Vol.2013-HCI-151 No.20
Vol.2013-SLP-95 No.20
2013/2/2

# 複数人会話におけるロボットによる視聴覚情報に基づくアクティブユーザの推定

中島 大一[1,a)]　駒谷 和範[1]　佐藤 理史[1]

**概要**：複数人会話では，積極的に会話に参加するユーザの位置や人数をシステムが推定できるのが望ましい．このような推定により，参加ユーザの状況に応じた発話を生成できる．本研究では，2 体のロボットに搭載されたマイクロフォンとカメラから得られる視聴覚情報を，その時点におけるアクティブユーザの推定に用いる．まず，2 体のロボットから得られる音源定位結果と顔検出結果を，確率密度関数として表現し，アクティブなユーザの存在する位置の確率分布を得る．それらを入力が得られる度に更新し，積極的に会話に参加するユーザの人数や位置を推定する．音源定位結果と顔検出結果それぞれに基づく推定結果を組み合わせることで，参加ユーザの様々な状況を判定し，それに基づく発話を生成する．評価実験では，ロボットとユーザの実際の会話において得られたデータを用いて，アクティブユーザを推定し，その時のユーザの状況に応じた発話を生成できることを示した．

## 1. はじめに

ヒューマノイドロボットを用いた複数人会話システムの開発を行っている．複数人会話システムとは，2 人以上のユーザと会話するシステムである [8]．本稿では，ロボットの動作の決定や，ロボットの入力に基づき推定を行うものをシステム，実際に動作や発話を行うものをロボットと呼ぶ．

複数人会話においては，参加ユーザが全員，積極的に発話を行えるのが望ましい．そのような会話を実現するためには，システムは参加ユーザの状況を踏まえた発話を行えることが重要である．例えば，ある時点までに，あまり発話を行っていないユーザを推定し，「何か聞きたいことはないですか?」と発話を行うことで，そのユーザを会話に参加させることができる．我々はこれまでに，発話したユーザがどこにいるのかを特定する発話者の特定に取り組んで来た [6]．発話者の特定により，発話したユーザに対してロボットが顔を向け応答を行うシステムを実現した．我々のシステムは一問一答型であり，ロボットはユーザの質問に対する発話しか行わない．参加ユーザの状況を踏まえた発話を行うためには，ある一定時間内に積極的に会話に参加したユーザの位置や，人数を推定する必要がある．本研究では，ロボットに搭載されたマイクロフォンやカメラから得られる視聴覚情報のみを用いて，そのような推定を行う手法を提案する．

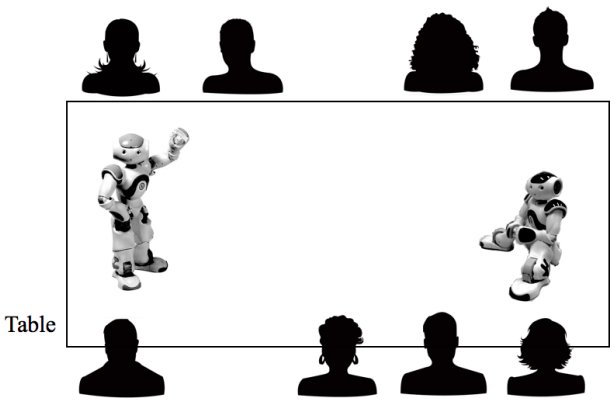


**図 1**　想定する会話状況．複数ユーザが机を囲み，机の上に配置したロボットと会話を行う．

本研究では，ある一定時間内に積極的に会話に参加したユーザを，アクティブなユーザする．このアクティブさを以下のように定義する．

- 時間とともに変化する．つまり，会話全体ではなく，ある一定時間内のアクティブさを対象とする．
- ロボットを含めた参加者全体で共有される内容に関して，発話を行うユーザはアクティブである．つまり，ロボットの発話内容に関係なく，一部のユーザ間で会話を行うユーザは，積極的に会話に参加しているとは言えない．
- 何度も発話を行うユーザはアクティブである．一度だけ発話しただけで，以後黙ってしまうユーザは積極的

[1] 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University
[a)] taichi_n@nuee.nagoya-u.ac.jp

**表 1** 音源定位結果と顔検出結果の組み合わせに基づくユーザの状態の推定

| | 音源定位結果あり | 音源定位結果なし |
|---|---|---|
| 顔検出結果あり | 積極的に会話に参加している | 発話する機会を伺っている |
| 顔検出結果なし | ユーザ同士で会話を行っている | 存在していない, または全く会話に参加していない |

に会話に参加しているとは言えない.

- 発話を行う参加者に対して顔を向けるユーザはアクティブである．全く関係のない方向を向いているユーザは，積極的に会話に参加しているとは言えない.

本稿では，ある時点においてアクティブさの高いユーザをアクティブユーザと呼ぶ．このようなアクティブユーザの位置や人数を，ロボットに搭載されたマイクロフォンとカメラから得られる視聴覚情報のみを用いて推定する．つまり，特別なセンサを準備したり，カメラやマイクロフォンが多く設置されたスマートルーム環境 [3] を前提としない．複数人会話を行う状況として，図 1 に示すような複数ユーザが机を囲み，その机の上に配置された 2 体のロボットと行う会話を設定した．この状況設定は，複数人会話におけるユーザの位置の決定を簡単化するものである．つまり，複数のユーザが机を囲んだ状況はユーザの位置を机の周辺に限定できる．そして，ユーザの移動は前提としない.

本研究では，視聴覚情報をアクティブなユーザが存在する位置の確率分布として表現し利用する．聴覚情報には，音源の到来方向を示す音源定位結果を，視覚情報には，カメラの視野角内のユーザの顔の位置を示す顔検出結果を用いる．本研究では，音源定位結果を発話を行うユーザの位置，顔検出結果をロボットに顔を向けるユーザの位置として解釈する．それらの結果を確率密度関数として表現し，アクティブなユーザの存在する位置を確率分布として得る．会話の中でユーザのアクティブさは時間とともに変化する．そのため，ユーザの位置の確率分布を時間ごとに更新し，ユーザのアクティブさを得る．得られたアクティブさのピークを検出することで，その時点のアクティブユーザの人数や位置を推定する.

さらに，我々は音源定位結果から得られるユーザのアクティブさと，顔検出結果から得られるユーザのアクティブさを別々に維持する．それらの推定結果の組み合わせにより，様々なユーザの状況を判定し，それに基づく発話を生成する．表 1 に，音源定位結果が得られる場合と得られない場合，顔検出結果が得られる場合と得られない場合の組み合わせにより判定されるユーザの状況を示す．表 1 より例えば，顔検出結果が得られるのにも関わらず，音源定位結果が得られないユーザは，ロボットの方を見て積極的に会話に参加しようとしているが，発話の機会が得られないユーザであると判定できる．この結果を併用することでロボットは，そのユーザに対して発話を促すなど，ユーザの状況に応じた発話の生成が可能となる.

本論文では，まず関連研究について述べ，我々の研究を位置づける．3 章では，アクティブユーザの推定方法について述べる．4 章では，評価実験として実際に収録したデータを用いてアクティブユーザが推定できることを示す.

## 2. 関連研究

今日まで開発されてきた複数人会話に参与するロボットやバーチャルエージェントは，それらの想定する会話状況に応じて，参加ユーザの状況の推定を行ってきた．図 1 に示す本研究で想定する会話状況では，以下の 3 つが前提として存在する.

(1) ユーザがカメラの視野角内に常に存在するとは限らない．これはロボットに搭載されたカメラの視野角が狭いためである.
(2) 狭い視野角を補うために，常に周りを見回し続けるのは不適当である．これはロボットが発話の当事者であり，このような挙動は会話において不自然であるためである.
(3) ユーザが机を囲んで行う会話を行う．つまり，会話中のユーザの移動を想定しない.

Lang らのシステムは，レーザや広角カメラ，マイクロフォンを用いて，システムに興味を持つユーザを推定する [4]．システムは，カメラの視野角内に存在し，かつ発話を行うユーザを，システムに興味をもつユーザであると認定し，そのユーザに注意を向ける．Bohus らは，複数のユーザが会話に参加したり，退出したりする会話において，画像情報に基づき，ユーザの参加状況を推定する手法を提案した [2]．これにより，会話に参加すると推定されるユーザに対して注意を向け，参加を促す発話を行う．これらの研究では，広角カメラを用いており，ユーザが常にカメラの視野角内に存在していると仮定している．Bennewitz らは，ロボットに搭載されたカメラを通した顔検出に基づき，参加者の存在について確率的な信頼度を維持する手法を提案した [1]．システムは，その信頼度に応じて，ある時点で重要なユーザを決定し，そのユーザに注意を向ける．さらに，定期的に周囲を見渡すことで，信頼度の更新を行う．この手法は，常にカメラの視野角内にユーザを捉えておく必要はないが，一度もカメラの視野角内に現れないユーザを，信頼度の対象にはできない.

本研究では，視覚情報のみでなく，音源定位結果も同時に用いる．これにより，カメラの視野角外のユーザの状況も推定できることが期待できる．さらに，2 つの情報を組

Vol.2013-HCI-151 No.20
Vol.2013-SLP-95 No.20
2013/2/2

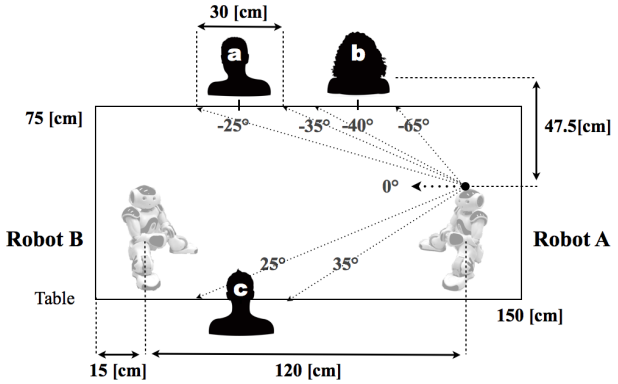


**図 2** ロボットとユーザの位置関係

み合わせることで，視覚情報のみでは得られないユーザの状況を得る．

## 3. ユーザがアクティブである度合の推定

本研究では，ロボットに搭載されたマイクロフォンやカメラから得られる音源定位結果と顔検出結果を，ある時点でのアクティブユーザの人数と位置の推定に用いる．会話の中でユーザのアクティブさは時間の経過とともに変化する．そのため，ユーザの位置の確率分布を時間ごとに更新し，ユーザのアクティブさの時間変化を表現する．ある時点に得られたアクティブさのピークを検出し，その時点におけるアクティブユーザの人数と位置を推定する．さらに，推定結果に基づきユーザの状況を判定し，それによりロボットが新たな発話を生成可能であることを示す．

図 2 に，我々の想定する，机を囲んで行う会話状況を示す．ここでは，Robot A の正面方向を 0 度とし，反時計回りを正方向とする座標系 (例えば，ロボットの左手方向は 90 度) を採用している．2 体のロボットそれぞれから得られる音源定位結果と顔検出結果を，すべてこの座標系に対応させる．この座標系は，机の周囲の位置を，Robot A から見た角度で示している．

### 3.1 音源定位結果に基づくユーザの存在確率

2 体のロボット (図 2 の Robot A，Robot B) に搭載されたマイクロフォンを通して，音源定位結果とそのパワーを得る．音源定位結果は，音源の到来方向を角度で示す．角度が得られれば，図 2 の状況から，それをユーザの位置へと一意に変換できる．Robot B の音源定位結果は，座標変換により Robot A の座標系に対応させる [6]．

音源定位結果から得られる，ユーザの存在する位置を確率分布で表現する．音源定位結果は常に正しい位置を示しているわけではなく，雑音等による誤検出が避けられない．そのため，雑音等による音源定位結果のパワーは小さいとし，パワーによる重み付けを行うことで，発話による正しい定位結果とそうでないものを区別する．ある時刻 $t$ において，音源定位結果 $\theta_{r,t}$ とパワー $p_{r,t}$ が得られた場合，定

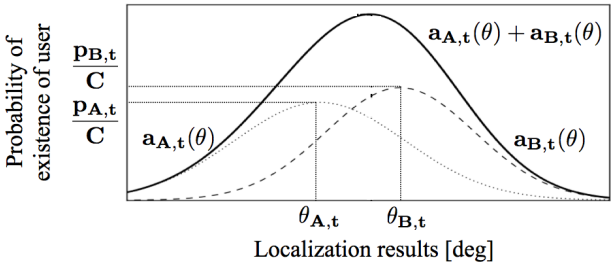


**図 3** 確率密度関数の足し合わせの例

位結果の曖昧さは正規分布に従うと仮定し，確率密度関数 $a_{r,t}(\theta)$ を定義する (式 1)．ここで，$r$ は ID を示す，例えば，Robot A の，ある時刻 $t$ における音源定位結果を，$\theta_{A,t}$ と表現する．式 1 において，$\sigma^2_{r,t}$ は分散であり，音源定位結果がどれだけ不確かであるかを示す．

$$a_{r,t}(\theta) = \frac{1}{\sqrt{2\pi\sigma^2_{r,t}}} \exp(-\frac{(\theta-\theta_{r,t})^2}{2\sigma^2_{r,t}}) \quad (1)$$

この確率密度関数の最大値 $a_{r,t}(\theta_{r,t})$ が，音源定位結果 $\theta_{r,t}$ のパワー $p_{r,t}$ に比例すると仮定する (式 2)．この仮定は，パワー $p_{r,t}$ が大きいほど，音源定位結果が $\theta_{r,t}$ である確率が高くなることを示す．ここで，式 2 の $\alpha$ は定数であり，実験的に決定する．

$$a_{r,t}(\theta_{r,t}) = \frac{1}{\sqrt{2\pi\sigma^2_{r,t}}} = \frac{1}{\alpha} p_{r,t} \quad (2)$$

式 2 より $\sigma_{r,t}$ を定める (式 3)．式 3 より，$\sigma_{r,t}$ はパワー $p_{r,t}$ に反比例する．つまり，パワーが大きいほど音源定位結果は散らばりが小さいとしている．

$$\sigma_{r,t} = \frac{\alpha}{\sqrt{2\pi}} \frac{1}{p_{r,t}} \quad (3)$$

ある時刻 $t$ に，2 体のロボットから同時に音源定位結果 $\theta_{A,t}$，$\theta_{B,t}$ とそのパワー $p_{A,t}$，$p_{B,t}$ が得られる場合，それぞれ確率密度関数 $a_{A,t}(\theta)$，$a_{B,t}(\theta)$ を定義する．そして，式 4 により，統合によるユーザの位置の確率分布 $a_{mix,t}(\theta)$ を求める．

$$a_{mix,t}(\theta) = \frac{1}{2}(a_{A,t}(\theta) + a_{B,t}(\theta)) \quad (4)$$

確率密度関数 $a_{r,t}(\theta)$ の例を図 3 に示す．グラフの横軸は音源定位結果を示し，縦軸はユーザの存在確率を示す．

### 3.2 顔検出結果に基づくユーザの存在確率

Robot A のカメラを通して顔検出結果を得る．顔検出結果は，ユーザの顔の位置の，視野の中心からの水平角度として得られる．これを，現在のロボットの首の角度と同時に用いることで，推定を行う座標系における，ユーザの顔の位置が得られる．

この顔検出結果を，ロボットに対して顔を向けているユーザの位置と解釈し，その位置を確率分布で表現する．ある時刻 $t$ において，ロボットの水平方向の首の角度 $H_t$

と視野角内での顔の検出位置 $\theta_{k,t}$ が得られたとき，確率密度関数 $v_{k,t}(\theta)$ を定義する (式 5)．ここで $k$ は同時に検出された顔の ID を示す．$1 \leq k \leq n$ とし，$n$ は時刻 $t$ にロボットの視野角内で検出された顔の数である．検出結果の位置の不確かさを示す $\sigma_v$ は，顔検出の性能に基づき実験的に決定する．

$$v_{k,t}(\theta) = \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp(-\frac{(\theta - (H_t + \theta_{k,t}))^2}{2\sigma_v^2}) \quad (5)$$

ある時刻 $t$ において，同時に検出された顔の ID ごとに $v_{k,t}(\theta)$ を定義し，式 6 によりユーザの存在確率 $v_{mix,t}(\theta)$ を求める．

$$v_{mix,t}(\theta) = \frac{1}{n}\sum_{k=1}^{n} v_{k,t} \quad (6)$$

### 3.3 確率分布の更新方法

音源定位結果と顔検出結果に基づき定義した確率分布を時間ごとに更新し，ユーザのアクティブさの時間変化を表現する．会話の開始時刻を $t = 0$ として，時刻 $t$ における音源定位結果に基づくアクティブさを $A_t(\theta)$，顔検出結果に基づくアクティブさを $V_t(\theta)$ とする．まず，時刻 $t = 0$ におけるユーザのアクティブさを式 7 で定義する．

$$\begin{aligned} A_0(\theta) &= V_0(\theta) = S(\theta) \\ S(\theta) &= \begin{cases} N & (-180 \leq \theta \leq 180) \\ 0 & (else) \end{cases} \\ &\text{ただし} \int_{-180}^{180} N \,\mathrm{d}\theta = 1 \end{aligned} \quad (7)$$

ここで $S(\theta)$ はステップ関数である．つまり，式 7 は，時刻 $t = 0$ において，ユーザのアクティブさはどの角度においても一様であることを示している．

ある時刻 $t$ において，音源定位結果に基づくユーザの存在位置を示す確率分布 $a_{r,t}(\theta)$，顔検出結果に基づく確率分布 $v_{k,t}(\theta)$ が得られた場合，$A_t(\theta)$ と $V_t(\theta)$ を式 8，9 により更新する．

$$A_t(\theta) = \lambda_1 a_{r,t}(\theta) + (1 - \lambda_1) A_{t-1}(\theta) \quad (8)$$

$$V_t(\theta) = \lambda_2 v_{k,t}(\theta) + (1 - \lambda_2) V_{t-1}(\theta) \quad (9)$$

ここで，$\lambda_1$, $\lambda_2$ $(0 \leq \lambda_1, \lambda_2 \leq 0)$ は，事前に蓄積されたアクティブさと時刻 $t$ に得られた確率分布の，足し合わせの重みを表す．$\lambda_1$,$\lambda_2$ が大きい程，その時点の検出結果が更新に優先され，小さいほど事前に蓄積されたアクティブさが優先される．

ある時刻 $t$ において，確率分布が得られない場合は，式 10，11 により，$A_t(\theta)$ と $V_t(\theta)$ を更新する．

$$A_t(\theta) = \lambda_1 S(\theta) + (1 - \lambda_1) A_{t-1}(\theta) \quad (10)$$

$$V_t(\theta) = \lambda_2 S(\theta) + (1 - \lambda_2) V_{t-1}(\theta) \quad (11)$$

$S(\theta)$ は一様分布のステップ関数である．つまり，音源定位結果または顔検出結果が得られない場合，ユーザの存在確率は一様であることを示す．事前に蓄積されたアクティブさと $S(\theta)$ を $\lambda_1$, $\lambda_2$ による重みで足しあわせることで，アクティブさは減衰する．$\lambda_1$, $\lambda_2$ が大きいほど，アクティブさの減衰は速くなり，小さいほど遅い．これらの値は，実験的に決定する．

### 3.4 アクティブユーザの推定方法

ある時刻 $t$ におけるアクティブさのピークを検出することで，その時点でのアクティブユーザの人数と位置を推定する．時刻 $t$ において，アクティブさ $A_t(\theta)$ と $V_t(\theta)$ が得られたとき，閾値以上の極大値の数をアクティブユーザの人数，その角度をユーザの位置とする．閾値を設定することで，雑音等の誤検出による小さなピークを削減する．$A_t(\theta)$ の閾値を $T_a$，$V_t(\theta)$ の閾値を $T_v$ とする．この閾値は，実験的に決定する．

### 3.5 推定結果から推定される状況とそれに基づく発話

音源定位結果によるアクティブさと，顔検出結果によるアクティブさを用いた推定結果の組み合わせにより，1 章の表 1 に示したようなユーザの状況を判定する．さらに，それに基づきロボットが生成可能な発話を示す．

ある時刻 $t$ において，アクティブさ $A_t(\theta)$ よりその時点のアクティブユーザの位置 $P_a$ が，$V_t(\theta)$ よりアクティブユーザの位置 $P_v$ が得られたとする．このとき，$P_a$ と $P_v$ の差が $D$ より小さい場合に，2 つは対応しているとする．ここでは $D = 10$ とした．

このとき，以下のような条件により，ユーザの状況が推定でき，それに基づき発話の生成が行える．

**条件 1** $P_a$ に対応する $P_v$ が存在しない場合．つまり，音源定位結果は得られるが，顔検出結果が得られない場合である．このとき，$P_a$ のユーザは発話を行っているが，ロボットに顔を向けていないと判定できる．そのため，その位置のユーザは，ユーザ同士で会話を行っていると推測できる．これより，ロボットは「私の話しを聞いていますか?」や「私と一緒に会話しましょう」と，そのユーザに会話への参加を促すことができる．

**条件 2** $P_v$ に対応する $P_a$ が存在しない場合．つまり，顔検出結果は得られるが，音源定位結果が得られない場合である．このとき，$P_v$ のユーザはロボットに顔を向けているが，発話を行っていないと判定できる．そのため，その位置のユーザは，会話には参加しているが発話の機会を伺っていると推測できる．これより，ロボットは「何か聞きたいことはないですか?」と，そのユーザに発話を促すことができる．

**条件 3** $P_a$ に対応する $P_v$ が存在する場合．つまり，音源定位結果も，顔検出結果も得られている場合である．

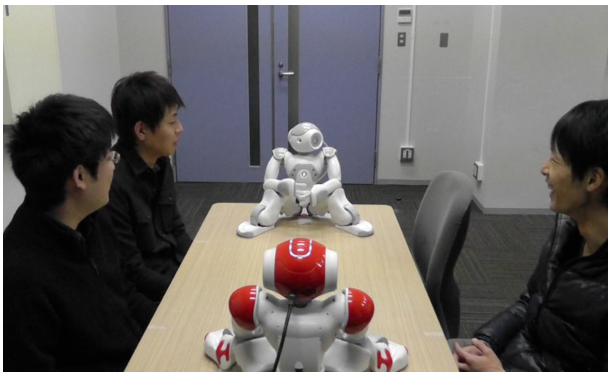

**図 4** システムとユーザの会話の様子．発話したユーザを特定し，顔を向けて応答を行う．

**表 2** ユーザの領域ごとの発話回数 (発話フレーム数)

| 領域 | 発話回数 (フレーム数) | | | | | |
|---|---|---|---|---|---|---|
| | システム | | それ以外 | | 合計 | |
| a | 4 | (718) | 3 | (1399) | 7 | (2117) |
| b | 3 | (347) | 1 | (81) | 4 | (428) |
| c | 6 | (1021) | 5 | (1547) | 11 | (2568) |
| 合計 | 13 | (2086) | 9 | (3027) | 22 | (5113) |

$P_a$ のユーザは発話を行い，かつシステムに顔を向けていると判定できる．そのため，その位置のユーザはその時点においてアクティブさが高いと推測できる．このとき，それ以外の位置から音源定位結果が得られた場合，それを今まで得られなかった未知の位置として棄却する，もしくは新たなユーザとして顔検出により，ユーザの存在の確認を行うことができる．

## 4. 評価実験

システムとユーザの実際の会話を収録したデータを用い，ユーザのアクティブさが推定できることを確認する．データには，ロボットのマイクロフォンを通した録音ファイルとシステムのログを利用する．

### 4.1 利用するシステム

本研究で利用するシステムは，我々の研究室で開発している，2 体のヒューマノイドロボット NAO[*1]による研究室紹介システムである [6]．図 4 にシステムとユーザのインタラクションの様子を示す．ユーザはロボットに，我々の研究室に関する質問ができる (例えば，「研究室の生活について教えて」)．2 体のロボットにはそれぞれ役割が設定されている．役割は主にユーザの質問に答える説明役と，ユーザとともに質問を行う質問役である．

システムは，複数ユーザの中から発話者を特定し，ロボットはそのユーザに顔を向けて応答を行う．一定時間ユーザの沈黙を検出したときは，質問役のロボットが，説明役のロボットに対して質問を行い，ロボット同士で会話を行う．ユーザへの応答中やロボット同士での会話中以外は，ロボットは正面を向いており，この状態の時のみシステムはユーザの発話を受理する．

[*1] http://www.aldebaran-robotics.com/en/

### 4.2 システムの設定

音源定位には，ロボット聴覚システム HARK [5] を用いた．HARK は MUltiple SIgnal Classification (MUSIC) 法 [7] に基づき，1 フレーム (0.01 秒) ごとに音源定位結果とそのパワーを出力する．MUSIC 法は，音源と入力に用いるマイクロフォン間のインパルス応答 (伝達関数) に基づき，音源を定位する．音の入力には，ヒューマノイドロボット NAO の頭部の前後左右に搭載された 4 つマイクロフォンを用いた．伝達関数を計算するためのインパルス応答は，ロボットより 1m の距離から，10 度間隔で 36 点計測した．したがって，音源定位結果の角度分解能は 10 度である．

画像情報の入力には，NAO の頭部に搭載されたカメラを利用した．カメラの視野角度は，水平方向が 47.8 度，垂直方向が 36.8 度である．顔検出には，NAO に付属の顔検出 API を利用する．API からは，ユーザの顔が検出できた場合に，その顔の位置が，カメラの視野角の中心からの水平角度として出力される．顔検出結果の分散 $\sigma_v$ は，6 人のユーザに対する 240 回の検出実験より，$\sigma_v = 1.00$ となった．さらに，API を用いることで，NAO のその時点での首の水平角度も取得する．

### 4.3 会話データの収録

データは，研究室紹介システムと参加者が実際に会話を行ってもらい収録した．参加者は，本研究室の学生 3 名である．図 2 に示すように，机 (150cm × 75cm) を準備し，その上に 2 体のロボットを配置した．2 体のロボット間の距離は 120cm で，参加者の座る位置との距離は 47.5cm とした．参加者は Robot A のマイクロフォンの中心から見て −30 度，−47 度，30 度の位置に座ってもらった．座席の中心から ±15cm をその参加者の領域とし，Robot A のマイクロフォンから見た領域は，それぞれ −35 度から −25 度 (a)，−65 度から −40 度 (b)，25 度から 35 度 (c) である．参加者には，「これは研究室紹介を行うシステムです．自由に会話をして下さい．」と教示を行い，特別な指示は与えていない．会話の時間は 200 秒とし，システム自身が会話を開始，終了させた．

収録したデータにおけるユーザの領域 (a, b, c) ごとのユーザの発話回数と発話フレーム数を表 2 に示す．発話回数は，システムに向けた発話とそれ以外に分けて集計した．それ以外の部分には，別のユーザへの発話や笑い，独り言が含まれる．集計は，録音データを用いて人手で行い，発話後に 400ms 以上の無音区間が存在した場合，1 発話と認定した．

**表 3** ユーザの領域ごとの音源定位の性能

| 領域 | 全発話フレーム数 | 定位フレーム数 | 正解定位フレーム数 | Precision | Recall | F 値 |
|---|---|---|---|---|---|---|
| a | 2117 | 2080 | 1047 | 0.503 | 0.495 | 0.499 |
| b | 428 | 424 | 398 | 0.94 | 0.93 | 0.93 |
| c | 2568 | 2502 | 1545 | 0.618 | 0.602 | 0.609 |
| それ以外 | 0 | 4561 | 0 | 0 | - | - |
| 合計 | 5113 | 9567 | 2990 | 0.313 | 0.585 | 0.407 |

**表 4** ユーザの領域ごとの顔検出の性能

| 領域 | 全正解フレーム数 | 検出フレーム数 | 正解検出フレーム数 | Precision | Recall | F 値 |
|---|---|---|---|---|---|---|
| a | 1035 | 296 | 296 | 1.00 | 0.29 | 0.45 |
| b | 2156 | 518 | 518 | 1.00 | 0.24 | 0.39 |
| c | 4303 | 340 | 293 | 0.86 | 0.068 | 0.13 |
| それ以外 | 0 | 1022 | 0 | 0 | - | - |
| 合計 | 7494 | 2176 | 1107 | 0.509 | 0.148 | 0.229 |

表 2 から，どのユーザも 3 回以上発話を行ったことがわかる．a と c のユーザは，システムに向けた発話と同等の回数で，それ以外の発話も行った．それ以外の発話のほとんどは，ロボットの発話内容に対する笑いである．a と c のユーザと比べて，b のユーザは発話が少ない．b のユーザは笑うなどの反応も少なく，システムに向けた発話のみを行った．

### 4.4 音源定位と顔検出の性能

表 3 にユーザの領域ごとの音源定位の性能を示す．表は，左から全発話フレーム数，定位フレーム数，正解定位フレーム数，Precision，Recall，F 値である．全発話フレーム数は，ユーザごとの発話フレーム数を人手で集計した値である．その領域に対する定位フレーム数は，ユーザの発話中にシステムが定位結果を出力したフレーム数である．正解定位フレーム数は，ユーザの発話中にシステムがその領域内を定位したフレーム数を示す．Precision は，定位フレーム数と正解定位フレーム数の比より，Recall は，全発話フレーム数と正解定位フレーム数の比より算出した．

表 3 から，どの領域も F 値 0.5 の音源定位性能が得られていることが確認できる．特に b のユーザの F 値は，他のユーザと比較して高い．これは b のユーザが，説明役の Robot A に近いためである．a のユーザの F 値が他のユーザより低いのは，a のユーザの声が小さかったためである．

表 4 にユーザの領域ごとの顔検出の性能を示す．全正解フレーム数は，ロボットのカメラの視野内にユーザが存在していたフレーム数を，人手で集計した値である．その領域に対する検出フレーム数は，カメラの視野内にユーザが存在していたと考えられるフレームにおいて，システムが検出結果を出力したフレーム数である．正解検出フレーム数は，カメラの視野内にユーザが存在していたと考えられるフレームにおいて，ユーザの存在する領域を検出したフレーム数である．Precision は，検出フレーム数と正解検出フレーム数の比より，Recall は，全正解フレーム数と正解検出フレーム数の比より算出した．

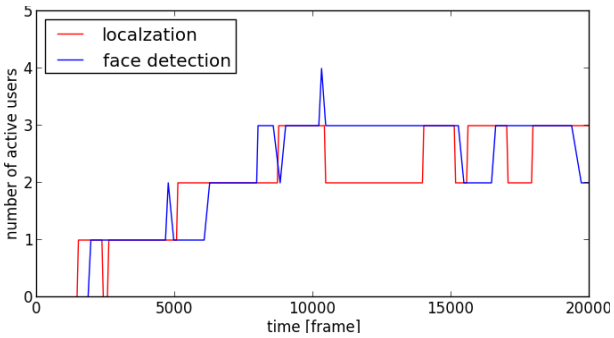


**図 5** アクティブユーザ数の時間変化

表 4 から，顔検出の Precision が高く，Recall が低いことが確認できる．これは，利用した顔検出 API の，検出結果を出力するか否かを決める閾値が厳しく設定されており，顔である可能性が十分に高い場合しか検出結果を出力しないためである．特に，c のユーザの Recall が低い．これは，ロボットが c の領域に顔を向けた際に，c のユーザが笑うなどの反応を行い，顔をロボットから背けることが多かったためである．

### 4.5 アクティブユーザの推定結果

収録データより得られたユーザのアクティブさと会話中の実際のユーザの行動を照らし合わせることで，本手法によりユーザのアクティブさの時間変化を表現できていることを確認する．音源定位結果と，顔検出結果によるアクティブさに基づく，アクティブユーザ数の時間変化を図 5 に示す．図の横軸は時刻で，縦軸はアクティブユーザの人数を示す．音源定位結果によるアクティブユーザ数の変化を赤線，顔検出結果によるアクティブユーザ数の変化を青線で示す．パラメータはそれぞれ，$\alpha = 300$, $\lambda_1 = 0.0001$,

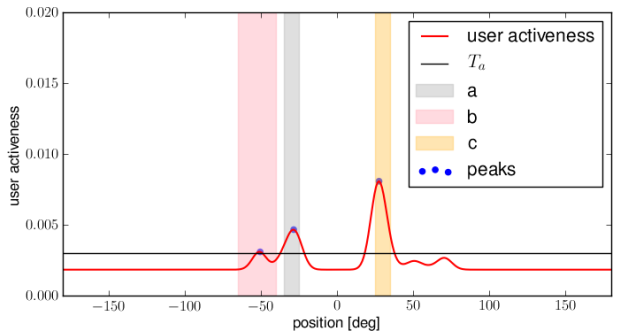


**図 6** 音源定位結果によるユーザのアクティブさ (10000 フレーム目)

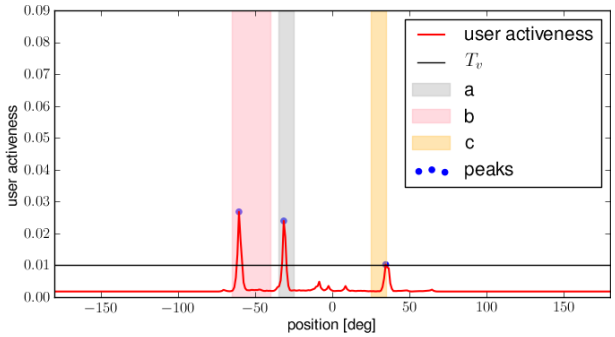


**図 7** 顔検出結果によるユーザのアクティブさ (10000 フレーム目))

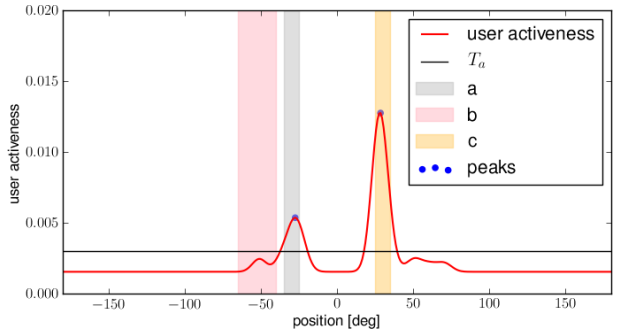


**図 8** 音源定位結果によるユーザのアクティブさ (13000 フレーム目)

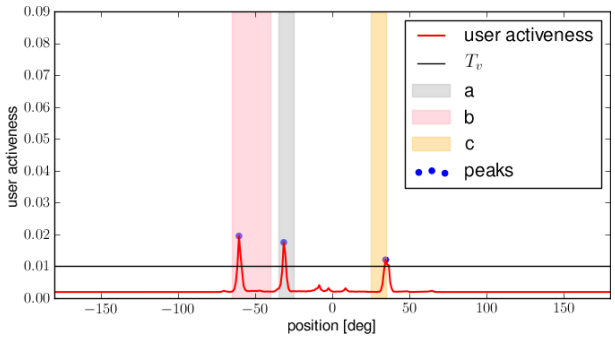


**図 9** 顔検出結果によるユーザのアクティブさ (13000 フレーム目))

$\lambda_2 = 0.0003$, $T_a = 0.003$, $T_v = 0.01$ とした.

図 5 より，会話開始から 10000 フレームまでは，音源定位結果によるアクティブユーザ数と，顔検出結果によるアクティブユーザ数は，ともに段階的に増加し，最終的には 3 人となっている．これは 3 人のユーザが順番に発話を行い，それに対して Robot A がユーザの存在する領域に顔を向け，カメラの視野内にユーザを捉えたためである．次に，10000 フレームから 15000 フレームまでは，音源定位結果によるアクティブユーザ数は 2 人と推定されている．これは，b のユーザが発話を控え，a と c のユーザのみが発話を行うもしくは Robot A の応答に対して笑うといった行動を続けたためである．一方で，顔検出結果によるアクティブユーザ数は，3 人のまま変化しない．これは，顔検出結果によるアクティブさの更新の重み $\lambda_2$ は $\lambda_1$ より大きいが，顔検出結果の分散が音源定位結果と比較して小さく，アクティブさが減衰しにくいためである．その後，15000 フレーム周辺で再び，音源定位結果によるアクティブユーザ数は 3 人となる．これは b のユーザが発話を再開したためである．

次に，各時点ごとのユーザのアクティブさより，アクティブユーザの位置を推定できること確認する．さらに，その時点での推定結果に基づき，ロボットが生成すべき発話を示す．

会話開始から 10000 フレーム目 (インタラクションが半分終了した時点) に得られた，音源定位結果によるユーザのアクティブさと，顔検出結果によるユーザのアクティブさをそれぞれ図 6，7 に示す．それぞれのグラフの横軸は Robot A からみた位置を示し，縦軸はユーザのアクティブさである．色づきの帯はそれぞれユーザが存在した領域を示し，グレーが a (−35 度から −25 度)，ピンクが b (−65 度から −40 度)，イエローが c (25 度から 35 度) である．青色のドットは検出されたピークの位置を示す．

音源定位結果によるアクティブさを示す図 6 から，3 名のユーザの存在する領域にアクティブさのピークが確認できる．これは，10000 フレーム目までに，3 名がそれぞれ 1 回以上，Robot A に対して発話を行ったためである．さらに，c に最も大きなピークが見られる．これは，この時点までに，c のユーザが最も多く発話を行ったためである．顔検出結果によるアクティブさを示す図 7 も，3 名のユーザが存在する領域にアクティブさのピークが確認できる．これは Robot A がユーザの発話に応答するため，それぞれのユーザの領域に 1 回以上顔を向けていたためである．また，b に最も大きなピークが見られる．これは，直前に Robot A が b のユーザに顔を向けたためである．

このとき，3 人のユーザが積極的に会話に参加していると判定できる．なぜなら，音源定位結果のアクティブさと顔検出結果によるアクティブさが対応する ($D = 10$) 位置に，アクティブさのピークが存在しているためである．これより，システムはアクティブさのピークの数を用いることで，未知の音源定位結果に対して賢く応答できる．

次に，13000 フレーム目に得られた，音源定位結果によるユーザのアクティブさと，顔検出結果によるユーザのア

クティブさをそれぞれ図 8，9 に示す．13000 フレームは，図 5 において，音源定位結果によるアクティブユーザ数が 2 人となり一定時間が経過した時点である．

音源定位結果によるアクティブさを示す図 8 から，2 名のユーザの存在する領域にアクティブさのピークが確認できる．c に最も大きなピークが見られるのは，直前に Robot A の応答に対して c のユーザが笑ったためである．顔検出結果によるアクティブさを示す図 7 から，3 名のユーザが存在する位置にアクティブさのピークが確認できる．

このとき，b のユーザは発話の機会を伺っていると判定できる．なぜなら，顔検出結果によるアクティブさから b にピークを検出できるが，それと対応する $(D = 10)$ 位置に，音源定位結果によるアクティブさによるピークが検出れていないためである．これより，ロボットは b のユーザに対して「何か聞きたいことはないですか?」と発話を促せる．

## 5. まとめと今後の課題

本論文では，2 体のロボットのマイクロフォンとカメラから得られる音源定位結果と顔検出結果に基づきアクティブユーザの推定を行う手法について述べた．評価実験より，システムとユーザの実際のインタラクションデータを用いることで，アクティブユーザを推定できることを示した．今後の課題は，オンラインで推定を行い，推定結果に応じた発話を生成することである．

## 謝辞

Nao と HARK を接続するプログラムは，京都大学の水本武志氏と協力して作成した．本研究の一部は，JST 戦略的創造研究推進事業さきがけの支援を受けた．

### 参考文献

[1] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. Integrating vision and speech for conversations with multiple persons. In *Proceedings of IEEE/RSJ the International Conference on Intelligent Robots and Systems (IROS)*, pages 2523–2528, 2005.

[2] Dan Bohus and Eric Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference*, pages 225–234, 2009.

[3] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the EACL*, 2006.

[4] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer. Providing the basis for human-robot-interaction: a multi-modal attention system for a mobile robot. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 28–35, 2003.

[5] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system 'HARK' - open source software for listening to three simultaneous speakers. *Advanced Robotics*, 5:739–761, 2010.

[6] Taichi Nakashima, Kazunori Komatani, and Satoshi Sato. Integration of multiple sound source localization results for speaker identification in multi-party dialogue system. In *Proceedings of International Workshop on Spoken Dialogue Systems*, 2012.

[7] Ralph O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34:276 – 280, 1986.

[8] David Traum and Jeff Rickel. Embodied agents for multiparty dialogue in immersive virtual worlds. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 766 – 773, 2002.