音源方向と顔画像による話者検出*

井　上　大　輔**
高　橋　伸　弥***
森　元　　逞***

Speaker Detection by Sound Source Direction and Face Image

Daisuke INOUE, Shin-ya TAKAHASHI and Tsuyoshi MORIMOTO

In this paper we propose a speaker dctection system combining estimation of sound source direction and face image tracking. This system first estimates a speaker direction by calculating cross-correlation of signals from two microphones, then rotates the camera to the direction, and finally detects a face region from camera image using skin color detection. We describe an algorithm of the speaker detection in the proposed system and introduce a hardware implementation of the prototype system. To show the effectiveness of this approach, we conduct some experiments for speaker detection. The experimental results show that the proposed system can improve speaker detection accuracy using speech signals and face images compared with that using speech signals only.

Key Words: Speaker Detection, Face Image Tracking, Human Interface

1. はじめに

音声対話システムにおいて，マイクを口元に近付けてからマイクに向かって話しかけるよりも，離れた位置からでも両手を自由に他の用途に使いながら対話することができれば便利である．しかし，離れた位置からの音声ではマイクロホン増幅器の感度を上げる必要があり，その結果，周囲の雑音や騒音も大きな音で拾ってしまう．また，反響音などの影響も加わって音声対話システムがうまく認識してくれないという問題が生じる．

ハンズフリーを目指すためには，口元から離れた位置で，高品質な音声認識を行う技術が必要である．上で述べた周囲の雑音も大きな音で拾ってしまうという問題を防ぐために，目的とする方向の音だけに感度が高く，周囲の不要な音に対しては感度が低いような指向性マイクロホンが利用されている．もし，目的とする音源の方向を音声対話システムが検出することができれば，この指向性マイクを音源方向に向けることにより，音声認識を精度良く行うことが可能になる．

実環境下で目的とする音源（話者）の方向を検出する方法については，これまで多くの研究が行われている[(1)～(5)]．例えば，マイクロフォンアレーを用いて音声特有の調波構造という情報を活用する方法[(1)(2)]や，アレー信号処理における相関性干渉信号を抑制するための空間平均化を利用する方法[(3)]などが提案されている．しかし，これらの方法は音声情報のみで音源方向を推定しており，周囲の雑音や反響音などの影響を受けやすいという問題点がある．また画像情報を併用する方法としては，マイクロフォンアレーと３眼カメラにより取得した視聴覚情報を組み合わせて，音源位置を推定する方法[(4)]や，音声信号の到来方向の推定と画像内の顔探索を統合する方法[(5)]が提案されているが，これらの方法は処理が複雑であり，カメラ，マイク（接続されているアンプ，AD

* 平成18年１月20日受付
** 電子情報工学専攻
*** 電子情報工学科

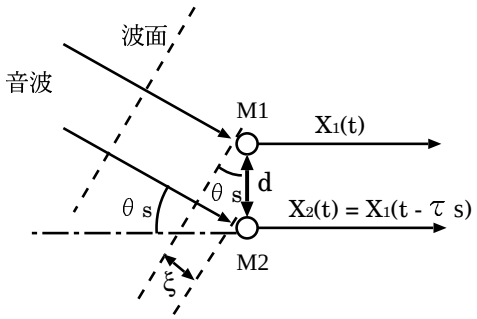


図1　２本のマイクで受音される音声信号

変換器を含む），駆動系をそれぞれネットワークを介して相互に接続された別々のパソコン（PC）に接続しなければならない．

そこで我々は，２本の小型のピンマイク，軽くて安価なUSBカメラ，および指向性マイクの３つをステッピングモータの上に配置するという比較的簡単なハードウェア構成を用いることとした．まず２本のピンマイクで音声信号の到来方向を大まかに推定し，モータを回転させてその方向にカメラを向ける．次にカメラから得られた画像内で顔探索を行う．このように音情報と画像情報を組み合わせることにより，話者を検出する．またモータの回転に伴い，指向性マイクも話者の方向を向くので，このマイクを用いれば話者音声を精度良く認識できる．

2.　話者検出の手法

2.1 マイクによる話者検出

まず，θ_s 方向から到来する音を距離 d だけ離れて設置された２本のマイクで受音することを考える．この様子を図１に示す[6]．

図１の θ_s 方向から到来した音波は，まずマイクM1において受音され，その後に距離 ξ だけ進んでマイクM2に到達する．図１によりこの距離 ξ は，

$$\xi = d \cdot \sin\theta_s \tag{1}$$

と表される．したがってM2での受音信号 $X_2(t)$ はM1での受音信号 $X_1(t)$ と比べて音波が距離 ξ だけ進行するのに要する時間 τ_s だけ遅れた信号となる．すなわち，c を音速として，

$$\tau_s = \frac{\xi}{c} = \frac{d \cdot \sin\theta_s}{c} \tag{2}$$

の関係が成り立つことから，音源方向は次式で計算できる．

$$\theta_s = \sin^{-1}\left(\frac{c \cdot \tau_s}{d}\right) \tag{3}$$

ここで，時間差 τ_s は，$X_1(t)$，$X_2(t)$ との相互相関関数

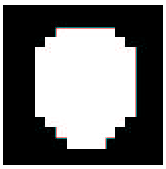


図2　顔画像検出に使用したテンプレート画像

$$\Phi_{12}(\tau) = \frac{1}{N}\Sigma X_1(t) \cdot X_2(t+\tau) \tag{4}$$

から求められる．ただし，$X_1(t)$，$X_2(t)$ はそれぞれ標本化されたものと考える．また t は離散時間を表し，Σ はN個の標本点総和を表す．

２つの音声信号にレベル差がないと仮定すると

$$X_2(t) = X_1(t-\tau_s) \tag{5}$$

$$\Phi_{12}(\tau) = \frac{1}{N}\Sigma X_1(t) \cdot X_1(t+\tau-\tau_s) \tag{6}$$

となることから，式(6)において相互相関関数 $\Phi_{12}(\tau)$ は，$\tau=\tau_s$ で最大値をとる．つまり $\Phi_{12}(\tau)$ の最大値を与える τ を式(3)に代入することによって音源方向を求めることができる．

2.2 カメラによる話者検出

マイクによる話者方向の推定には反響音等の影響により誤差が生じる場合がある．これに対し，カメラを用いて話者の顔画像を検出することで，誤差を修正することを考える．以下，顔画像を検出する方法について述べる．

2.2.1 テンプレートマッチング法を用いた顔画像検出方法

テンプレートマッチングとはあらかじめ用意したテンプレート画像を被探索画像上で移動させながら類似する領域を探す方法である．カメラの入力画像から肌色領域を検出し，２値化処理を施した後，あらかじめ用意した図２に示すテンプレート画像と比較して類似した領域を探索する．ただし，ここでは検出対象とする人物は正面を向いていると仮定する．

テンプレート画像を被探索画像上で１画素ずつ移動させ，最も類似した領域を探索していく．このとき，探索範囲は被探索画面の幅や高さから，テンプレート画像の幅や高さを差し引いた範囲である．ただし，被探索画像において探索したい対象が一部しか見えていない場合，検出することはできない．顔画像を検出する処理の流れを図３に示す．まず，カメラで取得した入力画像（図４(a)）をHSV表色系[注1]に変換する．これは，RGB表色系では光の影響を受けやすく，同じ色であっても光の反射によって違う色と認識してしまうが，HSV表色系では色相・彩度・明度で色を表現しているため光の影響を

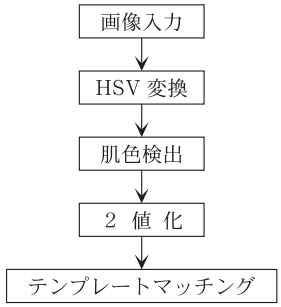
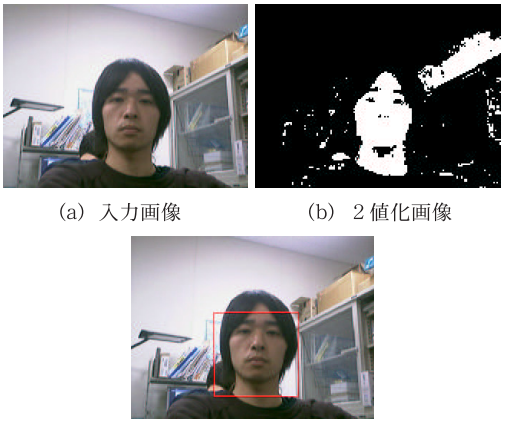


図3 顔画像を検出する処理の流れ



（a）入力画像　　（b） 2 値化画像

（c）結果画像

図4 顔画像の検出

受けにくいからである．

次に HSV に変換された画像内の肌色の部分を白色に，肌色以外の色を黒色に変換する（2 値化）．後述する実験環境では，肌色の範囲としてHの値を245～350とした．

図 4 (a) の入力画像を 2 値化した画像を図 4 (b) に示す．この 2 値化された画像と図 2 に示すテンプレート画像を 1 画素ごと比較していく．ここで，テンプレート画像の座標 (i,j) の値を $M_{i,j}$，被探索画像の座標 (x,y) の値を $I_{x,y}$ とすると，

$$\delta_{i,j}=\begin{cases}1 & (M_{i,j}=I_{x+i,y+j})\\ 0 & (otherwise)\end{cases} \tag{7}$$

$$s(x,y)=\sum_i\sum_j \delta_{i,j} \tag{8}$$

を求め，次式に示す評価値により，被探索画像上で最もテンプレート画像と一致する領域を求める．

$$\hat{s}=\max_{x,y}\ s(x,y) \tag{9}$$

入力画像に対する顔領域の探索結果を図 4 (c) に示す．なお，評価値にしきい値を設定し，評価値がそのしきい値より小さい値であれば，その領域には顔は存在しないとみなす．

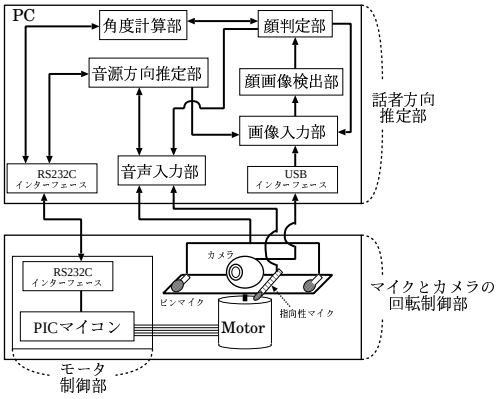


図5 全体構成

3. 話者検出システム

3.1 システム概要

本システムの全体構成を図 5 に示す．上部が話者方向推定部，下部が回転制御部であり，モータ制御部と PC は RS232C で通信を行う．

話者方向推定の流れを図 6 に示す．まず始めに 2 本のマイクで話者からの音声の入力を待つ．音声信号が入力されれば，図 7 に示すように最も振幅の大きな部分を一定のフレーム長で切り出す．そして相互相関関数より 2 本のマイクに到達する音声信号の時間差を求め，音源方向を推定する．次に，推定された音源方向にマイクとカメラを向ける．その後，カメラで取得した画像内で人の顔画像を探索し，顔がカメラの中央に来るようにモータを回転させる．

この方法を用いれば，音源方向を推定した際に反響音などの影響で多少誤差が生じてもカメラでその誤差を修正することで，精度良く話者を検出することができる．また，複雑な計算でないためリアルタイムな話者検出が可能である．

3.2 回転制御部の構成

回転制御部の構成と外観を図 8 に示す．ステッピングモータの上に音源方向検出用の 2 本のピンマイクが幅20 cm の間隔で配置され，その中央に顔画像検出用の USB カメラと音声認識用の指向性マイクが配置されている．回転制御部の諸元を表 1 に示す．

モータ制御部では図 9 に示すような処理を行う．まず PIC マイコンは電源が入るとデータを読み込み待ち状態になる．PC より角度が指示されると，角度に対応するステップ数を計算する．なお，今回使用したステッピン

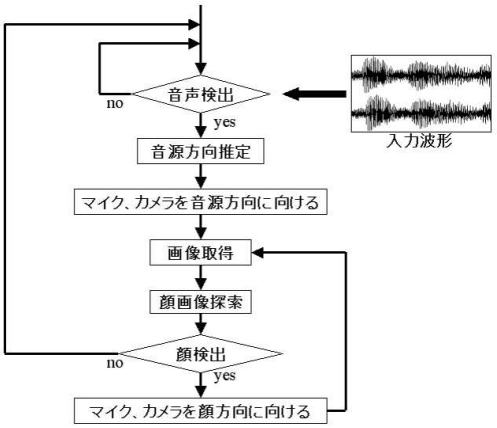


図6 音声入力から話者検出までの流れ

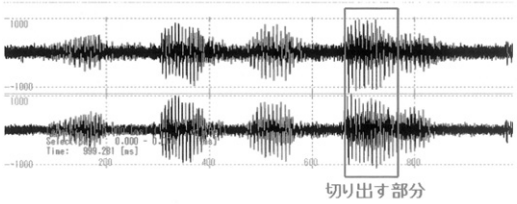


図7 ２本のマイクで録音された音声信号

グモータでは１ステップは1.8° である．次に，計算されたステップ数を１ステップずつモータに送信する．なお，トルク不足でモータが回らなくなるのを回避するために信号を送った後にある程度の待ち時間が必要になる．計算されたステップ数を送信し終えたら回転が完了したことを PC 側に送信し，再び入力待ち状態に戻る．

4. 実験

4.1 予備実験

4.1.1 無指向性マイクと指向性マイクを組み合わせることの有効性

まず，無指向性マイク（ピンマイク）と指向性マイクとにおいて，距離・角度を変えて音声認識率の比較を行った．実験条件を表２に示す．また，発話文の例を図10に示す．

音源を正面に置き，ピンマイクと指向性マイクで距離を変えて音声認識実験を行った結果を表３に示す．

距離が10cm の場合，どちらのマイクにおいても音声認識率は同じである．しかし，ピンマイクの場合，距離が離れるに従って音声認識率は極端に低下する．一方，指向性マイクでは，２ｍまでの距離であれば高い音声認識率が得られることが分かる．

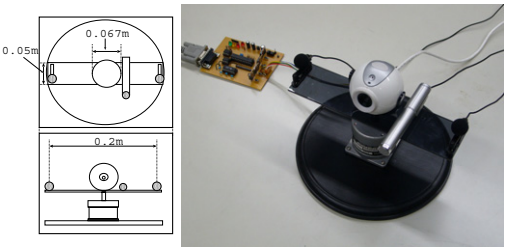


図8 回転制御部の構成と外観

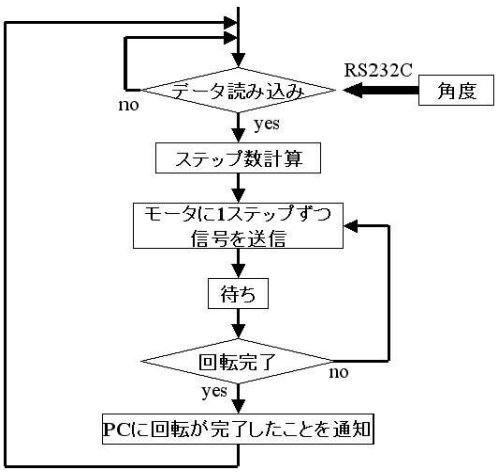


図9 PIC マイコンの処理

表1 回転制御部の諸元

PIC マイコン	PIC16F873-20/sp
シリアル通信	PS232C １系統
ステッピングモータ	多摩川精機製 TS3103N124
ピンマイク	SONY ECM-115
指向性マイク	SONY ECM-Z60
カメラ	Logicool QV-4000WH

表2 音声認識の実験条件

音声認識器	HVITE （HTK ）
音響モデル	HMM（４混合性別非依存トライフォン）
言語モデル	学習テキスト1000文を用いて作成したバイグラム言語モデル
音声データ	学習テキストから30発話（３名）（サンプリングレート：16kHz ）

次に，２ｍの距離で音源の方向を変えて音声認識実験を行った結果を図11に示す．

２ｍの距離ではピンマイクの音声認識率はかなり低い

表3 距離毎の音声認識率

マイクと音源の距離	音声認識率（ピンマイク）	音声認識率（指向性マイク）
10cm	92.4％	92.4％
1 m	80.0％	86.7％
2 m	64.4％	82.7％
3 m	59.6％	61.3％

カードで払えますか？
ここでタバコを吸ってもいいですか？
乗り換え切符をもらえますか？
タクシーを呼んでいただけますか？
日本語の新聞はありますか？
どうやって書類を記入したらいいですか？
ここで写真を撮ってもいいですか？
⋮

図10 会話文の例

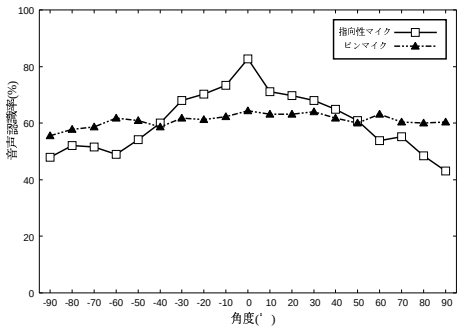


図11 音声認識率（距離２m）

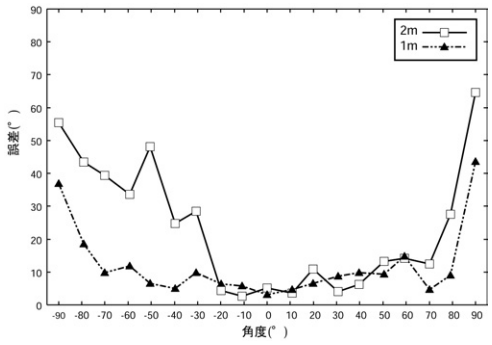


図12 話者検出結果（32kHz，音声のみ）

が，指向性マイクでは正面からの音声であれば高い音声認識率が得られている．このことから，無指向性マイクで音源方向を推定し，音声認識には指向性マイクを利用するという組み合わせが有効であることが分かる．なお，いずれのマイクを使用しても３m以上離れると音声認識率は低下してしまう．これは，反響音の影響によるものと思われ，別の対策が必要となる．

4.1.2 音声情報だけを用いる場合の検出誤差

サンプリングレートが32KHz，距離が１mと２mの場合において音声のみで話者検出を行った結果を図12に示す．ただし，実験条件は4.2節の話者検出実験と同じである．

図12から距離が離れると誤差が大きくなり，また角度が大きくなるに従って誤差が大きくなっていることが分かる．これは，音のレベル差や反響音の影響が増大するためであると思われる．

以上のように，音声情報だけで話者方向を推定することは難しい．

4.2 音声情報と顔画像情報を用いた話者検出実験

4.2.1 実験方法

話者検出実験を行った実験室(注2)の様子を図13に示す．この実験では，顔写真を音源の上に設置することで話者の代わりとした．複数の発話者が同時に発声するという状況も実際には起こりうるが，本研究では最も単純な場合として一人の発話者しか存在しない状況を想定している．

音声情報と画像情報の双方を用いて，本システムに対して音源の位置を約10°毎ずらしながら音源方向を検出する実験を行った．

ここで，２本のピンマイクの幅を20cm としているため，式(3)よりマイクの正面を0°とすると

$$-90^\circ < \sin^{-1}\left(\frac{340\times\tau_s}{0.2}\right) < 90^\circ \tag{10}$$

より

$$-1 < \frac{340\times\tau_s}{0.2} < 1 \tag{11}$$

という関係が成り立つ．よって，相互相関を計算する際は，およそ

$$-0.59_{ms} < \tau_s < 0.59_{ms} \tag{12}$$

の範囲内で最大値を与える τ_s を求めればよい．

サンプリングレートが低ければサンプル間の間隔が長くなるため，角度の分解能が低下する．そこで以降の実験では，サンプリングレートを48kHz で行うことにした．また，入力された音声信号において相関を計算するフレームのサイズは1000サンプル（約21ms）と2000サンプル（約42ms）の２ケースで行った．

4.2.2 実験結果と考察

まず，フレームサイズが1000サンプル（約21ms）の場合の実験結果を図14に示す．

図12の２mの結果と比べて，サンプリングレートの向上により角度の分解能は高くなっているはずであるが，音声のみでの話者検出精度に大きな変化が見られなかった．これはフレームサイズが1000サンプルでは短すぎることが原因ではないかと考えられる．そこで，フレームサイズを2000サンプル（約42ms）にして実験を行った．その結果を図15に示す．

図14と比較すると，音声のみで話者検出をした場合の誤差は全体として低下しているものの，角度が大きくなるに従ってかなりの誤差が生じている．この原因として音のレベル差と反響音の影響が考えられる．また，音声情報で音源方向を求めたときの誤差が大きい場合，画像情報を用いてもその誤差を軽減させることができない．これは，本研究で使用したUSB カメラの視野が約40°（±20°）であるため，音声情報で求めた角度に20° 以上の誤差が生じると，顔画像がカメラに映らないことが原因だと考えられる．

そこで，音声情報で求めた方向にシステムを向けた後，もしカメラで取得した画像内に顔画像が存在しなかった場合は左右（今回は±20°）にモータを再回転して顔画像を再度探すようにシステムを改良した．システム改良後の実験結果を図16に示す．

図16から分かるように，左右にモータを再回転する機能を追加した後では，音声情報と画像情報を組み合わせることにより，ほぼ正確に話者を検出できていることが分かる．よって，表３に示したように２m程度の距離であれば，高い精度で音声認識を行うことができることになる．

5. おわりに

まず音声信号の時間差から音源方向を大まかに推定し，次にその方向にカメラを向け顔画像を検出することで，話者を精度良く検出するシステムを作成した．このシステムは２本の小型ピンマイクとUSB カメラをステッピングモータの上に配置するという比較的簡単な構成となっている．

実験により，反響音や２本のマイクに到達する音のレ

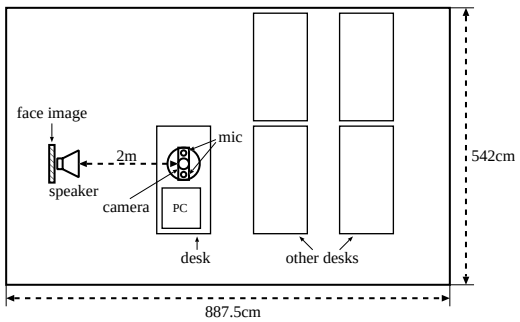


図13 実験室

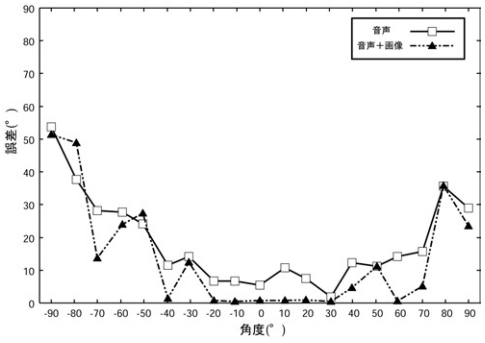


図14 48kHz，２mでの話者検出
（framesize：1000sample）

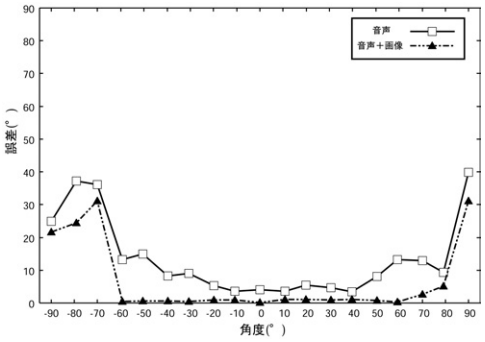


図15 48kHz，２mでの話者検出
（framesize：2000sample）

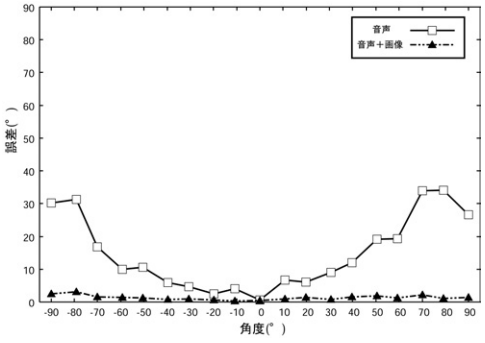


図16 48kHz，２mでの話者検出
（framesize：2000sample）

ベル差の影響で多少誤差が生じたとしても，話者の位置を高い精度で検出できることを示した．また，その方向に指向性マイクを向けることにより，２m程度の距離であればマイクを口元に置く場合に比べ，音声認識率を10ポイントほどの低下に抑えることができる．

今後は，話者の唇の動きを検出することで複数の話者が存在する場合に対応させることや，２m以上の距離にも対応させることなどを検討していきたい．

注

(1) H（色相），S（彩度），V（明度）で色を表現する表色系

(2) 特別な反響音対策を行っていない一般的な研究室

参 考 文 献

(1) 谷川，浜田：“２チャンネルマイクロホンアレーの仮想多チャンネル化による音声の到来方向推定法”，電子情報通信学会論文誌 A，Vol.J85-A, No.2, pp.153-161, 2002.

(2) 山田，中村，鹿野：“マイクロフォンアレーを用いた発話者方向検出によるハンズフリー音声認識”，情報処理学会論文誌，Vol.39, No.5, pp.1275-1284, 1998.

(3) 日岡，浜田：“反射音の存在する環境における音声の到来方向推定”，信学技報，EA2002-111, pp.35-40, 2003.

(4) 陳，日黒，金子：“ベイジアンネットワークに基づく視聴覚情報の統合を用いた画像からの３次元音源位置推定”，電気学会論文誌 C, Vol. 124, No.3, pp.720-728, 2004.

(5) H G. Okuno, *et. al.* :“Human-Robot Interaction Through Real-Time Auditory and Visual Multiple-Talker Tracking”, Proc. of 2001 IEEE/RSJ Int. Conference on Intelligent Robots and Systems, pp.1402-1409, 2001.

(6) 北脇：未来ねっと技術シリーズ「ディジタル音声・オーディオ技術」，オーム社，1999.