2008年5月30日
13:00~14:30 at 大会議室

NTT CS研 未来想論2008

1F　2A会場で展示中

実世界コミュニケーションシーンを理解する音声映像技術

～会話の流れを分析する　音声技術と映像技術の調和～

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
メディア情報研究部
大 塚 和 弘

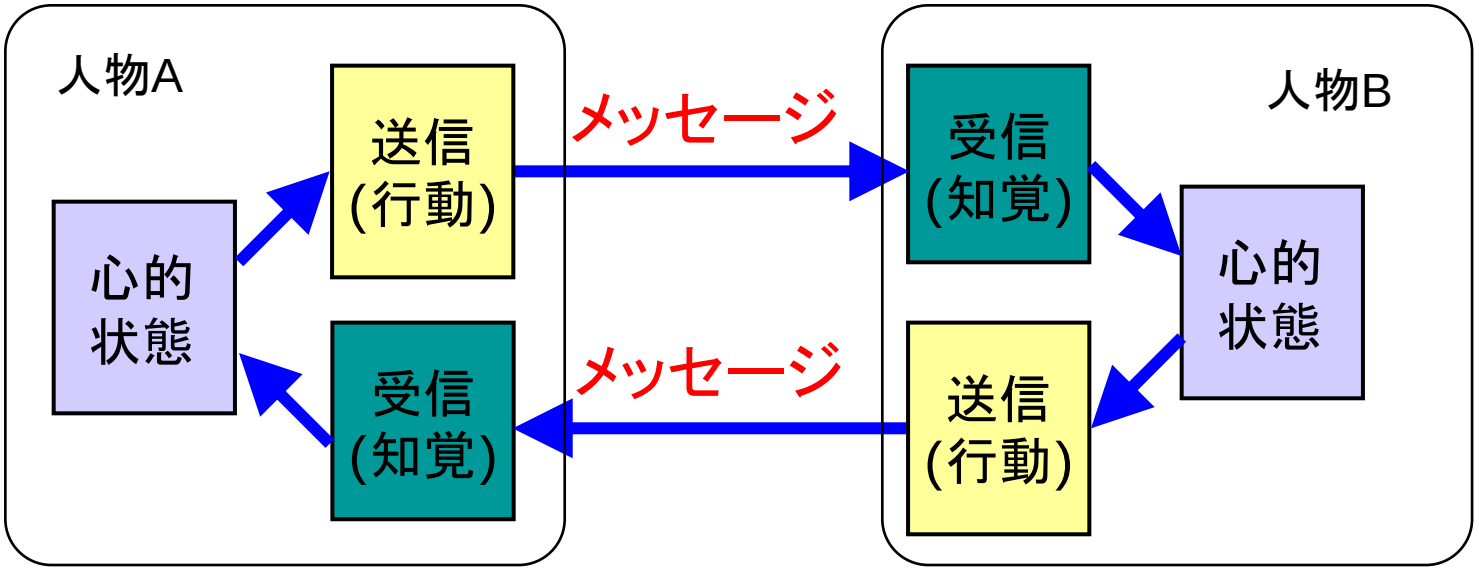
コミュニケーションの情報科学

そもそもコミュニケーションとは？

コミュニケーションの定義

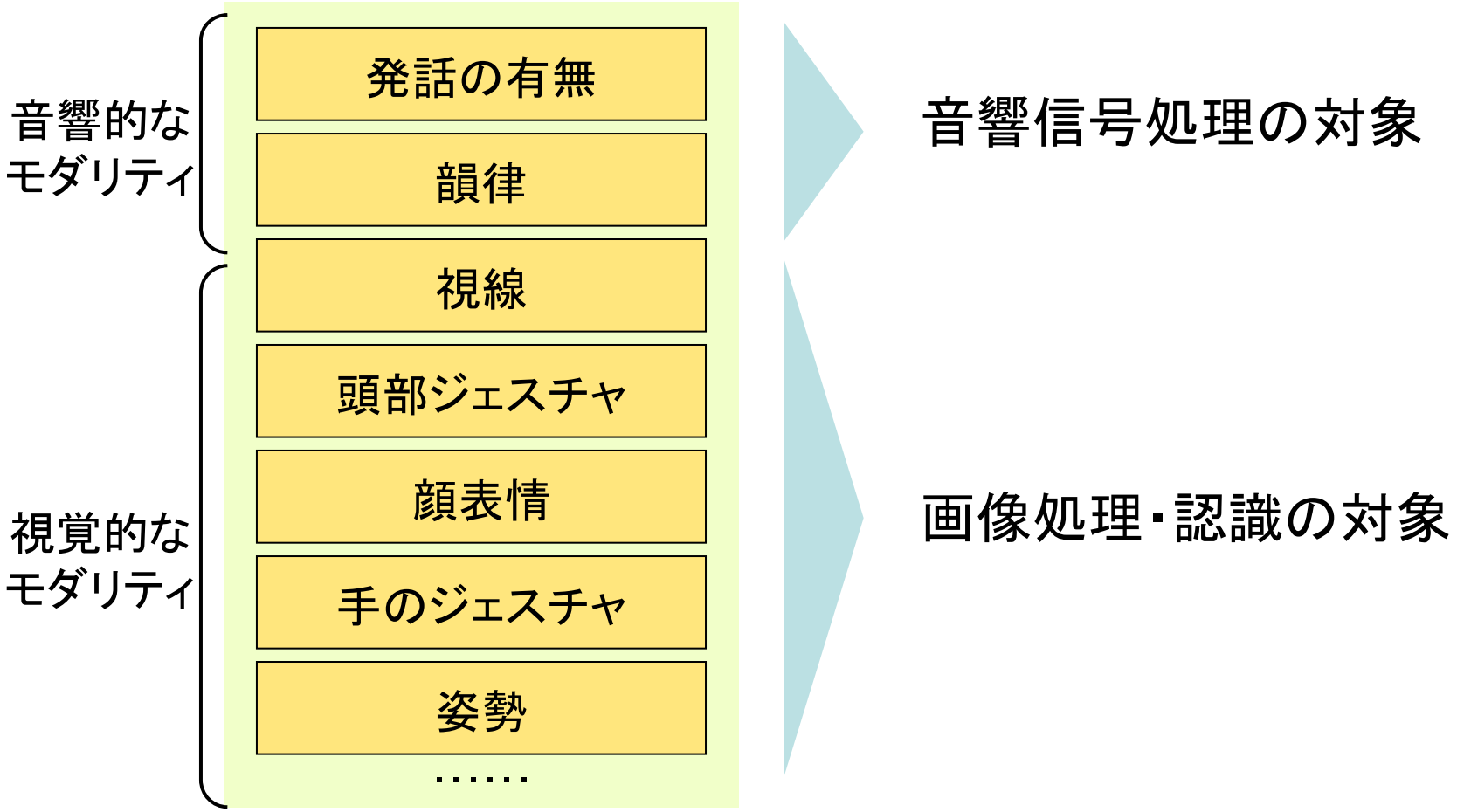
（1）人物間における**メッセージの送信・受信**の過程
（2）メッセージの送受信によって**心的状態に変化**を生じさせる現象

コミュニケーションの過程



非言語行動・非言語メッセージ

対面コミュニケーションにおいて，
非言語行動・非言語メッセージは重要な役割を持つ



コミュニケーション・シーンの理解

コミュニケーションシーン理解のタスク

:=観測した画像・音響信号等から6W1Hを自動的に明らかにすること

when	where	who	whom	what	how	why
いつ？	どこで？	誰が？	誰と？	何を？	どのように？	何故？

様々なレベルのタスク

行動レベル

例）いま，話しているのは誰か？．．．．話者推定の問題

例）彼は誰と話していているのか？．．．話者＋受け手の推定

例）誰のどの発言が彼を怒らせたのか？

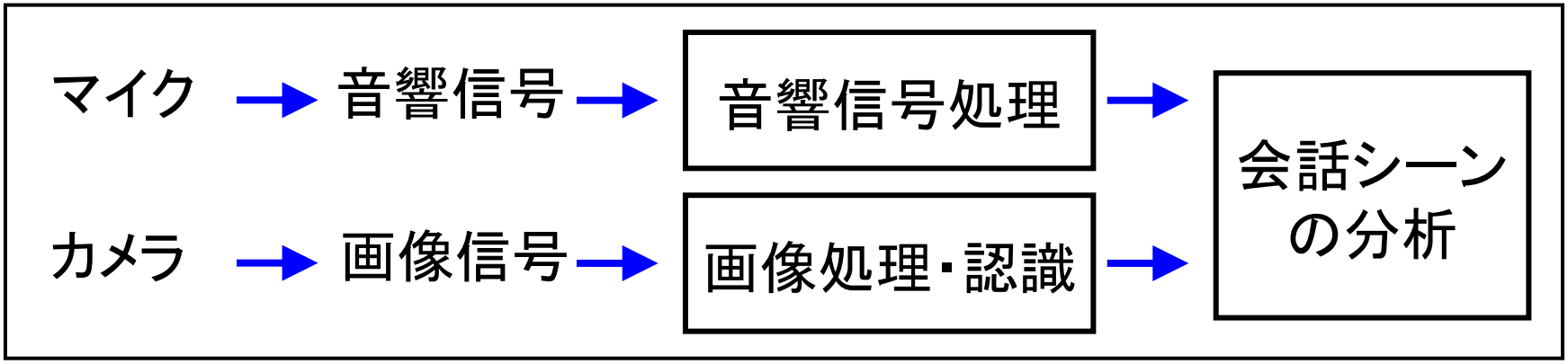
文脈・意味のレベル

例）何故，彼女は泣いているのか？

例）誰の意見がグループの結論に貢献したのか？

今回展示のデモシステムの概要

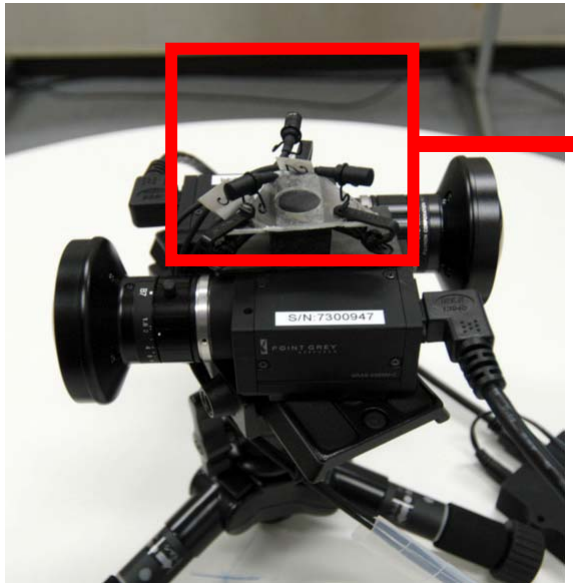
- 1)マルチモーダル・シーン分析
  - 「誰がいつ話しているか」を推定
  - 「誰が誰の方を向いているか」を推定
  - 「誰が注目を集めているか」を推定
- 2)コンパクトなカメラ・マイク統合システム
- 3)リアルタイムに動作する
- 4)いくつか会話シーンの可視化方法も提案

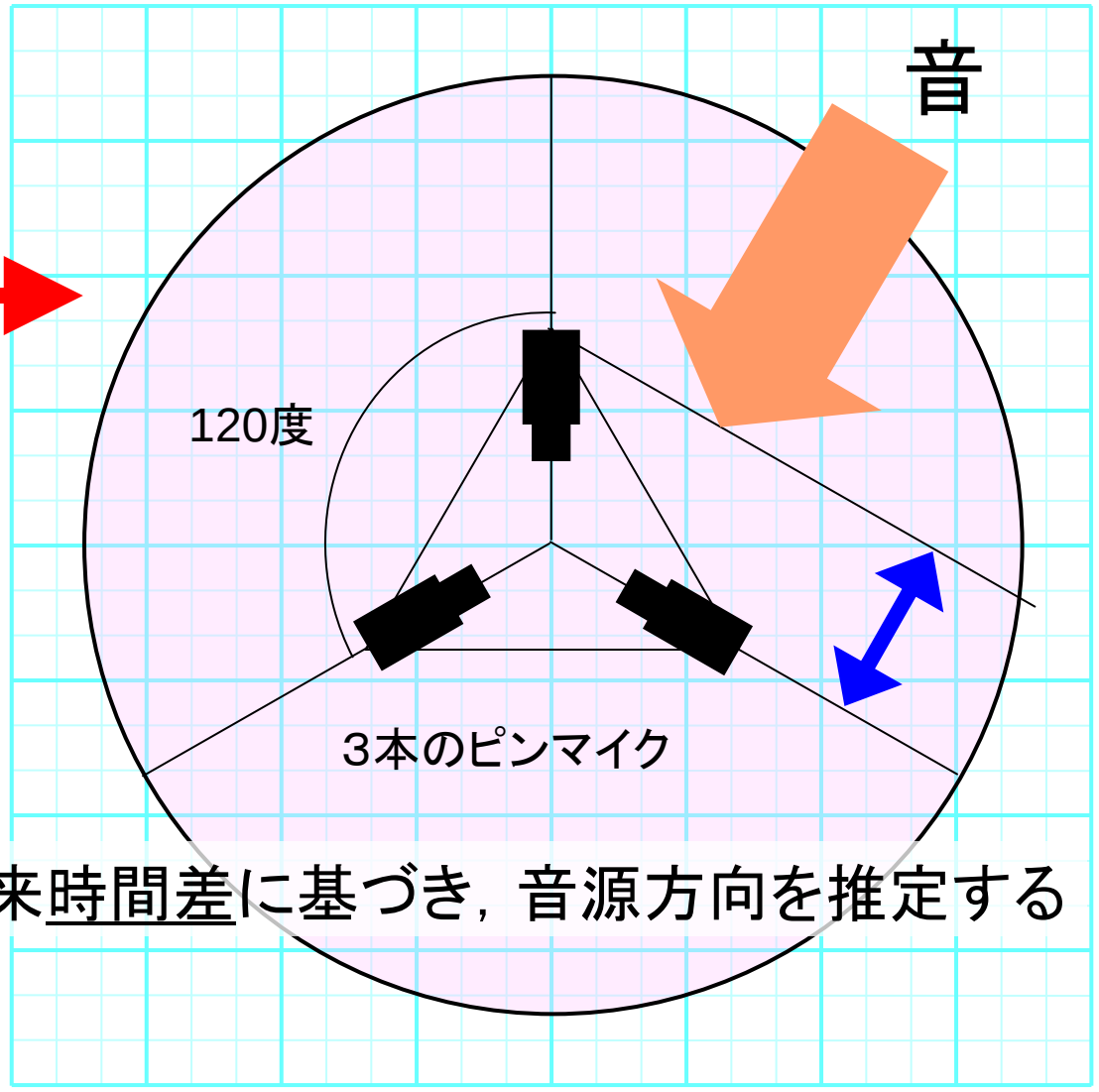


デモシステムの様子

全方位カメラ・マイクシステム

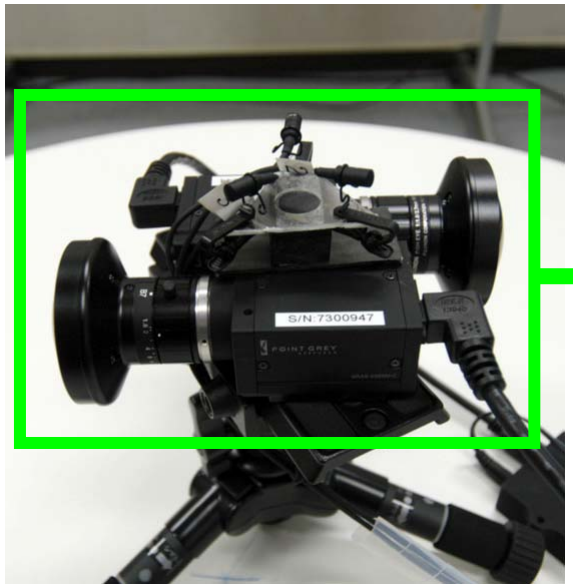
マイクロフォンアレー

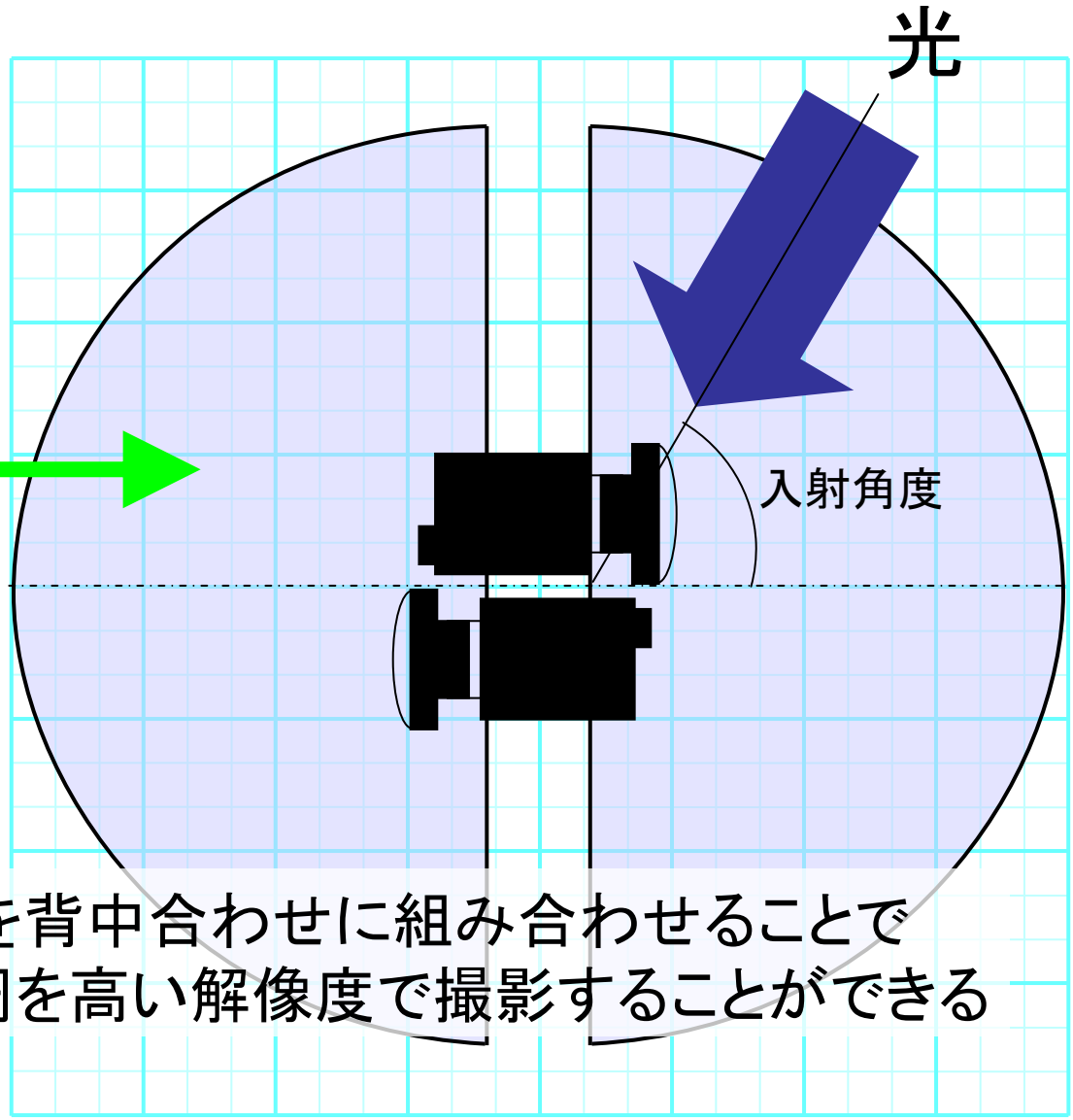




マイク間での音の到来時間差に基づき，音源方向を推定する

全方位カメラ

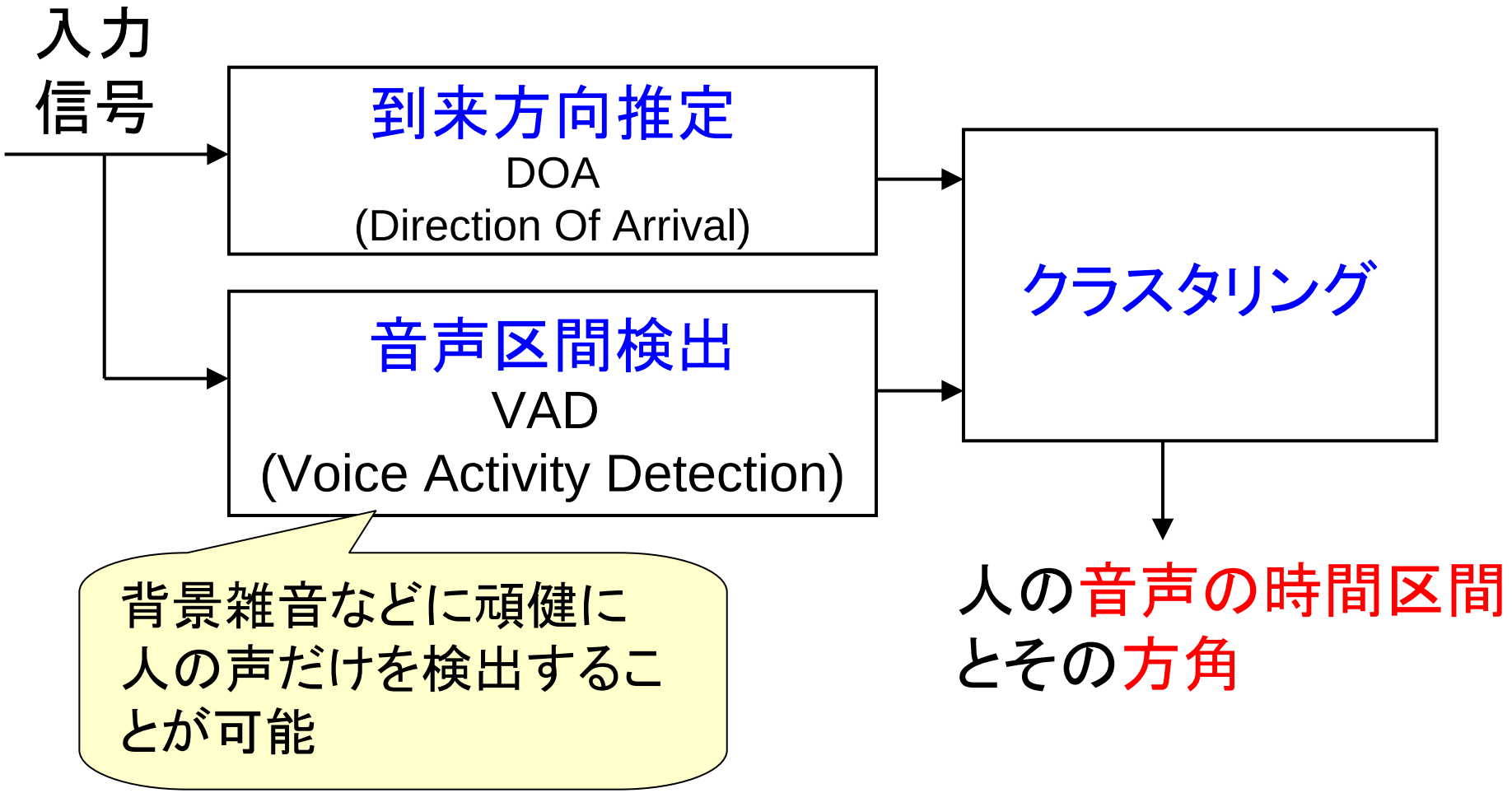




2つの魚眼レンズを背中合わせに組み合わせることで
ほぼ360度の全周を高い解像度で撮影することができる

音声系技術

音声到来方向推定と音声区間検出を組み合わせた技術



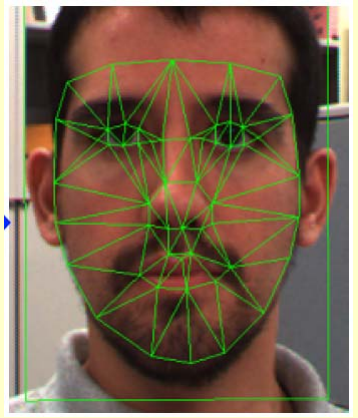
画像系技術

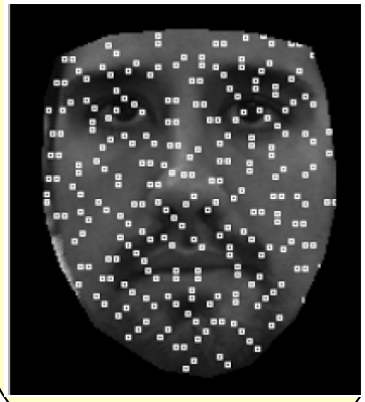
入力
画像

初期化
人ごとの顔のモデルを作成

追跡（トラッキング）
各人の顔の位置と向きを推定

顔の位置と向き

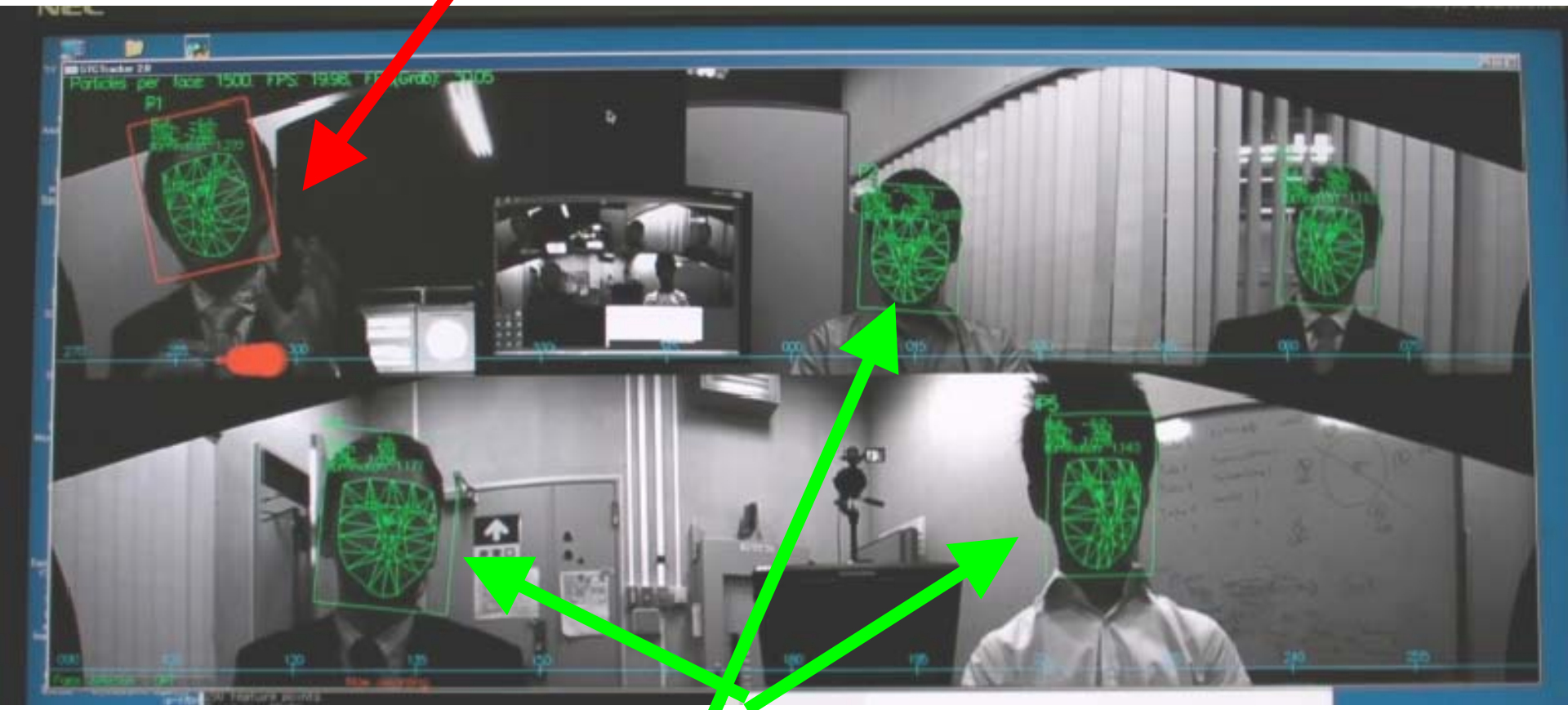




並列アルゴリズム（パーティクルフィルタ）を
並列ハードウェア（GPU=Graphics Processing Unit）
にて実行することで，約10倍の高速化を実現

デモシステムの動作の様子

・各参加者の発話を検出



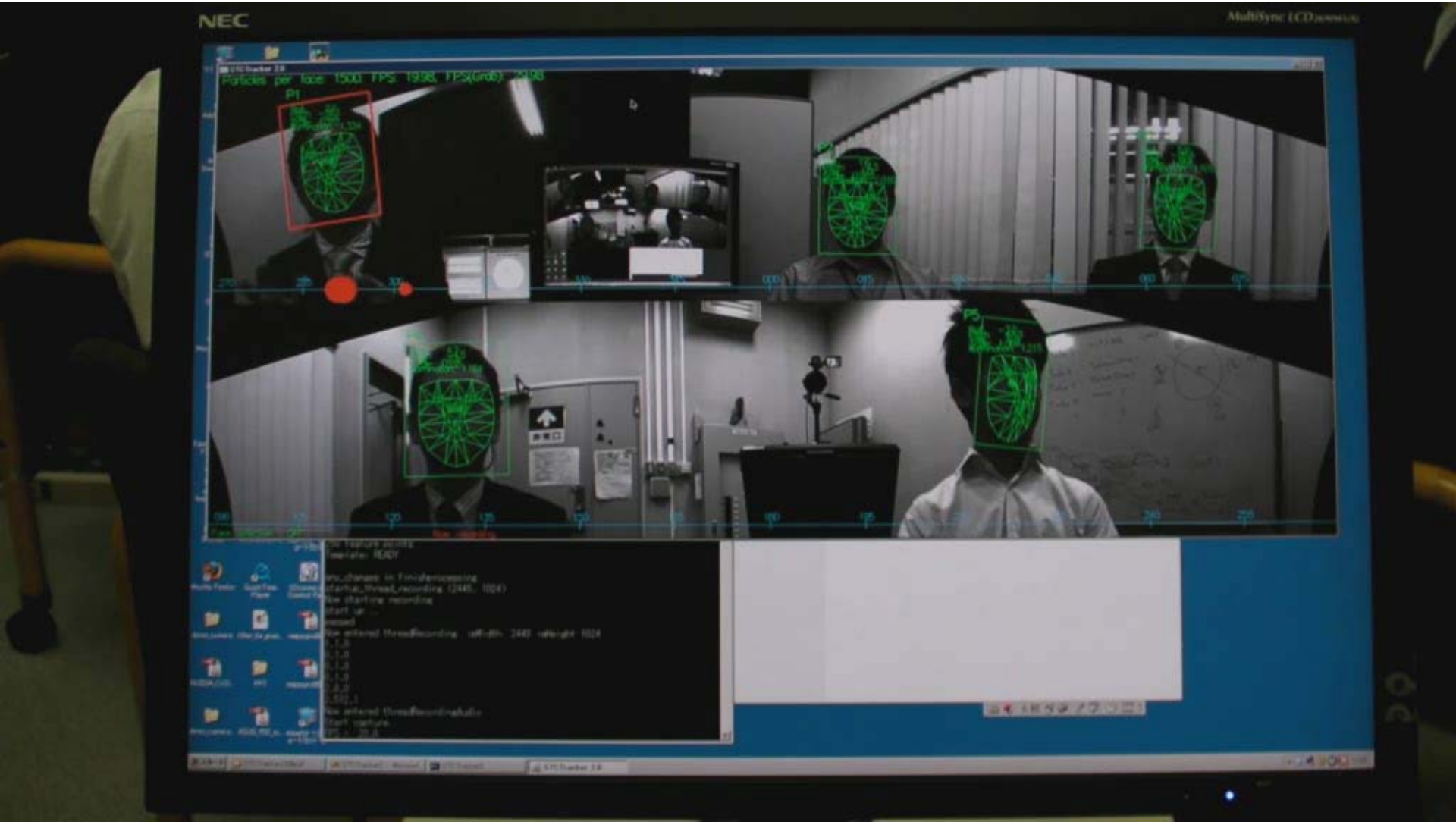
・会話参加者の顔の位置と向きを追跡（トラッキング）

デモシステムの動作の様子

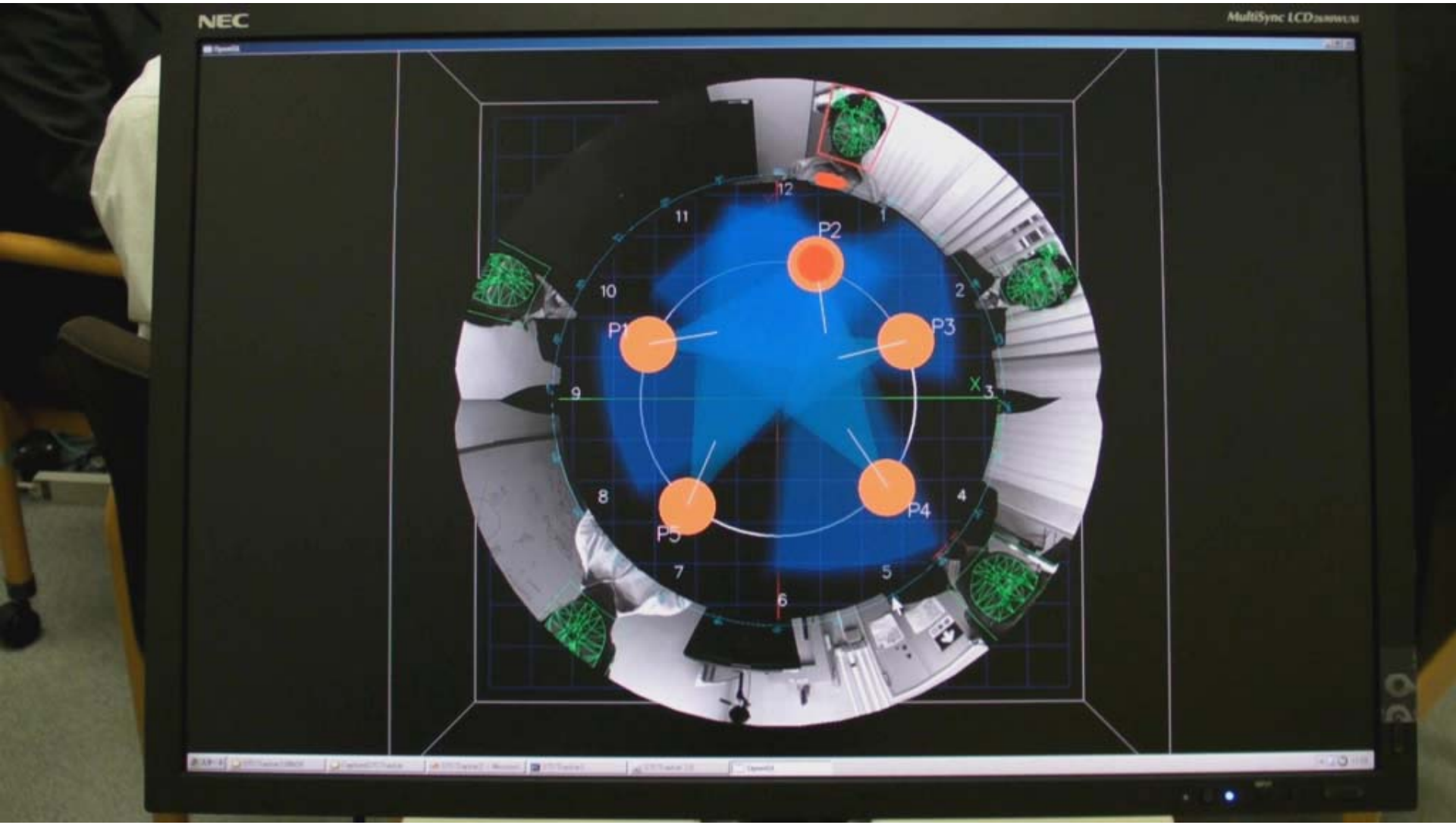
[ムービー 15秒]

発話検出･顔方向追跡の様子

[ムービー 15秒]



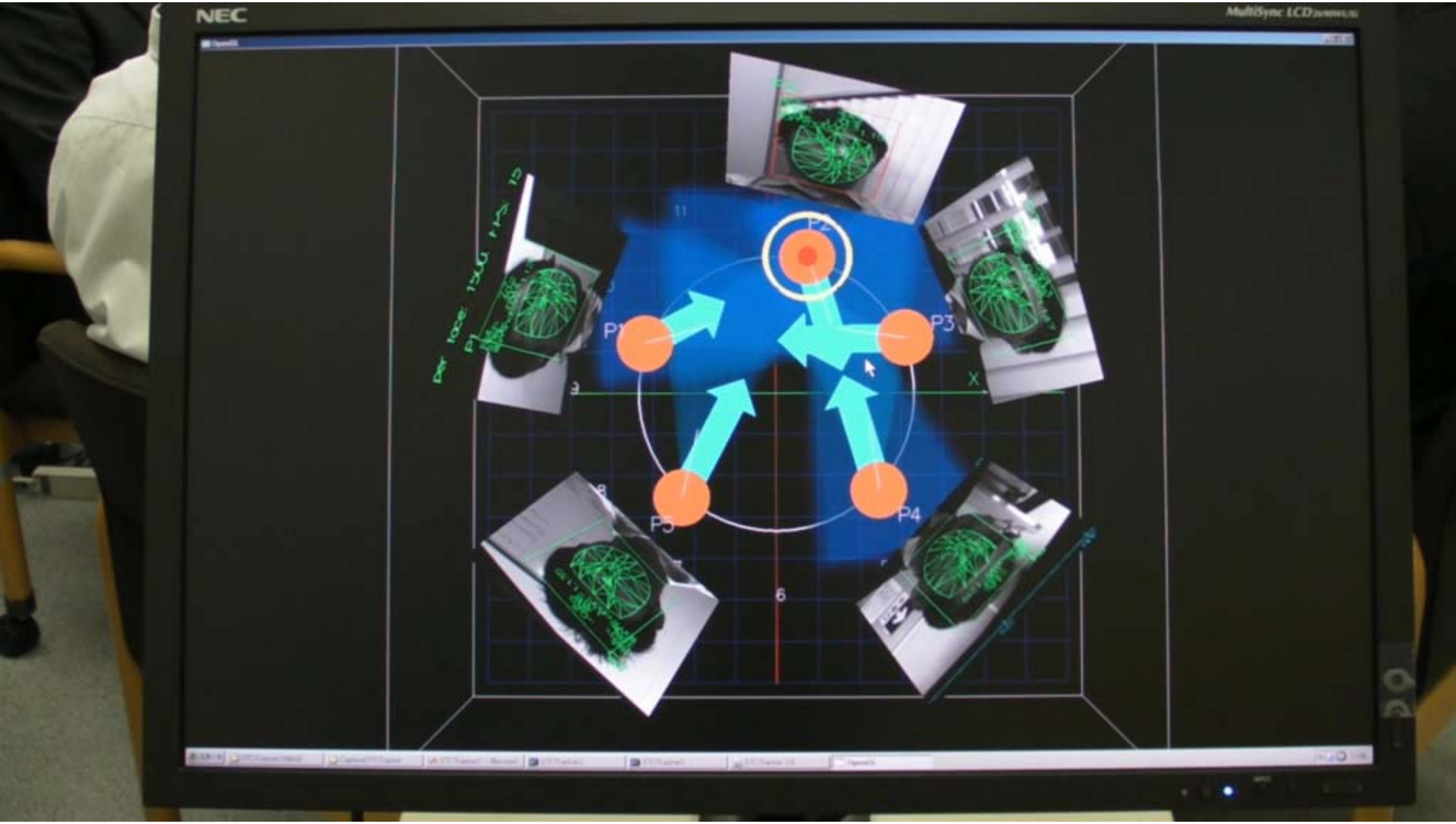
俯瞰的な表現



各人の顔の向いている方向と，（おおまかな）視野範囲を表示．

視線方向推定の例

[ムービー 20秒]



矢印で推定視線方向を示す．円で注目されている人を示す

カメラワークの例

[ムービー 40秒]

会話の模様を自由に視点を変えて見ることが出来る

議　論

- このような技術によって何が可能になるのか？
  - 「人のコミュニケーションをより良く知る」研究への貢献
  - 「人のコミュニケーションをより良くする」アプリケーションへの貢献
- どのような方向に技術開発を進めるのか？
- 複数分野の研究者との連携の必要性
  - 画像，音声，言語，心理，社会，認知科学など