社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

複数のセンサ情報に基づく話者状況の理解

西口　敏司†　　東　　和秀††　　亀田　能成†††　　美濃　導彦†††

† 京都大学大学院法学研究科
〒 606-8501 京都市左京区吉田本町
†† 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
††† 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

E-mail: †{nishigu,higashi,kameda,minoh}@mm.media.kyoto-u.ac.jp

あらまし　講義を電子的にアーカイブ化するとき、講義中の話者の状況を理解することは講義映像の撮影やインデクシングに有用である。話者の状況は、人物の位置、音源の位置及び身振りの有無などで構成されるので、その推定には様々なセンサを組合せて用いることが不可欠である。そこで本研究では、マイクロホンアレイ、観測カメラ、超音波定位センサを用いて話者の状況を推定する方法について述べる。本手法に基づく推定結果を人手で判断した話者状況と比較した結果、83.0%の割合で正しく話者状況を推定していることが確認できた。

キーワード　話者状況、自動撮影、インデクシング、複数センサ、センサフュージョン

A Method for Understanding Situation of Speakers Based on Multimodal Sensors

Satoshi NISHIGUCHI†, Kazuhide HIGASHI††, Yoshinari KAMEDA†††, and Michihiko MINOH†††

† Graduate School of Law, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
†† Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
††† Center for Information and Multimedia Studies, Kyoto University, Yoshida Nihonmatsucho, Sakyo-ku, Kyoto, 606–8501 Japan

E-mail: †{nishigu,higashi,kameda,minoh}@mm.media.kyoto-u.ac.jp

Abstract　It is useful to understand a situation of a speaker in a lecture for automatic camera control and video indexing in lecture archiving systems. However, it is difficult to understand the situation only by visual sensor because it consists of multimodal features. Therefore, we use multimodal sensors, such as microphone array, observation cameras and ultrasonic position sensor in order to understand multimodal information of a speaker. We propose a method for estimating a situation of a speaker by integrating data from multimodal sensors. The rate of correct estimation of a situation of a speaker was 83.0 % by our method in comparison with manual estimation.

Key words　Situation of Speaker, Automatic Shooting, Video Indexing, Multimodal Sensors, Sensor Fusion

1. はじめに

講義の自動撮影や講義のビデオアーカイブの検索では、講義室内の状況に応じてカメラ制御やインデクスの付加が行われる。すなわち、講義室内の状況の把握が重要である。講義において理解すべき状況の種類はいくつか存在するが、その中でも特に話者状況を理解することを考える。ここで話者状況とは、誰がどこでどのように話しているかを表現したものである。講義中の話者状況に応じて撮影カメラを制御すれば、説明している講師や質問している受講者の表情や身振りを撮影することができ、講師や受講者の様子をより詳しく知ることができる。

有益な話者状況を自動的に推定するために、どのような種類のセンサをどのような方法で組み合わせて利用するかが問題となる。従来の講義の撮影に関する研究では、講師や受講者の位置や動きの大きさに注目し、それを用いて被写体選択を行なっている [1] [2] [3] 。このような位置や動きの大きさに関する特徴のみでは、人物が話中かそうでないか判断できないため、上述のような話者状況を表現することができない。

そこで本研究では、マイクロホンアレイ、観測カメラ、超音波定位センサの 3 種類のセンサを利用する。観測情報の種類や観測空間の広さなど、それぞれのセンサの特徴を生かした情報を検出し、その結果を互いに補い合うように統合することによって話者状況を理解する手法を提案する。

2. 話者状況の理解

2.1 話者状況の定義

本研究では、図 1 に示すような対面講義を行う講義室を想定する。講義における話者は講師または受講者である。講義中に講師は図 1 中の**講師領域**を歩いて移動し、講義室前方に設置されたオンラインスライド表示用スクリーンや電子白板を指し示すことがある。また受講者は**受講者領域**内の座席に着席して移動しない。このような環境の下で、ある時刻における話者状況を表 1 のように定義する。対面講義では講義中に複数の人物が同時に話す状況はほぼないと考えられるので、ある時刻 t における話者状況 $C(t)$ は L_1, L_2, L_3, S_i, ϕ のいずれかであるとみなすことができる。

2.2 複数のセンサを利用した話者状況の推定

話者状況の推定に必要なセンサと、それぞれのセンサが推定する情報を表 2 に示す。各センサはそれぞれの特徴に応じた情報を取得する。取得した情報を統合することによって、ある時刻における話者状況が推定可能となる。以下、表 2 に示す各センサごとの情報取得方法と、取得した情報を統合して話者状況を推定する方法について述べる。

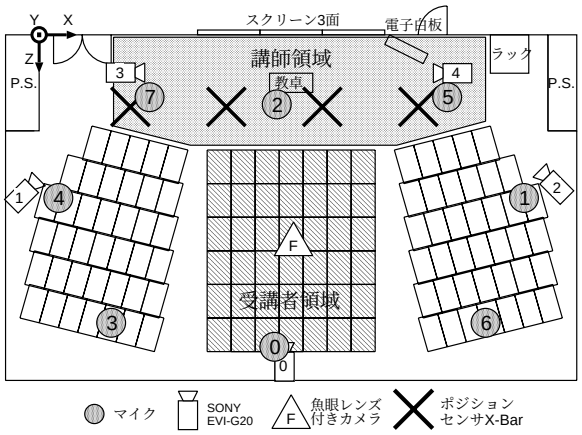

図 1　講義室の俯瞰図

2.2.1 マイクロホンアレイによる講師音声および受講者音声の位置推定

本研究では、マイクロホンアレイを用い、CSP 法 [4] で話者の位置推定を行う。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ことで音源から各マイクの間の距離の差を求める手法である。距離の差からマイク対を焦点とする双曲面を求めることができ、音源はその双曲面上に存在する。本研究では、講師領域では直立した人物の口元の高さを持つ平面と双曲面との交線上に講師音声の音源が存在し、受講者領域では着席した人物の口元

表 1　時刻 t における話者状況 $C(t)$

$C(t)$	状況の説明
L_1	講師が位置 p で、立ち止まって説明のみしている
L_2	講師が位置 p で、立ち止まってスクリーンまたは電子白板を指し示しながら説明している
L_3	講師が位置 p で、移動しながら説明している
S_i	受講者が座席 i で質問している
ϕ	誰も発話していない

表 2　複数のセンサと話者状況

センサ	話者状況				
	L_1	L_2	L_3	S_i	ϕ
マイクロホンアレイ	講師音声の位置			受講者音声の位置	音声無し
観測カメラ	講師の頭部・足元の位置			受講者の位置	—
超音波定位センサ	講師の胸部の位置			—	—
	指示無し	指示有り	指示無し	—	—

の高さを持つ平面と双曲面との交線上に受講者音声の音源が存在するものとする。音源位置を推定するために少なくとも 2 組以上のマイク対を利用し、複数の双曲線を求めてそれらの交点を音源とする。

2.2.2 観測カメラによる講師の頭部・足元位置推定

観測カメラを用いた講師の頭部と足元の位置推定は、位置、方向及び内部パラメータがキャリブレーション済みの観測カメラを用いて、講師を撮影した画像からその三次元位置を推定する [5] 。この方法では、講師の頭部と足元の高さ (Y 座標) を既知として与えることにより、X 座標、Z 座標の値を求めることができる。本研究では、講師を複数のカメラを用いて観測し、各観測カメラで得られる位置の平均を求めることによって、頭部と足元の位置を推定する。

2.2.3 超音波定位センサによる講師の胸部位置推定

超音波定位センサを用いた講師の胸部位置推定では、超小型ビーコンを講師の両肩にそれぞれ 1 個ずつ取り付け、超音波定位センサによりそれぞれの位置を検出する。両肩に取り付けたビーコンの位置の平均を講師の胸部の三次元位置とみなす。

2.2.4 観測カメラによる受講者位置の推定

受講者は座席に着席しているので、観測カメラ上で各座席ごとに、受講者の存在位置を背景差分及び時間連続フレーム間差分によって推定する。背景差分によって検出され、かつある時間以上動きがある座席に相当する位置に受講者が存在すると推定する。

2.2.5 超音波定位センサによる講師の指示の有無推定

オンラインスライド表示用スクリーン及び電子白板に対する指示の有無は、講師の手首と肘に設置した 2 個の超小型ビーコンの位置を検出することによって行う。検出されたビーコンの平均位置がスクリーンまたは電子白板に近く、かつ肘から手首に向うベクトルがスクリーンまたは電子白板と交差するとき、講師は腕を上げて指し示していると推定する。

2.3 複数のセンサ情報の統合による話者位置推定

それぞれのセンサで得られた複数の情報を統合して、発話している講師または受講者の位置を求める。位置推定の精度は撮影などに必要な程度で十分であるため、講師領域及び受講者領域をそれぞれ人物が占有する程度の大きさの**矩形領域**に分割し、それぞれの矩形領域単位で位置を推定する。講師位置に関しては、推定された頭部・足元または胸部の位置を床平面に投影し、投影された位置を含む矩形領域を講師の位置とする。

講師は移動するので、講師を観測可能な空間を広くするために 2 種類のセンサからの情報を統合して講師位置を推定する。一方、音源位置、講師位置、受講者位置はノイズ等の原因により、複数の位置に推定されることがある。そこで、音源位置として推定された位置に講師または受講者が存在するかどうかという情報を統合することによって、話者の位置を推定する。

まず、講師の位置の統合を考える。これは、超音波定位センサのデータが取得できないときは観測カメラによるデータを利用することによって実現する。これにより、常に講師の位置が推定可能となる。

次に話者の位置推定について考える。音源が存在する矩形領域を求めるため、まず、あるマイク対に関して推定された双曲線が通過する矩形領域に 1 点ずつ投票する。この操作を、すべてのマイク対について行う。そして、得られた得点がある値以上の矩形領域に講師音声の音源または受講者音声の音源が存在する可能性があると推定する。

そして、音源が検出された矩形領域であり、かつ講師または受講者が存在する矩形領域を話者位置であるとみなす。推定された話者位置が講師領域内に含まれる場合は、話者は講師であり、話者状況は L_1 から L_3 のいずれかであると推定する。また受講者領域に含まれる場合は、話者は受講者であり、話者状況は S_i であると推定することができる。

2.4 講師の移動の有無の推定

話者状況が L_1 から L_3 のいずれかのとき、L_2 かそれ以外かは講師の指示の有無で区別できる。L_1 か L_3 であるかの区別は、講師の移動の有無で判断する。講師が移動しているかどうかは、これまでに推定された講師の位置から求める。速度がある値以上のとき、講師は移動していると推定する。

3. 実験及び考察

3.1 実験環境

マイクロホンアレイの設置に関して、講義室全体の音声を獲得できるように、マイク 8 本を図 1 の灰色の丸印で示す位置に、軸を講義室の中心方向に向けて設置した。この 8 本のマイクロホンからの音声信号を PowerDAQ PD2-MF8-300/16 ボードに入力し、振幅を時間同期してデジタル値に変換した信号を CSP 法で処理した。講師領域内の音声位置の検出には講義室の後方にあるマイクを利用して (4, 3), (3, 0), (0, 6), (6, 1), (3, 6) の 5 組のマイク対を構成し

表 3　講師位置の検出精度

センサ種別	平均誤差 (cm)	誤差の分散 (cm)
超音波定位センサ	5.32	0.34
観測カメラ	22.80	3.38

た。受講者領域内の音声位置の検出には講義室の前方にあるマイクを利用して (4, 7), (7, 2), (2, 5), (5, 1), (7, 5) という 5 組のマイク対を構成した。また講師領域を 133 個の矩形領域に分割し、受講者領域を講義室の中心部に設置された座席に相当する 6 行 x7 列の計 42 個の矩形領域に分割した。隣接する矩形領域の中心同士の距離は、講師領域では X 方向、Z 方向共に 50cm、受講者領域では X 方向に 60cm、Z 方向に 90cm である。

講師の頭部の Y 座標の位置は 160cm とする。講師の位置を推定する観測カメラは、5 台の SONY EVI-G20 を図 1 に示す位置 (0, 1, 2, 3, 4) に設置して利用した。

受講者の位置を推定する観測カメラは、受講者同士のオクルージョンを避けるため、図 1 に F で示す天井の位置に魚眼レンズ付きカメラを光軸を真下に向けて設置した。

超音波定位センサは INTERSENSE IS-600 Mark2 X-Bar を利用し、図 1 の×で示す天井の位置に設置した。

3.2　結果及び考察

3.2.1　各センサの検出精度

講師の位置推定に関して、超音波定位センサの検出精度と観測カメラによる位置推定の精度を表 3 に示す。この表から、超音波定位センサを利用した位置検出のほうが検出精度が高いことがわかる。また、推定誤差はどちらも講師領域内の矩形領域の大きさより小さい。

また、受講者位置の推定に関して、20 人の受講者が着席している状況で約 5 分間位置を推定した結果を表 4 に示す。この表から、受講者位置に関して、74.9%の割合で正しく推定できることがわかる。

音声位置推定に関して、講師領域内の矩形領域及び受講者領域内の矩形領域別の正解率を表 5 に示す。受講者領域のほうがマイクロホンアレイに囲まれているため、検出精度が高い。

3.2.2　講義における話者状況の推定結果

約 7 分間の講義を 2 回行い、本手法に基づいて人手で話者状況を判断した結果と、本手法で自動的に推定した結果を比較した。推定結果を表 6 に示す。話者状況推定の正解率とは、講義全体の時間のうち正しく話者状況を推定していた時間の割合である。話者状況変化の正解率とは、ある話者状況が他の話者状況に変わったときの時刻とその前後の話者状況を正しく推定できた割合である。この表から、平均で 83.0%の割合で正しく話者状況を推定できることがわかり、また話者状況の変化に関しても平均で 68.0%の割合で正しくそのタイミングを推定できていることがわかる。

表 4　受講者位置推定の正解率

	正しく推定		誤って推定	
正解	着席	空席	着席	空席
推定結果	着席	空席	空席	着席
平均 (席)	14.1	17.3	5.9	4.7
平均 (%)	33.7	41.2	14.0	11.2
合計 (%)	74.9		25.1	

表 5　音声位置推定の正解率

領域種別	正解率 (%)
講師領域	72.9
受講者領域	92.4

表 6　話者状況の推定結果

推定種別	講義 1 での正解率 (%)	講義 2 での正解率 (%)	平均の正解率 (%)
話者状況推定	88.8	77.1	83.0
話者状況変化	68.6	67.4	68.0

4.　おわりに

本研究では、講義においてマイクロホンアレイ、観測カメラ、超音波定位センサで得られる情報を統合することによって話者状況の推定を行う手法を提案した。本手法を実装し、講義を実験対象として手法の有効性を検証した。本手法によって、話者状況が 83.0%の確率で正しく推定することができた。今後の課題としては、位置推定精度の向上による話者状況の定義の詳細化などが挙げられる。

文　献

[1] 大西正輝, 泉 正夫, 福永邦雄, “情報発生量による遠隔講義映像の自動生成とその評価”, 電子情報通信学会技術研究報告, PRMU98-176, pp.1-6, 1998

[2] K.Ishizuka, Y.Kameda, M.Minoh, “A Study for Distance Learning Service - TIDE Project -” IEEE International Conference on Multimedia and Expo, Vol.3, pp.1237-1240, 2000.

[3] 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”, 電子情報通信学会論文誌, Vol.J84-D-II, no.2, pp.248-257, 2001

[4] M.Omologo, P.Svaizer, “Acoustic event localization using a crosspower-spectrum phase based technique”, Proc.ICASSP94, pp.273-276, 1994

[5] 亀田 能成, 石塚 健太郎, 美濃 導彦, “状況理解に基づく遠隔講義のための実時間映像化手法”, 情報処理学会 研究報告 CVIM, Vol.2000, No.33, No.121-11, pp.81-88, 2000.