時間–空間マッピングによる音声ブラウジングツール

小林 稔[†,††] クリス シュマント[†]

音声データのブラウジングは，印刷文書をブラウジングするほど容易ではない．本論文では，カクテルパーティ効果と空間的な記憶を利用して音声データの時間軸に空間的にアクセスするブラウジングインタフェースを提案する．本システムのユーザは，人工立体音場内を移動しながら音声データを再生する複数の音源を介して音声データの内容を聞く．複数の音源が再生しているのは，長い音声データの複数の部分である．ユーザは，テープの早送りや巻戻しをするかわりに，これらの複数の音源の間で注意を切り替えることで音声データをブラウジングする．音源が動きながら音声データを再生する結果，音声データ中の話題ごとに固有の位置が割り当てられ，音声データの時間軸はユーザの頭のまわりの空間にマッピングされる．このマッピングによって，空間的な記憶を使って音声データ中の所望の話題にアクセスすることが可能となる．本論文では，このようなブラウジングシステムの実現に向けたシステム設計の過程を，開発した種々の立体音声へのインタフェース手法とあわせて報告する．

Mapping Time to Space for Audio Browsing

MINORU KOBAYASHI[†,††] and CHRIS SCHMANDT[†]

Browsing audio data is not as easy as browsing printed documents because of the temporal nature of sound. This paper presents a browsing environment that provides a spatial interface for temporal navigation of audio data, taking advantage of human abilities of simultaneous listening and memory of spatial location. In the virtual acoustic space of the system, users hear multiple moving sound sources playing different portions of one audio recording simultaneously. Instead of fast-forwarding or rewinding, users browse the audio data by switching their attention between the sound sources. The motion of the sound sources maps temporal position within the audio data onto spatial location around the users' head, so that the same portion of the audio recording always appears at the same location. Thus, listeners can use their memory of spatial location to find a specific topic. Users can also browse by pointing to a spatial location where his/her desired data may appear. Upon the user's request, the system creates a new sound source that begins playing the audio data from the point that corresponds to the spatial location. This paper describes the iterative design approach toward the audio browsing system, including the developement of user interface devices.

1. はじめに

印刷された本をパラパラとめくってブラウジングするとき，我々はページ全体に目を走らせ内容の量や構成を把握し，空間的な記憶をもとに特定の記事を探したりする．このような視覚メディアに対して，音声メディアは再生しなければ音として知覚することすらできず，ブラウジングは困難である．たとえば，録音された会議録から必要な部分を探して聞こうとするとき，我々はテープの早送り・巻戻し・再生を繰り返しながら必要な部分を探す．もしかしたら大切な部分を飛ばしてしまい聞き逃すかもしれない．本論文で提案するブラウジングシステム Dynamic Soundscape の目的は，会議の録音やラジオのニュース番組のような長い音声データを空間的に提示することで，空間的記憶能力を活用しながら音声データをブラウジングできる効果的なインタフェースを提供することである．

パーティなどの騒がしい環境で何人もの人が同時に話しているような場合でも，我々は特定の話し相手の話す内容を選択的に聞き取ることができる．また，背後で聞こえた音に気づいて注意を切り替えることができる．このような人間の能力を“カクテルパーティ効果”[1]とよぶ．本論文では，音声データの時間軸を空間にマッピングするシステムで，このカクテルパーティ効果を積極的に利用することによって，早送り巻戻しではない新しい音声ブラウジングインタフェースを実

† マサチューセッツ工科大学メディアラボ
MIT Media Laboratories
†† NTT ヒューマンインタフェース研究所
NTT Human Interface Laboratories

現する．

本論文ではまず，音声データのブラウジングインタフェースや人工的立体音響技術等の関連する研究を紹介し，本論文のインタフェースの2つの基本方針（同時並行再生と空間マッピング）について説明する．論文後半では，実際に機能するブラウジングシステムを実現するために行われたデザインの過程を示し，得られた知見を整理する．

2. 関連研究

2.1 AudioNotebook と Filochat

AudioNotebook[2]はコンピュータで機能拡張された紙のノートである．授業や会議の音声をページ上の書き込みと対応づけながら録音する．後からページ上のマークをクリックすると，そのマークを書き込んだときに録音された音声が再生される．AudioNotebookでは，ページ上の空間的位置に関連した記憶から，必要な音声情報を引き出すことができる．

Filochat[3]も録音された音声と書き込んだマークを対応づけて記録し，後でマークに対応する音声を引き出すことができる．Filochatは，書き込みは電子的に記録される電子ノートである．

AudioNotebookやFilochatは，音声データにアクセスするのに空間的な記憶を利用する点で本論文と関連が深い．しかし，AudioNotebookやFilochatがノートに書き込まれた視覚的マークを利用するのに対し，本論文では視覚的補助を用いず音声だけで空間的記憶を活用するブラウジングツールの実現を目標とする．

2.2 立体音声

本論文では音声データを空間に配置するのに，人工的に立体音声を生成しヘッドフォンを通して聞く人工立体音場を構成する技術を利用する．立体音声の生成技術は様々な研究がなされ，今現在も活発に研究が進められている分野であるが[4),5)]，本論文では既存の技術を利用し立体音声の利用技術を検討する．利用技術に関しては映像と音声を組み合わせたバーチャルリアリティなどでの利用が目立つ[5)]．会議室内の参加者の位置と音源の位置を対応づけたグループウェアへの応用や[6)]，声だけでなくキーボードを打つ音などを立体音場中に配置することで参加者の活動状態なども示すシステム[7)]なども開発されている．また，PittとEdwards[8)]によるステレオ音声を用いた盲人用のマウスによるオブジェクト操作インタフェースの研究がある．Pittらのシステムでは空間中のオブジェクトはつねに音を発しつづけ，その音量はマウスカーソルとオブジェクトの距離によって決まる．ユーザは所望の音が大きくなるようにマウスを操作することでオブジェクトにアクセスすることができる．

様々な利用技術が検討されてきたが，本論文で提案するような音声データの時間軸にアクセスするために立体音声を応用することは，これまで試みられていなかった．

2.3 SpeechSkimmer

SpeechSkimmer[9)]は，音声データをブラウズしたりスキム（ななめ読み）するツールである．早聞き再生や音声データ中の無声部分を取り除くことで，短時間で全体を聞くことができる．また自動的に話題の切れ目等の特徴点を抽出しジャンプすることで，音声データ中に含まれる様々な話題を“ななめ聞き”できる．

2.4 AudioStreamer

AudioStreamer[10)]はカクテルパーティ効果を利用した音声だけによるブラウジングツールである．3本の音声データが人工立体音場の3カ所に固定された音源から再生され，ユーザはそのうちの1つを聞いたり他に注意を切り替えたりしながら3本の音声データをブラウジングする．ユーザが1つの音源を選択すると，その音源の音量を上げ，その音源を選択的に聞き取ることを補助している．重要な話題が再生されるときには音量を上げたりベルを鳴らすことでユーザの注意を喚起する．

AudioStreamerによって，カクテルパーティ効果の音声ブラウジングへの応用の可能性が示された．本論文ではAudioStreamerのアイデアを発展させ，より効果的に音声データの時間軸にアクセスできるブラウジングツールを実現する．本論文のシステムとAudioStreamerの主な違いは，再生される音声データの数と音源の位置である．AudioStremaerでは3本の音声データを3カ所の固定位置の音源から再生するのに対して，本論文のシステムでは1本の音声データを移動する音源から再生する．移動音源から再生することで音声データの時間軸を空間にマッピングし，音声データの時間軸への空間的アクセスを可能にする．

3. Dynamic Soundscapeの概要

本論文のブラウジングシステムDynamic Soundscapeの基本的な考え方を説明するために，まずユーザがシステムからどのような音声を聞くのかを説明する．

図1は本システムの音空間の概念図である．“Speaker”は本論文のシステムの中心的な役割を果たす音声オブジェクトである．図中では説明のため

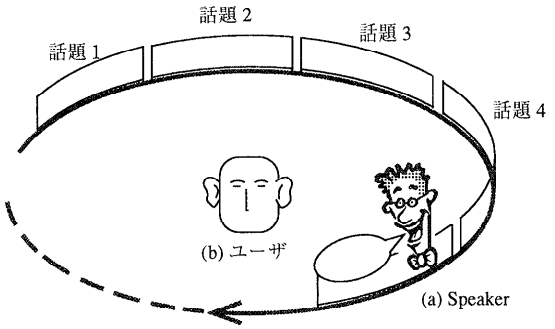

図 1 DynamicSoundscape の生成する音空間の概念図
Fig. 1 The concept of the auditory space created by the system. A Speaker (a) in the virtual audio space "speak"s audio data as it goes around the user (b).

に吹出しをつけた人の顔で表現しているが，実際には目には見えない音声オブジェクトである．ユーザは Speaker を通して，録音されたニュース番組等の音声データの内容を聞く．システムが始動すると，1 つの Speaker がユーザの頭を中心とした円形軌道上に生成される．Speaker は円形軌道上を一定方向に一定速度で移動しながら音声データを再生する．位置が時間の関数となっている移動 Speaker を通して音声データを再生することによって，音声データ中の時間と円形軌道上の位置の間の対応関係が形成され，音声データの時間軸は軌道上の空間にマッピングされる（図 1）．また，複数個の Speaker が同時に軌道上に存在し再生を行うこともある．1 つの音声データの複数の部分が同時に並行して再生されるので，ユーザは長い音声データのいろいろな部分を 1 度に聞いて，興味のある内容を再生している Speaker に注意を切り替えて聞くことができる．

すでに聞いた音声データを再度聞きたい場合には，その音声データが再生された位置をタッチパッド等のポインティングデバイスを用いて指し示す．すると，そこに新しい Speaker が生成され対応する音声データが再生される（**図 2**）．新しい Speaker が生成された後も，もとの Speaker は変わらずに音声データの再生を続けるので，複数の Speaker が 1 つの音声データの異なる部分を同時に再生することとなる．

音声データ中の先の部分にジャンプして再生することもできる．軌道上の適当な位置を指し示すことで新しい Speaker がその位置に生成され対応する音声データが再生される．この場合も，もとの Speaker は再生を続けるので，スキップした部分をもとの Speaker から聞き取ることができ，興味のある話題がもとの Speaker から聞こえれば，もとの Speaker に戻って聞くことができる．

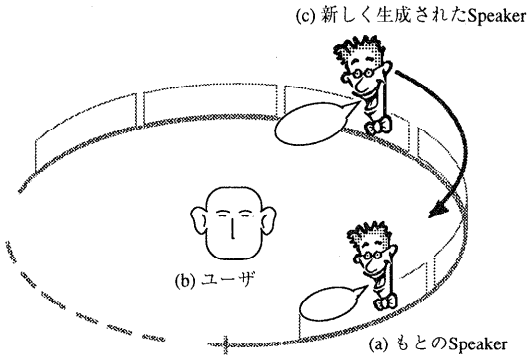

図 2 新しい Speaker の追加
Fig. 2 Adding a new Speaker: Upon user's request, a new moving Speaker (c) is created. The original Speaker (a) keeps going. The user hears multiple portions of the audio stream simultaneously.

Dynamic Soundscape は次の 2 つの基本アイデアに基づいている[11]．

(1) 音声データの時間軸の空間へのマッピング
(2) 単一の音声データの複数の部分の同時並行再生

(1) によって，ユーザは時間軸上に並んでいた音声データ中の個々の話題に空間的にアクセスすることが可能になり，空間的な記憶を活かせるようになる．我々の記憶や感覚の時間的属性は空間的属性に比べて曖昧なものである[12),13)]．たとえば "隣の人と話をしたのは何分前だったのか？" 時計を見ずに正確に思い出すのは意外と難しい．音声データの時間軸に空間的にアクセスできれば，"90 秒前に再生された部分" という代わりに，空間的な記憶を活かして "この辺で聞いた話" というアクセスが可能になる．

(2) によって，1 本の音声データの複数部分を同時に聞くことが可能になる．早送り巻戻しではなく，いろいろな話題を再生する Speaker の間で注意を切り替えることでブラウジングすることが可能になる．このようなブラウジングは，ページ全体を視野におさめ，いろいろな部分に焦点を移動させながら全体を見渡す印刷書類のブラウジングと似ているとも考えられる．

4. 実現へのアプローチ：初期システム

前章で示した Dynamic Soundscape を実現するにあたり，どのように設計するべきか未知の課題が多かったので，イテラティブな（繰返しの）デザインアプローチをとった．つまり，まず大まかにシステムを構築し，それを試した結果を反映させながら発展的にシステムを再設計してゆくこととした．本章では，最初に製作した初期システムを説明し，試用結果を基に

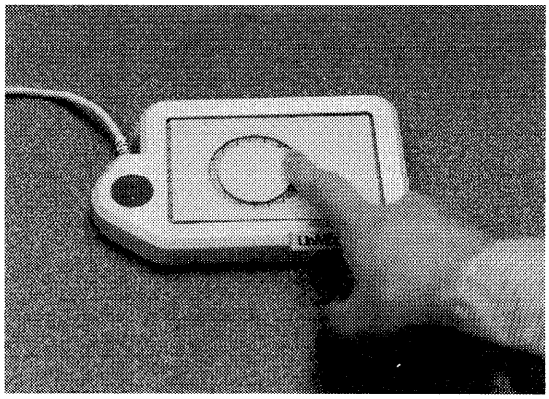

図3 タッチパッドインタフェース：円形軌道を指で触れて確かめられるようにテンプレートがついている．
Fig. 3 Touchpad interface: The template is attached to the surface of the touchpad, so the user can feel the shape of round path without seeing the device.

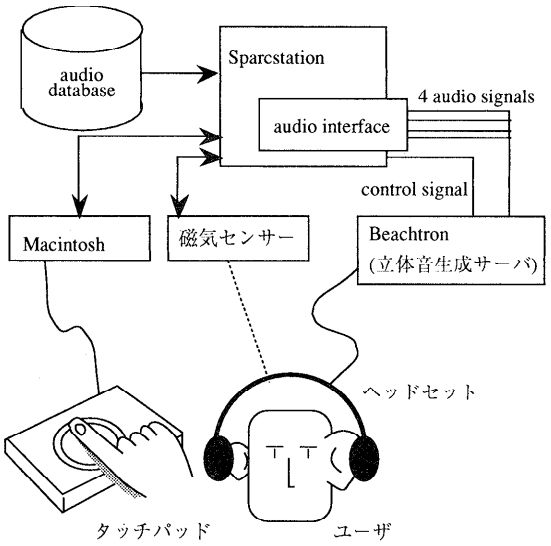

図4 システム構成
Fig. 4 System configuration.

有効なブラウジングツールへと改善するために解決すべき課題を整理する．

4.1 初期システム：Speakerの動作

音声データを空間にマッピングするSpeakerの動き方として円形軌道動作を選んだ[14]．円形軌道動作では音声データを2次元空間にマッピングするので，1次元の直線動作に比べ広いアクセス空間を利用できる．また，音源が時計回りで動くことをユーザに知らせておけば，音源の位置の前後を聞き分けられない場合でも右向きに進んでいる音源は前，左向きに進んでいる音源は後ろというように位置を判断することができる．Speakerは毎秒6度で上から見て時計回り方向に動くこととした．

4.2 初期システム：操作インタフェース

初期システムでは，単純なポインティングによるインタフェースだけを組み込んだ．ユーザはタッチパッド上の円形の溝（図3）の1点を押すことで，円形軌道上の点を指示する．指示された位置にSpeakerが存在しなければ，システムはその場所に新しいSpeakerを生成して対応する音声データの再生を始める．ハードウェアの制約から同時に存在できるSpeakerは最大4個である．それを超えて新しいSpeakerの生成が要求された場合は，最も古いSpeakerを停止し新しいSpeakerに割り当てる．また指示された点にSpeakerがすでに存在すれば，システムはその指示を焦点切替コマンドと解釈しそのSpeakerの音量を上げる．このような操作によって，ユーザは空間を指示することで音声データの時間軸にアクセスできる．

4.3 初期システム：システムの実装

システム構成を図4に示す．全体の制御機能と複数の音声データを同時再生するオーディオサーバの機能はSparcstation（Sun microsystems社製）が受け持つ．本システムではSparcstationの2つのステレオ出力を合計4つのモノラル出力として使用する．音声を空間定位させヘッドフォンを通して聞く立体音声を生成する処理は，立体音生成サーバに組み込まれた2枚のBeachtronカード（Crystal River社製）が行う．立体音生成サーバは4つのモノラル音声を人工立体音場の指定された位置に定位させる．

ヘッドセットには磁気位置センサが取り付けられ，ユーザの頭の位置と方向を検出する．音源の位置を特定するとき人間は頭を動かしながら音の変化を聞き取りそれを手がかりとするが，人工立体音場でも頭の方向や位置の情報を反映させて音を変化させることで立体感を強めることができる[15]．本システムでも，磁気センサで検出された頭の位置は立体音生成サーバで処理され立体音場に反映される．

その他に，マッキントッシュコンピュータを介してタッチパッドインタフェース（図3）などの入力デバイスがシステム接続されている．

4.4 試用結果

実装したシステムを用いて試用実験を行った．毎秒6度の速度で移動するSpeakerを通じて，録音した英語によるインタビューやラジオ番組などを，日常的に英語を使用する人に聞いてもらい意見を聞いた．その結果に基づき，有効なブラウジングツール実現のために解決しなければならない課題を整理した．

● **解決すべき課題1：話題の位置に関する記憶**

“ユーザは音声データ中の話題が再生された位置を覚えていることができる”というのが当初の期待であった．位置を記憶できて初めて，時間-空間マッピングが音声ナビゲーションに有効になるからである．しかしSpeakerが毎秒6度の速度で動く初期システムを試したユーザによると，音声データ中の話題や出来事の位置を思い出すことは難しかった．1つの話題を再生する間にSpeakerが動いてしまうので，ユーザは話題に対応する位置をはっきりと覚えることができなかった．

位置を覚えにくいことの原因は，動きそのものであるかもしれない．動体視力と静体視力が異なるように，聴覚についても動いている音源を聞く場合と静止している音源を聞く場合とでは能力が異なるのかもしれない．そうだとすると，動く音源を使う基本方針そのものを考え直さなければならない．別の原因としてはSpeakerの移動速度が不適切だったためとも考えられる．毎秒6度という速度はSpeakerが動いていることがユーザに分かるように設定されたものだが，その速度は位置を覚えるには早すぎるというのが大半のユーザの感想であった．Speakerの動作の再設計が，本システムを有効にするためには必要であると考えられる．

● **解決すべき課題2：人工立体音場での選択的聴取**

Speakerの数を増やすと，複数の音源から1つの音を選択的に聞き取るのが困難になった．その原因としては，立体音の生成が不適切で音像の空間定位が不正確であったことが考えられる．また特に本システムでは，複数のSpeakerが再生しているのは1つの音声データの複数の部分なので，同じ人の声が複数のSpeakerから同時に再生されていることも多かった．声の違いは選択的聴取の重要な要素であるので[16]，このことも選択的聴取を困難にしたと考えられる．我々がどのように複数の音から1つの音を聞き分けているのかを調べ，人工的立体音場での選択的聴取を容易にするインタフェースを開発することが必要がある．

● **解決すべき課題3：音源位置を特定する際の誤差**

システムが生成する人工立体音場を立体的に知覚できないユーザには，話題の位置を覚えることは不可能である．また，立体的に聞こえるが聞こえる位置に誤差のあるユーザは，間違った位置を記憶してしまう．立体音場を正確に生成するためには，頭部の形状等に基づき音波の伝わり方を表現するパラメータ（頭部伝達関数）を各ユーザごとに適切に設定することが必要である．これは耳や頭部の形状には個人差があり，音の位置の知覚はそれらの形状に関係しているからである．しかし，本システムではすべてのユーザが同一のパラメータを使用したので，システムが意図する位置とユーザが知覚する位置との間にはズレが存在した．このズレを埋め合わせる方策が必要である．

● **解決すべき課題4：音の位置の記憶の解像度**

音声データの位置に関する記憶は，必要な音声データの位置を正確に指示できるほど細かい解像度を持つものではなかった．たとえば“右前”“左斜め前”といった程度の粗い記憶で，けっして“正面から右へ37度”といった細かい解像度のものではなかった．よって，音声データにアクセスするために円形軌道上の位置を指示するとき，指示した位置は所望の音声データの位置に近いことはあっても正確に同じにすることは不可能であった．ユーザが望んでいるであろう位置を推測したり，間違った位置を指示してしまってもそれを修正できるインタフェースが必要である．

● **解決すべき課題5：非直接的な指示インタフェース**

誤差はポインティングデバイスを介して位置を指示するときにも発生する．たとえ正確に音の位置を知覚し記憶できたとしても，音を聞いた空間中の位置を，ポインティングデバイス上の位置に変換する際に誤差が発生する可能性がある．タッチパッドインタフェースの場合，ユーザは半径40インチの円周上のSpeakerの位置を指示するのに，半径1インチのタッチパッド上の点を押さなければならない．音を聞いたその場所を直接指示できるインタフェースが，このような誤差を解消するためには必要である．

5. 実現へのアプローチ：音声提示方法

課題1（話題の位置を覚えにくい）を解決するために，音声データの空間へのマッピング方法つまりSpeakerの動作を再設計した．また課題2（人工立体音場では選択的聴取が困難）を解決するために，微妙な頭の動きを利用したインタフェースを組み込んだ．

5.1 音声データの空間マッピング

初期システムで話題の位置を覚えにくかった原因として，動きそのものとスピードが早すぎたことの2つをあげた．もし動きそのものが原因ならば，再生中は動かない断続的な動き（静止して再生した後，移動して再び再生する）を用いれば問題は解決すると考えられる．スピードが早すぎたのならば，スピードを遅くすることで使いやすくなると考えられる．

これら2つのアプローチの有効性を試すために以下の3通りの動きを比較した．

(a) もとの連続的な動き（毎秒4.8度）

(b) 速度は同じで，断続的な動き（5秒間ごとに1

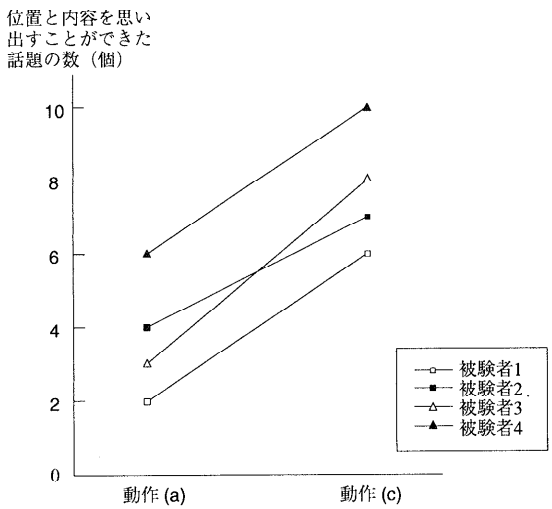

図5 動作 (a) と (c) を用いたときに各被験者が位置と内容を思い出すことができた話題の数

Fig. 5 The numbers of topics that the four subjects could remember the location.

回，毎秒 4.8 度の割合で移動する）

(c) ゆっくりとした連続的な動き（毎秒 1.2 度）

4 人の被験者にこれらの動作を試してもらった．各被験者は 5 分間のニュース番組の録音を上記 (a), (b), (c) の動き方をする Speaker から聞き，それぞれのセッションの後に覚えている話題と位置の関係を解答用紙に書き出してもらった．使用したニュース番組は英語のラジオニュースで，含まれる各話題の長さは概ね 30 秒であった．4 人の被験者は日常的に英語を使用する米国の大学生・大学院生である．

被験者の数が 4 人と少ないので定量的な結果を導くことはできないが，観察された一定の傾向を実験結果としてまとめる．

- 断続的な動作 (b) は，本実験の範囲では明確な効果は見い出せなかった．また，動作 (b) では Speaker の位置が突然大きく動くので，位置を追い続けるのが難しかった．これは Speaker が複数存在する場合にはさらに深刻になると考えられる．今回の実験では 4 人の被験者は途中で位置を覚えることを断念してしまった．
- ゆっくりとした連続的な動作 (c) を使ったときに，最も多くの話題とその位置を覚えていた．**図 5** に動作 (a) と動作 (c) を使った場合に各被験者が内容と位置を思い出すことのできた話題の数を示す．各被験者によって覚えられた数に違いはあるが，すべての被験者が動作 (c) を用いた場合に動作 (a) を用いた場合より多くの話題を思い出すことができた．本実験では，再生される録音素材の内容による影響を避けるために同一の録音を (a), (b), (c) 3 回の試行で使用した．そのため，学習効果によって 1 回目よりも 2 回目以降の方が多くの話題を覚えられる可能性もあった．しかし，動作 (a) を後から試した被験者（図 5 中の被験者 3, 4）ですら，最初に試した動作 (c) の方が多くの話題とその位置を思い出すことができた．また，実験後に感想を聞いたところ速い動作 (a), (b) が被験者にストレスを与えたのに対して，ゆっくりとした動作 (c) は被験者に最も好まれた動作であった．

これらの結果から 3 種の動作の中では動作 (c) が Speaker の動作として最も適当であると考えられる．本実験では，被験者に話題と位置を書き出してもらうのと同時に，どの程度の単位で音の位置を覚えていたかを質問した．人工的立体音をどの程度立体的に知覚したかにもよるが，自分の周囲の全周の概ね 1/4 から 1/12 を単位に覚えていたと答えた．これは，左右というよりは細かいが，10 度単位というよりは粗い解像度である．使用したニュース番組中の話題の一般的な長さは 30 秒であったが，これは動作 (c) では 36 度の範囲にマッピングされる．この角度は円周の 1/10 にあたるから，ユーザが音の位置を覚えている単位角度ごとに 1 つの話題が割り当てられていたことになる．記憶の単位空間ごとに，記憶の単位となる 1 つの話題を割り当てるというのは合理的とも考えられる．よって，動作 (c) を Speaker の動作として採用することとした．

5.2 選択的聴取能力の強化

自然の環境では複数の音源から必要なものを選択的に聴取することができるが，人工立体音場による初期システムではそれは難しかった．本システムでは複数の音源から同じ人の声が再生されていることも難しさの原因と考えられる．ここでは選択的聴取能力を強化するインタフェース，つまりユーザの興味を汲み取って聞こうとしている音を聞きやすくするインタフェースの実現を目指す．まず自然の環境で選択的に音を聞き取ろうとするときの人間の振舞いを観察し，それを基にインタフェースを設計した．

- **観察**

インタフェースを設計するための手がかりを得るために，複数の音源が存在する中で 1 つの音に聞き入るときの人間の振舞いを観察する簡単な実験を行った．本論文のシステムに近い環境を設定するために，3 個のスピーカを被験者を中心とした半径約 1 m の円周上に設置し，録音した会話を各スピーカから再生した．被験者には，指示に従い 3 個のスピーカのうちの 1 つ

図6 左後ろのスピーカの音を聞くために頭を傾けている様子
Fig. 6 Leaning head toward the speaker on the left.

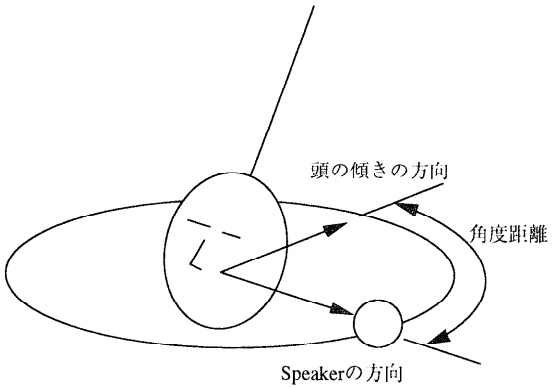

図7 頭の傾き方向とSpeakerの方向の間の角度距離
Fig. 7 Angular distance between the direction of leaning and the direction of Speaker.

を聞き内容を理解して報告するよう依頼した．被験者の振舞いは正面に設置したビデオカメラで撮影した．実物のスピーカを使った実験の後で，ヘッドフォンによる人工立体音場に同様の環境を設定し，振舞いを観察した．

同じような実験が，音源位置を特定するときの頭の動きを観察するために過去に行われている[17),18)]．それらの実験では，音源の位置を見つけるための戦略的な頭の動きが観察されている．この実験では，音源の位置を探すための動きではなく，音源から流れる音声の内容をよく聞き取るための動きに注目した．

まず実物のスピーカを用いた実験から行った．実験に参加したのは4人であったが観察された振舞いは多様であった．体全体を動かしたり手を耳に当てたりして聞き取ろうとする者もいれば，まったく体を動かさずに目を閉じて集中しようとする者もいた．しかしビデオテープを観察すると，体をまったく動かさない被験者でも頭を音源の方向に傾ける動き（図6）が観察された．傾きの方向は必ずしも音源の方向とは一致しないが近い方向であった．

次に行った人工立体音場を用いた実験では，被験者はほとんど頭を動かさなかった．人工立体音場でも頭の位置と方向は立体音生成に反映するようにプログラムされていたが，実験後多くの被験者が“頭を動かしても効果はないと思った”と報告した．人工立体音場では頭の位置や向きを調整することの効果がユーザに分かりにくかったと考えられる．

本実験では被験者が少数だったので“選択的聴取をする際には頭を傾ける”と一般化することはできない．また頭の位置や向きの調整法には個人差があると考えるのが自然である．本論文では人工立体音場で被験者が頭の向きを調整しなかった点に着目し，頭の向きの調整動作に対するシステムからのフィードバックを明確にすることで，聞きやすい位置を探す調整動作を促すインタフェースを構築する．

- **頭の動きを用いたインタフェースの設計**

観察された頭を傾け向きを調整する動作を利用して，人工立体音場内での選択的聴取を容易にするインタフェースを開発した．ヘッドフォンに取り付けられている磁気センサでユーザの頭の傾き方向を計測し，それによって各Speakerの音量を変化させた．各Speakerの音量の変化量はSpeakerの位置する方向とユーザが頭を傾けた方向の角度距離（図7）に比例するように設定し，ユーザが頭を傾ける方向がSpeakerの方向に近ければ近いほどそのSpeakerがよく聞こえるようにした．

音量の変化量は頭の傾き方向とSpeakerの方向の角度距離のみに比例し連続的に変化するので，ユーザは頭の傾きに対する音量の変化に反応しながら頭の向きを調整し，最もよく聞こえる方向を探ることができる．これは，自然の環境で聞きやすい頭の位置を探ってゆく動作に似ていると考えられる．

本システムでは最大8デシベルの音量変化を与えている．自然界では頭の向きを調整してもこれほど大きな変化は起こらない．本システムではこのような誇張したフィードバックを返すことで，人工立体音場では不明瞭であった頭の向きの調整動作による効果をユーザに明確に示し，選択的聴取を容易にした．

6. 実現へのアプローチ：インタラクション

初期システムでは，所望の音声データにアクセスするのに的確な場所を指示するのは困難であった．課題3（音の位置を誤って知覚する），課題4（音の位置の記憶の解像度が低い），課題5（非直接的指示インタフェース）の3点がこれに関連する．正確で細かい指

示を可能とするため3種のインタラクティブなインタフェースを開発した．

6.1 grab-and-move インタフェース

ユーザは音声データの細かな位置までは記憶できず所望の音声データの位置を正確に指示することは困難であった．5.1節で音声データの位置の空間的記憶を容易にするために遅い速度でSpeakerを移動することにした．その結果，単位空間に割り当てられる音声データの量は増加し，音声データへの細かいアクセスはさらに難しくなった．細かい操作を可能にし所望の音声データにアクセスできるようにするために，再生された音声を聞いた後で期待した音声データと違えば再生位置を調整できる“grab-and-move”（つかんで動かす）インタフェースを開発した．

grab-and-moveインタフェースでは，初期システムのインタフェースと同様にユーザは所望の音声データに対応する位置を指示することでデータにアクセスする．指示した位置にすでにSpeakerが存在する場合は，そのSpeakerをユーザの制御に移しユーザが位置を調整できるようにする（つかんだ状態）．つかまれているSpeakerはユーザが分かるように音量が大きくなる．所望の音声データと違うものが再生されたら，ポインティングデバイスを介してSpeakerの位置を動かし調整することができる．動かしている間Speakerはその位置に対応する音声データの小さなセグメントを繰り返し再生するので，ユーザは位置に対応する音声データの中身を聞きながらSpeakerの位置を調整することができる（**図8**）．正しい音声データがSpeakerから再生されたら，つかんだSpeakerを放すことでSpeakerはその位置から通常どおりの移動しながらの再生を始める．Speakerを放す方法は，各インタフェースデバイスによる．指示した位置にSpeakerが存在しない場合には新しいSpeakerを生成し，それをユーザがつかんだ状態にする．

ユーザが指示する位置情報は大雑把なものなので，指示された角度に忠実に再生位置を決定するよりも，話題の切れ目などの位置を選んで再生開始位置とした方が妥当である．本システムでは，前処理によって音声データ中の話題の切れ目である可能性の高い点のリストを生成しておく．再生開始位置を決定する際には指示された位置に最も近い話題の切れ目をそのリストから選択する．前処理プログラムはNewscomm[19]のために開発されたものを用いた．無声部分・話し手の交代・声の大きさなどに基づき音声信号処理によって話題の切れ目と考えられる位置を抽出する．

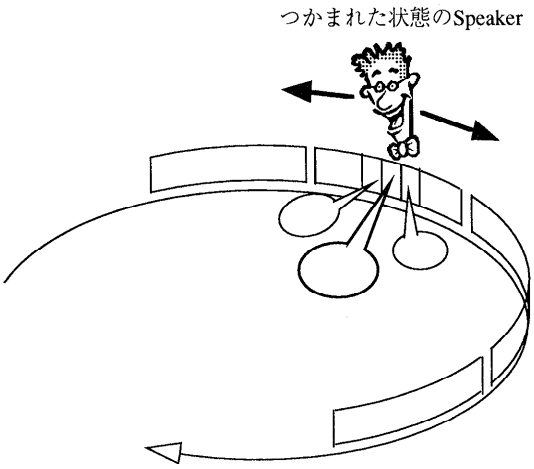

図8 Grab-and-move インタフェース

Fig. 8 Grab-and-move interface: Users can adjust the position of the Speaker after hearing the incorrect audio selection. While the user adjusting, the “grabbed” Speaker repeats the small segment of audio corresponding to the Speaker’s location.

6.2 audio cursor

システムを試用したほとんどのユーザにとって，本システムの生成する人工立体音場は不正確であった．程度の差こそあれ，ユーザが知覚する音源位置とシステムが意図した音源位置との間にはズレが存在した．音像定位をより正確にすることは他の研究に期待することとして，ここでは，存在するズレを埋め合わせる方法を検討した．

本システムでは新しい音オブジェクト“audio cursor”を導入した．audio cursorはユーザによってオンにされている間，特有のリズムを持つノイズ（zebetube.auとして知られる，チューブの中でスプリングが振動する音）を連続的に発し続ける．ユーザはaudio cursorの人工立体音場での位置をポインティングデバイスによって操作し，位置に対する音のフィードバックを得ることができる．そして，ユーザはaudio cursorを移動し音オブジェクトに音響的にオーバレイすることで正確に音オブジェクトにアクセスすることができる（**図9**）．タッチパッドインタフェースを使う場合は，ユーザがタッチパッドに触れるとaudio cursorがオンになり，指を動かしてaudio cursorを移動し，面を強く押すことで音オブジェクトをつかむことができる．

6.3 手による直接的指示インタフェース

音声データが提示される頭の周りの空間中の位置をタッチパッドなどのインタフェースデバイス上の位置に変換する際に生じる誤差を減らすために，手による

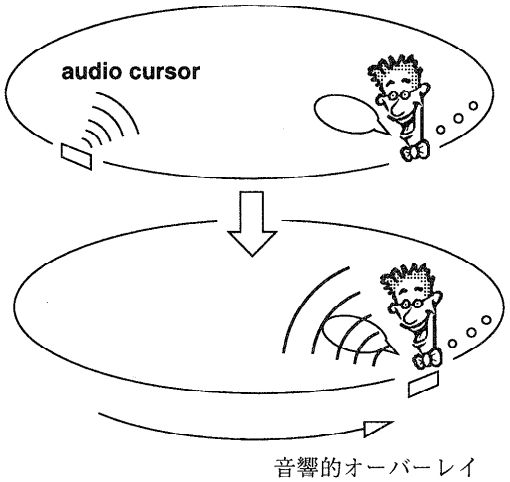

図9 Audio cursorと音響的オーバレイ
Fig. 9 Audio cursor and acoustic overlay.

直接的な指示インタフェース“Point-by-hand”インタフェースを開発した．これはハンドジェスチャによるインタフェースで，ユーザは所望の音声データを聞いた場所を手で直接指し示すことで音声データにアクセスできる．

ハンドジェスチャの入力デバイスとしてMITメディアラボで開発されたFishセンサを用いた．これはセンサが発生する電界にユーザの身体が与える影響を測定するセンサで，非接触で身体の動きを検出することができる[20]．本システムでは**図10**に示すように，Fishセンサの信号発信アンテナをユーザが座る椅子の座面に設置し，4個の球形の信号受信アンテナをユーザの頭上に配置する．このように配置することでFishセンサはユーザの手と4個の受信アンテナそれぞれとの距離を電界の強度として測ることができる．ユーザの体型等による誤差を除くための校正を行った後で，システムはユーザの手の水平面内の位置座標と大まかな高さを計算によって得ることができる．

Point-by-handインタフェースでは，ユーザは手を上げることでaudio cursorをオンにし，手を動かすことでaudio cursorを移動する．Speakerにアクセスするには，audio cursorの位置を所望のSpeakerに合わせてから腕を伸ばすように手を高く上げることでSpeakerをつかむことができる．つかんだSpeakerはユーザが手を下に下ろすまではつかまれた状態が続くので，手を動かすことでSpeakerを移動することができる．

7. 考　　察

各章の実験と同じ被験者に，5章と6章の改善を施した本システムを使用しアンケートに答えてもらった．その結果を**図11**，**図12**にまとめる．見学等で本システムを体験したユーザからの意見と合わせて報告し，考察を加えて今後の方向を示す．

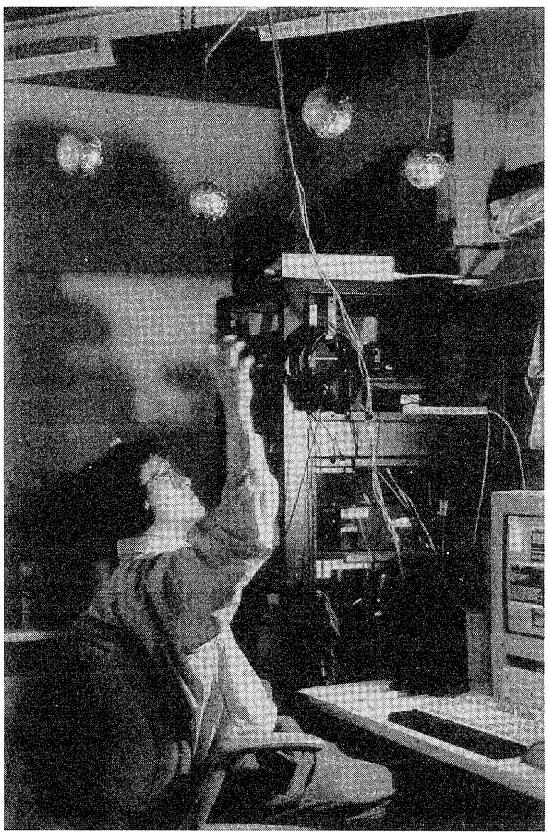

図10 Point-by-handインタフェースを使用している様子
Fig. 10 The “point-by-hand” interface in use: Four metal balls are the receivers. The transmitter is on the chair.

7.1 空間マッピングされた音声データに関する記憶

話題の中身や位置についてほとんど記憶できなかった初期システムとは対照的に，改善されたシステムではほとんどのユーザが空間的な記憶をたよりに音声データにアクセスすることができると報告した（図11質問A）．さらにSpeakerが適当な速度で動いている場合には，空間が音声データの内容を整理格納する棚のような役目を果たし記憶を助けているようであった．これは“シモニデスの記憶の宮殿”[21]に似た効果である．つまり“右前のこの辺で聞いた話題は何であっただろうか”，“その隣は何であっただろうか”のように空間を手がかりに記憶を呼び起こすことで，より多くの内容を思い出すことができる．このような効果が本システムでは音声だけのインタフェースで得られたと

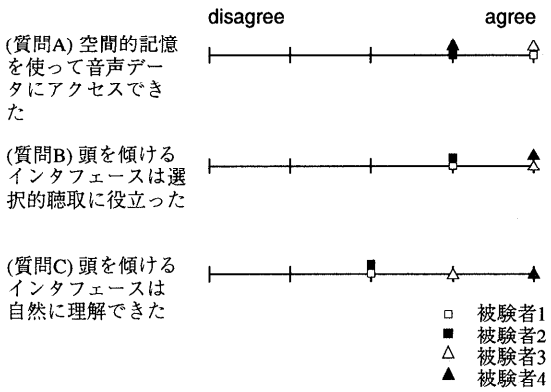

図 11 アンケートの結果
Fig. 11 The result of questionaire.

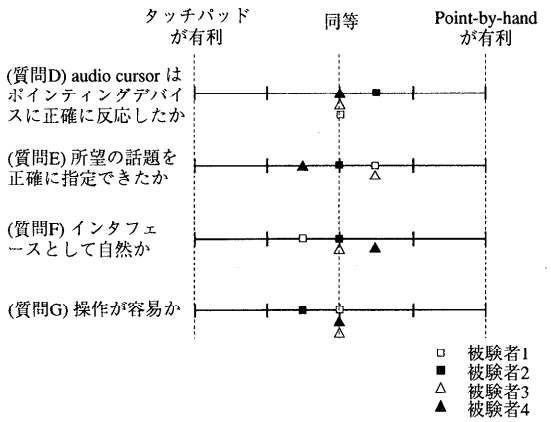

図 12 アンケートの結果（タッチパッドと Point-by-hand インタフェースの比較）
Fig. 12 The result of questionaire (comparison between the touchpad and the point-by-hand interface).

考えられる．

7.2 頭の傾きによる選択的聴取インタフェース

人工立体音場での選択的聴取能力を高めるために頭の傾きを利用したインタフェースを組み込んだ．これは，頭の向きに応じて変化する音の聞こえ方に反応して，よく聞こえる位置へと調整してゆく自然な過程をまねたものであった．ほとんどのユーザが本インタフェースによって選択的聴取がしやすくなったと報告した（図 11 質問 B）．また半数のユーザはその仕組みを自然と理解でき（図 11 質問 C），半数は説明を受けることでこのインタフェースを利用できるようになった．ユーザは頭を微妙に動かすことで素早く別の音源へと注意を切り替えることができるので，複数の Speaker の間で注意を切り替えることによるブラウジングが容易になった．

7.3 Point-by-hand インタフェース

図 12 にタッチパッドインタフェースと Point-by-hand インタフェースを用いた場合の使用感に関するアンケート結果をまとめる．大きな Point-by-hand インタフェースと小さなタッチパッドインタフェースを比較すると，指示の正確さや使いやすさの点では，ほとんどのユーザが同程度と感じた（図 12 質問 D～G）．アンケートに答えなかったユーザの意見も含めると，音像位置と指示位置が同じ Point-by-hand インタフェースが好まれる傾向があった．ただし，この結果には目新しさも影響していると思われる．

小さなタッチパッドインタフェースでは，音空間の位置をデバイス上の位置に変換しなければならないが，マウスなどで慣れているので抵抗感なく操作することができるということであった．

Point-by-hand インタフェースでは，audio cursor を操作する高さと Speaker をつかむ高さが微妙なのでそれを練習しなければならないが，慣れるのは難しくなかったということであった．

7.4 audio cursor

システムの生成する立体音の位置を正しく知覚できないユーザは，audio cursor を動かすことで音源の位置とその聞こえ方の関係を学ぶことができた．また，Point-by-hand インタフェースとあわせて用いた場合には，ユーザの手の位置に重なって audio cursor が提示されることになる．これは audio cursor を手で直接操作しているような感覚を呼び起こし，人工立体音の立体感を増強したと考えられる．

7.5 今後の展開

今回の実験では各話題の長さが 30 秒に固定されたニュース番組を用いたので，一定速度で移動する Speaker は各話題を一定の空間にマッピングすることができた．しかし一般には音声データ中に含まれる各話題の長さはまちまちである．たとえば，小説ならば段落は長いものも短いものもある．音声データの内容に応じて Speaker の移動速度を変化させ各話題を一定の空間にマッピングする方法の開発が，さまざまな音声素材に適応するには必要である．

注目していない Speaker からの音声が聞き取りにくいというユーザもいた．慣れたユーザはときどき頭を動かして素早く各 Speaker が再生している内容をチェックすることで注目していない Speaker からの情報にも注意を払った．このような複数の点に注意を配る動きは，視点を素早く動かしながら本のページ全体を見渡すような動きと似ている．慣れればこのように全体を見渡すことも可能になるが，それでも大切な

話題を聞き逃す可能性はある．それを避けるために，ユーザが注目していないSpeakerから大切な話題が再生されるときには，AudioStreamer[10]で行われたように，ベルを鳴らしたり音量を変化させてユーザの注意を喚起する仕組みも必要である．

8. 結　論

本論文では，音声データの時間軸に空間的にアクセスするインタフェースの実現に向けたデザインの過程を示した．音声データの時間軸を空間にマッピングし，また音声データの複数の部分を同時に再生することにより，会議の録音やラジオのニュース番組のような長い音声データを空間的にブラウジングすることが可能となった．マッピングの方法に関しては，素材に適応するマッピング方法などの研究を進めることが必要であるが，本論文では“1つの話題を記憶の単位角度（全周を4等分から12等分した程度の角度）に割り当てる”という指針を示した．

また数種のインタラクティブなアクセス方法も開発した．grab-and-moveインタフェースによって，細かい操作が可能になり音声データの位置に関する記憶の粗さを補うことができた．audio cursorは人工立体音を知覚するときの位置の誤差を補い，また音像を重ね合わせる操作によって音オブジェクトへの正確なアクセスを可能にした．Point-by-handインタフェースは音オブジェクトが存在する位置を直接指し示してアクセスすることを可能にした．またPoint-by-handインタフェースを用いてaudio cursorを手で操作することで，人工立体音と空間位置の関係をユーザが確かめることができた．頭の傾きを用いたインタフェースは，人工立体音場では困難であった選択的聴取を容易にした．

本論文は，空間的記憶を活用できる音声ブラウジングツールを構築するために音声データの時間軸を空間にマッピングする手法を示した．そして，音声の提示方法やインタフェース方式に改善を加えながら使用可能なシステムを構築することを通じて，本手法の実現可能性を示した．

参考文献

1) Arons, B.: A Review of the Cocktail Party Effect, *Journal of the American Voice I/O Society*, Vol.12, pp.35–50 (1992).
2) Stifelman, L.J.: Augmenting Real-World Objects: A Paper-Based Audio Notebook, *Extended Abstract of CHI '96*, Vancouver, Canada, pp.199–200, ACM (1996).
3) Whittaker, S., Hyland, P. and Wiley, M.: Filochat: Handwritten Notes Provide Access To Recorded Conversations, *Proc. CHI '94*, Boston, MA, pp.271–277, ACM (1994).
4) Mills, A.W.: Auditory localization, *Foundations of Modern Auditory Theory*, pp.303–348, Academic Press (1972).
5) Wenzel, E.M., Wightman, F.L. and Foster, S.H.: A Virtual Display System for Conveying Three-Dimensional Acoustic Information, *Proc. The Human Factors Society* (1988).
6) Cohen, M., Koizumi, N. and Aoki, S.: Design and control of shared conferencing environments for audio telecommunication, *Proc. the International Symposium on Mesurement and Control in Robotics*, pp.405–412, Society of Instrument & Control Engineers (1992).
7) Seligmann, D.D., Mercuri, R.T. and Edmark, J.T.: Providing Assurances in a Multimedia Interactive Environment, *Proc. CHI '95*, Denver, CO, pp.250–256, ACM (1995).
8) Pitt, I.J. and Edwards, A.D.N.: Pointing in an Auditory Interface for Blind Users, *Proc. 1995 IEEE International Conference on Systems, Man and Cybernetics*, pp.280–285, IEEE (1995).
9) Arons, B.: SpeechSkimmer: Interactively Skimming Recorded Speech, *Proc. UIST '93*, pp.187–196, ACM (1993).
10) Schmandt, C. and Mullins, A.: AudioStreamer: Exploiting Simultaneity for Listening, *Extended Abstract of CHI 95*, Denver, CO, pp.218–219, ACM (1995).
11) Kobayashi, M. and Schmandt, C.: Dynamic Soundscape: Mapping time to space for audio browsing, *Proc. CHI '97*, Atlanta, GA, pp.194–201, ACM (1997).
12) Mandler, J.M., Seegmiller, D. and Day, J.: On the coding of spatial information, *Memory & Cognition*, Vol.5, No.1, pp.10–16 (1977).
13) Schulman, A.I.: Recognition memory and the recall of spatial location, *Memory & Cognition*, Vol.1, No.3, pp.256–260 (1973).
14) Kobayashi, M.: Design of Dynamic Soundscape: Mapping time to space for audio browsing, Master's thesis, Massachusetts Institute of Technology, Cambridge, MA (1996).
15) Loomis, J.M., Hebert, C. and Chcinelli, J.G.: Active localization of virtual sounds, *Journal of Acoustical Society of America*, Vol.88, No.4, pp.1757–1763 (1990).
16) Cherry, E.C.: Some experiments on the recognition of speech, with one and two ears, *Journal of the Acoustic Society of America*, Vol.25,

pp.975–979 (1953).
17) King, W.J. and Weghorst, S.J.: Ear Tracking: Visualizing Auditory Localization Strategies, *Extended Abstract of CHI '95*, Denver, CO, pp.214–215, ACM (1995).
18) Thurlow, W.R., Mangels, J.W. and Runge, P.S.: Head Movements During Sound Localization, *Journal of the Acoustical Society of America*, Vol.42, No.2, pp.489–493 (1967).
19) Roy, D.K.: NewsComm: A Hand-Held Device for Interactive Access to Structured Audio, *Proc. CHI '96*, Vancouver, Canada, pp.173–180, ACM (1996).
20) Zimmerman, T.G., Smith, J.R., Paradiso, J.A., Allport, D. and Gershenfeld, N.: Applying Electric Field Sensing to Human-Computer Interfaces, *Proc. CHI '95*, Denver, CO, pp.280–287, ACM (1995).
21) Yates, F.A.: *The Art of Memory*, Routledge & Kegan Paul (1966).

(平成9年 7 月3日受付)
(平成9年10月1日採録)

小林 稔（正会員）

1964年生．1988年慶応義塾大学計測工学科卒業，1990年同大大学院修士課程修了．1996年マサチューセッツ工科大学修士課程修了．1990年日本電信電話株式会社入社．NTTヒューマンインタフェース研究所にて，映像通信とコンピュータによる協同描画システムを中心としたCSCWの研究，コミュニケーションシステムのための空間的ヒューマンインタフェースの研究に従事．電子情報通信学会，ACM，IEEE Computer Society 各会会員．

Chris Schmandt

Mr. Schmandt is a Principal Research Scientist and the director of the Speech Research Group at MIT's Media Laboratory. He has been at the Media Laboratory since its creation 5 years ago, and had spent the previous 5 years at its predecessor, the Architecture Machine Group. His research covers a broad range of conversational computer systems, with current emphasis on audio in portable computing and techniques for scanning large quantities of stored speech. He holds BS and MS degrees from MIT. He is the author of "コンピュータとのヴォイスコミュニケーション未来のコンピューティングに向けて"（サイエンス社，原著は Van Nostrand Reinhold 社）. He is a member of ACM and IEEE, and a member of the Board of Directors of American Voice Input/Ouput Society.