あなたの声を聞き分ける全方位聴覚技術

Omnidirectional Acoustic Sense Technology for Voice Differentiation

鈴木　薫　　古賀　敏之
■ SUZUKI Kaoru　■ KOGA Toshiyuki

東芝は，人とロボットの自然な対話のため，全方位性を備えたロボット用聴覚システムを開発した。全方位聴覚技術を使うことで，ロボットは，様々な方向から複数の人に話しかけられても各々の声を聞き分け，その話しかけられた方向に応答することができる。この技術は，ハフ変換を用いて周波数－位相差散布図上に同一到達時間差音を示す直線を検出することにより音源の検出と定位を行うもので，開発したロボットApriAlpha$_{TM}$に実装し，全方位からの音声を実際に聞き分ける動作を実現させた。

Toshiba has developed a new omnidirectional acoustic sense technology to facilitate natural interactions between humans and robots. We used the Hough transform to detect straight lines from the frequency phase difference space for the detection and localization of sound sources. An ApriAlpha$_{TM}$ robot equipped with this function could localize and recognize multiple speakers from unlimited different directions and reply to each speaker.

1　まえがき

東芝は，家庭内で利用者と音声で自然に対話し，家電操作の支援や天気予報などの情報サービスの提供を想定したロボット“ApriAlpha$_{TM}$”（**図1**）を開発している[1]。

人とロボットの自然な対話のためには，移動したり向きを変えたりして活動しているロボットに，利用者が自分のいる場所から話しかけて命令できるようにする必要がある。話しかけられたロボットは，利用者の音声がどの方向から来たかを判断し，命令された内容を聞き分け，その利用者の方を向いて応答する。この動作を実現するため，当社は，①様々な方向から到来する音声を検出し，②方向ごとに定位して，③個別に分離認識できる，ロボット用聴覚システムを開発した。

ここでは，このような全方位性を備えたロボット用聴覚システム[2]の概要を述べる。

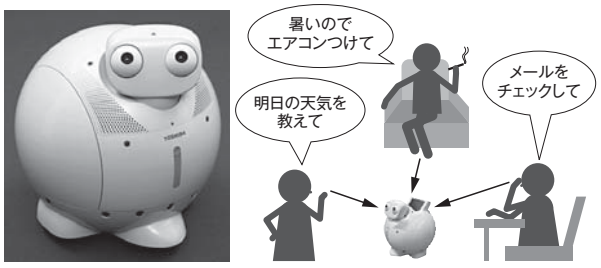

図1. ApriAlpha$_{TM}$とその利用シーン ― 複数のマイクを搭載し，四方八方からの音声命令に対して，その方向と内容を聞き分ける。
ApriAlpha$_{TM}$ robot and scene of its application

2　動作原理

このシステムは，技術的には音の周波数と位相差の関係に着目し，音源の数と方向の推定を周波数－位相差空間における直線検出問題に帰着させ，後述するように位相差の循環性を考慮した独自のハフ変換応用アルゴリズムで解く。更に，検出された直線をロボットに装備した複数のマイクの対の間で対応付けて音源候補の空間定位を行い，適応アレイ処理[3]によって音源音をその他の音声（生活雑音など）から分離して認識する。

2.1　音源の検出原理

マイク1とマイク2から成るマイク対を考える。音源がマイク間距離dに比べて十分遠く，途中に障害物がないと仮定するなら，音源を発してマイク対に到達する波面はほぼ平面である。この平面波を観測すると，両マイクを結ぶ線に対する音源方向の角度に応じて，両マイクで観測される音響信号に所定の到達時間差ΔTが観測される。

到達時間差ΔTは，音源の方向に応じて$\pm\Delta Tmax$の範囲で変化し得る。ここで$\Delta Tmax$は，音速をVsとして，$\Delta Tmax = d/Vs$として定められるΔTの最大値である。このとき音源の方向ϕは，マイク間を結ぶ線の中点を基点として垂直な方向を0とし，(1)式を用いて計算する。なお，ΔTは音源の方向により符号付きの量となり，ϕも符号付きの角度となる。

$$\phi = \sin^{-1}(\Delta T/\Delta Tmax) \qquad (1)$$

また，入力音声を高速フーリエ変換（FFT）で周波数分解し，周波数成分fのマイク間位相差をΔPhとすると$\Delta Ph = \Delta T \times 2\pi f$であることから，到達時間差$\Delta T$と$\Delta Ph$の関係は(2)式となる。

$$\Delta T = \Delta Ph / 2\pi f \qquad (2)$$

このシステムでは，ΔPhをマイク1におけるfの位相角$Ph1$とマイク2におけるfの位相角$Ph2$の差とし，その値を$\{\Delta Ph : -\pi < \Delta Ph \leqq \pi\}$に収まるように幅$2\pi$の剰余系として算定し，$x$軸を$\Delta Ph$，$y$軸を$f$とする2次元座標系上の点$(x, y)$としてプロットすることで，周波数−位相差散布図を得る。

ΔTが一定のときΔPhはfに比例することから，同じΔTを持つ周波数成分は散布図上で原点を通る直線状に並ぶ。この直線の傾きはΔTに応じて変化することから，ΔTを同じくする周波数成分は同一音源に由来すると仮定すれば，音源の数と方向の推定はこのような直線を散布図上で発見することに帰着できる。そして有力な直線を検出できれば，この直線の傾きから(2)式で求められるΔT，すなわち(1)式で示されるϕの方向に，この直線に寄与した周波数成分を持つ音源の存在を仮定できる。そこで，散布図上の点群から直線を検出する手段として直線ハフ変換[4]を用いる。

2.2　直線ハフ変換

2次元xy座標系上の点p(x, y)を通る直線は無数に存在する。原点から各直線に下ろした垂線のX軸からの傾きをθ，この垂線の長さをρとすると，ある点p(x, y)を通る直線の取り得るθとρの組は，$\theta\rho$座標系上で固有の軌跡（$\rho = x\cos\theta + y\sin\theta$）を描くことが知られている。この点から軌跡への変換を直線ハフ変換という。

複数の点を共通に通る直線は，各点の軌跡が一点で交差する場所として現れるため，所定の投票バッファに軌跡を投票して得票分布S(θ, ρ)を形成し，そこで高得票の位置を検出することで，多数の点を通る有力な直線を検出できる。これをハフ投票という。このとき，投票値を周波数成分fの対数パワーに比例するよう改良することで，パワーの大きい周波数成分を持つ音源を選択的に検出できるようになる。このとき，ΔPhとθの関係は(3)式となり，(2)式と(1)式を使って，傾きθから音源方向ϕを求めることができる。

$$\Delta Ph = f\tan(-\theta) \qquad (3)$$

2.3　ρ=0の制約と位相差の循環性

マイク1，2の信号が同相でA/D（Analog to Digital）変換される場合，検出しようとしている直線は必ずρ=0，すなわちxy座標系の原点を通る。したがって，音源の推定は得票分布S(θ, ρ)でρ=0となる，θ軸上の1次元の得票分布H(θ)=S$(\theta, 0)$からローカルピークを探索する問題に帰着する。

しかし，S$(\theta, 0)$上で直線検出が完了するためには，解析対象となる最高周波数で真の位相差が$\pm\pi$を逸脱しない（音源方向が正面付近に限られている）ことが条件となる。この条件はΔTが1/Fr秒（Frはサンプリング周波数）を超えないことで，ΔTが1/Frを超える場合には，次に述べるように位相差が循環性を持つ値としてしか得られないことを考慮しなければならない。

手に入れることのできる周波数成分ごとの位相角は，$-\pi$からπの間というように2πの幅でしか得ることができない。これは実際の位相差が両マイク間で1周期以上開いていてもデータとして得られた位相角からはそれを知ることができないということであり，ΔTに起因する真の位相差は，得られた位相差の$\pm 2\pi$や，更には$\pm 4\pi$や$\pm 6\pi$の値かもしれない。

ρ=0の制約では原点を通る直線だけを探すことになるので，このように2π周期で循環した位置に現れる点群を正しくカウントしていない。これは傾きの大きな直線ほど得票で不利になることを意味している。θ全域の公平な探索を実現するには，循環して現れる点群が構成する原点を通る直線に対する平行線も得票分布H(θ)に加えなければならない。この平行線の間隔$\Delta\rho$は，直線の傾きθの関数として(4)式で定義される符号付きの量となる。

$$\begin{aligned} \Delta\rho(\theta) &= \ \ 2\pi\cos\theta \quad : \theta > 0 \\ \Delta\rho(\theta) &= -2\pi\cos\theta \quad : \theta < 0 \end{aligned} \qquad (4)$$

2.4　位相差循環を考慮した直線群の検出

室内の雑音環境下でふたりの人物が，マイク対の正面約20度左と約45度右からほぼ同時に発話した際の直線群検出結果の例を**図2**に示す。

(c)のハフ投票結果で，白い破線は平行直線群を成す同一θについて$\Delta\rho$ずつ離れた投票位置を表している。θ軸と破線は等間隔に離れており，それぞれ$\Delta\rho(\theta)$のa(自然数)倍である(図ではa=2までを描画)。また，直線が循環しないθの領域(中央部)には破線を描画していない。

あるθ_0の得票H(θ_0)は，$\theta=\theta_0$の位置で縦に見たときのθ軸上と破線上の得票の合計値，すなわち，(5)式で計算される。

$$H(\theta_0) = S(\theta_0, 0) + \Sigma\{S(\theta_0, a\Delta\rho(\theta_0))\} \qquad (5)$$

この操作は，$\theta=\theta_0$となる原点を通る直線とその平行線の得票を合計することに相当する。この得票分布H(θ)を棒グラフにしたものが図2(d)である。

この得票分布H(θ)から，図2(e)に示す13個のローカルピークが検出される。このうち，上位二つのローカルピークが，マイク対の正面約20度左と約45度右からの各音声を検出した直線群に対応している。このように，$\Delta\rho$ずつ離れた箇所の得票値を合計して極大位置を探索することで，角度の小さい直線から角度の大きい直線まで検出できるようになる。

2.5　ストリーム追跡

前節までに述べたように，直線群はハフ投票ごとに時系列的に求められることになる。このとき，直線群のθは音源方向ϕと1対1に対応しているので，音源の静止または移動によらず，安定な音源に対応するθの時間軸上の軌跡は連続するは

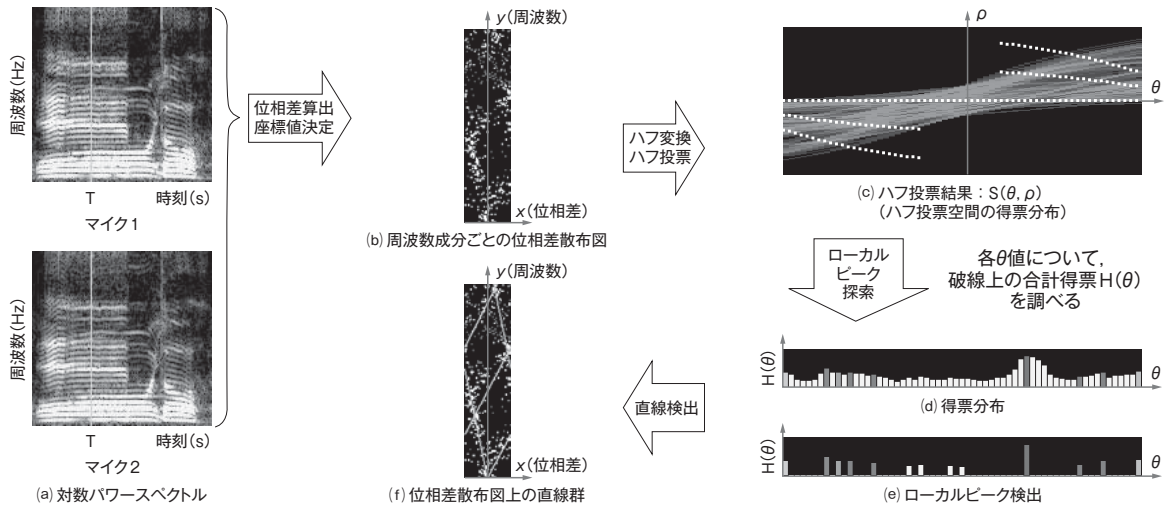

図2．2話者による同時発話時の直線検出結果 — 方向の異なるふたりの話者に対応した二つの直線群が検出される。

Results of line detection experiment in case of two simultaneous speakers

ずである。一方，検出された直線群の中には，ローカルピーク検出のしきい値設定の具合によっては背景雑音に対応する直線群が含まれていることがある。しかし，このような直線群の軌跡は連続していないか，連続していても短いことが予想されるため，直線群検出の周期ごとに求められるθの時間軸上の軌跡を追跡して，長く連続するグループを検出すれば有力な音源候補を選別できる。このグループをストリーム，グループ分けを行う処理をストリーム追跡と呼ぶ。

2.6　成分推定とストリーム照合

これまでの処理は，一つのマイク対でそれぞれ実行される。ロボットは複数のマイク対を並行して処理できるので，マイク対をまたぐ処理を行うことで更に音源についての情報を得ることができる。

異なるマイク対で検出されたストリームでも，同一音源に由来するかぎり，その継続期間と周波数成分は似ているはずである。既に述べたように，各マイク対で検出される音源の主な周波数成分は，その証拠となった直線群の近傍に分布する散布図上の点を選別することで近似的に得ることができる。このように音源の周波数成分を粗く推定し，推定された周波数成分を比較，照合することで，あるマイク対の直線群が，別のマイク対のどの直線群と似ているかを評価することができる。

ストリーム照合は，推定された音源の周波数成分及び継続期間を評価することで，同時期に似た周波数成分を持つ音源をマイク対間で対応付ける処理である。対応付けられる相手が見つからないストリームはノイズとして削除される。

2.7　音源定位

ストリーム照合によって対応付けられたストリームは，θから計算された各マイク対に対する音源方向φを，対応付けられたマイク対の数だけ持っている。これをマイク対の数nを使って表すと，(6)式となる。

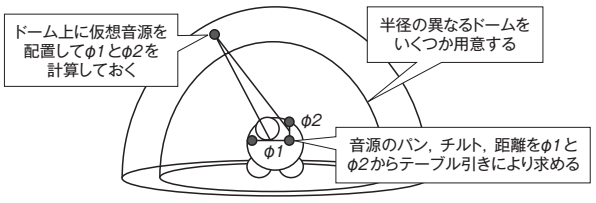

図3．仮想ドームによる音源定位 — 観測されたφ1とφ2の組合せにもっとも近い方向にある仮想音源を選び出すことで，音源の空間定位を行う。

Principle of sound source localization with virtual domes

$$音源方向情報 = \{\phi 1 \cdots \phi n\} \qquad (6)$$

この集合は，ある空間位置にある音源がそれぞれのマイク対から見てどの方向にあるかを示したデータである。そこで，ロボットを中心に仮想的なドームを考え，そのドーム表面に適度な間隔で離散した仮想的な音源を配置し，それぞれの仮想音源が各マイク対のどの方向にあるかをあらかじめ計算してテーブル化しておく。

二つのマイク対を使ったときの音源定位の概念を**図3**に示す。各マイク対から得られた音源方向φ1とφ2について，最小二乗誤差となるドーム上の仮想音源を探索して空間定位を行う。

3　システム構成と動作例

3.1　システム構成

2章で述べた処理によって全方位聴覚を実現するシステム構成を**図4**に示す。

ロボットはマイク1～6を装備している。システムは，一つ

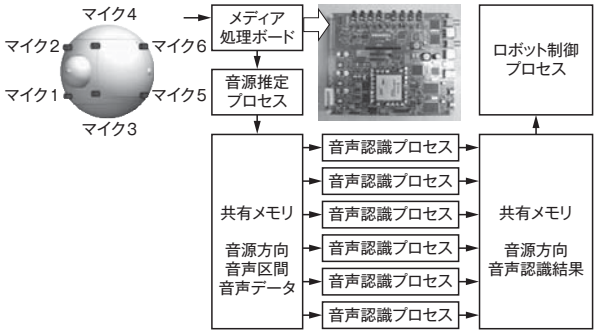

図4. システム構成 — 音源推定プロセスで推定された音源の音声は，音声認識プロセスで最大6個まで同時に認識可能であり，ロボット制御プロセスは，その認識結果に基づいてロボットを動作させる。

System configuration

の音源推定プロセスと最大六つの音声認識プロセスとで構成され，プロセス間は共有メモリで結ばれている。

3.2 音源推定プロセス

音源推定プロセスは，2章で述べた信号処理により，各マイク対に対する音源数と方向の推定処理及び，複数のマイク対を使った音源の空間定位処理を行う。

3.3 音声認識プロセス

音声認識プロセスは，認識エンジンの前段に適応アレイ処理を配したプロセスであり，マイク対に対する音源の方向ϕを使って適応アレイの追従範囲を設定し，音源音の抽出と抽出された音声の認識を行う。音声認識プロセスは，音源推定プロセスによって方向の異なる音源ごとに処理対象を割り当てられて認識を実行する。マイクからの音響信号は，新開発のメディア処理ボードによって全チャネル同期して取込みがなされるため，ローカルピーク探索時には散布図上で原点を通る平行直線群を探索すればよい。

音源が検出されると，音源推定プロセスがその音源方向に対して他の音源方向と重ならないユニークなマイク対を選択し，音声認識プロセスがこのマイク対からの入力音声を認識する。音声認識プロセス内では，選択されたマイク対からの音響信号を適応アレイ処理して雑音を除去した後，認識する。

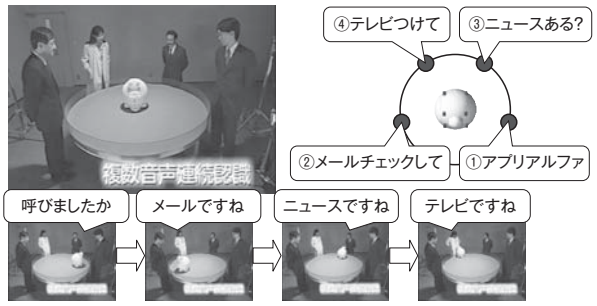

図5. 4話者による連続発話時の認識実験 — 複数の話者が順番に話したその方向と内容を認識して，話した順に応答する。

Demonstration of recognition of four consecutive speakers

3.4 複数話者発話時の全方位性確認の実験

4人の話者が順に発話したときのロボットの動作例を**図5**に示す。人とロボットは約1 m離れ，図中の①～④は発話順を表している。ロボットがこの発話順に話者の前まで移動して，発話内容に応答する音声を出力することを確認できた。

また，ふたりの話者がほぼ同時に発話したとき，ロボットは発話の開始順に話者の前まで移動して，発話内容に応答する音声を出力することも確認できた。

4 あとがき

ここでは，家庭内で利用されるロボットにとって不可欠な全方位での音源検出，音源定位，音源分離，及び音声認識を行うロボット用聴覚システムについて，その動作原理，システム構成，及び動作例を述べた。このシステムは周波数と位相差の関係に着目し，音源検出問題を直線検出問題に帰着させ，位相差の循環性を考慮したハフ投票アルゴリズムで解く。ロボットの動作のようすを4話者の連続発話時，2話者の同時発話時について示した。

なお，この開発は独立行政法人 新エネルギー・産業技術総合開発機構の次世代ロボット実用化プロジェクトに採択され実施したもので，2005年愛・地球博のプロトタイプロボット展及び常設展で実演を行った。また2006年には，アキバロボット運動会で日本語による展示を，ドイツで開催のIFA2006ではドイツ語と英語による展示を行い，ロボット型音声インタフェースを介した家電操作と情報サービスを，来場者自身の声で体験してもらっている。

文 献

(1) 尾崎文夫，ほか．“ロボット情報家電ApriAlphaの情報サービス機能”．第22回日本ロボット学会学術講演会．岐阜，2004-09．2E16.
(2) 鈴木 薫，ほか．“ハフ変換を用いた音源音のクラスタリングとロボット用聴覚への応用”．人工知能学会第22回AIチャレンジ研究会資料 SIG-Challenge-0522．静岡，2005-10．p.53-58.
(3) 天田 皇，ほか．音声認識のためのマイクロホンアレー技術．東芝レビュー，**59**，9，2004，p.42-44.
(4) 岡崎彰夫．はじめての画像処理．東京，工業調査会，2000，218p.

鈴木 薫 SUZUKI Kaoru
研究開発センター マルチメディアラボラトリー研究主務。
文字・図面・画像認識，音声処理，CG，HI，及びロボットの研究・開発に従事。情報処理学会，システム制御情報学会会員。
Multimedia Lab.

古賀 敏之 KOGA Toshiyuki
研究開発センター ヒューマンセントリックラボラトリー。
画像認識，音声処理，HI，及びロボットの研究・開発に従事。日本ロボット学会会員。
Humancentric Lab.