人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0420

ＡＩチャレンジ研究会 (第２０回)

Proceedings of the 20th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2004 年 12 月 6 日　　場　所　京都大学工学部　8 号館 中会議室
Kyoto University, Dec. 6, 2004

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

21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」
“Informatics Research Center for Development of Knowledge Society Infrastructure”

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-1 (12/6)

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

On the tracking and separation of moving sound sources using microphone array

浅野太，麻生英樹

Futoshi Asano and Hideki Asoh

産業技術総合研究所

AIST, Tsukuba

f.asano/h.asoh@aist.go.jp

Abstract

The problem of moving sound targets is important for auditory system of robots. In this article, applicability of techniques used for moving targets in the area of radar/sonar or computer vision, such as EM algorithm, Kalman filter and particle filter is considered.

1 はじめに

音源が静止している場合には，音源定位，音源分離について多数の研究があるが (例えば [1])，移動音源に対する追跡，分離の研究は少ない [2]．一方で，レーダ，ソナーなどの軍事応用や，画像などの分野では，移動体の追跡について，多数の研究例がある [3]．本稿では，これらの分野で使われている EM アルゴリズム，カルマンフィルタ，パーティクルフィルタなどを取り上げ，移動音源の追跡と分離への応用の可能性について考える．

2 信号のモデル

本報告では，周波数領域で信号を扱う．ここでは，次節以降の話を分かりやすくするために，周波数領域での観測信号とその統計量のモデルについて述べておく．

2.1 観測ベクトルのモデル

m 番目のマイクロホン入力の短区間フーリエ変換を $Y_m(\omega, n)$ とし，これを用いて，観測ベクトル $\mathbf{y}(\omega, n) = [Y_1(\omega, n), \cdots, Y_M(\omega, n)]^T$ を定義する．n と ω は，時間及び周波数のインデクスである．以降，周波数のインデクス ω は，簡単のため省略する．環境に L 個の音源及び背景雑音がある場合，観測ベクトルは，次式のようにモデル化される．

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

ここで，行列 $\mathbf{A}$ は，位置ベクトル $\mathbf{a}_l$ を用いて，$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_L]$ のように構成される．移動音源に対しては，$\mathbf{A}$ も時変となるので，$\mathbf{A}(n)$ と書くべきであるが，見易さのため，省略する．$\mathbf{a}_l$ は，音源から各マイクロホンまでの直接音の伝達関数を用いて，$\mathbf{a}_l = [A_{1,l}e^{-j\omega\tau_{1,l}}, \cdots, A_{M,l}e^{-j\omega\tau_{M,l}}]$ のように構成される．ここで，$A_{m,l}$ と $\tau_{m,l}$ は，l 番目の音源から m 番目のマイクロホンまでのゲイン及び到達時間である．ベクトル $\mathbf{s}(n) = [S_1(n), \cdots, S_L(n)]^T$ は，各音源のスペクトル $S_l(n)$ を構成要素に持つ．ベクトル $\mathbf{n}(n) = [N_1(n), \cdots, N_M(n)]^T$ は，各マイクロホンで観測される背景雑音のスペクトル $N_m(n)$ を構成要素に持つ．

2.2 共分散行列のモデルと尤度

統計量をとるための便宜上，N 個の観測ベクトルをひとまとめにした単位をブロックと呼び，このブロック内では音源の移動量が少なく，音源の位置は定常と見なせるものと仮定する．以降は，ブロック内での時間インデクスを n とし，ブロック単位の観測ベクトルを $\mathbf{Y}(t) = [\mathbf{y}(t, 1), \cdots, \mathbf{y}(t, N)]$ と定義する．ここで，$\mathbf{y}(t, n)$ は，t 番目のブロック内の n 番目の観測ベクトルを表す．ブロックのインデクス t は，見易さのため，省略する場合がある．

ブロック内での観測ベクトルの共分散行列 (以下サンプル共分散と呼ぶ) を以下のように定義する．

$$\mathbf{C}_y(t) = \sum_{n=1}^{N} \mathbf{y}(t, n)\mathbf{y}^H(t, n) \tag{2}$$

音源信号 $\mathbf{s}(t, n)$ と背景雑音 $\mathbf{n}(t, n)$ が無相関であるとすると，サンプル共分散は，以下のようにモデル化される．

$$\mathbf{K}_y = \mathbf{A}\mathbf{K}_s\mathbf{A} + \sigma\mathbf{I} \tag{3}$$

ここで，$\mathbf{K}_s = \mathrm{diag}(\gamma_1, \cdots, \gamma_L)$ は，各音源のパワーを表す．各音源信号も相互に無相関であると仮定している．また，背景雑音 $\mathbf{n}(t, n)$ は，白色ガウス性を仮定している．

データ $\mathbf{Y}(t)$ に対する尤度関数 [5] は，

$$L_y(\Theta, \mathbf{K}_s; \mathbf{Y}(t)) = \psi_y \prod_{n=1}^{N} \exp\left(-\frac{1}{2}\mathbf{y}^H(t,n)\mathbf{K}_y^{-1}\mathbf{y}(t,n)\right) = \psi_y \exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{C}_y\mathbf{K}_y^{-1}\right]\right) \quad (4)$$

ここで，$\psi_y = (2\pi)^{-MN/2}\left[\det(\mathbf{K}_y)\right]^{-N/2}$ は，パラメタによらない定数と仮定する (必ずしも厳密ではないが)．

最尤推定 (Maximum Likelihood) 法では，この尤度関数を最大とするよう，パラメタ $\Theta = [\theta_1, \cdots, \theta_L]$ 及び $\mathbf{K}_s = \mathrm{diag}(\gamma_1, \cdots, \gamma_L)$ を推定する．ここで，Θ 及び $\mathbf{K}_s$ は，音源の方向及びパワーである．パワー $\mathbf{K}_s$ は，方向 Θ の推定後，遅延和法などにより別途推定するとしても，(4) を最大にすることは，Θ に関する L 次元の最適化問題となり、一般には解くのが困難であったり，多大な計算量を必要とする．このため，後述の EM アルゴリズムやパーティクルフィルタでは，様々な工夫がなされている．

3 EM アルゴリズム

EM アルゴリズムによる音源定位法は，[4, 5] により提案され，これを移動音源の分離に拡張した手法が著者らにより提案されている [6]．EM アルゴリズムは，ブロック内の限られたデータに対し，尤度計算と最適化のループを繰り返し適用することにより，良い初期値さえ得られれば，従来の遅延和法や MUSIC 法など，最適化を 1 回しか行わない方法に比べ，推定精度の向上が期待される．また，音源定位のプロセスに，音源のモデル (1) に基づいた音源分離の過程が埋め込まれており，音源位置を各音源ごとに個別に推定する．これにより，複数の音源位置を同時に推定する従来法に比べ，音源間の相互干渉（共分散行列のクロス項）の影響が少なくなり，この点からも，推定精度の向上が期待される．

3.1 EM アルゴリズムによる音源定位

EM アルゴリズムを用いた音源定位法では，次式のように，入力ベクトルを各音源に対応した観測ベクトル $\mathbf{x}_l(n)$ に分解して考える．

$$\mathbf{y}(n) = \sum_{l=1}^{L}\mathbf{x}_l(n) = \sum_{l=1}^{L}\left[\mathbf{a}_l S_l(n) + \mathbf{n}_l(n)\right] \quad (5)$$

ここで，ベクトル $\mathbf{n}_l(n)$ は，雑音ベクトル $\mathbf{n}(n)$ の任意の分解である．分解した観測ベクトルの集合 $\{\mathbf{x}_l(n)\}$ は，EM アルゴリズムでは，*complete data* と呼ばれ，直接観測することは，できない．続いて，complete data $\mathbf{x}_l(n)$ に対しても，観測ベクトル $\mathbf{y}(n)$ と同様に，サンプル共分散行列及びモデル共分散行列を定義する．

$$\mathbf{C}_{xl} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_l(n)\mathbf{x}_l^H(n) \quad (6)$$

$$\mathbf{K}_{xl} = \gamma_l\mathbf{a}_l\mathbf{a}_l^H + \frac{\sigma}{L}\mathbf{I} \quad (7)$$

上述の定義を用いて，音源ごとのデータ $\mathbf{X}_l = [\mathbf{x}_l(1), \cdots, \mathbf{x}_l(N)]$ に対する尤度関数は，以下のようになる．

$$L_{xl}(\theta_l, \alpha_l; \mathbf{X}_l) = \psi_{xl}\exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{C}_{xl}\mathbf{K}_{xl}^{-1}\right]\right) \quad (8)$$

ここで，$\psi_{xl} = (2\pi)^{-MN}\left[\det(\mathbf{K}_{xl})\right]^{-N/2}$．これから分かるように，complete data の導入により，L 次元の最適化問題 (4) は，各音源に対する 1 次元の最適化問題に低減されている (正確には，$[\theta_l, \gamma_l]$ の 2 次元だが，次に述べるように音源のパワー γ_l は，問題を簡単化するため，方向 θ_l の推定後，別途推定する)．

ただし，上述のように $\mathbf{X}_l$ は観測不可であるため，EM アルゴリズムでは，尤度関数の評価に必要な $\mathbf{C}_{xl}$ の期待値を E-ステップで計算し，これを用いて，M-ステップでパラメタを推定し，更新されたパラメタを用いて、再度 $\mathbf{C}_{xl}$ の期待値を計算する $\cdots$，というループによりこの問題を解決している．以下に（一部簡略化した）EM アルゴリズムによる音源パラメタ推定法をまとめる [5]．

E-Step:

$$\mathbf{C}_{xl}^p \equiv E[\mathbf{C}_{xl}|\mathbf{C}_y; \hat{\mathbf{K}}_y^p] = \hat{\mathbf{K}}_{xl}^p - \hat{\mathbf{K}}_{xl}^p(\hat{\mathbf{K}}_y^p)^{-1}\hat{\mathbf{K}}_{xl}^p + \hat{\mathbf{K}}_{xl}^p(\hat{\mathbf{K}}_y^p)^{-1}\mathbf{C}_y(\hat{\mathbf{K}}_y^p)^{-1}\hat{\mathbf{K}}_{xl}^p \quad (9)$$

$$\hat{\mathbf{K}}_y^p = \sum_{l=1}^{L}\hat{\mathbf{K}}_{xl}^p \quad (10)$$

$$\hat{\mathbf{K}}_{xl}^p = \hat{\gamma}_l^p\mathbf{a}(\hat{\theta}_l^p)\mathbf{a}(\hat{\theta}_l^p)^H + \frac{\sigma}{L}\mathbf{I} \quad (11)$$

M-Step:

$$\hat{\theta}_l^{p+1} = \arg\max_{\theta_l}\frac{\mathbf{a}^H(\theta_l)\mathbf{C}_{xl}^p\mathbf{a}(\theta_l)}{|\mathbf{a}(\theta_l)|^4} \quad (12)$$

$$\hat{\gamma}_l^{p+1} = \frac{\mathbf{a}^H(\hat{\theta}_l^{p+1})\mathbf{C}_{xl}^p\mathbf{a}(\hat{\theta}_l^{p+1})}{|\mathbf{a}(\hat{\theta}_l^{p+1})|^4} \quad (13)$$

ここで，p は，EM アルゴリズムの反復回数を表す．

3.2 EM アルゴリズムにおける信号分離過程

EM アルゴリズムで特筆すべきは，音源のパラメタ推定の過程に，音源分離の過程が含まれていることである．$\mathbf{x}(t) = [\mathbf{x}_1^T(t), \cdots, \mathbf{x}_L^T(t)]^T$ のサンプル共分散行列 $\mathbf{C}_x$ の期待値は，Kalman Filter における知識 [7] などを用いて，次式のように整理することができる [6]．

$$E[\mathbf{C}_x] = \overbrace{(\mathbf{I} - \mathbf{GH})\hat{\mathbf{K}}_x}^{Model} + \overbrace{\mathbf{G}\mathbf{C}_y\mathbf{G}^H}^{Observation} \quad (14)$$

このうち，第 1 項がモデル共分散行列 $\hat{\mathbf{K}}_x$ に関する項，第 2 項がサンプル共分散行列 $\mathbf{C}_y$ に関する項である．(14) に現れる ゲイン $\mathbf{G}$ は，以下の写像を行う．

$$\hat{\mathbf{x}}(n) = \mathbf{G}\mathbf{y}(n). \quad (15)$$

$\hat{\mathbf{x}}(n)$ は，観測値における音源ごとの寄与（マイクロホン入力）$\mathbf{x}(n)$ の推定値であることから，ゲイン $\mathbf{G}$ に音源分

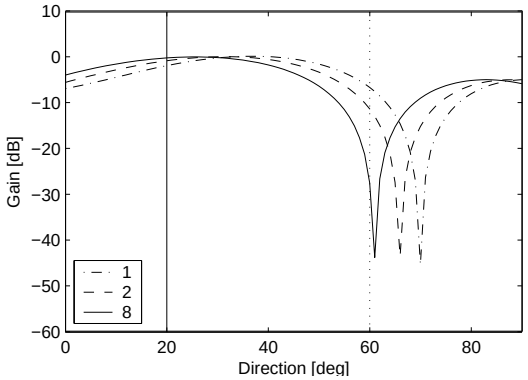


Figure 1: Directovity of Gain $\mathbf{G}_1$

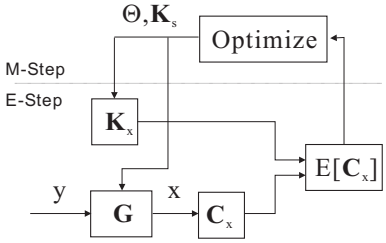


Figure 2: Block diagram of the EM Algorithm

離の機能があることがわかる．実際，$\mathbf{G} = [\mathbf{g}_1^T, \cdots, \mathbf{g}_L^T]^T$ のうち，第 l ブロック $\mathbf{g}_l$ は，次式のように近似され，

$$\mathbf{g}_l \simeq (\hat{\gamma}_l \hat{\mathbf{a}}_l)(\hat{\mathbf{a}}_l^H \hat{\mathbf{K}}_y^{-1}) \tag{16}$$

このうちの後半の項 $\hat{\mathbf{a}}_l^H \hat{\mathbf{K}}_y^{-1}$ と，次節で示す最小分散ビームフォーマとを比較すると，$\mathbf{g}_l$ は，最小分散ビームフォーマと同等の音源分離性能を持つことが分かる．

図 1 は，音源数 $L = 2$ の場合について $\mathbf{g}_1$ の方向-感度特性を示したものである．この図から，$\mathbf{g}_1$ には，音源 S2($60°$) に対して死角を形成する空間フィルタの働きがあることが分かる．また，EM アルゴリズムの反復により，死角の方向が，音源 S2 の方向の真値 (点線の縦線) に漸近していく様子が分かる．

図 2 に，EM アルゴリズムの機能をまとめる．まず，観測信号 $\mathbf{y}(n)$ は，ゲイン $\mathbf{G}$ により，各音源に対する観測ベクトル $\mathbf{x}(n)$ に分解される．これと，モデル共分散行列 $\hat{\mathbf{K}}_{xl}$ とから，サンプル共分散行列の期待値 $E[\mathbf{C}_{xl}]$ を推定する．期待値が求まったら，この期待値を用いてパラメタ $\mathbf{\Theta}$,$\mathbf{K}_s$ を最適化し，モデル共分散行列 $\hat{\mathbf{K}}_{xl}$ を更新する．このループを収束するまで反復する．

3.3 EM アルゴリズムを用いた音源分離

[6] では，EM アルゴリズムの中間生成物であるモデル共分散行列 $\hat{\mathbf{K}}_{xl}$ を利用して，音源分離を行う手法を提案している．

まず，音源分離の枠組みとして用いる最小分散 (MV)

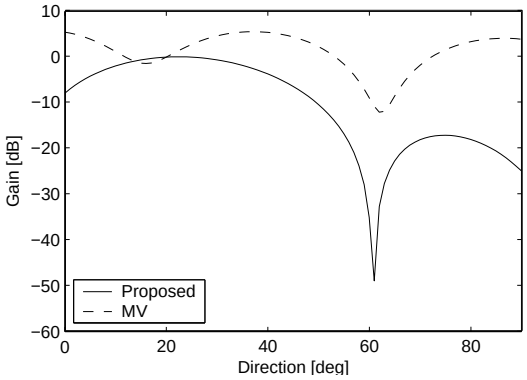


Figure 3: Directivity of the conventional and the proposed MV beamformer.

ビームフォーマを示す．

$$z(n) = \mathbf{w}^H \mathbf{y}(n) \tag{17}$$

$$\mathbf{w} = \frac{\mathbf{C}_y^{-1} \hat{\mathbf{a}}_l}{\hat{\mathbf{a}}_l^H \mathbf{C}_y^{-1} \hat{\mathbf{a}}_l} \tag{18}$$

ここで，$\mathbf{w}$ はビームフォーマ係数ベクトル，$\hat{\mathbf{a}}_l$ は，強調すべき第 l 番目の音源の位置ベクトルである．$\mathbf{C}_y$ は，2.2 節で述べた観測ベクトル $\mathbf{y}(n)$ に対するサンプル共分散であるが，移動音源の場合は，定常とみなせる十分な観測区間が得られないため，このサンプル共分散の推定精度が悪く，このまま $\mathbf{C}_y$ を用いて最小分散ビームフォーマを構成しても，移動音源に対する音源分離性能は低下する．

そこで，[6] では，$\mathbf{C}_y$ の代わりに，EM アルゴリズムの反復中に精度が改善された $E[\mathbf{C}_{xl}]$ あるいは $\mathbf{K}_{xl}$ を用いて $\mathbf{C}_y$ の代替物を生成し，用いることを提案している．例えば，$\mathbf{K}_{xl}$ を用いる場合は，$\mathbf{K}_y = \sum_{l=1}^{L} \mathbf{K}_{xl}$ を代替物として，次式のように最小分散ビームフォーマを構成する．

$$\mathbf{w} = \frac{\mathbf{K}_y^{-1} \hat{\mathbf{a}}_l}{\hat{\mathbf{a}}_l^H \mathbf{K}_y^{-1} \hat{\mathbf{a}}_l} \tag{19}$$

図 3 は，$\mathbf{K}_{xl}$ を用いた場合 (提案法) 及び $\mathbf{C}_y$ を用いた場合 (従来法) の最小分散ビームフォーマの方向感度特性である．この図から，従来の MV ビームフォーマでは，$N = 8$ 程度の平均では，雑音源の方向の減衰が 10dB 程度であるのに対し，提案法では，深く鋭い死角が形成されているのがわかる．

続いて，音源が移動する場合についてのシミュレーション結果を示す．音源の初期位置は，$[20, 60]°$ であり，半径 1.5m の円周上をそれぞれ，$[3, 2]$ km/s の速度で，等速度運動しているものとする．観測値は，(1) において，$\mathbf{A}$ を上記の運動に従って動的に変化させることにより，人工的に作成した．雑音 $\mathbf{n}(t)$ のレベルは，信号 $\mathbf{A}s(t)$ に対し，-20 dB とした．部屋の反射などは，考慮されていない．推定された音源の軌跡を図 4 に示す．図 5 は，提案法

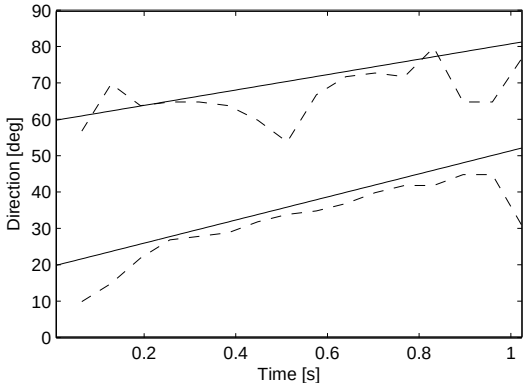


Figure 4: Trajectory estimated by the EM algorithm. Solid line: true; Dotted line: estimated.

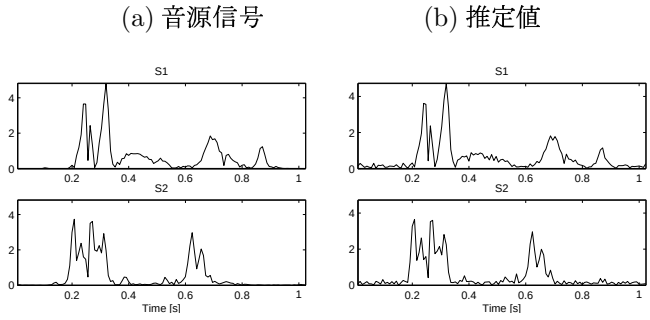


Figure 5: (a) Original waveform, (b) Separated waveform (at 1000Hz).

により，移動音源を分離した結果である (結果は単一周波数のみ).

4 Kalman Filter

4.1 観測系及び内部状態のモデル

Kalman Filter では，移動する音源のようなダイナミックなシステムを，適当な観測系を通して観測する場合を考え，観測値とシステムの内部状態 (例えば音源位置や速度などのパラメタ) を，以下のようにモデル化する．

$$\mathbf{y}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}(t) \quad (20)$$

$$\mathbf{x}(t+1) = \mathbf{F}(t+1,t)\mathbf{x}(t) + \mathbf{w}(t) \quad (21)$$

ここで，(20) は，観測方程式と呼ばれ，システムの内部状態 $\mathbf{x}(t)$ を，観測系 $\mathbf{H}(t)$ を通して観測することを表している．一方，(21) は，プロセス方程式と呼ばれ，システムの内部状態の時間的な遷移を表している．$\mathbf{v}(t)$ 及び $\mathbf{w}(t)$ は，観測系とシステム内部の雑音である．

簡単な音源追跡の問題に当てはめてみよう．内部状態 (音源のパラメタ) を音源の方向 θ 及び角速度 ω とし，時刻 t から $t+1$ への遷移は，以下のようにモデル化されるものとする．

$$\theta(t+1) = \theta(t) + \omega(t)\Delta_t \quad (22)$$

ここで，Δ_t は，時刻 $t+1$ と t との時間間隔である．また，観測値としては，MUSIC 法や最尤推定などにより，音源位置の推定値 $\hat{\theta}(t)$ が得られているものとする．以上から，観測方程式，プロセス方程式の各構成要素は，以下のようになる．

$$\hat{\theta}(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \theta(t) \\ \omega(t) \end{bmatrix} + v(t) \quad (23)$$

$$\begin{bmatrix} \theta(t+1) \\ \omega(t+1) \end{bmatrix} = \begin{bmatrix} 1 & \Delta_t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta(t) \\ \omega(t) \end{bmatrix} + \mathbf{w}(t) \quad (24)$$

動的な音源を対象とした場合，MUSIC 法などの静的な音源を対象とした方法では，安定した推定値が得られない場合がある．そこで，Kalman Filter では，推定値をそのまま鵜呑みにするのではなく，移動（状態遷移）のモデルを導入して，このモデルにより推定値の補正を行うことにより，より安定した推定値を得ようとするものである．実際的な効果としては，推定値の Smoothing のような効果が期待される．

4.2 Kalman Filter の解法

以下に，Kalman Filter による反復の手続きをまとめる [8].

Propagation:

$$\hat{\mathbf{x}}^-(t) = \mathbf{F}\hat{\mathbf{x}}(t-1) \quad (25)$$

$$\mathbf{P}^-(t) = \mathbf{F}\mathbf{P}(t-1)\mathbf{F}^T + \mathbf{Q} \quad (26)$$

Kalman Gain:

$$\mathbf{G}(t) = \mathbf{P}^-(t)\mathbf{H}^T \left[\mathbf{H}\mathbf{P}^-(t)\mathbf{H}^T + \mathbf{R}\right]^{-1} \quad (27)$$

Update:

$$\hat{\mathbf{x}}(t) = [\mathbf{I} - \mathbf{G}(t)\mathbf{H}]\,\hat{\mathbf{x}}^- + \mathbf{G}(t)\mathbf{y} \quad (28)$$

$$\mathbf{P}(t) = [\mathbf{I} - \mathbf{G}(t)\mathbf{H}]\,\mathbf{P}^-(t) \quad (29)$$

Propagation では，時刻 $t-1$ の推定値から，状態遷移のモデル $\mathbf{F}$ を用いて，時刻 t での状態を予測する．Update では，状態遷移による予測値 $\hat{\mathbf{x}}^-(t)$ と，現在の推定値 $\mathbf{y}(t)$ とから，推定値の修正を行う．Kalman Gain, $\mathbf{G}(t)$ は，予測値 $\hat{\mathbf{x}}^-(t)$ と現在の推定値 $\mathbf{y}(t)$ との重みのような役割を果たす．行列 $\mathbf{P}(t)$ は，次式で定義される推定誤差の共分散である．

$$\mathbf{P}(t) = E\left[(\mathbf{x}(t) - \hat{\mathbf{x}}(t))(\mathbf{x}(t) - \hat{\mathbf{x}}(t))^T\right] \quad (30)$$

$\mathbf{Q}$ 及び $\mathbf{R}$ は，雑音 $\mathbf{w}(\mathbf{t})$ 及び $\mathbf{v}(\mathbf{t})$ の共分散である．これらは，未知であることが多く，推定値と予測値の重みを変えるパラメタとして利用される．

Figure 6: Scene of the experiment.

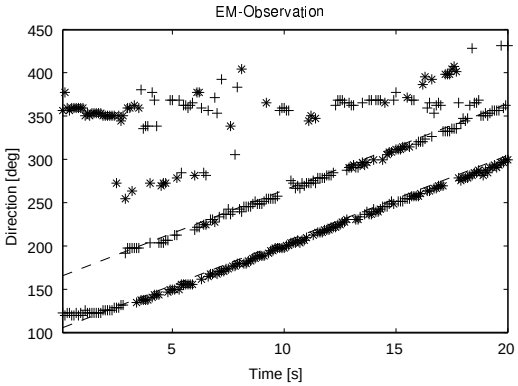
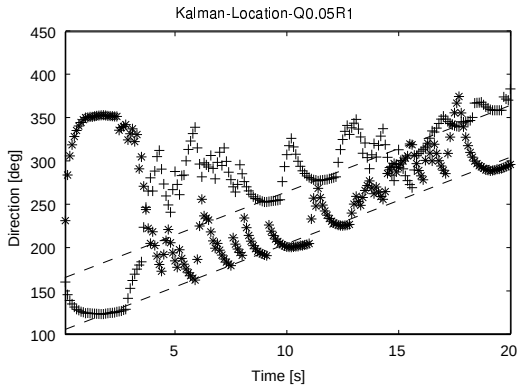


Figure 7: Trajectory estimated by the maximum likelihood method.

(a) R=1



(b) R=150

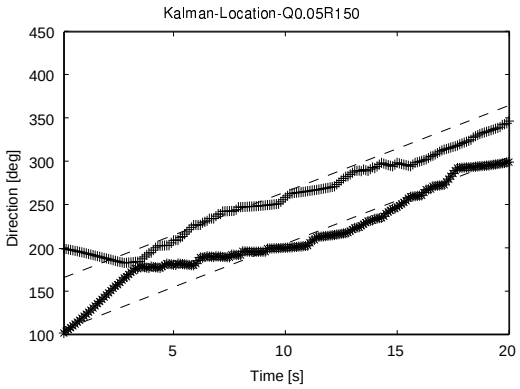


Figure 8: Trajectory smoothed by the Kalman filter.

4.3 適用例

ここで示す例では，産総研で開発したヒューマノイドHRP-2の頭部に搭載した8個のマイクロホンを使用した．音源は、ラウドスピーカ及び人であり，音源自体の位置は固定し，ロボットの頭部をターンテーブル (日東紡音響 TT-1000) に載せて，等角速度運動をさせた．図6に，実験の風景を示す.

図7は，最尤推定により，音源2個の位置推定を行った結果である．最尤推定では，5.1 節で示すように，(4) を各周波数で評価し，これを周波数軸上で積分して，最終的な尤度を算出した．点線は，設定速度 (10° /sec) を元に算出した音源の軌跡である．両者を比較すると，かなりの部分は，音源の軌跡と一致するが，軌跡以外の部分に推定値が分散している部分もある．これは主に，壁面からの反射や，音源のパワーが弱い部分（音声の子音など）での推定誤りと思われる．

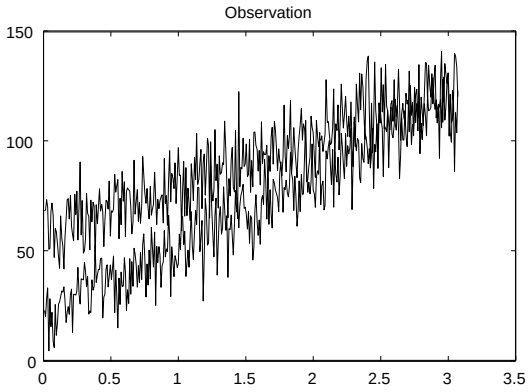
図8は，Kalman Filter による Smoothing の結果である．$R = 1$ の場合は，推定値に対する重みが大きく観測値をトレースするような結果となっている．一方，$R = 150$ の場合は，予測値に対する重みが大きくなり，軌跡は滑らかになっている．軌跡の真値とのずれはあるが，この例にかぎって言えば，Smoothing により，音源の大まかな軌跡は，ある程度推定され，時々観測される大きな推定誤差の影響は取り除かれている．この大まかな軌跡を初期値として，EM アルゴリズムなどを適用すれば，されに音源位置の推定精度が向上する場合もあろう.

図9は，人工的に作成したデータに Kalman Filter を適用した例である．観測雑音 $\mathbf{v}(t)$ は，ガウス雑音である．この場合は，データの分散は大きいが，実データの場合と比較して，良好な smoothing が得られている．この差は，主に雑音の分布のミスマッチによるものと考えられる.

4.4 Kalman Filter の拡張

Kalman Filter では，状態遷移 $\mathbf{F}$ を線形，雑音 $\mathbf{v}(\mathbf{t})$ 及び $\mathbf{w}(\mathbf{t})$ をガウス雑音と仮定することにより，定式化を大幅に簡単化している．しかし，問題によっては，これらの仮定で必ずしも十分でない場合もある．例えば，上述の例のような等速度運動では，線形の状態遷移で十分であるが，音源のように運動そのものは等速度運動していても，喋ったり喋らなかったりというように観測値が断続的になる場合は，線形の状態遷移では不十分である．また，同時に発話する人数が変化するような場合は，誤差分布も，単純なガウス分布では不十分な場合もある．画像やレーダ/ソナーなどの分野でも同様の問題があり，Kalman Filter は，様々な拡張をされてきた．表1に，その一部をまとめる．EKF では，非線形な状態遷移を1次の Taylor 展開を用いて線形化 (Linearliation) している．UKF 及び PF

(a) Observation



(b) Smooted by Kalman Filter

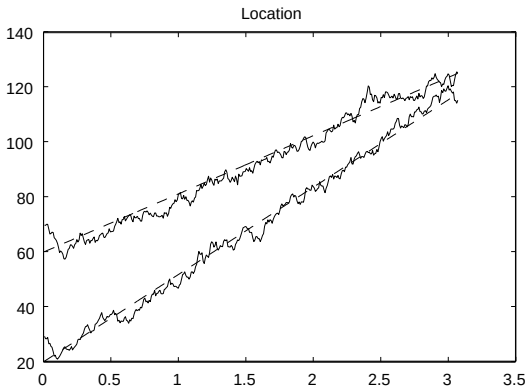


Figure 9: Kalman filter smoothing for artificial data.

表 1: Extention of Kalman filter.

	状態遷移	分布
Kalman Filter	線形	ガウス
Extended Kalman Filter (EKF)	非線形	ガウス
Unscented Kalman Filter (UKF)	非線形	ガウス
Particle Filter (PF)	非線形	非ガウス

は，サンプリング法と呼ばれ，分布上の任意（または特定）の点を選んで，状態遷移により予測値を推定し，これらの予測値から新しい時刻での分布を再構成するため，任意の状態遷移を用いることができる．さらに PF では，分布自体も任意の形状を選ぶことができる．このように自由度が増加する反面，計算量も増大するというトレードオフもある．

5 Particle Filter

5.1 状態遷移確率と尤度

ここでは，Kalman Filter における話を一般化して，一定の観測時間内 $[1 : T]$ における観測値 $\mathbf{Y}_{1:T} = [\mathbf{Y}(1), \cdots, \mathbf{Y}(T)]$ と，これに対応する内部状態（音源の位置や速度などのパラメタ）$\mathbf{X}_{1:T} = [\mathbf{X}(1), \cdots, \mathbf{X}(T)]$(観測不可) があるものとする．今，欲しいのは，観測値に対

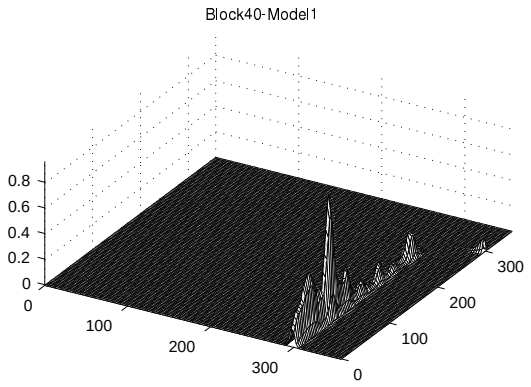


Figure 10: An example of likelihood function.

する内部状態の事後確率 $P(\mathbf{X}_{1:T}|\mathbf{Y}_{1:T})$ である．事後確率は，同時分布 $P(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T})$ が分かれば，ベイズの定理を用いて，計算可能である．ここでは，この同時分布が，

$$P(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}) = \prod_{t=1}^{T} P(\mathbf{Y}(t)|\mathbf{X}(t))P(\mathbf{X}(t)|\mathbf{X}(t-1)) \tag{31}$$

のように，状態遷移確率 $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ と，観測値に対する尤度 $P(\mathbf{Y}(t)|\mathbf{X}(t))$ の積に分解できるものと仮定する．

このうち，状態遷移確率 $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ については，Kalman Filter の例で示した線形な等速度運動のモデルなどのように，内部状態の遷移に関する先見的な知識に基づいて与えるのが一般的である．一方，尤度 $P(\mathbf{Y}(t)|\mathbf{X}(t))$ については，内部状態を表すパラメタから，観測信号のモデルを構成し，このモデルにより，観測信号がどの程度説明できるのかを，評価する．本報告では，2.2 節で示した尤度関数 (4) を用いる．この式から分かるように，観測された生データに直接フィッティングを行うのではなく，データを共分散行列の形にまとめ，サンプル共分散（観測値）とモデル共分散とのフィッティングを行う．

(4) では，尤度は，音源の方向及びパワーの関数となっているが，パラメタ空間の次元を減らすため，本報告では，音源のパワーについては，(13) と同様，次式の遅延和法により推定を行い、方向の関数となるようにしている．

$$\hat{\gamma}_l = \frac{\mathbf{a}^H(\theta_l)\mathbf{C}_y\mathbf{a}(\theta_l)}{|\mathbf{a}(\theta_l)|^4} \tag{32}$$

さらに，広帯域信号に対応するよう，周波数ごとに算出される尤度の積を用いている．

$$\bar{L}_y(\Theta) = \prod_{\omega=\omega_L}^{\omega_H} L_y(\Theta(\omega); \mathbf{Y}(\omega, t)) \tag{33}$$

図 10 に，尤度関数の例を示す．この例では，音源数 $L = 2$ であるので，$\Theta = [\theta_1, \theta_2]$ となり，尤度関数は 2 次元となる．

5.2 Particle Filter のアルゴリズム

上述の尤度関数をパラメタ Θ がとりうるすべての組み合わせについて計算すれば，同時分布 (31) を計算するこ

とができる．しかし，これには多大な計算コストを必要とする．例えば，音源 2 個、周波数帯域 [800,3000]Hz の場合に，２０秒程度のデータについて尤度を計算すると，Matlab（なのであまりあてにはならないが）と Pentium IV 3.6GHz 程度のマシンで，約 1 日の計算を必要とする．Particle Filter では，粒子 (particle) により，尤度関数をサンプリングし，このサンプリングした点のみにおいて式 (33) を評価することにより，計算の効率化を図っている．ただし，闇雲にサンプリングしたのでは，尤度関数の良好な近似とはならないので，尤度関数のうち，重要な点（尤度の高い点）の近傍をサンプリングする戦略をとる (Importance Sampling).

ここでは，[9] に従い，図 11 を用いて，最も簡単な Particle Filter のアルゴリズムを紹介しておく．まず，時刻 $t-1$ において，各粒子が，(a) の位置にあるものとする．(b) では，尤度関数をサンプリングし，サンプルした尤度を粒子の重みとする．尤度の高いものほど，重みは大きくなる．図中では，黒丸の大きさが重みの大きさを表している．(c) では，この重みに比例して，粒子の増殖/淘汰を行う．この増殖/淘汰により，分布の重要な部分で密なサンプリングが行われる．(d) では，状態遷移確率を用いて，時刻 t での粒子の状態を推定する．(e) では，新たな粒子位置での尤度関数のサンプリングを行う．このループを時刻 t を増加させながら繰り返す.

音源のトラッキングでは，音源が断続的であるため，音源の軌跡を見失うことが必然的に起こる．このような場合，尤度関数の分布は平坦になり，このため，それまで尤度の高いところに集中していた粒子が，パラメタ空間全域にばらまかれることになる．したがって，見失った軌跡を再発見する確率も増す．このようなメカニズムにより，断続的な音源のトラッキングに有効ではないかと期待している.

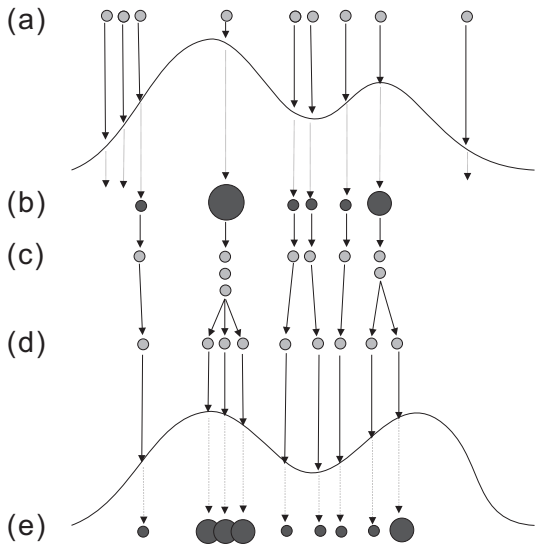


Figure 11: Explanation of particle filter algorithm.

5.3 適用例

実験環境は，4.3 と同様である．マイクロホンアレイは，ヒューマノイド HRP-2 の頭部にマウントされている 8 素子のものを用いた．このロボット頭部を等速度で回転させることにより，音源を相対的に移動させた．音源は，話者及びラウドスピーカであり，話者は，断続的に発声している．ラウドスピーカからは，定常的に音楽が流れている.

尤度の計算では，音響情報からの尤度 (33) に加え，画像に対する尤度 (ある方向の人らしさ) も加味し，音響情報と画像情報の統合を行った．図 12(a) は追跡対象（人とラウドスピーカ) の存在確率，(b) は発音対象の存在確率，(c) は人発話の存在確率 (最終的に求めたいもの) を示している．詳しくは，[10] 参照されたい.

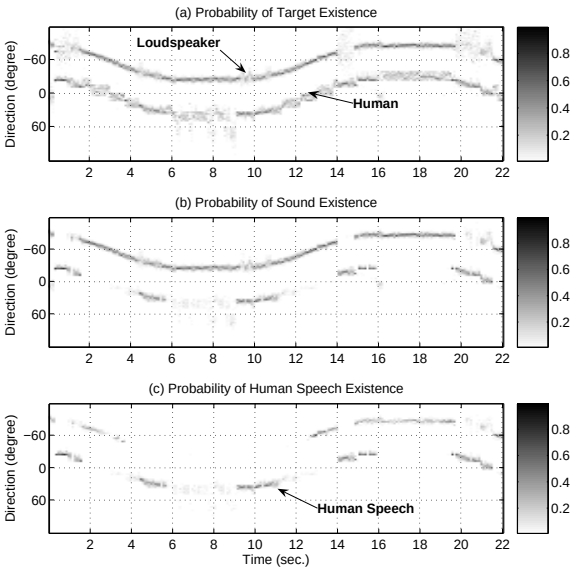


Figure 12: Estimation results by particle filter.

参考文献

[1] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *to appear in J. Applied Signal Processing*, 2004.

[2] K. Nakadai *et al.*, "Real-time auditory and visual multiple-object tracking for humanoids," in *Proc. IJCAI2001*, 2001.

[3] ," http://www.fusion2002.org.

[4] Meir Feder and Ehud Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 36, no. 4, pp. 477–489, 1988.

[5] Michael Miller and Daniel Fuhrmann, "Maximum-likelihood narrow-band direction finding and the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 38, no. 9, pp. 1560–1577, 1990.

[6] Futoshi Asano and Hideki Asoh, "Sound source localization and separation based on the em algorithm," in *Proc. SAPA2004*.

[7] 有本卓, **カルマンフィルター**, 産業図書, 1977.

[8] Simon Haykin, Ed., *Kalman filtering and neural networks*, Wiley Inter-science, 2001.

[9] Arnaud Doucent, Nando de Freitas, and Neil Gordon, *Sequential Monte Carlo methods in practice*, Springer, 2001.

[10] 麻生英樹，浅野太，山本潔, "パーティクルフィルタを用いた人発話の追跡," in *SI2004* **講演論文集**. 計測自動制御学会.

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-2 (12/6)

128 チャンネルスピーカーアレイによるサウンドスポット形成

Sound Spots Generation by 128-Channel Large Scale Speaker Array

溝口 博 [1,2]，玉井裕樹 [1,2]，加賀美聡 [2,3,1]，鳥羽高清 [1]，長嶋功一 [4]，高野太刀雄 [2]
Hiroshi Mizougchi[1,2], Yuki Tamai[1,2], Satoshi Kagami[2,3,1], Takakiyo Toba[1], Koichi Nagashima[4], and Tachio Takano[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，[4]R-Lab.
[1]Tokyo University of Science [2]Digital Human Research Center, AIST [3]JST [4]R-Lab. Inc.
[1]e-mail:hm@rs.noda.tus.ac.jp

Abstract

This paper presents a novel sound interface for *HAL*-like environmental man-machine system. The interface consists of multiple loud speakers. It utilizes and controls interference phenomenon of sound wave to concentrate sound, voice or music at multiple spot like small areas in the environment. The are is called *sound spot*. Location of these sound spots can be specified arbitrarily. The authors have implemented 128-channel large scale speaker array as a prototype of the interface. The current implementation has succeeded to transmit four different sound contents at four locations simultaneously. In other words, up to four people can separately enjoy to listen their own contents simultaneously even if they are in the same room.

1 はじめに

スタンリーキューブリック監督の映画「2001 年宇宙の旅」に登場するコンピュータ HAL9000 は，既に三十数年前の時点で，いわゆるユビキタスなコンピューティング環境として構想された先駆的なものであった．映画の中の HAL は最後には人間に反乱を起こしてしまう否定的な存在ではあるが，HAL の基本的な機能そのものは，人間と機械の共生システムの観点からみて，いまだに魅力的な存在である．HAL は視覚と聴覚を介して人間の行動を把握可能であり，人間と音声を介してコミュニケーションをはかることが可能である．ある意味では，人間と機械との共生環境の究極形の一種であると言えよう．

映画の中の架空の存在であった HAL を志向する環境型システムに関し，近年，いくつかの研究が現れてきている．典型例として，MIT CSAIL(旧 AI Lab.)の Intelligent Room [1][2]や AIRE [3][4], Media Lab.の Smart Rooms[5], 東京大学の Robotic Room [6]や Intelligent Space[7], 産総研(旧電総研)の SELF[8][9]や Enabling Environment[10], ジョージア工科大の Aware Home [11][12], Microsoft Research の Easy Living[13][14], 産総研の Learning by Doing Project [15][16]や MIT ホームオブザフューチャー・コンソーシアムの Home_*n* [17][18]などが挙げられる．これらのシステムの殆どは，一人の人間としかやりとりできない．中には二人以上の人間を相手にできるシステムもあるが，その場合でも，やりとりのコンテクストは一つだけに限定される．したがって，複数の人同士がそのコンテクストを共有して対話に参加することになる．上記のどのシステムも，同時に複数の人をそれぞれ別々に相手をすることはできない．

これら HAL 型環境の大きなメリットは，システムと人間との間で，文字通りハンズフリーなコミュニケーションが期待できる点にある．音声インタフェースは，そのような自然なコミュニケーションにとって最も有望でかつ有効な手段の候補たり得る．ただし，それはあくまでその環境内に一人しか人間が存在せず，コンテクストが一つの場合においてのことである．そのような場合にはうまく機能し，環境内の人間はハンズフリーなコミュニケーションを享受できる．しかし，環境内に複数の人間がいて，同時にそれぞれが独立した別々のコンテクストで対話を交わそうと欲すると，当然のことながら干渉が生じてしまい深刻な問題が生じる．完全ハンズフリーなコミュニケーションと，個別のコミュニケーションとを，同時に両立させることは不可能である．

そこで著者らは，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感度・高音圧分布「サウンドスポット」を作り出し，S/N 比の高い集音や伝送を実現，たとえその人々が動いてもサウンドスポットを追従させることが可能な技術の研究開発に取り組んでいる．「サウンドスポット」とは，その領域内だけで音が聞こえる，約 300mm 径程度の小さな円形領域のことである．具体的には，マイクロホンやスピーカーを多数並べたアレイにより，サウンドスポットを形成する．Fig. 1 に複数サウンドスポット同時形成のイメージを示す．

これまでにスピーカー128 台から成る大規模スピーカーアレイを構築し，それを用いて複数のサウンドスポット形成に成功した．しかも，個々のスポットの内容音声は独立で別々である．すなわち，同時に複数の人の「耳元で」それぞれ別のコンテクストで「語りかける」ことを可能とした．以下では構築したスピーカーアレイの実現技術とスポット形成技術について述べる．

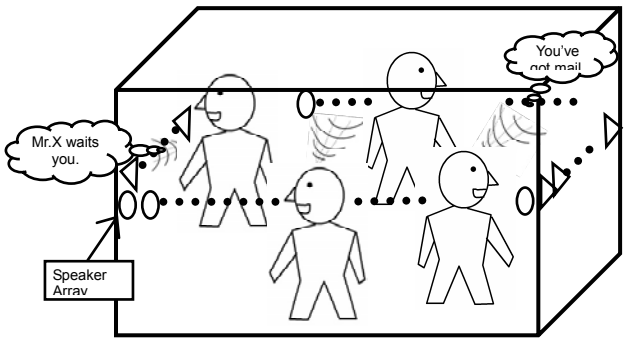


Figure 1: Image of simultaneous sound spots generation

2　サウンドスポット形成

図２に，サウンドスポットの基になっているサウンドビーム形成の概念を図示する．複数のスピーカーから同時に同じ音を出力した場合，焦点に到達する各スピーカーからの音は，焦点までの距離の相違により，それぞれ振幅も位相も異なる．焦点から各スピーカーまでの距離差から生じる音の到達時間差と減衰比とを求め，それらを補償する形の付加遅延時間と振幅とを各スピーカーの出力信号に付与する．これによって焦点位置における各スピーカーからの音の波の位相と振幅を一致させ，互いに強め合うようにする．この結果として, Fig. 3 左図に示すように, 一直線アレイの場合は, ビーム状の高音圧分布「サウンドビーム」が形成される．さらに４直線のアレイを直交させ正方形状に配置することにより, Fig. 3 右図に示すようなサウンドスポットが形成可能である．

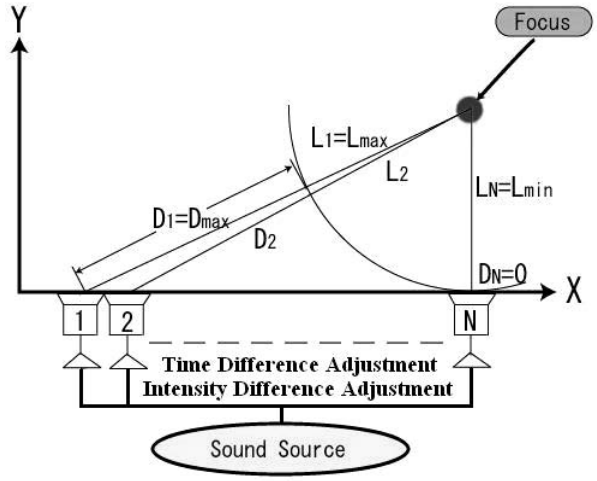


Figure 2: Concept of sound beam forming

3　シミュレーション

スピーカーアレイの構築に先立ち，シミュレーションを実施して前章の考え方の有効性と実現可能性の確認を行った．このシミュレーションでは，スピーカーアレイによって形成される音場の音圧分布マップを作成する. シミュレーションは, ６つの異なる周波数 125Hz, 250Hz, 500Hz, 1000Hz, 2000Hz, 4000Hz を用いる．議論を簡素化するため，音源は点音源であると仮定する．

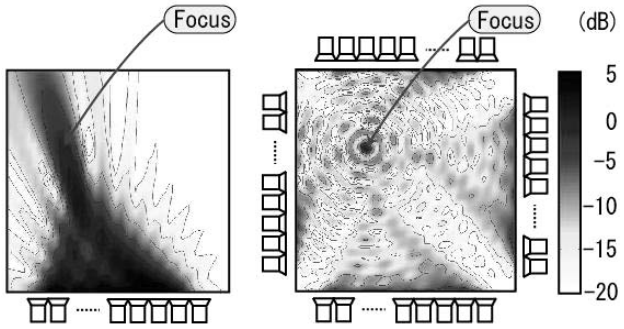


Figure 3: Images of sound beam and spot forming

3.1 音圧分布の算出式

Fig. 4 に直交２直線状スピーカーアレイの座標系を図示する．正方形状配置のスピーカーアレイは，このような直行２直線状アレイの重ね合わせとして取り扱えるため，直行２直線状アレイのシミュレーションができれば，正方形状アレイのシミュレーションも実現できる．したがって，ここでは直交２直線状アレイについての考察を行う．Fig. 4 中，x 軸上のスピーカーには Mi，y 軸状のスピーカーは Mj と番号付けられている．単純化のため, スピーカーは点音源であると仮定する．F(xf, yf)は焦点を，S(xs, ys)は音圧の測定点である．焦点から i 番目のスピーカーまでの距離を RFi とする．同様に測定点から i 番目のスピーカーまでの距離を RSi で表す．i 番目のスピーカーから発射された音 si(t)を，周波数 f, 振幅 ai, 位相 bi として次式のように表わす．

$$s_i(t) = a_i \sin 2\pi f(t - b_i) \qquad (1)$$

同様に j 番目のスピーカーから発射された音を sj(t)とする．

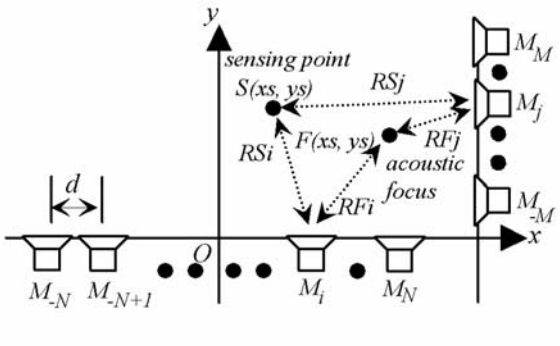


Figure 4: Coordinate system of orthogonal two lines speaker array

測定点 *S(xs, ys)*では, 各スピーカーから伝播した音の波が重なり合って合成波が得られる．ここで i 番目のスピーカーから伝播した音を xi(t)とすると，次式のよ

うに表される．

$$xi(t) = \frac{a_i}{RSi} \sin 2\pi f (t - b_i - \tau_i) \tag{2}$$

ここでτ_iは i 番目のスピーカーから測定点までの音の到達時間である．τ_iは次のように表される．

$$\tau i = RSi / v \tag{3}$$

ここで v は常温での音速である．(2 式より i 番目のスピーカーからの音の振幅が距離に反比例して減衰することが判る．同様に j 番目のスピーカーから伝播した音を xj(t)で表す．

いま，焦点 $F(x_f, y_f)$から最も遠いスピーカーまでの距離を RFmax.とする．他のスピーカーからの音は，$(RF\text{max} - RFi)/v$ 秒だけ早く焦点に到達し，位相差となる．したがって，各スピーカーの音を，$(RF\text{max} - RFi)/v$ 秒だけ遅らせて発射すれば，焦点 $F(x_f, y_f)$で位相が揃うことになる．また，i 番目のスピーカーの音を RFi 倍すれば，焦点 $F(x_f, y_f)$で各スピーカーからの音の振幅も揃うことになる．

このような付加遅延と振幅を各スピーカーの出力信号に与えた場合の，測定点 S での合成波 $y(t)$ は次のように表される．

$$y(t) = \sum_{i=-N}^{N} RF_i x_i(t) + \sum_{j=-M}^{M} RF_j x_j(t) \tag{4}$$

この式を展開すると以下のようになる．

$$\begin{aligned} y(t) &= \sum_{i=-N}^{N} \frac{RF}{RS_i} \sin 2\pi f (t - \frac{RS_i + RF_i \max - RF_i}{v}) \\ &\quad + \sum_{j=-M}^{M} \frac{RF_j}{RS_j} \sin 2\pi f (t - \frac{RF_j + RF_j \max - RF_j}{v}) \\ &= \left(\sum_{i=-N}^{N} \alpha_i(xs, ys; xf, yf) + \sum_{j=-M}^{M} \chi_j(xs, ys; xf, yf) \right) \sin 2\pi f t \\ &\quad + \left(\sum_{i=-N}^{N} \beta_i(xs, ys; xf, yf) + \sum_{\text{j=-M}}^{M} \delta_j(xs, ys; xf, ys) \right) \cos 2\pi f t \end{aligned} \tag{5}$$

ただし，$\alpha_i, \chi_i, \beta_i, \delta_i$は以下のとおりである．

$$\alpha_i(xs, ys; xf, yf) = \frac{RF_i}{RS_i} \cos 2\pi f (\frac{RS_i + RF_i \max - RF_i}{v}) ,$$

$$\beta_i(xs, ys; xf, yf) = \frac{-RF_i}{RS_i} \sin 2\pi f (\frac{RS_i + RF_i \max - RF_i}{v}) ,$$

$$\chi_j(xs, ys; xf, yf) = \frac{RF_j}{RS_j} \cos 2\pi f (\frac{RS_j + RF_j \max - RF_j}{v}) ,$$

$$\delta_j(xs, ys; xf, yf) = \frac{-RF_j}{RS_j} \sin 2\pi f (\frac{RS_j + RF_j \max - RF_j}{v}) .$$

式(5)を更に整理すると式(6)となる．

$$\begin{aligned} y(t) &= A(xs, ys; xf, yf) \sin(2\pi f t + B(xs, ys; xf, yf)) \\ A(xs, ys; xf, yf) &= \sqrt{(P+Q)^2 + (R+S)^2}, \\ B(xs, ys; xf, yf) &= \tan^{-1}\left(\frac{R+S}{P+Q}\right), \end{aligned} \tag{6}$$

ただし，

$$P = \sum_{i=-N}^{N} \alpha_i(xs, ys; xf, yf), \quad Q = \sum_{j=-M}^{M} \chi j(xs, ys; xf, yf),$$

$$R = \sum_{i=-N}^{N} \beta_i(xs, ys; xf, yf), \quad S = \sum_{\text{j=-M}}^{M} \delta j(xs, ys; xf, ys)$$

ここで，式(6)の $A(xs,ys;xf,yf)$ は，焦点が $F(xf, yf)$として与えられた場合の，空間中の任意の点(xs,ys)における音圧の値を表す．したがって，(xs,ys)の値を変化させ，空間中の多数の点において $A(xs,ys;xf,yf)$を求めることにより，音圧の空間分布を得ることができる．

3.2 シミュレーション結果

Fig. 5 に 128 チャンネル正方形状配置のスピーカーアレイのシミュレーション結果を示す．正方形の一辺の長さは 3230mm である．図では，焦点における音圧を 0dB とした相対値で音圧分布を示している．

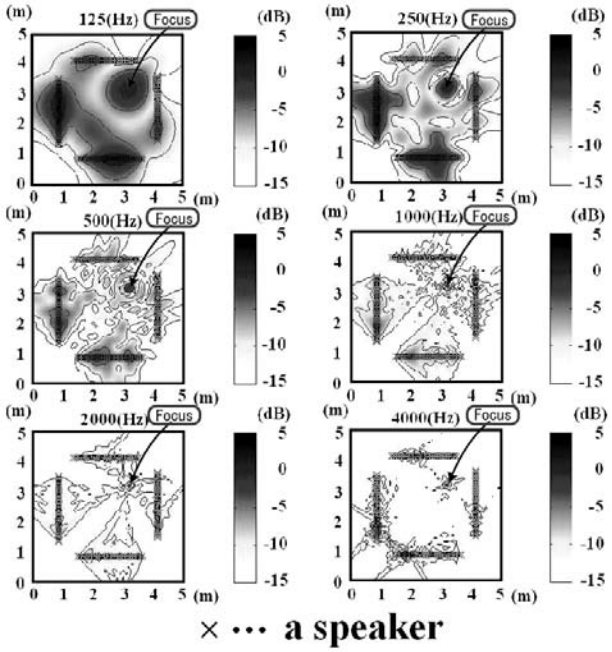


Figure 5: Simulation results

シミュレーション結果によれば，500Hz から 1000Hz の範囲ではサウンドスポットが効果的に形成されていることがみてとれる．500Hz 未満の低い周波数では，波長が長くなってしまうために，小さなスポットではなく，高い音圧の領域が大きく拡がってしまっている．また，高い周波数では，スポットが小さくなり過ぎるため，音の伝送は困難になることが予想される．ただし，上記のように音声帯域に相当する中域では適切なスポットが形成されるので，局所的な音声伝送が期待できる．効果的なスポット形成という意味では，低域と高域を除去するバンドパスフィルターが不可欠である．

4　128 チャンネルスピーカーアレイの構築

スピーカーアレイを実現をするためには，スピーカー群に供給すべき 128 チャンネル分の信号を同時にサンプリングする必要がある．しかも，CD 音質を得るためには十数 μ 秒といった短い周期で制御する必要がある．しかし，市販の DA 変換ボードは，たかだか 16 チ

ャンネル程度，しかも同時出力ではないものが殆どで，128 チャンネル同時出力といった仕様のものは存在しない．そこで著者らは，サンプリングレート 44.1KHz，即ち周期約 23 μ sec でのサンプリングで，128 チャンネルの信号が同時に出力可能な DA 変換ボードを新規に開発した．Fig. 6 に開発した 128 チャンネル同時出力 DA 変換ボードを示す．

また，システム構築の上では，十数 μ sec オーダでの周期の制御も不可欠である．このため著者らは，市販品も含め複数種類の実時間オペレーティングシステムを実動比較した．その結果，ART-Linux[20]のみが十数 μ sec オーダの等周期ループを安定して実行可能であることを発見し，これを採用した．他の実時間オペレーティングシステムでは，周期変動が数 μ sec から十 μ sec オーダに及ぶため，たかだか msec オーダの周期までしか安定して実現できない．ART-Linux を用いることで，44.1KHz の CD 音質のサンプリングレート実現がソフトウェアのみで可能となった．

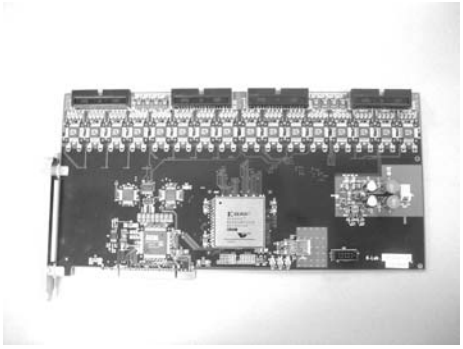


Figure 6: 128-channel D/A board

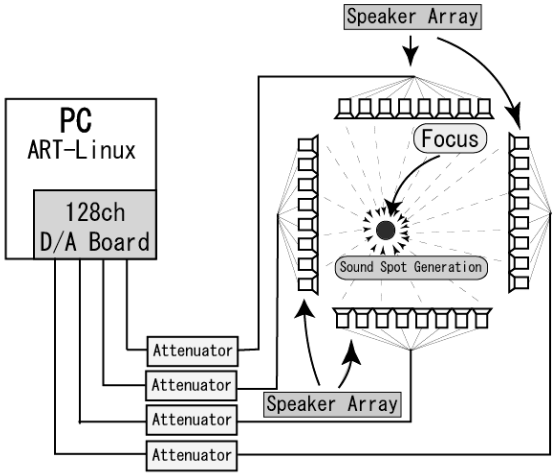


Figure 7: Block diagram of implemented speaker array

Fig. 7 に構築した 128 チャンネルスピーカーアレイシステムのブロック図を示す．上記の 128 チャンネル同時出力 DA 変換ボードと ART-Linux の採用により，1 台の汎用 PC 上のみで，128 台のスピーカー群の実時間制御が達成できている．Fig. 8 に構築したスピーカーアレイシステムの外観を示す．スピーカー間の距離は 70mm である．アレイの要素スピーカーには，YAMAHA YST-M10 を用いた．

Figure 8: 128-channel square speaker array

5 実験

構築したスピーカーアレイの評価は，実験を通じて実証的に行った．音圧分布の測定には，先端に音圧計を付けたコンピュータ制御のガントリークレーンを用いる．測定に際しては，格子点ごとに音圧を測定してゆくプログラムを開発した．Fig. 9 にガントリークレーンと音圧計の外観を示す．サウンドスポットの移動可能性については，直線状マイクアレイの各マイクからの出力信号の時間変化を用いて評価を行った．この目的のために 16 チャンネルの直線状マイクアレイを開発した．Fig. 10 に開発した 16 チャンネルマイクアレイの外観を示す．

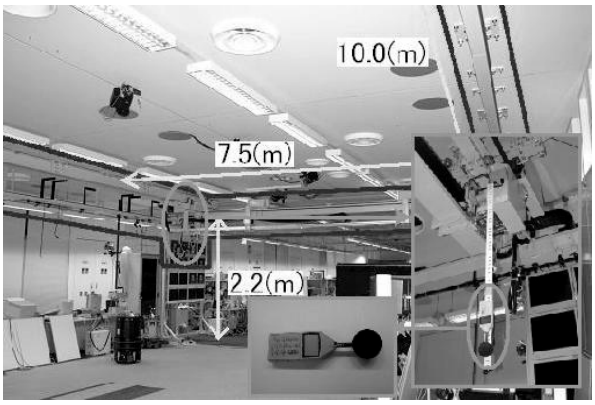


Figure 9: Gantry crane system and sound level meter

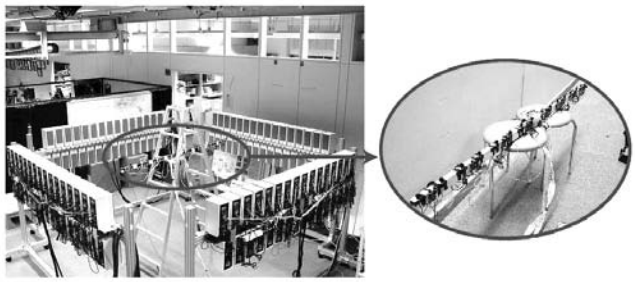


Figure 10: 16-channel linear microphone array

5.1 単一スポットの形成

Fig. 11 はサウンドスポットを一個形成した場合の音圧分布の実測値を示している．焦点の位置は座標(1.12, 1.12) (m) である．焦点付近の音圧が，他の部分の音圧より 10dB 以上大きくなっていることが判る．すなわち，焦点の箇所に，十分に満足できる抑圧比のサウンドスポットが形成されていることが確認できる．

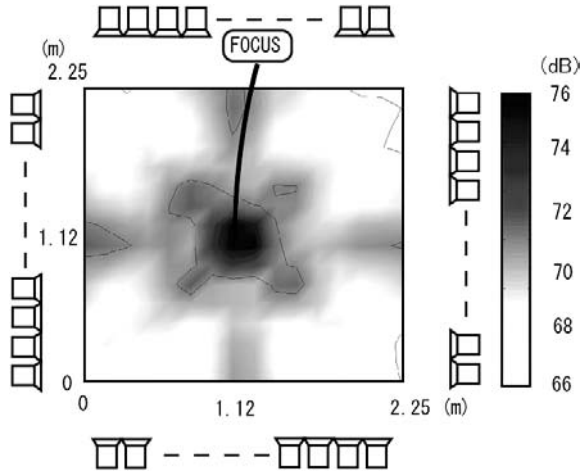


Figure 11: Experimental result of one sound spot forming

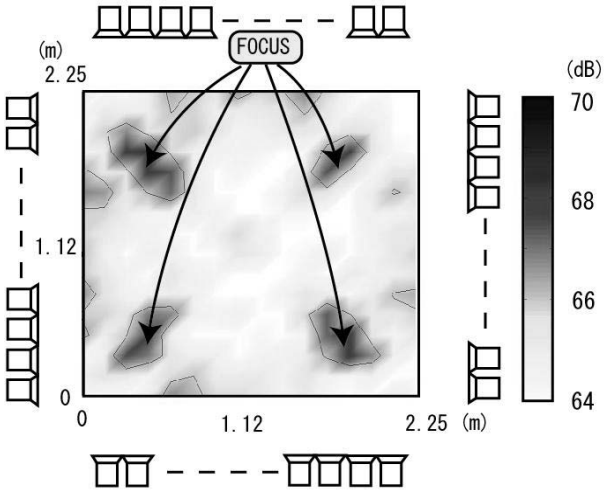


Figure 12: Experimental result of multiple sound spots forming

5.2 複数スポットの形成

Fig. 12 はサウンドスポットが複数個の場合の音圧分布の実測値を示している．４つの焦点の位置は，それぞれ(0.5, 0.5), (1.75, 0.5), (1.75, 1.75), (0.5, 1.75) (m)である．焦点付近の音圧は，周囲より約 5dB ほど高くなっている．したがって，４つの焦点の位置にそれぞれサウンドスポットが形成されていることが確認できる．この効果は測定値だけでなく聴感上でも確認でき，焦点の位置に立った時だけ，音がはっきり聞こえるので，その位置にサウンドスポットが形成されていることが判る．

5.3 スポットの移動可能性

Fig. 13 に，サウンドスポット移動可能性評価実験の，装置構成を示す．図に示すように，座標(0.5, 1.65)(m)から(2.5, 1.65)(m)まで直線状のマイクアレイが設置されている．初め，左端(0.5, 1.65)(m)の位置にあったサウンドスポットは，40 秒間かけて右端(2.50, 1.65)(m)の位置まで移動してゆく．マイクアレイの要素数は上述のように 16 である．各要素マイク間の間隔は等間隔で 120mm である．

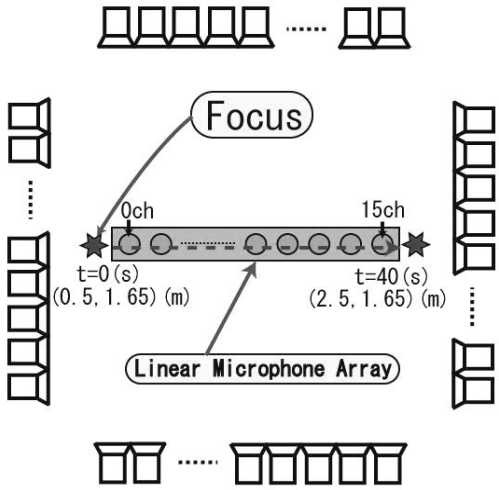


Figure 13: Configuration/Setup of experiment of sound spot movement

Fig. 14 にサウンドスポット移動可能性確認実験の結果を示す．この図は各マイクの出力電圧の，0.1sec ごとの時間平均値をプロットしたものであり，最大値を 0dB をした相対値で表している. 縦軸が x 軸の座標を，横軸が時間を示す．出力電圧最大を示すマイクの位置が，時間と共に推移してゆく様子がみてとれる．しかも，この軌跡が，サウンドスポットの移動と一致していることも確認できる. サウンドスポットそのものは，上述のとおりソフトウェア的に制御されている．ソフトウェアで付加遅延時間を変えてから，サウンドスポットの位置が変化するまでは 10msec オーダの時間で済む．本実験で示したようなゆっくりした動きであれ

ば，十分に追従可能である．

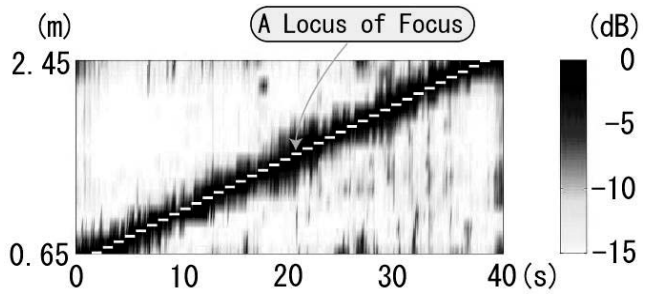


Figure 14: Experimental result of sound spot's movement

6 おわりに

本論文では，構築した 128 チャンネル大規模スピーカーアレイと，それを用いて行ったサウンドスポット形成に係わる実験について述べた．実験は単一のサウンドスポット形成と複数のサウンドスポット形成，およびサウンドスポットの移動可能性を示すものである．

実験を通じ，単一のスポットを介せば，特定の対象人物に対する音情報の選択的伝送が効果的に行えることを示した．周りがうるさい状況下では，いわゆるカクテルパーティー効果と結果において同等の効果が，コンテンツの意味内容とは独立の物理的現象として得られるものと期待できる．

実験ではまた複数個のスポットを同時に生成可能であることも示した．これにより，HAL 型の環境型システムに，同時に複数の独立したコンテクストで複数の人とやりとりできる手段を提供することになる．更に実験ではスポットの移動可能性も示した．追跡視覚やタグシステムと組み合わせれば，対象とする人物が移動しても，その人にスポットを追従させることができるものと期待できる．

謝辞

本研究の一部は文部科学省科学研究費補助金特定領域研究「情報学」の補助を受けて実施されたものである．記して謝意を表する．

参考文献

[1] M. Coen: "The Future of Human-Computer Interaction, or How I learned to stop worrying and love my Intelligent Room", IEEE Intelligent Systems, March/April 1999.

[2] M. Coen, L. Weisman, K. Thomas, and M. Groh: "A Context Sensitive Natural Language Modality for an Intelligent Room", Proceedings of International Workshop on Managing Interactions in Smart Environments (MANSE'99), pp.68-79, 1999.

[3] S. Peters and H. Shrobe: "Using Semantic Networks for Knowledge Representation in an Intelligent Environment", Proceedings of PerCom'03: 1st Annual IEEE International Conference on Pervasive Computing and Commnications, 2003.

[4] A. Adler and R. Davis: "Speech and Sketching for Multimedia Design", Proceedings of 2004 ACM International Conference on Intelligent User Interfaces (IUI 04), pp.214-216, 2004.

[5] A. Pentland: "Looking at People: "Sensing for Ubiquitous and Wearable Computing", Trans. on PAMI, Vol.22, No.1, pp. 107-119, 2000.

[6] T. Sato: "Robotic Room: Human Behavior Measurement, Behavior Accumulation and Personal/Behavior Adaptation by Intelligent Environment", Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003), pp.515-520, 2003.

[7] H. Hashimoto, J. H. Lee, and N. Ando: "Self-Identification of Distributed Intelligent Networked Device in Intelligent Space", Proceedings of the 2003 IEEE International Conference on Robotics and Automation (ICRA '03), pp.4172-4177, 2003.

[8] T. Hori, Y. Nishida, T. Suehiro, and S. Hirai: "SELF-Network : Design and Implementation of Network for Distributed Embedded Sensors", Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2000), pp. 1373-1378, Nov., 2000.

[9] Y. Nishida and T. Hori: "Noninvasive and Unrestrained Monitoring of Human Respiratory System by Sensorized Environment", Proceedings of the First IEEE International Conference on Sensors (Sensor 2002).

[10]Y. Nishida, H. Aizawa, T. Hori, N.H. Hoffman, T. Kanade, and M. Kakikura: "3D Ultrasonic Tagging System for Observing Human Activity", Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS2003), pp.785-791, Oct., 2003.

[11]C. D. Kidd, R. J. Orr, G.. Abowd, C. G. Atkeson, I. Essa, B. MacIntyre, E. Mynatt, T. Starner, and W. Newstetter: "The Aware Home: A Living Laboratory for Ubiquitous Computing Research", Proceedings of the Second International Workshop on Cooperative Buildings - CoBuild'99, Position paper, 1999.

[12]G.. Abowd, G.. A. Bobick, I. Essa, E. Mynatt, and W. Rogers: "The Aware Home: Developing Technologies for Successful Aging", Proceedings of AAAI Workshop and Automation as a Care Giver, 2002.

[13]B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer: "EasyLiving: Technologies for Intelligent Environments", Proceedings of International Symposium on Handheld and Ubiquitous Computing, 2000.

[14]B. Brumitt, J. Krumm, B. Meyers, and S. Shafer: "Ubiquitous Computing and Role of Geometry", IEEE Personal Communications, August 2000.

[15]Y. Nishida, K. Kitamura, H. Aizawa, T. Hori, M. Kimura, T. Kanade, and H. Mizoguchi: "Real World Sensorization for Observing Human Behavior and Its Application to Behavior-To-Speech", Proceedings of 2004 ACM International Conference on Intelligent User Interfaces (IUI 04) , pp. 289-291, Jan. 2004.

[16]M. Hiramoto, Y. Nishida, F. Kusunoki, and H. Mizoguchi: "Learning by Doing: Assist of Foreign Language Learning through a Sensorized Environment", Proceedings of the 10th International Conference on Virtual Systems and Multimedia (VSMM2004), Nov. 2004. (to appear)

[17]S. S. Intille: "Designing a Home of the Future", IEEE Pervasive Computing, April-June 2002, pp.80-86, 2002.

[18]S. S. Intille and K. Larson: "Designing and Evaluating Supportive Technology for Homes", The 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM2003) Conference Digest, pp. xvi-xxii, 2003.

[19]Y. Tamai, S. Kagami, H. Mizoguchi, and K. Nagashima: "Simultaneous Forming/Capture of Multiple Focuses Sound Beams", Proceedings of 2003 IEEE International Conference on Systems, Man, and Cybernetics (SMC'03), pp.4613-4618, 2003.

[20]ART-Linux
http://www.movingeye.co.jp/mi6/art-linux_feature.html

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challnege-0420-3 (12/6)

ロボットによる音源定位のための人工耳介

Artificial Pinnae for Sound Localization by Robots

公文誠, 下田倫子, 神澤龍市, 水本郁朗, 岩井善太
Makoto Kumon, Tomoko Shimoda, Ryuichi Kohzawa, Ikuro Mizumoto and Zenta Iwai
熊本大学
Kumamoto University
kumon@gpo.kumamoto-u.ac.jp

Abstract

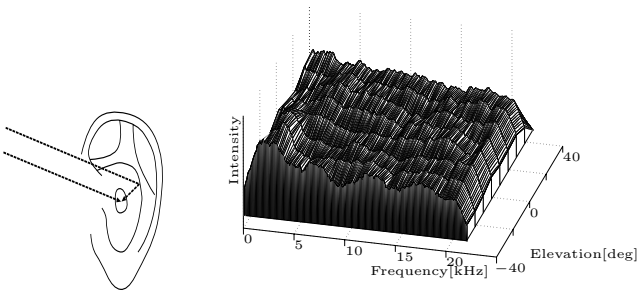
It is important for auditory robots to localize the sound source. Authors proposed an adaptive audio servo system which made a robot with two microphones direct to the sound source in horizontal plane. In order to extend this method to the sagittal plane, spectral cues by pinnae are considered in this paper. The shape of the artificial pinna which plays an important role to cause spectral cues is discussed. A simple physical acoustic model is used to design the shape and spectral responses of the designed pinnae are investigated.

1 はじめに

ロボットが人間との対話を行う際などに音源方向にロボットを向けることはマイクロフォンなど音情報を取得する部位を有効に利用する点で利点があり, 音を活用するロボットにとって必要な能力である. 音源定位を行う方法は種々提案されており, マイクロフォンアレーにより音源の位置を推定し, 推定された音源方向にロボットを動作させる逐次的なアプローチが一般的に用いられている. 例えば, 中島 [中島, 2004] らは耳介を有する 2 つのマイクロフォンとカメラを有するロボットにおいてまず音信号を受聴し, その信号から音源方向にロボットを駆動する制御指令値を推定するニューラルネットワークの学習方法を提案している. 一方, 著者 [Kumon, 2003] らも最も簡単なマイクロフォン 2 つを用いた構成を考え, リアルタイムに水平方向の音源にロボットを向ける制御手法として両耳間強度差 (IID) をもとにした適応オーディオサーボ系を提案している. 本報告ではこの手法を上下方向の音源にも拡張することを目的とし, その基礎的検討結果について報告する. この目的を実現する方法としてマイクロフォン外部に人工耳介を取りつけ, 耳介における音源方向による影響をスペクトルキュー (周波数領域での情報) として利用することを考える. スペクトルキューは耳介の形状に強く影響を受けるため, 音源の方向に合わせて適切にスペクトルキューを生じる耳介の構成を設計する必要がある. そこで本研究では耳介の形状を中心に検討することとした.

耳介において反射した音波は耳孔に直接到達する音波と干渉するが, 耳介が上下方向に対して非対称な形状をしているため, 耳孔において節や腹になる周波数は音波が耳に進入してくる方向に応じて変化する. この現象は, 人が音源の上下方向を判断する際に利用されている音の特徴の一つであると考えられている. Shaw[Shaw, 1968] は近傍の音源からの音が耳介において変調を受けていることを実際の実験において示した. Hebrank[Hebrank, 1974] はこの現象を説明する物理モデルとして Simple-Delay-and-Add(以下 SDAA) を提案し, 受聴される音の周波数特性 (振幅) と音源方向の関係を定性的に示した. また, Lopez-Poveda[Lopez-Poveda, 1996] は耳介からの反射波との干渉をより詳細に検討し, Gardner[Gardner, 1997] の KEMAR の結果を良く説明する物理モデルを提案している. ところで一般に人間の耳介の形状は複雑であるが, ロボットの音源定位に耳介を応用することを考えた場合, スペクトルキューの本質的な構造を捉えれば十分である. そこで, 本研究では Lopez-Poveda[Lopez-Poveda, 1996] の方法を簡素化したモデルを用いて, ロボットに用いる単純な人工耳介を設計・製作し, その特性を実際の実験によって検証する.

本報告ではまずスペクトルキューについて簡単に説明し, その後実際に設計した人工耳介について示し, その実験結果を紹介する.



(a) Pinna (b) HRTF of KEMAR[Gradner, 1994]

図 1: Schematic diagrams of Pinna Notch

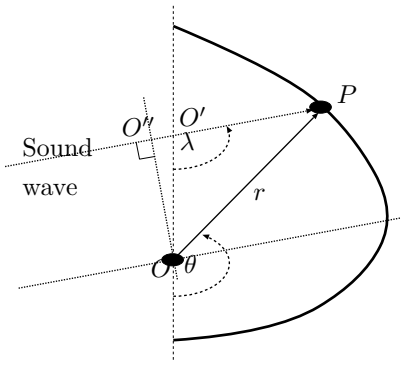


図 2: Proposed Pinna

2 スペクトルキュー

この節ではスペクトルキューについて簡単に説明する. 主に文献 [Garas, 2000] による.

不規則な形状をした耳介において反射した音波は耳孔に直接到達する音波と干渉するため, 音源からの音はその到来する方向に応じて変調される. つまり, 複雑な形状を有する耳介からの反射音によって引き起こされる耳孔での干渉は, 外耳を一種の音響フィルタとして機能させることになり, 音源の方向は周波数空間における特徴量として捉えることが出来る (図 1(a)). これをスペクトルキューと呼ぶ. 音波の干渉の結果, 周波数によって振幅特性がピークやノッチを形成する. ピークに比べノッチの方が音源方向との関係がはっきりとしているので, ノッチの周波数について着目することが多く, 特に第 1 ノッチのことを耳介ノッチ (pinna notch) と呼ぶ.

Shaw[Shaw, 1968] は実際の実験においてスペクトルキュー確認し, また, Gardner[Gradner, 1994] はダミーヘッド (KEMAR) において同様の実験を行っている. 耳介ノッチの例として KEMAR の振幅特性を図 1(b) に示す. 縦軸が振幅, 横軸が周波数, 奥に向かって音源の高さが高くなるよう示している. 下向きでは 4kHz 付近にあった耳介ノッチが音源が高くなるにつれ低い周波数に移動し途中で消え, また他のノッチが現われている. また 2kHz 付近に音源方向に依らないなだらかなピークが生じていることも分かる.

この現象を説明する簡単なモデルとして Hebrank[Hebrank, 1974] は SDAA を提案している. これは音源と耳孔を通る直線上での音波の干渉に限定してスペクトルキューを考えるもので, 耳介の一点だけの反射を考慮する簡単なものである. このモデルではノッチが生じる周波数 F_n は次の式で与えられる.

$$F_n(\lambda) = \frac{V}{4d(\lambda)}, \tag{1}$$

ただし λ と V, $d(\lambda)$ はそれぞれ音源の方向, 音速, 耳孔から耳介の後壁までの距離を表すものとする. (1) はノッチの周波数と音源方向の間の定性的な関係を示すことが出来るが, 耳介全体での反射の効果を考慮していないためその精度はあまり良くない. これに対し Lopez-Poveda[Lopez-Poveda, 1996] は耳介全体での反射と回折を考慮し, 耳孔での音圧をより正確に求める回折・反射モデルを提案している. Lopez-Poveda はこのモデルを用いて KEMAR における耳介ノッチ (および高次のノッチ) やより人間の耳介に近い形状の耳介でのスペクトルキューなどを推定することに成功している.

これらの研究から, 耳介の形状はスペクトルキュー, あるいは耳介ノッチを形成する周波数に密接な関連があることが示唆されるが, ロボットへ耳介の効果を利用しようと考える場合, 耳介の形状は十分に注意して設計すべき要素であると考えられる. そこで次節では 回折・反射モデルを簡略化したモデルによってシンプルな構造の耳介において耳介ノッチの特性について考察する.

3 人工耳介

3.1 耳介の形状

耳介の形状を耳孔を原点とする極座標で次のように表現するものとする.

$$r = r(\theta),$$

ここで r は耳孔から耳介上の角度 θ の点までの距離を表す (図 2). スペクトルキューを生じるためには耳介が音源の方向について非対称である必要があるので $r(\theta)$ は θ とともに変化しなければならない. そこで距離 r が θ の変化に対してある程度「大幅に変化する」よう関数 r を指数関数で定義しモデル化することとした. 従って耳介形状は次のように定義される.

$$r(\theta) = a\exp\theta + b \quad \text{for } \theta \in [0, \pi], \tag{2}$$

ここで a, b は調整可能な係数パラメータであり, 以下で与えるモデルを用いて数値計算を繰り返し行い決定した.

3.2 耳介の音響モデル

回折・反射モデル [Lopez-Poveda, 1996] はスペクトルキューの良い物理モデルを与えるが, パラメータ調整にあたり繰返し計算を行うには計算量を削減することが望ましい. 従って耳介の形状を決定するパラメータ a, b を求めるにあたり音響モデルとして回折・反射モデルをもとに反射だけを考慮した以下のモデルを用いることにした. 十分に音源方向と耳介ノッチの間に強い相関のある耳介が設計することが目的であり, 厳密な現象の解析ではないこと, また耳介があまり大きくなく, 対数螺旋のような単純な構造を有する場合においてこのモデルは有効であると期待される.

以下音響モデルについて説明する. 図 2 中 O は耳孔を表し音波は λ の方向から到達し耳介上の点 P において音が反射した場合を表している. 今 λ は $-\frac{\pi}{2}$ から $\frac{\pi}{2}$ の範囲にあるものとする. これは音源が耳の前方に位置する場合に相当する. 点 P の角度を θ で表すとすると, 耳孔 O から切片 O' までの距離は次の式で与えられる.

$$\xi(\theta,\lambda) = r(\theta)\cos\theta - r(\theta)\frac{\sin\theta}{\tan\lambda} \quad (\lambda \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2})). \quad (3)$$

点 P を通り音源と耳孔 O を結ぶ直線と平行な直線を考え, その直線の上に O を射影した点を O'' とする. この時反射した音波の経路 O''-P-O の距離は

$$\begin{aligned} d(\theta,\lambda) &= \sqrt{(r(\theta)\cos\theta)^2 + (\xi(\theta,\lambda) - r(\theta)\sin\theta)^2} \\ &\quad + \xi(\theta,\lambda)\cos\lambda + r(\theta) \end{aligned} \quad (4)$$

と与えられる. 音波の周波数を f, 音速を V と表わすと, 耳孔における反射音と直接音の間の位相差は

$$\phi(f,\lambda,\theta) = 2\pi f \frac{d(\theta,\lambda)}{V} \quad (5)$$

である.

以上を用いて, 耳孔における音圧は直接波と耳介上の全ての点からの反射波を加え合わせたものとして計算でき

$$I(f,\lambda) = \quad (6)$$
$$\sqrt{\left(\int_0^{\pi} \cos(\phi(f,\lambda,\theta))d\theta\right)^2 + \left(\int_0^{\pi} \sin(\phi(f,\lambda,\theta))d\theta\right)^2}.$$

と求められる. 従って (2) および (3), (4), (5), (6) より耳介による周波数応答を計算できる.

さて, 適当な形状を持った耳介を設計するために, いくつかの a, b の組み合わせに対して上記のモデルを用いて数値計算を行い, その結果から適当な特性を有するものを求めた. ここで以下の判断基準を用いた.

1. 耳介の全体の大きさ: 実際のロボットに用いる場合の実装上の制限
2. 耳介ノッチの存在する周波数帯域: あまりに低周波では時間分解能が悪く, 高周波では A/D 変換器などの変換速度の上で問題がある. また, 狭い帯域にノッチが存在している場合, 方向分解能が悪くなるため, ある程度広い帯域に渡って存在している必要がある.

このような特性を有する耳介を設計することは, 厳密には複雑な逆問題を解くことになり現実的ではない. そこで適当な a, b の組み合わせに対して応答を調べその応答が上記の要請を満足するものかどうかを調べる方法によって設計することとした. つまりここでの計算は必ずしも最適解を求めることが目的ではなく, 現実的に利用可能な耳介形状を得ることが目的であることに注意されたい.

以下では $a = 0.001$[m], $b = 0.01$[m] とするものの結果を図 3 に等高線図として示す. 横軸は周波数であり, 縦軸は音源の高さに対応する λ の値である. 図では λ は $\frac{\pi}{6}$ から $\frac{5\pi}{6}$ まで変化した場合について示しており, 音源が耳孔と同じ高さにある時 $\lambda = \frac{\pi}{2}$ とする. 低周波数側に強いピークが見られ, 4 つのノッチが音源の高さが上がるにつれて周波数の低い方へと変化してく様子が分かる. 特に 4 から 10kHz と 10 から 14kHz に 2 つのはっきりとしたノッチが見られ, この耳介は上記の要求事項 2 を満足している.

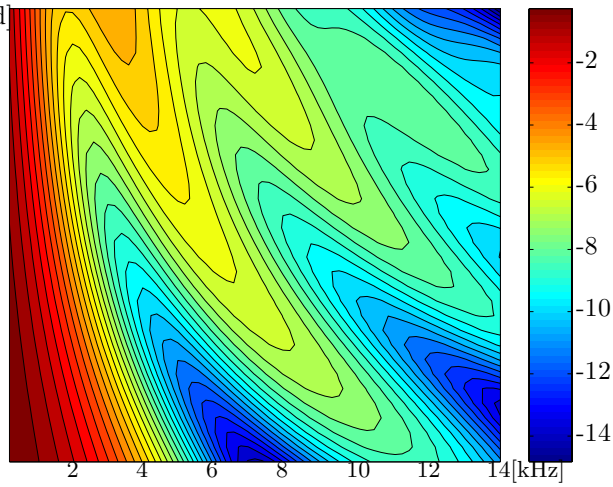


図 3: Spectral Response of the Proposed Pinna

ただし, この耳介は既に上で延べたように簡略化したモデルに基づいて設計したものでありその特性については実際の実験を通じて検証をする必要がある. そこで次節では実際にこのような形状を有する耳介を作成し周波数応答を測定した例を示す.

3.3 周波数特性

3.3.1 装置

実際に作成した耳介を図 4 に示す. 図 4(a) は上で設計した対数螺旋型のものであり, 比較のため人間の耳介に似せたもの (同図 (b)) も作成した. 前節では耳介の平面形

(a) Proposed Pinnae　(b) Pinnae like Humans

図 4: Pinnae used at Experiments

状について考察してきたが, 文献 [Lopez-Poveda, 1996] に倣い 6cm の奥行きを持たせることとした. また, 実装上の関係で開口部が 9mm 程度になるよう大きさをスケールし, 前節のものに比べ全体に大きなものを採用した. 耳介は厚さ 0.5mm のアルミニウムを曲げたものである.

図 5 に実験装置の模式図を示す. また, 提案した耳介を取り付けた状態実際の装置の全体図および頭部の正面図と側面図を図 6 に示す.　頭部内部にはマイクロフォ

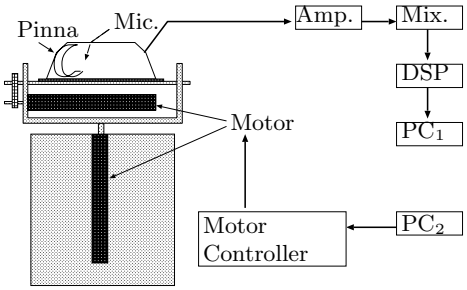


図 5: Block Diagram of the System

ンが埋め込んであり, その開口部に耳介を取り付けている. マイクロフォンで収録した音情報はアンプで増幅した後, デジタルミキサにてサンプリングし, コンピュータ PC_1 にて処理される. この PC_1 は周波数特性を測定するために音信号の記録と処理を行うために用いる. 頭部は二つのモータによって駆動され, ヨー軸およびピッチ軸まわりに回転することが出来る. モータにはエンコーダが取り付けられており, コンピュータ PC_2 からの指令信号に合わせて位置制御される.

(a) Front view

(b) Head Part(Front and Side)

図 6: Photos of the System

3.3.2 周波数特性

ロボットの前方 0.5m の所にイヤホンを音源として TSP 信号 [Suzuki, 1992] を用い, この応答と逆 TSP 信号を畳み込むことで耳介の周波数特性を測定した. 音源の位置を上下方向に変化させながら周波数応答を測定した.

図 7 に図 4(a) の耳介を用いた時の一方の耳の周波数応答を示す. 横軸は周波数であり縦軸は音の大きさを dB 単位で表したものである. 本研究ではノッチの周波数に着目しているので, 音の大きさは重要ではなくここではマイクロフォン電圧に比例する適当なデジタル量を示している. なおサンプリング周波数は 44.1kHz である.

音源の向き λ は耳から見て水平を $\frac{\pi}{2}$rad とし音源が下になるに従い正になる方向に取る. 本実験では $\lambda = \frac{\pi}{4}$ から $\lambda = \frac{3\pi}{4}$ までを $\frac{\pi}{6}$ 刻みで測定した.

測定された結果より, 全ての音源の向きについて 3 から 15kHz においてピークとノッチが形成されている. しかしながら, 全てのノッチが音源方向に依るものでないため, 耳介ノッチを特定するのは簡単ではない. そこで音源方向の変化に合わせて, 周波数の変化しているノッチを調べたところ図中 N_A, N_B ならびに N_C と印した

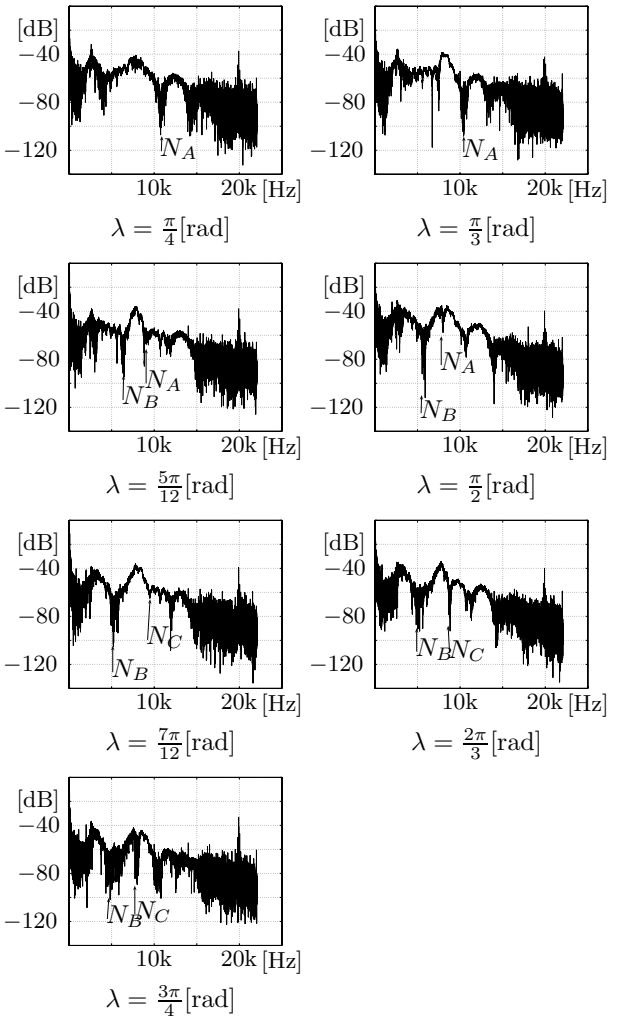


図 7: Spectral Cues of Proposed Pinna

3 つを見つけることが出来た. ただし, N_A は λ が小さい時に見ることが出来るが, 音源が水平よりも低くなると消失し, かわりに N_B が現れている. また, 制作時に耳介の大きさを調整したため, これらのノッチは上の設計時 (図 3) よりも低い周波数に生じている. しかしながら, 図 3 で期待された通り音源の方向 λ が増加するに従い, ノッチは低周波数側に移動しており所望の性能を有するものと期待できる. 以上より音源の上下を耳介ノッチを利用して認識する能力が期待できる.

次に比較のため, 人間の耳介を模した図 4(b) の周波数応答を測定した (図 8). この耳介は上で設計したものとほぼ同じ大きさであるが, 耳孔よりも前方にも反射する面があり, 形状が異なる. どの音源の方向に対してももピークとノッチを見ることが出来, λ に合わせて移動しているが, 例えば $\lambda = \frac{\pi}{3}$ および $\lambda = \frac{2\pi}{3}$ の時の応答は設計したものに比べ複雑になっている. ロボットに応用することを考えると, 音源定位の意味では必要以上に複雑な応答は望ましくなく, 図 4(b) の耳介は適当でないと言える.

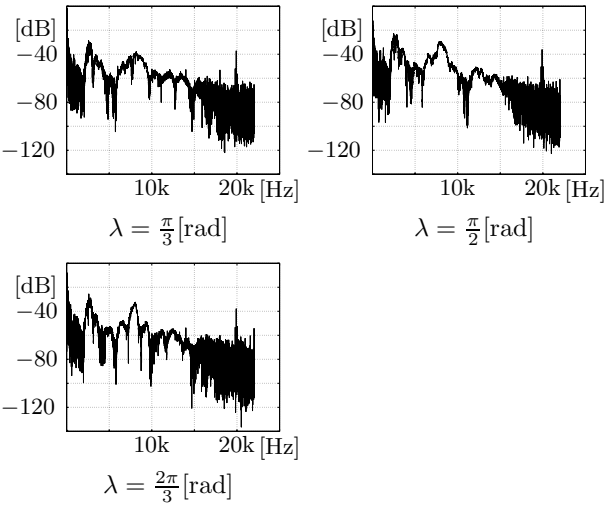


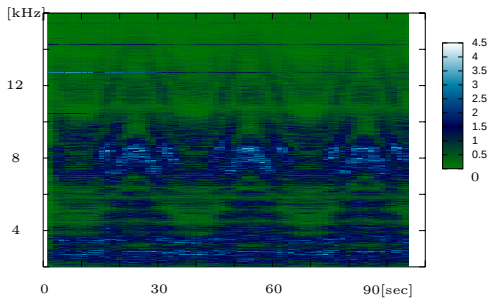
図 8: Spectral Cues of Pinna like Humans Ear

3.4 動的スペクトルキュー

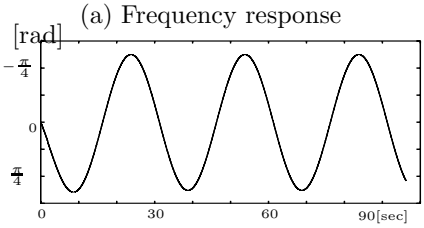
前節では, ロボット頭部を固定した静的な状態での周波数応答を示した. 本研究の目的はオンラインでの上下方向の音源定位であり, 最終的にはモータによって頭部を運動させながら同時に耳介ノッチを認識する必要がある. この時, モータや機構部分の駆動音などが音源定位に悪影響を与える可能性がある. そこで, 実際に頭部を駆動しながらマイクロフォンで音信号を収録し動作時の音響特性を調べることとした. 上と同様に音源をロボット前方 0.5m の所に固定し, 頭部を正弦波指令値に従い上下に位置制御することで, 相対的に音源を上下させる実験を行った.

実際の実験システムは図 5 の PC_2 に A/D 変換器を搭載し, 全ての信号処理を PC_2 で行った. PC_2 は Athlon XP 2500+ のコンピュータであり FFT 演算には FFTW[Matteo, 2003] を用いた. 実験ではマイクロフォン 2ch からの信号を 40,960Hz でサンプリングし 0.5sec 毎に短時間フーリエ変換することとした. サーボ信号はサーボモータの速度指令値が専用ボードにてパルス信号として生成される. 頭部の姿勢はモータに取り付けられたロータリエンコーダで計測されカウンタボード (パルスボードに内蔵) で処理され PC_2 上のソフトウェアにて制御信号を計算するものとした. また, 短時間で周波数応答を得るために音源からはホワイトノイズを生成した. ホワイトノイズは実験開始直後に印加した.

図 9(a) に測定された時間-周波数応答を示す. 横軸が時間, 縦軸が周波数を表す. また実際の頭部の向きを 同図 (b) に示す. ホワイトノイズを印加したが, イヤフォンの特性により実際には高周波側には十分なパワーが得られていない. しかしながら対象とするノッチの存在する帯域には応答が見られており, この信号で実験の目的



(a) Frequency response



(b)Head motion

図 9: Dynamic spectral cue

は達成できるものと考えられる. 実際, 観測されたノイズ信号の帯域 4 から 10kHz 付近には, 耳介ノッチが表われており, 正弦波状の頭部の動きに対応している. また, モータの駆動音などの雑音の影響はほとんど見られない. これは, モータがマイクロフォンから下方に離して設置されていること, 耳介によって前方からの音を特に収音するため, 下方からの騒音の効果を抑制する結果となっていることなどに依るものと考えられる. なお, 図 9(a) にはいくつかの周波数において騒音が存在しているが, これらは PC のファンなどの雑音源に依るものである.

一方, この実験の結果では複数のノッチが混在し, ノッチの周波数から音源方向への関係は多価関数になってしまっている. 従って, 単一のノッチの周波数から直ちに音源方向を定位することは出来ないため, 複数のノッチを利用する必要があると考えられる.

4 おわりに

本研究では, 2 つのマイクロフォンを有する聴覚ロボットが音源の上下方向を定位するため, 耳介によるスペクトルキューを利用することを検討した. これを実現するため, 現実的なスペクトルキューを得られる耳介の形状を簡単な音響モデルをもとに計算し, 実際に耳介を設計・製作した. またその特性を測定し, 所望の特性が得られていることが分かった. さらに実際のロボットの運動を想定し, マイクロフォンを内蔵したロボット頭部を動かした際の特性を測定したところ, 音源方向を認識できる可能性が示唆された.

今後は, 耳介ノッチをリアルタイムで検出する方法の確立と耳介ノッチの周波数の情報からロボットを制御するシステムの構築を行う予定である.

参考文献

[中島, 2004] 中島弘道, 向井利春. 前方の音源に対する音源定位学習システム. In 第 *22* 回日本ロボット学会学術講演会予稿集, 2004.

[Kumon, 2003] M.Kumon, T. Sugawara, K. Miike, I. Mizumoto, and Z.Iwai. Adaptive audio servo for multirate robot systems. In *Proceeding of 2003 International Conference on Intelligent Robot Systems*, pages 182–187, 2003.

[Shaw, 1968] E. A.G. Shaw and R. Teranishi. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *The Journal of the Acoustical Society of America*, 44(1):240–249, 1968.

[Hebrank, 1974] J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America*, 56(6):1829–1834, 1974.

[Lopez-Poveda, 1996] E. A. Lopez-Poveda and R. Meddis. A physical model of sound diffraction and reflections in the human concha. *The Journal of the Acoustical Society of America*, 100(5):3248–3259, 1996.

[Gardner, 1997] W.G. Gardner and K.D. Martin. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, (6):3907–3908, 1997.

[Garas, 2000] J. Garas. *Adaptive 3D Sound Systems*. Kluwer, 2000.

[Gradner, 1994] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone, 1994.

[Suzuki, 1992] Y. Suzuki. Study on the design of the time streched pulse. *Techical Report of IEICE*, EA92(86), 1992.

[Matteo, 2003] F. Matteo and S. G. Johnson FFTW. Web site.

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-4 (12/6)

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

Source Localization and Mixed Speech Recognition by using Four-line Directional Microphones Mounted on Head of Robot

持木南生也 関矢俊之 小川哲司 小林哲則

Naoya Mochiki Toshiyuki Sekiya Tetsuji Ogawa Tetsunori Kobayashi

早稲田大学 理工学部

Department of Computer Science, Waseda University

{mochiki,sekiya,ogawa,koba}@pcl.cs.waseda.ac.jp

Abstract

Sound source separation and localization methods using four-line directivity microphones mounted on a head of a robot are proposed. These methods are free from strict estimations of HRTF. The separation method utilizes a difference of the power specturm with the robot head acting as a sound barrier. It performes signal processing in three layers:two-line SAFIA, two-line Spectral Subtraction and their Integration. The localization method utilizes the idea that the direction of arrival depends on a sound gain difference from four microphones and traces it statistically. The experimental results prove that the proposed separation and localization methods are very effective.

1 はじめに

ロボット頭部側面に設置したマイクロホンによる音源定位手法およびハンズフリー音声認識手法について検討する．

ロボット頭部に設置したマイクロホンによる音源分離，定位では，中臺らによる，頭部伝達関数による手法がある [Nakadai, 2003]．この手法は，ロボットの頭部伝達関数に基づいて2系統の差を用いて，実時間，実環境での動作を実現している．しかし，ロボットの厳密な頭部伝達関数は，部屋の位置や残響時間によって複雑に変化するため，環境によっては，良好に動作しない可能性がある．

本稿では，今までに提案してきた厳密な頭部伝達関数を必要としない，音源分離手法について述べる [Mochiki, 2004]．また，加えて4つのマイクロホンから得られるスペクトル強度パターンに基づいた音源定位手法について述べる．

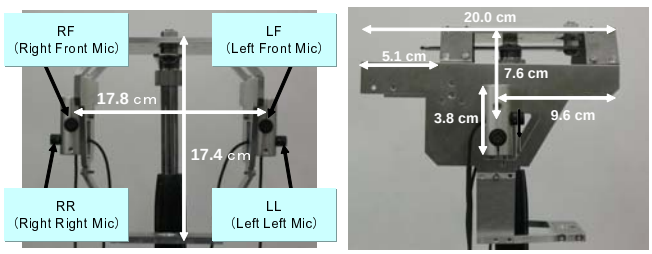
音源分離では，4系統の指向性マイクロホンを設置し，ロボット頭部が障壁として働くことにより生じる音圧の大小関係を利用することで，厳密な伝達特性の推定を必要としない，よりロバストな音源分離を実現する．この大小関係を利用し，階層的信号処理によって，音源分離を行う．

また，音源定位では，原音声の周波数特性に依らず，マイク間のスペクトル強度比が方向ごとに特徴的なパターンを示すことを利用し，統計的パターン認識手法に帰着させることで問題を解く．このような方法においては，学習環境と実際の動作環境との差異が問題となるが，このような差を補正するために，動作環境で得られた数方位からの数データを用いて，MLLR によりモデルを適応する．

以下，2. で, ロボット頭部におけるマイクロホンの設置条件を示す．そして，3. で音源分離手法の提案および実環境での同時発話を対象とした連続音声認識の結果について述べる．4. で音源定位手法の提案および実環境での単一音源定位の結果について述べ，5. でまとめとする．

2 マイクロホンの設置

指向性マイクロホンとして，Audiotechnica ATM15a を使用した．今回の実験では，ロボット本体の頭部ではなく，ロボット頭部の外殻のみを使って実験を行った．ロボッ



(a) 正面　　(b) 側面

Figure 1: ロボット頭部の外殻

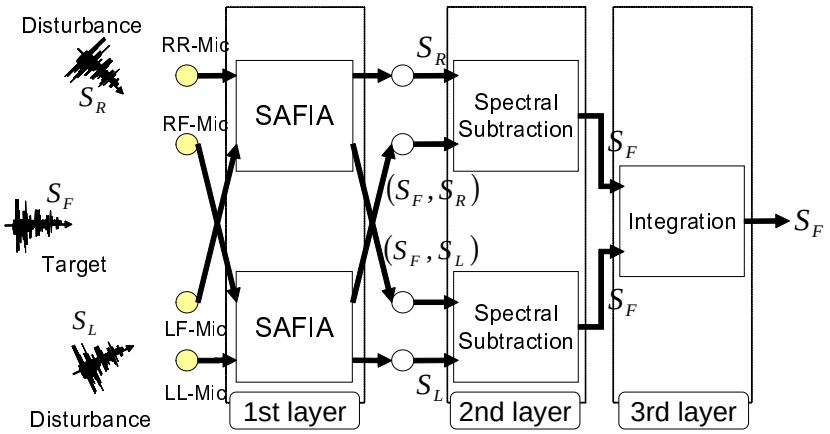


Figure 2: 分離手法のダイアグラム

トには, 両側面に 2 個づつ, 計 4 個のマイクロホンを Figure 1(a), 1(b) の様に設置した．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．

このように設置することで，分離や定位を行う際に優位な入力を得られる．RF-Mic は正面，右方向から到来する音声に対して，左方向から到来する音声は劣勢に受音され，逆に LL-Mic は左方向から到来する音声に対して，正面，右方向から到来する音声は劣勢に受音されると期待できる．

3 音源分離

3.1 提案手法

提案システムでは，3 階層に分けて信号処理を行う．提案する階層的音源分離システムの概要を Figure 2 に示す．

3.1.1 第 1 階層

Figure 3 に第 1 階層の処理を示す．第 1 階層では，RF-LL 間，LF-RR 間において SAFIA [Aoki, 2001] を行う．SAFIA とは，2 チャンネルの入力に対して，周波数成分毎にどちらのマイクロホンに対する入力が優位かを判定し，各周波数成分を優位なマイクロホンに近い音源に属するものとして帯域選択を行う手法である．

S_L のスペクトルが，S_F，S_R のスペクトルに対して，劣勢に含まれることを劣勢なスペクトルに $[\ ]^S$ をつけることで $(S_F, S_R, {S_L}^S)$ と定義する．例えば，LF-RR 間においては，ロボット頭部の側面を障壁として利用することで，LF-Mic は $(S_F, {S_R}^S, S_L)$ のスペクトルを受音し，RR-Mic は $({S_F}^S, S_R, {S_L}^S)$ のスペクトルを受音する．つまり，LF-RR 間において SAFIA を行うことで，(S_F, S_L) と S_R のスペクトルに分離することが可能となる．なぜなら，チャネル間で劣勢な帯域は除去されるためである．また，RF-LL 間の SAFIA により，(S_F, S_R) と S_L のスペクトルに分離することができる．

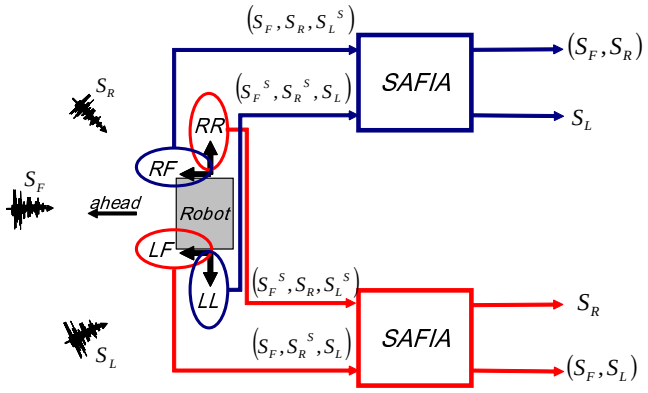


Figure 3: 第 1 階層 SAFIA

3.1.2 第 2 階層

第 2 階層では，第 1 階層で抽出された S_R もしくは S_L のスペクトルを利用し，Spectral Subtraction(SS) [Boll, 1979] により，S_F のスペクトルのみを抽出する．

例えば, 第 1 階層における RF-LL 間の SAFIA で得られた S_L を利用し, LF-RR 間の SAFIA で得られた混合スペクトル (S_F,S_L) に含まれる S_L を SS により除去し，S_F のみを抽出する．同様に，第 1 階層で得られた S_R を利用し，混合スペクトル (S_F,S_R) に含まれる S_R を SS により除去し，正面の目的音声 S_F のみを抽出することができる．

3.1.3 第 3 階層

第 2 階層の処理後，S_F の推定値を 2 つ得ることができる．1 つは，RF-Mic が寄与しているもので，もう 1 つは，LF-Mic が寄与しているものである．第 3 階層では，これら 2 つの S_F の推定値を統合し，S_F を再構成する．この処理により，さらに高精度な S_F が生成される．統合は，

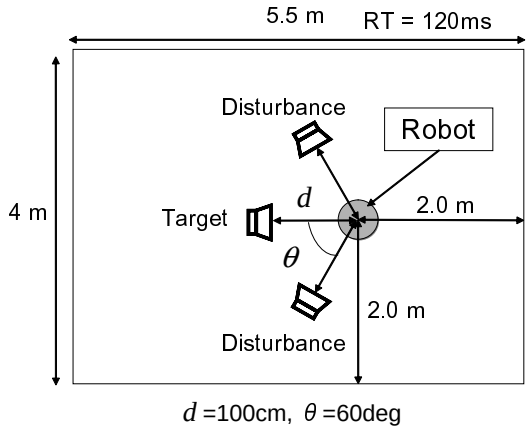


Figure 4: Recording enviroment

Table 1: 特徴量算出パラメータ

pre-emphasis	$1 - 0.97z^{-1}$
frame length	25ms
frame shift	10ms
window	Hamming window
acoustic feature	12th MFCC+ΔMFCC+Δ power

2 つの S_F を加算し，平均化することにより行う．このような平均化により，S_F は異なるノイズ成分から構成されるため，ノイズの分散を小さくすることが期待できる．

3.2 3 話者同時発話音声認識の実験

提案分離手法を音声認識の前処理として適用し，3 話者同時発話音声認識において評価を行う．

3.2.1 収録条件

標本化周波数 32kHz，16bit 量子化で収録を行った．発話者の代わりに音源として，3 個のスピーカを Figure 4 に示す配置に設置した．目的音声には日本音響学会の新聞読み上げ音声コーパス (ASJ-JNAS) の男性話者から 23 人計 100 文を選択した．妨害音声には, 同様に JNAS から認識対象外の男性話者の音声を用いた．スピーカから再生する音声は, それぞれの発話長がほぼ等しく, 目的音声と各妨害音声の発話の SNR が 0dB になるように音量を調整した．

3.2.2 分離条件

音声認識の前処理として，妨害音声を除去する．処理する際のフレーム長，FFT サイズは 64ms とし，フレームシフトは 16ms とした．窓関数にはハミング窓を用いた．

3.2.3 認識条件

分離音声に対して 20000 語彙の連続音声認識を行う．認識の際に用いた音響特徴量を Table 1 に示す．音響モデルは ASJ-JNAS の男性話者約 100 人のクリーン音声約 20000 文を用いて学習した．言語モデルは CSRC 提供の

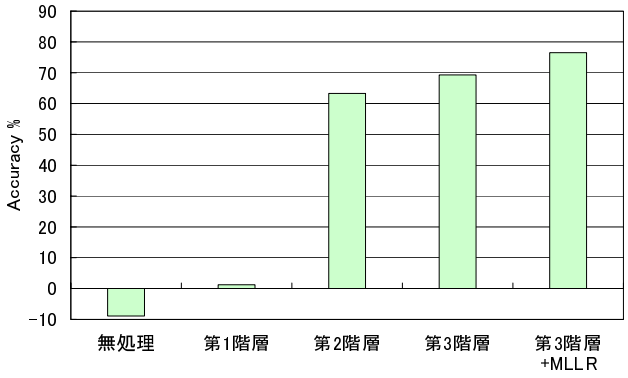


Figure 5: 3 話者同時発話音声の連続音声認識結果

語彙数 20000 語の trigram を使用し，認識器には当研究室開発のデコーダ [柴田, 2002] を用いた．

提案手法では，雑音成分を除去できるが，スペクトルの不連続性により回復した音声はミュージカルノイズが生じる．歪んだ音声を精度良く認識率するために，分離音声を用いて音響モデルの MLLR 適応を試みた．適応データとして，評価データから 80 文を使用し，残りの 20 文を認識データとする．認識データの選び方は 5 通りあるので，テストデータは，3.2.1 で述べた 23 話者の計 100 文である．

3.3 分離結果

3 話者の同時発話音声認識結果を Figure 5 に示す．第 2 階層までの処理を行うと，第 1 階層までの処理に比べて，約 63%のエラー削減に成功した．第 3 階層として統合処理を行うことで，さらに性能を向上させることができた．第 2 階層，無処理時に比べて，それぞれ約 16%，72%のエラーを削減することができた．

また，分離後生じる歪みに MLLR で音響モデルを適応させることで，クリーン音響モデルに対して，約 23%のエラー削減に成功した．

4 音源定位

原音声の周波数特性に依らず，マイク間のスペクトル強度比が方向ごとに特徴的なパターンを示すことを利用して，特徴量を抽出し，モデルの学習および認識を行う．また，残響の異なる環境でも頑健な定位を行うために，モデルの適応を試みる．本手法のダイアグラムを Figure 6 に示す．

4.1 提案手法

4.1.1 特徴量抽出

頭部伝達関数に由来するスペクトル強度パターンをフィルタバンクを用いて，圧縮したものを単語単位の特徴量とする．各単語音声は，N 個のマイクロホンにより受音される．i 番目のマイクロホンの入力信号に対して DFT

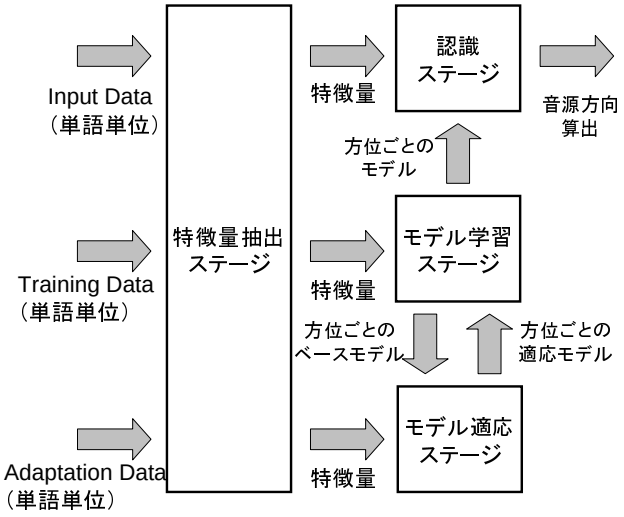


Figure 6: 音源定位ダイアグラム

を施したスペクトルを $X_i(k,t)$ とする．k 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 $X_i(k,t)$ に対して，ある 1 つのマイクロホンから得られるスペクトル $X_N(k,t)$ で正規化を行う．

$$Y_i(k,t) = \frac{|X_i(k,t)|}{|X_N(k,t)|} \quad (i = 1, \cdots, N-1) \tag{1}$$

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．

$$Y_i(k) = \sum_t Y_i(k,t) \tag{2}$$

この平均スペクトル $Y_i(k)$ 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．

$$m_i(l) = \sum_{k=lo}^{hi} W(k;l) \cdot Y_i(k) \ (l = 1, \cdots, L) \tag{3}$$

$$W(k;l) = \begin{cases} \frac{k-k_{lo}(l)}{k_c(l)-k_{lo}(l)} & \{k_{lo}(l) \le k \le k_c(l)\} \\ \frac{k_{hi}(l)-k}{k_{hi}(l)-k_c(l)} & \{k_c(l) \le k \le k_{hi}(l)\} \end{cases} \tag{4}$$

$$c_i(l) = \log m_i(l) \tag{5}$$

ただし，$k_{lo}(l)$，$k_c(l)$，$k_{hi}(l)$ 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，$(N-1) \times L$ 次元に圧縮される．この圧縮された特徴量をもちいて，パターン認識を行う．

4.1.2 モデルの学習および認識

学習データとして，各方位毎に単語音声を収録し，4.4.1 で述べた方法により $(N-1) \times L$ 次元の特徴量を求める．この特徴量から，各方位ごとに単一正規分布を構築する．

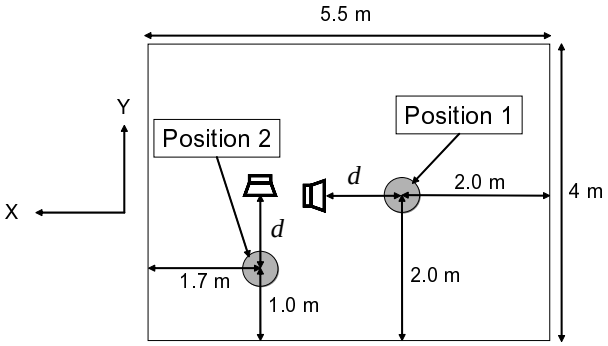


Figure 7: 収録配置図

認識する際は，入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て尤度を算出する．そして，最大の尤度を与えるモデルの方位を音源方向とする．

4.1.3 モデルの適応

本手法はパターン認識の枠組みを用いているため，残響など，モデルを学習した環境と認識を行う環境の違いにより，性能が劣化する恐れがある．残響の異なる環境でも頑健な定位を行うために，その環境で得られた数方位からの数データを用いてすべての方位のモデルを適応することを試みる．この目的を達成するための適応手法として望まれることは，少量の適応データで高い性能が得られること，また，全てのモデルに対する適応データを持つことなしに，すべてのモデルを適応することである．このような条件を満たす適応手法として，MLLR を用いる．

4.2 定位実験

4.2.1 収録環境

収録配置図を Figure 7 に示す．すべての収録は 32kHz, 16bit で標本化，量子化されている．ロボットの配置としては，以下に示す 3 つの配置で収録を行った．

配置 1 ロボットの位置は Figure 7 に示されているように Position1 とし，ロボットは X 軸の方向を向く．

配置 2 ロボットの位置は Position1 とし，ロボットは Y 軸の方向を向く．

配置 3 ロボットの位置は Position2 とし，ロボットは Y 軸の方向を向く．

また，残響時間は，120ms と 200ms の環境で収録した．したがって，配置が 3 通り，残響が 2 通りと，計 6 通りの収録パターンがある．次に，収録方位を説明する．収録した方位を Figure 8 に示す．Figure 8 に示すように，11 方位の収録を行った．ロボットから見て，正面を 0deg とし，右方向を正，左方向を負として定義した．ロボットの正面は，10deg ごとに密にモデルを学習した．それに対して，側面は，30deg ごとに疎にモデルを学習した．

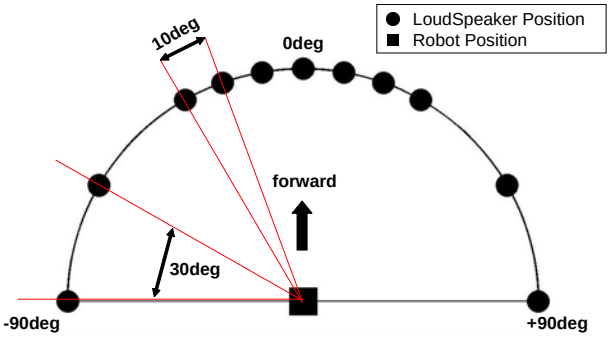


Figure 8: 収録の方位

収録音声は ATR 音素バランス単語 100 単語を男性 10 人が発話したものをスピーカから再生し，各方位に対して収録を行った．その中から，学習データを 90 文，認識データ 10 文とし，認識データの組合わせは 10 通りあるため，評価データは計 100 文からなる．適応データは，3 方位（60deg，0deg，-60deg）から学習データと同じ話者 1 人，同じ音声 5 文，を用いる．

4.2.2 特徴量抽出条件

特徴量を抽出する際の分析条件は，フレーム長 128ms，フレームシフト 32ms，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 8，周波数のレンジは 0∼4000[Hz] とした．マイク数 $N = 4$，バンク数 $L = 8$ なので，特徴量は 24 次元になる．

4.2.3 評価内容

実験の概要を以下に示す．

実験 1 まず始めに，環境に対してクローズの実験を行う．配置 1，残響時間 RT=120ms の環境で，スピーカとロボットの距離 d=100cm，150cm の場合の 2 通り収録して，モデルを構築する．これを残響時間 120ms のベースモデルとする．このモデルで，距離 d=100cm の環境で収録したもの認識する．

実験 2 ロボットの位置が移動した場合の実験を行う．配置 2 または配置 3，RT=120ms，d=100cm を評価データとし，ベースモデルで認識を行う．

実験 3 残響時間が変化した時の実験を行う．配置 1，RT=200ms，d=100cm を評価データとし，ベースモデルで認識を行う．また，配置 1，RT=200ms，d=100cm の適応データを用いて，適応モデルを構築し，認識を行う．

実験 4 ロボットの位置，かつ残響時間が変化した際の実験を行う．配置 1 または配置 2，RT=200ms，d=100cm を評価データとして，ベースモデルからそれぞれの適応モデルを構築し，認識を行う．

Table 2: 音源定位結果（実験 1）

(a) 認識対象，認識モデル：配置 1，RT=120ms

認識対象＼認識結果	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	0	100	0	0	0	0	0	0	0
-10	0	0	0	0	100	0	0	0	0	0	0
0	0	0	0	0	2	98	0	0	0	0	0
10	0	0	0	0	0	1	99	0	0	0	0
20	0	0	0	0	0	0	0	100	0	0	0
30	0	0	0	0	0	0	0	1	99	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 99.6 %

Table 3: 音源定位結果（実験 2）

(a) 認識対象：配置 2，RT=120ms
認識モデル：ベースモデル

認識対象＼認識結果	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	7	93	0	0	0	0	0	0	0
-10	0	0	0	4	96	0	0	0	0	0	0
0	0	0	0	0	2	98	0	0	0	0	0
10	0	0	0	0	0	2	98	0	0	0	0
20	0	0	0	0	0	0	3	97	0	0	0
30	0	0	0	0	0	0	0	7	93	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 97.7 %

(b) 認識対象：配置 3，RT=120ms
認識モデル：ベースモデル

認識対象＼認識結果	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	2	98	0	0	0	0	0	0	0
-10	0	0	0	7	93	0	0	0	0	0	0
0	0	0	0	0	2	98	0	0	0	0	0
10	0	0	0	0	0	0	100	0	0	0	0
20	0	0	0	0	0	0	4	96	0	0	0
30	0	0	0	0	0	0	0	36	64	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 95.4 %

(c) 認識対象：配置 3，RT=120ms
認識モデル：適応モデル

認識対象＼認識結果	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	2	98	0	0	0	0	0	0	0
-10	0	0	0	10	90	0	0	0	0	0	0
0	0	0	0	0	2	98	0	0	0	0	0
10	0	0	0	0	0	0	100	0	0	0	0
20	0	0	0	0	0	0	6	94	0	0	0
30	0	0	0	0	0	0	0	9	91	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 97.4 %

4.3 定位結果

実験 1 の実験結果を Table 2 に示す．学習されたモデルとロボットの位置，残響時間が同じ場合，平均正解率 99.6%と理想的な定位を実現した．このような定位が実現できているのも，原音声の周波数特性に依らず，マイク間のスペクトル強度比が方向ごとに特徴的なパターンを示すからであると考えられる．

実験 2 の実験結果を Table 3 に示す．ロボットの位置が異なり，残響時間が同じ場合，平均正解率はそれぞれ，97.7%，95.4%となった．ロボットの位置が異なる場合，性能の大幅な劣化は見られなかった．配置 3 の方位 30deg に

Table 4: 音源定位結果（実験 3）

(a) 認識対象：配置 1，RT=200ms
認識モデル：ベースモデル

	認識結果										
	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	30	70	0	0	0	0	0	0	0
-10	0	0	0	47	53	0	0	0	0	0	0
0	0	0	0	0	44	56	0	0	0	0	0
10	0	0	0	0	0	46	54	0	0	0	0
20	0	0	0	0	0	0	1	49	50	0	0
30	0	0	0	0	0	0	0	27	73	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 77.8 %

(b) 認識対象：配置 1，RT=200ms
認識モデル：適応モデル

	認識結果										
	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	96	4	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	2	98	0	0	0	0	0	0	0
-10	0	0	0	3	97	0	0	0	0	0	0
0	0	0	0	0	4	93	3	0	0	0	0
10	0	0	0	0	0	1	97	2	0	0	0
20	0	0	0	0	0	0	1	99	0	0	0
30	0	0	0	0	0	0	0	0	100	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 98.2 %

Table 5: 音源定位結果（実験 4）

(a) 認識対象：配置 2，RT=200ms
認識モデル：適応モデル

	認識結果										
	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	0	100	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	5	95	0	0	0	0	0	0	0
-10	0	0	0	10	90	0	0	0	0	0	0
0	0	0	0	0	3	97	0	0	0	0	0
10	0	0	0	0	0	3	94	3	0	0	0
20	0	0	0	0	0	1	6	93	0	0	0
30	0	0	0	0	0	0	0	2	98	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 97.0 %

(b) 認識対象：配置 3，RT=200ms
認識モデル：適応モデル

	認識結果										
	-90	-60	-30	-20	-10	0	10	20	30	60	90
-90	100	0	0	0	0	0	0	0	0	0	0
-60	1	99	0	0	0	0	0	0	0	0	0
-30	0	0	100	0	0	0	0	0	0	0	0
-20	0	0	1	99	0	0	0	0	0	0	0
-10	0	0	0	4	96	0	0	0	0	0	0
0	0	0	0	0	5	95	0	0	0	0	0
10	0	0	0	0	0	5	95	0	0	0	0
20	0	0	0	0	0	2	6	92	0	0	0
30	0	0	0	0	0	0	0	5	95	0	0
60	0	0	0	0	0	0	0	0	0	100	0
90	0	0	0	0	0	0	0	0	0	0	100

平均正解率 = 97.4 %

おいてのみ，劣化が見られたが，適応モデルを使用することで，エラーを削減することができた．

実験 3 の実験結果を Table 4 に示す．ロボットの位置が同じで，残響時間が異なる場合，ベースモデルで認識を行うと，平均正解率は 77.8%にまで性能が劣化した．ここで，3 方位 (60deg，0deg，-60deg) において，各方位から 5 単語，計 15 単語の適応データを用いて，適応モデルを構築し認識を行うと，平均正解率は 98.2%となり，約 92%のエラーを削減することに成功した．

実験 4 の実験結果を Table 5 に示す．ロボットの配置，残響時間がともに異なる場合，実験 3 と同様に，ベースモデルで認識を行うと，平均正解率は，それぞれ 89.0%，85.3%にまで劣化した．しかし，それぞれ異なる環境に対する適応モデルを構築し，認識を行うと，平均正解率はそれぞれ 97.0%，97.4%になった．これは，それぞれ約 73%，82%のエラーを削減したことに相当する．以上により，ロボットの位置，残響時間が異なる環境においても，MLLR を用いてモデルを適応することにより，高い識別性能が得られることが示された．

5 まとめ

ロボット頭部に設置した 4 つの指向性マイクロホンによる音源分離および音源定位手法を提案した．3 話者同時発話の音声認識により，音源分離性能を評価をしたところ，最高で 76.5%と高い認識性能が得られた．また，音源定位においては，原音声に依らず，マイク間のスペクトル強度比が方向ごとに特徴的なパターンを示すことを利用した統計的パターン認識に基づく手法を提案し，MLLR によりモデルを適応させることにより，残響など環境の変化に対してロバストに高い性能が得られることを確認した．今後，この手法を実際のロボットに実装する予定である．

参考文献

[Nakadai, 2003] K. Nakadai, D. Matusura, H. G. Okuno, H. Kitano: Applying Scattering Theory to Robot Audition System, Proc. IROS-2003, pp.1147-1152, Oct. 2003.

[Mochiki, 2004] 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. ICSLP2004, pp.821–824, 2004.

[Aoki, 2001] 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, J. Acousic. Soc. vol.22, No.2, pp149–157, 2001.

[Boll, 1979] S. F. Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. ASSP-33, Vol.27, pp.113-120, 1979.

[柴田, 2002] 柴田大輔，小林哲則: ワンパストライグラムデコーダにおける単語履歴の束ね処理に関する検討，日本音響学会秋季講演論文集，pp151–152，2002.

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-5 (12/6)

ロボットに装着したマイクロフォンアレイによる音源分離とミッシングフィーチャー理論に基づく音声認識

Sound Source Separation by Microphone-Array attached on Robot and Missing Feature Theory based Automatic Speech Recognition

山本 俊一[1] **Jean-Marc Valin**[1,2] **中臺 一博**[3] **奥乃 博**[1]

Shunichi Yamamoto[1], **Jean-Marc Valin**[1,2], **Kazuhiro Nakadai**[3] **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] **LABORIUS, Depart. of Electrical Engineering and Computer Engineering, Universite de Sherbrooke**

[3] 株式会社 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, okuno}@kuis.kyoto-u.ac.jp, jean-marc.valin@usherbrooke.ca, nakadai@jp.honda-ri.com

Abstract

This paper presents a humanoid audition system that gives a humanoid the ability to localize, separate and recognize simultaneous sound sources. A microphone array is used along with a real-time dedicated implementation of Geometric Source Separation (GSS) and a multi-channel post- filter that gives us a further reduction of interferences from other sources. An automatic speech recognizer (ASR) based on the Missing Feature Theory (MFT) recognizes separated sounds in real-time by generating missing feature masks automatically from the post-filtering step. The main advantage of this approach for humanoids resides in the fact that the ASR with a clean acoustic model can adapt the distortion of separated sound by consulting the post-filter feature masks. Recognition rates are presented for three simultaneous speakers located at 2m from the robot. Use of both the post-filter and the missing feature mask results in an average reduction in error rate of 42% (relative).

1 はじめに

将来，様々な面で人間をサポートするようなヒューマノイドロボットを実現するためには，社会で人間と同等に行動できるように人間と同等の認識能力を有する必要があろう．特に，人間のコミュニケーションにおいて音声は重要な位置を占めることから，実環境における音声認識はヒューマノイドロボットの基本的な聴覚機能といえる．一般に，実環境においてロボットに搭載されたマイクには様々な音源からの音が混在した混合音が入ってくる．しかし，現在の音声認識技術のほとんどは単一音源を仮定しているため，十分な認識精度が得られないという問題がある．この問題に対処するためには，混合音に対する音源定位，音源分離，分離音認識という主に 3 つの能力が必要である．このうち，音源定位と音源分離については，信号処理や音環境理解 (*Computational Auditory Scene Analysis*, CASA) の分野で研究が行われてきたものの，分離音認識については，これまでほとんど扱われていなかった．実際，実環境での音声認識が必要とされるヒューマン・ロボット・インタラクションの分野では，音声だけを収音するために口元に設置された接話型マイクを利用するのが一般的である．例えば，MIT の Kismet は，音声認識には接話型マイクを利用し，耳介付近に設置された 2 本のマイクは使用していない[4].

混合音として非音声雑音と音声が混在している場合については，AURORA プロジェクト[1, 12]などで，盛んに研究が行われている．雑音を含んだ音声を学習データに対して HMM パラメータを学習するマルチコンディション学習が，一般的な手法として挙げられる[9, 3]．この手法で得られた音響モデルには，特定条件下で予想される雑音が反映されているため，定常性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，ロボットは動的に雑音が変化する環境で動作する能力，および音声雑音（音声と音声の混合音）を扱う能力が求められる．このような問題を扱う研究としては，マイクロフォンアレイを用いたビームフォーミングによる音声分離が挙げられる．例えば，澤田らは，8 ch のマイクロフォンアレイで同時発話音声を分離し，音響モデル適応による分離音声認識を報告している [18]．また，非定常性雑音に対処するために，ミッシングフィーチャー理論（*Missing Feature Theory*, MFT）も利用されている[2, 13].

我々は，これまでに，2 本のマイクを用いた混合音声分離，およびマルチコンディショニング学習と MFT による

分離音声認識を実装・評価した．これにより，予め与えたクリーン音声から計算したミッシングフィーチャーマスクを利用して，MFT がロボットにおける分離音声認識に有効であることを確認した [15].

本稿では，これをさらに一歩進めて， 実環境においてロボットに装着したマイクロフォンアレイによる音源分離手法， およびクリーン音声や他の先見的情報を与えず，音源分離処理から得られるデータのみを利用したミッシングフィーチャーマスクの自動生成手法を報告する．

以降，2 章では MFT に基づく音声認識についてミッシングフィーチャーマスクの自動生成手法も含めて説明し，3 章では本稿で報告するロボット聴覚システムの概略を説明する．4 章では 多チャンネル post-filter について説明する．

2　ミッシングフィーチャー理論

MFT に基づく音声認識では，認識処理の際に，入力音声の特徴量のうち，ミッシングフィーチャー（雑音によって歪んでしまった特徴量）をマスクすることによって認識向上を図る．この際，2 つの課題を考慮する必要がある．

1. 音声認識で用いられる特徴量の設計
2. ミッシングフィーチャーマスクの自動生成

以下，音声認識特徴量の設計について，2.1 節で，ミッシングフィーチャーマスクの自動生成を 2.2 節で，ミッシングフィーチャーマスクを用いた音声認識を 2.3 節で詳細に述べる．

2.1　音声認識特徴量の設計

一般に音声認識システムでは，音声の特徴としてメル周波数ケプストラム係数（MFCC）が用いられる．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，それがたとえ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，音源分離手法の多くは，周波数領域において分離処理を行うので，少なからず，スペクトル歪みが生じる．このため，分離音声の認識で，特徴量として MFCC を利用した場合は，スペクトル歪みが全 MFCC に広がり，ミッシングフィーチャーマスクを推定することは困難である．従って，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スペクトル特徴量を用いる．実際には，MFCC を逆離散コサイン変換することによって得られる**メル周波数領域対数スペクトル**を用いる．スペクトル特徴量としては，ガンマトーンフィルタバンクの出力が用いられることも多い．しかし，対ノイズロバスト性を向上させるために，MFCC 算出時に行われるような特徴量の正規化が難しく，ロバスト性の面でパフォーマンスを確保することが難しい．

周波数領域の特徴量を利用することにより，ビームフォーミングの後処理である多チャンネル post-filter とも親和性が高いというメリットもある．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おり，これらの情報からミッシングフィーチャーマスクの自動生成が期待できる．

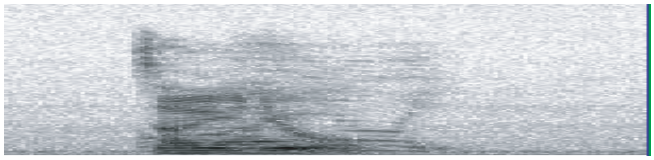
以下に，MFCC で行われるのと同等の正規化を行ったメル周波数領域対数スペクトルの導出の手順を示す．

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（CMS）を行う．
7. 逆離散コサイン変換を行う．
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する．

2.2　ミッシングフィーチャーマスクの自動生成

a priori マスクは，単に，分離音声の特徴量と対応するクリーン音声の特徴量を比較することによって生成されるミッシングフィーチャーマスクである．対応するクリーン音声の特徴量を事前に与えるため，理想的なミッシングフィーチャーマスクを生成することができ，高い音声認識率が得られる．言い換えれば，*a priori* マスクを利用した音声認識によって得られる認識率は，ミッシングフィーチャー理論に基づく音声認識の性能の上限値を表しているといえる [16, 15].

ミッシングフィチャーマスクを自動生成するには，分離音声のスペクトルのうち，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．先見的情報を与えず， 音源分離処理から得られるデータのみを利用して，このような情報を得るために，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 は，ビームフォーマーの出力音響信号を入力として雑音推定を行い，雑音を抑制した音響信号を出力するフィルタである．詳細なアルゴリズムは，4 章に記述する．ミッシングフィーチャマスクのうち，（微分値でない）特徴量に対応するマスク $M_k(i)$ はメル周波数帯域 i のフレーム k における多チャンネル post-filter の入力を $S_k^{in}(i)$，出力を $S_k^{out}(i)$，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



a) Spectrogram at multi-channel post-filter input

b) Spectrogram at multi-channel post-filter output

c) Mel-frequency missing feature mask

Figure 1: Missing feature mask computation

いとき 0）として定義する．また，閾値 T は実験的に求め，0.3 とした．

$$M_k(i) = \begin{cases} 1, & m_k(i) > T \\ 0, & otherwise \end{cases} \quad (1)$$

$$m_k(i) = \frac{S_k^{out}(i) + N_k(i)}{S_k^{in}(i)}$$

このように，多チャンネル post-filter の入力と出力だけでなく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．これは，背景雑音しか存在しなかった周波数帯域は音声認識から見ると，無音であることが信頼できる領域であるためである．

また，ミッシングフィーチャマスクのうち，特徴量の一次微分に対するマスク $\Delta M_k(i)$ 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．

$$\Delta M_k(i) = M_{k-2}(i) M_{k-1}(i) M_{k+1}(i) M_{k+2}(i) \quad (2)$$

特徴量とその一次微分に対応したマスクからなるミッシングフィーチャーマスクの次元数は，スペクトル特徴量と同じ 48 となる．最終的に生成されたミッシングフィーチャーマスクの例を Figure 1 に示す．

2.3 ミッシングフィーチャー理論に基づく音声認識

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のに対して，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．

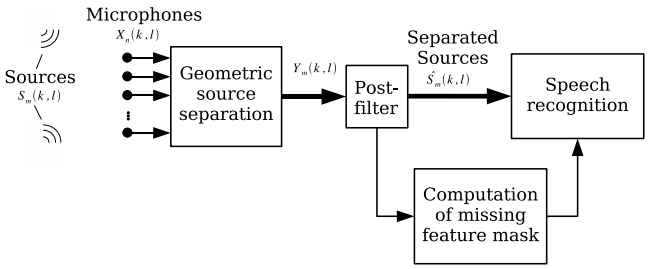


Figure 2: Overview of the system

特徴ベクトル $\boldsymbol{x}$，状態 S の時の通常の連続分布型 HMM の出力確率 $f(\boldsymbol{x}|S)$ は，次の式で表される．

$$f(\boldsymbol{x}|S) = \sum_{k=1}^{M} P(k|S) f(\boldsymbol{x}|k, S) \quad (3)$$

ここで，M は混合正規分布の混合数，$P(k|S)$ は混合比である．MFT に基づく音声認識では $f(\boldsymbol{x}|S)$ を確率密度関数 $p(x)$ に関して平均したものを出力確率とする．

$$\overline{f(\boldsymbol{x}|S)} = \int f(\boldsymbol{x}'|S) p(\boldsymbol{x}') d\boldsymbol{x}' \quad (4)$$

$$= \sum_{k=1}^{M} P(k|S) \int f(\boldsymbol{x}_r|k, S) f(\boldsymbol{x}_u|k, S) p(\boldsymbol{x}'_r, \boldsymbol{x}'_u) d\boldsymbol{x}'_r d\boldsymbol{x}'_u \quad (5)$$

ここで，$\boldsymbol{x} = (\boldsymbol{x}_r, \boldsymbol{x}_u)$ であり，$\boldsymbol{x}_r$ は信頼できる特徴，$\boldsymbol{x}_u$ は信頼できない特徴を表している．信頼できない特徴について事前知識が与えられていない場合には，$p(\boldsymbol{x}'_r, \boldsymbol{x}'_u) = \delta(\boldsymbol{x}'_r - \boldsymbol{x}_r)$ となるので，

$$\overline{f(\boldsymbol{x}|S)} = \sum_{k=1}^{M} P(k|S) f(\boldsymbol{x}_r|k, S) \quad (6)$$

となる．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るので，信頼できない特徴による影響を除去することができる．

3 システムの概要

混合音声認識システムは以下の 4 つのシステムから構成されている（Figure 2）．

1. 幾何学的音源分離（*Geometric Source Separation*, GSS）の一種として実装されている，線形音源分離
2. 多チャンネル post-filter
3. ミッシングフィーチャーマスクの計算
4. 分離音とミッシングフィーチャーマスクを利用した音声認識

マイクロフォンアレイはヒューマノイドロボットに設置された 8 本の無指向性マイクで構成されている．文献 [14] のアルゴリズムにより音源を検出し，音源定位を行う．

音源分離は，基本的には Parra と Alvino [11]によって提案された 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ている．

多チャンネル post-filter は，ビームフォーマーの post-filter 処理[5, 14] を複数音源を扱えるように拡張した手法である．この手法では，雑音を定常性雑音と非定常性雑音に分けて推定することにより，目的音源の強調を行っている．詳細に関しては，4 章で説明する．

多チャンネル post-filter は分離音における干渉音を抑制するだけでなく，特定の時刻，特定の周波数における雑音に関する手がかりを得ることができる．そこで，2.2 節で述べたように，多チャンネル post-filter の入出力と多チャンネル post-filter で推定された背景雑音からミッシングフィーチャーマスク推定を行っている．

MFT ベースの音声認識エンジンとして，CASA Toolkit (CTK) を用いる．CTK はトライフォンの音響モデルをサポートしており，ビームサーチアルゴリズムによる HMM のデコードが可能である．また，CTK は正規文法の言語モデルのみをサポートしており，統計的言語モデルは未サポートである．従って， 実験では正規文法の言語モデルを利用した孤立単語認識を行った．

4 多チャンネル post-filter

GSS アルゴリズムによる分離音を強調するために，Ephraim と Malah によって提案された最適化推定 [7, 8]に基づく周波数領域 post-filter を利用する．マイクロフォンアレイにおける post-filter は，これまでいくつかのアプローチが提案されている．そのほとんどは定常性雑音しかを扱っていた [17, 10] のに対して，最近，非定常性の干渉を考慮した post-filter が Cohen によって提案された [5].

我々は，Figure 3 に示すように，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う．さらに， 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，最終的な雑音推定を行っている．なお，Figure 3 において，$X_n(k,l)$ は n 番目のマイクから GSS への入力，$Y_m(k,l)$ は GSS で推定された m 番目の音源の信号，$\hat{S}_m(k,l)$ は多チャンネル post-filter 処理後の推定された m 番目の音源の信号を表している．$G_m(k,l)$ は重み関数で $\hat{S}_m(k,l) = G_m(k,l)Y_m(k,l)$ と定義される．

この多チャンネル post-filter では，干渉音源はすべて定位されているものとし，残響，音源定位誤り，マイクの周波数応答の相違，近接場効果などによるチャンネル間の漏洩は一定とする．

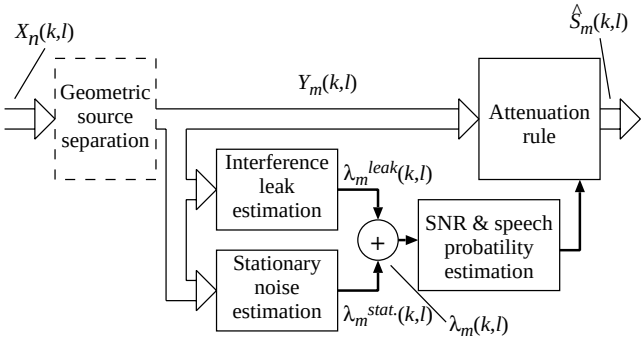


Figure 3: Overview of the multi-channel post-filter

4.1 雑音推定

推定された雑音の分散 $\lambda_m(k,l)$ は以下の式で定義される．

$$\lambda_m(k,l) = \lambda_m^{stat}(k,l) + \lambda_m^{leak}(k,l) \tag{7}$$

ここで，$\lambda_m^{stat}(k,l)$ は音源 m，フレーム l，周波数 k の定常性雑音の推定値であり，$\lambda_m^{leak}(k,l)$ は音源から漏洩した信号の推定値である．

定常性雑音 $\lambda_m^{stat}(k,l)$ は Cohen により提案されている Minima Controled Recursive Average (MCRA) により計算する [6]. 非定常性雑音 $\lambda_m^{leak}(k,l)$ を推定するために，他の音源からの干渉は係数 η（一般的には $-10dB < \eta < -5dB$）により除去することができるものとして，$\lambda_m^{leak}(k,l)$ を以下のように定義する．

$$\lambda_m^{leak}(k,l) = \eta \sum_{i=0, i\neq m}^{M-1} Z_i(k,l) \tag{8}$$

ここで，$Z_m(k,l)$ は m 番目の音源 $Y_m(k,l)$ の平滑化スペクトルであり，以下の式により再帰的に定義される（$\alpha_s = 0.7$）．

$$Z_m(k,l) = \alpha_s Z_m(k,l-1) + (1-\alpha_s)Y_m(k,l) \tag{9}$$

4.2 音声に対する抑制規則

音声が存在するという仮説 H_1 のもとに抑制規則を導入する．以後，特に明示しない限り m と l は省略し，各式は変数 m，l のもとに定義されるものとする．提案する雑音抑制規則は，振幅スペクトル $|X(k)|^{\frac{1}{2}}$ の最小二乗平均誤差推定に基づいている．音声の存在が不確定な場合，振幅スペクトルと対数振幅スペクトルのどちらを選択するかは，実験的に良い結果が得られる方を選択する（4.3 節参照）．

振幅の推定量は以下の式で定義される．

$$\hat{A}(k) = (E[|S(k)|^{\alpha}|Y(k)|])^{\frac{1}{\alpha}} = G_{H_1}(k)|Y(k)| \tag{10}$$

ここで，$\alpha = \frac{1}{2}$ とすると，$G_{H_1}(k)$ は音声が存在していると仮定した場合のスペクトル利得である．

任意の α におけるスペクトル利得は，文献 [8] の式 (13) から次のように定義される．

$$G_{H_1}(k) = \frac{\sqrt{\upsilon(k)}}{\gamma(k)} \left[\Gamma\left(1 + \frac{\alpha}{2} M(-\frac{\alpha}{2}; 1; -\upsilon(k))\right)\right]^{\frac{1}{\alpha}} \tag{11}$$

ここで，$M(a;c;x)$ は合流型幾何関数，$\gamma(k) \triangleq |Y(k)|^2/\lambda(k)$ は事後 S/N 比，$\xi(k) \triangleq E\left[|S(k)|^2\right]$ は事前 S/N 比，$\upsilon(k) \triangleq \gamma(k)\xi(k)/(\xi(k)+1)$ である [7].

事前 S/N 比 $\xi(x)$ は以下の式により再帰的に推定される．音声の存在が不確定な場合を考慮して，文献 [6] で提案されている手法を利用する．

$$\begin{aligned}\xi(k,l) &= \alpha_p G^2_{H_1}(k,l-1)\gamma(k,l-1) \\ &\quad +(1-\alpha_p)\max\gamma(k,l) - 1, 0\end{aligned} \tag{12}$$

4.3 音声の存在が不確定な場合の利得最適化

音声存在確率を考慮した振幅推定を行う．

$$\hat{A}(k) = (E[A^\alpha(k)|Y(k)])^{\frac{1}{\alpha}} \tag{13}$$

音源 m において，音声が存在するという仮定 H_1 と音声が存在しないという仮定 H_0 を考慮すれば，次の式が得られる．

$$\begin{aligned}E[A^\alpha(k)|Y(k)] &= p(k)E[A^\alpha(k)|H_1, Y(k)] \\ &\quad +[1-p(k)]E[A^\alpha(k)|H_0, Y(k)]\end{aligned} \tag{14}$$

ここで，$p(k)$ は周波数 k における音声存在確率である．

最適な利得は次の式から得られる．

$$G(k) = [p(k)G^\alpha_{H_1}(k) + (1-p(k))G^\alpha_{min}]^{\frac{1}{\alpha}} \tag{15}$$

ここで，$G_{H_1}(k)$ は，式 (11) で定義され，G_{min} は音声が存在しない場合に許される最小利得である．対数振幅スペクトルの場合と異なり，$G_{min} = 0$ としても問題が起こらない．$\alpha = \frac{1}{2}$ の場合，次のようになる．

$$G(k) = p^2(k)G_{H_1}(k) \tag{16}$$

$G_{min} = 0$ とすると，減衰には限界値が存在することになる．従って，信号が音声でないことが確実である場合には，利得が 0 に近づく傾向がある．これは，干渉が定常性雑音ではなく音声である場合には特に重要で，ミュージカルノイズが残る．

音声存在確率は次の式で計算される．

$$p(k) = \left\{1 + \frac{\hat{q}(k)}{1-\hat{q}(k)}(1+\xi(k))\exp(-\upsilon(k))\right\}^{-1} \tag{17}$$

ここで，$\hat{q}(k)$ は周波数 k に音声が存在する事前確率であり，以下のように定義される．

$$\hat{q}(k) = 1 - P_{local}(k)P_{global}(k)P_{frame} \tag{18}$$

ここで，$P_{local}, P_{global}, P_{frame}$ は，それぞれ，文献 [6]で定義されており，現在のフレームにおける局所的な周波数窓による音声らしさ，大局的な周波数窓による音声らしさ，全フレームにおける音声らしさにそれぞれ対応する．

Figure 4: SIG2 robot with eight microphones (two are occluded)

5 実験

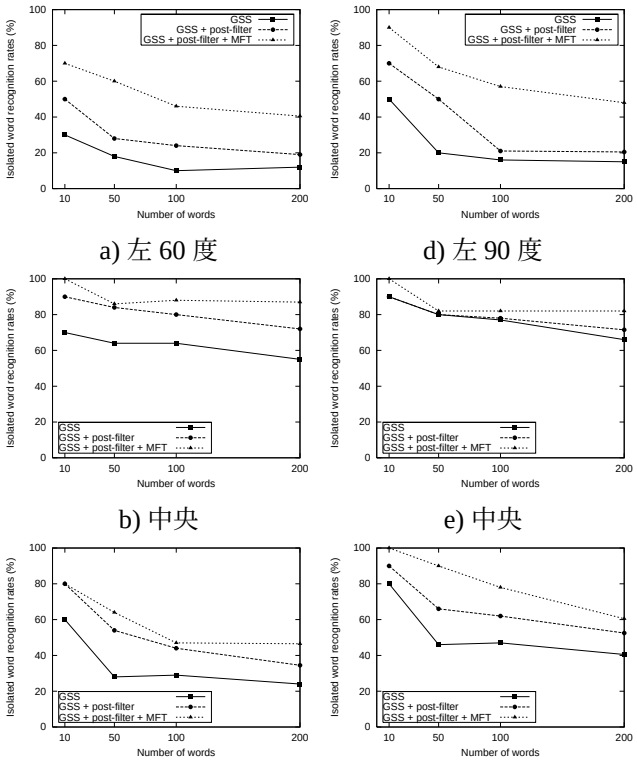
システムの評価を行うためにヒューマノイド SIG2 に 8 本のマイクを取り付け，三話者同時発話認識実験を行った．3 体のスピーカから異なる組み合わせで ATR 音素バランス単語を再生して，三話者同時発話を録音して孤立単語認識実験を行った．実験を行った部屋は 5 m × 4 m の大きさで，残響時間は 0.3 - 0.4 秒 (RT_{20}) である．実験で利用した SIG2 を Figure 4 に示す．SIG2 とスピーカの距離は 2 m で，左 60 度，中央，右 60 度の場合と左 90 度，中央，右 90 度の場合で録音した．孤立単語認識の語彙サイズは 10 語，50 語，100 語，200 語である．

音響モデルはクリーンな音声で学習したトライフォンを利用した．学習データには，合計 25 人の男女の ATR 音素バランス単語 216 語の音声を利用し，3 状態 8 混合のトライフォンを構築した．

比較のために，以下の 3 通りの音声認識実験を行った．

(1) GSS による分離音声を通常の音声認識
(2) GSS と post-filter 処理を行った分離音声に対して通常の音声認識
(3) GSS と post-filter 処理を行った分離音声に対して自動生成したマスクを利用して音声認識

単語正解率を Figure 5 に示す．語彙サイズ 200 語の場合に注目すると，post-filter 処理を行った分離音声に通常の音声認識を行った場合，(1) の単語正解率と比較して 17% の向上が見られた．また，post-filter 処理を行った分離音声を自動生成したマスクを利用して音声認識した場合，(1) の単語正解率と比較して 42% 向上した．すべての角度，語彙サイズで，単語正解率は (1) < (2) < (3) となった．この結果は，post-filter の情報から生成したミッシングフィーチャーマスクが分離音声認識に有効であることを表している．方向ごとの認識率を比較すると，中央が最もよく，左右の認識率は中央よりも低くなった．これは，3 方向の音声の再生音量の違いにより，各方向の分離音声の S/N 比が異なることが原因の一つであると考えられる．



a) 左 60 度　d) 左 90 度

b) 中央　e) 中央

c) 右 60 度　f) 右 90 度

Figure 5: 三話者同時発話認識結果（単語正解率 (%)）
スピーカの間隔は a)，b)，c) が 60 度，d)，e)，f) が 90 度である

6 まとめ

本稿では分離音認識に注目し，GSS と post-filter による音源分離とミッシングフィーチャー理論に基づく音声認識の統合を報告した．その結果，分離音に対しそのまま通常の音声認識を行うよりも，自動生成したミッシングフィーチャーマスクを利用することで，三話者同時発話の孤立単語認識の単語正解率が向上した．

今後の予定として，横方向の話者の音声認識率の改善，システム全体の実時間処理実現が挙げられる．GSS と post-filter で構成される音源分離システムとミッシングフィーチャーマスクの自動生成は，すでに実時間動作が可能であるため，ミッシングフィーチャー理論に基づく音声認識に関して実時間動作を可能にする予定である．

本研究は，科研費 基盤 (A) No.15200015，特定領域「情報学」No.1601625，および，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」の支援を受けた．著者の一人 Valin は (独) 日本学生支援機構の短期留学推進制度の支援を受けた．御討論いただいた京都大学情報学研究科尾形講師，駒谷助手，奥乃研究室の皆さん，HRI-JP の辻野氏，Sherbrooke 大学の Rouat 教授，Michaud 教授に感謝します．

参考文献

[1] AURORA. http://www.elda.fr/proj/aurora1.html" "http://www.elda.fr/proj/aurora2.html.

[2] J. Barker, M. Cooke, and P. Green. Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Proc. of Eurospeech-2001*, pages 213–216. ESCA.

[3] M. Blanchet, J. Boudy, and P. Lockwood. Environmentadaptation for speech recognition in noise. In *Proc. of EUSIPCO-92*, volume VI, pages 391–394.

[4] C. Breazeal. Emotive qualities in robot speech. In *Proc. of IROS-2001*, pages 1389–1394. IEEE.

[5] I. Cohen and B. Berdugo. Microphone array post-filtering for non-stationary noise suppression. In *Proc. of ICASSP-2002*, pages 901–904.

[6] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(2):2403–2418, 2001.

[7] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-32(6):1109–1121, 1984.

[8] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33(2):443–445, 1985.

[9] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multi-styletraining for robust isolated-word speech recognition. In *Proc. of ICASSP-87*, pages 705–708. IEEE.

[10] I.A. McCowan and H. Bourlard. Microphone array post-filter for diffuse noise field. In *Proc. of ICASSP-2002*, volume 1, pages 905–908.

[11] L. C. Parra and C. V. Alvino. Geometric source separation: Mergin convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, 2002.

[12] D. Pearce. Developing the ETSI AURORA advanced distributed speech recognition front-end & what next. In *Proc. of Eurospeech-2001*. ESCA.

[13] P. Renevey, R. Vetter, and J. Kraus. Robust speech recognition using missing feature theory and vector quantization. In *Proc. of Eurospeech-2001*, volume 2, pages 1107–1110. ESCA.

[14] 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 *Proc. of ICRA 2004*. IEEE.

[15] S. Yamamoto, K. Nakadai, H. Tsujino, and H. G. Okuno. Assessment of general applicability of robot audition system by recognizing three simultaneous speeches. In *Proc. of IROS 2004*. IEEE.

[16] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno. Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory. In *Proc. of ICRA 2004*, pages 1517–1523. IEEE.

[17] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proc. of ICASSP-88*, volume 5, pages 2578–2581.

[18] 澤田 知寛, 関矢 俊介, 小川 哲司, and 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**, pages 27–32, 2003.

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-6 (12/6)

コミュニケーションロボットにおける ノンバーバル情報を用いた状況依存型音声認識

Situated Speech Recognition based on Nonverbal Information for Communication Robots

岩瀬 佳代子[†‡]，塩見 昌裕[†*]，神田 崇行[†]，石黒 浩[†*]，柳田 益造[‡]

Kayoko Iwase[†‡]，Masahiro Shiomi[†*]，Takayuki Kanda[†]，Hiroshi Ishiguro[†*] and Masuzo Yanagida[‡]

ATR 知能ロボティクス研究所[†]，同志社大学大学院[‡]，大阪大学大学院[*]

ATR Intelligent Robotics and Communication Laboratories[†]，Doshisha University[‡]，Osaka University[*]

E-mail: {kayoko-i, m-shiomi, kanda, ishiguro}@atr.jp, myanagid@mail.doshisha.ac.jp

Abstract

This paper describes speech recognition based on nonverbal information for communication robots. It will explain how a robot can extract human emotions from nonverbal information and limit the number of possible situations. First, the results of an experiment on interaction between a human and a robot, lead a conclusion that there are two types of emotions: one, including emotions (joy, anger, fear, etc.) which depend on the context, and the other, including emotions (strain, etc.) which do not depend on it. Also, discussed are the possibility that the presence of strain emotion prevents displaying context dependent emotions, and results obtained by using emotion recognition based on the prosodic features of voice and facial information.

1. はじめに

近年，ロボットの開発技術が高度化し，工場などで作業を行うことのみを目的としたロボットではなく，視覚・聴覚・触覚などのさまざまな認識機能を搭載し，人間との自由なコミュニケーションの実現を目指したロボットが開発されるようになった．そのようなコミュニケーションロボットにおいて，特に，人間との自由な「音声対話機能」の充実が期待されている．そのためには，人間同士の会話において用いられるようなノンバーバル情報を利用することが重要であると報告されている[1][2][3]．しかし，現在のロボットは，バーバル情報を利用する音声対話は行っているが，ノンバーバル情報を利用した音声対話はあまり行われていない．

ところが，語用論によると，人間同士の音声対話では，聞き手と話し手の相互のやり取りによるバーバル情報以外のノンバーバル情報を認識し，状況を理解することが重要であるとされる[4][5]．また，A.Mehrabian[6]は，相手にメッセージを伝える際，バーバル情報から 7%，ノンバーバル情報から 93%（パラ言語情報から 38%, 顔の表情から 55%）の割合で，メッセージに含まれる感情が伝達されると提唱し，ノンバーバル情報の重要性を示している．したがって，ロボットが人間の感情を認識して音声対話を行う必要がある．

伊藤ら[1]は，パラ言語情報から人間の感情を認識するシステムについて報告している．[1]の実験より，人間とロボットの対話において表出されやすい感情は，持続的感情であり文脈に依存しにくい感情としての「緊張」，一時的感情であり文脈に依存しやすい感情としての「喜び」と「困惑」であることが検証された．また，一時的感情は，発話単位での変化が見られ，文脈に依存しやすい感情であり，持続的感情は発話単位で変化しにくく，文脈に依存しにくい感情である．

本稿では「緊張」の感情に注目する．持続的であり文脈に依存しにくい「緊張」の感情が表出している場合，この感情が文脈に依存する感情（たとえば，喜び）を妨げるという仮説を立てた．文脈に依存する感情の表出が妨げられると，本研究の主旨である，ノンバーバル情報を利用した音声認識が意味をなさなくなるという点で問題となる．この仮説は，第 4 章の対話実験により検証された．つまり，ロボットが発話や動作によって緊張を緩和した後，文脈に依存する感情の認識を行うことにより，人間とロボットの音声対話においてノンバーバル情報を有効に利用できると示唆される．

本検証を元に，第 5 章では，ATR 知能ロボティクス研究所で開発されたロボット Robovie[7]に，ノンバーバル情報による既存の感情システム，顔の表情から感情を認識するシステム[8] ，ならびに，パラ言語情報から感情を認識するシステム[1]を実装する．まず，文脈に依存しない緊張の感情検出を，次に喜びなどの文脈に依存する感情の検出を行うことにより，現在の状況を認識する．さらに，認識した状況に応じた単語辞書を用いて認識する単語や語彙を絞り込み，音声認識システムによる認識を行うという，「状況依存型音声認識システム」を提案する．

2. ロボット対話における感情表出の仮説

本研究の目的は，ノンバーバル情報を利用して人間の感情を認識することにより，ロボットと人間の音声対話を自然なものと導くことである．本稿では特に，人間との対話において，ロボットが感情を認識するためのアプローチについて提案する．

[1]より，人間とロボットの対話において表出されやすい感情は，持続的で文脈に依存しにくい感情である「緊張」であることが示されている．本稿では，心理学による感情モデルに基づき，この緊張の感情が人間の感情表出を抑制するという仮説を立てる．

2.1 心理学における感情モデル

心理学の分野では，古くから人間の感情についての研究がされている．また，人間の感情表出に関しては，顔の表情や姿勢，音声，ジェスチャーなどの行動および自律反応など，ノンバーバルな情報を対象とした研究がされている．その中でも特に，感情と表情に関する研究は非常に多くの知見を得ている．例えば，Ekman の基本感情，Russell による感情の円環モデルなどが挙げられる[9].

Ekman は, 人間には基本的な感情として 6 つの感情（喜び，驚き，怒り，嫌悪，恐れ，悲しみ）が存在すると提唱している．さらに，基本感情の基準として「刺激に対し急速に，生体が意識する前に生じること」,「通常は極めて短時間（数秒以内）で終結すること」などを挙げている．

また，Russell は,「快-不快」,「覚醒-眠気」の 2 つの次元上に感情を表現することができると提唱している．現在の感情は, 中心から円環方向へのベクトルの向きによって示される．

2.2 人間－ロボット対話における感情表出の仮説

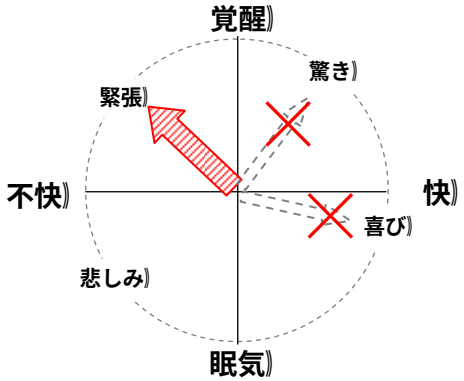


図 1 本稿における感情表出の仮説

前節の感情モデルより考えられることは，ある感情が長時間強く表出してしまうと，その他の感情の方向へベクトルの向きが変化しなくなることである．特に，本稿で注目する「緊張」の感情は，持続的な感情であるため，その感情が表出されてしまうと，Russell による感情の円環モデルより，図 1 に示すようにベクトルの向きが緊張から変化せず，その他の感情が表出されにくくなってしまうという仮説が立てられる．また，緊張の感情は文脈に依存しない感情であるため，現在の発話に含まれる感情として認識することは困難である．したがって，緊張の感情が強く表出することにより，現在の感情の判別が困難になることが示唆される．

以上の仮説より，本稿では緊張の感情を検出し，緊張が検出されればそれを緩和するような行動や発話を行うこと，また緩和された場合に文脈に依存する感情を認識し，状況を認識することが可能になると考えられる．

3. 感情認識システム

本章では，本稿で用いる感情認識システムについて紹介する．

3.1 表情による感情認識システム

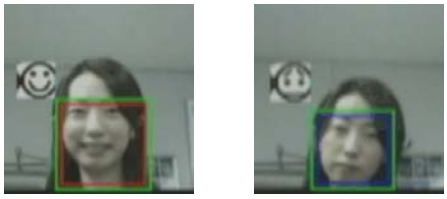


図 2 表情認識システムの判別例
左：喜びの感情判別，右：悲しみの感情判別

顔表情から感情を認識するシステムは，G. Littlewort ら[8]によって開発されたシステムである．FACS（Facial Action Coding System）を利用し，Ekman の基本 6 感情（2.1 節の基準より，文脈に依存する感情と考えることができる）と普通の 7 つの表情（怒り，嫌悪，恐れ，喜び，悲しみ，驚き，普通）を判別するシステムである．このシステムの判別例を図 2 に示す．顔を囲っている外側の緑色の枠は顔が発見できたことを，内側の赤色や青色の枠と顔のマークは表情を判別できたことを示す．また，G. Littlewort らによると，このシステムの「喜び」の感情の判別成功率は 87.0%であった．

3.2 パラ言語情報による感情認識システム

パラ言語情報を利用した感情認識システム[1]は，特徴量として，基本周波数，パワー，発話間間隔など 29 の特徴量を用い, C5.0 または SVM を用いて, 持続的な感情である「緊張」, また，一時的な感情である「喜び」と「困惑」の感情の有無を検出するシステムである．持続的な感情とは，対話を通して大きく変化しない，文脈にも依存しない感情であり，一時的な感情とは発話ごとに変化しやすく，文脈に依存する感情である[1]．伊藤らの実験によると，このシステムの感情判別率は，SVM を用いた場合の喜び感情の判別率が 74.1%，困惑感情の判別が 79.6%，また，C5.0 を用いた場合の緊張の感情の判別率が 87.0%を示している．

4. 対話実験による感情表出の調査

本章では，ノンバーバル情報を音声認識の性能向上に利用するため，第 3 章で紹介した感情認識システムを用いて認識すべき人間の感情の種類を，45 名を被験者とする対話実験を通して調査を行った．まず，4.1 節では人間とロボット（Robovie，図 3 参照）の対話実験によるデータの収集について，4.2 節では収集データに対する感情のラベル付けについて，そして，4.3 節，4.4 節ではそれらの結果より，音声認識の結果から相手の発話内容を絞り込むために有効な感情の種類について検証する．

4.1 実験設定

人間とロボット（Robovie），1 対 1 の対話実験（図 4 参照）を，以下に示す条件の下で行った．

＜実験被験者＞
大学・大学院生　男女　45 名

＜実験環境＞
研究所の実験室内で，図 4 のように人間とロボットが 1 対 1 で向かい合い，簡単な対話を行う．

＜実験条件＞
ロボットは被験者に，いくつかの問いかけを同じ内容で繰り返し行う．それに対し被験者は，

（1）自由に回答する
（2）肯定的に回答する
（3）否定的に回答する

という条件を与えられる．

＜対話例＞
R（ロボット）:「おはよう.」
S（被験者）:「お，おはようございます.」
R:「僕はロボビーだよ.」
S:「えっと，私は，○○です.」
R:「一緒に遊ぼうよ.」
S:「良いですよ.」
R:「じゃんけんしようよ.」
S:「よし，じゃんけんしましょう.」
R:「ロボビーかわいいでしょ？」
S:「(笑う）･･･うん，かわいいですね.」
R:「バイバイ.」
S:「はい，またね.」

また，実験中，ロボットの目のカメラから入力した画像とロボットのマイクから入力した音声を，デジタルビデオに記録した．次節からは，この記録したデータを用い，Robovie との対話における人間の感情について調査する．

図 3　Robovie

図 4　対話実験の風景

4.2 表情のラベリング

ロボットとの対話において，人間がどのような感情が表出するか，また，緊張の感情が文脈に依存する感情表出に及ぼす影響を検証するため，前節の実験により収集した画像（顔表情）データに対し，表情のラベリングを行う．

本節では，利用する感情認識システムが認識可能な感情の種類などを考慮してラベリングの評価対象とする感情を選択し，実験被験者以外の第三者にそれぞれの感情に対し，評価尺度法によるラベリングを行った．

4.2.1 評価対象の感情

本研究では，前章で紹介した感情認識システムを利用するため，ラベリングの評価対象の感情として，Ekman の 6 基本感情に注目する．この感情の分類は，感情認識においてよく用いられる分類方法である．

また，伊藤ら[1]によって人間とロボットの対話において重要とされている「緊張」の感情に注目し，第 2 章において以下のような仮説を立てた．文脈に依存しない感情である緊張の感情表出が強く持続している場合，図 1 に示すように，Russell の感情の円環モデルにおいて，不快-覚醒の間にベクトルが向いたままになってしまい，文脈に依存する感情の表出を妨げてしまうと考えられる．そのため，ロボットが現在の状況を認識しにくくなると考えられる．

以上より，評価対象の感情として，「怒り」「嫌悪」「不安」「喜び」「悲しみ」「驚き」「緊張」の 7 感情を用いることとする．

4.2.2 評価尺度法による表情のラベリング

対話実験において記録した画像データについて，Robovie から問いかけ終了後 200～400msec の被験者の表情を静止画として切り出した．全体の画像データ数は，72 フレームである．評価方法は評価尺度法であり，前述した 7 つの感情について，図 5 に示すような「とてもある」から「全くない」までの 6 段階の尺度を用いた．また，このラベリングの対象者は，対話実験の被験者ではない第三者 4 名（男 3 名，女 1 名）である．

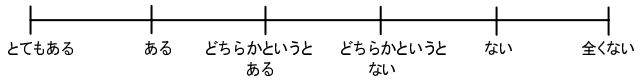


図 5　表情のラベリングに用いた評価尺度

さらに，ラベリングの集計として，図 5 のそれぞれの段階に（とてもある）3 点，2 点，1 点，−1 点，−2 点，−3 点（全くない）の点数をつけ，ラベリング対象者 4 名(a,b,c,d)の点数$(f_{a,emo}, f_{b,emo}, f_{c,emo}, f_{d,emo})$の平均値 $\bar{f}_{emo}$ を以下の式により求めた．また，その表情に対する感情は，7 つの感情の平均値 $\bar{f}_{emo}$ の最大値 $Max(\bar{f}_{emo})$ をとるものとした．

$$\bar{f}_{emo} = (f_{a,emo} + f_{b,emo} + f_{c,emo} + f_{d,emo})/4$$

$$\left(\begin{array}{l} emo \text{ は，怒り，嫌悪，恐れ，喜び，} \\ \quad \text{悲しみ，驚き，緊張のいずれか} \end{array} \right)$$

4.3　ロボットとの対話における感情表出の調査

前節で求めた $Max(\bar{f}_{emo})$ から，ロボットとの対話において人間が表出した感情の種類について調査を行った．以下の表に，対話実験における表情のラベリングによる結果を示す．

表 1　表情ラベリングの集計（返答内容への影響）

感情	肯定的返答	否定的返答
怒り	4.3%	15.0%
嫌悪	16.3%	25.0%
恐れ	0%	2.5%
喜び	52.3%	17.5%
悲しみ	8.7%	0%
驚き	0%	0%
緊張	18.5%	40.0%

表 2　表情ラベリングの集計（緊張の有無への影響）

感情	$\bar{f}_{緊張} > 0$（緊張あり）		$\bar{f}_{緊張} \le 0$（緊張なし）	
	肯定的返答	否定的返答	肯定的返答	否定的返答
怒り	0%	0%	6.3%	27.3%
嫌悪	2.9%	0%	21.9%	45.4%
恐れ	0%	7.1%	0%	0%
喜び	29.4%	7.1%	59.4%	27.3%
悲しみ	5.9%	0%	12.4%	0%
驚き	0%	0%	0%	0%
緊張	61.8%	85.8%	0%	0%

表 1 は，ロボットの問いかけに対し，肯定的／肯定的な返答の場合に分け，各感情について $Max(\bar{f}_{emo})$ の割合を示したものである．ここで期待されることは，発話毎に一時的な感情が表出し，肯定的な返答をする場合は喜びなどの肯定的な感情，また，否定的な返答をする場合は喜び以外の否定的な感情が表出するということである．しかし，表 1 おいて，肯定的な返答の場合は喜びの感情が最も多く表出しているが，否定的な返答をしている場合は持続的な緊張の感情が最も多く表出している．また，全体的に緊張の感情が多く表出しており，文脈に依存する感情があまり表出されていない．したがって，ロボットが人間の感情を認識し，状況を認識することが困難になることが示唆される．

次に，緊張の感情の表出が文脈に依存する感情の表出に与える影響を調査する．表 2 では，評価尺度法によって求めた緊張の平均値 $\bar{f}_{緊張}$ について，$\bar{f}_{緊張} > 0$ の場合と $\bar{f}_{緊張} \le 0$ の場合に分け，さらに，返答内容の肯定的／否定的分けて集計を行った．ここで，$\bar{f}_{緊張} > 0$ は全体の 3 分の 1 を占めている．表 2 より，$\bar{f}_{緊張} > 0$ の場合，ほとんどの場合で以下の式に示すように，緊張が 7 感情の中での最大となった．

$$Max(\bar{f}_{緊張}) > Max(\bar{f}_{emo}) \qquad emo \neq \text{緊張}$$

これは，前に示唆したように，文脈に依存しにくい緊張の感情が表出している場合，文脈に依存しやすい感情が表出しにくいこと，さらに，状況を認識しにくいことが検証されている．また，$\bar{f}_{緊張} \le 0$ の場合は，肯定的返答の場合は喜び，否定的返答の場合はむしろ，嫌悪のような否定的な感情が表出する確率が高くなった．

4.4　状況の認識に有効な感情の検証結果

対話実験より，人間とロボットの対話において表出しやすい感情は，緊張の感情であることが検証された．また，緊張の感情は文脈に依存しにくい感情であること，持続する感情であることから，文脈に依存する感情が表出しにくくなり，状況を認識することが困難になることが検証された．

ロボットに人間の感情を認識させ，状況を認識させるためには，緊張の感情が表出している場合はそれを緩和する必要があると考えられる．緊張を緩和した後，文脈に依存する感情の認識を行うことにより，状況を認識しやすくなることが示唆され，さらに，ロボットの音声認識の性能向上が期待される．

5.　状況依存音声認識システムの提案

本章では，第 3 章に挙げた感情認識システムの利用と，第 4 章の対話実験による検証結果に基づき，ATR 知能ロボティクス研究所で開発されたコミュニケーションロボット Robovie（図 3）に実装する，感情認識の結果を利用した「状況依存音声認識システム」を提案する．まず，全体的なシステムの構成について 5.1 節に述べ，5.2 節ではその一部の感情認識部の処理方法，および，感情認識の結果を利用した音声認識の提案手法について述べる．

5.1　全体のシステム構成

本稿，ならびに参考文献[10]で提案する状況依存音声認識システムの全体の流れを，図 6 に示す．Robovie の周りの状況を

雑音の性質や発話内容，対話の相手の感情として認識し，それに最適な音響モデルや単語辞書を用いて音声認識をするシステムである．すなわち，現在の状況において，相手が発話すると考えられる単語を絞り込み，最も適当な単語辞書などを用いて音声認識を行うのである．既にある状況としては，相手・自分の状態，周囲の環境などが挙げられ，Robovie はさまざまなセンサーにより，これらの情報を取り入れることが可能である．さらに，Robovie は，認識した状況からさらに新しい状況生成行動（発話や動作）により，自らが状況を作り出すことも可能である．

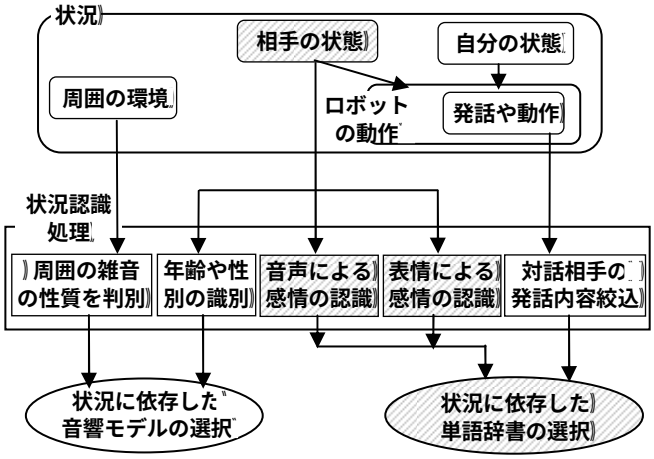


図 6　状況依存音声認識システム

[10]において，図 6 における，「自分自身（Robovie）の発話や動作を認識することにより，状況に適切な単語辞書を選択して音声認識の性能が向上させることが可能である」という状況依存の可能性を示した．本稿では，図 6 の斜線部分の処理，対話の相手の状態から感情を認識することにより，音声認識で用いる単語辞書を限定するという，感情認識部の処理を提案する．

5.2　提案手法　　－感情認識部－

4 章で検証されたように，人間がロボットと対話する場合，緊張の感情多く表出し，それが他の感情の表出を妨げる可能性がある．これに基づき，感情認識を 2 段階に分け，人間の緊張の感情の有無と文脈に依存する感情の検出を行い，状況を認識してそれに適切な音声認識を行うという手法を提案する．

図 7 は，図 6 の感情認識処理の部分を表したものである．Robovie に搭載する感情認識システムには，第 3 章で紹介したパラ言語情報と顔表情による感情認識システムを利用している．パラ言語情報による感情認識については「緊張」と「喜び」，表情による感情認識については Ekman の 6 感情の認識結果を用い，両方の認識結果を照合して感情の認識を行う．緊張の感情が検出されなければ，一時的感情の検出，特に，喜びの感情の有無を検出し，状況を認識して音声認識を行うモデルである．また，緊張があると判断した場合，現時点では状況の認識を行わず，ロボットが緊張を緩和させるような発話や行動をすることにより，一時的感情が表出しやすい状況を作り出す．

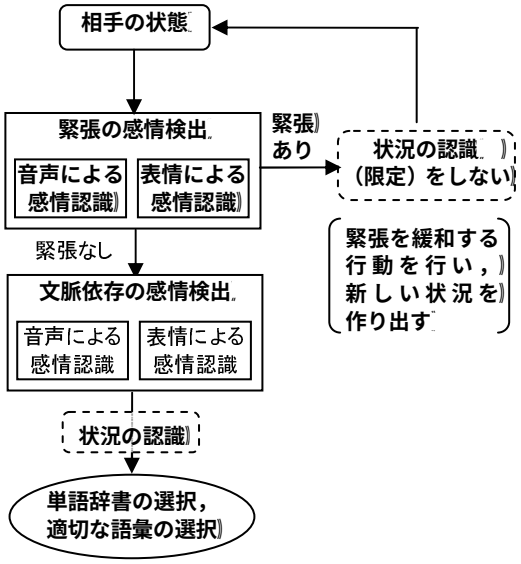


図 7　感情認識モデル

6.　まとめと今後の課題

本稿では，ATR 知能ロボティクス研究所で開発されたロボット Robovie に搭載する音声対話機能として，視覚や聴覚のから取り入れたノンバーバル情報を利用する，状況依存型音声認識システムを提案した．また，ノンバーバル情報の中で特に，顔の表情とパラ言語情報の重要性を示し，それらを利用した感情認識システムを利用して，状況依存型音声認識システムの感情認識部を構成する．

また，対話実験より，人間とロボットの対話において緊張の感情が表出しやすいことが検証された．この感情は持続的な感情であり，また，文脈に依存しにくい感情であるため，表出している場合は文脈に依存する感情の認識が困難になることがわかった．そのため，まず「緊張」の感情の有無を検出し，それがない場合，また緩和された場合に文脈に依存する「喜び」などの感情の認識を行うことが必要であることを示唆した．

今後の課題として，感情認識部を搭載した状況依存型音声認識システムとして，音声認識性能の評価を行う．

謝辞

本研究は情報通信研究機構の研究委託「超高速知能ネットワーク社会に向けた新しいインタラクション・メディアの研究開発」により実施したものである．

また，感情認識システムの利用に関してご指導頂いた，京都大学 河原達也教授，伊藤亮介氏，San Diego 大学 J. R. Movellan 氏，また，データ収集等にご協力頂いた ATR，同志社大学の関係者各位に謝意を表する．

参考文献

[1] 伊藤亮介, 駒谷和範, 河原達也, 奥乃博：“ロボットとの音声対話におけるユーザの心的状況の分析”, 情報処理学会研究会資料, SLP-45-18, （2003.2）

[2] 佐藤賢太郎，広瀬啓吉，峰松信明：“生成過程モデルに基づくコーパスベース感情音声合成とその評価”，情報処理学会研究会資料, SLP-50-8, （2004.2）

[3] 森山剛，斎藤英雄，小沢慎治：“音声表現における感情表現語と感情表現パラメータの対応付け”, 電子情報通信学会論文誌, D-II, Vol.J82-D-II, No.4, pp.703-711, （1999.4）

[4] S. C. Lebinson 著，安井稔 奥田夏子 訳：“英語用語論”, 第 6 章, 研究社出版，1990 年

[5] 松尾太加志：“コミュニケーションの心理学”，第 1 章・第 2 章, ナカニシヤ出版，2000 年

[6] A. Mehrabian 著，西田司 津田幸男 岡村輝人 山口常夫 訳：“非言語コミュニケーション”，第 5 章，聖文社，1986 年

[7] 神田崇行, 石黒浩, 小野哲雄, 今井倫太, 前田武志, 中津良平：“研究用プラットホーム」としての日常活動型ロボット”Robovie”の開発”,電子情報通信学会論文誌, D-I, Vol.J85-D-I, No.4, pp.380-389, （2002.4）

[8] G. Littlewort, M. S. Bartlett, I. Fasel, J. Chenu, T. Kanda, H. Ishiguro and J. R. Movellan：“Towards social robots: Automatical evaluation of human-」robot interaction by face detection and expression classification”, International Conference on Advances in Neural Information Processing Systems, Vol.16, MIT Press, （2003.12）

[9] 濱治世，鈴木直人，濱保久：“感情心理学への招待 ―感情・情緒へのアプローチ―”，第 1 章，サイエンス社，2001 年

[10] 岩瀬佳代子, 伊藤亮介, 神田崇行, 河原達也, 石黒浩, 柳田益造：“日常活動型ロボットの状況依存音声認識”, 情報処理学会関西支部」支部大会 講演論文集, B-04,（2003.10）

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-7 (12/6)

指向性スピーカを用いた人・ロボットコミュニケーション手法の検討

Towards new human-humanoid communication by using ultrasonic directional speaker

中臺 一博, 辻野広司

Kazuhiro Nakadai and Hiroshi Tsujino

(株) ホンダ・リサーチ・インスティチュート・ジャパン,

HONDA Research Insutitute Japan, Co. Ltd.

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

Abstract

自然な人・ロボットコミュニケーションを実現する上で、ロボットの動作中、発話中など高雑音下でロバストな聴覚機能の実現は大きな課題である。本稿では、このうち発話中のロバストな聴覚機能を実現するために、発話デバイスとして指向性スピーカを用いた手法を報告する。指向性スピーカは、超音波と空気の非線形性を利用することによって、指向性の高いビーム状の音場を生成することが可能なスピーカである。また、高い指向性を利用して、ささやくように特定の人物だけに情報を伝えるという新しいコミュニケーション機能実現を併せて試みる。実際に、Honda ASIMO の口の位置に指向性スピーカを実装し、 発話中の孤立単語認識を行った結果、最大 50%の単語正解率の向上を達成した。

1 はじめに

近年、ロボットの分野では、自然な人・ロボットコミュニケーションを目指した研究が盛んに行われている。人間同士のコミュニケーションにおいて言語の果たす役割が大きいことから、人・ロボットコミュニケーションについても、言語はく本質的であるといえる。これは、ロボットに対し、高度に「聞く」、「話す」ことができるという機能が求められていることを示している。

ロボットの聴覚機能を実現するという点では、これまで、「ロボット聴覚」[1]を提案し、研究を行ってきた[13]。我々は、ロボット聴覚を向上させるためには、聴覚と動作を結びつけるアクティブな聴覚が鍵であると考えている。しかし、こうしたアクティブな動作は必然的にモータノイズを発生させる。一般にロボットのマイクは、他の音源に比べてモータに近い位置に設置されているため、たとえモータノイズの絶対パワーが他の音源に比べて小さい場合であっても、マイクが収音するモータノイズのパワーは相対的に大きくなってしまう。従って、動作中の音源定位やモータノイズをキャンセルするための音源分離手法が報告されており[12, 17]、 さらに、分離した音声を認識するように拡張された研究も報告されている[19, 6]。

一方、ロボットの発話機能に関しては、主に下記の３つの構成要素を考慮する必要がある。

1. 自然な人・ロボット間の会話を実現するための非言語情報を含む「対話機能」
2. 自然で柔軟な音声信号を生成する「音声合成機能」
3. 音声を出力する「発話デバイス」

対話機能については、多くの課題が残されているものの、音声だけでなく、ジェスチャ、アイコンタクト、韻律といったマルチモーダルな情報を扱うことができる対話ロボットに関する報告例は多い[9, 10, 5]。

音声合成機能に関しては、波形編集法、声道モデル法など音声合成の分野で多くの手法が提案されている[15]。例えば、PSOLA (pitch synchronous overlap-add) のように波形編集で音声合成を行う手法は、比較的自然で柔軟な音声を合成できる手法として知られている[16]。

発話デバイスに関しては、少数の報告[21]を除き通常のスピーカが用いられている。この種のスピーカは、一般に無指向性であるため、あらゆる方向に音声が伝わるという特徴がある。また、音声の出力パワーはスピーカユニットの位置が一番大きく、スピーカユニットから離れるにつれ減衰していく。このスピーカをロボットの発話デバイスに適用した場合、上述のモータノイズの問題と同様の理由から、発話中に他の音源からの音声を認識することは難しい。加えて、このようなスピーカでは、音声を対話を行う相手に届くように出力するため、出力パワーはモータノイズのパワーに比べて大きい。つまり、信号対雑音 (S/N) 比が小さくなり、このような状況下では、たとえロボット聴覚の分野で研究されているような技術を用い

[1]ロボット聴覚への関心も年々高まっており、IROS 2004 では、初めてロボット聴覚のオーガナイズドセッションが開催された。(http://winnie.kuis.kyoto-u.ac.jp/SIG/ 参照)

ても、現時点では、相手からの音声を認識することは難しい。従って、発話中は聴覚機能をオフにしたり、ロボットのマイクではなく、ヘッドセットなどを用いて発話者の口元にマイクを設置したりすることによってこの問題を避けている研究がほとんどである。

しかし、人間は、話しながら聴くという能力を備えていることから、人とコミュニケーションを行うロボットも、発話中に音声を認識する機能を持つ必要がある。このような機能を実現するために、本稿では、指向性スピーカに着目する。このスピーカは次節で詳細に述べるが、長年にわたって研究され、近年、ようやく商品化されるに至った技術である。この指向性スピーカを用いて、発話中の音声認識機能、および、ささやくように特定の相手のみに音声で情報を伝える機能を扱う。

2 指向性スピーカ

一般に、指向性スピーカは、可聴音による通常のスピーカを用いるものと超音波スピーカを用いるものの２つに大きく分類できる。

前者は、一般的なスピーカを用いる方法であり、様々な手法が提案されている。典型的な手法はホーンや音響管をスピーカの前面に設置するというものであり、拡声器など様々な商品が出回っている。スピーカアレイを用いる手法も広く知られている[11, 8]。基本的に、スピーカアレイは各スピーカから出力される音響信号の位相と振幅を制御することにより高い指向性を得る手法である。

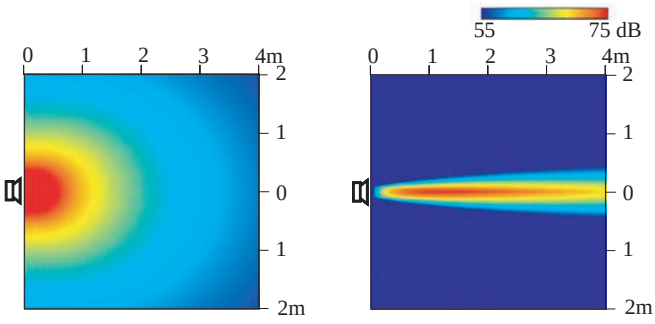


Figure 1: Simulation Result of Speaker Directivity at 1 kHz (三菱電機エンジニアリング（株）提供)

後者は、一般にパラメトリックスピーカアレイといわれている手法である。パラメトリックスピーカの原理が 1963 年に Westervelt によって報告され[18]、以降、50 年以上にわたって、実用化に向けた様々な研究が行われてきた[20, 4, 14]。近年になり、その成果が実り、ようやく、製品が入手可能になった技術である[2, 3]。

この種の指向性スピーカは、超音波の共変調と空気の非線形性を利用して、非常に指向性の高いビーム状の音場を実現する。一般的な無指向性のスピーカとパラメトリックスピーカアレイを用いた指向性スピーカの指向性のシミュレーション結果を、それぞれ、図 1a) および b) に示す。これから、パラメトリックスピーカアレイは下記のような二つの特徴を持っていることがわかる。

- 超音波を搬送波として利用しているため、指向性が高い。
- 空気の非線形性が有効になるまでにある程度音波が空気中を進む必要があるため、可聴音は、スピーカユニットから 0.5 – 1.0 m 離れたところから発生する。つまり、スピーカユニットから 0.5 m 以内には、可聴音がほとんど発生しないことを示している。

Figure 2: Directional Speaker installed in ASIMO

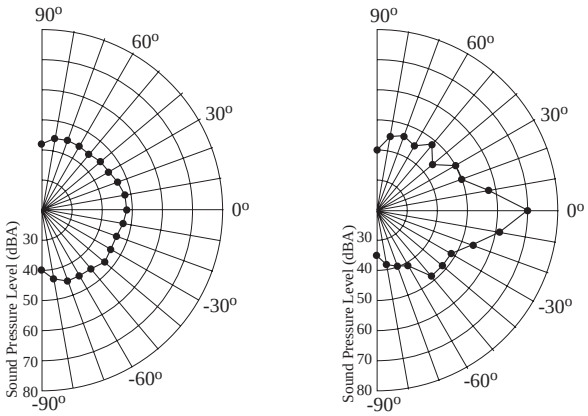


Figure 3: Measured Result of Speaker Directivity at 1 kHz

本稿で用いる指向性スピーカもパラメトリックスピーカである。図 2 は、口の位置に指向性スピーカを実装したホンダ ASIMO の頭部写真である。図 3 は、この指向性スピーカと無指向性スピーカ (GENELEC 1029A) の指向性を実際に計測した結果を示している。計測を行った部屋は、3 m× 5 m の大きさで残響時間が 0.08 秒程度の部屋である。騒音計 HIOKI 3430 をスピーカから 1.0 m の距離に設置して音圧を計測した。音圧は、スピーカの正面方向を 0 度として ±90° の範囲で 10° おきに計測した。計測の指標には、人間の聴覚の感度に近くなるように周波数ごとのパワーの重み付けを行っている dBA を用いた。

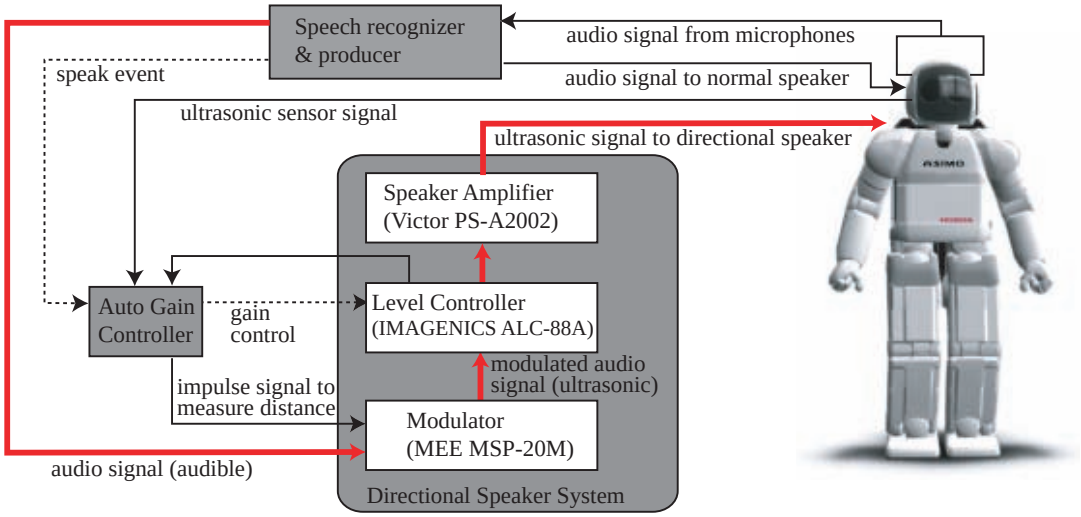


Figure 4: New Communication System Architecture by Directional Speaker

ASIMO に実装された指向性スピーカは、指向性を向けた方向に対して約 20dBA パワーが増加が見られ、これは、図 1b) のシミュレーション結果とよく一致している。この指向性スピーカの音圧は、横方向では不安定である。これは、壁、床、天井の反射波による影響と考えられる。

このように、指向性スピーカを用いるとビーム状の高い指向性が得られる。しかし、超音波を搬送波として利用しているため、信号の減衰率が小さく、反射波が一定のパワーを保ったまま、ロボットのマイクに届いてしまう。従って、話しながら聞く機能を実現するためには、搬送波のゲインコントロールが必要である。次節では、搬送波のゲインコントロール機能を備えたロボット用指向性スピーカコミュニケーションシステムについて述べる。

3 指向性スピーカによるコミュニケーションシステム

指向性スピーカを用いた人・ロボットコミュニケーションのプロトタイプシステムを構築した。構築したシステムのアーキテクチャを図 4 に示す。システムは、「ヒューマノイドロボット」,「指向性スピーカ制御部」,「自動ゲイン制御部」,「音声認識・生成部」という 4 つの構成要素からなっている。

3.1 ヒューマノイドロボット

ホンダ ASIMO をテストベットとして用いた。ASIMO は、本体内部に通常のスピーカと左右の耳の位置に一対のマイクを備えている。今回のシステム用に、指向性スピーカを図 2 に示すように口の位置に実装した。指向性スピーカの内部には、パラメトリックスピーカアレイの他に、超音波センサ（受音器のみ）が実装されている。

3.2 指向性スピーカ制御部

指向性スピーカ制御部は、変調器、音響レベル制御器、スピーカアンプの 3 つの機器からなっている。変調器には、三菱エンジニアリング（株）製の MSP-20M を用いた。この変調器は、入力可聴音によって変調された超音波の搬送波を出力する。搬送波の周波数は、音質・音量の面で最もパフォーマンスのよい 40 kHz 近辺に設定した。変調された搬送波は、音響レベル制御器 (IMAGENICS ALC-88A) に出力される。音響レベル制御器では、自動ゲイン制御部からのコマンドに従って、搬送波のゲインを制御する。音響レベル制御器の出力はスピーカアンプと自動ゲイン制御部の両方に送られる。自動ゲイン制御部へ出力された超音波信号は、対象物までの距離を推定するためのリファレンス用の信号として用いられる。スピーカアンプには、Victor PS-A2002 を用いた。ここで、増幅された超音波信号は、ASIMO に実装された指向性スピーカに送られ、実際に超音波が出力される。

3.3 自動ゲイン制御部

自動ゲイン制御部は、超音波センサによって取得した距離情報に基づき、超音波のパワーを可聴音が目的の人物のみに届くように制御する。人物までの距離は、音響レベル制御器からの超音波信号と指向性スピーカ内に実装されている超音波受信センサから信号の時間差を利用して推定する。ゲイン制御のアルゴリズムを以下に示す。

1. インパルス信号を自動ゲイン制御器から音響レベル制御器に 100 ms 毎に出力する。ただし、音声認識生成部からの発話イベントを受け取った場合は、その内容に応じて、出力を ON/OFF する。
2. インパルス信号によって変調された超音波が、指向性スピーカ制御部で生成され、自動ゲイン制御部に

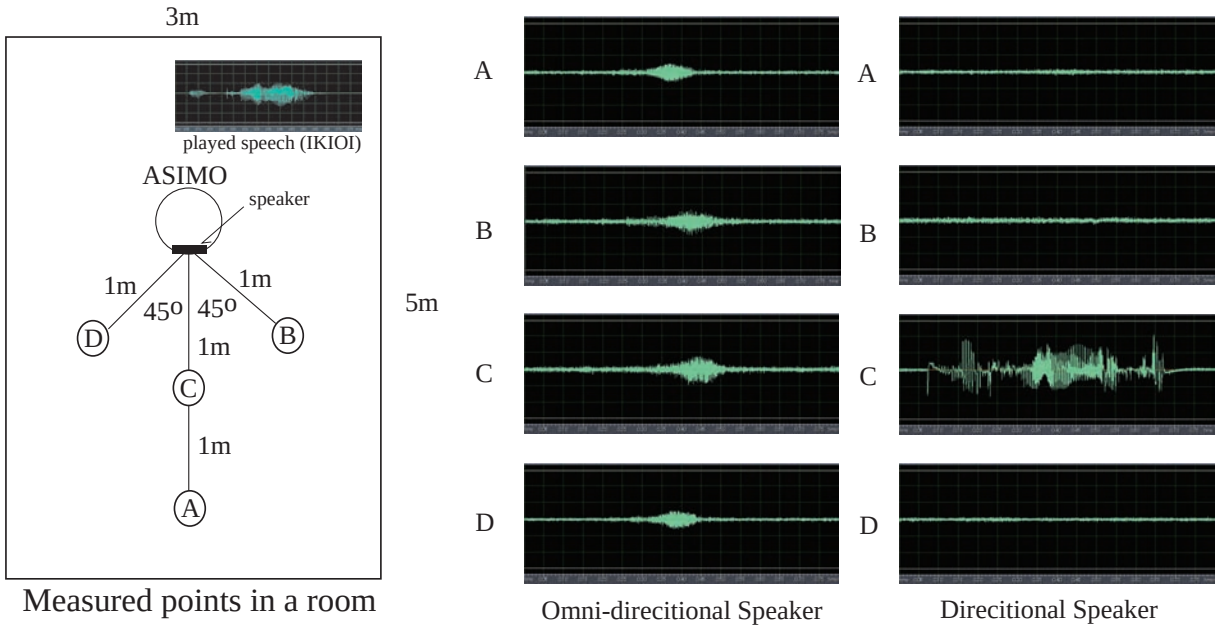


Figure 5: Basic Performance of Communication System

リファレンス信号 s_o として送られる。また、同時に、指向性スピーカへも送られ、超音波が出力される。

3. 指向性スピーカの超音波センサが、ロボットの前にいる人物によって反射した超音波を受信し、自動ゲイン制御部によって、この反射信号 s_r とリファレンス信号 s_o が同時に 192 kHz のサンプリングレートで取り込まれる。
4. インパルス信号の立ち上がり時刻 o_o, o_r を s_o, s_r からゼロクロス法により抽出する。 ロボットと人物間の距離 d は、音速 v (340 m/s) を用いて、以下のように定義される。

$$d = (o_r - o_o) \times v \tag{1}$$

5. 推定された距離に応じて、最適なゲイン値を選択する。最適値は、予め 1 m 間隔で実験的に得た値である。最終的に、選択されたゲイン値を設定するコマンドが RS-232C 経由で音響レベル制御器に送信される。

距離推定の誤差は約 50 cm である。現状の実装では、パラメトリックスピーカ自体を距離測定用の超音波発信器として利用している。インパルス信号が指向性スピーカ制御部に送信される際、変調器によって、インパルス信号が変調され、信号が歪んでしまうため、誤差が大きくなっている。この問題については、今後、送信器と受信器が一体となった超音波センサを別途用意し解決を図る予定である。

3.4 音声認識・生成部

音声認識・生成部は、ASIMO のマイクで収音した音声の認識、および指向性スピーカもしくは無指向性スピーカへ音声信号を送信する。音声認識エンジンには, 京大で開発された Julian を用いた[7]。出力する音声は、事前に録音されたものを用いた。また今回は、指向性スピーカと無指向性スピーカの選択は手動で行った。指向性スピーカからの音声出力の開始/終了時には、それぞれ、距離測定処理を OFF/ON する発話イベントを自動ゲイン制御部に送信する。音声合成や対話処理との統合については今後の課題としたい。

3.5 構築システムの動作例

構築システムの動作例を図 5 に示す。「勢い」という単語を指向性、もしくは無指向性スピーカから出力し、それを図 5 の左図に示される A–D の各地点で計測した。また、左図の波形は、「勢い」の元波形である。中図と右図は、それぞれ、無指向性スピーカ、指向性スピーカから音声が出力された場合の A–D 地点における音声波形を示している。A 点と C 点を比較すると、指向性スピーカの高指向性を保ちつつ、ゲインコントロールもうまく働いていることがわかる。指向性スピーカの音は、実際には、人間の耳にはそれほど歪んだ音には感じられないが、C 点の波形は、元波形と比較して歪んでいる。これは、マイクの周波数特性が人間の耳の周波数特性と異なるためであると考えられる。

4 評価

発話中の聴覚機能を ASIMO と ASIMO の前方にある音源が同時に異なる単語の音声を出力した場合の孤立単語認識によって評価を行う。

	ASIMO ear position	loudspeaker position
omni-directional speaker inside AISMO	70 dBA	62 dBA
directional speaker with max power	58 dBA	70 dBA
directional speaker with optimal power	56 dBA	62 dBA
ASIMO switched on	55 dBA	51 dBA
background noise	23 dBA	23 dBA

Table 1: Sound Pressure Levels in Evaluation

4.1 実験内容と条件

音源用のスピーカには GENELEC 1029A (以後、音源用スピーカ) を用い、ASIMO の正面 1 m の位置に設置した。部屋の残響時間は、1 kHz で 0.08 秒 (RT30) である。

実験は、ATR の音素バランス単語 216 語を下記の 3 つの条件で、音源用スピーカから出力し、孤立単語認識を行った。

1. 指向性スピーカから音声を同時に出力する。
2. 指向性スピーカから音声を同時に出力する。ただし、出力ゲインは音源用スピーカの位置に立っているユーザにのみ伝わるように最適に制御する。
3. ASIMO 内部の無指向性スピーカから同時に音声を出力する。ただし、出力パワーは、音源用スピーカの位置で条件 2 と同じパワーになるよう制御する。

表 1 は、上述の各条件で、音源用スピーカから音声を出力しない場合（指向性もしくは無指向性スピーカからのみ音声が出力される場合）の ASIMO のマイク位置（耳位置）および、音源用スピーカの位置での音声のパワーを示している。指向性スピーカから出力される音声のパワーは、無指向性スピーカの場合と異なり、スピーカの位置より、耳位置（つまりスピーカユニットの側）の方が小さい。

音源用スピーカの出力は、70 dBA から 90 dBA まで 5 dBA 刻みで変更した。ASIMO の耳の位置に音源用スピーカから出力された音が到達するまでに、15 dBA の減衰が見られるため、耳の位置でのパワーの変化は、55 dBA から 75 dBA までとなる。

音声認識用の音響モデルは、ASIMO の電源を ON にして、ロボット以外の騒音源がない状態で、音源用スピーカから ATR 音素バランス単語 216 語の音声を出力し、ASIMO のマイクで収録した各音声を Hidden Markov Model Toolkit (HTK) [1] を用いて、triphone として学習することによって得た。

4.2 実験結果

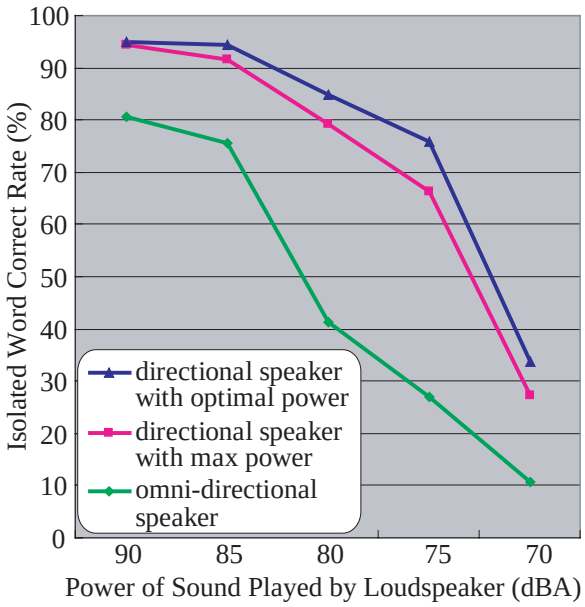


Figure 6: Isolated Word Recognition Result

図 6 に孤立単語認識の結果を示す。横軸はスピーカから出力される音声のパワー (dBA) を、縦軸は孤立単語正解率 (%) を示している。認識結果は、よい順に、最適にゲイン制御を行った場合の指向性スピーカ、ゲイン制御を行わない指向性スピーカ、無指向性スピーカとなった。音声のパワーが 90 dBA の場合、指向性スピーカの単語正解率は、約 95% に達し、無指向性スピーカでは約 80%であった。 無指向性スピーカでは、音声認識は、80 dBA 以下の場合急激に悪化する。指向性スピーカでは、パワーが 70 dB になった場合、同様の傾向が見られた。

4.3 実験の考察

表 1 に示されるように、最適にゲイン制御された指向性スピーカと無指向性スピーカは、両方とも、音源用スピーカの位置で同程度の音声出力レベルであるにもかかわらず、孤立単語認識率には大きな差が見られ、最大で 40% 以上 (出力が 80 dBA の場合) になっている。また、音源用スピーカの位置で、ゲイン制御を行わない指向性スピーカの出力の方が、無指向性スピーカの出力より大きいにもかかわらず、指向性スピーカの孤立単語認識率の方が高くなっている。以上より、一般に、発話中の聴覚機能を実現するための発話デバイスとして、指向性スピーカは無視構成スピーカより性能がよいといえる。

音源用スピーカ出力が 70 dBA まで低下すると、指向性スピーカの孤立単語認識率が悪化した。これは、背景雑音のためだと考えられる。前述したように ASIMO の耳の位置での音声パワーは音源用スピーカ出力が 70 dBA の場合、55 dBA である。表 1 によれば、ASIMO の電源 ON 時

の背景雑音も 55 dBA である。これは、S/N 比が 0 dB であることを示している。従って、背景雑音が強く音声認識結果に影響していると考えられる。この問題を解決するには、S/N 比を改善する前処理として音源分離を行うといった対応が必要であろう。

音源分離に関しては、これまでに特定の方向からの音を抽出するアクティブ方向通過型フィルタを提案した[12]。また、アクティブ方向通過型フィルタと音声中の歪みをマスクすることによって音声認識の向上が可能なミッシングフィーチャ理論の統合を報告した[19]。このような技術を用いたコミュニケーションシステムの向上は今後の課題である。

5 おわりに

本稿では、指向性スピーカを用いた人・ロボットコミュニケーションを提案し、プロトタイプシステムを実装した。発話中の聴覚処理を実現するという観点から、発話中の孤立単語認識実験を通じてコミュニケーションシステムの有効性を示した。

構築したコミュニケーションシステムは、指向性スピーカのゲインをうまくコントロールすれば、スポットライトのようにある特定のエリアのみに音場を生成することが可能であることを示した。これにより、ささやき声のような秘匿性の高いコミュニケーションの実現が可能となる。また、指向性スピーカと無指向性スピーカの両方を組み合わることによって、人間と自然で豊かなコミュニケーションを行うシステムの構築が可能となるであろう。どのようにこれらのスピーカを使い分け、対話を行うかについては今後の課題である。

謝辞

森清文氏を始めとした三菱電機エンジニアリング（株）のメンバー、京都大学奥乃教授、本田技術研究所の吉田雄一氏および、HRI のメンバーに感謝する。

参考文献

[1] http://htk.eng.cam.ac.uk/.

[2] http://www.holosonics.com/products.html.

[3] http://www.mee.co.jp/pro/sales/kokodake/kokodake.html.

[4] K. Aoki, T. Kamakura, and Y. Kumamoto. Parametric loudspeaker–characteristics of acoustic field and suitable modulation of carrier ultrasound. *Electronics and Communications in Japan*, 74(9):76–80, 1991.

[5] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *Proc. of the Sixteenth International Joint Conference on Atificial Intelligence (IJCAI-99)*, pages 1146–1151, 1999.

[6] K. Itou, F. Asano, M. Goto and H. Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. *Eurospeech 2001*, pages 1013–1016, ESCA.

[7] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. Japanese dictation toolkit – 1997 version –. *Journal of Acoustic Society Japan (E)*, 20(3):233–239, 1999.

[8] S. Komiyama, Y. Nakayama, K. Ono, and S. Koizumi. A loudspeaker-arry to control sound image distance. *Acoust. Sci. & Tech.*, 24(5):242–249, 2003.

[9] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services. In AAAI, editor, *Proceedings of AAAI-99*, pages 621–627, 1999.

[10] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface — a robot who communicates with multi-user. *EUROSPEECH-99*, pages 1723–1726. ESCA.

[11] H. Mizoguchi, Y. Tamai, K. Shinoda, S. Kagami, and K. Nagashima. Invisible messenger: Visually steerable sound beam forming system based on face tracking and speaker array. *IROS 2004*.

[12] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano. Real-time speaker localization and speech separation by audio-visual integration. *ICRA-2002*, pages 1043–1049. IEEE.

[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] F.J. Pompei. The use of airborne ultrasonics for generating audible sound beams. *J. Audio Eng. Soc.*, 47:726–731, 1999.

[15] M. Schröder. Emotional speech synthesis: A review. In *Eurospeech 2001*, pages 561 – 564.

[16] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and Nakajima S. A japanese tts syntem based on multi-form units and a speech modification algorithm with harmonics reconstruction. *IEEE Transactions on Speech and Processing*, 9(1):3 – 10, 2001.

[17] 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. ICRA 2004, IEEE.

[18] P. J. Westervelt. Parametric acoustic array. *J. Acoust. Soc. Am.*, 35(4):535–537, 1963.

[19] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno. Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory. *ICRA-2004*, pages 1517–1523, IEEE.

[20] M. Yoneyama, J. Fujimoto, Y. Kawamo, and S. Sasabe. Audio spotlight: An application of nonlinear interaction of sound waves to a new type of loadspeaker design. *J. Acoust. Soc. Am.*, 73(5):1532–1536, 1983.

[21] 今井倫太, 櫻井一人. 狭指向性スピーカを用いたロボットの対話における音声の指向性に関する有用性の実証実験と評価. **第** *66* **回情処全大**, 3P–6, 2004.

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

人工知能学会研究会資料
JSAI Technical Report
SIG-Challnege-0420-8 (12/6)

聴覚フィードバック系を有する人間形発話ロボットの開発

Development of Human-like Talking Robot Having Auditory Feedback System

○福井孝太郎 (早稲田大学理工学部)
西川員史 (早稲田大学理工学部，日本学術振興会特別研究員)
桑江俊治，秋山隆行 (早稲田大学理工学部)
髙信英明 (工学院大学工学部，早稲田大学ヒューマノイド研究所)
持田岳美 (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)
誉田雅彰 (早稲田大学スポーツ科学部)
高西淳夫 (早稲田大学理工学部・ヒューマノイド研究所)

* Kotaro FUKUI, Kazufumi NISHIKAWA, Toshiharu KUWAE, Takayuki AKIYAMA (Waseda University), Hideaki TAKANOBU (Kogakuin University), Takemi MOCHIDA (NTT), Masaaki HONDA (Waseda University), Atsuo TAKANISHI (Waseda University)

Abstract—This paper describes an autonomous control method of an anthropomorphic talking robot WT-4 (Waseda Talker No.4) to mimic continuous human speech sounds by auditory feedback. WT-4 consisted of 1-DOF lungs, 4-DOF vocal cords and articulators (the 7-DOF tongue, 5-DOF lips, 1-DOF teeth, nasal cavity and 1-DOF soft palate), and could reproduce human-like articulatory motion; the total DOF was 19. In this method, the trajectory of each robot parameter was controlled so that the acoustic parameters (pitch, sound power, formant frequencies that are resonant frequencies of the vocal tract and have the peak of the output spectrum, and the timing of the switch between voiced and voiceless sounds) generated from the robot were close to those of human speech sounds. The trajectory of each robot parameter was optimized by inputting the acoustic parameters. This method will help to clarify the human speech mechanism and to create a new speech production system.

1. はじめに

音声言語の生成に関しては，多くの研究がなされているが，未だ脳における発声の運動計画処理機構から運動器官における音声生成の運動までを包括的に研究された例はなく，また人の発声運動は十分に解明されていないのが現状である．1998 年より，科学技術振興機構 (JST) 戦略的創造研究推進事業

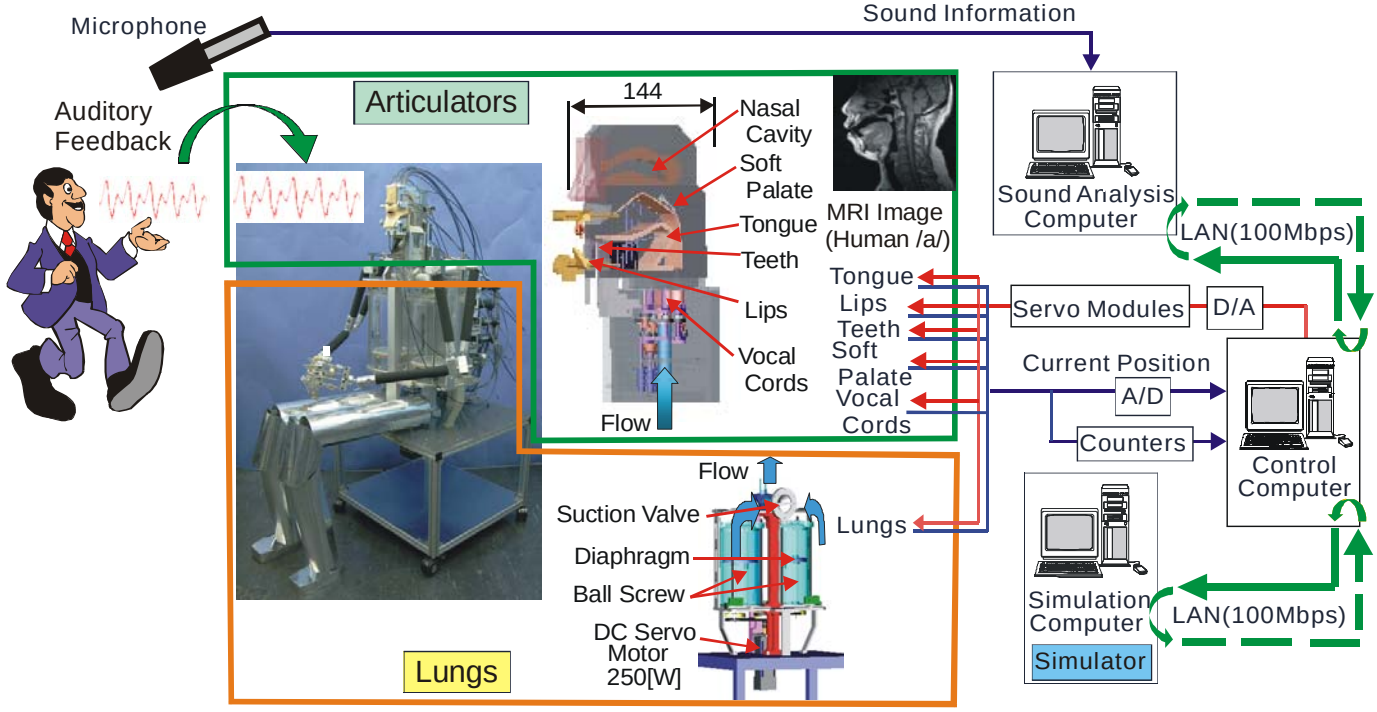
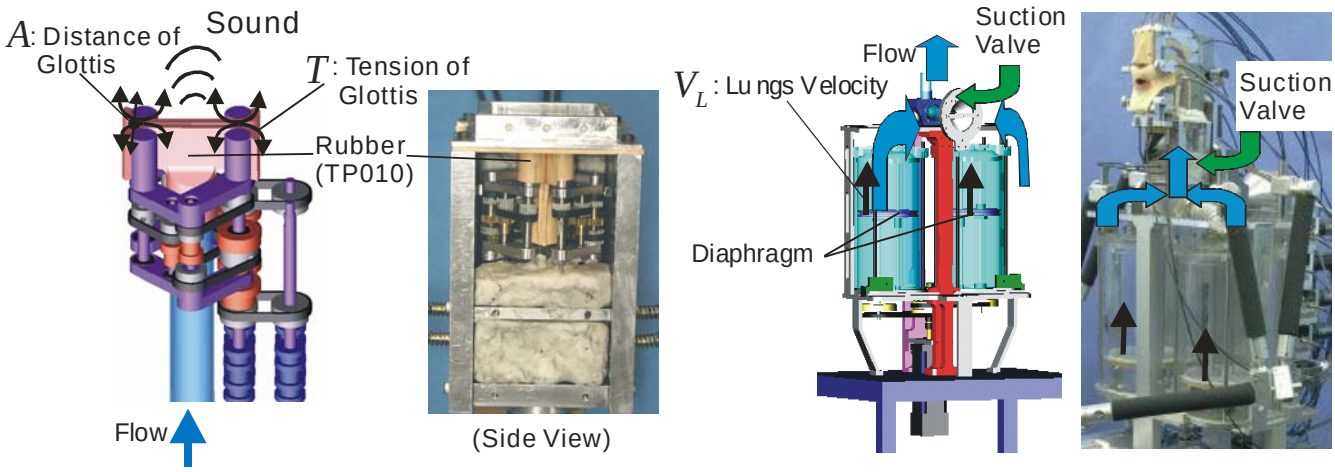


Fig. 1 Mechanical overview and control systems of talking robot WT-4 to mimic human speech sounds by auditory feedback



(a) Vocal cords (b) Lungs

Fig. 2 WT-4's vocal cords and lungs mechanisms and the three manipulated parameters (A, T, V_L)

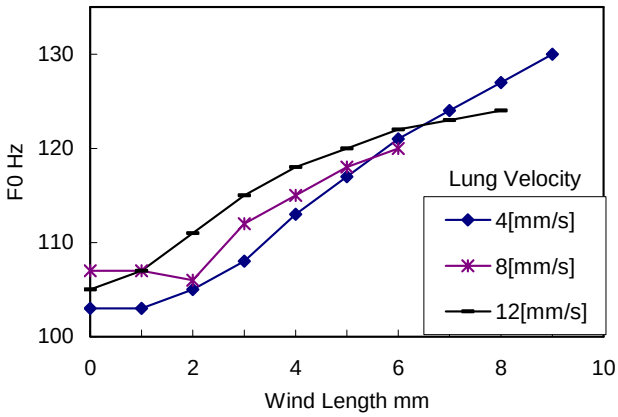
(CREST) 「脳を創る」研究プロジェクト「発声力学に基づくタスクプラニング機構の構築」として，日本電信電話株式会社 (NTT)を中心に全国 10 の医学，音響学および工学の研究機関で人間の発声に関する共同研究が開始された．その中でわれわれは，人間の発声運動を再現する実機械モデルとしての発話ロボットの開発を担当している．

本研究は，発声器官 (肺・声帯)および調音器官(舌・唇・歯・鼻腔)を有し，人間の発声動作を模擬した発話ロボットを開発し，これを用いて人間と同様の発声を実現することにより，計算機シミュレーションのみでは解明困難な発声系のメカニズムをロボット工学的な視点から明らかにすることを目的としている．

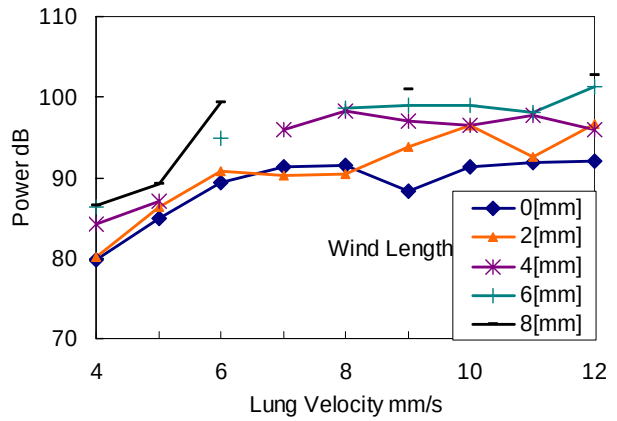
われわれは，人間に近い自然な発声を目指し，WT-2 (Waseda Talker No.2)[(1)]・WT-3 (Waseda Talker No.3)[(2)]を改良し，2004 年に図 1 に示すような人間形発話ロボット WT-4 (Waseda Talker No.4)を開発した．WT-4 は肺・声帯・口腔および鼻腔からなる全 19 自由度の制御機構を有し，声道長さは約 175mm と人間と同程度の大きさを持つ．

また，聴覚フィードバック系を構築し，音響特徴量を用いて発話ロボットの制御パラメータを最適化し，ロボットの制御に利用した．さらに人間の連続発声に対する動的な音響特徴量を抽出し，同手法によりロボットを用いてそれを再現する聞き真似発話を実現した．

本論文では，音響特徴量を用いた発話ロボット WT-4 の制御パラメータの最適化と聞きまね発話の実現について述べる．



(a) Relation between the tension of the vocal cords and the pitch

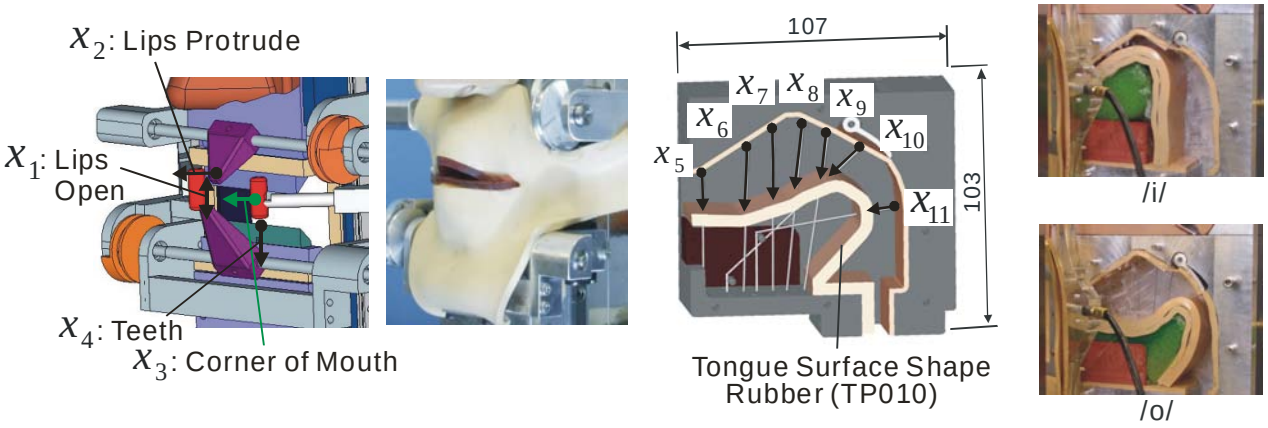


(b) Relation between the lung velocity and the sound power

Fig. 3 Relation of the robot parameters of the vocal cords and lungs and the acoustic parameters

2. 発話ロボット WT-4 の機構

発話ロボットWT-4 (Waseda Talker No.4)は，図 1，2，4 に示すように肺 (1自由度)，声帯 (4自由度)，調音器官である口唇 (5自由度)，歯 (1自由度)，舌 (7自由度)，鼻腔，軟口蓋 (1自由度)の全19自由度を有し，声帯から口唇までの声道長さは175[mm]であり成人男性と同程度の大きさを持つ．その発話器官は，動きを確保しつつ大規模に変形し，かつ空気・音の密閉性を確保しなければならず，そのため弾性体の



(a) Lips (b) Tongue

Fig. 4 WT-4's lips and tongue mechanisms and the eleven manipulated parameters ($x_1, x_2, \ldots, x_{11}$)

超低硬度ゴムTP010[3]で構成されている．上記器官により，明瞭性の高い母音および破裂性子音・摩擦性子音・鼻子音の生成を実現し，日本語の五十音すべての発声が可能となった．

さらに，聴覚フィードバックによるロボットの制御を行っており，第３章に制御パラメータの最適化手法について述べる．

3. 発話ロボットの制御パラメータの最適化手法

聴覚フィードバック系を構築し，音響特徴量を用いて発話ロボットの制御パラメータを最適化し，ロボット制御に利用可能とした．ただし，音響特徴量はマイクで音声を録音し，音声分析ソフトウェア "Praat" (http://www.praat.org/)を用いて分析を行った．下記にその手法について述べる．

3.1 音響特徴量

制御量である音響特徴量として以下のものを用いる．

1)基本周波数 f_0

2)音の強さ P

3)第１・第２フォルマント周波数 f_1・ f_2

4)有声/無声音の切替え時間

この中で，音声の抑揚としての基本周波数・音の強さは，発声器官系である肺・声帯に大きく依存して制御されるパラメータであり，フォルマントは舌や口唇などの調音器官によって大きく制御されるパラメータである．

3.2 ロボット制御パラメータ

発声器官系の操作量は，図２に示すように 1)声帯の張力T，2)声門間距離A，3)肺速度V_Lの３つである．ただし，無声音の発声時，Tは無視される．肺・声帯の制御パラメータT ・V_Lと音声の抑揚 f_0 ・ P に関する実験を行い，図３に示すようにTと f_0，V_L と P には弱線形性があることを確認した．

声道形状を決定する調音器官系の操作量 (調音パラメータ)は，図４に示すように 11 個を定義する．なお，フォルマント周波数を制御量としない無声音については，これらすべてのパラメータが無視される．しかし，11 個の調音パラメータとフォルマント周波数 f_1・ f_2 は冗長で非線形性の関係があり，フォルマント周波数から調音パラメータを決定することは困難である．そこで，下記の手法を用い，ロボットの制御パラメータの最適化を行った．

3.3 最適化アルゴリズム

ロボットの制御パラメータをx，ロボットが発声した音響パラメータをy，目標となる音響パラメータを$\hat{y}$とすると，

$$x=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{11} \\ T \\ V_L \end{bmatrix} \begin{matrix} \left.\begin{matrix}\\\\\\\\\end{matrix}\right\}\text{: Vocal Tract Parameters} \\ \text{: Vocal Cords Tension} \\ \text{: Lungs Velocity} \end{matrix} \tag{1}$$

$$y=\begin{bmatrix} f_0 \\ P \\ f_1 \\ f_2 \end{bmatrix} \begin{matrix} \text{: Pitch Frequency} \\ \text{: Sound Power} \\ \text{: First Formant Frequency} \\ \text{: Second Formant Frequency} \end{matrix} \tag{2}$$

$$\hat{y}={}^t\begin{bmatrix} \hat{f}_0 & \hat{P} & \hat{f}_1 & \hat{f}_2 \end{bmatrix} \tag{3}$$

である．

評価関数$S(x)$として

$$S(x)=\left\| W\cdot\left(\log(\hat{y})-\log(y)\right)\right\|^2 \tag{4}$$

を定義し，ロボットを発声させ，これが最小となるxを図５に示すようにGauss-Newton法[4]の反復改良によって求める．

ただし，

$$\log(y)={}^t\left[\log(f_0)\ \ \log(P)\ \ \log(f_1)\ \ \log(f_2)\right]\in R^4 \quad (5)$$

とする．

ここで，音響パラメータを対数で評価しているのは値の正規化のためであり，

$$\text{重み係数}\ \ W=\begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix} \quad (6)$$

の値は評価関数の重みのみの意味を持つ．

さて，k 回の反復改良によって推定値 $x^{(k)}$ が与えられているとき，その近傍で x の各要素をそれぞれ独立に微小変化させてロボットを発声させ，ヤコビ行列

$$J^{(k)}=\frac{\partial \log\left(y^{(k)}\right)}{\partial x^{(k)}} \quad (7)$$

を観測する．すなわち，

$$J_{ij}^{(k)}=\frac{\Delta \log\left(y_i^{(k)}\right)}{\Delta x_j}\quad \begin{matrix} i=1,2,\cdots,4 \\ j=1,2,\cdots,13 \end{matrix} \quad (8)$$

となる．なお，各ロボットパラメータにおける微小変化分は，

$$\begin{aligned} &\text{(Lips and Tongue)} && \Delta x_i=\begin{cases} 5[mm] & x_i\le 10 \\ -5[mm] & x_i>10 \end{cases}\ (i=1\cdots 11) \\ &\text{(Tension of Glottis)} && \Delta T=1[mm] \\ &\text{(Lung Velocity)} && \Delta V_L=1[mm/s] \end{aligned} \quad (9)$$

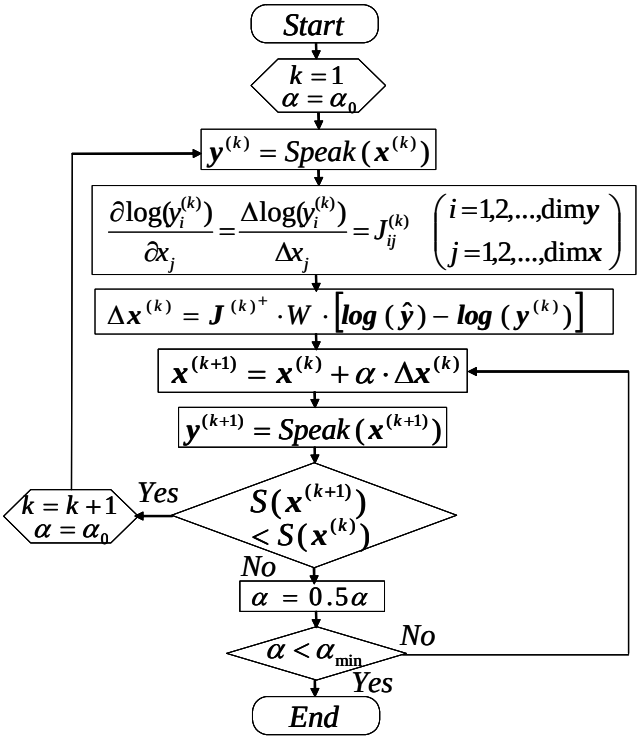


Fig. 5 Optimization algorithm

とした．そして $J^{(k)}$ の一般逆行列 $J^{(k)+}$ を用い，

$$\Delta x^{(k)}=J^{(k)+}+W\cdot\left[\log(\hat{y})-\log(y)\right] \quad (10)$$

$$x^{(k+1)}=x^{(k)}+\alpha\cdot\Delta x^{(k)} \quad (11)$$

により改良した推定値 $x^{(k+1)}$ を得る．ここで，α は収束を安定化させる内部パラメータで，反復の各ステップにおいて評価関数値が改良前よりも減少するように調節する．

以上，この最適化アルゴリズムを用い，$S(x)$ が最小となる（y が $\hat{y}$ に漸近した）x を求める．この最適化の実験を第４章に述べる．

Table 1 Target acoustic parameters $\hat{y}$

Parameter	Value		
	Target 1	Target 2	Target 3
F0 Hz	105	105	105
Power dB	75	75	75
F1 Hz	500	500	650
F2 Hz	1500	1900	1300

Table 2 Initial robot parameters $x^{(0)}$ (Fig. 2-3)

Parameter		Value		
		Initial 1	Initial 2	Initial 3
X_1	mm	10	10	15
X_2	mm	0	0	0
X_3	mm	0	-5	0
X_4	mm	10	5	10
X_5	mm	10	4	14
X_6	mm	10	2	19.5
X_7	mm	10	2	25
X_8	mm	10	6.5	26
X_9	mm	10	8	10
X_{10}	mm	10	15	10
X_{11}	mm	10	23	10
A	mm	0	0	0
V_L	mm/s	7	7	7

(a) Initial 1

(b) Initial 2

(c) Initial 3

Fig. 6 Initial vocal tract shape

４．最適化実験

4.1 母音

WT-3 を用いて母音発声時の音響パラメータのフィードバックによるロボットパラメータの最適化実験を行った．

a) 実験条件

$$W = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 1.0 \end{bmatrix}$$

目標音響パラメータ $\hat{y}$ を表１に，初期ロボットパラメータ $x^{(0)}$ を表２にそれぞれ示す．それぞれの初期パラメータとして大きく異なるものを与えており，(b)は舌先に，(c)は舌後部にそれぞれ狭めを持っている．

b) 実験結果

表１と表２の初期値と目標値を組み合わせ実験を行った．代表的な実験結果として初期値１と目標値１の組み合わせを表３および図７に，初期値２と目標値２の組み合わせを表４および図８にそれぞれ示す．ここで，最適化の際に基本周波数の最適化を優先させたため，音圧の誤差は基本周波数の誤差より

も大きくなっている．しかし，実験により推定された値は目標音響パラメータに十分に接近していることが確認できる．以上のように本最適化手法が母音発声における音響パラメータからロボットパラメータへの逆変換に有効であることを確認した．

4.2 連続発話実験

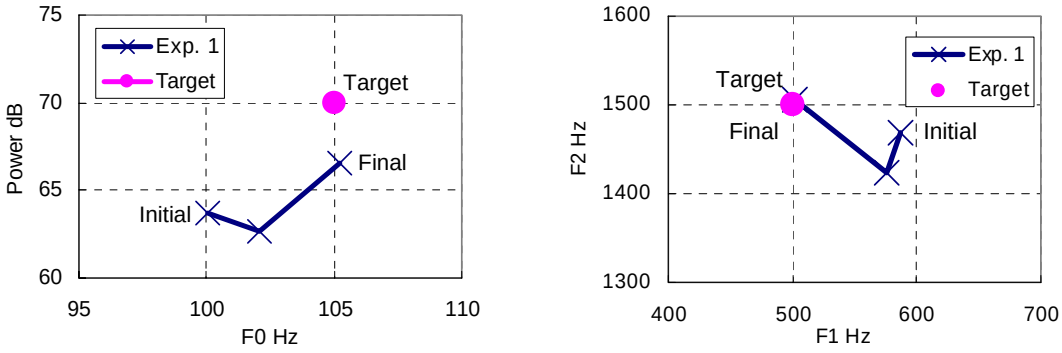
人間の連続発声に対する動的な音響特徴量を抽出し，第３章の最適化手法を用い，フレーム毎の WT-4

Table 3 Organized experiment 1 (Target 1 and Initial 1)

Parameter	Value			
	Initial		Estimated	Target
F0 Hz	100.0	→	105.2	105
Power dB	63.7		66.5	75
F1 Hz	586.4		501.4	500
F2 Hz	1467.7		1504.9	1500

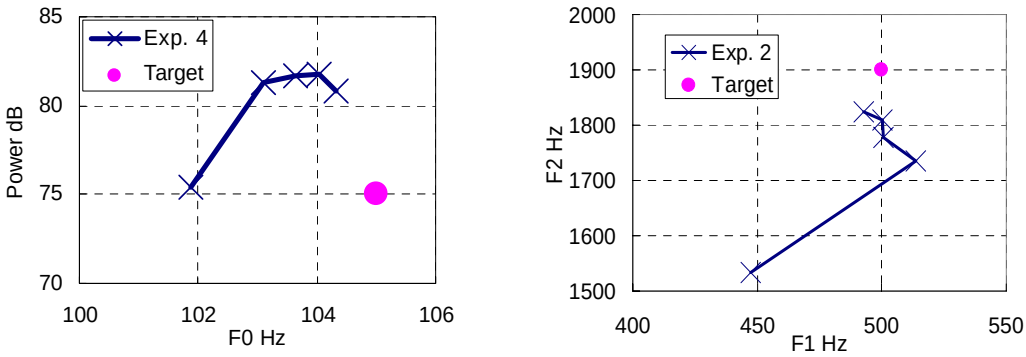
Table 4 Organized experiment 2 (Target 2 and Initial 2)

Parameter	Value			
	Initial		Estimated	Target
F0 Hz	101.9	→	104.3	105
Power dB	75.4		80.8	75
F1 Hz	447.5		492.8	500
F2 Hz	1534.0		1824.6	1900



(a) F0-Power　　(b) F1-F2

Fig. 7 Organized experiment 1 (Target 1 and Initial 1)



(a) F0-Power　　(b) F1-F2

Fig. 8 Organized experiment 2 (Target 2 and Initial 2)

の制御パラメータ最適化による人間の連続音声の聞きまね発話を試みた．その実験条件は下記である．

a) 実験条件

1) 目標音声: /hassei/ (成人男性の発声音声)

2) フレーム間隔: 50 [ms]

ただし，母音の有声音区間のみ最適化を行い，子音の無声音区間は WT-4 の子音発声時の制御パラメータを参考とした．

b) 実験結果

最適化実験の結果の基本周波数 f_0 と第1・第2フォルマント周波数 f_1・f_2 の変化を図 9 に示す．図より人間の音声の特徴をまねたロボットの基本周波数・フォルマント変化を確認できる．しかし，声帯の不安定性などの問題のため，有声化後の2・3フレームは誤差を有しており，今後声帯の改良を行う．

５．結論と今後の展望

発話ロボット WT-4 の聴覚フィードバック系を構築し，WT-4 の生成音声から抽出される音響特徴量を用いてロボットの制御パラメータを最適化し，制御に利用可能とした．音響特徴量としては基本周波数，音の強さ，フォルマント周波数，有声/無声音の切替え時間を用い，全 19 自由度のロボットの制御パラメータの最適化を行う．さらに人間の連続発声に対する動的な音響特徴量を抽出し，同手法によりロボットを用いてそれを再現する聞き真似発話を実現した．

本システムは，聴覚フィードバックによる人間の発話獲得動作を再現することを目的としており，今後は同システムを発展し，発話の脳内情報生成メカニズムの解明を目指す．

謝辞

本研究は，科学技術振興機構(JST) 戦略的創造研究推進事業(CREST)の援助を受けた．研究に協力して頂いた共同研究プロジェクトの研究者各位，また機構部製作に協力して頂いたオキノ工業株式会社の沖野晃久氏，3D-CAD ソフトウェアを提供して頂いたソリッドワークス・ジャパン株式会社，テフロン被覆ワイヤを提供して頂いた中興化成株式会社に感謝致します．

参考文献

1) 西川，林，桑江，棚橋，髙信，持田，誉田，高西：人間形発話ロボットにおける母音および子音発声の実現，第 20 回日本ロボット学会講演会予稿集 (2002).

2) 西川，小河原，池尾，藤田，髙信，持田，誉田，高西：人間に近い声帯・声道形状変更機構を有する新型発話ロボットの開発，第 21 回日本ロボット学会講演会予稿集 (2003).

3) 東京ゴム株式会社製，ショア硬さ: 1 (JIS-A)，引張強さ: 5.9 [MPa]，材質: EPDM (Ethylene Propylene Diene Monomer

4) W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery: Numerical Recipes in C, Cambridge University Press (1992).

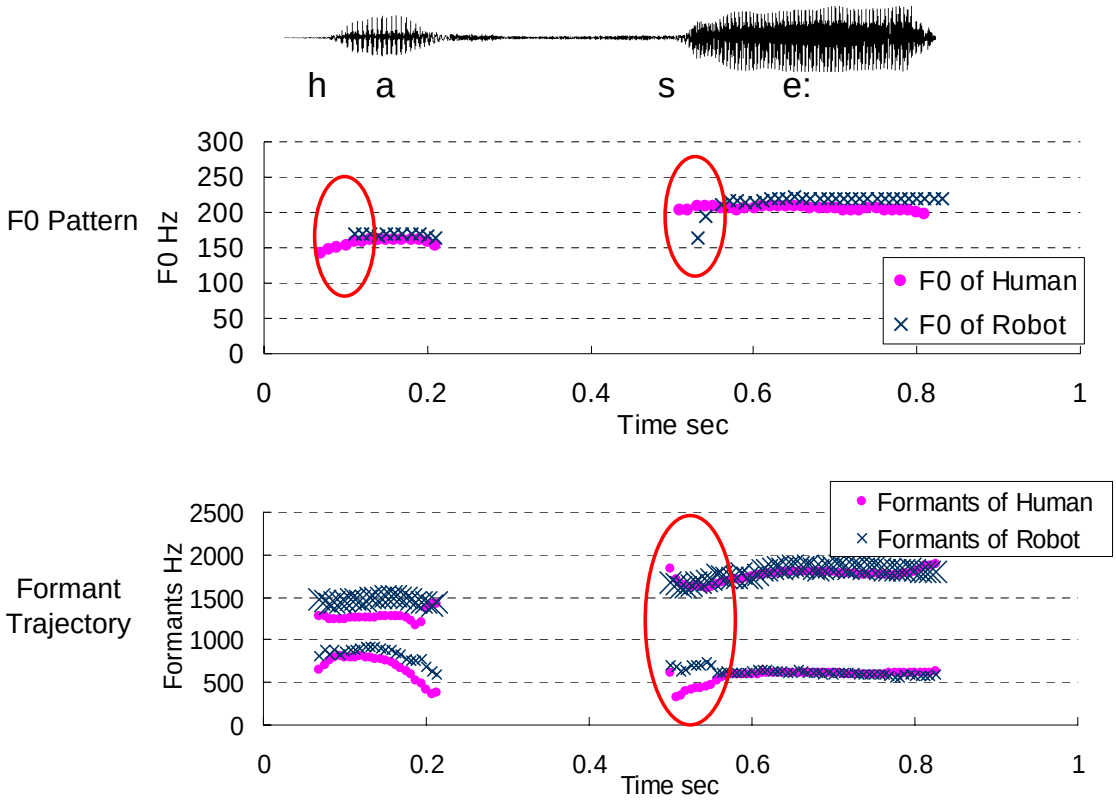


Fig. 9 Organized experiment to mimic human speech sounds “hassei”

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

ＡＩチャレンジ研究会
主 査
奥乃 博
京都大学大学院 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町
075-753-5376 Fax: 075-753-5977
okuno@i.kyoto-u.ac.jp

幹 事
浅田 稔
大阪大学大学院 工学研究科
知能・機能創成工学専攻
中臺 一博
(株) ホンダ・リサーチ・インスティチュート
・ジャパン
光永 法明
(株) ATR 知能ロボティクス研究所

Executive Committee
Chair
Hiroshi G. Okuno
Dept. of Intelligence Science and Technology,
Gradulate School of Informatics
Kyoto University
Yoshida-Honmachi Sakyo, Kyoto 606-8501 JAPAN

Secretary
Minoru Asada
Dept. of Information and Intelligent Engineering
Graduate School of Engineering
Osaka University
Kazuhiro Nakadai
Honda Research Institute Japan
Noriaki Mitsunaga
ATR Intelligent Robotics and Communication Laboroatories

SIG-AI-Challenges home page (WWW):
http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/