思考状態と発話停止点を利用した会議の動画ダイジェスト生成支援

宮田章裕[†] 林剛史[†] 福井健太郎[†]
重野寛[†] 岡田謙一[†]

我々は，思考状態および発話停止点を利用した会議の動画ダイジェスト自動生成を提案する．映像中のシーン変化やカメラワーク，テロップを利用する従来手法では，日常的な会議映像のダイジェストを生成することは困難であった．そこで，本手法では,「思考」と「発話」という情報を利用して日常的な会議映像のダイジェストを生成する．具体的には，脳波情報を利用して思考状態を MS-Level という独自の指標で数値化し，発話停止点を利用して素材映像から単位映像への分割を行う．そして，単位映像の中から MS-Level が高いものを抽出・連結する．プロトタイプシステムを利用した評価実験では，比較システムより的確なダイジェストを生成できることを確認した．

Conference Movie Summarization Assistance Using Mental States and Speech Breakpoints

Akihiro Miyata,[†] Takefumi Hayashi,[†] Kentaro Fukui,[†]
Hiroshi Shigeno[†] and Ken-ichi Okada[†]

In this paper, we propose a method to summarize movies of a conference automatically by the use of participants' mental states and speech breakpoints. Currently, many research groups have developed techniques for video summarization, but these ways are not qualified for editing conference movies. To address this issue, we propose a method to summarize conference movies by the use of participants' mental states and speech breakpoints. We define MS-Level (Mental State Level) derived from one's EEGs as an indicator of mental states. Video Segmentation is conducted by detecting speech breakpoints of participants. After segmentation, high MS-Level scenes are extracted and constitute a digest movie. We ran experiments to evaluate our proposition using a prototype system, and conclude that our proposal will contribute to a better summarization of conference movies.

1. はじめに

本手法の目的は,「思考」と「発話」という情報を利用して日常的な会議映像のダイジェストを生成することである．これまでに様々な動画ダイジェスト生成手法が提案されているが，会議映像を対象としたとき，これらの手法は適切でない場合が多い．なぜならば，日常的な会議映像はたいてい固定カメラ 1 台で撮影しただけの平坦な映像であり，シーン変化やカメラワークを動画編集のキーとして利用できないからである．日常的に繰り返される会議の映像にテロップを付与するとも考え難いので，これを利用した動画編集も不可能である．

ここで，会議の性質を考えたとき，それは「思考」と「発話」を繰り返すプロセスであるということに気付く．つまり，会議の動画ダイジェストを生成する際には,「思考」と「発話」の情報をキーにする手法が考えられる．

そこで，我々は，思考状態および発話停止点を利用した会議の動画ダイジェスト生成支援を提案する．「思考状態」は脳波情報を利用して導出した MS-Level（Mental State Level）という独自の指標で数値化する．また,「発話停止点」とは会議中に発話がまったく起こらない状態のことである．発話停止点を利用して素材映像から単位映像への分割を行い，単位映像の中から MS-Level が高いものを抽出・連結してダイジェストを生成する．

以降，2 章では会議の動画ダイジェストの必要性や研究例ついて述べ，3 章では脳波の発生原理や研究例について述べる．4 章では提案概念について述べ，5 章ではプロトタイプシステムを紹介する．6 章では予備実験，7 章では提案概念を評価するために行った本実験について述べ，最後に，8 章では結論と今後の展望

† 慶應義塾大学大学院理工学研究科
Graduate School of Science and Technology, Keio University

を述べる．

2. 会議の動画ダイジェスト

会議とは，人間が行う協調作業の代表例であり，情報交換，問題解決，相対する意見の調整などが行われる．ここで当然，会議を記録するという行為が考えられるが，議事録のようにテキストによる記録は内容が正確である反面，記録が困難な情報も多い．たとえば，会議の雰囲気や議長の表情，発言者の語調などをすべてテキストで記録することは容易ではない．この問題を解決するためには，会議の様子を動画で記録する方法が有効である．動画は，画像や音声など多くの情報を同時に，容易に記録することができる．

しかし，動画は一定の長さを持った連続的なメディアであるため，テキストなどのメディアに比べて閲覧に多くの時間がかかる．また，会議にはしばしば不毛な時間帯があり，時間と労力をかけて動画のすべてを閲覧する必要はない場合が多い．そこで，会議映像の中から重要なシーンを抽出し，つなぎ合わせてダイジェストを生成することが必要である．

これまでに，映像の色変化や動きの大きさからシーンの境界を検出する研究[1),2)]や，映像文法に基づいてカメラワークを分析することでダイジェスト生成を行う研究がある[3)]．これらの手法は，対象となる映像にシーン変化やカメラワークが十分に含まれている場合は有効である．また，映像に含まれるテロップを利用して映像の各シーンの重要度評価を行う研究もある[4)]．テロップの属性（フォント，サイズ，表示位置）には映像作成者の意図が現れている場合が多く，ニュースなどのテロップが含まれる映像が対象の場合には有効な手法である．

3. 脳　　波

脳波とは，脳の活動にともなって頭皮上に生じる電位のことであり，人間からつねに発生し続けている生体情報の 1 つである．脳波は，その周波数（f Hz）帯域から δ 波（$f < 4$ Hz），θ 波（$4\,\text{Hz} \leq f < 8\,\text{Hz}$），$\alpha$ 波（$8\,\text{Hz} \leq f \leq 13\,\text{Hz}$），$\beta$ 波（$13\,\text{Hz} < f$）の 4 つに分類することができる．α 波は，目を閉じたり，落ち着いた状態やぼんやりとした状態のときに頭頂部や後頭部で優勢となる．一般にはリラックスしているときには α 波が現れ，覚醒しているときには α 波が減衰するとされている[5)]が，α 波がまったく観測されない被験者も少なくない[6)]．β 波は，速波と呼ばれ，脳波の所見上は意識レベルの高い状態（興奮，緊張，集中など）で観察でき，前頭部で顕著に観測される[6),7)]．β 波帯付近の脳波は思考を要する作業を行うときに強く出現し，思考を要しない作業時にはあまり出現しないという報告もいくつかある[5),8)〜10)]．これを利用して，脳波情報から導出した思考状態を互いにアウェアできる遠隔コミュニケーションシステムや[11)]，被指導者の思考状態を複合現実空間上に提示する指導者支援システム[12)]が提案されている．また，ウェアラブルカメラで記録した個人体験映像をインデキシングする際に脳波情報を利用する研究[13)]や，ダイジェストを生成する際に撮影者の脳波情報をキーとして利用する研究[14)]もある．さらに，感性スペクトル解析法では，脳波情報の中から各感情に対応する特定パターンを抽出して感情解析を行っている[15)]．

4. 思考状態と発話停止点を利用した会議の動画ダイジェスト生成の提案

2 章で述べたように，ダイジェストを生成する際には，映像に含まれるシーン境界やカメラワークを抽出する手法が一般的である．しかし，日常的な会議映像は固定カメラ 1 台で撮影しただけの平坦な映像である場合が多く，シーン変化やカメラワークが元々含まれていないことが多い．カメラマンを用意するとしても，会議に関係ない人が立ち回り撮影していると議論に集中できないおそれがある．テロップを利用して各シーンの重要度評価を行う手法もあるが，会議は日常的に頻繁に繰り返されるものであり，これを撮影するたびに手間と時間をかけてテロップを埋め込むとは考え難い．

そこで，我々は，**思考状態および発話停止点を利用した会議の動画ダイジェスト生成支援を提案する．**

4.1 脳波計測による思考状態の導出

本手法では，**簡易脳波計で測定した脳波情報に基づいて思考状態を導出する．**この理由は以下のとおりである．

- 脳波はつねに発生し続けている情報である．
- 表情やジェスチャと異なり，通常は脳波を意図的に操作しないし，操作することも容易ではない．
- 簡易脳波計は現実的なコストで導入できるし，人体に悪影響を与えることもなく，計測中に会話したり動いたりすることも可能である．
- 一般的な簡易脳波計は前頭部の計測しかできないが，前頭部で顕著に検出できる β 波帯近辺の脳波[6),7)]は思考を要する作業を行うときに強く出現し，思考を要しない作業時にはあまり出現しないとされており[5),8)〜10)]，思考状態と強い相関がある指標だといえる．

簡易脳波計は IBVA Technologies 社の IBVA を利用する．この装置は頭部に小型センサを装着するだけで計測を行うことができ，脳波計から PC へのデータ転送が無線式であるので使用者は自由に動き回ることができる．なお，プライバシの観点から，脳波を計測されることに抵抗を感じる参加者もいると思われるので，システムとしてはこの点に配慮する必要がある（具体的な対策については 4.3 節参照）．

我々は，「どの程度頭を働かせているか」といった思考状態を表す指標として **MS-Level（Mental State Level）**を定義した．MS-Level の導出手順は以下のとおりである．

(1) 会議中の各参加者の脳波を周波数分解し，使用しない帯域のデータを除去する．使用する帯域に関してもノイズ除去を行う．

(2) 参加者全員の脳波強度が同じ範囲内に収まるようにスケーリングする．

(3) 思考状態を的確に表している周波数帯域のデータを平均し，これを **1 サンプルの脳波データ**と定義する．

(4) 各瞬間において，最新 N サンプルの脳波データに重み関数を掛けて加算し，その値をその瞬間の **MS-Level** と定義する．

(1) で除去する帯域は，まばたきや筋電などのノイズが非常に多い低周波数帯（0–2 Hz）と，脳波自体が比較的微弱な高周波数帯（40–60 Hz）である．また，残った帯域に関しても，筋電などにより瞬間的に非常に大きなノイズが入るという脳波の性質を考慮して，数学的に外れ値となるようなデータをノイズと見なして除去する．

(2) では，会議中の脳波強度の最低値と最高値が全参加者の間で等しくなるように脳波データのスケーリングを行う．脳波の絶対的な強度には個人差があり，これを吸収することが目的である．

(3) では，ノイズが多い帯域などを除去した残りの脳波データの中から，思考状態を的確に表している $F_{low}-F_{high}$ Hz のデータを抽出し，この区間のデータを平均したものを **1 サンプルの脳波データ**と定義する．なお，3 章で述べたように，β 波は思考状態と密接な関係があると認知されているが，β 波と呼ばれる帯域はかなり広く，思考を要する作業でも内容の違いによって優勢になる帯域が若干異なることを経験上確認している．さらに，会議という作業に特化して脳波変化を調査した事例は少なく，会議中の思考状態を的確に表しているのはどの帯域なのか，過去の知見だけからは判断しかねる．このため，我々の以前の研究[16]を行った時点では，利用する β 波の周波数帯を一部に限定するのではなく，一般に β 波と呼ばれている帯域（13–40 Hz）をすべて含む 12–40 Hz の帯域を利用する手法をとっていた．しかし，この広い帯域の中には思考を的確に表していない帯域が含まれている可能性がある．

そこで，今回我々は，会議を構成している代表的な要素を想定したタスクを被験者に課し，各タスク中に測定した脳波データを分析することで，的確に思考状態を表している帯域 $F_{low}-F_{high}$ Hz を導出することにする（詳細は 6.1 節参照）．

(4) では，最新 N サンプルの脳波データを利用して思考状態を導出する．このように，現在だけではなく過去のサンプルも利用する理由は，思考は時間的に離散したものではなく時間幅を持ったプロセスであると思われるからである．また，ここまでで取り除けなかった瞬間的なノイズの重みを減らすという意味もある．我々の以前の研究[16]でも最新数サンプルのデータを取得し，その中で一定の閾値を超えたものの割合で思考状態を表現していた．しかし，閾値さえ超えていれば強度が大きいサンプルも小さいサンプルも同一視していたので，思考の変化を的確に表せていたとはいい難い．

そこで，今回我々は，最新 N サンプルの脳波データの各強度に，新しいデータほど重みが大きくなるような重み関数を掛けて加算し，これをその瞬間の **MS-Level** と定義する．以下に MS-Level の導出方法を示す．

$$MS{-}Level(x)=\sum_{i=0}^{N-1}(N-i)Power(x-i)$$

この際，各パラメータは以下のものを表している．

- N：使用する脳波データの最新サンプル数
- x：脳波のサンプル ID
- $Power(x)$：ID が x の脳波サンプルの強度
- $MS{-}Level(x)$：脳波のサンプル ID が x のときの MS-Level

なお，N の最適値は作業内容によって大きく変動すると思われる．たとえば，自動車を運転するときのように，秒単位で重要なイベント（信号の色が変わる，歩行者が飛び出してくる，など）が発生する作業であれば，N を小さくして思考を短いプロセスととらえた方が都合がよいだろう．逆に，人生を記録した映像の各シーンを評価する場合[13]は，N を大きくして数分から数時間前の思考の影響を考慮する必要があるかもしれない．

つまり，N は一意に決定するのではなく，目的に応じて適切な値を導出するべきである．我々の目的は会議映像のダイジェストを生成することなので，N もそれに適した値を利用する（詳細は 6.3 節参照）．このとき，各瞬間の MS-Level を導出する際に過去サンプルを利用することで,「驚き」や「閃き」などのような瞬間的な思考の変化がとらえ難くなる可能性があるが，今回はある程度長期的な思考状態の変化のみを取り扱うことにした．

脳波から思考状態の推定を行う関連研究として，相澤らは，ウェアラブルカメラを着用して日常生活を送り，その中で記録した個人体験映像をインデキシングする際に，映像と同時記録した脳波情報を利用して個人の主観を反映させている[13]．この手法では，長い間，興奮，注意，集中の状態にあると持続的に β 波が現れる現象を利用している．中村らは，ダイジェストを生成する際に映像撮影者の脳波情報をキーとして利用している[14]．この手法でも，撮影者の β 波が増加しているときは撮影者が集中・興奮していると判定され，その時間帯の映像がリプレイ映像の候補となる．

このように，β 波の出現を検知して興奮や注意，集中といった状態を判定する研究はいくつかあり，この点からも脳波（特に β 波帯域）を利用して思考状態を推定する本研究の手法は妥当性があるといえる．

4.2 発話停止点を利用した映像分割

本手法では，**発話停止点で映像分割を行う．**具体的には，会議参加者が誰も発言していない時間帯（沈黙時間帯）が T sec を超えた場合は，発話が停止した（発話停止点）と判定して映像を分割する．分割後に残った各映像を「単位映像」と定義する．T sec を超える長さを持つ沈黙時間帯はどの単位映像にも含まれることはなく，ダイジェスト映像の候補からも外れることとなる．このような方法で映像分割を行う理由は以下のとおりである．

- カメラワークやテロップなどとは異なり，発話の有無はどんな会議映像からも必ず検出できる指標である．
- 発話の有無だけを評価するので，言語・性別・年齢に非依存である．
- 発話停止点で映像分割を行うので，誰かが発言している最中で映像が分割されるのを防ぐことができる．
- 会議は論理的に高度なコミュニケーションであり，発言内容を言語認識やニューラルネットワークで解析すると処理が複雑になるばかりでなく，どうしても判定に不正確さが生じる．その点，あくまで「発話の有無」だけで映像分割を行うこの方法は処理が単純であるし，正確である．
- 実際の会議では沈黙していても視線や雰囲気を察し合ったり,「沈黙の駆け引き」が行われたりするので，沈黙時間帯も無価値ではない．しかし，会議映像として後から参照した場合，これらのノンバーバル情報を映像から把握することは困難であるので，長い沈黙時間帯はダイジェストの候補から外しても問題がないといえる．

4.3 ダイジェスト生成の方法

本手法では，**MS-Level の平均値が閾値を超える単位映像を結合してダイジェストを生成する．**「MS-Level の平均値」とは，各単位映像区間における参加者全員の MS-Level の平均値である．閾値は任意に設定可能であり，閾値が高くなるほど選択される単位映像が少なくなり，圧縮率の高いダイジェストが生成される．

なお，特定の話題や話者に対する関心・無関心が他人に公開されると人間関係に悪影響を及ぼしかねないなど，プライバシ上の問題でユーザに受け入れられない可能性がある．そのため，システムとしては，個人の MS-Level は本人しか参照できないようにする．つまりこの場合は，自分の MS-Level だけに基づいたエゴセントリックなダイジェストが生成されることになる．ただし，これだけではシステムとして不十分なので，会議参加者全員の MS-Level を平均して匿名性を高めたものならば閲覧できるようにし，全員の MS-Level に基づいたダイジェストも閲覧できるようにしている．このように，個人のプライバシを考慮することが，本提案における我々のデザインコンセプトである．

また，この方法は単位映像の意味内容を直接的に評価していないので，会議の冒頭・終了シーンがダイジェストに含まれない可能性がある．通常，会議の冒頭シーンには「メンバの紹介」や「議題の発表」，終了シーンには「内容のまとめ」や「結論」があり，会議の内容を把握するためには必須のシーンといえる．そこで，**MS-Level の値にかかわらず，会議の冒頭・終了シーンは必ずダイジェストに含むようにした．**

このようにして生成されたダイジェストは，思考状態という観点から会議映像を振り返ることができるので参加者にとって有益である．たとえば,「よく頭が働いていたシーン」など，通常の議事録やダイジェストからは見つけ出すことが難しいようなシーンを見つけ出すことが可能である．

5. プロトタイプの実装

5.1 会議の記録

会議中の参加者の脳波は IBVA によって測定され，各時間帯における脳波情報が PC に無線で送信されて記録される．データ速度は約 0.87 samples/sec である．

会議中の発言はマイクによって録音され，各時間帯における発話の有無が PC に送信されて記録される．精密な音声処理技術は本手法の対象外であるため，今回は簡易な発話検出システムの実装にとどめ，音声認識による発話者特定や厳密なノイズ除去などは行っていない．そのため，一部で正確に発話や沈黙が検出できない場面も生じたが，これはすべて手動で修正を行った．

会議中の映像はカメラによって撮影され，MPEG 形式で PC に記録される．人手を掛けてカメラワークを行ってもよいが，基本的にカメラは会議参加者全員が映る位置に固定したまま撮影を行う．

5.2 ダイジェストの生成

ダイジェストの生成・閲覧手順は以下のとおりである．

(1) 会議映像を読み込む．
(2) 沈黙時間帯の閾値を指定する．
(3) MS-Level の閾値を指定する．
(4) ダイジェストを再生する．

(1) では，映像データ全体を読み込む．ダイジェスト生成前の映像をすべて閲覧することも可能である．

(2) では，発話データを読み込んだ後，沈黙時間帯の閾値 T sec を指定する．4.2 節で述べたように，この閾値を超える沈黙時間帯で映像が分割され，単位映像が定義される．

(3) では，各時間帯における全参加者の MS-Level の平均値を読み込んだ後，閾値を指定する．4.3 節で述べたように，区間中の MS-Level の平均値がこの閾値を超える単位映像が結合されて，ダイジェストが生成されることになる．

(4) では，**図 1** に示すようなインタフェースを利用してダイジェストを再生する．A の部分で点灯している区間の 1 つ 1 つがダイジェストに採用された単位映像である．B の部分は映像の再生位置を示すシークバーであり，A が点灯している時間帯だけが再生される仕組みになっている．A が点灯していない時間帯はダイジェストに含まれておらず，シークバーが自動的にジャンプして映像は再生されない．(2) と (3) の手順を繰り返して各閾値を変更するたびにダイジェストに採用される区間が再計算され，その結果は A の部分に点灯する区間として表現される．また，C の部分で映像の再生・停止などを行う．

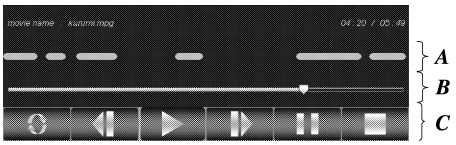

図 1 ダイジェスト閲覧コントローラ
Fig. 1 The digest controller.

なお，会議中の発話が少なくダイジェストが短かすぎるような場合には，ダイジェスト閲覧者に「たいして議論が行われなかった」という印象を与えかねない．しかし，ダイジェスト映像とダイジェスト閲覧インタフェースはつねに同時に表示されているため，仮にダイジェスト映像が短かすぎた場合でも，A が点灯していない部分が多いことを視認すれば「発話は少ないが何らかの議論があった可能性」に気付くことはできるし，A が点灯していない部分の様子が知りたければ，簡単な操作でその時間帯の映像を閲覧することができる．あるいは，(2) の操作で「発話がない」と判定される沈黙時間帯の閾値を大きくすることで，「沈黙」と判定される時間帯を減らすことができるので，沈黙が頻繁に起こる会議でもダイジェスト生成の際に映像がむやみにカットされるのを防ぐことができる．

6. 予 備 実 験

ここまでで未決定であった各パラメータの最適値を導出するために，以下のような予備実験を行った．

なお，未知の部分が多い脳波という情報を扱うため，極力多くの被験者に対して実験を行い，その結果予備実験 1 では 60 人，予備実験 2 では 18 人，予備実験 3 では 17 人の被験者数となった．予備実験 1 は脳波の基本的な性質を調査するためのものであり，今後の研究に生かすという観点からも特に多くの被験者に対して行っている．各実験で被験者数が異なっているが，各実験間（7 章の本実験を含めて）で結果を直接比較するわけではないため，大きな問題はないと思われる．

6.1 予備実験 1：使用する脳波帯域の決定

6.1.1 実 験 内 容

この実験の目的は，MS-Level を導出する際に利用する脳波の周波数帯域 $F_{low} - F_{high}$ Hz を決定することである．この実験では，会議を構成している代表的な要素を想定したタスクを被験者に課し，各タスク中に測定した脳波データを分析することで，的確に思考状態を表している帯域を決定する．会議の構成要素を想定した「論理思考」,「質疑応答」,「会話」,「リラック

表 1 各タスクの内容
Table 1 The subject tasks.

タスク名	内容
論理思考	論理問題（与えられた条件を整理して解を求める問題）を記述式で解く．
質疑応答	一般常識を問う知識問題に口頭で回答する．
会話	3 人で相談しながら論理問題（「論理思考」と同じ形式）を解く．
リラックス	何も考えず安静を保つ．発言や動作をいっさい行わない．

表 2 「最も頭を使ったタスク」のアンケート結果（n = 60）
Table 2 The questionnaire results (n = 60).

	論理思考	質疑応答	会話	リラックス
回答数（人）	26	19	13	2

ス」の各タスクを被験者 **60 人**（19–25 歳の学生，男 57 人，女 3 人）に課した．なお，各タスクの内容は **表 1** のとおりである．

6.1.2 実 験 結 果

実験終了直後に，被験者全員に対して「最も頭を使ったタスク」を問うアンケートを実施したところ，**表 2** のように「論理思考（26 人）」が最も多く，「リラックス（2 人）」が最も少ない回答であった．そこで，MS-Level を導出する際には，脳波強度が「論理思考」中に大きくなり，「リラックス」中に小さくなるような周波数帯域を利用するのが妥当であるといえる．ただし，0–2 Hz は眼球運動によるノイズが非常に多く，40 Hz 以上は脳波自体が微弱であったため，今回は利用しない．また，12 Hz 以下は「リラックス」時に強度が大きくなっているため（「リラックス時には α 波（8–12 Hz）が現れる」という報告[5] と一致），この帯域も利用しない．そもそも，頭頂部および後頭部で優勢な α 波を前頭部に装着する IBVA で計測することは精度に不安が残るし，α 波自体が計測されない人も少なくない[6]．残りの帯域について，「論理思考」時の脳波強度を調べると**図 2** のようになる．横軸の 1 つ 1 つが各周波数帯域を示しており，たとえば，「12–14」は「12 Hz から 14 Hz までの帯域」という意味である．図 2 を見ると，「論理思考」時の脳波強度が最大になるのは「12–28」，つまり，12 Hz から 28 Hz までの周波数帯域を利用した場合であるので，$\boldsymbol{F_{low}} = 12$, $\boldsymbol{F_{high}} = 28$ となる．なお，脳波強度は被験者ごとに 0–1 の範囲にスケーリングされているので単位はない．また，調査は 12–40 Hz の全帯域について行っており，図 2 に示されているのはその一部である．

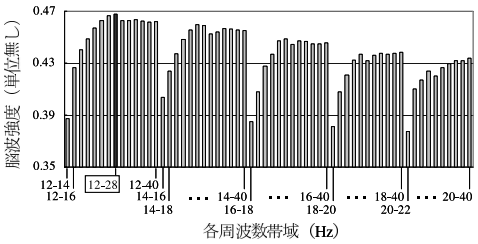

図 2 「論理思考」時の脳波強度
Fig. 2 EEG intensities of “logical thinking”.

6.2 予備実験 2：自然な映像分割を行うための沈黙時間帯閾値の決定

6.2.1 実 験 内 容

この実験の目的は，沈黙時間帯の閾値 T sec の最小値を決定することである．4.2 節でも述べたように，沈黙時間帯が T sec を超えた場合は発話停止点と判定され，映像が分割されることになる．この閾値は任意に設定可能であるが，あまり小さすぎると発言中に含まれるわずかな無音時間帯で映像が分割されてしまい，視聴者に不自然な印象を与えてしまうおそれがある．そこで，閾値の最小値を決定するために，どのくらい沈黙時間が続けば，場面を分割しても視聴者に不自然な印象を与えないか調査するためのタスクを被験者 **18 人**（21–27 歳の学生・会社員，男 16 人，女 2 人）に課した．具体的なタスクの内容は以下のとおりである．

(1) 2 文が連続して読み上げられるシーン（音声のみ）を 3 つ提示する．その際，音声ファイルを編集して 2 文の間隔を様々に変えて提示した．なお，各シーンは過去に行われた実際の会議記録から抽出したものであり，すべて 2 文の間で話題転換が行われている．

(2) 2 文の間で話題転換があるという情報を与えたうえで，被験者に「これ以上の沈黙間隔があればシーン分割してもよい」と思えるパターンを回答してもらう．

6.2.2 実 験 結 果

3 シーンすべてに対して回答を集計したところ，平均で **1.85 sec** の沈黙時間があればシーンを分割しても不自然ではないという結果を得た．

6.3 予備実験 3：使用する脳波データの最新サンプル数の決定

6.3.1 実 験 内 容

この実験の目的は，MS-Level 導出時に使用する脳波データの最新サンプル数 N を決定することである．4.1 節で述べたように，N は会議映像のダイジェスト生成に適した値である必要がある．そこで今回は，ダ

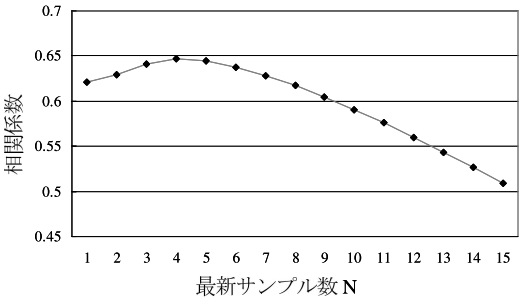

図 3 正解判定と MS-Level 判定の相関係数
Fig. 3 Correlation factors between correct judgements and MS-Level judgements.

イジェスト生成で最も重要な工程といえる「シーンの重要度判定」において，**正解判定**と **MS-Level 判定**を比較し，両者が最も似通った結果になるような N を最適値と判定する．その際，正解判定は人手（被験者）で生成することにした．これまでに様々なダイジェスト生成方法が提案されているが，やはり，人間が手間と時間をかけて行った判定が「正解」に一番近いといえるからである．

正解判定を生成するために，著者ら **3 人**が **10 分間**会議（創造会議）を行う会議映像を用意し，**1.85 sec**（6.2 節参照）以上沈黙が続く部分で我々が映像を分割した．この結果，10 分の会議映像から **30 個**の単位映像が定義され，各シーンの長さは平均**約 16.45 sec** となった．そして，被験者 **17 人**（21 – 24 歳の学生，男 17 人）に会議映像を見せ，各単位映像の重要度を **0 – 3 の 4 段階**で判定してもらった．最後に，30 個の単位映像それぞれに対して，被験者 17 人が判定した重要度の平均値を求め，これを正解判定とした．

また，それぞれの単位映像に対して，MS-Level の平均値を利用して重要度（**0 – 3**）を導出し，これを MS-Level 判定とした．なお，MS-Level の平均値とは各単位映像中における会議参加者全員の平均であり，30 個の単位映像の中で MS-Level の平均値が最大のものは「重要度 3」，最低のものは「重要度 0」のように判定した．

6.3.2 実験結果

N の値を 1 – 15 まで変化させたところ，**図 3** のように $\boldsymbol{N = 4}$ のときに正解判定と MS-Level 判定の相関係数が最大になった．なお，4 サンプルの脳波データは約 3.5 sec 分に相当する．

7. 本実験：提案手法の有用性の評価

7.1 実験内容

この実験の目的は，本提案で生成したダイジェストの有用性を評価することである．評価の方法として，6.1 – 6.3 節で導出したパラメータを利用して生成した MS-Level 判定が，被験者が 6.3 節と同様の方法で生成した正解判定にどれだけ近いか調査した．なお，予備実験 3 と本実験では被験者の一部が重複しているため，会議映像は予備実験 3 で使用したものと同形式であるが内容が異なるもの（創造会議）を使用した．これは **32 個**の単位映像（平均**約 16.37 sec**）に分割されている．また，予備実験と同様，極力多くの被験者を集めるという方針で実験を行い，結果的に正解判定の生成を依頼した被験者は **25 人**（21 – 48 歳の学生・研究員，男 22 人，女 3 人）であった．

さらに，MS-Level 以外の異なる判定方法との比較が必要と判断したため，MS-Level 判定の代替手法として**発話密度判定**も行った．これは，単位映像長に対する発話時間の割合を**発話密度**と定義して，発話密度が最大のものは「重要度 3」，最低のものは「重要度 0」というように判定を行う．複数の会議参加者が同時に発話しているときは，発話している人数に比例して密度が高くなる．なお，発話密度で判定を行う単位映像は MS-Level で判定したものとまったく同じである．この方法を比較対象として選んだ理由は，発話密度判定が最も合理的かつ一般的な手法と思われるからである．つまり，会議中において「発言数が多いシーン」や「同時発話が多く起こるシーン」を議論が活発になっているシーンととらえることは合理的であるし，発話情報はカメラワークなどと違いどんな会議映像にも含まれているので，一般的な会議映像に広く利用することができる．今回は発話量の集計を厳密に手動で行ったため，発話検出のミスはない．

なお，本実験で利用した会議映像を解析したところ，議長とその他 2 人のメンバの MS-Level が最大値の 50%を超えている時間帯の割合はそれぞれ 32.6%，35.0%，35.5%で，発話している時間帯の割合は 42.2%，9.1%，27.7%，発話数（相槌などを除いた一連の発話の数．発話が途切れなくても，明らかに話題転換が行われていれば 2 つの発話として数えた）は 42 回，9 回，13 回となった．また，全員の MS-Level の平均値が最大値の 50%を超えていた時間帯は 26.4%，誰か 1 人でも発話していた時間帯は 68.9%，複数の人が同時に発話していた時間帯は 9.6%であった．同形式の会議を別途 4 回行い，同様に解析して平均値を求めたところ，議長とその他 2 人のメンバの MS-Level が最大値の 50%を超えている時間帯はそれぞれ 34.2%，32.0%，29.8%で，発話している時間帯は 50.3%，12.7%，24.5%，発話数は 39.2 回，13.2 回，

11.9 回であり，全員の MS-Level の平均値が最大値の 50%を超えていた時間帯は 28.1%，誰か 1 人でも「発話」していた時間帯は 71.9%，複数の人が同時に「発話」していた時間帯は 11.9%であった．この結果は本実験で用いた会議映像と大きな差がないため，本実験で利用した会議映像が特別に偏った特徴を持つものではないことが分かる．

7.2 実験結果と考察

実験の結果，各方法による判定結果の相関分布は**図 4**（図中の直線は近似曲線）のようになり，正解判定と MS-Level 判定の相関係数は **0.558**，正解判定と発話密度判定の相関係数は **0.387** となった．つまり，発話密度判定よりも，MS-Level 判定の方が正解判定と似通っているので，**本提案による MS-Level 判定の方が，発話密度判定より有用であるといえる．**

一例として，32 個の単位映像のうち重要度が上位の **10 個**を抽出・連結したダイジェストで検証したところ，正解判定と MS-Level 判定では **81.3%**（26/32 個），正解判定と発話密度判定では **56.3%**（18/32 個）の単位映像でダイジェストへの「採用・不採用」の判定が一致した．この結果からも，MS-Level 判定の方が正解判定に似通っていることが改めて確認できる．

このような結果になった理由は，MS-Level の方が発話密度よりも，会議の各シーンの重要度と密接な関係があるからと思われる．実際，正解判定を行った被験者に対してヒアリングを行ったところ，発話密度が低くても MS-Level が高いシーンは「発言数は多くないが有意義な会話をしている」と判定し，発話密度が高くても MS-Level が低いシーンは「発言数は多いが無駄話である」と判定しているケースが多いことが分かった．

また，MS-Level 判定の精度を下げている要因を調査したところ，「MS-Level は高いがほとんど発言がないシーン」を重要度が高いと判定していることが主な原因であった．ダイジェスト生成を目的とした場合，発話がほとんどないシーンは情報量が少ないので重要度は低いといえるし，被験者の多くもそのような判定を行った．つまり，ダイジェストにこのようなシーンが含まれていると，視聴者は違和感を覚える可能性がある．このような場合，視聴者は簡単な操作で「発話がない」と判定される沈黙時間帯の閾値を小さくすることができ，発言数が少ないシーンをダイジェストから外すことが可能である．また，この問題を根本的に解決するためには，MS-Level 単独でシーンの重要度判定を行うのではなく，「MS-Level が高くても発話密度が低ければ重要度は低い」といったように，MS-Level と発話密度を連携させて判定を行うべきであり，これは今後の検討課題である．

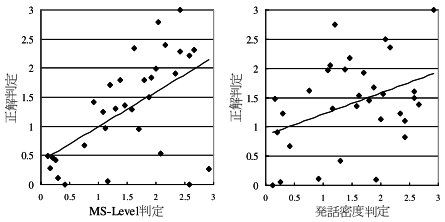

図 4 各方法による判定結果の相関分布
Fig. 4 Correlated distributions of judgement results.

8. 結　　論

我々は，思考状態および発話停止点を利用した会議の動画ダイジェスト自動生成を提案した．「思考状態」は脳波情報を利用して導出した MS-Level という独自の指標で数値化した．ダイジェストを生成する際には，発話停止点を利用して素材映像から単位映像への分割を行い，単位映像の中から MS-Level が高いものを抽出・連結する手法をとった．プロトタイプシステムを利用した評価実験では，比較システムより高い精度で「正解」に近いダイジェストを自動生成できた．よって，我々の提案方式は会議映像のダイジェスト生成に有用であったといえる．

とはいえ，7.2 節で述べたように，正解判定と MS-Level 判定の相関係数が 0.558 となっており，まだ実用に耐えうる精度ではない．そこで今後は，単位映像の重要度判定にも発話情報を利用するなど，脳波と発話情報の連携も視野に入れつつ，より精度の高いダイジェスト映像生成方法を検討していく方針である．また，本手法では脳波情報を会議の非同期解析に利用したが，リアルタイムコミュニケーションシステムや自己フィードバックシステムなど，様々な分野に脳波情報を利用する研究も現在行っている．

参 考 文 献

1) DeMenthon, Kobla and Doermann: Video Summarization by Curve Simplification, *Proc. 6th ACM International Conference on Multimedia*, pp.211–218 (1998).
2) 鈴木，中嶋，坂野，三部，大塚：動き方向ヒストグラム特徴を用いた映像データからのカット点検出法，電子情報通信学会論文誌，No.4, pp.468–478 (2003).
3) 天野，上原，熊野，有木，下條，春藤，塚田：映像文法に基づく映像編集支援システム，情報処理学会論文誌，Vol.44, No.3, pp.915–924 (2003).

4) Kuwano, Taniguchi, Arai, Mori, Kurakake and Kojima: Telop-on-demand: video structuring and retrieval based on text recognition, *ICME 2000*, pp.759–762 (2000).
5) 宮田：現代心理学シリーズ 2 脳と心，培風館 (1996).
6) 小杉，武者：電子情報通信工学シリーズ生体情報工学，森北出版株式会社 (2000).
7) 加藤，大久保：初学者のための生体機能の測り方，日本出版サービス (1999).
8) Giannitrapani: The Role of 13-Hz Activity in Mentation, *The EEG of Mental Activities*, pp.149–152 (1988).
9) Rasey, Lubar, McIntyre, Zoffuto and Abbott: EEG Biofeedback for the Enhancement of Attentional Processing in Normal College Students, *Journal of Neurotherapy*, Vol.1, No.3, pp.15–21 (1998).
10) 八木：現代心理学シリーズ 6 知覚と認知，培風館 (1996).
11) Fukui, Miyata and Okada: Implementation of Avatar Mediated Communication Environment with Thinking Awareness, *SCIS & ISIS2004 THE-7*, pp.116–120 (2004).
12) 宮田，宮狭，本田，岡田：脳波情報および複合現実感を利用した指導者支援の提案，*DICOMO 2004*, pp.587–590 (2004).
13) 相澤，石島，椎名：ウェアラブル映像の構造化と要約：個人の主観を考慮した要約生成の試み，電子情報通信学会論文誌，No.6, pp.807–815 (2003).
14) 中村，市村，岡田，松下：撮影グループの生体反応を相互利用した映像コンテンツ作成，*DICOMO 2004*, pp.373–376 (2004).
15) Musha, Terasaki, Haque and Ivanitsky: Feature Extraction from EEG Associated with Emotions, *Artif Life Robotics*, pp.15–19 (1997).
16) 宮田，福井，本田，重野，岡田：会議を撮影した動画メディアの思考状態インデキシングの提案，情報処理学会論文誌，Vol.45, No.11, pp.2509–2518 (2004).

(平成 17 年 2 月 28 日受付)
(平成 17 年 12 月 2 日採録)

宮田　章裕（学生会員）

2003 年慶應義塾大学理工学部情報工学科卒業．2005 年同大学大学院理工学研究科開放環境科学専攻修士課程修了．現在，グループウェア等の研究に従事．2004 年 DICOMO シンポジウム最優秀プレゼンテーション賞受賞．

林　　剛史（学生会員）

2005 年慶應義塾大学理工学部卒業．現在，同大学大学院理工学研究科開放環境科学専攻修士課程に在学中．グループウェア等の研究に従事．

福井健太郎（正会員）

2001 年慶應義塾大学理工学部情報工学科卒業．2003 年同大学大学院理工学研究科修士課程修了．現在，同大学院理工学研究科後期博士課程に在学中．グループウェア等の研究に従事．2004 年 IEEE Computer Society Best Paper Award 受賞．

重野　　寛（正会員）

1990 年慶應義塾大学理工学部計測工学科卒業．1997 年同大学大学院理工学研究科博士課程修了．1998 年同大学理工学部情報工学科助手（有期）．現在，同大学理工学部情報工学科助教授．工学博士．計算機ネットワーク・プロトコル，モバイル・コンピューティング，マルチメディア・アプリケーション等の研究に従事．著書『〜ネットワーク・ユーザのための〜無線 LAN 技術講座』（ソフト・リサーチ・センター），『コンピュータネットワーク』（オーム社）等．電子情報通信学会，IEEE，ACM 各会員．

岡田　謙一（フェロー）

慶應義塾大学理工学部情報工学科教授，工学博士．専門は，CSCW，グループウェア，HCI．情報処理学会誌編集主査，論文誌編集主査，GW 研究会主査等を歴任．現在，MBL 研究会運営委員，BCC 研究グループ幹事，日本 VR 学会 CS 研究会委員長．情報処理学会論文賞（1996 年，2001 年），情報処理学会 40 周年記念論文賞，日本 VR 学会サイバースペース研究賞，IEEE SAINT'04 最優秀論文賞を受賞．情報処理学会フェロー，IEEE，ACM，電子情報通信学会，人工知能学会会員．