2008－CVIM－165 (9)
2008／11／27

多人数会話シーン分析に向けた実時間マルチモーダルシステムの構築

—マルチモーダル全方位センサを用いた顔方向追跡と話者ダイアリゼーションの統合—

大塚 和弘　荒木 章子　石塚 健太郎　藤本 雅清　大和 淳司

日本電信電話株式会社
NTT コミュニケーション科学基礎研究所

E-mail : otsuka@eye.brl.ntt.co.jp, {shoko, ishizuka, masakiyo}@cslab.kecl.ntt.co.jp, yamato@brl.ntt.co.jp

あらまし　本稿では，複数人の対面会話シーンの分析に向けた実時間マルチモーダルシステムを提案する．このシステムでは，基本的な会話の状態を知るために，「誰がいつ話しているか」という話者の同定（話者ダイアリゼーションと呼ぶ），及び，「誰が誰をみているか」という視覚的な注意の焦点の推定を実時間で行うことを目標とする．まず，会話シーンを観測するために，2 台の魚眼レンズ付きカメラと 3 本のマイクからなる全方位マルチモーダルセンサを提案する．次に，全周画像上にて会話参加者の顔の位置と方向の推定を行う．ここではその方法として STCTracker(疎テンプレートコンデンセーション追跡法) と呼ばれる方法を採用し，これを GPU(グラフィックスプロセッシングユニット) と呼ばれる並列ハードウェア上にて実行する．また，マイクからの音響信号に対して，音声区間検出と音声到来方向推定を組み合わせた話者ダイアリゼーションを行う．さらに分析の結果を三次元的に可視化する方法も提案する．画像と音響の処理にそれぞれ一台の PC を用い，5 人会話に対して平均 27.1[frame/sec] にて動作することを確認した．

A Realtime Multimodal System toward Multiparty Conversation Scene Analysis

— Integrating Face Pose Tracking and Speaker Diarization using Multimodal Omnidirectional Sensors —

Kazuhiro Otsuka, Shoko Araki, Kentaro Ishizuka, Masakiyo Fujimoto, and Junji Yamato

NTT Communication Science Laboratories,
NIPPON TELEGRAPH AND TELEPHONE CORPORATION

Abstract　This paper presents a realtime system for analyzing group meetings that uses a novel omnidirectional camera-microphone system. The goal is to automatically discover the visual focus of attention (VFOA), i.e. "who is looking at whom", in addition to speaker diarization, i.e. "who is speaking and when". First, a novel sensing device is presented; it consists of two cameras with two fisheye lenses and a microphone array. Second, from omnidirectional images captured with the cameras, the position and pose of people's faces are estimated by STCTracker (Sparse Template Condensation Tracker); it realizes realtime tracking by utilizing GPUs (Graphics Processing Units). The face position/pose data is used to estimate the focus of attention in the group. Using the microphone array, robust speaker diarization is carried out by a VAD (Voice Activity Detection) and a DOA (Direction of Arrival) estimation. This paper also presents new 3-D visualization schemes for the results of an analysis. Using two PCs, one for vision and one for audio processing, the system runs at 27.1[frame/sec] on average for 5-person meetings.

1　はじめに

我々人間は，情報の共有や感情の伝達，集団での意志決定などを目的とし，日々様々なコミュニケーションを行なっている．こうした人と人とのコミュニケーションをより円滑にする情報技術を開発するためには，そもそも人間がどのようにコミュニケーションを行っているか知ることが重要である．近年，「環境知能」[1] と呼ばれる，環境に埋め込まれた機械知性が人間の活動を見守り，状況に応じて問い掛け，サポートするような新しい情報技術のパラダイムが提唱されているが，人間同士のコミュニケーションシーンの分析，及び，その結果に基づくコミュニケーション環境の構築は，そのような環境知能の一つとして捉えることができる．また，その具現化の例としては，遠隔映像会議システム，会議映像アーカイブ・要約システム，社会的ロボット・エージェントなどへの展開が期待される．

人間のコミュニケーションの中で，複数の人物が対面の状況で行う「会話」は，最も基本的なコミュニケーションの形態といえる．対面会話において，会話参加者は言語情報だけではなく，非言語情報の交換を行っている．非言語情報には，視線や顔表情，ジェスチャ，手振りや身振り，声のトーンなどが含まれ，これらは重要な役割を果たし

ている [2]．著者らは，これら会話参加者が発する情報を画像・音響信号として観測し，そこから参加者間でのメッセージ送受信の過程や心的状態の変化を自動的に推測するというタスクを会話シーン分析と呼び，その実現を目指して研究を進めている．

近年，画像や音声などのマルチモーダル情報に基づいて，ミーティングなどの対面会話の認識・理解を目指す研究分野が急速に発展している [3]．しかし，今日まで多くの研究は，記録済データのオフライン処理を前提としている．著者らは，遠隔会議システムや会議支援システムなどの応用のためには実時間での分析は必須であると考え，本稿で提案のシステムの開発に取り組んでいる．また，ミーティングアーカイブの構築など，実時間動作が必ずしも必須ではない応用に対しても，その効率化のために高速なシステムは有用であると考える．

以上を踏まえて，本稿では，複数人の対面会話シーンの分析に向けた実時間マルチモーダルシステムの提案を行う．本稿にて提案するシステムでは，初期の目標として，話者ダイアリゼーション，及び，視覚的な注意の焦点の推定に取り組む．ここで話者ダイアリゼーションとは，「誰がいつ話しているか」という話者を推定する問題を指す．また，視覚的な注意の焦点とは，「誰が誰を見ているか？」という視線の方向のことを指す．これらの推定により「誰が誰に向かって話をしているか」といった会話の基本的な構造を知ることが可能となる．

本システムは，小規模の円卓ミーティングを対象とした全方位マルチモーダルセンサを新たに導入する．このセンサは 2 台のカメラと 3 本のマイクロホンから構成される．各カメラには魚眼レンズが装着されておりほぼ全周の画像を得ることができる．この画像上にて会話参加者の顔の位置と方向の推定が行われ，その情報は注意の焦点の推定に用いられる．本システムでは，顔の位置・方向の推定のため，STCTracker(疎テンプレートコンデンセーション追跡法) と呼ばれる方法を採用し，これを GPU(Graphics Processing Unit) 上にて並列実装することで追跡の実時間化を実現した．また，マイクからの音響信号を入力として，音声区間検出 (VAD=Voice Activity Detection) 並びに音声到来方向 (DOA=Direction Of Arrival) 推定を行う．これら画像処理と音響処理の結果を統合することで会話の状態（各人物の発話の状態と注意の焦点）が推定される．

このシステムは画像と音の処理にそれぞれ 1 台の PC を用い，5 人会話に対して平均 27.1[frame/sec] にて動作する．著者らの知る限り，本稿で提案するシステムは，全方位マルチモーダルセンサ，及び，顔方向追跡を組み込んだ多人数会話シーンの分析システムとして最初に実時間での稼働を実現したシステムである．

本稿は以下のように構成されている．まず，第 2 節において関連研究を概観し，続く第 3 節で我々のシステムを提案する．第 4 節にてシステムの構成を紹介し，第 5 節において実験結果とその評価結果を記述する．最後に第 6 節において，本稿のまとめと今後の課題について述べる．

2 関連研究

全方位カメラによる撮影　会話シーンの撮影には，いわゆる「全方位カメラ」がよく用いられている．その構成としては，双曲面ミラーなど反射光学系と一台のカメラの組み合わせ [4, 5, 6, 7, 8] や，全周をカバーするよう複数のカメラを一つの筐体内に配置したもの [9, 10, 11]，ミラーと複数カメラの組み合わせ [12] などが知られている．ミラーとカメラ 1 台を用いた全方位カメラはその光学的特性により人物の顔領域の画像解像度が低いという問題がある．一方，複数のカメラを用いる場合には，それら画像の合成により高解像度の全周画像が得られるが，画像の境界にて不連続部分が生じ，その部分では正確な顔追跡などが難しいという問題がある．本研究では，画像の解像度を最大化し，かつ不連続箇所数を最小化する構成として，高解像度 CCD カメラと魚眼レンズを用いた新しい全方位カメラの構築を試みる．

顔の検出と方向推定　会話中の人物の位置や顔の向きなどを知る手段として，画像上での顔検出や顔追跡の利用が有望視されている．顔の位置のみを検出した例として，文献 [13, 5, 6, 8, 10] などがあげられる．また，顔の方向まで推定する方法としては，左右中央など粗く離散化した顔方向 [14, 7] や，連続な顔方向角 [15, 16, 4] を検出・推定する方法があげられる．多人数会話シーン分析の文脈にて，これら顔方向の推定を実時間で行ったシステムはこれまで報告されていない．本稿の提案システムでは，以下で述べる視線方向の推定のため，連続な顔方向角を顔の追跡によって高精度かつ実時間にて推定する方法を組み込んでいる．

視覚的注意の焦点の推定　対面会話において，視線は，他者のモニタリングや態度・興味の表出，ターンテーキングなど会話の流れの制御において重要な役割を果たしており [17]，会話の状況を理解する上で視線の方向（視覚的注意の焦点）を自動計測することは重要な課題とされる．また，「誰が誰に話し掛けているか」「誰が誰の話を聞いているか？」「誰が誰に反応しているか？」といった「誰が誰に」(who to whom) という情報は，会話中のインタラクションを記述する上で必須であり，それを特定する上で視覚的な注意の焦点（方向性）は重要な手掛かりである．しかしながら，会話中の視線方向を直接的かつ非侵襲的に計測することは困難であるため，視線の代用として，顔の方向から視線方向を推定する方法が研究されている [4, 18, 16, 19]．本稿のシステムも同様のアプローチをとる．

マイクアレーによる話者ダイアリゼーション　従来，話者ダイアリゼーションを行うため，6 本のマイク [12]，8 本のマイク [9, 7, 10, 11] や 16 本のマイク [8] からなるマイクアレーがよく用いられている．それに対して，著者らのグループでは，平面上での音源方向を推定するための最小のマイク本数である 3 本のマイクを用いてダイアリゼーションを行う方法を提案しており [20]，本システムでもそれを採用する．マイクの本数が少ないことでコンパクトなセンサーが構成可能である点が利点である．

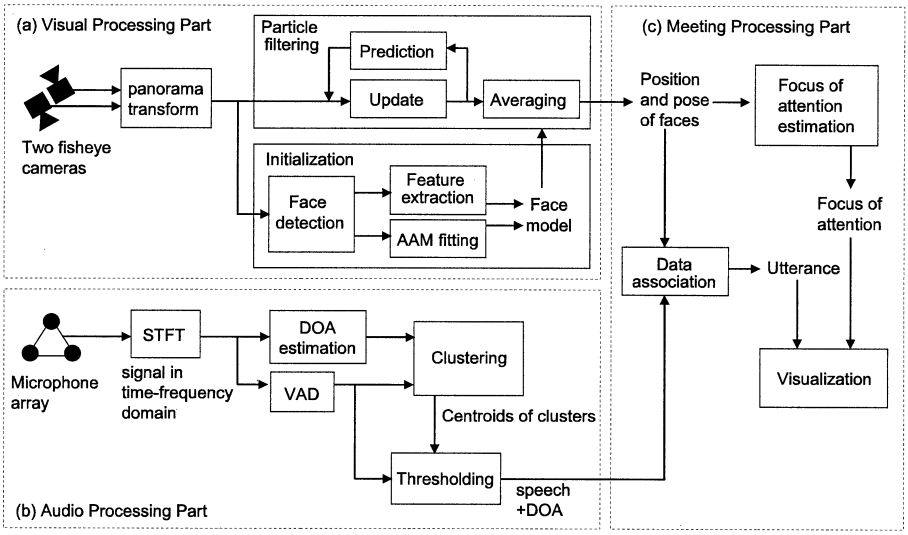

図 1: Diagram of system

全方位マルチモーダルシステム　全方位カメラとマイクアレーを用いて会話の場面を記録・分析するシステムが複数提案されている．遠隔映像会議などを念頭に実時間動作するシステムとして，マイクロソフトの RingCam[9], RoundTable[12], USC の SmartRoom[8] などが知られている．また，会話場面の記録，及び，後日の閲覧・分析を主眼においたシステムとして，AMI ミーティングルーム [11], 産総研の MARC[10], 阪大の MMLogger[7] などがあげられる．前者のシステムについては，発話検出に基づく話者画像の提示などが実時間で実現されている．しかし，顔方向追跡やさらに上位の会話シーン分析の機能は含まれていない．また，後者のシステムでは，顔方向推定などが行われるが，実時間での稼働は報告されていない．これら従来のシステムに対して，本稿で提案するシステムでは，発話検出に加えて，顔方向推定による視覚的注意の焦点の推定まで実時間で行う点にて新規性を有する．

3　マルチモーダル会話シーン分析システム

図 1 に，提案システムの構成を示す．このシステムは (a) 画像処理部，(b) 音響処理部，(c) 会話処理部から構成される．画像処理部は，全方位カメラから得られる画像上において顔方向の追跡を行い，顔の位置と方向の推定を行う．また，音響処理部においては，話者ダイアリゼーションを行う．この処理は，音声区間検出，及び，その音声の到来方向の推定からなる．最後に会話処理部では，これら推定結果を統合することで，発話者とその発話区間の特定を行う．また，各人の頭部方向からは各参加者の視覚的注意の焦点（視線方向）の推定を行い，会話の場において注目されている人物の同定を行う．これら結果は実時間でディスプレイ上に表示される．

3.1　画像処理

画像処理部は，全方位カメラによる撮影，並びに，得られた画像上での顔方向追跡から構成される．

3.1.1　魚眼レンズを用いた全方位カメラ

図 3 に，全方位マルチモーダルセンサの外観を示す．このセンサは 2 つのカメラと 3 本のマイクロホンから構成される．各カメラには魚眼レンズが装着されている．この魚眼レンズはおよそ半球の領域をカバーできるため，これらを 2 台背中合わせに配置することで，おおよそ全球の領域が撮影できる．図 4(a) には，撮影された画像の一例を示す．提案システムでは，会話参加者の顔領域を含むような水平の帯状の部分のみを撮影する．なお，図 4(a) のように，このセンサにて得られた画像には 2 箇所の不連続部分が存在するため，会話参加者の座席配置には配慮が求められる．本システムの魚眼レンズの射影方式は，等距離射影（通称 $f \cdot \theta$）である．この方式では，世界座標系上の一点が画像平面上の一点に投影されるとき，その画像中心からの距離は入射角 θ に比例する．この射影方式に基づいて，魚眼レンズで得られた画像はパノラマ画像へと変換される（図 4(b)）．

3.1.2　顔の位置と方向の推定

本システムでは，人物の顔の位置とその方向（姿勢）を推定する方法として，STCTracker(Sparse Template Condensation Tracker=疎テンプレートコンデンセーション追跡法)[21] を採用する．この方法は，元々は岡山大にて提案された方法 [22] であるが，著者らによって，顔に特化したテンプレートの三次元化，自動初期化，GPU 実装など独自の改良・拡張 [21] が施されたものである．これまで，著者らによりオフラインの会話シーン分析での有効性が確

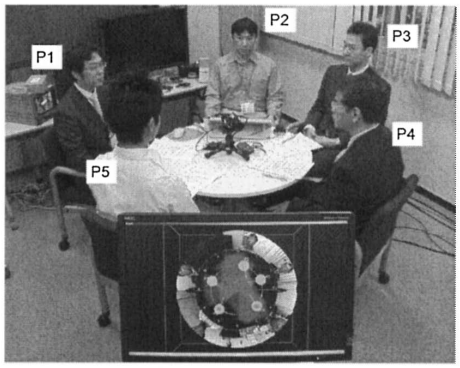

図 2: Meeting scene. LCD on near side shows the result of realtime processing.

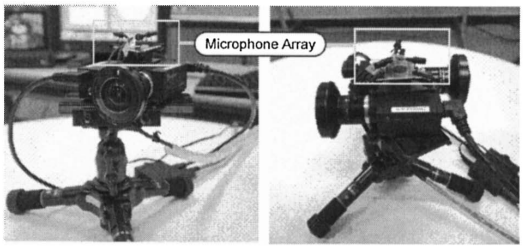

図 3: Omnidirectional camera-microphone system

認されている [23]. 本手法の特徴としては，頭部の水平方向の回転に対する頑健性 (〜約 60 度) やその精度 (平均絶対誤差 < 4[deg])，及び，GPU 実装による実時間性 (CPU と比較して約 10 倍高速) があげられる [23].

STCTracker は，疎テンプレート照合を基本原理とする．疎テンプレート照合とは，通常のテンプレート照合とは異なり，疎な画素の集合のみを用いる照合法である．テンプレートの状態は，画像上での顔の位置 (2 自由度)，各軸周りの回転 (3 自由度)，スケール (1 自由度)，及び，照明変動の補正項 (1 自由度) の合計 7 次元のベクトルとして定義され，パーティクルフィルタにより逐次的にその事後確率密度分布が推定される．

図 1 の左上に STCTracker の構成図を示す．STCTracker は，初期化とパーティクルフィルタの部分から構成される．初期化の段階では，画像上から正面顔を検出し，顔のテンプレートを構成し，パーティクルの初期分布を生成する．まず，Viola & Jones による顔検出器 [24] を用いて正面顔を検出し，検出された顔画像に対して，Active Appearance Model による顔モデルのあてはめが行われる．その結果，得られる顔部品や輪郭の 2 次元座標を用いて，既存の平均 3 次元顔形状モデルを変形させることで検出された顔の 3 次元形状を近似的に得る．また，画像上の顔領域内において特徴点の抽出が行われ，特徴点の 3 次元座標，及び，輝度値の集合として疎テンプレートが構成される．

パーティクルフィルタは更新，予測，平均化の各段階により構成される．更新の段階では，各パーティクルの重みの計算が行われる．この重み付きパーティクルの集合としてテンプレートの状態の事後分布が表現される．ここでパーティクルの重み計算は，その並列性を利用して GPU 上にて実行される．予測の段階ではリサンプリング処理と，次時刻におけるパーティクル分布の予測が行われる．この更新の段階と予測の段階は，各画像フレームについて繰り返し実行される．平均化の段階では，更新の段階より得られるパーティクル分布より点推定値（分布の平均値）が計算され，顔の位置・姿勢として出力される．

3.2 音響信号処理

図 1 左下部 (b) の音響処理部において，マイクアレーにより得られる音響信号を処理することで話者ダイアリゼーションが実行される．図 3 に示すように，このマイクアレーは，三角形の頂点上 (一辺 4cm) に配置された 3 本のマイクロホンにより構成される．話者ダイアリゼーションの方法として，著者らのグループが提案している文献 [20] の方法を用いる．図 1 の左下部 (b) にその流れを示す．まず，短時間フーリエ変換 (STFT) により観測信号の時間周波数表現を得る．次にこれを入力として，VAD により発話活動の検出（人の声と雑音との判別）が行われる．更に人の声の到来方向 (DOA) を推定し，その方向のクラスタリングにより (潜在的) 話者の発話の状態を得る．

従来，音声到来方向の推定法として，GCC-PHAT 法 [25] が ICSI[26] や CHIL[27] などで一般的に用いられている．しかし，GCC-PHAT 法では，一つの音声フレームでは一つの音源方向のみ存在するという拘束条件が必要であるため，重複発話などの場面では話者を正確に求めることが難しい．この問題を回避するため，本研究では著者らの提案による時間周波数領域における到来方向推定 (TFDOA=Time-Frequency DOA)[20] を用いることとした．なお，人物数などの情報は未知とする．本システムで用いている音響処理の特徴としては，VAD による耐雑音性，及び，ノンパラメトリックな DOA 推定によって人数が不問 (マイク本数より多い人数も可) である点が上げられる．

3.2.1 音声区間検出 (VAD=Voice Activity Detection)

本システムでは，VAD の方法として，“Multi Stream Combination of Likelihood Evolution of VAD” (MUSCLE-VAD)[28] を用いる．この方法は，発話・非発話の弁別器として 2 種類の方法を組み合わせたものである．一つは PARADE と呼ばれる信号の周期成分と非周期成分との比率に基づく方法であり，もう一つは，スイッチングカルマンフィルタ (SKF=Switching Kalman Filter) を用いた方法である．PARADE は，突発的なノイズに対して頑健であり，SKF を用いた方法は，定常ノイズと非定常ノイズの双方に対して頑健である．よってこれら 2 つの方法を用いた MUSCLE-VAD 法は，幅広い種類のノイズに対して頑健であることが知られている．

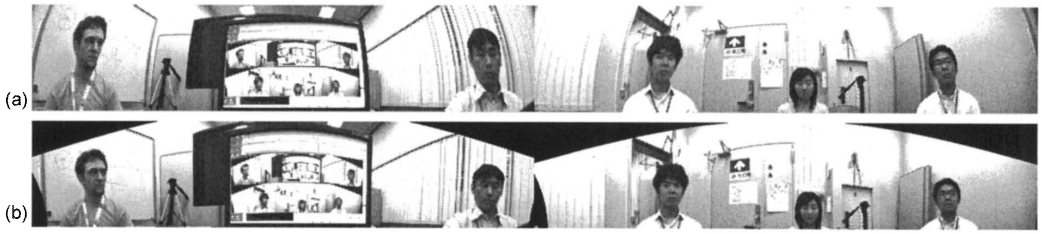

図 4: Camera images, (a)Fisheye images, (b)Panorama images. Two images from each camera are aligned side by side to form 360-degree view (image size = 4896×512 pixels).

3.2.2 音声到来方向 (DOA=Direction of Arrival) 推定

本稿では，GCC-PHAT 法の代わりに，時間周波数領域 DOA(TFDOA=Time-freqency domain DOA) 法を用いる．TFDOA 法は，まず，各時間周波数スロットにおいて，話者の方位角情報を含んだ DOA ベクトルを出力する．次に，前節の VAD で得られた音声区間における DOA ベクトルをオンラインクラスタリング (leader-follower 法) によりクラスタリングする．各クラスタが各々一つの音源 (=各話者) に対応するため，これにより各話者の発話区間を推定できる．このオンラインクラスタリングにおいて，各時間周波数の DOA と，既存クラスタ中心の DOA との距離が閾値以上の場合に，新たな話者を検出する．このため人物数が未知であっても適切なクラスタリングを行いダイアリゼーションを実行することが可能である．

3.3 会話処理

会話処理部では，画像処理と音響処理の結果を入力とし，会話の状態の推定を行う．現時点では，会話の状態を記述するための最も基本的な要素として，話者の特定と視覚的注意の焦点の推定を実施している．話者の特定は，画像より推定された各人物の顔の位置 (方位角) と音響信号より推定された人の声の到来方向の情報の統合により行われる．この問題は，一種の data association 問題として捉えることができる．現在は，単純な方位角の比較と閾値処理により，話者を特定している．

また，視覚的注意の焦点の推定には，画像より得られる顔の位置 (カメラに対する方位角) と顔の方向の情報が用いられる．本研究では，注意の焦点として「他の参加者のうち一人を見ている」あるいは「誰も見ていない」という離散化された対人視線方向を推定の対象とする．ここでは，ある人物がある他者に視線を向けている時の顔方向の分布を表す尤度関数を導入し，最尤法により視線方向を推定する．この尤度関数として文献 [18] でも採用しているガウス分布を採用する．なお，本システムでは，人物の位置として，カメラに対する相対的な方位角は得られるが，空間中の 3 次元座標は得られない．そのため，円卓の周囲に会話参加者が着席している状況を想定し，カメラと各人物の間の距離が等しいという仮定をおく．

図 5 は 2 人物間の相対的な位置関係を図示したもので

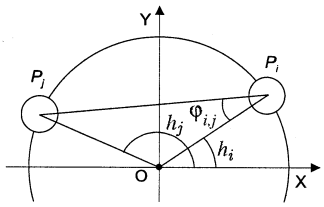

図 5: Spatial configuration of participants and their relative angles

あり，カメラが座標系の原点 O に置かれている．この図において各人の位置は方位角 $h_i, (i = 1, \cdots, N)$ で表される (N は，人物数である)．この方位角はパノラマ画像上での顔の水平方向の座標値より得られる．顔の水平方向の回転角 r_i は，顔がカメラ方向に正対する時に 0 となる．また，ある人物 P_i の顔が他者 $\mathrm{P}_j, (i \neq j)$ の方に真っ直ぐ向けられているときの人物 P_i の顔の回転角を $\varphi_{i,j}$ と印す．この角度 $\varphi_{i,j}$ は，

$$\varphi_{i,j} = -\tan^{-1}\left[1/\tan\left((h_i + h_j)/2\right)\right] \qquad (1)$$

のように得られる．この角度 $\varphi_{i,j}$ を用いて，人物 P_i が他者 P_j を見ているときの人物 P_i の顔方向 r_i の尤度関数を

$$L(r_i|X_i = j) := N(r_i|\kappa \cdot \varphi_{i,j}, \sigma^2), \qquad (2)$$

として定義する．ここで $X_i = j$ は，人物 P_i の視線方向が人物 P_j を向いていることを表す．また，$N(\cdot|\mu, \sigma^2)$ は，平均 $\mu = \kappa \cdot \varphi_{i,j}$，分散 σ^2 のガウス分布を表す．κ は定数 (本システムでは 1) を表す．また，誰も見ていない時の頭部方向の尤度関数として一様分布を用いる．このような尤度関数を用いて最尤法により各時刻における各人の視線方向を推定する．また，各人が何人の他者から見られているかをカウントすることにより，グループにおける注意の焦点を検出する．

4 システムの構成

図 6 に本システムのハードウェア構成を示す．本システムは，画像処理 (含む，会話処理)，及び，音響処理に各 1 台の PC を用いる．画像処理用の PC は，CPU Intel Core2Extreme QX9650 3.0GHz，GPU NVIDIA GeForce9800GX2(または GeForce GTX280)，OS WindowsXP SP2，カメラ Point-GreyResearch Grasshopper (B/W 500 万画素, 2/3" CCD)，魚

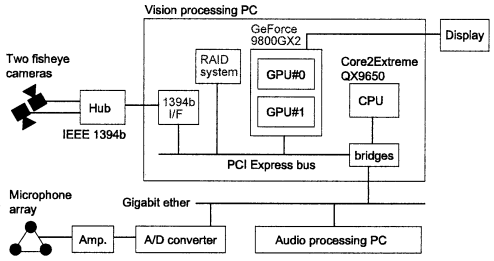

図 6: Hardware configuration

眼レンズ FUJINON FE185C086HA-1(f=2.7mm) により構成される．音響処理用 PC には，AMD Athlon 64 2.4 GHz，OS Linux を使用した．カメラと PC は IEEE1394b により，両 PC 間はギガビットイーサネットワークにより接続される．プログラミング言語は，基本的な画像処理には Microsoft Visual C++ 8，GPU 計算には NVIDIA CUDA 1.1，音声区間検出には C 言語，音声到来方向推定には MATLAB6.5 をそれぞれ使用した．

各カメラで得られる最大画像サイズは，2448 × 2048 画素であるが，本システムでは，縦のサイズを抑えた部分画像 (2448 × 512 画素) を取得する．この部分画像の切り出し位置は，各人物の顔画像が収まるように事前に設定される．この部分画像の切り出しにより，画像データの伝送帯域が減り，30.0fps(=frame/sec) の転送速度が達成される．画像は 8 ビットのグレースケールである．2 つのカメラは時間同期される．魚眼画像からのパノラマ展開，及び，顔方向追跡は GPU 上にて実行される．また，実時間分析と併せて，パノラマ画像は 30.0fps にてハードディスクへ記録される．音響処理のサンプリングレートは 16 kHz，STFT の窓幅は 64ms，フレームシフトは 32ms とした．各音響フレームについて，発話活動が検出される．

5 実験

提案したシステムの性能を検証するため実験を行った．ここでは参加者 5 名による円卓ミーティングを対象とした．図 2 にミーティングの環境と参加者を示す．カメラと人物の顔の間の距離は約 1.5m であった．図 2 中のディスプレイには，実時間での処理結果が表示される．システムの実時間性を検証するため，このディスプレイと各参加者の双方を捉える位置にシステムとは別のカメラを設置し撮影を行った．図 2 と図 7(及び，図 8) は，それらビデオ映像より切り出したものである[1]．図 7 では，2 台のカメラから得られたパノラマ画像上に顔の位置と方向がメッシュにより表現されている．また，声の到来方向が水平座標軸上の円として表され，会話処理の結果として話者が赤い枠で記される．顔方向追跡には，一人物当たり 1500 個のパーティクルを使用した．この 5 人会話において約 20fps にて動作することが確認された．システムの遅延時間は，画像処理に関しては約 170ms，音響処理に関しては平均約 80ms であった．なお，この実験では GPU に GeForce9800GX2 を使用したが，GeForce GTX280 を使用した場合，参加者数 3 人，5 人，8 人に対して，それぞれ平均 29.8fps，27.1fps，19.9fps にて動作することを確認している．

[1]デモムービーは著者のホームページ [29] にて公開中

5.1 可視化

本システムでは，会話シーン並びに分析の結果を外部の観察者に分かりやすく提示するために，新たな会話シーンの可視化法を導入している．図 8(a) は会話シーンを俯瞰するようパノラマ画像を円柱に投影して表示した例である．中央の円が各人物を表す．また，図 8(a) 中の半透明の三角形は各人の大まかな視野範囲を表示したものである．また，発話者には赤い丸が点灯する．もう一つの可視化の例を図 8(b) に示す．こちらは各人物の顔領域を切り出して表示したものであり，図 8(a) の表示法よりも顔のサイズが大きいため，顔表情の把握がより容易であり，かつ，人物間の相対的位置関係を保存した表示であるため，インタラクションの様子が分かりやすい．また，図 8(b) の例では，各人の視線方向が矢印で表示され，また，2 人以上から注目を集めている人物が白い円で印される．

また，本システムでは，3 次元マウス (3Dconnexion 社製 SpaceNavigator) を用いた視点操作インタフェースも提供される．ユーザは，SpaceNavigator のノブの回転により人物の選択を行い，また，ノブの上下動により俯瞰から一人物のズームアップまで自在に視点を変えることができる (図 8(c)，図 8(d))．このようなインタフェースにより，会話場面の状況理解がより容易となることが期待される．

5.2 定量的評価

システムの性能を定量的に検証するため，本稿では予備的に上記の 5 人会話 (約 3 分間) を対象とした評価を行った．まず，話者ダイアリゼーションの評価結果を表 1 に示す．ここでは評価尺度として NIST[30] により提唱されている，Diarization Error Rate (DER)，Missed Speaker Time (MST)，False Alarm Speaker Time (FAT)，Speaker Error Time (SET) を用いた．DER は，

$$\mathrm{DER} = \frac{\text{Wrongly estimated speaker time length}}{\text{Entire speaker time length}} \times 100[\%]$$

のように定義される．表 1 から，正確なダイアリゼーションが行われたことが読み取れる．これらスコアは，過去の実験評価 [20] と比較して，同じ程度か更に値が良い．これは主に今回対象とした会話が比較的フォーマルなミーティングで，発話の重複が少なかったことに起因する．

次に視覚的注意の焦点 (VFOA=Visual Focus of Attention) の精度を評価するために，上述の可視化法とユーザインタフェースに基づくアノテーションツールを開発した．このツールを用いて，視線方向のアノテーションを (画像フレーム単位に) 付与した．この作業は (非会話参加者) 一名

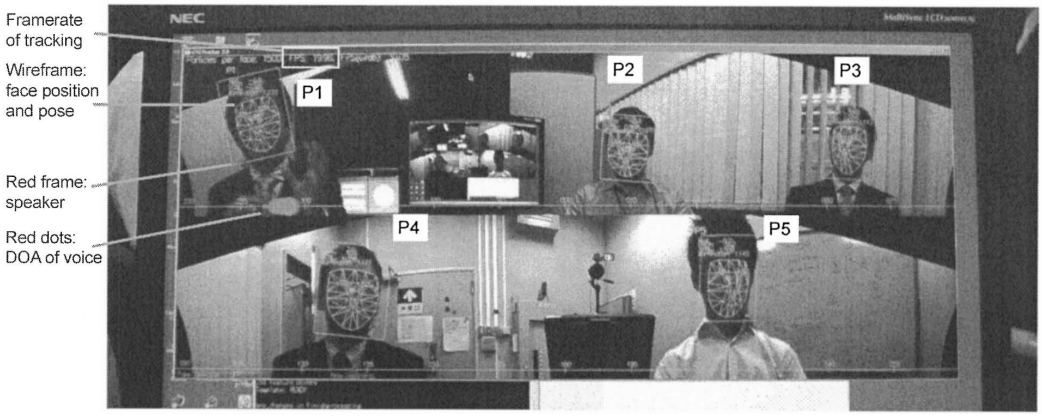

図 7: Screenshot of system monitor displaying face tracking and VAD results

Table 1: Evaluation results of speaker diarization[%]

DER	MST	FAT	SET
4.0	0.9	3.0	0.1

Table 2: Average accuracy of gaze directions[%]

All	P1	P2	P3	P4	P5
55.9	69.5	55.9	20.6	57.6	74.6

により行われた．表 2 には推定された視線方向と人手で付与された視線方向が一致していた割合を示す．視線方向の推定精度は，人物 P3 を除いて，過去の研究報告 [18, 16] と比較しても，妥当な水準にあると言える．誤りの主な原因は，頭部方向と視線方向の本質的な差，つまり，人は頭を動かさずに眼球のみ動かすことができることに由来する．全ての参加者 P1∼P5 に対して，視線方向の誤りの約半数は，視線回避 (誰も見ていない) に関連するものであった．また，残りの誤りの内，約 45%は，推定された視線の先の隣の人物を見ていたケースに相当する．また，人物 P3 の精度がとりわけ低い原因は，この人物の視線が終始落ち着き無く彷徨っていた事による．

6 議論と結び

本稿では，実時間で稼働する会話シーン分析のためのマルチモーダルシステムの提案を行った．このシステムは，画像上での顔方向追跡と音響信号に基づく話者ダイアリゼーションとを統合したものである．円卓ミーティングの状況を観測するために，魚眼レンズ 2 台とマイクアレーを組み合わせた全方位マルチモーダルセンサを開発した．画像上での高速な顔方向追跡のため，GPU 上で動作するパーティクルフィルタを実装した．これにより会話中の各参加者の顔の位置と方向を実時間で推定することが可能となった．また，3 本のマイクから構成されるマイクアレーを用い，音声検出と音声到来方向推定を組み合わせたロバストな話者ダイアリゼーションを実装した．画像と音声の各処理にそれぞれ 1 台の PC を用い，5 人会話に対して平均 27.1fps にて動作することを確認した．

今後の課題は以下の通りである．まず，顔方向追跡をより頑健なものに改良する必要がある．特に追跡の失敗時の再初期化の性能向上，及び，追跡可能な頭部方向の範囲を拡大が望まれる．また，音響信号処理の課題としては，反響に対する頑健性の向上や発話開始時の検出性能の向上などがあげられる．さらに，異なる条件（参加者，人数，座席配置など）にてより多くの実験・評価を行う必要もある．加えて，可視化法の主観的な評価を行うことも重要な課題である．

本稿で提案したシステムは，実時間マルチモーダル会話シーン分析の先駆けとして当分野の発展に寄与するものと考える．現時点では，初歩的な会話シーン分析に留まっているが，今後，様々な発展の可能性を有する．特に，実時間での分析が必須とされる遠隔会議システムや会議支援システムなどへの適用に期待が高まる．

References

[1] 外村, 前田 (編)：“環境知能のすすめ”, リミックスポイント (2008).

[2] M. Argyle: “Bodily Communication – 2nd ed.”, Routledge, London and New York (1988).

[3] D. Gatica-Perez: “Analyzing group interactions in conversations: a review”, Proc. IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems ’06, pp. 41–46 (2006).

[4] R. Stiefelhagen, J. Yang and A. Waibel: “Modeling focus of attention for meeting index based on multiple cues”, IEEE Trans. Neural Networks, **13**, 4 (2002).

[5] F. Wallhoff, M. Zobl, G. Rigoll and I. Potucek: “Face tracking in meeting room scenarios using omnidirectional views”, Proc. ICPR2004 (2004).

[6] D. Douxchamps and N. Campbell: “Robust real time face tracking for the analysis of human behaviour”, Proc. MLMI2007, pp. 1–10 (2007).

[7] 横江, 伊藤, 馬場口：“参加者のインタラクションを可視化したマルチメディア議事録の作成”, 情処研報 CVIM-162-20, pp. 121–126 (2008).

[8] C. Busso, et al: “Smartroom: Participant and speaker localization and identification”, Proc. ICASSP’05, pp. II–1117–1120 (2005).

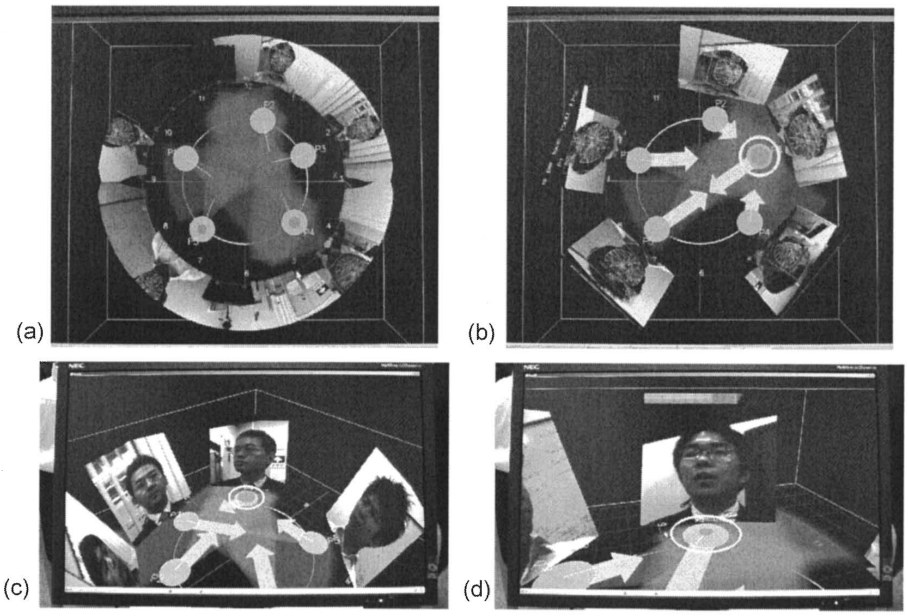

図 8: Visualization schemes, (a)Cylindrical visualization, (b)Piecewise planar visualization, (c)Viewpoint maneuver to middle range, (d)Viewpoint maneuver to close-up range

[9] R. Cutler, et al.: "Distributed meetings: A meeting capture and broadcasting system", ACM Multimedia'02, pp. 503–512 (2002).

[10] 松坂, 緒方, 麻生, 浅野："多人数インタラクションの工学的応用–認識・理解システムの構築とその利用について–", 信学技報 HCS-106(219), pp. 13–18 (2006).

[11] S. Renals, T. Hain and H. Bourlard: "Interpretation of multiparty meetings the AMI and AMIDA projects", Proc. HSCMA2008, pp. 115–118 (2008).

[12] Microsoft. http://www.microsoft.com/UC/products/roundtable.mspx.

[13] M. Voit and R. Stiefelhagen: "Tracking head pose and focus of attention with multiple far-field cameras", Proc. ICMI2006, pp. 281–286 (2006).

[14] D. Gatica-Perez, J.-M. Odobez, S. Ba, K. Smith and G. Lathoud: "Tracking people in meetings with particles", Technical Report IDIAP-RR 04-71 (2004).

[15] L. Chen, et al.: "VACE multimodal meeting corpus", Proc. MLMI2006, pp. 40–51 (2006).

[16] K. Otsuka, J. Yamato and H. Murase: "Conversation scene analysis with dynamic Bayesian network based on visual head tracking", Proc. ICME'06, pp. 949–952 (2006).

[17] A. Kendon: "Some functions of gaze-direction in social interaction", Acta Psychologica, **26**, pp. 22–63 (1967).

[18] 大塚, 竹前, 大和, 村瀬："複数人物の対面会話を対象としたマルコフ切替えモデルに基づく会話構造の確率的推論", 情報処理学会論文誌, **47**, 7, pp. 2317–2334 (2006).

[19] S. O. Ba and J.-M. Odobez: "A study on visual focus of attention recognition from head pose in a meeting room", Proc. MLMI2006, pp. 75–87 (2006).

[20] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada and S. Makino: "A DOA based speaker diarization system for real meetings", Proc. HSCMA2008, pp. 29–32 (2008).

[21] O. Mateo Lozano and K. Otsuka: "Real-time visual tracker by stream processing", Journal of Signal Processing Systems, **DOI 10.1007/s11265-008-0250-2**, (2008).

[22] 松原, 尺長："疎テンプレートマッチングとその実時間物体追跡への応用", 情報処理学会論文誌 CVIM, **46**, SIG9(CVIM11), pp. 60–71 (2005).

[23] K. Otsuka and J. Yamato: "Fast and robust face tracking for analyzing multiparty face-to-face meetings", Proc. MLMI2008, LNCS 5237, pp. 14–25 (2008).

[24] P. Viola and M. Jones: "Robust real-time face detection", IJCV, **57**, 2, pp. 137–154 (2004).

[25] C. H. Knapp and G. C. Carter: "The generalized correlation method for estimation of time delay", IEEE Trans. ASSP, **24**, 4, pp. 320–327 (1976).

[26] X. Anguera, C. Wooters and J. Hernando: "Acoustic beamforming for speaker diarization of meetings", IEEE Trans. Audio, Speech and Language Processing, **15**, pp. 2011–2022 (2007).

[27] D. Macho, et al.: "Automatic speech activity detection, source localization, and speech recognition on the chil seminar corpus", ICME'05, pp. 876–879 (2005).

[28] M. Fujimoto, K. Ishizuka and T. Nakatani: "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme", Proc. ICASSP2008, pp. 4441–4444 (2008).

[29] http://www.brl.ntt.co.jp/people/otsuka/realtime_systemJ.html

[30] NIST Speech Group: "Spring 2007 (RT-07) rich transcription meeting recognition evaluation plan", Technical Report rt07-meeting-eval-plan-v2, NIST (2007).