5T-8

音声方向情報を考慮した全方位映像型 TV 会議システム

斉藤純一 † 米田祐也 †† 橋本浩二 † 柴田義孝 †
† 岩手県立大学ソフトウェア情報学部 †† 岩手県立大学ソフトウェア情報学研究科

1 はじめに

近年，全方位カメラの登場により広視野を持つ映像の効率的な取り込みや表示映像の臨場感の向上や映像の効率化などが可能となり，これに加えてコンピュータの高性能化，映像関連技術の向上，利用可能なネットワーク帯域の拡大により，高精細な全方位カメラを用いた TV 会議システムが実現可能となった．

しかしながら，既存のシステム [1][2] は全方位の映像を取り込むために映像中の話者特定が困難であること，使用するマイクは人数に合わせて増やすため，本数に応じた設置コストが生じること，無指向性マイクを用いた場合は話者の位置特定には不向きであるなどの問題が存在していた．

そこで本研究では複数の有指向性マイクを使用することで，話者の位置情報を取得し，全方位映像中の話者の映像抽出による話者特定，及び話者方向の音声送信を可能とし，利用者に分かり易い TV 会議システムを開発し，そのプロトタイプを評価する．本論文では システムとプロトタイプを評価することにより得られた有効性について検証する．

2 システム構成

本研究で開発するシステムの構成は図 1 に示すように，ネットワークを介した複数地点間による全方位 TV 会議である．利用者は遠隔で行われている会議の模様をリアルタイムで受信し，３６０度の全表示および部分拡大表示をすることが可能である．

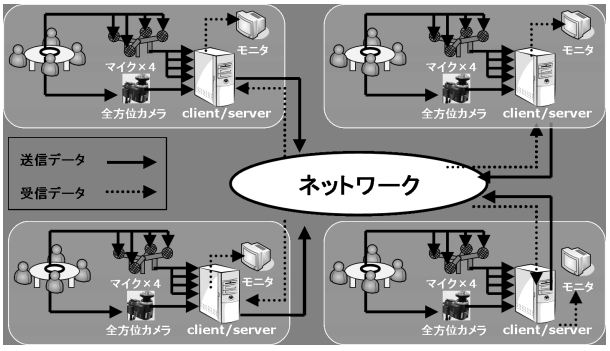

図 1: システム構成図

全ての会議空間ではサーバ/クライアントを兼ねるホストマシンに加えて映像取得用の全方位レンズを装着した DV カメラが一台，全方位音声の取得と話者の決定に必要なデータを取得するために複数の有指向性マイクが設置される．

全方位映像はクライアント側で環状映像からパノラマ映像へ展開処理を行い，本稿で導入する方向決定アルゴリズムを適用することで現在進行中の TV 会議の映像中における注目すべき話者の部分拡大映像を抽出した再生が可能である．加えて，複数指向性マイクから得られた音声を用いることで，必要な話者方向の音声のみを再生することも可能である．

3 システムアーキテクチャ

本システムのアーキテクチャは図 2 に示されるように MidField System[3] の上位層に位置し，２階層２プレーンで構成される．

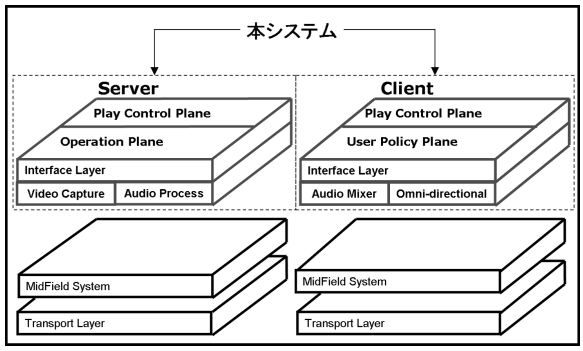

図 2: システムアーキテクチャ

サーバ側の Interface Layer では Operation Plane が利用者の操作情報管理を行い，クライアント側の User Policy Plane では利用者の要求に合わせてミキシングや映像展開の命令を実行する．Play Control Plane では映像音声の再生機能を果たす．Video Capture，Audio Capture ではそれぞれ映像のキャプチャ部分の実行，音声のキャプチャ及び話者の方向の計算を行う．Audio Mixer では音声のミキシングの機能を持ち，Omni-directional では全方位展開のミドルウェア [4] を用いることで全方位カメラで取得した環状映像をパノラマ映像や部分拡大映像，そして補正映像といった 3 種類の映像への変換が可能となり，これを用いることで話者方向の映像の部分拡大が実現可能となる．

また，これらの映像と音声の通信には MidField System を使用する．MidField System とはトランスポート層の上位層として 3 階層，4 プレーンとして構成され，相互通信の通信路にある適切なノードでトランスコーディング機能を動的に稼働させる仕組みにより，通信端末の処理能力や利用可能なネットワーク帯域幅に応じて適切なフォーマットによる通信が可能となる等，アプリケーションに対して，柔軟なマルチメディア通信を実現するための機能を提供する．

4 音声方向決定アルゴリズム

本稿で用いられる方向決定アルゴリズムは有指向性マイク 4 本を全方位カメラの周りに９０度間隔に配置して適用する．マイク番号は０度から９０度毎に１ ～ ４マイクとして定義されている．

本システムでは有指向性マイクを用い，マイク正面を中心に音声を拾うことで，前方以外の不要な音声もしくは雑音等の不要な音声を最小限に抑えることができ，音声方向決定の精度を高める事が可能となっている．

方向決定では，一定間隔に配置されたマイクそれぞれから音声を取得し，取得した音声の波形データから一定間隔

TV Conferencing System provided with Omni-directional image operation using Audio Direction Information

† Junichi Saito
†† Yuya Maita
† Koji Hashimoto
† Yoshitaka Shibata
Faculty of Software and Information Science, Iwate Prefectural University (†)
Graduate School of Software and Information Science, Iwate Prefectural University (††)

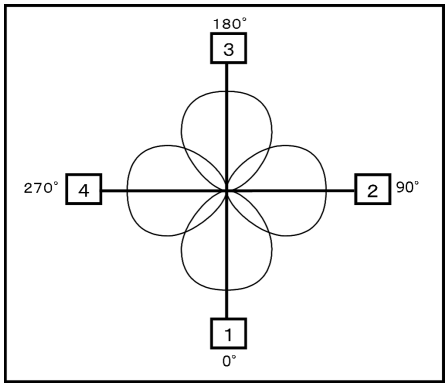

図 3: マイク配置と指向性特性

で更新される波形の大きさの平均を割り出し，最も大きいマイク２本を検出する．この時マイクは基本的に隣り合うものとする．マイク番号が隣り合わない場合はその瞬間のデータを使用せず一つ前に使用したデータを採用した後，次の入力に備える．複数話者が同時に発話した場合，より大きい入力が行われたマイク方向の話者を優先して選択するものとする．

最も大きい音声入力を Hm，２番目に大きい音声入力を Lm とし，最も大きい音声入力のあったマイク番号を m1，２番目に大きい音声入力のあったマイクの番号 m2 とする．

音声入力が開始されると，全てのマイク番号と音量をセットで取得し，1 番目と２番目に大きい音声データの値とマイク番号を算出する．

基準となる角度（マイク番号が小さい方のマイクの存在する角度) を θ_{base} とすると, 求める角度 θ は以下の式で表現することができる．

$$\theta = \begin{cases} \theta_{base} + \left(\frac{90}{Hm+Lm}\right) \times Hm & (m1 \geq m2) \\ \theta_{base} + \left(\frac{90}{Hm+Lm}\right) \times Lm & (m1 < m2) \end{cases}$$

例外として，２本算出されたマイク番号が１と４であった場合のみ，マイク番号１は位置的には大きくて，番号的には小さいということになるため，マイク番号は１のほうが大きいものとして計算する．得られたデータを基に話者の存在する角度を計算する．これにより，部分拡大映像に伴う話者方向の音声のみの再生が可能となる．

5 プロトタイピングと評価

モジュール構成図 4 は 4ch 入力の音声から発話者の角度を算出する Angle Calculator，方向データと音声データの送信を管理する Audio Sender，方向データと音声データの受信を管理する Audio Receiver，ユーザの要求に対して映像と音声を加工する User Policy，映像と音声の再生を管理する Play Control，録音機能を管理する Record から成る．

評価としてはプロトタイプとして２地点間リアルタイム映像による円卓会議中継システムの実装を行う．利用者はネットワークを通してそれぞれの空間での注目すべき発話者の映像と音声を迷うことなく見聴きすることが可能となる．

システム構成はサーバー/クライアント PC 及び PAL レンズを装着した DV カメラ，最大８チャンネルの音声を同時にサンプリング可能な 8 チャンネル音声入力ボード，有指向性マイク 4 本から成り，マシンスペックは，CPU が Pentium4 3.73GHz，memory が 2.00GB，OS は WindowsXP を使用し，これは開発環境も同様である．開発言語には C++を用い，8 チャンネル音声入力ボードの制御には ASIO API を用いる．

最終的な評価は作成したプロトタイプを用いて，映像と音声の同期，送信される音声の正確性，遅延の測定，音声方向決定アルゴリズムの精度の測定を行う．実験は 4 人程度の TV 会議を想定して行う．90 °間隔にマイクを配置したものを基準とし、話者の位置を変ながら，多様な話者の位置や状況における音声方向決定アルゴリズムの有効性を検証する．さらに話者を歩かせた場合にどのような結果が得られるかも検証する．

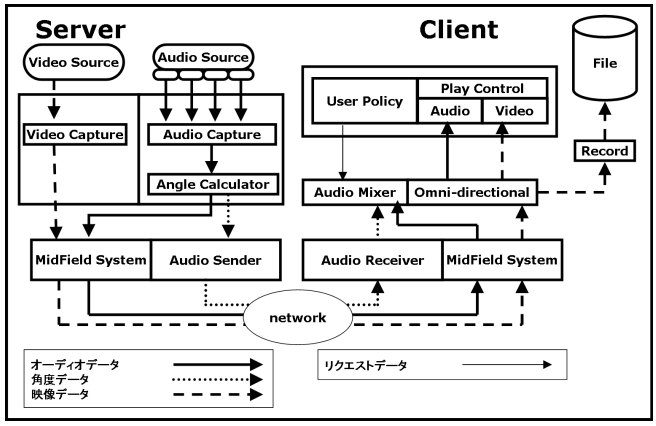

図 4: モジュール構成

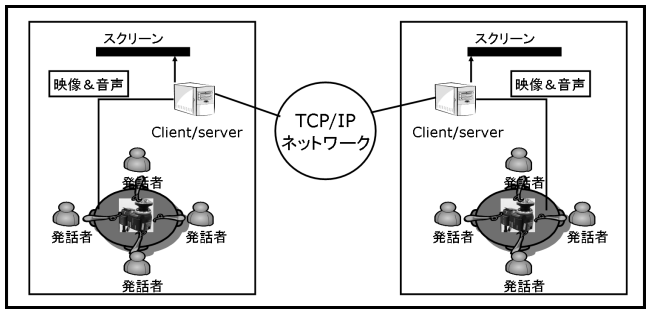

図 5: プロトタイプシステム構成

6 まとめ

本稿では，有指向性マイクから取得したデータによる話者方向決定を応用した全方位 TV 会議システムについて述べた．話者の位置特定を予め行い必要部分な映像又は音声のみを再生することで，利用者は全方位の中から注目すべき点を探す手間を省く事が可能である．また，現在は本稿で提案したアルゴリズムの評価を行うためにプロトタイプシステムの構築を行っている．

参考文献

[1] M. Fiala, D. Green. and G. Roth, A Panoramic Video and Acoustic Beamforming Sensor for Videoconferencing. published in the IEEE International Workshop on Haptic Audio Visual Environments and theirApplications(HAVE'2004). Ottawa, Ontario, Canada. October2-3, 2004. NRC47364

[2] B. Kapralos, M. Jenkin, and E. Milions. Audiovisual localization of Multiple speakers in a video teleconferencing setting. InIntl.Jour.Imaging Systems and Technology, volume13(1), pages95-105,2003

[3] Koji Hashimoto and Yoshitaka Shibata：Design of a Middleware System for Flexible Intercommunication Environment, International Conference on Advanced Information Networking and Applications(AINA2003), pp.59-64, 2003

[4] 米田祐也，橋本浩二，柴田義孝：全方位映像通信のためのミドルウェアの研究，情報処理学会研究報告 2005-DPS-122 pp.93-98, 2005.