社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE
電子情報通信学会 EA 研究会 2005-6-24

仮想の“音の部屋”によるコミュニケーション・メディア *voiscape* のための音声 3D 化と残響の計算

金田 泰
日立製作所 中央研究所
〒185-8601 東京都国分寺市東恋ヶ窪1-280
E-mail: kanada@crl.hitachi.co.jp

あらまし 3D 音響技術によってつくられた仮想的な“音室”内を移動して相手を選択しつつ会話ができるコミュニケーション・メディア voiscape を開発している．Voiscape の第 2 のプロトタイプ VPII においては，FIR 法によって低遅延な HRTF フィルタ計算をおこなうとともに，移動可能な範囲としての音室を音響計算上の部屋とみなし，その壁による初期反射をシミュレートした．この初期反射によって音の頭外定位と距離感の表現を可能にした．また，ユーザの移動を追跡し必要な補間処理をおこなった．これによって，話者識別が容易で，複数の会話コンテクストが共存することができ，また音室内の移動が自然でノイズがすくない音声コミュニケーション環境を実現した．
キーワード IP 電話，音声通信，音声会議，3 次元オーディオ，3D 音響，Voiscape．

Computation of Spatialization and Reverberation For A Virtual “Sound Room” Based Communication-Medium Called *voiscape*

Yasusi Kanada
Central Research Laboratory, Hitachi, Ltd.
Higashi-Koigakubo 1-280, Kokubunji, Tokyo 185-8601, Japan
E-mail: kanada@crl.hitachi.co.jp

Abstract We are developing a communication medium called voiscape, which enables taking to people while selecting persons to talk by moving in a virtual “sound room”. In the second prototype of voiscape called VPII, the FIR Method is used for low-delay HRTF filtering, the sound room — the range of motion — is identified with the room in acoustical calculation, and early reflections by the sound room walls are simulated. The early reflections produce out-of-head sound localization and sound distance expression. We also implemented motion-tracking and interpolation algorithms into the spatialization method. VPII enabled a voice communication environment, in which speaker identification is easy, multiple conversation-contexts can be created in a room, and motions of users and objects in a sound room are natural and causes only small noises.
key words IP telephony, Voice communication, Audio conferencing, Spatial audio, 3D sound, Voiscape.

1. はじめに

人間どうしのコミュニケーションの基本は音声による会話である．現在もっとも人気がある音声コミュニケーション・メディアは電話であるが，電話はけっして理想的なメディアであるとはいえない．なぜなら，第 1 に，電話では基本的には 1 対 1 でしか会話することができない．第 2 に電話においてはスピーカが 1 個しかないため，耳が 2 個あることによる人間のすぐれた聴覚能力を一部しかいかすことができない． 第 3 に，電話は会話するあいだだけ相手と接続し，会話がおわると切断してしまうために，相手が電話してもよい状態にいるかどうかもわからないし，切断されているあいだに重要なことがおこっても意識的に伝達 (電話) しないかぎりはつたえられない．

これらの欠点のおおくは電話のかたいインフラとくにネットワークの制約からうみだされている．電話のネットワークが IP ネットワークによって置換されようとしているいまこそ，これらの欠点をなくした，あたらしいメディアを開発するべきときである [Kan 03]．電話にかわるべきあたらしいメディアを報告者は voiscape と呼んでいる．報告者が予想する voiscape のすがたはつぎのとおりである．第 1 に，voiscape においては多対多の自然な会話が可能になる．すなわち，voiscape は基本的に会議メディアである．第 2 に，両耳で音声をきくことにより人間の聴覚能力をいかすことができ，話者識別や複数の会話のききわけなどが可能になる [Kan 05]．多対多の自然な会話が可能になるのは，単にインフラがそれに対応したからではなく，両耳で音声をきくことによって方向感・距離感がえられ，いわゆるカクテルパーティ効果 [Che 53] がえられるからである．第 3 に，voiscape においては意識的な接続・切断は不要である．これは，IP ネットワークはパケット交換ネットワークであって常時接続が基本だという利点をいかすものである．

Voiscape においては，電話のように特定の相手と接続して話をするのではなく，音で仮想的な部屋 (音室とよぶ) を表現し，そのなかで会話する．ユーザは音室のなかで自由に移動することができ，音室内で進行している複数の会話やストリーミング再生のなかから，すきなものにちかづいて，きいたり，会話したりすることができる．また，音室内でちかくにいるひとの声はおおきくはっきり，とおくにいるひとの声はちいさくぼやけて，それぞれの方向からきこえる．

さらに，音室は複数存在し，あらかじめ使用権をえている音室のなかからすきなものを選択することができる．

Voiscape の最初のプロトタイプ [Kan 03][Kan 04b] においては Java のライブラリ JMF, Java3D を使用してこのような環境をつくることをこころみたが，音質や遅延の点で満足できるプロトタイプをつくることができなかった [Kan 04a]．そこで，C++ をつかって Linux 上に第 2 のプロトタイプ VPII を開発した．

携帯性を実現するため，voiscape においては音声をスピーカーによる多チャンネル再生ではなくバイノーラル再生するのが原則だとかんがえている [Kan 03]．バイノーラル再生を実現するためのもっとも容易な方法はヘッドフォンまたはヘッドセットを使用する方法であり，VPII においてもヘッドセットの使用を原則としている．

VPII においては，地図とカーソルキーによって仮想の場所における選択的にコミュニケーションを可能にするユーザインタフェースや，SIP 拡張のイベント通知機構 SIMPLE (SIP for Instant Messaging and Presence Leveraging Extensions) にもとづく音室管理法といった技術を開発した [Kan 05]．また，低遅延であり，初期反射のシミュレーションにより音の頭外定位と距離感の表現を可能にし，さらにユーザの移動を追跡し必要な補間処理をおこなう 3D 音響技術を開発したが，ここではそれについて報告する．

第 2 章では開発したプロトタイプの構成をしめす．第 3 章では VPII が使用している HRTF (Head Related Transfer Function) について説明する．また，音を頭外に定位させ距離感 (distance cue) をつくるおおきな要因は残響だといわれているので，それについて第 4 章において説明する．Voiscape においてはユーザが自由に仮想空間内を移動できるため，音源と聴取者自身の両方の移動にともなって 3D 音場が動的に変化する．自然な 3D 音場の変化を実現するにはくふうが必要だが，それについて第 5 章で説明する．第 6 章において結果をまとめ，最後に結論をのべる．

2. プロトタイプの構成

voiscape の第 2 プロトタイプ VPII (Voiscape Prototype II) の構成を簡単に説明する．VPII の全体に関するより詳細な説明は Kanada [Kan 05] が記述している．

2.1 全体構成

Voiscape のための典型的なアーキテクチャとして分散型と集中型とがある．第 1 プロトタイプは分散型にちかい構成をとったが，VPII は集中型の構成をとる．集中型構成においては voiscape システムはつぎの各要素によって構成される．

- **ユーザエージェント**: ユーザが使用する端末としては PDA (Linux 版の Sharp Zaurus) または Miscosoft Windows を搭載した PC を使用し，通信には無線 LAN (IEEE802.11b) を使用する．端末に搭載される端末ソフトウェアであるユーザエージェントはメディアサーバとのあいだで音声を送受信するとともに，音室管理サーバや音室リストサーバとセッション制御メッセージ等を交換する．当面は標本化周波数として電話程度の再生帯域を実現する 8 kHz, コーデックとして ITU-T G.711 を使用する．
- **管理サーバ群**: 管理サーバ群は音室管理サーバ，音室リスト管理サーバ，SIP レジストラなどによって構成される．ユーザは複数の音室のなかから 1 個を選択して入室するが，この制御には SIP (Session Initiation Protocol) を使用する．また，ユーザエージェントはユーザの音室内での位置や方向などの情報をつねに音室管理サーバに送付するが，そのためには SIP 拡張である SIMPLE を使用する．また，音室リストサーバは音室リストを管理し，音室の生成・抹消などに関与する．
- **メディアサーバ**: 第 1 プロトタイプにおいては音声は端末間で直接 VoIP (Voice over IP) 通信によって伝達したが，VPII においては 3D 化とミキシングを集中的におこなうため，メディアサーバを介して通信する．メディアサーバは管理情報を音室管理サーバからうけとり，それにしたがって各ユーザエージェントとのあいだで音声通信をおこなう．メディアサーバについては次節においてさらに説明する．

2.2 メディアサーバの構成

メディアサーバの機能と構造の概要をのべる (**図 2.1** 参照)．

VPII のメディアサーバは，音室内の各ユーザエージェントから 1 チャンネルの音声を VoIP によって入力し，音声 3D 化とミキシングとをおこなった結果の 2 チャンネル (バイノーラル) 音声を VoIP によって各ユーザエージェントに出力する．プロトコルとしては RTP (Real-time Transport Protocol) を使用する．出力先のユーザごとにことなる音声 3D 化の処理が必要なので，音声 3D 化だけでなくミキシングも出力先のユーザごとにおこなう．そのため，ユーザ数が n であり，すべてのユーザが接続されているときには，音声 3D 化は n ($n-1$) 回，ミキシングは n 回おこなう (図 2.1)．

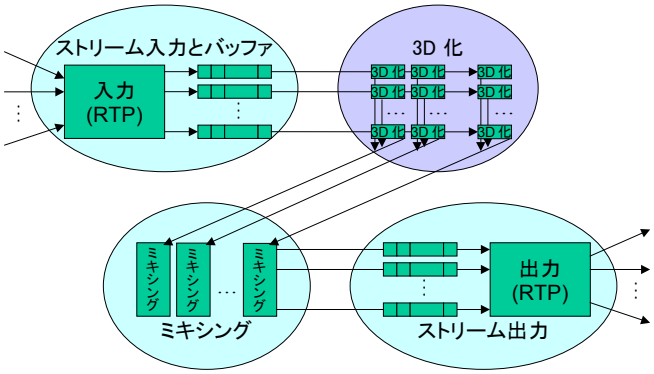

図 2.1 メディアサーバの構造

メディアサーバは音室管理サーバからユーザの入退室や移動に関する情報をうけとり，それらにしたがって音声 3D 化とミキシングとをおこなう．すなわち，ミキシングするかどうかはユーザが音室内にいるかどうか，またミキシングするべき距離にいるかどうかなどの状態によって決定され，また音の方向や距離はユーザの移動によって動的に変化する．たとえば，あらたにユーザが入室したときやポリシーによってきめられた通信可能な距離にはいったときには，そのユーザエージェントから入力される音声が他のユーザエージェントに出力される音声にそれぞれミキシングされる (すなわち，ミキシングする音声の数が増加する)．また，ユーザが退室したときやポリシーによってきめられた通信可能な領域からでたときには，逆にその音声が他のユーザへの音声にミキシングされないようにする (ミキシングする音声の数が減少する)．ユーザの移動は SIP (Session Initiation Protocol) によって間欠的に通知される [Kan 05]．

以下，音声 3D 化とミキシングに関する部分についてのべる．この部分においては入力された音声を音室管理サーバからつたえられたユーザの位置や方向の情報を使用して音室内に位置づける．音声 3D 化部は 1 チャンネル (モノーラル) 音声を入力し，2 チャンネル (バイノーラル) 音声を出力する．処理は基本的には 1 パケット (20 ms 分の音声データ) ごとにおこなうが，HRTF (Head-Related Transfer Function) や残響の計算において遅延が導入されるため，前回の計算において遅延されたデータを保管して使用する．ミキサは 3D 化された複数の音声をあわせて 2 チャンネルの単一音声

にする．音室管理サーバから指定されたくみあわせで音声をミキシングする．生成された音声はかならずしもただちにユーザエージェントに出力できないので，いったん出力バッファにためる．

3. 頭部伝達関数

Voiscape においては仮想空間内の音源の方向と距離を表現する必要がある．方向を表現するパラメタとしては，1960 年代あるいはそれ以前から ITD (interaural time difference) および ILD (interaural intensity difference) がつかわれたが，その後はより正確に方向が表現ができる頭部伝達関数 (HRTF, Head Related Transfer Function) がよく使用されてきた．VPII でもこれを使用している．

3.1 HRTF による 3D 化の計算法

HRTF またはその時間領域表現である HRIR (Head Related Impulse Response) を使用してある音源からのモノーラル信号を 3D 化するには，音源の方向によってことなる，すなわち方位角ごと仰角ごとにことなる HRTF (または HRIR) を選択し，ちょうどその方向のものがなければ補間をおこない，そうしてえられた HRTF に元信号とを入力して複数のチャネルに関してフィルタリング計算をおこない，結果としてえられた複数チャネルの信号をヘッドフォンまたはスピーカーによって再生する．

HRTF または HRIR のデジタル・フィルタとしての表現法およびその計算法としてはつぎの 3 つの方法がある: 1) FIR (有限インパルス応答) の時間領域における計算，2) FIR の周波数領域における計算，3) IIR (無限インパルス応答) による計算．1) においては，時間領域においてたたみこみ計算をおこなうとフィルタ長の 2 乗の計算量を必要とする．そのため，標本化周波数が 8 kHz のときはよいが音楽再生で通常使用される 44.1～48 kHz にすると膨大な計算を必要とする．これに対して 2) においては，データ長を n とすると $n \log n$ に比例する計算量ですむ．そのため，オーディオ再生用にはこの方法が多用されている．しかし，フーリエ変換は時間を捨象するため周波数領域において信号を加工すると容易に因果律に反する効果がとりこまれる．また，遅延をさけることが困難である．3) はこれらの問題がなく，リアルタイム性がとくに重要な voiscape には適しているが，設計がむずかしい．そこで，1), 2) のうちで遅延がすくない 2) がよいと判断して，当面これを採用することにした．

3.2 HRTF の測定結果とその利用法

HRTF の測定には相当な時間がかかるため，測定になまみの人間を使用すると苦痛をあたえる．また，特定の頭部・耳殻などの形状に依存しない結果をえるためには人間をつかった測定はかならずしも適切ではない．そのため，おおくの研究において HRTF はダミーヘッドを使用して測定されてきた．その代表例が Gardner [Gar 94a] が MIT メディアラボにおいておこなった測定結果である．音響測定用のダミーヘッド・マイクロフォンとしては KEMAR (Knowles Electronic Manikin for Acoustic Research) とよばれるものがもっとも有名であり，Gardner も KEMAR を使用している．VPII においてはこのデータを使用している．現在はそのなかのダミーヘッドによる測定結果を使用しているが，よりたかい臨場感をえるためには個人差を HRTF に反映させる必要があり，そのためには各被験者の測定結果をうまくとりいれる必要があるとかんがえられる．

3.3 プロトタイプにおける HRTF の計算法とその分析

VPII においては CIPIC データベースにふくまれているダミーヘッドによる標本化周波数 44.1 kHz による測定結果 (HRIR) にチェビシェフ・フィルタをかけてダウンサンプリングし，8 kHz における HRIR をえて使用している．この方法においては波形が保存されずながく尾をひく波形に変換されるが，周波数応答を優先した．

もとの測定結果には仰角をかえた測定結果もふくまれているが，音源が水平方向にあるときのデータだけを使用している．このデータにおいては方位角が 5° ごとに測定されている．音源の方位角がこれらの方位からはずれているときは補間をおこなうことがのぞましいが，補間のアルゴリズムはかならずしも単純でなく，リアルタイムに適用しやすい比較的単純なアルゴリズムを使用するとかならずしも正確な結果をえることができない．また，方位角は比較的こまかく 5° ごとに測定されている．そこで，現在は補間をおこなわず，方位角を 5° ごとに量子化して，HRIR をもとのまま使用している．

VPII においては，従来の電話や会議システムのおおくが 8 kHz の標本化をおこない，コーデックとしてとくに G.711 を多用していることと，携帯電話などのモバイルネットワークは遅延や QoS を犠牲にせずに広帯域化するのが困難であることなどから標本化周波数を 8 kHz とした．しかし，もし満足な音声 3D 化ができないならば，それをたとえば 22.05 kHz や 24 kHz に変更する必要がある．

HRTF はひろく使用されているが，標本化を 8 kHz でおこなっているものはすくない．その理由としては，HRTF の主要な用途が音楽再生であり，そのためには通常 44.1 kHz 以上の標本化周波数が使用されるという理由もある．しかし，音源の方向感をえるうえでは 4 kHz をこえる周波数の音が重要であり，それが再生できない 8 kHz の標本化周波数では HRTF の目的が十分に達せられないという理由がおおきいとかんがえられる．たとえば，背後からくる音は 8 kHz 付近に (HRTF に由来する) 谷があるといわれるが，この谷を再生するには 16 kHz をこえる標本化周波数が必要である．ただし，Begault らによる遠隔会議などのための狭帯域の 3D 化を効果的にする方法を追求した研究 [Beg 99] もある．

4. 距離の表現と残響

4.1 頭内定位の解消と距離感付与

音源の距離を表現するには，まず距離による音の減衰を表現する必要がある．残響がなければ音の振幅は音源からの距離に反比例する．音のエネルギーは距離の 2 乗に反比例する．

また，音は常温で約 340 m/s の速度でつたわるので，たとえば音源からの距離が 10 m なら 30 ms の遅延がある．これは認知可能なので，ばあいによっては遅延をシミュレートする必要がある．しかし，WAN を使用した音声通信においては通常 100 ms をこえる遅延があるので，それにくらべると通常会話する距離における遅延は十分にちいさく，遅延を増加させる意味はないとかんがえられる．

距離を表現するために減衰や遅延をシミュレートしたとしても，ヘッドフォンによる音は認知的には通常，頭内に定位する．この音を頭外に定位させるには残響のシミュレーションが必要である．

すなわち，残響のシミュレーションは音楽をよりよくきかせ，部屋の雰囲気をだすためにつかわれるが，VPII においては頭内定位を解消して音に距離感をあたえる目的で残響を付加する．残響によって仮想音源の距離が表現されることは，たとえば Shinn-Cunningham [Shi 00a] が実験的に確認している．また，音現の距離が残響のある環境と無響環境とでは，前者のほうが 2.3～3.8 倍ながく認知されることを Begault [Beg 92] が実験的に確認している．

室内においては直接音が音源からの距離に反比例して減衰するのに対して，残響は音源からの距離によらずほぼ一定である．そのため，音源からの距離が増加するにつれて間接音と直接音との比 (R/D ratio [Beg 00]) は増大する．この R/D 比が人に音源の距離に関する感覚をおこすとかんがえられている [Bro 99]．しかし，実空間の R/D 比を完全にシミュレートするのがかならずしもよいわけではなく，Gardner [Gar 99] によれば，経験的には 3D 音響では

残響を(実空間よりおおきい) 10 dB 減衰させるのがよいという.

いずれにしても, 残響の量や特性はは部屋ごとにことなり, R/D 比も部屋によってことなるので, それらがもし固定的に距離の感覚にむすびついていると仮定すると, 正確に距離を把握できないことになる. Shinn-Cunningham [Shi 00b] は, ひとがそれをおぎなうために学習をおこなっていることを実験的に確認している. Voiscape においてはグラフィカル・ユーザ・インタフェースにおいて距離を把握することができるので, ひとがそこから距離を学習する余地がある. この点の追究は今後の課題である.

4.2 残響の構造

室内における残響はつぎの 2 つの部分からなりたっているといわれている (図 4.1 参照) [Gar 94b].

- **初期反射** (early reflection): 室内では, 直接音がきこえたあと数 ms から 100 ms くらいのあいだに, 条件によっては, 壁, 天井, 床などからの数 10 個の反射を他の音から分離してきくことができる. これが初期反射である. 部屋の形状が直方体であれば 1 回反射は 6 個だけだが, より複雑な形状または家具などがある部屋においては反射音の数がふえ, また壁などで複数回反射した音もきこえる.
- **後期残響** (late reverberation): 直接音がきこえてから 150 ms 以上すぎたころには, 音は多数回反射し, 反射音の数もふえているため, もはや個々の音をくべつしてきくことはできない. また, 音は等角反射するだけでなく壁・天井などで散乱されるため, 残響の構造はさらに複雑になる. これらによって構成されるのが後期残響である. このような後期の残響は, 方向・位相がランダムで指数関数的に減衰する音によってをモデル化される.

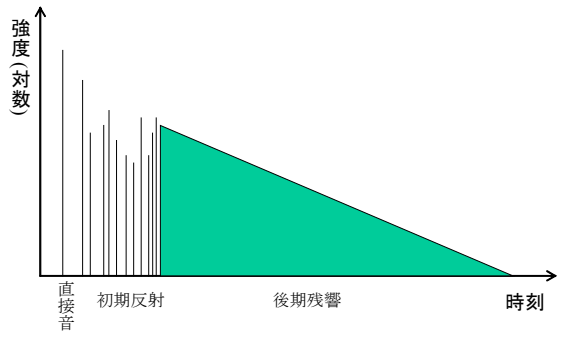

図 4.1 室内残響の構造

Begault ら [Beg 01] は, 方向感の正確さ (azimuth error) についても頭外定位についても, 初期反射だけの残響と後期残響まであわせた完全な残響とのいずれにも明確な効果があり, かつそれらを比較して効果にほとんど差がないことを実験結果としてえている. また Griesinger [Gri 00] によれば, 個別の音のひろがりは直接音がきこえてから 50 ms のあいだにかなりきまり, 50 ms から 150 ms のあいだの音は, ひとはエネルギーとしては感じるがその時刻や方向などを変化させても鈍感だという. しかし, 一方で, 初期反射は方向感をにぶらせるともいわれる.

4.3 プロトタイプにおける残響計算

150 ms 以降の部分すなわち後期残響は直接音からは分離してきこえるので, それはへやのひろさなど音環境に関する感覚をあたえる [She 82] とかんがえられるが, 基本的には直接音の属性をきめるものではないとかんがえられる. また, 残響がおおきいと音声の明瞭度が低下するといわれるが, 明瞭度を低下させるおもな原因は後期残響だとかんがえられる. Voiscape においては部屋の残響をシミュレートすることが目的ではなく, 音声に方向感と距離感とをあたえるのが 3D 音響を採用した目的であるから, 初期反射と後期残響とを独立に制御できる上記のモデルを採用するならば, 後期残響はなくすかまたは最低限におさえるのが適切だとかんがえられる. そのため, VPII においてはつぎのような方針をとった.

- **VPII においては後期残響をとりいれず, 初期反射だけをとりいれる.**

初期反射の計算法としてはつぎの 3 つをはじめとして, さまざまな方法がある.

- Image source method [All 79]: 部屋の壁, 天井, 床を鏡面とみなし, 反射音を鏡面の反対側にある音源の像からの音として計算する方法である. この方法は部屋の面における乱反射がすくないときには適している.
- **光線追跡法** (ray tracing method) [Kro 68]: 音が進行する直線をたどりながら計算する, グラフィクスにおける光線追跡法とおなじ方法である. 光線追跡法は 乱反射があるときは反射音それぞれの直線をたどるため, 計算量がおおきい.
- **光束追跡法** (beam tracing method): 光線追跡法と同様に音の進行する方向に計算をすすめるが, 線の束ごとに計算をおこなう. そのため, 光線追跡法よりすくない計算量でより正確な計算ができる可能性がある.

これらの方法はたとえば Funkhouser [Fun 03] がサーベイしている.

これらの方法は, 部屋のおおきさや形状にもとづいてできるだけ正確なシミュレーションをおこなうことをめざしている. VPII においても部屋のおおきさや形状をシミュレートしているが, これによって移動範囲としての部屋と残響計算のための部屋とを一致させたことが VPII の音声 3D 化法のひとつの特徴である.

しかし, 部屋のおおきさや形状にもとづくシミュレーションが認知的に効果をあげるのかどうかはほとんど実験的にたしかめられていないようである.[1] EAX (Environmental Audio Extensions) [Cre 01] をはじめ, 音声 3D 化をおこなうおおくのシステムにおいては, 部屋の形状やユーザの位置などの情報をあたえないため, 部屋の形状やその中での位置は残響の計算において考慮していない. しかし, もし部屋のおおきさや形状をシミュレートするのが効果的であるなら, voiscape においてとじた空間を使用することはコミュニケーションの観点だけからでなく音響心理上も重要だということになる.

VPII においては 2 次元の image source 法を使用している. すなわち, 天井と床は無反射だと仮定し, 直方体の形状をした音室の 4 つの壁による 12 個の反射を計算している (図 4.2). 図 4.2 においては中央に本来の音室があり, その周囲にその音室の 12 個の鏡像がえがかれている. これらの鏡像のそれぞれのなかに音源の像があるが, そこからの音が聴取者に直進するとして, この音像からの距離と方向をもとめる. ただし, 壁の

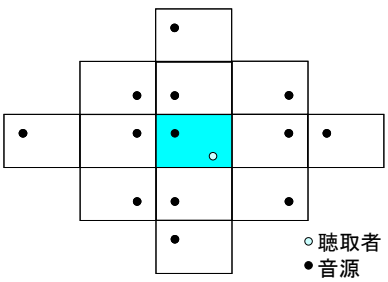

図 4.2 2 次元鏡像法による初期反射の計算

[1] 残響に関する研究のおおくは音響心理ではなく建築音響の観点からなされていて, その評価において心理実験をおこなっていない. そのなかで心理に重点をおいた Shinn-Cunningham らの研究は重要だが, まだその成果はまだかぎられている.

反射率を α ($0 \le \alpha \le 1$)とすると，壁で n 回反射される音の標本には α^n を乗じる．鏡像を 12 個にとどめた理由は，辺が 10 m 以上の音室においてはこの 12 個以外の鏡像からの音が聴取者にとどくのは直接音がとどいてから 50～100 ms 以上たってからであって初期反射の時間をこえていることと，反射回数と距離が増大するため減衰がおおきいこととである．ただし，よりちいさい音室においては，本来は初期反射の時間内によりおおくの反射が聴取者に到達することになる．

反射率 α の値の決定において考慮するべき点を列挙する．

- α をおおきくして十分な R/D 比をえることによって，距離を感じるのに十分な残響がえられるようにする．
- α が過大なために方向感がにぶらされることがないようにし，距離による音の減衰をへらす (間接音をふやす) ことにより距離感が減少することがないようにする必要がある．
- α が過大なために音声の明瞭度低下や不自然さが生じることがないようにする．

反射音はそれぞれことなる方向から人頭に達するため，直接音とはことなる HRTF を適用する必要がある．しかし，多数の反射音にそれぞれことなる HRTF をたたみこみ計算 (または FFT と乗算) すると膨大な計算が必要になる．それをさけるため，VPII においては反射音の計算にはその方向にかかわらず正面に音源があるときの HRTF を適用し，ITD, IID を計算して，左右の耳に達する音の差を表現している．この方法によって，計算量は直接音と同程度におさえながら，反射音に方向をあたえることが可能になった．

5. 動的変化への対処

この章では，ユーザの移動によって発生する問題を分析し，VPII におけるその解決策をしめす．

5.1 動的変化によって発生する問題と従来の解決法

被験者や音源が移動すると，つぎのような問題が生じうる．

1. **音量の急激な変化によるクリックノイズの聴取**: 音源との距離が急に変化して音量や遅延が急に変化すると，クリックノイズがきかれる．とくに，VPII においては位置情報が間欠的につたえられるだけなので，位置情報をうけとったときに急にユーザの位置を変更するとノイズが発生することになる．
2. **方向の急激な変化による喪失**: 音源の方向が急激に変化すると，移動後の音源がもとはどこにあったものかがわからなくなる．

第 1 の問題を解決するには，補間によってユーザどうしの距離や距離に依存する音量と遅延とを急に変化させないようにすればよい．また，第 2 の問題を解決するには，やはり補間によって方向が急激に変化しないようにすればよい．すなわち，いずれの問題もユーザ間の相対的な位置の変化を補間するとともに，音量や遅延を補間することによって解決することができる．

仮想音場の変化にともなう音量と遅延の補間に関しては Savioja [Sav 99] が言及している．Savioja らによる仮想音場システム DIVA においてはいずれも線形補間をおこなっている．遅延の補間においては遅延時間の変化にともなって標本が不足したり過剰になったりするが，標本を複写したり廃棄したりして対応している．この補間によってドップラー効果が生じるが，それは生じるべき効果である．しかし Savioja は詳細な点には言及していないので，VPII においては補間法を新規に考案した．その方法を以下の節でのべる．

5.2 ユーザ位置・方位角の補正

ユーザ位置と方位角の補正に関して説明する (**図 5.1** 参照)．音を聴取するユーザ l (以下，局所ユーザとよぶ) と対話相手のユーザ r (以下，遠隔ユーザとよぶ) の両方の位置を補正する．この補正によって 3D 化をおこなう時刻は変化しない．したがって，補正はみかけ上の移動を遅延させるようにはたらく．補正前の位置を時刻 t の関数として $x(u, t)$ ($u = l$ または $u = r$) とあらわし，補正後の位置を $x'(u, t)$ とあらわし，補正前の局所ユーザの方位角を $\theta(u, t)$，補正後の局所ユーザの方位角を $\theta'(u, t)$ とする．時刻 t は連続値をとることができるが，ユーザ r に関する 3D 化開始時の時刻を t_i (i = 1, 2, ...) とし，x' はこれらの時刻においてだけ定義する．3D 化は約 20 ms 間隔で実行されるので，$t_i - t_{i-1}$ は約 20 ms となる．

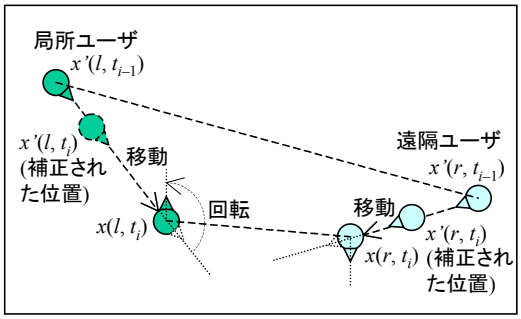

図 5.1 ユーザ位置の補正

時刻を t_i において補正をおこなうのは位置の変化 $x(l, t_i) - x'(l, t_{i-1})$ または $x(r, t_i) - x'(r, t_{i-1})$ の絶対値がある一定値 (現在は 0.1 m) よりおおきいか，局所ユーザがおおきく回転したとき，すなわち $\theta(l, t_i) - \theta'(l, t_{i-1})$ の絶対値がある一定値 (現在は $\pi / 72$) よりおおきいときだけである．

補正後の位置・方位角は前回の補正位置・方位角から現在の位置にむかって移動または回転させることによってもとめる (図 5.1)．変化がおおきくなりすぎないように移動量と回転量をおさえているが，あまりおさえすぎると遅延がおおきくなるので，極端におおきな変化がないかぎりは数 10 回以下 (数 100 ms 以下) で補正されるようにしている．なお，VPII においては音源にすべて点音源すなわち指向性のない音源を使用しているため，補正計算において音源の方位角は考慮する必要がない．

なお，遠隔ユーザの位置も局所的に計算するので，遠隔ユーザの端末上で計算した位置と局所ユーザの端末上で計算した値とはかならずしも一致しない．

5.3 直接音の補間

直接音に関しては，Savioja [Sav 99] と同様に線形補間をおこなう．補正前の時刻 t_i におけるユーザ r からの音声の距離による減衰値を $a(r, t_i)$ ($0 \le a \le 1$) とすると，標本 $s_1, s_2, \ldots, s_N$ (標本数 N = 160) の補正後の値は $\delta = (a(l, t_i) - a(l, t_{i-1})) / N$ を使用して

$$(1+\delta)s_1,\ (1+2\delta)s_2,\ \ldots,\ (1+N\delta)s_N$$

とする．この値を使用して，移動がないときと同様に HRTF のたたみこみをおこなうことにより，ノイズが聴取されることはなくなった．

5.4 残響の補間

残響に関しては，その計算においてつぎの 2 か所で補間が計算される．

- 音量を補間しながら HRTF を計算する部分
- HRTF 計算後の標本を鏡像ごとに遅延時間と音量とを変化させ補間しながら，直接音とミキシングする部分

まず HRTF の計算における補間についてのべる．初期反射に

関しては鏡像の位置にかかわらず同一の HRTF を適用しているので，直接音と同様のあつかいをすることはできない．なぜなら，移動による鏡像との距離の変化は，鏡像ごとにことなるからである．鏡像ごとにことなる HRTF をもとめるのであれば ITD, IID を使用する方法の利点はうしなわれ，膨大な計算が必要になる．そのため，HRTF の計算においては補間をおこなわず，移動がないときとまったく同様に計算している．これにより，移動があるときに直接音をのぞいた初期反射だけを聴取するとクリックノイズがきかれるのをさけることができない．しかし，初期反射を直接音とともにきくとノイズはあまりめだたない．したがって補間しない方式をとっている．

つぎに，鏡像ごとの音量と遅延の補間についてのべる．音量に関しては直接音に関するのと同様の方法で補間をおこなっている．これによってノイズをおさえることができる．しかし，上記のように HRTF の計算において補間をおこなっていないため，完全にはノイズをなくせない．遅延に関しては本来は局所ユーザの移動にともなって残響も変化するはずだが，この変化を現在はシミュレートしていない．すなわち，ユーザは，移動前にきいた直接音に対応する残響に関しては，移動しなかったのとおなじ音をきくことになる．この方法においてはドップラー効果が生じることもない．これに対して直接音に関しては補正された位置からの音をききく．この簡易化によって聴覚的にどのような影響があるかはわかっていない．

6. 結果

VPII 開発からえられた結果をまとめる．

- **HRTF の計算法**: 原データとして KEMAR の測定結果を使用し，標本化周波数 44.1 kHz の HRTF とチェビシェフ・フィルタを使用して 8 kHz の HRTF をもとめた．この HRTF の使用するとほぼただしく方向を識別できるが，他の方法との比較はまだおこなっていない．
- **残響の計算法**: 残響 はimage source 法による初期反射だけをとりいれて，音のひろがりや距離感を表現することができた．限定的な実験の結果，反射率が 0.4 では距離感の表現が不十分だった．0.8 では音声は明瞭だが不自然さがあり，0.7 程度が最適と判断した．しかし，音のひろがりや距離感の効果には個人差があり，頭内定位を指摘する被験者もいた．ユーザが移動可能な範囲としての音室と残響計算でも使用したことによる効果や，この計算に 5.3 節の簡略化された HRTF の計算をとりいれた効果ははっきり確認できていない．
- **動的変化への対処法**: ユーザや音源が移動したり回転したりしたとき，位置や方位角を補正し，音量や遅延を補間している．その結果，通常使用する状態では移動や回転によりユーザを不快にするほどのノイズが発生することはふせぐことができた．
- **実行性能**: 2.8 GHz Pentium 4 の PC において，もっとも計算負荷がたかい HRTF のたたみこみ計算をする部分の実行が (1 パケットぶんすなわち 20 ms ぶんのデータ処理に MMX 等のベクトル演算命令を使用せずに) 38 μs かかるが，これは 1 GFLOPS の計算速度を実現している．反射の計算と初期化部分をあわせた音声 3D 化全体では約 60 μs かかるが，これは 20 ms の時間内に 300 回以上の音声 3D 化計算をおこなうことができることを意味している．すなわち，1 個の CPU で 18 人のユーザをふくむ音室の音声 3D 化計算がおこなえる (18×(18－1) = 306).

7. 結論

VPII においては，初期反射のシミュレーションにより音の頭外定位と距離感の表現を可能にし，さらにユーザの移動を追跡し必要な補間処理をおこなう 3D 音響技術を開発した．これによって，話者識別が容易で，複数の会話コンテクストが共存でき，音室内の移動が自然でノイズがすくない音声コミュニケーション環境を実現した．しかし，これらの技術はまだあらけずりであり，今後，認知的な評価等にもとづいて洗練していく必要がある．

参考文献

[All 79] Allen, J. B. and Berkley, A., “Image Method for efficiently Simulating Small-Room Acoustics”, *J. Acoustical Society of America*, Vol. 65, No. 4., pp. 943–950, April 1979.

[Beg 99] Begault, D. R., Virtual Acoustic Displays for Teleconferencing: Intelligibility Advantage for “Telephone-Grade” Audio, *J. Audio Engineering Society*, Vol. 47, No. 10, pp. 824–828, October 1999.

[Beg 00] Begault, D. R., “3-D Sound for Virtual Reality and Multimedia”, NASA/TM-2000-XXXX, NASA Ames Research Center, April 2000, http://human-factors.arc.nasa.gov/ihh/spatial/papers/pdfs_db/-Begault_2000_3d_Sound_Multimedia.pdf

[Beg 01] Begault, D. R., “Direct Comparison of the Impact of Hed Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source”, *J. Audio Engineering Society*, Vol. 49, No. 10, pp. 904–916, October 2001.

[Bro 99] Bronkhorst, A. W. and Houtgast, T., “Auditory Distance Perception in Rooms”, *Nature*, 397, pp. 517–520, 1999.

[Che 53] Cherry, E. C., “Some Experiments on the Recognition of Speech, with One and with Two Ears”, *J. Acoustical Society of America*, Vol. 25, pp. 975–979, 1953.

[Cre 01] Creative Technology, “Environmental Audio Extensions: EAX 2.0, Version 1.3”, http://www.sei.com/algorithms/eax20.pdf.

[Fun 03] Funkhouser, T., Tsingos, N., and Jean-Marc Jot., “Survey of Methods for Modeling Sound Propagation in Interactive Virtual Environment Systems”, Presence, 2003.

[Gar 94a] Gardner, B. and Martin, K., “HRTF Measurements of a KEMAR Dummy-Head Microphone”, MIT Media Lab Perceptual Computing – Technical Report #280, 1994.

[Gar 94b] Gardner, W. G., “The Virtual Acoustic Room”, Masters Thesis, MIT, 1994.

[Gar 99] Gardner, W. G., “3D Audio and Acoustic environment Modeling”, HeadWize technical Papers Library, http://headwize2.powerpill.-org/tech/gardner_tech.htm, 1999.

[Gri 00] Griesinger, D., “Reflections on Surround”, Sound on Sound, March 2000, http://www.soundonsound.com/sos/mar00/articles/dave.htm

[Kan 03] 金田 泰, “仮想の ‘音の部屋’ によるコミュニケーション・メディア *Voiscape*”, 電子情報通信学会 技術研究報告 (MVE / VR 学会 EVR 研究会), 2003-10-7.

[Kan 04a] 金田 泰, “仮想の ‘音の部屋’ によるコミュニケーション・メディア voiscape の JMF と Java 3D を使用した実装”, 電子情報通信学会 技術研究報告 (DPS/CSEC 研究会), 2004-3-5.

[Kan 04b] Kanada, Y., “Multi-Context Voice Communication Controlled by using an Auditory Virtual Space”, *2nd Int’l Conference on Communication and Computer Networks* (*CCN* 2004), pp. 467–472, 2004.

[Kan 05] Kanada, Y., “Multi-Context Voice Communication In A SIP/SIMPLE-Based Shared Virtual Sound Room With Early Reflections”, *NOSSDAV* 2005, 出版予定, 2005.

[Kro 68] Krockstadt, U. R., Calculating the Acoustical Room Response by the Use of a Ray Tracing Technique, J. Sound and Vibrations, Vol. 8, No. 18, 1968.

[Sav 99] Savioja, L., “Modeling Techniques for Virtual Acoustics”, Helsinki University, 1999.

[She 82] Sheeline, C. W., “An Investigation of the Effects of Direct and Reverberant Signal Interaction on Auditory Distance Perception”, Ph.D. Dissertation, Stanford University, 1982.

[Shi 00a] Shinn-Cunningham, B., “Distance Cues for Virtual Auditory Space”, *1st Pacific Rim Conference on Multimedia*, pp. 227–230, IEEE, December 2000.

[Shi 00b] Shinn-Cunningham, B., “Learning Reverberation: Consideration for Spatial Auditory Displays”, *International Conference on Auditory Display* (*ICAD* 2000), pp. 126-134, April 2000.