人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-04

深度センサとマイクロフォンアレイを用いた聴覚アウェアネスの提示

Proposal of auditory awareness using by depth sensor and microphone array

井山貴裕 [1]
Takahiro IYAMA

杉山治 [2]
Osamu SUGIYAMA

坂東宜昭 [1]
Yoshiaki BANDO

糸山克寿 [1]
Katsutoshi ITOYAMA

吉井和佳 [1]
Kazuyoshi YOSHII

奥乃博 [3]
Hiroshi G. OKUNO

[1] 京都大学大学院情報学研究科　　[2] 東京工業大学先進理工学研究科

[3] 早稲田大学実体情報学プログラム

Abstract

本稿では深度センサとマイクロフォンアレイを用いた音源位置推定に基づく聴覚アウェアネス可視化システムについて述べる．従来の音環境の可視化システムは MUSIC スペクトルをカメラ画像上に重畳するものであり，空間的· 時間的な聴覚アウェアネスが欠けている．空間的· 時間的な聴覚アウェアネスを提示するため，聴覚アウェアネス可視化のための三層モデルを設計し，本モデルに基づく可視化システムを開発する．本モデルでは，深度センサを用いることで空間的な聴覚アウェアネスを，音源を追跡し音源の時間変化を求めることで時間的な聴覚アウェアネスを提示する．また被験者実験により，本モデルに基づいて可視化された動画を視聴しながら，各レイヤごとに音源の発音したことを認識するまでの時間を比較し，本モデルの各レイヤごとの差異や有効性を確認した．

1　序論

環境の探索．監視システムの機能向上のためには，画像と音の情報統合に基づく充実した聴覚アウェアネスの提示が必要不可欠である．聴覚アウェアネスとは，音源の方向や位置，音量，種類，状態変化など，音源に対する総合的な気づきを意味する．単独マイクロフォンでは，当然ながら空間的な聴覚アウェアネスが提示できないのに加え，複数の音源が同時に発音した場合には音源種類の提示も困難になる．マイクロフォンアレイを用いると，音源方向など一部の聴覚アウェアネスは聴覚アウェアネスを提示できるが，奥行きも含めた音源の位置や音源状態変化などの提示は困難である．したがって，充実した聴覚アウェアネスの提示のためには，マルチチャネル信号処理技術で得られる音情報と RGB カメラや深度センサなどの画像情報の統合が不可欠である．

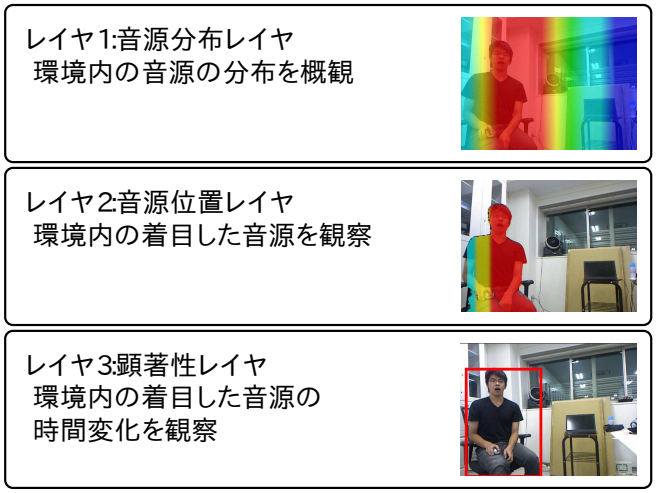


図 1: 聴覚アウェアネス可視化の三層モデルの構成

本研究では，聴覚アウェアネス可視化のための三層モデルを設計し，三層モデルに基づく聴覚アウェアネス可視化システムを設計する (図 1)．本モデルは音源分布レイヤ，音源位置レイヤ，顕著性レイヤの 3 つのレイヤから構成される．音源分布レイヤは環境中の音源の分布の様子を概観する機能を提供する．音源位置レイヤは着目した音源物体の音情報，すなわち空間的なアウェアネスを提示する．顕著性レイヤは着目した音源の時間変化の様子，すなわち時間的な聴覚アウェアネスを提示する．ユーザはこれらのレイヤを自由に選択しながら，着目した音源を観察することができる．

本稿の構成は以下の通りである．第 2 章では，従来の音情報可視化手法とその問題点について述べる．第 3 章では，聴覚アウェアネス可視化のための三層モデルについて述べ，第 4 章では，本モデルに基づく可視化システムについて述べる．第 5 章では，被験者実験により三層モデルの有効性を確認し，第 6 章でまとめを行う．

2　関連研究

聴覚アウェアネス可視化システムの開発のため，音環境の可視化· 深度センサを用いたマルチメディア統合に関する従来法を挙げ，本研究の位置づけを明確にする．

まず，音環境の可視化に関して神保ら[Jimbo et al., 2008]は，192 個のマイクロフォンアレイと CMOS カメラを使用し，RGB 画像上へ音高の帯域ごとの強さを重畳表示している．この可視化手法は，音源の分布を提示するが，空間的な聴覚アウェアネスである音源の位置や時間的な聴覚アウェアネスである音源の時間変化の提示は行っていない．

次に，深度センサを用いたマルチメディア統合に関して Even ら[Even et al., 2013]は，マイクロフォンアレイとレーザーレンジファインダを使用し，SLAM で作成した地図上に音源の位置を重畳表示している．この可視化手法は，音源の強さとレーザレンジファインダによって音源の位置を提示するが，音源の時間変化である時間的な聴覚アウェアネスの提示は行っていない．

これらの研究を受け井山ら[Iyama et al., 2014]は，マイクロフォンアレイと深度センサを使用し，聴覚アウェアネスを三層モデルで定義し，これを可視化するシステムを開発した．この三層モデルは，環境内の音の分布を概観する機能を提供する音源分布レイヤ，着目した音源の位置やパワーを抽出する音源位置レイヤ，新しい音源の出現や音源のパワーの大きな変化を抽出する顕著性レイヤから構成される．そのため，空間的· 時間的なアウェアネスの提示も行なっている．しかし，そのモデルの有効性が評価されていなかった．本稿では，聴覚アウェアネスの三層モデルを拡張し，可視化システムを開発し，その評価を行うことで三層モデルの有効性を確認する．

3　聴覚アウェアネス可視化のための三層モデル

聴覚アウェアネス可視化のための三層モデルは，音源分布レイヤ，音源位置レイヤ，顕著性レイヤの 3 つのレイヤから構成される．充実した聴覚アウェアネスの提示のため，音源位置レイヤは空間的な聴覚アウェアネスを，顕著性レイヤは時間的な聴覚アウェアネスを提示する．ユーザはこれらのレイヤを自由に切り替えながら，音環境の観察を行うことができる．各レイヤは，2 つの処理から構成される．はじめに，レイヤの入力データの可視化可能なデータへの変換や高次レイヤへのデータの受け渡しを行う．次に変換したデータから可視化画像の生成を行い，ユーザに提示する．次節以降で各レイヤの役割と処理について述べる．

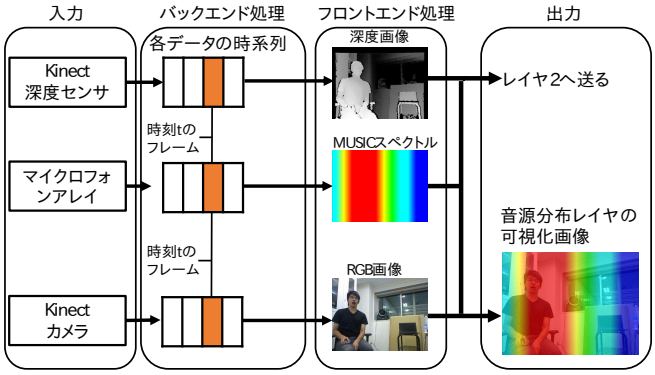


図 2: 音源分布レイヤの処理と可視化結果

3.1　音源分布レイヤ (レイヤ 1)

音源分布レイヤは環境内の音の分布を概観する機能を提供する．ユーザは環境内の音源の分布を MUSIC スペクトル[Asano et al., 2001]の色画像を RGB 画像に重畳した画像として提示される．ユーザが効率的に環境を概観するため，MUSIC スペクトルの色画像の濃淡や可視化する MUSIC スペクトルの範囲を変更することができる．ユーザが環境を観察したいとき，MUSIC スペクトルの色画像を RGB 画像に重畳した画像から環境全体を概観することができる．

音源分布レイヤのデータフローと処理は図 2 の通りである．入力データは RGB カメラから取得した RGB データ，深度センサから取得した深度データ，マイクロフォンアレイから取得した MUSIC スペクトルである．入力デバイスから取得したデータの時間同期を行い，各入力データの色画像への変換と，MUSIC スペクトルの色画像を RGB 画像に重畳した画像を生成する．MUSIC スペクトルの色画像は MUSIC スペクトルのパワーに対応した色を割り当てることで生成する．音源分布レイヤのこれらの処理によって，ユーザは環境内の音の分布を概観することができる．

3.2　音源位置レイヤ (レイヤ 2)

音源位置レイヤは環境内のユーザが着目した音源の音情報，すなわち空間的な聴覚アウェアネスをユーザに提示する．ユーザは着目した音源の RGB 画像上のみに MUSIC スペクトルの色画像を重畳した画像を提示される．ユーザは音源分布レイヤを用いて環境を概観した後，音源位置レイヤを用いて着目した音源のみの音情報を観察することができる．

音源位置レイヤのデータフローと処理は図 3 の通りである．まず，音源分布レイヤから送られる深度データからユーザが着目する音源物体の形状を推定する．着目する音源物体の形状は深度データに領域成長法[Ballard et al., 1982]を用いて算出する．そして，各入力データの色画像への変換と，MUSIC スペクトルの色画像を RGB 画像の着目した音源物体上に重畳した画像を生成する．音源位

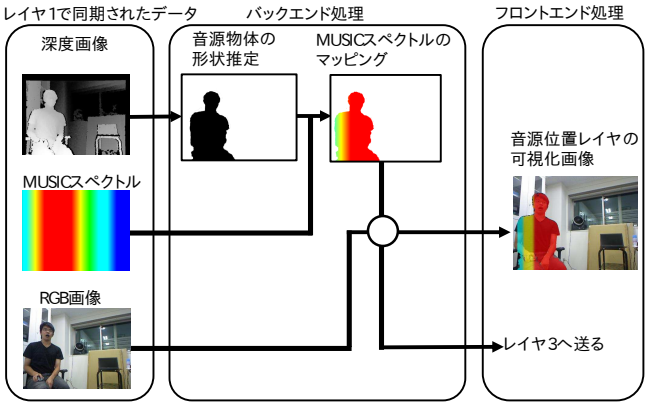


図 3: 音源位置レイヤの処理と可視化結果

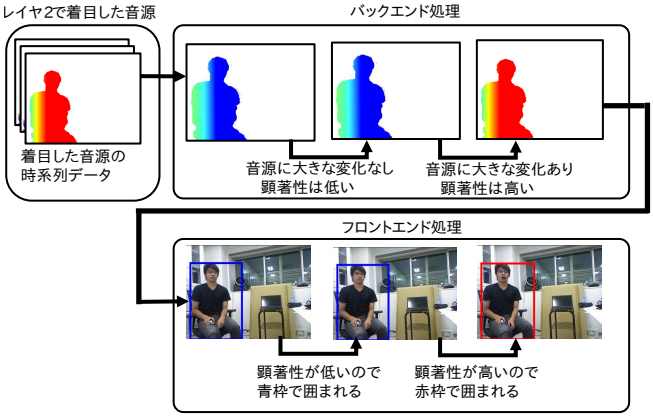


図 4: 顕著性レイヤの処理と可視化結果

置レイヤのこれらの処理によって，ユーザは環境内の着目した音源の音情報のみを観察することができる．

3.3 顕著性レイヤ (レイヤ 3)

顕著性レイヤは環境内のユーザが着目した音源の顕著性，すなわち時間的な聴覚アウェアネスをユーザに提示する．音源の顕著性とは，音響情報と音源の位置· 形状情報の時間変化によって定義される．音を発していなかった音源が音を発し始める場合や環境内に新しく音源が出現した場合は顕著性が大きくなる．一方，音源が音を発していない場合や，音源が発生している音に変化がない場合は顕著性が小さくなる．ユーザは RGB 画像に着目した音源の顕著性の大きさに対応した色枠を重畳した画像を提示される．ユーザは環境内の着目した音源の時間変化の様子を観察することができる．

顕著性レイヤのデータフローと処理は図 4 の通りである．まず，着目した音源物体の顕著性を算出する．顕著性 d_c はフレーム t とフレーム $t-1$ の音源物体の MUSIC スペクトルの変化量 l_m と位置の変化量 l_d の加重平均として求められる

$$d_c = \alpha \cdot l_d + (1.0 - \alpha) \cdot l_m. \tag{1}$$

l_m，l_d はカルバック· ライブラーダイバージェンスによって次のように求められる

$$\begin{cases} l_d = \frac{1}{2}\left[\log\frac{|\Sigma_{d_{t-1}}|}{|\Sigma_{d_t}|} + tr\{\Sigma_{d_{t-1}}^{-1}\Sigma_{d_t}\} \right. \\ \qquad \left. + (\mu_{d_t} - \mu_{d_{t-1}})^T\Sigma_{d_{t-1}}^{-1}(\mu_{d_t} - \mu_{d_{t-1}}) - 3\right] \\ l_m = \frac{1}{2}\left[\log\frac{\sigma_{m_{t-1}}^2}{\sigma_{m_t}^2} + \frac{\sigma_{m_t}^2}{\sigma_{m_{t-1}}^2} + \frac{(\mu_{m_t} - \mu_{m_{t-1}})^2}{\sigma_{m_{t-1}}^2} - 1\right] \end{cases}$$

ここで，Σ_{d_t}，$\Sigma_{d_{t-1}}$ はフレーム t，$t-1$ の深度データの共分散行列，μ_{d_t}，$\mu_{d_{t-1}}$ は深度データの平均，σ_{m_t}，$\sigma_{m_{t-1}}$ は MUSIC スペクトルの分散，μ_{m_t}，$\mu_{m_{t-1}}$ は MUSIC スペクトルの平均である．そして，顕著性の大きさに基づく色画像の生成と生成した色画像を RGB 画像に重畳した画像を生成する．顕著性レイヤのこれらの処理によって，ユーザは環境内の着目した音源の時間変化の様子を観察することができる．

表 1: インタフェースのパラメータ

レイヤ	パラメータ	概要	範囲
1	透明度 t	MUSIC スペクトルのパワー画像の透明度	$0 < t < 1$
	可視化の最小値 p	可視化する MUSIC スペクトルの最小値	$0 < p$
2	領域成長法の閾値 d	領域成長法で統合するかの閾値	$0 < d$
3	重み α	顕著性の重みパラメータ	$0 < \alpha < 1$

4 インタフェース設計

ユーザが要求する可視化結果を得るため，GUI の操作やパラメータの変更はスライダやボタンでなく，可視化画像に対するジェスチャを用いて行う．例えば，ユーザが着目したい音源対象を選択する操作は，画面内の画像を直接選択できるほうがより直感的である．本インタフェースはジェスチャの入力としてマウスの左クリック，右クリック，中クリック，マウスホイールを使用する．ユーザが要求する可視化結果を得るためのインタフェースの操作を簡易で直感的にできるよう設計する．

GUI は図 5 に示すように，画像表示部，ステータス部から成る．画像表示部に，三層モデルで生成した画像を組み合わせたものが表示される．ユーザは画像表示部上で三層モデルで使用するパラメータを変更することによって，表示する画像を変更できる．表 1 はユーザが変更できるパラメータの一覧である．レイヤごとにパラメータは存在し，ユーザは三層モデルにおける特徴量を柔軟に組み合わせることができ，自由に画像表示部に表示される描画結果を変更することができる．ステータス部は現在のレイヤやパラメータの変更内容などを表示する部分である．これによりユーザは現在のレイヤ状態やパラメータ状態を確認しながら操作することができる．以下では，各レイヤにおける操作について述べる．

図 5: インタフェースのデザイン

図 6: マウスホイールによる透明度の変更

4.1 音源分布レイヤのインタフェース設計

音源分布レイヤでは，ユーザは重畳される MUSIC スペクトルの色画像の濃淡と可視化する MUSIC スペクトルの帯域を変更できる (図 6)．これらの機能によって，ユーザは環境や要求に応じた可視化結果を得ることができる．例えば，音の分布を鮮明に観察するときは濃く，音の弱い部分が必要ないときは帯域の最小値を上昇させることができる．

これらのパラメータの変更は，いずれも増減であるため，マウスホイールの操作により変更する．マウスホイールを上に回転させると，色画像の透明度や可視化する MUSIC スペクトルの最小値が増大し，下に回転させると減少する．これら 2 つのパラメータのどちらを変更するかの切り替えは，右クリックで行えるようにする．

4.2 音源位置レイヤのインタフェース設計

音源位置レイヤでは，ユーザは着目する音源対象の選択と領域成長法の類似度の閾値パラメータの変更できる (図 7)．これらの機能によって，ユーザは着目したい音源を選択できる．

着目する音源の選択は，可視化領域内の着目する画像をマウスの左クリックにより行う．領域成長法の閾値パラメータが大きいと，より広い範囲に存在する複数物体を同一領域とみなし，小さいと，領域をより細かく分割する．このパラメータの変更は増減であるため，マウスホイールの操作により変更する．マウスホイールを上に回転させると，閾値パラメータは増大し，下に回転させると減少する．

図 7: クリックによる音源の選択

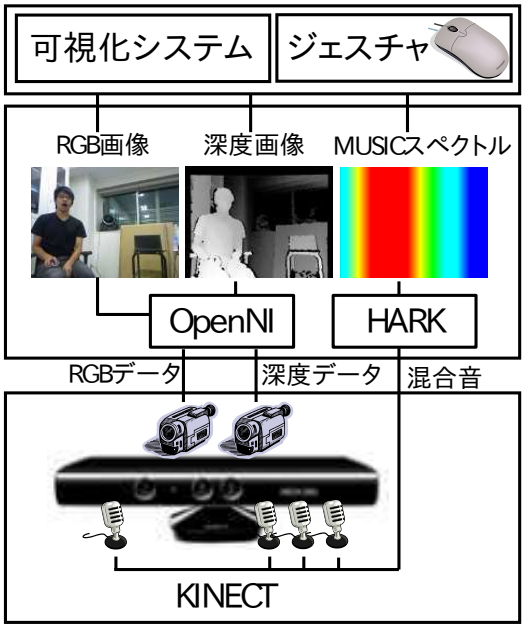


図 8: 聴覚アウェアネスの可視化システムの詳細

4.3 顕著性レイヤのインタフェース設計

顕著性レイヤでは，ユーザは顕著性を算出する際の音響情報と深度情報の重みの変更できる．音響情報の重みが大きくなるほど，着目した音源の音響情報の変化が顕著性に大きく影響し，深度情報の重みが大きくなるほど，着目した音源の移動量や形状の変化が顕著性に大きく影響するようになる．

これらのパラメータの変更は，いずれも増減であるため，マウスホイールの操作により変更する．また，音響情報と深度情報のどちらを変更したいかは，ユーザの要求に応じて変わるので，重みを変更する情報をマウスの右クリックで変更する．

各レイヤ間の移動については以下のように設計する．音源分布レイヤから音源位置レイヤへの移動は，音源位置レイヤで着目する音源を左クリックしたときに移動する．音源位置レイヤから顕著性レイヤへの移動は，音源位置レイヤで着目している音源の領域内を左クリックしたときに移動する．顕著性レイヤから音源位置レイヤ，音源位置レイヤから音源分布レイヤへの移動はマウスの中クリックを行うことで移動する．

5 実験

聴覚アウェアネス可視化のための三層モデルを用いて音環境の可視化を行い，三層モデルによって聴覚アウェアネ

(a) 音源が 2 個のときの音源の配置　(b) 音源が 4 個のときの音源の配置　(c) 音源が 8 個のときの音源の配置

図 9: 音源の配置と可視化動画の一例

スが提示されているかを被験者実験によって評価した．

5.1　システム構成

システム構成は図 8 の通りである．本システムへの入力データは Kinect を用いて取得した RGB 画像，深度画像，および多チャンネル音響信号である．RGB 画像と深度画像は OpenNI ライブラリ [sim, 2014]を，多チャンネル音は HARK(Honda Research Institute Japan Audition for Robots with Kyoto University)[Nakadai et al., 2010]を通じてそれぞれ取得され，システムに渡される．各レイヤのデータ処理や GUI への様々な画像の描画には Processing を用い，三層モデルの特徴量を柔軟に変化させることができるようシステムを設計する．

5.2　実験設定

実験では，空間的な聴覚アウェアネスの有効性を確認する．空間的な聴覚アウェアネスを考慮しない音源分布レイヤによる可視化結果と，音源位置レイヤによる可視化結果について，どちらが音の発生に即座に気づくかを比較した．音の発生から被験者の認識までの時間を比較するために，それぞれのレイヤで可視化された 30 秒程度の動画を被験者に視聴させ，あらかじめ指定した物体が音を発したと認識した時間を記録した．

視聴する動画は各レイヤについて，音源が 2 個，4 個，8 個の 3 種類，計 6 種類用意した (図 9)．各実験と音源の数，可視化するレイヤの対応は表 2 の通りである．音源の再生デバイスとしてはすべて同一のスピーカを使用した．使用した音源は，ATR の音素バランス文[Kurematsu et al., 1990]，RWC 音楽データベース[Goto et al., 2002]のクラシック曲，ホワイトノイズ，サイン波のテストトーンである．三層モデルに必要な各データは，Kinect から取得した 30fps の深度データ・RGB データと 4ch 同期，16bit 量子化，16kHz の音響信号を用いた．各動画の各音源の音の発生時刻は図 10 の通り．

実験手順は，音源が 2 個の動画の各レイヤによる実験，次は音源が 4 個の動画，最後に音源が 8 個の動画の実験という手順で行った．各実験でどちらのレイヤによる動画を視聴した順序による実験結果の偏りをなくすため，6

表 2: 各実験の音源数と可視化レイヤの対応

実験種類	音源数	可視化レイヤ
実験 1-1	2	音源分布レイヤ
実験 1-2	2	音源位置レイヤ
実験 2-1	4	音源分布レイヤ
実験 2-2	4	音源位置レイヤ
実験 3-1	8	音源分布レイヤ
実験 3-2	8	音源位置レイヤ

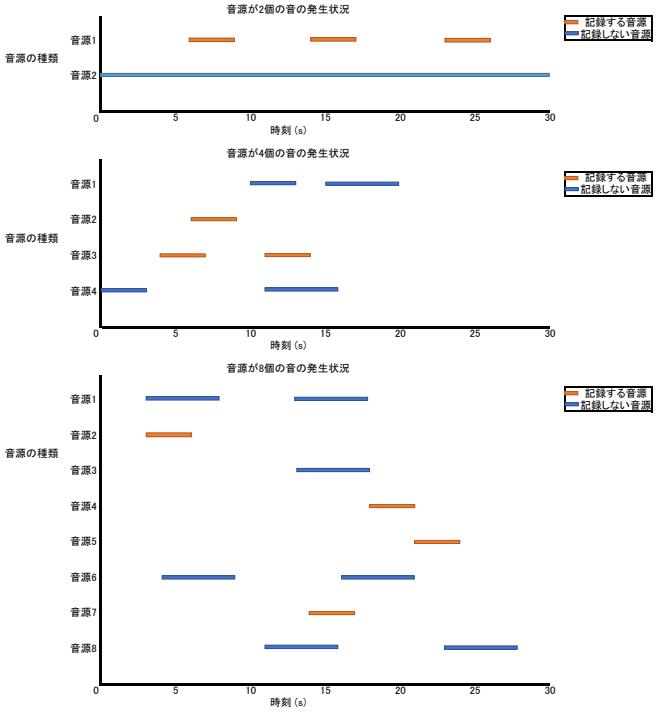
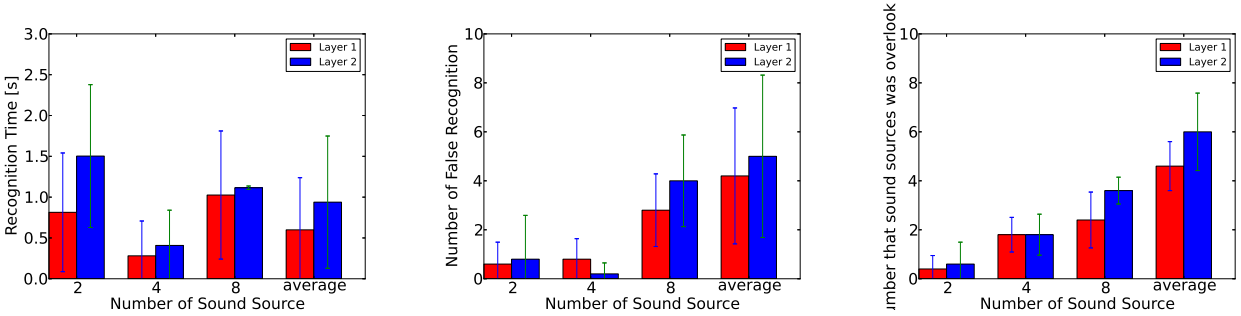


図 10: 各動画の各音源の音の発生時刻

人の被験者を 2 つのグループに分けて実験を行った．あるグループは，音源分布レイヤによる可視化動画による実験を行った後に音源位置レイヤによる可視化動画による実験を行い，一方のグループは音源位置レイヤによる可視化動画による実験を行った後に音源分布レイヤによる実験を行った．また，どの音源からどのような種類の音が発生するのかは事前に知らせておらず，常に未知の音を聞く状態にした．記録した時間が正しい範囲は，音源の再



(a) 各音源数の認識時間　(b) 各音源数の誤認識数　(c) 各音源数の見逃し数

図 11: 各音源数・レイヤの認識時間・誤認識数・見逃し数

生時間が 3 秒であるため正解時間から 3 秒間とした．

5.3 実験結果

図 11 に示すように，音源分布レイヤの場合の認識時間は音源位置レイヤの場合の認識時間よりも早い．平均認識時間はすべて正しく記録した時間の値のみから計算されており，見逃しや誤認識の時間は含まれていない．また，合計誤認識数，合計見逃し数ともに音源分布レイヤの場合のほうが少ない．合計誤認識数・合計見逃し数は音源数が 2 個，4 個，8 個の場合の合計である．

5.4 考察

平均認識時間・合計誤認識数・合計見逃し数の全ての値に関して，音源分布レイヤの結果が音源位置レイヤの結果より数値が小さく，どのような音が発生するか未知の場合では全体を概観する機能のほうが適していると考えられる．これは音源物体の大きさが画面に比べて小さく，音源定位の結果が少しでも誤った場合，正しく物体上に MUSIC スペクトルの色画像が重ねられないためであると考えられる．そのため，今後の実験においては実際の使用順序にしたがって，発生する音の種類を既知とした実験を行うべきだと考えている．また，正確なキャリブレーションや定位精度向上の手法に取り組む必要がある．

6 結論

本研究では，音源分布レイヤ，音源位置レイヤ，顕著性レイヤから構成される聴覚アウェアネス可視化の三層モデルを設計し，Kinect を用いた聴覚アウェアネス可視化システムを実装した．音源分布レイヤは環境内の音の分布を概観する機能を，音源位置レイヤは着目した音源情報を抽出する機能を，顕著性レイヤは音情報の時間変化，すなわち，新しい音源の出現や音源のパワーの大きな変化といった顕著性を抽出する機能を提供する．三層モデルに基づくデータ処理や可視化を行い，各レイヤのパラメータをジェスチャを用いて変更することで，音環境を分析するための直感的な操作が可能なインタフェースを開発した．被験者実験によって，三層モデルによって聴覚アウェアネスが提示されているかの実験を行った．その結果，発生する音の種類が未知の状況下では音源分布レイヤによる可視化が音源位置レイヤによる可視化より，高速な音源の認識や少ない誤認識を行うことができることを確認した．

今後，発生する音の種類を既知にするなど実験の情報を増やし，実際の使用に近い環境における実験を行い，再度モデルの有効性の評価や音源定位の精度向上などシステムの処理部分の改善を行う予定である．

謝辞 本研究の一部は科研費 No.24220006 と No.24700168 の支援を受けた．

参考文献

[Asano et al., 2001] F. Asano et al. Real-time sound source localization and separation system and its application to automatic speech recognition. In *INTERSPEECH*, pages 1013–1016, 2001.

[Ballard et al., 1982] D. H. Ballard et al. *Computer Vision*. Prentice Hall, 1982.

[Even et al., 2013] J. Even et al. Creation of radiated sound intensity maps using multi-modal measurements onboard an autonomous mobile platform. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3433–3438, Nov 2013.

[Goto et al., 2002] M. Goto et al. RWC Music Database: Popular, Classical and Jazz Music Databases. In *ISMIR*, volume 2, pages 287–288, 2002.

[Iyama et al., 2014] T. Iyama et al. Visualization of auditory awareness based on sound source positions estimated by depth sensor and microphone array. In *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on*, pages 1908–1913, Sep 2014.

[Jimbo et al., 2008] N. Jimbo et al. Visualization of sound environment using multi channel acoustic measurement system. In *Acoustic Society Symposium,2008*, pages 1509–1510, Sep 2008.

[Kurematsu et al., 1990] A. Kurematsu et al. Atr japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9(4):357–363, 1990.

[Nakadai et al., 2010] K. Nakadai et al. Design and Implementation of Robot Audition System 'HARK' – Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[sim, 2014] simple-openni - openni library for processing. https://code.google.com/p/simple-openni/, 2014.