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

日立製作所 中央研究所

金田 泰

背景

- 音声は人間どうしのコミュニケーション・メディアの起源であり，現在ももっとも重要である．
- さまざまな音声コミュニケーション・メディア (VCM)
  - 電話
    - “不便な”ユーザインタフェースが 130 年間もかわらずにきた．
  - 遠隔会議システム
    - 電話の不便さを一部解消した．
    - 他の不便さを導入した．
  - 他の VCM
    - トランシーバ
    - アマチュア無線
    - …

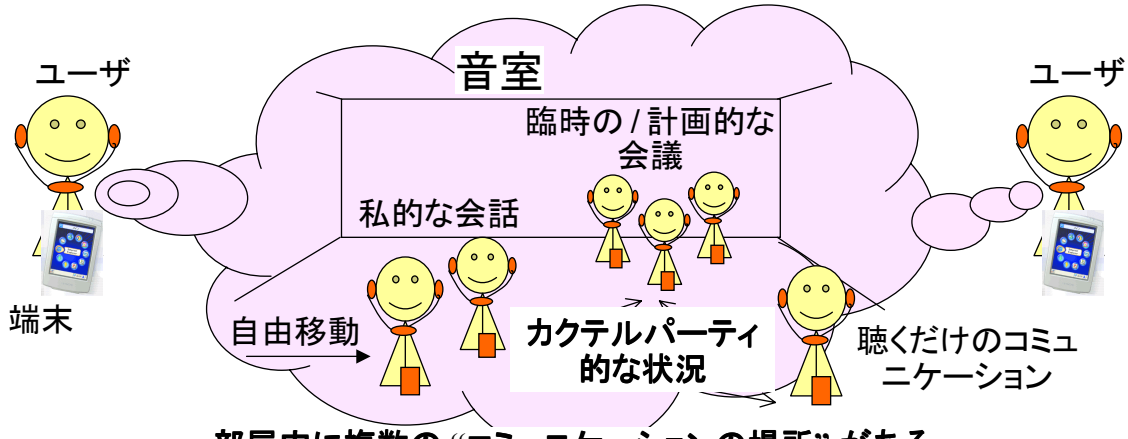
A telephone set in 1878
(http://www.atcaonline.com/phone/coffin.html)

背景 (つづき)

- **VCMを革新するべし**
  - ◆顔をつきあわせての会話では，さまざまなコミュニケーション・パターンが可能.
    - ▮ 例: 2人以上による自由な会話.
  - ◆VCMをとおしたコミュニケーション・パターンは限定的.
- **VCMにおける具体的な問題**
  - ◆話者識別問題
    - ▮ 話者の同定や話者を記憶することが困難
      — とくに音声だけの環境では.
  - ◆複数話者問題
    - ▮ 顔をつきあわせての会話では，しばしば並列の会話がおこる.
    - ▮ VCMではこれは実現困難.

voiscapeとは?

- **“音室”(音の部屋)**
  - ◆仮想空間を音の方向や距離によって表現する
    (3D音響によって表現する).
  - ◆音室内のひとは自由に移動できる.
- **voiscapeは音室を使用するVCMである.**
  - ◆音室内に“コミュニケーションの場所”がつくられる.



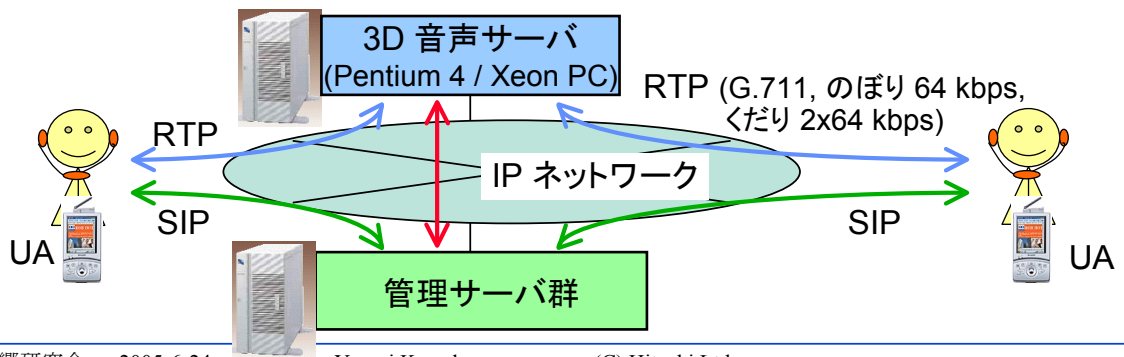
部屋内に複数の“コミュニケーションの場所”がある.

voiscape のプロトタイプ

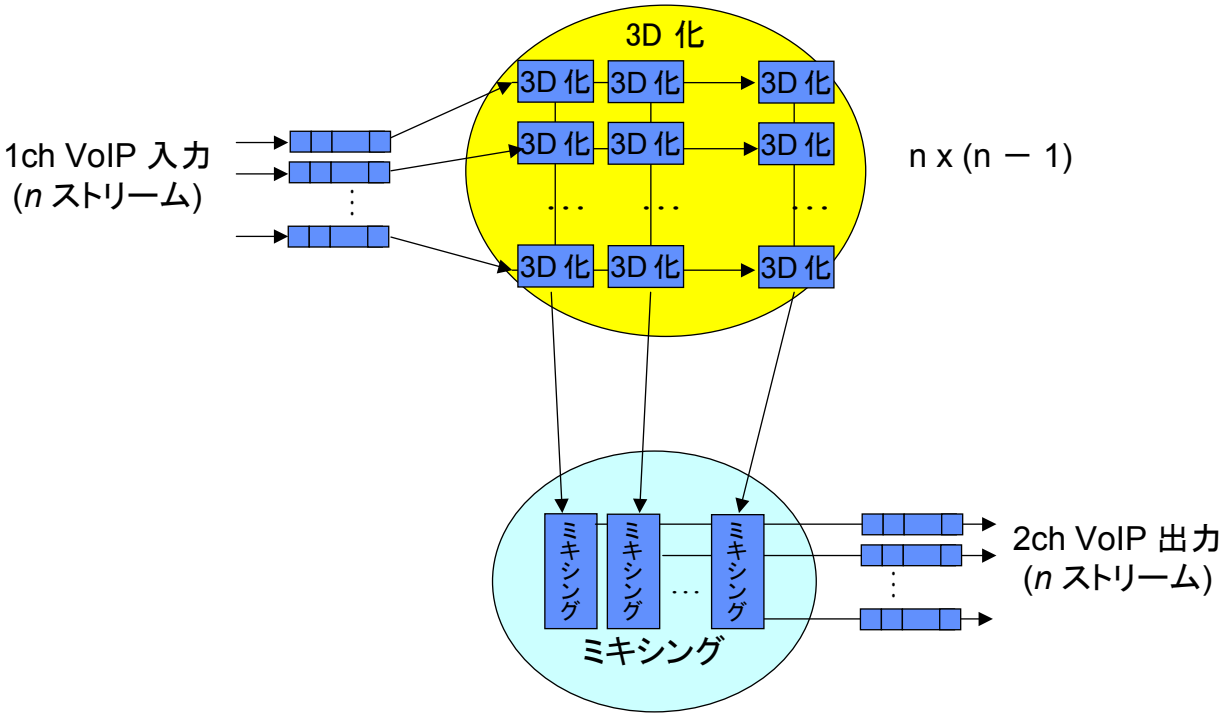
- **Jasper: 最初のプロトタイプ [CCN 2004].**
  - Java ベース (JMF, Java3D, LWJGL (light-weight Java Game Library))
  - くみこみの VoIP と 3D 音響を使用 — 音質がよくなかった.
- **VPII (Voiscape Prototype II): 第 2 のプロトタイプ [ここで報告].**
  - C++ と C にもとづく — よりよい性能をえるため.
  - VoIP (RTP) と 3D 音響はゼロから開発.

VPII のアーキテクチャ

- **VPII の 3 大要素**
  - ユーザエージェント (UA)
    - 端末ソフト: Linux PDA (Zaurus) または Windows PC で動作.
    - Ethernet または無線 LAN を使用.
  - 管理サーバ群 (RMS, RLS, SIP レジストラ)
    - 部屋, ユーザ位置, 音室リストの管理 — SIP と SIMPLE を使用.
      - SIMPLE = SIP for Instant Messaging and Presence Leveraging Extensions.
  - 3D 音声サーバ (or メディアサーバ)
    - 3D 化とミキシング.
    - 現在, DSP は使用していない.

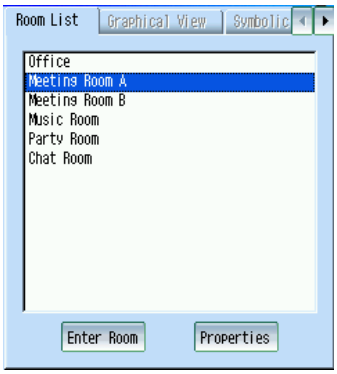


3D 音声サーバの処理構造

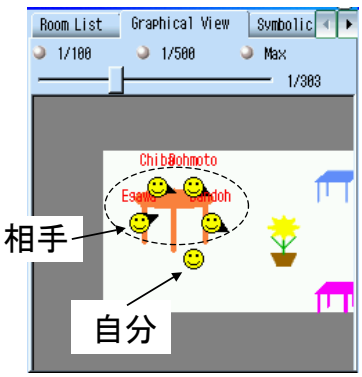


VPIIのユーザインタフェース

■ ユーザはリストから音室を選択.
- ◆RLS (音室リストサーバ) が音室リストをUAに送信.



■ UAが音室を表示.
- ◆聴覚表示 (auditory display) が主表示.
- ◆視覚表示 (地図) は補助表示.
- ◆これらのくみあわせ
  - ▮ ユーザは音声とアイコンのマッピングをとる.



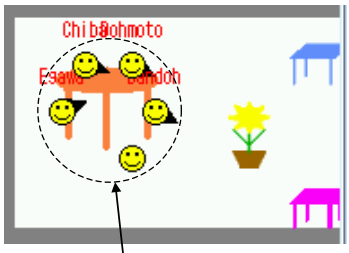
VPIIのユーザインタフェース (つづき)

■ ユーザはカーソルキーか他のポインティング・デバイスをつかって移動する.
  ◆この動作は実世界での動作と独立.

VPIIの特徴

■ 低遅延・動作追跡型3D音響
  ◆HRIR (頭部伝達関数対応のインパルス応答)と初期反射とを計算.
  ◆双方向通信のため, 3D化による遅延を最小化.
  ◆ユーザの動作を実時間で再生音に反映.

■ 仮想の場所にもとづく選択的なコミュニケーション
  ◆ユーザは“コミュニケーションの場所”を地図とアイコンをつかって選択.
  ◆アイコンは“標識”としてつかえる.



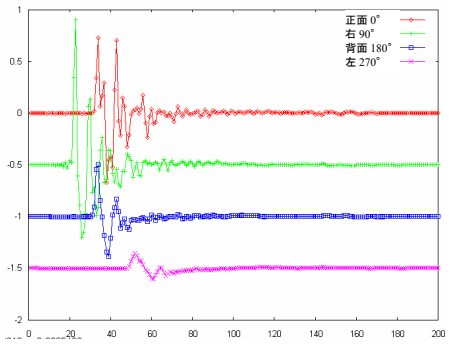
■ SIP/SIMPLEにもとづく音室管理
  ◆ユーザの位置・方向を部屋の “プレゼンス” の一部としてあつかう.
  ◆SIP/SIMPLEをプレゼンス・イベント(動作)の通知に使用.
    ▮ SIP = Session Initiation Protocol (IETF 標準)
    ▮ SIMPLE = SIP for Instant Messaging and Presence Leveraging Extensions

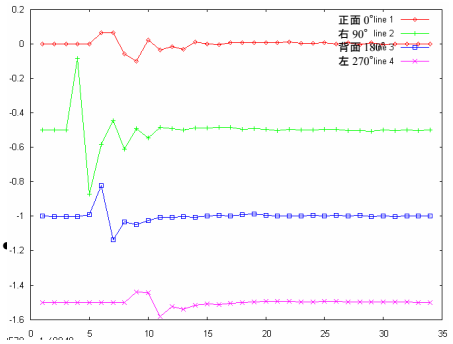
VPIIの標本化周波数

- 標本化周波数を 8 kHz とした.
- 8 kHz を使用した理由
  - ◆無理のない帯域幅・低遅延の実現
    - 広帯域で圧縮率のたかい MP-3, AAC などは遅延・負荷がおおきい.
    - ITU-T G.711 (8 kHz) は遅延なしにどんな端末でも実現できる.
  - ◆時間領域での実時間信号処理
    - フーリエ変換 (FFT) をつかうと遅延が発生する.
  - ◆基本的に音声だけをあつかうので狭帯域
    - **音声には高域成分はすくない**
      — 8 kHz 標本化でほとんどの情報をつたえられる.

VPIIの3D音響におけるHRTFの計算

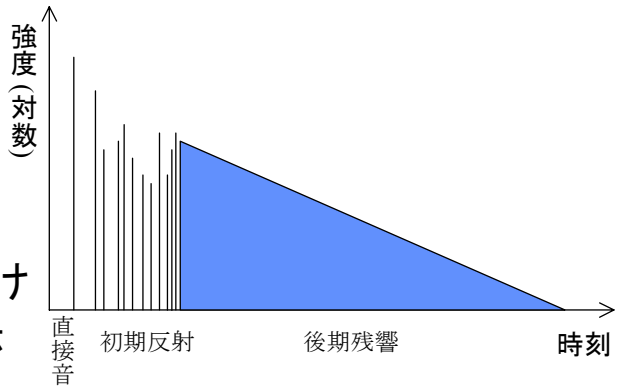
- **時間領域でHRIR (頭部インパルス応答) をたたみこみ**
  - **◆有限インパルス応答 (FIR) を使用.**
- **測定データとその変換**
  - **◆CIPICデータベースからKEMARによる測定結果を入手**
  - **◆44.1 kHzで測定したデータを8 kHzにダウンサンプリング**
    - **周波数応答を優先し，位相は犠牲にした.**
  - **◆5°ごとに測定された水平方向のデータだけを使用.**
    - HRIR (HRTF) **じたいは補間していない.**





VPIIの3D音響における残響計算

■ 残響はつぎの2つからなる.
- ◆初期反射
- ◆後期残響



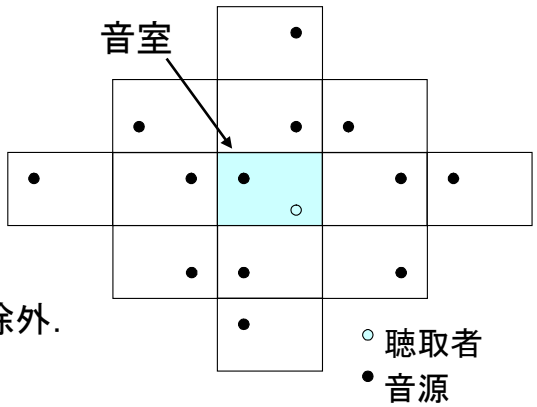
■ VPIIにおける残響の計算法
- ◆音室の壁による初期反射だけを2次元のimage source法によって計算.

■ 初期反射を計算している理由
- ◆頭外定位させる.
- ◆距離感をあたえる.

■ 後期残響を計算していない理由
- ◆明白な利点がない(?): 頭外定位や距離感にはきかない.
- ◆むしろ有害(?): 音声を不明瞭にするし, 計算量がおおい.

VPIIの3D音響における初期反射の計算法

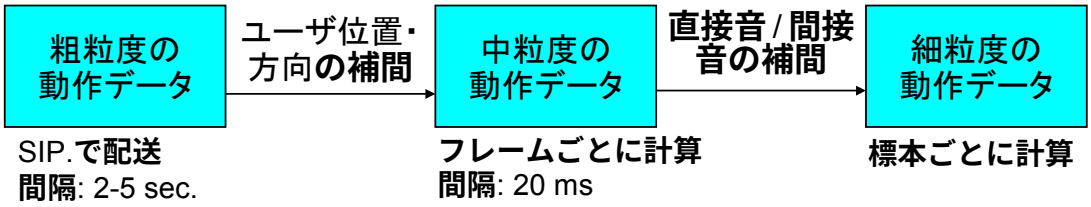
■ 2次元image source法
- ◆壁による12個の反射を計算.
  - ▮ 音室とその鏡像を上からみた図:
  - ▮ 遅延 150 ms をこえる反射音は除外.



■ 計算量をへらすためのくふう
- ◆初期反射はITD, IIDを制御することによって3D化している.
  - ▮ ITD = interaural time difference (両耳間時間差)
  - ▮ IID = interaural intensity difference (両耳間強度差)
- ◆初期反射の方向によらず同一のHRIRを使用.

VPIIの3D音響における動作追跡

- **急速なユーザ動作から発生する問題**
  - クリック・ノイズ
  - ユーザが識別できなくなること: 移動前後で同一性がわからなくなる.
- **問題解決のための3つの補間法**
  - ユーザ位置・方向の補間
  - 直接音の補間
  - 反射音の補間



- **反射音の補間計算はVPIIでは省略している.**
  - 理由1: 反射計算量の削減をだめにする.
  - 理由2: 発生するノイズはちいさい.

結果 (非公式の評価)

- **定位**
  - 大半のひとは頭外定位をみとめた.
  - 垂直方向の定位はあいまい — 個人差がおおきい.
  - 水平の定位も不正確 (?) — 初期反射のため?
- **残響の計算法**
  - 反射率は0.7程度が適切 (15m × 10m程度の音室において).
    - 反射率0.4では距離感の表現が不十分.
    - 反射率0.8〜では音声は明瞭だが不自然さがある.
  - 反射計算における計算量削減の効果は確認できていない.
- **動作追跡**
  - ユーザを不快にするほどのノイズの発生はおさえられた.
- **実行性能**
  - 1フレーム (20ms) のデータ処理時間 (2.8GHz Pentium 4)
    - HRIRのたたみこみ: 38μs, 3D化全体: 60μs.
  - 18人のユーザをふくむ音室のメディア処理が1 CPUで可.

結言

■ 結論

- ◆voiscapeむきの3D音響技術を開発した.
  - ▮ 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た.
  - ▮ ユーザの移動を追跡し，必要な補間処理をおこなっている.
- ◆複数の“コミュニケーションの場所”をふくむ環境を実現した.
  - ▮ 音室内で並列の会話ができるようになった.
  - ▮ ユーザの移動が自然でノイズがすくない.

■ 今後の課題

- ◆認知的な評価
- ◆3D音響技術の洗練

デモのかわりに録音ずみの再生音のサンプルを用意しています.