視聴覚情報を用いた画角外音源に対応可能なカメラシステム*

荒木 潤一　長西 将弘*1　苣木 禎史*1　宇佐川 毅*1
(熊本大学 工学部)　*1(熊本大学 大学院自然科学研究科)

1　はじめに

人は視覚や聴覚等の五感を用いて身の回りの環境理解を行っている。近年，視覚および聴覚情報処理を模擬しヒューマノイドロボットへ実装する研究が盛んに行われている。Kim らはステレオカメラ 1 つおよびマイクロホン 3 つからなるアレーを用いて対象物の定位をおこなうロボット IROBAA を提案している [1]。一方で，監視カメラにおける視聴覚情報処理を模擬したシステムの構築に関する研究はあまりなされていない。これまでに，少ない素子数での方向角，仰角を推定およびその方向の音源分離が可能な周波数領域両耳聴モデル (Frequency Domain Binaural Model : FDBM) を応用した画角外音源に対応可能な監視カメラ向けの音源方向推定手法が提案されている [2]。

本論文では，視聴覚情報処理を模擬した空間情報抽出システム構築の初期検討として，聴覚情報処理を模擬した音響信号処理による監視カメラの画角外音源に対する方向推定性能の評価を行う。

2　音源方向推定アルゴリズム

2.1　2 素子アレイによる推定

まず，任意の 2 素子間で得られる入力信号の周波数領域における位相差及びレベル差情報を用いた，音源の方向推定アルゴリズムを図 1 に示す。ここで，ϕ'，ψ' はそれぞれ音響信号処理によって推定された音源方向の方向角および仰角を示している。本来，視聴覚情報の融合を模擬したシステムにおいては，画像情報における対象物の方向角 ϕ_{img}，仰角 ψ_{img} および音響信号処理における音源方向の方向角 ϕ_{snd}，仰角 ψ_{snd} を区別をする必要があるが，本報告においては音響情報のみを扱うため ϕ_{snd}，ψ_{snd} 簡略化して ϕ，ψ を用いる。

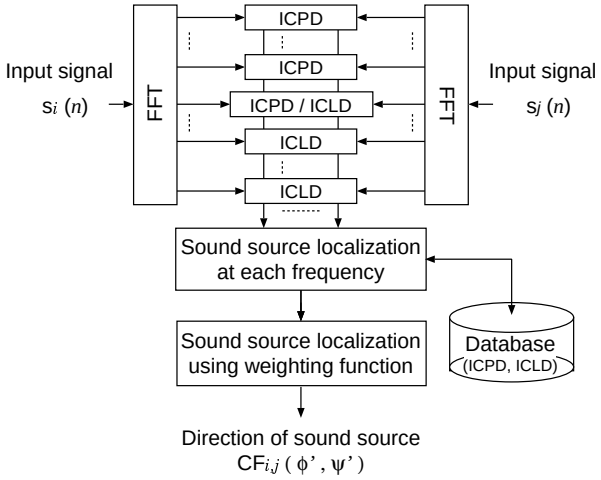


図 1: 2 素子アレイによる音源方向推定手法のブロック図

はじめに，2 入力の観測信号に対して FFT による帯域分割を行う。各チャネルの観測信号 $s_i(n)$, $s_j(n)$ をフーリエ変換することで得られるスペクトル $S_i(k)$, $S_j(k)$ を用い，チャネル間におけるクロススペクトル $C_{i,j}(k)$ を求める。ここで，i および j はチャネル番号 $(i \neq j)$，k は周波数帯域のインデックスを表す。

$$C_{i,j}(k) = S_i(k)S_j(k)^* \tag{1}$$

但し，$*$ は複素共役を示す。各周波数毎のチャネル間位相差 (Inter-Channel Phase Difference：ICPD) $\theta_{i,j}(k)$ は，クロススペクトル $C_{i,j}(k)$ を用い，

$$\theta_{i,j}(k) = \tan^{-1}\left\{\frac{\mathrm{Im}[C_{i,j}(k)]}{\mathrm{Re}[C_{i,j}(k)]}\right\} \tag{2}$$

---

*A camera system based on audiovisual information for out-of-view angle objects.
By Junichi Araki, Masahiro Naganishi, Yoshifumi Chisaki and Tsuyoshi Usagawa (Kumamoto University)

より求められる。また，チャネル間レベル差 (Inter-Channel Level Difference：ICLD) $\xi_{i,j}(k)$ は，パワースペクトルを $C_{i,i}(k)$ とすると，

$$\xi_{i,j}(k) = 20\log\left|\frac{C_{i,j}(k)}{C_{i,i}(k)}\right| \tag{3}$$

で与えられる。

今，方向角 ϕ，仰角 ψ，周波数帯域 k におけるデータベースとして，両チャネル間位相差情報 $\theta_{map\ i,j}(k,\phi,\psi)$ および両チャネル間レベル差情報 $\xi_{map\ i,j}(k,\phi,\psi)$ が与えられているとする。このとき，観測信号より周波数インデックス毎に求められた ICPD $\theta_{i,j}(k)$ および ICLD $\xi_{i,j}(k)$ をデータベースと比較し，式 (4) および式 (5) を満たす組み合わせ $(\phi,\ \psi)$ を，周波数帯域 k における音源方向の候補として得る。ここで，$D_{\theta,i,j}(k,\phi,\psi)$ および $D_{\xi,i,j}(k,\phi,\psi)$ は ICDP および ICLD より得られる周波数帯域 k における方向推定情報，$\alpha_{\theta,i,j}$ および $\alpha_{\xi,i,j}$ はデータベースとの差分の閾値を表す。

$$D_{\theta,i,j}(k,\phi,\psi) = \begin{cases} 1 & if\ \ \alpha_{\theta,i,j}(k) > |\theta_{map\ i,j}(k,\phi,\psi) - \theta_{i,j}(k)| \\ 0 & else \end{cases} \tag{4}$$

$$D_{\xi,i,j}(k,\phi,\psi) = \begin{cases} 1 & if\ \ \alpha_{\xi,i,j}(k) > |\xi_{map\ i,j}(k,\phi,\psi) - \xi_{i,j}(k)| \\ 0 & else \end{cases} \tag{5}$$

算出された各方向推定情報 $D_{\theta,i,j}(k,\phi,\psi)$ および $D_{\xi,i,j}(k,\phi,\psi)$ は，次式により統合される。

$$\begin{aligned} D_{i,j}(k,\phi,\psi) &= \beta(k)\cdot D_{\theta,i,j}(k,\phi,\psi) \\ &+ (1-\beta(k))\cdot D_{\xi,i,j}(k,\phi,\psi) \end{aligned} \tag{6}$$

ここで $\beta(k)$ は周波数荷重係数であり，位相回転を考慮し，低域においては ICPD，高域においては ICLD が強調されるように定義される。

次に，式 (6) より得られる周波数毎の方向推定情報 $D_{i,j}(k,\phi,\psi)$ にその周波数帯域 k におけるパワーを考慮した重み関数 $E_{i,j}(k)$ を乗じ，全ての周波数における総和を求める。なお，$E_{i,j}(k)$ は下式により算出する。

$$E_{i,j}(k) = \frac{|S_i(k)| + |S_j(k)|}{2} \tag{7}$$

これにより，チャネル間の全周波数帯域における音源の方向推定情報 $CF_{i,j}(\phi,\psi)$ を得る。

$$CF_{i,j}(\phi,\psi) = \sum_k E_{i,j}(k) D_{i,j}(k,\phi,\psi) \tag{8}$$

そして，得られた $CF_{i,j}(\phi,\psi)$ の値が極大値となる (ϕ,ψ) を求め，その対応する方向角・仰角の組み合わせを任意の 2 素子アレイにより得られる音源方向とする。

2.2 N 素子アレイによる実環境下での推定

本報告においては入力信号においてパワのある区間検出のために，背景雑音のパワに基づく閾値が設定できる環境を想定しており，その閾値を越えた場合のみ方向推定処理を行う。前述した 2 素子アレイによる手法を一般化した，N 素子アレイによる音源方向推定について示す。i 番目および j 番目のマイクロホン間により得られる推定情報を $CF_{i,j}(\phi,\psi)$ とすると，N 素子アレイにより推定される音源方向は式 (9) により決定される。

$$(\phi',\psi') = \{(\phi,\psi) \mid max\{CF(\phi,\psi)\}\} \tag{9}$$

ここで，$CF(\phi,\psi)$ は，

$$CF(\phi,\psi) = \sum_{i,j}^{p} \zeta_{i,j}\cdot CF_{i,j}(\phi,\psi) \tag{10}$$

であり，(ϕ',ψ') は推定された音源の方向角および仰角，p は 2 素子アレイの組み合わせ数 ${}_NC_2$，$\zeta_{i,j}$ は各チャネル間の推定情報 $CF_{i,j}(\phi,\psi)$ を統合するための重み係数である。なお，本論文では $\zeta_{i,j}$ は 全て 1 とする。

3 音源方向推定実験

3.1 実験条件

本論文で用いる旋回型監視カメラ TOA (株) C-CC 511 (天井埋め込み金具 TOA (株)

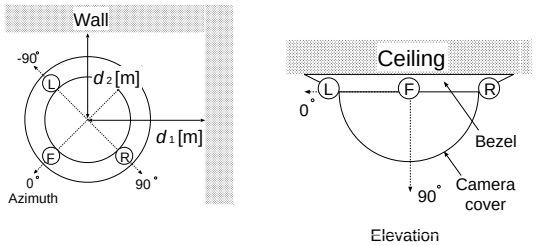


図 2: 3 素子マクロホンを取り付けた監視カメラと壁面の位置関係

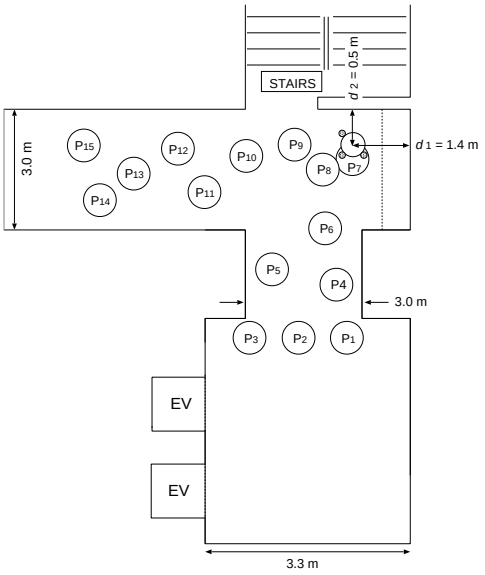
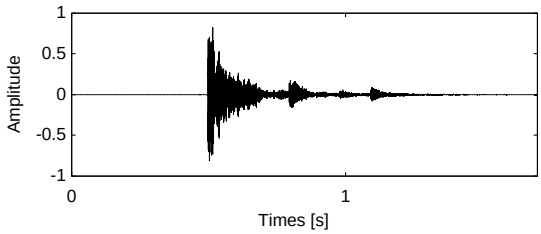
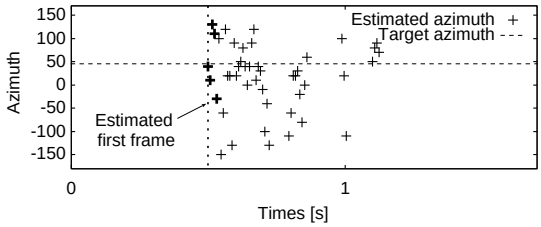


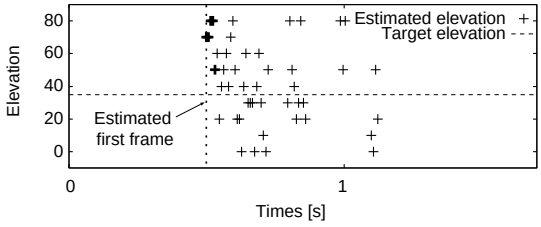
図 3: 音源位置とカメラ位置の関係図



(a) 位置 P_1 における マイクロホン L での 1 回目の試行における受音信号



(b) 推定結果（方向角）



(c) 推定結果（仰角）

図 4: 受音波形および方向推定結果の一例

C-BC 511 U 使用時) におけるマイクロホン配置と壁面との位置関係を図 2 に示す。本研究で用いるマイクロホンは 3 素子である。画角外での音響イベントを想定し，図 3 に示すような監視カメラと対象物との位置関係で空缶を落とし，その音を検出しその方向が画角内に入るように監視カメラを制御することを想定している。図 3 の $P_1 \sim P_{15}$ の位置で空缶を落下する。試行回数は 10 回である。壁面からの監視カメラの音響中心までの距離は，おおよそ $d_1 = 1.4$ m，$d_2 = 0.5$ m であり，床面からの高さは 2.5 m である。また，暗騒音レベルは監視カメラの位置で 62 dB であり，空缶を落下させた際の観測音の大きさは約 85 dB である。音源方向推定で用いる ORTF (Object Related Transfer Function) データベースは無響室において測定した TSP 信号を同期加算して導出した。このデータベースは方向角で $-180° \sim 170°$(10° 間隔)，仰角で $0° \sim 80°$ (10° 間隔) の情報を持つ。音響信号処理における FFT フレーム長は 512 , 周波数分解能は 31.25 Hz である。

3.2 音源方向の推定性能評価

各位置 $P_1 \sim P_{15}$ における音源方向の推定性能評価を行う。評価はフレーム毎での方向推定結果と各位置 $P_1 \sim P_{15}$ のカメラからの設置角との誤差をもとに行う。また，推定アルゴリズムにおけるフレームシフトは 1/4 とし，各試行回数での受音信号における時系列でのトリガー処理は行っていない。なお，評価を行ったフレーム数は全ての試行回数において信号の立ち上がりを検出した初期フレームを含めて 5 フレームとした。

図 4 (a) に，位置 P_1 での 1 回目の試行におけるマイクロホン L での受音信号を示す。0 s から 0.5 s までの間が，暗騒音の区間であり，対象音源の信号は，暗騒音に対して十分な振幅を有していることがわかる。また，図 4 (b)(c) は音源方向の推定結果であり，閾値を越えた 45 フレーム分の結果を示

表 1: 各位置における推定結果と設置角との誤差の平均値 [deg.]

Position	P_1	P_2	P_3	P_4	P_5
Azimath	6.0	20.0	8.8	11.1	10.2
Elevation	10.0	12.1	8.5	11.7	9.66
Position	P_6	P_7	P_8	P_9	P_{10}
Azimath	19.2	63.0	18.0	25.9	10.2
Elevation	7.8	26.3	5.1	5.3	5.0
Position	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}
Azimath	8.7	11.3	7.0	4.2	7.0
Elevation	11.5	5.6	10.0	5.2	9.0

表 2: 各フレームにおける誤差の最も小さい方向を推定した確率 [%]

Frame number	1	2	3	4	5
% of best estimation *	61.3	22.0	10.7	3.3	9.3

* Due to multiple frames shows the best estimates, the total of % exceeds 100.

している。

表 1 は各位置 $P_1 \sim P_{15}$ 毎でそれぞれ 10 回試行した際の推定結果 5 フレームのうち，最も推定誤差の小さいフレームを選び方向角，仰角において試行回数 10 における平均値を求めたものである。各試行においての方向角および仰角の推定誤差をみると，必ずしも同一フレーム内で両者が最小となるわけではないが，今回は各設置位置と推定した方向の同一平面状における位置との距離が最も小さいフレームを対象として評価を行う。

本研究で使用している監視カメラの画角は方向角では 47.3°，仰角では 36.5° である。よって，位置 P_7, P_9 以外の全ての位置において画角内に収まることが確認できる。P_7 については設置位置の仰角が 87.3° でほぼカメラの真下に位置している。しかしながら，音源方向推定に用いるデータベースは仰角において 0° $\sim$ 80° (10° 間隔) の範囲のため推定を行う事自体が困難であると考えられる。一方，P_9 においては仰角に比べて方向角での誤差が顕著にみられるが，試行回数毎に検討すると方向角において全て設置位置よりも壁面方向の推定を行っており，反射音の影響が原因ではないかと考えられる。

次に，表 2 に全ての位置での全試行回数 150 回において推定フレーム毎で各試行における推定誤差が最も小くなった確率を示す。本報告では入力信号における時系列でのトリガー処理は行ってはいないが，初期フレームあるいは第 2 フレームにおいて推定誤差が小さい傾向が示唆されている。本実験における観測音はインパルス信号に近いものであり，入力の立ち上がり部分において強いパワを持っている可能性が高い。よって，誤差の最も小さい方向を推定した確率が高くなっていると考えられる。

4 まとめ

本論文では，視聴覚情報処理を模擬した対象物定位手法の初期検討としてカメラの画角外音源にインパルス信号に近い信号を用いて，受音信号の初期フレームを含む 5 フレームに対する推定性能の評価を行った。結果より本実験で用いた信号に関しては比較的高い確率で音源方向を画角内に収めることが可能であった。しかし，より現実的な環境での使用を考えた場合，音声等さまざまな受音信号の立ち上がり検出およびその認識を行う事が求められる。よって，今後の課題として反射音の存在する様々な環境においても対応可能な受音信号の立ち上がり検出方法の検討が挙げられる。

謝辞

本研究に際して，TOA 株式会社 栗栖清浩氏，前田和昭氏との有益なディスカッションに感謝する。本研究の一部は，科学研究費補助金基盤 (C) No. 18500135，および東北大学電気通信研究所共同プロジェクト (H19) で行った。

参考文献

[1] Hyun-Don Kim, Jong-Suk Choi and Munsang Kim, “Human-Robot Interaction in Real Environments by Audio-Visual Integration” *International Journal of Control, Automation, and Systems*, Vol. 5, No. 1, pp.61-69, Feb., 2007

[2] 長西 将弘，高田 俊亨，苣木 禎史，宇佐川 毅，“回折を用いた音源方向推定機能を有するカメラシステムにおける反射音の影響に関する検討” **日本音響学会** *2007* **年春季研究発表会講演論文集**, pp.599-600, Mar., 2007