社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－SLP－63 (2)
2006／10／20

SOSとマイクロフォンアレイの統合による会議記録システムの開発

木村 文彦　近藤 功一　田村 哲嗣
速水 悟　山本 和彦
岐阜大学大学院工学研究科
〒 501-1193 岐阜市柳戸 1-1
TEL: 058-293-2763
E-mail: fumi@hym.info.gifu-u.ac.jp

あらまし

会議を記録する場合、録音機器を用いて録音することが考えられるが、発言した話者の特定が困難であり、視覚的に話者を識別することもできない。そこで本稿では、話者画像と音声情報を取得し、後に発話の有無や話者画像を表示し、会議を再現するシステムを構築した。全方向のカラー画像と 3 次元情報をリアルタイムに取得できる SOS（Stereo Omnidirectional System）と、3 次元空間でリアルタイムに音源方向を推定できるマイクロフォンアレイを統合し、リアルタイムで音源方向を推定し、全方向画像上に音源位置を示した。また、会議を模擬した状況で話者を特定する精度を検証した。

Development of Meeting Record System by Integration of SOS and Microphone Array

Fumihiko Kimura　Kouichi Kondo　Satoshi Tamura
Satoru Hayamizu　Kazuhiko Yamamoto
Graduate School of Engineering, Gifu University
Yanagito, Gifu-shi, 1-1, Japan
TEL: +81-58-293-2763
E-mail: fumi@hym.info.gifu-u.ac.jp

Abstract

To record a meeting, a sound recorder is used in most cases. However, specification of the speaker who spoke is difficult and cannot identify a speaker visually, either. In this paper, an audio-visual integration system is presented which uses a microphone array to presume the directions of a sound source and also uses SOS to acquire color images of all the directions and three-dimensional information on real time. This system acquires a speaker and voice information simultaneously and displays the existence of utterance and the speaker images to reproduce the meeting afterwards. The accuracy which specifies the speaker of the system in a simulated meeting was verified.

1 はじめに

われわれが社会活動をするにあたって会議は非常に重要なものであり、会議を記録しておき、後に参加していない人が視聴することができるシステムが求められる。

会議を記録するにあたって、まず考えられるのがカセットレコーダーなどの録音機器を用いて録音することである。しかし会議の内容を後で再生するためには、時間がかかる。また話者の様子や身振りなどの画像情報が含まれていない。このような問題に対してわれわれは画像と音声を同時に取得し、記録する会議システムを開発している。

これまでマイクロフォンアレイを用いた 3 次元空間における発話者・音源の方向をリアルタイムで推定するシステムを構築[1,2]したが、本研究ではこのシステムと全方向ステレオシステム（SOS）[3]を統合した。会議システムの関連研究としてはマイクロフォンアレイ信号処理を用いて、発話中の相槌や他の話者の割り込みを分離し、音声認識の有効性を確認した研究[4]やハンズフリーの電話会議において遅延和アレー法による話者方向推定と適応型アレーによる指向性形成を用いて話者ごとに自動音量調整を行う装置の開発など[5,6]がある。また、マイクロホンアレーと全方位画像を用いた、発話者方向推定の研究[7]もある。

本研究では、全方向の画像を一挙に映し出すことができ、距離情報も得ることができる全方向ステレオシステム（SOS）と音源方向を推定できるマイクロフォンアレイを統合することで画像情報と音声や音源方向といった音声情報を同時に取得するシステムを開発した。さらに取得した画像情報と音声情報をメタデータ化した上で保存し、後に取得データから会議を再現することを可能とした会議記録システムを構築した。

2 会議記録システムの構築

2.1 SOS とマイクロフォンアレイの統合

マイクロフォンアレイで推定した音源方向を SOS に送信することで、リアルタイムで全方向画像上に音源位置を枠で表示させた。また、同時に音声と音源方向、話者画像を取得することができ、ある時間帯での発話音声の話者の特定が可能となった。

2.2 メタデータ化

マイクロフォンアレイで取得した音声と音源方向、時間、SOS で取得した話者画像をメタデータ化して保存する。メタデータとはデータについての情報を記述したデータであり、多くの記録された会議からのデータの検索を容易にしたり、個々の会議の様子を一覧表示したりするためのものである。

2.3 会議記録システム

メタデータ付けして保存しておいた音声、音源方向、時間、話者画像を用いて、会議を再現する。システム全体の流れを図 1 に示す。

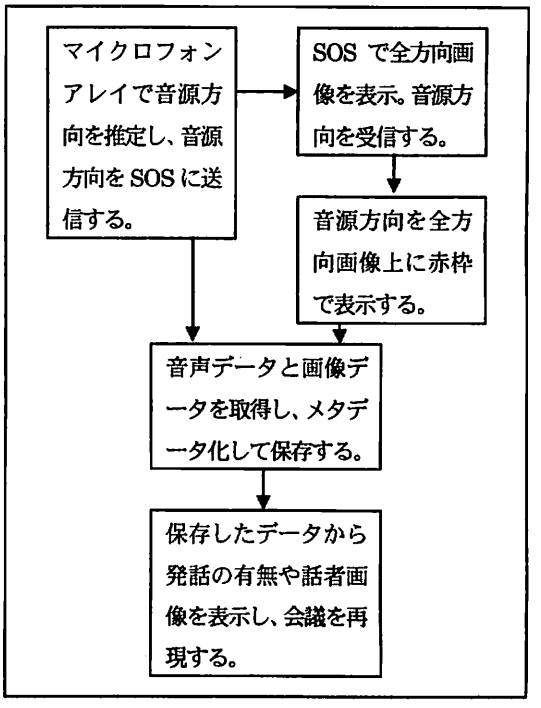

図 1 会議記録システムの流れ

3 SOSとマイクロフォンアレイの説明

3.1 全方向ステレオシステムの説明

SOSは36個のカメラを用いて正12面体の各面上に1組3眼のステレオビジョンユニットを配置することによって、リアルタイムに観測点を中心とした全ての方向の3次元カラー情報を均一な解像度で同時に得ることができる。SOSの外観を図2に示す。

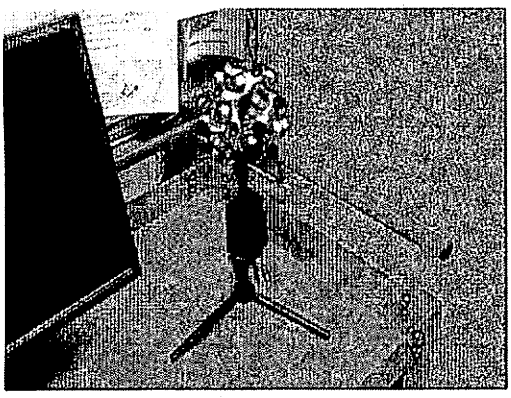

図2 SOSの外観

また、1個のステレオユニットから得られた視差画像を図3に示す。視差とは各カメラで取得した画像の差異から遠近や奥行きを認識することである。図3は視差が大きい点ほど輝度値を高く表示している。黒い部分は、視差情報が得られなかった部分を示している。このようにSOSはカラー画像に加え、距離画像も取得できていることがわかる。その元画像を図4に示す。

図5は各ユニットで得られた画像を張り合わせて、全方向画像として表示したものである。

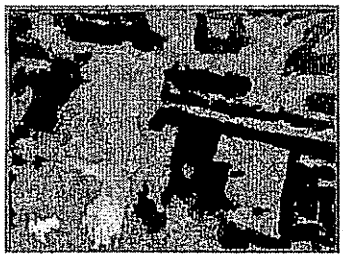

図3 視差画像

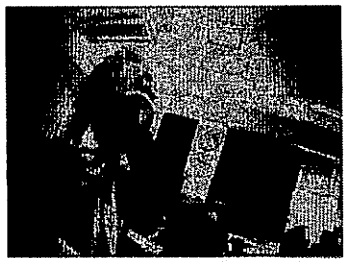

図4 図3の元画像

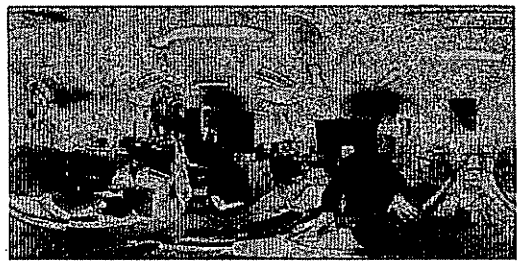

図5 全方向画像

3.2 マイクロフォンアレイの説明

マイクロフォンアレイは正20面体の頂点に12個のマイクを取り付けてある。マイクロフォンアレイと対になっているものが多チャネルリアルタイム音響信号処理装置、rasp[8]である。このシステムは複数のマイクロフォンで音を収集し、これより複数の音源の位置を推定し、この情報を元に目的となる音源の信号を、他の環境雑音などと分離するための装置である。この2つの装置でリアルタイムでの音源方向の推定が可能である。マイクロフォンアレイを図6に示す。

このマイクロフォンアレイでは音源方向推定のための計算法としてMUSIC法を用いている。MUSIC法は相関行列の固有値・固有ベクトルを用いた到来方向推定法であり、マイクの位置が直線や円形にしないといけない等の物理的制約にとらわれることなく音源を推定することができる。また、音の到来方向に対しては非常に鋭いピークをもち、その結果、複数の音が存在する場合の分離性能や雑音耐性にも優れているという特徴をもっている。raspの動作例を図7に示す。

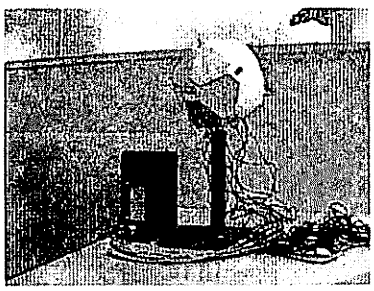

図6　マイクロフォンアレイの外観

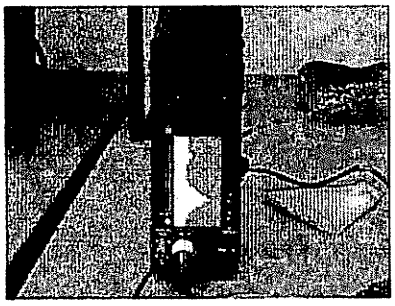

図7　raspの動作例

4　実験

4.1　実験条件

本システムが実際の会議において発話した話者を特定する精度を調べるために、評価実験を行った。実験を行った部屋の大きさは10m×10m程度の大きさである。その部屋の中心にマイクロフォンアレイとSOSを置いた。そして、図8のようにマイクロフォンアレイの位置から距離1.5mの角度60°、120°、270°の場所に椅子を置き話者3人をそれぞれ座らせた。そこで模擬的な会議を行い、その状況においてマイクロフォンアレイでは音声データ、SOSで画像データを取得した。その後に取得データを用いた本システムでの会議の再現状況、発話した話者を特定する精度を調べた。音源方向推定は2音源を推定したうち第1候補のみを用いた。また、音源方向推定の処理間隔は1秒である。音源方向推定の精度は水平方向が5°刻みの72方向、垂直方向は上、中、下の3方向であるが、実験では、水平方向の15°刻みで評価した。

話者には以下に示す条件での会議を模擬してもらった。

・パターン1…3人のうち1人がリーダー役、2人が部下になってスケジュール調整をする。リーダー役が部下のスケジュールをきいて会議の日時を決定する。(総時間1分15秒)

・パターン2…パターン1と同じ配役で、プログラムの開発状況について話し合う。その後今後の開発の進め方について話し合い決定する。

(総時間4分54秒)

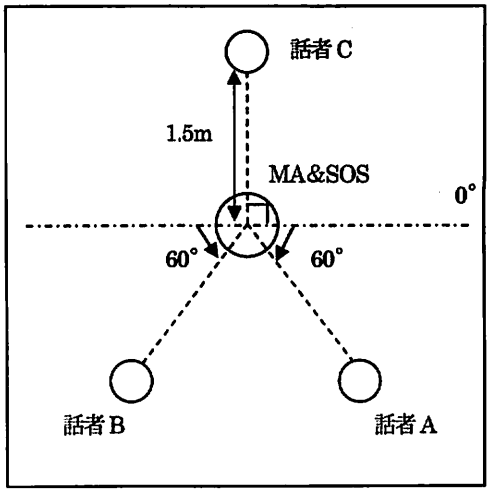

図8　話者の位置

4.2　実験結果

表1　話者方向検出率

	発話パターン	検出率（%）
I	パターン1	81.3
II	パターン2	73.5

表1は本システムが検出した話者方向の検出率を示したものである。話者方向として正確に検出できたかどうかは、音声を聴いてその時点で判断した話者の方向と本システムが検出した音源方向が15°の範囲内であった場合を正解とした。

話者検出率を算出した式は次のとおりである。

検出率[%]=

$$\frac{\text{話者方向を正確に検出 できた時間}}{\text{会議の全時間}} \times 100$$

表2　話者Bの検出率

	発話パターン	検出率（%）
Ⅲ	パターン1	100
Ⅳ	パターン2	84.7

表2は、比較的、話している時間が長かった話者Bのみの検出率を示したものである。

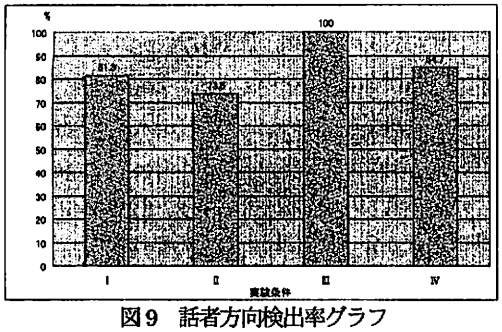

図9　話者方向検出率グラフ

また、リアルタイムシステムを開発し、SOSによって映し出された全方向画像上に、音源方向を表示させることができた。そして、取得した音声、音源方向、話者画像から会議を再現するアプリケーションを作成した。全方向画像上の音源位置表示を図10に、会議を再現するアプリケーションの画面を図11に示す。

図11のアプリケーションは、発話の有無を音源方向別に表示させ、音声の再生にあわせて時点をあらわすバーが移動するようにした。動作はボタンで再生、一時停止、停止ができる。また、その各時点で発話している話者画像を右上のウィンドウに表示するようにした。そして、全方向画像上の発話者の位置を枠で表示し、一目で誰が発話しているのかがわかるようにした。

図10　全方向画像上の音源位置表示

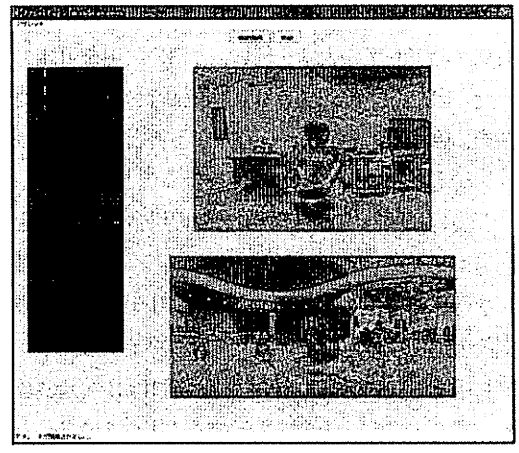

図11　会議を再現するアプリケーション画面

4.3　考察

実際の会議に近い状況での音源方向を推定したため、1人の話者が発話している際に相槌をうつことや、相手の話しに割り込むことが起こるために検出率が低下していると考えられる。その他、発話している話者以外の人が物音をたてたり、咳こんだりすると、そのような音は発話音声に比べてパワーが大きいために推定方向を誤ってしまうと思われる。ただし今回の実験の場合、2音源を音源方向推定した第1候補の方向のみを使用して検出率を計算しているため、第2候補の方向も含めて考えると、検出率がもう少し改善すると思われる。また、話者の発話が短かったり、小さかったりした場合に特定しづらいことがわかった。今回の実験では1秒間隔で音源方向推定の処理をしているため、1秒以下の発話は方向を推定が難しい場合があると思われる。

5 まとめ

全方向画像を一挙に映し出し、距離情報も得ることができるSOSと、音源の方向をリアルタイムに高精度で推定できるマイクロフォンアレイを統合することで視覚的に音源の位置を確認することが可能になり、画像情報と音声情報を同時に取得することができるようになった。データ検索が容易な形にするため、取得データをメタデータ付けして保存した。後に保存データを用いて、時点ごとの発話の有無や話者画像の表示を行い、会議の再現を可能とした。また、評価実験において、話者を識別する範囲を、話者の左右の動きを考慮することで会議記録システムとしての一定の精度を得ることができた。

今後の課題として、音源が2つの場合における音源方向の全方向画像上への表示と発話の有無の表示を行うこと、会議を再現するアプリケーションの機能の追加を予定している。また、マイクロフォンアレイとSOSの装置自体を統合すれば3次元の音源位置を全方向画像上に表示することができる。いずれはSOSでの肌色検出の画像処理を行い、話者の位置を画像の側からも補正すること、あるいは音があっても画像上人物が存在しなければ、発話ではなく物音として処理することなどが考えられる。また、SOSは距離情報も得られるため、マイクロフォンアレイでは難しい音源までの距離を画像側から得ることができる。このように、マイクロフォンアレイとSOSによって得られた音声情報と画像情報を統合して利用することで、お互いに補完しあう仕組みを目指している。

参考文献

[1] Fumitaka Ban, Satoru Hayamizu: "Estimation of Sound Source Direction Using a Real-time Microphone Array System in a 3-D Environment"Proc.VSMM2004,pp.460-466 (2004)

[2] 坂文貴, 速水悟: "マイクロフォンアレイを用いた実環境音の認識による音源定位", 日本音響学会講演論文集Ⅰ, 春季, pp.615-616(2005)

[3] 山本和彦, 棚橋英樹, 桑島茂純, 丹羽義典: "実環境センシングのための全方向ステレオシステム(SOS)", 電気学会論文誌 C, Vol.121-C, No.5, pp.876-881(2001)

[4] 浅野太, 緒方淳, 松坂要佐, 山田実一, 中村雅巳: "会議収録データにおける発話イベントの構造化と分離について", 日本音響学会講演論文集, 秋季, pp.29-30(2006)

[5] 小林和則, 羽田陽一, 日和崎祐介, 大室仲, 入島勉, 中山圭一, 阿部匡伸: "方向別AGC機能のIP電話会議装置への実装", 電子情報通信学会講演論文集, D-14-12, pp.136(2006)

[6] Ziyou Xiong, Xiang Sean Zhou, Qi Tian, Yong Rui, Thomas S. Huang "VIDEO OF MEETINGS", Signal Proc Magazine Vol. 23 No. 2 March , pp.21-22(2006)

[7] 傳田遊亀, 西浦敬信, 山下洋一: "マイクロホンアレーと全方位画像を用いたマルチモーダル発話者方向推定", 日本音響学会講演論文集, 春季, pp.211-212(2006)

[8] 浅野太, 古河弘光, 釜島力: "マイクロフォンアレイ信号処理用ハードウェアの試作", 日本音響学会講演論文集Ⅰ, 秋季, pp.499-500(2003)