会話シーン分析のための音声映像技術

複数の人物が対面で会話を交わす場面を対象として，その会話の状況，例えば，「いつ誰が誰に向かって話しているか？」「誰が注目を集めているか？」といったことを自動的に分析するマルチモーダル会話シーン分析システム，および本システムで使われている画像技術，音声技術を紹介します．

大塚 和弘（おおつか かずひろ） /荒木 章子（あらき しょうこ）

NTTコミュニケーション科学基礎研究所

コミュニケーションを科学する

私たちは，日常会話や会議，電話など，日々いろいろなコミュニケーションを行っています．数あるコミュニケーションの中でも人と人とが対面で交わす「会話」はもっとも基本的なものであり，日々，情報の伝達・共有や他者の意図・感情の理解，グループでの意志決定などが会話を通じて行われています．これまで遠隔地間における人と人とのコミュニケーションを支援する技術としてさまざまな情報通信技術が研究開発され，広く用いられていますが，いまだ多くの場面で対面の会話ほどスムーズなコミュニケーションを実現するには至っていません．

NTTコミュニケーション科学基礎研究所では，時間や距離に伴うコミュニケーションの障壁を解消するためには，そもそもその人と人とがどのようにメッセージや想いを伝達するのかというコミュニケーションのプロセスを明らかにすることが重要であり，それが未来の情報通信技術の創造へとつながっていくものと考えています．我々はコミュニケーションという現象を科学するため，コミュニケーションの場面をセンサや計算機を用いて，定量的かつ客観的に計測・分析できるような技術の実現を目指して研究開発を進めています．

非言語コミュニケーション

日常のコミュニケーションにおいて，我々人間はさりげなく自分の考えを口にして，また，相手の話に耳を傾けています．こうしたコミュニケーションの行動は人間からみると何気ないものにみえますが，これを計算機で自動的に計測・分析することは容易ではありません．その難しさの一因として人間の間でやり取りされる情報の多様性があります．例えば，対面状況の会話において，人は言語的な情報だけでなく，さまざまな非言語情報の交換を行っています．このような非言語情報には，視線や顔の表情，ジェスチャ，身振り・手振り，声のトーンなどが含まれており，これらが総合的に作用することでさまざまな情報が伝達・交換されています．これら情報は身振りや手振りであれば目を通して視覚的に知覚され，また，声のトーンであれば音の情報が耳を通して知覚されます．我々は，人の目や耳の代わりにカメラやマイクなどを用いることで，人の行動を画像・音声情報として観測し，その情報を計算機を用いて分析することで，人のコミュニケーションの過程を明らかにすることができるのではと期待しています．このような視点に立って，我々，NTTコミュニケーション科学基礎研究所では，画像や音声などメディア技術の研究開発を進めており，それら技術を統合することでコミュニケーションシーンを理解できる機械の実現を目指しています．

会話シーン分析

我々は，画像や音声などの情報に基づいて自動的に会話の状況を認識・理解するという技術分野を「会話シーン分析」と名付けました．この会話シーン分析の目標は，入力情報から会話の6W1Hの記述を自動的に得ることと考えています．6W1Hとは，When（いつ），Where（どこで），Who（誰が），Whom（誰に対して），What（どのようなメッセージを），How（どのような振る舞いによって），Why（なぜ）発したのか？　また，それがどう受け止められたのか？　ということを指します．これらの要素の組合せによって，行動レベルから文脈・心的なレベルまでさまざまなレベルの問題が定義できます．

例えば，WhoとWhenを組み合わせ

ることで，「誰がいつ喋ったのか？」という話者検出の問題が定義できます．また，それにWhomを加えることで，「誰が誰に話し掛けたのか」という話者＋受け手の推定の問題になります．また，Howは表情やジェスチャなどメッセージの伝え方に関する問いです．さらにWhyの問いに対しては会話の文脈や感情など内面にかかわる高次な情報の推論が必要です．

会話シーン分析にはこのように非常に幅広い問題が含まれますが，我々はまず行動のレベルの問題として，メッセージの伝達の方向性，つまり，「誰が誰に向かって話しているか」を基本問題としてとらえて検討を始めており，徐々にジェスチャや顔の表情など感情や態度などにも検討範囲を広げているところです．本稿では，我々の研究成果の一例として，「誰が誰に向かって話しているか？」「誰が誰に注目しているか？」という行動レベルの会話の状態を実時間で推定するシステムについて紹介します．

マルチモーダル会話シーン分析システム

現在，NTTコミュニケーション科学基礎研究所では，対面の状況で人と人が会話する場面を自動的かつ実時間で分析できるシステムとして「マルチモーダル会話シーン分析システム」の開発を進めています(1)．このシステムは，8人程度までの比較的小規模のミーティングを想定して，テーブル上に置かれたコンパクトな全方位カメラ・マイク統合システムにより画像情報，音響情報を収集し，その情報を処理・統合することで会話の状況を分析し，その結果をディスプレイ上に表示するシステムです．具体的には，「誰がいつ話しているか？」「誰が誰の方を向いているか？」「誰が誰に注目しているか？」といった会話の状態をリアルタイムにモニタリングすることができます．本システムのようにマルチモーダル情報を統合し，リアルタイムに会話の状態を分析するシステムは世界でも初めてです．

■システムの構成

このシステムの構成図を**図1**に示します．このシステムは①画像処理部，②音響処理部，③会話処理部から構成されます．①画像処理部では，カメラから得られる画像上から人物の顔の位置と方向を推定します．②音響処理部では，マイクから得られる音響信号から人の声が発せられた時刻，およびその声が来た到来角度を推定します．③会話処理部では，画像処理部と音響処理部の結果を統合して会話の状態を推定し，結果の表示を行います．

■全方位マルチモーダルセンサ

図2は，このデモシステムの動作の様子を示す図です．ここでは5名の人物が円卓の周囲に座り会話をしています．その場面に対するシステムの処理結果が図中手前のディスプレイにリアルタイムに表示されます．図2の円卓中央部に，全方位カメラ・マイクシステムが配置されています．このカメラ・マイクシステムは，2つのカメラと3つのマイクから構成されています（**図3**）．各カメラには魚眼レンズが装着されており，各々およそ半球の領域が撮影できます．このカメラを背中合わせに配置することで，およそ全周の領

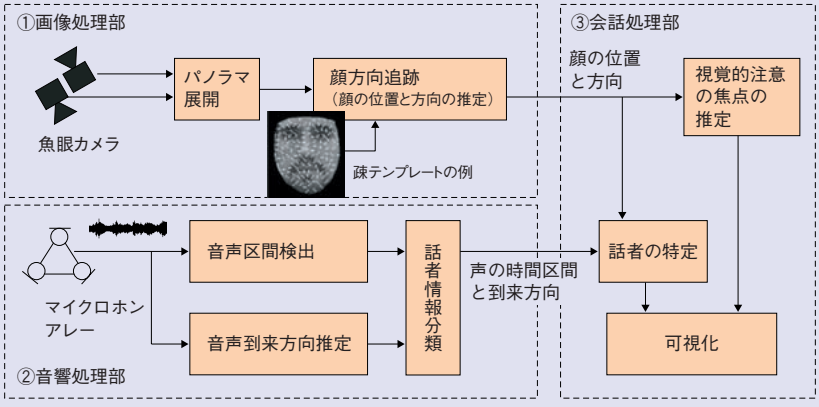


図1　会話シーン分析システムの構成図

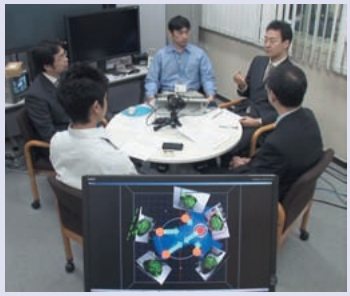


図2　会話の様子

図3　全方位カメラマイクロホンシステム

域が撮影できます．カメラの上部には３つのマイクが三角形の頂点状に配置されており，マイクアレーを構成しています．

■画像処理部

全方位カメラにより撮影された画像から魚眼レンズ特有の歪みを取り除いたパノラマ画像へと変換を行います．そのパノラマ画像から各人物の顔の位置と向きの推定が行われます．本システムではその方法として，「疎テンプレートコンデンセーション追跡法」と呼ばれる方法を用いています．この方法は，画像上での人の顔の明るさの情報を疎テンプレートと呼ばれる空間的に疎な点の集合として保持しておき，新しく入力された画像上でそのテンプレートがどの位置，どの角度，どの大きさのときにもっともよく当てはまるか照合・探索し，そのときのテンプレートの位置や角度，大きさ（状態）によって人物の顔の位置や方向を求めるという方法です．この探索は時刻ごと，つまり，ある時刻の画像で求めたテンプレートの状態が次の時刻でどの辺りに移動しているかという予想を行い，１人ひとりの顔を追跡することができます．

また，ここでは，コンデンセーション（パーティクルフィルタとも呼ばれる）法を用いています．この方法は，テンプレートの状態の候補を多数生成し（数百個～数千個），それら１つひとつについて，テンプレートの当てはまりの度合いを計算する方法です．このコンデンセーション法は，評価する仮説数が膨大であるために計算時間が掛かるという問題がありましたが，我々は新たにGPU（Graphics Processing Unit）と呼ばれる並列ハードウェア上で実行する並列アルゴリズムを考案しました．これにより従来，CPUのみを用いた方法と比較して約10倍の速度向上を達成し，複数の人物（８人程度まで）の位置と顔の向きを実時間で推定できるようになりました．

■音響処理部

音響処理部では，全方位センサのマイクロホンアレーにより得られる音響信号を入力として，そこから人の声のした時刻とその方向の推定を行います．本システムでは，大きく分けて音声区間検出，音声到来方向推定，話者情報分類から構成されます．図１にその構成を示します．

音声区間検出部では，人の声と雑音とを区別して，人の声がした時間区間のみを検出します．我々のグループでは，この発話の検出の原理として，人の声に含まれる周波数成分，および周期的な成分と非周期的な成分の比が，雑音とは異なり独特であるという性質を用いた独自の方法を開発しました．雑音の性質を時々刻々推定するので，雑音の時間的変化に追従することが可能です．

また，音声到来方向推定部では，マイクから見て人の声が到来した方角を推定します．その原理は，３本あるマイクに音が到着する時間差から方角を推定するというものです．この原理を使った方法は従来から知られているのですが，これまでは複数の人が同時に話す場合や人の声以外の雑音がある場合などに正しく推定することができませんでした．この問題に対して，我々のグループでは，先ほどの音声区間検出の結果を用いることによって，雑音の影響を受けず人の声だけの方向を推定することに成功しました．また，マイク間で音の時間差を計る際に，人ごとの声の性質の違い（声に含まれる音の高さの分布）を考慮することで，同時に複数の人が話す場合にもそれぞれの人の声の到来方向を正確に推定できるようになりました．最後に話者情報分類部において，音声発話検出と音声到来方向の結果に対してクラスタリング処理を行うことで，その場にいる人数も自動的に検出することができます．

■会話処理部

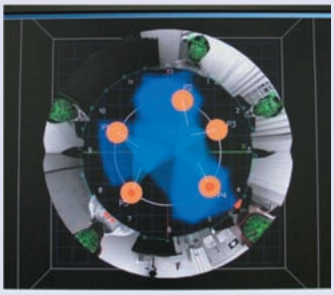
会話処理部では，画像処理と音響処理の結果から会話の状態を推定します．現段階では，話者の特定と視覚的注意の方向（視線の方向＝誰を見ているか？）の推定を行っています．前者は，画像上の各人の顔の位置と，音声の到来方向とを照らし合わせて，センサに対して同じ方角の顔と声を対応付けすることで行います．後者は，全員が円卓の周囲に着席しているという仮定のもと，画像より得られる顔の方位角と方向より，おおよその視覚的注意の方向を推定しています．

■結果の可視化

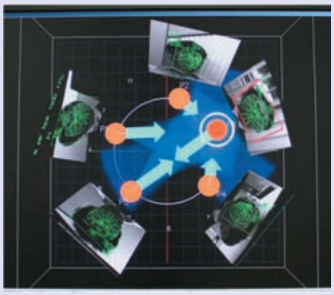
図４，５は本システムの処理結果の一例です．図４は，２つのカメラに対応するパノラマ画像をそれぞれ上下に表示しており，顔追跡の結果が緑色のメッシュで表されています．さらに，横軸上の赤い丸が音声到来方向を，顔の周囲の赤い枠が発話状態を表しています．

図５(a)，(b)は，会話シーン分析の結果の表示画面例です．ここでは会話の場面を上部から俯瞰するように表示しています．また，画面中央に会話シーン分析の結果が図示されています．各人物の位置はオレンジ色の丸として，また，発話をしている人物は赤丸として表示されます．また，図中，青い三角錐は，顔方向から示唆される各人物の視野範囲を表し，矢印は対人視線方向の推定結果を表しています．さらに，２人以上から注目を集めている人

図 4　顔の位置・方向の推定の結果，および音声区間検出の結果



(a)　円柱表示



(b)　切出し表示

(c)　視点切替

(d)　ズームアップ

図 5　会話シーンの可視化の一例

物には白い丸が付加されています．

さらに図 5 (c)，(d)のように，3 次元マウスによりユーザの所望の人物をクローズアップして表示することも自在にできます．

今後の応用・発展性

今回紹介しましたマルチモーダル会話シーン分析システムを今後，さらに発展・拡張していくことで，将来的には，会議のマルチメディア議事録の自動作成や，テレビ会議システムの自動カメラワーク，ひいては人間どうしの会話に参加できるロボットやエージェントなど，より円滑なコミュニケーションを可能とする情報通信技術へつながっていくものと期待しています．今後は，画像や音声各々の要素技術の性能向上を目指すとともに，うなずきや顔の表情，声のトーンなどより豊かな人間の非言語・言語行動をとらえられる技術の研究開発も進めていく予定です．

■参考文献

(1)　大塚・荒木・石塚・藤本・大和：“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～，”信学技報MVE2008-68，2008.

(左から) 大塚　和弘/ 荒木　章子

今回紹介した技術・システムは我々のグループで研究していることのほんの一部です．今後，人と人とのコミュニケーションをより深く理解できるような技術の研究を進めていきます．

◆問い合わせ先

NTTコミュニケーション科学基礎研究所
メディア情報研究部
メディア認識研究グループ
TEL　046-240-3639
FAX　046-240-4707
E-mail　otsuka@eye.brl.ntt.co.jp
URL　http://www.kecl.ntt.co.jp/rps/index-j.html