講義自動撮影における話者位置推定のための視聴覚情報の統合

正員 西口敏司
非会員 東和秀
非会員 亀田能成
非会員 角所考
非会員 美濃導彦

電気学会論文誌C
(電子・情報・システム部門誌)
平成16年3月号抜刷
IEEJ Trans. EIS, Vol. 124, No. 3, 2004

論　文

講義自動撮影における話者位置推定のための視聴覚情報の統合

正　員　西口　敏司*　非会員　東　和秀**
非会員　亀田　能成***†　非会員　角所　考***
非会員　美濃　導彦***

Audio and Visual Information Integration for Speaker's Localization in Automatic Shooting of Lecture

Satoshi Nishiguchi*, Member, Kazuhide Higashi**, Non-member, Yoshinari Kameda***, Non-member, Koh Kakusho***, Non-member, Michihiko Minoh***, Non-member

It is useful for automatic video shooting in a lecture room to estimate the location of a speaker in the lecture room. The captured videos are used for distance learning and lecture archiving systems. In order to estimate the location of a speaker in a wide lecture room, multiple cameras and multiple microphones are used. However, it is difficult to estimate the precise location of a speaker using only visual or acoustic sensors because of calibration problems, noise, and other interference. Therefore, we propose a method that integrates audio and visual information from a speaker in the lecture room. A lecturer's cell and a student's cell ared introduced as a unit of estimation of the location of a speaker. We defined 120 cells in a real lecture room and our multi-modal method were applied to the cells. The estimation accuracy of the location of a speaker is sufficient for automatic video shooting of a speaker in a lecture room by our integrating method.

キーワード：話者位置，情報統合，自動撮影，複数マイク
Keywords: speaker localization, multi-modal information integration, automatic shooting, multiple microphones

1. はじめに

教育現場へのIT(Information Technology)導入の試みとして，遠隔講義や，講義の様子を電子的なメディアとして記録する講義のアーカイブ化などが盛んとなっている。この対象となる講義は，通常，大学などでよく行われている一斉授業型の講義であり，一人の講師が多数の受講者を相手に，板書やスライドを用いて教示内容を説明する一方，それに対して受講者が質問するという形態となる。このうち，板書やスライドは，オンライン型の電子白板やスライドを利用すれば，伝送や記録は比較的容易であることから，遠隔講義や講義のアーカイブ化の実現には，講師や受講者の様子を伝えるための映像の自動撮影が重要な技術的課題となり，このための研究が活発に行われている[(1)(2)]。

講師や受講者の映像を伝える際に最も重要となる撮影対象は，発話中の講師や質問中の受講者といった話者であり，その自動撮影には，各時刻での話者位置を求める必要がある。本稿では，このための話者位置推定手法について議論する。

講義の自動撮影に関する従来研究の多くは，観測カメラから得られる視覚情報のみに基づいて撮影すべき人物を選択しており[(3)(4)]，音源としての話者を撮影するには，利用している情報が適切であるとはいえない。また，講義室にカメラを設置する場合，講義室の壁面には窓や扉による物理的制約が多いことに加え，映像撮影の際には背景となるため，その見苦しさを考慮すれば，カメラは天井付近にのみ設置することが望ましい。ところが，講義室のような大きな空間において，そのようなカメラ配置で得られる映像から，撮影すべき人物の3次元位置を高精度で安定して推定することは，必ずしも容易ではない。

* 京都大学大学院 法学研究科
〒606-8501 京都市左京区吉田本町
Graduate School of Law, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
** 京都大学大学院 情報学研究科
〒606-8501 京都市左京区吉田二本松町
Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
(現：株式会社オージス総研)
*** 京都大学学術情報メディアセンター
〒606-8501 京都市左京区吉田二本松町
Academic Center for Computing and Media Studies, Kyoto University
Yoshida Nihonmatsucho, Sakyo-ku, Kyoto, 606–8501 Japan
† (現：筑波大学 機能工学系)

一方，音響処理の分野では，従来から既に，マイクロホンアレイを利用した音源定位手法[5]~[7]が提案されており，これを利用した会議の自動撮影の試みも報告されている[8]。しかし，このような手法を講義の自動撮影にそのまま適用しても，必ずしも有効とは限らない。マイクロホンアレイによる音源定位では，一般に，マイクロホンから音源までの距離がマイクロホン間の距離に対して十分短いことが要求されるが，講義室のような広い空間では，観測カメラの場合と同様に，マイクロホンアレイは天井近くに設置した方が望ましい。その場合，音源とマイクロホンの間の距離が長くなるため，音源位置を高精度で特定することは難しい。実際，上の文献[8]においても，4名程度の少人数による会議の自動撮影を対象に，マイクロホンアレイを話者の目前1.5m以内に配置している。

さらに講義室内には，映像教材の音声，空調機器や視聴覚機器のファンによるノイズ，携帯電話の着信音など，話者以外の音源が存在し得るため，特に上のようにマイクロホンアレイの設置条件の制約のために音源定位や音源分離の精度が期待できない状況では，このような聴覚情報のみによるアプローチも，話者位置推定のための手法としてはやはり限界があるといえる。

そこで本研究では，数十名～数百名が収容できる程度に規模が大きく，かつ話者以外の音源が存在し得るような空間における話者位置推定の実現を目指す。講義室における講師や学生の存在場所に関する制約を導入することによって，講義室の天井近くに配置されたマイクロホンアレイとカメラから得られる視覚情報，聴覚情報それぞれから，人物位置推定及び音源位置推定を実現し，さらに両者の結果を統合する手法を提案する。

視覚情報と聴覚情報の併用は前出の文献[8]でもなされているが，そのねらいは話者位置推定の実現ではなく，撮影指針としてどの被写体を撮影すべきかという各被写体の重要性を決めるための情報として，視覚情報としての映像特徴と聴覚情報としての音源位置を併用するというものである。従って，このときの映像特徴と音源位置はそれぞれ十分な信頼性を持って独立に獲得できることが前提となっており，各々を獲得するための映像処理，音響処理が誤りを含む場合は考慮されていない。

これに対し本研究では，実際に講義が行われる環境を適用対象とするため，カメラに基づく人物位置推定とマイクロホンアレイに基づく音源位置推定それぞれにある程度の誤りが含まれる可能性があることを前提とする。これら2つの推定結果を統合し，人物かつ音源としての話者位置を，講義の自動撮影に必要な精度で推定することを目標とする。

統合方法として，まず本研究では，講義室における話者を人物でありかつ音が出ているものであると定義する。しかし，人物であることを判断するための動きに関する特徴と音響に関する特徴は，それぞれ全く性質が異なる。また，実際の講義室には，動きだけ大きなものや，音だけ大きなものが話者以外に存在する場合など，話者の持つそれぞれの特徴が，それぞれの性質が持つ側面から，必ずしも大きな値を持っているとは限らない。そこで，それぞれの情報を統合する際に，人物が存在していると判断された位置であり，かつ音が存在している位置であるとみなす単純な方法では，正しく話者の位置を推定できない可能性がある。そこで本研究では，各情報の統合の際に，それぞれの推定結果をそのまま用いるのではなく，それぞれの推定手法の精度に関する確信度を定義し，この確信度に基づいて情報統合することによって，自動撮影に必要な程度の精度で話者の位置を推定する手法を提案する。

本論文は以下のような構成である。まず2.章では，講義の自動撮影のための話者位置推定に関して，想定する話者や，講義室の環境について述べ，要求される推定精度について議論する。3.章では，この議論を踏まえ，聴覚情報と視覚情報を統合して話者位置を推定するための具体的な手法を提案する。4.章では，この手法を用いて，実際の講義室で話者位置推定精度を評価した結果について報告し，5.章で本研究のまとめと今後の課題について議論する。

2. 講義自動撮影のための話者位置推定

一斉授業型の講義では，一度に数十名～数百名が受講者となるため，講義室は一辺が数メートル～数十メートルのかなり広い空間となる。講義室の前部には，通常，黒板やスライド投影スクリーンなどの教材提示用設備，および教卓が設置されており，その周辺に1人の講師が立って，教材を使いながら教示内容を講述することになる。一方，講義室の中・後部には受講者が座る座席が設置されており，受講者はこのいずれかに着席して講義を受講する。そこで本研究では講師と受講者の存在し得る領域を，黒板やスクリーン，教卓等のある講義室前部，および座席のある講義室中・後部，にそれぞれ限定し，これらの領域を**講師存在領域**，**受講者存在領域**と呼ぶ。このような講義室内のレイアウトの例を図1に示す。

上のような形態の講義では，講義中に複数の人物が同時

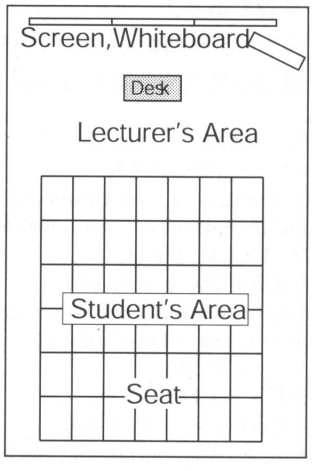


図1　対面型講義室の例 (俯瞰図)
Fig. 1.　Example of a lecture room.

に話し続けるという状況はほとんどないため，本研究では，任意の時刻における発話者の人数は最大1人であると仮定する。ただし，発話者以外にも，教材の音声や携帯電話の着信音などの音源は存在し得る。また，受講者存在領域には，受講者の鞄など，人間以外の物体が存在し得る。したがって，このような状況の下で発話者である対象を特定するための条件とは，その対象が“人間でかつ音源”であることといえる。

上のような条件を満足する対象を見つけるには視覚情報と聴覚情報の併用が不可欠であるが，そのためのカメラとマイクロホンは，機器自身が撮影映像に映り込んで見苦しくならないように，天井の近くに設置する必要がある。このようなカメラとマイクロホンの設置位置の制約により，得られる人物位置，音源位置には，どちらにも誤りが含まれる可能性を考慮に入れておく必要がある。

一方，人間のカメラマンが人物の映像を撮影する場合，通常は，人物の，胸から上，腰から上，全身，をそれぞれ画面内に捉えるバストショット，ミドルショット，ロングショットのいずれかの構図を採用する[9]。このことから，話者を自動撮影するための話者位置推定において，最も高い位置推定精度が要求されるのはバストショットを実現する場合であり，数十センチ程度の位置推定精度が要求される。ただし，受講者を撮影する際には，バストショットでは隣の受講者も一緒に写ってしまう場合があり，そのような場合に，発話している受講者だけが視野に入るようになるまで視野を狭めて撮影するには，受講者の座席サイズにほぼ等しい位置推定精度が要求される。

本研究の目的は，講義の自動撮影のための話者位置推定であり，視聴覚情報の統合，講義室に関する制約の利用とにより，以上のような条件を満足する処理を実現することを試みる。次章では，このための具体的な手法について述べる。

3. 視聴覚情報の統合による話者位置の推定

〈3・1〉 話者位置の表現　講義室内の話者位置を表現するために，講義室床面の隅を原点，床面上の壁との交線を x，y 軸，床面に垂直に z 軸をとった世界座標系を考える。2. 章での議論から，講師は講師存在領域内で直立しており，受講者は受講者存在領域内で着席しているので，講師及び受講者の床面から口元までの高さをそれぞれ z_l，z_s とすると，話者位置は，講師存在領域内の平面 $z=z_l$，もしくは受講者存在領域内の平面 $z=z_s$ 上に存在するものと考えられる。さらに，講義の自動撮影に必要な位置推定の精度は，講師については数十センチ，受講者については座席サイズとなることから，講師については，講師領域内の $z=z_l$ 平面を1辺が数十センチの正方形のセルに分割する一方，受講者については，受講者領域内の $z=z_s$ 平面上に各座席位置に対応する矩形のセルを設定し，それぞれを単位として話者位置を推定する。以下では，講師存在領域，受講者存在領域中で，世界座標系の x, y 軸に沿って (X_l, Y_l) 番目，(X_s, Y_s) 番目に位置するセルを，それぞれ $L(X_l, Y_l)$，$S(X_s, Y_s)$ で表し，**講師セル**，**受講者セル**と呼ぶ。ただし，X_l, Y_l, X_s, Y_s は0以上の整数である。

2. 章で述べたように，講義の自動撮影において，話者は人物でかつ音源である対象として特徴付けられることから，話者位置推定の際には，カメラからの視覚情報に基づいて各セルに人物が存在する可能性を評価する。

これまでにも，画像による動物体追跡の方法が提案されている。これらの研究において，動物体をブロックマッチングによって位置や形状を検出する方法[10]や，肌色情報に基づく手法[11]などが提案されている。一方，本研究で検出すべき情報の一つである講師の位置の場合，姿勢の情報は必要はなく，受講者の位置に関しては，オクルージョンの問題を回避するために，観測カメラを受講者の頭部が見える位置に設置する必要があるため，肌色領域に基づく手法などを適用することも困難である。そこで本研究では，画像処理手法としては単純な背景差分法やフレーム間差分法を用いて，そこで得られた情報と，講義室における講師や受講者の動きの特徴を利用して，講師および受講者が存在する位置を推定する手法を提案する。

一方，マイクロホンアレイからの聴覚情報に基づいて各セルに音源が存在する可能性を評価し，これらの結果を統合することによって，各セルに人物と音源が同時に存在する可能性を評価する。これらの処理の具体的手順について次に説明する。

〈3・2〉 視覚情報に基づく講師位置の推定　前節のように，講師の口元が講師領域中の $z=z_l$ 平面上に存在すると仮定すれば，講師の頭頂部はさらにその上の $z=z_l'$ の平面上に存在することになる（z_l' は講師の身長）。そこで，これを制約として利用することにより，位置及び向きが既知のカメラによる撮影画像中での講師の頭頂部の位置が検出できれば，その3次元位置を復元することができる。

カメラ画像から講師領域を抽出する処理は，2. 章で議論したように，講師は講師領域中に単独で存在すると仮定しているので，本稿では講義室の天井に講師領域を視野に捉えたカメラを設置し，ここから得られる画像上での頭部位置を，以下のような背景差分に基づいて抽出する。

時刻 t におけるカメラ画像および講師が存在しない状況で予め撮影しておいた背景画像に平滑化処理を施した後の画素位置 (i,j) における濃淡値を，それぞれ $I_L(i,j,t)$，$\hat{I}_L(i,j)$ とする。これらの画像を用いた背景差分 $B_L(i,j,t)$ を次式で定義する。ただし，T_l は差分値の二値化のための閾値である。

$$B_L(i,j,t)=\begin{cases}1, & \text{for} \quad |I_L(i,j,t)-\hat{I}_L(i,j)|\geq T_l\\ 0, & \text{otherwise}\end{cases} \quad \cdots\cdots(1)$$

得られた差分画像 B_L に対して孤立点除去，膨張処理を施した画像を B_L' とする。B_L' の各画素 (i,j,t) に対し，各 j における i 軸に沿った画素値ヒストグラム $H(j,t)$ を次式のように算出する。

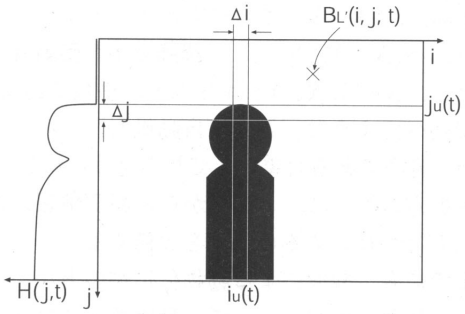


図 2　講師の頭頂部の推定

Fig. 2.　Estimating the crown of a lecturer's head.

$$H(j,t) = \sum_i B'_L(i,j,t) \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

このヒストグラムの j を Δj 画素分の幅を持つ区間に分け，式 (3) のように，各区間毎のヒストグラムの和が T_h を越える最小の j 座標を求め，講師の頭頂部の j 座標 $j_u(t)$ とする。さらに，B'_L の $j = j_u(t)$ の位置に沿って $\Delta i \times \Delta j$ 画素の大きさを持つ矩形領域を設定し，各矩形領域のうち，式 (4) のように，画素値の合計が最も多い矩形領域の i 座標を講師の頭頂部の i 座標 $i_u(t)$ とする。以上の処理の様子を図 2 に示す。

$$j_u(t) = \mathrm{argmin}_j \Big| \sum_{k=j}^{j+\Delta j} H(k,t) - T_h \Big| \cdots\cdots\cdots (3)$$

$$i_u(t) = \mathrm{argmax}_i \sum_{k=i}^{i+\Delta i} \sum_{l=j_u(t)}^{j_u(t)+\Delta j} B'_L(k,l,t) \cdots\cdots (4)$$

世界座標系における時刻 t での講師の頭頂部の 3 次元位置を $(x_l(t), y_l(t), z'_l(t))$ とすると，この位置は，画像座標中における頭頂部の位置 $(i_u(t), j_u(t))$ と，カメラのレンズ中心を結んだ直線と平面 $z = z'_l$ との交点として求めることができる (図 3)。

この位置と同じ x, y 座標を含む $z = z_l$ 平面上の講師セルを，$L(\tilde{X}_l, \tilde{Y}_l)$ とすると，時刻 t において x, y 軸に沿った講師セルの位置 $\mathcal{X}(t), \mathcal{Y}(t)$ は，$\mathcal{X}(t) = \tilde{X}_l, \mathcal{Y}(t) = \tilde{Y}_l$ となる。

〈3・3〉 視覚情報に基づく受講者位置の推定　受講者に関しては，講師とは異なり，受講者存在領域中に複数の受講者が密集して存在する。従って，講義室の天井に，講師や受講者が存在する領域を斜め上方向から取り囲むように設置するような通常のカメラ配置では，得られる画像中で受講者同士のオクルージョンが発生し，受講者一人一人の位置を実時間で推定することは困難な問題となる。このため，1 章で挙げたような映像情報に基づく講義撮影の研究[3][4] においても，受講者については，全体としての動きの特徴などを利用するというアプローチが多く，受講者一人一人の位置を個別に推定する処理までを実現しているものは少ない。しかし，本研究のように，受講者が話者となった場合も含めた話者位置推定を，視覚情報による人物位置推定結果と聴覚情報による音源位置推定結果の統合によって実現しようとする場合には，視覚情報によって受講者一人一人の位置を推定できる手法の存在が前提となる。そこで本研究では，このための手法として，新たに以下のような手法を導入した。

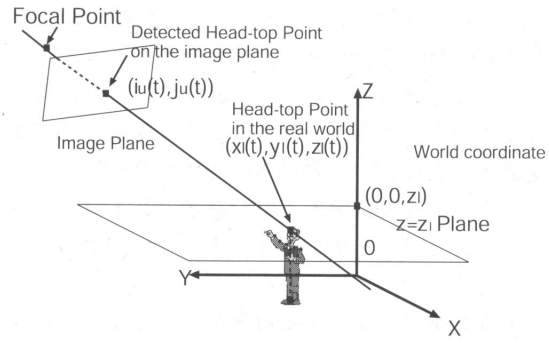


図 3　講師の位置の推定

Fig. 3.　Estimating the location of a lecturer's head in the world coordinate.

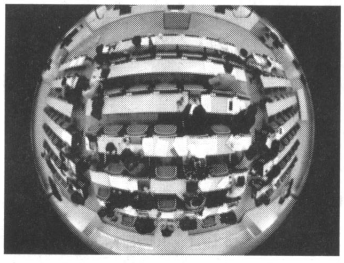


図 4　受講者存在領域の魚眼画像

Fig. 4.　Fish-eye image of student's area.

カメラ画像中での受講者同士のオクルージョンを避けるため，受講者存在領域の中心の天井に光軸を鉛直下方向に向けたカメラを設置し，これによって受講者存在領域を撮影する。

このとき，受講者存在領域全体を観測するため，観測カメラには魚眼レンズを装着する。このカメラによる撮影画像例を図 4 に示す。この図から分かるように，魚眼レンズ付きカメラで得られる画像は周辺部が歪んだものとなるが，受講者の動作認識を行うほどの解像度は必要ではなく，また受講者セルは画像中で重複せず，位置も変化しないので，各受講者セルに対応する画像中の領域を予め手動で設定しておく (図 5)。後述するように，各受講者の存在の判定や動きの大きさを評価する際には，設定した各矩形領域の大きさに依存する量を用いて正規化する。

上述のような撮影方法により，撮影画像中での受講者のオクルージョンが回避されるため，受講者一人一人の位置を推定する処理は，受講者セルに対応する各画像領域に受講者が存在するかどうかを判定する処理に帰着される。ここで，各画像領域をそれぞれ別の画像と考えれば，これは単独の人物の位置を推定する処理と同一である。このため，講師が存在するセルの推定と同様，各画像領域に対する背景差分の利用が有効と考えられるが，受講者セルは，講師

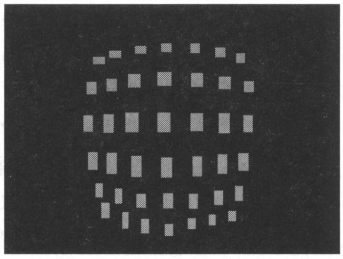


図 5　受講者セルに対応する画像領域
Fig. 5.　Rectangle area of student's cell.

セルとは異なり，(1) 受講者以外に鞄などの物体が画像中に存在したり，机と受講者の衣服の色が似ている場合がある，(2) 受講者は講義の始めから終りまで全く動かないということはないが，常に動き続けているわけではない，といった特徴をもつため，以下のように，背景差分及びフレーム間差分の時間的継続性の 2 つに基づいて，受講者セルに対応する各画像領域に受講者が存在する可能性を評価する。

なお，以下の説明では，受講者が存在しない状況で予め魚眼レンズ付きカメラで撮影しておいた背景画像と時刻 t での画像に，平滑化処理を施した後の画素位置 (i,j) における濃淡値をそれぞれ $I_S(i,j,t)$，$\hat{I}_S(i,j)$ と表記し，I_S，$\hat{I}_S$ 上で受講者セル $S(X_s,Y_s)$ に対応する画像上の領域を構成する画素集合を $\mathcal{I}_S(X_s,Y_s)$，画素数を $n_S(X_s,Y_s)$ と表記する。

まず，$S(X_s,Y_s)$ における背景差分の大きさ $R_S(X_s,Y_s,t)$ を次式で定義する。

$$R_S(X_s,Y_s,t)=\frac{\sum_{(i,j)\in\mathcal{I}_S(X_s,Y_s)} B_S(i,j,t)}{\max_t \sum_{(i,j)\in\mathcal{I}_S(X_s,Y_s)} B_S(i,j,t)} \quad \cdots\cdots (5)$$

$$B_S(i,j,t)=\begin{cases}1, & \text{for} \quad |I_S(i,j,t)-\hat{I}_S(i,j)|\geq T_s\\ 0, & \text{otherwise}\end{cases} \quad \cdots\cdots (6)$$

ただし，T_s は差分値の二値化のための閾値である。

背景差分では，対象物体が静止物体であれば背景差分の大きさは一定であることが仮定できるが，受講者のように身動きをする動物体では，背景差分の大きさも変化する。また，差分が検出される画素の数は，同じ動きでも受講者セルの位置によって，その絶対値は異なる。そこで，式 (5) では，該当受講者セルで過去に最も多くの画素で背景差分が検出された際の画素数で割ることによって，$R_S(X_s,Y_s,t)$ が 0 から 1 の値を取るように正規化している。

次に，時刻 t におけるフレーム間差分の時間的な継続性 $M_S(X_s,Y_s,t)$ を定義するために，予め時刻 t におけるフレーム間差分の大きさを表す $F_S(X_s,Y_s,t)$ を，次式で定義する。ただし，T_e は差分値の二値化のための閾値である。

$$F_S(X_s,Y_s,t)=\frac{\sum_{(i,j)\in\mathcal{I}_S(X_s,Y_s)} E_S(i,j,t)}{n_S(X_s,Y_s)} \quad \cdots\cdots (7)$$

$$E_S(i,j,t)=\begin{cases}1, & \text{for} \quad |I_S(i,j,t)-I_S(i,j,t-1)|\geq T_e\\ 0, & \text{otherwise}\end{cases} \quad \cdots\cdots (8)$$

魚眼レンズ付きカメラで撮影できる受講者は数十人であり，1 人が占める画素の数は数十画素と少ない。そこで本研究では，受講者の動きの大きさではなく，動きの有無のみを考慮する。これを式 (9) に示す。ただし，T_A は動きの有無を判断するための閾値である。また，受講者は常に動き続けるわけではないので，過去 T_t(1 を単位とする離散値) の期間の動きの有無の和を，現時刻 t の動きの評価値とみなす。ただし，評価値が上がる程度を制御するために，固定値 $+\alpha(0.0<\alpha\leq 1.0)$ を掛け，値を 0 と 1 の間に正規化するため，1 と比較して小さいほうの値を採用する (式 (10))。

$$D_S(X_s,Y_s,t)=\begin{cases}1, & \text{for} \quad F_S(X_s,Y_s,t)\geq T_A\\ 0, & \text{otherwise}\end{cases} \quad \cdots\cdots (9)$$

$$M_S(X_s,Y_s,t)=\min(1.0,\sum_{t-T_t}^{t}\alpha D_S(X_s,Y_s,t)) \quad \cdots\cdots (10)$$

各受講者セルごとの受講者存在可能性 $P_S(X_s,Y_s,t)$ は式 (11) から式 (14) によって求める。ここで，$W_R(X_S,Y_S,t)$ は，時刻 t における $R_S(X_S,Y_S,t)$ に対する重みであり，$W_R(X_S,Y_S,t-1)$ は，時刻 $t-1$ における $R_S(X_S,Y_S,t-1)$ に対する重みである。また β は重みの変化量を表し，0 から 1 の値を持つ固定値である。これは $R_S(X_s,Y_s,t)$ と $M_S(X_s,Y_s,t)$ の重み付き加算であるが，過去に動きが続いていた場合，そのセルに人物が存在する可能性が高く，鞄などの物体である可能性が低い。そこで，背景差分による結果よりもフレーム間差分に関する結果のほうが評価値が高いときは，背景差分の結果を重視するように重みを変更する (式 (11))。

$$W_R(X_s,Y_s,t)=\begin{cases}W_R(X_s,Y_s,t-1)+\beta, & \\ \quad \text{for } R_S(X_s,Y_s,t)\leq M_S(X_s,Y_s,t) & \cdots (11)\\ W_R(X_s,Y_s,t-1), \quad \text{otherwise} & \end{cases}$$

$$W_M(X_s,Y_s,t)=1-W_R(X_s,Y_s,t) \quad \cdots\cdots (12)$$

$$P_S(X_s,Y_s,t)=W_R(X_s,Y_s,t)R_S(X_s,Y_s,t) \quad \cdots (13)$$
$$+W_M(X_s,Y_s,t)M_S(X_s,Y_s,t)$$
$$\cdots\cdots (14)$$

〈3·4〉　聴覚情報に基づく音源位置の推定　音源位置あるいは音源方向の推定方法としては，MUSIC(MUltiple

SIgnal Classification) 法による音源方向推定法[12]，同期乗算による音源位置推定法[6]，ステレオビジョンにおけるエピポーラ幾何の概念を聴覚に拡張した聴覚エピポーラ幾何に基づく方法による音源方向推定法[13]，白色化相互相関法 (CSP 法: Cross-Power Spectrum phase analysis) による音源方向推定法[7][14]~[18] など，様々な手法が提案されている。

音源位置推定によく用いられる手法は，MUSIC 法であるが，この方法は，(1) 複数の観測音の音源に対処可能である，(2) フーリエ変換した後の信号を処理するため，波形に依存しない解析方法である，という長所を持つ一方，(1) 測定環境における音の伝達関数が既知でなければならない，(2) 音源から発せられる音と相関のある雑音に弱い，等の性質がある[19]。

一般に，一斉講義型の講義室は，講義によって出席する受講者の数も変わるため，毎回受講者の数が変化する講義室において音の伝達関数を求めておくことは困難である。また，講義室は閉じた空間であり反響音等のノイズも無視できない。一方，CSP 法はある音源から 2 つのマイクに到達する音の位相差を推定する手法である。この手法は，同じ空間に複数個の音源が存在する状況では音源方向を推定できないという短所があるが，予め音の伝達関数を求めておく必要がないため，各講義毎に受講者の数や位置が変わる状況でも適用可能であるという利点を持つ。

本研究では，一斉授業型の講義においては，同時に複数の人物が話すことがないという状況と，音の伝達関数が講義ごとに変り得る実際の講義室に適用とするという目的から，CSP 法を利用した音源位置推定法を採用し，音源位置の推定を行う。

CSP 法は 2 本のマイクロホン (**マイク対**) への音到来時間差を計算する手法である。時刻 t において，ある音源から 2 本のマイクロホン M_p, M_q に届いた音声波形を一定時間サンプリングし，得られた 2 系列の振幅値間の相関の大きさを表す CSP 係数 $C_{p,q}(h,t)$ を求めると，これが最大となる位相差 h から，次式のように 2 つのマイクロホンへの到達時間差 $\tau_{p,q}(t)$ を求めることができる。ただし，f_s はサンプリングレートである。

$$\tau_{p,q}(t) = \frac{\delta_{p,q}(t)}{f_s} \quad \cdots\cdots (15)$$

$$\delta_{p,q}(t) = \arg\max_h (C_{p,q}(h,t)) \quad \cdots\cdots (16)$$

CSP 法で求めた音の到達時間差が等しい位置は，3 次元空間上でマイク対を構成する 2 つのマイクロホンを焦点とする双曲面となり，音源はその双曲面上にあると推定できる。一方，〈3·1〉節で述べたように，話者位置は平面 $z = z_l$，$z = z_s$ 上の $L(X_l, Y_l), S(X_s, Y_s)$ を単位として推定するので，マイクロホンが 2 対あれば，それぞれから得られる双曲面の交線と平面 $z = z_l$，$z = z_s$ 上との交点を求めることによって，音源位置を推定できる。ただし，音源が存在し得る位置は講義室内の広い範囲にわたっていることから，音源位置によってマイク対と音源との距離が大きく異なり，これが位置推定精度に影響する可能性がある。そこで本研究では，位置推定精度が音源位置に依存することを避けるために，講義室の周辺部の天井に複数のマイクロホンを講義室を取り囲むように設置し，隣接する各マイク対から得られる双曲面と平面 $z = z_l$，$z = z_s$ との交線上の位置を，講師セル及び受講者セルを単位として，すべてのマイク対にわたって投票することによって，音源位置が存在する可能性を評価する。このときの講師セル $L(X_l, Y_l)$，受講者セル $S(X_s, Y_s)$ に対する投票数をそれぞれ $e_L(X_l, Y_l, t)$ 及び $e_S(X_s, Y_s, t)$ と表記する。

〈3·5〉 視聴覚情報の統合による話者位置推定 人物でかつ音源としての話者位置を推定するには，理論的には，〈3·2〉節，〈3·3〉節で得られる人物位置と，〈3·4〉節で得られる音源位置が一致する位置を求めれば良いことになる。ただし，〈3·2〉節，〈3·3〉節で求めた人物位置推定手法では，背景差分やフレーム間差分を用いて講師領域や受講者領域を抽出するために，その性質に関するいくつかの仮定を設けており，これが十分に成立しない状況等においては，位置推定の結果に誤りが含まれることになる。また，〈3·4〉節で求めた音源位置についても，音源位置や雑音の状況などによっては，位置推定の結果に誤りが含まれる。これらの誤りが存在する場合には，人物位置と音源位置が一致しない可能性がある一方，話者が存在しない状況では，逆に人物位置と音源位置は一致しない方がむしろ当然である。以上のことから，本研究では，〈3·2〉節~〈3·4〉節の手法に基づいて，各講師セル，受講者セルに対する講師，受講者，音源の存在可能性の分布を連続値で表現し，それらの値に基づいて，話者の有無及びその位置を式 (17) から式 (21) のように定義する。

(1) 講師存在可能性の表現

〈3·2〉節の手法で求まった講師セル位置 $(X(t), Y(t))$ には，頭部位置検出の誤りや，観測カメラからのキャリブレーション誤差などの理由により，正解位置を中心とした推定誤りが存在する。そこで本稿では，講師セル位置 $(X(t), Y(t))$ に含まれる誤りをガウス分布で表現し，各講師セルにおける講師存在可能性 $P_L(X_l, Y_l, t)$ を次式で表す。ただし，v は講師セルを単位とする，誤りの分布の大きさである。

$$P_L(X_l, Y_l, t) = \exp\left(-\frac{(\mathcal{X}(t) - X_l)^2}{v^2}\right) \times \exp\left(-\frac{(\mathcal{Y}(t) - Y_l)^2}{v^2}\right) \quad \cdots\cdots (17)$$

(2) 受講者存在可能性の表現

各受講者セルの受講者存在可能性に関しては，〈3·3〉節で求めた $P_S(X_s, Y_s, t)$ 自身が各セルにおける受講者の存在する可能性を表すので，これをそのまま用いる。

(3) 音源存在可能性の表現

講師セル及び受講者セルにおける音源存在可能性 $P_{L_a}(X_l, Y_l, t)$，$P_{S_a}(X_s, Y_s, t)$ を次式で表す。時刻 t におけるセルの投票数が閾値 T_E 以上のとき，その時刻に

おける最も大きな投票数で割って正規化した値を音源存在可能性であるとみなす。ただし，〈3·4〉節で述べた音源位置推定法では，双曲線上に存在する音源位置でないセルにも必然的に投票を行うが，音源でないときは投票数が小さいことから，投票数が閾値 T_E を越えていないセルでは，音源存在可能性はないものとする。

$$P_{L_a}(X_l,Y_l,t) = \begin{cases} \dfrac{e_L(X_l,Y_l,t)}{\max\limits_{X_l,Y_l,X_s,Y_s}(e_L(X_l,Y_l,t),e_S(X_s,Y_s,t))} \\ \quad \text{for} \quad e_L(X_l,Y_l,t) \geq T_E \\ 0, \quad \text{otherwise} \end{cases} \quad \cdots\cdots (18)$$

$$P_{S_a}(X_s,Y_s,t) = \begin{cases} \dfrac{e_S(X_s,Y_s,t)}{\max\limits_{X_l,Y_l,X_s,Y_s}(e_L(X_l,Y_l,t),e_S(X_s,Y_s,t))} \\ \quad \text{for} \quad e_S(X_s,Y_s,t) \geq T_E \\ 0, \quad \text{otherwise} \end{cases} \quad \cdots\cdots (19)$$

(4) 話者存在可能性の表現

(1)～(3) の結果の積を用いて，講師セル及び受講者セルにおける話者存在可能性 $P_{L_w}(X_l,Y_l,t)$，$P_{S_w}(X_s,Y_s,t)$ を次式によって表現する。

$$P_{L_w}(X_l,Y_l,t) = P_L(X_l,Y_l,t)P_{L_a}(X_l,Y_l,t) \quad \cdots (20)$$

$$P_{S_w}(X_s,Y_s,t) = P_S(X_s,Y_s,t)P_{S_a}(X_s,Y_s,t) \cdots (21)$$

(5) 話者位置の推定

最も大きい話者存在可能性が，ある閾値 T_w を越えているとき，そのセルが話者位置であると推定する。閾値を越えていないとき，話者はいないと判断する。

4. 実　　験

〈4·1〉 実験環境　前章で提案した手法の有効性を検証するために，実際の講義室を利用した話者位置推定の実験を行った。

実験で用いた講義室の大きさは，幅が 15.5m，奥行が 9.6m，天井の高さが 3.2m，座席数は 120 席である (図 6)。この講義室の講師存在領域全体を一台の観測カメラでカバーすることは困難であるため，講師領域全体を 3 つの領域に分け，それぞれの領域を担当する 3 台のカメラを図中 0,1,2 に設置した。また，受講者位置推定のための魚眼レンズ付カメラ 1 台 (図中 3)，音源位置推定のためのマイクロホン 8 本 (図中灰色丸印) を設置した。講師セルの大きさは 1 辺 50cm とし，講師存在領域を 6 行 13 列，合計 78 個の講師セルに分割した。受講者セルの大きさは座席サイズに合わせて縦 90cm，横 60cm とし，受講者存在領域に 6 行 x7 列，計 42 個の受講者セルを設定した。講師セル及び受講者セルの高さは，講師役，受講者役の被験者の口元の高さに合わせて，それぞれ 157cm，115cm とし，講師位置推定に用いる講師の身長は，講師役被験者に合わせて 177cm とした。

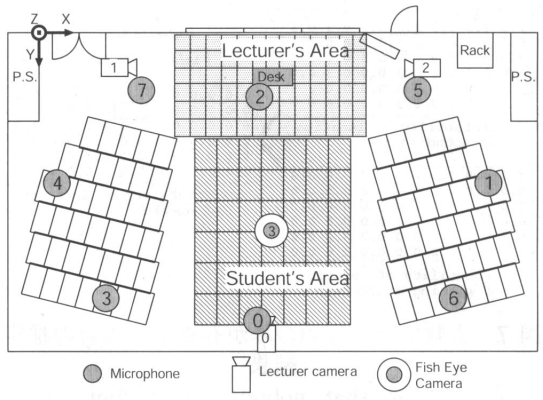


図 6　各種センサの位置
Fig. 6.　Location of sensors.

表 1　実験システムで用いた閾値
Table 1.　The values of threshold in our experiments.

式番号	閾値記号	値
(1)	T_l	50
(3)	T_h	15
(3),(4)	$\delta i, \delta j$	各 5
(6)	T_s	100
(8)	T_e	10
(9)	T_A	0.01
(10)	T_t	10
(10)	α	0.1
(11)	β	0.05
(15)	f_s	8KHz
(17)	v	3
(18),(19)	T_E	3
〈3.5〉節の (5)	T_w	0.56

また，3. 章において定義した，各式に適用した閾値を表 1 に示す。

式 (1) における T_l と式 (6) における T_s は，それぞれ背景差分に用いる閾値であり 0 から 255 の値をとりうる。経験的に受講者の方が差分値が大きくなりやすいため，受講者用閾値の方を大きくして，誤検出を少なくしている。

また，式 (8) における T_e は，ある画素における動きの有無を判断するための閾値であり 0 から 255 の値をとりうるが，これに関しても，静止物体は検出されず，動きのある受講者に関しては検出される値を用いた。

式 (3),(4) における $\delta i, \delta j$ は，画像上の講師の頭頂部を求めるための，画像上での矩形領域の大きさを持たせるが，これも，講師の頭髪部分を含む程度の画素数を用いた。

また，式 (3) における T_h は，上記 $\delta i \times \delta j$ のうち，どの程度背景差分によって検出された画素が含まれれば頭部の高さ位置であるとみなすかを表す量であるが，これは，経験的に 6 割の値を用いた。

式 (9) の T_A は，受講者セルを構成する画素のうち，どの程度の数の画素に動きが検出された場合にその受講者セル内に動きのある物体が存在するかを表した量であり，これに関しては，経験的に 1(%) となるように設定している。

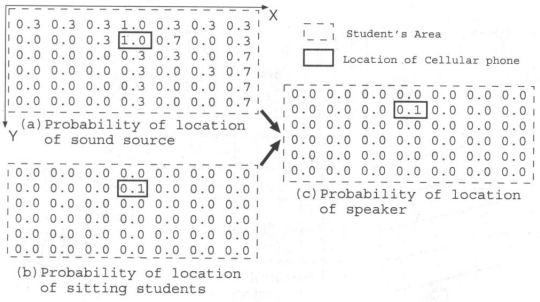


図 7　人物でない音源のみが存在する場合の推定結果

Fig. 7.　A case that nobody speaks but sound source is detected.

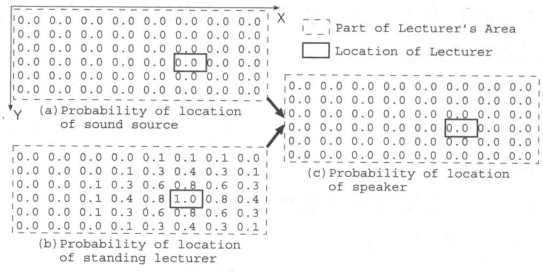


図 8　話者でない人物のみが存在する場合の推定結果

Fig. 8.　A case that nobody speaks but there is ambient noise.

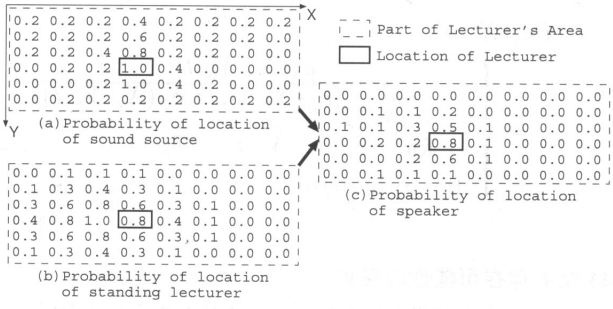


図 9　話者が存在するときの推定結果

Fig. 9.　A case that somebody speaks.

式 (10) の T_t は，各受講者セルにおける動きの評価に関して過去何回分の評価値を用いるかという量である。本研究において実装したシステムでは 1 秒間に約 2 回の評価が可能であり，過去 5 秒間の期間を評価可能な値を用いた。また，式 (10) の α は，約 5 秒間動きが検出され続ければ，$M_S(X_s, Y_s, t)$ の値が 1.0 となるような値を用いた。

式 (11) における β は，背景差分による結果よりもフレーム間差分に関する結果のほうが評価値が高い時間が総計で約 5 秒間続いたときに最も背景差分の結果を重視するような値を用いた。

式 (15) における f_s は，音声を録音するときの周波数であるが，音響の分野で一般的に用いられる 8KHz を用いた。

式 (17) における v は，講師位置推定の結果得られた講師セルからどの程度の広さの講師セルまでを許容するかを表す量であるが，本稿では，講師領域の奥行き方向の講師セルの数 6 の半分である 3 を用いた。

式 (18),(19) の T_E に関しては，少なくとも 3 つのマイク対による投票があった場合に音源位置の候補があるものとした。

最終的に話者の存在を判断する閾値 T_w は，人物の存在確率が 0.75，音源の存在確率が 0.75 の場合にその積が 0.56 となることから，0.56 を用いた。

〈4・2〉 人物かつ音源としての話者位置推定　視聴覚情報の統合によって，人物かつ音源としての話者位置を推定できることを確認するために，講義室に，(1) 人物でない音源のみが存在する状況，(2) 発話していない人物のみが存在する状況，(3) 話者が存在する状況，の 3 つの状況において，3. 章で述べた人物位置推定，音源位置推定，および両者の結果の統合による話者位置推定，それぞれの結果を求めた。以下にその結果の具体例を示す。以下，講師存在領域に関する図においては，紙面の都合によりその一部を切り出して示す。

〈4・2・1〉 人物でない音源のみが存在する状況　受講者セルに携帯電話を置き，着信音を鳴らした状態で，各受講者セルに対する受講者存在可能性 $P_S(X_s, Y_s, t)$，音源存在可能性 $P_{S_a}(X_s, Y_s, t)$，話者存在可能性 $P_{S_w}(X_s, Y_s, t)$ を求めた。図 7 は，黒枠で示した $S(3, 1)$ のセルに携帯電話を置いた場合の結果である。このような状況では，音源位置推定は，同図 (a) のように，携帯電話の存在するセルを中心に高い可能性を示す。一方，受講者存在可能性は，携帯電話が存在する位置で背景差分が零にはならないため，全体としては零とはならないが，動きがまったくないため，同図 (b) のように値が小さい。以上の結果の統合により，話者位置推定の評価値は同図 (c) のように携帯電話を置いたセルも含めて小さくなり，話者は存在しないと判定される。

〈4・2・2〉 発話していない人物のみが存在する状況　講師セルの 1 つに被験者が発話せずに直立している状態で，上と同様に各講師セルに対する講師，音源，話者，それぞれの存在可能性を求めた。図 8 は，黒枠で示したセル $L(5, 3)$ の位置に被験者が存在する場合の例である。このような状況では，講師存在可能性は，同図 (b) のように被験者が存在しているセルの周辺で高い評価値を示す。一方，音源推定は，このような状態でも空調のファン音が存在するため，CSP 法による音源位置の投票結果は零とはならないが，音源位置が集中しておらず，投票値が小さいため，〈3・5〉 節の (2) で述べたような処理の効果により，同図 (a) のように可能性自体は零となっている。以上の結果の統合により，話者存在可能性は同図 (c) のように被験者の座席位置のセルも含めて低くなり，話者は存在しないと判定される。

〈4・2・3〉 話者が存在する状況　講師セルの 1 つに被験者が発話しながら直立しているときの各講師セルに対する存在可能性を上と同様に求めた。図 9 は，被験者である話者が黒枠で示したセル $L(3, 3)$ に存在している場合の例

表 2　人物位置推定の誤差 (単位：セル)
Table 2. Estimation error of the location of a person.

Lecturer's Area											
1.0	1.0	0.2	0.5	1.0	0.0	0.0	0.0	0.0	0.8	1.0	1.0	1.0
2.0	1.6	0.2	0.0	0.0	0.9	0.0	0.0	1.0	0.0	0.9	1.0	0.1
3.0	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0	1.0	0.2
1.0	0.9	0.1	0.1	0.0	1.0	1.0	2.0	0.2	1.0	0.5	0.0	1.0
3.0	2.0	1.0	0.0	1.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	2.0
7.0	6.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0

Student's Area						
0.0	0.0	0.2	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.3
0.6	0.5	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.3	0.0	0.7	0.0
0.0	1.9	0.0	0.0	0.0	0.4	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0

表 3　音源位置推定の誤差 [空調停止時](単位：セル)
Table 3. Estimation error of the location of a sound source with air-conditioning.

Lecturer's Area												
0.2	0.3	0.0	0.5	0.1	0.1	0.1	0.0	0.3	0.0	0.1	0.0	0.0
0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
0.2	0.4	0.1	0.0	0.1	0.1	0.0	0.2	0.1	0.0	0.1	0.2	0.1
0.2	0.3	0.2	0.0	0.4	0.2	0.0	0.1	0.2	0.2	0.0	0.0	0.1
0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.2
0.1	0.0	0.1	0.3	0.1	0.0	0.0	0.6	0.0	0.0	0.0	0.1	0.0

Student's Area						
0.6	0.0	0.3	0.2	0.0	0.1	0.4
0.0	0.1	0.0	0.0	0.2	0.0	0.3
0.0	0.1	0.2	0.1	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.1	0.0	0.0	0.0
0.1	0.0	0.0	0.0	0.0	0.0	0.2

表 4　音源位置推定の誤差 [空調稼働時](単位：セル)
Table 4. Estimation error of the location of a sound source without air-conditioning.

Lecturer's Area												
0.6	0.2	0.6	1.4	5.0	0.6	1.0	0.8	1.1	0.7	0.6	0.3	0.0
1.2	0.5	0.8	0.6	0.7	0.7	1.0	0.7	0.7	0.9	0.7	0.4	0.2
0.6	0.5	0.7	0.6	0.9	0.6	1.0	0.6	0.5	0.5	0.9	0.4	0.3
0.5	0.4	0.6	0.2	0.5	0.7	0.5	0.5	0.5	0.6	0.2	0.4	0.4
0.9	0.1	0.8	0.4	0.4	0.8	0.3	1.4	0.5	0.5	0.0	0.2	0.3
0.3	0.1	1.6	0.5	0.1	0.0	0.1	0.8	0.4	0.1	0.1	0.3	0.0

Student's Area						
0.9	0.0	0.9	0.3	1.3	1.0	2.5
0.0	0.2	1.0	0.2	0.4	0.8	1.0
0.0	0.7	0.3	0.6	0.1	0.3	0.1
2.0	0.1	0.0	0.1	0.0	0.0	0.0
0.0	0.1	1.2	0.2	0.0	0.0	0.1
1.7	0.0	0.1	0.0	0.2	0.0	0.7

表 5　話者位置推定の誤差 (単位：セル)
Table 5. Estimation error of the location of a speaker.

Lecturer's Area												
0.5	0.3	0.0	0.5	4.6	0.2	0.4	0.3	0.8	0.2	0.1	0.0	0.0
0.1	0.9	0.5	0.0	0.0	0.4	0.6	0.0	1.0	0.3	0.2	0.1	0.0
0.1	0.0	0.2	0.0	0.0	0.1	0.6	0.0	0.0	0.0	0.6	0.2	0.0
0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.3
0.1	0.1	0.7	0.0	0.1	0.0	0.1	0.3	0.0	0.1	0.0	0.0	0.5
0.3	0.6	1.0	0.3	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.2	0.0

Student's Area						
0.0	0.0	0.4	0.0	0.7	0.0	0.0
0.0	0.2	0.0	0.0	0.0	0.0	0.0
0.0	0.2	0.0	0.4	0.0	0.0	0.0
0.0	0.1	0.0	0.1	0.0	0.0	0.0
0.0	0.2	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0

である。このような状況では，同図 (a),(b) のように，音源と講師の存在可能性が共に，話者の存在するセルの周辺で高くなる。ただし，推定の誤りのために，(a) のように正しい位置からずれた場所に最大値が存在したり，(b) のように最大値が複数存在したりすることも多い。しかし，そのような場合でも，両者の結果が統合されることにより，(c) のように，正しい話者位置で評価値が最大となり，話者位置推定に成功する。

〈4·3〉 視聴覚情報の統合による誤り低減の評価

〈4·2·3〉 節で示したような，視聴覚情報の統合による話者位置推定の精度向上がどの程度有効に働いているのかを調べるために，講師セル，受講者セルのそれぞれの位置で被験者が発話を行っている状態を，各セルで 30 回ずつ観測し，正解位置と推定位置のずれをセルを単位とした市街地距離で評価した。各セルに正解位置が存在するときの人物位置推定，音源位置推定，両者の統合による話者位置推定，のそれぞれによる推定位置の誤差を，30 回の観測データに対して平均した値を表 2 ～5 に示す。

なお，この値を算出するにあたって，講師位置，受講者位置，音源位置はそれぞれ，〈3·5〉 節の (1)，(2)，(3) の方法で求めた存在可能性が最大となる位置の重心を用いた。

上の結果において，表 2 に示されている通り，講師位置推定による誤差が 1.0 以上の講師セルは 31 個，受講者位置推定による誤差が 1.0 以上の受講者セルは 2 個であった。受講者セルについては，被験者がくずれた姿勢をとることがなかったため，誤差が少なく，正しく位置が推定できる場合が多かったが，講師セルに関しては，カメラからの距離が遠い位置に存在する，講師存在領域の周辺部ほど，推定誤差が大きくなった。

音源位置推定について，空調設備を稼働しない状況と，稼働した状況で実験を行った。空調設備を稼働しない状況での音源位置推定結果を表 3 に示す。また，空調設備を稼働した状況でも音源位置推定結果を表 4 に示す。これらの表から，音源位置推定においては，空調設備による雑音の影響を受けないほうが全体的に推定精度が高いことがわかる。また，空調設備を稼働した状況では特に受講者存在領域の右上および左下に位置する受講者セルの誤差が大きいが，これは，その付近の天井にエアコンの吹き出し口があ

り，これによる雑音の影響を受けている可能性がある。ここでは実際に講義が行われる状況を想定しているので，表4について考える。この表で誤差が1.0以上であったセルは，講師存在領域では9個，受講者存在領域では8個であった。

以上の結果に対し，これらの結果の統合による話者位置推定結果では，誤差が1.0以上であったセルは，表5に示すように，講師存在領域では3個，受講者存在領域では0個であった。講師セル (8, 1) に関しては，誤差が低減されていないが，この実験を行った際，この講師セルについては，常に隣りの講師セルに間違えて推定され，かつその隣の位置を含む複数のセルで音源位置推定の評価値が最も高く推定され続けた結果生じた現象である。この結果から，講義室のような広い空間を対象とした場合に問題となりやすい，カメラからの距離の長さや空調設備のノイズの存在などによる人物位置推定や音源位置推定の誤りが全体として低減され，講義の自動撮影に必要な精度による話者位置推定が実現されているといえる。

5. おわりに

本稿では，一斉授業型の講義が行なわれる大規模な講義室において，講義を自動撮影するための話者の位置を推定する手法を提案した。講義の自動撮影における話者位置推定では，話者ではない人物や，人物ではない音源が存在し得るため，人物かつ音源として話者位置を特定するためには視聴覚情報の併用が必要となる。また，観測カメラやマイクロホンアレイの設置条件のため，これらを用いた人物位置推定や音源位置推定には誤りが含まれる場合が多い。そこで本研究では，人間のカメラマンが人物を撮影する場合の構図に基づいて講義の自動撮影に必要な位置推定精度を定めた上で，観測カメラからの視覚情報に基づく講師位置・受講者位置の推定手法，および，マイクロホンアレイからの聴覚情報に基づく音源位置の推定手法について検討し，これらの結果の統合によって，上の精度を満たす話者位置推定を実現した。筆者らの所属する大学では，昨秋より，週5コマ程度の講義をアーカイブ化し，学内LANに接続された計算機を利用して学生が自由に閲覧できるサービスを試験的に始めているが，このための講義の自動撮影の際に，本手法による話者位置推定結果を利用している。自動撮影の結果は，話者位置推定結果だけではなく，得られた話者位置の情報から，どのカメラで話者を撮影するのか，またどのような大きさで話者を撮影するかといった撮影ルールの内容や撮影カメラの実時間制御などにも大きく影響されるため，個々の処理を個別に評価することは難しいが，概ね問題ない撮影ができている。

今後は，上のような実運用の際のデータを収集し，様々な講義状況に対する本手法のロバスト性をより大規模に検証すると共に，逆に十分なロバスト性や位置推定精度を実現するためのカメラやマイクロホンの設置台数や配置について分析する必要があると考えている。

（平成15年6月26日受付，平成15年11月10日再受付）

文　献

（1） M. Onishi, M. Izumi, and K. Fukunaga: “Automatic Production of Video Images for Distance Learning System Based on Distributed Information”, *IEICE Japan*, Vol.J82-D-II, No.10, pp.1590–1597 (1999-11) (in Japanese)
大西正輝・泉　正夫・福永邦雄：「情報発生量の分布に基づく遠隔講義撮影の自動化」，信学論誌 (D-II), **J82-D-II**, 10, pp.1590–1597 (1999-10)

（2） T. Sakiyama, N. Ohno, M. Mukunoki, and K. Ikeda: “Video Stream Selection According to Lecture Context in Remote Lecture”, *IEICE Japan*, Vol.J84-D-II, No.2, pp.248–257 (2001-2) (in Japanese)
先山卓朗・大野直樹・椋木雅之・池田克夫：「遠隔講義における講義状況に応じた送信映像選択」，信学論誌，**J84-D-II**, 2, pp.248–257 (2001-2)

（3） Y. Kameda, K. Ishizuka, and M. Minoh: “A Live Video Imaging Method for Capturing Presentation Information In Distance Learning”, *IEEE Int. Conf Multimedia and Expo* (ICME2000), Vol.3, pp.1237–1240 (2000)

（4） M. Onishi, M. Murakami, and K. Fukunaga: “Computer-Controlled Camera Works at Lecture Scene Based on Situation Understanding and Evaluation of Video Images”, *IEICE Japan*, Vol.J85-D-II, No.4, pp.594–603 (2002-4) (in Japanese)
大西正輝・村上昌史・福永邦雄：「状況理解と映像評価に基づく講義の知的自動撮影」，信学論誌 (D-II), **J85-D-II**, 4, pp.594–603 (2002-4)

（5） D. Rabinkin, R. Renomeron, J. French, and J. Flanagan: “Estimation of Wavefront Arrival Delay Using the Cross-Power Spectrum Phase Technique”, *J. Acous. Soc. Am.*, Vol.100, No.4 Pt.2, p.2697 (1996-10)

（6） K. Kobayashi, H. Hokari, and S. Shimada: “Estimation of Plural Talker Locations Using Randomly Positioned MIcrophones(Method of Sub-Array Selection)”, *IEICE Japan*, Vol.J82-A, No.2, pp.193–200 (1999-2) (in Japanese)
小林和則・島田正治・穂刈治英：「複数マイク自由配置による複数話者位置推定 (マイク選択方法の提案)」，信学論誌，**J82-A**, 2, pp.193–200 (1999-2)

（7） M. Omologo and P. Svaizer: “Acoustic event localization using a crosspower-spectrum phase based technique”, *Proc. ICASSP94*, pp.273–276 (1994)

（8） M. Onishi, T. Kagebayashi, and K. Fukunaga: “Production of Videoconferencing Images by Computer Controlled Camera Based on Integration of Audiovisual Information”, *IEICE Japan*, Vol.J85-D-II, No.3, pp.537–542 (2002-3) (in Japanese)
大西正輝・影林岳彦・福永邦雄：「視聴覚情報の統合による会議映像の自動撮影」，信学論誌 (D-II), **J85-D-II**, 3, pp.537–542 (2002-3)

（9） D. Arijon: “Grammar of the film language”, Focal Press Limited (1976) (in Japanese)
ダニエル・アリホン：映画の文法，紀伊國屋書店 (1980)

（10） T. Kaneko and O. hori: “Robust Object Tracking Method Using Small Region Block Matching”, *IEICE Japan*, Vol.J85-D-II, No.7, pp.1188–1200 (2002-7) (in Japanese)
金子敏充・堀　修：「小領域のブロックマッチングを複数用いたロバストなオブジェクト追跡法」，信学論誌 (D-II), **J85-D-II**, 7, pp.1188–1200 (2002-7)

（11） A. Matsumura, Y. Iwai, and M. Yachida: “Tracking People by Using Skin-Color Information”, *IPSJ SIG Notes*, Vol.2002, No.34, 2002-CVIM-233-6, pp.133–138 (2002-5) (in Japanese)
松村朱里・岩井儀雄・谷内田正彦：「肌色情報を用いた複数人物追跡」，情報処理学会 研究報告，**2002**, 34, 2002-CVIM-233-6, pp.133–138 (2002-5)

（12） S.U. Pillai: Array Signal Processing, Springer-Verlag, New York (1989)

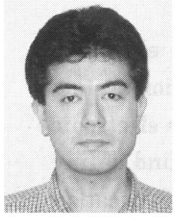
（13） K. Nakadai, T. Lourens, H.G. Okuno, and H. Kitano, “Active audition for humanoid”, Proc. 17th National Conference on Artificial Intelligence(AAAI-2000), pp.832–839, AAAI (2000)

（14） C.H. Knapp and G.C. Carter: “The generalized correlation method for estimation of time delay”, *IEEE Trans. Acoust., Speech and Signal Process.*, Vol.ASSP-24, No.4, pp.320–327 (1976-4)

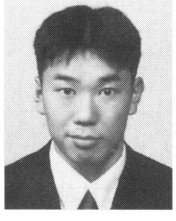
（15） M. Brandstein, J. Adcock, and H. Silverman: “A closed-form method for finding source locations from microphone-array time-delay estimates”, *Proc. ICASSP95*, pp.3019–3022

(1995)
(16) M. Omologo and P. Svaizer: "Acoustic source locationin niosy and reverberant environment using CSP analysis", *Proc. ICASSP96*, pp.921–924 (1996)
(17) P. Svaizer, M. Matassoni, and M. Omologo: "Acoustic source location three-demensional space using crosspower spectrum phase", *ICASSP97*, pp.231–234 (1997)
(18) M. Omologo and P. Svaizer: "Use of the Crosspower-Spectrum Phase in Acoustic Event Location", *IEEE Trans. on Speech and Audio Processing*, pp.288–292 (1997)
(19) M. Kunita, D. Arita, and R. Taniguchi: "Pan-Tilt Camera Control using Sound Source Localization", *IPSJ SIG Notes*, Vol.2002, No.34, 2002-CVIM-233-6, pp.41–48 (2002-5) (in Japanese)
國田政志・有田大作・谷口倫一郎:「音源壱同定によるカメラの首振りに関する研究」, 情報処理学会 研究報告, **2002**, 34, 2002-CVIM-233-6, pp.41–48 (2002-5)

西口敏司 (正員) 1971 年 9 月 13 日生。2001 年京大大学院情報学研究科博士課程了。同年同大学法学部助手。現在, 複数のセンサ情報を利用した講義室空間の状況把握, 講義の自動撮影およびアーカイブ化に関する研究に従事。

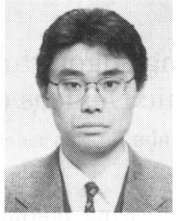


東和秀 (非会員) 1977 年 4 月 9 日生。2002 年京大大学院情報学研究科知能情報学専攻修士課程了。同年, 株式会社オージス総研入社。在学中, 映像音声処理の研究に従事。



亀田能成 (非会員) 1968 年 10 月 2 日生。1996 年京大大学院工学研究科博士課程了。同年同大学大学院工学研究科助手。1998 年同大総合情報メディアセンター助手。2001～2002 年マサチューセッツ工科大学客員研究員。2002 年同大学術情報メディアセンター助手。2003 年筑波大学機能工学系講師。工博。モデルベースドビジョン, 講義アーカイブ・遠隔講義, 三次元状況理解の研究に従事。

角所考 (非会員) 1964 年 6 月 3 日生。1993 年阪大大学院工学研究科博士課程了。1993～1994 年スタンフォード大客員研究員。1994 年阪大助手。1997 年京大総合情報メディアセンター助教授。2002 年同学術情報メディアセンター助教授。工博。視覚メディア処理, コミュニケーションに関する研究に従事。



美濃導彦 (非会員) 1956 年 1 月 2 日生。1983 年京大大学院工学研究科博士課程了。同年同大学工学部助手。1987～1988 年マサチューセッツ州立大客員研究員。1989 年京大工学部助教授。1995 年同教授。1997 年同大総合情報メディアセンター教授。2002 年同学術情報メディアセンター教授。工博。画像処理, 人工知能に関する研究に従事。