［特別寄稿］

ロボティックホームのための サウンドスポット形成技術

Sound Spot Forming Technology for Robotic Home

溝口　博
■MIZOGUCHI Hiroshi

人の周りにいて人と共存すべきホームロボットの究極形は，家屋や部屋そのものがロボットであるような“ロボティックホーム”であろう。

著者らは，ロボティックホームのような環境型システムへの適用を想定し，対象とする人の周りでだけスポット的に音のやり取りができるような，新しい非束縛型ヒューマンインタフェースの実現に取り組んでいる。これは“人の存在を認識”してその人に注意を向け，いわば“聞き耳をたてる”ような形で音声を拾い，“耳もとで語りかける”ような形で音を聴かせる技術の確立を目指すものである。ここでは，その一環として構築したスピーカ128台の大規模スピーカアレーと，これを用いて成功した別内容音声の複数箇所同時送出実験について述べる。

One of the ultimate forms of home robot is the “robotic home,” a system in which a room or house itself is a robot. Pioneering projects in this field include the Robotic Room of the University of Tokyo, the Aware Home of the Georgia Institute of Technology, and Home_*n* of the Massachusetts Institute of Technology.

The author and his colleagues are working on sound spot forming by a speaker array as a novel hands-free human interface applicable to such environmental systems. A sound spot is a small area of higher sound pressure level. A huge speaker array system of 128 loudspeakers was constructed and an experiment was conducted using this system. The results of the experiment demonstrated that the number of sound spots is not limited to one, but that four spots of different sound contents can be simultaneously formed. This paper describes the constructed speaker array and the experiment.

1　まえがき

人の周りにいて人と共存し，手助けや手伝いをしてくれるホームロボットへの期待と要望は，高齢化社会の到来を前に年々高まってきている。そのようなロボットの一つの究極形は，家屋や部屋そのものがロボットであるような環境型ロボット“ロボティックホーム”であろう。そのイメージは，例えばスタンリー キューブリック監督の映画「2001年宇宙の旅」に描かれるHAL9000である。小説や映画の中の想像の産物にとどまらず，関連の研究開発も，1990年代半ばの東京大学のRobotic Roomやマサチューセッツ工科大学のSmart Roomsを初めとして，マイクロソフト研究所のEasy Living，東京大学のIntelligent Space，ジョージア工科大学のAware Home，マサチューセッツ工科大学のHome of the Future（未来の家屋）研究コンソーシアムのHome_*n*と近年活性化してきている。

パーベイシブ コンピューティングやユビキタス コンピューティング，あるいはセンサ ネットワークといった技術の進展とあいまって，このようないわゆる“センサリッチ（sensor rich）”な環境型システムを通じた対人相互作用の実現が期待されている。特に，人がシステムを操作するタイプの従来型ヒューマンインタフェースに対し，人を見守り，人と音声を介して直接対話したり，人の身ぶりを理解したりできるような識別型ユーザーインタフェースPUI（Perceptual User Interface）への期待と要求が高まってきている。

しかしながら，音声理解や音声対話以前の問題として，広い範囲で動き回る人に対し，離れた所から，雑音なく高品質で音声を授受する手段さえも確立されていないのが現状である。“人に優しい”識別型ユーザーインタフェースを備えるロボティックホーム実現のためには，まず，人の存在を認識してその人に注意を向け，いわば“聞き耳をたてる”ような形で音声を集音し，“耳もとで語りかける”ような形で音を聴かせる技術の確立が急務である。

そこで著者らは，対象とする複数の人の頭部周辺に，それぞれスポット状の高感度・高音圧分布の“サウンドスポット”を作り出し，SN比（信号と雑音の比）の高い集音や伝送を実現して，たとえその人が動いてもサウンドスポットを追従させることが可能な技術の研究開発に取り組んでいる。対象とする複数の人々に対し，同時に別々の内容の音を聴かせたり拾ったりできることを目指す。具体的には，マイクロホンやスピーカを多数並べたアレーにより，サウンドスポットを形成する。

これまでにスピーカ128（32×4）台から成る大規模スピーカ

アレーを構築し，それを用いて別内容音声の複数箇所同時送出実験に成功した。すなわち，同時に複数の人の耳もとで，それぞれ別の内容を語りかけることを可能とした。また，“人の存在を認識”して注意を向ける技術にも取り組んでいる。複数台のテレビ（TV）カメラと実時間顔追跡視覚とを組み合わせ，対象とする人が広い範囲で動いても，それに追従してその人の位置座標を得ることに成功した。ここでは，これらの研究内容について述べる。

2 サウンドスポット形成

まず，構築した128チャンネル スピーカアレーについて述べる。これは，32台ずつのスピーカを正方形状に配置したもので，サウンドスポットの形成位置を実時間で動的に変更可能という特長を持つ。以下，スピーカアレーの原理，構築したシステムの構成，構築システムを用いて行った実験及び結果の順で説明する。なお，ここではサウンドスポットの位置を“焦点”と呼ぶ。

2.1 サウンドスポット形成の原理

複数のスピーカから同時に同じ音を出力した場合，波の干渉の影響で様々な場所で増強や減衰が生ずる。スピーカアレーでは，焦点から各スピーカまでの距離の差から生ずる音の到達時間の差と減衰比とを求め，それらを補正する形の付加遅延時間と振幅とを各スピーカの出力に付与する。これにより，焦点位置における各スピーカからの音の波の位相と振幅を一致させ，互いに強め合うようにする。結果として，サウンドスポットが形成される。

今，スピーカの個数をN，i番目のスピーカから焦点までの距離をLi，Liの最大値をLmax，音速をVとすると，i番目のスピーカに付加するべき付加遅延時間Diと振幅比Aiは，それぞれ次の式で求めることができる。

$$D_i = \frac{(L_{\max} - L_i)}{V} \qquad A_i = \frac{L_i}{L_{\max}}$$

これらの計算結果を各スピーカから出力する値に付加することで，焦点だけで位相と振幅のそろった合成波が得られることになり，サウンドスポットが形成される。

2.2 スピーカアレーシステムの構成

構築した128チャンネル スピーカアレーシステムのブロック図を**図1**に，外観を**図2**に示す。このシステムを実現するためには，128台のスピーカを同時に制御する必要がある。また，CD相当の音質を得るためには，十数μs単位での制御が必要となる。しかし，市販のD/A（Digital to Analog）変換ボードはたかだか16チャンネル程度，しかも，出力も同時ではないものがほとんどで，128チャンネル同時出力などといった仕様のものは存在しない。そこで著者らは，CDと同じサンプリングレート44.1 kHz，すなわち周期約23 μsでのサンプリングが可能な，128チャンネル同時出力D/A変換ボードを新規に開発した。

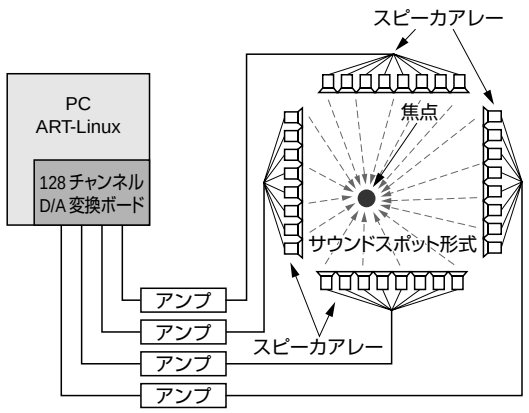


図1．構築したスピーカアレーのブロック図 ― 実時間OSを搭載した1台の汎用PCで128台のスピーカを制御している。23 μs周期の等周期ループがソフトウェアだけで実現されている。
Block diagram of implemented speaker array system

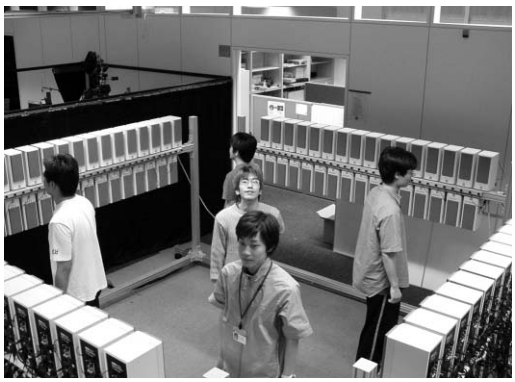


図2．構築したスピーカアレーの外観 ― 一辺が約3 mの正方形配置である。一辺に32台のスピーカが付いている。
View of 128-ch speaker array system

また，システム構築のうえでは，十数μsオーダでの周期の制御も不可欠である。このため著者らは，市販品も含め複数種類の実時間オペレーティングシステム（OS）を実動比較した。その結果，ART-Linuxだけが十数μsオーダの等周期ループを安定して実行可能であることを確認し，これを採用した。ほかの実時間オペレーティングシステムでは，周期変動がμsから十μsオーダに及ぶため，たかだかmsオーダの周期までしか安定して実現できない。ART-Linuxを用いることで，CDと同じ44.1 kHzのサンプリングレートがソフトウェアだけで実現可能となった。

アレーの要素スピーカには，市販のアンプ内蔵型パソコン（PC）用外付けスピーカを用いた。使用したスピーカは，安価ながら低域から高域まで周波数特性が良く，高音質の実現に

寄与している。構築したアレーの要素スピーカどうしの間隔は70.0 mmで，正方形一辺の長さは3.23 mである。

2.3　サウンドスポット形成実験

ここでは，構築したスピーカアレーを用いたサウンドスポット形成実験について述べる。目には見えない“音”を対象とする場合，その効果を定量的かつ視覚的に把握することは容易ではなく，なんらかの工夫が必要となる。この研究課題では，スピーカアレーによって作られる音場の音圧を全自動で測定できるシステムを，PC制御が可能なガントリークレーンと騒音計を用いて開発した。実験では，このシステムを用い，128チャンネル正方形配置型スピーカアレーに囲まれた2.25 m × 2.25 mの範囲を，150 mm間隔の256（16 × 16）点で測定した。音源としては，男性ニュースキャスターのアナウンス録音を用いた。グラフ化する際に用いた計測結果は，各点における10秒間の平均音圧である。

図3は，実際にこのシステムによって作られたサウンドスポット付近の音圧を，前記の自動測定システムで実測した音圧分布図である。焦点は（1.125，1.125）mの座標位置に設定した。図3を見ると，焦点とそのほかの位置で10 dB以上の音圧差が生じていることを確認できる。また，実際に焦点位置に立って聞いてみると，ほかの位置よりも大きくはっきりとした音声を聞くことができた。また，ソフトウェア上で新たな焦点位置を設定すると，その位置に焦点が移動することも，実際に聞いてみて聴感上確認した。以上のことから，ここで構築した128チャンネル正方形配置型スピーカアレーシステムを用いてサウンドスポットを形成可能であることが確認できた。

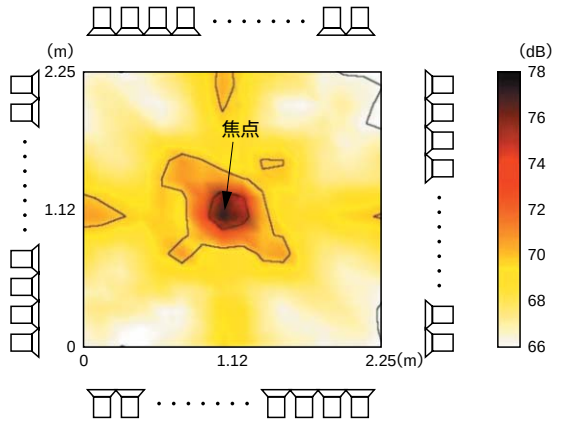


図3．1焦点の場合の音圧分布実測値 ― 焦点（サウンドスポット）が1個の音圧分布図である。周囲と10 dB以上の音圧差が生じている。
Spatial distribution of sound pressure level in case of one focus

図4は，焦点を（0.5，0.5），（1.75，0.5），（1.75，1.75），（0.5，1.75）mの4点に設定し，各焦点でサウンドスポットを形成し，前記のシステムで実測した音圧分布図である。各サウンドスポットは焦点に近い二つのアレーで形成している。音源は，音源による音圧のレベル差をなくすために，図3の実験と同じ男性ニュースキャスターのアナウンス録音を用いた。

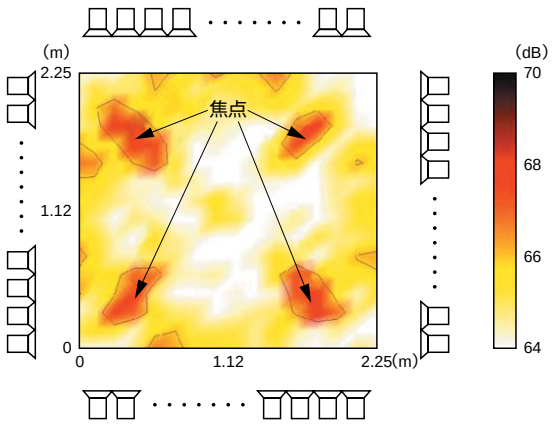


図4．4焦点の場合の音圧分布実測値 ― 焦点（サウンドスポット）が4個の音圧分布図である。各焦点と周囲とで数dBの音圧差が生じている。
Spatial distribution of sound pressure level in case of four focuses

図4を見ると，各焦点とその他の位置で4～5 dBほどの音圧差が得られている。実際に焦点位置に立って聞いてみると，ほかの位置よりも大きくはっきりとした音声を聞くことができた。また，各焦点で別の音声のサウンドスポットを形成するように設定し，実際に聞いてみたところ，各焦点において，それぞれ異なった音声を聞くことができた。以上のことから，対象となる人物が複数に増えた場合でもサウンドスポットを形成可能であることが確認できた。

3　実時間顔追跡視覚

スピーカアレーで囲まれた空間内に定めたサウンドスポット形成位置は，スピーカアレー座標で指定すればよい。視覚による顔発見と追跡を利用してその位置測定を行えば，顔のある座標位置，すなわち人の頭の位置にサウンドスポットが形成されることになる。

サウンドスポットの焦点範囲（音声の聞こえる範囲）は音声の周波数にもよるが，図4の実験結果を見る限り，平均的におよそ300 mm程度である。この範囲内に人の耳が入っていなければ音声を十分に聞かせることができない。幸い人の耳は顔の両脇についている。したがって，顔を発見し，その位置を測定することで音声を聞かせることが可能と考えられる。

3.1　カメラアレー

視覚による顔追跡は，テンプレートマッチング法を用いて実現した。テンプレートマッチングによる顔発見方法では，顔画像の情報量が十分でないと，顔発見の精度や成功率が低くなってしまう。つまり，画像の情報量，すなわち十分な解

像度が得られていないと顔発見ができない。しかし，カメラが対象人物を発見するべき範囲は，スピーカアレーのサウンドスポットの焦点形成が可能な範囲であり，これだけの広範囲を，顔発見が可能な高い解像度を保ったまま，一つのカメラで撮影することは困難である。そこで，複数のカメラを配置し，分散視覚化する。これにより，顔発見による位置測定が可能となる。ここでは，2台のカメラを並べて設置し，カメラアレーと呼ぶことにした。

3.2　位置情報の統合

各カメラがそれぞれ独立して顔発見位置の測定を行えば，位置情報は複数になる。サウンドスポットの焦点形成位置と

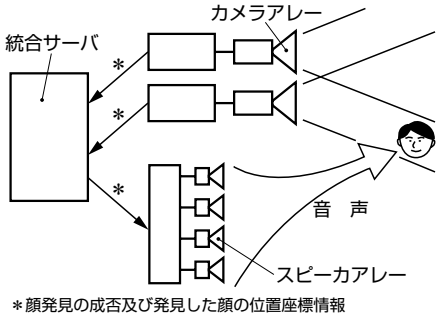
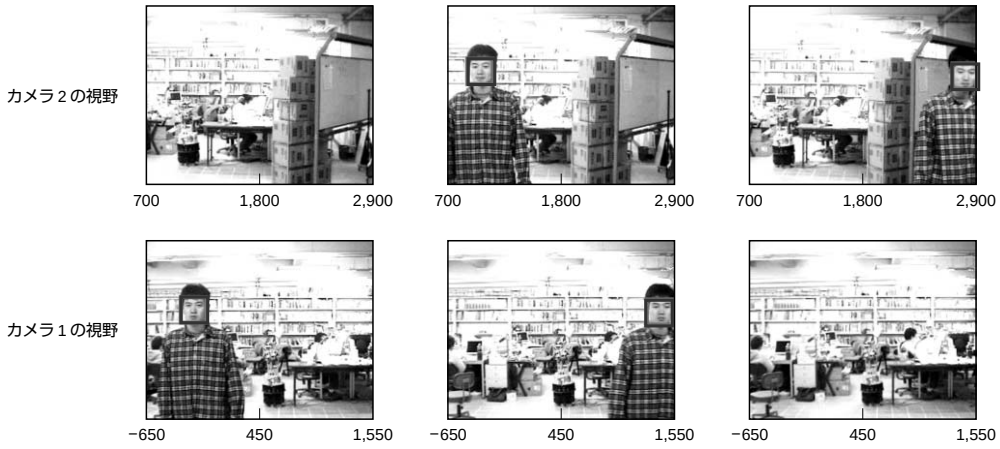


図5．カメラアレーとスピーカアレーとの関係 ― 各カメラからの顔発見・追跡情報は，統合サーバを介してスピーカアレーに伝えられる。それを基に焦点（サウンドスポット）の位置が更新される。

Block diagram of visually steered sound spot forming

図6．実装したカメラアレー ― カメラ2台から成る最小構成例である。

View of implemented camera array



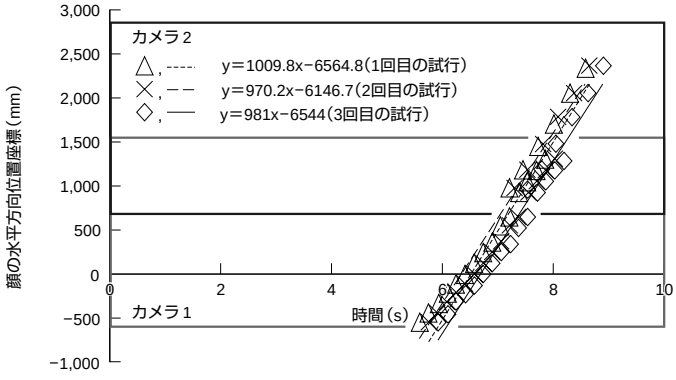


図7．実時間顔追跡実験の結果 ― (a)はカメラの視野を示す。追跡中の顔が枠で示されている。(b)は追跡した軌跡を示す。実時間で追跡できていることがわかる。

Experimental results of real-time face tracking

して与えるべき座標を一つに統合する必要がある。そこで**図5**に示すように，各カメラを管理し各位置情報の統合を行うサーバを用意する。これを統合サーバと呼ぶことにする。

統合サーバは，一定周期で各カメラの顔発見情報を集める。それらの測定値を調整し，一つの位置情報に統合してスピーカアレーに与える。カメラの数を2台とした場合，同時に顔発見に成功した場合は平均し，そうでない場合は顔発見に成功した値を優先する方法を用いた。

3.3 実時間顔追跡実験

顔追跡視覚の実験例として，移動する対象人物の追跡を行ったようすを**図6**に示す。2台のカメラがそれぞれ対象人物の位置測定を行えているかを確認する実験である。特に，いずれのカメラにおいても顔が確認された状況下では，両方の顔発見プログラムから同時に位置情報が送られてくるので，統合サーバが機能していることが重要となる。**図7**に示すグラフが実験結果である。3回の試行において，いずれの場合も約1.0 m/sで移動する対象に追従できていることがわかる。

4 あとがき

ここでは，ロボティックホームへの適用を想定して著者らが研究開発中のサウンドスポット形成技術について述べた。具体的には，これまでに実施した①128チャンネル大規模スピーカアレーの構築，②それを用いて行った別内容音声の複数箇所同時送出実験，及び③複数台のカメラと顔追跡視覚との組合せによる広範囲な実時間顔追跡実験について述べた。

①と②は耳もとで語りかける技術の一環として位置づけられる。正方形状に配置した128チャンネルの大規模スピーカアレーを用いることで，別内容音声のサウンドスポットを4か所に同時に形成することに成功した。一方，③は人の存在を認識する技術の一環として位置づけられる。カメラアレーと実時間顔追跡視覚とを用いることで，対象となる人物が動いても，広い範囲でその人の顔を追跡し，顔の位置情報を得ることができた。

以上の①～③を連携させることで，サウンドスポットの対人追従が実現可能であると期待される。現状では，スピーカアレーによるサウンドスポット形成は二次元的なものである。これを三次元化すること，顔追跡視覚で奥行情報を取得できるようにすること，複数の人にも対応できるようにすることなども今後の課題である。

謝　辞

ここで述べた研究のうち，初期の一部はIPA（情報処理推進機構）未踏ソフトウェア創造事業の支援を受けて行われた。また現在，一部は文部科学省科学研究費補助金特定領域研究「情報学」の支援を受けて行われており，一部は産業技術総合研究所との共同研究として行われている。以上，記して謝意を表します。

溝口　博　MIZOGUCHI Hiroshi, D.Eng.
東京理科大学 理工学部機械工学科教授，工博。
産業技術総合研究所 客員研究員。人間と機械の実世界中でのインタラクションの研究に従事。IEEE，日本ロボット学会会員。
Tokyo University of Science