視聴覚情報の統合による話者位置検出システム

5B－1

松尾 直司　　北川 博紀　　長田 茂美　　棚橋 純一
株式会社 富士通研究所

1．はじめに

ユーザが自然に，かつ容易に利用できるインタフェースとして，音声処理や画像処理等の様々な手段を統合するマルチモーダルインタフェース[1]がある．マルチモーダルインタフェースでは，ユーザとパーソナルコンピュータ等の機械とのインタフェースの対話様式である複数のモダリティが同時または逐次的に用いられ，複数情報の入出力と解釈が行われる．このマルチモーダルインタフェースを用いることにより，音声や視線やジェスチャ等によって，ユーザはパーソナルコンピュータ等とのインタラクションが可能になり，より自然なヒューマンインタフェースを実現することができる．

著者らは，マルチモーダルインタフェースの一つとして，視聴覚情報を統合することにより，照明条件の変動や雑音などの外乱に対して頑健な話者位置検出技術の開発を行っている[2]．これは，カメラやマイクロホンアレイを基準とする座標上の各位置における顔の存在する確からしさと音源の存在する確からしさを統合することにより，話者位置を検出する技術である．例えば，この話者位置検出技術と目的信号強調技術[3]を組み合わせて，音声認識システムに用いることにより，音声を明瞭に取り込むことができ，雑音環境における認識率向上が可能になる．

本稿では，視聴覚情報の入力センサであるカメラとマイクロホンアレイの位置関係が未知の場合でも，予めカメラを基準としたマイクロホンアレイの位置と回転角度を検出し，その結果を基に視聴覚情報の統合を行うことを特徴とする話者位置検出システムを提案する．

A Speaker Position Detection System using Audio-Visual Information
N. Matsuo, H. Kitagawa, S. Nagata and J. Tanahashi
FUJITSU LABORATORIES LTD.

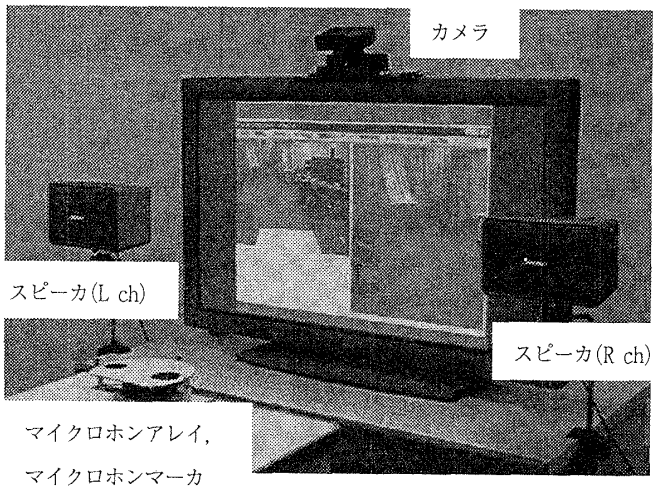

図1　試作システム

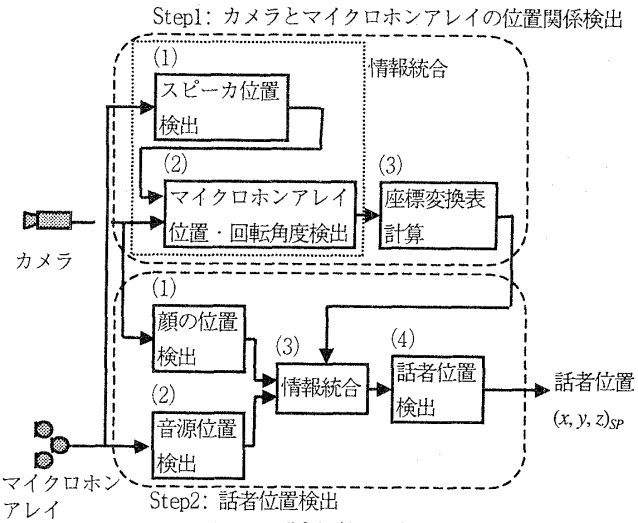

図2　話者位置検出処理

2．話者位置検出システム

2．1．概要

試作システムを図1に，全体の処理手順のブロック図を図2に示す．このシステムは画像入力用のカメラと音声入力用のマイクロホンアレイを備えている．また，2台のパソコンと，2 chスピーカ(L ch, R ch)を備えている．2台のパソコンはネットワークにより接続されており，1台目のパソコンで聴覚情報処理を行い，2台目のパソコンで視覚情報処理と情報統合処理を行う．

本システムは，視覚情報処理により，図3に示す情報統合用座標と共通の視覚情報処理用座標上で顔の存在する確からしさを求め，聴覚情報処理により，図4に示すマイクロホンアレイを基準とした聴覚情報処理用座標上で音源の存在する確からしさを求める．これらの結果を情報統合用座標上で統合し，話者位置$(x, y, z)_{SP}$を求める．ここで，情報統合用座標において，カメラとスピーカの位置は固定で，マイクロホンアレイは任意の位置に設置可能とする．従って，カメラとスピーカの位置は既知で，マイクロホンアレイの位置は未知である．故に，顔の存在する確からしさと音源の存在する確からしさを情報統合用座標上で統合するためには，予めカメラとマイクロホンアレイの位置関係を検出する必要がある．本システムでは，この位置関係を検出するために，図2の Step 1 に示す聴覚情報と視覚情報の統合処理を行う．この統合処理においては，視覚情報処理による詳細なマイクロホンアレイ探索範囲を限定することによる処理時間の短縮と検出率の向上を目的として，死角の少ない聴覚情報を利用して，カメラを基準としたマイクロホンアレイの位置を粗く検出し，視覚情報により位置と回転角度の詳細な値を求める．次に，Step2 に示す情報統合処理により，Step1 で求めたカメラとマイクロホンアレイの位置関係を基にして，顔の存在する確からしさと音源の存在する確からしさを統合し，話者位置を検出する．これらの処理により，カメラとマイクロホンアレイの位置関係が未知の場合でも，情報統合による話者位置検出が可能になる．

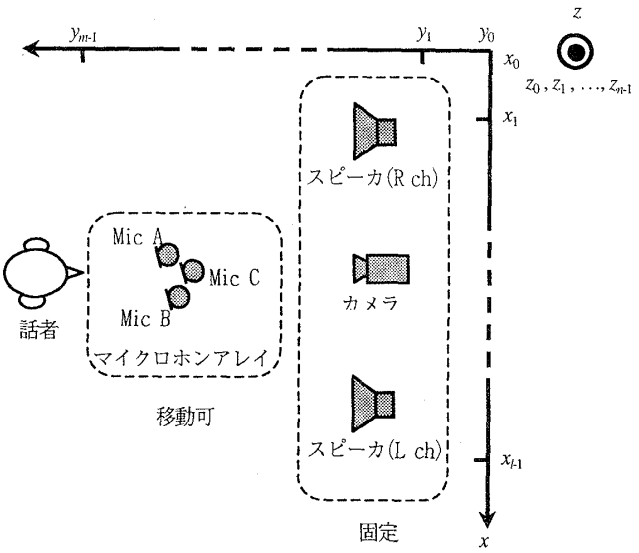

図3　情報統合処理と視覚情報処理用座標

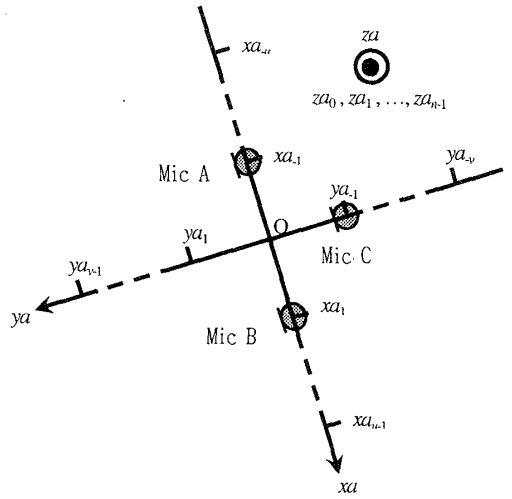

図4　聴覚情報処理用座標

2．2．カメラとマイクロホンアレイの位置関係検出

図2に示す Step1 において，次のカメラとマイクロホンアレイの位置関係検出処理を行う．

(1)スピーカ位置検出

図5にスピーカ位置検出処理のブロック図を示す．マイクロホンアレイを用いた音源位置検出処理により，図4に示した聴覚情報処理用座標上でのスピーカ位置を検出する．ここで，マイクロホンアレイの構成マイクロホンは，3個の無指向性マイクロホン Mic A, Mic B, Mic C で，図4に示すように *xa ya* 平面上に配置した．

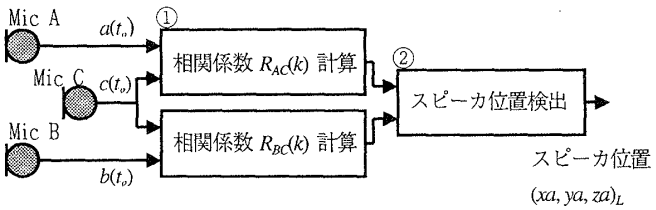

図5　スピーカ位置検出処理

次に，図5の各処理内容について説明する．

①相関係数 $R_{AC}(k)$ 計算，相関係数 $R_{BC}(k)$ 計算

図3に示したステレオスピーカの L ch より音を出力し，Mic A と Mic C の入力信号の相関係数 $R_{AC}(k)$ と，Mic B と Mic C の入力信号の相関係数 $R_{BC}(k)$ を計算する．

②スピーカ位置検出

相関係数 $R_{AC}(k)$と $R_{BC}(k)$を基に，三角測量の手法により，Lch スピーカの位置$(xa, ya, za)_L$を検出する．

同様に R ch スピーカから音を出し，聴覚情報処理用座標上で，R ch スピーカの位置$(xa, ya, za)_R$を求める．

(2)マイクロホンアレイ位置・回転角度検出

図6にマイクロホンアレイの位置・回転角度検出処理のブロック図を示す．(1)のスピーカ位置検出で求めたスピーカ位置を基に，視覚情報処理用座標上におけるマイクロホンアレイの位置を粗く求め，次に視覚情報処理により，マイクロホンアレイの位置と回転角度を詳細に検出する．

次に，図6の各処理内容について説明する．

①マイクロホンアレイ位置検出

聴覚情報処理用座標上のスピーカ位置より，同じ座標上のカメラ位置を求める．このカメラ位置を基に，視覚情報処理用座標上で，マイクロホンアレイ位置を粗く求める．

②カメラ方向制御

上記①で求めたマイクロホンアレイの位置にカメラを向ける．これにより，カメラの視野内にマイクロホンアレイをとらえる．

③マイクロホンアレイ位置・回転角度検出

視覚情報処理により，アレイ上面に取り付けたマイクロホンマーカの入力画像を基に，視覚情報処理用座標上のマイクロホンアレイの詳細な位置と回転角度を検出する．この処理に用いるマイクロホンマーカを，図7に示す．このマーカには白地に直径の異なる二つの黒円を配置し，2値画像より2円の重心を求め，カメラからの方向を求める．また，2円の重心間距離よりカメラとマーカまでの距離を，重心間を結ぶ線分の傾きより，図8に示すマーカの回転角度を検出する．なお，撮影方向によらず2円を区別できるように2円で異なる直径を用い，面積の大きい黒円を基準として回転角度を求める．

(3)座標変換表計算

(2)のマイクロホンアレイ位置・回転角度検出で求めたマイクロホンアレイの位置と回転角度を基に，

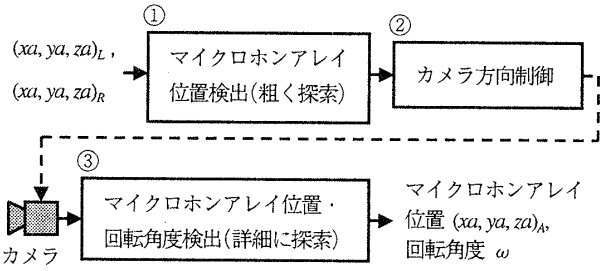

図6　マイクロホンアレイ位置・回転角度検出処理

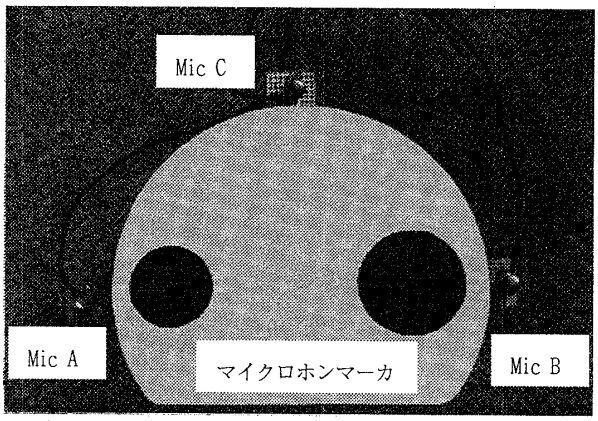

図7　マイクロホンアレイとマイクロホンマーカ

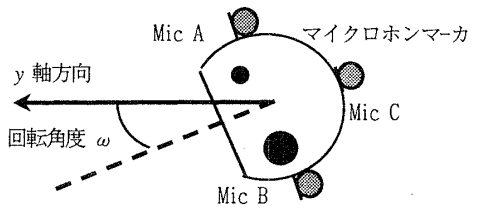

図8　マイクロホンアレイの回転角度

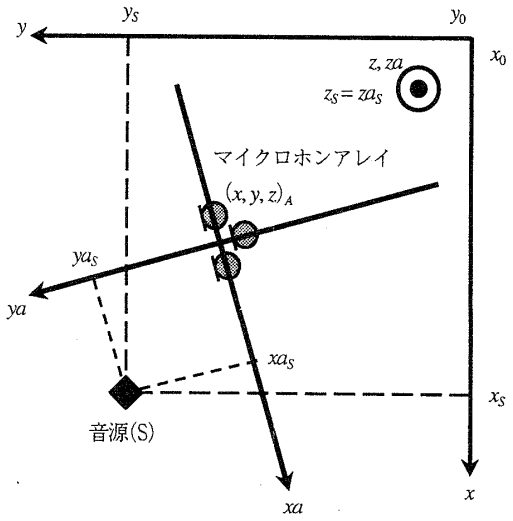

図9　座標変換

図9に示すように，平行移動とz軸周りの回転により，聴覚情報処理用座標上の音源位置$(xa, ya, za)_S$に対応する情報統合用座標上の位置$(x, y, z)_S$を求める．この座標変換を式(1)に示す．ここでは，話者位置検出時の処理量削減のため，式(1)を用いて，予め聴覚情報処理用座標上の各位置に対応する情報統合用座標上の位置を計算し，座標変換表を準備する．一例として，マイクロホンアレイの位置が$(x, y, z)_A$=(20, 10, 5)，回転角度がω = 0度のときの座標変換表を表1に示す．

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_S = \begin{bmatrix} \cos(-\omega) & -\sin(-\omega) & 0 \\ \sin(-\omega) & \cos(-\omega) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} xa \\ ya \\ za \end{bmatrix}_S + \begin{bmatrix} x \\ y \\ z \end{bmatrix}_A \quad (1)$$

2．2．話者位置検出

図2に示すStep2において，次の話者位置検出処理を行う．

(1)顔の位置検出

図10に顔の位置検出処理のブロック図を示す．モデルとなる顔画像との色ヒストグラムの照合[4]を行い，図3の視覚情報処理用座標における顔の存在する確からしさを求める．

次に，図10の各処理内容について説明する．

①HSI変換

入力画像のRGBデータをHSIデータ(Hue, Saturation, Intensity)へ変換する．

②色ヒストグラム計算

画像を複数の部分領域に分け，部分領域ごとにHSI空間におけるヒストグラムを計算する．

③照合

テンプレートであるモデル画像の色ヒストグラムと入力画像の各部分領域の色ヒストグラムの一致度を計算し，その結果を，顔の存在する確からしさ$r_v(x, y, z)$とする．

(2)音源位置検出

図5のスピーカ位置検出と同様の処理を行う．Mic AとMic Cの入力信号の相関係数と，Mic BとMic Cの入力信号の相関係数を計算し，それらの積を，図4の聴覚情報処理用座標における音源の存在する確からしさ$r_a(xa, ya, za)$とする．

表1　座標変換表

聴覚情報処理用座標	(-10, -10, 0)	…	(0, 0, 0)	…	(9, 9, 4)
情報統合用座標	(10, 0, 5)	…	(20, 10, 5)	…	(29, 19, 9)

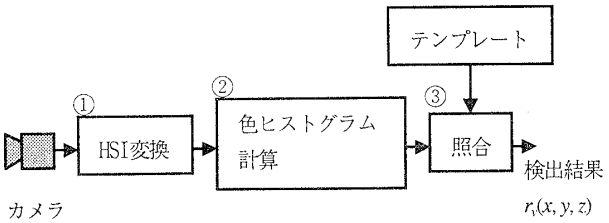

図10　顔の位置検出処理

(3)情報統合

表1に示すような座標変換表を用いて，聴覚情報処理用座標における音源の存在する確からしさ$r_a(xa, ya, za)$を，情報統合用座標上の値$r'_a(x, y, z)$に変換する．この音源の存在する確からしさ$r'_a(x, y, z)$と顔の存在する確からしさ$r_v(x, y, z)$の積より，話者の存在する確からしさ$r(x, y, z)$を求める．

(4)話者位置検出

情報統合用座標における話者の存在する確からしさ$r(x, y, z)$の値が最大になる座標を，話者位置$(x, y, z)_{SP}$とする．

3．実験

今回提案した話者位置検出システムの評価のために，カメラとマイクロホンアレイの位置関係を検出する実験と話者位置を検出する実験を行った．

今回の実験において，図3の情報統合用座標は，次のように設定した．

- x軸：x_0, ～, x_{39}　(10 cm間隔)
- y軸：y_0, ～, y_{39}　(10 cm間隔)
- z軸：z_0, ～, z_9　(20 cm間隔)
- カメラ位置：(x_{20}, y_0, z_7)
- スピーカ(L ch)位置：(x_{13}, y_3, z_5)
- スピーカ(R ch)位置：(x_{27}, y_3, z_5)

また，図4の聴覚情報処理用座標は，次のように設定した．

・xa 軸：xa_{-10}, ～, x_9　　(10 cm 間隔)
・ya 軸：ya_{-10}, ～, y_9　　(10 cm 間隔)
・za 軸：za_0, ～, za_4　　(20 cm 間隔)

ここで，サンプリング周波数は8 kHz とした．

3．1．カメラとマイクロホンアレイの位置関係検出実験

座標変換表を計算するために必要となる，情報統合用座標におけるマイクロホンアレイの位置と回転角度の検出精度を調べる実験を行った．

・実験方法

図11 に×印で示す9ヶ所にマイクロホンアレイを置き，回転角度を ω=-90, -45, 0, 45, 90 度としたときの，次の二つの実験を行った．

- 実験1

聴覚情報処理によって検出したスピーカ位置を基に，情報統合用座標におけるマイクロホンアレイの位置を粗く求め，その方向にカメラを向け，カメラ視野内にアレイをとらえることができる割合を調べる．

- 実験2

視覚情報処理による，マイクロホンマーカを用いたマイクロホンアレイ位置と回転角度の検出精度を調べる．ここで，マイクロホンアレイの位置は，カメラを基準としたマイクロホンアレイまでの距離と方向の検出精度で示す．

・実験結果

- 実験1の結果

各回転角度のときの，マイクロホンアレイをカメラの視野内にとらえることができた割合を表2に示す．全体の平均は，約84 %であった．

表2の結果より，マイクロホンアレイの回転角度によって検出率が異なることが分かった．これは，図4に示す構成マイクロホン Mic A, Mic B, Mic C の配置により，音源位置検出精度が音源方向に依存するためである．従って，今後，マイクロホンアレイの形状の検討を行う予定である．

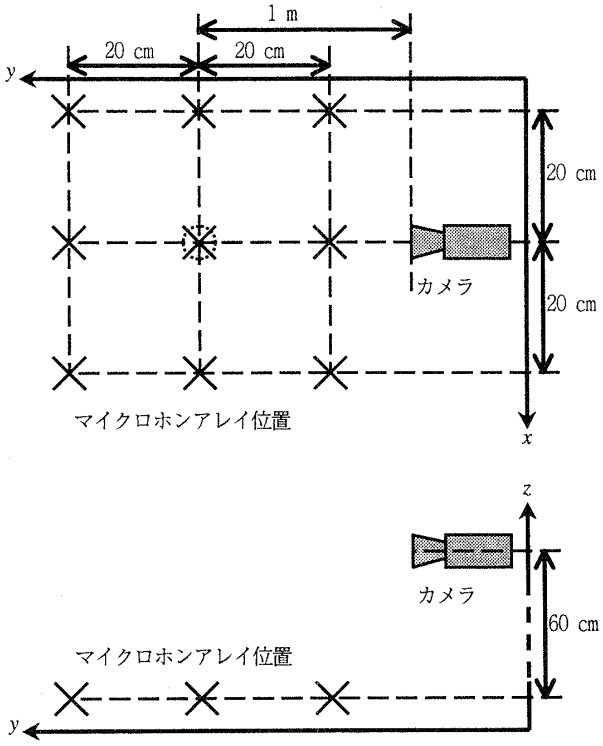

図11　実験におけるマイクロホンアレイ位置

表2　マイクロホンアレイ位置検出の実験結果

回転角度(度)	-90	-45	0	45	90
検出率(%)	100	82	76	76	84

- 実験2の結果

マイクロホンアレイの位置と回転角度の検出結果を次に示す．

位置の検出精度：距離 ±10 cm
　　方向 ±1度
回転角度の検出精度：±3度

このマイクロホンアレイの位置と回転角度の検出誤差は，次の話者位置検出実験における検出誤差の許容範囲を人体の大きさを考慮して±30 cm とした結果，話者位置検出率に影響が出ない値であることが分かった．

3．2．話者位置検出実験

外乱の存在する環境における話者位置検出率を調べる実験を行った．

・実験方法

図 11 の点線の円で示す位置にマイクロホンアレイを設置し，回転角度をω= 0 度とした場合の，話者位置検出精度を調べた．また，情報統合の効果を確認するため，視覚情報のみ，聴覚情報のみでの検出実験も同時に行った．

実験では，予めカメラとマイクロホンアレイの位置関係を検出して座標変換表を作成した．また，話者位置検出に関しては，人体の大きさを考慮して±30 cm を検出誤差の許容範囲とした．

外乱は次の3種を用いた．また，比較のため，外乱が無い場合でも実験を行った．

- 視覚情報の外乱
  - 蛍光灯・電球の切り替えによる照明条件の変動
  - 人物のポスターを置くことによる複雑な背景
- 聴覚情報の外乱
  - スピーカ出力の音声(64dB(A))による雑音

・実験結果

実験結果を図 12 に示す．外乱の無い状況での検出率は，情報統合を行った場合 96 %，視覚情報のみの場合 100 %，聴覚情報のみの場合 100 %であった．これに対し，外乱のある状況では，情報統合を行った場合 90 %，視覚情報のみの場合 79 %，聴覚情報のみの場合 46 %の検出率となった．従って，情報統合を行った場合，外乱のある環境での話者位置検出率が，視覚情報または聴覚情報のみを用いる場合に比べ，高いことが確認できた．

4．まとめ

今回，マイクロホンアレイとカメラの位置関係が未知の場合でも，カメラに対するマイクロホンアレイの位置と回転角度を検出し，その結果を利用して視覚情報と聴覚情報を統合するシステムを開発した．

今後は，方式改良を行い，マイクロホンアレイの位置と回転角度の検出率と話者位置の検出率の向上を図る予定である．また，パソコン等のインタフェースへの適用を検討する予定である．

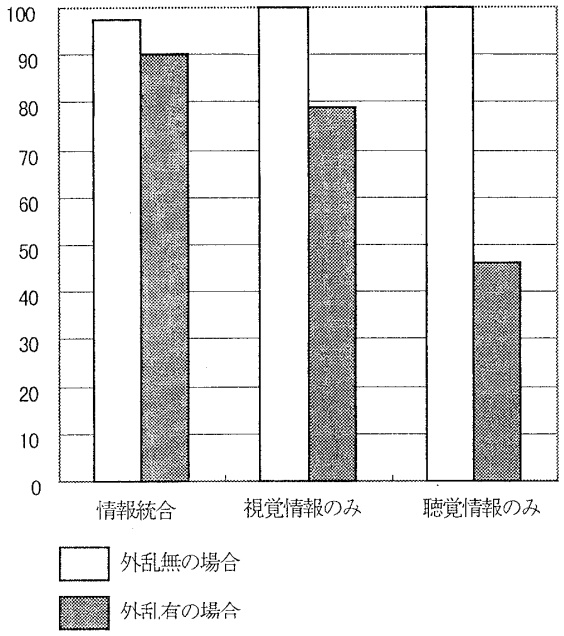

図12　話者位置検出の実験結果

参考文献

[1]畑岡，菊地，“音声を利用したマルチモーダルインタフェース，”信学誌 Vol.80 No.10, pp.1031-1035, 1997.

[2]松尾，北川，長田，“音声情報と画像情報の統合による話者位置検出システム，” 第 13 回ヒューマン・インタフェース・シンポジウム論文集，pp.469-474, 1997.

[3]Naoshi Matsuo and Shigemi Nagata ‘‘Study on Directional Microphone Technology using Estimated Signal,’’ In *Proc. ITC-CSCC '97*, pp. 425-428, 1997.

[4]M. J. Swain and D. H. Ballard, ‘‘Indexing via color histograms,’’ In *Proc. Image Understanding Workshop*, pp. 623-630, 1990.