装着型センサを用いた会議ログの構造化システム

大西 鮎美[1]　村尾 和哉[2]　寺田 努[3,4]　塚本 昌彦[4]

概要：　会議や報告会など複数人が会話する場において，会話内容に対する各参加者の興味度や理解度の推定や，場の盛上りや長考状態の検出を行い，それらの情報を録音音声や録画映像と組み合わせることで，各参加者の役割や貢献度を推定し，重要な箇所を自動抽出できるようになる．また本研究では，加速度センサを用いて会議における参加者の動作を検出して会議全体の状況を認識し，タグ付けを行うシステムを提案する．提案システムでは，会議参加者全員の頭部に装着した加速度センサを用いて各参加者の発話やうなずきなどの頭部動作を検出する．提案手法による発話検出およびうなずく，左右を見る，首をかしげるの 3 種類の頭部動作検出の精度評価の結果，発話の認識精度は 44.4%，うなずく動作の認識精度は 38.9%，左右を見る動作の認識精度は 39.4%であった．また，本研究では，加速度および角速度センサを用いて会議参加者の発話, および 3 種類の動作を認識して録画映像にタグ付けを行うアプリケーションである Meeting Review Tree (MRT) を構築した．提案システムでは，ミーティングの構造を会議中における報告者の遷移を表す第一層，報告中での発話者の交代情報を表す第二層，うなずきや細かな発話情報を表す第三層に階層化してタグ付けする．評価実験の結果，階層化した第一層の認識精度は 57.0%，第二層は 61.0%であった．

1. はじめに

組織やグループに所属する人間にとって，会議は決定事項の伝達，プロジェクトの方針の決定，新たなアイディアの創出，情報の共有などのための重要な場である．会議での決定事項や議論した内容，出された意見などは重要な情報であり，参加者が後日参照したり，欠席者と情報共有するために議事録を作成することが一般的である．しかし，全ての発言の中から重要な箇所を取り出しつつ，それに対する参加者の様子を記録してまとめ上げる作業には多大な労力と時間を要する．これまでに，議事録を自動的に生成するシステムとして VoiceGraphy[1]，AmiVoice 議事録作成支援システム [2]，衆議院の音声認識による議事録のシステム [3] などが提案されている．

しかし，これらはいずれも会議におけるすべての会話を記録して音声認識によって文字に起こしているだけであり，多くの賛成が得られた発言や反論の多かった発言など会議において重要であると考えられる発言に注目して振り返るには生成された議事録を見直さなければならず，時間や手間がかかる．会議における重要な発言を抽出するためには，発言者以外の賛成や反論を認識し，複数人での会話における状況認識を行い，会議全体の状況を判断する必要がある．複数人での会話における状況認識に関する研究は数多く行われている [5][6][7] が，うなずきの検出などにとどまっており状況認識を行うまでには至っていない．会話状況を認識するにはうなずきの対象や程度，否定や疑問などを表現する首ふり，前のめりになって話を聞いていない姿勢などを検出する必要があると考えられる．

本研究では，会議参加者全員の頭部に装着した加速度センサを用いて会議における複数人の行動を検出して会議全体の状況を認識し，タグ付けを行うシステムを提案する．

本論文の構成を以下に述べる．2 章で関連研究を紹介し，3 章では提案システムについて述べる．4 章で評価実験を行い，5 章で提案システムを用いたアプリケーションについて説明し，最後に 6 章で本論文をまとめる．

2. 関連研究

2.1 会話中の動作認識

会議などの複数人数での会話時におけるユーザの行動認識に関する研究はこれまでに数多く行われている．Morency らは画像認識によってうなずきの検出を行い，ロボットなどに非言語行動を認識させる研究を行っている [5]．この

1 神戸大学工学部
School of Engineering, Kobe University
2 立命館大学情報理工学部
School of Information Science and Engineering, Ritsumeikan University
3 科学技術振興機構さきがけ
PRESTO, Japan Science and Technology Agency
4 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

研究では，頭部動作の認識のために一人につき一台以上のカメラで動画を撮る必要がある．また，カメラの向きも重要であるため，会議の場所の変更への対応が難しい．加速度センサを用いてうなずきの自動検出を行っている研究として，山本らはうなずきを興味の指標として会話や会議の重要な部分の抽出を行うことを目的とし，装着型センサをヘッドセットに取り付け，センサデータからうなずきの検出を行っている [6]．この研究では会話におけるうなずきの意味を詳細に調査しているが，会議における重要箇所の抽出や会議全体の状況認識は行っていない．Wohler らは，二人での会話中のうなずき，首振り，首かしげ，注目という動作をモーションセンサを用いて認識している [7]．しかし，この研究では個人の動作のみを認識しており，会話の状況については検証していない．河原の研究では，ポスターセッションを kinect とマイクロフォンアレイを用いて収録し，深度情報，音声情報，画像情報から視線と相槌を検出し，発話の予測を行っている [12]．しかし，この研究ではマイクロフォンアレイと kinect およびカメラ 6 台を設置する必要があり，会議の場所の変更に対応することが難しい．

角らは IMADE 環境 [9] を用いて，被験者の動作，視線，音声，生体データなどの会話中のさまざまな情報を採取している．IMADE 環境を用いると会議の記録や分析が詳細に行えるが，これらのセンサを会議参加者すべてに装着させるにはコストがかかる．また複数のカメラを用いるため，会議の場所の変更などにも対応できない．本論文で提案する会議ログ取得システムは，教育現場や自然発生的な会議での利用を想定しているため，場所に依存しないように，参加者に小型装着型センサのみを取りつけて，装着型の加速度センサの値のみから動作判定を行う．

2.2 会話構造の解析

会話中の個人の情報を組み合わせて，誰から誰へ向けられた話であるかを特定するなどの会話構造の分析に関する研究も数多く行われている．会話構造の分析として，高梨は多職種ミーティングにおける懸念導入の分析を行っている [8]．この研究では，ミーティング時の「気になる/する」というフレーズに着目して，このフレーズが話題導入としての役割を果たしているとし，関心の度合いを定量化する指標を提案している．しかし，フレーズの検出は手作業で行っている．

井上らは状況説明の会話において，ジェスチャが会話の組織化にどのように関わっているのかを IMADE 環境下で測定して一人の発話者の説明が完了しないうちに異なる人物が説明の続き（説明の引き取り）を行っている場合の会話構造を分析している [10]．しかし，音声データと画像データを手動で解析しているため，会話構造の自動検出はできず手間が大きい．また，発話者と異なる人物が会話内容を引き取る契機やその際のジェスチャについて考察しているが，評価を行っていないため会話構造との関連性は不明である．岡田らの研究では，IMADE 環境下でモーションキャプチャシステムとアイマークレコーダ，加速度センサを用いてうなずきと腕のジェスチャを計測して説明者と聞き手の非言語パターンを自動抽出し，会話間の動作のない部分を言いよどみと定めることで，言いよどみの前後の会話構造を分析している [11]．しかし，うなずき以外の動作は認識しておらず，より詳細に会話構造を解析するためには，複数の動作を対象とする必要がある．また，斎賀らのうなずきの自動検出による会話制御機能の分析 [15] では装着型センサを用いて行動認識を行っているが，うなずきなどの動作の検出にとどまっており，状況認識を行うまでには至っていない．

中田らの研究ではうなずきに加えて，ポスターセッションによく用いられる指さし動作と発話やあいずちの関係から会話構造の分析を行っている [13]．この研究では，加速度データからうなずきを自動検出しているが，発話区間の判定や指さし動作の検出は手作業で行われている．熊野らの研究では，4 人でのミーティング時の動画から笑顔と微笑みを認識し，誰が誰に笑いかけているのかなどの 4 人の関係性を分析している [14]．しかし，この研究では，認識する動作が笑顔と微笑みのみであり，高精度な認識を行うためには，各参加者の正面にカメラを設置する必要があり，手間やコストがかかる．斎賀らは加速度センサを用いて会話中の話者および聞き手のうなずきの検出，話者の視線計測およびうなずきの機能の分析を行っている [15]．しかし，対象とする動作は視線注視とうなずきのみであり，他の動作は認識していない．また，うなずきの認識は行っているが，それを利用した集団の状況認識は行っていない．大塚らは複数人物の対話中の画像中の人物の頭部追跡に基づいた視線に着目し，会話構造のモデル化を行っている [17], [18]．しかし，会話構造の分析に複数人会話時の視線の移動のみを用いており，うなずきなどの視線移動以外の動作に着目していない．また，大塚らは動画の画像認識を用いて 4 人での会話における発話や頭部ジェスチャ，視線移動の検出を行い，誰から誰へ向けた話なのかを推定している．しかし，会話の参加人数が増えると視線の判定が難しくなることが予想されるが，他の人数での会話に対する考察は行われていない．

このように，さまざまな重要度の会話が行われる会議などにおいて議事録を作成する際，うなずきのみなどの単一の動作からその場の状況を詳細に推定することが難しい．従って，詳細な会話構造の推定には複数種類の動作を認識して動作と会話の関係を解析する必要がある．本研究で提案する会議ログのタグ付けシステムは，加速度センサと角速度センサから個人の発話状態や頭部動作を認識して得られた行動情報を複数人分組み合わせて解析することで，発

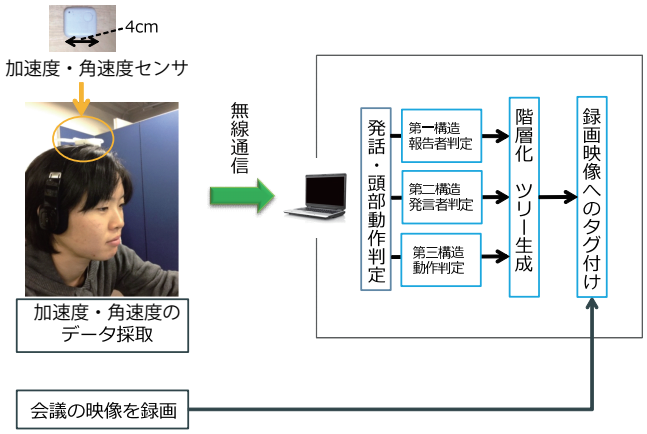

図 1　システム構成

言者および発言に対する周囲の動作が活発であるような会話の重要箇所のタグ付けを行う．

3. 提案システム

本章では会議の録音音声や録画映像にタグを付与するシステムについて述べる．提案システムは頭部に装着した加速度・角速度センサを用いて頭部の動作を認識し，認識結果に応じて録音音声や録画映像にタグ付けを行う．

3.1 想定環境

本論文は参加者が一人ずつ順番に報告をしていく形式のミーティングを想定しており，このような形式のミーティングを報告型ミーティングと定義する．参加者は複数人の報告者および進行役や書記，アドバイザを想定しているが，報告者以外の参加は必須ではないとする．報告型ミーティングにおいて最も重要な箇所は，各参加者が報告中に発話している部分である．音声認識によって発話と発話者を認識するアプローチが考えられるが，十分な精度での発話者の認識には 1 回の発言が十分な長さである必要であり [19]，短い発言が入り乱れた状況では認識が困難であると考える．そこで，本研究では参加者の頭部に装着した加速度センサと角速度センサを用いて頭部の動きを検出し発話部分を推定する．また，発話に対する他の参加者の振る舞いも重要であるため，同意と否定・疑問の 2 種類の状況を想定し，これらの状況を表現する動作として，うなずく動作と首をかしげる動作の 2 種類を検出する．さらに，参加者はミーティング全体の様子を確認する際に左右を見るため，左右を見る動作の検出も行う．

3.2 システム構成

提案システムの構成を**図 1** に示す．提案システムは参加者の頭部に装着された 3 軸無線加速度・角速度センサから各参加者の PC へ送信してデータを採取し，発話検出および頭部動作検出を行う．また，参加者の正面からの映像を録画しており，録画映像に対してタグ付けを行う．使用

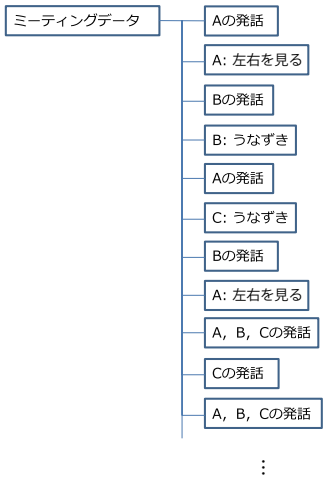

図 2　単一の時系列上に付与されたタグ

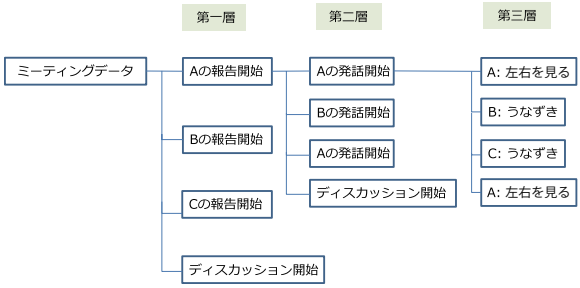

図 3　提案手法で付与する階層化されたタグ

した加速度・角速度センサはワイヤレステクノロジー社製 WAA-006[20] で，サンプリング周波数は 50Hz である．本論文ではヘッドセットにセンサを取り付けているが，将来的には眼鏡に装着したり，耳に挟むことを想定している．

検出した発話や頭部動作を録画映像にタグ付けする方法を述べる．**図 2** に示すように検出結果を単一の時系列上に並べてタグ付けを行うだけでは全体像としての大まかなミーティングの流れが分からないため，システムの利用者がミーティングの状況を想像することが難しい．また，タグが多すぎて録画映像から自分の見たい箇所を探す作業に時間がかかる．そこで，場面の把握を容易にするため，**図 3** に示すようにミーティングの構造を階層化する．第一層として A の報告→ B の報告→ C の報告→全員によるディスカッション→終了という大きなミーティングの流れをタグ付けし，第二層として A の報告の区間内での A の発話→ B の発話→全員でのディスカッションというように，ある報告者の報告中の他の参加者の発話をタグ付けする．さらに第三層として A の発話中の各参加者のうなずきなどの頭部動作 をタグ付けする．第三層を詳細に解析することで，議論の重要性や参加者の興味度，参加者の役割や議論に対する貢献度などを推定できると考えており，今後取り組む予定である．本研究では，アドバイザや進行役は会議終了後に手動で設定するものとする．

表 1　予備実験の環境

日付	10/2	10/8	10/17	10/25	10/31	11/07	11/14	11/25	12/19	1/10	1/17
参加人数	5	6	7	7	7	5	5	7	6	3	6
時間	2:00	1:50	1:50	1:30	1:30	2:00	1:40	1:50	1:50	2:00	1:20

図 4　ミーティングの様子

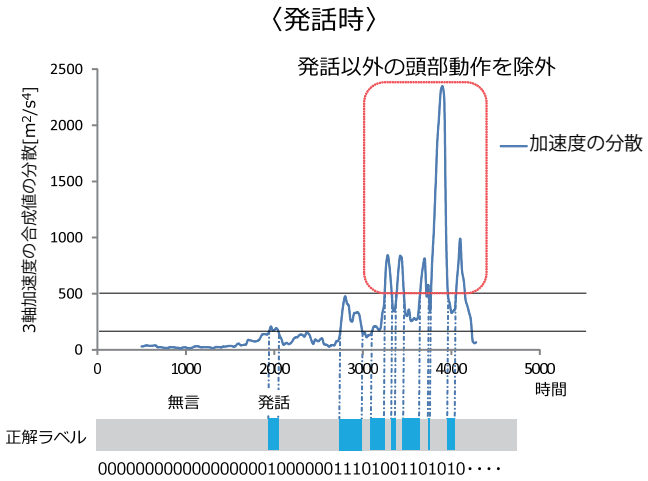

図 5　無言状態と発話状態の加速度の合成値の分散値波形

3.3　提案手法

3.3.1　予備実験

ミーティング中に行われるうなずく，首をかしげる，左右を見るの 3 種類の頭部動作および発話を検出するアルゴリズムを構築するために，データ採取を目的とした予備実験を行った．アドバイザ 1 名，進行役 1 名を含む最大 7 名が参加する約 2 時間のミーティングを 11 回実施し，各参加者の加速度データ，角速度データ，正面ビデオ映像，および会議全体の様子の確認のために 360 度カメラのビデオ映像を採取した．なお，それぞれの実験時の日付，参加人数およびミーティングの実施時間を**表 1**，ミーティングの様子を**図 4** に示す．ミーティングではアドバイザおよび進行役以外が 1 名ずつ順に報告を行った．報告者の数は最少 3 名，最多 7 名であった．なお，表 1 における参加者とは，ミーティングにおいて加速度・角速度データを取得できた人数を示しており，実際の参加者よりも少ない場合がある．

3.3.2　発話の判定方法

図 5 は静止状態と発話状態の加速度の合成値の分散値の波形を示したものである．分散の計算に用いるウィンドウサイズは 200 ミリ秒とした．静止状態では参加者は何もせず座っており，加速度に大きな変化が見られなかった．一方，発話状態では被験者は手ぶりなどのジェスチャも交えながら話すため，加速度は発話中長時間にわたって大きく変化した．これらの波形における一定時間の安定と変化から静止状態，発話状態を判別できると考え，以下の 3 種類の特徴量を用いて発話判定を行い，それぞれについて認識率を評価した．

- 手法 1　X，Y，Z 軸加速度をそれぞれ accX，accY，accZ とする．3 軸加速度の合成値 $G_{acc} = \sqrt{accX^2 + accY^2 + accZ^2}$ の過去 200 ミリ秒の分散値 V_{acc} が閾値を超えたときを発話状態，分散値が閾値を下回ったときを無言状態と判定する．発話以外の頭部動作を除外するため，閾値に上限を設けている．

$$if\ (V_{acc} \leq 150\ \&\&\ V_{acc} \geq 500) \rightarrow Speaking \quad (1)$$

- 手法 2　X，Y，Z 軸角速度をそれぞれ angX，angY，angZ とする．3 軸角速度の合成値 $G_{ang} = \sqrt{angX^2 + angY^2 + angZ^2}$ の過去 200 ミリ秒の分散値 V_{ang} を閾値に用い，手法 1 と同様に発話判定を行う．発話以外の頭部動作を除外するため，閾値に上限を設けている．

$$if\ (V_{ang} \leq 200\ \&\&\ V_{ang} \geq 500) \rightarrow Speaking \quad (2)$$

- 手法 3　加速度と角速度の合成値を両方閾値として用いて発話判定を行う．手法 1，手法 2 と同様に，発話以外の頭部動作を除外するため，閾値に上限を設けている．

$$if\ ((V_{acc} \leq 150\ \&\&\ V_{acc} \geq 500)\ \&\& \\ (V_{ang} \leq 200\ \&\&\ V_{ang} \geq 500)) \rightarrow Speaking \quad (3)$$

これらの手法については評価実験で比較を行った．また，発話の判定結果から無言状態を 0，発話状態を 1 とラベリングを行った．なお，動作判定は発話判定とは別処理として並行して行うこととする．

次に，参加者それぞれについて行ったラベリング結果の過去 10 秒間の値を加算していき，加算結果の平均値 L_{ave} が閾値を超えた人物を報告型のミーティングの第一層における話者とし発話の開始部分にタグをつける．発話ラベリングの加算結果の平均値が閾値を超えたときに話者と判定し，他の人物が閾値を超えるまでをその人物の話者区間とする．

$$if\ (L_{ave} \geq 0.35) \rightarrow Speaker \quad (4)$$

話者と判定された人物の発話ラベリングの加算結果が閾

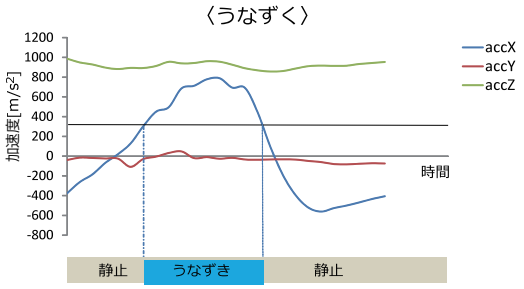

図 6　うなずく動作の判定手法

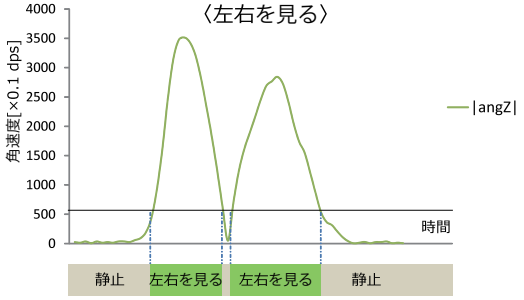

図 7　左右を見る動作の判定手法

値を下回った場合の話者区間終了の判定は行っていない．これは，会議の議事録は会議内容のはじまりの部分から再生し，どこで再生を終了するかどうかは視聴者が判断すると考えたためである．また，会議参加者全員の発話ラベリングの加算結果が閾値を超えている区間はディスカッションと判定した．このように，第一層の大まかなミーティングの流れを検出し参加者の報告開始のタグ付けを行った後，第一層の区間を第二層としてさらに発話ごとに分割する．話者判定と同様に，1 サンプルごとの各参加者の発話判定からラベリング結果が 1 のときは発話とカウントし，区間内で発話のカウント数が閾値を超えた人物を検出する．検出された人物のラベリング結果において加算を行う区間を短くすることで，さらに細かい変化を検出できる．したがって，過去 2 秒間でさらに階層化し，発話ラベリング結果の 0，1 の値を加算していくことで，その値が閾値を超えた参加者を閾値を超えている区間の話者とする．

$$if\ (L_{ave} \geq 0.7) \rightarrow Speaker \tag{5}$$

3.3.3　動作の判定方法

動作判定ではうなずく，左右を見る，首をかしげる頭部動作および静止状態の検出を行う．予備実験において，うなずく，左右を見る，首をかしげる動作を行っているとき，ミーティング参加者には**図 6**，**図 7**，**図 8** に示すような加速度・角速度の波形が見られた．これらの波形をもとに，各動作を検出する条件を決定する．うなずく動作を行うと，首を前向きに屈めることによって図 6 のうなずきの波形に示すように地面と並行で正面前向きの加速度（x 軸）の値が大きくなる．そのため，x 軸の加速度が閾値を超えて再び閾値以下となるまでの区間をうなずく動作の区

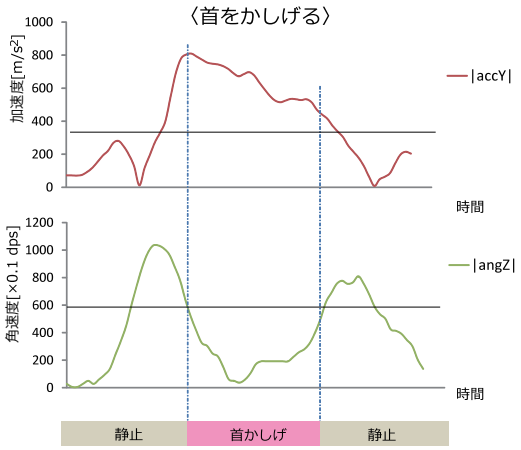

図 8　首をかしげる動作の判定手法

間と判定する．左右を見る動作を行うと，首が左右に回転するため図 7 の左右を見る波形に示すように地面と垂直方向 z 軸を回転軸とした角速度の絶対値が変化する．そのため，z 軸を回転軸とする角速度の値が閾値を超えてから再び閾値以下となるまでの区間を左右を見る動作として検出する．左右を見る動作に関しては，横を向いてから正面に顔の向きを戻す際に一度閾値を下回り，再度閾値を超えるが，会話中など一度顔を横に向けてから一定時間そのまま横向きの状態を保持し，正面を向くなどのパターンがあるため，1 回の首振りごとに検出することとした．首をかしげる動作を行うと，首が左右に傾くため，図 8 の首をかしげる動作の波形に示すように地面と平行で左右方向の加速度（y 軸）の絶対値が大きくなる．しかしこの条件だけでは左右を見る動作と判別が困難な場合があるため，左右を見る動作と異なり首は回転させないため，地面に垂直な角速度（z 軸）の絶対値が閾値以下の条件も同時に満たす区間を首をかしげる動作として検出する．以下に頭部動作の判定条件を示す．

$$if\ (accX \geq 300\ ||\ angY \geq 300) \rightarrow Nod \tag{6}$$

$$\begin{aligned} &else\ if\ ((accY1 \leq -400\ ||\ accY1 \geq 400)\ \&\& \\ &(angZ1 \geq -500\ ||\ angZ1 \leq 500)) \\ &\rightarrow Look\ around \end{aligned} \tag{7}$$

$$\begin{aligned} &else\ if\ (angZ1 \leq -500\ ||\ angZ1 \geq 500) \\ &\rightarrow Put\ head\ on\ one\ side \end{aligned} \tag{8}$$

4.　評価実験

4.1　評価環境

提案手法による発話検出および頭部動作検出の精度を評価するために，4 名が向かい合う形で机の周りに着席し，一人ずつあらかじめ決めておいたテーマについて A，B，C の順に約 3 分程度ずつ報告を行い，その後役分程度全員でディスカッションを行う形式で実験を行った．なお，被験

表 2　発話認識率

被験者	加速度			角速度			加速度＋角速度		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
A	0.210	0.600	0.310	0.746	0.507	0.604	0.197	0.634	0.300
B	0.217	0.440	0.290	0.847	0.394	0.538	0.212	0.449	0.289
C	0.227	0.622	0.332	0.440	0.306	0.361	0.099	0.574	0.168
D	0.153	0.309	0.107	0.620	0.173	0.271	0.101	0.336	0.155

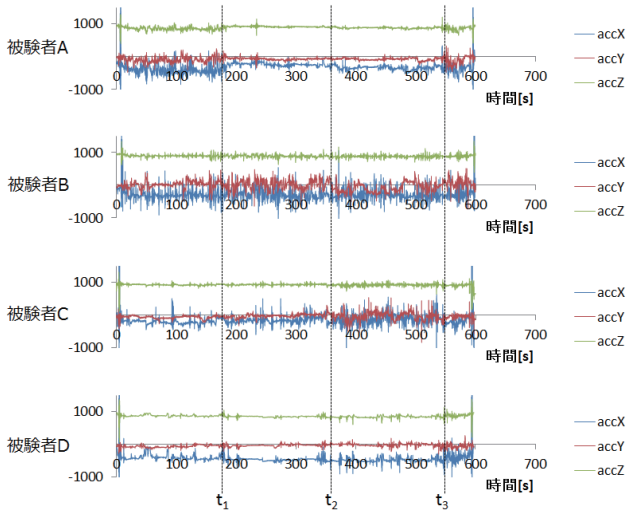

図 9　ミーティング中の加速度

者 D はこの実験における進行役である．各被験者は 3.3.1 節の予備実験と同様に頭部に加速度・角速度センサを装着してデータを採取した．また机の中央には正解のラベルを取得するために 360 度カメラをおいて動画を記録した．提案システムによって発話検出および頭部動作検出を行い，再現率，適合率，F 値を算出した．

4.2　評価結果と考察

4.2.1　発話動作の検出

発話動作の検出結果について述べる．発話動作の検出手法として提案した加速度データから判定する手法，角速度データから判定する手法，加速度データと角速度データから判定する手法の検出結果の再現率，適合率，F 値を**表 2** に示す．また，各被験者の加速度および角速度データを**図 9** と**図 10** にそれぞれ示す．図 9 および図 10 中の t_1，t_2，t_3 は話者が交代した時刻である．表 2 より，発話動作を加速度データから判定する手法では，再現率が低く，0.2 前後であった．この原因として，図 9 に示すように，ある報告者が報告を行っている間も他の参加者の加速度データが大きく変化しており，誤った参加者が発話しているとタグ付けされたことが考えられる．

一方，角速度データから発話動作を判定する手法では，再現率が大幅に改善しており，表 10 からも角速度データは発話者に対応して大きく変化していることが分かる．また，加速度と角速度の両方から判定する手法は再現率が低

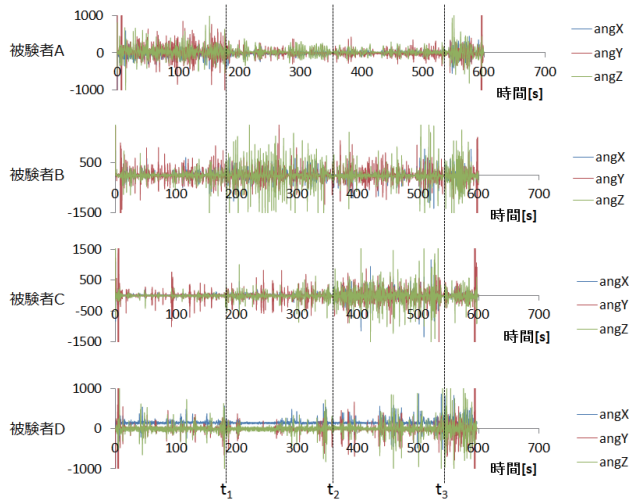

図 10　ミーティング中の角速度

表 3　話者判定の認識率評価

階層	再現率	適合率	F 値
第一層	0.63	0.52	0.57
第二層	0.55	0.7	0.61

表 4　判定結果の遅延を補正した場合の認識率

階層	再現率	適合率	F 値
第一層	0.70	0.68	0.68
第二層	0.55	0.7	0.61

下する代わりに適合率が向上している．これは，発話を検出する条件が増えて条件が厳しくなり，正しく検出されなかった発話が増えたためであるといえる．発話判定を用いた階層化において，第一層の話者と第二層の発話者を判定する際，一定時間内の発話動作の検出結果の割合を用いるため，F 値が高い方が望ましい．そのため，発話動作の判定には角速度データを用いることが有効であると考えられる．表 2 の結果より，角速度を用いたときのミーティング参加者 D の適合率が 17%と低い理由として，D は進行役であったため発言回数が少なく，ミーティング全体の様子をうかがうために左右を見るなどの加速度および角速度の分散値が大きくなる動作を発話時以外に多く行っていたことが考えられる．また，参加者 C の F 値が参加者 A，B と比較して低い理由として，報告者は自身の報告中，常に発話し続けているわけではなく，発言の途中における沈黙があったことが挙げられる．C は自身の報告中に瞬間的に何度も沈黙しており，発話から沈黙および沈黙から発話の瞬間の波形の乱れが発話の誤検出の原因であると考えられる．これらは，認識結果の平滑化を行うことで検出精度が改善すると考えている．また，静止状態から発話状態に遷移するときにウィンドウサイズによる認識遅延が起こり，判定結果と正解データの時間的ずれが認識率の低下の要因の一つと考えられる．

階層化の評価結果について，大まかな話者判定である第

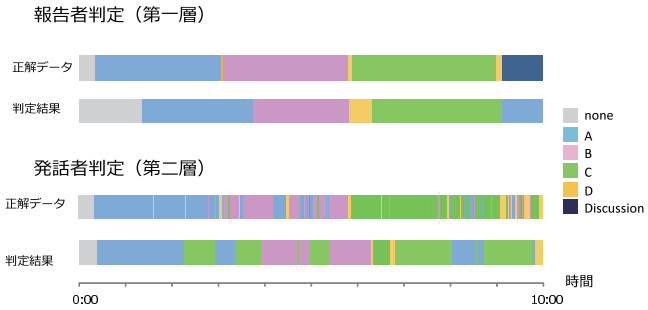

図 11　階層別ミーティング中の話者判定結果

一層と第二層の区間ごとの話者判定の再現率，適合率および F 値を**表 3** に示す．表 3 より，第一層の報告者判定の F 値は 0.57，第二層の発話者判定の F 値は 0.61 であった．認識率が低下した原因について考察するために**図 11** に第一層と第二層の判定結果を示す．図中の横軸は時間を表しており，帯の各色はその時刻の話者を表している．図 11 より，第一層の報告者判定は検出に遅れが生じたことにより，認識率が低下したと考えられる．また，最後に参加者 A，B，C，D の四人が同時に発話となるディスカッションを行ったが，提案手法では A の報告と判定された．提案手法は会議参加者全員の発話ラベリングの加算結果が閾値を超えている区間をディスカッションと判定し，一人でも閾値を超えず発話状態でない場合，複数人からの発言があったとしても参加者のいずれか 1 人の報告と判定するため，ディスカッションの際，発話頻度が高く，発話時に動作を交えていた A が報告者と判定されたと考えられる．

第二層の細かな発話の判定結果について，正解データの約 4 分経過時点から B の発話が多くなっているが，これに対し，判定結果では約 30 秒の判定の遅れが生じていた．このような判定の遅れは同様に正解データの 6 分経過時点での C の判定でも見られる．話者判定には発話判定の過去 10 秒間のラベリング結果を用いていることから，正確な検出には発話が始まってからある程度の時間を要するため，判定結果に遅れが生じたと考えられる．第一層のタグ付けを行う際にはこの遅延を考慮して差し引くことで補正することができる．しかし，第二層では，10 分のミーティングの始まりの部分について，A の報告開始はほとんど遅れは生じず，発話開始が一致している．この違いが生じた理由について，第二層は，第一層より閾値を低く設定しており，ミーティングのはじめは特徴量として用いている角速度の合成値の分散が 4 人とも小さい状況で A のみが話し始めたため，判定の遅れはほとんど見られず正確であったが，複数人数が会話に参加し，複数人数の角速度の合成値の分散が閾値に近い値となった場合に報告者の誤判定および判定の遅れが生じていると考えられる．また，第二層については，正解データを見ると正解データの方が細かく話者が交代している．これは，第二層の判定に発話ラベルの 2 秒間の加算結果を用いており，細かな発話の移り変わりが認識

表 5　頭部動作の認識結果

被験者	頭部動作	再現率	適合率	F 値
A	うなずく	0.974	0.151	0.262
	左右を見る	0.731	0.250	0.388
	首をかしげる	-	-	-
B	うなずく	1.00	0.300	0.460
	左右を見る	0.968	0.129	0.228
	首をかしげる	-	-	-
C	うなずく	0.991	0.257	0.408
	左右を見る	0.946	0.250	0.396
	首をかしげる	-	-	-
D	うなずく	0.911	0.275	0.427
	左右を見る	1.00	0.391	0.563
	首をかしげる	-	-	-

できないためである．しかし，発話ラベルの加算結果をさらに短い時間にするという方法では，うなずく動作や左右を見る動作が誤検出されると考えられる．

次に，第一層と第二層の判定の遅延を補正するため，第一層は 10 秒，第二層は 2 秒の時間を判定結果から差し引いて再評価を行った．差し引いた時間は，それぞれ判定に用いたウィンドウサイズ分である．再評価結果を**表 4** に示す．表 4 より，第一層は，再現率，適合率，F 値ともに補正前より高くなったが，第二層は，認識率に変化が見られなかった．第二層は，判定結果より正解データの方が細かく話者が交代しており，第一層より誤判定が多いため，ウィンドウサイズ分の遅延の補正が，認識率改善につながらなかったと考えられる．

4.2.2　頭部動作の検出

頭部動作の認識結果を**表 5** に示す．まず，首をかしげる動作については，10 分の評価実験内では動作が行われていなかった．これはミーティングの内容が関係していると考えられるため，今後さまざまな内容のミーティングを行い検証していく必要がある．すべての被験者においてうなずく動作と左右を見る動作の再現率は高いが適合率が低くなっている．これは，閾値が低く，実際には動作した数以上の検出結果が現れたためであるが，再現率と適合率にはトレードオフがあるため，認識結果の利用目的によって閾値を調整する必要がある．本研究では，対象としている議事録の性質上，重要な部分が取り出されない可能性が高いが議事録を早く閲覧できるケースより，議事録の閲覧に時間がかかるかもしれないが重要である部分は必ず検出できる方が良いと考える．そのため，誤ったタグが多く発生するかもしれないが再現率をを高くする方が適切であると考える．

各動作について詳しく見ると，うなずきの適合率が低くなった理由として，少しうつむいた状態で止まっていたり，座りなおす動作がうなずきとして誤判定された点が考えられる．左右を見る行動については，少しでも左右に首をまわすと判定されるように閾値設定を行ったが結果的に検出

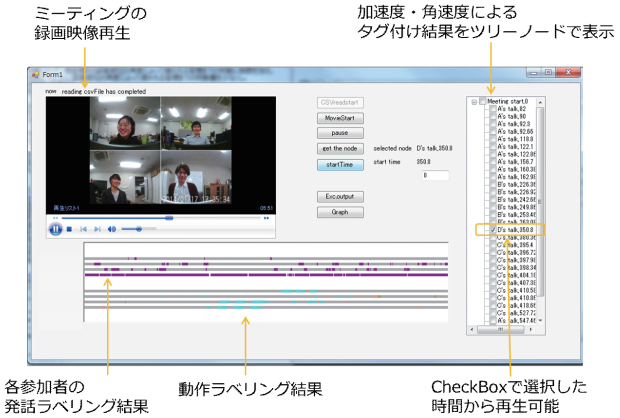

図 12 アプリケーション画面

数が多くなりすぎて適合率を下げる要因となった．また，各動作について発話検出と独立して頭部動作の検出を行ったが，発話中に左右を見つつ，うなずきながら話すなどさまざまな動作が混在しており，動作の誤判定が多くみられた．参加者についてみると，参加者 B，C，D が再現率が 0.9 を超えているのに対し，A の左右を見る動作は再現率が 0.731 と低い．これは，ミーティング中に座っている椅子が回転式であり A は頻繁に座り方を変えていたことが原因であると考えられる．

5. アプリケーション

提案手法を用いてタグ付けを行った会議映像を視聴する議事録アプリケーション Meeting Review Tree (MRT) を構築した．MRT のスクリーンショットを**図 12** に示す．MRT で会議参加者全員の加速度・角速度データおよび録画映像を読み込むことで，提案手法によって録画映像に 3 層のタグ付けが行われ，各層をツリー化して閲覧できる．MRT 利用者はツリーの第一層から視聴したい報告者の部分を選択し，その中での特定の参加者の発言やうなずきの多い箇所など再生したい箇所のチェックボックスを選択することで，該当箇所の録画映像を視聴できる．また，その前後の各参加者の発話タグと頭部動作の認識結果を確認できる．さらに，各参加者の発話のラベリング結果と動作のラベリング結果はグラフで表示されており，ユーザはこのグラフから再生箇所を選択することも可能である．

6. まとめ

本研究では，加速度および角速度センサを用いて会議参加者の発話動作やうなずく，左右を見る，首をかしげる動作を認識して録画映像にタグ付けを行うシステムを提案した．評価実験の結果より，発話の認識精度（F 値）は 0.44，うなずく動作の認識精度（F 値）は 0.39，左右を見る動作の認識精度（F 値）は 0.39 であった．また，階層化した第一層の認識精度（F 値）は 0.57，第二層は 0.61 であった．提案システムでは，認識結果をもとに主となる話者，その間の発話者，参加者の頭部動作の 3 層でのタグ付けを行い，会議映像の振り返りが容易となるように支援している．また，提案手法を実装してタグ付けされた会議映像を視聴できる議事録アプリケーション Meeting Review Tree(MRT) を構築した．今後の課題として，発話動作および頭部動作の認識精度の向上が挙げられる．また，発話動作や頭部動作から参加者の会話内容の理解度や関心度を推定する手法を提案し，会議映像視聴者にとって重要である箇所を自動抽出して提示するシステムを構築する予定である．

謝辞 本研究の一部は，科学技術振興機構戦略的創造研究推進事業 (さきがけ) および文部科学省科学研究費補助金挑戦的萌芽研究 (24650565) によるものである．ここに記して謝意を表す．

参考文献

[1] VoiceGraphy, available from ⟨ http://jpn.nec.com/voicegraphy/.⟩.
[2] Advanced Media Inc., available from ⟨http://www.advanced-media.co.jp/solution/proceeding/rewriter_kaigi.html/.⟩.
[3] 河原達也: 議会の会議録作成のための音声認識-衆議院のシステムの概要-, 情報処理学会研究報告 (音声言語情報処理), Vol. 2012, No. 5, pp. 1–6 (Oct. 2012).
[4] MR300, available from ⟨http://www.kingjim.co.jp/sp/mr360/.⟩.
[5] L. P. Morency, I. de Kok, and J. Gratch: Context-Based Recognition During Human Interactions: Automatic Feature Selection and Encording Dictionary, *Proc. of the 10th International Conference on Multimodal Interfaces (ICMI 2008)*, pp. 181–188 (Oct. 2008).
[6] 山本 剛, 坂根 裕, 竹林洋一: マルチモダールヘッドセットを用いたうなずき検出と会話の重要箇所把握, (情報処理学会研究報告. HI, ヒューマンインタフェース研究会報告), Vol. 2003, No. 94, pp. 13–19 (Sep. 2003).
[7] N. Wohler, U. Grosekathofer, A. Dierker, M. Hanheide, S. Kopp, and T. Hermann: A Calibration-Free Head Gesture Recognition System with Online Capability, *Proc. of the 20th International Conference on Pattern Recognition (ICPR 2010)*, pp. 3814–3817 (Aug. 2010).
[8] 高梨克也: 多職種ミーティングにおける懸念導入表現「気になる/するのは」の多角的分析, 言語処理学会第 19 回年次大会発表論文集, pp. 658–661 (Mar. 2013).
[9] 角 康之, 西田豊明, 坊農真弓, 來嶋宏幸: IMADE: 会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤, 情報処理学会論文誌, Vol. 49, No. 8, pp. 945–949 (Aug. 2008).
[10] 井上 卓, 角 康之, 高梨克也: 状況説明会話における説明者の発話とジェスチャの引き取り, インタラクション 2011 論文集, 2SCL-21, pp. 457–460 (Mar. 2011).
[11] 岡田将吾, 坊農真弓, 角 康之, 高梨克也, 新田克己: 非言語パターンの自動抽出による状況説明会話における言い淀みシーンの分析, インタラクション 2012 論文集, 3EXB-53, pp. 1007–1012 (Mar. 2012).
[12] 河原達也: スマートポスターボード: ポスター会話のマルチモーダルなセンシングと認識, 電子情報通信学会技術研究報告, SP2012-51 (July 2012).
[13] 中田篤志, 角 康之, 西田豊明: 非言語行動の出現パターンによる会話構造抽出, 電子情報通信学会論文誌, Vol. J94-D(1), pp. 113–123 (Jan. 2011).

[14] S. Kumano, K. Otsuka, D. Mikami, and Y. Junji: Recognizing Communicative Facial Expressions for Discovering Interpersonal Emotions in Group Meetings, *Proc. of the 11th International Conference on Multimodal Interfaces (ICMI-MLMI 2009)*, pp. 99–106 (Sep. 2009).

[15] 斎賀弘泰, 角 康之, 西田豊明: 多人数会話におけるうなずきの会話制御としての機能分析, 情報処理学会研究報告, Vol. 2010-UBI-26, No. 1, pp. 1–8 (May 2010).

[16] I. McCowan, D. Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang: Automatic Analysis of Multimodal Group Actions in Meetings, *the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI 2005)*, Vol. 27, No. 3, pp. 305–317 (Mar. 2005).

[17] 大塚和弘, 大和淳司, 村瀬 洋: 複数人物の対面会話シーンを対象とした画像中の人物頭部追跡に基づく会話構造のモデル化と確率的推論, 画像の認識・理解シンポジウム (MIRU 2006), pp. 84–91 (July 2006).

[18] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase: Quantifying Interpersonal Influence in Face-to-Face Conversations based on Visual Attention Patterns, *Proc. of 4th ACM Conference on Human Factors in Computing Systems (CHI 2006)*, pp. 1175–1180 (Apr. 2006).

[19] 小島慎也, 岩野公司, 古井貞熙: マルチストリーム HMM を用いた特徴量の次元別重み付き話者照合の検討, 情報処理学会研究報告 (音声言語情報処理) Vol. 2007, No. 129, pp. 43–47 (Dec. 2007).

[20] Wireless Technologies Inc., available from ⟨`http://www.wireless-t.jp/.`⟩.