：遅延和法）

θ方向から到来する音の遅延量（時間差）とは正負逆の遅延を、各チャンネルの受音信号に印加する。その後、総和をとった信号の二乗和を、その方向から来る音の推定量とする。具体的には、到来音パワー推定量$P_D(\theta)$は、入力信号ベクトル$\mathbf{x}$に対して、

$$P_D(\theta)=E[(\mathbf{d}(\theta)^*\mathbf{x})^2]=E[\mathbf{d}(\theta)^*\mathbf{x}\mathbf{x}^*\mathbf{d}(\theta)] \\ =\mathbf{d}(\theta)^*E[\mathbf{x}\mathbf{x}^*]\mathbf{d}(\theta)=\mathbf{d}(\theta)^*\mathbf{R}_x\mathbf{d}(\theta) \tag{21}$$

と表される。ただし、*は転置共役ベクトルを表す。

(b) MV（最小分散法）

θ方向以外から到来する音成分を最小化する適応型アレーの出力をその方向から来る音の推定量

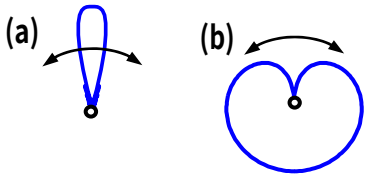


Fig.2 Steering　(a)beam, (b) null.

$P_M(\theta)$とする方法。$P_{MV}(\theta)$は、

$$P_{MV}(\theta)=\frac{1}{\mathbf{d}(\theta)^*\mathbf{R}_x^{-1}\mathbf{d}(\theta)} \tag{22}$$

と表される。

(c) MUSIC

入力相関行列$\mathbf{R}_x$の固有値を計算し、大きいほうから、想定される音源の数(K個) の固有値を取り除く。残った固有値に対応する固有ベクトルを$\mathbf{v}_q$，ｑ＝K+1, ･･･, M と表したとき、次の行列$\mathbf{R}_n$を定義する。

$$\mathbf{R}_n=\sum_{q=K+1}^{M}\mathbf{v}_q\mathbf{v}_q^* \tag{23}$$

この行列$\mathbf{R}_n$を用いて、具体的に、到来音パワー推定量$P_{MS}(\theta)$は、

$$P_{MS}(\theta)=\frac{1}{\mathbf{d}(\theta)^*\mathbf{R}_n\,\mathbf{d}(\theta)} \tag{24}$$

と表される。

３．音源方向を推定する２つの考え方

指向性受音器を用いて音源方向を検出する場合、次の２つの方法が考えられる。図 2(a)に示すような鋭い指向性（beam）を利用する方法と、図 2(b)に示すような死角（感度がゼロの方向：null）を用いる方法である。

鋭い指向性は、ある特定の方向の音だけを受音するので、出力はその方向の音の大きさを反映する。よって、指向性の向きを、対象とする範囲で変化させれば（beam-steering）、音源方向が推定できる。一方、死角の方向を変化させ（null-steering）た時には、死角方向と音源方向とが一致すると、出力が減少するので、音源方向が推定できる。

鋭い指向性は遅延和アレーで実現できる。これはアレーの加算処理である。また、死角の制御はアレーの減算処理で実現できる[9]。このように、アレーの加算処理および減算処理を基本として音源方向の検出を行うことができる。

(a)加算形の方向検出

２つのマイクの受音信号を $x_1(t), x_2(t)$ とする。

$x_2(t)$ に遅延τを付加して加算して得られる遅延和アレーの出力は $x_1(t)+x_2(t-\tau)$ で、そのパワー（二乗期待値）は、

$$E[\{x_1(t)+x_2(t-\tau)\}^2]$$
$$= E[x_1{}^2(t)]+E[x_2{}^2(t)]+E[x_1(t)\,x_2(t-\tau)]$$

となる。第 1,2 項は、τの値によらない定数項であり、第 3 項は相関関数を表している。すなわち、２マイクロホンの遅延和アレーの出力のパワーは、定数項を除けば、相互相関関数と一致する。

よって、相関関数は加算形の処理に属することがわかり、遅延和アレーと共通の性質をいくつか持っている。

(b) 減算形の方向検出

多少強引な分類であるが、MV 法や MUSIC などの「高分解能法」とよばれる方法は減算形に属する。指向性の谷はビームより狭角度に作れるので、到来音の推定も鋭いピークで行うことができる。ただし、マイクロホンの数 M に対して指向性の谷はM－１個しか作れないので、ある周波数における音源数がマイクロホンの数を上回ると性能は低下することが理解できる。

４．アンテナ分野との相違

２節で述べた基本技術の多くは、アンテナなどの分野で開発された技術である。電波も音波も、波という意味では同一なので、それらの技術がスムースに導入されてきた。しかし、アンテナの分野では信号は狭帯域であるのに対しわれわれの分野（以下、音響分野）では信号は 10 オクターブもの広帯域信号である。その結果、アンテナの分野の常識とは若干異なることもある。

一例として、空間のサンプリング定理と呼ばれるものがある。これは、マイクロホン間隔が空間波長の半分を超えないようにと定めるものである。半分を超えると、空間的折り返し、すなわち、目的方向以外にも複数のビームを持つ指向性が形成されてしまう。その結果、到来音の推定結果にも誤ったピークが発生してしまう。人間の聴覚でもこのサンプリング定理が満たされない 1500Hz 以上の周波数での方向検出には位相差（時間差）情報を利用しないことも、このことの重要性を意味しているように思える。

しかし周知のように、広帯域信号の方向検出において、必ずしも空間サンプリング定理を満足する必要はない。正しい音源方向は周波数によらず一定であるのに対して、折り返しによって現われる誤ピークの方向は、周波数によって変化するので、周波数ごとに求めた推定結果を周波数で平均化すれば、誤ピークの影響を軽減することができる[安倍]。

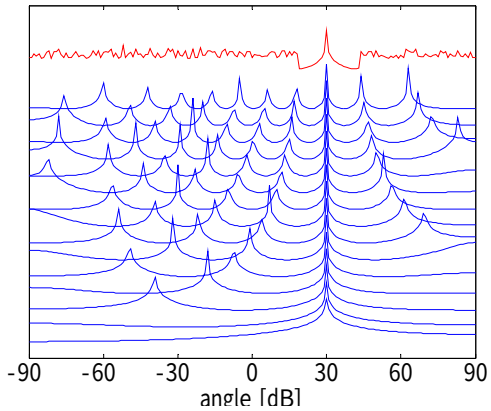


Fig.3 Beam aliasing and frequency averaging.

図 3 は、このことを示したシミュレーション結果である。30 ° 方向から到来する白色雑音を 2 マイクで受音し、最小分散法で方向推定を行った。各線は、下から、100Hz, 300Hz, 500Hz, ･･･ と 200Hz おきの周波数成分の結果である。どの周波数においても 30 ° 方向は正しく推定されている。しかし、マイク間距離が 60cm と大きめであるため、500Hz 以上の推定結果においては、空間折り返し誤差による誤推定のピークが見られる。誤推定ピークは正しいピークと同程度の大きさを持っている。

この折り返しの誤ピークは周波数によって発生する方向（横軸上の位置）が推移していくことがわかる。そして、それらを周波数軸上で平均した結果を最上段の赤線として示した。実際の音源方向以外は平均化されて小さな値となり、音源方向の推定が可能となることがわかる。

広い意味で、推定結果を周波数ごとに求めて平均化する方法には、上記の直接的な方法のほかにもさまざまな形態があり[朝の]、方向推定では代表的な考え方であるといえる。

なお、選択収音技術においては、折り返しピークによって、所望方向以外の音を拾ってしまうと、品質の劣化につながるため、折り返しは無視できない場合も多い。

５．実環境における課題

ここまでは、理想的な条件で方向検出技術を考えてきたが、実環境で音源方向推定をする際にはさまざまな条件が追加されてくる。以下にその主なものを述べる。

1) 周囲騒音（雑音）

2) 反射音
3) 複数音源
4) 音源の移動
5) 近距離音場

1) 周囲騒音（雑音）

最も基本的な問題であり、アンテナなどにおいても十分考慮されている。しかし、他分野では主としてセンサ雑音に注意が払われ、複数の信号に含まれる雑音は無相関として扱われる場合が多い。その場合、相関行列では微小な対角成分として扱われる。しかし、音響の分野で、特に低周波数では、雑音成分であってもマイク間で相関が高い。したがって、それに応じたモデルをたてるなどの配慮が必要な場合も有る。

2) 反射音

反射音は目的音と相関の高い不要音で、目的音源方向とは異なる方向から到来して推定結果に大きな影響を及ぼす。アンテナの分野でも同様な問題が存在し、少数の反射音に対しては、相関行列を空間的に平均することで、相関除去を行う方法などが提案されている[2]。しかし、しかし、実際の室内で、音源－マイク間距離が大きくなると多数の反射音が影響を及ぼすようになるので、そのような手法での解決は困難である。

もちろん、音源－マイク間距離が 1m 程度であれば、反射音の影響は小さく、また、ロボットと人間との対話距離もその程度が現実的ではある。しかし、数 m 離れた場所におけるイベントの検出や、近くに反射物体がある場合などを考えると、反射音の影響の軽減は望まれる。

反射音の影響を取り除く自然な考え方は、反射音が到達する前の、反射音の影響を受けていない直接音部分に注目して方向検出を行うことである。具体的には、閾値処理により信号の初期部分を切り出して方向検出を行う方法[11]、ピークホールド処理により後続する反射音をマスクすることでその影響を軽減する方法[12][13]などが提案されている。ピークホールド処理の実験例については、６節で紹介する。

3) 複数音源

MV 法や MUSIC などの減算形の処理では、音源数はマイクロホン数－１ 以下である必要がある。しかし、加算形の処理ではそのような制約は無い。

図 4 は、-45 ° 0 ° 45 ° の３方向から到来する３つの音声を 60cm 離した２つのマイクで受音し（シミュレーション）、相関関数により方向推定した結果である。相関の平均時間が長く（2 秒）音声の周期性の影響を回避しているので、２つのマイクであっても３つの到来方向が明確に検出できている。

一方、図 5 の太青線は、マイク長を 20cm とした時の結果である。45 ° の音源方向はほぼ検出できているが、-45 ° 0 ° 方向の検出は失敗している。同図に、

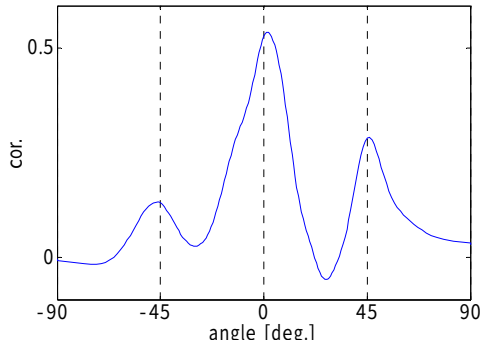


Fig. 4 Cross-correlation between 2 micropohne output (microphone distance is 60cm).

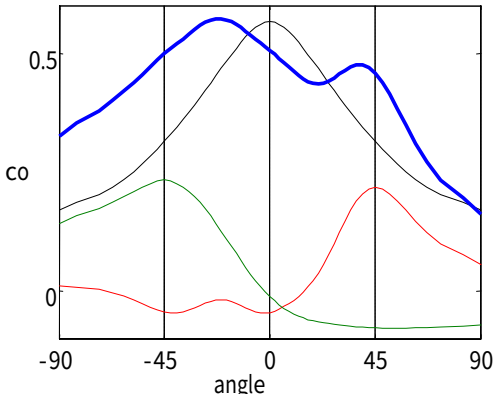


Fig. 5 Cross-correlation between 2 micropohne output (microphone distance is 20cm).

各音声が単独で存在する場合の結果を細線で示した。それぞれの曲線は-45 ° 0 ° 45 ° でピークを持っており、正しい方向が検出できている。太線の 25 ° 付近のピークは、-45 ° 0 ° の２つ音源のピークが融合してできたピークであることがわかる。

このことより、相互相関関数を用いる方法では、単一音源は良好に検出できるが、複数の音源がある場合は、複数のピークが融合して誤った結果となる場合がある。この問題を解決する方法のとしては、白色化相互相関を用いたピークの先鋭化が有効と考えられる。また、第 2 の方法としては、マイク間距離を広くすることも有効である。相関法は遅延和アレーと同様なので、マイク間距離を広げることはアレーの全長を広くすることになり、指向性ビームを鋭くすることになるので、各音源に対応する相関波形を先鋭化できる。

4) 移動音源

移動音源を対象とする場合には、固定と見なせる音源と違って、短時間のデータに基づいた推定が必要とされる。そしてその結果、推定誤差がより多く発生する。そのような問題に対しては、Kalman フィ

ルタによるスムーシングや、Particle フィルタによるトラッキングなどの適用が試みられている。詳細は、昨年度の本研究会の講演論文[14]を参考にされたい。

5) 近距離音場

アレーサイズに比べて音源からの距離が近い場合は、近距離音場と考えられ、波面も平面波ではなく球面波のモデルを導入する必要な場合もある。

また、近距離音源に対しては、音源の「方向」ではなく、「位置」の検出が求められる場合がある。複数のアレーで「方向」を検出し、その方向の交点として「位置」を求めるのが通常である。この方法ではしかし、複数の音源が存在する場合では必要以上の「位置」の候補が発生し、その対策が検討されている。[15]

6) その他

本稿の枠組みでは紹介し切れなかったが、ロボットの HRTF を用いたパターン認識的なアプローチ[16][17]や、演算量低減のためのアプローチ[18]、音声の時間－周波数領域におけるスパース性の利用（ある時間のある周波数成分を見れば、複数音声でも単一音声と見なせる）[19][20]、など興味深い報告がなされている。

６．ピークホールド処理を用いた音源方向推定

最後に、反射音の影響を軽減する目的で検討を進めているピークホールド処理[21]について紹介する。

6.1 ピークホールド処理の考え方

図6にピークホールド処理の模式図を示す。(a)は受音信号波形を表す。τ_sの時間差をもった直接音のほかに、異なった時間差を持った反射音が含まれる。(b)は図(a)の信号$x_1(t), x_2(t)$の相互相関関数を表す。反射音の影響で多数のピークを持っており、雑音の影響などでτ_s 以外のピークが誤って検出されることがある。

この反射音の影響を軽減するために直接音にピークホールド処理を行って、後続する反射音をマスクする。図(c)は図(a)にピークホールドを行った波形を表す。その後、ピークホールド波形の時間差分をとって（図(d)）、(e)はそれらの相互相関関数を表す。図(e)より、反射音の影響が除去され、直接音の時間差τ_sが明確になっていることが分かる。

なお、複数回の発音に対応可能とするためには、ピークホールド値に，室内残響と同程度の指数減衰を持たせるものとする。

6.2 対数処理

複数の初期反射音が近接した時刻に到着する場合、図 7 に示すように反射音のパワーが直接音のパワーより大きくなることがある。この影響を軽減するために、図 8 のようにピークホールド処理をした波形に、対数操作を行う。図の数値例で示すように、ピ

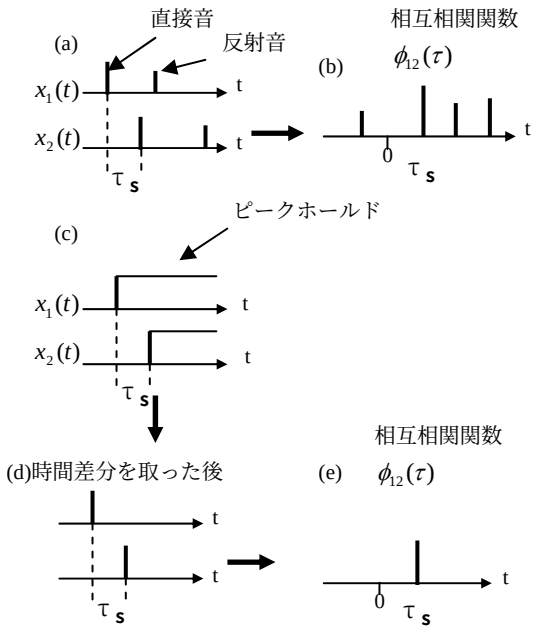


Fig.6 Peak-hold processing.

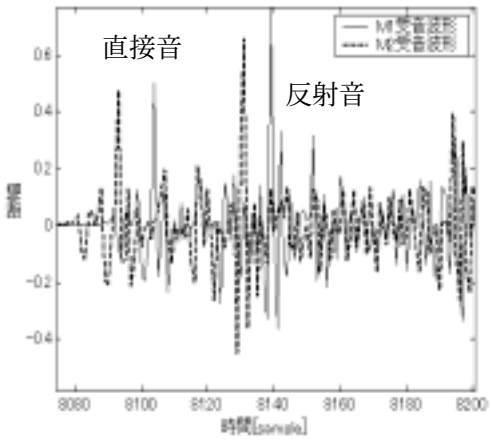


Fig.7 Direct and reflected sound.

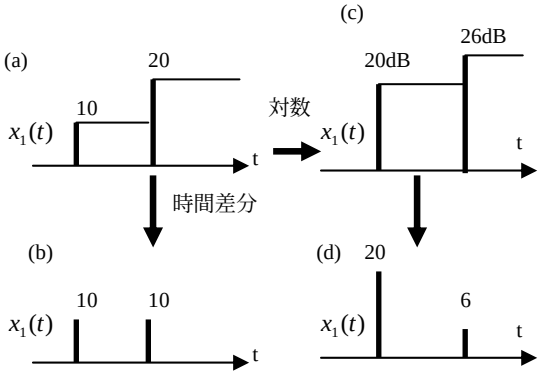


Fig.8 Logarithmic operation.

ークホールド処理をした波形に対数をかけると、後続の振幅の大きい反射音の影響を軽減することが期待できる。

6.3 立ち上がり検出と相関処理

図9に拍手音を２マイクで受音し、その信号のピークホールド波形を対数値として表した例を、青・赤の線で示す。雑音がほぼ定常であると仮定し、短時間T[ms]でAth[dB]以上の振幅変化があった時刻を音源信号の立ち上がりとみなした。（今回は拍手音を仮定しているのでT=10ms, Ath=10dBとした）図9よりわかるように、反射音の影響はほぼマスクされているので、立ち上がり時刻の検出にあまり精度は要求されない。そして、立ち上がり時刻の前後数10ms程度の信号を切り出し、時間差分を取った後、相互相関関数を計算して、時間差τ_sを検出した。

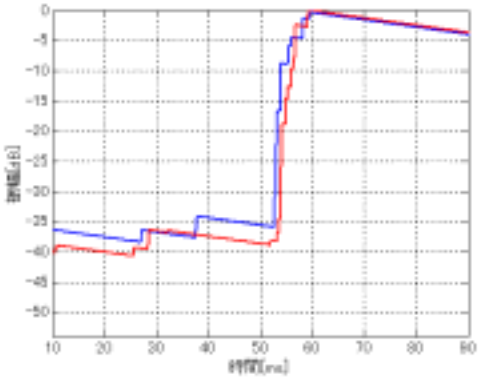


Fig.9 Example of peak-hold wave form.

6.4 音場実験

この手法の反射音に対する効果を確認するために、反射音の影響が大きいと考えられる図10の音場条件で、音源方向検出実験を行った。音源は拍手音を用い、マイクロホン間隔dは0.6mとした。実験室（5.0×9.0×2.4[m]、残響時間400ms）のマイクロホンの近くには反射板を図のように設置した。音源－マイクロホン距離は1,2,4mとした。

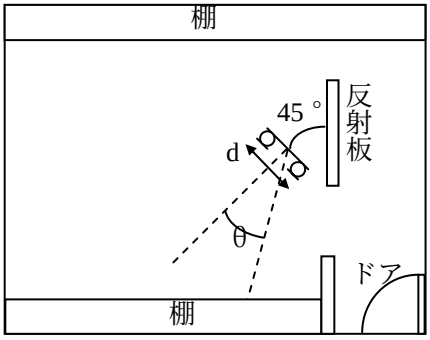


Fig.10 Experimental set up.

6.5 実験結果

２つの受音信号波形から、相互相関関数を計算する方法を従来法とし、ピークホールド法と比較を行った。

図11は、従来法による結果を表すもので、音源距離が2,4mの時に正しい角度を検出できていない。これは、音源が遠ざかることで, 直接音が小さくなり、反射板からの反射音の影響が大きくなったためと考えられる。これに対して図12のピークホールド法の結果を見てみると、音源距離が大きくなった場合にも、ほぼ正しい角度を検出していることがわかる。今回の結果は拍手音という衝撃的な音が対象ではあるが、ピークホールド法を用いると、反射音の影響を軽減して音源方向検出の性能が向上できる可能性を示すものと考えられる。

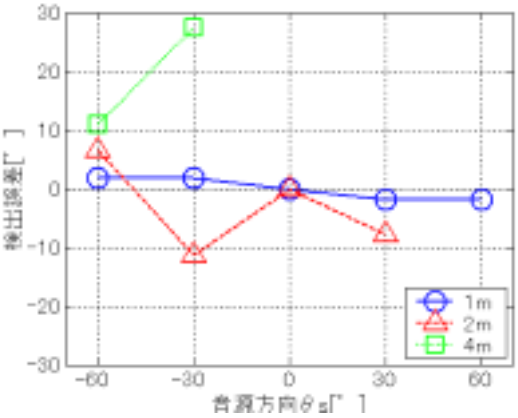


Fig.11 Experimental result (conventional).

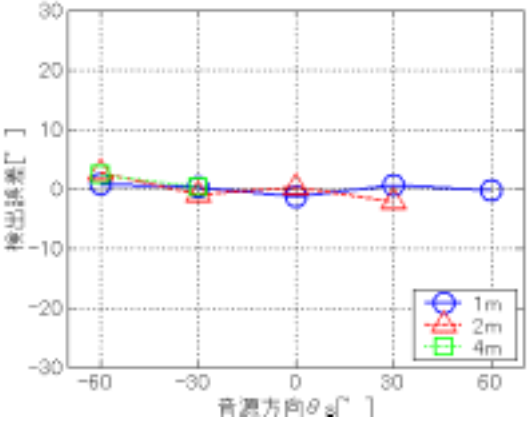


Fig.11 Experimental result (with peak-hold).

7．むすび

本稿では音源方向検出の基本的技術について概説した。本稿で述べた技術は、「ロボットの耳」の末端のボトムアップ処理に相当する。

現実の環境での音源検出においては、本稿で書いたようにさまざまな変動要因、妨害要因が予想される[16]。その結果、「耳の処理」だけで完全な音源方向検出を達成するのは困難と考えられる。そこで、画像による話者認識処理と組み合わせたり、学習に基づいた知能判断処理を組み込んだり、またそれらに基づいたトップダウン的な「耳の処理」を導入し

たりと、統合的な処理と判定メカニズムの構築が重要と考えられ、その検討が進められている[16][22]。

そのような中で、「耳の処理」単独の性能は限界近くまで開発されているのではないかとも考えられる。ただしかし、各種の方法には、さまざまな音響環境・用途に応じた得意不得意があって、それを環境と対応付けて明確にしていくことは残された課題であると考えている。そして、一つの処理方式ですべての状況に対応するのではなく、状況に応じて最適な処理方式を選択、利用していく処理形態が「末端の耳の処理」としての望ましい形だと考える。

参考文献

[1]S.U.Pillai, “Array signal processing,” New York, Springer Verlag, 1989.
[2]菊間, “アレーアンテナによる適応信号処理”, 科学技術出版 (1998)
[3] G. C. Carter, Ed., “Special issue on time delay estimation,” IEEE Trans. Acoust. Speech, Signal Processing, ASSP-29, June (1981).
[4] C. H. Knapp and G. C. Carter: “The generalized correlation method for estimation of time delay,” IEEE Trans. on Acoust., Speech and Signal Proc., ASSP-24, 4, pp.320-327 (1976).
[5] 林, 他:”信号の白色化による航空機騒音識別手法の改良とハードウェアによる実現,” 信学技法EA89-38 (1989).
[6] M. Omologo, et al. : “Acoustic event localization using a crosspower-spectrum phase based technique,”IEEE ICASSP94, II-273-276 (1994).
[7] M. S. Brandstein: Time-delay estimation of reverberated speech exploiting harmonic structure,” J. Acoust. Soc. Am., 105, 5, pp.2914-2929 (1999).
[8]大賀, 他:”音響システムとディジタル処理”, pp. 197-208, 通信学会, (1995).
[9]金田:”マイクロホンアレーによる指向性制御,” 音響学会誌, 51, 5, pp.390-394 (1995).
[10]永田、他:”多数センサによる音源位置の推定,” 音響学会誌, 46, 7, pp.531-540 (1990).
[11]黄、他:”生体に示唆を得た音源定位システム－反響のある環境での単一音源定位,” 信学論(A), J71-A, 10, pp.1780-1789 (1988).
[12]金田:”室内残響下における広帯域音源の方向推定”, 日本音響学会講演論文集(秋), pp.547-548, (1991.10).
[13]小林、他:”雑音と反射音に対してロバストな話者方向推定法,” 日本音響学会講演論文集(春), pp.535-536 (2001.3).
[14]浅野、他:”マイクロホンアレイを用いた移動音源の追跡と分離について,” 第20回AIチャレンジ研究会資料, pp.1-8 (2004).
[15]西浦、他:”マイクロホンアレーを用いたＣＳＰ法に基づく複数音源位置推定,” 信学論(A), J83-D-Ⅱ, 8, pp.1713-1721 (2000).
[16]奥乃、中臺:”ロボット聴覚の現状と課題,” 日本音響学会講演論文集(春), pp.633-636 (2005.3).
[17]小林:”ロボットに搭載したマイクロホンによる音像定位･音源分離,” 日本音響学会講演論文集(春), pp.637-640 (2005.3).
[18] M. Sato, et al.: ”Near-field sound-source Localization based on signal binary,” IEICE Trans. Fundamentals, ESS-A, 8, pp.2078-2085 (2005).
[19] 井原、他:”周波数振分けによるマルチチャンネル混合音声の分離と音源定位,” 信学論(A), J86-A, 10, pp. 998-1009 (2003).
[20] 陶山、他:”2段階のデータ選別による複数音源定位,” 信学論(A), J79-A, 6, pp. 1127-1137 (1996).
[21] 木皿、他:”拍手音に対するピークホールド音源検出法の有効性について,” 日本音響学会講演論文集(秋), (2005.9).
[22] 浅野、他:”音響と画像の情報統合を用いた話者追跡と音源分離,” 第18回AIチャレンジ研究会資料, pp.19-26 (2003).

SIMO-ICA を用いた音響テレプレゼンスのためのブラインド音情景分解

Sound Scene Decomposition for Audio Tele-Presence Using SIMO-ICA

高谷智哉, 猿渡洋, 鹿野清宏

Tomoya TAKATANI, Hiroshi SARUWATARI, Kiyohiro SHIKANO

奈良先端科学技術大学院大学 情報科学研究科

Nara Institute of Science and Technology, Graduate School of Information Science

{tomoya-t, sawatari, shikano}@is.naist.jp

http://www.aist-nara.ac.jp/~tomoya-t/index.html

Abstract

In this paper, we address a blind decomposition problem of binaural mixed signals observed at the ears of humanoid robot, and we introduce a novel blind signal decomposition algorithm using Single-Input Multiple-Output-model-based ICA (SIMO-ICA). The SIMO-ICA consists of multiple ICAs and a fidelity controller, and each ICA runs in parallel under the fidelity control of the entire separation system. The SIMO-ICA can separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones in the robot ear. Thus, the separated signals of SIMO-ICA can maintain the spatial qualities of each sound source, i.e., they represent the decomposed sound scenes. Obviously the attractive feature of SIMO-ICA is highly applicable to not only speech recognition but also, e.g., humanoid-robot-based auditory tele-existence technology. The experimental results reveal that the spatial quality of the separated sound in SIMO-ICA is remarkably superior to that of the conventional method, particularly for the fidelity of the sound reproduction.

1 はじめに

ブラインド音源分離 (BSS) は各入力チャネルで観測された観測信号の情報のみを用いて音源信号を推定する手法である．近年，独立成分分析 (ICA)[1] を用いた BSS が主流となっており，音響信号の混合過程に相当する畳み込み混合の分離を取り扱った手法 [2, 3, 4] が多く提案されている．これらの手法は干渉音の抑圧を行う事が可能であるが，独立な各音源信号をモノラル信号として抽出する．その結果, 出力信号は各音源の定位感や残響感などの空間的性質を失っている．従って，その性質を必要としないハンズフリー音声認識やハンズフリー音声通話システムへの応用が期待されるが，バイノーラル信号処理 [5], 高品質な音場再現システム [6], またそれらを必要不可欠な基礎理論とするヒューマノイドロボットを用いた音響 tele-presence (tele-existence)[7, 8] システムへ応用することは困難である．

上述の問題点を改善するために, 我々は混合された音響信号をそれらの各要素である Single-Input Multiple-Output (SIMO) モデルに基づく信号に分解する SIMO モデルに基づく ICA (SIMO-ICA) [9, 10, 11, 12] の研究を行ってきた．ここで, SIMO モデルとは単一の音源からの信号を複数点で受音する伝達系のことを言い，SIMO モデルに基づく信号とは SIMO モデルで観測される信号群のことを言う．SIMO-ICA は複数の ICA 部と単一の fidelity controller により形成され, 各 ICA は分離システム全体の音質制御の下で並列に動作する．SIMO-ICA は観測信号から音源信号をモノラル音として推定するのではなく, 観測信号を SIMO モデルに基づく信号群に分解する．従って, SIMO-ICA の出力信号は各音源の定位感や残響感などの空間的性質を維持したアレー信号であり, この信号により個々の音源毎の音情景を表現することが可能である．我々はこの特徴をいかして，次のような統合手法の研究も行ってきた．

統合法 1 SIMO-ICA とブラインド MINT 法を統合したブラインド音源分離・残響抑圧法 [13]

統合法 2 SIMO-ICA と適応ビームフォーミングを統合したブラインド音源分離法 [14]

統合法 3 SIMO-ICA とバイナリーマスキング処理を統合したリアルタイムブラインド音源分離法 [15]

統合法 1 では，音声などの有色信号を音源とする等決定問題 (音源数＝マイクロホン数) において，干渉音の抑圧だけでなく，伝達系の影響を除去することを実現している．また，統合法 2 においては，SIMO-ICA により得られた出力信号群に適応ビームフォーミングを適用し，各手法単体より更なる干渉音の抑圧に成功している．更に，統合法 3 においては，リアルタイム性の高い信号処理を組み合わせることにより，リアルタイム高性能ブラインド音源分離が可能となっている．各手法の詳細については参考文献を御覧いただきたい．

本稿では，SIMO-ICA アルゴリズムを用いて，ロボット頭部の影響を含む混合バイノーラル音を単一音源からのバイノーラル音に分解する音響 tele-presence システムを提案する．提案システムの出力信号は，SIMO-ICA の特長より，音源方位などの空間情報が含まれている．残響環境下において head and torso simulator (HATS) を用いて収録した混合バイノーラル音を用いた音情景分解実験より，SIMO-ICA の性能は従来法より優れていることが示された．また, 出力信号の両耳間時間差 (ITD) 及び両耳間レベル差 (ILD) の結果より，SIMO-ICA の出力信号には各音源の定款や残響感などの空間的性質が維持されていることが確認された．

2 混合過程と従来 BSS 法

2.1 混合過程

本研究では, マイクロホン数 K を 2, 音源数 L を 2 とする．一般に，音源信号が線形に混合された観測信号は以下で表される．

$$\begin{aligned}\boldsymbol{x}(t) &= \sum_{n=0}^{N-1}\boldsymbol{a}(n)\boldsymbol{s}(t-n)=\boldsymbol{A}(z)\boldsymbol{s}(t)\\ &= \begin{bmatrix} A_{11}(z) & A_{12}(z)\\ A_{21}(z) & A_{22}(z)\end{bmatrix}\begin{bmatrix} s_1(t)\\ s_2(t)\end{bmatrix}\\ &= \begin{bmatrix} A_{11}(z)s_1(t)+A_{12}(z)s_2(t)\\ A_{21}(z)s_1(t)+A_{22}(z)s_2(t)\end{bmatrix}\end{aligned}\tag{1}$$

但し, $\boldsymbol{a}(n)=[a_{kl}(n)]_{kl}$ はフィルタ長 N の混合フィルタ行列, $\boldsymbol{A}(z)=[A_{kl}(z)]_{kl}=[\sum_{n=0}^{N-1}a_{kl}(n)z^{-n}]_{kl}$ は $\boldsymbol{a}(n)$ の Z 変換である．ここで, z^{-1} は単位遅延演算子であり, $z^{-n}\cdot s(t)=s(t-n)$ と表記する．また，$[X]_{ij}$ は i 行 j 列に要素 X をもつ行列を表す.

2.2 モノラル出力型 ICA を用いた BSS 法

本稿では, 安定なフィルタ設計が可能な FIR フィルタを各要素とする分離フィルタ行列を用い，サブバンド信号処理を用いることなくフルバンドの観測信号のみを用いて分離フィルタ行列を最適化する時間領域 ICA を取り扱う．ICA の出力信号は次式で表される．

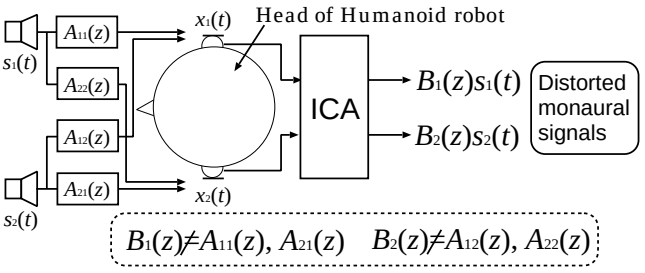


Figure 1: 従来 ICA における入出力の関係. $B_l(z)(B_l(z)\neq A_{kl}(z))$ は任意の FIR フィルタであるため, 出力信号はスペクトル歪みを有する.

$$\boldsymbol{y}(t)=[y_1(t),y_2(t)]^{\mathrm{T}}=\sum_{n=0}^{D-1}\boldsymbol{w}(n)\boldsymbol{x}(t-n),\tag{2}$$

但し, $\boldsymbol{w}(n)$ は分離フィルタ行列であり, D は分離フィルタのフィルタ長を示している．分離フィルタ行列は $\boldsymbol{y}(t)$ の結合確率密度関数と $y_l(t)$ の周辺確率密度関数間の Kullback-Leibler divergence (KLD) を最小化することによって最適化される．音声などの有色信号の分離問題を扱う反復学習則は Choi et al. によって次式のように与えれている [16].

$$\begin{aligned}\boldsymbol{w}^{[j+1]}(n) &= \boldsymbol{w}^{[j]}(n)-\alpha\sum_{d=0}^{D-1}\Big\{\text{off-diag}\left\langle\boldsymbol{\varphi}(\boldsymbol{y}^{[j]}(t))\right.\\ &\qquad\left.\boldsymbol{y}^{[j]}(t-n+d)^{\mathrm{T}}\right\rangle_t\Big\}\boldsymbol{w}^{[j]}(d),\end{aligned}\tag{3}$$

但し, α は更新係数, $[j]$ は更新回数，$\langle\cdot\rangle_t$ は時間平均である．また，off-diag $\boldsymbol{X}$ は行列 $\boldsymbol{X}$ の全ての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 $\boldsymbol{\varphi}(\boldsymbol{y}(t))=[\varphi(y_1(t)),\varphi(y_2(t))]^{\mathrm{T}}$ は $[\tanh(y_1(t)),\tanh(y_2(t))]^{\mathrm{T}}$ である.

従来 ICA の問題点として, 次の 2 点を挙げることができる.

- 独立な各音源信号をモノラル信号として抽出する
- 出力信号の音質は規定されず, スペクトル歪みを含んでいる.

本研究では，単一音源からの音を複数のマイクロホン群で受音した SIMO モデルに基づく信号の推定を目的とする．従って, K(マイクロホン数)$\times$ L(音源数) の出力信号数が必要である．しかしながら, 従来法では各音源に対してモノラル信号を出力するため, 出力信号数は L である．また，従来 ICA 手法の出力信号は任意のスペクトル歪みを有している．これは，音源信号 $s_i(t)$ に独立性が成立するとき, 音源信号に任意の線形フィルタ $b_i(n)$ を畳み込んだ信号 $B_i(z)s_i(t)$ もまた独立性が成立するため，出力信号

間の独立性のみを評価したコスト関数を扱う従来 ICA では，これらの識別が困難であることに起因する．従って，仮に従来 ICA を並列に動作させて出力信号数を $K \times L$ としても，SIMO モデルに基づく信号を得ることはできない．

従来ICA の出力信号の音質の任意性の問題点を改善するために, Matsuoka et al. は Minimal Distortion Principle [17] に基づいた拡張 ICA を提案している．この手法は，従来 ICA で用いる出力信号間の KLD と出力信号ベクトルと観測信号ベクトルの差のフロベニウスノルムを同時に最小化することにより，任意の FIR フィルタを $B_l(z)$ を規定することを目指した手法である．しかしながら，このコスト関数は規範の異なるコスト関数の和を最小化するため，それらを調整するバランシングパラメータを与える必要がある．また，そのパラメータの最適値は音源の性質に依存するため, 事前に設定することは困難である．更に, この手法の出力信号はモノラル信号であるため，空間の位相情報等は失われている．

2.3 従来の SIMO モデル出力型 ICA を用いた BSS 法

上述の問題点を改善するために, 各音源に対してモノラル信号を推定後, それらを観測信号の空間に射影する手法が提案されている．射影の設計の一例として, 分離フィルタ行列 $\boldsymbol{w}(n)$ の逆行列を利用する手法がある ([18])．この処理には分離フィルタ行列の正則性が必要であるが，それは必ずしも保証されていない．従って，特異点においてはモノラル信号の分離に成功していても射影に失敗し SIMO モデル信号を得ることができない [19, 20]．第二の設計手法として, 特定のモノラル音源信号 $y_l(t)$ を抽出した後, それを k 番目のマイクロホンで観測された信号の空間に射影するデフレーション型の手法が挙げられる [21, 22]．しかしながら, 分離フィルタ行列が Null space であるとき, 得られるモノラル出力信号は零信号であり，またその出力信号を射影することはできない．第三の手法として, IIR フィルタを用いた ICA[23] も提案されているが，この手法には分離フィルタの不安定性の問題が存在する．また，Cardoso は観測信号を観測点における SIMO モデルに基づく信号を抽出することを目的とした Multidimensional ICA (MICA) [24] を提案している．しかしながら，MICA アルゴリズムは瞬時混合問題においてのみ適用が可能であり，畳み込み混合問題への拡張が課題となっている．

従って, 安定な FIR フィルタを用いて観測信号中のすべての SIMO 要素を更新アルゴリズム中で同時に推定する新しい SIMO 出力型 ICA 法の開発が必要である．

3 提案法: SIMO-ICA

上述の 2 つの問題点を同時に解決する手法として，我々は観測信号を SIMO モデルに基づく信号に分離する SIMO-ICA を提案している [9, 10, 11, 12]．SIMO-ICA は, 2 音源 2 素子の場合，単一の ICA と情報幾何理論に基づいて構成される単一の fidelity controller (FC) によって構成される．従って，システム全体が情報幾何理論に基づく学習アルゴリズムで構成されており，それらのバランスを決定するパラメータは 1 でよい [11, 12]．SIMO-ICA における ICA の出力信号は次式のように表される．

$$\begin{aligned} \boldsymbol{y}_{(\mathrm{ICA})}(t) &= [y_1^{(\mathrm{ICA})}(t), y_2^{(\mathrm{ICA})}(t)] \\ &= \sum_{n=0}^{D-1} \boldsymbol{w}_{(\mathrm{ICA})}(n)\boldsymbol{x}(t-n), \end{aligned} \tag{4}$$

ここで, $\boldsymbol{w}_{(\mathrm{ICA})}(n)$ は ICA の分離フィルタ行列である．FC の出力信号は, 次式により計算され，各要素は互いに独立になるように最適化される．

$$\begin{aligned} \boldsymbol{y}_{(\mathrm{FC})}(t) &= [y_1^{(\mathrm{FC})}(t), y_2^{(\mathrm{FC})}(t)] \\ &= \boldsymbol{x}(t-\frac{D}{2}) - \boldsymbol{y}_{(\mathrm{ICA})}(t). \end{aligned} \tag{5}$$

今後, $\boldsymbol{y}_{\mathrm{FC}}(t)$ をヴァーチャル ICA の出力信号として扱い, ヴァーチャル ICA の分離フィルタ行列を次のように定義する．

$$\boldsymbol{w}_{\mathrm{FC}}(n) = \boldsymbol{I}\delta(n-\frac{D}{2}) - \boldsymbol{w}_{\mathrm{ICA}}(n). \tag{6}$$

ここで，$\delta(n)$ は $\delta(0)=1$, $\delta(n)=0$ $(n \neq 0)$ を満たすデルタ関数である．また, 式 (6) を用いて，式 (5) を以下のように書き表すことできる．

$$\boldsymbol{y}_{\mathrm{FC}}(t) = \sum_{n=0}^{D-1} \boldsymbol{w}_{\mathrm{FC}}(n)\boldsymbol{x}(t-n). \tag{7}$$

ICA とは違い，FC の分離フィルタ行列 $\boldsymbol{w}_{(\mathrm{FC})}(n)$ は $\boldsymbol{w}_{(\mathrm{ICA})}(n)$ の従属関数であり, それ単体で分離フィルタを持たないため, 我々はこの分離フィルタに対して，“ヴァーチャル” という単語を用いている．式 (5) を次のように展開することで, FC の意義は明確になる．

$$\boldsymbol{y}_{(\mathrm{ICA})}(t) + \boldsymbol{y}_{(\mathrm{FC})}(t) - \boldsymbol{x}(t-D/2) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{8}$$

つまり，出力ベクトルの和 $\boldsymbol{y}_{(\mathrm{ICA})}(t) + \boldsymbol{y}_{(\mathrm{FC})}(t)$ が全ての SIMO 要素の和 $[\sum_{l=1}^{L} A_{kl}(z)s_l(t-D/2)]_{k1} (= \boldsymbol{x}(t-D/2))$ と等価であることを意味している．ここで, $D/2$ の遅れは非最小位相システムを扱うために使用している．

もし, 独立な音源信号が式 (4) によって分離され, 同時に式 (5) が互いに独立であるとき, 分離フィルタ行列は唯一の解に収束する．この証明については, [10] を参照して

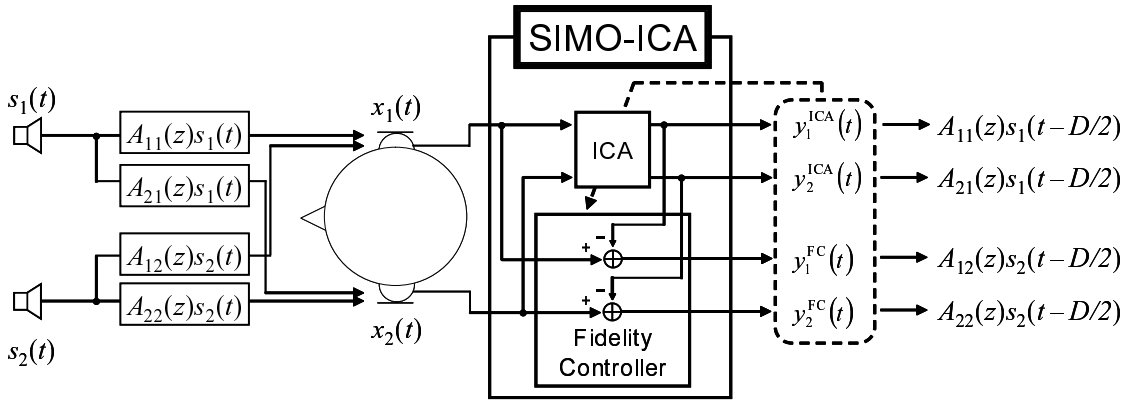


Figure 2: SIMO-ICA における入出力の関係．独立な音源信号が式 (4) によって分離され, 同時に式 (5) が互いに独立であるとき, 分離フィルタ行列は唯一解に収束する.

いただきたい. また, この唯一解を得たとき，出力信号は以下となる.

$$\begin{bmatrix} y_1^{\mathrm{ICA}}(t) \\ y_2^{\mathrm{ICA}}(t) \end{bmatrix} = \begin{bmatrix} A_{11}(z)s_1(t-D/2) \\ A_{22}(z)s_2(t-D/2) \end{bmatrix}, \quad (9)$$

$$\begin{bmatrix} y_1^{\mathrm{FC}}(t) \\ y_2^{\mathrm{FC}}(t) \end{bmatrix} = \begin{bmatrix} A_{12}(z)s_2(t-D/2) \\ A_{21}(z)s_1(t-D/2) \end{bmatrix}, \quad (10)$$

or

$$\begin{bmatrix} y_1^{\mathrm{ICA}}(t) \\ y_2^{\mathrm{ICA}}(t) \end{bmatrix} = \begin{bmatrix} A_{12}(z)s_2(t-D/2) \\ A_{21}(z)s_1(t-D/2) \end{bmatrix}, \quad (11)$$

$$\begin{bmatrix} y_1^{\mathrm{FC}}(t) \\ y_2^{\mathrm{FC}}(t) \end{bmatrix} = \begin{bmatrix} A_{11}(z)s_1(t-D/2) \\ A_{22}(z)s_2(t-D/2) \end{bmatrix}. \quad (12)$$

上述の解を得るために，ICA の分離フィルタの反復学習式 (3) に，式 (5) の KLD の $\boldsymbol{w}_{\mathrm{ICA}}(n)$ に関する natural gradient[4] に nonholonomic 拘束 [16] を適用した学習項を加えればよい．従って，更新学習式は以下で与えられる.

$$\begin{aligned} &\boldsymbol{w}_{(\mathrm{ICA})}^{[j+1]}(n) \\ &= \boldsymbol{w}_{(\mathrm{ICA})}^{[j]}(n) - \alpha \sum_{d=0}^{D-1} \Big[\Big\{ \text{off-diag} \left\langle \boldsymbol{\varphi}(\boldsymbol{y}_{(\mathrm{ICA})}^{[j]}(t)) \right. \\ &\quad \left. \boldsymbol{y}_{(\mathrm{ICA})}^{[j]}(t-n+d)^{\mathrm{T}} \right\rangle_t \Big\} \boldsymbol{w}_{(\mathrm{ICA})}^{[j]}(d) \\ &\quad - \Big\{ \text{off-diag} \left\langle \boldsymbol{\varphi}\Big(\boldsymbol{x}(t-\frac{D}{2}) - \boldsymbol{y}_{(\mathrm{ICA})}^{[j]}(t)\Big) \right. \\ &\quad \left. \Big(\boldsymbol{x}(t-n+d-\frac{D}{2}) - \boldsymbol{y}_{(\mathrm{ICA})}^{[j]}(t-n+d)\Big)^{\mathrm{T}} \right\rangle_t \Big\} \\ &\quad \Big(\boldsymbol{I}\delta(d-\frac{D}{2}) - \boldsymbol{w}_{(\mathrm{ICA})}^{[j]}(d)\Big) \Big], \end{aligned} \quad (13)$$

4 実験と結果

4.1 実験条件

本実験では，ヒューマノイドロボットの頭部を模擬した受音系として，Brüel & Kjær 社の Head And Torso Simulator (HATS) の両耳外耳道入口にコンデンサマイクロホンを取り付けたもの使用した（Figure 3 参照）．また，2 つの独立な音源信号は，2 方位 $\theta_1 = \{-90°, -75°, -60°, -45°, -30°, -15°, 0°\}$, $\theta_2 = \{0°, 15°, 30°, 45°, 60°, 75°, 90°\}$ から放射されるものとした．音源と HATS 間の距離は 1.5 m であり，混合時の SNR は 0 dB である．残響時間は約 200 ms である．音源信号として，ASJ 研究用連続音声コーパスより選択した男女 2 名ずつの話者 4 名による 12 セットの音声を使用した．この音声のサンプリング周波数は 8 kHz であり，学習及び分離データの長さは 3 秒である．また，分離フィルタ行列のタップ数は 512 であり, 更新回数は 5000 回である．比較対象の従来法として，式 (3) で表される手法を用いた．ICA の初期フィルタには，$\pm 60°$ を音源方位とする頭部伝達特性 [5] の逆行列を用いた．評価値として，出力信号と真の SIMO モデルに基づく信号との距離で表現される SIMO-model Accuracy (SA) を用いる．SA の定義は以下である.

$$\begin{aligned} \mathrm{SA} = \frac{1}{4}\Big\{ &10\log_{10}\frac{\| A_{11}(z)s_1(t-\frac{D}{2}) \|^2}{\| y_1^{(\mathrm{ICA})}(t) - A_{11}(z)s_1(t-\frac{D}{2}) \|^2}, \\ &+10\log_{10}\frac{\| A_{21}(z)s_1(t-\frac{D}{2}) \|^2}{\| y_2^{(\mathrm{FC})}(t) - A_{21}(z)s_1(t-\frac{D}{2}) \|^2}, \\ &+10\log_{10}\frac{\| A_{12}(z)s_2(t-\frac{D}{2}) \|^2}{\| y_1^{(\mathrm{FC})}(t) - A_{12}(z)s_2(t-\frac{D}{2}) \|^2}, \\ &+10\log_{10}\frac{\| A_{22}(z)s_2(t-\frac{D}{2}) \|^2}{\| y_2^{(\mathrm{ICA})}(t) - A_{22}(z)s_2(t-\frac{D}{2}) \|^2} \Big\}. \end{aligned} \quad (14)$$

4.2 実験結果と考察

Figure 4, 5 に異なる音源方位毎の従来法と提案法の分解実験の結果を示す. この結果より, ほとんどすべての音源方位組み合わせにおいて, SIMO-ICA の SA は従来法の SA を上回っていることが確認できる. 従って, SIMO-ICA の分解性能は従来 ICA 法の分解性能より優れていること

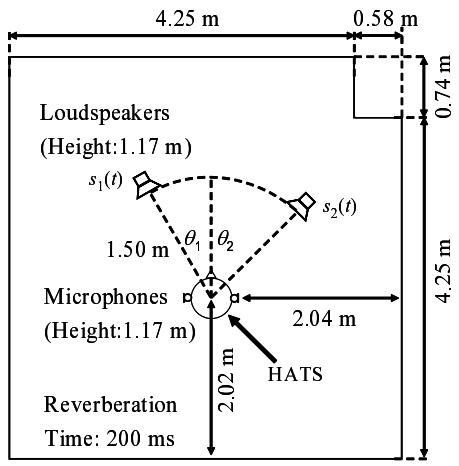


Figure 3: 実験に用いた残響室の見取図.

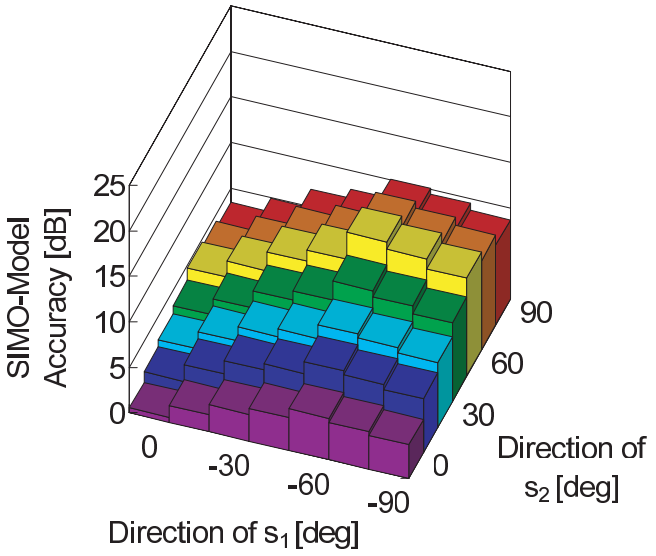


Figure 4: HATS を用いて収録した混合バイノーラル音の分解実験における従来 ICA 法の SA の結果.

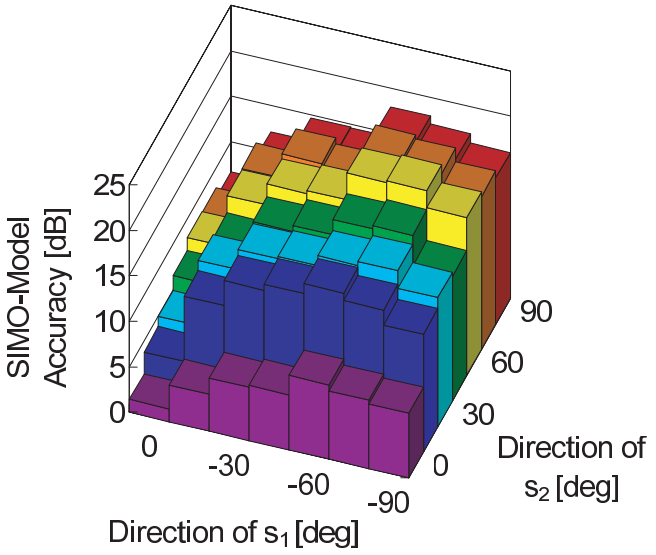


Figure 5: HATS を用いて収録した混合バイノーラル音の分解実験における提案法 SIMO-ICA の SA の結果.

が示された.

一般的に, 人間が音の到来方向を知覚するためには, 両耳間の音の違い, 特に両耳間時間差 (ITD) 及び両耳間レベル差 (ILD) が必要であると言われている [5]. SIMO-ICA の分離信号に含まれる音源毎の空間情報を検証するため, 出力信号の ITD 及び ILD を計算した. 検証の際には, 音源信号の周期構造の影響を除くため, 音源信号をインパルス応答信号 $\delta(t)$ として, ITD 及び ILD を評価した. このとき, 推定すべき SIMO モデル信号はインパルス応答 $A_{ij}(z)\delta(t)$ で与えられ, SIMO-ICA で推定されるインパルス応答は以下で与えられる.

$$[h_1^{(\mathrm{ICA})}(t), h_2^{(\mathrm{ICA})}(t)] = \sum_{n=1}^{D-1} \boldsymbol{w}_{(\mathrm{ICA})}(n)\boldsymbol{A}(z)\begin{bmatrix}\delta(t+D/2-n)\\ \delta(t+D/2-n)\end{bmatrix}, \quad (15)$$

$$[h_1^{(\mathrm{FC})}(t), h_2^{(\mathrm{FC})}(t)] = \sum_{n=1}^{D-1} \boldsymbol{w}_{(\mathrm{FC})}(n)\boldsymbol{A}(z)\begin{bmatrix}\delta(t+D/2-n)\\ \delta(t+D/2-n)\end{bmatrix}, \quad (16)$$

但し, $\delta(t+D/2-n)$ は分離フィルタ行列の時間遅れ $D/2$ の影響を取り除くために用いた.

Figure 6 (a) に, 音源 $s_1(t)$ からの真のインパルス応答間の ILD と SIMO-ICA で推定されたインパルス応答間の ILD を示す. また, Figure 6 (b) に, 音源 $s_2(t)$ からの真のインパルス応答間の ILD と SIMO-ICA で推定されたインパルス応答間の ILD を示す. これらの結果より, 音源方位が θ_1=−15°, 0° 且つ θ_2=0°, 15° を除いて, SIMO-ICA は事前情報を用いることなくインパルス応答の正確な ILD を表現していることが確認された.

Figure 7 (a) に, 音源 $s_1(t)$ からの真のインパルス応答間の ITD と SIMO-ICA で推定去れたインパルス応答間の ITD を示す. また, Figure 7 (b) に, 音源 $s_2(t)$ からの真のインパルス応答間の ITD と SIMO-ICA で推定されたインパルス応答間の ITD を示す. 実験結果より, 音源方位が正面付近を除いて SIMO-ICA は時間差を維持していることが示された.

更に, 音の到来方向は次式を用いて計算することが可能である.

$$\mathrm{DOA}_{s_l} = \sin^{-1}\frac{c\tau_{s_l}}{d}, \quad (17)$$

但し, c は音速, τ_{s_l} は音源 s_l に関する ITD, d は HATS の両耳間の距離を示す.

Figure 8 に ITD を用いて到来方向推定を行なった結果を示す. この結果より, SIMO-ICA の出力信号には音源の

方位情報も維持していることが確認できる.

以上より, SIMO-ICA は各音源に関する空間情報を損なうことなく, 混合バイノーラル信号を各音源に関する SIMO モデル信号に分解することが可能であり，これにより従来実現が困難であった音 tele-presence システムの構築が可能である.

4.3 Sound Demonstrations

提案法 SIMO-ICA による分解実験のサウンドデモンストレーションを以下の WEB サイト上で公開している.

`http://www.aist-nara.ac.jp/~tomoya-t/demo_index.html`

本デモンストレーションサウンドはバイノーラル信号であるため, スピーカではなく, ヘッドフォンもしくはイヤフォンでの受聴をお勧めする.

5 Conclusion

本稿では，SIMO-ICA を混合バイノーラル音のブラインド分解問題に適用した音響 tele-presence システムを提案した．SIMO-ICA は観測信号を各音源に対してモノラル信号として推定するのではなく, 各マイクロホンで観測された SIMO モデルに基づく信号に分解する拡張 ICA であるため，本システムの出力信号は単一音源からのバイノーラル音となる．有効性の検証のため, HATS を用いた混合バイノーラル音の分解実験を試みた．実験結果より，SIMO-ICA の分解性能は従来法より優れており, また SIMO-ICA の出力信号は音源の方位情報等の空間的性質を維持していることが確認された.

参考文献

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

[2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol.22, pp.21–34, 1998.

[3] H. Saruwatari, T. Kawamura, and K. Shikano, “Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming,” *Proc. of Eurospeech*, pp.2603–2606, Sept. 2001.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, Ltd, West Sussex, 2002.

[5] J. Blauert, *Spatial Hearing (revised ed.)*, Cambridge, MA: The MIT Press, 1997.

[6] Y. Tatekura, H. Saruwatari, and K. Shikano, “Sound reproduction system including adaptive compensation of temperature fluctuation effect for broad-band sound control,” *IEICE Trans. Fundamentals*, vol.E85–A, no.8, pp.1851–1860, Aug. 2002.

[7] I. Toshima, H. Uematsu, T. Hirahara, “A steerable dummy head that tracks three-dimensional head movement: TeleHead,” *Acoustical Science and Technology*, vol.24, no.5, pp.327–329, 2003.

[8] S. Tachi, K. Komoriya, K. Sawada, T. Nishiyama, T. Itoko, M. Kobayashi and K. Inoue, “Telexistence cockpit for humanoid robot control,” *Advanced Robotics*, vol.17, no.3, pp.199–217, 2003.

[9] T. Takatani, T. Nishikawa, H. Saruwatari, K. Shikano, “High-fidelity blind separation of acoustic signals using SIMO-model-based Independent component analysis,” *IEICE Trans. Fundamentas*, vol.E87–A, no.8, pp.2063–2072, 2004.

[10] T. Takatani, T. Nishikawa, H. Saruwatari, K. Shikano, “High-fidelity blind source separation of acoustic signals using SIMO-model-based ICA with information-geometric learning,” *Proc. of IWAENC*, pp.251–254, 2003.

[11] T. Takatani, T. Nishikawa, H. Saruwatari, K. Shikano, “”Comparison between SIMO-ICA with least squares criterion and SIMO-ICA with information-geometric learning,” *Proc. of International Congress on Acoustics*, pp.I-329–332, 2004.

[12] T. Takatani, S. Ukai, T. Nishikawa, H. Saruwatari, K. Shikano, “Evaluation of simo separation methods for blind decomposition of binaural mixed signals,” *Proc. of IWAENC*, pp.233–236, 2005.

[13] H. Saruwatari, H. Yamajo, T. Takatani, T. Nishikawa, K. Shikano, “Blind separation and deconvolution for convolutive mixture of speech combining SIMO-model-based ICAand multichannel inverse filtering,” *IEICE Trans. Fundamentas*, vol.E88–A, no.9, pp.2387–2400, 2004.

[14] H. Saruwatari, S. Ukai, T. Takatani, T. Nishikawa, K. Shikano, “Two-stage blind source separation combining SIMO-model-based ICA and adaptive beamforming,” *Proc. of EUSIPCO*, TueAmPO2, 2005.

[15] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata T. Morita "Real-time implementation of two-stage blind sources separation combining SIMO-ICA and binary masking," *Proc. of IWAENC*, pp.229–232, 2005.

[16] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," *Proc. of International Workshop on ICA and BSS*, pp.371–376, Jan. 1999.

[17] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. of International Conference on ICA and BSS*, pp.722–727, Dec. 2001.

[18] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proc. of 1998 International Symposium on Nonlinear Theory and its Application (NOLTA)*, vol.3, pp.923–926, Sep. 1998.

[19] T. Nishikawa, H. Saruwatari, and K. Shikano, "Stable learning algorithm for blind separation of temporally correlated acoustic signals combining multistage ICA and linear prediction," *IEICE Trans. Fundamentals*, vol.E86-A, no.8, pp.2028–2036, 2003.

[20] H. Saruwatari, H. Yamajo, T. Takatani, T. Nishikawa, and K. Shikano, "Blind separation and deconvolution of MIMO-FIR system with colored sound inputs using SIMO-model-based ICA", *Proc. IEEE Workshop on SSP*, pp.421–424, Sept. 2003.

[21] C. Simon, P. Loubaton, C. Vignat, C. Jutten, G. d'Urso, "Separation of a class of convolutive mixturesa contrat: a contrast function approach," *Proc. of ICASSP*, 1429–1432, March 1999.

[22] J.K. Tugnait, "Identification and deconvolution of multichannel linear non-gaussian processes using higher order statistics and inverse filter criteria," *IEEE trans. on signal processing*, Vol. 45. pp 658–672, 1997.

[23] N. Charkani, and Y. Deville, "A convolutive separation method with self-optimizing non-linearities," *Proc. of ICASSP*, pp.2909–2912, Mar. 1999.

[24] J.-F. Cardoso, "Multidimensional independent component analysis," *Proc. of ICASSP*, vol. 4, pp.1941–1944, May 1998.

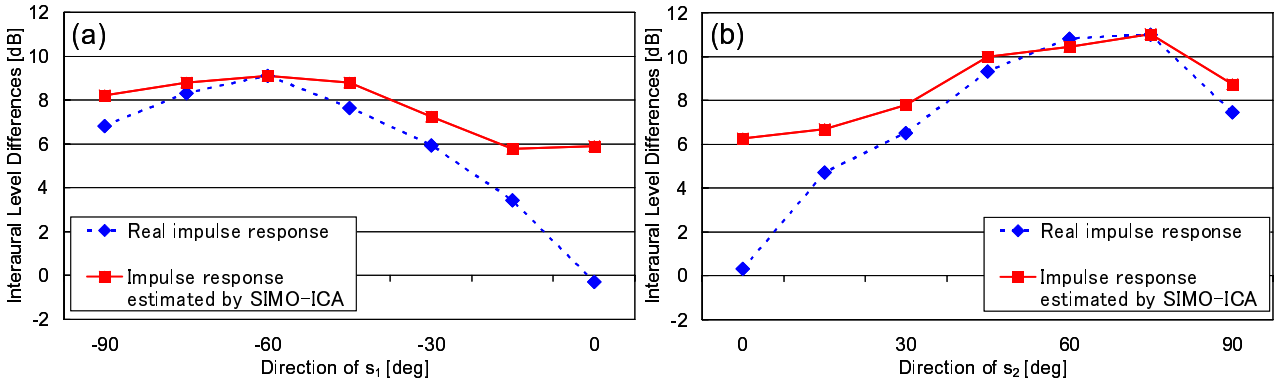


Figure 6: (a) 出力信号 $h_1^{(\mathrm{ICA})}(t)$ と $h_2^{(\mathrm{FC})}(t)$ の ILD の結果, (b) 出力信号 $h_2^{(\mathrm{ICA})}(t)$ と $h_1^{(\mathrm{FC})}(t)$ の ILD の結果.

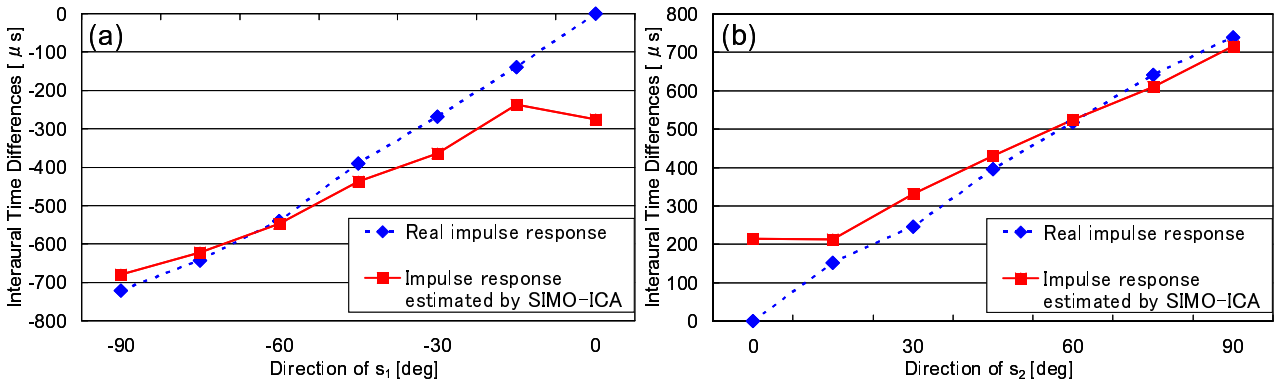


Figure 7: (a) 出力信号 $h_1^{(\mathrm{ICA})}(t)$ と $h_2^{(\mathrm{FC})}(t)$ の ITD の結果, (b) 出力信号 $h_2^{(\mathrm{ICA})}(t)$ と $h_1^{(\mathrm{FC})}(t)$ の ITD の結果.

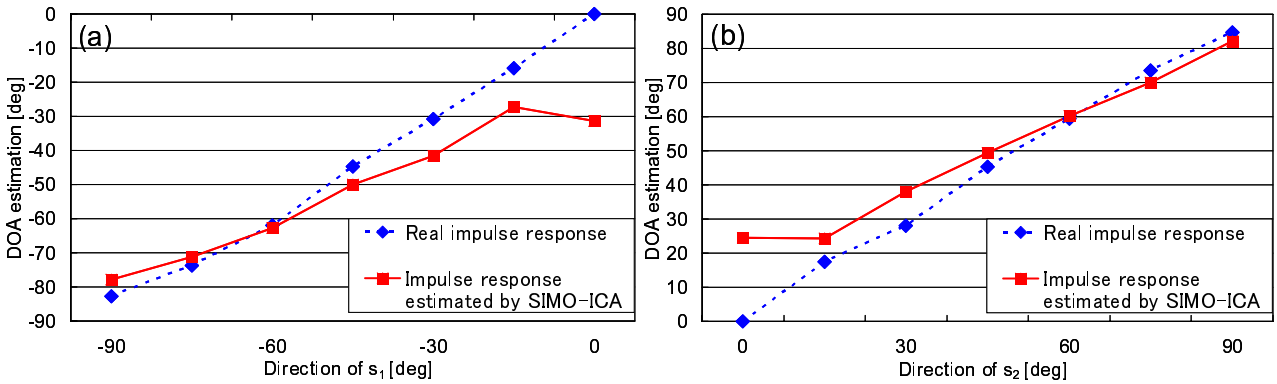


Figure 8: (a) Figure 7 (a) の ITD を用いて推定した音源 s_1 の到来方位推定結果, (b) Figure 7 (b) の ITD を用いて推定した音源 s_2 の到来方位推定結果.

多音源に対する周波数領域ブラインド音源分離
Blind source separation of many sounds in the frequency domain

澤田 宏，向井 良，荒木 章子，牧野 昭二
Hiroshi Sawada, Ryo Mukai, Shoko Araki and Shoji Makino
日本電信電話（株） NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation
{sawada, ryo, shoko, maki}@cslab.kecl.ntt.co.jp

Abstract

This paper describes the frequency-domain approach to the blind source separation (BSS) of convolutively mixed acoustic signals. The advantage of the frequency-domain approach is that convolutive mixtures in the time domain can be approximated as multiple simple mixtures in the frequency domain. However the permutation ambiguity should be solved to group the frequency components of the same source together. This paper presents effective methods to align the permutation ambiguity. Based on the methods, we succeeded in separating many sources in real-world situations.

1 はじめに

複数の音が混ざり合った複数マイクでの観測信号から目的の音を取り出す音源分離技術には，雑音下での音声認識など，様々な応用が期待できる．もし，目的音源の方向などを事前に知っていれば，ビームフォーミング [1] により分離はある程度達成できる．しかし，それらの事前情報が得られない，あるいは得られたとしても正確でない場合には，事前情報を必要としない，いわゆるブラインド音源分離 (BSS: Blind Source Separation [2]) の技術が重要となる．

独立成分分析 (ICA: Independent Component Analysis [3]) は，BSS にとって主要な統計的処理の一つであり，源信号の非ガウス性と独立性に着目して分離を達成する．音声など有益な音信号は，多くの場合非ガウス性を持ち，ICA が効率良く適用できる．ただし，実環境で音が混ざる場合は，単なる混合ではなく，時間遅れと残響を伴った畳み込み混合となるため，いかにして ICA を適用するかが問題となる．

第一の方法は，時間領域において，畳み込み混合を直接 ICA で解くものである [4–6]．これは，正しい収束点では精度良く分離を達成できるが，複雑な畳み込み混合を扱うため，収束までの計算時間が大きいという困難さがある．第二の方法は，観測信号を短時間フーリエ変換し，周波数領域で ICA を適用するものである [7–10]．周波数領域では，畳み込み混合が周波数ビン毎の単純混合に近似できるため，ICA 自体の収束は速い．しかし，ばらばらに分離された各音源の周波数成分を音源毎にグループ化するという，いわゆる permutation の問題を解かなければならない．

我々は，第二の周波数領域での方法に関して精力的に研究を行っており [11–16] ，特に，permutation の問題に対する効率的な手法を開発した．その結果，実環境において，2 音源 2 マイクの基本形を初めとし，3 次元的に配置された 6 音源の分離や，無数の背景雑音の中での主要 3 音源の分離を達成した．次章以後，その技術を説明し，実験結果を示す．

2 周波数領域ブラインド音源分離の流れ

N 個の音源 $\mathrm{s}_1(t), \ldots, \mathrm{s}_N(t)$ が空間で畳み込み混合され，M 個のマイク $\mathrm{x}_1(t), \ldots, \mathrm{x}_M(t)$ で観測されるとする．

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{l} \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l) \qquad (1)$$

ここで，h_{jk} は音源 k からマイク j までのインパルス応答である．従来の BSS 研究では，音源数 N を知っており，なおかつそれがマイク数 M 以下であるという状況に限る場合が多い．本稿で提案する手法は，そのような状況では当然有利ではあるが，それ以外の状況でも動作するように設計されている．ただし，例えば音源数 N がマイク数 M より大きい場合，すべての音源を分離する [16] ことはせず，主要な音源に対応するものから順に M 個の分離信号を出力する．

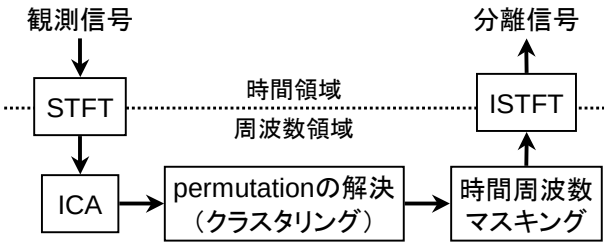


Figure 1: 本稿で説明する周波数領域 BSS の流れ

本章では，以下，周波数領域ブラインド音源分離の流れを説明する．まず，マイクでの観測信号に対してフレーム長 L の短時間フーリエ変換（STFT: Short-Time Fourier Transform）を適用する．

$$x_j(f,\tau) = \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(\tau+r)\,\mathrm{win}(r)\,e^{-\jmath 2\pi f r} \qquad (2)$$

ここで，$f \in \{0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s\}$ は周波数，$\mathrm{win}(r)$ はハニング窓 $\frac{1}{2}(1+\cos\frac{2\pi r}{L})$ など両端が 0 に収束する長さ L の窓，τ は時間を表現する新たな変数である．STFT により，式 (1) の畳み込み混合は，各周波数 f での単純混合

$$x_j(f,\tau) \approx \sum_{k=1}^{N} h_{jk}(f)s_k(f,\tau) \qquad (3)$$

に近似される．ここで，$h_{jk}(f)$ は音源 k からマイク j までの周波数応答，$s_k(f,\tau)$ は式 (2) を用いて同様に得られる音源の時間周波数表現である．ベクトル表記 $\mathbf{x} = [x_1,\ldots,x_M]^T$，$\mathbf{h}_k = [h_{1k},\ldots,h_{Mk}]^T$ により要素をまとめると，式 (3) のベクトル表現

$$\mathbf{x}(f,\tau) \approx \sum_{k=1}^{N} \mathbf{h}_k(f)s_k(f,\tau) \qquad (4)$$

が得られる．周波数領域で解釈すると，ブラインド音源分離の目的は，観測信号ベクトル $\mathbf{x}(f,\tau)$ から，音源 k 毎にすべての周波数成分 $\mathbf{h}_k(f)s_k(f,\tau)$ を求めることにある．

そのためにまず，周波数毎に独立成分分析（ICA）

$$\mathbf{y}(f,\tau) = \mathbf{W}(f)\,\mathbf{x}(f,\tau), \qquad (5)$$

を適用する．ここで，$\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_N]^H$，$\mathbf{w}_i = [w_{1i},\ldots,w_{Mi}]^T$ は $N\times M$ の分離行列，$\mathbf{y} = [y_1,\ldots,y_N]^T$ は分離信号の要素をまとめたベクトルである．もし音源数 N を知っていてかつ $N < M$ であれば，主成分分析により次元数削減を行うことは有効であるが，それ以外の場合は $N = M$，すなわち正方行列として $\mathbf{W}$ を求める．なお，複素数に対する ICA のアルゴリズム，および，そこで用いられる非線形関数に関する議論は，[11] を参照されたい．

ICA により分離信号 y_i の非ガウス性が高められるため，音源が非ガウス性であり，互いに独立であれば，周波数毎に分離が達成される．しかし，ICA の解には permutation の任意性（分離信号の順序を入れ換えても ICA の解となる）と scaling の任意性（分離信号を定数倍しても独立性は保たれる）が存在するため，これらを解決する必要がある．permutation の解決については次章で詳しく説明する．scaling は，次章の (7) 式により基底ベクトルを求めた後，ある着目したマイク J に対応する要素 $a_{Ji}(f)$ を用いて，

$$y_i(f,\tau) \leftarrow a_{Ji}(f)y_i(f,\tau), \qquad (6)$$

として解決できる．これは，マイク J での観測信号に scaling を合わせることであり，minimal distortion principle [4] や projection back [8] と呼ばれるものと等価である．

次に，図 1 の流れに示す時間周波数マスキングを適用する．式 (5) に示す ICA による分離は，線形フィルタによる分離のため，マイク数 M が音源数 N 以上でない場合には，干渉音の残留成分が分離信号にどうしても残る．時間周波数マスキングは，この残留成分を減らす効果がある．その方法の詳細は [15] を参照されたい．本稿では，4.2 章でその効果のみを示す．

以上の処理を経た後，ISTFT: Inverse STFT

$$\mathrm{y}_i(\tau+r) = \frac{1}{L\cdot\mathrm{win}(r)} \sum_{f\in\{0,\frac{1}{L}f_s,\ldots,\frac{L-1}{L}f_s\}} y_i(f,\tau)\,e^{\jmath 2\pi f r}$$

により，分離信号の周波数成分を集めて時間領域の信号に戻す．

3 permutation の解決

3.1 基底ベクトル

ICA の解の任意性を解決するにあたり，分離行列 $\mathbf{W}$ の逆行列（$N < M$ の場合は疑似逆行列）を計算することが非常に有益である．以下，$N = M$，すなわち $\mathbf{W}$ が正方行列である場合を説明する．本稿では，分離行列 $\mathbf{W}$ の逆行列により得られるベクトル $\mathbf{a}_i$ を基底ベクトル (basis vector)

$$[\mathbf{a}_1,\cdots,\mathbf{a}_M] = \mathbf{W}^{-1},\ \mathbf{a}_i = [a_{1i},\ldots,a_{Mi}]^T. \qquad (7)$$

と呼ぶ．これは，この逆行列を式 (5) の両辺に掛け合わせることで，観測信号ベクトルがその線形和

$$\mathbf{x}(f,\tau) = \sum_{i=1}^{M} \mathbf{a}_i(f)y_i(f,\tau) \qquad (8)$$

で表現されることによる．この式は周波数領域 BSS において非常に重要である．ICA の解が良好に得られていれば，式 (8) のある i に関する項 $\mathbf{a}_i y_i$ が式 (4) のある k に関する項 $\mathbf{h}_k s_k$ に対応するからである．その対応関係を求めることが permutation を解くことに相当する．その際に使えるのは，基底ベクトル $\mathbf{a}_i$ と分離信号 y_i の情報である．$\mathbf{W}$ も使えるが，$\mathbf{a}_i$ と逆行列の関係にあるため，基本的には同じ情報になる．$\mathbf{a}_i$ と y_i の双方を使うことでより精度

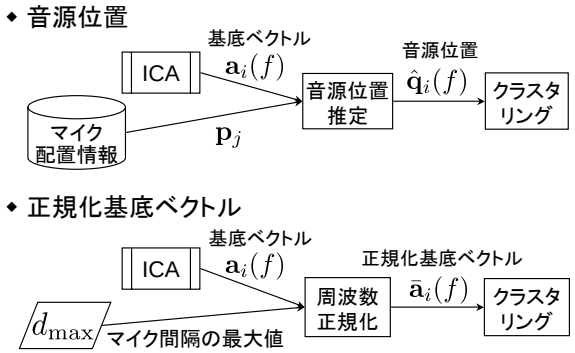


Figure 2: permutation 解決のための 2 種類の方法

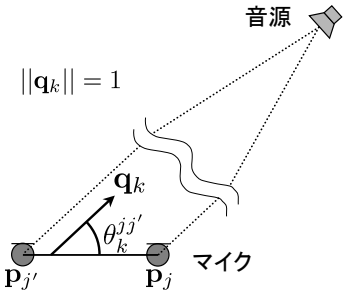


Figure 3: 遠距離場モデル

良く permutation を解決できる [12]．以下では，基底ベクトル $\mathbf{a}_i$ の情報をどのように用いるかを説明する．図 2 に示すように 2 種類の方法があるが，以下の章でそれぞれ説明する．

3.2 推定音源方向のクラスタリング

一つ目は，基底ベクトルから音源位置を推定し，その値をクラスタリングすることで permutation を解決する方法である．本章では，話を簡単化するため，遠距離場モデル

$$h_{jk}(f) \approx \exp\left[\jmath 2\pi f c^{-1} \mathbf{p}_j^T \mathbf{q}_k\right] \tag{9}$$

により音源 k からマイク j への周波数応答 $h_{jk}(f)$ を近似して音源方向 (DOA: Direction-Of-Arrival) を推定する．ここで，c は音の速度，$\mathbf{p}_j$ はマイク j の位置を示す 3 次元ベクトル，$\mathbf{q}_k$ は音源 k の方向を示す長さ 1（$||\mathbf{q}_k|| = 1$）の 3 次元ベクトルである．図 3 に示すように，二つのマイク j と j' を考えると，

$$\frac{h_{jk}(f)}{h_{j'k}(f)} \approx \exp\left[\jmath 2\pi f c^{-1} (\mathbf{p}_j - \mathbf{p}_{j'})^T \mathbf{q}_k\right] \tag{10}$$

$$= \exp\left[\jmath 2\pi f c^{-1} ||\mathbf{p}_j - \mathbf{p}_{j'}|| \cos\theta_k^{jj'}\right] \tag{11}$$

が得られる．このように，音源方向は，座標系に対して決まるもの $\mathbf{q}_k$ と，マイクペア (j, j') に対して相対的に決まる角度 $\theta_k^{jj'}$ の 2 種類の表現がある．なお，より一般的な近距離場モデルを仮定して，方向だけでなく距離も含めた位置を推定する方法に関しては [13, 14] を参照されたい．

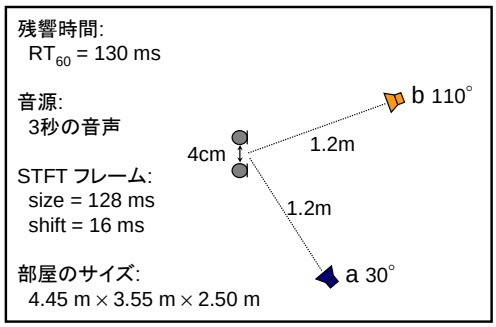


Figure 4: 実験条件（音源方向推定）

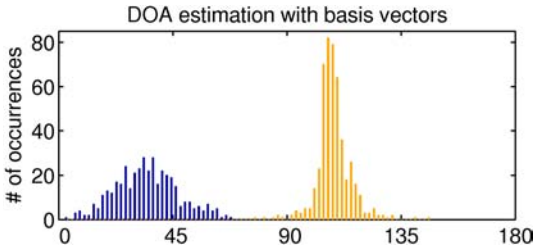


Figure 5: 音源方向推定結果

ICA の解が良好に得られて式 (4) と式 (8) の項に対応関係があると仮定すると，基底ベクトル $\mathbf{a}_i(f)$ の要素 $a_{ji}(f)$ と $a_{j'i}(f)$ の比は以下のように表現できる．

$$\frac{a_{ji}(f)}{a_{j'i}(f)} = \frac{a_{ji} y_i}{a_{j'i} y_i} \approx \frac{h_{jk} s_k}{h_{j'k} s_k} = \frac{h_{jk}(f)}{h_{j'k}(f)} \tag{12}$$

ここで，添字 i と k が異なる可能性があることが，permutation の任意性を示している．式 (12) と式 (11) の偏角に着目すると，角度の推定値

$$\hat{\theta}_i^{jj'}(f) = \arccos \frac{\arg[a_{ji}(f)/a_{j'i}(f)]}{2\pi f c^{-1} ||\mathbf{p}_j - \mathbf{p}_{j'}||} \tag{13}$$

を求めることができる．

図 4 に示す実験条件で周波数領域 BSS を実行し，式 (13) により音源の角度を推定した結果を図 4 に示す．STFT のフレーム長は $L = 1024$ であったため，その約半分の周波数ビンでの推定をヒストグラムで示している．二つのクラスタが存在することがわかり，そのクラスタリング結果に基づいて permutation を解決することができる．

次に，座標系に対して決まる方向 $\mathbf{q}_k$ を考える．式 (12) と式 (10) の偏角に着目すると，

$$2\pi f c^{-1} (\mathbf{p}_j - \mathbf{p}_{j'})^T \mathbf{q}_k \approx \arg[a_{ji}(f)/a_{j'i}(f)] \tag{14}$$

が得られる．u 個のマイクペア $(j_1, j_1'), \ldots, (j_u, j_u')$ に対し同様に考えると，連立方程式

$$2\pi f c^{-1}\, \mathbf{D}\, \mathbf{q}_k = \mathbf{r}_i(f) \tag{15}$$

が得られる．ここで，

$$\mathbf{D} = [\mathbf{p}_{j_1} - \mathbf{p}_{j_1'}, \ldots, \mathbf{p}_{j_u} - \mathbf{p}_{j_u'}]^T,$$
$$\mathbf{r}_i(f) = [\arg(a_{j_1 i}/a_{j_1' i}), \ldots, \arg(a_{j_u i}/a_{j_u' i})]^T.$$

である．実際の環境では推定誤差等の影響により，この連立方程式を厳密に満たす解 $\mathbf{q}_k$ は存在しにくい．そのため，Moore-Penrose 疑似逆行列 $\mathbf{D}^+$ を用いてその推定値

$$\hat{\mathbf{q}}_i(f) = \frac{\mathbf{D}^+ \mathbf{r}_i(f)}{2\pi f c^{-1}}, \quad \hat{\mathbf{q}}_i(f) \leftarrow \frac{\hat{\mathbf{q}}_i(f)}{||\hat{\mathbf{q}}_i(f)||} \tag{16}$$

を近似的に得る．4.1 章では，3 次元配置の音源とマイクに対して周波数領域 BSS を実行し，式 (16) に従って音源方向を推定してクラスタリングした結果を紹介する．

3.3 正規化基底ベクトルのクラスタリング

次に，基底ベクトル $\mathbf{a}_i(f)$ の周波数依存性をできるだけ除去した正規化基底ベクトル $\bar{\mathbf{a}}_i(f)$ をクラスタリングすることで permutation を解決する方法を示す．これは，上記に示した音源方向による方法ほど直観的ではないが，マイクの配置情報が不要であるという利点がある．まず手順を説明する．正規化基底ベクトル $\bar{\mathbf{a}}_i(f) = [\bar{a}_{1i}(f), \ldots, \bar{a}_{Mi}(f)]^T$ の要素 $\bar{a}_{ji}(f)$ は，

$$\bar{a}_{ji}(f) \leftarrow |a_{ji}(f)| \exp\left[\jmath \frac{\arg[a_{ji}(f)/a_{Ji}(f)]}{4 f c^{-1} d_{\max}}\right] \tag{17}$$

により計算される．ここで，J はある基準マイクの添字，$d_{\max}$ はある正の実数値であり，基準マイク J と他のマイクとの距離の最大値とすれば良い．この式により，周波数依存性が除去される．次に，scaling の任意性を除去するために，長さを 1 に正規化する．

$$\bar{\mathbf{a}}_i(f) \leftarrow \bar{\mathbf{a}}_i(f) \,/\, ||\bar{\mathbf{a}}_i(f)|| \tag{18}$$

これらの操作により，正規化基底ベクトル $\bar{\mathbf{a}}_i(f)$ は，式 (9) の遠距離場モデルに従うと，周波数に依存せず音源の方向 $\mathbf{q}_k$ およびマイクの位置 $\mathbf{p}_1, \ldots, \mathbf{p}_M$ のみに依存する．実際，式 (14) の関係を利用すると，

$$\bar{a}_{ji}(f) \approx \frac{1}{\sqrt{M}} \exp\left[\jmath \frac{\pi}{2} \frac{(\mathbf{p}_j - \mathbf{p}_J)^T \mathbf{q}_k}{d_{\max}}\right],$$

であることがわかる．なお，より一般化された近距離場モデルに従った場合でも，周波数 f に依存しないことが証明できる [15].

次に，正規化基底ベクトル $\bar{\mathbf{a}}_i$ のクラスタリングを行い，クラスタ $C_1, \ldots, C_M$ を求める．クラスタ C_k のセントロイド $\mathbf{c}_k$ は，$|C_k|$ をクラスタ C_k のメンバ数として，

$$\mathbf{c}_k \leftarrow \sum_{\bar{\mathbf{a}} \in C_k} \bar{\mathbf{a}}/|C_k|, \quad \mathbf{c}_k \leftarrow \mathbf{c}_k/||\mathbf{c}_k||, \tag{19}$$

と計算する．クラスタリングの基準は，クラスタのメンバ $\bar{\mathbf{a}} \in C_k$ とセントロイド $\mathbf{c}_k$ との自乗距離の総和 $\mathcal{J}$ を最小化することである．

$$\mathcal{J} = \sum_{k=1}^{M} \mathcal{J}_k, \; \mathcal{J}_k = \sum_{\bar{\mathbf{a}} \in C_k} ||\bar{\mathbf{a}} - \mathbf{c}_k||^2. \tag{20}$$

Figure 6: 3 次元に配置された 6 音源と 4cm 立方体の各頂点に配置された 8 個のマイク

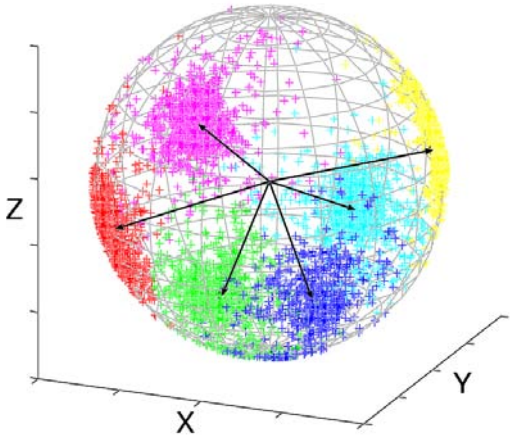


Figure 7: 音源方向推定結果

クラスタリングが終了すれば，各周波数 f で出力の順序を入れ換える順列 Π_f は

$$\Pi_f = \mathrm{argmin}_{\Pi} \sum_{k=1}^{M} ||\bar{\mathbf{a}}_{\Pi(k)}(f) - \mathbf{c}_k||^2. \tag{21}$$

として計算できる．4.2 章では，無数の背景雑音の中での主要 3 音源の分離において，正規化基底ベクトルがどのようにクラスタリングされるかを示す．

4 実験

これまでに述べた手法を用いて，実環境で多音源を分離した実験結果を 2 種類示す．

4.1 3 次元的に配置された 6 音源の方向推定と分離

図 6 に示すような 3 次元的配置の 6 音源と 8 マイクを用いて周波数領域 BSS を実行した．音源は，8 秒の英語音声を 6 個用いた．BSS 処理に要した時間は，2GHz の Pentium M を搭載したノート PC で 25 秒程度であった．

式 (16) によりすべての周波数ビンで音源方向を推定した結果を図 7 に示す．音源方向の推定値であるベクトル

Table 1: 6 音源分離: SIR 改善量 (dB)

	SIR_1	SIR_2	SIR_3	SIR_4	SIR_5	SIR_6	平均
入力 SIR	−11.6	−9.0	−9.0	−6.6	−6.9	−2.5	−7.6
SIR 改善量	19.2	21.2	25.4	21.0	20.5	16.2	20.6

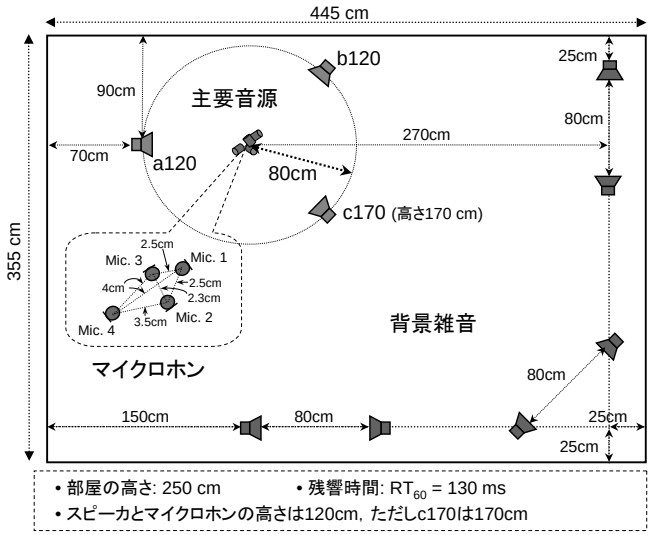


Figure 8: 実験条件（主要 3 音源の分離）

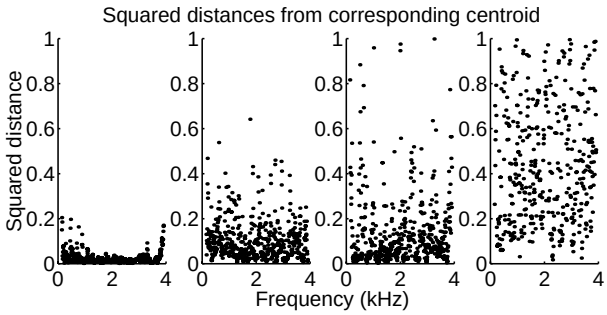


Figure 9: 正規化基底ベクトルのクラスタリング結果（主要 1 音源の場合）

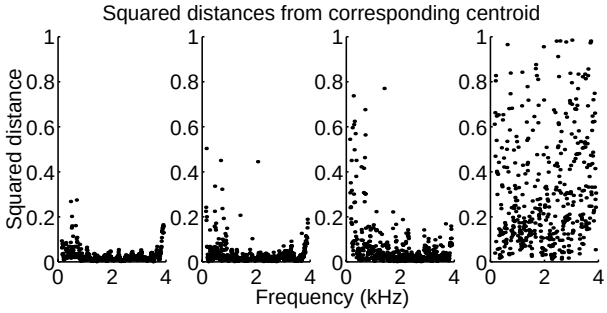


Figure 10: 正規化基底ベクトルのクラスタリング結果（主要 3 音源の場合）

$\hat{\mathbf{q}}_i(f)$ の長さは 1 に正規化されているため，単位球の表面に推定値が乗る．これら推定値に対してクラスタリングを行い，その結果を 6 色で示した．方向の推定値にある程度の分散はあるが，permutation が正しく解ける程度には推定できている．この情報を元に permutation を解き，分離信号を作成した．SIR: Signal-to-Interference Ratio による分離性能の評価を表 1 に示す．SIR の平均改善量は 20dB 程度であり，高精度に分離が達成できたと言える．

4.2 無数の背景雑音の中での主要音源の分離

次に，図 8 に示す状況で実験を行った．ここでは，マイクから遠い背景雑音が 6 個あり，雑踏など無数の背景雑音が存在する状況を模倣している．マイクの近くには 3 個の主要音源があり，これらの音を分離することを目的とした．9 個の音源すべてに音声を用いた．

マイクの数が 4 個のため，各周波数ビン毎に 4 つの周波数成分を ICA で求めた．permutation の問題は，3.3 章に示した正規化基底ベクトルのクラスタリングで行った．クラスタリング結果を図 9 と図 10 に示す．マイクの数と同じ 4 個のクラスタがある．正規化基底ベクトルは M 次元の複素ベクトルであるため，クラスタの様子を可視化することは難しいが，ここでは，それぞれのクラスタについて，セントロイドとメンバとの自乗距離を示している．

4 つのクラスタのうち，どのクラスタがマイクに近い主要音源に対応し，どのクラスタが背景雑音に対応するかを決定するために，我々はクラスタの分散 $\mathcal{J}_k/|C_k|$ に着目し，分散が小さいクラスタを主要音源として選んでいる．図 9 は，マイクに近い主要音源のうち 1 つだけを鳴らし，他の二つを鳴らさなかった場合のクラスタリング結果である．左端のクラスタの分散が小さく，主要音源に対応していることがわかる．図 10 は，3 個すべての主要音源を鳴らした場合のクラスタリング結果である．分散が小さいクラスタが 3 個あり，これらが主要音源に対応することがわかる．このようにクラスタの分散から，主要音源の数を推定することもできる．

次に，時間周波数マスキングの効果を図 11 に示す．これは，マイクに近い主要音源のうち 1 つだけを鳴らした場合の一例である．主要音源 1 個と背景雑音 6 個の合計 7 個の音源があるため，4 個のマイクを用いた ICA による分離では，図 11 の 2) に示すように干渉音の残留成分がどうしても残る．[15] の方法で時間周波数マスキングを適用すると，3) に示すようにその残留成分が抑圧され，4) の目的信号のみのスペクトログラムに近付いた．

最後に，3 個すべての主要音源を鳴らした場合の分離性能を，SIR 改善量で表 2 に示す．音源の組合せを変化させて 10 回試行した平均である．4 個のマイクに対して 9 個の音源（主要音源 3 個と背景雑音 6 個）という非常に厳しい条件ではあるが，10dB 以上の SIR 改善量が得られた．時間周波数マスキングを併用することで，その改善量は更に高まった．分離音のサンプルは，http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/dominant/ で聞くことができる．なお，このような処理をリアルタイム

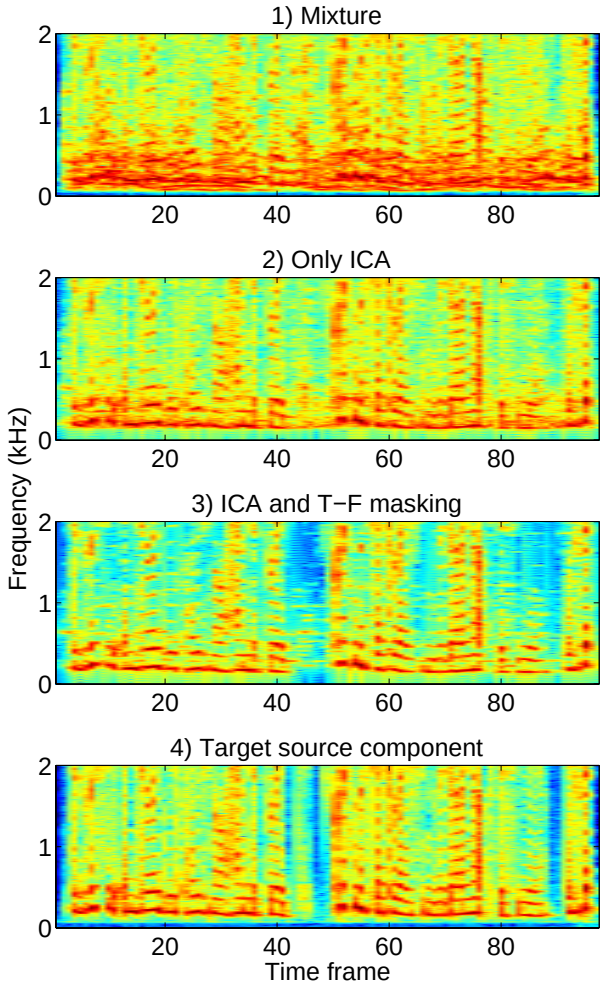


Figure 11: スペクトログラムの例: 1) 混合信号，2) ICA による分離信号，3) ICA と時間周波数マスキング (T-F masking) による分離信号，4) 目的信号のみ（正解）

Table 2: 9 音源中の主要 3 音源分離: SIR 改善量 (dB)

主要音源の位置	a120	b120	c170
入力 SIR	−3.9	−3.6	−5.9
ICA のみ	12.5	13.6	14.5
ICA と時間周波数マスキング	15.1	16.5	17.6

で行うシステムも実現している．

5 おわりに

周波数領域 BSS について説明し，音源数が多い状況での 2 種類の実験でその有効性を示した．ポイントとなる技術は permutation の解決であるが，本稿で説明したように，音源方向や正規化基底ベクトルをクラスタリングすることで効率的に解決できる．なお，マイク数よりも多い音源をすべて分離するものとして，ICA を用いない周波数領域 BSS が知られているが，そのような手法でも，本稿で述べた permutation の解決法と同様の技術が適用できる [16].

参考文献

[1] B. D. Van Veen and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, pp. 4–24, Apr. 1988.

[2] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley & Sons, 2000.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[4] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” in *Proc. ICA 2001*, Dec. 2001, pp. 722–727.

[5] S. C. Douglas and X. Sun, “Convolutive blind separation of speech mixtures using the natural gradient,” *Speech Communication*, vol. 39, pp. 65–78, 2003.

[6] S. C. Douglas, H. Sawada, and S. Makino, “A spatio-temporal FastICA algorithm for separating convolutive mixtures,” in *Proc. of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. V, Mar. 2005, pp. 165–168.

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

[8] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[9] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, “Combined approach of array processing and independent component analysis for blind separation of acoustic signals,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 3, pp. 204–215, May 2003.

[10] 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, Nov. 2003.

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, “Polar coordinate based nonlinear function for frequency domain blind source separation,” *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, Mar. 2003.

[12] ——, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[13] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Frequency domain blind source separation for many speech signals,” in *Proc. of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004 / LNCS 3195)*. Springer-Verlag, Sept. 2004, pp. 461–469.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, “Frequency-domain blind source separation,” in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer, Mar. 2005, pp. 299–327.

[15] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Blind extraction of a dominant source from mixtures of many sources using ICA and time-frequency masking,” in *Proc. of 2005 IEEE International Symposium on Circuits and Systems (ISCAS 2005)*, May 2005, pp. 5882–5885.

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” in *Proc. 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*, Sept. 2005.

SIMO-ICA とバイナリマスク処理を組み合わせた 2 段型リアルタイムブラインド音源分離

Two-Stage Real-Time Blind Source Separation Combining SIMO-ICA and Binary Mask Processing

森 康充†，高谷 智哉†，猿渡 洋†，鹿野 清宏†，稗方 孝之‡，森田 孝司‡

Y. Mori†, T. Takatani†, H. Saruwatari†, K. Shikano†, T. Hiekata‡, T. Morita‡

†奈良先端大・情報 **Nara Institute of Science and Technology**

‡(株) 神戸製鋼所 **Kobe Steel,Ltd.**

E-mail: †{yoshim-m, tomoya-t, sawatari, shikano}@is.naist.jp,

‡{t-hiekata, takashi-morita}@kobelco.jp

Abstract

We newly propose a real-time two-stage blind source separation (BSS) for binaural mixed signals observed at the ears of humanoid robot, in which a Single-Input Multiple-Output (SIMO)-model-based independent component analysis (ICA) and binary mask processing are combined. SIMO-model-based ICA can separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of SIMO-model-based ICA can maintain the spatial qualities of each sound source, and this yields that binary mask processing can be applied to efficiently remove the residual interference components after SIMO-model-based ICA. The experimental results obtained with a human-like head reveal that the separation performance can be considerably improved by using the proposed method in comparison to the conventional ICA-based and binary-mask-based BSS methods.

1 はじめに

ブラインド音源分離 (BSS) とは，観測された混合信号の情報のみを用いて元の音源信号を推定する技術である．この技術は，音源分離処理の時点で学習区間や明示的な音源の到来方位 (DOA) といった事前情報を必要としない教師なしフィルタリング技術に基づいている．BSS はこうした魅力的な特徴をもっているため，信号処理の多くの分野で BSS 技術は大きな関心を集めている．音声信号処理におけるひとつの有望な例として，たとえば，ロボットの両耳で観測した混合信号を分離するといった，人型ロボットの聴覚システム [1] があげられる．これは，インテリジェントロボット技術には欠くことのできない要素である [2, 3].

近年独立成分分析 (ICA) [4] に基づく BSS の研究において，音声信号の分離に関して様々な手法が提案されている [5, 6, 7, 8]．本稿では，実際の音響アプリケーションにおいてしばしば発生する，高残響下での BSS 問題について取り扱う．この場合，従来の ICA では非常に長い分離フィルタが必要とされるが，そのフィルタの学習は容易ではないため，分離性能は十分ではない．従来の改善策の一つは，ICA と，スペクトル減算 [9] といった，他の教師あり信号強調技術を部分的に組み合わせることである．しかしながら，従来の ICA の枠組みでは，各音源に関してモノラル信号を出力するため，アレー信号 (多チャネル信号) 入力を前提とする従来の高精度信号強調手法を適用することが困難であった．

この問題を解決するため，人型ロボットの聴覚処理に適切な，新しい 2 段型 BSS アルゴリズムを提案する．この手法では，BSS 問題を 2 段階で解決する：第一段目においては，Single-Input Multiple-Output (SIMO) モデルに基づく ICA (SIMO-ICA) [10, 11] を適用し，次段においてその SIMO-ICA の出力する SIMO 信号ごとに時間–周波数領域におけるバイナリマスク処理 [12, 13, 14] を行う．ここで，“SIMO” という用語は特定の伝達系を表し，入力が一つの音源で出力が複数マイクロホンで観測される信号である．SIMO-ICA は混合信号を，モノラルではなく，SIMO モデル信号と呼ばれる，独立な音源を各マイクロホンで観測した時点での信号に分離することが可能である．このため，SIMO-ICA の分離信号は，各音源の空間的特性を保つことが可能である．SIMO-ICA の後にバイナリマスク処理を加えることにより，非目的音の消し残り成分を効率的に取り除くことが可能である．実験結果から，提案法は実際の残響環境下でも，音声同士の BSS が十分に動作することが分かった．

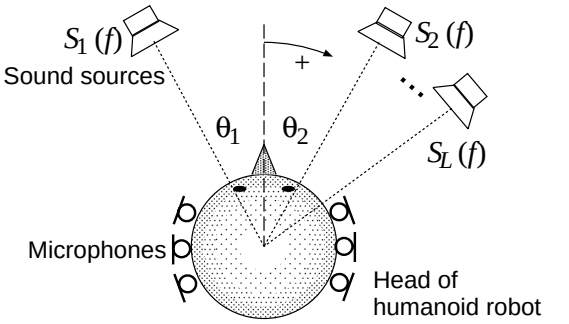


Figure 1: 音源とロボット頭部のマイクロホン配置

2 混合過程と従来の BSS

2.1 混合過程

本稿では，マイクロホン数を K，音源数を L とする．L 個の音源からの到来方位を $\theta_l (l = 1, \cdots, L)$ とし (Figure 1 参照)，$K = L$ の場合を考える．

複数の音源信号が線形混合された場合の観測信号は，周波数領域において以下の式で与えられる．

$$\boldsymbol{X}(f) = \boldsymbol{A}(f)\boldsymbol{S}(f) \tag{1}$$

ここで $\boldsymbol{X}(f) = [X_1(f), \ldots, X_K(f)]^{\mathrm{T}}$ は，観測信号ベクトル，$\boldsymbol{S}(f,t) = [S_1(f), \ldots, S_L(f)]^{\mathrm{T}}$ は，音源信号ベクトルである．また，$\boldsymbol{A}(f) = [A_{kl}(f)]_{kl}$ は混合行列であり，$[\cdot]_{ij}$ は i 行 j 列要素が $\cdot$ である行列を表す．混合行列 $\boldsymbol{A}(f)$ は複素行列であり，これはマイクロホンアレーの配置と部屋の残響を含む遅延のモデルを表現するためである．

2.2 従来の ICA に基づく BSS

従来の周波数領域 ICA (FDICA) では，まず，観測信号の短時間分析を離散フーリエ変換 (DFT) を用いてフレーム毎に行う．これにより，観測信号ベクトルは $\boldsymbol{X}(f,t) = [X_1(f,t), \ldots, X_K(f,t)]^{\mathrm{T}}$ と表現できる．次に，時間–周波数信号の複素分離行列 $\boldsymbol{W}(f) = [W_{lk}(f)]_{lk}$ を用いて，分離信号 $\boldsymbol{Y}(f,t) = [Y_1(f,t), \ldots, Y_K(f,t)]^{\mathrm{T}}$ を次の式により周波数毎に求める．

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

また $\boldsymbol{W}(f)$ の最適化は，以下の反復学習式により行われる．

$$\boldsymbol{W}^{[i+1]}(f) = \eta \left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \boldsymbol{Y}(f,t)^{\mathrm{H}} \right\rangle_t \right] \boldsymbol{W}^{[i]}(f) + \boldsymbol{W}^{[i]}(f) \tag{3}$$

ここで $\boldsymbol{I}$ は単位行列，$\langle \cdot \rangle_t$ は時間平均演算子，$\boldsymbol{X}^{\mathrm{H}}$ は複素共役転置，$[i]$ は i 番目の反復における値，η は更新係数である．本稿では，非線形関数 $\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))$ を以下のように定義する [15]：

$$\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \equiv \left[e^{\jmath \cdot \arg(Y_1(f,t))}, \ldots, e^{\jmath \cdot \arg(Y_L(f,t))} \right]^{\mathrm{T}} \tag{4}$$

また，$\arg(\cdot)$ は複素数の偏角を求める演算子である．反復学習の後，パーミュテーション問題を，例えば [8, 16] により解決する．

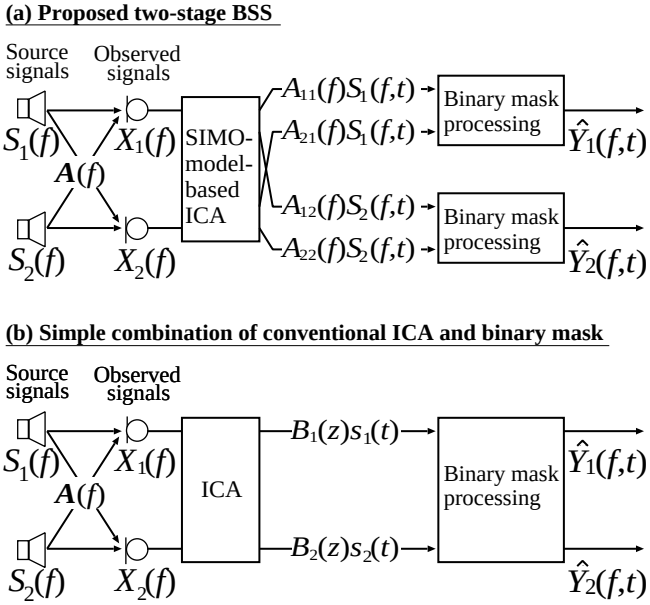


Figure 2: 信号の入出力関係：(a) 提案する 2 段 BSS，(b) 従来の ICA とバイナリマスク処理の単純接続 ($K = L = 2$ の場合に相当)

2.3 従来のバイナリマスクに基づく BSS

バイナリマスク処理 [12, 13, 14] は，ICA に基づかない BSS 手法の一つである．この手法は，聴覚マスキング現象を模擬したものであり，より強い信号は弱い信号をマスクするというものである．バイナリマスクの決定は，耳 (マイクロホン) に接近しているパワーの強い音源成分を選択的に抽出することで行う．この処理は時間–周波数領域で行い，対象音源が支配的な領域はそのまま処理をせず，他の領域ではマスクをかける．ここで，l 番目の音源が l 番目のマイクロホンに近いと仮定し，$L = 2$ の時を考えると，l 番目の分離信号は次の式で与えられる．

$$\hat{Y}_l(f,t) = m_l(f,t) X_l(f,t) \tag{5}$$

ここで $m_l(f,t)$ はバイナリマスク演算子で，$|X_l(f,t)| > |X_k(f,t)|$ $(k \neq l)$ の時は $m_l(f,t) = 1$，その他の時は $m_l(f,t) = 0$ と定義される．

この手法はわずかな計算量しか必要とせず，リアルタイム処理が可能である．しかし，この手法は音源のスペクトル成分間にスパース性，つまり，時間–周波数領域において音源同士で成分の重なりが無いことを仮定しているが，一般的な音響音声信号の混合問題ではこの仮定は満たされないことが多い (実際，音声と一般的な広帯域定常雑音は多くの成分において重なりを持っている)．

3 提案する 2 段型リアルタイム BSS 手法

3.1 概要

近年我々の研究グループにより開発された SIMO-ICA [10, 11] は，混合信号を，モノラル信号ではなく SIMO モデル信号と呼ばれるマイクロホン観測時点でのアレー信号に分

離することができる．このため，SIMO-ICA から得られる各音源に対応する SIMO 成分にバイナリマスク処理を適用することが可能である．提案法の構成を Figure 2(a) に示す．SIMO-ICA の後段に置いたバイナリマスク処理により，少量の計算量で効率的に消し残り成分を取り除くことが出来る．

この 2 段型 BSS の優位性は，SIMO-ICA とバイナリマスク処理の独自の接続方法にある．提案法の新規性を説明するために，従来のモノラル出力型 ICA とバイナリマスク処理を接続した単純 2 段手法 (Figure 2(b)) [17] との比較を以下で行う．

一般的に，従来の ICA は音源信号 $Y_l(f,t) = B_l(f)S_l(f,t) + E_l(f,t)$ $(l = 1, \cdots, L)$, (ここで $B_l(f)$ は任意の歪を表すフィルタ，$E_l(f,t)$ は ICA の不十分な学習に起因する分離残差成分) のみを出力する．残差成分 $E_l(f,t)$ は後処理であるバイナリマスク処理により取り除かれることが期待される．しかし，この組み合わせはとても不確実で，時間–周波数領域でスペクトルに重なりが存在すると動作しない．例えば，ある周波数サブバンドですべての音源が 0 でないスペクトル成分を持っている (スパース性が成り立っていない) 場合，$Y_1(f,t)$ と $Y_2(f,t)$ に対するバイナリマスク処理は正しく決定できず，出力結果は目的音声分が大きく刈り取られたような歪んだ信号となってしまう．このため，従来の ICA とバイナリマスク処理の単純接続は，BSS 問題を解決するには有効ではない．

一方，提案法では初段に SIMO-ICA を適用している．SIMO-ICA により，各音源に対する SIMO 信号 $A_{kl}(f)S_l(f,t)$ は，混合過程の時間差や残差成分をも保って出力される．言うまでも無く，得られた SIMO 成分は，観測マイクロホン時点での音量差を保ったまま分離がされているため，バイナリマスク処理を適用可能である．そのため，SIMO-ICA の後段にバイナリマスク処理を接続する手法は，スパース性の仮定に関係なく効率よく消し残り成分を取り除くことができる．

3.2 アルゴリズム

時間領域 SIMO-ICA [10] は，著者の一人により提案され，ICA の更新時に直接 SIMO モデルに基づく信号を得ることができる．本稿では，時間領域 SIMO-ICA を周波数領域 SIMO-ICA (FD-SIMO-ICA) に拡張する．FD-SIMO-ICA は $(L-1)$ 個の FDICA と単一の *fidelity controller* (FC) から成り，各 ICA は分離システム全体の再現精度を保ちながら並列に動作する．FD-SIMO-ICA における l 番目の ICA $(l = 1, \ldots, L-1)$ は以下で定義される．

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = \left[Y_k^{(\mathrm{ICA}l)}(f,t)\right]_{k1} = \boldsymbol{W}_{(\mathrm{ICA}l)}(f)\,\boldsymbol{X}(f,t) \tag{6}$$

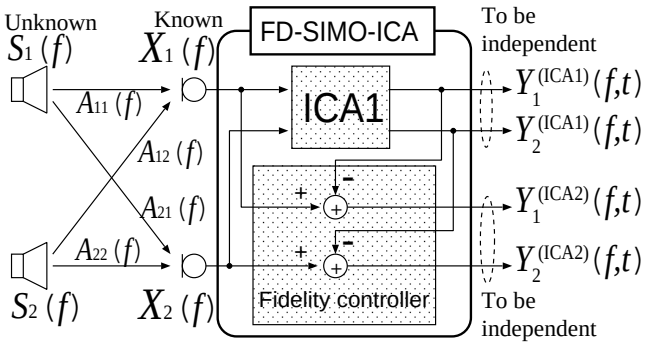


Figure 3: $K = L = 2$ の場合における提案 FD-SIMO-ICA の信号の入出力関係

ここで $\boldsymbol{W}_{(\mathrm{ICA}l)}(f) = [W_{ij}^{(\mathrm{ICA}l)}(f)]_{ij}$ は，l 番目の FDICA における分離フィルタ行列である．また，FC は以下の式で表される信号を出力する．

$$\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t) = \boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) \tag{7}$$

以後，$\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$ を，仮想的な L 番目の ICA の出力とみなす．ここで「仮想的な」という単語を使う訳は，L 番目の ICA はその他の ICA と違い独自の分離フィルタを持たず，$\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$ は $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ $(l = 1, \cdots, L-1)$ に従属しているためである．右辺の第二項 $(-\sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t))$ を左辺に移項すると，式 (7) は，全 ICA の出力ベクトルの和 $\sum_{l=1}^{L} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)$ が，全 SIMO 成分の和 $[\sum_{l=1}^{L} A_{kl}(f)S_l(f,t)]_{k1} (= \boldsymbol{X}(f,t))$ になるための拘束条件になっていることが分かる．

式 (6) により独立な音源が分離され，同時に式 (7) により得られた信号が互いに独立であれば，出力信号は次のような一意な SIMO モデル信号に収束する．

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = \mathrm{diag}\left[\boldsymbol{A}(f)\boldsymbol{P}_l^{\mathrm{T}}\right]\boldsymbol{P}_l\boldsymbol{S}(f,t) \tag{8}$$

ここで $\boldsymbol{P}_l$ $(l = 1, \ldots, L)$ は $\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}$ となるような排他的置換行列群である．この証明は [10] を周波数領域に拡張することにより得られる．式 (8) により与えられる解は，各 l 番目の音源に関する必要十分な SIMO 成分，$A_{kl}(f)S_l(f,t)$，を与えることは明白である．そのため，SIMO-ICA の分離信号は，各音源信号の空間特性を保つことが可能である．$L = K = 2$ の場合，ICA の出力は次の式で与えられる．

$$\begin{aligned}&\left[Y_1^{(\mathrm{ICA1})}(f,t),\ Y_2^{(\mathrm{ICA1})}(f,t)\right]^{\mathrm{T}}\\ &= \left[A_{11}(f)S_1(f,t),\ A_{22}(f)S_2(f,t)\right]^{\mathrm{T}},\end{aligned} \tag{9}$$

$$\begin{aligned}&\left[Y_1^{(\mathrm{ICA2})}(f,t),\ Y_2^{(\mathrm{ICA2})}(f,t)\right]^{\mathrm{T}}\\ &= \left[A_{12}(f)S_2(f,t),\ A_{21}(f)S_1(f,t)\right]^{\mathrm{T}},\end{aligned} \tag{10}$$

ここで，$\boldsymbol{P}_1 = \boldsymbol{I}$，$\boldsymbol{P}_2 = [1]_{ij} - \boldsymbol{I}$ としている．

式 (8) を得るため，式 (7) の Kullback-Leibler Divergence の $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ に関する Natural Gradient を，l 番目 ($l = 1, \ldots, L-1$) の ICA における分離フィルタの Non-holonomic 反復学習式 [5] に追加する必要がある．したがって，FD-SIMO-ICA における l 番目 ($l = 1, \ldots, L-1$) の ICA 部の新しい反復学習アルゴリズムは以下で与えられる．

$$
\begin{aligned}
&\boldsymbol{W}_{(\mathrm{ICA}l)}^{[j+1]}(f) \\
&= \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f) - \alpha \Bigg[\Big\{ \text{off-diag} \Big\langle \boldsymbol{\Phi}\left(\boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)\right) \\
&\quad \boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)^{\mathrm{H}} \Big\rangle_t \Big\} \cdot \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f) \\
&\quad - \Big\{ \text{off-diag} \Big\langle \boldsymbol{\Phi}\big(\boldsymbol{X}(f,t) - \sum_{l'=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l')}^{[j]}(f,t)\big) \\
&\quad \cdot \big(\boldsymbol{X}(f,t) - \sum_{l'=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l')}^{[j]}(f,t)\big)^{\mathrm{H}} \Big\rangle_t \Big\} \\
&\quad \cdot \Big(\boldsymbol{I} - \sum_{l'=1}^{L-1} \boldsymbol{W}_{(\mathrm{ICA}l')}^{[j]}(f)\Big) \Bigg] \qquad (11)
\end{aligned}
$$

ここで，α は更新係数で，非線形ベクトル関数 $\boldsymbol{\Phi}(\cdot)$ は [15] と定義する．

$$
\begin{aligned}
\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \equiv \big[&\tanh(|Y_1(f,t)|) e^{\jmath \cdot \arg(Y_1(f,t))}, \cdots, \\
&\tanh(|Y_L(f,t)|) e^{\jmath \cdot \arg(Y_L(f,t))} \big]^{\mathrm{T}}. \qquad (12)
\end{aligned}
$$

また，$\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ の初期値は全て異なっている必要がある．

FD-SIMO-ICA の後にバイナリマスク処理を行う．FD-SIMO-ICA の出力が式 (9), (10) の場合，音源 1 に対応する出力信号は以下のように与えられる．

$$
\hat{Y}_1(f,t) = m_1(f,t)\, Y_1^{(\mathrm{ICA1})}(f,t) \qquad (13)
$$

ここで，$m_1(f,t)$ はバイナリマスク演算子で，$|Y_1^{(\mathrm{ICA1})}(f,t)| > |Y_2^{(\mathrm{ICA2})}(f,t)|$ のとき $m_1(f,t) = 1$ となり，それ以外は $m_1(f,t) = 0$ となる．また，音源 2 に対応する出力信号も以下のように与えられる．

$$
\hat{Y}_2(f,t) = m_2(f,t)\, Y_2^{(\mathrm{ICA1})}(f,t) \qquad (14)
$$

ここで，$m_2(f,t)$ はバイナリマスク演算子で，$|Y_2^{(\mathrm{ICA1})}(f,t)| > |Y_1^{(\mathrm{ICA2})}(f,t)|$ のとき $m_2(f,t) = 1$ となり，それ以外は $m_2(f,t) = 0$ となる．$L = K > 2$ の場合も同様の手法で簡単に一般化できる．

4 実環境下での実験

4.1 実験条件

Figure 4 に示す実験室において，2 音源 2 マイクロホンで収録した音声信号を用いて，2 音源分離実験を行った．残響時間は 200 ms，収音装置として，ロボット聴覚を模すため Brüel & Kjær 社製の Head And Torso Simulator (HATS; Figure 5 参照) を用いた．2 音声信号が (θ_1, θ_2) と

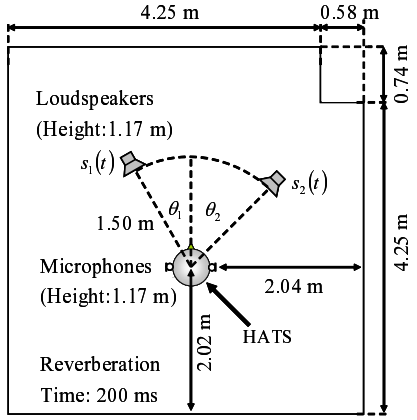


Figure 4: 実験を行った部屋のレイアウト

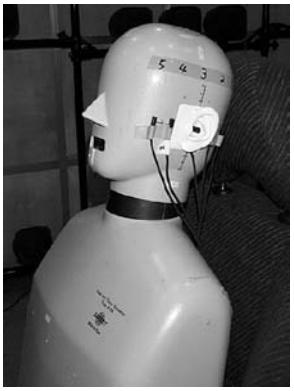


Figure 5: 実験で用いた Head And Torso Simulator

いう異なった方位から到来すると仮定し，方位の組み合わせは $(\theta_1, \theta_2) = (-60°, 60°), (-60°, 0°)$ の 2 通りを実験した．音源信号は，男女 2 名ずつの話者 4 人による組み合わせと，Human Speech Like Noise と呼ばれる定常雑音と話者との組み合わせ，の 2 種類を用いた．この音源データの長さは 3 秒，サンプリング周波数は 8 kHz である．分離フィルタ行列のタップ数は 1024，初期値は ±15° もしくは ±30° の HRTF の逆行列を用いた．

4.2 実験結果

実験では A∼D の 4 手法を比較した：(A) 式 (5) で与えられるバイナリマスク処理，(B) 式 (2) で与えられる従来の ICA，(C) 従来の ICA とバイナリマスク処理の単純接続，(D) 今回の提案法．ここでは，どの手法を用いた分離処理の際にも，音源の DOA 情報，部屋の伝達関数，マイクロホンの配置，HATS(ロボット頭部) の音響特性といった事前情報を一切与えていない．こうした情報は，特に本体やユーザが動き回るロボット環境では，使用することが不可能である．

分離性能の評価値として，出力と入力 Signal-to-Noise Ratio (SNR) の dB 上の差分を表す Noise Reduction Rate (NRR) [8] を用いる．SNR は干渉信号音をノイズとみなすことで計算する．

Figure 6 に，異なる話者配置，異なる初期値ごとの話者同士の場合の NRR の結果を示す．値は 12 話者組み合

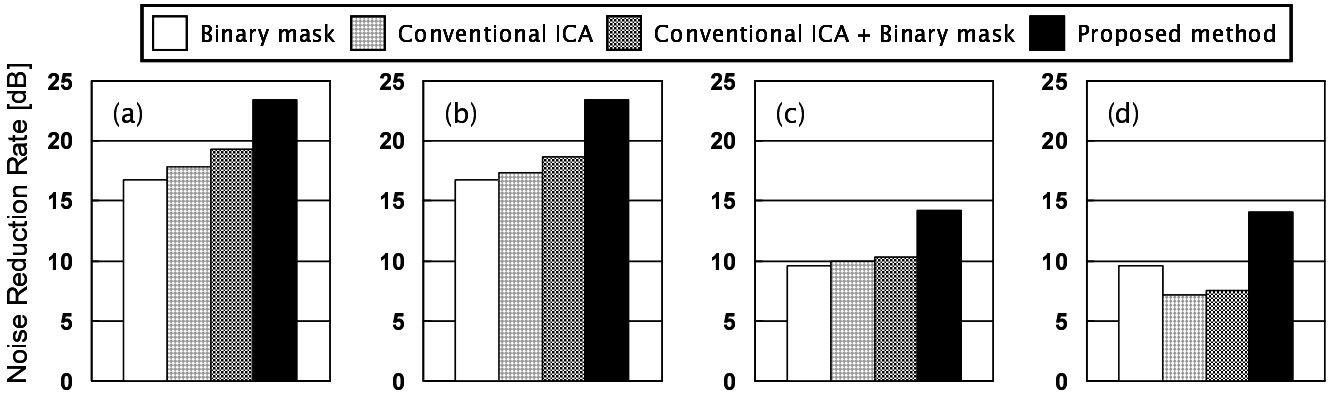


Figure 6: 話者同士の混合時の NRR：(a) 音源方位 $(-60°, 60°)$ &初期値 $\pm 30°$, (b) 音源方位 $(-60°, 60°)$ &初期値 $\pm 15°$, (c) 音源方位 $(-60°, 0°)$ &初期値 $\pm 30°$, (d) 音源方位 $(-60°, 0°)$ &初期値 $\pm 15°$,

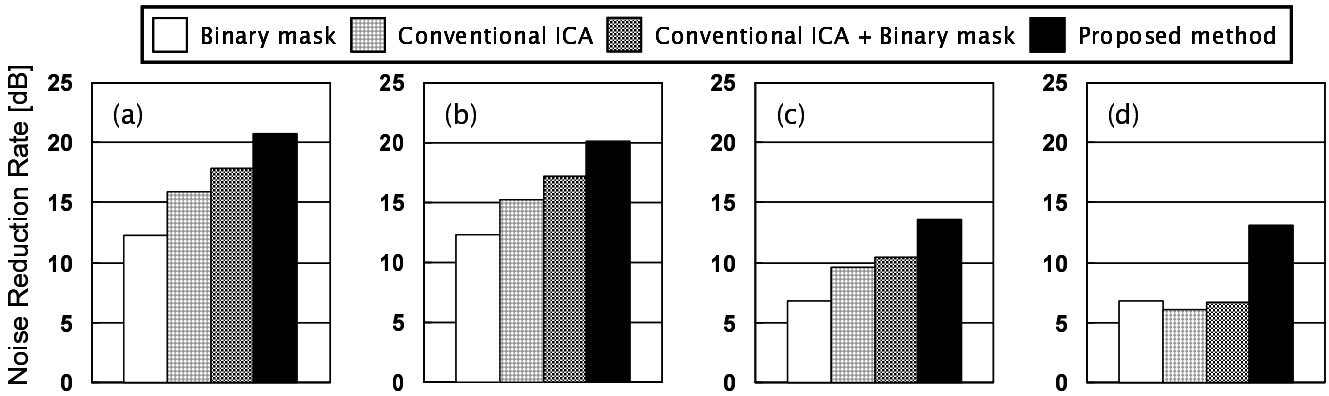


Figure 7: 話者と雑音の混合時の NRR：(a) 音源方位 $(-60°, 60°)$ &初期値 $\pm 30°$, (b) 音源方位 $(-60°, 60°)$ &初期値 $\pm 15°$, (c) 音源方位 $(-60°, 0°)$ &初期値 $\pm 30°$, (d) 音源方位 $(-60°, 0°)$ &初期値 $\pm 15°$

わせの平均である．また，Figure 7 に，スペクトルスパース性が成り立っていない場合に相当する，話者と定常雑音の混合の場合の NRR 結果を示す．これらの結果より，提案する 2 段 BSS 手法は，話者方位や雑音，初期値によらず，一貫して分離性能を大きく改善することができることを確かめることができた．また，提案法は，従来のバイナリマスク処理とは違い，スパース性が成立しない混合問題においても性能の改善が見られることに意味がある．この事実は，提案する SIMO-ICA とバイナリマスク処理の接続法が有効であることを示唆している．

5 リアルタイム実装

我々はすでに DSP を用いてリアルタイム動作する 2 段型 BSS デモシステム (TI-C67, 200 MHz, 150 g; Figure 8 参照) を実装している．提案法をリアルタイム実装したブロック図を Figure 9 に示す．信号は以下の手順で処理される．

1. 入力信号はフレーム毎に高速フーリエ変換 (FFT) を用いて時間-周波数系列に変換される．
2. SIMO-ICA は 3 秒間のデータを使い分離フィルタ行列の推定を行う．推定された分離行列は次の 3 秒のデータに対して用いられる．これは，SIMO-ICA の学習には多くの計算量が必要で，学習中の 3 秒のデータに対して最適化されたフィルタをデータ自身に適用することが不可能なためである．
3. SIMO-ICA より得られた分離信号に対しバイナリマスク処理を行う．SIMO-ICA とは違い，バイナリマスクはリアルタイムに信号を処理する．
4. バイナリマスク処理を行った信号に逆 FFT を適用することで，時間領域の波形に変換する．

SIMO-ICA の分離フィルタの更新はリアルタイムではなく 3 秒の遅延が生じるが，システム全体で見ると，バイナリマスク処理が遅延なしで動作しているため，リアルタイムで系に追従しているように見える．一般的に，従来の ICA で生じるフィルタ更新遅延の影響はリアルタイムシステムに適用するには問題になるほど大きい．しかし，提案法における SIMO-ICA 部分のフィルタ更新遅延の影響は，リアルタイム動作可能なバイナリマスク処理を導入することにより大きく軽減されている．

6 まとめ

本稿では，SIMO モデルに基づく ICA とバイナリマスク処理の効果的な組み合わせによる新しい BSS 手法を提案した．その有効性を評価するため，残響環境下での分離実験を行った．実験結果から，提案した 2 段 BSS を用いると，NRR が大いに改善されることが示された．それに

Figure 8: 開発したリアルタイム BSS モジュール

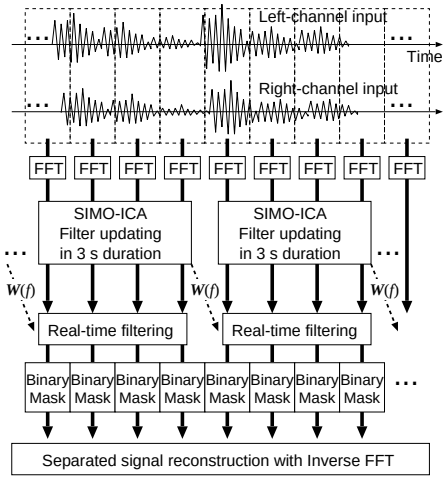


Figure 9: 提案法のリアルタイム処理時の信号の流れ

加え，提案法が，従来の ICA やバイナリマスク処理単体，それらの従来法の単純組み合わせ手法の性能を上回ることが示された．

参考文献

[1] K. Nakadai, D. Matsuura, H. Okuno, and H. Kitano, “Applying scattering theory to robot audition system: robust sound source localization and extraction,” *Proc. IROS-2003*, pp.1147–1152, 2003.

[2] R. Nishimura, T. Uchida, A. Lee, H. Saruwatari, K. Shikano, and Y. Matsumoto, “ASKA: Receptionist robot with speech dialogue system,” *Proc. IROS-2002*, pp.1314–1317, 2002.

[3] R. Prasad, H. Saruwatari, and K. Shikano, “Robots that can hear, understand and talk,” *Advanced Robotics*, vol.18, pp.533–564, 2004.

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

[5] N. Murata and S. Ikeda, “An on-line algorithm for blind source separation on speech signals,” *Proc. NOLTA98*, vol.3, pp.923–926, 1998.

[6] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol.22, pp.21–34, 1998.

[7] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources,” *IEEE Trans. Speech & Audio Processing*, vol.8, pp.320–327, 2000.

[8] 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, pp.1135–1146, 2003.

[9] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech & Signal Process.*, vol.ASSP-27, no.2, pp.113–120, 1979.

[10] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, “High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning,” *Proc. IWAENC2003*, pp.251–254, 2003.

[11] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, “High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis,” *IEICE Trans. Fundamentals*, vol.E87-A, no.8, pp.2063–2072, 2004.

[12] R. Lyon, “A computational model of binaural localization and separation,” *Proc. ICASSP83*, pp.1148–1151, 1983.

[13] N. Roman, D. Wang, and G. Brown, “Speech segregation based on sound localization,” *Proc. IJCNN01*, pp.2861–2866, 2001.

[14] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, ”Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” *Acoustical Science and Technology*, vol.22, no.2, pp.149–157, 2001.

[15] H. Sawada, R. Mukai, S. Araki, and S. Makino, “Polar coordinate based nonlinear function for frequency domain blind source separation,” *IEICE Trans. Fundamentals*, vol.E86-A, no.3, pp.590–596, 2003.

[16] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *Proc. Int. Sympo. on ICA and BSS*, pp.505–510, 2003.

[17] M. Aoki and K. Furuya, ”Using spatial information for speech enhancement,” *Technical Report of IEICE*, vol.EA2002-11, pp.23–30, 2002 (in Japanese).

適応雑音推定処理を備えた空間的サブトラクションアレーによる実環境下でのハンズフリー音声認識

Hands-Free Speech Recognition Using Spatial Subtraction Array with Adaptive Noise Estimation Prosessing under Real Environment

木内千絵，高谷智哉，猿渡洋，鹿野清宏

Chie Kiuchi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano

奈良先端科学技術大学院大学

Nara Institute of Science and Technology

chie-k@is.naist.jp

Abstract

We newly propose an improved spatial subtraction array (SSA) with an adaptive noise estimation processing, which aims at the achievement of robust hands-free speech recognition in real environments. The previously proposed SSA can recognise a target speech with a high accuracy under a laboratory environment. However the conventional SSA used an ideally designed null beamformer (NBF) for noise estimation, and consequently it cannot take into account the reverberation effect which arises in an actual environment. The proposed SSA introduces adaptive beamformer (ABF) for the accurate noise estimation, and thereby remarkably improves the noise subtraction performance even under real reverberant conditions. The speech recognition experiments reveal that the word accuracy of the proposed SSA is superior to that of the conventional SSA as well as the conventional delay-and-sum beamformer and adaptive beamformer.

1 はじめに

高精度なハンズフリー音声認識の実現を目標として，新しいマイクロホンアレー信号処理技術である空間的サブトラクションアレー (SSA) が提案されている [1]．SSA は，従来音声強調手法である Delay-and-Sum アレー (DS)[2] や Griffith-Jim 型適応アレー (GJ)[2],[3] と異なり，アレー信号処理を音声認識システムに特化させた手法である．DS は，各マイクロホンアレーで受音した信号に対し目的方位に同位相化を行い，その同位相化された信号を足し合わせることで，目的方位の信号を強調する．ただし，DS において，低周波数帯域で位相差情報を得るには，広いマイクロホン素子間隔が必要となる．そのため，高精度な音声強調を行うには大規模なマイクロホンアレーが必要となってしまう．一方，GJ は DS よりも小規模なマイクロホンアレーで高精度な雑音抑圧が可能である．GJ は，DS によって得られる音声強調を行った信号から適応フィルタによって推定される雑音を減算することで，雑音抑圧を行う．GJ は適応フィルタを用いるので，高精度な雑音推定が可能であるが，GJ で使用している適応フィルタは，フィルタの学習に多大な演算量が必要となるので高速処理が困難であるといった問題がある．また，DS や GJ は時間波形を出力するため，スペクトル特徴量等を入力とする音声認識処理系にとっては冗長な処理を含む．一方 SSA は，音声認識を行う際の特徴量である Mel Frequency Cepstrum Coefficient (MFCC)[4] を直接出力し，波形再構成等の冗長な処理を行わない．また，パワースペクトル上でスペクトル減算に基づく雑音抑圧を行うので，頑健な雑音抑圧が可能である．さらに，フィルタバンク数程度のパラメータで動作するので，高速処理が可能である．

我々は，雑音推定フィルタに死角制御型ビームフォーマ (NBF)[5] を用いた NBF 型 SSA を既に提案しており [1]，実験室で収録された雑音を重畳した残響・雑音付加音声データ (実験室データ) を用いた認識実験では NBF 型 SSA の有効性を確認している．本稿では，実環境で収録された雑音を重畳した残響・雑音付加音声データ (実環境データ) を用いて NBF 型 SSA が実環境でも有効であるのかを検証し，さらに認識精度向上の為，雑音推定部に適応雑音推定処理を備えた新しい SSA(ABF 型 SSA) を提案する．特に，異なる実験環境における音声認識実験を行い，NBF 型 SSA と ABF 型 SSA の違い及び有効性を検証する．

2 従来法: GJ 型適応アレー [2],[3]

本研究では，SSA の比較対象手法として，SSA の主パス部分に使用している DS と，SSA の従来法である GJ を選択している．以下では，特に GJ について概説する．

2.1 GJ の原理

GJ における信号の流れを図 1 に示す．ここで，$x_j(t)$ $(j = 1, \cdots, J)$ は各マイクロホンアレー素子で受音された観測信号であり，J はマイクロホンの素子数を示す．GJ は，ユーザ音声の強調を行う主パス，雑音推定を行う参照パス，そして主パスの信号から参照パスの信号を減算する箇所から成る．GJ の出力である $z(t)$ は，以下の式で計算された後，逆フーリエ変換を行うことによって，波形再構成の後に時間波形として出力される．

$$Z(k) = Y_{\mathrm{DS}}(k) - \boldsymbol{A}_{\mathrm{ADF}}^{\mathrm{T}}(k)\boldsymbol{Y}_{\mathrm{SUB}}(k) \qquad (1)$$

ここで，T は転置を表し，k は離散スペクトルの周波数のビン番号である．$Y_{\mathrm{DS}}(k)$ は，主パスで DS によって音声強調された信号である．また $\boldsymbol{Y}_{\mathrm{SUB}}(k) = [Y_1^{\mathrm{SUB}}(k), \cdots, Y_{J-1}^{\mathrm{SUB}}(k)]^{\mathrm{T}}$ は，同位相化された隣り合う信号同士を，参照パスで減算することによって得られた雑音推定信号ベクトルである．さらに，$\boldsymbol{A}_{\mathrm{ADF}}(k) = [A_1^{\mathrm{ADF}}(k), \cdots, A_{J-1}^{\mathrm{ADF}}(k)]^{\mathrm{T}}$ は，参照パス内の適応フィルタベクトルである．この適応フィルタの係数は，最小二乗法 (LMS) によって出力のパワー $|Z(k)|^2$ を最小化するように求められる．これは，以下の式によって表される．

$$\boldsymbol{A}_{\mathrm{ADF}}^{(h+1)}(k) = \boldsymbol{A}_{\mathrm{ADF}}^{(h)}(k) + \mu \cdot Z^{(h)}(k)\boldsymbol{Y}_{\mathrm{SUB}}^{(h)}(k) \qquad (2)$$

ここで，μ はステップサイズパラメータであり，(h) は更新回数を示す．

2.2 GJ の問題点

GJ は式 (2) に示されるように，雑音のみが存在しユーザ音声が発せられていない無音区間において，一周波数ビンあたり $J-1$ 次元の適応フィルタ係数を収束するまで更新する必要がある．そのため，数千〜数万個のパラメータ更新といった多大な演算量が必要になる．また主パスと参照パスにおける減算は，振幅スペクトルと位相情報を共に必要とするため条件が複雑となり，あまり頑健な雑音抑圧ができないといった問題がある．さらに，従来法である DS や GJ は時間波形を出力するため，スペクトル特徴量等を入力とする音声認識処理系にとって冗長である．そこで本研究では，GJ を音声認識システムに特化するように拡張した SSA を提案する．

3 SSA[1]

3.1 原理

SSA における信号の流れを図 2 に示す．ここで，$x_j(t)$ $(j = 1, \cdots, J)$ は各マイクロホン素子で受音された観測信号であり，J はマイクロホン素子数を示す．$Y(k)$ はユーザ方位 θ に同位相化した遅延和アレー出力のスペクトル，$Z(k)$ は雑音推定後のスペクトルである．

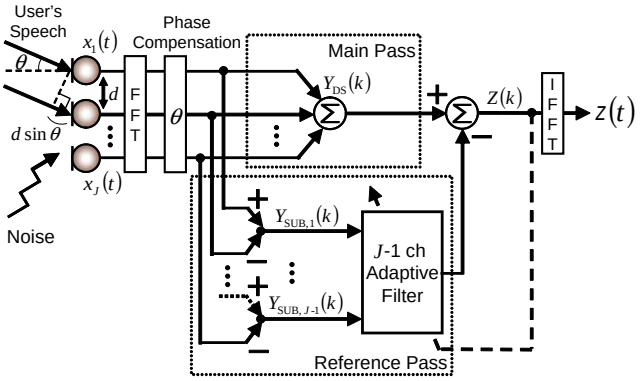


Figure 1: GJ における信号の流れ

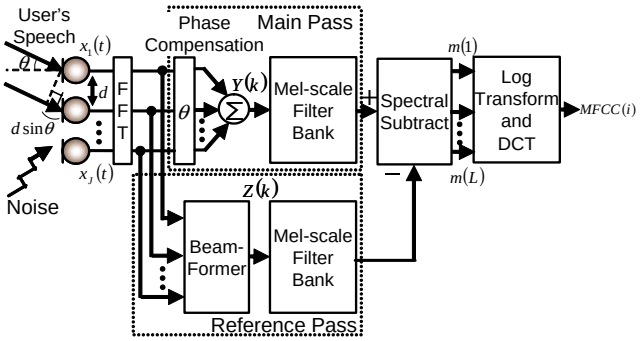


Figure 2: SSA における信号の流れ

SSA は，GJ と同様に音声を強調する主パスと雑音を推定する参照パスから成る．主パスでは，DS を用いて各マイクロホン素子で受音された信号をユーザ方位に同位相化し，その各スペクトルを足し合わせることでユーザ音声を強調している．参照パスでは，ビームフォーマを用いてユーザ音声を抑圧し雑音を推定している．また，SSA は最終出力として時間波形ではなく，音声認識を行う際の特徴量である MFCC を出力する．従って，両パスにおいてフィルタバンク分析 [4] を行う必要があり，各パスにメルフィルタバンクを挿入している．SSA はフィルタバンク分析を行った後，主パスで得られたユーザ音声を強調したスペクトル $Y(k)$ から参照パスで得られた雑音を推定したスペクトル $Z(k)$ を減算する．

3.2 フィルタバンク分析 [4]

SSA で最終出力として MFCC を出力する際，フィルタバンク分析を行う必要がある．そのため，以下に示す L 個の三角窓 $W(k;l)$ $(l = 1, \cdots, L)$ を周波数軸上に配置する．図 3 にメルフィルタバンクの配置図を示す．

$$W(k;l) = \begin{cases} \dfrac{k - k_{\mathrm{lo}}(l)}{k_{\mathrm{c}}(l) - k_{\mathrm{lo}}(l)} & \big(k_{\mathrm{lo}}(l) \leq k \leq k_{\mathrm{c}}(l)\big) \\ \dfrac{k_{\mathrm{hi}}(l) - k}{k_{\mathrm{hi}}(l) - k_{\mathrm{c}}(l)} & \big(k_{\mathrm{c}}(l) \leq k \leq k_{\mathrm{hi}}(l)\big) \end{cases} \qquad (3)$$

ここで $k_{\rm lo}(l)$, $k_{\rm c}(l)$, $k_{\rm hi}(l)$ はそれぞれ l 番目のフィルタの下限，中心，上限の周波数番号であり，隣り合うフィルタ間で以下の関係を持つ．

$$k_{\rm c}(l) = k_{\rm hi}(l-1) = k_{\rm lo}(l+1) \tag{4}$$

さらに，$k_{\rm c}(l)$ はメル周波数軸上で等間隔に配置される．$k_{\rm c}(l)$ に対するメル周波数 $Mel_{k_{\rm c}(l)}$ は以下の式によって計算される．

$$Mel_{k_{\rm c}(l)} = 2595\log_{10}\left\{1 + \frac{k_{\rm c}(l)\cdot f_s}{1400(N-1)}\right\} \tag{5}$$

ここで f_s はサンプリング周波数，N は FFT 長である．

3.3 雑音抑圧処理

フィルタバンク分析を行った後，以下の式 (6) のように主パスで得られたユーザ音声を強調したスペクトル $Y(k)$ から参照パスで得られた雑音を推定したスペクトル $Z(k)$ を減算する．また減算時に，パワースペクトルが負になった場合のミュージカルノイズを回避するために，式 (7) のフロアリング処理を行う．

$$m(l) = \sum_{k=k_{\rm lo}(l)}^{k_{\rm hi}(l)} W(k;l)\{|Y(k)|^2 - \alpha(l)\cdot\beta\cdot|Z(k)|^2\}^{\frac{1}{2}} \quad (\text{ if } \ |\mathrm{Y(k)}|^2 - \alpha(\mathrm{l})\cdot\beta\cdot|\mathrm{Z(k)}|^2 \geq 0\ ) \tag{6}$$

$$m(l) = \sum_{k=k_{\rm lo}(l)}^{k_{\rm hi}(l)} W(k;l)\{\gamma\cdot|Y(k)|\} \quad (\text{ otherwise }) \tag{7}$$

ここで，$m(l)$ はメルフィルタバンク上で雑音抑圧処理を行うことによって得られた l 番目帯域の振幅スペクトル和である．β は L 個の三角窓に対して一定の減算係数である．また，γ も各三角窓に対して一定のフロアリング係数である．$\alpha(l)$ は，雑音の成分が 0 になるよう各次元で調節を行うパラメータであり，ユーザが発話せず雑音のみが存在する区間で以下のように計算される．

$$\begin{aligned}\alpha(l) &= \left(\frac{m_{\rm Y}(l)}{m_{\rm Z}(l)}\right)^2 \\ &= \left(\frac{\sum_{k=k_{\rm lo}(l)}^{k_{\rm hi}(l)} W(k;l)|Y(k)|}{\sum_{k=k_{\rm lo}(l)}^{k_{\rm hi}(l)} W(k;l)|Z(k)|}\right)^2\end{aligned} \tag{8}$$

ここで，$m_{\rm Y}(l)$, $m_{\rm Z}(l)$ は，それぞれ主パス，参照パスで得られた振幅スペクトル和である．

一般的に，音声認識は位相情報を陽には用いないため，パワースペクトル上のみで減算を行う SSA は音声認識に対して有効である．また，$m(l)$ は通常 24 個から成り，SSA はその各 $m(l)$ に対するパラメータを調節するだけで雑音抑圧処理を行うことができる．従って，GJ のように数百

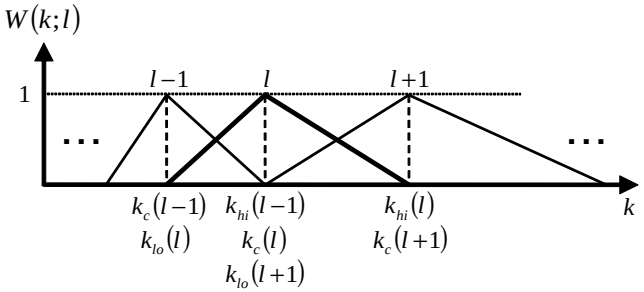


Figure 3: メルフィルタバンク

〜数千次元のパラメータを収束するまで学習する必要がない．その結果，SSA では高速処理が可能となる．SSA で出力される MFCC パラメータは，式 (6) 及び (7) で得られた $m(l)$ の対数値を離散コサイン変換することで求められる．

$$MFCC(i) = \sqrt{\frac{2}{L}}\sum_{l=1}^{L}\log_e\{m(l)\}\cos\left\{\left(l-\frac{1}{2}\right)\frac{i\pi}{L}\right\} \tag{9}$$

ここで，i は MFCC の次元数を表す．

3.4 死角制御型ビームフォーマによる雑音推定処理 [5]

SSA は，主パスにおけるユーザ音声の強調とは別に，参照パスにおいて雑音のスペクトルを推定し，その推定したスペクトルを主パスで強調されたユーザ音声のスペクトルから減算することによって，効率的にユーザ音声のスペクトルを推定する．NBF 型 SSA では，その雑音推定フィルタとして，目的方位に死角を形成することで目的音声を抑圧し雑音を推定する NBF を使用している．

NBF 型 SSA の参照パスにおける雑音推定処理を定式化する．ここで，$\boldsymbol{X}(k)$ が観測信号ベクトル $[x_1(t),\cdots,x_J(t)]^{\rm T}$ を離散フーリエ変換したものだとすると，参照パスにおいて NBF フィルタを適用した各周波数ビンの出力 $Y_{\rm NBF}(k)$ は以下の式 (10) で表すことができる．

$$Y_{\rm NBF}(k) = \boldsymbol{A}_{\rm NBF}^{\rm T}(k)\boldsymbol{X}(k) \tag{10}$$

$\boldsymbol{A}_{\rm NBF}(k)$ は NBF で設計されるフィルタであり，以下の式で設計される．

$$\begin{aligned}\boldsymbol{A}_{\rm NBF}(k) &= [A_1^{\rm NBF}(k),\cdots,A_J^{\rm NBF}(k)]^{\rm T} \\ &= \left\{[1,0]\cdot[\boldsymbol{e}_{\rm NBF}(k,\theta_{\rm N}),\boldsymbol{e}_{\rm NBF}(k,\theta_{\rm U})]^{+}\right\}^{\rm T}\end{aligned} \tag{11}$$

ここで，$+$は擬似逆行列，$\theta_{\rm N}$ は雑音の方位，$\theta_{\rm U}$ はユーザ方位であり，$\boldsymbol{e}_{\rm NBF}(k,\theta)$ は，式 (12) で計算されるステアリングベクトルである．

$$\begin{aligned}\boldsymbol{e}_{\rm NBF}(k,\theta) &= \big[\exp(j2\pi\frac{k}{K}f_s d_1\sin\theta/c) \\ &\qquad \cdots,\exp(j2\pi\frac{k}{K}f_s d_J\sin\theta/c)\big]^{\rm T}\end{aligned} \tag{12}$$

ここで，c は音速，d_1〜d_J はマイクロホン素子座標である．

3.5 NBF 型 SSA の問題点

NBF フィルタは，無残響かつマイクロホン素子誤差がない状態を仮定して計算機上で設計されるため，実際の環境に適応したフィルタは設計できない．一般に，無残響もしくは残響の少ない環境であれば，ユーザ方位からのみ到来する信号がマイクロホンアレーで受音されるため，NBF フィルタでユーザ音声を抑圧することは可能である．しかし実環境では，残響等の影響により各周波数ごとに到来する信号の方位が異なるため，NBF フィルタではユーザ音声を十分に抑圧することは困難である．また，NBF は素子誤差に非常に敏感であり，誤差の影響により死角が目的方位からずれてしまうという問題もある．

3.6 既知雑音重畳を用いた音韻モデルとの適合化

残響雑音下では SSA で雑音除去を行っても，残留雑音やフロアリング処理による歪が存在し完全な雑音抑圧は困難である．つまり，認識を行う際にクリーンモデルを用いてしまうと音韻モデルとの不一致により，挿入誤りが多く生じてしまう．そこで，山出らによって提案された既知雑音とマッチドモデルを用いる手法 [6] を導入している．ただし，SSA は MFCC を直接出力するのでメルフィルタバンク上で重畳するように拡張してある [1].

4 提案法: ABF を用いた SSA (ABF 型 SSA)

4.1 適応雑音推定処理

3.5 節の問題点を解決し認識精度を向上させるため，本研究では，NBF 型 SSA の参照パスに適応雑音推定処理として適応ビームフォーマ (ABF) を導入した新しい SSA アルゴリズム (ABF 型 SSA) を提案する．本研究ではこの適応フィルタの設計に，目的方位の利得を 1 に保ちながら非目的音出力を最小とする Frost 型アレー [7] を使用する．一般の Frost 型アレーの使用においては，ユーザ音声の強調が目的のため目的方位はユーザ音声であるが，本研究における ABF の役割は参照パスにおける雑音スペクトルの推定であるため，ユーザ音声が非目的音に相当する．つまり，雑音方位の利得を 1 に保ちながらユーザ音声を最小にするフィルタを設計する．

従来の Frost 型 ABF フィルタの役割と SSA での Frost 型 ABF フィルタの役割を，図 4 及び 5 にそれぞれ示した．従来の Frost 型 ABF では，図 4 上段の図のようにユーザ音声が発せられておらず雑音のみが存在する区間でその雑音を学習し，それを最小化するようなフィルタを設計する．この場合は，ユーザ音声が目的音であり雑音が非目的音となる．そして，図 4 下段の図のように実際にユーザ音声と雑音の両方が存在する区間で，設計したフィルタを適用する．すると，フィルタは入力されてくる雑音を最小化するように設計されているので，ユーザ音声つまり目的音が得られる．

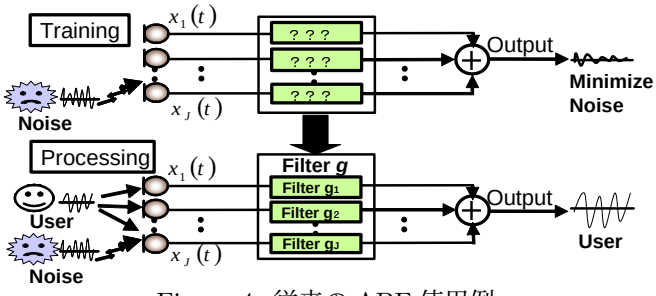


Figure 4: 従来の ABF 使用例

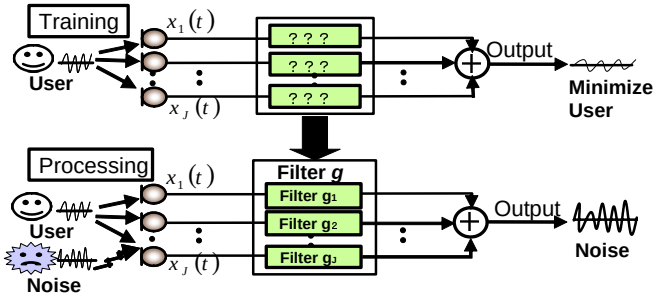


Figure 5: SSA の参照パスにおける ABF 使用例

一方 SSA で使用する Frost 型 ABF は，図 5 上段の図のように雑音がなくユーザ音声のみが存在する区間でそのユーザ音声を学習し，それを最小化するようなフィルタを設計する．この場合は，SSA の参照パスが雑音推定の役割を果たすので，雑音が目的音でありユーザ音声が非目的音である．そして，図 5 下段の図のように実際にユーザ音声と雑音の両方が存在する区間で，設計したフィルタを適用する．すると，フィルタは入力されてくるユーザ音声を最小化するように設計されているので，雑音つまり SSA の参照パスにおける目的音が得られる．

4.2 ABF の定式化

ABF を SSA に応用するための定式化を行う．SSA の参照パスにおける目的音の到来方位つまり雑音方位を θ_{N}，また SSA の参照パスにおける非目的音信号つまりユーザ音声をマイクロホンアレーで観測し，短時間 DFT によって時間-周波数系列にしたものを $\boldsymbol{X}(k,n)$ 及び ABF で設計するフィルタを $\boldsymbol{G}(k)$ とする．ここで，$\boldsymbol{X}(k,n)$ は，雑音が存在せずユーザ音声のみが発せられている区間つまり，非目的音のみが信号を発している区間で観測される．また，非目的音信号 $\boldsymbol{X}(k,n)$ にフィルタ $\boldsymbol{G}(k)$ を適用して得られる出力 $Y(k,n)$ は，式 (13) で表すことができる．

$$Y(k,n) = \boldsymbol{G}(k)^{\mathrm{T}} \boldsymbol{X}(k,n) \tag{13}$$

$$\boldsymbol{G}(k) \;=\; [G_1(k), \cdots, G_J(k)]^{\mathrm{T}} \tag{14}$$

ここで，n はフレーム番号である．ABF の設計は, 非目的音のみが信号を発している区間 (非目的音区間) で非目的音を学習し，目的方位 θ_{N} からの利得を 1 に保ちながら出力 $Y(k,n)$ を最小化するフィルタ $\boldsymbol{G}(k)$ を求める条件付最

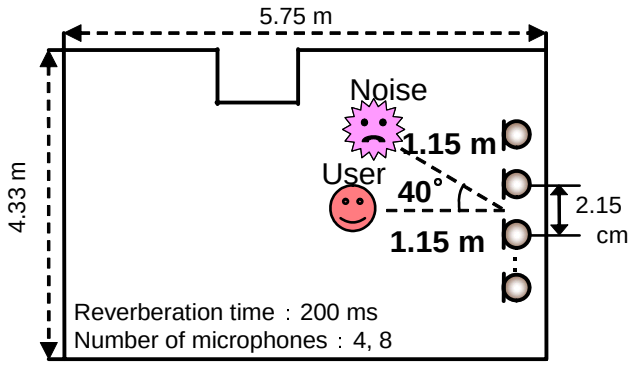


Figure 6: 実験環境 1

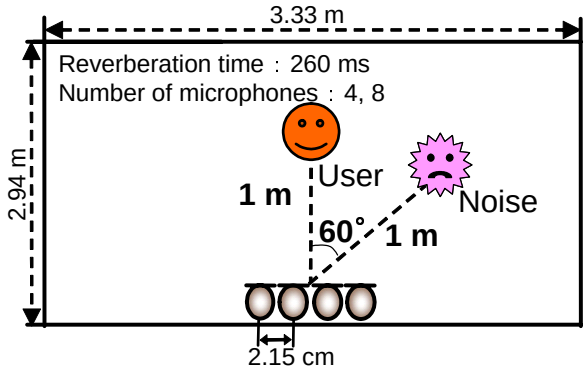


Figure 7: 実験環境 2

小化問題であり，その解は以下の式 (15) で与えられる．

$$\boldsymbol{G}(k) = \frac{\boldsymbol{d}_{\theta_{\mathrm{N}}}(k)^{\mathrm{H}}\boldsymbol{R}(k)^{-1}}{\boldsymbol{d}_{\theta_{\mathrm{N}}}(k)^{\mathrm{H}}\boldsymbol{R}(k)^{-1}\boldsymbol{d}_{\theta_{\mathrm{N}}}(k)} \tag{15}$$

$$\boldsymbol{R}(k) \equiv \langle \boldsymbol{X}(k,n)\boldsymbol{X}(k,n)^{\mathrm{H}} \rangle_{t \in B} \tag{16}$$

$$\boldsymbol{d}_{\theta_{\mathrm{N}}}(k) \equiv \begin{bmatrix} \exp[j2\pi \frac{k}{K} f_s d_1 \sin(\theta_{\mathrm{N}})/c] \\ \vdots \\ \exp[j2\pi \frac{k}{K} f_s d_J \sin(\theta_{\mathrm{N}})/c] \end{bmatrix} \tag{17}$$

ここで，$\boldsymbol{R}(k)$ は非目的音の相関行列，$\boldsymbol{d}_{\theta_{\mathrm{N}}}(k)$ は目的方位 θ_{N} に関するステアリングベクトル [2] である．B は非目的音区間のフレーム番号の集合である．また，H は複素共役転置を表す．

5 実験

5.1 実験環境

実験を行った室内環境を図 6 及び 7 に示す．図 6 は音響実験室，図 7 は実際のマンションの部屋である．以後，それぞれを「実験環境 1」及び「実験環境 2」と呼ぶ．

5.2 実験 1: 指向特性

実際に設計した NBF フィルタと ABF フィルタの指向特性を比較する．指向特性は，周波数と方位そしてゲインの 3 次元で表示している．図 8 及び 9 に NBF フィルタと ABF フィルタの指向特性を示した．フィルタは，0 °方位からユーザが発話していると仮定して，死角を形成して

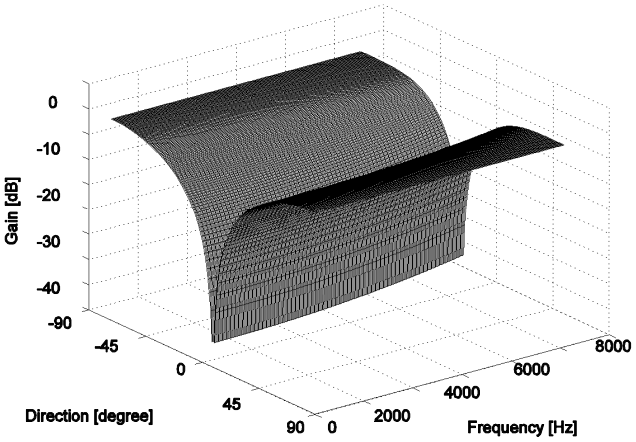


Figure 8: NBF の理想指向特性例

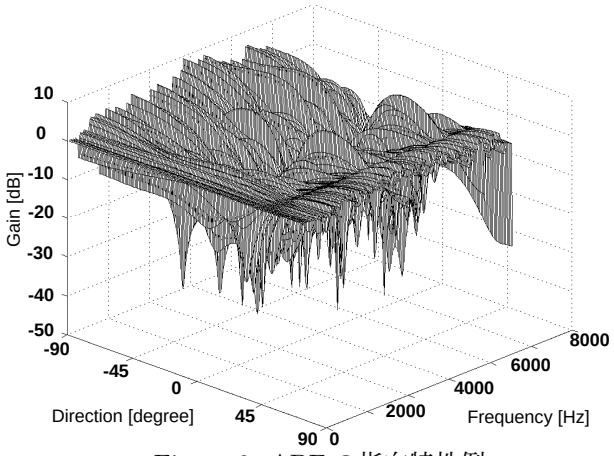


Figure 9: ABF の指向特性例

いる．NBF フィルタは，図 8 に示すように，残響がなくかつマイクロホン素子誤差もない状態を仮定して計算機上で設計されるので 0 °方位のみに鋭い死角を形成する非常に整った指向特性を持つことがわかる．一方 ABF フィルタは，図 9 に示すように，実際に実環境で収録されたデータを使用して設計されるので，残響やマイクロホン素子誤差も考慮された指向特性を持つことがわかる．従って，NBF のような鋭い死角は持たず，その環境に適した指向特性を形成する．

5.3 実験 2: 音声認識実験

図 6 及び 7 の二つの異なる実験環境において，NBF 型 SSA と ABF 型 SSA の認識精度を比較するための音声認識実験を行った．また，従来法として DS 及び GJ も同様に実験を行った．クリーン音声データベースに図 6 及び 7 の環境で計測されたインパルス応答を畳み込み，音声に SNR が平均 10 dB になるように雑音を重畳した音声を評価データとして用いた．減算係数 β 及びフロアリング係数 γ については，それぞれのパラメータを変化させ音声認識実験による単語認識精度を基に最適なものを選んだ．ここで，認識に使用する音韻モデルは PTM[8](2000 状態，64 混合) の既知雑音重畳モデル [6] を使用した．

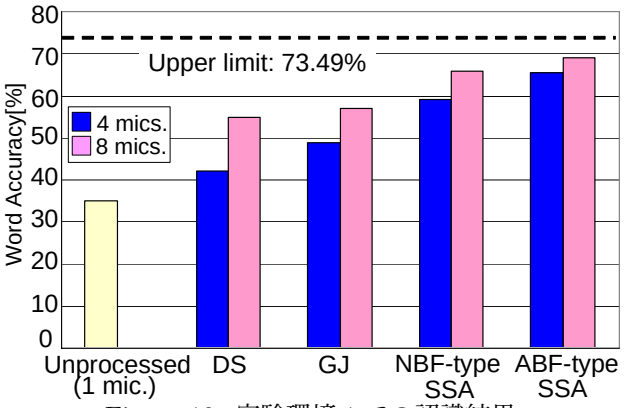


Figure 10: 実験環境 1 での認識結果

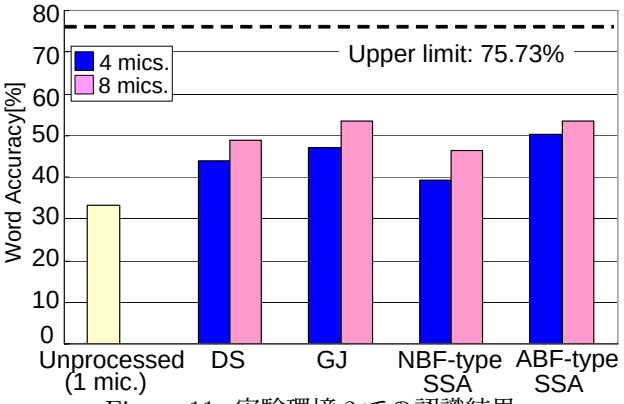


Figure 11: 実験環境 2 での認識結果

5.4 認識精度比較

図 10 及び 11 に，各環境における各手法の音声認識精度を示す．また，各図における点線は，各環境において雑音信号が発生していない場合の単一マイクロホンでの認識精度 (Upper limit) を示している．実験環境 1 において，NBF 型 SSA は DS 及び GJ 以上の認識精度を達成し，ABF 型 SSA でも NBF 型 SSA，DS 及び GJ 以上の認識精度を達成した．しかし，実験環境 2 において，NBF 型 SSA は DS 及び GJ に劣る認識精度となった． この原因として，実験環境 1 と実験環境 2 における残響の違いが考えられる．実験環境 2 のほうが実験環境 1 よりも残響時間が長く残響の影響は大きい．従って，NBF 型 SSA の参照パスでは実験環境 2 においてユーザ音声が十分に抑圧されておらず，その結果として目的音声に歪が生じ，認識精度が GJ や DS に劣ってしまったと考えられる．一方，ABF 型 SSA は NBF 型 SSA，DS 及び GJ と同等もしくはそれ以上の認識精度を達成した．これは，残響の影響があっても ABF 型 SSA の参照パスではユーザ音声が十分に抑圧されており，正確に雑音推定ができているからだと考えられる．以上より，ABF 型 SSA は，適応雑音推定処理によって環境に適応した雑音推定が可能であることが確認された．

6 まとめ

本稿では，参照パスに適応雑音推定処理を備えた新しい空間的サブトラクションアレー (ABF 型 SSA) を提案した．ABF 型 SSA は適応雑音推定処理を備えているため，環境に適応した雑音推定が可能である．実験 1 の指向特性結果より，NBF フィルタと ABF フィルタの指向特性の違いが明確になった．実データを用いて設計する ABF は，残響等の影響もフィルタに含めて設計する．従って ABF フィルタは，NBF フィルタのように鋭い死角は形成せず，残響特性に応じた指向特性を形成する．また，実験 2 の結果より NBF 型 SSA は残響の影響により認識精度が劣化するのに対し，ABF 型 SSA では頑健に動作していることが確認された．従って，本提案法である ABF 型 SSA はハンズフリー音声認識において非常に有効な手法であると言える．

今後の課題としては，実環境に存在する残響の影響を考慮し，認識に残響・既知雑音重畳モデルを使用するなど，音声認識面での精度向上も考えていかなければならない．

謝辞 この研究の一部は，文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われたものである．

参考文献

[1] Y. Ohashi, T. Nishikawa, H. Saruwatari, A. Lee, K. Shikano, “Noise-Robust Hands-free Speech Recognition Based on Spatial Subtraction Array and Known Noise Superimposition,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.533–537, 2005.

[2] 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,” コロナ社, 1995.

[3] 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.

[4] 鹿野 清宏, 伊藤 克亘, 河原達也, 武田一哉, 山本 幹雄, “音声認識システム,” オーム社, 2001.

[5] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, 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.

[6] 山出 慎吾, 馬場 朗, 芳澤 伸一, 李 晃伸, 猿渡 洋, 鹿野 清宏, 電子情報通信学会論文誌, vol.J-87-D-II, no.4, pp.933–941, 2004.

[7] O.L. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol.60, no.8, pp.926–935, 1972.

[8] A. Lee, T. Kawahara, K. Takeda, K. Shikano, “A new phonetic tied-mixture model for efficient decoding,” *Proc. ICASSP*, vol. III, pp.1269–1272, 2000.

脳型情報処理から見たロボット聴覚：「脳とからだをもった耳」

Robot audition from the viewpoint of brain-like information processing

辻野広司

Hiroshi Tsujino

㈱ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

tsujino@jp.honda-ri.com

Abstract— A more “intellectual” function is expected to the present, the robot that movement with a complicated action was enabled. Conventional robot intelligence assumed definite input and output information and realized an intellectual function by symbolizing informational relations to some extent. On the other hand, the pursued intellectual functions premise high dimensional input/output and online processing, and there are relatively few subjects to cut by function realization by what we symbolize definitely. “Robot audition” is a research field to investigate this new generation of intellectual appropriately because of its nature of high dimensionality and online interactive characteristics. In this paper, we introduce the ideas from a brain-like information processing, especially a spiking neural processing on the basis of evidences in auditory information processing in a brain.

1. はじめに

ロボットハードウェア及びロボット制御技術の進展により、ヒューマノイドのような複雑な行動を伴ったロバストな移動が可能になり、今後のロボットにはより知的な機能が期待されている[30]。

従来行われていたロボット知能研究は、限定された入出力情報を前提にしたものであり、それらの情報関係を記号化することで知的な機能を実現していた。限定したとしても入力情報の量は十分大量であり、適当な知識量のもとではかなり知的な行動を実現できた[19]。しかし、用いる知識の増加や外界からの入力情報の不完全性を扱う次元になると、必要とされる計算時間が指数関数的に増大するという問題が生じた。Brooks[2]はあえて複雑な知識構造を用いず、駆動される行動を基準にした感覚－運動知識を構成する行動ベースアーキテクチャによりこの問題を解決したが、複雑な構造の扱いへの道筋を示すことはできなかった。Jordan[9]のグラフィカルモデルは、不確かで複雑な構造を持つ事象を取り扱うことを可能にし、ロボット応用も期待されるが、オンライン性・実時間性などに課題をもつ。

実環境と常に対峙するロボットに強く求められている知能は、多次元入出力・オンライン・実時間処理を特徴としたものである。ロボットの聴覚処理は、これら特徴を最も強く求められる機能の代表であろう。また、ロボットが家庭に入ってくるようになり，ロボットと人とのコミュニケーションや音による環境知覚は機能的にも重要になってきている[29]。そのように考えると、今後発展するロボットハードウェアとその多次元性・オンライン性・実時間性・システム性を考慮した全く新しい学問領域としての「**ロボット聴覚**」研究の展開が求められているのであろう。それは、従来、通常計算機の音声信号処理を載せ変えただけで行っていたような聴覚処理とは異なるものになるのではなかろうか。

脳型情報処理は生物で行われている中枢情報処理をヒントにした情報処理技術の構築を目指した領域である。アプローチとしては新しいものではなく、むしろ常に必要とされる研究領域であると考えている。たとえば、人工知能という名前のオリジナルとして著名な 1955 年の Dartmouth 会議[17]においても議論され、結果として神経回路などの技術が生まれた。脳型情報処理は各時代で得られる脳科学の知見を用いているため、常に新しさはもつ。しかし、それは一方で理論的積み重ねの困難さという弱点をはらみ、技術的価値をもった工学成果は神経回路以来乏しい。そのような中で、近年脳の構造的特徴をとらえた計算理論が提案され[26][12]、積み重ねが可能な理論構築の枠組みが構成されつつある。

我々は生物の情報処理として身体性・多次元性・オンライン性・システム性などの特徴に着目し、脳における自己組織の原理仮説[30]、脳システムや統合の計算モデル[22]を提案してきた。本論文ではこれらの仮説やモデルに関しては特に述べないが、脳科学の知見を簡単にレビューした後、そのベースとなるスパイク型神経情報処理の観点を中心に脳の聴覚処理を考察する。

2. 脳における聴覚処理

2.1. 概観

ロボットの聴覚機能を考えた場合、言語は重要な機能であり、解明すべき課題であるが、生物研究に

おける方法論に関しては制約が多い。最大の制約は言語を扱える動物が「ヒト」しか見当たらないという点である。

しかし、近年は fMRI などの非侵襲の脳計測手法が発達し、ヒトの脳の活動がある程度計測できるようになった。そのような画像化手法と脳損傷の観測をあわせることで、新たな知見が加速度的に増えている。言語処理に関わる脳領域に関しても、その領域は Broca 野,Wernicke 野だけでなく、かなり広範な部分が重要な機能を果たしていることがわかりつつある[20]。

つまり、脳における言語情報処理を解明するためには、どこかが活動しているというような場所の情報解析より各脳部位活動との時空間的関係や変化の情報の解析が必要である。そのような観点で、聴覚情報を軸に、脳を眺めてみる。Fig.1 に感覚信号・運動信号と脳の主要部位との関係を図示した。

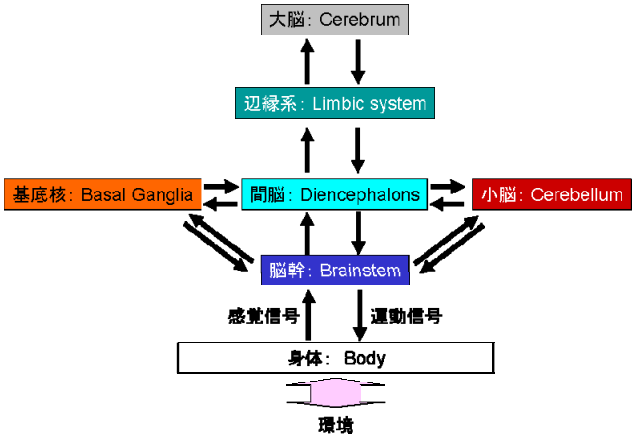


Figure 1. Hierarchical structure of a brain.

身体から得られる感覚情報は脳幹を通し、基底核や間脳に至り、さらに辺縁系、そして最後に大脳新皮質へ至る。信号を伝達するのには時間がかかるため、このようなシステムでは、実時間の処理や実行は下位で行われ、実時間性の少ないものが上位に配置されると考えられる。

聴覚処理では実時間性は必須である。言語対話などの能力を外した動物聴覚においては、実時間性は敵や餌を見つける上で極めて重要である。言語対話でも、対話の概のシナリオをもって話すこともあるが、大抵の対話は反射的だったりその場的だったりする。瞬時に入力情報を認識し発話することが大切であるが、一方で意味は後からわかったりする。聴覚処理では Fig. 7 の下位部の重要性が高い。

ヒトの脳は進化により発達したものであり、その構造は進化的に古いものの上に新しい脳を重ねるようにしてできている。Fig. 1 の場合、下位部にあたる脳幹、基底核、小脳、間脳は魚類から引き継いだ脳構造であり、両生類、爬虫類で基底核がさらに発達すると共に原始的な辺縁系ができ、哺乳類は大脳皮質をもつに至った。ヒトの対話は、動物聴覚にヒトのみがもつのであろう広範囲の皮質内連合情報処理が加わったことにより実現されている。

このような古い脳と新しい脳の階層化は、聴覚以外にも共通したものであるが、脳における聴覚処理が他の感覚情報と異なる点は、古い聴覚情報処理経路が種の進化を経ても保持されている点にある。対極は視覚である。視覚の古い情報経路である視蓋は残存してはいるが、主要経路とはなっていない。視覚情報は進化の過程で、脳幹を通らずに間脳にある視床経由で新皮質に入る経路ができ、新と旧の並列経路を用いるようになっている。聴覚の場合、哺乳類が言語を話すようになって間もないので、古い脳の動物聴覚を保持しているとも考えられるが、それだけではなさそうである。

その一つが、古い脳での情報処理の複雑性にある。聴覚情報の場合、蝸牛からの聴神経は蝸牛核、上オリーブ核、外側毛帯核、下丘を経て視床に向かう。一方、視覚情報の古い経路は、網膜神経節細胞からの信号が中脳視蓋で処理され、すぐに視床に向かう。ところが、この中脳視蓋の処理だけで、両生類などの動物は敵と餌を見分けたり、障害物を検知したりするのである[8]。さらに複雑な回路をもつ古い脳での聴覚処理は、想像を超えた処理を行っている可能性が高く、そのため、もはや捨てることができない回路となっているのであろう。

二つめは、新しい脳である新皮質との協調処理である。前述の複雑性も関係するが、古い脳の聴覚処理においては、蝸牛核、上オリーブ核、外側毛帯核、下丘といった核群間でのインタラクションが頻繁になされる。それらを通して、音の定位、マスク、分離、分類などがなされているが、その相互作用処理が新皮質を輪の中に入れることを容易にすると共に、輪の中に入った新皮質との関係をより強固なものにしていると考えられる。

三つめが、身体に最も近いが所以の身体性である。身体と感覚情報や運動情報をやりとりする最前線が脳幹である。そこでは、感覚情報以外にも、身体の状況を表すドーパミン、セロトニン、アセチルコリン、ノルアドレナリンなどを放出する細胞が群居している。中脳終端では、視覚、聴覚、運動といった情報が上丘（視蓋）で結びついている。基底核では、感覚情報を運動情報と結びつけ無意識に運動できるしくみがあり、前述のように言語処理に強く関わっている可能性が示されている。

次節では、この古い脳の聴覚処理を中心にいくつかの知見を紹介する。

2.2. 古い脳での聴覚処理

聴覚に関する研究は古くはピタゴラスによる「音が空気の振動である」という洞察に始まるだろうか。その後、16 世紀に Vesalius, Fallopio, 18 世紀に Corti らにより耳の解剖学的構造が明らかにされ、19

世紀初めに Wever,Bray らにより聴神経が音の周波数に同期して放電するという電気生理的知見が得られるに至った。この聴神経に関しての研究は 1960-70 年ころより本格化し、1980 年代には非常に多くの知見を得られた。1990 年代は聴神経の投射先である蝸牛核に関する知見が増え、近年では、下丘や皮質に関する知見が出てきている。

2.2.1. 脳幹における聴覚処理

蝸牛核に至る聴神経までの処理の一般的理解は以下である。

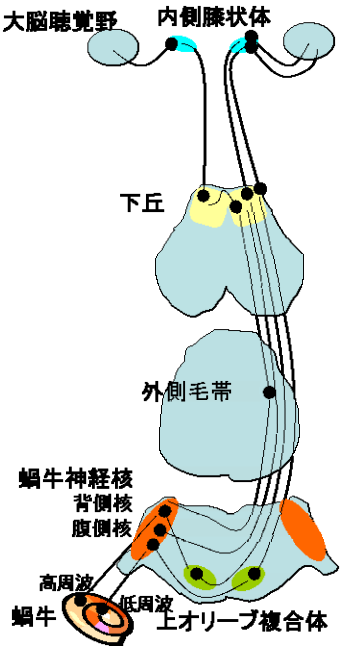


Figure 2. Pathway of the ascending auditory system.

Fig.2 の下部のように、まず、蝸牛にある基底膜に鼓膜からの振動が伝わると、この膜が振動。高い音は基底膜のうち蝸牛の入り口に近い方を、低い音は奥の方を大きく振動させる。この結果、どの場所がどのくらい振動するかによって、音の周波数分析ができる。基底膜には硬い毛の生えた有毛細胞という特殊な神経細胞が、びっしりと並んでおり、有毛細胞は基底膜のその場所が振動すると神経パルスの列を発生し、有毛細胞につながっている聴神経に伝える。神経パルスの大きさはほぼ一定(Fig.4)だが、頻度は基底膜の振動が大きいほど多くなり、周波数成分の強さが神経の信号に変換される。さらに、4～5kHz 以下の音の場合、神経パルスは基底膜が一回振動する間の特定のタイミングで発生されるので、周波数成分の位相も聴神経に伝えられる。周波数成分の強さと位相がわかるので、スペクトル解析を行っているといわれている。

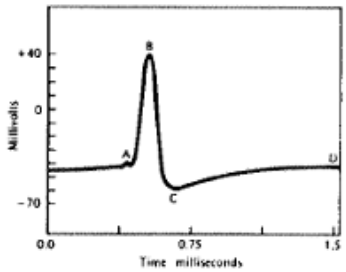


Figure 4. Characteristic of auditory nerve.

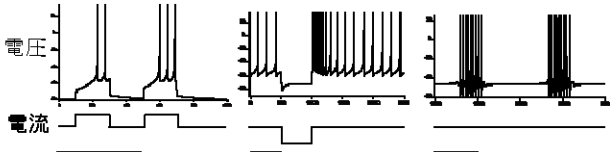


Figure 5. Firing behaviors of neurons.

実際の神経細胞の発火形式は複雑で、環境、入力、細胞の種類などで Fig.5 のように様々な形をとる。

聴神経の投射を受ける蝸牛核でも複雑な発火パターンが観測されている。このような発火パターンは樹状突起形状、シナプス特性、細胞体膜特性などを起因とする、膜電位の昇降や休止電位変化の微分係数などの変化により生じるものである。従来は情報処理としては無視されていたこのような神経細胞の発火形式だが、これらも情報表現の形式と考えると神経情報処理の潜在能力が格段に向上する。

蝸牛核では上記の複雑発火とともに音周波数マップが形成され、上オリーブ複合体へ投射され IID（両耳間強度差）, ITD（両耳間時間差）が求められるとされている(Fig.2 下部参照)。

聴神経の活動様式を知った上で、脳幹における聴覚処理の発達をみることも参考になる。

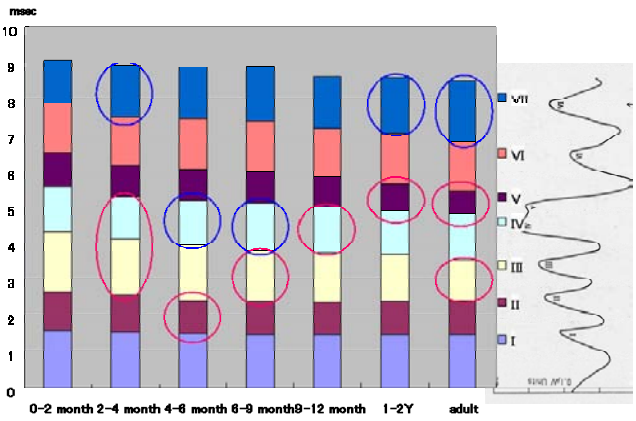


Figure 6. Development of ABR.

人間の場合、脳幹における聴覚のハードウェア構造（細胞と細胞間のおよその結合）は胎生時にほとんどできあがっている。生後は軸策のミエリン化と結合の詳細化がなされていると考えてよい[1]。

Fig.6 は幼児の ABR (Auditory Brainstem Response) の変化を図表化したものである。ABR は音刺激に対する脳幹の反応で、クリック音刺激を与え頭皮上から微弱な電位変化を計測し、通常は 10ms 以内に陽性波が出現する(Fig.6 の右にあるような時系列の反応)。Ⅰは聴神経、Ⅱは蝸牛核、Ⅲは上オリーブ核、Ⅳは外側毛帯、Ⅴは下丘、Ⅵは内側膝状体、Ⅶは聴覚野の反応であるとされている。図の縦軸は反応時間を表し、横軸は月齢および年齢を表す。たとえば、一番左の棒グラフは 0－2 ヶ月の幼児であり、新皮質の反応まで 9msec 強を用している。赤丸は前月齢より反応時間が短縮した部分を示し、青丸は反対に反応時間が延びた部分を示している。つまり、2－4 ヶ月では 0－2 ヶ月に比べ、上オリーブ核、外側毛帯の反応時間が短縮し、皮質の反応時間が延びている。

幼児期は聴覚経路の身体化と詳細化がなされるので、処理が短縮していることはそのような回路が最適化されたこと、処理が延びたときは追加の回路が形成されたことに相当すると考えられる。

すると、幼児期の聴覚は 2－4 ヶ月までに第一期の発達が脳幹で生じ、その発達に伴い新皮質の回路が形成され始めると推定できる。実際、たとえば音源の定位は生後直後にある程度できるのであるが、その後 1，2 ヶ月はうまくできなくなり、3，4 ヶ月で再びできるようになる[3]。Ⅲの上オリーブは特に定位に関連する部分なので、この間に身体化がなされているのだろう。遅れて 4－6 ヶ月から新しい質の聴覚情報の回路が蝸牛核で形成され、それは月齢と共に上オリーブ、外側毛帯、下丘と伝播していく。回路形成が下丘にいたる 1－2 才のころから、新皮質で第 2 次の回路形成が進む。実は 2－4 ヶ月のころの幼児は４ＫHz 未満の低周波音や広い周波数スペクトルをもった音に強く反応する[25]。このことは、４ＫHz 未満の音に対し位相に合った発火を行う聴神経が回路形成を誘導しているとも考えられる。聴神経の中でも自発発火率の低い L 型といわれるものは少数だがダイナミックレンジが大きくフォルマント等の大局構造の表現が可能なため、この L 型の発火活動が回路形成をこの誘導を行っている可能性が高い。このような機能は、生後母親の音声を優位に検知するのに役立つものとして、ある程度組み込まれているのかもしれない。この 2－4 ヶ月ころに、声を発することができる程度に発声器官が発達する。この時期以降の第 2 次発達は、自分の声への対応、視覚などとの情報融合、高周波表現に向けた発達が行われているのだろう。

2.2.2. 基底核や小脳における聴覚処理

Fig.2 からもわかるように、脳幹により聴覚情報の前処理が行われ、大脳により言語の認識や発声の高次処理が行われるとされている。大脳でも特にブローカ野とウェルニケ野がそれぞれ運動性、感覚性の言語処理中枢とされている。しかし、近年計測手法の発展に伴い、言語や概念の扱いはこの 2 つの領野だけでなく、脳の多くの部分が内容に応じて分担し関与していることがわかりつつある[20]。Damasio[4]は 169 人の脳損傷患者と 55 人のコントロールに対して、名詞、カテゴリ化を中心に脳画像を用いた網羅的解析を行い大脳各部位と言語処理の関係を示した。 Watkins[24]は遺伝的言語障害をもつ KE 家系では基底核の尾状核が小さいことを計測し、言語処理と基底核の強い関与を示した。小脳の損傷が構音生成障害を起こすことも知られている。ブローカ野が損傷しても基底核に損傷がなければ言語機能は回復するが、逆に基底核に損傷があるとブローカ野が無事でも言語機能は回復しないという知見もある[13]。

Ullman[23]は神経科学的認知心理の観点から、基底核を手続き的な文法知識の場、小脳を獲得された手続き情報の修正の場と考え、新皮質の各領野との機能的関係を認知モデル化し、Dominey[6]は言語のような継時的情報の処理モデルとして、新皮質と基底核からなる計算モデルを提案し、幼児からの言語習得過程を説明した。しかし、脳幹の処理に比べ、このような機能的モデルの研究は少ない。

3. 脳型情報処理と聴覚

脳を参考に新しい情報処理方式を開拓しようとするアプローチは、脳において観測される現象に注目し、その現象を人工的に再現するモデル化を行い、次に情報科学的洞察を行い、情報科学的意味をもった情報処理方式を提案するものである。従って、具体化しようと着目する現象により、開発される方式は様々である。神経細胞といった小規模な素子による並列分散活動に着目したものが、パーセプトロン、連想記憶、バックプロパゲーションなどの神経回路である。その場合、その他の要素である脳の構造や神経細胞の挙動などは特に関知しない。

並列分散活動に加えて、我々は解剖学的関係、解剖学的構造、神経細胞の特性（放電パターン、可塑性）に着目している。前節で述べたように、脳の各部位はある程度の機能的意味づけがされているので、それらの解剖学的関係から、システム構造が推定できる。また、解剖学的構造からは、神経が結合して構成される回路構造の青写真が描ける。そして、そこに神経細胞の特性を入れ込むことで、回路の中で行われている情報処理への仮説が形成できる。

すでに我々はこれら着目点から、脳における散逸的自己組織化[30]、双方向の仮説制御システム[12]，スパイキングニューロンによる時間符号化[11]，変動を組み込んだスパイキングニューロン学習モデル[22]などを提案してきた。本稿では、スパイキングニューロンによる情報処理モデルの考えをもとに聴覚処理を考察する。

3.1. スパイキングニューロン

神経細胞の多くは Fig.4 のように細胞の電位が一時的にプラスからマイナスになる活動電位（スパイク）という現象を示す。この現象を含め、神経細胞レベルの物理的モデル化はかなり厳密に行われている[10]。スパイキングニューロンのモデルはその中でも、形式神経細胞[18]と異なり、スパイクの現象に着目した神経情報処理モデルである。この情報処理の特徴は動的特性にあり、大きくは確率的モデルである Spike Response Model と決定論的モデルである Integrate and Fire Model に分かれる。確率的には後者は前者の特殊な例とも解釈でき、目的により使い分けが必要である。数学的記述や生物学的意味に関しては[5] [16]などの書籍を参照されたい。

このようなモデルにより再現可能なスパイクイベントは、Fig.5 のような時間イベントである。このようなイベント系列により情報を表現することにより、時間的に動的な情報処理が可能になる。たとえば、２つの細胞のスパイク間の時間関係は２つの情報の**位相関係とその動的変化**を表し、１つの細胞におけるスパイクの時間間隔は情報の**周波数とその動的変化**を表し、スパイクや膜電位の時間パターンは情報の**強弱や動的特性**を表現できる。さらに、スパイクは情報を伝達するだけでなく、制御信号としても用いることができるため、時間軸を用いて情報を統合したり分離したりすることもできる。また、電位に変化を与えることで、ニューロンの処理周期を変えたり、他のニューロンと周期を合わせたりする**処理制御**を行うことができる。

生物はこれらの処理をサブミリ秒の時間解像度で行っており、神経細胞による処理は、柔軟性、適応性の基盤となっているのだろう。しかし、一方でこのような処理を現在のコンピュータで行おうとすると非常に効率が悪い。なぜなら、コンピュータの情報表現はビット列のみであり時間情報はない。しかも、神経細胞の処理は非同期超並列を前提にしているのに対し、コンピュータは同期逐次処理を前提にしているからである。

3.2. スパイキングニューロンを用いた聴覚処理

前述のようなスパイキングニューロンによる情報処理の特性は、聴覚のように情報が時間にのっており、しかも実時間性を必要とする対象に非常に適している。特にロボット聴覚においては、音源やロボットが移動するため、動的特性への要求度は高い。

ここでは、例題を用いて簡単なモデルを作成する。

例題は反響抑制である。ロボットが環境中で音源定位する場合、ロボットは移動するため、環境・位置により異なる反響にいかに対応するかが大きな問題となる。人間もあんまりひどい反響があると音声認識が困難になる。こうした反響音への対応は電話などでも重要な技術であるため、適応フィルタを中心に開発は進んでいるが、音源位置、環境変化が限定されたなかでの技術となっている。

行動実験による反響（時間差をもった音）に関する生物の聴覚特性からは以下のような知見が報告されている[14][15][27]。

まず、２つの音が１ミリ秒以内で来た場合は１つの音とされる。例えば、対面する２つのスピーカーから同じ音を１ミリ秒以内で聞かせると、仮想音源が定位される。1 ミリ秒を超え、20 ミリ秒程度ずれた場合は、１つの音に聞こえるが、先に聞こえた音の位置に定位する。20 ミリ秒を超え 30 ミリ秒くらいまでは２つの音に聞こえるが、定位位置は先に聞こえた音に引っ張られ、30 ミリ秒を超えるようになって、2 つの音を 2 つの位置に定位できる。

このことから生物の場合、20 ミリ秒以内の反響は先行する音を優先し、他の反響音は時間的に引き込まれるか抑制されていると考えられる。Pollak[21]は、下丘はその下位で処理された様々な特徴を統合する場と考え、外側毛帯と上オリーブ複合体が反響音抑制に関与していると提案している。外側上オリーブ核は同側の外側毛帯背側にグリシンによる抑制性の投射、対側の外側毛帯背側に興奮性投射、外側毛帯背側は対側に GABA による抑制性の投射があることから、下記のような機能図を描くことができる。

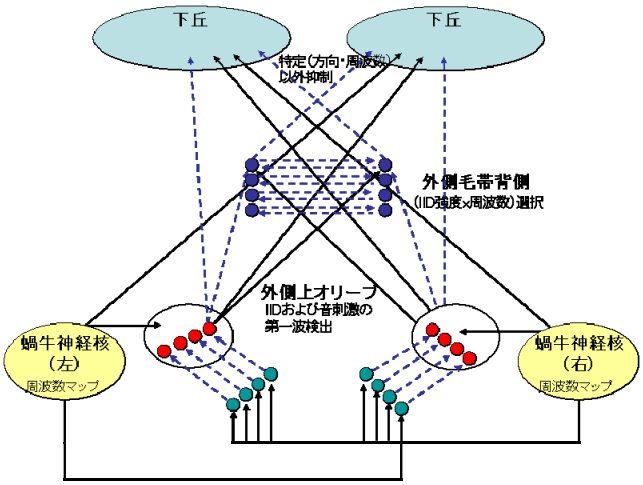


Figure 7. Functional scheme of brainstem auditory system.

この機能図では、相対的に「強度の強い音に関する IID 情報と周波数」が到来した場合、その音源方向が一時的(20 ミリ秒ほど)に優先され、それ以外を抑制（マスク）する。実線は興奮性、破線は抑制性の結合である。結合はすべてを記していない。

これだけを動作させるのであればアルゴリズムを書いてもよいのだが、スパイキングニューロンで構成することでより単純かつロバストになる。なぜなら、Fig.7 で記したようなインターフェースは特定の機能的側面を見ただけのものなので、少し問題が変わると対応できない。スパイキングニューロンのイ

ンターフェースはスパイクイベントと決まっており、その時間情報に位相、強度、周波数など必要な情報が準備されているので、受け側で選択し利用すればいい。また、個々の処理が単純なため、冗長ではあるが多数のニューロンを用意し関係付けることで、時間的にも空間的にも補完的な処理が可能になる。

このモデルは学習発達型でも利用可能である。発達型にするのであれば、2.2.1 などを参考に、低周波選択で位相発火型の聴神経をベースに内側上オリーブ－下丘間の結合を学習し、下丘に方向マップを概ね作成した後に外側オリーブや外側毛帯を導入し、上下で挟むことで学習を進めればいい。

また、このモデルにおいてマスクを形成する外側毛帯の活動は、一時の入力に対し 20 ミリ秒ほどの継続的発火を伴うものなのだが、そのような特性も細胞特性に組み込むだけでよい。IID などは、計算方式はアルゴリズムで記述するものと異なるが内容に大きな差はないが、同じ計算原理で Fig.7 のモデルをすべて動かせることが何よりも重要である。しかし、3.1 で述べたように原理は実時間向きでも、処理が実時間でないことは大きな課題であり、ロボット応用に向けてはさらなる技術蓄積が必要である。。

4. おわりに

脳型情報処理の立場から、ロボット聴覚を考察した。本稿ではスパイキングニューロンの処理に関してのみ述べたが、3 節冒頭で述べたように、脳型情報処理はシステム、回路、素子の 3 要素からなり、それらの相乗効果が大きい。三位一体の研究展開が重要である。ロボット聴覚は情報処理として求められる動的特性、学習能力の観点から、脳型情報処理の研究において最適な課題といえる。今後、両研究領域の相互発展を期待したい。

参考文献

1) Berg, BO., Principles of Child Neurology, McGraw-Hill, New York NY, 1995.
2) Brooks, RA. : A Robust Layered Control System for a Mobile Robot, IEEE Journal of Robotics and Automation 2 (1), 14–2, 1986.
3) Clifton RK., The development of spatial hearing in human infants, in Werner LA, Rubel EW (eds): Developmental Psycholoacoutics., American psychological Association, Washington, DC, 135-157, 1992.
4) Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., Damasio, A., Neural systems behind word and concept retrieval, Cognition, 92, 179-229, 2004.
5) Dayan, P., Abbott, LF., Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, MIT Press, Cambridge MA, 2001.
6) Dominey PF.,Hoen M.,Blanc JM.,Lelekov-Boissard T., Neurological basis of language and sequential cognition: evidence from simulation, aphasia, and ERP studies., Brain and Language, 86(2),207-25, 2003.
7) D'Esposito M, Alexander MP., Subcortical aphasia: distinct profiles following left putaminal hemorrhage. Neuorology, 45, 38–41, 1995.
8) Ewert, J.-P. and Arbib, M.A., Eds., Visuomotor Coordination: Amphibians, Comparisons, Models and Robots, New York: Plenum Press, 1989.
9) Jordan, MI. (Ed.), Learning in graphical models., MIT Press, Cambridge MA, 1999.
10) Koch, C. (ed.), Biophysics of Computation, Oxford University Press, New York, 1999.
11) Koerner, E., Gewaltig, M-O., Koerner, U., Richter, A. and Rodemann, T., A model of computation in neocortical architecture., Neural Networks, 12:989–1006, 1999.
12) Koerner, E., Tsujino, H. and Masutani, T.: A Cortical-type Modular Neural Network for Hypothetical Reasoning, Neural Networks 10, 791-814, 1997.
13) Lieberman, P., On the nature and evolution of the neural bases of human language, Yearbook of physical anthropology, 45, 36-62, 2002.
14) Litovsky RY, Colburn HS, Yost WA, Guzman SJ., The precedence effect, J Acoust Soc Am.,106(4 Pt 1),1633-54, 1999.
15) Litovsky RY, Shinn-Cunningham BG., Investigation of the relationship among three common measures of precedence: fusion, localization dominance, and discrimination suppression, J Acoust Soc Am., 109(1),346-58, 2001.
16) Maas W., Bishop, CM. (eds), Pulsed neural networks, MIT Press, Cambridge MA, 1998.
17) McCarthy, J., Minsky ML., Rochester, N., Shannon CE., A proposal for the Dartmouth summer research project on artificial intelligence", 1955.
18) McCulloch, W.S. and Pitts, W.H., A logical calculus of the ideas immanent in neural nets, Bulletin of Mathematical Biophysics, 5 : 115-133, 1943.
19) Nilsson, NJ. : Shakey The Robot, Technical Note 323. AI Center, SRI International, 1984.
20) Poeppel, D., Hickok, G., Towards a new functional anatomy of language, Cognition, 92(1-2), 1-12, 2004.
21) Pollak GD, Burger RM, Klug A., Dissecting the circuitry of the auditory system, Trends Neuroscience, 26(1),33-9, 2003.
22) Tsujino, H., Output-driven operation and memory-based architecture principles embedded in a real-world device, Journal of Integrative Neuroscience, 3(2), 133-42, 2004.
23) Ullman, MT., Contribution of memory circuits to language: the declarative/procedural model, Cognition, 92, 231-270, 2004.
24) Watkins KE, Vargha-Khadem F, Ashburner J, Passingham RE, Connelly A, Friston KJ, Frackowiak RS, Mishkin M, Gadian DG., MRI analysis of an inherited speech and language disorder: structural brain abnormalities., Brain, 125(Pt 3), 465-78, 2002.
25) Werner LA., Gillenwater JM., Pure-tone sensitivity of 2-to 5-week-old infants, Infant Behavior and Development, 13(355), 355-375, 1990.
26) Wolpert D, Kawato M: Multiple paired forward and inverse models for motor control. Neural Networks 11,1317-1329, 1998.
27) Yang X, Grantham DW., Echo suppression and discrimination suppression aspects of the precedence effect, Percept Psychophys, 59(7),1108-17, 1997.
28) 井上博充 ： 人間型ロボットが拓く未来社会と新産業の創成, 日本ロボット学会誌, 22 (1), 2-5 , 2004.
29) 奥乃博, 中臺一博, ロボット聴覚の課題と現状, 情報処理, 44(11), 104-113, Nov. 2003.
30) 松本元、辻野広司: 脳のこころ、「情と意の脳科学」、松本元・小野武年共編、培風館, 2002.

パーソナルロボット PaPeRo における近接話者方向推定と 2 マイク音声強調

Near-Field Sound-Source Localization and Adaptive Noise Cancellation in a Personal Robot, PaPeRo

○佐藤　幹 (NEC メディア情報研究所)
杉山　昭彦 (NEC メディア情報研究所)
大中　慎一 (NEC メディア情報研究所)
* Miki SATO(NEC.), Akihiko SUGIYAMA(NEC.), Shin'ichi Ohnaka(NEC.)

m-sato@dh.jp.nec.com, aks@ak.jp.nec.com, s-ohnaka@cp.jp.nec.com

Abstract—This paper presents implementation and evaluation of speech interface for a personal robot, PaPeRo, based on sound-source localization and noise cancellation. Sound-source localization incorporates a new formula taking near-field conditions into account for offsetting errors caused by the relative altitude of the speech source to the microphones. In noise cancellation, a novel stepsize control assuming a wide range of signal-to-noise ratios of the input signal helps achieve both small residual noise and distortion in the noise-cancelled signal. Evaluation results with recorded signals in the real environment demonstrates 40% higher source-localization performance and as much as 65% higher speech recognition rates in noisy environment.

1. はじめに

近年、人間と共生することを目的としたパートナー型ロボットの研究が盛んに行われている [1]。これらのロボットは、通常、音声コマンドによって、離れた位置から制御される。背景雑音や妨害信号の影響を低減して、正確に音声コマンドを認識するために、指向性マイクロホンが広く使われている。このため、音声の到来する方向を推定し、推定方向にマイクロホンの指向性を一致させることが重要となる。

遠隔会議などの通信応用と異なり、人間とロボットの対話では、話者の口、すなわち音源とマイクロホンは、同一平面上にあると見なすことはできない。しかし、ロボットにおける話者方向推定では、暗黙のうちに音源とマイクロホンが同一平面上にあると仮定してきた。この仮定が話者方向推定結果に与える影響は、人間とロボットとの距離が近くなるほど大きくなる。すなわち、近接音場を想定した方向推定が重要となるのである。

一方、マイクロホンの指向性だけで抑圧できない雑音や妨害信号は、音声強調処理によって、その影響を軽減する。応用毎に異なる要求条件に応じて、1 つ又は多数のマイクロホンを用いた雑音及び妨害信号の抑圧が、広く行われている [2]。人間とロボットの対話においては、2 つのマイクロホンを用いた適応ノイズキャンセラが、マイクロホン数、雑音除去性能、及び歪の観点から見て、良い妥協策である。

適応ノイズキャンセラは、音声用と雑音用の 2 つのマイクロホンを用いて、雑音の消去を行う。符号化や音声認識の前処理に用いるために、係数更新ステップサイズを音声対雑音比（SN比）に応じて制御することで、高い雑音消去性能と小さな音声歪を両立することができるノイズキャンセラ［3］が提案されている。このノイズキャンセラは、ヘッドセットなどのように、音声用マイクロホンが話者の口元にあることを想定しているため、様々な距離から話しかけられるロボットに適用することはできない。音声用マイクロホンと口との距離に応じて、SN比が広範囲に変化するためである。

本稿では、音声対話機能をもつ自律移動型パーソナルロボット PaPeRo[4]における、近接音場を想定した話者方向推定と、広範囲な SN 比に対応できるノイズキャンセラについて紹介する。2 節で、PaPeRo の構成と音声インタフェースについて説明する。3 節では近接話者方向推定、4 節ではノイズキャンセラをとりあげる。5 節では評価結果を用いて性能を明らかにし、6 節で今後の課題について述べる。

Figure. 1: PaPeRo の外観

2. パーソナルロボット PaPeRo

2.1. ハードウェア

パーソナルロボットPaPeRoの外観を、Fig. 1に示す。PaPeRoは、高さ385mm、幅248mm、奥行245mm、重量5.0kgの自律移動型ロボットである。胴体正面に４個、左右にそれぞれ１個、背面に１個の無指向性

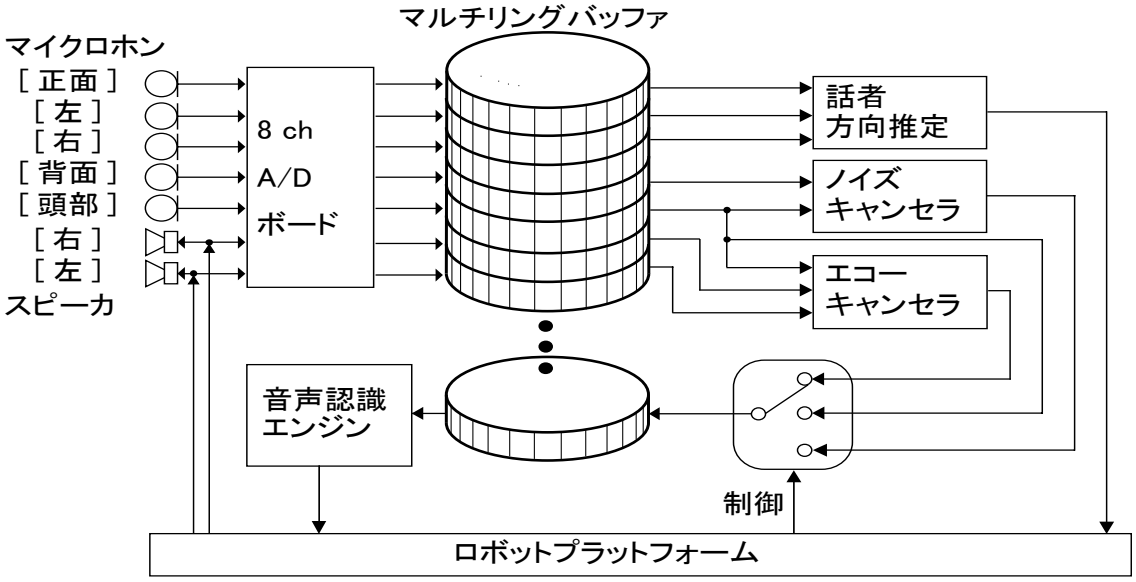


Figure. 2: PaPeRo の音声インターフェース

マイクロホン、及び頭部に１個の指向性マイクロホン、さらにCCDカメラ、超音波センサ、赤外線センサ、プッシュスイッチを搭載し、４個のモータによって動作する。これらのセンサ及びモータは、Windows XPで動作するPentiumM 1.6GHz CPUによって制御する。

2.2. 音声インターフェース

PaPeRoの音声インターフェースは、Fig. 2に示す構成を有する。前記マイクロホンに入力された信号を8チャンネルAD変換ボード（東京エレクトロンデバイス、TD-BD-8CSUSB）を用いてディジタル化し、マルチリングバッファに格納する。格納された信号から必要な信号を選択して、所望の処理を行う。話者方向推定では、正面、左、右の3マイクロホンに入力された信号を使用する。推定された方向の情報は、ロボットの方向を制御するロボットプラットフォームに供給される。ノイズキャンセラでは、頭部と背面の2つのマイクロホンに入力された信号を使用する。雑音が低減された出力は、出力用リングバッファに格納される。ロボットプラットフォームは、必要に応じて出力用リングバッファから音声認識エンジンに信号を供給し、音声認識が行われる。

2.3. AD変換ボード

パーソナルロボットPaPeRoに内蔵可能な小型の8チャンネルAD変換ボード（TD-BD-8CSUSB）を、東京エレクトロンデバイスと共同で開発した。ボードの外観を、Fig. 3に示す。8チャンネルのマイクロホン入力信号を同時にサンプリングし、USB2.0またはUSB1.1インターフェースで、パーソナルコンピュータ（パソコン）等へ取り込むことができる。ボードは、ASIOドライバに準拠しており、USB経由で供給された電力による動作（バスパワー動作）が可能である。バスパワーとは別に、外部電源供給の端子も有している。サンプリング周波数は、48kHz、44.1kHz、22.05kHz、11.025kHz、又は8kHzのいずれかを選択することができる。マイクロホン入力アンプ及びAD変換利得調整機能を内蔵しており、60dB以上のSN比を達成している。

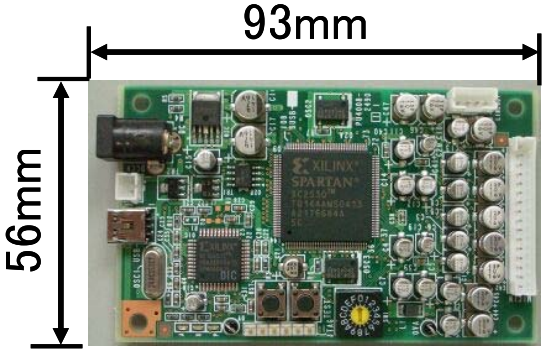


Figure. 3.　8 チャンネルAD変換ボード

3.　近接音場における話者方向推定［5］

3.1. 方向推定の原理

話者方向推定は、複数のマイクロホンに入力された信号の時間差に基づいて行う。Fig. 4に示すように、マイクロホンを結ぶ直線と直角の方向に対して、角度 θ をなす方向から音が到来する例を考える。2つのマイクロホンへの入力信号をx1(t)、x2(t)、これらの時間差を τ とする。時間差 τ は、入力信号x1(t) とx2(t)に関する相互相関の最大値として求めることができる。この τ を用いて、θ は(1)で与えられる。

$$\theta = \sin^{-1}\left(\frac{c\,\tau}{M}\right) \qquad (1)$$

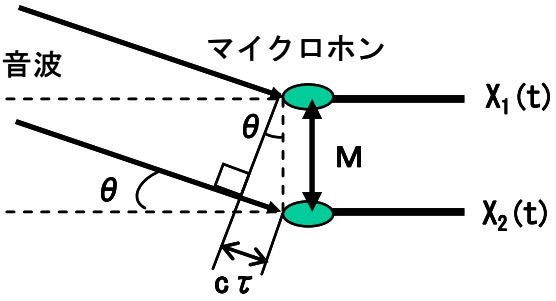


Figure. 4: 音の到来方向と受音信号

ただし、Mはマイクロホン間隔、cは音速である。

入力がサンプリング周波数fsの離散信号である場合は、τに対応するサンプル数の差k及びfsを用いて、θは(2)で表される。ただし、kは整数である。

$$\theta = \sin^{-1}\left(\frac{ck}{Mf_s}\right) \qquad (2)$$

3.2. 近接音場における方向推定

人間とロボットの対話では、話者の口とマイクロホンが同一水平面上に位置せず、距離も十分に遠くないため、(1)又は(2)を用いた方向推定を直接適用することができない。このため、近接音場特有の方向推定が必要となる。

Fig. 5に、人間とロボットの対話における、音源（話者の口）とマイクロホンの位置関係を示す。θとφはそれぞれ、音源の高さhが0に等しいとき及び0でないときに対応した、音声の到来方向を示す。一般に、φ＞θの関係が成立する。

2つのマイクロホンに入力された信号の時間差をτ_nとすると、音の到来方向φは、2つのマイクロホンの中心から音源までの水平距離dと、音源の高さhを用いて(3)で表される。

$$\phi = \sin^{-1}\left\{\frac{c\tau_n}{M}\cdot\frac{c\tau_n/2 + l}{\sqrt{d^2+h^2}}\right\} \qquad (3)$$

ただし、lは音源からマイクロホンまでの距離のうち小さい方の値で、

$$l^2 = h^2 + d(d - M\sin\theta) + M^2/4 \qquad (4)$$

で表される。人間とロボットの対話において、h、d、lは数メートルであるのに対し、$c\tau_n$、Mは高々数センチメートルである。この事実に基づいて、(4)は次式で近似することができる。

$$l^2 \approx h^2 + d^2 \qquad (5)$$

同様の近似及び(5)を(3)に適用すると、

$$\phi = \sin^{-1}\left\{\frac{c\tau_n}{M}\right\} \qquad (6)$$

を得る。音声の到来方向θは、(6)のφを用いて、(7)

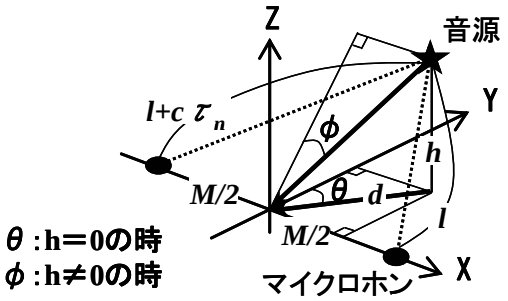


Figure. 5: 音源の高さと到来方向

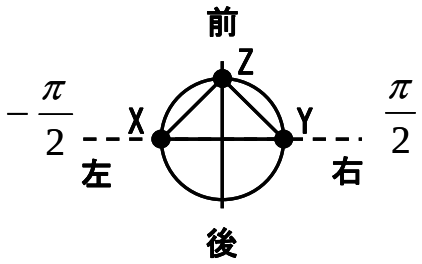


Figure. 6: マイク配置

で求めることができる。

$$\theta = \sin^{-1}\left(\frac{\sqrt{d^2+h^2}}{d}\sin\phi\right) \qquad (7)$$

3.3. 前後方向の決定

2 つのマイクロホンを用いた方向推定では、マイクロホンを結ぶ直線に関して対称な 2 つの方向（前後方向）を区別できない。そのため、PaPeRo では、Fig. 6 に示すように配置した 3 つのマイクロホンを使って、前後方向を決定する。図において、XZ=YZ=M_{short}、XY=M_{long}、∠ZXY=∠ZYX=γとする。また、ϕ_{PQ}はマイクロホン P 及び Q によって得られた推定方向を表し、$-\pi/2 \leqq \phi \leqq \pi/2$ とする。

前後方向決定は、方向推定精度に関する以下の 2 つの性質に基づいて行う。

(a) 方向推定分解能は、マイクロホン間隔が大きいほど向上する
(b) 方向推定精度は、φが±π/2 に近づくほど低下する

これらの特徴を考慮して、ϕ_{XY}を主推定方向、ϕ_{XZ}、ϕ_{YZ}を補助推定方向として用いる。前後方向の決定は、例えば$-\pi/2 < \phi_{XY} < 0$の場合、ϕ_{XY}とϕ_{XZ}の差と、$-\pi-\phi_{XY}$と$-\pi-\phi_{YZ}$の差を比較する。前者の方が小さい場合はϕ_{XY}を、後者の方が小さい場合は$-\pi-\phi_{XY}$を方向推定結果φとする。他の3つの場合についても、(8)に従って方向φを決定する。

$$\phi = \begin{cases} \phi_{XY} & \text{for } \phi_{XY} < 0 \text{ and } |\phi_{XY} - \phi_{XZ} + \gamma| < |-\phi_{XY} + \phi_{YZ} - \gamma| \\ & \text{or } \phi_{XY} > 0 \text{ and } |\phi_{XY} - \phi_{XZ} - \gamma| < |-\phi_{XY} + \phi_{YZ} + \gamma| \\ -\phi_{XY} - \pi & \text{for } \phi_{XY} < 0 \text{ and } |\phi_{XY} - \phi_{XZ} - \gamma| < |-\phi_{XY} + \phi_{YZ} + \gamma| \\ & \text{or } \phi_{XY} > 0 \text{ and } |\phi_{XY} - \phi_{XZ} + \gamma| < |-\phi_{XY} + \phi_{YZ} - \gamma| \end{cases} \quad (8)$$

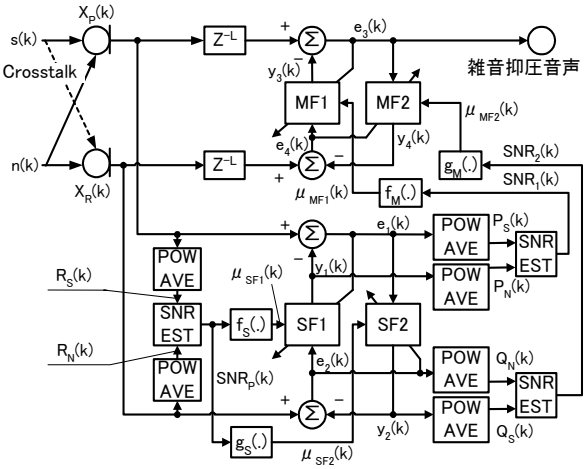


Figure. 7: ノイズキャンセラの構成

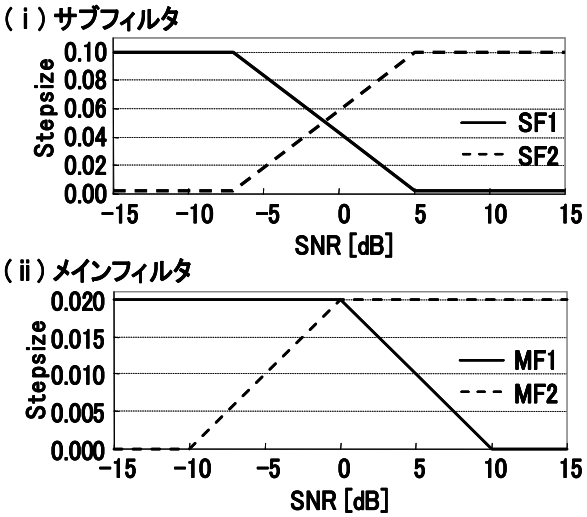


Figure. 8: 推定 SN 比と係数更新ステップサイズ

4. ノイズキャンセラ [6]

4.1. ノイズキャンセラの原理

ノイズキャンセラは、雑音用マイクロホンに入力された信号を適応フィルタで処理することによって擬似雑音を生成し、音声用マイクロホンに入力された信号から減算することによって、音声に混入する雑音を消去する。PaPeRo では、雑音用マイクロホンに混入する音声信号（クロストーク）の推定・消去も合わせて行うため、クロストーク推定用の適応フィルタも備えている。Fig. 7 に、PaPeRo におけるノイズキャンセラの構成を示す。

メインフィルタ(MF1、MF2)2 つ、サブフィルタ(SF1、SF2)2 つ、合計 4 つの適応フィルタによって構成され、それぞれ擬似雑音と擬似クロストークを生成する。雑音用マイクロホンの入力信号 $X_R(k)$から MF2 の出力する擬似クロストーク $y_4(k)$を減算したクロストーク消去信号 $e_4(k)$を、MF1 に入力する。MF1 の出力する擬似雑音 $y_3(k)$を、音声用マイクロホンの入力信号 $X_P(k)$から減算した雑音消去信号 $e_3(k)$を、強調音声として出力する。MF1、MF2 の係数更新ステップサイズは、それぞれ SF1、SF2 の出力を用いて制御する。SF1、SF2 の係数更新ステップサイズは、入力信号 $X_P(k)$と $X_R(k)$を用いて制御する。

4.2. ステップサイズ制御

適応フィルタの係数更新ステップサイズは、係数更新に対する妨害信号が大きいときに、小さな値とする。例えば、SF1 と MF1 のステップサイズ $\mu_{SF1}(k)$、$\mu_{MF1}(k)$は、入力信号 $X_P(k)$の SN 比が高いとき、すなわち、$X_P(k)$における音声成分が支配的であるときに、小さな値にする。一方、SN 比が低いときは、雑音に対する追従性を向上させるために、大きな値に設定する。SF2 と MF2 のステップサイズ $\mu_{SF2}(k)$、$\mu_{MF2}(k)$に関しては、音声と雑音が入れ替えるだけで、同様の制御を行う。

MF1 のステップサイズは、$X_P(k)$の SN 比推定値 SNR1(k)を用いて制御する。SNR1(k)は、SF1 の出力する擬似雑音 $y_1(k)$及び雑音消去信号 $e_1(k)$の平均電力 $P_S(k)$、$P_N(k)$の比として求める。MF2 のステップサイズに関しても同様に、擬似音声信号 $y_2(k)$とクロストーク消去信号 $e_2(k)$に基づいて求めた、$X_R(k)$の SN 比推定値 SNR2(k)を用いて制御する。SF1 と SF2 のステップサイズは、入力信号における SN 比の推定値 $SNR_P(k)$を用いて制御する。$SNR_P(k)$は、入力信号 $X_P(k)$と $X_R(k)$の平均電力 $R_S(k)$、$R_N(k)$の比として求める。メインフィルタとサブフィルタにおける推定 SN 比とステップサイズの関係を、Fig. 8 に示す。

5. 評価結果

5.1. 話者方向推定

PaPeRo を用いて、以下の 3 方式を比較評価した。

① 非近接音場方向推定＋多数決による前後決定
② 近接音場方向推定＋多数決による前後決定
③ 近接音場方向推定＋3 マイクロホンによる前後決定

評価環境をFig. 9に、パラメータ値を表1に示す。

人間とロボットの対話において、d、hは通常未知なので、表1に示す代表的な値に設定した。PaPeRo

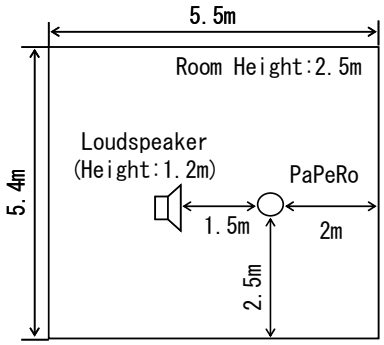


Figure. 9: 評価環境

Table 1: 評価パラメータ

M_{long}	21[cm]
M_{short}	14[cm]
fs	16[kHz]
c	340[m/s]
d	1.5[m]
h	1.0[m]

をπ/4ずつ回転させ、8方向から10回ずつ収録音声をスピーカで再生したときの方向推定正解率を求めた。ただし、人物検出可能なカメラ画角の制約によって、±π/9以内のずれまでを正解として許容した。得られた方向推定結果を、Fig. 10に示す。

①と②の結果を比較すると、近接音場方向推定によって、正解率を16%改善できたことがわかる。(8)に示した前後方向の選択手法を用いることにより、正解率をさらに23%改善することができた。近接方向推定と(8)を合わせて用いることにより、正解率は約40%改善したことになる。PaPeRoを用いた音源方向推定の正解率は、85%に達した。

5.2. ノイズキャンセラ

5.2.1. 雑音消去性能

サブフィルタにおけるステップサイズ制御の有（提案法）と無（従来法）に対するノイズキャンセラの雑音消去性能を比較した。男性音声の収録音声をスピーカで再生し、距離0.5mに配置したPaPeRoを使って収録したデータを用いて評価した。雑音は、距離1.0m、方向180度、音量57dBで再生したテレビの音を用いた。収録は、幅5.5[m]、奥行5.0[m]、高さ2.4[m]のカーペット敷きの部屋で、ロボットシナリオを動作させて行った。MF1、MF2のステップサイズ制御結果をFig. 11に、出力信号の雑音抑圧量をFig. 12に、音声歪を図8に示す。各グラフ上に発話の有無の状態を示す。雑音抑圧量 $R_3(k)$、音声歪 $D_3(k)$は、式(9)、(10)で求める。

$$R_3(k) = 10\log_{10}\left[\frac{\Sigma_{j=0}^{N-1} e_3^2(k-j)}{\Sigma_{j=0}^{N-1} X_P^2(k-j)}\right] \quad (9)$$

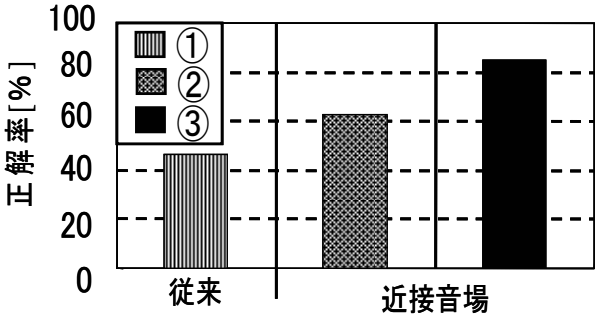


Figure. 10: PaPeRoによる方向推定正解率

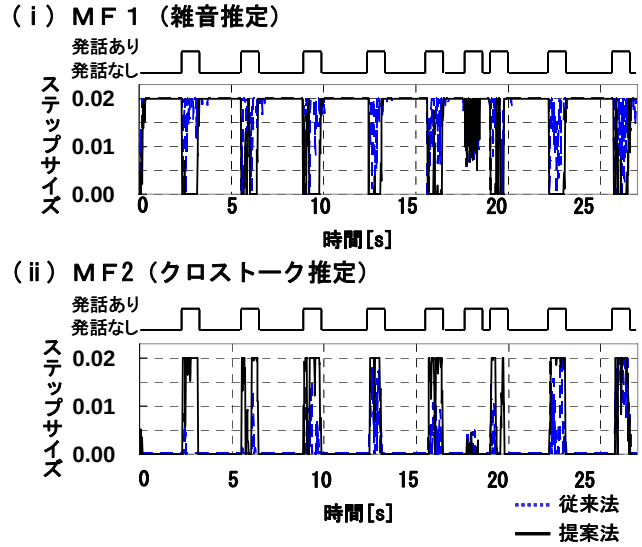


Figure. 11: ステップサイズ

$$D_3(k) = 10\log_{10}\left[\frac{\Sigma_{j=0}^{N-1}\{e_3(k-j) - S(k-j)\}^2}{\Sigma_{j=0}^{N-1} X_P^2(k-j)}\right] \quad (10)$$

Fig. 11を参照すると、提案法を用いた場合、MF1のステップサイズは、妨害信号となる音声信号がある区間で、小さな値となることが確認できる。一方、MF2のステップサイズは、発話区間で、大きな値となることが確認できる。Fig. 12、Fig. 13を参照すると、雑音抑圧量は、最大20dB、音声歪は、最大20dB改善した。

5.2.2. 音声認識性能

PaPeRoを用いて、Fig. 14に示す環境で、ノイズキャンセラ有と無に対する音声認識性能を比較した。男女子供30名による1500単語の収録音声をスピーカで再生し、正面方向、距離0.5m及び1.5mに配置したPaPeRoにおける認識率を評価した。距離1.0m、方向30、60、90、135、180度の5方向から、音量57dB、67dBで再生したテレビの音を雑音とした。音声認識には、PaPeRoの認識語600単語の辞書を有

する隠れマルコフモデルに基づく離散単語認識シス

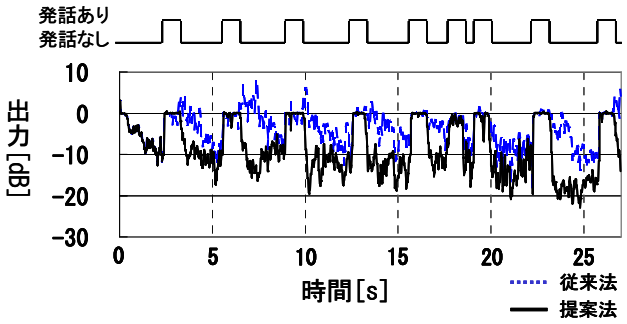


Figure. 12: 出力信号の雑音抑圧量

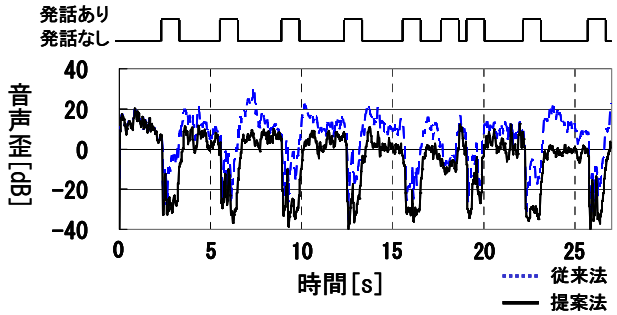


Figure. 13: 音声歪

テムを用いた。得られた認識率を、Fig. 15 に示す。棒グラフは、ノイズキャンセラによる認識率改善の最大値と最小値を示す。

例えば、話者距離 0.5m、雑音音量 57dB に対する結果を参照すると、雑音方向が 90 度より後方のときは、無雑音と同等の認識率を達成している。このとき、認識率の最大改善値は 65%に達した。他の 3 例においても、雑音方向の前方への回り込み、話者ーロボット間距離の増加、雑音音量増大のいずれかが存在すると、認識率が低下することが確認できる。

6. 今後の課題

今後は、複数話者が存在する環境で、各々の話者方向推定を行うことが課題となる。また、音声認識では、雑音環境下での遠距離発話認識率の向上が課題である。これら課題の解決には、本稿で紹介した手法の改良、他の音響信号処理技術の統合、さらに非音響センシングを統合した、より高精度な音声・雑音制御が必要となる。そのためには、フィールドテストを通じたデータの収集・評価、その分析を通じた問題点の明確化と対策が重要となる。

7. おわりに

PaPeRo における、近接話者方向推定と広範囲な SN 比に対応できるノイズキャンセラについて紹介した。実環境評価により、近接音場方向推定が従来よりも 40%高い正解率を達成することを示した。また、ノイズキャンセラの実環境評価結果を用いて、雑音抑圧量が最大 20dB、音声歪が 20dB、音声認識率が最大 65%改善し、57dB の雑音に対して無雑音と同等の

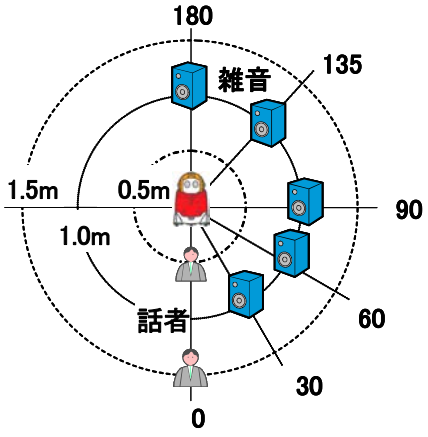


Figure. 14: 実験環境

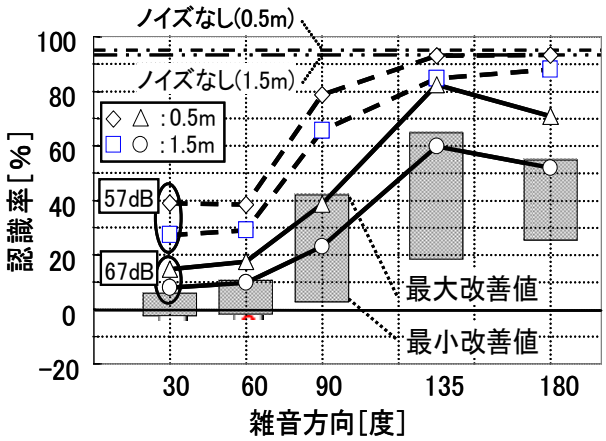


Figure. 15: 雑音消去による音声認識率の違い

認識率を達成できることを示した。本研究の一部はＮＥＤＯ実用システム化推進事業の助成を受けて行っており、この技術に基づいたチャイルドケアロボットは、2005年愛知で開催された愛・地球博において、技術実証運用を行った。

参考文献

1) Special Issue on Entertainment and Amusement Robot Tecnologies, J. of Robotics and Mechatronics, Vol.14, No.1, Feb. 2002.
2) M. Brandstein and D. Ward, “Microphone Arrays,” Springer Verlag, Berlin, 2001.
3) S. Ikeda and A. Sugiyama, “An Adaptive Noise Canceller with Low Signal-Distortion in the Presence of Crosstalk,” IEICE Trans. Fund, pp.1517-1525, Aug. 1999.
4) Y. Fujita,“Personal Robot PaPeRo,” J. of Robotics and Mechatronics, Vol.14, No.1, Jan.2002.
5) M. Sato, A. Sugiyama, O. Hoshuyama, N. Yamashita, and Y. Fujita, “Near-Field Sound-Source Localization Based on a Signed Binary Code,” IEICE Trans. Fund, pp.2078-2086, Vol.E88-A, No.8, Aug. 2005
6) M. Sato, A. Sugiyama and S. Ohnaka, “An Adaptive Noise Canceller with Low Signal -Distortion based on Variable Stepsize Subfilters for Human-Robot Communication,” IEICE Trans. Fund, pp.2055-2061, Vol.E88-A, No.8, Aug. 2005

人工知能学会
JSAI Technical Report
SIG-CHallege-0522-8 (10/14)

コミュニケーションロボット・DAGANE

DAGANE: A Communication Robot

原直 西野隆典 伊藤克亘 宮島千代美 武田一哉

Sunao Hara, Takanori Nishino, Katunobu Itou, Chiyomi Miyajima, and Kazuya Takeda.

名古屋大学大学院情報科学研究科

Graduate School of Information Science, Nagoya University

hara@sp.m.is.nagoya-u.ac.jp

Abstract

This paper presents a multi-user and multilingual communication robot DAGANE (Dialogue AGent Applied to Navigation Enhancement) that can naturally communicate with two or more people simultaneously through spoken dialogue and gesture. DAGANE mainly consists of three modules including a speech recognizer, a dialogue manager and a robot. The speech recognizer has language-dependent acoustic models and network grammars of three languages: Japanese, English, and Chinese. The dialogue manager has a language-independent state transition model to understand the context and semantic class of the user utterances. The design of the robot focuses on a friendly looking face and a lovely character so that people can communicate with the robot in a relaxed and affable mood. DAGANE was demonstrated at the Prototype Robot Exhibition sponsored by NEDO (the New Energy and Industrial Technology Development Organization) at the Aichi World Expo 2005 and performed a guidance of local tourist attractions and food of the three prefectures in Tokai area.

1 はじめに

音声認識や音声対話の研究が進み，様々な音声対話システムが構築されている．それらの音声対話システムは，情報提供をタスクとするものがほとんどである．さらに，情報提供の方法としては，ディスプレイモニタや音声を用いた情報提供に限られている．

情報を提供しようとした場合，情報の内容によっては，言葉だけでは理解しづらいことも多い．例えば，携帯電話のマニュアルを見てみよう.「フロントスタイルのとき，操作の取り消しはサイドクリアキーでおこないます」という記述がある．これを言葉だけで伝えようとしても，フロントスタイルやサイドクリアキーが何を指すのかわからない人が多いだろう．また，同じようなボタン/キーについても，製造会社ごとに名称が異なることが多い．しかし，携帯電話の実物を使って，対象となるボタン/キーを直接指して「このキーで操作の取り消しをおこないます」と身振りを交えて説明した方がずっとわかりやすいだろう．

このような身振りによる情報提示も併せたマルチモーダル対話システムとして，対話機能を持ったコミュニケーションロボット DAGANE (Dialogue AGent Applied to Navigation Enhancement) を構築した．このロボットは，愛・地球博の NEDO (新エネルギー・産業技術総合開発機構) によるプロトタイプロボット展 (2005 年 6 月 9 日から 19 日まで) に参加した．

2 システム概要

コミュニケーションロボット DAGANE は，東海三県の観光地，名物という限定されたタスクドメインの説明を行なうシステムである．万国博覧会という展示会場の性質を考慮して，複数言語 (日本語，中国語，英語) と複数話者 (3 名まで) に対応できるように設計した．

また，システム設計の面では，以下の点を基本方針とした．モジュラリティを高めること，各モジュールは堅牢性を高めるためになるべく既存の記述を用いること，対話管理ソフトウェアは移植性を高めること．

以下，システムの詳細について述べる．

2.1 タスク

タスクの内容は観光案内とする．案内する地域は東海三県 (愛知，岐阜，三重) とし，案内する内容は観光地と名

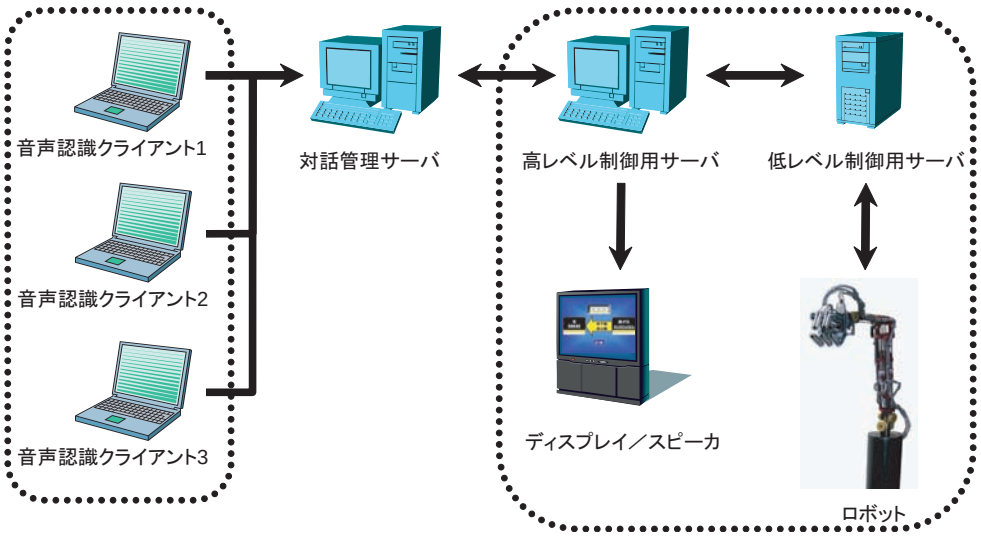


Figure 1: システム構成

物の説明とする．

対象となるユーザとしては，万博会場ということを考え次のように想定した．数はトータルで数千人程度，年齢層は，子供から老人まで，性別は男女であり，音声対話システムの初心者である．また，国籍は多岐にわたり，案内内容については知識がないことを前提とする．

想定ユーザにあわせて，対話戦略はシステム主導で行なうこととした．ただし，複数話者の扱いなどに関しては，なるべく自然になるような機能を持たせることとした．

システムの稼働環境は，背景雑音の音量が非常に大きい，大きな建物の中とする．

これらの項目を考慮して，マイクの選択や，案内内容の作成，対話戦略などを設計する必要がある．

2.2 ハードウェア

ハードウェアからシステム構成を見た場合，ロボット，音声認識クライアント，対話管理サーバの三つのモジュールから構成される (Fig. 1)．

ロボットは，7 自由度のアーム型ロボット (Fig. 2) と低レベル制御用サーバ，高レベル制御用サーバからなる [1, 2]．このモジュールは，動作に関するコマンドを対話管理サーバから受け付けると，ロボットの動作をさせることができる．低レベル制御用サーバは，ロボットのモータを直接制御するサーバであり，高レベル制御用サーバは，対話管理サーバとの通信，動作コマンドの発行，音声応答の出力，画像応答の表示を行なう．なお，音声応答は，ナレータの発声を録音したデータが発話ごとに再生される．

音声認識クライアントは，ノート PC であり，ユーザ一人ごとに一台を割当てる．内蔵 AD に接続したマイクロホンで音声を収録し，音声認識結果に応じたデータを対話管理サーバに送る．

対話管理サーバは，音声認識クライアントからの入力

Figure 2: アーム型ロボット

に応じて次の応答を決定し，その応答に対応するシナリオファイルにしたがって，ロボットモジュールに対しコマンドを発行する．ソフトウェアの詳細は 3 節で説明する．

2.3 対話シナリオ

対話の典型的なやり取りを次に示す．ただし，U はユーザ発話を表し，D は DAGANE の応答を表すものとする．

U：明治村にはどう行けばいいですか?
D：まずリニモに乗って地下鉄藤が丘駅に行きます
D：次に地下鉄東山線で，藤が丘駅から名古屋駅まで行きます
D：名鉄犬山線に乗り換えて，名古屋駅から犬山駅まで行きます
D：犬山駅から名鉄バスで明治村まで行きます
U：いくらかかりますか?

D: 片道 1580 円です

このような対話の流れを，複数ユーザに同時に対応できるようにするには，前の発話と同一の事柄について別のユーザが質問をした場合や，別の事柄に対して同時に別々のユーザが質問をした場合などで，どのように対話を管理すれば良いのかという問題が生じる．また，上記の対話の流れの場合,「いくらかかりますか」のように，省略を含む発話もユーザから自然に発話される．

これらの問題点の一部は，文脈を扱うことにより実現することができる．そこで対話管理システムに文脈を扱うような機能を実装した．この機能については 3 節で説明する．

2.4 キャラクタ

対話システムのキャラクタがユーザに与える印象は，対話そのものにも大きな影響を与えるため，非常に重要な要素である．

本システムでは，ロボットのアーム部分を，Fig. 3 のような着ぐるみで覆った．これにより，親しみやすさをアピールした．また，アーム部分が直接ユーザにぶつからないようにする安全面での効果もある．

応答音声は，日本語，中国語，英語をそれぞれ別の声質の似た女性ナレータにより収録したものを用いた．

Figure 3: DAGANE キャラクタ

3 ソフトウェア

3.1 対話管理

対話管理は，状態遷移モデルに基づいている．このモデルでは，現在の状態で文脈を表現する．

状態遷移モデルによる対話処理の例を Fig. 4 に示す．初期状態で,「明治村」という発話が入力されると，場所に関しては確定するが，明治村の何について聞きたいか (本システムでは「話題」と呼ぶ) は確定していない．一方，初期状態で「いくらかかるの」という発話が入力された場合は，話題は確定するが，場所は確定しない.「場所確定」状態から,「いくらかかるの」という発話が入力されると，場所も話題も確定した状態になる．場所も話題も確定した状態で,「どうやって行けばいいの」などの発話が入力された場合には，その状態に留まり続ける．

また,「ありがとう」などの挨拶は，どの状態においても，その状態に留まる発話 (状態遷移をしない発話) として定義されている．

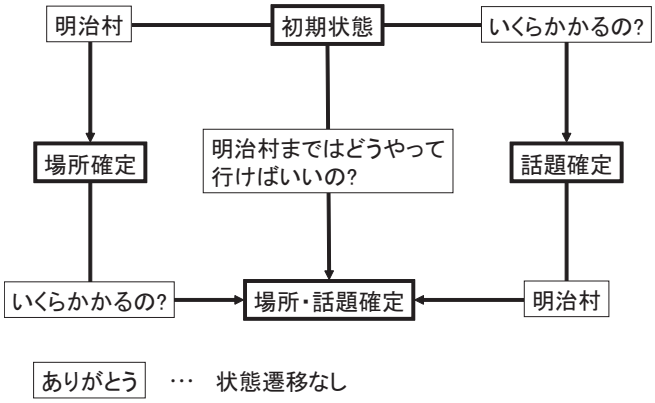


Figure 4: 状態遷移モデルによる対話処理

このような状態遷移モデルを利用すると，例えば，場所確定状態で,「いくらかかるの」という省略発話を扱うことができる．

ここで，この状態遷移モデルを用いて複数ユーザに対応するためには，ユーザごとに状態を保持する必要がある．この状態を保持する変数をユーザスロットと呼ぶ．ユーザスロットが保持する情報を Table 1 に示す．

しかし，このユーザスロットだけでは，自然な対話を行うことができない場合がある．例えば，あるユーザが「明治村への行き方」を尋ねた後で，二人同時に「値段はいくらか」を尋ねたとする．応答は一人ずつ行なうことになっているが，この場合，片方のユーザに説明したら，その後に，別のユーザに説明する必要はない．

このような対話管理を実現するために，グローバルスロットと優先話者という変数を用意した．

優先話者はロボットが対応している話者であり，グローバルスロットには，その話者が行なっている対話の状態が保持される．グローバルスロットと同じ事柄の対話の場合は，そのまま応対がなされる．グローバルスロットと異なる事柄の対話の場合は,「ちょっと待っててね」と発話して待たせておき，グローバルスロットの対話が一通り終了してから応対がなされる．この間，当該ユーザのユーザスロットは変化させずに保持する．

このシステムはある程度，案内を説明するのに時間がかかる．このようなシステムの場合，応答前に沢山の入力がたまってしまう場合がある．こういった場合は，長い応答

Table 1: ユーザスロット

スロット名	内容	例
場所	最近の会話の対象の場所	明治村
名物	最近の会話の対象の名物	天むす
話題	最近の会話の対象の話題	○○への行き方
シナリオ	最後に再生したシナリオファイル	
状態	現在の状態	場所確定

の後で，たまっていた入力に対する応答を行なっても何に対する応答かわからなくなることがある．こういった応答を防ぐため，応答前に一定数以上の入力がたまってしまった場合は，応答している入力以外は破棄することとした．

3.2 応答生成

応答の生成には，ある程度の単位で動作を指定するスクリプトを用意した．そのスクリプトファイルを次々に再生することで応答が行なわれる．シナリオファイルの例を次に示す．なお，D:はディスプレイへの表示，M:はロボット動作の指示，V:は音声出力を行なうコマンドであり，コマンドに続けてパラメータを記述する．

```
D:NAGOYAJO01.jpg        #画面表示
M:TurnToUser?.act       #user の方向を向く
V:NAGOYAJO.wav          #音声出力 (名古屋城への行き方の説明)
S:                      #同期
M:W20                   #待機 (20 秒)

R:                      #巻き戻しポイント
D:NAGOYAJO02.jpg        #画面表示
V:W10                   #待機 (10 秒)
V:GotoFujigaoka.wav     #音声出力 (藤が丘までの行き方)
M:act_disp_64.act       #画面を指しながら動く
M:H2                    #ホームポジションに戻る

S:                      #同期
M:W20                   #待機 (20 秒)
I:                      #割り込みポイント
```

このように，異なったモダリティの単位応答を記述することで，ある程度まとまった大きさの応答を記述している．このように，各モダリティは次々に駆動されるだけであり，非同期に再生される．しかし，各モダリティがバラバラでは，意味のある応答に見えない．そこで，ある程度の同期を可能にするためのコマンドを用意した．

```
D:NAGOYAJO02.jpg        #画面表示
V:W10                   #待機 (10 秒)
V:GotoFujigaoka.wav     #音声出力 (藤が丘までの行き方)
M:act_disp_64.act       #画面を指しながら動く
M:H2                    #ホームポジションに戻る
S:                      #同期
D:NAGOYAJO03.jpg        #画面表示
```

この例の場合，Fig. 5 に示すようなタイミングで，各モダリティの出力が再生される．このように S のコマンドで，同期を実現しており，D，V，M のうち一番最後の動作が終了するまで，次の D:NAGOYAJO03.jpg を実行することを待つ．

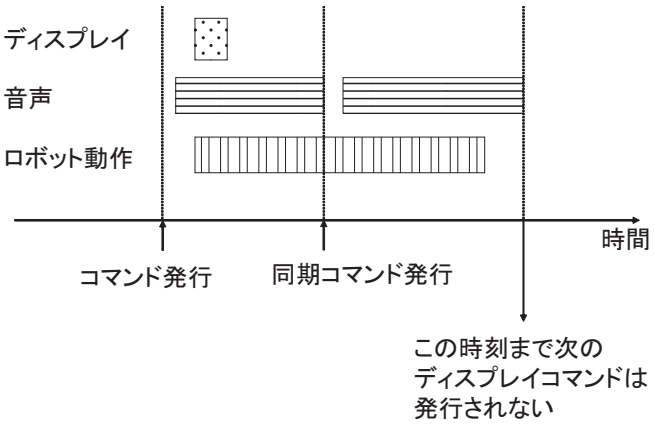


Figure 5: 同期の例

割り込みがなされた場合には，途中で応答を中止する．しかし，どこまで進んだかを記録する必要があるので，割り込み処理を発行するタイミングをシナリオファイルに明示的に記述することにしている．

```
V:GotoFujigaoka.wav     #音声出力 (藤が丘までの行き方)
M:act_disp_64.act       #画面を指しながら動く
M:H2                    #ホームポジションに戻る

S:                      #同期
M:W20                   #待機 (20 秒)
I:                      #割り込みポイント
```

例えば，ロボットの動作 act_disp_64.act を実行中に，割り込み発話が入力された場合でも，その動作が終わ

り，ホームポジションに戻り 20 秒待つところまでは動作が行われる．また，優先話者以外は，応答の途中に割り込みできないようにしている．

また，シナリオファイルはかなり長くなる場合もあるため，言い直し場所を指定できるようにした．「もう一度言って」というような聞き返しの発話があった場合には，最も近い割り込みポイントで再生を止め，直近の言い直し場所から再度再生される．

各国語の音声は全て同じ名前が付いており，認識に用いられた言語と同じ言語の音声ファイルが再生されるようになっている．

3.3 音声認識

音響モデルは，日本語成人，日本語子供，中国語，英語男性，英語女性の 5 つを用意した．このモデルは，ユーザに合わせたものをあらかじめ設定しており，一旦設定したら，その後は切り替えない．

言語モデルは，文法により記述されており，日本語，中国語，英語が用意されている．また，意味 ID ごとに文法を作成してあるため，認識結果として意味 ID を得ることができる．そのため，認識結果を構文解析/意味解析する必要はない．語彙サイズは 200， 意味 ID の数は 254 であった．

また，グローバルスロットの値から判断して，意味的に許容されない ID の認識結果が得られた場合は，認識誤りであるとして無視する．これによって，ある程度の突発的な雑音に対する誤動作も防ぐことができた．

4 プロトタイプロボット展

愛・地球博の NEDO (新エネルギー・産業技術総合開発機構) によるプロトタイプロボット展 (2005 年 6 月 9 日から 19 日まで) に参加した．(Fig. 6)

Figure 6: プロトタイプ展参加風景

会場において，一般の来客を対象に体験デモを実施した．会期中のほとんどの時間帯は，かなり混雑した状態であったが，何を話しかければよいのかがわかっている話者に関しては，かなり良好な対話をおこなえた．

体験デモのユーザ発話 (2 日間分，496 発話) を収録しその SNR を分析した (分析手法は文献 [3] を参照)．SNR の分布を Fig. 7 に示す．平均は 23.7dB であり，雑音レベルは十分に低いことがわかる．

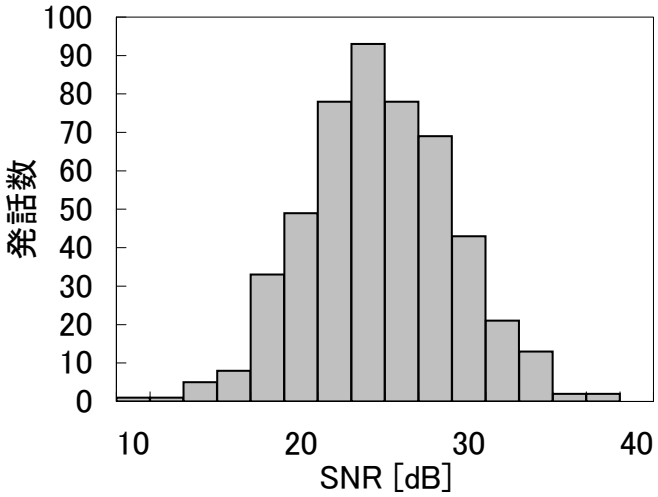


Figure 7: 万博会場ユーザ発話の SNR 分布

5 むすび

複数の人間との対話を同時に行なうことができる音声対話システムの開発，および対話システムのロボットへの組み込みを行なった．対話の文脈や発話の意味を考慮することで，対話の流れを管理することが可能となっただけではなく，雑音などの影響から生じる誤動作を防ぐことが可能となった．また，親しみやすいキャラクターとすることにより，ロボットとの対話の際にユーザが身構えることが少なくなり，ユーザに普段どおりの発話を促すことが可能となった．今後，認識語彙やシナリオを増加させることによる内容の自由度向上だけでなく，音声認識インタフェースの研究対象としてのロボットの活用が重要である．

謝辞

DAGANE は名古屋大学とビジネスデザイン研究所が共同で開発しました．ロボットの基本ソフトウェアは MIT Media Laboratory に提供していただきました．中国語音響モデルは Microsoft Research Asia のコーパスを利用して作成させていただきました．関係各位に感謝いたします．

参考文献

[1] Kai-yuh Hsiao, Nikolaos Mavridis and Deb Roy, “Coupling Perception and Simulation: Steps Towards Conversational Robotics,” Proc. IEEE/RSJ

International Conference on Intelligent Robots and Systems (2003).

[2] Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis, "Mental Imagery for a Conversational Robot," IEEE Transactions on Systems, Man, and Cybernetics, Part B, Volume 34 , Issue 3, pages 1374-1383 (2004).

[3] K. Takeda, T. H. Dat, H. Fujimura, and F. Itakura, "SNR and local noise power estimations based on Gaussian mixture modeling on the log-power domain," Proc. ICASSP'05, pp.I-881-884, 2005.

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

Clustering of sound-source signals using Hough transformation, and application to omni-directional acoustic sense for robots

○鈴木 薫, 古賀 敏之, 廣川 潤子, 小川 秀樹, 松日楽 信人
株式会社 東芝 研究開発センター ヒューマンセントリックラボラトリー
* Kaoru SUZUKI, Toshiyuki KOGA, Junko HIROKAWA,
Hideki OGAWA, and Nobuto MATSUHIRA
Humancentric Laboratory, Corporate Research & Development Center, Toshiba Corporation

kaoru3.suzuki@toshiba.co.jp

Abstract— In this paper, we proposed a new method of omni-directional acoustic sense with which a robot could localize and recognize multiple sounds from unlimited direction even under a noisy environment. We used Hough transformation to detect straight lines from the frequency phase difference space for detection and localization of sound sources. Experimental results with our robot, ApriAlpha were shown to verify the efficacy of this method.

1. はじめに

筆者らは、家庭内環境で利用者と音声でインタラクションすることを想定したロボット『ApriAlpha ™』を開発している。家庭内環境には様々な雑音が存在する。他方で、ロボットはこれら雑音にさらされながら命令権限のある利用者の音声をどの方向からでも受け付けて個々に認識し、サービスを提供できなければならない。また、ロボットはより高度に状況を認識するために、利用者の命令音声に限らず、室内のドアの開閉音、シャワーの断続音、ガラスの破壊音などの環境音を聞き分け、その発生源の位置を特定できると都合が良い。そのため、ロボット用の聴覚システムは、（1）四方八方から到来する音声を、（2）音源毎に分離抽出して、（3）個別に定位し認識できる必要がある。本稿では、このような全方位性を持つロボット用聴覚（Fig.1）の 1 方式について報告する。

Fig.1　A scene of omni-directional acoustic sense

上述したような聴覚機能の代表的な研究成果として、音源を制約せずに空間相関行列を解く方法[1]や、音源音声が調波構造を持つことを利用する方法[2]が報告されている。前者は音源数を超える数のマイクを利用する場合に優れた方式であり、後者は人間の音声を扱う場合に、より少ないマイクでマイク数以上の音源を扱える方式である。本稿で報告する方式は上記後者のアプローチに類似しているが、音源検出に調波構造を用いる代わりに、周波数と位相差の関係に着目し、音源の数と方向の推定を周波数－位相差空間における直線検出問題に帰着させてハフ変換により解く。検出された直線を複数のマイク対について対応付けて音源候補の空間定位を行い、適応アレイ処理によって音源音を分離して認識する。

以下、本稿では、開発中の聴覚処理方式の動作原理を説明するとともに、4 話者順次発話時と 2 話者同時発話時で全方位性を確認した実験について報告する。

2. 本方式の動作原理

2.1. 音源方向 ϕ・到達時間差 ΔT・周波数毎の位相差 ΔPh の関係

マイク 1 と 2 から成るマイク対を考える。音源がマイク間距離 d に比べて十分遠く、途中に障害物がないと仮定するなら、音源を発してマイク対に到達する波面はほぼ平面となっている。この平面波を観測すると、両マイクを結ぶベースラインに対する音源方向の角度に応じて、両マイクで観測される音響信号に所定の到達時間差 ΔT が観測される。到達時間差 ΔT は $\pm \Delta Tmax$ の範囲で変化し得る。$\Delta Tmax$ は、音速を Vs として、$\Delta Tmax = d/Vs$ として定められる到達時間差の理論上の最大値である。このとき、音源の方向 ϕ をマイク間ベースラインの中点を基点にベースライン垂直方向を 0 として式 1 を用いて計算する。なお、ΔT はマイク対を構成するマイクの一方に対する他方の到達時間差となるため符号付きの量であり、ϕ も符号付きとなる。

$$\phi = \sin^{-1}(\Delta T / \Delta Tmax) \quad \cdots \text{（式 1）}$$

マイク 1 とマイク 2 で到達時間差 ΔT を持つ音響信号を FFT によって周波数成分毎に分解してその位相差 ΔPh を眺めると、両者に比例関係が認められる。

例えば、同一時間差ΔT に対して、周波数 f の波は 1/2 周期、すなわちπだけの位相区間を含むとすると、2 倍の周波数 2f の波では 1 周期、すなわち 2πの位相区間を含む。このように、同一時間差ΔT に対する位相差ΔPh は周波数に比例して大きくなる。したがって、同一音源から発せられてΔT を共通にする各周波数成分の位相差を、Fig.2 に例示するように横軸を位相差、縦軸を周波数とする 2 次元座標系上にプロットすると、各周波数成分の位相差を表す座標点が 1 本の直線上に並ぶ。ΔT が大きいほど、すなわち両マイク間で音源までの距離が異なるほど、この直線の傾きは大きくなる。このとき、ΔPh とΔT の関係は式 2 のようになる。

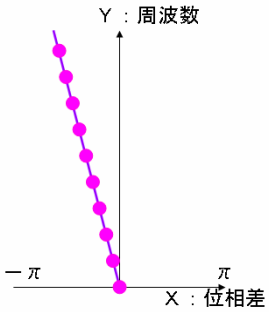


Fig.2　Phase-Frequency linearity

$$\Delta Ph = \Delta T \times 2\pi f$$
$$\rightarrow \Delta T = \Delta Ph / 2\pi f \quad \cdots \text{（式 2）}$$

マイク対で得られる 2 つの周波数分解データ a と b を比較して、同一周波数成分毎に両者の位相値の差を計算して ab 間位相差を求める。ある周波数成分 f の位相差ΔPh(f)は、マイク 1 における位相値 Ph1(f)とマイク 2 における位相値 Ph2(f)の差を計算し、その値が{ΔPh(f)：－π＜ΔPh(f)≦π}に収まるように 2π の剰余系として算定する。

次いで、算定されたΔPh(f)と f を 2 次元の XY 座標系上の点(x, y)としてプロットすると、Fig.2 に示したような位相差プロット図を得ることができる。既に述べたように、同一時間差ΔT に対応する位相差ΔPh(f)は周波数 f に比例するので、もしこのプロット図上に点群を結ぶ直線が検出できれば、この直線の傾きから式 2 で求められるΔT、すなわち式 1 で示されるφの方向に音源の存在を検出することができる。

2. 2.　直線ハフ変換

本システムにおける音源の数と方向を推定する問題は、Fig.2 のようなプロット図上で有力な直線を発見することに帰着できる。また、音源毎の周波数成分を推定する問題は、検出された直線に近い位置に配置された周波数成分を選別することに帰着できる。そこで、点群から直線を検出する手段として直線ハフ変換[3]を用いる。

Fig.3 に模式的に示すように、2 次元座標上の点 p(x, y)を通り得る直線は図中に例示するごとく無数に存在するが、原点 O から各直線に下ろした垂線の X 軸からの傾きをθ、この垂線の長さをρとして表現すると、1 つの直線についてθとρは一意に決まり、ある点 p(x, y)を通る直線の取り得るθとρの組は、θ ρ座標系上で固有の軌跡（$\rho = x \cdot \cos\theta + y \cdot \sin\theta$）を描くことが知られている。このような、(x, y)座標値からそこを通り得る直線の(θ, ρ)軌跡への変換を直線ハフ変換という。複数の点を共通に通る直線は各点の軌跡が 1 点で交差するため、所定の投票バッファに軌跡を投票することで、多数の点を通る有力な直線を高得票位置に検出することができる。これをハフ投票という。なお、このとき、直線が左に傾いているときθは正値、垂直のとき 0、右に傾いているとき負値であるとし、また、θの定義域は{θ：－π＜θ≦π}を逸脱することはない。

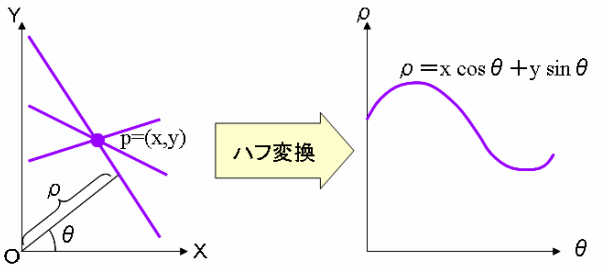


Fig.3　Principle of Hough transformation

2. 3.　実際の音声による位相差の方向別傾向確認

Fig.4 に方向を変えた実際の音声を使って得たパワースペクトルと位相差スペクトルを例示する。図の上段の 2 つがマイク 1 と 2 による入力音声波形、中段の 2 つがマイク 1 と 2 によるパワースペクトル、下段が位相差スペクトルである。各スペクトルの輝度はパワーの対数で計算し、位相差の色は位相角に応じた色円環上の色相で決定している。

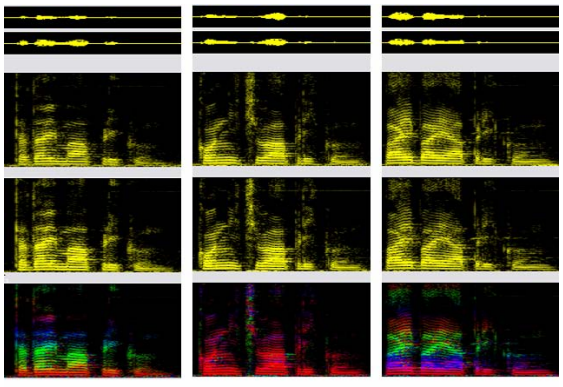


Fig.4　Phase difference that has a trend according to sound source direction

図中（a）は、マイク対の右から発声したときの結果を示したものである。位相差は周波数の高い領域に行くにつれて橙色から紫色へ変化するグラデーション、すなわち色円環上で一定方向への回転を示し

ている。

図中（b）は、マイク対の正面から発声したときの結果を表したものであり、位相差スペクトルは回転せずに 0 付近（赤色）に留まっている。

図中（c）は、マイク対の左から発声したときの結果を示したものである。位相差は周波数の高い領域に行くにつれて赤紫色から橙色へ変化するグラデーション、すなわち色円環上で（a）の逆方向への回転を示している。

いずれの場合も音源は移動していないため、周波数成分毎の位相差は時間軸方向でほぼ一定に安定している。すなわち、位相差スペクトルは音源の方向を知るための情報として信頼できる。そして、この位相差スペクトルの各周波数成分値 ΔPh は前述の直線の傾き θ から式 3 によって知ることができる。なお､位相差 ΔPh が負値となるとき､θ は正値となる。そのために、θ の符号を反転させている。

$$\Delta \mathrm{Ph}(\theta, f) = f \cdot \tan(-\theta) \quad \cdots \text{（式 3）}$$

2.4. ρ=0 の制約

マイク 1 と 2 の信号が同相で A/D 変換される場合、検出されるべき直線は必ず ρ=0、すなわち XY 座標系の原点を通る。したがって、音源の推定問題は、ハフ投票バッファ上の得票分布 S(θ, ρ)で ρ=0 となる θ 軸上の 1 次元の得票分布 S(θ, 0)からローカルピークを探索する問題に帰着する。

Fig.5 に実際の単独発話音声を使って直線を検出した例を示す。この例は、室内雑音環境下で 1 人の人物がマイク対の正面約 20 度左から発話した実際の音声を用いて処理した結果である。図中にマイク 1 と 2 のパワースペクトル及び位相差スペクトル（a）、図中（b）に FFT 結果から得た周波数成分毎の位相差プロット図を、図中（c）に位相差プロット図から得たハフ投票結果を、図中（d）に ρ=0 上の得票分布を、図中（e）に検出されたローカルピーク（直線候補）を、図中（f）に得票 1 位のピークをプロット図上に描画した直線（赤）をそれぞれ示す。

マイク対で取得された音声は、周波数成分毎のパワー値と位相値のデータに変換される。これを受けて、周波数成分毎の位相差が求められ、その(x, y)座標値が算出される。この座標値の集合をプロットした図を確認すると、原点から左に傾いた直線に沿う点群分布が認められる。このような分布を示している各点の(θ, ρ)軌跡がハフ投票バッファに投票されて得票分布 S(θ, ρ)を形成する。

θ 軸上の得票分布 S(θ, 0)を H(θ)として抜き出して棒グラフにしたものが図中（d）である。この得票分布 H(θ)には幾つか極大部が存在している。図中（e）にローカルピークの探索結果を示す。このようにすることで十分な得票を得た直線の θ を割り出すことができる。この例では、検出された 4 つのピークのうち、閾値処理によって 1 位のみが検出されている。

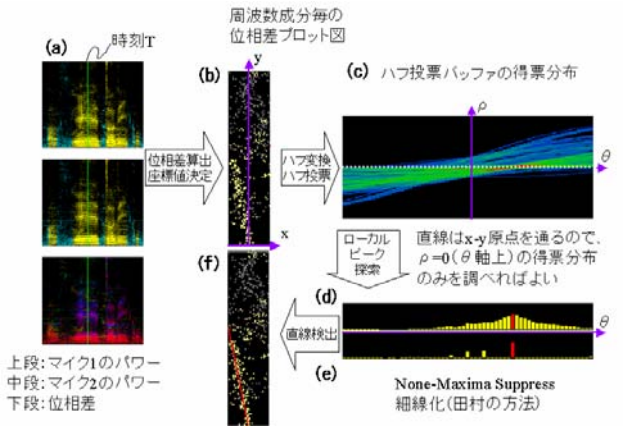


Fig.5 An experimental result of line detection with ρ =0 constraint in a case of single speaker

2.5. 位相差の循環性

ところで、両マイク間の位相差 ΔPh(f)が Fig.2 に示したように全域で周波数に比例するのは、解析対象となる最低周波数から最高周波数まで通して真の位相差が$\pm\pi$を逸脱しない場合に限られる。この条件は ΔT が、最高周波数 Fr/2[Hz]（サンプリング周波数 Fr の半分）の 1/2 周期分の時間、すなわち 1/Fr[秒]以上とならないことである。もし、ΔT が 1/Fr 以上となる場合には、次に述べるように位相差が循環性を持つ値としてしか得られないことを考慮しなければならない。

手に入れることのできる周波数成分毎の位相値は例えば$-\pi$からπの間というように 2πの幅でしか得ることができない。これはその周波数成分における実際の位相差が両マイク間で 1 周期以上開いていても、データとして得られる位相角からそれを知ることができないことを意味する。要するに、ΔT に起因する真の位相差は、データとして得られた位相差の$\pm 2\pi$や、さらに$\pm 4\pi$や$\pm 6\pi$の値である可能性がある。これを模式的に示すと Fig.6 のようになる。図中の○が ΔT を同じくする真の位相差だが、観測されるのは左に 2π 平行移動して現れる●となる。

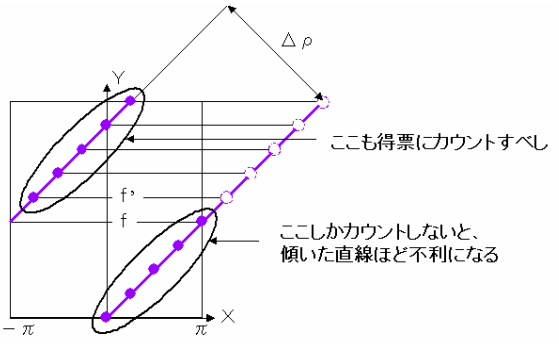


Fig.6 Circulation of phase difference

ρ=0 の制約では原点を通る直線のみを探すことになるので､このように 2π の剰余系として循環した位置に現れる点群をカウントしていない。これは傾きの大きな直線ほど得票で不利になることを意味している。そのことは Fig.5（d）に示した得票分布にお

いて、θの絶対値が大きくなる左右端付近にほとんど得票値のないことから見てとれる。θの全域に渡る公平な探索を実現するためには、循環して現れる平行線群を得票分布 H(θ)に加えなければならない。なお、平行線の間隔Δρは直線の傾きθの関数Δρ(θ)として式4で定義される符号付きの値となる。

$$\Delta\rho(\theta) = 2\pi\cdot\cos\theta \quad : \theta > 0$$
$$\Delta\rho(\theta) = -2\pi\cdot\cos\theta \quad : \theta < 0 \quad \cdots \text{（式4）}$$

2.6. 位相差循環を考慮した直線群の検出

Fig.5 で検出した直線は XY 座標原点を通る直線である。しかし、実際には位相差の循環性によって、Fig.6 に示すように、原点を通る直線がΔρだけ平行移動して反対側から循環してくる直線もまた同じ到達時間差を表す直線である。この直線のように原点を通る直線を延長して X の値域からはみ出した部分が反対側から循環的に現れる直線を、原点を通る直線の「循環延長線」、基準となった原点を通る直線を「基準直線」とそれぞれ呼ぶことにする。

Fig.7 に位相差の循環性を考慮して直線群を検出した例を示す。この例は、Fig.5 の例と同じ音声を用いて処理した結果である。

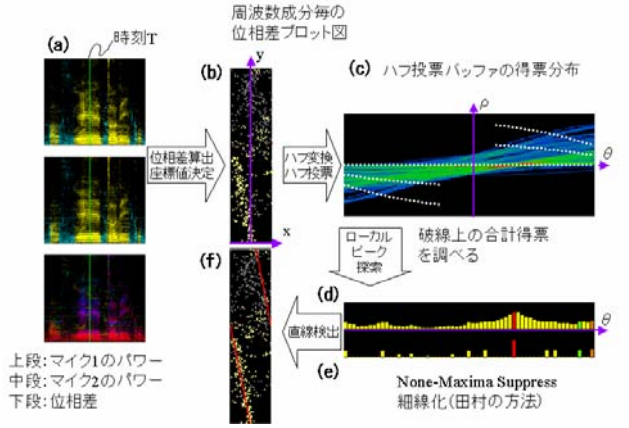


Fig.7 An experimental result of line detection with ρ=0 constraint and considering circulation of phase difference in a case of single speaker

直線群を表す同一θについてΔρずつ離れた箇所をハフ投票バッファ（c）上で白線表示している。このとき、θ軸と白線はそれぞれΔρ(θ)の自然数 a 倍で等間隔に離れている。なお、直線が循環しないθの領域（中央部）には白線を描画していない。

あるθ0の得票 H(θ0)は、θ＝θ0の位置で縦に見たときのθ軸上と白線上の得票の合計値、すなわち $H(\theta 0)=S(\theta 0,\ 0)+\Sigma\{S(\theta 0,\ a\Delta\rho(\theta 0))\}$ として計算される。この操作はθ＝θ0となる基準直線とその循環延長線の得票を合計することに相当する。この得票分布 H(θ)を棒グラフにしたものが図中（d）である。Fig.5（d）と異なり、θの絶対値が大きくなっても得票がなくなっていない。これは、得票計算に循環延長線を加えたことで全てのθについて同じ周波数帯を使うことができるようになったからである。得票で1位のピーク位置によって、プロット図（f）上に平行な2本の直線が描かれる。

2.7. 同時発話時の直線検出

Fig.8 に、室内雑音環境下で 2 人の人物がマイク対の正面約 20 度左と約 45 度右からほぼ同時に発話した実際の音声を用いて処理したときのスペクトル（a）、FFT 結果から得た周波数成分毎の位相差プロット図（b）、位相差プロット図から得たハフ投票結果（c）、得票 H(θ)とローカルピーク検出結果（d, e）、プロット図上に描画した直線群（f）を示す。

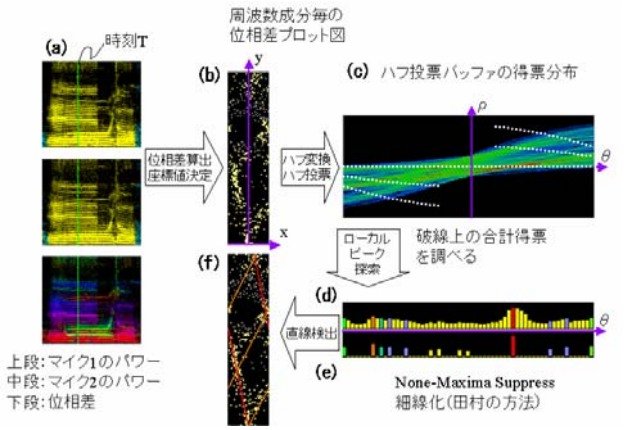


Fig.8 An experimental result of line detection with ρ=0 constraint and considering circulation of phase difference in a case of two simultaneous speakers

得票分布 H(θ)からは図中（e）に示す 13 個のローカルピークが検出される。このうち、上位 2 つのローカルピークがマイク対の正面約 20 度左からの音声を検出した直線群と、マイク対の正面約 45 度右からの音声を検出した直線群に対応している。このようにΔρずつ離れた箇所の得票値を合計して極大位置を探索することで、角度の小さい直線から角度の大きい直線まで検出できるようになる。

2.8. ストリーム追跡

上述した通り、直線群はハフ投票毎に時系列的に求められることになる。このとき、直線群のθは音源方向φと 1 対 1 に対応しているので、音源が静止していても移動していても、安定な音源に対応するθの時間軸上の軌跡は連続しているはずである。一方、検出された直線群の中には、ローカルピークの閾値設定具合によって背景雑音に対応する直線群が含まれていることがある。しかしながら、このような直線群の軌跡は連続していないか、連続していても短いことが期待できる。すなわち、直線群検出周期毎に求められるθの時間軸上の軌跡を追跡して長く連続するグループを検出することで、有力な音源を選別することができる。このグループをストリーム、グループ分けを行う処理をストリーム追跡と呼ぶことにする。

Fig.8 に実際の音声を使ったストリーム追跡結果の例を示す。横軸が時間、縦軸がθである。この例は、

室内雑音環境下で 2 人の人物が異なる方向から同時発声したときのストリーム追跡結果である。0～4 番までの 5 つのストリームが追跡により検出されている。このうち、1 番と 3 番が正解、0 番と 4 番が背景での人の話し声、2 番が偽のストリームである。

なお、音源が大きく移動しないと仮定すれば、ストリームを構成する各時刻の直線群による θ を平均したものをストリームの θ とすることができる。

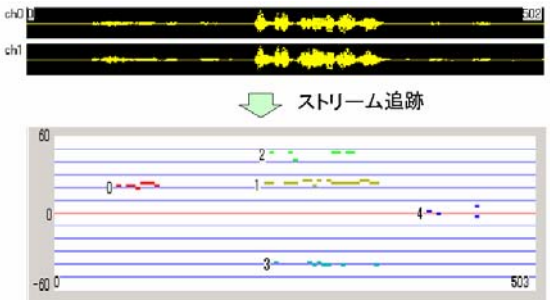


Fig.8　An experimental result of stream tracking

2.9.　成分推定とストリーム照合

以上で説明した処理は、1 つのマイク対においてそれぞれ実行される処理である。ロボットは複数のマイク対を実現できるので、マイク対をまたぐ処理によって、さらに音源についての情報を得ることができる。

異なるマイク対で検出されたストリームでも、同一音源に由来する限り、その継続期間と周波数成分は似ているはずである。既に述べたように、各マイク対で検出される音源の主な周波数成分は、その証拠となった直線群の近傍に分布するプロット図上の点を選別することで推定することができる。このように音源の周波数成分を粗く推定し、推定された周波数成分を例えば単純類似度法などで比較照合することで、あるマイク対のある直線群が、別のマイク対のどの直線群と似ているかを評価することができる。

ストリーム照合は、推定された音源の周波数成分と継続期間を評価することで、同時期に似た周波数成分を持っている音源をマイク対間で対応付ける処理である。対応付けられるべき相手が見つからないストリームはノイズとして削除される。

2.10.　音源定位

ストリーム照合によって対応付けられたストリームは、各マイク対に対する音源方向 φ（θ から計算）を、対応付けられたマイク対の数だけ持っている。これをマイク対の数 n を使って表すと、音源方向情報＝{φ1…φn}となる。この集合は、ある空間位置にある音源が、それぞれのマイク対から見てどの方向にあるかを示したデータである。そこで、ロボットを中心に仮想的なドームを考え、そのドーム表面に適度な間隔で離散した仮想的な音源を配置し、各仮想音源が各マイク対のどの方向にあるかを予め計算してテーブル化しておく。Fig.9 は 2 つのマイク対を使ったときの音源定位の概念を示した図である。各マイク対から得られた音源方向 φ1 と φ2 について最小二乗誤差となるドーム上の仮想音源を探索する。

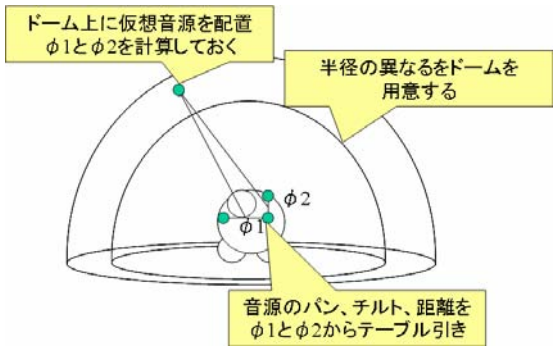


Fig.9　Principle of sound source localization with φ1 and φ2

3.　動作確認実験

3.1.　システム構成

2 章で説明した処理によって全方位聴覚を実現するシステム構成を Fig.10 に示す。

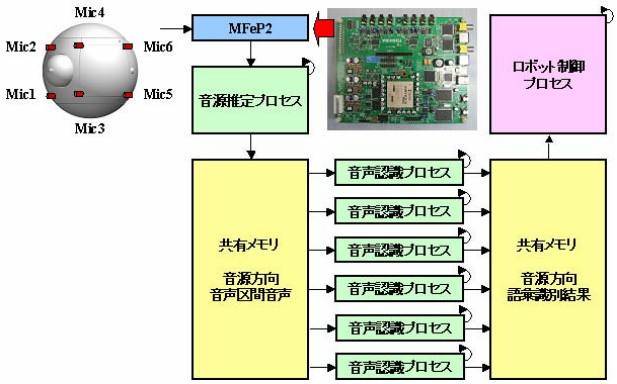


Fig.10　System block diagram

ロボットは Mic1～6 の 6 つのマイクを装備している。システムは 1 つの音源推定プロセスと最大 6 つの音声認識プロセスとで構成されており、プロセス間は共有メモリで結ばれている。音源推定プロセスは前述した音響信号処理を行って、各マイク対に対する音源数と方向の推定処理と、複数のマイク対を使った音源の空間定位処理を行う。音声認識プロセスは認識エンジンの前段に適応アレイ処理を配したプロセスであり、マイク対に対する音源の方向 φ を使って適応アレイの追従範囲を設定し、音源音の抽出と抽出された音声の認識を行う。音声認識プロセスは音源推定プロセスによって方向の異なる音源毎に処理対象を割り当てられて認識を実行する。マイクからの音響信号は新開発のメディア処理ボード（MFeP2）によって全チャンネル同期取り込みがされるため、ローカルピーク探索時には原点を通る直線群（基準直線とその循環延長線）を探索すればよい。

3.2.　処理の流れ

Fig.11 にこのシステムの処理の流れを示す。音声は入力されると FFT 処理を施される。入力音声からレベルによって音声区間の始端が検出されると終端検出までの間が音声区間とされ、ハフ変換からストリーム追跡までの処理が進められる。終端が確定すると、ストリーム照合以降の処理が実行されて、認識結果が出力される。

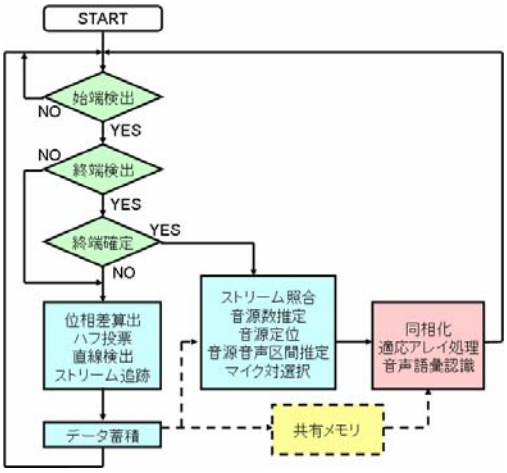


Fig.11 Flow diagram

3.3. マイク対選択･同相化･適応アレイ処理･音声語彙認識

音源が検出されると、音源推定プロセスがその音源方向に対して他の音源方向とかぶらないユニークなマイク対を選択し、音声認識プロセスがこのマイク対からの入力音声を認識する。音声認識プロセス内では、選択されたマイク対からの 2ch 音響信号を同相化することで、音源があたかもマイク対の正面にあるかのような信号を生成する。このように正面向きに補正された音響信号を同じく正面向きに狭追従範囲を与えられた適応アレイで処理することで、適応アレイの設定限界に制約されることなく、どの方向からの音声も処理できるようにする。適応アレイ通過後の音響信号は認識エンジンで処理され、得られた認識結果は共有メモリ上に格納されて利用される。

3.4. 4 話者順次発話時の全方位性の確認

以上の処理を実装して、4 人の話者が順に発話したときの音源定位･音声認識実験を行った。Fig.12 に実験の様子を示す。発話内容は 4 種類で、適応アレイの追従範囲を±15 度とし、マイクは Mic3～6 の 4 個、マイク対は Mic3-4、Mic5-6、Mic3-5、Mic4-6 の 4 組を使用した。実験の結果、ロボットは発話順に話者方向を向き、その前まで移動して、発話内容に応じた応答音声を出力した。

3.5. 2 話者同時発話時の全方位性の確認

2 人の話者がほぼ同時に発話したときの音源定位･音声認識実験を行った。Fig.13 に実験の様子を示す。発話内容は 2 種類で、適応アレイの追従範囲を±15 度とし、マイクは Mic3～6 の 4 個、マイク対は Mic3-4、Mic5-6、Mic3-5、Mic4-6 の 4 組を使用した。実験の結果、ロボットは発話開始順に話者方向を向き、その前まで移動して、発話内容に応じた応答音声を出力した。

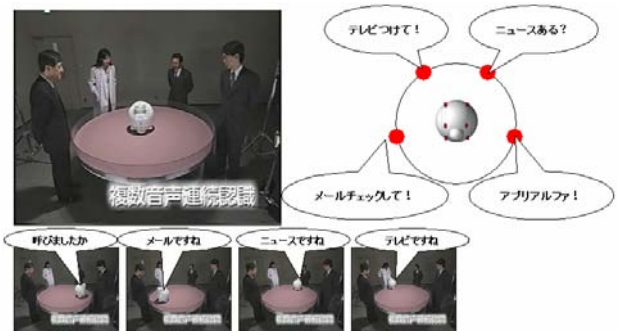


Fig.12 Demonstration of four sequential speakers

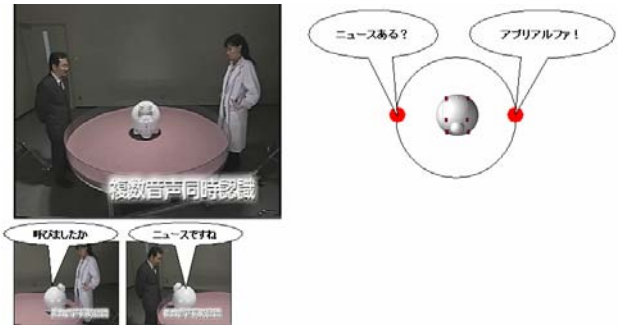


Fig.13 Demonstration of two simultaneous speakers

4. おわりに

本稿では、家庭内で運用されるロボットにとって不可欠と思われる全方位聴覚の 1 方式を報告した。

今回は 4 話者順次発話時と 2 話者同時発話時の音源定位と音声語彙認識における全方位性を実験によって確認するに留まったが、提案方式が機能し得ることを検証することができた。

なお、本開発は NEDO（新エネルギー･産業技術総合開発機構）次世代ロボット実用化プロジェクト（プロトタイプ開発支援事業）に採択され実施したものであり、2005 年愛地球博 NEDO プロトタイプロボット展（6/9～6/19、モリゾー･キッコロメッセ）、および常設展（8/23～9/4、ロボットステーション）で展示実演を行った。そこで、本システムの複数話者順次発話対応機能を実演したところ、80dB 程度の周囲雑音まで動作可能であることを確認している。

参考文献

[1] 浅野太, “音を分ける”, 計測と制御, 第 43 巻, 第 4 号, pp.325-330, Apr.2004

[2] 中臺一博他, “視聴覚情報の階層的統合による実時間アクティブ人物追跡”, 工知能学会 AI チャレンジ研究会, SIG-Challenge-0113-5, pp.35-42, Jun.2001

[3] 岡崎彰夫, “はじめての画像処理”, 工業調査会刊, Oct.2000

人間共生ロボット”EMIEW”の聴覚機能

Auditory Ability of Human-Symbiotic robots ”EMIEW”

戸上真人 天野明雄 新庄広 鴨志田亮太 ((株) 日立製作所 中央研究所)
玉本淳一 柄川索 ((株) 日立製作所 機械研究所)
Masahito TOGAMI,Akio AMANO,Hiroshi SHINJO,
Ryota KAMOSHIDA(Hitachi, Ltd., Central Research Laboratory),
Junichi TAMAMOTO,Saku Egawa(Hitachi, Ltd., Mechanical Engineering Research Laboratory)
{mtogami,amano,shinjo,ryota-k}@crl.hitachi.co.jp,
{saku.egawa.kv,junichi.tamamoto.xs}@hitachi.com

Abstract

Sound source localization and distant talk recognition are essential functions for human-symbiotic robots. We describe methodology of sound source localization based on sound sources overlap judge and adaptation method for minimum variance beam-former based on frequency segregation in this paper. These auditory functions are implemented in ”EMIEW”(Excellent Mobility and Interactive Existence as Workmate).

1 はじめに

接客ロボット, 介護ロボット, 家事手伝いロボットなどの人間共生型ロボットの実現に向けて, 音声インターフェースへの期待が大きくなってきている。タッチパネルやリモコンなどと比較し, 音声インターフェースは, 離れたところから, 特別な道具を使わずに情報伝達が可能な, 簡便なユーザーインターフェースである。そして, 特にロボットにおいては, 音声インターフェースの簡便さに対する期待だけでなく, 呼びかけられた方向を振り向いたり, ロボットが人間の音声を認識し, 人間の言葉を話すという行為そのものが, 人間とロボットの心理的な距離を縮める効果があり, 人間に対する親和性の良さという点でも期待が大きい。

そのようなことから, 我々は, 来訪者の受付案内やオフィスにおける物流サポートなど, 人と同じ環境で生活し人と一緒に仕事ができるロボットとして開発を進めている,EMIEW(Excellent Mobility and Interactive Existence as Workmate) の開発目標の一項目に, ロボット聴覚機能の実現を挙げている。

EMIEW は聴覚機能の他に, 人間と共に生活するために必要となる以下の機能を持っている [1]。

- 高速移動機能（人の早足に相当する時速６ｋｍ）
- 障害物回避機能（人ごみでも人を避けて, 移動が可能）
- 高品位音声合成機能 [2]
- 顔画像抽出機能

本稿では,EMIEW の聴覚機能について, 音源定位機能及び妨害音抑圧機能について述べる。

音源定位機能については, 複数の音源が存在する環境でも音源定位が可能である音源重複度判定に基づく修正遅延和アレイ法を提案する。従来の遅延和アレイ法 [3] は, 複数の音源が存在する環境で高精度に音源定位することは困難であった。またロボットへの実装を考えた場合, ハードウェアの処理能力の制約から固有値計算などの重い処理を必要とする音源定位方式は実装が困難となる。提案する方式は, 音声のスパース性を前提にし, 方向毎に周波数成分を振り分ける。そして音源重複度判定に基づき複数の音源が重複している時間・周波数成分を判定し, 単一の音源だけが優勢な時間・周波数成分のみを使ってパワースペクトル計算を行うことを特徴とする。

妨害音抑圧機能については, 修正遅延和アレイ法に基づき妨害音が優勢な帯域のみを抽出し, その帯域のみで計算した相関行列を用いてフィルタリングすることを特徴とする最小分散ビームフォーマの適応方式を提案する。そして, 提案する適応方式の音の伝達モデルの誤差に対するロバスト性などについて評価を行う。

以下,2 章では,EMIEW の聴覚機能の構成を述べる。3 章では, 前提とする入力信号のモデル化を行い,4 章

で音源重複度判定に基づく修正遅延和アレイ法を説明する。5 章では, 修正遅延和アレイ法に基づく最小分散ビームフォーマの適応方式を説明する。

2　EMIEW の聴覚機能の構成

Fig.1 に EMIEW の外観を示す。EMIEW は首周りに 6 つのマイク, 両耳に 2 つのマイクを持っている。

Figure 1: Appearance of EMIEW

聴覚機能の構成を Fig.2 に示す。

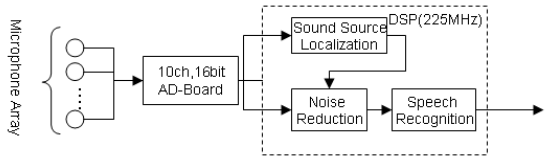


Figure 2: Construction of Auditory Functions

8 つのマイクロホンの入力信号は,EMIEW の筐体内にある 16 ビットの AD ボードに取り込まれる。その後,DSP(225MHz) 上で 360 °の全方位音源定位やある方向の音のみを抽出する妨害音抑圧処理を施す。抽出された音は音声認識に渡される。音声認識結果や音源方向情報は, メイン制御プログラムに渡される。

3　入力信号のモデル

本稿で前提とする入力信号のモデルについて述べる。

3.1　空間伝達モデル

空間内に存在する各音源の信号は, 空間的な場所に応じて異なった空間伝達過程でマイクに到達する。各音源が点音源で直接音のみ考慮する場合, 空間伝達過程は音源とマイクまでの距離による位相遅れと減衰によってモデル化することができる。この場合, $\boldsymbol{a}_d(f)$ を音源方向 d からの空間伝達モデルとすると,

$$\boldsymbol{a}_d(f) = \left[\frac{1}{r_{d,1}}\exp^{-i2\pi f\tau_{d,1}} \cdots \frac{1}{r_{d,M}}\exp^{-i2\pi f\tau_{d,M}}\right]^T \tag{1}$$

ここで, $r_{d,i}$ は音源 d とマイク i の距離で,c を音速とすると,

$$\tau_{d,i} = \frac{r_{d,i}}{c} \tag{2}$$

である。

3.2　入力信号のモデル

マイクロホンアレイの入力信号に短時間フーリエ変換を施した後の信号を $\boldsymbol{x}(\tau : f)$ と記述する。τ は短時間フーリエ変換のフレームインデックスである。

N 個の音源が存在する場合, 入力信号は以下のようにモデル化することができる。

$$\boldsymbol{x}(\tau : f) = \boldsymbol{A}(f)\boldsymbol{S}(\tau : f) \tag{3}$$

ここで,

$$\boldsymbol{A}(f) = \Big[\boldsymbol{a}_1(f) \cdots \boldsymbol{a}_N(f)\Big], \tag{4}$$

$$\boldsymbol{S}(\tau : f) = \Big[s_1(\tau : f) \cdots s_N(\tau : f)\Big]^T \tag{5}$$

であり,$\boldsymbol{a}_i(\tau : f)$ は音源 i の空間伝達モデル,$s_i(\tau : f)$ は音源 i の原信号である。

3.3　スパース性

1990 年代後半より, 音声のスパース性と呼ばれる性質を利用した新しい妨害音抑圧手法が盛んに研究されるようになってきた [4][5]。音声のスパース性とは, 短時間（数十 ms 程度）では, 音声のパワーはある一部の帯域に集中するという性質である。音声のスパース性は,「短時間に複数の音源が同じ周波数成分を保持することは確率的に低い」と言い換えることができる。この音声のスパース性に基づき, 式 (3) は, 次のように近似することができる。短時間フーリエ変換のフレームサイズは, 数十 ms 程度とする。

$$\boldsymbol{x}(\tau : f) \approx s_{j(\tau:f)}(\tau : f)\boldsymbol{a}_{j(\tau:f)}(f) \tag{6}$$

ここで, $j(\tau : f)$ は, フレーム τ に周波数 f を保持する音源のインデックスである。

4 音源定位

提案する音源定位法は, まず時間・周波数毎に遅延和アレイの出力値から入力信号をどこか一つの角度に振り分ける。そして振り分けられた成分が, 単一の音源のみからなるかどうか音源重複度判定で判定する。単一の音源のみからなると判定された成分を方向毎に積み上げ, パワースペクトルを計算する。そのパワースペクトルをピークサーチすることで音源方向の推定値を得ることができる。

4.1 従来の遅延和アレイ法

音源方向の探索範囲を Λ とする。遅延和アレイ法は,$d \in \Lambda$ 毎に, 式 (7) に示す方向パワースペクトルのピークサーチを行い, 音源方向を見つける手法である。

$$P(d) = \sum_f \Big(\boldsymbol{a}_d(f)^* \boldsymbol{R}(f) \boldsymbol{a}_d(f) \Big) \tag{7}$$

$\boldsymbol{R}(f)$ は空間相関行列で,

$$\boldsymbol{R}(f) = \sum_\tau \Big[\boldsymbol{x}(\tau : f) \boldsymbol{x}(\tau : f)^* \Big] \tag{8}$$

が用いられる。式 (3) と式 (8) を式 (7) に代入すると,

$$P(d) = \sum_f \sum_\tau \Big| \sum_{i=1}^{N} \boldsymbol{a}_d(f)^* \boldsymbol{a}_i(f) s_i(\tau : f) \Big|^2 \tag{9}$$

となる。$\boldsymbol{a}_d(f)^* \boldsymbol{a}_i(f)$ は, 探索方向と音源方向との一致度を表す係数であり, 大きさが 0 から 1 までの値を取る。探索方向と音源方向が完全に一致した場合のみ,1 となる。ここで, 探索方向と音源方向が一致しない場合に,$\boldsymbol{a}_d(f)^* \boldsymbol{a}_{j(\tau:f)}(f)$ が必ずしも 0 になるわけではないところに注意する必要がある。つまり, サーチ方向と音源方向が一致しない場合に, サーチ方向のパワースペクトルに音源方向のパワースペクトルがノイズとして, 混入してしまう。そして, このことは, 遅延和アレーの音源定位精度を劣化させる要因となる。

4.2 修正遅延和アレイ法

入力信号について, 式 (6) の近似が成立することを仮定する。この場合, 式 (9) は, 以下のように変形できる。

$$P(d) = \sum_f \sum_\tau \Big| \boldsymbol{a}_d(f)^* \boldsymbol{a}_{j(\tau:f)}(f) s_{j(\tau:f)}(\tau : f) \Big|^2. \tag{10}$$

式 (6) の近似が成立し, 音源 $j(\tau : f)$ のみが, 時間 τ の周波数 f 成分を保持するのであれば, 本来 $d = j(\tau : f)$ 以外の場合,$\Big| \boldsymbol{a}_d(f)^* \boldsymbol{a}_{j(\tau:f)}(f) s_{j(\tau:f)}(\tau : f) \Big|^2$ は 0 となるべきである。また $\boldsymbol{a}_d(f)^* \boldsymbol{a}_{j(\tau:f)}(f)$ は,$d = j(\tau : f)$ の場合に最大となるため,$j(\tau : f)$ は, 入力信号から次のように推定することができる。

$$j(\hat{\tau : f}) = \mathrm{argmax}_{j \in \Lambda} \Big| \boldsymbol{a}_j(f)^* \boldsymbol{x}(\tau : f) \Big| \tag{11}$$

そこで, 式 (10) を次のように修正する。

$$P(j) = \sum_f \sum_\tau \begin{cases} 0 & j \neq j(\hat{\tau : f}) \\ \Big| \boldsymbol{a}_j(f)^* \boldsymbol{x}(\tau : f) \Big|^2 & j = j(\hat{\tau : f}) \end{cases}. \tag{12}$$

式 (12) の方向パワースペクトルに基づく音源定位手法を修正遅延和アレー法と呼ぶことにする。修正遅延和アレー法は, 遅延和アレイ法の問題点であった音源方向以外の方向へのパワースペクトルの混入問題を解決することができると考えられる。本稿の実験では, 式 (12) のパワースペクトルの対数を足し合わせた

$$P(j) = \sum_{\tau, f} \begin{cases} 0 & j \neq j(\hat{\tau : f}) \\ \log \Big| \boldsymbol{a}_j(f)^* \boldsymbol{x}(\tau : f) \Big|^2 & j = j(\hat{\tau : f}) \end{cases}. \tag{13}$$

を用いる。

4.3 音源重複度判定

音声のスパース性が完全に成立する場合は, 修正遅延和アレイ法は, 高い音源定位性能を示すことが予想される。しかし, 実際には, 複数の音源が重複する時間・周波数成分が存在し, その時間・周波数成分について, 修正遅延和アレイ法は, 正しい音源方向を推定することができない。この時,$j(\hat{\tau : f})$ がどの方向を指すかは, 各音源の原信号 $s_i(\tau : f)$ のパワー比や空間伝達過程に依存して変わる。パワー比によっては, 音源の全く存在しない方向を $j(\hat{\tau : f})$ が指すこともありうる。

このようなことから, 音声のスパース性が成立しない周波数では,$j(\hat{\tau : f})$ は信頼性の低い情報であり, 音声のスパース性が成立しない周波数成分は, 音源定位に利用しないほうが良いと考えられる。

そこで, 音声のスパース性が成立するかどうかを次の式で判定することにする。

$$e(\tau : f) = \frac{\Big| \boldsymbol{x}(\tau : f) - \boldsymbol{a}_{j(\hat{\tau : f})}(f)^* \boldsymbol{x}(\tau : f) \boldsymbol{a}_{j(\hat{\tau : f})}(f) \Big|^2}{\Big| \boldsymbol{a}_{j(\hat{\tau : f})}(f)^* \boldsymbol{x}(\tau : f) \boldsymbol{a}_{j(\hat{\tau : f})}(f) \Big|^2}. \tag{14}$$

式 (14) の分子は,$j(\hat{\tau : f})$ 方向にビームを向けた遅延和アレイが抑圧した妨害音のパワーである。音源が重複していない場合, 遅延和アレイが抑圧した妨害音のパワーは 0 となり, 式 (14) は 0 となる。音源が重複している場合, 遅延和アレイが推定方向以外の到来音を抑圧するため, 式 (14) は 0 より大きくなる。

音源重複度の指標である式 (14) を使って式 (13) で,$j \neq j(\hat{\tau : f})$ または $10 \log_{10} e(\tau : f) \geq P_{th}$ の時,$\sum$

の中を 0 とするように改良する。P_{th} は音源重複度の閾値であり, 予め定める必要がある。式 (14) の音源重複度の指標を用いる修正遅延和アレイ法を音源重複度判定型修正遅延和アレイ法と呼ぶ。

4.4 音源定位の評価

提案する修正遅延和アレイ法及び音源重複判定型修正遅延和アレイ法の音源定位性能の評価を行う。用いるマイク数は 3 つとした。また音源数を 4 つとした。音源 4 つのうち 1 つを目的音, 残りの 3 つを妨害音とする。音源の間隔は,2 度〜5 度まで 1 度おきに変化させた。音源とマイクロホンアレー中心との距離は 1m とした。マイク間隔は 5cm とした。評価で用いる音声試料には,ATR 音韻バランス 150 文に含まれる各話者 50 文ずつ, 男女 4 話者分を用いた。4 話者のうちの 1 話者を目的音とし, その他の話者を妨害音とする計 4 通りの組み合わせで, 評価した。評価用データは計算機上のシミュレータで作成したデータである。シミュレータでは反響・残響が無いものとし, 直接音のみ混合した。音源定位時に用いる音源重複度の閾値 P_{th} は, 予備実験より -40(db) が妥当な値であると判断し, 設定した。また音源定位では, サーチ方向のパワースペクトルが, 前後の方向のパワースペクトルより, 大きい場合にピークと判定する。そしてピークと判定された方向をパワースペクトルの降順で並び替え, 最初から 4 つ目までを音源方向として出力するようにした。サンプリングレートを 11025Hz とし, フレームサイズは 512 ポイントとした。

推定した音源方向が, 真の音源方向から ± 1 度のずれであれば, 正しく方向推定できたものとする。正しく方向推定できた割り合い（音源定位正解率）を評価基準とした。

遅延和アレイと, 式 (15) の最小分散ビームフォーマ (MVBF) と比較評価を行う。

$$P(j) = \sum_f \log \frac{1}{\left(\boldsymbol{a}_j(f)^* \boldsymbol{R}(f)^{-1} \boldsymbol{a}_j(f)\right)^2}. \qquad (15)$$

Fig.3 に評価結果を示す。

提案した音源重複判定型修正遅延和アレーが評価した音源間隔の全ての場合で, 最も良い性能を示した。音源間隔が 2 度の結果では, 音源重複判定型修正遅延和アレイは, 遅延和アレイに対して, 約 75%の改善,MVBF に対して, 約 61%の改善, 修正遅延和アレイに対して, 約 22%の改善となった。このことより, 音声のスパース性に基づく遅延和アレイの修正法, 及び音源重複判定尺度に基づく複数の音源が重複した時間・周波数成分の判定法の有効性が示された。

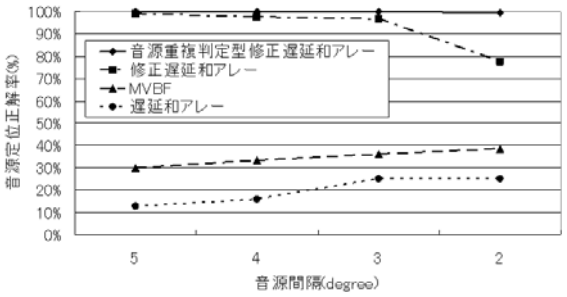


Figure 3: Evaluation of Sound Source Localization

5 妨害音抑圧法

従来の妨害音抑圧法である最小分散ビームフォーマ [3] は, 目的音方向の音の空間伝達モデルを与える必要があるが, 想定している目的音方向と実際の目的音方向のずれ, 環境毎の反響・残響特性の違いなどで, 実際の空間伝達過程はモデルと異なったものとなることが多い。従来の適応方法は, 目的音の伝達モデルのずれが生じると, 目的音を妨害音とみなし目的音方向に死角を形成するようにフィルタを制御するという問題がある。

5.1 提案する適応方法

提案手法は, 全ての周波数成分で適応するのではなく, 予め周波数成分を妨害音が優勢な成分と目的音が優勢な成分に振り分け（周波数成分の振り分け）妨害音が優勢な成分のみを使ってフィルタを適応する。周波数成分の振り分けは, 音源定位の章で述べた修正遅延和アレイ法を応用する。

提案する適応方法に基づく妨害音抑圧システムのブロック図を Fig.4 に示す。

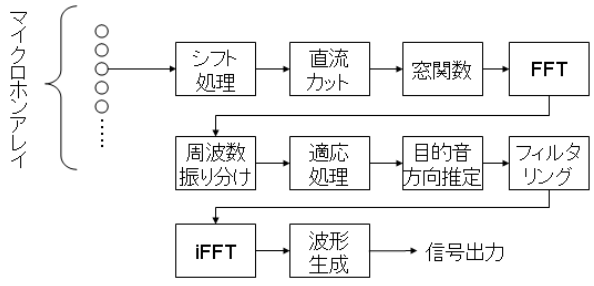


Figure 4: Block Diagram of Supposed Noise Reduction System

以下, 提案手法で特徴的な, 周波数振り分け, 適応処理, 目的音方向推定について述べる。

5.1.1 周波数振り分け

提案する適応方法で用いる周波数振り分け処理は, 修正遅延和アレイ法に基づく。式 (11) により, 時間・周波数毎に音源方向の推定値である $j(\tau\hat{:}f)$ を求める。

目的音方向及び目的音方向周辺の音源方向を含む Λ の部分集合を $\Lambda_{subject}$ とし,Λ の中で $\Lambda_{subject}$ に含まれない方向の集合を妨害音方向と定義し,Λ_{noise} で表す。$j(\tau\hat{:}f)$ が Λ_{noise} に含まれる場合はその成分を妨害音が優勢な成分であると判定し,$\Lambda_{subject}$ に含まれる場合はその成分を目的音が優勢な成分であると判定する。周波数振り分けの結果, 出力される妨害音信号は, 以下のように記述される。

$$\boldsymbol{n}(\tau:f)=\begin{cases}0 & if\ j(\tau:f)\in\Lambda_{subject}\\ \boldsymbol{x}(\tau:f) & if\ j(\tau:f)\in\Lambda_{noise}\end{cases} \tag{16}$$

5.1.2 適応方法

最小分散ビームフォーマの適応方法の 1 つである SMI による直接解法 [6] では, 音源方向の情報を保持するマイク間の相関行列の逆行列 (空間相関逆行列) を使ってフィルタ係数を算出する。空間相関逆行列は, 以下の式に基づき, 入力信号から逐次更新することが可能である。

$$\boldsymbol{R}(\tau+1:f)^{-1}=\frac{1}{\beta}\boldsymbol{R}(\tau:f)^{-1}- \frac{(1-\beta)\boldsymbol{R}(\tau:f)^{-1}\boldsymbol{x}(\tau+1:f)\boldsymbol{x}(\tau+1:f)^{*}\boldsymbol{R}(\tau:f)^{-1}}{\beta^{2}+\beta(1-\beta)\boldsymbol{x}(\tau+1:f)^{*}\boldsymbol{R}(\tau:f)^{-1}\boldsymbol{x}(\tau+1:f)}. \tag{17}$$

提案手法では, 入力信号では無く, 式 (16) で求めた妨害音信号を使い, 以下の式で空間相関逆行列を逐次更新する。

$$\boldsymbol{R}(\tau+1:f)^{-1}=\frac{1}{\beta}\boldsymbol{R}(\tau:f)^{-1}- \frac{(1-\beta)\boldsymbol{R}(\tau:f)^{-1}\boldsymbol{n}(\tau+1:f)\boldsymbol{n}(\tau+1:f)^{*}\boldsymbol{R}(\tau:f)^{-1}}{\beta^{2}+\beta(1-\beta)\boldsymbol{n}(\tau+1:f)^{*}\boldsymbol{R}(\tau:f)^{-1}\boldsymbol{n}(\tau+1:f)}. \tag{18}$$

空間相関逆行列を用いて, 最小分散ビームフォーマのフィルタを,

$$\boldsymbol{w}(\tau:f)=\frac{\boldsymbol{R}(\tau:f)^{-1}\boldsymbol{a}_{sub}(f)}{\boldsymbol{a}_{sub}(f)^{*}\boldsymbol{R}(\tau:f)^{-1}\boldsymbol{a}_{sub}(f)}$$

と表すことができる。ここで,$\boldsymbol{a}_{sub}(f)$ は目的音方向推定処理で求める目的音方向の空間伝達モデルである。

5.1.3 目的音方向推定

修正遅延和アレイ法を使いフレーム単位に目的音方向を推定する。

$$P(j)=\sum_{f}\begin{cases}0 & j\neq j(\tau:f)\ \text{のとき}\\ \log\left|\boldsymbol{a}_{j}(f)^{*}\boldsymbol{x}_{t}(f)\right|^{2} & j=j(\tau:f)\ \text{のとき}\end{cases} \tag{19}$$

$\Lambda_{subject}$ に含まれる方向 j について, 式 (19) で定義されるパワースペクトル $P(j)$ が最大となる方向を目的音方向とする。

5.2 評価

提案する適応方法に基づく妨害音抑圧法の出力信号の SNR による評価と, 実環境での遠隔音声認識性能の評価を行う。比較対象は, 全ての周波数成分で適応する従来の最小分散ビームフォーマと, 周波数振り分けの結果を使い,

$$\boldsymbol{S}(\tau:f)=\begin{cases}0 & if\ j_{\tau:f}\in\Lambda_{noise}\\ \boldsymbol{x}(\tau:f) & if\ j_{\tau:f}\in\Lambda_{subject}\end{cases}$$

で求めた目的音が優勢な成分 $\boldsymbol{S}(\tau:f)$ を逆 FFT する方式 (周波数振り分け型) の 2 つとする。

5.2.1 SNR による評価

提案手法の目的音方向の空間伝達モデルのずれに対するロバスト性及び妨害音抑圧性能について, 式 (20) で定義される SNR による評価を行う。

$$SNR=10\log_{10}\frac{\sum_{t}|S_{t}|^{2}}{\sum_{t}|\hat{S}_{t}-S_{t}|^{2}} \tag{20}$$

ここで,S_t は目的音の原信号であり,$\hat{S}_t$ は目的音の推定信号である。用いるマイク数は 6 とした。音源数は 2 とした。2 つの音源のうち 1 つを目的音, もう 1 つを妨音とする。本評価では, 計算機上のシミュレータで作成した空間伝達過程を使い目的音と妨害音を混合した。マイク配置は EMIEW の首周りのマイク配置とする。目的音源の音声試料は, 孤立 10 数字発話で, 妨害音源用音声試料は, 展示会場騒音とする。目的音方向と妨害音方向の角度差は 30 °とした。目的音源はマイクロホンアレイの正面に配置した。目的音方向の空間伝達モデルは, 距離 1m として, 式 (1) で作成する。マイクロホンアレイと音源の距離は 0.5m とした。目的音方向の空間伝達モデルが距離 0.5m ずれた場合のロバスト性を評価することになる。目的音の探索範囲 $\Lambda_{subject}$ は,-10 °〜+10 °までとし,10 °刻みとする。妨害音の探索範囲 Λ_{noise} は,-180 °〜-20 °及び+20 °〜+170 °とし, 同じく 10 °刻みとする。評価結果を Fig.5 に示す。従来の最小分散ビームフォーマの SNR に比べ, 提案法の SNR は,15db 程度上回っている。提案法は, 目的音方向の空間伝達モデルのずれに対して, ロバストであることが分かる。また周波数振り分け型と比較しても,3db 程度上回っている。

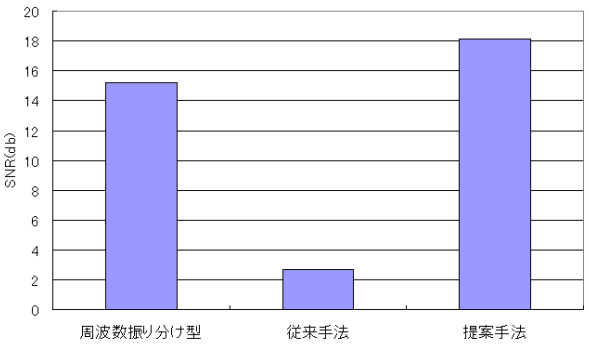


Figure 5: Evaluation of Robustness

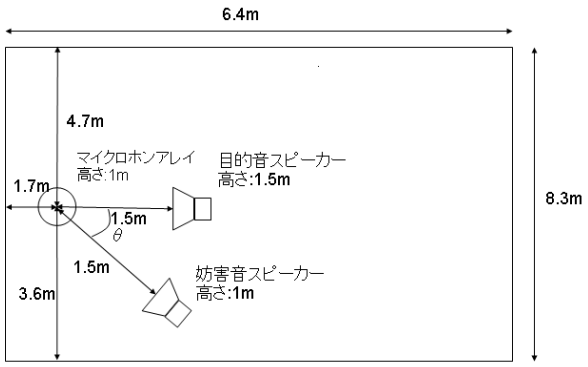


Figure 6: Recording Environment

5.3 遠隔音声認識性能の評価

実環境での音声認識性能の評価を行う。音声試料は, 孤立 10 数字発話とする。話者数は 80 話者, 発話数は,800 発話とする。妨害音源は 1 つで, 方向を 30 °〜180 °まで 30 °刻みで変化させた。目的音と妨害音は S/N が 5db になるように調節した。3 つの手法の出力信号を同一の音声認識エンジン (LPC ケプストラムベース) に入力し, 音声認識率を測定する。

評価データの収録環境を Fig.6 に示す。

評価結果を Fig.7 に示す。提案手法は, 妨害音方向

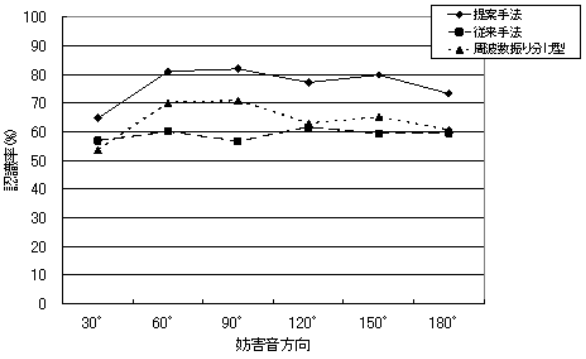


Figure 7: Evaluation of Speech Recognition

60 °〜150 °までで 80%程度の認識率となり, 従来手法に対して 20%程度上回ったことから, 妨害音が優勢な成分で適応することが有効であると考えられる。また, 音声認識は信号の歪みに敏感であるが, 本提案手法は歪みが小さいことから, 周波数振り分け型より認識率が高いという認識結果になったと考えられる。

6 まとめ

人間共生ロボット”EMIEW”の聴覚機能について, 音源定位機能と妨害音抑圧機能について述べた。音源定位機能としては, 複数の音源が重複している帯域を判別する音源重複度判定に基づく音源定位法を提案し, 従来手法と比べ, 音源定位正解率が向上することを確認した。妨害音抑圧機能としては, 修正遅延和アレイの出力結果を利用し, 妨害音が優勢な成分のみで適応処理を行う最小分散ビームフォーマの適応方法を提案し, 従来手法と比べ, 空間伝達モデルのずれに対してロバストであることと, 音声認識率が向上することを確認した。

本研究は, 独立行政法人：新エネルギー・産業技術総合開発機構 (NEDO 技術開発機構)「次世代ロボット実用化プロジェクト プロトタイプ開発支援事業」の一環として行われたものである。

参考文献

[1] 細田祐司他, ”人間共生ロボット”EMIEW”の開発-開発コンセプトと全体システム-” 第 23 回日本ロボット学会学術講演会,2005 年 9 月

[2] N.Nukaga,R.Kamoshida and K.Nagamatsu, ”Unit selection using pitch synchronous cross correlation for Japanese concatenative speech synthesis,” 5th ISCA Speech Synthesis Workshop,pp.43-48,2004

[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] M.Aoki, M.Okamoto, S.Aoki, H.Matsui, T.Sakurai, and Y.Kaneda, ”Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” Acoust.Sci & Tech. vol.22,no.2,pp.149-157,2001.

[5] Ö.Yılmaz, and S.Rickard, ”Blind Separation of Speech Mixtures via Time-Frequency Masking,” IEEE Trans.SP,Vol.52,No.7,2004.

[6] 菊間信良, ”アレーアンテナによる適応信号処理,” 科学技術出版,1998.

認知神経科学から見たロボット聴覚：「聴知覚のダイナミクス」

Cognitive Neuroscience: The dynamics of auditory perception

○柏野　牧夫 (日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 /
科学技術振興機構 ERATO 下條潜在脳機能プロジェクト)
* Makio KASHINO (NTT Communication Science Laboratories, NTT Corporation /
ERATO Shimojo Implicit Brain Function Project, JST)

kashino@avg.brl.ntt.co.jp

Abstract—Human auditory perception changes dynamically according to the history of input acoustic signals. We have been studying the dynamics of auditory perception by integrating various approaches including psychophysics, brain imaging, neurophysiology, and mathematical modeling. Here I describe three types of perceptual phenomena we have examined, and discuss their neural mechanisms and functional significance.

1. はじめに

聴取者が知覚する音の世界は，耳に入力された音響信号の物理的特性とは単純に対応しない．両者の間には，系統的ではあるがきわめて複雑な関係がある．この対応関係を分析すれば，背後にある情報処理メカニズムに関する重要な手がかりが得られる[1)]．

ここで注目したいのは，音響信号と聴知覚との対応関係は固定的なものではなく，それまでに入力された音響信号の履歴（すなわち聴取者の「経験」）に応じてダイナミックに変化するということである．このような変化には，ミリ秒から数分オーダーのごく短期的で一過性のものから，数日から数年というオーダーの長期的で固定的なものまでさまざまなものがある（後者はしばしば可塑性あるいは学習と呼ばれる）．筆者らは，とくに前者について，心理物理学，神経生理学，脳活動計測，数理モデル等の手法を統合することによって分析し，聴知覚の形成過程や，その神経メカニズムの解明を図っている．本稿では，その中から3種類の話題を紹介する．ひとつめは，同一の単語を反復聴取すると聞こえ方が変化するという錯覚に関するものである．あとのふたつは，知覚対象が定位されるべき空間的，時間的枠組の変化に関するものである．

これらの聴知覚のダイナミックな変化は，音響信号と知覚との乖離，すなわち錯覚をもたらす処理のエラーのように思われるかも知れない．しかし実は，環境や処理ハードウェアに由来する種々の制約の中で適切な知覚を実現するための巧妙な戦略とみることができる．

2. 反復単語の知覚的変化

2.1. 単語変形効果

最初に紹介するのは，物理的には同一の音であるのに，反復すると聞こえ方がどんどん変わっていくという現象である．このような現象は，単純な音系列のグルーピングにおいても観察できるが[2)]，ここでは音声（speech）を素材とする現象を取り上げる．まず，短い単語を録音する．例えば「バナナ」とやや早口で発声する．次に，それを切れ目なくループ（反復）再生する．すると，「バナナ」のはずが，「ナッパ」になったり，「ハナ」になったり，さらには二人の声に分かれたり，機械的な音が聞こえてきたりと，人によって中身はさまざまだが，普通は1分間も聞けばかなりの変化が体験できる．次々と違う内容が聞こえる人もいれば，比較的少数の聞こえ方が交互に現れるという人もいる．時には全く何も変化しないという人もいるが，別に異常ではない（ただし，加齢につれて変化が起こりにくくなるというデータもある）．この現象自体は古くから知られていて，単語変形効果（verbal transformation）と呼ばれている[3)]．視覚でも，同一の刺激に対して見ている間に知覚が変わる，いわゆる多義的知覚と呼ばれる現象は，ネッカー・キューブ，両眼視野闘争などいろいろある．単語変形効果は，それらと似ているところもあるし，違うところもある．「バナナ」をループすると，「ナバナ」，「ナナバ」など音節のまとまり方に多義性ができるが，聞こえてくる内容はそれよりもバリエーションが多く，もともと含まれていない音素が聞こえることも珍しくない．

2.2. 錯覚に伴う脳活動

なぜこのような奇妙な現象が生じるのか，そのメカニズムはよくわかっていない．その手がかりを得るために，筆者らは，機能的磁気共鳴画像法（fMRI）を用いて，反復単語を聴取しているときの脳活動を計測した[4)]．ここで興味があるのは，「聞こえ方が変化する」という事象に関連した脳活動である．単純に音を聞いていることによる活動や，「バナナ」や「ナッパ」など特定の聞こえ方に対応する活動ではない．そこで，聞こえ方が変化するたびに実験参加者にボタンを押してもらい，それに関連した脳活動の部位を推定した．さらに，対照実験として，同じ反復単語に加えてときどき提示される短い純音を検出してボタンを押すという課題を別のセッションで行い，その脳活動も計測した．純音の提示タイミングは，同一の被験者の，同一反復単語の聞こえ方が変化する頻度に基づいて設定したが，聞こえ方の変化とは同期していない．

実験の結果，聞こえ方の変化を検出する課題で有意な活動が見られた部位は，左右聴覚野，左右前頭前野（PFC），左前頭葉腹側部（IFC; ブローカ領域），左島皮質，前帯状皮質（ACC），右頭頂間溝，視床前部などである（Fig. 1）．一方，反復単語中の純音を検出する課題では，左右の聴覚野から後方に頭頂葉にかけての部位，左島皮質，視床前部の一部などに活動が見られた．詳細な分析は省くが，両条件で顕著な違いが見られたのは前頭葉である．単語変形効果における知覚の変化は，聴覚野だけではなく，前頭葉を中心とし広範囲に分散した脳部位の連携によって生み出されていると考えられる．

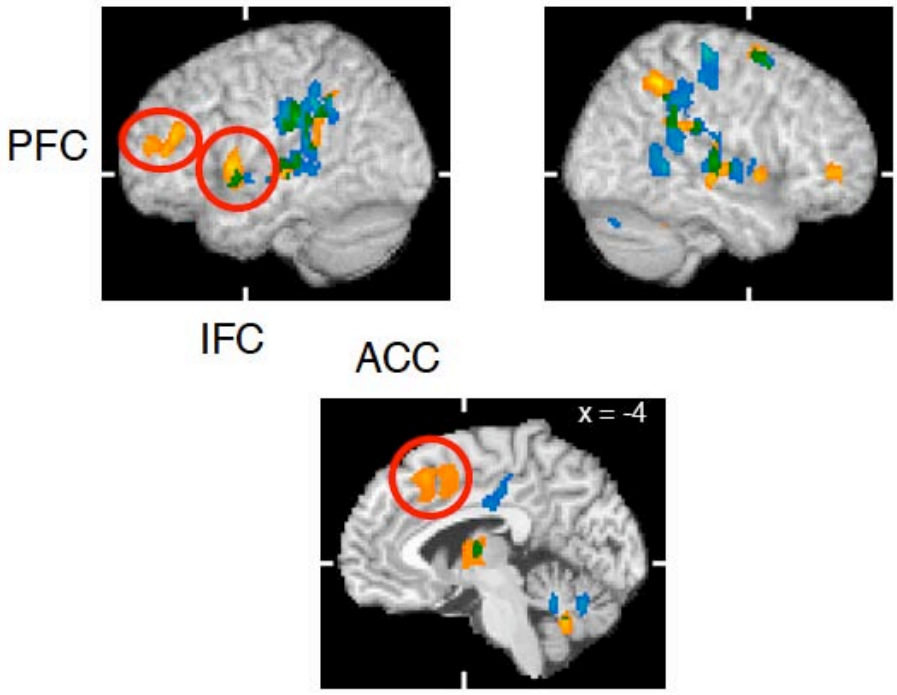


Fig. 1. Brain activity synchronized with verbal transformations (orange) and tone detection (blue) (N=12). Overlapped areas are green[4].

2.3. 機能的意義

単語変形効果は，同一単語の反復提示というきわめて不自然な状況で生じる錯覚であるが，その背後には，実環境で役に立つ情報処理メカニズムが見え隠れする．実環境では，耳に入力される音響信号は，いつでも完全であるとは限らない．情報の欠落や変形のため，多義的であったり，不分明であったりすることがしばしばある．同一単語の反復提示は，このような不完全な入力信号のモデルになっている．まず，反復によって音節のまとまり方が多義的になっている．さらに，特徴抽出過程の順応によって感覚情報が時々刻々変形していく．このような場合には，ボトムアップの感覚情報だけではなく，脳内に蓄えられた情報に基づくトップダウンの予測生成が重要な役割を果たすのではないだろうか．

このような考え方に合致するデータが脳活動計測の結果から得られている．単語の聞こえ方が変化したタイミングで，いわゆるブローカ領域の活動が見られたことは先述した通りである．ブローカ領域は，伝統的に，言語音声の生成に関与すると考えられてきた．しかしこの実験の課題はもっぱら知覚課題であって，生成課題ではない．そこで思い出されるのが，「音声の知覚は，音響信号から，それを生成した発話者の調音ジェスチャーを推定することに基づいてなされる」という音声知覚の運動理論である[5]．単語変形効果におけるブローカ領域の活動は，曖昧な音声を分節化するにあたって，音声生成に関する内部モデルによる予測が用いられていることを示唆しているのではないだろうか．この仮説を裏付けるように，ブローカ領域の活動は，知覚の変化を頻繁に報告した人の方がそうでない人に比べて顕著であった（Fig. 2）．運動理論は今日なお賛否両論あるが，筆者らの実験結果は，そのひとつの証拠と言えるかも知れない．

一方，次々にトップダウンの予測を生成するだけでは，入力信号の適切な解釈に至るとは限らない．ボトムアップの感覚情報によって，常にトップダウンの予測の妥当性を検証することが必要である．また，システムの安定性を保つためには，あまりにも頻繁に解釈が変化することは避けなければならない．このような，トップダウンの「暴走」を防ぐメカニ

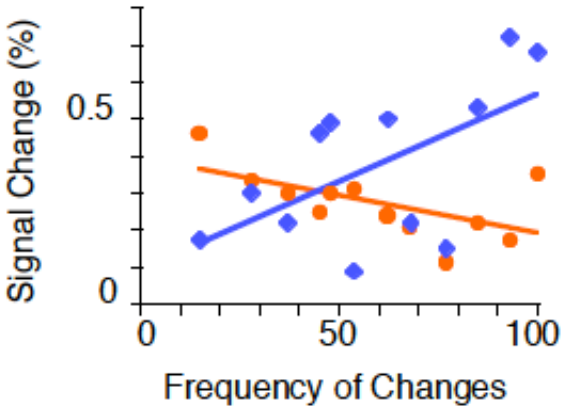


Fig. 2. Signal intensity in the left ACC (orange) and left IFC (blue) as a function of the number of perceptual changes for each participant [4].

ズムの存在も，脳活動計測のデータから見てとれる．前部帯状皮質の活動は，ブローカ領域の活動と逆に，知覚の変化が少ない人ほど顕著なのである（Fig. 2）．

内部モデルに基づく予測と感覚情報に基づく検証のダイナミックな相互作用によって聴知覚が形成される過程の全貌を明らかにするには，まだ検証すべき点が多い．単語変形効果のような特殊な錯覚を多角的に分析することによって，その手がかりが得られると期待される．

3. 知覚的空間の適応的変化

3.1. 聴覚定位残効

次は知覚対象そのものの特性ではなく，それが定位される空間的枠組みについての話である．物理的な手がかりと知覚される音源の位置との関係は，基本的に単一の音源について詳細に調べられてきた．しかし実際の環境では，音源はいろいろな位置に複数存在していることが多い．そのような場合，音源の位置を判断するときに，お互いが影響し合うことがある．筆者らは，ある両耳間時間差（ITD）を持つ音を提示した直後に別のITDを持つ音を提示して，その定位を聴取者に判断させる実験を行った．すると，後に提示した音は，単独で提示した場合に比べて，直前の音から遠ざかる方向にずれて知覚されることがわかった（Fig. 3）．これを聴覚定位残効と呼ぶ[6]．このずれは角度に換算して40度ほど両者が離れているとき最大となり，その量は15〜20度にもおよぶ．また，先行音と後続音の周波数が近いときにしか残効が生じない． 後に，広帯域雑音をスピーカで提示した場合にも，上下左右方向に同様の効果が生じることが示された[7]．ある音が存在することによって，知覚的な空間が系統的に歪むということもできる．この効果は，当初は比較的長時間の先行音を用いて観測されたが，1秒以下の短い先行音でも生じることがわかっている[8]．

3.2. 神経メカニズム

聴覚定位残効は，どのような脳内メカニズムによって生じているのだろうか．我々は，麻酔下のスナネズミを用い，下丘と呼ばれる脳幹の部位から神経応答を記録する神経生理実験を行った[9]．下丘は聴覚経路の中の要のような部分で，周波数や両耳間の時間差やレベル差に選択性をもつニューロン群が存在している．音源定位の手がかりとしては，まず両耳間の位相差（IPD；実験に用いたのは純音刺激なので時間差と等価）のみに注目した．また，先行音と検査音の時間長はそれぞれ200 ms と50 msと，人間の心理物理実験[6]よりはかなり短い値を用いた．

まず，単独の純音のIPDをさまざまに変えて単一のニューロンの反応強度（単位時間あたりの平均スパイク数）を測定すると，典型的には，あるIPDで非常に強く反応し，そこからIPDが離れていくと反応が減少するようなIPD選択性が見られた．次に，先行音のIPDをさまざまに変えながら同様の測定を行うと，先行音によって後続音に対する反応が規則的に影響を受けることがわかった．多くのニューロンでは，最も強く反応するIPDの値は先行音の有無によってほとんど変化せず，同調の鋭さも変わらなかったが，反応強度が先行音によって弱められるという現象が見られた．しかも，先行音のIPDがそのニューロンの同調している値に近いほど，反応強度の低下が著しかった．一言で言えば，先行音に対する反応が強いほど，後続音に対する反応が弱くなる，つまりニューロンの感度が低下するのである．

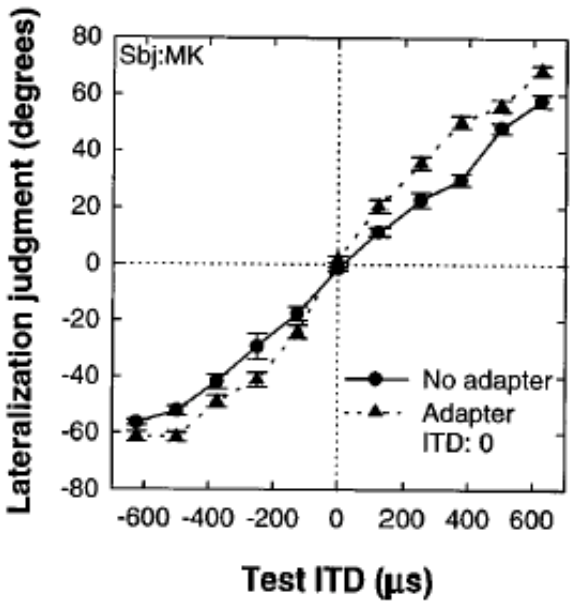


Fig. 3. The auditory localization aftereffect.[5] Mean lateralization judgments are shown as a function of the test ITD for a trained subject when no adapter was presented and when the adapter ITD was 0. The localization aftereffect is the difference between the two conditions. Error bars indicate the standard error of the mean.

このような個々のニューロンの挙動と聴覚定位残効のような知覚現象とを結びつけるには，ニューロンの集団の挙動を考える必要がある．そこで我々は，2種類のIPD符号化モデルをテストした．ひとつめは，さまざまなIPDに同調した数多くのニューロン全体の反応パターン（重心）でIPDが符号化されるというものである（Fig. 4）．もうひとつは，左右半球のいずれかに緩やかに同調したチャンネル間の興奮のバランスでIPDが符号化されるというものである．

いずれの場合にも，個々のニューロンは，先行音の存在によって感度を変化させると仮定する．入力が大きければ感度を下げ，小さければ逆に上げるという具合である．これらのモデルに，実際に測定されたニューロンの反応特性を適用して計算機でシミュレーションしてみると，いずれのモデルでも，後続音の位置が先行音から遠ざかる方向に系統的にずれるというパターンのモデル出力が得られた．これ

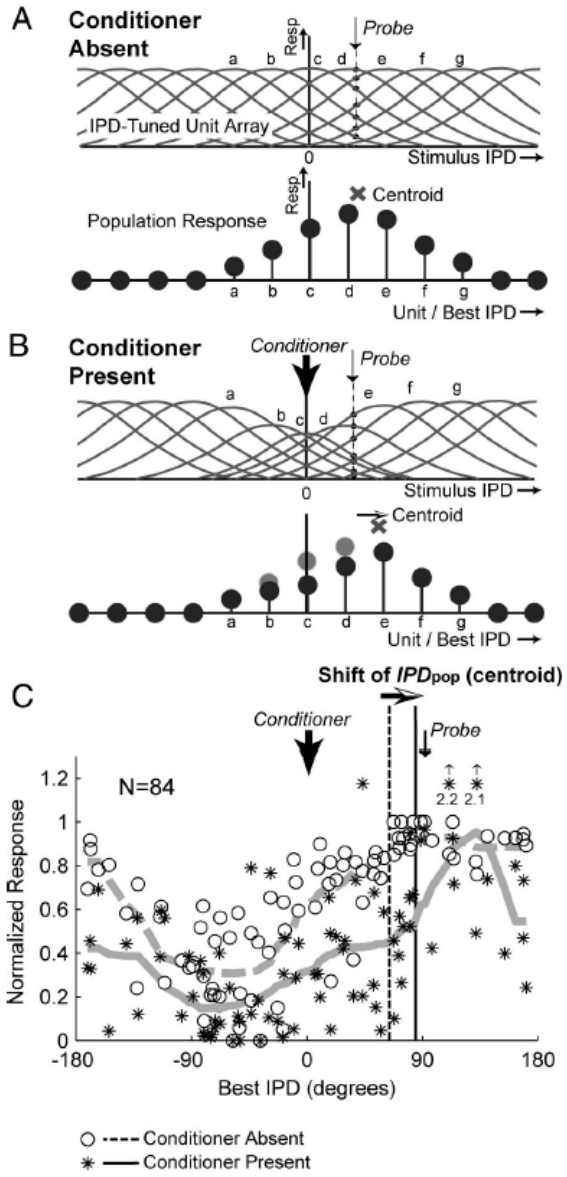


Fig. 4. Illustration of the population-vector model for IPD (interaural phase difference) representation, and an example of population responses.[9] A: an array of units tuned to various IPDs, curves representing the units' IPD functions of individual units (above) and the response strength distribution along the best IPD axis for a given probe IPD (below). Panel represents cases without a conditioner (adapter). B: as above, but with a conditioner. C: normalized responses of the unit array to a probe with an IPD of 90°, plotted against the unit's best IPD. Each open circle or asterisk represents the response of one unit in the absence of, and in the presence of, a 0-IPD conditioner, respectively. Two asterisks accompanied by upward-pointing arrows indicate the data points outside the ordinate rage, for which the normalized response values are indicated by the numbers below the symbols. Thick gray lines are drawn to guide the eye to the overall trends, by computing running averages using a Hanning window spanning 90° (broken line: conditioner absent; solid line: conditioner present). Vertical dashed (conditioner absent) and solid (conditioner present) lines indicate the centroids of the response distribution (i.e., IPD representations).

は人間で観測された定位残効の特性と共通している．種の違い，刺激の時間特性の違いなどがあるので性急に結論はできないが，下丘の段階までに，最終的な知覚内容に対応した神経情報がつくり出されている可能性がある．

3.3. 機能的意義

先行音によって後続音の位置がずれて知覚されるというのは，単に聴覚系が不正確だということを意味するのだろうか．

実は，このような処理特性には情報処理上の利点がある．わずかに位置の異なるふたつの音を弁別する実験を行うと，先行音がある場合には，ない場合に比べ，先行音の付近だけ位置の弁別力が向上するのである．そのかわり，先行音から離れたところでは弁別力は低下する[10, 11]．先行音の付近の空間が伸びるということは，ちょうどそこを虫眼鏡で拡大して見ているようなもので，新しい音と先行音との差分にはきわめて敏感になる．

定位残効に見られるような適応的な符号化の機能的な意義を，もう少し詳しく考えてみよう．一般的に言って，個々のニューロンは，ある範囲内では入力の値が大きくなると単位時間あたりのスパイク数が増えることによって入力値を表現している．しかしその範囲はあまり広くなく，上限付近では飽和し，下限付近では雑音に埋もれる．ところが，我々のモデルのように，入力値が大きいときに感度を下げ，小さいときに上げると，実効的には広い範囲の入力に対応してその差を表現することが可能となる．個々のニューロンがそのように振る舞えば，その全体としては，直前の入力の近くでわずかに異なった入力に対する興奮パターンの差が大きくなり，弁別力が向上することになる．これは知覚系一般に適用できる情報処理原理であろう．

以上のように，聴覚における空間情報処理は，ダイナミックで適応的なものである．聴取者自身も気づかぬうちに，入力される音に応じてあちこち処理の焦点を動かして，効率のよい処理を実現しているのである．

4. 知覚的同時性の適応的変化

4.1. 視聴覚同時性残効

どの感覚モダリティでも従来あまり研究されてこなかったことだが，時間についても先行刺激が後続刺激に影響することが最近わかってきた．聴覚でも音の時間順序判断が先行音の影響を受けるが[12]，ここでは視覚と聴覚にまたがった同時性判断における残効の例を紹介する[13]．

視覚刺激として黒い背景上の白いリング状のフラッシュ，聴覚刺激として短い純音を用い，両者をさまざまなタイミングで提示して，同時か否かを観察者に判断してもらう（Fig. 5）．この際，その前に数分間，視覚刺激よりも聴覚刺激の方が早いペアを観察すると，視覚刺激の方が先に提示されたときに同時と知覚されるようになる．逆に聴覚刺激の方が視覚刺激よりも早いペアを観察した後では，聴覚刺激の方が先に提示されたときに同時と知覚されるようになる．このような主観的同時点の移動すなわち残効の量は，種々の条件にもよるが，順応した時間ずれの15%程度になることもある．また，視聴覚が同時と判断される検査刺激の時間ずれの範囲も，順応刺激の時間ずれの方向に広がる（Fig. 6）．つまり，一定の視聴覚の時間ずれを経験すると，そのずれを小さくする方向に同時性の判断基準が移動するのである．

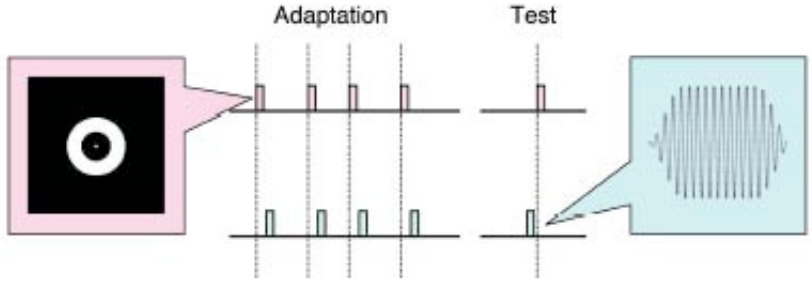


Fig. 5. The time course of the stimulus sequence used to test the effects of audiovisual lag adaptation on simultaneity judgments[13]. The left-hand box shows the configuration of the visual stimulus, and the right-hand box shows the waveform of the auditory stimulus.

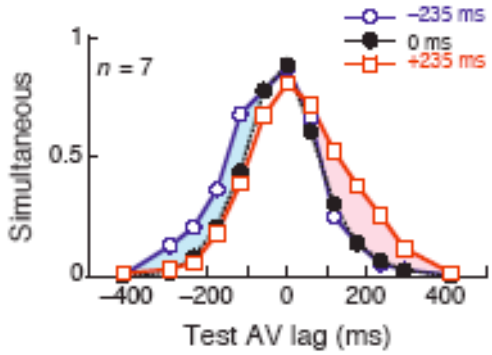


Fig. 6. The effects of audiovisual (AV) lag adaptation on simultaneity judgments[13]. The probability of ‘simultaneous’ response is shown as a function of the test audiovisual lag. The response probability for each lag was computed for each participant, and then averaged across participants. The main effect of adaptation is an increase in the probability of simultaneity on the side of adapted lag (shaded areas).

同様の残効は，直接的に視聴覚の同時性を判断しない課題でも観察できる．まず，残効の説明の前に，基本となる視聴覚現象を説明する．ふたつの小円が左右から近づいてきて，中央で交差するという視覚パターンを提示すると，ほとんどの場合小円が直進して通り過ぎているように見える．ところが，交差する瞬間に短い音を鳴らすと，小円が衝突して左右に反発しているように見える率が高まる[14]．この現象も，他の視聴覚統合現象と同様，短い音を鳴らすタイミングと小円が交差するタイミングとが合っていないと効果がなくなる．つまり，視聴覚情報の同時性は，両者を結びつける強い手がかりとなる．さて，ここからが残効である．小円が交差する瞬間と，音が鳴る瞬間とが一定時間ずれているパターンをしばらく観察すると，反発効果が最大となる視聴覚のタイミングがそちらの方向にずれるのである[13]．このときのずれのパターンは，上述した，リングと純音の同時性を直接的に判断する場合の結果と非常によく似ている．つまり，視聴覚の時間ずれに対する順応効果は，主観的な同時性判断だけではなく，知覚上の機能的な同時性にも影響することがわかる．

4.2. 機能的意義

このような順応の機能的意義を考える上で重要なのは，視聴覚情報の同時性を判断するのは，脳にとっては原理的に非常に難しい情報処理課題だということである．第一に，光と音では空気中を伝わる速さが全く違う．光は秒速約30万km，外界のイベントから目に到達するまでにかかる時間は無視できる．一方音は秒速340 m程度なので，対象が10 mばかり先にあれば，光よりも30 ms程度遅れて耳に到達することになる．もちろん，この時間差は対象の距離によって変化する．第二に，仮に目と耳に同時に刺激が届いたとしても，そこから神経の活動が視聴覚それぞれの経路を通って大脳皮質の視覚野と聴覚野に到達するまでの時間は無視できないくらい異なる．マカクザルの脳に電極を挿入して計測したデータや，人間の脳波や脳磁界の計測データによれば，神経活動が大脳皮質に到達する時間は聴覚の方が視覚よりも数十msも速い．この時間は注意の向け方や刺激の

強さなどによっても変化する．これらの二種類の要因によって，外界のイベントから光と音が同時に発せられたとしても，脳内では，視聴覚情報の間に単純に予測できない時間のずれが生じることになる．視聴覚の同時性判断が原理的に難しいといったのはこのためである．

では，この難しい情報処理課題を，脳はいかに解決しているのであろうか．ここで紹介した視聴覚の時間ずれに対する順応効果の意味するところは，入力された視聴覚情報間の時間のずれがある程度一定であれば，そのずれを小さくする方向に同時性の基準が移動するということである．つまり，脳は，固定的な同時性の基準を持っているのではなく，直近に経験した感覚情報をもとに，同時性の基準を適応的に変化させているのである．これが，物理的および神経的な非同期にもかかわらず，適切に視聴覚情報の同時性を判断するための脳の戦略であると考えられる．

5. おわりに

本稿で紹介した3種類の知覚現象は，知覚される対象も，それが定位されるべき時間や空間の枠組みも，物理的なものと同一ではなく，直近の経験に基づいて絶えず変化しているダイナミックなものであることを如実に物語っている．この機能があるがゆえに，環境の変化や知覚メカニズムに内在する原理的困難などにかかわらず，適切に情報を統合し，周囲の状況把握やそれに基づいた行動を行うことができるのである．

さて，このような知見は，各種ロボットの聴覚系を設計する上で，どのように役立つであろうか．自律ロボットに関しては，人間の知見が直接利用できるとは考えない方がよい．ロボットの聴覚系には，人間の聴覚系とは別の，それぞれ固有の目的，環境，ハードウェアの制約があるはずだからである．ただ，人間の話も，ともかくうまく動作しているひとつの例として，何かのヒントにはなるかも知れない．

一方，人間とロボットが一体となって何かを行うようなシステムの場合には，言うまでもなく人間のメカニズムを十分考慮しなければならない．例えば，遠隔地に，視覚センサ，聴覚センサ，自分の身体の動きを反映して動くアクチュエータを設置して，ネットワーク経由で遠隔操作する場合を想定してみよう．視覚情報，聴覚情報，運動情報の間にどの程度時間的，空間的なずれがあると知覚や行動が妨げられるであろうか．最初は不自然でも，しばらく体験しているうちに脳が適応し，問題なく活動できるようになるだろうか．ロボット技術と通信技術の発達によって，生身の人間の身体がもつ時間的，空間的制約から解放されることも現実味を帯びてきたが，果たして脳はそのような新しい身体に適応し，リアリティを獲得することができるであろうか．これはまさに，知覚系，運動系の短期的，長期的なダイナミクスの問題なのである．

謝辞

聴覚および視聴覚のダイナミクスに関する共同研究者である岡田美苗，水谷伸，Peter Davis，近藤洋史，古川茂人，牧勝弘，藤崎和香，西田眞也，下條信輔，小林まおり（順不同）の各氏に感謝します．

参考文献

1) 柏野 牧夫 (1998). 聴覚：環境に適応する無意識の知性. 日本音響学会誌 **54**, 508-514.
2) Okada, M., Mizutani, S., and Kashino, M. (2005). The dynamics of auditory streaming. *The 27th MidWinter Meeting of ARO*.
3) Warren, R.M. (1999). *Auditory Perception – A new analysis and synthesis*. Cambridge, UK: Cambridge University Press.
4) Kondo, H. and Kashino, M. (2005). Distributed brain activation involved in the changes of auditory perceptual organization: an fMRI study on the verbal transformation illusion. *The 27th MidWinter Meeting of ARO*.
5) Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* **74**, 431-461.
6) Kashino, M. and Nishida, S. (1998). Adaptation in sound localization revealed by auditory aftereffects. *J. Acoust. Soc. Am*. **103**, 3597-3604.
7) Carlile, S., Hyams, S., and Delaney, S. (2001). Systematic distortions of auditory space perception following prolonged exposure to broadband noise. *J. Acoust. Soc. Am*. **110**, 416-424.
8) Kashino, M. (1999). Interaction in the perceived lateralization of two sounds having different interaural time differences. *J. Acoust. Soc. Am*. **105**, 1343.
9) Furukawa, S., Maki, K. Kashino, M. and Riquimaroux, H. (2005). Dependence of the interaural phase difference sesitivities of inferior collicular neurons on a preceding tone and its implications in neural population coding. *J. Neurophysiol*. **93**, 3313-3326.
10) Kashino, M. (1998). Adaptation in sound localization revealed by auditory after-effects. In A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Eds.), *Psychological and physiological advances in hearing*. London: Whurr Publishers, Pp. 322-328.
11) Getzmann, S. (2004). Spatial discrimination of sound sources in the horizontal plane following an adapter sound. *Hear Res*. **191**, 14-20.
12) Okada, M. and Kashino, M. (2003). The role of frequency-change detectors in auditory temporal order judgment. *NeuroReport* **14**, 261-264.
13) Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audio-visual simultaneity. *Nature Neurosci*. **7**, 773-778.
14) Sekuler, R., Sekuler, A. B., and Lau, R. (1997). Sound alters visual motion perception. *Nature* **385**, 308.

対話音声における韻律と声質の特徴を利用したパラ言語情報の抽出の検討

Using prosodic and voice quality features for paralinguistic information extraction in dialog speech

○ 石井カルロス寿憲（(株)国際電気通信基礎技術研究所　知能ロボティックス研究所）
石黒浩（(株)国際電気通信基礎技術研究所　知能ロボティックス研究所）
萩田紀博（(株)国際電気通信基礎技術研究所　知能ロボティックス研究所）

* Carlos Toshinori ISHI, Hiroshi ISHIGURO, Norihiro HAGITA (ATR Intelligent Robotics and Communication Laboratories)

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

Abstract - The use of voice quality features in addition to classical prosodic features is proposed for automatic extraction of paralinguistic information (like speech acts, attitudes and emotions) in dialog speech. Perceptual experiments and acoustic analysis are conducted for monosyllabic utterances spoken in several speaking styles, carrying a variety of paralinguistic information. Acoustic parameters related with prosodic and voice quality features potentially representing the variations in speaking styles are evaluated. Experimental results indicate that prosodic features are effective for identifying some groups of speech acts with specific functions, while voice quality features are useful for identifying utterances with an emotional or attitudinal expressivity.

1　はじめに

人間とロボットの間で対話音声を介して円滑なコミュニケーションが成立するためには、言語情報の理解以上に、発話意図や話者の態度・感情などのパラ言語情報の理解も重要となる。人間同士の対話では、「えー」、「あー」、「うーん」などのような非語彙的な発話が、話し相手の発言に対するリアクションとして頻繁に用いられ、何らかの行為、態度、感情などのパラ言語情報を伝達する。また、このような非語彙的な発話では音素情報が殆ど含まれていないため、パラ言語情報の表現は、韻律情報及び声質情報に多く含まれると考えられる。

これまでのパラ言語情報の抽出に関する多くの研究は、基本周波数（F0）・パワー・持続時間など、イントネーションやリズムに関連する韻律特徴(prosodic features)を重視して来たが、自然発話を分析した最近の研究では、声質情報の重要性も示されている[1,2,3,4]。特に表現豊かな発話音声(expressive speech)では、気息性や非周期性などを含んだ non-modal な声質が現れやすく、F0 さえ測定できない場合も多いので[5]、韻律情報以外に、声質情報を知ることは重要となる。

「声質」（“voice quality”）は、話者特有の声の特徴や、声道・鼻腔での特徴的な声の質など広義で扱うことが可能だが、本稿では、狭義での声帯振動のモード（発声様式：phonation style）によって特徴付けられる声の質のことを指す[6]。

声質は、modal (地声), breathy 及び whispery (気息性のある声), vocal fry または creaky (基本周波数が非常に低く、パルス的な声), harsh または ventricular (雑音的で耳障りのある声), pressed (喉頭を力んだ声) 及び、これらの声質の組み合わせによって表現できる[6]。

著者の過去の研究[7,8,9,10]では、韻律及びさまざまな声質に関連する音響パラメータが提案され、本研究ではそれらのパラメータを使用し、図 1 に示すような韻律と声質の特徴を利用した構造を提案し、さまざまなパラ言語情報の抽出を試みる。

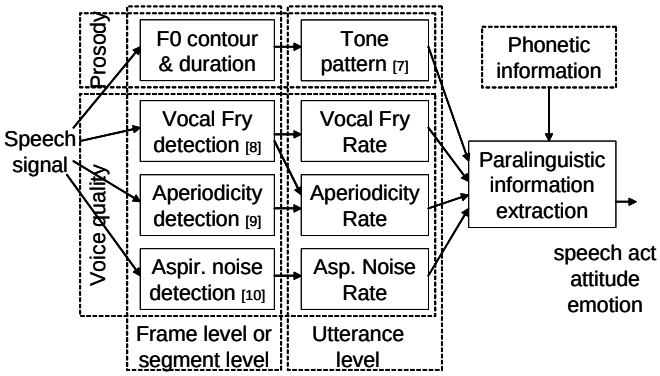


Fig. 1 Block diagram of the proposed framework for paralinguistic information extraction.

2　音声データと知覚データ

2.1 発話行為の音声データ

ある発話が伝達可能なパラ言語情報の種類は、図 1 でも考慮しているように、その発話の内容（Phonetic information）にも依存することが考えられるが、本稿では、対話音声でリアクションとして頻

繁に現れ、言い方（発話様式：Speaking style）の違いによって、伝達されるパラ言語情報が豊富な発話「え」に着目して検討する。ここで、「え」は、「ええ」、「えーー」、「え？」、「え！？」、「えっ」など、文字だけでは表現しきれないさまざまな発話様式を含んだものとする。

予備的な実験として、自然発話に現れるさまざまな発話様式の発話「え」を取り出し、それぞれが伝達するパラ言語情報をリストアップした結果、以下のようなリストにまとめられる。格好内の項目は意味的なニュアンスが多少異なるかもしれないが、ひとまとまりのカテゴリーとして扱うには問題ないと判断した。

肯定（肯定・承諾）、*同意*（同意・納得）、*相槌*（相槌・うなずき）、*考え中*（考え中・時間稼ぎ）、*戸惑い*（戸惑い・躊躇・困惑・迷い・悩み）、*同情*（同情・共感）、*感心*、*羨望*、*聞返し*、*意外*、*驚き*（驚き・びっくり）、*不満*、*非難*（非難・否定・拒絶）、*嫌悪*（嫌悪・いや）、*疑い*（疑い・不審）。

このリストには、行為的・機能的な役割を表すものや態度的なもの、感情的なものも含まれているが、明確な分類は難しいので、本稿ではこれらの項目をすべて「**発話行為**」と呼ぶことにする。

分析用のデータとしては、自然発話音声を扱うことが望ましいが、ここではパラ言語情報の観点からバランスのよいデータを求めるために、さまざまな発話行為を表現した発話音声を新たに収録する。そのために、指定の発話行為を表現した発声を誘導するような台本を作成し(付録の A 発話を参照)、この台本を先ず話者 1 名が発声したものを収録する。各発話行為において、例文は 2 個用意した。

次に、予め収録された各“誘導”発話を被験者に聞かせ、指定の発話行為を発話「え」にて表現するよう被験者に発声してもらう。より自然な発声が得られるように、発話「え」に後続して、指定の発話行為を強めるための短い発話も考案した（付録の B 発話を参照)。ただし、「え」と後続発話の間には短いポーズを入れるよう指示する。また、「え」で表現し難い場合は「へ」と発声することを許す。その他、追加発声として、自然発話では頻繁に現れるが、このような意図した発声では現れにくい喉頭を力んだ発声[11]を「え」と「へ」で発声してもらう。

さまざまな発話行為を意図して発声した 6 名の話者(15 才から 35 才の男性 2 名、女性 4 名)の音声データから、発話「え」もしくは「へ」の部分のみを切り出した総 207 発話を分析対象とする。

2.2 発話行為の知覚データ

発話行為の知覚ラベルを付与する理由は二つ挙げられる。一つ目は、特定の発話行為を意図して発声された発話「え」が文脈なしでどの程度聞き手に伝わっているのかを調べることである。もう一つの理由は、文脈によって同じ発話様式でも異なった発話行為が表現可能なので、その表現性の曖昧さを調べることである。ここでは 2.1 で収録した音声データから「え」または「へ」の部分のみを切り出した発話を聞いて、どの発話行為が知覚されるのかを付与する。

切り出された 207 発話をランダムに並べ替え、訓練無されていない被験者 4 名が各発話を聞いて、（文脈無しで）その発話のみから知覚される発話行為の項目を 2.1 で紹介した発話行為リストから選択した。ただし、文脈無しではリストの中から 1 個だけの発話行為が定まらない場合もあると考えられるし、リストの発話行為の複数の項目にも当てはまる場合も考えられるので、複数の項目を選択可能とした。4 名のうち、3 名以上が一致したものを発話行為の知覚データとして扱う。表 1 に発声時に意図した発話行為（１番目の列）と、知覚された発話行為との一致（2 番目の列）及び不一致（3 番目の列）の結果をまとめる。

Table 1 Matches, mismatches and ambiguities between intended and perceived speech act (SA) items.

Total number of intended SA	N. of matches	Number of mismatches or ambiguities
肯定(12)	(12)	同意(12) 相槌(12)
同意(9)	(9)	肯定(9) 相槌(9)
相槌(12)	(8)	肯定(6) 同意(7)
聞返し(12)	(11)	意外(1) 驚き(1)
感心(12)	(10)	羨望(3) 驚き(2) 意外(1)
驚き(12)	(10)	意外(6) 非難(1)
考え中(10)	(8)	戸惑(1) 不満(1) 嫌悪(1)
嫌悪(12)	(8)	非難(6) 不満(2) 疑い(1)
不満(12)	(7)	非難(5) 疑い(4) 嫌悪(2)
羨望(12)	(5)	不満(3) 意外(3) 驚き(2)
非難(12)	(5)	嫌悪(3) 疑い(2) 驚き(2) 意外(2)
疑い(12)	(4)	不満(5) 非難(4) 驚き(2)
意外(12)	(4)	驚き(4) 聞返(2) 疑い(2)
戸惑い(12)	(2)	考え中(5) 不満(6)
同情(12)	(2)	不満(4) 感心(3) 意外(2)
力んだ「え」(7)	-	嫌悪 (5) 考え中(2)
力んだ「へ」(5)	-	感心 (5)

先ず、意図して発声した発話行為がどの程度聞き手に正しく伝わったのかを示す 2 番目の列に注目してみると、*肯定*、*同意*、*相槌*、*聞返し*、*感心*、*驚き*、

*考え中*は文脈なしでも正しく伝わっており、*嫌悪*と*不満*はある程度伝わっているといえる。しかし、*戸惑い*、*同情*、*意外*、*非難*、*羨望*においては、発話の多数が他の発話行為として知覚された。これらの項目の不一致あるいは曖昧さを 3 番目の列で見ると、*戸惑い*の多くは*考え中*、*不満*と知覚され、*意外*の多くは*驚き*と知覚された。意外だと感じた場合、驚いてしまうという状況は十分あり得るので、この二つの項目が同時に現れることは十分考えられる。また、戸惑いながら考える、不満を感じて戸惑うというのもあり得る。*同情*の場合は*不満*、*感心*、*意外*など、異なった種類の項目との不一致が多く、文脈無しで「え」の発話様式のみからこれらの項目を認識することは難しいと考えられる。*羨望*の場合は、*不満*、*意外・驚き*と知覚され、これらも種類が異なるので、後続の発話（つまり、文脈）によって発話行為が明確になるものと考えられる。

本研究では、文脈無しの発話「え」のみからどの程度発話行為が認識できるのかという問題を重視する。従って、3 節の音響分析のターゲットとして、意図された発話行為の分類ではなく、知覚された発話行為による分類を用いる。例えば、不満を意図して発声されなかった発話も、3 名以上が不満と知覚したものは、不満のグループに入れることとする。

各発話行為が知覚された発話数を図 2 に示す。被験者間の判定の違い、または複数選択を許した結果が発話行為間の重なりとして表される。知覚の面で曖昧な分類が得られた 50 発話を除いた 157 発話をこれ以降の評価対象とする。

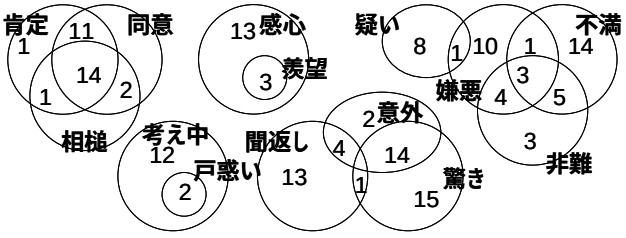


Fig. 2 Grouping of the speech act items according to the perceptual data results.

2.3 声質の知覚データと発話行為との関係

発話様式（speaking style）は、韻律特徴及び声質特徴の組み合わせで表現することを提案する。ここでは声質特徴の知覚データを付与し、声質と発話行為との関係を調べることと、声質に関連する音響パラメータを評価することを目的としている。

声質は知覚的には明確な分類が難しいので、ここでは声質の分類に経験のある被験者 1 名（著者本人）が音声を聴取し、波形やスペクトログラムを見ながら付与したものを扱う。

声質情報は、modal (***m***, 地声), whispery (***w***, 気息性のある声), aspirated (***a***, 発話末に現れる強い息盛れ), creaky (***c***, 非常に低くパルス的な声), harsh (***h***, 雑音的で耳障りのある声), pressed (***p***, 喉頭を力んだ声) 及び、これらの声質の組み合わせ（***hw***, ***pc*** など）によって分類した。

図 2 に示したように、知覚によって分類された発話行為の項目と、ここで知覚された声質との関係を表 2 にまとめる。

Table 2 Number of utterances of the perceived voice qualities, for each perceived speech act group.

	m	***w, a***	***hw, h***	***pc, p***	***c***
肯定･同意･相槌	23	6			
考え中･戸惑い	9	2			3
感心･羨望	10	2		4	
聞返し	12	1			
驚き･意外	14	12	10		
疑い	1	5	3		
嫌悪	2	1	4	4	
非難･嫌悪･不満	4	4	5		
不満	12	2			

表 2 の結果から、比較的強い non-modal な声質（***h, hw, a, w, pc***）が知覚された発話は比較的強い感情や態度を表現する発話行為（*驚き・意外*、*疑い*、*嫌悪・非難*、*関心・羨望*）に現れることが導ける。気息性（***w***）にかんしては、*肯定・同意・相槌*でも多少（6 発話）知覚されたが、これは丁寧さを表現するために生じたものと考えられる[12]。これらの結果はパラ言語情報の抽出における声質情報の重要性を示している。

3 音響パラメータによる発話行為の識別

前節では発話行為の項目と声質の関係を知覚の観点から調べた。本節では、さまざまな発話様式を表現するための韻律及び声質に関連する音響パラメータを紹介し、知覚された発話行為の識別性を調べる。

3.1 韻律に関連する音響パラメータ

韻律特徴の基本パラメータとなる F0 の抽出は、LPC 逆フィルタによる残差波形の自己相関関数の最大ピークに基づいた処理を行っているが、特に non-modal な区間では誤った値が抽出しやすいので、これらの誤りが後続処理への悪影響を防ぐために、自己相関関数で F0 の sub-harmonic に対応するピークも、ある閾値を満たさなければならないという制約を考慮した[8]。

韻律パラメータとして、[7]で提案した *F0move* と発話の持続時間を用いる。*F0move* は、ピッチ知覚を考慮し、音節内のピッチの動き（方向と度合い）を

semitone 単位で表すパラメータである。具体的には音節を 2 等分し、各区間において代表的な F0 の値を抽出し、これらの差分をとったものである。[7]では、各区間の代表的な F0 としてさまざまな候補が評価され、ここでは、ピッチ知覚に最も対応した前半区間の平均値（*F0avg2a*）と後半区間のターゲット値（*F0tgt2b*）を用い、*F0move* ＝ *F0tgt2b* – *F0avg2a* を扱う。F0 抽出法や F0 のターゲット値の具体的な求め方については、[7]をご参照ください。

持続時間に関しては、発話「え」は単音節なので、人手によって区切られた情報をそのまま使うことも可能だが、発話前後に無音区間が多少入ってしまう場合があるので、母音区間のみを抽出するためにパワー情報を利用し、発話前後のパワーが発話の最大パワーより 20 dB 落ちている位置まで、境界を自動的に矯正した。これによって得られた境界を用いて発話の持続時間（*duration*）を測定する。

図 3 は韻律パラメータ（*F0move* vs. *duration*）による発話行為の分布を示す。

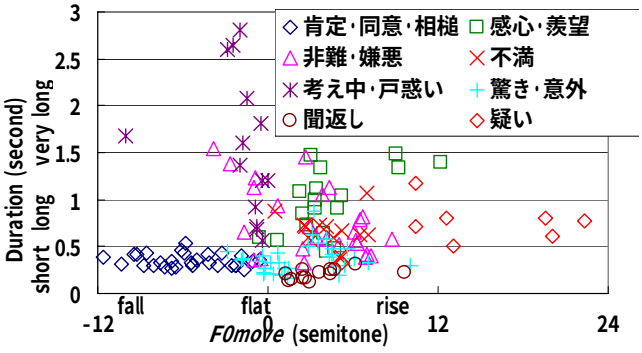


Fig. 3 Distributions of the prosodic parameters for each perceived speech act group.

図からは韻律特徴は、*肯定・同意・相槌*（短下降型）、*聞返し*（短上昇調）、*疑い*（動きの幅が広い上昇調）、*考え中・戸惑い*などフィラー的な曖昧な表現（平坦、長下降調）、それ以外の否定的または曖昧な表現（長上昇調、長平坦調）のように、主に機能的な項目を識別するのに有効であることが導ける。長上昇調ではさまざまな項目（*非難・嫌悪*、*感心・羨望*、*不満*、*驚き･意外*）が混合しており、韻律特徴のみでの識別は難しい。また、短上昇調の中でも、*聞返し*と*驚き･意外*の識別は明確ではない。この結果から、韻律特徴のみでの発話行為の識別には限界があることを示している。

3.2 声質に関連する音響パラメータ

本節では、声質に関連する音響パラメータを 3.2.1～3.2.3 で紹介し、韻律特徴のみでは識別できない発話行為の項目を声質特徴にてどの程度識別出来るのかを 3.2.4 で示す。

3.2.1 Vocal Fry (creaky)区間の検出

ここでは、最近提案した Vocal Fry (creaky) 区間検出アルゴリズム[8]を使用する。Vocal Fry のパルス性と非常に低い基本周波数（パルス・レート）の特徴を反映するため、“very short-term”（フレーム長 5 ms を 2.5 ms ごとに求めた）パワー軌道からパワー･ピークを声帯パルスの候補として検出し、隣り合うピークの周期性と類似性の制約をチェックし、Vocal Fry による声帯パルスであるかを判断するというアルゴリズムである（図 4 参照）。アルゴリズムは主に 3 つのパラメータに基づいている：パワー･ピークを検出するためのパワー(*PPw*: Peak Power)の閾値、自己相関関数に基づいたフレーム内の周期性（*IFP*: Intra-Frame Periodicity）の閾値、ピーク周辺の波形の相互相関に基づいたパルス間の類似性（*IPS*: Inter-Pulse Similarity）の閾値。アルゴリズムとパラメータの詳細については[8]をご参照ください。本研究では、*PPw* > 7 dB, *IFP* < 0.8, *IPS* > 0.6 の条件を設定した。

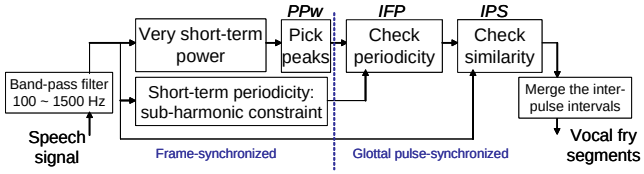


Fig. 4 Simplified block diagram of the vocal fry detection.

3.2.2 非周期・ダブル周期(aperiodicity; double-periodicity)区間の検出

Vocal fry 及び Harsh 発声は、声帯振動の周期性が不規則になる特徴を持っている。これらの不規則性は、声帯パルスの非周期性またはダブル周期性として現れる。ここでは、[9]で提案した非周期・ダブル周期に関連するパラメータを使用する。これらのパラメータは本来 Creaky (Vocal Fry) 区間を検出するために提案したものであるが、Harsh 発声による非周期性・ダブル周期性も反映されることが観察されたので、3.2.1 で Vocal Fry 区間として検出されなかった区間を Harsh とみなすことを試みる。

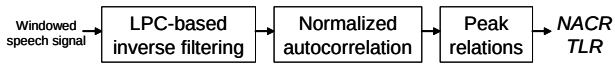


Fig. 5 Simplified block diagram of the parameters for aperiodicity/double-periodicity detection.

図 5 に簡単なブロック図を示す。パラメータは入力音声信号に声道の逆フィルタをかけて求めた音源波形の正規化自己相関関数の最初の 2 ピークの関係を表現する。一つ目のパラメータは、*NACR* (Normalized Auto-Correlation Ratio)と呼び、最初の 2

ピークの正規化自己相関値の比率である。もう一つのパラメータは、*TLR* (Time-Lag Ratio)と呼び、ピーク位置の比率を 2 で欠けたものである。 $NACR > 1$ または $TLR \neq 1$ の場合、ダブル周期性または非周期性を表す。パラメータの詳細に関しては[9]を御参照ください。

3.2.3 息漏れ（気息性のある：aspiration noise）区間の検出

息漏れ雑音（気息音）とは、“breathy voice”や“whispery voice”で起きる、声門での十分な狭めによって生成される気流雑音（“turbulent noise”）のことを差す。生成の面では、breathy と whispery は区別されるが[6]、音響的にも知覚的にも、カテゴリカルな分類は難しい[13]。また、気息音は harsh 発声と共に現れる場合もある（harsh whispery voice [6]）。

気息性を検出する手法として、[10]で提案したものを使用する。手法は２つのパラメータによって息漏れ検出を行う。主なパラメータは *F1F3syn* (F1 and F3 band synchronization)と呼び、第 1 と第 3 フォルマント(F1, F3)周辺の周波数帯域でフィルタリングした信号の同期性を定量化したものである。同期率は F1 と F3 帯域の波形振幅の相互相関関数によって求める（図 6 参照）。気息性がない場合、*F1F3syn* は 1 に近づき、気息性がある場合は 0 に近づく。二つ目のパラメータは *A1-A3* と呼び、F1 と F3 帯域のパワーの差を表し、*F1F3syn* の有効性を制限する役割を持つ。*A1-A3* が比較的大きい場合（つまり F3 帯域のパワーが F1 帯域のパワーと比較的弱い場合）F3 帯域の雑音は知覚されていない可能性があり、同期率を図る意味がないということを示す。

F1 帯域は 100 ~ 1500 Hz、F3 帯域は 1800 ~ 4500 Hz に固定する。本手法の詳細に関しては、[10]を参照してください。ここでは $F1F3syn < 0.4$ 及び $A1\text{-}A3 < 25$ dB の条件でフレームごとに気息音を検出する。

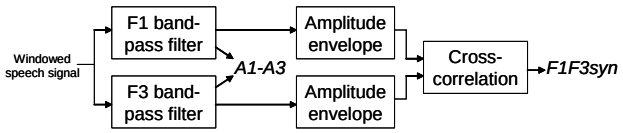


Fig. 6 Simplified block diagram of the parameters for aspiration noise detection.

3.2.4 声質パラメータによる発話行為の識別

以上のパラメータより、フレームごと、あるいは区間ごとの情報が得られるが、以下のものを発話ごとのパラメータとして提案する。

- Vocal fry rate (***VFR***)：発話全体に対し、Vocal fry (creaky)が検出された区間の割合。
- Aperiodicity rate (***APR***)：発話全体に対し、非周期またはダブル周期が検出され、Vocal Fry とは検出されなかった区間の割合。
- Aspiration noise rate (***ANR***)：発話全体に対し、息漏れ雑音が検出された区間の割合。

声質の認識の予備的な実験結果より、これらの発話レベルの声質パラメータの閾値を 0.1 と設定する。$VFR > 0.1$ の発話は ***c*** (creaky)、$APR > 0.1$ の発話は ***h*** (harsh)、$ANR > 0.1$ の発話は ***w*** (whispery)、それ以外のものは ***m*** (modal)と認識した結果を発話行為ごとに分配して表 3 にまとめる。

Table 3 Number of utterances of the detected voice qualities, for each perceived speech act group.

	m	***w***	***hw, h***	***c***
肯定･同意･相槌	24	4		1
考え中･戸惑い	11	1		2
感心･羨望	11	2		3
聞返し	12	1		
驚き･意外	18	12	6	
疑い	2	5	2	
嫌悪	2	1	1	4
非難･嫌悪･不満	6	5	2	
不満	12	2		

表 3 に示した結果より、強い気息性及び強い非周期性を持つ声質(***w, hw, h***)は、*驚き*、*意外*、*嫌悪*、*疑い*など、比較的強い感情や態度を表す項目を検出するのに有効といえる。この結果は、表 2 に示した声質の知覚データとの結果とある程度一致しているが、***h, hw*** の検出が不十分であることが分かる。これは、harsh 声質を正しく検出するためには、3.2.2 で提案した手法が不十分であることを示しており、改善が必要である。***c*** に関しては、強い感情を表す*感心*と*嫌悪*では ***pc***（力んだ creaky）であり、*考え中・肯定*で現れた ***c*** は柔らかい creaky であることを確認した。力みを検出するため、更なる音響特徴が必要である。また、「へ」と発声されたものは、*感心・羨望*として知覚される傾向が観られ、音韻情報も発話行為を識別するのに重要と示している。

また、韻律パラメータに関しても、F0 抽出には注意したが、主に harsh と creaky の区間で、F0 の抽出誤りが *F0move* に反映されてしまう発話が*嫌悪・非難*で少数現れたので、F0 抽出には更なる注意が必要である。

4 おわりに

さまざまな発話様式で発声された発話「え」を分析した結果、韻律特徴は肯定的な表現、聞返し、フィラー、否定的な表現など、機能的な発話行為を識

別するのに有効である一方、声質特徴（強い気息性、または強い非周期性を含んだ声）は驚き、嫌悪、疑い、感心など、比較的強い感情や態度を表す発話行為を認識することに有効と示した。今後、評価用のデータを増やし、主に声質に関連する音響特徴を改善し、韻律特徴との適切な組み合わせで発話行為の識別を評価する予定である。

謝辞

本研究は総務省の研究委託により実施したものである。アドバイスもしくは機材のサポートにご協力いただいた、榊原健一氏、パーハムモクタリ氏、北村達也氏、IRC の皆様に感謝する。音声収録及び知覚実験に協力いただいた皆様に感謝する。

参考文献

1) Erickson, D., “Expressive speech: production, perception and application to speech synthesis,” *Acoust. Sci. & Tech*., Vol. 26 (4), 317-325, 2005.

2) Maekawa, K., “Production and perception of ‘Paralinguistic’ information,” Proc. *Speech Prosody 2004*, 367-374, 2004.

3) Klasmeyer, G., Sendlmeier, W. F., Voice and Emotional States. In *Voice Quality Measurement*, Singular Thomson Learning. Ch. 15, pp. 339-358, 2000.

4) Gobl, C., Ní Chasaide, A., The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, pp. 189-212, 2003.

5) Hess, W., “Pitch Determination of Speech Signals”, Vol. 3 of *Springer Series of Information Sciences*, Springer-Verlag, Berlin, Heidelberg, New York, 1983.

6) Laver, J., Phonatory settings. In *The phonetic description of voice quality*. Cambridge University Press, Ch. 3, pp. 93-135, 1980.

7) Ishi, C.T., Mokhtari, P., Campbell, N., “Perceptually- related Acoustic-Prosodic Features of Phrase Finals in Spontaneous Speech,” Proc. *Eurospeech 2003*, 405-408, 2003.

8) Ishi, C.T., Ishiguro, H., Hagita, N., “Proposal of Acoustic Measures for Automatic Detection of Vocal Fry,” Proc. *Eurospeech 2005*, 481-484, 2005.

9) Ishi, C.T., “Analysis of Autocorrelation-based parameters for Creaky Voice Detection,” Proc. *Speech Prosody*: 643-646, 2004.

10) Ishi, C.T., “A New Acoustic Measure for Aspiration Noise Detection,” Proc. *ICSLP 2004*, Vol. II, 941-944, 2004.

11) Sadanobu, T., “A Natural History of Japanese Pressed Voice”, *J. of Phonetic Society of Japan*, Vol. 8 (1): 29-44, 2004.

12) Ito, M., “Politeness and voice quality – The alternative method to measure aspiration noise,” Proc. *Speech Prosody 2004*, 213-216, 2004

13) Kreiman, J., Gerratt, B., Measuring Vocal Quality, In *Voice Quality Measurement*, Singular Thomson Learning. Ch. 7, pp. 73-102, 2000.

付録：発話行為の音声収集に用いた台本

A：今日は雨かな？
B：（肯定　　）、雨だよ。
A：韓国料理は好き？
B：（肯定　　）、好きだよ。
A：今日は雨やね。
B：（同意　　）、そうやね。
A：お昼、ファミレス行こうか。
B：（同意　　）、行こう行こう。
A：　今日は雨みたい。
B：（相槌　　）、そうやね。
A：今日、また電車遅れてるみたいよ。
B：（相槌　　）、そうやってね。
A：今日は雨やし、バーベキュー中止しよっか？
B：（戸惑い　　）、どうしよう。
A：体の調子が悪いから、今日の予定はやめとこか？
B：（戸惑い　　）、じゃーどうしようかー。
A：今日は rainy だよ。
B：（聞き返し　　）？なんて？
A：明日の朝、7時に出発するよ。
B：（聞き返し　　）？　何時って？
A：今日は夕食の準備しておいてね。
B：（不満　　）、なんでよ。
A：この仕事、頼むで。
B：（不満　　）、なんで。
A：私の趣味は草刈だよ。
B：（意外　　）、うそ！
A：私、格闘技見るの好きやねん。
B：（意外　　）、そうなんや！
A：私はブッシュ大統領を支持するよ。
B：（非難　　）、なんでまたー
A：私、蛇飼ってるんねん。
B：（非難　　）、なんで蛇なん！？
A：私はゴキブリが好きだよ。
B：（嫌悪　　）、キモー！
A：満員電車が好きやねん。
B：（嫌悪　　）、どこがいいん？。
A：　今日から１ヶ月間、海外旅行へ行ってきまーす！
B：（羨望　　）、いいなー。
A：このネックレス、昨日彼氏が買ってくれてん。
B：（羨望　　）、ええなー。
A：ロボビーは完璧にしゃべれるようになったよ！
B：（感心　　）、すごいなー！
A：あの人はどんな曲でもピアノで演奏できるんだって。
B：（感心　　）、すごいなー！
A：ロボビーは完璧にしゃべれるようになったよ！
B：（疑い　　）、ありえへん！
A：私、ポルトガル語、ペラペラやねん。
B：（疑い　　）、うそや～。
A：今日抽選で当たりました。
B：（驚き　　）、すごい！
A：昨日空港で中島みゆきに会ってん！
B：（驚き　）、ほんまに？
A：もう３日も寝ないで仕事してるんだよ。
B：（同情　　）、大変やんなー。。
A：階段から落ちて、骨折してん。
B：（同情　　）、かわいそうやな。
A：１２８＋６３はいくつ？
B：（考え中　　）、。。。
A：３３０を１１で割ると？
B：（考え中　　）、。。。

大規模マイクロホンアレイによる室内移動音源の追跡と方向推定

Sound Source Tracking with Orientation Estimation by Using A Large Scale Microphone Array

中臺 一博 † **中島 弘史** ‡ **山田 健太郎** † **長谷川 雄二** † **中村 孝広** † **辻野 広司** †

Kazuhiro Nakadai†**, Hirofumi Nakajima**‡**, Kentaro Yamada**†**, Yuji Hasegawa**†**, Takahiro Nakamura**†**, Hiroshi Tsujino**†

†(株) ホンダ・リサーチ・インスティチュート・ジャパン ‡ 日東紡音響エンジニアリング (株)

†**Honda Research Institute Japan Co., Ltd.** ‡**Nittobo Acoustic Engineering Co., Ltd.**

{nakadai,yamaken,yuji.hasegawa,moo,tsujino}@jp.honda-ri.com, nakajima@noe.co.jp

Abstract

This paper addresses sound source tracking with orientation estimation by using a 64 ch microphone array. The microphone array system localizes a sound source and estimates its directivity pattern based on an weighted delay-and-sum beamforming method. The directivity pattern estimation has two advantages such as detection of actual human voice by comparing the estimated directivity pattern with pre-recorded ones, and estimation of sound orientation by detecting the angle with the highest power in the directivity pattern. The preliminary results show the effectiveness of the method in sound tracking and in orientation estimation.

1 はじめに

自然な人・ロボットコミュニケーションを実現する一つの要素研究として、「ロボット聴覚」が挙げられよう。ロボット聴覚は、実環境で実時間聴覚処理を実現することを目的とし、これまで、雑音下、残響下、かつ音響環境が動的に変化する中で、ロボットに設置した２本のマイクを用いた音源定位、分離、分離音の認識[22]等が報告されている。しかし、従来のロボット聴覚には、以下の２点の制約があった。

1. 水平方向に置かれた２本のマイクでは、理論的に音源の水平角しか推定できない。
2. 音源定位、分離、音声認識のパフォーマンスは話者とマイクが離れるにつれ悪化する。

一つ目の制約は、「人間とロボットの距離をロボットの聴覚処理からは得ることができない」ということを意味している。しかし、近接学[6]では、コミュニケーションにおいて、対人距離は、感情やコミュニケーションの方法に影響を与えることが報告されている。つまり、距離情報は人・ロボットコミュニケーションにおいても重要な要素であると考えられ、ロボットは対人距離に応じてコミュニケーションの方法を変更する必要があるといえる。このため、音源方向だけではなく、音源までの距離が推定できる手法が望まれる。

二つ目の制約は、主に部屋の雑音や反響に起因している。聴覚と動作を統合するアクティブ・オーディション[13] は、音源に近づくように行動することによってこの問題を解決するアプローチである。一方で、マイクから離れた人の声を認識できるような雑音に頑健なシステムが構築できるとすれば、認識の向上という意味において音源に向かって移動する必要はなくなり、ロボットの省エネ化や処理の高速化という点で効果的である。

また、音源定位・分離・認識以外にも、日常環境で使用されるロボットが備えるべき有用な機能があろう。例えば、正確な音源同定が可能であれば、ロボットは、音声と、非音声信号を区別したり、同じ音声であっても、実際にその場で人間が発声した音声（以後、肉声とする）なのか、テレビやラジオなどからの音声であるのかを区別したりして、状況やユーザの声を的確に認識することができよう。しかし、現状では、こうした機能の実現は難しく、例えば、多くの音声認識システムでは、入力音が常に音声信号であることを仮定している。この仮定は、必ずしも話者の口元にマイクを設置できないような実環境では強い制約となってしまう。この問題に対し、音オントロジー[15]は、様々な種類の音を扱う枠組みを提案しているが、現状では問題解決への道のりは遠い。話者同定技術を用いれば、音声と非音声を区別することはある程度可能となるが、肉声とテレビ・ラジオの音声を区別することは難しい。顔検出や唇検出といった視覚処理[8] もこの問題に対する有効な解決法となり得るが、顔検出では、テレビや写真の顔を誤検出してしまう、唇検出では、カメラの解像度の問題で、検出できる条件が限定されてしまうといったように、完全な解決を図ることは難しい。従って、肉声であるかどうかを、音響的な特徴から推定できれば、ロボットにとって有効な情報となる。

1.1 課題解決のアプローチ

これらの問題を扱う際に、以下の３つのアプローチが考えられる。

1. 精度よく音声信号を取得するため、ロボットへマイクロホンアレイを適用する。
2. 曖昧な聴覚情報を補完するため、視覚や動作など他のモダリティを併用する。

3. 人とロボットの距離に依存しない聴覚処理を行うため、環境へ設置したマイクロホンアレイを用いる。

最初のアプローチは、近年、移動アレイを用いた音響信号処理の研究課題として注目を浴びている。Valin らは、8 ch のマイクロホンアレイをロボットに搭載し、 GSS に基づく実時間音源定位・分離を報告している[20]。原らは、8 ch のマイクロホンアレイをヒューマノイドに搭載し、ヒューマノイドが動作している場合の 音楽と音声の混合音からの音響ストリーム抽出を報告している [7]。

二つ目のアプローチは、情報統合によってロバスト性を向上させる考え方である。実環境では、すべてのセンサ情報は曖昧性を含んでいるため、実環境アプリケーションを考える上で、このアプローチは本質的であるといえる。実際、我々は、これまでに、視覚、聴覚、動作を統合し、三話者同時認識を報告することによって、情報統合の有効性を示した[14]。また、原らは、パーティクルフィルタを用いた視聴覚統合による音源追跡を報告している[7]。

最後のアプローチは、マイクロホンアレイが静止していることを仮定している。この種のマイクロホンアレイは信号処理の分野で研究が進んでおり、遅延和ビームフォーミング [4], 適応ビームフォーミング[5, 10]、独立成分分析 (ICA) [9, 17]など様々な手法が提案されている。

最初の 2 つのアプローチは聴覚処理の向上という点で本質的であるが、音源がマイクから離れた場合の問題に対する解を提供していない。従って、本稿では、三つ目のアプローチに基づき屋内に設置したマイクロホンアレイを扱う。過去の研究では、環境設置型の大規模のマイクロホンアレイを扱った報告も複数見られる[21, 18, 1]が、こうした研究では、音源定位・分離のみを扱っていた。本稿では、信号処理アルゴリズムを拡張し、音源の指向特性推定を可能にすることによって、音源定位・分離だけでなく、音源向きの推定、および、肉声検出への応用を報告する。

2 マイクロホンアレイのアルゴリズム

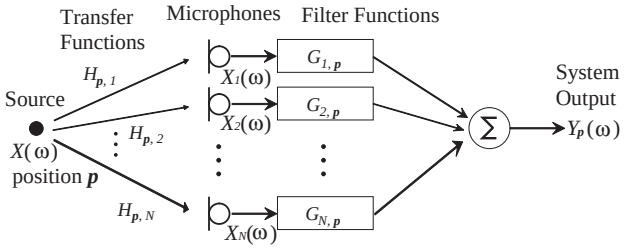


Figure 1: Weighted Delay-and-Sum Beamforming

本稿では、独立成分分析や適応ビームフォーミングといった手法と比べ、比較的、計算量が少ない重み付き遅延和ビームフォーミング (WDS-BF) を用いる。WDS-BF に対して以下の 2 点の拡張を行った。一点目は、従来の遅延和ビームフォーミングでは、一般に線形や円形などマイク間の時間差が計算しやすいような形状のアレイを用いているのに対し、自由にマイクのレイアウトを可能にした。二点目は、音源の定位だけでなく、指向特性推定を可能にした。これにより、指向特性のピークを追跡すれば、音源位置だけでなく音源向きの追跡が可能になったり、実際の人間の声の指向特性を予め DB として蓄えておけば、テレビやラジオなどスピーカから出力された声と区別したりといった応用が期待できる。

関連した研究は、筆者らの知りうる限りでは、Meuse らの研究が挙げられる[11]。彼らは、スピーカの放射パターンモデルを用いて、スピーカの大きさや向きのパラメトリック推定を報告しているが、放射パターンモデルを仮定しているため、この手法を直接、指向特性推定や肉声検出へ適用することは難しい。

2.1 重み付き遅延和ビームフォーミング（WDS-BF）

図 1 に WDS-BF の構成図を示す。この時、システム出力スペクトルは下記のように定義できる。

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

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

ここで、$X(\omega)$ は、$\boldsymbol{p} = (x, y)$ に位置する音源 S のスペクトル、 $H_{\boldsymbol{p},n}(\omega)$ は S から n 番目のマイクへの伝達関数、$X_n(\omega)$ は、n 番目のマイクで収音された信号のスペクトル、$G_{n,\boldsymbol{p}}(\omega)$ は n 番目のマイクで収音された信号から $\boldsymbol{p}$ における信号スペクトルを推定するフィルタ関数を示す。

通常の遅延和ビームフォーミング (DS-BF) では、$G_{n,\boldsymbol{p}}(\omega)$ は、マイクの位置関係を利用して、音源位置でのシステム出力が、できるだけ正確に音源信号に近づくように計算することによって求められる。従って、マイクのレイアウトは単純な方が好まれる傾向にある。

本稿で扱う WDS-BF では、音源方向推定を行うため、音源方向パラメータ θ を導入する。これは、式 (1), (2) において、位置ベクトル $\boldsymbol{p} = (x, y)$ を、$\boldsymbol{p'} = (x, y, \theta)$ と置き換えることにより実現する。

文献[12] に記述されているように、自由音場での点音源を仮定した場合、$H_{\boldsymbol{p'},n}(\omega)$ は、以下のように定義できる。

$$H_{\boldsymbol{p'},n}(\omega) = A(\theta)\frac{v}{r\omega} e^{\frac{ir\omega}{v}} \tag{3}$$

$$r = \sqrt{(x_n - x)^2 + (y_n - y)^2} \tag{4}$$

ここで x_n, y_n は、n 番目のマイクロホンの x, y 座標、v は音速、$A(\theta)$ は、音源 S の指向特性を示す。

この時、システムゲイン D は以下のように定義できる。

$$D(\boldsymbol{p'}, \boldsymbol{p'_s}) = \frac{Y_{\boldsymbol{p'}}(\omega)}{X_{\boldsymbol{p'_s}}(\omega)} = \sum_{n=1}^{N} G_{n,\boldsymbol{p'}}(\omega) H_{\boldsymbol{p'_s},n}(\omega) \tag{5}$$

ここで、N はマイク数を示す。

次に、x, y, θ を $x_1, \cdots, x_p, \cdots x_P, y_1 \cdots, y_q, \cdots y_Q$, and $\theta_1, \cdots, \theta_r, \cdots, \theta_R$ と離散化し、位置インデックスを $m = (p + qP)Q + r$ と定義すると、式 (5) は、下記の行列演算として定義できる。

$$\begin{aligned}
\boldsymbol{D} &= \boldsymbol{H}\boldsymbol{G}, \\
\boldsymbol{D} &= [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_m, \cdots, \boldsymbol{d}_M]^T, \\
\boldsymbol{d}_m &= [D_{m,1}, \cdots, D_{m,k}, \cdots, D_{m,M}], \\
\boldsymbol{G} &= [\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n, \cdots, \boldsymbol{g}_N]^T, \\
\boldsymbol{g}_n &= [G_{n,1}, \cdots, G_{n,k}, \cdots, G_{n,M}], \\
\boldsymbol{H} &= [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_m, \cdots, \boldsymbol{h}_M]^T, \\
\boldsymbol{h}_m &= [H_{m,1}, \cdots, H_{m,k}, \cdots, H_{m,N}].
\end{aligned} \tag{6}$$

ここで $\boldsymbol{D}, \boldsymbol{H}, \boldsymbol{G}$ は、それぞれ、指向ゲイン行列、伝達関数行列、フィルタ行列を示す。

最小ノルム重みに基づいた WDS-BF を構築する場合は、位置インデックス m に対する フィルタ行列 $\boldsymbol{G}$ は、近似フィルタ行列 $\hat{\boldsymbol{G}}$ として、下記の式により計算できる。

$$\hat{\boldsymbol{G}} = [\boldsymbol{H}]^{+}\boldsymbol{D} = \frac{\boldsymbol{H}^{H}}{|\boldsymbol{H}|^{2}}\boldsymbol{D} \tag{7}$$

ここで $\boldsymbol{H}^{+}$, $\boldsymbol{H}^{H}$ は、それぞれ $\boldsymbol{H}$ の擬似逆行列とエルミート転置行列を示す。

音源の指向特性を推定するために、$H_{m,k}$ に対する $D_{m,k}$ および $A(\theta_r)$ を下記のように定義する。

$$D_{m,k} = \begin{cases} 1 & if \quad \theta_r = \theta_a \\ 0 & otherwise \end{cases} \tag{8}$$

$$A(\theta_r) = \begin{cases} 1 & if \quad |\theta_r - \theta_a| \leq \theta_s \\ 0 & otherwise \end{cases} \tag{9}$$

ここで θ_a は、目的音源方向、θ_s は 閾値パラメータ ($180/R$ 度) で、例えば、8 方向を推定する場合は、$22.5°$ となる。

2.2 音源定位と指向特性推定

WDS-BF を用いた音源定位のアルゴリズムは以下の通りである。

1. サンプリングレート 16 kHz で、同時に 64 チャネル収音を行う。
2. 各チャネルの信号に対し、1,024 点 FFT による周波数解析を行い、パワースペクトル $Sp_n(\omega)$ に変換する。
3. 背景雑音より、20 dB 以上パワーの大きいサブバンドを抽出し、$(\omega_1, \cdots, \omega_l, \cdots, \omega_L)$ とする。信号信頼度 $\mathrm{SCR}(\omega_l)$ を下記のように定義する。

$$\mathrm{SCR}(\omega_l) = \frac{Sp(\omega_l) - N_o(\omega_l)}{Sp(\omega_l)} \tag{10}$$

$$Sp(\omega_l) = \frac{1}{N}\sum_{n=1}^{N} Sp_n(\omega) \tag{11}$$

ここで N_o は、背景雑音の平均スペクトルを示す。

4. $\boldsymbol{p'_m}(x_p, y_q, \theta_r)$ におけるフィルタ出力 $Y_{\boldsymbol{p'_m}}(\omega_l)$ を式 (1) より、方向別スペクトル強度 $I(\boldsymbol{p'_m})$ を、下式より、計算する。

$$I(\boldsymbol{p'_m}) = \sum_{l=1}^{L} \mathrm{SCR}(\omega_l) \left| Y_{\boldsymbol{p'_m}}(\omega_l) \right|^2 \tag{12}$$

5. 方向成分加算スペクトル強度 $I_s(x_p, y_q)$ を以下の式により計算する。

$$I_s(x_p, y_q) = \sum_{r=1}^{R} I(\boldsymbol{p'_m}) = \sum_{r=1}^{R} I(x_p, y_q, \theta_r). \tag{13}$$

6. 選択サブバンドの位置 (x_{p_a}, y_{q_a}) を下記の式により推定する。

$$(p_a, q_a) = \underset{p,q}{\mathrm{argmax}}\, I_s(x_p, y_q) \tag{14}$$

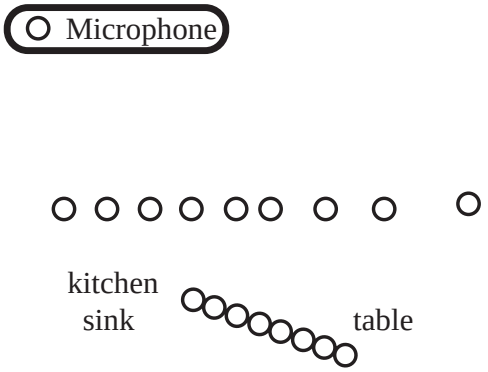


Figure 2: Actual Room with Microphone Array

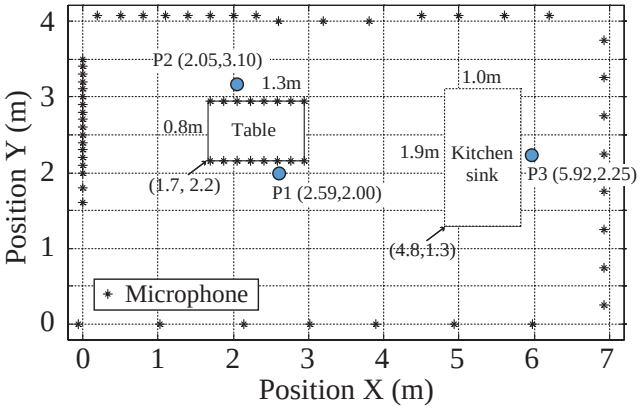


Figure 3: Layout of Microphones

7. (x_{p_a}, y_{q_a}) における指向特性 $\mathrm{DP}(\theta_r)$ を下式より計算する。

$$\mathrm{DP}(\theta_r) = \left\{ \frac{I(x_{p_a}, y_{q_a}, \theta_r)}{I_s(x_{p_a}, y_{q_a})} \;\middle|\; r = 1, \cdots, R \right\}. \tag{15}$$

音源方向 θ_{r_a} を計算する。

$$r_a = \underset{r}{\mathrm{argmax}}\, \mathrm{DP}(\theta_r) \tag{16}$$

8. (x_{p_a}, y_{q_a}) における推定信号に由来する n 番目のマイクロホンへの入力信号 $\mathrm{Xa}_n(\omega)$ を式 (2) より計算する。残差信号を下記により計算する。

$$X'_n(\omega) = X_n(\omega) - \mathrm{Xa}_n(\omega) \tag{17}$$

9. 3) から 8) までの処理を、音響信号が検出されなくなるまで繰り返す。

3 マイクロホンアレイシステム

前節で述べた WDS-BF に基づく音源定位を 構築した 64 ch マイクロホンアレイに実装した。マイクロホンアレイシステムを設置した部屋を図 2 に示す。部屋の大きさは 4.0 m × 7.0 m であり、内部に、キッチン、テーブルと 4 脚の椅子が設置されている。三方の壁は、吸音壁になっており、残りの一方は反響の大きいガラス壁となっている。

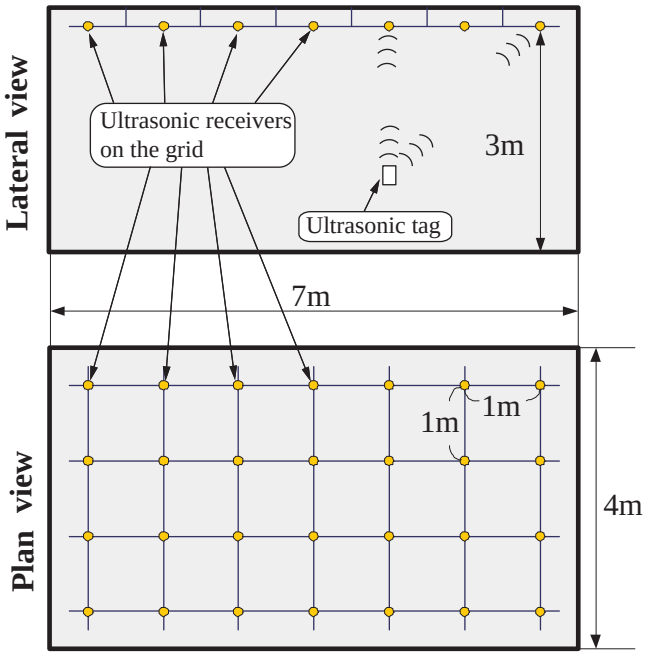


Figure 4: Layout of Ultrasonic Sensors

部屋の上から見た場合のマイクのレイアウトを図 3 に示す。アスタリスクがマイク位置を示している。マイクの高さは、壁に設置されているものは 1.2 m、テーブルに設置されているものは 0.7 m である。離散化の刻みは、位置が $25\,cm$、方向が $45°$ である。つまり、2 節で述べた、P, Q, R は、それぞれ、27, 15, 8 となる。従って位置インデックスの総数 M は 3,240 となる。マイクのレイアウトは、推定できる方向の数を最大にするように設計した。

4 評価実験

音源定位、指向特性推定、音源追跡の 3 点に関して評価を行った。

音源定位の評価については、単一音源、同時二音源の定位という 2 つの実験を行った。

単一音源では、スピーカ (GENELEC 1029A) から出力された白色雑音、スピーカから出力された録音音声、2 名 (A 氏, B 氏) の肉声の計 4 種類の音源を用いて、図 3 に示した P1, P2, P3 の 3 点で測定を行った。測定は、各点で 150 回ずつ行い、平均定位誤差とその標準偏差計算した。なお、フィルタベクトル $\boldsymbol{g}$ は、伝達関数 $\{\boldsymbol{h}_m | m = 1, \cdots, M\}$ の擬似逆行列から計算した。

同時二音源の定位では、 P2 と P3 にスピーカユニットが、 それぞれ 270°、180° を向くように設置した。ここで、方向は反時計回りで、(1,0) ベクトルの方向を 0° 方向と定義した。音源は、スピーカから出力される録音音声を用い、同時に 2 つのスピーカから出力した。

指向特性推定の評価は、スピーカから出力された白色雑音、スピーカから出力された録音音声、A 氏の肉声の計 3 種類の音源の指向特性推定実験を通じて行った。 音源位置は P1 で、音源方向は 180° とした。 3 つの実験、すべてにおいて、フィルタベクトル g を計算するための 伝達関数 $H_{\boldsymbol{p}',n}$ はスピーカから出力したインパルス応答測定結果から計算したものを用いた。

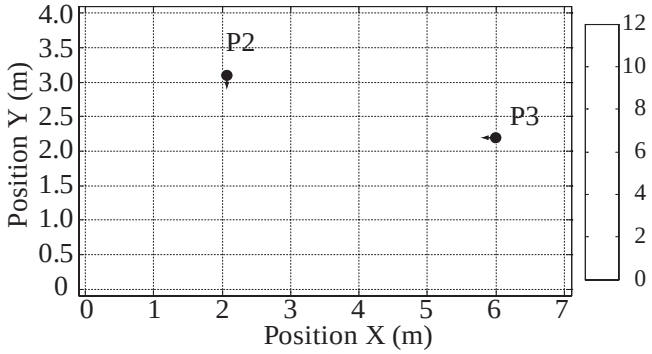


Figure 5: Localization of Two Simultaneous Speech Signals

音源追跡の評価を行うために、スピーカを P1 から P2 経由で P3 まで動かした。この際、移動軌跡のリファレンスデータを取得するため、超音波タグを用い、スピーカ移動時にデータを収録した。

式 (3) により得られる伝達関数 $H_{\boldsymbol{p}',n}$ から計算したフィルタベクトル g を用いて、20 ms 毎に、スピーカの位置と方向を推定した。

4.1 タグ位置測定システム

マイクロホンアレイが設置されている部屋には、産総研で開発された超音波三次元タグシステム (UTDTS) [16] が設置されている。

UTDTS は、大きく、単数もしくは複数の超音波タグと複数の超音波レシーバから構成されている。UTDTS は、タグの超音波出力時刻とレシーバへの入力時刻の差分を検出し、差分情報を三角測量と同様の手法で三次元情報に変換することにより、インドア GPS 機能を実現している。

図 4 に、設置した UTDTS を示す。 28 個の超音波レシーバがグリッド状に設置されている。位置と方向が計算できるように 3 個の超音波タグをスピーカに取り付けた。この構成では、部屋の中心では 1 – 8 cm 程度、壁の周辺で 6 – 13 cm 程度の誤差で定位が 可能である。

4.2 実験結果

単一音源の定位結果を表 1 に示す。誤差と標準偏差の単位はメートルである。同時二発話の定位のヒストグラムを図 5 に示す。

指向特性推定の結果を図 6 に示す。図の横軸と縦軸は音源方向（度）とそれに対するパワー比（%）となっている。細い実線は、無響室で計測した白色雑音に対して g を用いて計算した 1 kHz における指向特性を示している。破線は、[3] に報告されている人間の肉声の指向特性に対して g を用いて計算した 1 kHz における指向特性を示している。

太い実線は、本稿で述べた指向特性推定アルゴリズムを用いて推定した結果を示している。図 6a), b), c) はそれぞれ、スピーカから出力された白色雑音、スピーカから出力された録音音声、肉声に対する指向特性推定結果を示している。

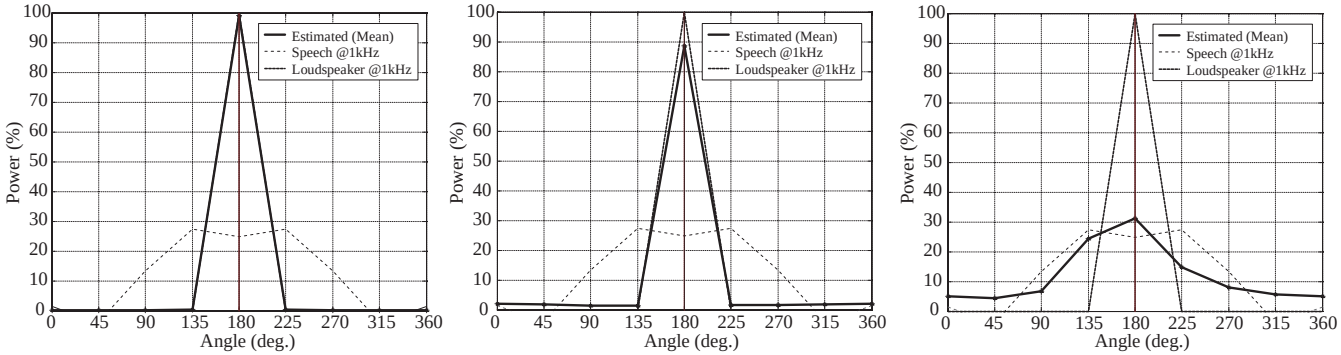
図 7a), b) に、スピーカの音源追跡結果を示す。太い実線は、音源の軌跡を示す。矢印は、各時刻でのスピーカの向きを示す。定位結果と UTDTS の観測結果から計算された位置と向きの平均誤差と標準偏差を表 2 に示す。

Table 1: Localization Error of A Single Sound Source (*m*)

Speaker device	Sound Source	P1		P2		P3	
		Avg.	S.D.	Avg.	S.D.	Avg.	S.D.
Loudspeaker	White noise	0.16	0.19	0.05	0.20	0.45	0.19
Loudspeaker	Recorded voice	0.15	0.71	0.40	0.39	1.80	1.80
Human	Mr. A	0.09	0.47	0.53	0.50	1.69	1.79
Human	Mr. B	0.04	0.57	0.36	0.53	1.52	1.64

Table 2: Tracking Error of a Moving Sound Source (white noise)

	Error	
	Avg.	S.D.
Localization (m)	0.24	0.19
Orientation (deg.)	9.8	94.3



a) White noise by a loudspeaker b) Human voice played by a loudspeaker c) Real human voice of Mr.A

Figure 6: Directivity Pattern Estimation of Three Kinds of Sound Sources

4.3 考察

音源定位の評価実験から、構築したマイクロホンアレイシステムは、反響を考慮せず、自由音場を仮定して計算したフィルタベクトルを用いた場合でも、部屋の中央で、定位誤差が 20 – 60 cm と比較的小さいことがわかる。しかし、壁に近い P3 では、定位誤差が大きくなる。これは、反響の影響というよりは、むしろ P1, P2 に比べ、P3 周辺では近傍のマイク数が少ないことが原因であると考えられる。同時二音源の場合には、定位がスピーカユニット周辺に集中していることから、両音源がうまく定位されていることがわかる。ただし、ガラス壁付近では、誤定位も見受けられる。これは、主に反響の影響であると考えられる。従って、測定した伝達関数を用いて、フィルタベクトルを導出すれば、反響の影響は軽減し、定位結果は向上すると考えられる。

指向特性推定の評価実験から、指向特性は正確に推定されていることがわかる。スピーカを用いた場合、音源の種類に依らず、スピーカの指向特性に近い指向特性が推定されている。

肉声の場合、指向特性は、スピーカの指向特性と大きく異なっており、肉声の理想的な指向特性に近い推定結果が算出されている。これにより、肉声と録音音声の区別が可能になることが期待できる。音源方向は、指向特性のパワーが最大となる方向を抽出することによって得ることができる。どの場合でも、音源方向は正確に推定されている。本稿では、この実験は伝達関数を得るためにインパルス応答を測定したが、測定には時間がかかるため、実際の使用を考えると、測定が不要な伝達関数の推定を行う必要がある。

音源追跡の評価実験では、図 7b) の追跡結果は、暴れているように見えるが、追跡の平均誤差は表 2 から、20 cm であり、離散化の間隔が 25 cm であることを考えるとそれほど悪い結果ではないと考えられる。カルマンフィルタなどの処理を加えることにより、精度の改善が期待できる。

一方、音源方向の平均誤差は、表 2 から、9.8° であるが、標準偏差が 90° を超えている。実質的に音源向きが前か後か程度の精度しか得られていない。特にガラス壁近辺の精度が悪いため、今後の課題として反響を考慮した手法の導入が必須であろう。

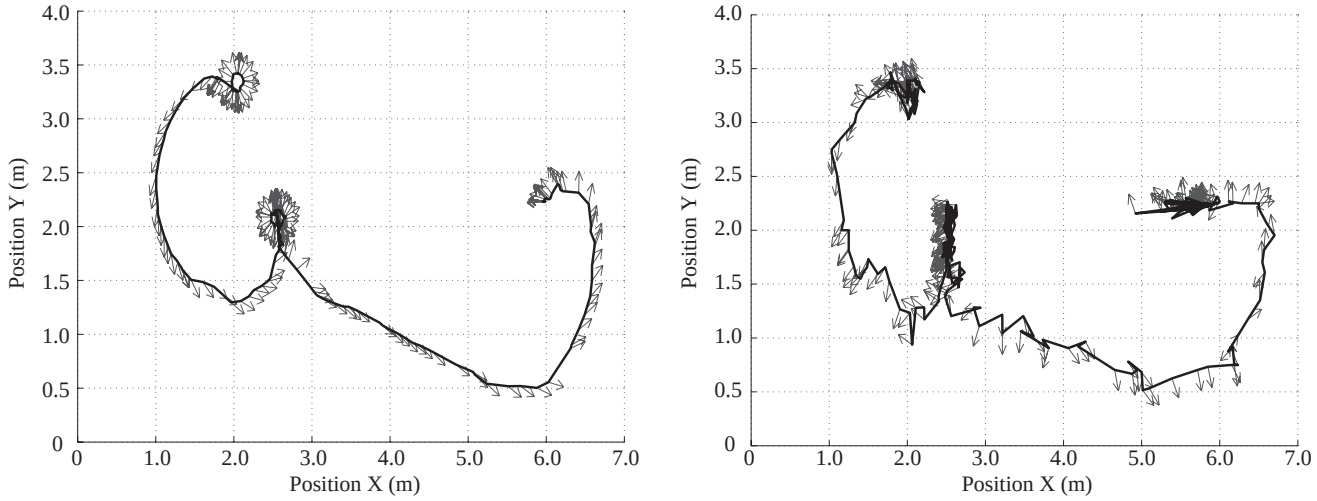
5 今後の課題

本稿で構築したマイクロホンアレイでは、マイクの位置校正は手動であるので、[2] のような自動校正手法が必要であろう。本稿で述べた WDS-BF は、予めフィルタベクトルを計算しているため、温度変化や屋内の物体の移動など環境の動的変化に対応することは難しい。MUSIC などの適応的手法や温度適応[19]などの導入、マイク間の位相誤差を低減するため、サブアレイに分割して処理を行うといった手法が必要であろう。

音源定位や追跡だけでなく、音源分離や分離音の音声認識についても今後扱う予定である。また、現状はオフライン処理で実装されているが、オンライン処理化および実時間処理化も今後の課題である。

6 結論

本稿では、大規模マイクロホンアレイを対象として、自由なマイクレイアウトを可能にし、音源の指向特性を推定できるような重み付き遅延和ビームフォーミングを提案した。 実際に、64 チャンネルのマイクロホンアレイシステムを構築し、複数音源の同時定位、指向特性推定を用いた肉声検出、音源位置と音源方向の同時追跡といった評価を通じてシステムの有効性を示した。今後は、センサが



a) Ultrasonic Three Dimensional Tag System　　b) Microphone Array System

Figure 7: Tracking of A Moving Sound Source with the Heading

埋め込まれた環境内で動作するロボットの環境理解と統合したい。

謝辞

本研究を進めるにあたり有益な議論・情報をいただいた NOE 鶴秀生氏、AIST 浅野太、麻生秀樹両氏、京大奥乃博教授、山本俊一氏に感謝する。

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] R. Biswas and S. Thrun. A passive approach to sensor network localization. In IEEE, editor, *Proc. of the IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS 2004)*, pages 1544–1549, 2004.

[3] H. K. Dunn and D. W. Farnsworth. Exploration of pressure field around the human head during speech. *Journal of Acoustical Society of America*, 10(1):184–199, 1939.

[4] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi. Autodirective microphone systes. *Acustica*, 73(2):58–71, 1991.

[5] L.J. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, AP-30(8):27–34, 1982.

[6] E. T. Hall. *The Hidden Dimension*. Anchor books doubleday, 1966.

[7] 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.

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

[9] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[10] Y. Kaneda and J. Ohga. Adaptive microphone-array system for noise reduction. *IEEE Transactions on Acoustics Speech Signal Processing*, ASSP-34(6):1391–1400, 1986.

[11] P.C. Meuse and H.F. Silverman. Characterization of talker radiation pattern using a microphone-array. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-94)*, volume II, pages 257–260, 1994.

[12] P.M. Morese and K.U. Ingard. *Theoretical Acoustics*. McGraw-Hill, 1968.

[13] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pages 832–839. AAAI, 2000.

[14] 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.

[15] T. Nakatani and H. G. Okuno. Sound ontology for computational auditory scene analysis. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 1004–1010. AAAI, 1998.

[16] Y. Nishida, H. Aizawa, T. Hori, N.H. Hoffman, T. Kanade, and Kakikura M. 3D ultrasonic tagging system for observing human activity. In IEEE, editor, *Proceedings of the 2003 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS 2003)*, pages 785–791, 2003.

[17] 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*, 2003(11):1135–1146, 2003.

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

[19] Y. Tatekura, H. Saruwatari, and K. Shikano. Sound reproduction system including adaptive compensation of temperature fluctuation effect for broad-band sound control. *IEICE Trans. Fundamentals*, E85-A(8):1851–1860, 2002.

[20] 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.

[21] 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, 2004.

[22] S. Yamamoto, K. Nakadai, H. Tsujino, and H. G. Okuno. Assessment of general applicability of robot audition system by recognizing three simultaneous speeches. In IEEE, editor, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004)*, pages 2111–2116, 2004.

ヒューマノイドロボット HRP-2 におけるロバスト音声インターフェース
Robust Speech Interface for Humanoid HRP-2

○原功，浅野太，麻生英樹，緒方淳，比留川博久，金広文男 (産業技術総合研究所)
山本潔(筑波大大学院)
* Isao HARA, Futoshi ASANO, Hideki ASOH, Jun OGATA, Hirohisa HIRUKAWA,
Fumio KANEHIRO (AIST Japan.), Kiyoshi YAMAMOTO(Univ. of Tsukuba)
isao-hara@aist.go.jp, f.asano@aist.go.jp, h.asoh@aist.go.jp, jun.ogata:aist.go.jp, hiro.hirukawa@aist.go.jp,
f.kanehiro@aist.go.jp, kyama@mmlab.cs.tsukuba.ac.jp

Abstract— For human-robot interaction in the real world, a communicative function based on speech is important. To realize such a function in anyplace, it is significant for the robots to extract target speech spoken by humans from mixture of sounds by their own resources. Consequently we have developed a robust speech interface on the humanoid robot HRP-2 using the real-time signal processing board and a microphone array system, applied a method of detection and separation of speech events. Furthermore, we have implemented a dialogue based home appliances and a humanoid control system. In this paper, we report the robust speech interface, and an experimental result of a dialogue based control is also described.

1. はじめに

近年，家庭やオフィスのような生活環境においてサービスを提供する様々なロボットの開発が行われており，2025 年には家庭における家事支援や高齢者の自立支援，介助・介護等の家庭環境における人間の生活を支援する次世代ロボットの実用化が期待されている[1]．人間と共存した環境下でサービスを行う次世代ロボットにとって，音声を用いた自然なコミュニケーションを実現する機能は，重要な知覚機能のひとつである．しかしながら，我々が活動している生活環境のほとんどの場面では，様々な雑音源が存在するために，ロボットが知覚する音声は，雑音や反響音を含んだ混合音となる．そのため，単純にマイクロホンと従来の音声認識システムを組み合わせてロボットに搭載するだけでは，人間とロボットが自然な音声対話を実現することは難しい．これに対し，生活環境内で人間と共に動作するロボットの音声インターフェースとして，複数のマイクロホンを使用し，混合音から音声のみを分離する機能や雑音を含んだ音響モデルによって音声認識をロバストにする機能を有する様々なシステムの研究開発が進められている[2) 3) 4) 5)]．我々も，さまざまな環境下で安全・安定に動作し，人間と自然なコミュニケーションが可能な人間型ロボットの実現のために，マイクロホンアレイを用いたロバスト音声インターフェースシステムの開発を行ってきた[6) 7) 8)]．このシステムでは，(a)音響情報を視覚情報との統合による発話区間の検出，(b)適応ビームフォーマを用いた音源分離および(c)音声認識におけるモデル適応の 3 つのロバスト化技術を融合し，音響処理専用のハードウェア上に実装することで，実時間の実環境音声認識を実現している．

本稿では，ヒューマノイドロボット HRP-2 に搭載したこのロバスト音声インターフェースについて述べる．また，ロボット上に実装されたロバスト音声インターフェースを用いたロボット動作制御およびネットワークに接続された情報家電機器制御を行う対話制御システムについて述べる．

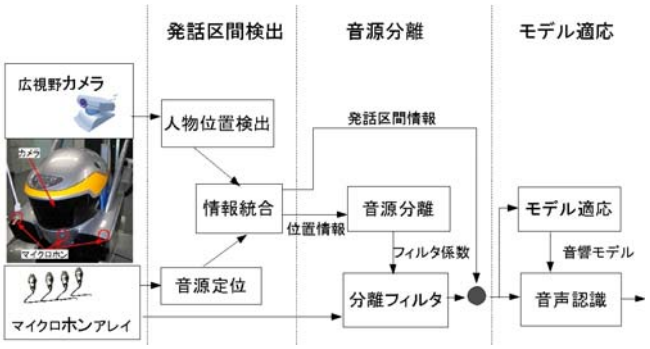


Fig.1 An overview of the robust speech interface

2. ロバスト音声インターフェース

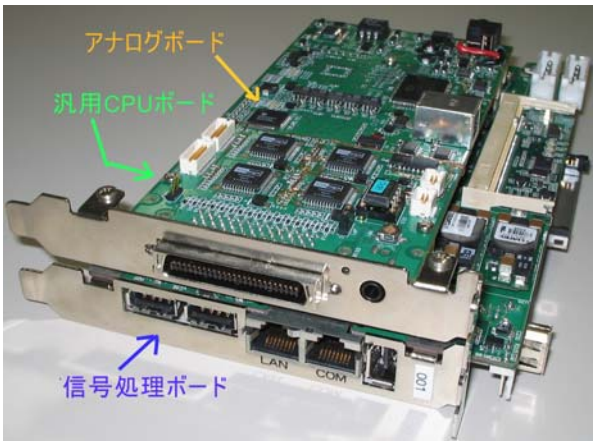
Fig. 1 に HRP-2 に搭載したロバスト音声インターフェースシステムの概要を示す．このシステムでは，(a)ロボット頭部に実装した広視野カメラ（画角：約 160 度）による人物位置推定情報と 8 個のマイクからなるマイクロホンアレイによる音源位置推定情報を用いた**発話区間検出**，(b)適応ビームフォーマによる**音源分離**による雑音の除去および(c)音声認識システムの音響モデルを残留雑音が乗った音声に合わせることで認識制度向上を図った**モデル適応**の 3 つの処理部から構成されている．(a)の発話区間検出では，マイクロホンアレイによる音源位置推定情報と画像処理による人物位置検出情報から，空間上の同一位置から発生した音響情報を発話と定義することで，雑音源が存在する実環境のおける話者による発話区

間と非発話区間との識別を行う．次に，(b)の音源分離では，(a)の発話区間検出で用いた話者の位置方向に対して適応ビームフォーマを用いることで，他の方向から発生した雑音等の除去を行い，混合音から音声のみを抽出している．この(a)および(b)の２つの処理により時間領域および空間領域での雑音除去が行われるが，処理後の音響情報には，発話区間の推定誤差や残響成分などの消し残しなどから残留する雑音が存在する．このために，音声認識プロセスにおいて，(c)の音響モデルの適応処理を施すことで，残留雑音に対する適応を行う．これにより，安定した実環境音声認識を可能にしている．

この一連の処理において，マイクロホンアレイおよびカメラからの大量の情報を実時間で処理しなければならず，当初，知覚機能処理用に実装されていた計算資源(Pentium III-S 1.4GHz)では不十分であった．そこで，多チャンネルのマイクロホンからの音響信号処理を効率的に処理することが可能なハードウェア RASP-２（Fig.2 参照）を開発し，ヒューマノイドロボット HRP-2 の体内への実装を行った．RASP-2 は，PCI ハーフサイズの基盤上に実装され，2 スロット分のスペースに収まるように設計されており，(i)16 チャンネルの A/D コンバータおよび 2 チャンネル D/A コンバータからなる**アナログボード**，(ii)PowerPC 450MHz を搭載した PrPMC タイプの**汎用 CPU ボード**, (iii)信号処理用 FPGA, IEEE1394, USB2, LAN 等のインターフェースを実装した**信号処理ボード**から構成されている．

ロバスト音声インターフェースにおける発話区間検出および音源分離の処理は，知覚機能処理用CPUボード（人物検出等の画像処理）とRASP-2（マイクロホンアレイ信号処理）上に分散して実装され，RMCP プロトコル [9] を介して実時間の分散処理を行っている．

次に，ロバスト音声インターフェースの 3 つのロバスト化技術の詳細について述べる．



Fig,2　The real-time signal processing board, RASP-2

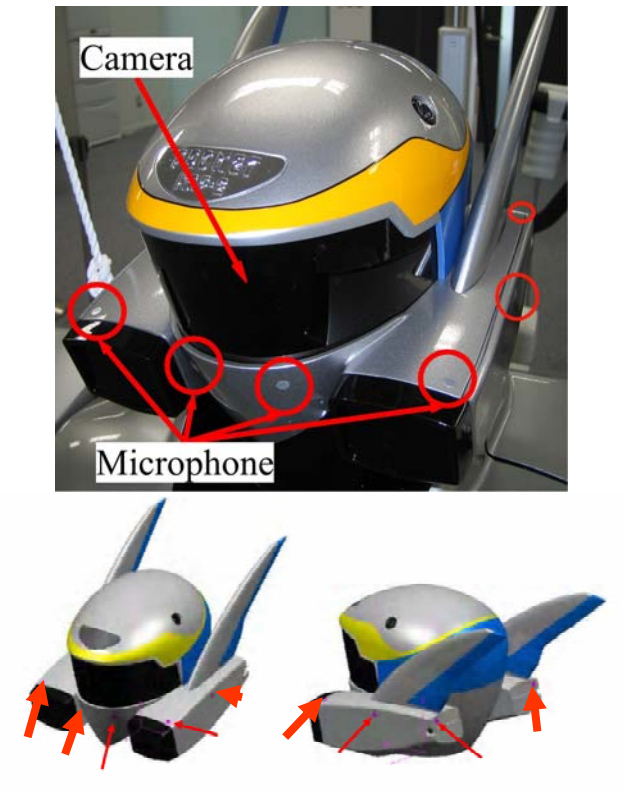


Fig.3 Microphone array and a camera on HRP-2

2.1 視覚情報と音響情報を用いた発話区間検出

生活環境内で得られる様々な雑音や反射音を含んだ音響信号から，それのみを用いてユーザが発話した部分を正確に切り出すことは非常に困難である．そのため，広視野カメラからの画像からユーザの位置情報推定し，マイクロホンアレイを用いた音源位置推定結果と統合することでユーザの発話区間の検出を行う．Fig.3 に HRP-2 に実装したマイクロホンアレイと広視野カメラを示す．また，Fig,4 に，ロバスト音声インターフェースで用いている画像情報，音響情報および発話検出用状態表示モニタを示す．

音源位置の推定には，サブスペース法（MUSIC: Multiple Signal Classification）[10] を空間相関行列の固有値を用いた重みつき平均により広帯域に拡張した方法を用いている． Fig.4(A)は，この手法を用いて得られた空間スペクトルである，この 空間スペクトルのピークを検出することで音源位置を推定することができる．Fig.4 の図では，正面（0 度）付近に音源があることを示している．

広視野カメラを用いた画像処理によるユーザの位置推定では，肌色情報と正面の顔のテンプレートマッチングによる人物発見とカーネル法を用いた追跡処理 [11] を組み合わせた方法を用いている．人物発見のプロセスでは，まず，画像中の肌色矩形領域を検出し，正面顔の平均画像とのテンプレートマッチング

を行い，閾値処理を行うことで顔領域を検出する．Fig.4(B)中の赤い矩形領域は，検出された顔領域である．この領域に対し，カーネル追跡アルゴリズムを適用し，ユーザが移動した場合にも，高速に追従することができる．この人物発見と追跡に関しては，現在, 同時に 3 人まで検出するように設定している．画像上で検出された顔領域の中心位置をユーザの位置としている．この処理モジュールは，知覚機能処理用CPUボード上に実装されており，約 15fpsの実時間処理を実現している．

上記の方法で得られた音源位置とユーザ位置を比較し, 両者の中で物理的に同じ位置(方向)に存在する場合，この方向音源からの音を発話として検出する．実際に得られる音源位置と人物位置の情報は推定誤差を含んでいるばかりか推定精度も異なっており，それぞれの位置情報が正確に重なることはほとんどない．そのため，両者の情報を柔軟かつ妥当に比較し，一致状態を推定するために，確率ネットワークであるベイジアンネットワークを用いている．Fig.4(C)は，発話区間検出に用いたベイジアンネットワークである．上部(親ノード)のノードは，発話状態か否かを示すノードであり, 下部(子ノード)の左側の 19 個のノードは，離散化された音源位置(角度)を示している．また，下部の左側の 10 個のノードは，画像上の人物位置を離散化したものである．下部の赤く表示されたノードは，それぞれ，音源の位置，人物の位置を表しており，上部ノードが赤い場合には，発話があったことを示している．

親ノードと子ノード間のそれぞれの条件付確率は，あらかじめ同システムを用いて，ユーザが単独で発話した場合の計測データにから学習したものを用いている．これによって，音響座標系と画像座標系との対応関係は，陽にキャリブレーションを行わずに異なるセンサ情報を統合することができる．

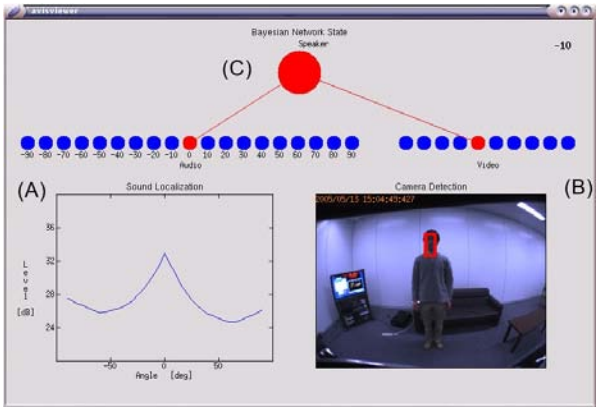


Fig.4　GUI for the robust speech interface
(A) Spatial spectrum of acoustic information
(B) View of a human tracking process
(C) Bayesian Network to detect speech events

2.2 適応ビームフォーマによる音源分離

前述の発話区間の検出では，連続した音響情報から発話部分に相当する区間の検出を行うことができた．しかしながら，テレビなどの大きな雑音源がある生活環境で得られた音響情報は，依然として音声とその背景にある雑音との混合音であり，このままでは十分な音声認識結果を得ることができない．そのために，マイクロホンアレイ処理によって得られた雑音源と話者の方向を用いることで音響信号から雑音の除去を行う．

この雑音の音源分離には，適応ビームフォーマによる分離フィルタ [12] を用いた．この分離フィルタ係数 **W**は，

$$\mathbf{W} = \frac{\mathbf{K}^{-1}\mathbf{g}}{\mathbf{g}^{H}\mathbf{K}^{-1}\mathbf{g}} \qquad (1)$$

で表されたものを用いている．**K** は非発話区間における雑音の空間相関行列であり， **g** は話者の伝達特性を含む位置ベクトルである．生活環境のような動的な環境下で音源分離を行うためには，上記のフィルタ係数は，話者の発話がないときに **K** を，話者が発話しているときにはその音源位置をもとに **g** を更新する必要がある．本システムでは，前述の発話区間の推定で得られた時間空間上の情報を用いて，分離フィルタ係数を更新し続けるようにしている．

式(1)の分離フィルタを用いて処理された分離音声は，連続した発話ごとに区切られ，それぞれ 1 発話の音声信号として音声認識システムへ送られる．

2.3 音声認識におけるモデル適応

前述の発話区間検出処理と音源分離処理による時間的・空間的な雑音除去した音声信号であっても，完全にすべての雑音除去をすることができず，反射などの影響による残留雑音が存在することが多い．現在，われわれが用いている音声認識システムJulian [13] では，通常，雑音のない音声情報によって音響モデルを構築し, 音声認識処理を行っている. そのため，残留雑音の影響によって音声認識システムで用いる音響モデルと入力音声との間でミスマッチが発生し，音声認識の性能が著しく低下する. そこで, 事前に，音声認識システムで使用する音響モデルに対し，適応処理を施す．一般に，音響モデルの適応におけるパラメータ推定方法は, MLLR[14]などのようにモデルパラメータ間での情報の共有化を利用した線形変換に基づく方法や, MAP推定 [15] などのように適応学習における事前知識を効率的に利用した方法が用いられることが多い．本システムでは，MLLR-MAP [16] による音響モデル適応を行った．これは，MLLRによりモデルパラメータの変換を行った後に，それを事前情報としてMAP推定を行ったものである．現在の本

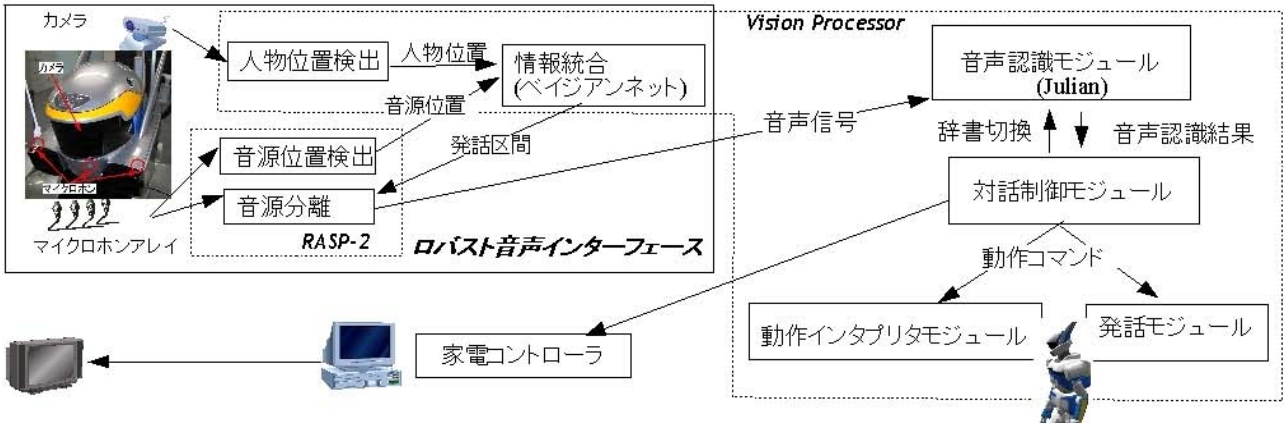


Fig.5　The dialogue system for the humanoid HRP-2

システムでは，3 人分の 20 単語を雑音源がある状態で録音したデータを基に音響モデルの適応処理を行ったものを使用している．これらの処理により音声認識システムの性能の向上を図っている．

3. HRP-2 における音声対話制御

上で述べたロバスト音声インターフェースを用いてヒューマノイドロボット HRP-2 の音声対話制御システムの実装を行った．Fig.5 にロバスト音声インターフェースを含めた HRP-2 の音声対話システム全体の概要を示す．情報家電コントローラを除いて，ロボット内部に実装した音響処理ハードウェア(RASP-2)および知覚機能処理用 CPU ボード(Vision Processor)上に実装されている．各モジュールは，独立した実行モジュールであり，ロバスト音声インターフェース部では RMCP を介して，その他のモジュール間では TCP ソケットを介して接続されている．これによって，適用するタスクやロボット内の計算資源に応じて，柔軟に分散処理を行うことが可能になっており，必要に応じてロボット体外のネットワーク上の計算資源を利用することも可能になっている．また，このような比較的疎なモジュール間の結合を用いることで，容易にシステムの拡張を行うことができる．現在，本システムで実装されている対話制御モジュールでは，ロボットやネットワークに接続された情報家電の制御することを対象としているために，音声コマンドと制御コマンドの対応付けを基本にしている．また，音声コマンドとして，孤立単語や比較的短い単文を想定し，制御対象機器の拡張などを用意にするために，音声認識システム Julian のサーバーモードの機能を利用し，対話制御モジュールから，音声認識用の辞書等を動的にロードし，動的辞書切換が可能になっている．これによって，制御対象に応じて認識モードを設定することができ，認識語彙の限定や辞書の切換による音声認識精度の向上を図ることができる．

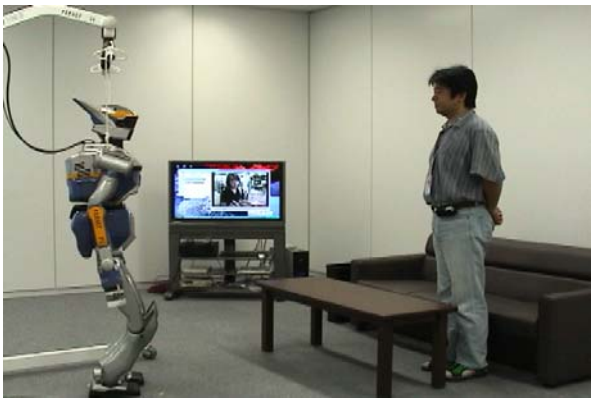
音声認識結果からロボットや外部情報家電を制御するためのコマンドへの対応づけにおいては，音声コマンドへの柔軟性を持たせるために，音声認識結果を正規表現または認識結果の列挙という形で記述することとし，対話コマンドの記述の大幅な削減と可読性を高めている． Fig.6 に，対話制御モジュールで用いているコマンド記述フォーマットを示す．この例からわかるように，対話制御のスクリプトは，XML 形式で記述し，**<rule>**タグの部分が 1 つの音声コマンドのセットとなっており，**<key>**タグに音声認識された結果および**<command>**タグによる対応する制御コマンドセットの形式で記述する．1 つの音声認識結果に対する機器制御コマンドに関しては，複数記述することが可能になっており，属性を指定することで TCP ソケットを用いた制御エージェントへのコマンド発行，音声認識辞書の切換等の内部関数呼び出しおよびスクリプト言語 Python インタプリタ呼び出しによる動的な制御コマンドの発行を行うことが可能になっている．

```
<rule>
 <key>音声認識結果1(正規表現)</key>
 <key>音声認識結果2(正規表現)</key>
 <command type="func">
       音声認識辞書切替(内部関数呼び出し)
 </command>
 <command type="net">制御コマンド</command>
 <command type="script">
    Python スクリプト
 </command>
</rule>
```

Fig.6　A script of the rule on the dialogue system

4. 音声対話制御実験

HRP-2 に実装したロバスト音声インターフェースの有効性を確認するために，Fig.7 に示すように雑音源としてテレビがある状況下でHRP-2 の動作制御，テレビのチャンネル等制御および家電コントローラであるPC上のWindows Media Playerを制御する対話制御実験を行った．Fig.7 に実験を行ったHRP-2,雑音源であるテレビおよび話者のそれぞれの配置を示す．テレビの雑音は，ロボット頭部のマイクロホンアレイ付近でS/Nが概ね 0dBになるように調整している．音声認識システムで用いる音響モデルとしては，連続音声認識コンソーシアムソフトウェア 2003 年度版のPTM(Phonetic Tied Mixture)型tri-phoneモデル[17] を用い，前述した事前に教師ありの適応を行っている．Fig.8 に実験で使用した発話シナリオを示す．HRP-2 の動作実験中もテレビの雑音がある状態で行ったが，ほぼ 9 割近くの認識率で音声による対話制御を行うことができた．本実験中で音声認識が失敗に終わるもののほとんどは，「音量を上げて」と「音量を下げて」など 1 語しか違いがないものや数字を含んだ文のときであった．これに対しては，言い方を変更や音声認識用の辞書の変更等で，より認識率を向上させることが期待できる．



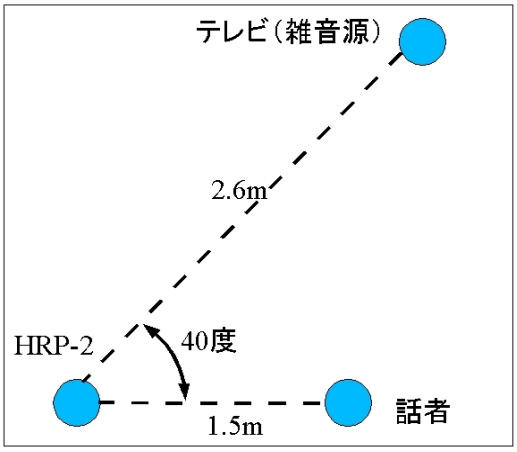


Fig.7　The arrangement on the experiment

1. こんにちは
2. 1 歩前進
3. テレビの操作
4. 電源を入れて
5. 音量を上げて
6. 音量を下げて
7. NHK 教育
8. 日本テレビ
9. NHK
10. テレビ朝日
11. ビデオの操作
12. 2 番目を再生
13. 早送り
14. 停止
15. 巻き戻し
16. 再生
17. ロボットの操作
18. 右手を上げて
19. 左手を上げて
20. 左手を下げて
21. 右を見て
22. こっちを向いて
23. 比留川さんにこれを届けて
24. ありがとう
25. さようなら

Fig.8　A list of speeches on the experiment

5. おわりに

本稿では，マイクロホンアレイを用いたロバスト音声インターフェースをヒューマノイド HRP-2 に実装を行い，ロボット本体の動作制御，外部の情報家電機器制御を行う音声対話システムの実装を行った．ロボット本体という限られたスペース内で音声対話機能を実現するために，実時間音響処理用ハードウェア RASP-2 を開発し，音響センシング，画像センシングおよび対話機能をモジュール化し，それらを分散配置することで，ロボット内の計算資源でほぼ対話システムを実装することができた．情報統合による発話検出，適応ビームフォーマによる音源分離および音響モデル適応の 3 つのロバスト化技術を融合させることで，S/N がほぼ 0dB のような高雑音化の環境においても，安定した音声によるロボット制御や情報家電制御のタスクが実行することができた．これによって，これらのロバスト化技術がロボットの実世界音声インターフェースとして有効であることが確認された．

しかしながら，今回実現したロボットの対話機能は，

ロボットと一人の話者との対話を前提としたものであり，対話制御の対象としたものは，ロボット本体の簡単な動作と情報家電の制御であった．しかしながら，我々の日常生活環境では，複数のユーザとロボットとの対話のような場面が用意に考えられ，そのような場面では，ユーザ同士の会話にロボットが反応し，誤動作を起こす可能性がある．今後は，現在使用している画像処理部分に，ユーザの視線や顔の方向検出を導入し，より詳細な対話制御を行うように，ロボット内の計算資源の拡張を含めたシステムの拡張をおこなう．また，音声認識の失敗による誤動作を修正するために，緊急時対応の音声コマンドの導入を進めていく予定である．

参考文献

1) http://www.nedo.go.jp/roadmap/index.html
2) K.Nakadai,K.Hidai,H.Mizoguchi, H.G.Okuno and H.Kitano, “Real-Time Auditory and Visual Mutiple-Object Tracking for Humanoid”, Proc of IJCAI2001, pp1424-1432, 2001.
3) http://www.incx.nec.co.jp/robot/
4) 松日楽信人,小川英樹,吉見卓,”人と共存する生活ロボット”,東芝レビュー, Vol.60,No.7, pp112-115,2005
5) http://www.hqrt.hitachi.co.jp/merl/robot/
6) F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura and H. Asoh,: “Detection and Separation of Speech EventUsing Audio and Video Information Fusion and Its Application to Robust Speech Interface", Eurasip Journal on Applied Signal Processing, 2004, 11, pp.1727-1738 ,2004.
7) I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa and K. Yamamoto: “Robust Speech Interface Based on Audioand Video Information Fusion for Humanoid HRP-2",Proc. of IROS 2004, pp. 2404-2410. 2004.
8) K. Yamamoto, F. Asano, I. Hara, J. Ogata, M. Goto, H. Furukawa, T. Kamashima and N. Kitawaki,: “Real-time Implementation and Evaluation of Speech Event Detection and SeparationBased on the Fusion of Audio and Video Information", Proc.s of GSPx 2004 , 2004.
9) M. Goto, R. Neyama and Y. Muraoka,: “RMCP:Remote Music Control Protocol | Design and Applications |", Proc. of the 1997 Int. Computer Music Conference, pp. 446-449 ,1997.
10) F. Asano, Y. Motomura, H. Asoh, T. Yoshimura,N. Ichimura and S. Nakamura: “Fusion of Audio and Video Information for Detecting Speech Event", Proce. of Fusion 2003, pp. 386-393, 2003.
11) D. Comaniciu, V. Ramesh and P. Meer: “Kernel-based object tracking", IEEE Trans. on Pattern Analysis Machine Intelligence, 25, 5, pp. 564-575,2003.
12) D. H. Johnson and D. E. Dudgeon,: Array Signal Processing ,Prentice Hall, 1993.
13) 河原達也，李晃伸:”連続音声認識ソフトウエア Julius”, 人工知能学会誌, Vol.20, No.1, pp.41-49, 2005.
14) C.J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, 9, 2, pp. 171-185 ,1995.
15) J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Transactions on Speech and Audio Processing, 2, 2, pp. 291-298.1994.
16) E. Thelen, X. Aubert and P. Beyerlein: “SpeakerAdaptation in the Philips System for Large Vocabulary Continuous Speech Recognition", Proc. of ICASSP '97, pp. 1035-1038 ,1997.
17) 河原達也，武田一哉,伊藤克亘，李晃伸，鹿野清宏，山田篤:”連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要", SP2003-169, NLC2003-106 (SLP-49-57)，電子情報通信学会技術研究報告, 2003.

ロボット頭部に設置したマイクロホンによる環境変動に頑健な音源定位

Sound Source Localization robust to variations of environments using microphones mounted to head of robot

久保 俊明, 持木 南生也, 小川 哲司, 小林 哲則

Toshiaki Kubo Naoya Mochiki Tetsuji Ogawa Tetsunori Kobayashi

早稲田大学 理工学部

Department of Computer Science, Waseda University

Abstract

A sound source localization method using statistical pattern recognition is extended so that it works robustly in various environments.

In our previous work, we proposed new types of sound source localization methods using robot mounting microphones, which are free from HRTF (Head Related Transfer Function) estimation. This method is performed with statistical pattern recognition which employs the ratio of spectra amplitude obtained for pairs of microphones as feature parameters. It works well whatever the sound source is, because the feature is completely sound-source-invariant. However, it is slightly sensitive to the variations of environments.

In order to solve this problem, HLDA (Heteroscedastic Linear Discriminant Analysis) is applied to extract environment-invariant features. Experimental results show perfect performance of the proposed method with HLDA feature extraction.

1 はじめに

ロボット上での実装に適した音源定位手法を，環境変動に対し頑健な形で実現する．

音源定位は自由空間上に設置されたマイクロホンにおいて，位相差情報を利用することが一般的である．しかし，これらの手法をロボットに適用しても，ロボット頭部に設置したマイクロホンは，ロボット頭部による回折波の影響を受けるため，正常な動作を見込めない．我々はこの問題を解決するために，4 つのマイクロホンをロボットの頭部に設置し，頭部による回折波の影響を積極的に利用する手法を提案してきた [1][2][3]．マイクロホン間のスペクトル強度比は，回折波の影響により原音声の周波数特性に依らず，音源の方向ごとに特徴的なパターンを示す．提案手法では，この音源方向毎の時系列パターンを統計的パターン認識を用いることで識別し，音源定位を実現する．このような方法においては，統計モデルの学習環境と実際の動作環境との差異が問題となる．前報 [3] においては，この差を補正するために，動作環境で得られた数方位からの少量のデータを用いて，MLLR によりモデルの適応を行なうことで定位の誤りを削減できることを示した．しかし，ロボットの移動に伴い環境は随時変動するため，その度に適応を行うのは実質的には困難である．そこで本稿では，HLDA (Heteroscedastic Linear Discriminant Analysis) [5] を用いて特徴量から識別に寄与しない残響などの環境情報を削除することで，環境が異なる場合にもロバストに動作可能な音源定位手法について検討を行う．

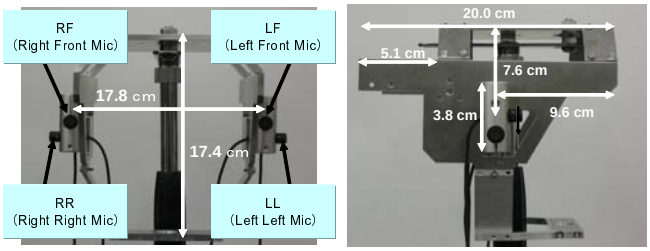
以下，2. で音源定位手法について述べ，3. で HLDA を利用した環境の変動にロバストな音源定位手法について述べる．さらに 4. で音源定位実験について述べ，5. でまとめとする．

2 音源定位手法

本節では，基本となる音源定位手法の概要を説明する．

2.1 マイクロフォンの設置

ロボットには，両側面に 2 個づつ，計 4 個の指向性マイクロホンを Figure 1 のように設置した．以下，ロボットの正面を向く方向のマイクロホンをそれぞれ，RF-Mic(Right-Front-Microphone)，LF-Mic(Left-Front-Microphone) と呼び，ロボットの側面に対して垂直な方向のマイクロホンをそれぞれ，RR-Mic(Right-Right-Microphone)，LL-Mic(Left-Left-Microphone) と呼ぶ．尚，以下の実験では指向性マイクロホンとして，Audiotechnica ATM15a を



(a) Front view (b) Side view

Figure 1: Setting of microphones

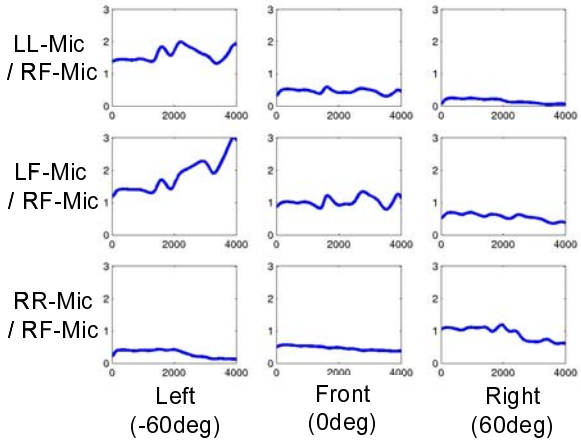


Figure 2: Example of Ratios of Amplitude Spectra

使用した.

2.2 マイクロホン間のスペクトル強度比の性質

各マイクロホンへの入力信号を用いて，マイクロホン間の強度比を求める．入力信号 $X(\omega)$ は，頭部伝達関数 $G(\omega)$ と原音声 $S(\omega)$ の積でそれぞれ表すことができる．ここでは，ω は角周波数を表す.

$$
\begin{aligned}
|X_{LL}(\omega)| &= |G_{LL}(\omega)| \cdot |S(\omega)| \\
|X_{RR}(\omega)| &= |G_{RR}(\omega)| \cdot |S(\omega)| \\
|X_{LF}(\omega)| &= |G_{LF}(\omega)| \cdot |S(\omega)| \\
|X_{RF}(\omega)| &= |G_{RF}(\omega)| \cdot |S(\omega)|
\end{aligned}
$$

ここで，$X_{RF}(\omega)$ を基準とし，正規化を行う.

$$
\begin{aligned}
\frac{|X_{LL}(\omega)|}{|X_{RF}(\omega)|} &= \frac{|G_{LL}(\omega)|}{|G_{RF}(\omega)|} \\
\frac{|X_{RR}(\omega)|}{|X_{RF}(\omega)|} &= \frac{|G_{RR}(\omega)|}{|G_{RF}(\omega)|} \\
\frac{|X_{LF}(\omega)|}{|X_{RF}(\omega)|} &= \frac{|G_{LF}(\omega)|}{|G_{RF}(\omega)|}
\end{aligned}
$$

原音声 $S(\omega)$ に依らず，マイク間のスペクトル強度比は，伝達関数の強度比で表すことができる．伝達関数は方位に関する関数になっているので，マイク間のスペクトル強度比も方向に関する関数になっていることがわかる (Figure 2).

2.3 特徴量抽出

前節で述べたスペクトル強度比の性質を利用した h 音源方向識別用の特徴量は以下の処理によって求める．処理は単語を単位として行う．マイクロホン i の入力信号に対して DFT を施したスペクトルを $X_i(w,t)$ とする．w は離散周波数，t はフレームのインデックスを表す．このとき，得られた $X_i(w,t)$ に対して，マイクロホン RF から得られるスペクトル $X_{RF}(w,t)$ で正規化を行う.

$$
N_i(w,t) = \frac{|X_i(w,t)|}{|X_{RF}(w,t)|} \quad (i = LL, RR, LF)
$$

次に，1 単語の全フレームのデータを用いて，平均スペクトルを算出する.

$$
Y_i(w) = \frac{1}{T} \sum_t N_i(w,t) \quad (t = 1, \cdots, T)
$$

この平均スペクトル $Y_i(w)$ をフィルタバンクを用いて圧縮する．フィルタバンクは，L 個の窓を周波数軸上に等間隔に配置する等間隔三角窓を使用する．単語単位の特徴量 C は以下のように求められる.

$$
c_i(l) = \sum_{w=lo}^{hi} W(w;l) \cdot \log N_i(w) \ (l = 1, \cdots, L)
$$

$$
W(w;l) = \begin{cases} \frac{w - w_{lo}(l)}{w_c(l) - w_{lo}(l)} & \{w_{lo}(l) \le w \le w_c(l)\} \\ \frac{w_{hi}(l) - w}{w_{hi}(l) - w_c(l)} & \{w_c(l) \le w \le w_{hi}(l)\} \end{cases}
$$

ただし，$w_{lo}(l)$，$w_c(l)$，$w_{hi}(l)$ はそれぞれの l 番目のフィルタの下限，中心，上限のスペクトルチャネル番号である．この処理により，単語単位の特徴量 C は，$3 \times L$ 次元に圧縮される.

2.4 音源定位

前節の述べた特徴量を用いて，統計的パターン認識の枠組で音源定位を行う．即ち，識別対象とする方位を定め，方位毎に特徴量の分布のモデルを用意する．予め，学習データを収集し 2.3 の方法で特徴量を求め，これを用いて分布を学習する．定位時には，人力された単語から特徴量を抽出し，各方位のモデルに対し尤度計算を行い，最大の尤度を与えるモデルの方位を音源方向とする.

3 HLDA を利用した環境の変動にロバストな音源定位

上述の音源定位手法はパターン認識の枠組みを用いているため，ロボットが動作する部屋の残響やロボットの位置など，モデルを学習した環境と実際に認識を行う環境の違いにより性能が劣化する可能性がある．我々はこれまでに，残響の異なる環境でも頑健な定位を行うために，実際にロボットが動作する環境で得られた数方位からの少量のデータを用いて，MLLR によりモデルを適応する手法を用いてきた [3]．しかし，ロボットが移動することで環境が随時変動することを考えると，その度に適応のための音声データを採取するのは実質的には困難であり，このようにモデルを適応する方法は現実的ではない．

そこで，特徴量から残響などの環境情報を削除し，識別に寄与する情報のみを抽出することを試みる．本稿では，これを識別情報抽出 (Useful Information Extraction; UIE) と呼ぶ．識別情報抽出では，座標変換を行なうことによって，特徴量を識別に有効な識別情報 (useful dimension) と残響や位置の情報など，本質的には識別に寄与しない環境情報 (nuisance dimension) に分離する．このように，識別に寄与する情報のみを特徴量として用いることで，環境の変動にロバストな音源定位が実現できる．本稿では，座標変換の手法として HLDA[5] を用いる．

HLDA では，useful dimension に対する平均と分散は全てのクラスで違う値，nuisance dimension に対する平均と分散は全てのクラスで共通の値を持つという拘束条件の下で，座標変換後のモデルに対する尤度 (式 (1)) を最大にするように変換行列を求める．

$$
\begin{aligned}
&L(\mathbf{x}_i|\mu_j,\mathbf{\Sigma}_j,\theta) \\
&\quad = -\frac{1}{2}\sum_{i=1}^{N}\{(\theta^T\mathbf{x}_i-\mu_{g(i)})^T\mathbf{\Sigma}_{g(i)}^{-1}(\theta^T\mathbf{x}_i-\mu_{g(i)}) \\
&\qquad + \log((2\pi)^n|\mathbf{\Sigma}_{g(i)}|)\} + \log|\theta| \qquad (1)
\end{aligned}
$$

θ は変換行列を示し，μ_j，$\mathbf{\Sigma}_j$ はクラス j の平均と分散を示す．また，$g(i)$ はデータ x_i がクラス $g(i)$ に属してることを意味する．

4 音源定位実験

本手法の有効性を示すために，音源定位実験を行った．

4.1 収録環境

収録環境を Figure 3 に示す．ここでは，以下に示す 3 通りにロボットを配置し収録を行った．

配置 1 位置: P_1，方向: X 軸方向を向く．

配置 2 位置: P_1，方向: Y 軸方向を向く．

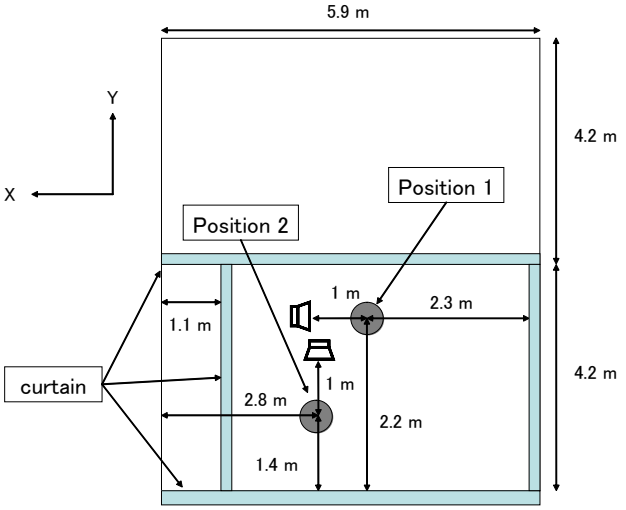


Figure 3: Recording enveiroment

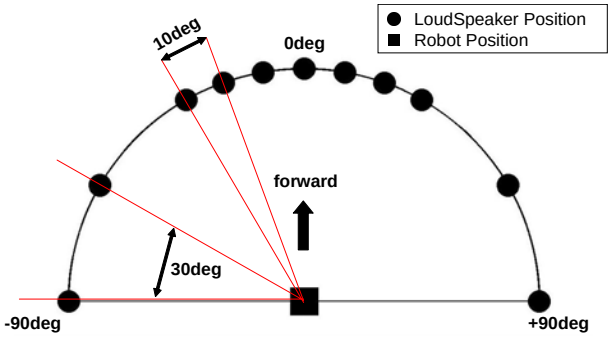


Figure 4: Direction of arrival to be recognized

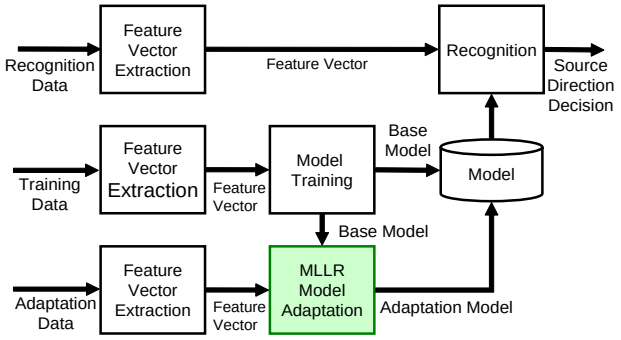
配置 3 位置: P_2，方向: Y 軸方向を向く．

また，残響時間に関しては，部屋のカーテンの開閉により 10 通り (226ms，232ms，237ms，238ms，246ms，267ms，282ms，318ms，326ms，395ms) の環境で収録を行った．したがって，配置が 3 通り，残響が 10 通りの計 30 通りの収録パターンがある．

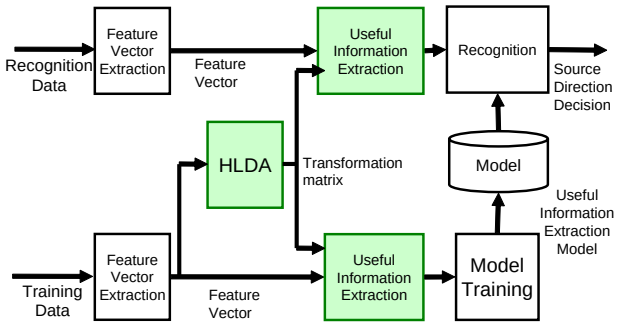
方位に関しては，Figure 4 に示すように，11 方位 (-90deg，-60deg，-30deg，-20deg，-10deg，0deg，10deg，20deg，30deg，60deg，90deg) に対して収録を行った．このとき，ロボットから見て正面を 0 deg とし，右方向を正，左方向を負として角度を定義した．

4.2 音声データ

ATR 音素バランス単語 10 単語を男性 10 人が発話したもの (計 100 単語) をスピーカから再生し，各方位に対して収録を行った．収録に用いた音声は全て，32 kHz で標本化，16bit で量子化されている．このとき，各方位に対して学習データを 90 単語，評価データを 10 単語とし，組合せをかえて 10 通りの実験を行う．よって，評価データは計 100 単語から成る．このとき，学習データと評価



(a) Baseline system and MLLR-based system



(b) HLDA-based system

Figure 5: System diagram

データは異なる話者，発話内容になるように組み合わせを選択し，評価を行った．

特徴量を抽出する際の分析条件はフレーム長 128 ms，フレームシフト 32 ms，窓関数はハニング窓とした．フィルタバンクに関しては，バンク数を 8，周波数のレンジは $0 \sim 4000$ [Hz] とした．マイク数が $N = 4$，フィルタバンク数が $L = 8$ なので，2.4 で述べた通り特徴量は 24 次元になる．

4.3 統計モデル

統計モデルは，4.2 で述べた 24 次元特徴量を用いて作成される単一ガウス分布を基礎とする (ベースモデル)．このベースモデルに対し，実際にロボットが動作する環境のデータを用いて適応を行った適応モデルと，3 で述べた識別情報抽出により変換された特徴量を用いて作成された識別情報抽出モデルについて評価を行う．各々の統計モデルを用いた実験のダイアグラムを Figure 5 に示す．

4.3.1 ベースモデル (Base Model; BM)

配置 2 種類，残響時間 2 種類 (238ms，318ms) の 4 環境で収録したデータを用いてモデルを構築した．これをベースモデルとする．収録した 3 種類の配置のうち，2 種類を学習環境，残り 1 種類を評価環境とし，組み合わせを変えて評価を行った．このとき，モデルは単一ガウス分布，全共分散行列とした．

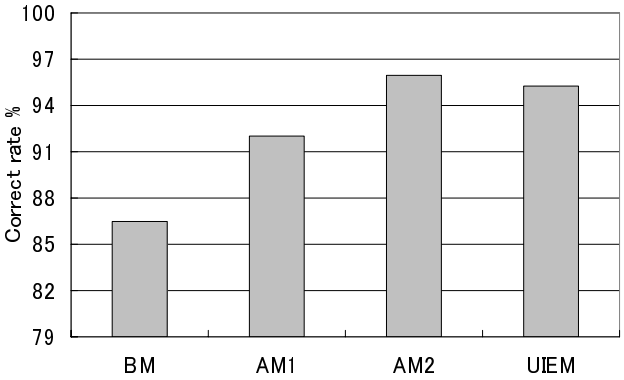


Figure 6: Experimental results of DOA estimation

4.3.2 適応モデル (Adaptaion Model; AM)

実際にロボットが動作する環境において得られた少量データから，ベースモデルを MLLR により適応を行う．適応データは，3 方位 (60 deg，0 deg，−60 deg) から学習データと同じ話者 1 人，同じ音声 5 単語 (計 15 単語) を用いた．評価データと同一の残響，異なる配置のデータを用いて適応したモデルを AM1，残響，配置ともに評価データと同一のデータを用いて適応したモデルを AM2 と呼ぶ．AM1 は部屋の残響のみに対する適応であり，同じ部屋である限り一度適応を行えば良い．また，ロボットが移動することで環境が随時変動する場合，その度に適応を行うという AM2 に基づくシステムは実際的には実現困難である．

4.3.3 識別情報抽出モデル (UIE Model; UIEM)

HLDA を用いた識別情報抽出によって構築したモデルを UIEM と呼ぶ．ここでは，ベースモデルの学習データと同じデータを用いて変換行列を求めた．useful dimension を固定し，その際に得られる変換行列を用いて同データを変換してモデルを作成した．モデルはベースモデル，適応モデルと同様，単一ガウス分布，全共分散行列とした．このように，UIEM は変換行列の推定において評価環境の情報を全く用いておらず，この点が適応モデルとは異なる．

4.4 音源定位結果

Figure 6 に 11 方位の音源定位結果を示す．BM は 86.5%，AM1 は 92.0%，AM2 は 95.9%，UIEM は 95.3% の性能を示した．これより，AM2 が最高の性能を与えることがわかる．しかし AM2 に基づくシステムでは，環境が随時変化する度に適応データを取得する必要があり，現実的とは言えない．それに対し，同一の部屋であれば一度適応データを取得すれば良い AM1 の性能は BM に対し若干の向上するに留まり，AM2 には及ばない．したがって，ロボットが移動する場合，時々刻々適応を行う必要が

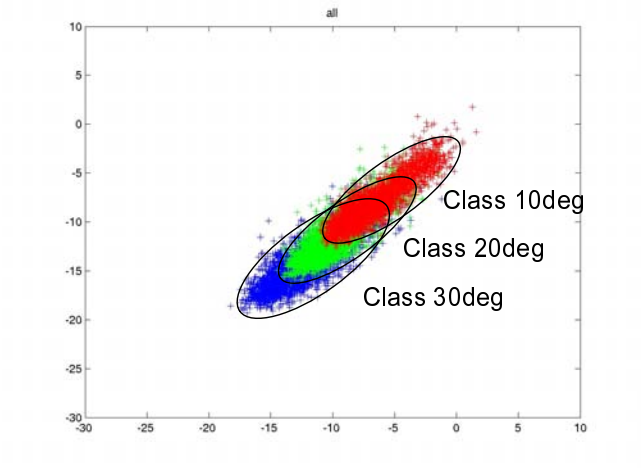


Figure 7: Data distribution before HLDA

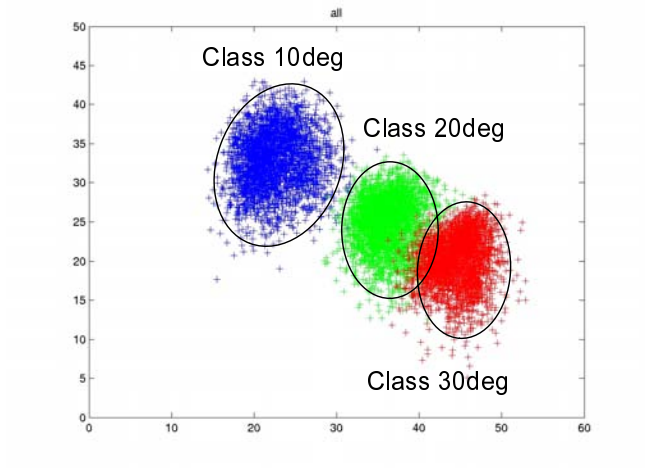


Figure 8: Data distribution after HLDA

あることがわかる．一方，UIEM は環境の情報を用いることなく，AM2 と同等の性能を与える．ここでは，useful dimension は 14 次元の場合の性能を示した．結果として UIEM は BM に対して 65% の誤りを削減した．

このように，HLDA に基づく座標変換を行うことにより，様々な環境に対してデータを収集する必要なしに，ロバストに動作する音源定位システムが構築可能であることが示された．

HLDA の効果として，HLDA を施す前と後の異なる環境に対する座標空間上におけるデータの分布の変化を調べた．HLDA を施す前のデータを Figure 7 に示す．Figure 7 における各プロットは，横軸の値が座標変換を施さない空間における特徴量の 17 次元目，縦軸の値が 18 次元目を示している．また，HLDA を施した後のデータ分布を Figure 8 に示す．Figure 8 における各プロットは，横軸の値が座標変換後の特徴量の 1 次元目，縦軸の値が 2 次元目を示している．

HLDA を施さない空間上のデータにおいては，部屋の残響の変化に伴いデータが変動しており，それが識別に影響を及ぼす方向に対するものである．それに対し，HLDA による座標変換後の空間では，部屋の残響が変化しても，識別に影響が少ない方向にデータが変動している．これは，環境に依存しない，真に識別に寄与する情報の抽出に対して，HLDA に基づく座標変換が有効に機能することの裏づけとなる．

5 まとめ

マイクロホン間のスペクトル強度比を用いた統計的パターン認識に基づく音源定位手法の改善を試みた．学習環境と実際に定位を行なう環境が異なるときに性能が劣化する問題に対し，HLDA を利用することで，環境の違いを考慮しない場合の誤りを 65% 削減することができ，環境の変動に対してもロバストな定位性能が得られることを示した．

参考文献

[1] 持木南生也，関矢俊之，小川哲司，小林哲則，“ロボット頭部に設置した 4 系統指向性マイクロフォンによる音源定位および混合音声認識,” 人工知能学会研究会資料，SIG-Challenge-0420-4，pp.21-27，Dec.2004.

[2] N.Mochiki, T.Sekiya, T.Ogawa and T.Kobayashi, “Recognition of Three Simultaneous Utterance of Speech by Four-line Directivity Microphone Mounted on Head of Robot,” Proc.ICSLP, pp.821-824,2004.

[3] 持木 南生也，関矢 俊之，小川 哲司，小林 哲則，“ロボット頭部に設置した 4 系統指向性マイクロホンによる音源定位,” 日本音響学会春季発表会講演論文集, pp.821-824, 2005.

[4] K.Nakadai, D.Matusura, H.G.Okuno, H.Kitano, “Applying Scattering Theory to Robot Audition System,” Proc.IROS, pp.1147-1152, Oct.2003.

[5] N.Kumar, “Investigation of silicon-auditory models and generalizationof linear discriminant analysis for improved speech recognition,” Ph.D. dissertation, Johns Hopkins Univ., Baltimore, MD, 1997.

[6] M.J.Hunt and C.Lefebre, “A comparison of several acoustic representations for speech recognition with degraded and undegraded speech,” Proc.ICASSP, pp.262-265, 1989.

[7] P.Brown, “The acoustic-modeling problem in automatic speech recognition,” Ph.D. dissertation, IBM

T.J.Watson Res.Center, Yorktown Heights, NY, 1987.

384ch 壁面・天井スピーカアレイによる複数音焦点形成

Sound Spots Forming with the 384ch Wall and Ceiling Speaker Array

○石井 最澄[1,2]，佐々木 洋子[1,2]，大友 佑紀[1]，加賀美 聡[2,3,1]，溝口 博[1,2]
* Yoshizumi Ishii[1,2], Yoko Sasaki[1,2], Yuuki Ootomo[1], Satoshi Kagami[2,1,3],and Hiroshi Mizoguchi[1,2]

[1]東京理科大学，[2]産業技術総合研究所，[3]科学技術振興機構

[1]Tokyo Univ. of Science, [2]Digital Human Research Center, AIST, [3]JST

j7505015@ed.noda.tus.ac.jp

Abstract—The paper describes 384 channel wall and ceiling speaker array which forms multiple sound spots. The panel-shaped arrangement of speakers gives lower side lobes and obtains high sound pressure difference between the sound spot and other points. The measurement shows the system makes possible about 15dB in sound pressure difference. We also considered signal output method well-suited for multiple sound spots generation by using beam forming simulation and devised split-up output. It proved sound pressure difference in the space and formed multiple sound spots efficiently.

1. はじめに

著者らは人と機械が音声を用い，相互にやり取りできる機械を目指し，特定の人又は場所に音を聞かせる研究を行っている．

音に指向性を持たせるための研究開発としては，これまでに，スピーカを平面状に多数並べた装置[1,2]や遅延和 (Delay and Sum Beam Forming: DSBF) 法を用いて指向性を実現したもの[3),4)]が開発されている．また，ＪＲ東海と三菱電機エンジニアリングは共同で超音波を利用した「超指向性音響システム」を開発した．このシステムの原理は，人の耳には聞こえない超音波を搬送波として用い，壁や床などに反射させることで可聴音に変化させ，音を聞かせる[5)]ものである．

著者らは，これまでに２軸，３軸の直線上にスピーカをならべたスピーカアレイを構築し，遅延和法を用いて対象とする複数の人の頭部周辺にそれぞれスポット状の高音圧領域「サウンドスポット」を作り出した[6,7)]．しかし，2 軸，3 軸スピーカアレイでは，高い音圧差が得られない，サイドローブができるなどの欠点があった．

そこで今回，音圧差の向上とサイドローブの低減を目的とした，大規模壁面スピーカアレイを構築した．スピーカの数を大幅に増やす(384ch)ことで，一つ一つのスピーカからの出力を小さくすることができ結果，音圧差の向上が期待される．Fig.1 にサウンドスポット形成の様子を示す．大規模壁面スピーカアレイが形成する，個々のスポットの音声は独立で別々である．以下 2 節では，サウンドスポット形成法に関し，スピーカの配置とアレイ要素の選択的利用法について述べる．3 節では原理に基づいたシミュレーションを実施し，複数焦点の形成と周波数依存性，アレイ要素選択的利用の効果を示す．4 節では構築したスピーカアレイシステムの実現技術について述べる．5 節では構築したスピーカアレイシステムの評価実験を行う．

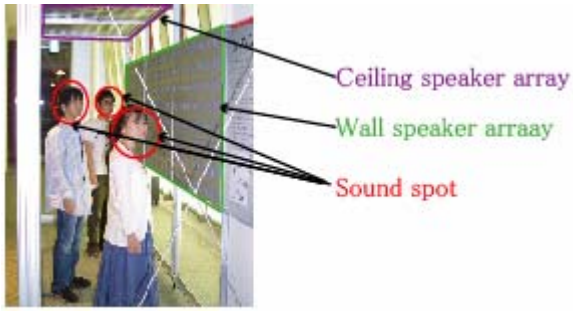


Fig. 1　Wall and ceiling speaker array generating three sound spots

2. サウンドスポット形成

遅延和法を用いて，サウンドスポット形成を行う．各スピーカから，焦点までの距離差を使い，音の位相と振幅が焦点でそろうようにする．

2.1 スピーカの配置

縦に 8 つ，横に 8 つ，計 64 個のスピーカを 1 つのパネルとして(Fig.2)，壁面 3 枚，天井 3 枚の計 6 枚のパネルを Fig.3 のように配置し，遅延和法を用いた．直線状スピーカアレイでの焦点形成では，高音圧のビームを２次元でしか制御できなという制約がある．そこで，多チャンネルのスピーカを２次元平面上に配置することで，高音圧のビームを３次元で制御することを可能とした．さらに平面アレイを，天井及び壁面に配置することで，３次元音響焦点(スポット)を形成することを可能とした．スピーカの間隔は，Fig.2 のとおり縦 140[mm]，横 140[mm]とした．これについては，第 3 節で詳しく述べる．

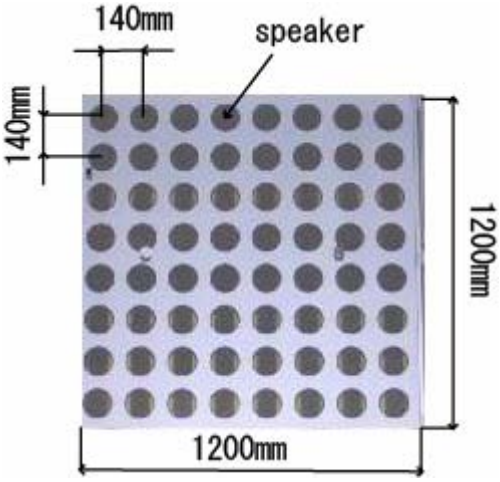


Fig. 2 One panel arrangement

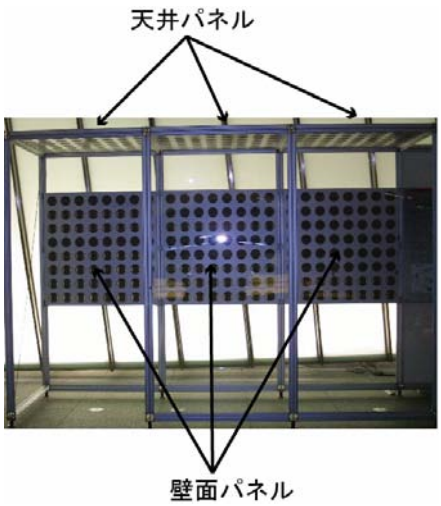


Fig. 3 System outlook

2.2 空間分離出力法

従来のシステムでは，すべてのスピーカからそれぞれの焦点に信号を出力していた．これは以下，合成法と呼ぶ．しかし，今回はスピーカを壁面状に大規模に並べるため，焦点とスピーカ間の最大距離が従来よりもはるかに長くなる．遅延和法においては，焦点で位相と振幅をそろえるため，焦点から遠い位置にあるスピーカは，焦点に近いスピーカよりも，距離差に比例した分だけ，大きな音を出さなければならない．結果，その大きな音が焦点以外の場所でも影響を与える．

以上を踏まえて，焦点に対し，焦点付近のスピーカからのみ信号を出力する方法(以下，空間分離法と呼ぶ)を提案する．式を用いた詳しい説明は次節で行う．以上の二つの方法について本論文では，シミュレーションと実測を行う．そして，シミュレーションと実測との比較から，提案手法を評価する．

3. 焦点形成のシミュレーション

スピーカアレイシステムにおいて遅延和法を用いた場合，スピーカの間隔，個数，出力周波数に応じてビームフォーミングの指向特性は変化する．そこで，本節では，実際にスピーカアレイを構築する前に，３次元ビームフォーミングと３次元的な局所的高音圧領域形成の可能性を調べる．そして，最適な指向特性の高音圧領域を形成できる条件を求めるために，様々な条件でシミュレーションを行う．また，焦点での収束具合，焦点とそれ以外の点での音圧差を比較する．

なお，シミュレーションでは，焦点での音圧を基準(0dB)とし，パネルの中心に焦点を作った．また，議論を簡単にするため，スピーカは点音源とみなした．1000[Hz]の正弦波の波長は 340[mm]なのでスピーカの外形サイズ(66.4×107[mm])より，波長は十分大きく，スピーカの大きさによる音場への影響は無視できると仮定する．

3.1 音圧分布の算出式

スピーカアレイ付近の点 C の座標を(x, y)，点 C から i 番目のスピーカまでの距離をR_i，振幅を１，正弦波の周波数をω,音速をV，i 番目のスピーカから焦点までの距離をL_iとする．この時，点 C での合成波Q_cは次式で表すことができる．

$$Q_c(t)=\sum_{i=1}^{N}\frac{Li_i}{Ri_i}\sin\left(\omega\left(t+\frac{L_{\max}+R_i-L_i}{V}\right)\right) \qquad (1)$$

3.2 シミュレーション範囲

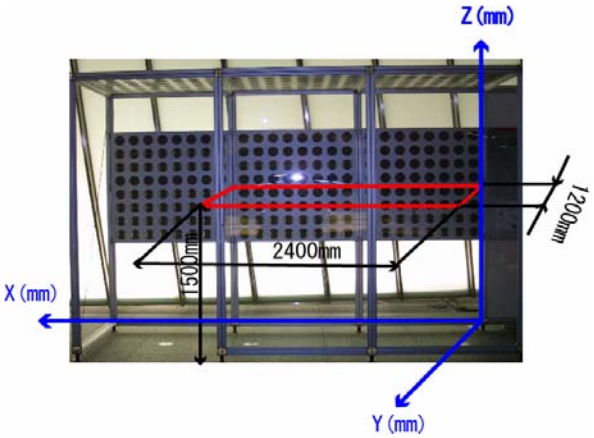


Fig.4 Simulation range and coordinates

Fig.4 に構築したスピーカアレイのシミュレーション範囲をしめす．Z 軸 1500[mm](焦点の高さ)で X 軸方向に 2400[mm]，Y 軸方向に 1200[mm]の長方形の範囲内で，焦点位置を(600, 600, 1500[mm]), (1800, 600, 1500[mm])として，シミュレーションを行う(Fig.4 の

赤い長方形内)．以下にシミュレーション結果を示す．

3.3 シミュレーション結果

a) 合成法

Fig.5 に音圧分布のシミュレーション結果を示す．焦点位置は黒×で示す．シミュレーションには，500, 1000, 1500, 2000, 2500, 3000[Hz]の正弦波を用いた．

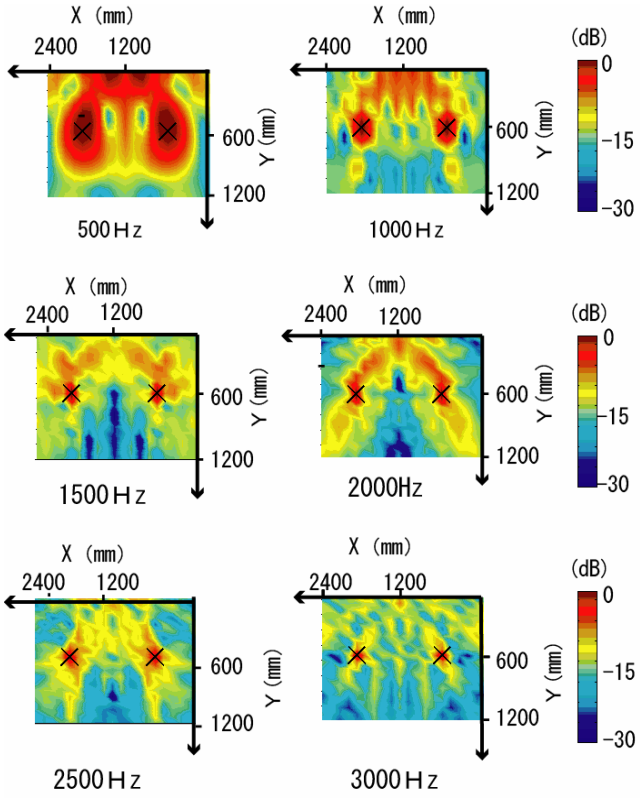


Fig.5 Simulation result: Original method

Fig.5 から 1000～2000[Hz]においてはサウンドスポットが効果的に現れていることがわかる．500Hz 以下の低周波帯域の音では波長が長いために収束せず，高音圧の領域が広がってしまっている．また 2000Hz 以上の高周波帯域ではスポットが小さくなりすぎる．

さらに，焦点以外のところでもある程度の強さの音が聞こえている．これでは，「特定の人にのみ聞かせる」という目的に対しては不十分である．そこで以下の空間分離法を適用する．

b) 空間分離法

前節の焦点形成の方法は，式(1)より距離によらず全てのスピーカから，各焦点にスポットを作るための音をだしていた．しかし，ここでは，焦点とそれ以外の場所での音圧差を大きくするために，隣接した天井パネルと壁面パネルを一組として，その焦点の信号を出力する．(1)式より，各スピーカから焦点までの距離の最大値が，合成法では約 Li=1900mm であるのに対し，空間分離法をとることで，約 Li=224mm となり，焦点以外での音圧の影響を少なくすることができる．その結果，指向特性があがり，焦点で効果的に音声を聞かせることができる．Fig.6 にシミュレーション結果を示す．

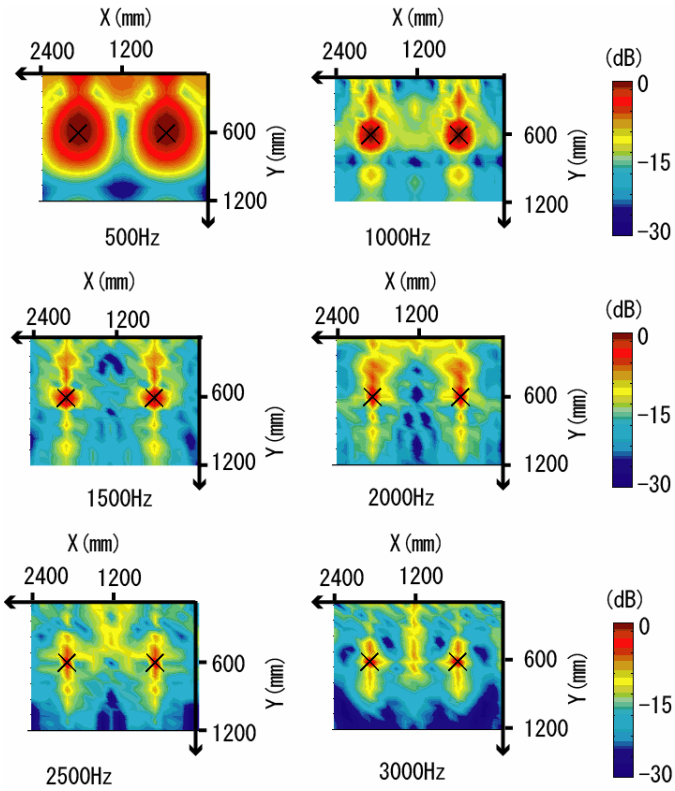


Fig.6 Simulation result : Proposed method with specially segmented speakers

Fig.6 から，500Hz 以下の低周波数帯域においては，高音圧領域が大きく広がっているものの，Fig.5 と比較すると，空間分離法では焦点で効果的にサウンドスポットができていることが確認できる．

4. 構築したスピーカアレイシステム

4.1 スピーカアレイシステムの構成

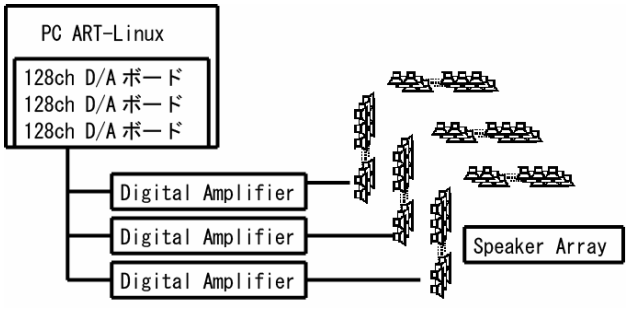


Fig. 7 System component

構築したスピーカアレイシステムのブロック図を Fig.7 に示す. 制御には実時間 OS である ART-Linux[8] と 128 チャンネル同時出力 D/A ボードを搭載した PC を用いた. PC とスピーカアレイの間にはデジタルアンプを介している.

4.2 超薄型平面スピーカ

本システムではスピーカに Fig.8 に示す FPS 社の超薄型平面スピーカ「FPS0304N3R1」を採用した.

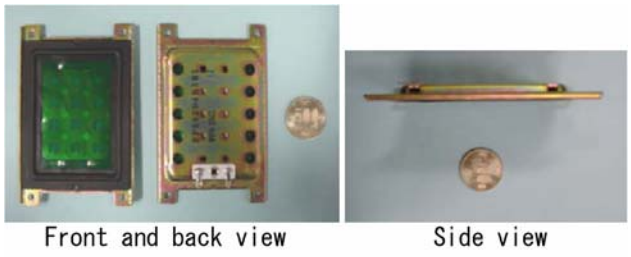


Fig.8 Thin-model speaker

スピーカ間隔は Fig. 2 に示すように 140[mm]である. これは, 1000～2000[Hz]の周波数帯でシミュレーションを行った結果, スピーカの間隔 140[mm]以下の場合に, 焦点において高音圧領域が効果的に収束したためである.

4.3 128 チャンネル同時出力 D/A ボード

スピーカアレイを実現するためには,各スピーカを μs オーダで同期させ, 等周期で制御する必要がある. そこで, 多チャンネル高速同時サンプリング用に開発された PCI 128ch D/A ボード [9]を 3 枚使用した (Fig.9). この D/A ボードは, 128ch 14bit 分のデータを 5μs 程度の時間内に DMA(Direct Memory Access) 転送することができる. 実時間 OS の ART-Linux を用いた等周期ループにより, 本システムでは 22[kHz] のサンプリングレートが実現できている.

Top View　　Bottom view

Fig. 9　PCI 128ch D/A board

4.4 デジタルアンプ

本シシステムで使用するスピーカにはアンプが内蔵されていないため, デジタルアンプを用いてシステムを構築した. デジタルアンプは, アナログアンプに比べて, 少スペースであり, また消費電力が少ないという特長を持つ. 本システムのアンプ回路[10),11)]では PWM 信号発生回路や D 級出力段を内蔵している米国トライパス社の TA2024 というデジタルオーディオアンプ IC を使用した [12,13,14)] .

5. スピーカアレイの評価実験

5.1 計測

パベック電子社のマイクロフォン MC-105, 同社の 16chA/D 変換装置, 同社のマイクロフォンアンプ MA-2016 を組み合わせて計測用マイクロホンアレイを構成した. この装置を用いて x 軸方向 0～2400[mm], y 軸方向 0～1200[mm], z 軸方向 940～2140[mm]までの範囲を 100[mm]間隔で測定する (Fig.10).

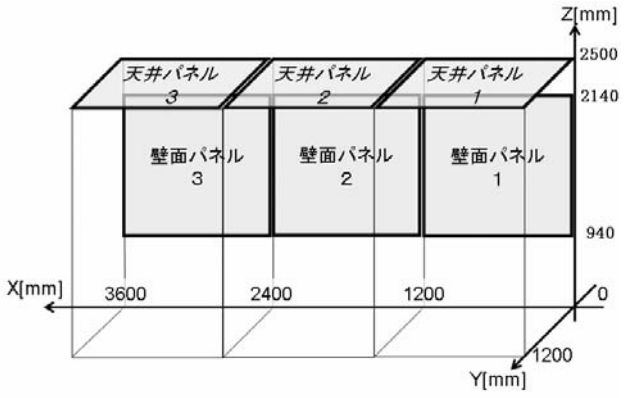


Fig.10 Experimental set up and dimensions

5.2 音場計測結果

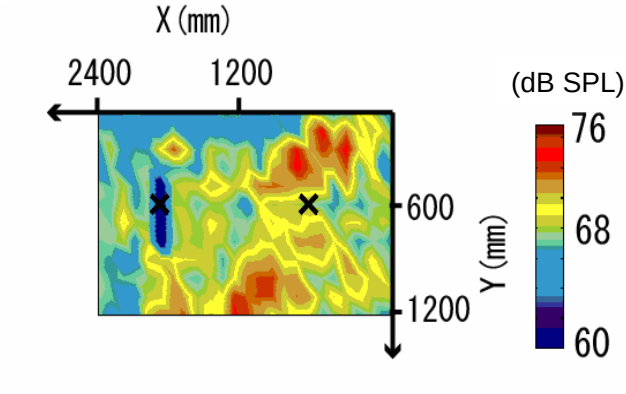


Fig.12. Simulation result: Original method

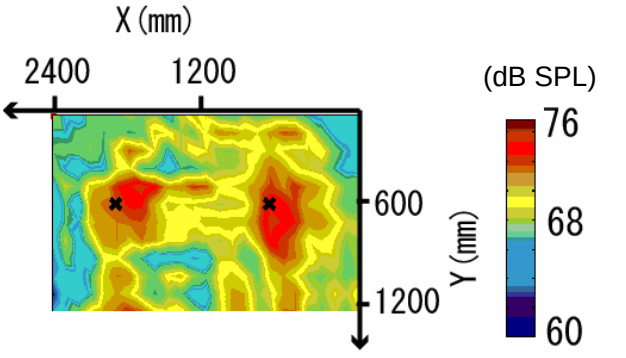


Fig.13 Simulation result: Proposed method with specially segmented speakers

Fig.12 に合成法によって焦点形成した場合の，Fig.13 に空間分離法による場合の測定結果を示す．焦点は(600,600,1500)[mm], (1800,600,1500)[mm]の位置に形成した．音源には 1000[Hz]の正弦波を用いた．Fig.12，Fig.13 は，床面からの高さ z=1500[mm]での水平面(Fig.4 の赤い長方形の内側)上の音圧分布である．合成法による音圧分布図(Fig.12)では，焦点付近に高い音圧の領域がある．しかし，焦点でスポット状の高音圧領域は形成していないことより，サウンドスポットを形成しているとはいい難い．一方，空間分離法では，スポットを作っているとはいえないが，焦点で高い音圧分布をしている．以上より，空間分離法のほうが合成法より効果的に高音圧領域を形成することができる．(注： Fig.12 の左側は測定ミスのため，焦点付近で音圧が非常に弱い状態になっている．このため右半分だけ参照されたい．)

シミュレーション結果と比較すると，合成法の実測結果では， y ＞600mm において高音圧領域を形成している．この点を除けば，同じような音圧分布をしているといえる．合成法の実測とシミュレーションを比較すると，シミュレーションの方は丸いスポットを形成しているが，実測の方は，y 軸方向に長い楕円に近いスポットを形成している．また，合成法，空間分離法のどちらにおいても，シミュレーションのほうが，焦点とそれ以外の場所の音圧差が大きい．これは，シミュレーションが反射の影響を考えていない等の原因が考えられる．

そこで，著者らは反射がどの程度スポット形成に影響を及ぼしているのかを確かめるために計測範囲の壁面及びスピーカアレイ外周を吸音材で覆い音圧分布の再測定を行った．Fig. 14 に結果を示す．

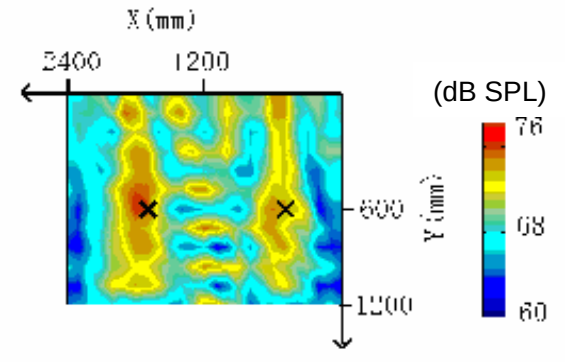


Fig.14 Measurement with sound-absorption material

Fig.14 と F ig.13 を比較すると Fig.14 の方が，焦点とそれ以外の点での音圧差が大きい．また，図に示される音圧のビーム形状を比較すると，Fig.14 の方が，ビームが鋭いことが分かる．シミュレーションと比較すると，Fig.14 の方が，シミュレーションに近い音圧分布図になっている．このことよりスピーカアレイシステムは環境に依存することが分かる．今後は環境に依存しないようなスピーカアレイシステムを構築する必要がある．具体的には，スピーカにマイクを追加して，音を集音し，環境の伝達関数を求め，スピーカから環境の伝達関数を考慮した，出力を行うなどが上げられる．

6. 考察

今回構築した 384ch 壁面スピーカアレイの音圧測定実験では，シミュレーションにおいて最適な効果を確認することができた正弦波を用いた．現状の方法では，低周波では音が広がりすぎてしまい(Fig.15)．一方 3000Hz 以上の高周波では，収束範囲が狭すぎるという欠点をもつ(Fig.15)．

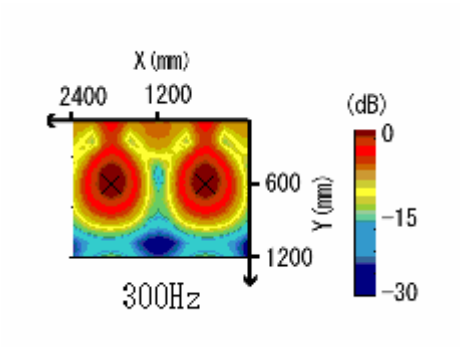


Fig.15 Sound spots on low frequency

人の声などを出力する時は，低周波数帯域の信号を除いても内容を理解することができるので，ハイパスフィルタで低周波信号成分を取り除くなどが考えられる．

7. おわりに

シミュレーションと実測から，スピーカを平面に並べることで，音をビーム状ではなくスポット状に近い形(楕円状)に制御できることを確認した．また，今回スピーカの数を大幅に増やしたことで，焦点位置とそのほかの位置で15dBの音圧差を形成することができた．しかし，低周波数帯域では音が収束せず，明確にサウンドスポットを作ることができなかった．高周波数帯域(今回のシステムでは2000Hz以上)では，音源間隔に比べて波長が小さくなることから，焦点以外でも高音圧の領域が現れた．今後は，高周波数帯域と低周波数帯域の音場の制御を工夫することが課題である．

謝辞

東京理科大学の雨宮豊氏と玉井裕樹氏，産業技術総合研究所の高野太刀雄氏，R-Lab社の長嶋功一氏と椛澤光隆氏は，本論文で述べたスピーカアレイの開発に多大な貢献をした．東京理科大学の酒谷広太，大嶋聖人，林宏樹，三竹伸生各氏は音圧測定に貢献をした．本研究の一部は科学技術振興機構さきがけ研究 21 の，一部は文部科学省科学研究費補助金の支援により行われた．記して謝意を表す．

参考文献

1) 吉川茂，藤田肇；基礎音響学，講談社サイエンティフィック(2002)．
2) HARRY F. OLSON, P_HD. Acoustic Sound 翻訳 西巻正朗，森司，古川誠二郎，近藤巖，横山巧:無線従事者教育協会(1959)．
3) 城戸健一，曽根敏夫，柴山幹夫，山口公典，中鉢憲賢:日本音響学会講座① 基礎音響工学，コロナ社(1977)．
4) 二村田忠元，奥田襄介，城戸健一，曽根敏夫：電気音響工学，オーム社(1963)．
5) 中村健太郎：音の仕組み，ナツメ社(1999)．
6) 玉井裕樹，加賀美聡，溝口博，長島功一，高野太刀雄：超多チャンネルスピーカアレイによるサウンドスポット形成の動特性評価，日本機械学会ロボティクス・メカトロニクス講演会‘04 講演論文集 pp. 1P1-H-58(1)-(4)(2004)．
7) 雨宮豊，玉井裕樹，加賀美聡，溝口博，長島功一，高野太刀雄：超多チャンネルマイクアレイによる生活環境下での２次元音源定位，日本機械学会ロボティクス・メカトロニクス講演会‘04 講演論文集 pp. 1P1-H-58(1)-(4)(2004)．
8) 石綿 陽一，松井俊浩，國吉 康夫：，高度な実時間処理機能を持つLinuxの開発，第16回日本ロボット学会学術講演会予稿集，p335-356(1998)．
9) 雨宮豊，玉井裕樹，溝口博，加賀美聡，長島功一，高野太刀雄：超多チャンネルマイクアレイによる生活環境下での２次元音源定位，日本機械学会ロボティクス・メカトロニクス講演会‘04 講演論文集 pp. 1P1-H-58(1)-(4) (2004)．
10) 中島平太郎：デジタルオーディオ読本，オーム社，(1991)．
11) 坂巻佳壽美：デジタル信号処理，工業調査会(1998)．
12) 臼井支郎，船田哲男，梅崎太造，戸田尚宏，萩原克行，横田康成，奥水大和:インターユニバーシティー 信号解析，オーム社(1991)．
13) 石田義久，鎌田弘之： デジタル信号処理のポイント，産業図書(1989)．
14) 足立修一：MATLABによるデジタル信号とシステム，東京電気大学出版局(2002)．

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

人工知能学会
JSAI Technical Report
SIG-CHallege-0522-17 (10/15)

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

Evaluation of Missing Feature Theory Based Automatic Speech Recognition for Simultaneous Speech Sentences

山本 俊一[1] **Jean-Marc Valin**[2] **中臺 一博**[3] **中野 幹生**[3] **辻野 広司**[3]
駒谷 和範[1] **尾形 哲也**[1] **奥乃 博**[1]

Shunichi YAMAMOTO[1], **Jean-Marc VALIN**[2], **Kazuhiro NAKADAI**[3], **Hiroshi TSUJINO**[3],
Kazunori KOMATANI[1], **Tetsuya OGATA**[1], **and Hiroshi G. OKUNO**[1]

[1] 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

[2] Dept. of Electrical Engineering and Computer Engineering, Université de Sherbrooke

[3] (株) ホンダ・リサーチ・インスティチュート・ジャパン, Honda Research Institute Japan, Co., Ltd.

{shunichi,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp, jean-marc.valin@usherbrooke.ca,

{nakadai,nakano,tsujino}@jp.honda-ri.com

Abstract

A robot in the real world usually hears mixtures of sounds. To achieve such a robot audition system, the integrtion of sound source separation (SSS) and automatic speech recognition (ASR) is necessary. We propose to use the missing feature theory (MFT) as an interface with high interoperability for the integration. The main advantage of this approach resides in the fact that the ASR with a clean acoustic model can adapt the distortion of separated speech by consulting a missing feature mask (MFM). In our MFT-based robot audition system, we developed a microphone array SSS system to output separated speech with a MFM generated without any prior knowledge, and we used Multiband Julius, which supported stochastic language models and recognized speech fast, for the MFT-based ASR to recognize the separated speech by using the MFM. We evaluate the robot audition system working with the humanoid *SIG2*. As a result, we showed the improvement in word correct rates and processing speed through speech recognition of the mixtures of three sentences as well as those of three isolated words.

1 はじめに

ヒューマノイドが人間と自然なインタラクションを行う上で，音声認識は重要な機能の一つである．実環境では，通常，単一音源からの音ではなく，複数の音が混在した混合音が聞こえる．つまり，実環境において混合音を認識することはヒューマノイドの基本的な聴覚機能であると言える．音声に非音声雑音が混在している混合音については，AURORA プロジェクト [1, 2]などで，研究が行われている．こうした状況に対応する一般的な手法として，雑音を含んだ音声に対して HMM パラメータを学習するマルチコンディション学習が挙げられる [3, 4]．この手法で得られた音響モデルには，特定条件下の雑音が反映されているため，想定範囲内の雑音には効果的であり，実際に，カーナビや電話サービスといった音声認識アプリケーションで用いられている．

一方，実環境では音声に音声雑音が混在している混合音を扱わなければならない場合もある．このような問題を扱う研究としては，音源分離に重点を置いた手法や分離音の認識に重点を置いた手法がこれまでに報告されてきた．前者としては、澤田らの 8 ch のマイクロホンアレイを用いた手法 [5] が挙げられる．澤田らは，ビームフォーミングと多段階の後処理を組合せて高精度な同時発話音声を分離を行い，音響モデル適応のみを用いて分離音声を精度よく認識できる手法を報告している．後者としては，中臺らのマルチコンディション学習を同時発話認識に応用した研究 [6] が挙げられる．中臺らは，ヒューマノイドの耳部に備えた 2 本のマイクを用いて方向通過型フィルタによる音源分離を提案しているが，この手法は高速である反面，分離性能が主に音源方向によって変ってしまうという特徴を持っている．そこで，このように音源分離に多少の歪みがある場合でも話者・方向依存の音響モデルを構築し，各音響モデルを用いた音声認識結果を統合することによって認識精度を向上させる，アンサンブルシステムを報告している．

我々は，実際に実環境で耐えうる混合音の認識を行うためには，音源分離と音声認識を個別に研究するのではなく，お互いの欠点を補い合えるような有機的な統合を可能にする親和性の高いインタフェースが必要であると考えて

いる．このようなインタフェースを実現するものとして，雑音によって歪んでしまった部分（ミッシングフィーチャ）を抽出し、マスクすることによって認識向上を図るミッシングフィーチャ理論（*Missing Feature Theory*, MFT）[7, 8] に着目し研究を進めている．実際に，これまでに，2 本のマイクを用いた混合音声分離，およびクリーン音声を先見情報として用いたミッシングフィーチャマスク（*Missing Feature Mask*, MFM）生成により孤立単語の分離音声認識の実装・評価を行い， 音声雑音のような非定常性雑音に対しても，ミッシングフィーチャ理論が，音源分離と音声認識のインタフェースとして有効に働くことを示した [9]．その反面，雑音によって歪んでしまった部分を先見情報を用いずにどのように抽出するのか，孤立単語認識以外にも有効であるか未確認であったといった問題点があった．そこで，本稿では，8 本のマイクを用いたマイクロフォンアレイ音源分離と音源分離からの情報を利用した MFM 自動生成を提案し，複数同時発話の連続音声認識に適用する．

2 音源分離と音声認識の統合における課題

音源分離と音声認識はそれぞれ別個な技術として独立に研究されてきた．例えば，音源分離では，分離音を出力するだけで，分離歪みの情報を利用できるような形で音声認識に提供するような報告はほとんどない．また，MFT を用いた音声認識では，ミッシングフィーチャを自動的に検出する手法に関する報告は，非音声雑音が重畳された単一話者の音声を対象としたものがほとんどであった [7].

このため，MFT を用いて音源分離と音声認識を統合する際には，以下のような課題を解決する必要がある．

1. 適切な音声認識特徴量の検討
   一般に音声認識の特徴量として，メル周波数ケプストラム係数（*Mel-Frequency Cepstral Coef- ficient*, MFCC）が用いられる．これに対し，歪みの検出は，一般にスペクトル領域で行う方が好ましい．そこで，MFT を利用する場合は，音源分離によって歪んだ音声に適した特徴量を音声認識で利用できるよう検討する必要がある．

2. MFM の自動生成
   これまで，クリーン音声を先見情報として MFM 生成に利用してきたが，実環境で，こうした先見情報を利用することは難しい．MFM 自動生成は，前述したように非音声雑音が重畳された単一話者の音声を対象としたものがほとんどであったが，混合音の場合，チャネル間のクロストークなども考慮して音源分離で，MFM 自動生成の手がかりとなる情報を推定する必要がある．

3. 高速かつ MFT に基づく音声認識を行える音声認識
   従来から利用してきた音声認識エンジン CASA Toolkit (CTK) [10] は，MFT を利用できるが，低速であり，孤立単語認識しかサポートしていなかった．MFT を利用でき，実時間動作が可能な音声認識エンジンを用いる必要があると共に，孤立単語認識以外への有効性の検証を行う必要がある．

以下，上記 3 点に関して，詳細，および本稿における解決のアプローチについて説明する．

3 音声認識特徴量

本稿で扱う 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 [11]を用いる．MFCC は入力音声がクリーンな場合は有効であるが，入力スペクトルに歪みがあると，たとえそれが特定の周波数領域での歪みであっても， MFCC の全係数に影響を与えてしまい，ロバスト性が低下する．また，本手法の音源分離の後処理に利用する多チャンネル post-filter は，周波数領域で背景雑音推定や，他の音源からの干渉成分のスペクトル推定を行っている．従って，これらの 2 つの理由から，MFM を自動生成するには周波数領域の特徴量である方が適している．

以下に，MFCC の計算で用いられるのと同様の正規化を行ったメル周波数対数スペクトル特徴量の導出の手順を示す．

(1) 音響信号を 16 ビット，16 kHz でサンプリングし，窓幅 25 ms，シフト幅 10 ms の FFT を行う．

(2) メル周波数領域で等間隔に配置した 24 個の三角形窓によりフィルタバンク分析を行う．

(3) 24 個のフィルタバンクの出力の対数を取り，メル周波数対数スペクトルを得る．

(4) 対数スペクトルを離散コサイン変換する．

(5) ケプストラム係数の 0, 13-23 次の項を 0 にする．

(6) ケプストラム平均除去（*Cepstral Mean Subtraction*, CMS）を行う．

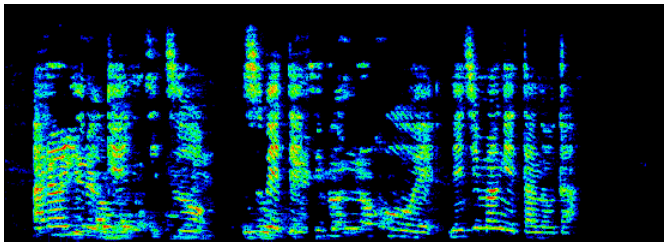
(7) 逆離散コサイン変換を行ってスペクトル領域に戻す．

(8) 各次元毎に一次微分を計算する．

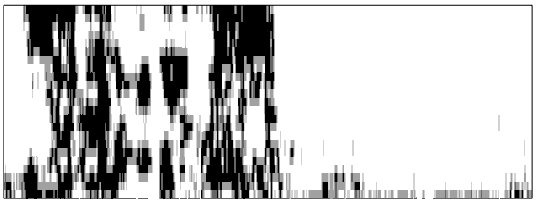
(9) 微分値と合わせて，計 48 次元の特徴量として抽出し，メル周波数対数スペクトル特徴量を得る．

4 ミッシングフィーチャマスク自動生成

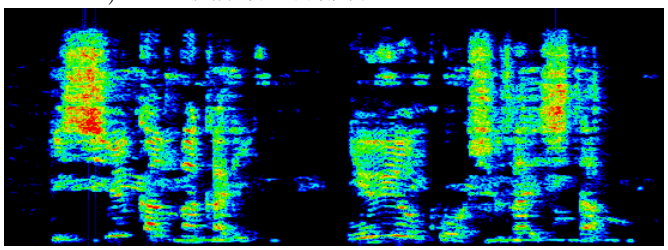
MFM 自動生成には，分離音声から抽出したメル周波数対数スペクトルのうち，どの周波数帯域が歪んでいるかという情報が必要である．周波数帯域の歪みを検出するた



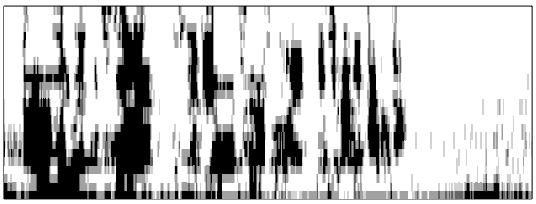
a) 左 90 度話者の分離音スペクトログラム



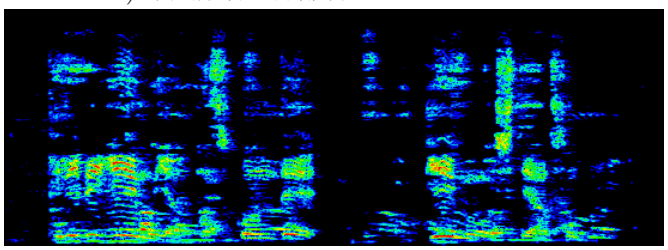
b) 左 90 度話者の分離音用 MFM



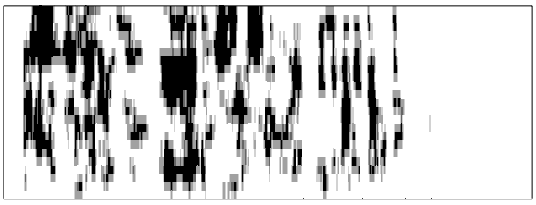
c) 中央話者の分離音スペクトログラム



d) 中央話者の分離音用 MFM



e) 右 90 度話者の分離音スペクトログラム



f) 右 90 度話者の分離音用 MFM

Figure 2: 自動生成された MFM の例

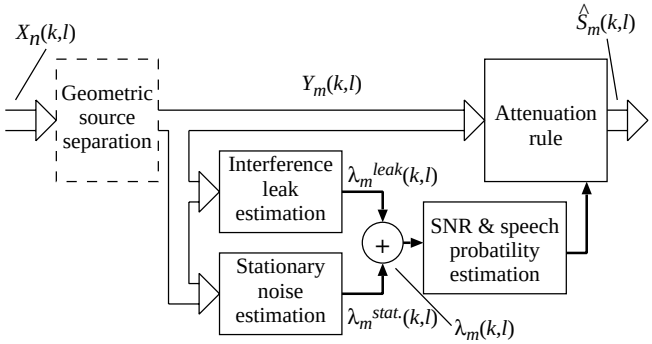


Figure 1: 多チャンネル post-filter の概要

めに，先見的情報は仮定せずに音源分離処理から得られるデータのうち，多チャンネル post-filter の入力および，出力音響信号，推定された背景雑音のスペクトルを利用する．

多チャンネル post-filter [12]は，幾何学的音源分離（*Geometric Source Separation*, GSS）の post-filter 処理[13]を複数音源を扱えるように拡張した手法である．Figure 1 に示すように，この手法は，GSS のチャンネル出力雑音を定常性雑音と非定常性雑音に分けて推定を行っている．定常性雑音は，主に背景雑音であるとし，背景雑音推定を行う．非定常性雑音は，GSS の過程で他のチャンネルから漏洩したものであると仮定して，適応的に他チャンネルからの干渉成分のスペクトル推定を行っている．最終的に，定常性雑音推定と非定常性雑音推定を統合することにより，雑音推定を行っている．なお，Figure 1 において，$X_n(k,i)$ は n 番目のマイクから GSS への入力，$Y_m(k,i)$ は GSS で分離された m 番目の音源の信号、$\hat{S}_m(k,i)$ は多チャンネル post-filter 処理後の m 番目の音源の分離信号を表している。$G_m(k,i)$ は重み関数で $\hat{S}_m(k,i) = G_m(k,i)Y_m(k,i)$ と定義される。

MFM のうち，微分値でない特徴量（$i = 1, \cdots, \frac{N}{2}$）に対応するマスク $M(k,i)$ は，メル周波数帯域のフレーム k における多チャンネル post-filter の入力を $Y(k,i)$，出力を．$\hat{S}(k,i)$，多チャンネル post-filter で推定された背景雑音を $N(k,i)$ とした場合， 以下のように 2 値のマスク（信頼できるとき 1，信頼できないとき 0）として定義する．また，閾値 T は実験的に求め，0.25 とした．

$$M(k,i) = \begin{cases} 1, & \left(\hat{S}(k,i) + N(k,i)\right)/Y(k,i) > T \\ 0, & otherwise \end{cases}$$

このように，推定された背景雑音を利用するのは，背景雑音が大部分を占める周波数帯域は信頼度が高くなるようにするためである．つまり音声認識から見ると，背景雑音しか存在しなかった周波数帯域は，無音であることが「信頼できる」領域であるとするということである．

また，MFM のうち，特徴量の一次微分（$i = \frac{N}{2} +$

$1, \cdots, N$）に対するマスク $M(k,i)$ は，以下のように定義する．この場合も，2 値のマスクとなる．

$$M(k,i) = \prod_{j=-2, j\neq i}^{2} M\left(k+j, i-\frac{N}{2}\right)$$

特徴量とその一次微分に対応したマスクからなる MFM の次元数は，スペクトル特徴量と同じ $N = 48$ となる．生成された MFM の例を Figure 2 に示す．Figure 2 の a), c), e) はそれぞれ,「あらゆる現実をすべて自分の方へねじまげたのだ」,「一週間ばかりニューヨークを取材した」,「テレビゲームやパソコンでゲームをして遊ぶ」という三話者同時発話を分離した音声のスペクトログラムである．また，b), d), f) はそれぞれ，a), c), e) に対して生成された MFM である．これらの図は，横軸は時間を表し，縦軸は周波数（MFM の図ではメル周波数）を表している。また， MFM の図では，白い部分は信頼できる特徴量を，黒い部分は信頼できない特徴量を表している．

5 音声認識エンジン

MFT に基づく音声認識は一般の音声認識と同様に，隠れマルコフモデル（*Hidden Markov Model*, HMM）に基づいている．一般の音声認識システムでは，状態遷移確率と出力確率から与えられた信号系列を最も高い確率で出力する状態遷移系列を求めるが，MFT に基づく音声認識システムでは，このうち出力確率の計算方法が一般の音声認識とは異なっている．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，N を特徴量の次元数とする．このとき，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．

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

しかし，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，信頼できない特徴量ほど出力確率に貢献しないように設計する．つまり，信頼できる特徴だけが出力確率の計算に用いられ，信頼できない特徴による影響を除去しなければならない．これを実現するために，特徴量の各成分に対する信頼度を表す MFM ベクトル $M(i)$ を用いて以下のように定義する．

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

この式によると，信頼できない特徴量に対するすべての音素 HMM の尤度が等しくなるので認識には影響しない．さらに，正解の音素 HMM の尤度を低下させるような信頼度の低い特徴量をマスクすることにより，正解の音素 HMM の尤度が相対的に低くなるのを防ぐことができる．

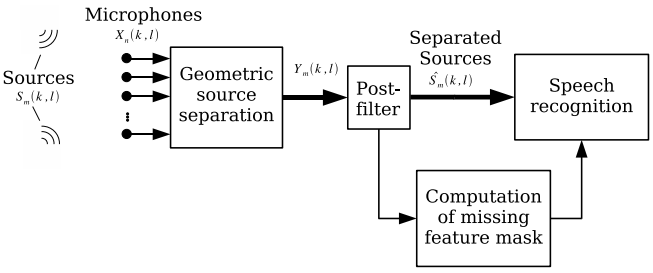


Figure 3: システムの概要

MFT に基づく音声認識エンジンの実装として，これまで利用していた CTK の代わりにマルチバンド版 Julius [11]を利用した．CTK は HMM のデコードアルゴリズムとして Viterbi アルゴリズムを用いており，音響モデルとしてモノフォンとトライフォンを，言語モデルとして有限状態文法をサポートしていた．一方，マルチバンド版 Julius は大語彙音声認識エンジン Julius [14]を MFT に基づく音声認識が行えるように改良したもので，単純なモノフォンやトライフォンだけでなく，状態を共有したトライフォンや，分布を共有したモデルなどをサポートしている．言語モデルとしても，有限状態文法だけではなく N-gram の統計言語モデルもサポートしている．また，Julius は 2 パスによる HMM のデコードを行い，リアルタイムで音声認識ができるように実装されている．マルチバンド版 Julius も速度の低下はあるものの CTK より高速に動作する．

6 同時発話認識システム

混合音声認識システムは以下の 4 つの処理部から構成されている（Figure 3）.

(1) GSS の一種として実装されているビームフォーマ

(2) 多チャンネル post-filter

(3) MFM 自動生成

(4) MFM を利用した MFT による分離音声認識

音源分離は，GSS に基づく線形音源分離法を用い，さらに，確率的勾配法を適用し，推定に利用する時間幅を短くすることによって高速化したものを利用する[12]．音源分離で利用するマイクロフォンアレイはヒューマノイドに設置された 8 本の無指向性マイクで構成されている．

多チャンネル post-filter では，前節で説明したように，GSS による分離音の目的音源を強調し，分離音だけではなく，post-filter 処理の際に推定した背景雑音スペクトルを出力する．

MFM 自動生成では，多チャンネル post-filter の入出力と推定した背景雑音スペクトルを元に MFM を生成する．

Figure 4: SIG2

Figure 5: 実験風景

Table 1: ATR 音素バランス文の例

文	あらゆる/げんじつ/を/すべて/じぶん/の/ほう/へ/ねじまげ/た/の/だ/。
音素	arayuru/geNjitsu/o/subete/jibuN/no/ho:/e/nejimage/ta/no/da/
文	いっ/しゅうかん/ばかり/ニューヨーク/を/しゅざい/し/た/。
音素	iq/shu:kaN/bakari/nyu:yo:ku/o/shuzai/shi/ta/
文	テレビ/ゲーム/や/パソコン/で/ゲーム/を/し/て/あそぶ/。
音素	terebi/ge:mu/ya/pasokoN/de/ge:mu/o/shi/te/asobu/

MFT に基づく音声認識では，分離音の特徴量と自動生成された MFM を利用して音声認識を行なう．

7 実験

システムの評価を行うためにヒューマノイド SIG2 に 8 本のマイクを取り付け（Figure 4），三話者同時発話認識実験を行った．実験を行った部屋（Figure 5）は 4 m×5 m の大きさで，残響時間は 0.3 – 0.4 秒 (RT20) である．SIG2 とスピーカの距離は 2 m で，スピーカの間隔は 30 度，60 度，90 度間隔の場合の 3 パターンで録音した．実験に用いた音声は ATR 音素バランス文 50 文で，3 体のスピーカから異なる組み合わせで文を出力し，三話者同時発話の音声認識実験を行った．音素バランス文の例を Table 1 に示す．

提案手法と比較するために，特徴量として MFCC を用いた場合の実験も行い，以下のような場合の三話者同時発話認識実験を行った．なお，音源分離に必要となる音源定位結果は所与であるとした．

(1) 特徴量 MFCC を用いて Julius によって分離音声を認識

(2) メル周波数対数ペクトル特徴量を用いて Julius によって分離音声を認識

(3) メル周波数対数ペクトル特徴量を用いて MFT に基づく Julius によって分離音声を認識

7.1 音響モデル

実験では，HMM に基づく音響モデルとして，クリーン音声で学習した単一のトライフォンを利用した．この音響モデルは 3 状態 4 混合の HMM である．学習データとして，日本音響学会の新聞記事読み上げ音声コーパスを利用した．このコーパスは，毎日新聞記事と ATR 音素バランス 503 文を 306 人の話者（男女それぞれ 153 名）が読み上げたデータである．話者 1 名あたり約 150 文，コーパス全体では約 45,000 文の発話が含まれている．

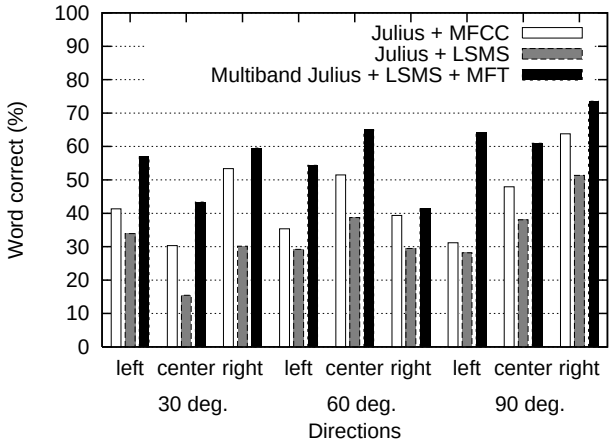
7.2 言語モデル

言語モデルは，新聞記事コーパスと ATR 音素バランス文から学習した統計言語モデルを利用した．実際に実験に用いたのは次の 2 つの言語モデルである．一つ目は，ATR 音素バランス文 50 文で学習した言語モデルであり，語彙サイズは約 400 語彙である．二つ目は，ATR 音素バランス文による言語モデルと毎日新聞記事による言語モデルを融合したもので，語彙サイズは約 2 万語彙である．この毎日新聞記事による言語モデルは連続音声認識コンソーシアム（CSRC）によって配布されているものである．

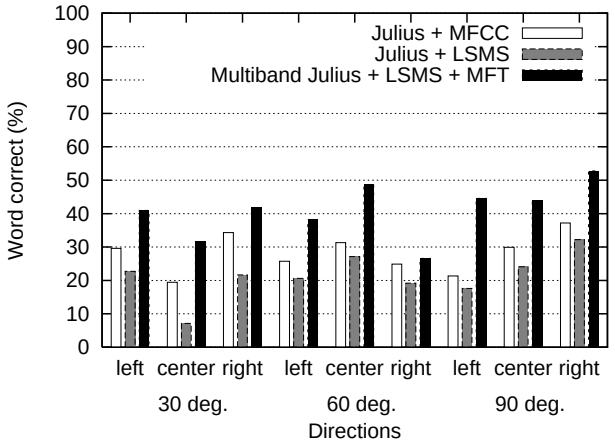
7.3 実験結果と考察

三話者同時発話認識結果の単語正解率を Figure 6 a)，b) に示す．全ての場合において，提案手法の単語正解率は通常の音声認識を行う場合よりも向上している．ATR 音素バランス文による言語モデルの場合の単語正解率は，スピーカ間隔が 30 度の場合で，平均が 53.3%，右方向の場合が最大で 59.5%，スピーカ間隔が 60 度の場合で，平均が 53.6%，右方向の場合が最大で 65.0%，スピーカ間隔が 90 度の場合で，平均が 66.2%，右方向の場合が最大で 73.5%であった．単語正解率が最大となったのは，右 90 度の場合で 73.5%であった．毎日新聞記事と ATR 音素バランス文による言語モデルの場合の単語正解率は，スピーカ間隔が 30 度の場合で，平均が 38.1%，右方向の場合が最大で 41.8%，スピーカ間隔が 60 度の場合で，平均が 37.8%，中央の場合が最大で 48.6%，スピーカ間隔が 90 度の場合で，平均が 47.1%，右方向の場合が最大で 52.8%であった．単語正解率が最大となったのは，右 90 度の場合で 52.8%であった．これらは，ヒューマノイド上での音声認識において MFCC に基づく通常の音声認識よりも，提案手法の方が適していることを示している．

処理速度に関しては，Pentium 4 (2.53 GHz) の Linux PC において，MFT に基づく Julius は 315 秒の分離音声を処理するのに 373 秒かかり，通常の Julius では 314 秒かかった MFT に基づく Julius は通常の Julius の 84%の速度であった．



a) ATR 音素バランス文による言語モデル（約 400 語）



b) 毎日新聞記事と ATR 音素バランス文による言語モデル（約 20,000 語）

Figure 6: 三話者同時発話認識結果の単語正解率

8 おわりに

音声に音声雑音が混在しているような混合音も扱うことができる音声認識を目指し，音源分離と音声認識の統合を可能にする親和性の高いインタフェースとして，ミッシングフィーチャ理論に着目し，1) MFT を用いる場合の適切な音声認識特徴量としてスペクトル特徴量の提案，2) マイクロフォンアレイによる音源分離から得られるチャネル間リーク情報を利用した MFM の自動生成，3) MFT に基づく高速な連続音声認識システムとしてマルチバンド版 Julius の利用の提案を行った．実験では三話者同時発話文の連続音声認識を行い，結果として，提案手法の有効性を通常の音声認識に対する単語正解率の向上により確認した．しかし，今回の実験で得られた単語正解率は 60% 弱であり，十分高いとは言えない．今後，音源分離性能，MFM 自動生成に更なる改善を行う予定である．また，移動音源への対応や，ヒューマノイドが動く場合への対応，また音源とヒューマノイドが同時に動く場合への対応についても検討する予定である．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (A)，特定領域研究「情報学」，京都大学 21 世紀 COE，(財) 電気通信普及財団の研究補助を受けている．また，マルチバンド版 Julian の利用を許可していただいた，東京工業大学の古井研究室と東京大学の西村義隆氏に感謝する．御討論いただいた京都大学奥乃研究室の面々，Sherbrooke 大学の Rouat 教授，Michaud 教授に感謝する．

参考文献

[1] AURORA. http://www.elda.fr/proj/aurora1.html" "http://www.elda.fr/proj/aurora2.html.

[2] D. Pearce. Developing the ETSI AURORA advanced distributed speech recognition front-end & what next. *Proc. of Eurospeech-2001*. ESCA, 2001.

[3] M. Blanchet, J. Boudy, and P. Lockwood. Environmentadaptation for speech recognition in noise. *Proc. of EUSIPCO-92*, pp.391–394, 1992.

[4] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multistyletraining for robust isolated-word speech recognition. *Proc. of IEEE ICASSP-87*, pp.705–708, 1987.

[5] 澤田知寛, 関矢俊介, 小川哲司, 小林哲則. 階層的音源分離に基づく混合音声の認識. **第 *18* 回 *AI* チャレンジ研究会報告**, pp.27–32, 2003.

[6] K. Nakadai, D. Matasuura, H.G. Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, 44(1-4):97–112, October 2004.

[7] 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. *Proc. of Eurospeech-2001*, pp.213–216.

[8] P. Renevey, R. Vetter, and J. Kraus. Robust speech recognition using missing feature theory and vector quantization. *Proc. of Eurospeech-2001*, volume 2, pages 1107–1110. ESCA, 2001.

[9] 山本俊一, 中臺一博, 辻野広司, 奥乃 博. ロボット聴覚システムの音源分離と音声認識のインターフェースへのミッシングフィーチャー理論の適用. **日本ロボット学会誌**, Vol.23, No.6, pp.743–751, 2005.

[10] CASA Toolkit. http://www.dcs.shef.ac.uk/ jon/ctk.html.

[11] 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音声認識の検討. In **日本音響学会 *2004* 年春季研究発表会講演論文集**, volume 1, pages 117–118, 2004.

[12] J.-M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. *Proc. of IEEE/RSJ IROS 2004*, 2004.

[13] I. Cohen and B. Berdugo. Microphone array postfiltering for non-stationary noise suppression. *Proc. of ICASSP-2002*, pp.901–904, 2002.

[14] T. Kawahara and A. Lee. Free software toolkit for japanese large vocabulary continuous speech recognition. *Proc. of ICSLP-2000*, pp.476–479, 2000.