顔の動作に追従したインタフェースを持つ音環境可視化システム

久保田 祐史† 吉田 雅敏‡ 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 ‡NTT アドバンステクノロジ株式会社

1. 音の情報爆発での課題

ディジタル技術やネットワーク社会の発展と共に, ディジタル音楽流通や講義内容配信など, ネットワークを通じてディジタルアーカイブ化された音響データのやり取りが活発化している. このような音情報を収集し, 活用していく上で, 情報爆発の観点から次の 2 つの課題がある.

1. **音情報の量的爆発の促進**: これまで音響データのアーカイブ対象は音楽や音声が中心であり, 音情景までアーカイブすることはほとんど取り組まれていない.
2. **種々の音の質的複雑化を軽減**: 音では, 画像のサムネイルやスナップショットなどを活用した一覧性や弁別性を達成することが難しく, うまく情報を提示できていない.

我々はこのような 2 つの課題に対して, 前者には混合音を聞き分ける技術で, 後者には音を見せる技術で対応をしてきた. すなわち, サラウンドマイクロフォンを通じて, 混合音を収録し, 音源定位, 音源分離により, それぞれの音源を録音するとともに, "Overview First, Zoom and Filter, then Details on Demand" (以下, O-ZF-D と略す) という 3 つのレベルにおける設計方針 [1] で, 音の可視化システムを設計, 開発されている [2].

O レベルではユーザーに全体の概略を示すべく, 音源の存在する方向を話者ラベルごとに色付けしてタイムグラフを表示. *ZF* レベルでは音源の存在する方向を 3D 空間上にビーム表示し, 全音源の再生を行うことで音源の存在を提示した. *D* レベルでは音源の詳細情報を提示するべく, ユーザーが指定した音源のみの再生, すなわち分離音再生を行う.

本稿では, この本可視化システムの課題を述べるとともに, ステレオカメラによる顔の動作追跡を用いたインターフェースについて報告する.

2. 音環境可視化システムの課題

これまでに開発してきた可視化システムの使用経験から次のような問題点が明かになった.

- **GUI の問題点**: 本システムは全ての操作をマウスによって行う. このため, システムを利用している間は, 視線と手の拘束性が高く, 特に音情報を提示しているディスプレイを見ながら, 所望の GUI パネルを選択し, マウスでクリックするのは即時的な操作が行えず, 音情報探索の阻害となる.

我々は日常生活の中で物を見るとき, 見難い文字や模様があれば顔を近づけ, 全体像を確認しようとする時は顔を遠ざける行動を取る. このようなアナロジーを本可視化システムのインタフェースに適用できると, 閲覧と同時に操作が行える上, 自然なインターフェースで, 提示内容の切替が可能となり, システムの利便性が向上すると期待される. そこで本稿では, 音に対する人間の顔の動作に着目し, 音源を探したり音を聞き分けたりするメタファーに基づいたインターフェースの設計を行った.

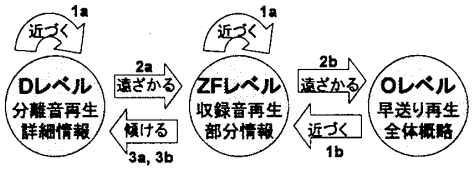

図 1: *O-ZF-D* レベルの遷移図

3. 顔の動作のメタファーに基づく設計

ある音源が存在する時, 我々は大きく分けて次の 3 つの顔の動作により, 音源に対して行動をとっている.

1. 顔を近づける ⇒ 音源に興味を示す.
   a. 微小な音を聞こうとする.
   b. 音源を確かめる.
2. 顔を遠ざける ⇒ 音源を避ける.
   a. 音源から退避する.
   b. 全体の音環境を把握しようとする.
3. 顔を傾ける ⇒ 一部の音源に注意を払う.
   a. 特定方向の音源に注意を向ける.
   b. 近い音源の定位の曖昧性を解消しようとする.

これらの顔動作による *O-ZF-D* レベルの遷移図を図 1 に示す. まず、*O* レベルでは, 顔を近づけることで, メタファー 1b に基づき, 収録音再生 (*ZF* レベル) に切り替える. 逆に収録音再生時に顔を遠ざけることでメタファー 2b に基づいて, *O* レベルの提示に切り替える. また, 顔を傾けた場合, メタファー 3a, 3b に基づいて, ユーザがどの音源に対して興味を持っているのかを示し, その音源の分離音再生 (*D* レベル) を行う. *D* レベルでは, 顔を遠ざけると, メタファー 2a に基づき, *ZF* レベルの提示を行う. また、*ZF, D* レベルで顔を近づけた場合, メタファー 1a に基づいて, 現在再生している音源の音量を大きくする.

このように 3 つの顔の動作に追従して, 情報視覚化の各レベルにおける提示内容の切替を行う.

4. 音環境可視化システムの拡張

4.1 システム構成

ディスプレイに対する顔の動作により, 前章のメタファーに基づいた可視化システムの操作を行うシステムを構築した. 本システムの構成を図 2 に示す. システムは二つのクライアントとサーバシステムで構成され, クライアント・サーバ間は LAN によって相互に接続される. 録音クライアント, 音再生モジュール, 出力インターフェースは従来システム [2] のものを使用した.

Auditory Scene Visualization with Tracking of Face Moves *Yuji Kubota (Kyoto Univ.), Masatoshi Yoshida (NTT-AT), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

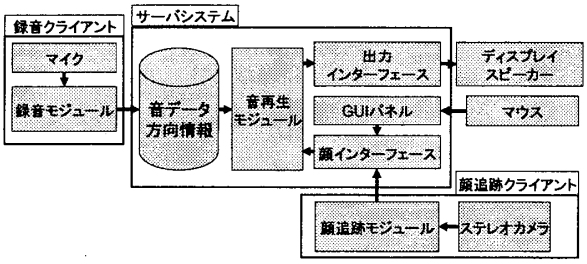

図 2: システム構成

4.2 顔追跡モジュール

1. 入力: 可視化システムをモニタするディスプレイ上に, 固定設置したステレオカメラから取得した 1 対のフレーム画像 k.
2. 出力: 顔の 3 次元位置 $(X_{obj}(k), Y_{obj}(k), Z_{obj}(k))$.
   (a) **顔領域検出**: 和田 [3] が開発した最近傍識別器を利用して, 肌色領域を検出.
   (b) **顔領域の追跡・重心座標の算出**: 検出した領域の中で最大のものに対し, 大池ら [4] が開発した改良 mean shift 法を用いて顔領域の追跡, 重心座標の算出を行う.
   (c) **3 次元位置計測**: 顔領域の重心座標を対応点として決定し, カメラからの顔の 3 次元位置 $(X_{obj}(k), Y_{obj}(k), Z_{obj}(k))$ を計測する.

4.3 顔インターフェース

1. 入力: GUI パネル操作 (再生 (PLAY), 停止 (PAUSE), 音量 (Vol_{sys}), 提示画面の視点位置 (θ_{sys})), 顔の 3 次元位置 $(X_{obj}(k), Y_{obj}(k), Z_{obj}(k))$.
2. 出力: 再生モード, 再生音源方向 (θ_{dir}), 音量 (Vol).

- **原点設定**: 再生 (PLAY) ボタンが押された時, 入力された顔の 3 次元位置を原点 $(X_{obj}(0), Y_{obj}(0), Z_{obj}(0))$ として設定し, 収録音再生モードを出力.
  停止 (STOP) ボタンが押された時, 原点を初期化し再生停止モードを出力.
- **音量調整**: メタファー 1a に基づき, 原点に対して現在の顔の位置 $(X_{obj}(k), Y_{obj}(k), Z_{obj}(k))$ が, どれだけディスプレイに近づいてるかに基づいて音量 (Vol)[dB] を出力する. メタファー 2b に基づき, 無音になった (音量が-15dB 以下になった) 時, 早送り再生モードを出力. また, メタファー 1b より, この状態から無音ではなくなった場合, 収録音再生モードを出力する.
- **顔の傾き計算**: メタファー 3a, 3b に基づき, 顔の傾きから一部の音源再生を行う. カメラに対する顔の角度 ($\theta_{obj}(k)$) を

$$\theta_{obj}(k) = \tan^{-1}\left(\frac{X_{obj}(k)}{Z_{obj}(k)}\right)$$

として算出. 原点とする顔の角度と現在の顔の角度の差分 $(\theta_{obj}(k) - \theta_{obj}(0))$ が, 閾値 $(\frac{\pi}{12})$ 以上離れていた場合, 分離音再生モードと再生音源方向 ($\theta_{dir} = \theta_{obj} + \theta_{sys} + \pi$) を出力. この閾値は顔の微小な動きにより、誤って分離音再生モードに切り替わることを防ぐために導入した.

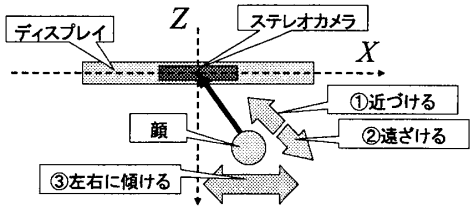

図 3: システムの使用状況を上部から見た様子

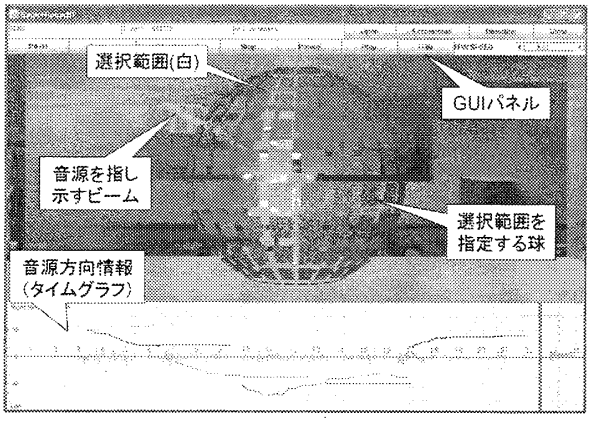

図 4: 分離音再生モードの出力インターフェース

4.4 実装

マイクに Holophone 社の 7.1 次元用サラウンドマイクロフォン H2PRO を, ステレオカメラに Point Gray Research 社の Bumblebee ステレオカメラを採用し, プログラム言語に JAVA 言語を用いて実装を行った. 図 3 に上部から見た本システムを使用している様子を示す. ディスプレイ上に提示された音情報を見ながら, 顔を近づける (①) ことで音量の増大を, 遠ざける (②) ことで音量の減少と早送りを行う. 左右に動かして傾ける (③) ことで, 図 4 のように球体が表示され, カメラへの視線方向を選択し, その方向に存在する音源の分離音再生を行う.

5. まとめと今後の課題

本稿では, 音源を探したり音を聞き分けたりする, 顔の動作のメタファーに基づいたインターフェースの設計により, 操作の柔軟性を達成した. 今後はこのインターフェースを長期的に使用してフィードバックを得ることで, より直感的な操作を可能とする顔インターフェースの検討を進める. また, 本操作インターフェースに基づいて, 音源同定結果や, 分離音が音声であった場合は音声認識結果などより詳細な音情報を提示し, 音の情報爆発に対応できる音環境可視化システムの実現を目指す.

謝辞 本研究の一部は, 科研費, GCOE の支援を受けた.

参考文献

[1] B. Shneiderman: *Designing the User Interface (3rd Ed)*, Addison-Wesley, 1998.
[2] 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.
[3] 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.
[4] 大池他: 鮮明な画像撮影のための高速追従型アクティブカメラ, MIRU2004, pp.I-113-114, 2004.