対面の複数人対話を撮影対象とした 対話参加者の視線に基づく映像切替え方法とその効果

竹 前 嘉 修[†,☆] 大 塚 和 弘[†] 武 川 直 樹[††]

本論文では，会議のような対面で行われる複数人の対話を撮影対象として，映像を介しその対話内容を対話に参加していない人に分かりやすく伝えるために，対話参加者の視線に基づく映像切替え方法を提案する．提案方法は，最も多くの対話参加者が視線を向けている人物の映像を選択し，時間的に画面を切り替える方法である．我々は，視聴者が映像を介し対話内容を理解するために必要な情報として，誰が誰に話しかけているかという「対話の方向性」と話し手の話しかけに対する「相手の反応」の 2 つの要素が重要であると考え，その観点から提案方法の有効性を検証する．まず，提案方法によりどのような映像が生成されるのかを確認するために，3 人から 5 人の複数人対話を収録し，最も多くの対話参加者が視線を向けている人物の推移を分析する．その結果から，提案方法により「話し手」と「相手」の映像が交互に選択・提示されることを示し，映像を通じて「対話の方向性」，および，「相手の反応」が伝達される可能性を確認する．次に，提案方法により編集された映像を実際に視聴した際にこれらの情報が伝達されることを検証するために，音声に基づく映像切替えやマルチ画面などの映像表現を比較対象として，被験者実験により評価を行う．その結果から，提案方法により編集された映像が，従来の映像表現と比べ，「対話の方向性」と「相手の反応」をより正確かつ分かりやすく視聴者に伝達できることを示す．

Effectiveness of Video Editing Based on Participants’ Gaze for Multiparty Conversation in Face-to-face Situations

Yoshinao Takemae,[†,☆] Kazuhiro Otsuka[†] and Naoki Mukawa[††]

This paper presents a video editing rule based on participants’ gaze for clearly conveying the contents of the conversation in face-to-face multiparty conversations, such as meetings, to non-attendees. We propose a new video editing rule that selects the shot of the participant that most participants are gazing at. *Who is talking to whom* and *addressee’s response to speakers* are extremely crucial pieces of information that allow viewers to more fully understand the conversation as captured on video. From this point of view, this paper conducts experiments and evaluations in two stages. First, we analyze the time transitions of participants’ gaze direction using captured 3-person to 5-person multiparty conversations in order to verify how videos are edited by the proposed method. The results show that the proposed method triggers alternate camera selection of the speaker and the addressee and effectively convey the desired information. Second, to verify the effectiveness of videos edited by the proposed method, we then present experiments that compare the proposed method to conventional visual representations such as multiple view shots and camera selection based on utterance. These results show that the proposed method can more clearly and accurately convey *who is talking to whom* and *addressee’s response* to the viewers.

1. はじめに

会議は企業などにおける不可欠なコミュニケーションの一形態である．しかし，移動時間がかかることや別の予定とかち合うことなどの理由により，会議への出席を必要とする人が必ずしもすべての会議に参加できるとは限らない．そのため，TV 会議システムや会議の映像アーカイブは，それらの問題を解決し，時間や空間を隔てた人と人のコミュニケーションを効率/円滑にするための手段の 1 つとして期待されている．

† 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
NTT Communication Science Laboratories, NTT Corporation
†† 東京電機大学
Tokyo Denki University
☆ 現在，NTT サイバーソリューション研究所
Presently with NTT Cyber Solutions Laboratories, NTT Corporation

従来の典型的な TV 会議システムなどでは 1 台の固定カメラや全方位カメラ[1),2)] による撮影が行われている．しかし，このようなカメラを用いて会議に出席している多数の参加者を撮影する場合，参加者 1 人 1 人の顔領域が画像上で小さくなる．その結果，映像を見る視聴者にとって参加者の表情や視線などの非言語情報を十分に読み取ることが難しくなる[3)]．

このような 1 台のカメラによる撮影の問題を解決するための方法として，複数のカメラを用いる方法がある．これは会議の各参加者を各々のカメラで撮影し，それらの映像を 1 つのモニタ上に並べて提示する方法である．しかし，この方法では対話の状況に応じて視聴者がどの参加者の映像を見るべきかを決定するための認知的な負荷が大きくなり，視聴者が対話内容を理解することに専念できない可能性がある．また，各参加者の画面間での幾何的整合性が完全に保存されていないため，視聴者が映像から参加者の位置関係を理解することが困難になると考えられる．そのため，視聴者にとって誰が誰に話しかけているかということが分かりにくいと予想される．

また，会議の各参加者を各々のカメラにより撮影した映像から 1 つの映像を選び，時間的に画面を切り替えて表示する方法がある．これは映像切替え，映像選択，複数カメラのスイッチングなどといわれている．たとえば，Cutler らの “Distributed Meetings” では，対話参加者の音声情報に基づく映像切替え方法が用いられている[2)]．この方法は，1 台の全方位カメラにより対話参加者全員を撮影した映像から，音声情報に基づき「話し手」の映っている領域を切り出し，それを時間的に切り替えて提示する方法である．しかし，この方法では，「話し手」の映像のみが選択されるため，「話し手」の話しかけているときの「相手」が誰であるのかを十分に伝達できないことが予想される．さらに，「話し手」の話しかけに対し「相手」が発話をともなわず表情や視線などの変化により反応する場合，「相手」の映像が選択されないため，視聴者がその「相手の反応」を十分に読み取ることができないと予想される．

このほかの映像切替え方法として，井上らは，TV 討論番組におけるショットの種類とショットの持続時間を分析し，そのショットの遷移確率に基づいて，対話参加者の映像を切り替える方法を提案している[4)]．この方法に基づく映像には「話し手」の把握や視聴者の注意の保持に効果があることが確認されている．しかし，この方法では実際に行われている対話とは別の TV 番組のショットの遷移確率が用いられているため，実際の対話内容に応じた適切な映像切替えは保証されないと考えられる．

一方，映画や TV 番組の制作においては，複数の人物が対話している場面を視聴者に分かりやすく伝達するために，制作者の長年の経験や知識に基づいた人手による映像の編集が行われている．この映像の編集には，映像切替えが用いられることが多い．その一例として，“Shot/Reverse Shot” と呼ばれる映像切替え方法がある[5)]．これは各登場人物の顔領域が大きく映された映像を交互に切り替える方法である．この方法により編集された映像はある人物の関心が向いている相手や特定の登場人物の反応を強調することができる[5)〜8)]．このような効果の期待できる理由として，映像切替えには視聴者の注意を制御し，視聴者に能動的な視聴を促す機能があることなどがあげられる[9),10)]．このような効果のため，映像切替えという映像表現は複数の人物が対話している場面を編集するための手段として有効であると考えられる．しかし，このような効果的な編集は，専門知識のある制作者による対話全体の展開や対話状況の理解に基づき行われるため，自動化の実現が難しい．

従来，会議の自動映像編集やアーカイブ構築などを目的とした研究においては，「話し手」の様子を視聴者に分かりやすく伝達することに主眼が置かれており，その点においては効果をあげていた．しかしながら，会議の状況や内容を理解するうえで参加者にとって重要な情報としては，この「話し手」の様子以外にも，「話し手」が話しかけている「相手」が誰であるのかという情報や，その「相手」がどのような反応を示しているのかという情報，その他の参加者の表情や視線などの非言語情報，参加者間の人間関係などの様々な情報も重要であると考えられる．しかし，従来研究においては，これらの情報の伝達という観点はあまり着目されておらず，この観点を陽に映像編集方法の設計指針として取り入れた研究はいまだ十分に進んでいないのが現状である．我々は，会議の状況・内容を映像によって分かりやすく伝達するためには，上記のような情報伝達の観点を取り入れた映像編集が不可欠であると考える．本研究では，そのなかでも最も基本的な要素の 1 つとして考えられる，誰が誰に話しかけているかという「対話の方向性」，および，「話し手」の話しかけに対する「相手の反応」の 2 つの要素に着目し，これらを視聴者に正確かつ分かりやすく伝達するための映像編集技術を確立することを目標に研究を進めている．本論文では，その一環として，対話参加者の視線の振舞いを用いて映像の切替えを行うという新

しい映像編集のアプローチを提案し，実験によりその有効性の検証を行う．なお，本論文では，この視線に基づアプローチの原理的な有効性の確認を目標とするため，人手によって抽出された視線データを用いた分析・実験を行う．

我々が視線に基づくアプローチを提案するに至った理由は以下の考えに基づいている．まず，対話参加者は，対話の各時点において，対話の状況を適切に理解し対話に参加するため，観測が必要と思われる人物を逐次判断し，その人物に視線を向けてその人物の表情などの情報を取得していると考えられる．我々はこの性質に着目し，「対話の方向性」や「相手の反応」といった情報を機械的に直接映像などから抽出・認識しなくても，対話参加者の着目している人物を映像として表示することで，結果的に，視聴者に対しても「対話の方向性」や「相手の反応」が分かりやすい映像ができるという仮説を立てた．また，より多くの参加者が見ているところには，対話の理解により重要な情報が含まれているとも考えられる．このような仮説に立脚し，本論文では，最も多くの参加者が視線を向けている人物の映像を選択して切り替えて提示するという方法を提案する．

また，本論文では，提案方法の妥当性を検証するために以下の 2 段階の実験・評価を行った結果を示す．まず，はじめに，提案方法によりどのような映像ができるのかを確認し，その映像により「対話の方向性」と「相手の反応」が伝達される可能性を検証するために，最も多くの参加者が視線を向けている人物の推移を分析する．その結果から，提案方法により「話し手」と「相手」の映像が交互に選択されることを示す．次に，提案方法により編集された映像により「対話の方向性」や「相手の反応」が視聴者にどの程度正確かつ分かりやすく伝達されるのかを検証するために，被験者を用いた映像評価実験を行う．

従来，編集された映像の視聴者に与える影響を調査するために，いくつかの方法による映像評価が行われている[11)〜15)]．たとえば，Kraft は，映像切替えを含む映像と含まない映像の違いを「活動的」「力強い」などの印象の観点から評価している[11)]．また，井上らは，TV 会議での使用が想定される数種類の映像の違いを,「話し手」の分かりやすさや，画面上の人物の表情や身振りの分かりやすさ，などの観点から評価している[12)]．これらの研究に対し，本論文では，「対話の方向性」および「相手の反応」という従来にはなかった視点からの映像評価を行うため，新しい映像の実験・評価方法を提案する．また，従来研究の評価実験で用いられる評価方法の多くは，被験者の主観評価を用いるものがほとんどであったのに対し，提案法では，それらの情報がどの程度正確かつ分かりやすく視聴者に伝達されるのかを評価するために，被験者の主観評価に加えて被験者の回答の正答率を測定し評価するという点が提案法の特徴の 1 つとしてあげられる．

本論文は以下のように構成される．2 章では提案方法について述べる．3 章では対話中の対話参加者の視線の振舞いを分析することによって，提案方法により「対話の方向性」および「相手の反応」が視聴者に伝達される可能性を検証する．4 章では提案方法により編集された映像の有効性を映像評価実験により検証する．5 章では，提案方法の一般性，および，研究の課題と方向性について議論する．最後に 6 章でまとめを行う．

2. 視線に基づく映像切り替え方法の提案

本章では，対話参加者の視線に基づく映像切替え方法を提案する．まず，本研究のアプローチについて述べる．次に，提案方法について詳細に述べる．最後に提案方法により編集された映像の例を示す．

2.1 本研究のアプローチ

本研究では，映像編集方法を構成するために対話参加者の視線の振舞いに着目する．我々が視線に着目した理由は以下の考えに基づいている．対話参加者が対話を理解するためには，言語情報とともに，視線や表情などの非言語情報を取得することが必要であると考えられる．なぜならば，このような非言語情報を取得することにより，対話参加者は互いに「相手の反応」や心理状態などを読み取ることができるからである[17)]．さらに，対話参加者は,「話し手」のみを見るのではなく，「話し手」と「相手」を交互に見ることによって，誰が誰に話しかけているかという「対話の方向性」を読み取る性質があると考えられる．

このような「相手の反応」や「対話の方向性」の情報を視覚を通して取得・解釈することにより，対話参加者は対話の理解を促進させていると考えられる．このため，参加者が対話中の各時点において視線を向けているところには，その時点において重要な参加者の視覚情報（非言語情報）が含まれている傾向がある．この結果，より多くの参加者の見ているところには，対話の理解のために取得するべき参加者の視覚情報がより含まれていると推測される．そのため，対話の進行とともに，最も多くの対話参加者の見ている人物の映像を選択することにより，視聴者に対しても，「対話の方向性」および「相手の反応」の分かりやすい映

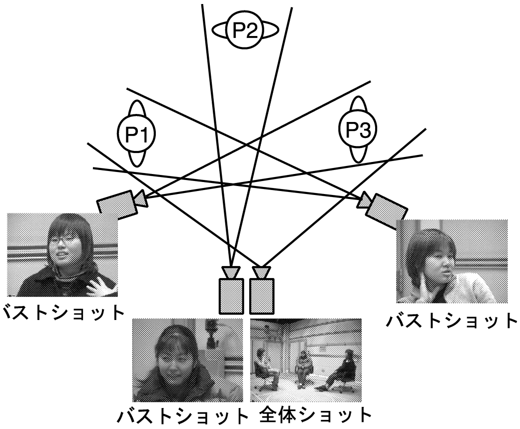

図 1 3 人対話における参加者の位置関係と収録するショットの例．各参加者（P1，P2，P3）のバストショットと全体ショットの例を示す

Fig. 1 Examples of participants' locations and captured shots in 3-person situations. They show bust shots of each participant (P1, P2, P3) and whole view shot.

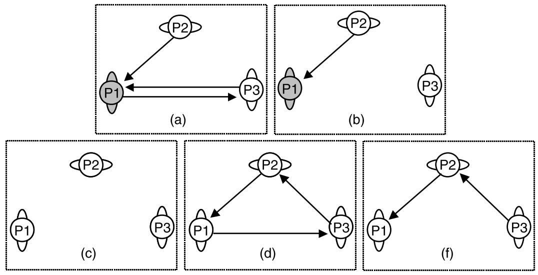

図 2 3 人対話における対話参加者（P1，P2，P3）の視線方向の例．矢印は各参加者の視線方向を示す．矢印がない場合は参加者が視線をそらしていることを示す．(a)，(b) は参加者 P1 に視線が集中している場合を示す．視線の集中の対象である参加者 P1 はグレー色に塗りつぶされている．(c)，および，(d) はそれぞれ視線の発散を示す．(f) は視線の競合を示す

Fig. 2 Example of gaze direction of each participant (P1, P2, P3) in 3-person situation. Arrows indicate each participant's gaze direction. (a) and (b) show participants gazing at just one person P1; P1 is shown in gray. A participant with no arrow has averting her/his eyes. (c) and (d) shows the dispersion of gaze. (f) shows several participants receive the same number of gazes, i.e. gaze competition.

像ができると期待される．

2.2 提案方法

前節に述べた考えのもと，対話参加者の視線の多数決に基づく映像切替え方法（以下（G）と記す）を提案する．本研究では，提案方法を，対話の開始時点から終了時点までの対話全体にわたり適用する．以下では，ある 3 人の対話を例に提案方法を説明するが，提案方法は 4 人以上の複数人対話においても同様に適用可能である．まず提案方法を適用するために必要な映像の収録および視線方向のラベル付けについて述べる．次にそのデータを用いて映像切替え方法について述べる．

2.2.1 映像の収録

各対話参加者のバストショット（胸上から顔領域を映したもの），および，対話参加者全員を同一画面に収めた全体ショットを収録する．**図 1** に 3 人対話における参加者の位置関係と収録するショットの一例を示す．バストショットを用いる理由は，各参加者の表情や視線などの非言語情報を視聴者に明瞭に伝達するためである．

2.2.2 視線方向のラベル付け

映像の各フレームごとに各参加者の視線方向のラベルを付ける．具体的には，ある参加者の顔方向に視線を向けている場合にはその参加者の番号を，各参加者の顔方向以外に視線を向けている（視線をそらしている）場合には 0 の番号を付ける．ただし，今回は，提案方法の妥当性の検証を目的とするため，収録した各参加者のバストショットを用いて目視によりラベル付けを行う．

2.2.3 映像切替え方法

図 2 を用いて提案する映像切り替え方法を説明する．図 2 に 3 人対話における各参加者（P1，P2，P3）の視線方向の例を示す．この図において矢印は各参加者の視線方向を示す．ここで，提案方法の説明に用いる用語について述べておく．

他の参加者から受けている視線の数のことを**視線の投票数**と呼ぶ．視線の投票数の最大値を持つ人物が 1 人のみ存在する現象のことを**視線の集中**と呼ぶ．また，その視線の投票数の最大値を持つ人物のことを**視線の集中の対象**と呼ぶ．各参加者の視線の投票数がそれぞれ 0 である（参加者全員が他者の顔方向から視線をそらしている）現象，および，各参加者の視線の投票数がそれぞれ 1 である（参加者全員が互いに違う人物の顔方向に視線を向けている）現象のことを**視線の発散**と呼ぶ．視線の投票数の最大値を持つ人物が 2 人以上存在する現象のことを**視線の競合**と呼ぶ．ただし，各参加者の視線の投票数がそれぞれ 1 である場合は除く．

提案方法では，参加者の 3 つの視線の振舞いに応じて以下の方法を適用する．

(1) 方法 1 視線の集中が観測される場合には，視線の集中の対象である人物のバストショットを選択する．たとえば，図 2 (a) では，P1 のみに P2 と P3 の視線が集まっているため，P1 のバストショットを選択する．また，図 2 (b) では，P1 が P2 から視線を向け

図 3　3 人対話のある区間における視線の振舞いと映像の例．(a) は各参加者（P1，P2，P3）のバストショット，P1 の非言語行動の変化，P2 と P3 の発話内容を示す．(b) は各参加者の視線方向の推移を示し，矢印が各参加者の視線方向を示す．視線の集中の対象である人物はグレー色に塗りつぶされている．(c)，(d) は，それぞれ，提案方法（G）により編集された映像，話者画面（S）による映像を示す．(c) において，視線の集中の推移により切り替わるショットは太線で囲まれている

Fig. 3　Gaze patterns in one part of conversation in 3-person situation and examples of final video sequences. (a) shows the transition of bust shots of each participant (P1, P2, P3), nonverbal behavior of P1, transcripts of P2 and P3. (b) shows the transition of gaze direction. Arrows indicate each participant's gaze direction. The person being gazed at by the other people is shown in gray. (c) and (d) shows the video sequence produced by the proposed method and the speaker shot, respectively. The shots representing a change in camera selection are shown by bold lines in (c).

られているため，P1 のバストショットを選択する．

（2）方法 2　視線の発散が観測される場合には，全体ショットを選択する．

（3）方法 3　視線の競合が観測される場合には，その直前の参加者の視線の振舞いに応じて以下のようなショットの選択を行う．

（3-1）　直前に，視線の集中が観測される場合：直前の視線の集中の対象でない人物の中で，現在，他者から視線を向けられている人物のバストショットを選択する．このような選択を行う理由は，視線の集中の対象である人物が他者から視線をそらされることにより，その人物への重要性が低くなると考えたためである．

（3-2）　直前に，視線の発散または視線の競合が観測される場合：ショットの切替えを行わず，全体ショットが継続される．

なお，映像の切替わり時における視聴者の認知的な負荷を軽減するため，1 つの画面の最低持続時間 ΔT を設定する．この時間内において，参加者の視線方向が変化する場合にはそれを無視し，映像の切替えは行わない．1 つ前の切替えから ΔT 経過以後に初めて参加者の視線方向が変化する時点で，その視線の振舞いに応じた映像の切替えを行う．

2.3　提案方法により編集された映像の例

提案方法（G）の編集過程を示す．以下に，ある 3 人対話の 7 分間について提案方法（G）を適用し作成された映像から，ある区間を切り出した映像について説明する．

図 3 に 3 人対話のある区間における視線の振舞いと映像の例を示す．図 3 (a) は各参加者（P1，P2，P3）

のバストショット，P1 の非言語行動の変化，P2 と P3 の発話内容を示す．図 3 (b) は各参加者の視線方向の推移を示す．矢印は各参加者の視線方向を示し，視線の集中の対象である人物はグレー色に塗りつぶされている．図 3 (c) は提案方法（G）により編集された映像を示す．図 3 (d) は話者画面（S）（音声に基づく映像切替え，4.2 節参照）の映像を示す．

この対話では，3 人が「人に順位をつけることが必要かどうか」について議論し，意見をまとめようとしている．議論の途中，P1 の考える「順位の定義」が他の参加者の考える「順位の定義」と食い違っている．そのため，P2 と P3 が P1 に自分らの考える定義を受け入れるように説得するが．P1 はそれを受け入れようとはしないという場面である．この場面では，ある人物の発話とそれに対する別の人物の反応の対が 2 つ観測される．まず，（a1）P2 が P1 に意見を述べる（図 3 (1)），（a2）P1 は視線をそらす，強張った表情をするなどのみで発話のともなわない無言の反応をする（図 3 (2)），の対が観測される．次には，（b1）P1 の無言の反応を見た P3 が P1 に説得する（図 3 (3)），（b2）P1 は首を傾げるなどのみで発話のともなわない無言の反応をする（図 3 (4)），の対が観測される．

このような場面では，図 3 (b) に示すように，視線の集中の対象は，P2（「話し手」）→ P1（「相手」）→ P3（「話し手」）→ P1（「相手」）に推移することが分かる．このような視線の振舞いの結果，図 3 (b) に示すように，提案方法（G）では対話の進行とともに「話し手」と「相手」が抽出され，「話し手」と「相手」の映像を交互に選択・提示することができる．このように「話し手」と「相手」の映像が交互に選択されることにより，視聴者は「話し手」や「相手」，および，「相手の反応」を容易に読み取ることができると期待される．

3. 複数人対話における視線の分析

本章では，「対話の方向性」および「相手の反応」が視聴者に伝達される可能性を検証する．我々は，「対話の方向性」と「相手の反応」が視聴者に伝達されるためには「話し手」や「相手」の映像が交互に選択・提示されることが必要であると考える．そこで，提案方法によってどのような映像が生成されるのかを確認するために，対話参加者の視線に基づき分析を行う．具体的には，「話し手」が発話している発話区間とその発話に対し「相手」が反応している反応区間の 2 種類の区間に着目し，それぞれの全区間において，視線の集中の対象が「話し手」と「相手」の間に推移する区間の頻度を分析する．

以下では，次のような順序で説明する．まず「話し手」と「相手」を定義する．次に対話の収録，および，参加者の視線方向と発話区間のラベル付けについて述べる．さらに分析対象とした発話区間および反応区間の抽出方法について述べる．最後に分析方法と分析結果について述べる．

3.1 「話し手」と「相手」の定義

本論文では，ある人物が特定の 1 人に話しかけている発話の形式のみを対象として，ある人物が発話している発話区間とその発話に対し別の人物が反応している反応区間を合わせた区間（以下，発話–反応区間と呼ぶ）において，最初に発話している人物を「話し手」と定義する．また，「話し手」が話しかけている特定の 1 人の人物を「相手」と定義する．

なお，本論文では，対話参加者の音声のパワーに基づき「話し手」を特定する．「話し手」の話しかけている「相手」については，現時点では「相手」を特定できる客観的な指標がないため[18]，対話には参加していない 2 人の評定者の主観的な判断に基づき特定する（3.4 節参照）．

3.2 対話の収録

今回は複数人の討論を対象とした．3 人，4 人，5 人をそれぞれ一組とする討論を扱った．各参加人数における組数は，それぞれ，4 組，1 組，1 組であった．各組の参加者（女性，19 歳から 23 歳，平均 21.5 歳）は友人同士の関係にあった．討論の議題は「人に順位をつけることが必要かどうか」「日本において安楽死を法的に認めるべきかどうか」などであった．収録前に，議題の候補について，各参加者に自分の考えやその理由を記入させ，意見の分かれた議題を実験者が選択した．その後，参加者は，選択された議題に対して，討論を実施した．

また，参加者の座席は，参加者全員がお互いの顔をよく見渡せ，かつ，会議の場面に適した社会距離 120〜360 [cm] [19] を保つように半円上に配置した．**図 4** に，3 人対話，4 人対話，5 人対話の収録時における対話参加者の位置，および，カメラ配置をそれぞれ示す．各参加者は常時各座席に着席した状態であった．複数のカメラにより，各参加者のバストショット，および，参加者全員を同一画面内に収めた全体ショットを撮影した（VGA サイズ，30 [fps]）．また，各対話参加者に装着されたピンマイクにより音声を収録した．収録された対話の長さは合計で，3 人対話が 85 分，4 人対話 21 分，5 人対話が 23 分であった．

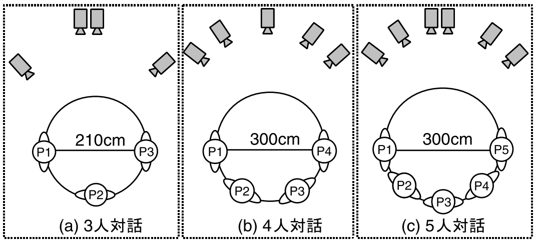

図 4 対話参加者の位置関係とカメラ配置．(a)，(b)，(c) は，それぞれ，3 人対話，4 人対話，5 人対話時の例を示す

Fig. 4 Participants' locations and camera position. (a), (b) and (c) shows examples of 3-person situation, 4-person situation, and 5-person situation, respectively.

3.3 視線方向および発話区間のラベル付け

収録された各参加者のバストショットを用いて人手により，映像の各フレームごとに参加者の視線方向をラベル付けした．このラベル付けは，討論に参加していない評定者 1 人（女性，28 歳）により行われた．また，発話区間については以下のように自動検出を行った．まず，収録された各参加者の音声の対数パワーを 100 [ms] ごとに算出した．その後，各参加者ごとに，対話全体における音声の対数パワーの平均を算出した．さらに，この値を用いた閾値処理により各参加者の無音区間を抽出し，300 [ms] 以上の無音区間に挟まれた区間を各参加者の発話区間として検出した．

3.4 分析対象の発話区間および反応区間

分析対象とした発話区間とその直後の反応区間，および，それらの抽出方法について述べる．

まず，発話区間については，前節で述べた方法により自動検出された各参加者の発話区間の中で，以下の 4 つの条件を同時に満足する区間を対象とした．

条件 1：「話し手」の発話が 5 秒以上連続する区間．これらの区間は自動検出された発話区間の長さを調べることにより抽出される．

条件 2：討論に参加していない 2 人の評定者（女性，平均 29 歳）の判断した「相手」が完全に一致した区間．この 2 人の評価の一致率は 68%であった．具体的には以下の方法で行った．まず対話全体について各参加者のバストショットと全体ショットの合成された映像（音声を含む）を作成した．次に 2 人の評定者が個別に対話全体の映像を視聴し，該当の発話区間における「相手」を主観的に判断した．このとき各評定者は対話全体および該当区間の映像を任意に何度でも視聴することが許された．2 人の評価が完全に一致しないものは分析の対象外とした．将来的には，評定者の評価と参加者の言語・非言語情報との関連性を明らかにすることにより，「相手」を自動的に抽出することが期待される．

条件 3：「話し手」が特定の 1 人（「相手」）に話しかけている区間．

条件 4：「話し手」と「相手」以外の人物による同時発話がない区間☆．これらの区間は自動検出された発話区間の重なりを調べることにより抽出される．

なお，発話区間の範囲については，「話し手」の発話開始時点から終了時点までの区間とした．また，「話し手」の発話と「相手」の発話による反応の一部は重なる場合が多い．その場合においても同様の区間とした．

次に，反応区間については，上記の 4 つの条件を満足した発話区間の直後の区間を対象とした．前節で自動検出された各参加者の発話区間に基づき，以下の 2 つの区間を抽出した．

(1)「相手」の反応が発話をともなう場合：直前の発話区間の終了時点から「相手」の発話が終了するまでの区間．

(2)「相手」の反応が発話をともなわない場合：直前の発話区間の終了時点から「相手」以外の人物が発話を開始するまでの区間．

以上の方法により抽出された発話–反応区間の数は合計で 32 であった．

3.5 分 析 方 法

すべての発話–反応区間の，全発話区間，および，全反応区間のそれぞれにおいて，視線の集中の対象が「話し手」と「相手」の間に少なくとも 1 回以上推移することが生じた区間の割合（%）を求める．

3.6 分 析 結 果

発話区間における割合は 79%であった．これは，かなり高い割合で視線の集中の対象が「話し手」と「相手」の間に推移したと考えられる．一方，反応区間における割合は 57%であった．反応区間においても，ある程度の割合で視線の集中の対象が「話し手」と「相手」の間に推移することが分かった．反応区間における割合が発話区間の割合と比べ低い理由は以下のように考えられる．「相手の反応」が発話をともなわない反応区間においては，「相手」以外の参加者はその「相手」の表情や視線などの視覚的な情報を取得しようとその「相手」に視線を向ける．この結果，その「相手」のみに視線が集中するためと考えられる．

以上のことから，対話中，視線の集中の対象が「話し手」あるいは「相手」のどちらかのみではなく，「話

☆ 複数人会話においては，同時発話の生じる場合も多い．本論文では，提案方法を，対話全体にわたり適用することを想定しているため，同時発話の区間においても，視線の集中の対象である人物のショットを選択する．

し手」と「相手」の間に推移することが分かった．このことから，対話中において，視線の集中の対象である人物の映像を選択すること（提案方法）により「話し手」や「相手」の映像が切り替わり表示され，「対話の方向性」や「相手の反応」を伝達する効果が期待できる．

4. 映像評価実験

本章では，提案方法（G）の有効性を確認するために，被験者を用いた映像評価実験を行う．比較評価には，複数人会話の撮影・提示に一般的に用いられている音声に基づく映像切替えなどの 3 種類の映像表現を用いる．本章では，(1)「話し手」が一方的に「相手」に意見を述べている，(2)「話し手」が「相手」に明示的に意見の要求や疑問を提示しその直後に「相手」が応答している，の 2 種類の基本的な会話の要素に焦点を当て，2 つの実験を行う．実験 1 では，(1) の発話区間を対象として，「対話の方向性」を評価する．実験 2 では，(2) の発話–反応区間（発話区間と反応区間を合わせた区間）を対象として，「対話の方向性」と「相手の反応」を評価する．これらの実験では，それぞれの項目について，正確性（正確さ）と明瞭性（分かりやすさ）の観点からの評価を行う．以下では，まず，被験者に提示する映像の作成方法について述べ，次に，提案方法の比較対象とした 3 種類の映像表現，および，実験 1 および実験 2 で用いた対象シーンの選定方法について述べる．その後，実験 1 および実験 2 の方法および結果について述べ，最後に実験の結果をまとめる．

4.1 提示映像の作成方法

本実験で被験者に提示する映像は以下の方法により作成する．まず，3.2 節で収録した対話の全区間について，比較対象を含む各々の映像表現の映像を作成する．次に，これらの映像から，3.4 節で述べた方法により抽出された発話区間および発話–反応区間に該当する短時間的な区間（6 秒から 20 秒前後の）の映像を切り出す．本論文では，「対話の方向性」と「相手の反応」を評価するために，被験者の主観評価に加えて被験者の回答の正答率の評価を行うが，このような評価において，被験者には対話中に多数存在する「話し手」の発話とその発話に対する「相手の反応」の対の各々を記憶することが求められる．そのため，本来の対話参加時には不必要な心的負荷が被験者にかかることはまぬがれない．そこで，そのような負荷を最小限にとどめ，各時点における瞬時的な被験者の理解の度合いを抽出するために，本実験では短時間的な映像区間を切り出し，その提示と回答を対とする実験を行う．なお，短時間の映像の区間のみからその区間における「対話の方向性」および「相手の反応」が視聴者に正確かつ分かりやすく伝達されれば，対話全体を視聴した際においてもそれらの情報が同じように伝達されると考える．

4.2 比較対象の映像表現

提案方法（G）の映像の比較対象として用いた 3 種類の映像表現（全体画面（W），マルチ画面（M），話者画面（S））について述べる．なお，本研究では，映像の視聴時に視聴者が使用するディスプレイの数は 1 台が一般的であると考え，映像を提示する画面サイズは固定とする．そのサイズ内で各画面の縦横比を保つように各映像表現の映像を作成する．

全体画面（W） 対話参加者全員を同一画面に収めた全体ショットである．**図 5** (a) に 3 人対話時の例を示す．

マルチ画面（M） 参加者のバストショットを参加者の位置関係と保ちつつ横一列に並べたものである．

(a) (b) (c) (d)

図 5 全体画面（W）とマルチ画面（M）の例．(a) は 3 人対話における全体画面（W）の例を示す．(b)，(c)，(d) はそれぞれ 3 人対話，4 人対話，5 人対話時のマルチ画面（M）の例を示す

Fig. 5 Examples of whole view shot (W) and multiple view shot (M). (a) shows an example of whole view shot (W) in 3-person situation. (b), (c) and (d) shows examples of multiple view shot (M) in 3-person situation, 4-person situation and 5-person situation, respectively.

図5(b)，(c)，(d) にそれぞれ3人対話，4人対話，5人対話時の例を示す．

話者画面（S）　発話している人物のバストショットを選択する方法である．以下映像の作成手順について述べる．まず，収録された各参加者の音声のパワーに対し閾値処理を行い，各参加者の発話している区間を検出する．次に，映像の各フレームごとに音声のパワーの最も大きい参加者をそのフレームにおける話し手として選ぶ．そして，時間的に話し手のバストショットに切り替える．ただし，1つの画面の最低持続時間 ΔT を1秒とする．この時間内に話し手が変わっても，それを無視し，他の人物のバストショットに切り替えない．1つ前の切替えから ΔT 経過時点において，その時点での話し手のバストショットに切り替える．図3(d) にその映像の例を示す．

4.3　対象シーンの選定方法

（1）実験2の対象シーン：3.4節で述べた方法により抽出された発話区間の中から，言語情報を利用して人手により「話し手」が「相手」に意見を要求している発話区間を選定した．具体的には「話し手」が「どう思う？」「どうして必要なの？」などのように発話している場面である．このような発話区間とその直後の反応区間（発話–反応区間）を実験2の対象シーンとした．なお，これらの発話–反応区間に含まれる「相手の反応」には，「私もそう思う」などのように発話のともなう反応と，強張った表情などのみで発話のともなわない無言の反応の2種類がある．

（2）実験1の対象シーン：3.4節で述べた方法により抽出された発話区間の中で，(1) に該当しなかった発話区間を実験1の対象シーンとした．

4.4　実験1：発話区間における「対話の方向性」の評価

4.4.1　方　　法

被験者　被験者は討論には参加していない男女57人（男性27人，女性30人，20歳から28歳，平均年齢25.2歳）である．被験者は4群に分けられ，第1群は提案方法（G）の映像を，第2群は全体画面（W）を，第3群はマルチ画面（M）を，第4群は話者画面（S）をそれぞれ視聴する．各群の被験者数は，1群が15人，2群が13人，3群が14人，4群が15人である．

提示映像　収録した対話の全区間について作成された各々の映像表現の映像から，該当区間（発話区間）の映像（平均10.3秒）を切り出した．シーン数は，3人対話6シーン，4人対話6シーン，5人対話3シーンである．提案方法（G）における1画面の最低持続時間 ΔT は話者画面（S）と同様に1秒とした．なお，提示映像は液晶プロジェクタを用いてスクリーンに投射した．被験者とスクリーンとの距離は，映像の評価において標準的とされる画面高H（100 [cm]）の4倍の長さとした（**図6**）．

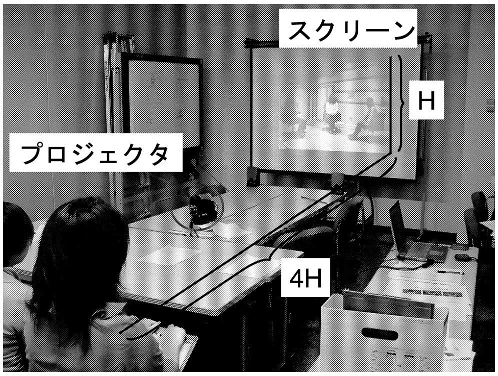

図6　映像の提示環境
Fig. 6　Environment for presenting videos.

質問紙　話し手の発話区間における「対話の方向性」を正確性と明瞭性の観点から評価するために，**表1**のような設問を設定した．Q1-1-1およびQ1-1-2は，それぞれ，「話し手」を同定する際の正確性，および，明瞭性を問う課題である．また，Q1-2-1およびQ1-2-2は，それぞれ，「話し手」の話しかけている「相手」を同定する際の正確性，および，明瞭性を問う課題である．正確性を問う課題に対しては，該当する人物を選択させた．また，明瞭性を問う課題に対しては，7段階の評定値（−3：非常に分かりにくい，−2：分かりにくい，−1：やや分かりにくい，0：どちらともいえない，1：やや分かりやすい，2：分かりやすい，3：非常に分かりやすい）を選択肢として提示し，選択させた．

正解の作成　Q1-1-1では，該当の映像区間の各フレームにおいて音声のパワーの最も大きい参加者を正解とした．Q1-2-1については3.4節で述べた方法により得られた「相手」を正解とした．

手続き　まず，評価を開始する前に，被験者には，参加者の顔と各参加者の位置関係を示した紙を配布した．次に，被験者には，参加者の顔と各参加者の位置関係を記憶するように教示し，各参加者が個別に発言している映像を視聴させた．この後，複数回の試行が行われた．各試行では，1区間の映像が1回提示され，映像の視聴後すぐに被験者は設問紙に回答するように指示された．回答時間は映像の視聴後1分とした．また，映像の提示順序は，各群でランダムに提示されるようにした．

表 1　実験 1 の設問内容
Table 1　Questionnaire for experiment 1.

設問番号	設問内容
Q1-1-1 :「話し手」の同定の正確性	話し手は誰だったと思いますか？
Q1-1-2 :「話し手」の同定の明瞭性	話し手は誰だったか，よく分かりましたか？
Q1-2-1 :「相手」の同定の正確性	話し手が中心的に話しかけていた人は誰だったと思いますか？
Q1-2-2 :「相手」の同定の明瞭性	話し手が中心的に話しかけていた人は誰だったか，よく分かりましたか？

表 2　実験 1 における分散分析の結果
Table 2　ANOVA results for experiment 1

設問番号	映像表現の主効果	対話人数の主効果	交互作用
Q1-1-1	$F(3,855) = 2.40$; $p = .66$	$F(2,855) = 1.15$; $p = .31$	$F(6,855) = .82$; $p = .54$
Q1-1-2	$F(3,855) = 15.09$; $p = .0001^{***}$	$F(2,855) = 1.58$; $p = .21$	$F(6,855) = 3.11$; $p = .005^{**}$
Q1-2-1	$F(3,855) = 7.55$; $p = .0001^{***}$	$F(2,855) = .96$; $p = .38$	$F(6,855) = 1.98$; $p = .064$
Q1-2-2	$F(3,855) = 8.92$; $p = .0001^{***}$	$F(2,855) = .25$; $p = .77$	$F(6,855) = 1.41$; $p = .89$

(***: $p < .001$, **: $p < .01$, *: $p < .05$)

4.4.2　結果と考察

被験者による回答の結果を，正確性については平均正答率として，また，明瞭性については，平均評価値として集計した．各設問について，4 群の被験者から得られた回答を従属変数とし，映像表現（(G), (W), (M), (S) の 4 種類）× 対話人数（3 人，4 人，5 人の 3 種類）の 2 要因分散分析を適用した．分散分析の結果を**表 2** に示す．各設問の結果を**図 7** に示す．さらに，分散分析の結果，有意差があった場合には Tukey の多重比較を行った．以下「話し手」および「相手」の同定の結果について考察する．

「話し手」の同定

（1）Q1-1-1 :「話し手」の同定の正確性　各映像表現の平均正答率は 90% 以上であり，映像表現間で有意差はなかった．

（2）Q1-1-2 :「話し手」の同定の明瞭性　提案方法（G）の平均評価値はすべての対話人数において 2.3 点以上であった．3 人対話のときは映像表現間に有意差はなかった．4 人および 5 人対話のときは，全体画面（W）よりも，提案方法（G），マルチ画面（M），および，話者画面（S）の平均評価値は有意に高かった ($p < .001$)．提案方法（G），マルチ画面（M），および，話者画面（S）の間には有意差はなかった．

上記の結果から，対話人数が多いほど，「話し手」を正確かつ分かりやすく伝えるためには，4 つの映像表現の中では，提案方法（G），マルチ画面（M），および，話者画面（S）がより有効であることが示唆された．この理由は，これらの映像表現では「話し手」の顔領域のみを映した映像が提示され，被験者が容易に参加者の口の動きを認識できたためと考えられる．全体画面（W）の平均評価値が対話人数の増加とともに低くなった理由は，対話人数が増加すると，画面上の参加者の顔領域のサイズがさらに小さくなり，被験者にとって参加者の口の動きの認識が難しくなったためではないかと考えられる．

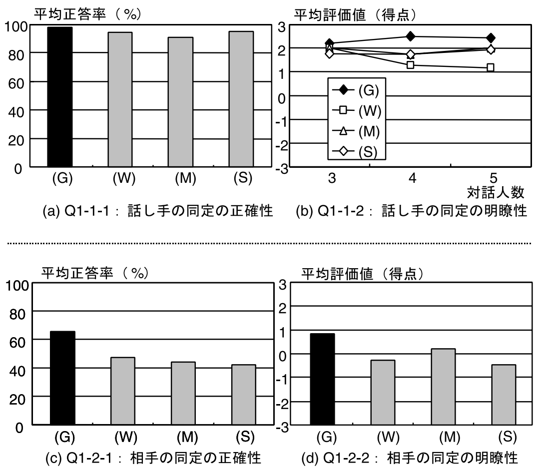

図 7　実験 1 の結果．(a) は Q1-2-1 における各映像表現の平均正答率を示す．(b) は Q1-1-2 における各映像表現の平均評価値を示す．(c) は Q1-2-1 における各映像表現の平均正答率を示す．(d) は Q1-2-2 における各映像表現の平均評価値を示す．(G)，(W)，(M)，(S) はそれぞれ，提案方法，全体画面，マルチ画面，話者画面を示す．また，(b) において横軸は対話人数を示す

Fig. 7　Survey results for experiment 1. (a) shows average rates of correct answers in Q1-1-1. (b) shows average scores in Q1-1-2. (c) shows average rates of correct answers in Q1-2-1. (d) shows average scores in Q1-2-2. (G), (W), (M) and (S) represent the proposed method, whole view shot, multiple view shot, and speaker's shot, respectively. The horizontal axis plots participant number in (b).

表 3 実験 2 の設問内容
Table 3 Questionnaire for experiment 2.

設問番号	設問内容
Q2-1-1：「話し手」の同定の正確性	最初に発言していた人は誰だったと思いますか？
Q2-1-2：「話し手」の同定の明瞭性	Q2-1-1 の発言者は誰だったか，よく分かりましたか？
Q2-2-1：「相手」の同定の正確性	Q2-1-1 の発言者の発言の最後において，話しかけていた相手は誰だったと思いますか？
Q2-2-2：「相手」の同定の明瞭性	Q2-1-1 の発言者が話しかけていた相手は誰だったか，よく分かりましたか？
Q2-2-3：「相手の反応」の同定の正確性	Q2-1-1 の発言者の話しかけに対して，話しかけられた人はどのような反応をしていたと思いますか？
Q2-2-4：「相手の反応」の同定の明瞭性	Q2-1-1 の発言者の話しかけに対して，話しかけられた人の反応はよく分かりましたか？

「相手」の同定

（3）Q1-2-1：「相手」の同定の正確性 提案方法（G）の平均正答率は 66%であった．これらの結果は他の映像表現よりも有意に高く評価された（$p < .001$）．

（4）Q1-2-2：「相手」の同定の明瞭性 提案方法（G）の平均評価値は 0.9 点であった．これらの結果は他の映像表現よりも有意に高く評価された（$p < .001$）．

上記の結果から，発話区間において「話し手」の話しかけている「相手」を正確かつ分かりやすく伝達するためには，提案方法（G）が 4 つの映像表現の中ではより有効であることが示唆された．提案方法（G）が評価された理由は以下のように考えられる．実験で用いた 15 シーンの中で，「話し手」と「相手」の映像が選択されたものは 12 シーンであった．これらの 12 シーンでは，提案方法（G）の平均正答率は 75%，平均評価値は 1.2 点であった．残りの 3 シーンでは「話し手」以外の参加者の視線が「話し手」に集中したため，「話し手」のみの映像が選択された．これらの 3 シーンでは，各映像表現の評価は同程度であった．このことから，提案方法（G）が評価された理由は「話し手」と「相手」の映像が交互に選択されることに起因したためと考えられる．他の映像表現の評価が低い理由は以下のように考えられる．全体画面（W）では，映像中の参加者の顔領域が小さく，話し手の視線方向が十分に認識できず，被験者にとって話しかけている相手を判断しにくかったためと考えられる．マルチ画面（M）では，映像から参加者の位置関係を理解するのが困難になり，話し手の視線が向いている相手を誤って認識してしまったためと考えられる．話者画面（S）では「話し手」の映像のみが選択され，「相手」を推測することが困難であったためと考えられる．

4.5 実験 2：発話–反応区間における「対話の方向性」と「相手の反応」の評価

4.5.1 方 法

被験者 実験 1 と同じ被験者である．

提示映像 実験 1 と同様に，収録した対話の全区間にわたり作成された各々の映像表現の映像から，該当区間（発話–反応区間）の映像（平均 9.6 秒）を切り出した．なお，「相手の反応」については，3.4 節で述べた同様の方法により 2 人の評定者が特定し，両者の評価が一致したもののみを扱った．シーン数は，3 人対話 9 シーン，4 人対話 2 シーン，5 人対話 2 シーンであった．ただし，発話のともなわない無言の反応は 3 人対話のもののみ 3 シーンである．

質問紙 発話–反応区間における「対話の方向性」と「相手の反応」を正確性と明瞭性の観点から評価するために，**表 3** のような設問を設定した．Q2-1-1 と Q2-1-2 は，それぞれ，「話し手」を同定する際の正確性および明瞭性を問う課題である．Q2-2-1 と Q2-2-2 は，それぞれ，「話し手」の話しかけている「相手」を同定する際の正確性と明瞭性を問う課題である．Q2-2-3 と Q2-2-4 は，それぞれ，「話し手」の話しかけている「相手の反応」を同定する際の正確性と明瞭性を問う課題である．実験 1 と同様に，正確性を問う課題（Q2-2-3 を除く）に対しては該当する人物を，明瞭性を問う課題に対して 7 段階の評定値をそれぞれ選択させた．Q2-2-3 については「肯定」「否定」「無反応」「表情の変化」「うなずき」「首の横振り」「首の傾げ」「微笑」「その他」から該当するものを選択させた（複数回答可）．

正解の作成 Q2-1-1 については実験 1 の Q1-1-1 と同様の方法により作成された．Q2-2-1 および Q2-2-3 ついては実験 1 の Q1-2-1 と同様に 2 人の評定者の評価により作成された．

手続き 被験者への教示は実験 1 と同じである．ただし，被験者には表 3 に示した設問を提示した．

4.5.2 結果と考察

設問 Q2-1-1 および Q2-1-2 については，映像表現（4 種類）× 対話人数（3 種類）の 2 要因分散分析を適用した．設問 Q2-2-1 から Q2-2-4 については，発話のともなわない無言の反応が 3 人対話のもののみであったため，映像表現（4 種類）× 反応の種類（発話のともなう反応，発話のともなわない無言の反応の 2 種類）の 2 要因分散分析を適用した．実験 1 と同様に，

表 4　実験 2 における分散分析の結果
Table 4　ANOVA results for experiment 2.

設問番号	映像表現の主効果	対話人数の主効果	交互作用
Q2-1-1	$F(3, 741) = 8.53$; $p = .0001^{***}$	$F(2, 741) = 9.55$; $p = .0001^{***}$	$F(6, 741) = 6.60$; $p = .0001^{***}$
Q2-1-2	$F(3, 741) = 31.64$; $p = .0001^{***}$	$F(2, 741) = 4.67$; $p = .009^{**}$	$F(6, 741) = 5.58$; $p = .0001^{***}$
設問番号	映像表現の主効果	反応の種類の主効果	交互作用
Q2-2-1	$F(3, 513) = 6.92$; $p = .0001^{***}$	$F(1, 513) = 1.33$; $p = .24$	$F(3, 513) = 3.33$; $p = .019^{*}$
Q2-2-2	$F(3, 513) = 6.88$; $p = .0001^{***}$	$F(1, 513) = 1.22$; $p = .26$	$F(3, 513) = 3.64$; $p = .012^{*}$
Q2-2-3	$F(3, 513) = 2.83$; $p = .038^{*}$	$F(1, 513) = .92$; $p = .33$	$F(3, 513) = 3.64$; $p = .013^{*}$
Q2-2-4	$F(3, 513) = 7.75$; $p = .0001^{***}$	$F(1, 513) = 14.66$; $p = .0001^{***}$	$F(3, 513) = 3.48$; $p = .016^{*}$

(***: $p < .001$, **: $p < .01$, *: $p < .05$)

分散分析の結果，有意差があった場合には Tukey の多重比較を行った．分散分析の結果を**表 4** に示す．各設問の結果を**図 8** に示す．以下「話し手」,「相手」，および,「相手の反応」の同定の結果について考察する．

「話し手」の同定

（1）Q2-1-1：「話し手」の同定の正確性　3 人および 4 人対話のときは各映像表現の平均正答率は 85%以上であり映像表現間に有意差はなかった．5 人対話のときは提案方法（G）および話者画面（S）の平均正答率は 98%以上であり他の映像表現よりも有意に高かった（$p < .01$）．

（2）Q2-1-2：「話し手」の同定の明瞭性　3 人および 4 人対話のときは映像表現間に有意差はなかった．対話人数が 5 人のときは提案方法（G）と話者画面（S）の平均評価値は 2.3 点以上であり他の映像表現よりも有意に高かった（$p < .01$）．このとき，提案方法（G）と話者画面（S）の間には有意差はなかった．

上記の結果から，対話人数が多いほど,「話し手」を正確かつ分かりやすく伝えるためには映像切替えの提案方法（G）と話者画面（S）がより有効であることが示唆された．これらの方法が評価された理由は，すべてのシーンにおいて「話し手」の顔領域を大きく映した映像が選択され，被験者は容易に参加者の口の動きを認識することができたためと考えられる．また，実験 1 の結果（図 7 (b) 参照）と異なり，マルチ画面（M）の評価値が 5 人対話のときに下がった理由は，(1) 画面上の各参加者の顔領域が小さくなり，被験者が各参加者の口の動きを判断するのに時間がかかってしまったこと，(2) 実験 2 の設問数が実験 1 と比べて多く，被験者への負荷が増えたこと，などが重なったためと考えられる．

「相手」の同定　各設問に対し，発話のともなう反応と発話のともなわない無言の反応の 2 種類に分けて考察する．

（3）Q2-2-1：「相手」の同定の正確性　発話のともなう反応の場合には，映像表現間に有意差はなかっ

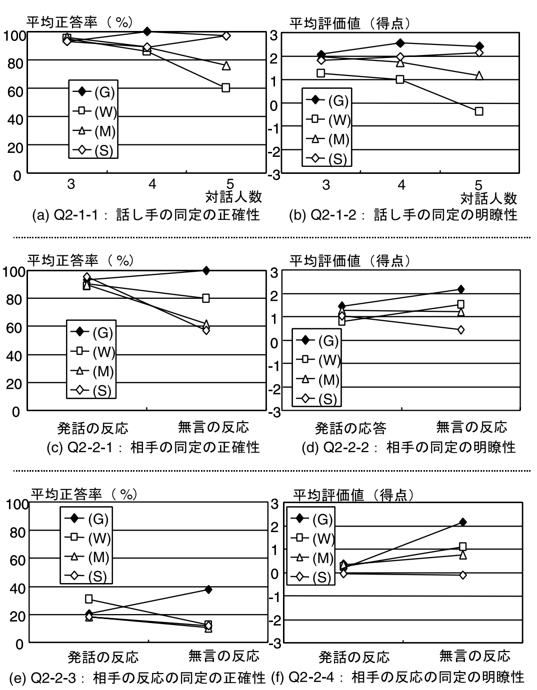

図 8　実験 2 の結果．(a) は Q2-1-1 における各映像表現の平均正答率を示す．(b) は Q2-1-2 における各映像表現の平均評価値を示す．(c) は Q2-2-1 における各映像表現の平均正答率を示す．(d) は Q2-2-2 における各映像表現の平均評価値を示す．(e) は Q2-2-3 における各映像表現の平均正答率を示す．(d) は Q2-2-4 における各映像表現の平均評価値を示す．(G)，(W)，(M)，(S) はそれぞれ，提案方法，全体画面，マルチ画面，話者画面を示す．(a) および (b) において横軸は対話人数を示す．(c)，(d)，(e)，(f) において横軸は相手の反応の種類を示す

Fig. 8　Survey results for experiment 2. (a) shows average rates of correct answers in Q2-1-1. (b) shows average scores in Q2-1-2. (c) shows average rates of correct answers in Q2-2-1. (d) shows average scores in Q2-2-2. (e) shows average rates of correct answers in Q2-2-3. (f) shows average scores in Q2-2-4. (G), (W), (M) and (S) represent the proposed method, whole view shot, multiple view shot, and speaker's shot, respectively. The horizontal axis plots participant number in (a) and (b). The horizontal axis plots response with utterance and that with silence in (c), (d), (e) and (f).

た．これは被験者が参加者の音声情報に基づき「相手」を判断したためと考えられる．一方，発話のともなわない無言の反応の場合には，提案方法（G）の平均正答率は 100%であり他の映像表現よりも有意に高かった（$p < .01$）．

（4）Q2-2-2：「相手」の同定の明瞭性　発話のともなう反応の場合には，映像表現間に有意差はなかった．一方，発話のともなわない無言の反応の場合には，提案方法（G）の平均評価値は 2.3 点でありマルチ画面（M）と話者画面（S）よりも有意に高かった（$p < .001$）．また，提案方法（G）と全体画面（W）の間には有意差はなかった．

上記の結果から，「相手」が発話のともなわない反応をしている場合において「話し手」の話しかけている「相手」を正確かつ分かりやすく伝達するために，提案方法（G）が 4 つの映像表現の中ではより有効であることが示唆された．この理由は以下のように考えられる．本実験に用いた無言の反応のシーンでは，「相手」が視線を「話し手」からそらし強張った表情をするなどの「否定」の反応を示していた．Exline は，対話の参加者が対立している状況において参加者は相手の反応を正確に読み取ろうと相手をよく見る性質があることを述べている[20]．本実験で用いたシーンは対立状況に等しいと考えられるため，Exline の説明と同様に，「相手」以外の参加者が「相手」の表情などの視覚的な情報を正確に取得しようとその人物に視線を向けたと考えられる．その結果，参加者の視線が「相手」に集まり，「相手」の映像が選択されたことが提案方法（G）の評価につながったと考えられる．

なお，提案方法（G）では，対立状況でない場合においても，「相手」以外の参加者の視線が「相手」に集中するため，「相手」の映像が選択されることが期待できる．これは，「話し手」が「相手」に意見を要求しているときは，「相手」以外の参加者が「相手」の表情などの視覚的な情報を取得しようとその「相手」に視線を向けるという性質があるためである．また，以下の実験結果からもこのような視線の性質が示唆される．本実験で用いた発話の反応をともなう 10 シーンのうち 8 シーンでは，「相手」が「私もそう思う」「うんうん」「そうそう」「そうだね」などの発言（相鎚を含む）をともなっていた．これらは複数人会話においてよく観測される場面である．実験の結果，この 8 シーンのうち 7 シーンにおいて，上記の視線の性質に基づき「相手」に他者の視線が集中したため，「相手」の映像が選択された．

「相手の反応」の同定　各設問に対し，発話のともなう反応と発話のともなわない無言の反応の 2 種類に分けて考察する．

（5）Q2-2-3：「相手の反応」の同定の正確性　発話のともなう反応の場合には映像表現間に有意差はなかった．この理由は，「相手」の「私もそう思う」「私は絶対にない」などの発言内容に基づき，被験者がその「相手の反応」を推測したためと考えられる．一方，発話のともなわない無言の反応の場合には，提案方法（G）の平均正答率は 39%であった．この値はそれほど高くないものの，他の映像表現よりも有意に高かった（$p < .001$）．

（6）Q2-2-4：「相手の反応」の同定の明瞭性　発話のともなう反応の場合には映像表現間に有意差はなかった．一方，発話のともなわない無言の反応の場合には，提案方法（G）の平均評価値は 2.2 点であり話者画面（S）よりも有意に高かった（$p < .001$）．提案方法（G），全体画面（W），マルチ画面（M）の間には有意差はなかった．

上記の結果から，発話のともなわない無言の反応の場合，「相手の無言の反応」を正確かつ分かりやすく伝達するためには，提案方法（G）が 4 つの映像表現の中ではより有効であることが示唆された．提案方法（G）が評価された理由は，「相手」の同定のところで説明した理由と同様に，「相手」以外の参加者の視線が「相手」に集まり，その結果，提案方法（G）により「相手」の表情などの読み取りやすい映像が選択されたためと考えられる．また，映像切替えという映像表現により人物の反応が視聴者に強調されること[5)〜8)]も提案方法（G）の評価につながった可能性があると考えられる．

4.6　実験のまとめ

実験 1 および実験 2 の結果から，提案方法は，発話区間における「相手」の同定の正確性と明瞭性，および，発話–反応区間における「相手」および「相手の反応」の同定の正確性において，他の映像表現よりも有意に高い評価を得ることができた．また，その他の項目においても，他の映像表現と比べ同程度かそれ以上の評価を得ることができた．したがって，映像を介し「対話の方向性」と「相手の反応」を視聴者に正確かつ分かりやすく伝えるためには，提案方法（G）が 4 つの映像表現の中ではより有効であることが示唆された．

以上のことから，発話区間と発話–反応区間の 2 種類の短時間的な映像を用いた限定的な実験であるものの，映像を介し「対話の方向性」と「相手の反応」を視聴者に正確かつ分かりやすく伝えるために，視線に

基づくアプローチが有効であることが分かった.

5. 議 論

本論文では, 3 人から 5 人の複数人の討論を対象として, 発話区間と発話–反応区間の 2 種類の短時間的な映像を用いて, 視線に基づく映像切替えのアプローチの有効性を確認した. 本章では, この結果を活用し, 実際の会議の場面における自動映像編集システムの構築へと提案方法を発展させるために課題となる提案法の一般性や実装の方針などについて論じる.

5.1 提案方法の実装

本論文では, 提案方法の原理の有効性を検証することに主眼をおき, 人手により視線方向のラベル付けを行った. この方法を実際に自動化・実装するためには, 画像認識などに基づく視線方向の自動計測, 設置カメラの台数の削減, などが課題となる.

視線計測に関しては, 近年, 画像認識に基づく非接触計測技術の研究が精力的に行われている[21),22)]. 従来は, 被測定者の頭部可動範囲がきわめて狭いという問題があったが, 頭部の自動追尾機能の導入などにより改良が進められている. また, 提案方法に要求される視線計測は, 誰が誰の顔方向に視線を向けているかという人間の目視によっても抽出できる程度の低分解能のものであるため, 比較的, 実装上の難易度は低いと思われる.

また, 視線方向と顔方向には密接な関係があるため, 顔方向を視線方向の近似として用いることも考えられる. 視線方向と顔方向の関係性については, たとえば, Stiefelhagen ら[24)] は, 4 人の会話 (参加者が机を囲む円卓型の座席配置) において, 参加者の顔方向と視線方向が水平方向に関しては 87%一致することを報告している. 近年, 画像から人物の 3 次元中の頭部位置・顔方向を推定する研究[23)] もさかんに行われており, 人物から離れた場所に設置されたカメラから, 会話を阻害することなく頭部位置・顔方向を実時間で計測できる水準まで技術的に到達している. このことから, 3 人, 4 人という比較的少人数の対話であれば, 視線を顔方向として近似することで, 本論文の提案方法は比較的容易に実装することができると期待される[☆].

一方, 今回の実験では, 各参加者のバストショットおよび全体ショットをそれぞれ別個のカメラで撮影している. 提案方法をより簡易に実装するためには, 設置カメラの台数が少ないことが望ましい. そのためには, たとえば, Cutler ら[2)] の “Distributed Meetings” のように, 1 台の全方位カメラを用いることなどが考えられる. 全方位カメラにより対話参加者全員を撮影した映像から, 参加者の頭部の位置情報に基づいて, 各参加者の映っている領域を切り出すことにより比較的容易に各人物のバスとショット, ならびに, 全体ショットを得ることができると考えられる.

5.2 対話全体への拡張性

今回の実験では, 発話の形式として,「話し手」が特定の 1 人に話しかけているもののみを扱った. これ以外にも 3 人以上の複数人対話における発話には, (1)「話し手」が自分以外の参加者全員に話しかけているもの, (2) 複数の人物が同時に発話しているもの, などがある. (1) については, 本実験と同規模の実験を行った結果, 提案方法は他の映像表現と同程度の評価であった. (2) については, 実験を行っていないものの, 提案方法では, 他者から最も注目され, より重要な発言を行っていると目される「話し手」の映像が選択・提示されるため, 視聴者の対話の理解を損ねる映像は生成されないと思われる. 以上のことから, これらの発話の形式を含む対話全体においても, 提案方法は他の映像表現よりも「対話の方向性」および「相手の反応」を正確かつ分かりやすく伝達できることが期待できる. 今後は, 上記の発話の形式などにおける視線の特性をより詳細に分析しその特定を提案方法に加味するとともに, 評価実験によりその効果を検証することが課題としてあげられる. また, 今回は, 4.1 節で述べた理由から, 短時間的な映像区間のみを評価対象としたが, 対話全体の効果を実際に検証するために, 対話全体の映像を用いた評価実験を行う必要がある.

5.3 対話人数の拡張性

今回の実験では, 3 人から 5 人の複数人会話を対象とした. 提案方法は, 対話参加者の視線の多数決に基づいているため, 原理的には, 対話人数に依存せず適用可能であることが期待できる. しかしながら, 対話人数が多いほど, 参加者が密集した状況になる. そのため, 多様な人物配置への対応 (後述), より高精度な視線計測, 対話人数の増加にともなうカメラ台数の抑制, などが課題になる.

5.4 人物配置の拡張性

今回の実験では, 会議参加者を半円上に並べたが現実の会議では, 多様な参加者の配置が考えられる. 一般に設置位置が大幅に異なるカメラの映像でスイッチングすると, 視点が急激に変化して参加者の位置関係や方向感覚を失い, 視聴者がどこからどの対象を見ているかを認識できないという問題が生じることが知ら

☆ 筆者らは, 現在, 画像認識に基づく頭部位置および顔方向推定技術[23)] を用いた自動映像編集システムの構築・実験を進めている[25)].

れている[26)]．そのため，本実験では，イマジナリーライン（人物と人物を結んだ線）を挟んですべてのカメラを同じ側に配置することで，参加者の位置関係の混乱の要因を排除し，純粋に映像切替えによる効果のみの評価を行うことをねらいとして実験を行った．提案方法では，対話参加者の視線の集中の推移に基づくため，原理的には対話参加者の位置関係に依存せず機能することが可能である．しかしながら，様々な人物配置をとる実際の会議においては，このような位置関係の混乱を視聴者に与えないようなカメラ配置や映像表現を検討する必要がある．その1つとしては，井上ら[27)] のように，参加者の位置関係を明確にする映画の映像理論を取り入れることなどが考えられる．

5.5 対話内容の評価

今回の実験では，短時間的な映像を用いて，映像を介し被験者に伝達される情報の中で，「対話の方向性」と「相手の反応」という限定的な要素のみを評価した．次の段階として，ある人物の発話と別の人物の反応が連続的に連なった対話全体の映像を用いて，提案方法により対話内容が分かりやすく被験者に伝達できているかどうかについて検証する必要がある．そのために必要な評価尺度の検討も今後の課題である．

6. ま と め

本論文では，複数人の対話を撮影対象として，「対話の方向性」と「相手の反応」を正確かつ分かりやすく視聴者に伝えるための新しいアプローチとして，対話参加者の視線に基づく映像切替え方法を提案した．この提案方法は対話の進行とともに最も多くの対話参加者が視線を向けている人物の映像を選択するものである．まず，提案方法によりどのような映像が生成されるのかを確認するために，3人から5人の複数人対話を収録し，視線の集中の対象の推移を分析した．その結果，提案方法により「話し手」および「相手」の映像が交互に選択・提示されることが分かり，本方法によって「対話の方向性」および「相手の反応」が伝達される可能性を確認した．次に，提案方法により編集された映像を実際に視聴した際に，これらの情報が伝達されるかどうかを検証するために，全体画面，マルチ画面，音声に基づく映像切替えの3種類の映像表現を比較対象として被験者実験を行い，正確性と明瞭性の2つの観点から評価を行った．その結果，提案方法により編集された映像は，比較対象の映像表現と比べ，「対話の方向性」と「相手の反応」をより正確かつ分かりやすく視聴者に伝達できることが示唆された．以上のことから，複数人物の対話を対象とした映像編集方法として，視線に基づくアプローチが有効であることを確認した．また，実際の会議への適用に向けた研究の課題と方向性を示した．

謝辞 日頃，ご指導いただく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牧野昭二部長，および，同研究部の皆様に深く感謝いたします．また，本研究に関する議論をしていただいたメディア情報研究部メディア認識研究グループの大和淳司氏，大野健彦氏（現NTTサイバーソリューション研究所），深山篤氏（現NTTサービスインテグレーション基盤研究所），の諸氏に深く感謝いたします．

参 考 文 献

1) Lee, D.-S., Erol, B., Graham, J., Hull, J.J. and Murata, N.: Portable Meeting Recorder, *Proc. 10th ACM International Conference on Multimedia*, pp.493–502 (2002).
2) Cutler, R., Rui, Y., Gupta, A., Gadiz, J., Tashev, I., Wei He, L., Colburn, A., Zhang, Z. and Silverberg, S.: Distributed Meetings: A Meeting Capture and Broadcasting System, *Proc. 10th ACM International Conference on Multimedia*, pp.503–512 (2002).
3) 徳　勲，友保康成，渋谷　雄，田村　博：テレビ会議技術の課題と利用法についての考察，*Proc. 8th Symposium on Human Interface*, pp.207–212 (1993).
4) 井上智雄，岡田謙一，松下　温：TV番組のカメラワークに基づいたTV会議システム，情報処理学会論文誌，Vol.37, No.11, pp.2095–2104 (1996).
5) スティーブン・D・キャッツ，津谷祐司（訳）：映画監督術—SHOT BY SHOT，フィルムアート社 (1996).
6) 森田敏夫，伊藤清次：TV番組の製作技術—基礎からノウハウまで，兼六館出版 (1991).
7) 富野由悠季：映像の原則—ビギナーからプロまでのコンテ主義，キネマ旬報社 (2002).
8) ダニエル・アリホン：映画の文法—実作品にみる撮影と編集の技法，紀伊國屋書店 (1976).
9) （協）日本映画・TV編集協会（編）：図解映像編集の秘訣—映画とTV番組コマーシャルから学ぶ映像テクニックのすべて，玄光社 (1999).
10) Glenny, M. and Tayler, R. (Eds): *S.M. Eisenstein Selected Works Volume 2, Towards a Theory of Montage*, British Film Institute (1991).
11) Kraft, R.: The Role of Cutting in the Evaluation and Retention of Film, *J. of Experimental Psychology: Learning, Memory, and Cognition*, Vol.12, No.1, pp.155–163 (1986).
12) 井上智雄，岡田謙一，松下　温：テレビ会議に

おける映像表現の利用とその影響，情報処理学会論文誌，Vol.40, No.10, pp.3752–3761 (1999).
13) 加藤大一郎，石川秋男，津田貴生，福島 宏，山田光穂：カメラワーク分析と映像の主観評価実験，映像情報メディア学会誌，Vol.53, No.9, pp.1315–1324 (1999).
14) Reeves, B. and Nass, C.: *The Media Equation*, CSLI Publication (1996).
15) Huston, A. and Wright, J.C.: Child Processing of Television: The Informative Function of Formal Features, *J. of Communication*, Vol.31, No.3, pp.32–48 (1981).
16) Kendon, A.: Some Function of Gaze-Direction in Social Interaction, *Act. Psychologica*, Vol.26, pp.22–63 (1967).
17) 松尾太加志：コミュニケーションの心理学—認知心理学・社会心理学・認知工学からのアプローチ，ナカニシヤ出版 (1999).
18) 榎本美香，伝 康晴：3 人会話における参与役割の交替に関わる非言語行動の分析，人工知能学会研究会資料，SIG-SLUD-A301, pp.25–30 (2003).
19) 大坊郁夫：しぐさのコミュニケーション—人は親しみをどう伝え合うか，サイエンス社 (1998).
20) Exline, R.V.: Exploration in the Process of Person Perception: Visual Interaction in Relation to Competition, Sex and Need for Affiliation, *J. of Personality*, Vol.31, pp.1–20 (1963).
21) 大野健彦，武川直樹，吉川 厚：2 点補正による簡易キャリブレーションを実現した視線測定システム，情報処理学会論文誌，Vol.44, No.4, pp.1136–1149 (2003).
22) 松本吉央，Alexander Zelinsky，小笠原司：ステレオビジョンを用いた顔・視線計測システムの開発，画像の理解・認識シンポジウム（MIRU2000）講演論文集，Vol.II, pp.217–222 (2000).
23) Morency, L.-P., Rahimi, A., and Darrell, T.: Adaptive View-Based Appearance Model, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.803–810 (2003).
24) Stiefelhagen, R. and Zhu, J.: Head Orientation and Gaze Direction in Meetings, *Ext. Abstracts CHI 2002*, pp.858–859 (2002).
25) Takemae, Y., Otsuka, K. and Yamato J.: Automatic Video Editing System Using Stereo-Based Head Tracking for Multiparty Conversation, *Ext. Abstract CHI 2005*, pp.1817–1820 (2005).
26) 宮崎英明，亀田能成，美濃導彦：複数のカメラを用いた複数ユーザに対する講義の実時間映像化手法，電子情報通信学会論文誌，Vol.J85-D-I, No.7, pp.1598–1605 (1999).
27) 井上亮文，吉田竜二，平石絢子，岡田謙一，松下 温：映画の映像理論に基づく対話会議シーンの自動撮影手法，情報処理学会論文誌，Vol.45, No.1, pp.212–221 (2004).

(平成 16 年 7 月 15 日受付)
(平成 17 年 5 月 9 日採録)

竹前 嘉修

平成 11 年慶應義塾大学理工学部電気工学科卒業．平成 13 年同大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．平成 13〜16 年 NTT コミュニケーション科学基礎研究所．現在，NTT サイバーソリューション研究所．映像編集，非言語コミュニケーションに興味を持つ．平成 14 年度電子情報通信学会学術奨励賞受賞．電子情報通信学会会員．

大塚 和弘（正会員）

平成 5 年横浜国立大学工学部電子情報工学科卒業．平成 7 年同大学院工学研究科博士課程前期修了．同年日本電信電話（株）入社．現在，NTT コミュニケーション科学基礎研究所研究主任．コンピュータビジョン，時系列画像解析，コミュニケーションシーンの分析に興味を持つ．第 55 回全国大会大会優秀賞，平成 9 年度電子情報通信学会学術奨励賞，IAPR 10th Int. Conf. Image Analysis and Processing Best Paper Award 各賞受賞．電子情報通信学会会員．

武川 直樹（正会員）

昭和 49 年早稲田大学理工学部電子通信工学科卒業．昭和 51 年同大学院修士課程修了．同年日本電信電話公社（現 NTT）入社．平成 6〜12 年（株）NTT データ，平成 12〜14 年 NTT コミュニケーション科学基礎研究所．この間，画像符号化，画像処理，画像認識，コンピュータビジョン，ロボットビジョン，ヒューマンインタフェースの研究に従事．平成 15 年より，東京電機大学教授．ノンバーバルコミュニケーション，ヒューマンコンピュータインタラクション，映像コミュニケーション，画像・映像解析の研究に興味を持つ．工学博士．IEEE，ACM，電子情報通信学会各会員．