人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-06

複数のマイクロホンアレイを用いた理科室における音源アクティビティの分析

Analysis of sound source activity in science classes using multiple microphone arrays

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
Jani Even（ＡＴＲ知能ロボティクス研究所）
塩見昌裕（ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Jani EVEN, Masahro SHIOMI, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, even@atr.jp, m-shiomi@atr.jp, hagita@atr.jp

Abstract – We are developing a dialogue behavior recognition platform, which is able to detect who is talking, where and when, based on 3D sound direction estimation by multiple microphone arrays, and human tracking technologies. We installed the developed system in a science room of an elementary school, and collected data including real science classes during a period of one month. In the present paper, we present preliminary analysis results on the sound activities of the science room.

1　はじめに

我々は、3次元空間での音源方向と、人位置の推定情報を組み合わせることにより、誰が、いつ、どこでしゃべっているのかを推定する対話行動認識プラットフォームを開発している [1]。

このようなシステムを利用することにより、教室内や会議などのように、複数の人が時に席を移りながら会話や協調作業をする際のデータの観察が容易になることが期待できる。

マイクロホンアレイ処理による音源定位に関する研究はこれまで多くされてきたが[2-6]、マイクロホンアレイを単体で扱うことが多い。その中でも、3次元空間での音源定位に関するものは比較的少ないが、実環境で対象となる音源のアレイに対する仰角が固定できない場合は、方位角のみならず仰角も推定することが重要となる。また、音源とアレイとの距離に関しては、理論上は推定可能であるが、角度推定に比べて精度は低く、処理時間も膨大となってしまう。

また、教室のような広い空間の音源をカバーするためには、その空間の複数の箇所にマイクロホンを配置する必要がある。[5]のように、一つのキャプチャで同期させた96個のマイクロホンを空間内に配置する方法もあるが、コストパフォーマンスの問題も生じる。

上述の問題点を踏まえ、我々は複数のマイクロホンアレイを用いて空間的に情報を統合し、3次元空間で精度よくかつ効率よく音源定位を行う枠組みを提案した[7]。壁や天井での反射の利用も試みてきた[7]。レーザ距離センサも利用し、マイクロホンアレイと組み合わせた枠組みも検討してきた[8]。

本研究では、マイクロホンアレイ処理や人位置検出においてこれまで開発してきたシステムを、小学校の理科室に設置し、実際の理科の授業が行われたデータを収集した。本論文では、システムの紹介と、理科室で観測された音源のアクティビティについて、予備的な分析結果を報告する。

本論文は以下のように構成される。次ぐ2章では、開発したシステムの概要を説明する。3章では、小学校理科室でのデータ収集と分析結果について述べる。4章で考察と今後の課題を記す。

2　開発したシステムの概要

図１に理科室の様子を示す。理科室に机は全部で8つあるが、そのうち実際授業に使用されているのが前方の6つであるため、6つのマイクロホンアレイをこれらの机の上に設置した。それぞれの机に対するアレイの位置は、学校側と相談の上、生徒たちの視界の妨げとならないよう、かつ先生が頭をぶつけないように、2メートル程度の高さに、机の流し台の真上に、天井からマイクロホンアレイを吊るした（図１上部参照）。また、人位置検出に使用するセンサとして、Kinectを多数天井に設置した。

図１．データ収集を行った理科室の様子

図２に開発したシステムの概要図を示す。まず、複数のマイクロホンアレイにおいて、それぞれ3次元空間の音源方向推定（方位角および仰角の推定）を行う。多くの音源定位の研究では、方位角のみが推定されるが、教室のように人の数が多い場合、同じ方向に複数の音源が存在する確率が高くなり、仰角の推定も重要となる。3次元空間における方位角および仰角を求めるため、マイクロホンアレイとして、16個のシリコンマイクが直径30㎝の半球面上に配置するようなアレイフレームを作成した。図３にマイクロホンアレイのマイク位置情報を示す。

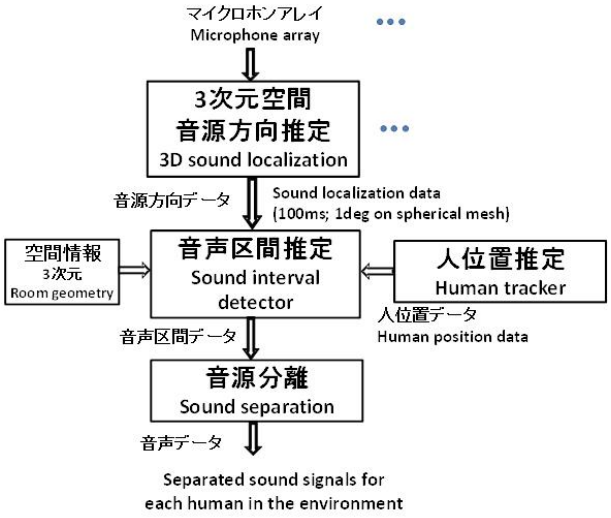


図２．開発したシステムの概要図

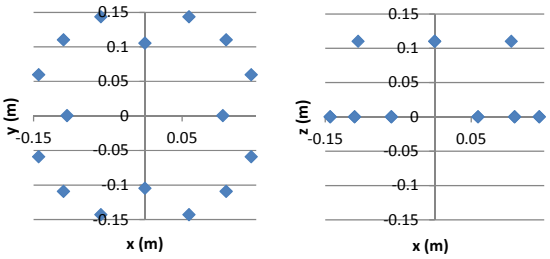


図３．マイクロホンアレイのマイク位置情報

音源方向推定部には、著者らが開発した実時間処理で3次元空間での音源方向を5度の空間的分解能および100msの時間分解能で推定するシステムを用いた[4]。音源方向推定は、空間的分解能が高いMUSIC（Multiple Signal Classification）法に基づいている（付録を参照）。周波数帯域は、アレイの形状を踏まえて、1000～5000Hzを使用している。アレイは2メートルの高さに設置しているため、方位角は0～360度で、仰角は0~-90度とした。実時間処理でMUSIC 法に基づいた3次元空間での音源方向推定を可能にするため、フレーム長を64点(4ms)としている。音源数の推定も難しいため、3に固定して、MUSICスペクトルで2.5dBの閾値を上回ったピークのみを探索している。

人位置検出部には、天井に設置した多数のKinectセンサによる3次元の人位置推定を用いている[9]。レーザ距離センサによる2次元の人位置推定も一つの選択肢であったが[8]、理科室で対象となる生徒の数が多く、センサも天井に設置した方が望ましいため、Kinectセンサによる手法を採用した。

音声区間推定部では、音源方向と人位置情報を基に、その人が発話しているか否かを判断する。部屋の空間情報とアレイの位置情報を基に、それぞれのアレイから得られた音源方向と、人位置推定部から得られる人の位置情報を重ね合わせる。検出された音源方向が、検出された人の口元の位置と重なった場合、その人が発話している確率が高いとみなす。本研究で用いた3次元の人位置検出は、空間内の2次元位置と身長を推定することが可能であるが、身長の推定は比較的精度がよくないため、口元の位置を、子供が座っている場合の80cmから大人が立っている場合の170cmに制限した。

人位置は33~66msごとに推定され、音源方向は100msごとに推定されるため、100msの時間分解能で音声区間が検出できる。

最後に、検出されたそれぞれの音源区間に対し、音源に最も近いマイクロホンアレイを用いて、検出された方向にビームを当て、音源分離を行う[8]。

3　データ収集および分析結果

3.1 データ収集

およそ１ヵ月に渡り（2013 年 2 月）、開発したシステムを用いて理科教室の授業時間を含むデータ収集を行った。各クラスの生徒の数はおよそ 30 名で、先生はクラス担当と理科担当の 2 名である。本論文では、そのうちの 1 日の授業における予備的な分析結果を示す。

図４に 6 つのアレイにより理科室で測定された音源方向推定結果の例を示す。点線は 1 メートル間隔で表示している。それぞれのアレイから出る直線は検出された方向を示し、線が出ていない丸は検出された人位置を表している。左図は、教室の前方で先

生が説明している場面で、右図は、実験中、2 列目と 3 列目の左側の机の生徒が同時に声を発している瞬間を示している。音源方向の線の色は高さ情報を表している。緑は 0 ~ 0.5 m、水色は 0.5 ~ 1.0 m、青は 1.0 ~ 1.5 m、ピンクは 1.5 ~ 2.0 m に対応する。複数のアレイから推定された方向が特定の位置で重なっていることが確認できる。左図ではピンク色で交わり、先生の口元の高さが 1.5m 以上であることに対応している。右図では、いずれも水色と青の境界周辺で線が交わっていることが分かるが、子供が椅子に座った時の口元の高さが 1m 弱であることに対応している。これらの例より、それぞれのアレイによる音源方向推定は、方位角のみならず、仰角も精度よく推定できていることが確認できる。

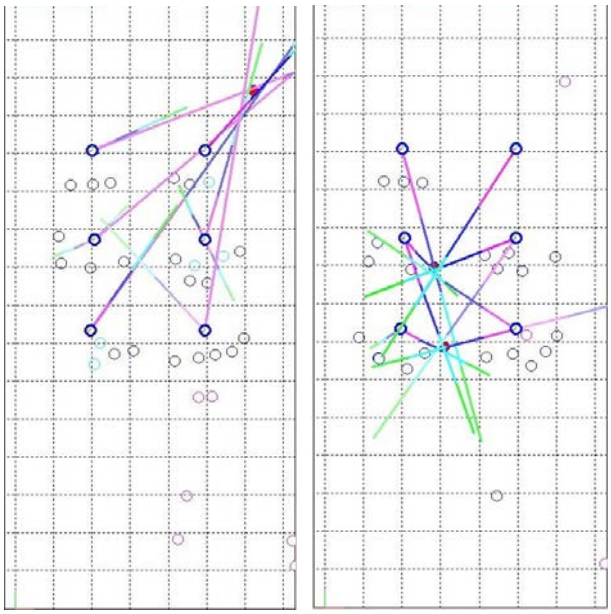


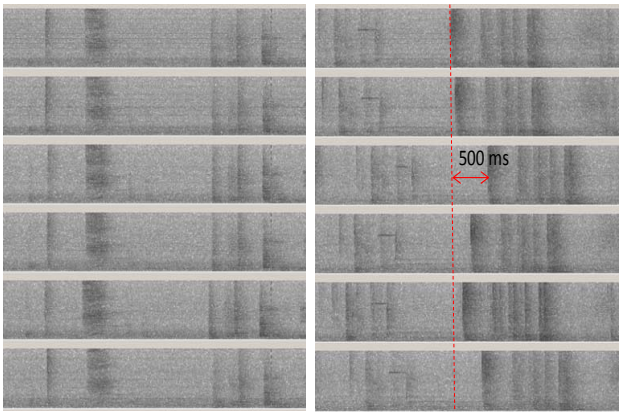
図４．6 つのマイクロホンアレイにより、理科室で測定された音源方向推定結果の例

人位置検出においては、位置検出の精度はそれなりに出ていたが、追跡に失敗することが多く、特定の人と音声発話を対応付けるまでは至っていない。特に生徒達の距離が近くなると追跡が難しく、一旦検出がされず数秒後に再度検出されて別の ID が割り当てられるようなケースが多かった。人位置追跡においては、現在研究開発が進められており、本論文では、アレイデータのみから得られる教室内の音源アクティビティについて分析結果を示す。

3.2 多キャプチャのデータの同期における注意事項

オフライン処理に関する問題点として、多チャンネルオーディオキャプチャデバイスのクロックが異なるため、長時間録音すると、徐々にキャプチャ間で時間ずれが生じることを観測した。

図５にキャプチャ間の時間ずれの例を示す。



(a) 8:50　　(b) 14:20

図 5. 異なった時刻における 6 つのマイクロホンアレイのスペクトログラム（0~6kHz）の例：キャプチャデバイス間のクロックの違いによる時間ずれ

午前 8 時 50 分頃にシステムを起動した際には 6 つのキャプチャのスペクトログラムで突発的な雑音による縦線が揃っていることが分かるが、午後 2 時 20 分頃にシステムを終了した際には、キャプチャ間で最大 500 ms 程度の時間ずれが生じていることが観測された。音源方向推定は 100 ms の分解能であることを踏まえると、この時間ずれは無視できない。オンライン処理では、それぞれのキャプチャからデータが届いた時刻を基に同期を行えば、ネットワーク遅延のみで多キャプチャのデータ同期には比較的影響は小さいが、オフライン処理の場合は、上述のキャプチャ間のクロックの違いにより、時間補正を行う必要がある。

3.3 理科教室の音源アクティビティの分析結果

図６に、６つのアレイにおける音源方向推定結果の例を示す。先生と生徒達が実験についてインタラクションを行っている際の 20 秒間の区間を表示している。各パネルの縦軸は方位角を示し（上半分は 180 ~ 0 度、下半分は 0 ~ -180 度）、色が仰角の違いを表している（赤が-90 ~ -67.5 度、ピンクが-67.5 ~ -45 度、青が-45 ~ -22.5 度、水色が-22.5 ~ 0 度；-90 度が真下方向、0 度が水平方向を差す）。丸は検出された音源方向を表す（時間間隔は 0.1 秒である）。

検出された音源方向において、各パネルの下半分で、仰角を示す色がピンクか青の横線は、各机の周りに座っている生徒達の音源アクティビティに対応している。各パネルの上半分の水色の横線は、教室の前方で先生が発話している区間、または机の前方の音源アクティビティに対応している。この場面では、すべての机の周りで、数名の生徒達が発言していることが分かる。

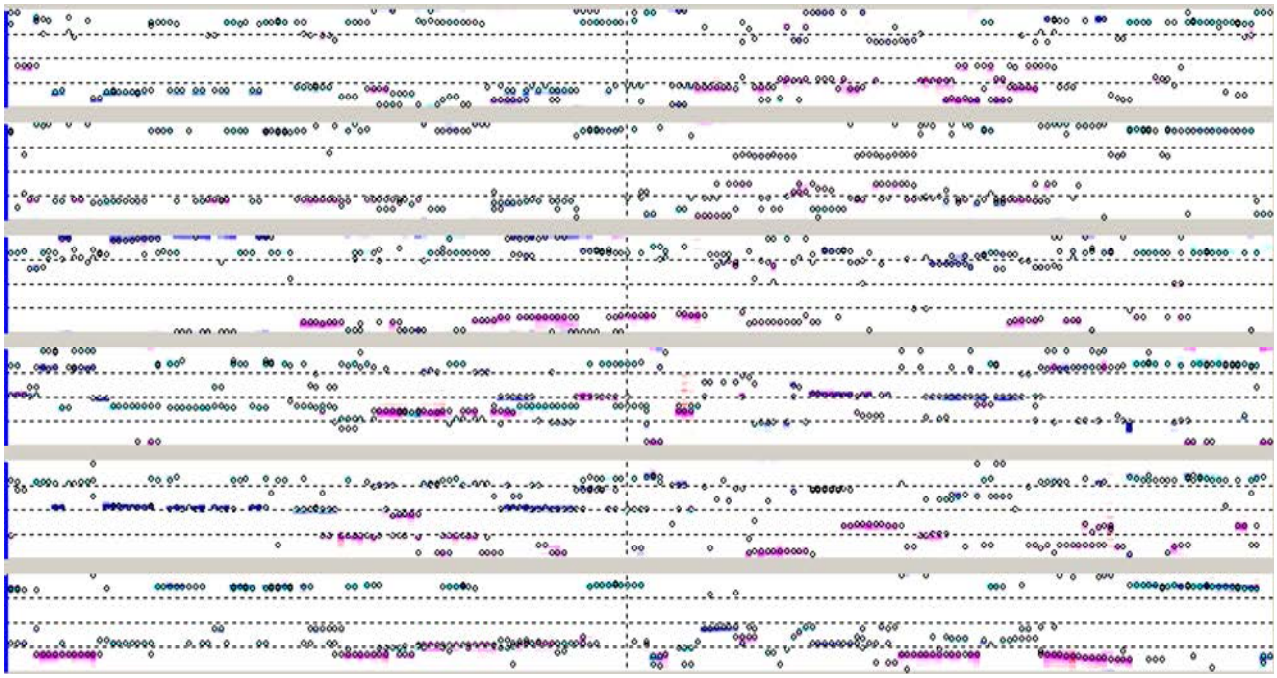


図６　６つのアレイにおける音源方向推定結果の例：先生と生徒達が実験についてインタラクションを行っている場面（20 秒間）

図４に示した例は、ある瞬間（0.1 秒以内）の音源アクティビティを空間的に表示したものであり、図６に示した例は、20 秒間における音源アクティビティの変化を示したものである。しかし、膨大なデータが蓄積された際に、もう少し長いスパンでデータを表示することも重要であると考えられる。

そこで、各机の周りの音源アクティビティのおおまかな流れを観測するため、5 分刻みに特定方向の範囲内に発生している音源アクティビティを集計（定量化）することとした。

音源アクティビティの集計には、0.1 秒ごとに算出される音源方向推定結果を用いて、仰角を-25 度～-85 度の角度領域において、5 分間（300 秒）の区間に対し、対象の方位角の範囲内（10 度間隔）に音源が検出された回数を 0.1 秒で掛ける。また、0.1 秒以内の突発的な音によるもの（足音や机に物を置いたときの音など）は、孤立した点を削除することにより、音源アクティビティの集計から除外している。

仰角においては、0 度が水平方向で-90 度が真下の方向を差すが、-25 度に制限することにより、隣の机の音源アクティビティの影響を避けるようにしている。また、-85 度の制限は、多チャンネルキャプチャの同位相の雑音による誤検出を避けるためであり、アレイの真下方向に位置する流し台周辺の音源アクティビティを観測しないこととなる。

図７に各机のマイクアレイで計測された 1 日分の収録に対する音源アクティビティの時系列ヒストグラムを示す。横軸の時間分解能を 5 分刻みとし、縦軸は方位角で分解能を 10 度刻みとしている。それぞれの時刻と方位角における音源アクティビティの集計秒数を 15 秒刻みで色別に表示している。

アレイの位置および向きにより、方位角が 0～180 度（各パネルの上半分）は、教壇側の音源アクティビティを反映し、-180～0 度（各パネルの下半分）は生徒達が座っている机の周りの音源アクティビティを反映している。

図７には、午前中４クラス（8:50～9:35、9:40～10:25、10:35～11:20、11:25～12:10）、お昼休みを挟んで午後の１クラス（13:05～13:50）を含む音源アクティビティが表示されている。

まず、8:50 までの授業前のアクティビティはすべてのアレイで低いことが分かる。左上のアレイでは、80 度周辺に強いアクティビティを持つ音源が観測されているが、これは教室の左前の角にヒーターが作動し、その定常雑音が観測されたものである。

授業中、教室前方の両アレイで、正の角度（0~180）で 15 秒以上のアクティビティが発している区間が観測できるが、これは先生が教壇周辺で説明をしている時間帯となる。

また授業時間内に、全アレイにおいて、負の角度（-180~0）の領域で 15 秒以上のアクティビティが発している区間が複数観測できる。これは理科の実験中、机の周りの生徒達のアクティビティを反映している。机とクラスによって、アクティビティが高い方向が異なることが分かる。

クラスとクラスの間の休憩時間およびお昼休み時間では、音源アクティビティが低くなっていることが観測できる。またお昼休み時間には右前のアレイで 130 度周辺の方向に強いアクティビティが観測されている。これは校内に流れていた音楽が教室の前方の右側のドアから漏れてきていたことを反映している。

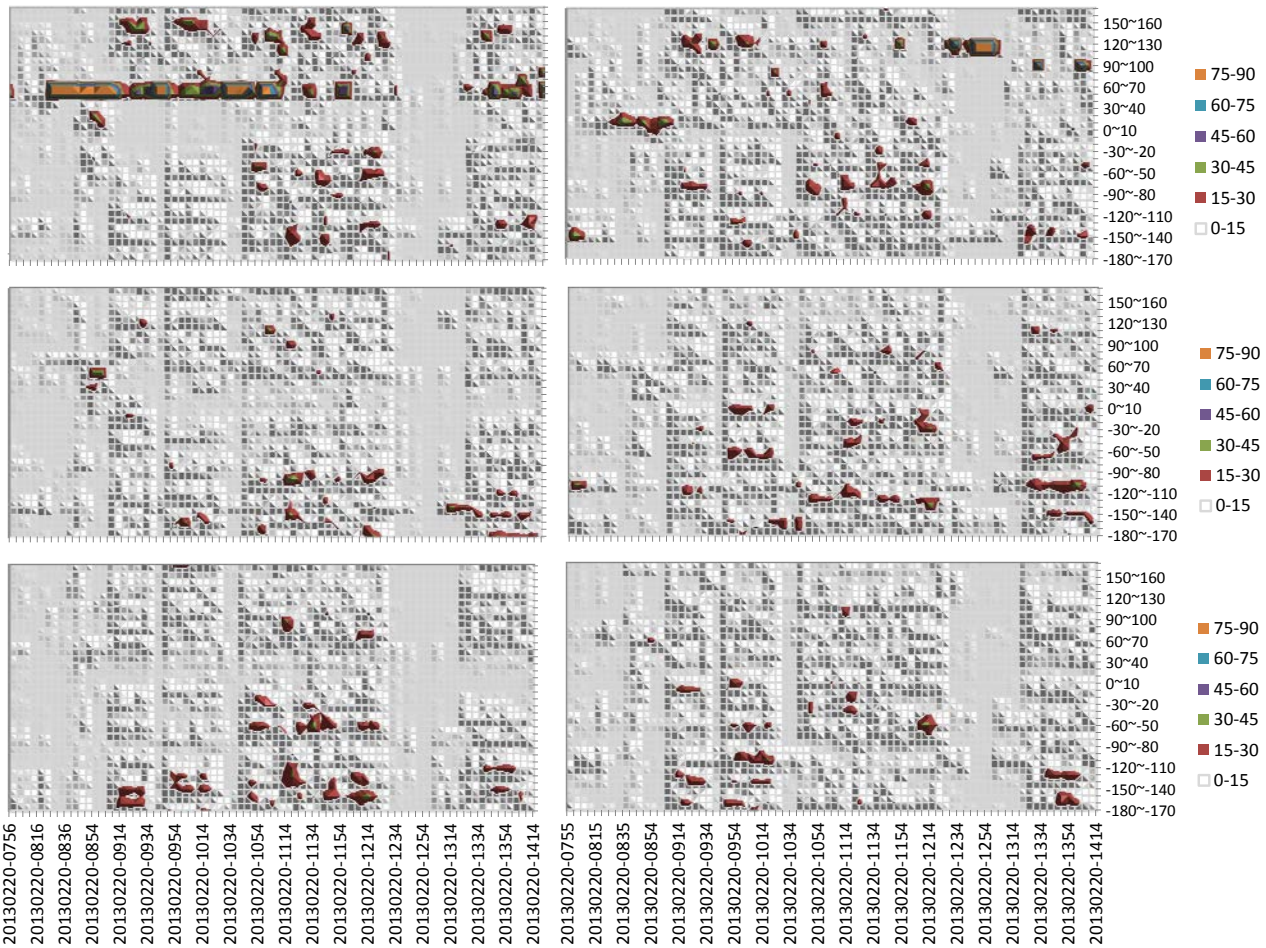


図７．６つのアレイ（上図は教室の前方から 1 列目の机、中央図は 2 列目の机、下図は 3 列目の机）による音源アクティビティの時系列ヒストグラム（横軸：日付-時間（YYYYMMDD-HHMM の形式）を 5 分刻みで；縦軸：方位角を 10 度刻みで；色別で 5 分以内の音源アクティビティの時間を 15 秒刻みで表示）

4　考察

本論文では、小学校の理科室の６つの机に設置したマイクロホンアレイによる音源アクティビティの分析を行った。

図７のようなおおまかな音源アクティビティの表示より、理科室内のおおまかな状況が把握可能であり、特定の時間帯におけるより詳細な音源アクティビティの探索が容易となる。また図６のような表示で詳細な音源アクティビティの区間が可能となり、図４のように空間的にどこで音が鳴ったのかが表示できる。似たような音や似たような声では、空間情報がその識別に重要である。

おおまかな音源アクティビティの分析より、クラスと机によって、音源アクティビティが変化することが観測された。例えば、左前の机のように特に目立った音源アクティビティがない机も観測されたが、音声アクティビティの高い生徒をこの机に席替えして、議論を活発化させるなど、クラス活動の助けとして利用できることも考えられる。あるいは、ロボットが先生のお手伝いとして教育現場への活用が可能になれば、アクティビティの低いグループを音環境知能システムが感知し、ロボットが積極的にそのグループに近づいて支援するような用途も考えられる。

現時点では、音源の方向のみに基づき、音声以外の音もアクティビティとして集計されている可能性もあり、机の周りのおおまかな分析に留まっている。しかし、これらの方向と人位置検出が結びつける段階まで研究開発が進めば、先生および生徒達の音声アクティビティが測定可能となる。これは今後の課題となる。

付録：MUSIC 法

M個のマイク入力のフーリエ変換$X_m(k,t)$は、式(1)のようにモデル化される。

$$\mathbf{x}(k,t) = [X_1(k,t),..., X_M(k,t)]^{\mathrm{T}} = \mathbf{A}_k\, \mathrm{s}(k,t) + \mathbf{n}(k,t) \quad (1)$$

ベクトル$\mathrm{s}(k,t)$はN個の音源のスペクトル$S_n(k,t)$から成る：$\mathrm{s}(k,t) = [S_1(k,t), \ldots, S_N(k,t)]^{\mathrm{T}}$ 。kとt はそれぞれ周波数と時間フレームのインデックスを示す。ベクトル$\mathbf{n}(k,t)$は背景雑音を示す。行列 $\mathbf{A}_k$ は変換関数行列であり、(m,n)要素はn番目の音源からm番目のマ

イクロホンへの直接パスの変換関数である。$\mathbf{A}_k$ のn列目のベクトルをn番目の音源の位置ベクトル（steering vector）と呼ぶ。

まず、式(2)で定義される空間相関行列$\mathbf{R}_k$を求め、式(3)に示す$\mathbf{R}_k$の固有値分解により、固有値の対角行列$\mathbf{\Lambda}_k$および固有ベクトルから成る$\mathbf{E}_k$が求められる。

$$\mathbf{R}_k = E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)] \qquad (2)$$

$$\mathbf{R}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^{-1} \qquad (3)$$

固有ベクトルは $\mathbf{E}_k=[\mathbf{E}_k^{s} \mid \mathbf{E}_k^{n}]$ のように分割出来、$\mathbf{E}_k^{s}$と$\mathbf{E}_k^{n}$はそれぞれ支配的なN個の固有値に対応する固有ベクトルと、それ以外の固有ベクトルである。

MUSIC空間スペクトルは式(4)と(5)で求める。rは距離、θとφはそれぞれ方位角と仰角を示す。式(5)は、スキャンされる点 (r,θ,φ) における正規化した位置ベクトルである。

$$P(r,\theta,\varphi,k) = \frac{1}{|\ \tilde{\mathbf{a}}_k^H(r,\theta,\varphi)\mathbf{E}_k^n\ |^2} \qquad (4)$$

$$\tilde{\mathbf{a}}_k(r,\theta,\varphi) = \frac{\mathbf{a}_k(r,\theta,\varphi)}{\left\|\mathbf{a}_k(r,\theta,\varphi)\right\|} \qquad (5)$$

空間スペクトル（本稿ではMUSIC応答と呼ぶ）は、MUSIC空間スペクトルを式(6)のように平均化したものである。

$$\overline{P}(r,\theta,\varphi) = \frac{1}{K}\sum_{k=k_L}^{k_H} P(r,\theta,\varphi,k) \qquad (6)$$

k_Lとk_Hは、周波数帯域の下位と上位の境界のインデックスであり、$K = k_H - k_L + 1$。音源の方位は、MUSIC応答のN個のピークから求められる。

謝辞

本研究は、MEXT 科研費 21118003 及び 21118008 の助成を受けたものである。実験にご協力いただいた京都府精華町立東光小学校の皆様、および実験に参加いただいた児童・保護者の皆様にお礼申し上げる。

参考文献

1) 宮下敬宏, J. Even, P. Heracleous, 石井カルロス, 塩見昌裕, 萩田紀博.「対話行動認識プラットフォームを利用したオーバーラップする発話での話者同定」 日本ロボット学会第30回記念学術講演会講演論文集, RSJ2012, 4M1-4, 2012
2) Y. Sasaki, S. Kagami, H. Mizoguchi, T. Enomoto "A predefined command recognition system using a ceiling microphone array in noisy housing environments," in *Proc. of IROS 2008*, Nice, France, 2008, pp. 2178–2184.
3) K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. of IROS 2009*, St. Louis, USA, 2009, pp. 664–669.
4) C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, St. Louis, USA, 2009, pp. 2027–2032.
5) H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, Y. Hasegawa, "Real-time sound source orientation estimation using a 96 channel microphone array," in *Proc. of IROS 2009*, St. Louis, USA, pp. 676-683.
6) R. Chakraborty, C. Nadeu, T. Butko, "Detection and positioning of overlapped sounds in a room environment," in *Proc. of Interspeech 2012*, Portland, USA, 2012.
7) C. Ishi, J. Even, N. Hagita, "Using multiple microphone arrays and reflections for 3D localization of sound sources," in *Proc. of IROS 2013*, Tokyo, Japan, 2013
8) J. Even, C. T. Ishi, P. Heracleous, T. Miyashita, N. Hagita: "Combining laser range finders and local steered response power for audio monitoring," Proc. IROS 2012: 986-991, 2012.
9) H. Kidokoro, T. Kanda, D. Brscic, and M. Shiomi, "Will I bother here? - A robot anticipating its influence on pedestrian walking comfort," Proc. HRI2013, 2013