3Y-2

顔追跡による音環境可視化システムのアウエアネスの改善

久保田 祐史 † 白松 俊 † 駒谷 和範 † 尾形 哲也 † 奥乃 博 †

† 京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

人間の聴覚は緊急性には優れているものの, 視覚情報ほど明確な理解が難しい. 特に録音された音を聞き取るとき, (1) 時間的一覧性の欠如, (2) 複数音声の全てを聞き取りの対象とする音源分離知覚における知能限界 [1] による同時複数発生音の弁別の困難性, (3) 音源方向や音環境の客観視などの空間情報の欠如による音の臨場感の損失, という 3 つの問題にしばしば直面する.

我々はこのような問題に対し, ロボット聴覚システム HARK[2] により収録された音源ごとの位置や発話内容を習得し, それらを “Overview first, zoom and filter, then details on demand ” という設計方針 [3] に基づいて, 視覚的な情報に置き換えることにより, 正確かつ直感的な理解が可能とする可視化システムを開発してきた [4, 5].

一方, (3) の損失問題は, ユーザの挙動を考慮した情報提示を行っておらず, 残されていた. 本稿では, (3) の問題を聴覚的アウエアネスの欠如によるものと捉え, これを改善させることで音の臨場感の向上を達成し, 音環境理解支援を行うシステムを設計, 開発したので報告する.

2. 聴覚的アウエアネスの課題

アウエアネスとは人間や動物が, 自分が置かれた状況や周囲の出来事に, 行動を起こし注意を払うことで視覚や聴覚を通じて認知・知覚することを指す心理学的概念である. 聴覚的アウエアネスとは, 身体動作やマウス操作などを含むユーザの行動により, 提示された音情報に注意を払うことで新たな音情報の知覚を得ることに相当し, 身の回りの音環境理解に非常に重要である. 特に聴覚器官を有する顔の動きは, 音源の存在する方向, 特に前後関係を知覚する働きに強い関わりがあることが報告されている [6]. 聴覚的アウエアネスの改善により音の臨場感の向上を図るための課題は次の 2 点である.

1. **音環境提示**: 効果的な音情報提示の機能設計.
2. **身体性を有するインターフェース**: 身体動作によって注意を払う音情報の選択と提示手法の設計.

音情報提示に対しては, これまで開発してきた音環境可視化システム [4, 5] をベースに拡張した. 音環境の AuditoryScene XML による表現法を設計し, AuditoryScene XML による音環境提示法として音声認識結果のカラオケ表示, 音源同定結果の表示, 従来オーディオ機器と異なる早送り機能を追加した.

身体性を有するインターフェースについては, 顔の動作に応じて音情報の提示を行う GUI を設計・実装した. これを通じて”没入感”をユーザに感じさせ, 録音された環境の空間情報を与え, 音の臨場感の向上に取り組んだ.

3. 身体性を有するインターフェースの設計

本稿では, あたかもユーザ自身がマイクロホン位置に居るかのようにシステム内に没入させ, 可視化された音環境の提示を行うインターフェースの設計を行った.

Face-Tracked Auditory Scene Visualization towards Auditory Awareness
Yuji Kubota, Shun Shiramatsu, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

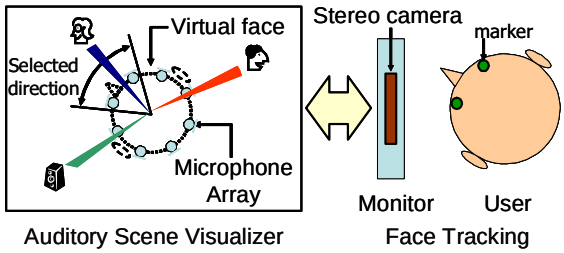

図 1: 身体性インターフェースの設計概要.

3.1 顔動作に基づくインターフェースの設計

音楽が流れている部屋で 2 名の話者が会話している状況を, 本システムで可視化を行った様子を図 1 に示す. ユーザが右の方向に顔を向けると女性の音声の, 逆に左側の方向に顔を向けると音響機器より再生されている音楽の情報が提供, 再生される. このようにユーザは可視化された情報から所望の音源を判断し, その方向に顔を向けることにより対応する音源の選択を行う. 顔の向きの検出は, 音環境可視化システムが表示されているモニター上に設置されたステレオカメラにより行う.

3.2 身体性を考慮したアウエアネス改善の関連研究

人間は左右 2 つの耳を用いて音を聞き取ることにより, 両耳に入力される両耳間レベル差 (Interaural Level Difference: ILD) や両耳間位相差 (Interaural Phase Difference: IPD) の変化を捉え, これらを手がかりに音源の位置を知覚することができる. 中でも Thurlow らは音像知覚に関し, 受聴時に頭部を固定するよりも, 頭部を移動させた方が音像の定位感が向上し, その際には横方向に 30°〜40°, 上方向に最大で 14° まで顔を移動することが最も多かったと報告している [7]. Algazi らはこの特性を利用し, 原音に忠実な音場空間再現を目指し, 顔動作に追従したバイノーラル再生を提案した [8].

しかし, この手法では 1 章で述べた (2) の知能限界による困難性が未解決であるなど, 実際の環境以上の聴覚的アウエアネスを与えることができない.

4. 音環境可視化システムの拡張

本システムは次の 3 つのサブシステムから構成される (図 2).

1. 音環境理解クライアントシステム
   (a) 音響信号録音モジュール,
   (b) 音源定位モジュール,
   (c) 音源分離モジュール,
   (d) 分離音認識モジュール.
2. 顔追跡クライアントシステム
3. 3 次元音環境可視化サーバーシステム

音環境理解クライアントシステムは, オープンソースとして公開されているロボット聴覚システム HARK[2] を用いて, AuditoryScene XML 形式で出力する. 色判別による顔追跡 [5, 9] により, ユーザの顔に貼られた 2 点のマーカー (図 1) の水平位置を検出し, ユーザの顔の向きを出力する. これらの入力を受け, 3 次元音環境可視化システムにて音情報の提供を行う.

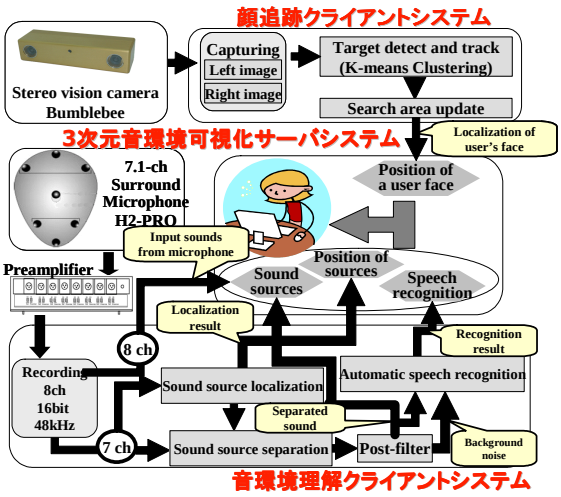

図 2: アーキテクチャ全体図

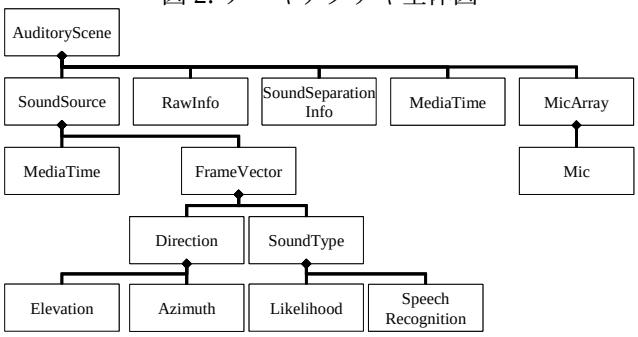

図 3: AuditoryScene XML 階層図.

4.1 AuditoryScene XML

本システムはロボット聴覚システムより得られた音情報を体系的なデータとして表現し, 各音情報の相互参照を実現させるため, AuditoryScene XML の設計を行った. AuditoryScene XML の階層図を図 3 に, 各要素と属性を以下に示す.

- **RawInfo**: 録音ファイルの設定. 録音ファイルの保存位置とサンプリングレート.
- **SoundSeparationInfo**: 音環境理解技術の設定項目.
- **MediaTime**: 録音の開始, 終了時刻.
- **MicArray**: マイクロホンアレイの設定, 数.
- **Mic**: マイクロホン 1 チャンネルごとの位置情報.
- **SoundSource**: 各音源ごとの音響信号データ. 音源 ID とファイルの保存位置.
- **FrameVector**: 音環境理解技術で得られたフレームサイズごとの各音源ごとのフレームベクトル.
- **Direction**: 各音源ごとの水平角度, 垂直角度.
- **SoundType**: 各音源ごとの音源の種類 (音声/非音声).
  - **Likelihood**: SoundType である尤度.
  - **SpeechRecognition**: SoundType が音声である場合, カラオケファイルフォーマット形式での音声認識結果.

4.2 インターフェース

本システムの操作画面を図 4 に示す. 操作パネル (①) には PLAY, PAUSE, STOP, RECORD, FFW ボタンと早送り速度, 音量調整バーがあり, 通常の音再生機器と同じ感覚で使用することが出来る. PLAY ボタンを押すことで音が再生され, 各音源の方向が 3D 画面中のマイクホロ

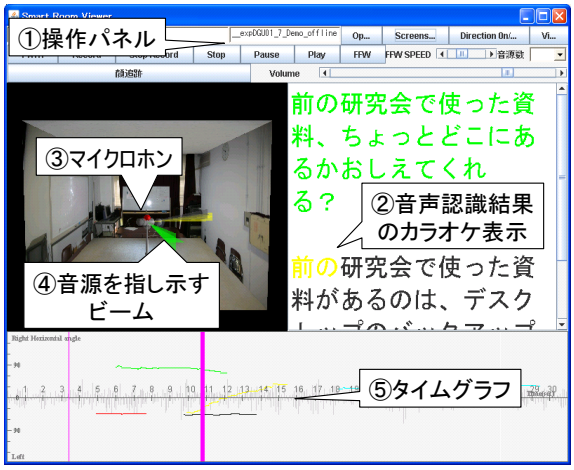

図 4: 音環境可視化インターフェース

ンアレイ (③) に向かってビーム (④) で表示され, 音声内容のカラオケ風表示が同期して行われる (②).

タイムラインやカラオケ風表示の文字をクリックすることにより, その時点までスキップが行われる. また, 音源数の指定を行うことで指定された音源数以下のシーンのスキップ再生が行われる.

顔追跡ボタンを押すことで, マイクロホンアレイ (③) 位置に視点が移動し, 3 章で述べた顔の動きに基づいた音の選択が行われる. また, View ボタンを押すことで表示されている部屋の視点角度が 90° ずつ回転し, 全方位の音源の選択が可能となる.

5. まとめと今後の課題

本稿では, 音の臨場感をアウエアネスの観点から向上させるために, 以下の手法を設計し実装した. (1) 音環境の AuditoryScene XML による表現法. (2)AuditoryScene XML による音環境表示法として新たに音声認識結果のカラオケ表示, 従来オーディオ機器と異なる早送り手法. (3) 目と耳をもつ顔の動きからユーザの音環境への注目を観測し, 対応する情報の提示手法. これらにより音環境の豊かなアウエアネスを与えることを実現した.

謝辞 本研究は, 科研費, GCOE, HRI-JP の支援を受けた.

参考文献

[1] 川島他: 同時複数音声の分散的聴取における知能限界, 日本音響学会誌, Vol.65, No.1, pp.3-14, 2009.
[2] 中臺他: ロボット聴覚オープンソースソフトウェア HARK の概要と評価, 日本ロボット学会第 26 回講演会, 1A2-04, Sep. 2008 http://winnie.kuis.kyoto-u.ac.jp/HARK/
[3] B. Shneiderman: *Designing the User Interface (3rd Ed)*, Addison-Wesley, 1998.
[4] 吉田他: 音環境を可視化する録音再生システム, 情処第 69 回全大, 6ZB-2, Mar. 2007.
[5] Y. Kubota, *et al.*: Auditory Scene Visualization With Face Tracking Towards Auditory Awareness, *ISM*, pp.468-476, 2008.
[6] P. Mackensen: Auditive localization. Head movements, an additional cue in localization. PhD dissertation, Tech. Univ. Berlin, 2004.
[7] W.R. Thurlow, *et al.*: Effect of induced head movements on localization of direction of sounds. *J.Acoust.Soc.Am.*,vol.42,pp.480-488,1967.
[8] V.R. Algazi, *et al.*: Motion-Tracked Binaural Sound. *J. Audio Eng. Soc.*, vol.52, No.11, pp.1142-1153, 2004.
[9] 和田: 最近傍識別器を用いた色ターゲット検出, 情報処理学会論文誌, Vol.44, No.SIG17-014, 2002.