修士論文

ウェアラブルセンサおよび対面カメラを用いた
ミーティング中のマイクロ行動の認識手法

曽根田悠介
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 安本 慶一 教授
ユビキタスコンピューティングシステム研究室（情報科学領域）

令和 2 年 3 月 13 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士 (工学) 授与の要件として提出した修士論文である。

曽根田悠介

審査委員：

安本 慶一	教授	(主指導教員, 情報科学領域)
中村 哲	教授	(副指導教員, 情報科学領域)
荒川 豊	客員教授	(副指導教員, 情報科学領域)
松田 裕貴	助教	(副指導教員, 情報科学領域)

ウェアラブルセンサおよび対面カメラを用いたミーティング中のマイクロ行動の認識手法*

曽根田悠介

内容梗概

近年では，アクティブラーニングと呼ばれる教育手法が注目され，その中でも少人数でのコミュニケーション能力の向上が重要視されている．先行研究では，円滑なコミュニケーションに重要とされているマイクロ行動（頷き，笑い，身振りなど）を認識する手法が数多く提案されている．しかし，使用するデバイスの金額が高額であったり，会話内容を解析する必要があるためプライバシーの保護の観点でリスクがあるなど実際の環境に導入するには困難と考えられる．

本論文で提唱するマイクロ行動の認識手法は,「実際のミーティングで導入可能な汎用的なシステムであること」,「マイクロ行動の認識精度が高いこと」を目指す．ミーティング中に発生するマイクロ行動を慣性センサと動画のそれぞれから認識する手法を提案する．慣性センサは被験者の頭部に装着し，動画は 360 度カメラを使用して円卓の中央から撮影したものを使用する．認識するマイクロ行動は慣性センサに関しては「Talk(発言)」「Smile(笑い)」「Nodding(頷き)」の 3 つ，動画に関しては「Talk」「Nodding」の 2 つである．動画からの Smile の認識は先行研究で数多く報告されているため本研究では行わない．被験者は 5 分間のディスカッションを行い，自分自身に発生したマイクロ行動についてラベリングを行い，そのラベルデータを正解ラベルとする．慣性センサと動画の有用性の比較および有用と判断した手法から実際のミーティングを想定した場合の認識を行い考察した．

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 2 年 3 月 13 日.

慣性センサと動画のどちらの場合についても機械学習 (Random Forest) によるマイクロ行動の認識を行なった．Talk の F 値は慣性センサ:62.4%，動画:69.5%，Nodding の F 値は慣性センサ:68.0%，動画:62.9%であった．Talk および Nodding の認識率は慣性センサと動画では大きな違いはなかった．慣性センサから認識したときの Smile の F 値は 50.2%と他のマイクロ行動の認識率よりも低いことが確認された．動画による Smile の認識は先行研究でも数多く報告されており，その認識精度は高く，また導入の容易さから動画による認識手法の方が有用であると判断した．次に，動画から一連の時系列データに対して機械学習によるマイクロ行動の認識を行なった．各マイクロ行動のサンプル数の偏りをオーバーサンプリングの手法である ADASYN によってバランシングすることで，認識精度が向上したことが確認された．

キーワード

マイクロ行動認識，マルチモーダルセンシング，機械学習，ヒューマンコミュニケーション，会議支援

Micro Behavior Recognition using Wearable Sensor and Facial Camera during Meeting*

Yusuke Soneda

Abstract

In recent years, educational methods called active learning have attracted attention, improvement of communication skills with a small group has been emphasized. Related studies have proposed methods for recognizing micro behaviors (nodding, laughing, gestures, etc.) that are important for smooth communication. However, it is considered difficult to introduce proposed methods into a real environment because the device is too expensive and there is a risk from the viewpoint of privacy because the conversation contents need to be analyzed.

In this thesis, we propose the micro behavior recognition method aims at "a general-purpose system that can be introduced in actual meetings" and "high accuracy recognition of micro behavior". We propose a method for recognizing micro behaviors that occur during a meeting from the IMU and the video. IMU is attached to the participant's head, and the video is shooted by a 360-degree camera from the center of the round table. We recognize 3 micro behaviors "Talk", "Smile" and "Nodding" from IMU data. And we recognize 2 micro behaviors "Talk" and "Nodding" from video data. We do not recognize Smile from video data in this study because it has been reported in related studies. Participants discuss for 5 minutes, label the micro behaviors for theirs, and use the label data as the correct label. We compared the usefulness of the IMU and the video, and recognized the micro behaviors in continuous time-series data.

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 13, 2020.

In both the case of the inertia meter and the video, we recognize the micro behaviors by machine learning (Random Forest). The Fmeasure of Talk was 62.4% for IMU and 69.5% for video, and the F-measure for Nodding was 68.0% for IMU and 62.9% for video. Talk and Nodding recognition rates did not greatly differ between IMU and video. It was confirmed that the F-measure of Smile recognized from the IMU was 50.2%, lower than the recognition rate of other micro behaviors. Some studies have reported Smile recognition methods using images. The recognition accuracy was high, and it is easy to use in the actual scene, so we judged that the video recognition method was more useful than IMU. Next, micro behavior recognition was performed on a series of time-series data from the video using machine learning. We confirmed that the recognition accuracy was improved by balancing the bias of the sample number of each micro behavior by ADASYN, which is an oversampling method.

Keywords:

Micro Behavior Recognition, Multi Modal Sensing, Machine Learning, Human Communication, Support Meeting

目 次

図 目 次

表 目 次

1. 序論

近年，日本では働き方改革と叫ばれるように，従来の働き方が見直されている．ある企業の調査によると [1]，日本の一般的な企業において，1 日の勤務時間のうちミーティングが占める時間は平均で 68.2 分と報告されている．1 日の勤務時間のうち多くの時間がミーティングに費やされていることから，働き方改革を推進していくにはミーティングにおける質の定量化や向上が重要になると考えられる．さらに Steven ら [2] の調査によると，効率的でないもしくは面白みのないミーティングは特に年齢の若い社員のエンゲージメントの低下を誘発し，ミーティングへの遅刻を引き起こすことが報告されている．この様に非生産的なミーティングは多大な時間の浪費をまねき，アメリカでの経済的な損失は約 4 兆円に相当すると報告されている．また，教育においてはアクティブラーニングと呼ばれる教育方法が注目され，生徒が主体的にコミュニケーションを取り，自主的に課題を解決する能力が求められている [3]．その際に，少人数でのディスカッションにおけるコミュニケーション能力は非常に重要視されている．上記のように，様々な環境で良好なコミュニケーションを行うための能力が近年において求められていることがわかる．しかし，コミュニケーションに関する能力は定性的に評価されることが多く，定量的にコミュニケーションを評価・支援を行うには実際の人対人のコミュニケーションについて詳しく解析を行う必要がある．

ミーティングの支援に関する研究について，ミーティング中の発言内容を自動で要約し議事録の作成を行うといった言語情報に注目した研究が複数報告されている [4][5]．また，ICSI Meeting Corpus[6] や AMI Meeting Corpus[7] といった，円滑な会議の進行や会議録の作成に関する研究の支援を目的とし，人手によるアノテーションが付与された会議録データが公開されている．これにより，ミーティングを振り返る時間や第三者にミーティングの内容を共有するために必要な時間を短縮することが可能になることが期待される．しかし，実際のオフィスでのミーティングでの発言内容には個人情報や機密事項を含む可能性があり，実際にオフィスや教育現場での発言内容の解析はプライバシー保護の観点からリスクが発生する．

そこで，ミーティング中に発生する頷きや身振りといったジェスチャーや顔の

微小な表情の変化，視線の動きなどに着目する．このように言語情報に頼らないコミュニケーション方法は非言語コミュニケーションと呼ばれる．Ekman[8] や Peter[9] の研究によると，このような非言語コミュニケーションは円滑なコミュニケーションを行うためには非常に重要であると報告されている．適切な非言語コミュニケーションはコミュニケーションを円滑にする一方，意識・無意識に関わらず誤った非言語コミュニケーションは相手に不快感を与える可能性がある．本論文では，発言内容の解析は行わず，特に頭部に現れるマイクロ行動（Talk，Smile，Nodding）に着目し，慣性センサと 360 度カメラによって撮影された動画から機械学習 (Random Forest) を用いた認識を行う．特徴量と正解データの抽出には，スライディングウィンドウ法を用いた．このときのウィンドウサイズは約 1 秒，オーバーラップは 50%とした．

慣性センサからの認識結果から算出された F 値は，Talk : 62.4%，Smile : 50.2%，Nodding : 68.0%と確認された．同様に，動画からの認識結果から算出された F 値は，Talk : 69.5%，Nodding : 62.9%と確認された．慣性センサと動画を用いた Talk と Nodding の認識精度には大きな差は見られなかった．慣性センサによる Smile の認識精度は他のマイクロ行動に比べて低かった．画像を用いた Smile の認識は先行研究でも複数報告されており，また慣性センサは被験者の人数分用意する必要があるのに対して，360 度カメラ 1 つから複数人の認識が行えることから，動画による認識手法の方がより有用であると判断した．

次に，有用と判断した動画による認識手法から，行動変化点を考慮し時系列に予測結果を可視化し考察を行なった．特に Nodding の認識精度については，ADASYN と呼ばれる擬似的にサンプル数を増加させるオーバーサンプリングにより，各マイクロ行動のサンプルの不均衡状態を解消することによって精度が向上することを確認した．

本論文の構成は以下の通りである．第 2 章では，関連研究として非言語コミュニケーションがコミュニケーションにおいて与える影響に関する研究を取り上げ，マイクロ行動のコミュニケーション中における重要性について整理する．また，既存のミーティング中に発生するマイクロ行動の認識に関する研究を取り上げ，既存手法における課題を整理する．第 3 章では，実際のミーティング中に発生す

るマイクロ行動をセンシングするために作成したシステムの説明とマイクロ行動の認識手法について記述する．第 4 章では，マイクロ行動に関するデータを収集するための実験内容とそれぞれのマイクロ行動に関する統計量について記述する．第 5 章では，実験から取得された慣性センサと動画のそれぞれのデータからマイクロ行動の認識を行い，慣性センサと動画の認識結果を比較する．第 6 章では，前章でより有用とした判断したデバイスを使用し，正解ラベルが付与されていない実際のミーティングのデータを想定した状態でマイクロ行動の認識を行う．第 7 章では，慣性センサと動画から認識を行なった結果を用いて考察を行う．最後に，第 8 章にて本論文の結論と今後の発展性について述べる．

2. 関連研究

2.1 コミュニケーションにおける参加構造

Clark は [10] 複数人でのコミュニケーションにおける参加構造モデルを図 1 のように提唱した．コミュニケーションに対するエンゲージメント密度から，それぞれの参加者がコミュニケーション内でどのような立ち位置にいるかを定義している．ratified participants とは，そのコミュニケーションに批准している人物を意味し，コミュニケーションに参加するべき立ち位置にいる．Harmonized とは，全員がエンゲージメント密度が高い状態でコミュニケーションに参加している状態を表し，図 1 における A，B，C1 で表現されている人物は Harmonized な状態である．Un-Harmonized とは，一部の人物が会話に参加できずにエンゲージメント密度が低い状態でコミュニケーションに参加できていない状態を表し，図 1 における C2 で表現されている人物は Un-Harmonized な状態である．この Un-Harmonized 状態の人物が発生している状況は，コミュニケーションとして好ましくない状態と考えられる．

松山らは [11]Clark の提唱するコミュニケーションの参加構造を会話内容と笑顔といった非言語コミュニケーションから推定し，ファシリテーション役のロボットが会話を用いてエンゲージメント密度を調整する研究を行なった．実験の結果，エンゲージメント密度を計算することでロボットは適切なタイミングでファシリテーションを行うことが可能になり，被験者はロボットの行動に妥当性を感じ被験者のエンゲージメント密度を適切に調整することが可能であった．しかし，発言内容を解析する必要があるため個人情報を取り扱うリスクが生じ，また非言語でのコミュニケーションの解析にはディスカッションに参加する人物一人につき RGBD カメラ (Microsoft Kinect) を用いる必要があるなど実際でのオフィスや教育現場での導入は困難である．

Jokinen は [12] 発話の移り変わりと視線情報を用いてコミュニケーションの参加構造について解析を行なった．会話中での役割や集中度を計測するのに，視線の情報と発言権の移り変わりを用いるのは有用であると結論付けている．しかし，視線情報を設置型の視線系を用いて計測している一方，笑いや頷きといった非言

語コミュニケーションについては分析されておらず，コミュニケーションへのエンゲージメント密度を計測するには不十分であると考えられる．

このように，コミュニケーションにおける参加構造を会話内容と笑顔や視線といった非言語コミュニケーションから求める研究は複数存在する．ミーティングにおけるエンゲージメントを計測する上で，非言語コミュニケーションの解析することは有効であると考えられる．

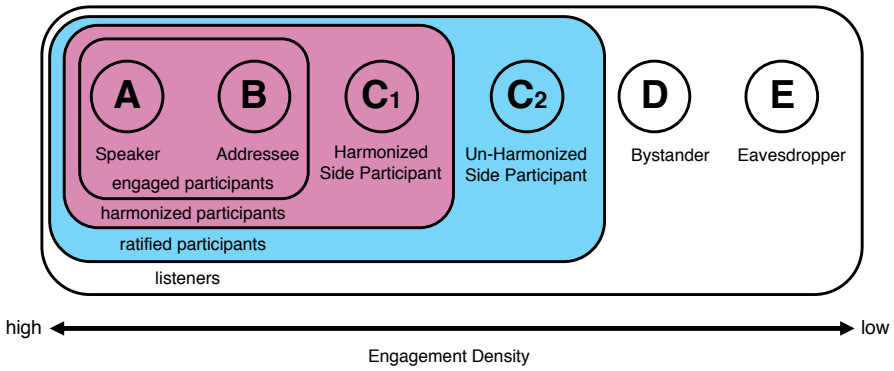


図 1 Clark の提唱するコミュニケーションの参加構造

2.2 非言語コミュニケーションが与える影響について

2.2.1 微表情 (Micro Expression) に関する関連研究

Ekman らは [8] 文化や国に依存せず，顔に表出する感情は 6 つに分類することが可能であると報告した．その後，さらに Ekman ら [13] は顔の表情をさらに詳細に表現するため，FACS(Facial Action Coding System) と呼ばれる顔の動きを分類するシステムを開発した．FACS は，顔面の筋肉を解剖学的な視点から 46 の動作単位 (Action Unit，AU) を要素に客観的に顔の動きを記述する方法である．この AU の組み合わせにより 6 つの基本的な表情をさらに詳細に表現することが可能になった．

顔画像から AU を推定する研究も数多く報告されている [14][15]．また，AU を推定するために，ラベル付けされた顔画像のデータセットも多く公開されている [16][17]．このように，顔画像から顔の微表情を認識を行う研究は近年において

も盛んに研究されている．顔の微表情がコミュニケーションに与える影響は大きいと考えられるため，微表情の認識精度の向上により非言語コミュニケーションの解析が発展していくと期待される．

2.2.2 視線に関する関連研究

Garau らは [18] モニタに表示された人間型のエージェントを用いて，視線がコミュニケーションに与える影響について調査を行なった．視線をランダムに動かすエージェント，視線を適切なタイミングで動かすエージェント，実際の人間の映像，モニターには表示はなく音声のみの 4 通りの方法を用いて，被験者は擬似的に会話を行い，その印象についてアンケートを行う．その結果，被験者は視線をランダムに動かすエージェントと音声のみ場合よりも視線を適切なタイミングで動かすエージェントに対しての方が，より自然なコミュニケーションを行えたと回答した．適切な視線の動かし方はコミュニケーションにおいて重要であり，誤った視線はコミュニケーションを行う相手に対して違和感や不自然感を与えてしまう可能性が考えられる．このように視線は円滑なコミュニケーションを実現するために重要な役割を果たしている．

2.2.3 仕草や姿勢に関する関連研究

Peter は [9] 仕草や姿勢がどのような意図を表出しているのか，もしくは相手にどのような印象を与えるかについて研究を行なった．例えば，腕組みや足組は相手に対して退屈である・反対意見を持っている印象を有意に与えることが確認された．このように仕草や姿勢は意識・無意識に関わらずコミュニケーションに影響を与えることが報告されている．

Duncan は [19][20][21] 一連の研究で，話し手または聞き手が発言権の交代に関する社会的シグナルとそれらの役割についての法則を発見した．例えば，話し手がハンドジェスチャーを行うことによって聞き手が発言権を要求するのを阻止する行動 (speaker gesticulation signal)，聞き手が頷くことによって話し手の発言を聞き続けたいことを表明する行動 (auditor back-channel signal) などがある．この

ように，いくつかのジェスチャーは社会的シグナルを意味し，コミュニケーションを円滑に行うために重要であるとされている．

上記の複数の研究によると，非言語コミュニケーションが人対人でのコミュニケーションに大きな影響を及ぼすことが報告されている．コミュニケーションの質の推定や参加者にフィードバックを行う上で，非言語コミュニケーションの解析や認識は重要であると考えられる．

2.3 ミーティングに関するセンシング

Samrose らは [22] オンラインミーティングに着目し，明確な答えがなく全員が均等に会話に参加することを目標とするテーマ（例：沈み行く船から脱出する際にどの道具を持ち出すべきか？）についてディスカッションを行い，チャットボットのフィードバックによって被験者にどのような行動変容が発生するか調査した．パソコンに付随するインカメラの映像から感情の認識や各被験者のパソコンから取得される音声の解析を行い，ディスカッション中の発言率や表情についてチャットボットは各被験者にフィードバックを行う．その結果，一度目のディスカッションで発話が多かった被験者は 2 度目のディスカッションの際には発言を控える傾向にあり，反対に発言が少なかった被験者は積極的に発言するようになり全体が均等に発言を行うように有意に変化することが確認された．しかし，Samrose らが提案する実験は限定的なテーマに関してのみ評価が可能であり，オフィスワークや教育現場の中で実際に想定されるようなディスカッション内容を想定されていない．

大西らは [23] 頭頂部に慣性センサを装着し，3 軸の加速度・角加速度から発話，頷き，見渡しの 3 種類の行動について機械学習を用いて認識を行なった．提案された手法では発話の認識精度は 97.5%，頷きは 52.4%，見渡しは 53.6%と報告されている．しかし，前提として参加者は研究に関する報告を一定の順番で行うといった限定された状況でのみ実験が行われ，他の形式のミーティングでの評価が行われておらず汎用的なシステムではない．

Morency らは [24] 動画と会話のコンテキスト情報（疑問文であるか，説明であるかなど）を用いて頷きまたは首を振るの動作について認識を行なった．この時

の実験環境は，質問を行うロボットに対して被験者が回答を行うというものである．動画と会話のコンテキスト情報を組み合わせた場合の方が，動画のみから認識を行った時よりも精度が向上した．しかし，被験者の発する行動は頷きと首を振るといった簡単な行為のみに限定されるような実験であり，実際の環境で発生するような発言や笑いといった行動が含まれていない．また，ロボットの動きは実験者により制御されているため，人対人のコミュニケーションには適用できない．

Yu らは [25] 複数のセンサからディスカッション中の発言内容や頷きといった行動の認識を行うためのシステムを開発した．特に頭部の動作に着目し，光学式モーションキャプチャシステムから頭部の位置情報や傾きの検出を行なった．提案されたシステムを用いることによって，顔の角度を正確に捕捉することが可能となり，頷きの認識精度は 76.4%，首を横に振る動作の認識精度は 80.0%であった．しかし，高額な光学式モーションキャプチャを用意する必要があり，また音声の解析も行うため各被験者につき 1 つずつマイクが必要など一般的な環境で導入するには困難である．

上記の研究のように，ディスカッションについてのセンシングに関する先行研究は多く存在する．特に，非言語コミュニケーションについては会話の中で円滑なコミュニケーションのために重要と考えられている頷きの認識が試みられている．しかし，認識精度には向上の余地があり，また導入におけるコストが高い設備が必要であることが確認された．さらに，ミーティングの条件を限定することによって認識精度を向上を試みる研究もいくつか報告されているが，様々な場面でのミーティングの解析を行うためにはより汎用的な状況で導入できる認識システムが求められる．これらの先行研究は，データセットが公開されておらず，非言語コミュニケーションに関する研究が発展していくためにはデータセットの公開が重要であると考えられる．本研究では，このような背景から実際のミーティングを想定した状況で，慣性センサと 360 度カメラを用いてミーティング中に発生するマイクロ行動の認識を行い，認識結果がどのように異なるか考察を行う．

3. ミーティング中に発生するマイクロ行動に注目したセンシングシステムおよび認識手法

本研究では，マイクロ行動を認識するために複数のセンサを用い，ディスカッション中に発生する仕草や行動に関するセンサデータを収集した．特に，非言語コミュニケーションの中で重要と考えられ，各被験者に共通して観測された頭部に発生するマイクロ行動「Talk(発言)」「Smile(笑い)」「Nodding(頷き)」に着目した．本章では，データセットを作成するにあたって使用したデバイスや構築したシステムとマイクロ行動の認識方法について記述する．

3.1 使用デバイスおよびデータ収集システム

ミーティング中に発生するマイクロ行動を測定するために，RICOH THETA V[26]，Pupil Labs Mobile Eye Tracking Headset[27]（以下 Pupil と省略），LPMS-B2[28] の 3 種類のセンサデバイスを使用した．

THETA V は全方位を撮影可能な 360 度カメラであり，その動画のフレームレートは 29.97fps， 画質は 3840×1920 ピクセルである．Pupil は装着型の視線測定計であり，図 3 に示すように装着者がどこを見ていたのかを計測する．図 3 の場合，緑色の円が Pupil を装着している被験者の視線を表し，右の人物を注視していることが分かる．LPMS-B2 は図 2 に示すように 3 軸の加速度と 3 軸の角加速度を取得出来る慣性センサである．センサの周波数は設定可能であり，本実験では 100Hz に設定した．THETA V は机の中央に設置，Pupil は図 4 のように各被験者が装着し，LPMS-B2 は図 5 のように各 Pupil の柄の部分に装着した．LPMS-B2 は頭部の加速度と角加速度を取得する．

これらのセンサを全て手動で起動させデータの収集を行う場合，各センサデータは独立して計測を開始するためタイムスタンプにずれが生じる．各センサを図 6 に示すように，WiFi，Bluetooth， 有線を用いて制御用のコンピュータに接続することによって同期してセンサからデータを収集できるシステムを構築した．このシステムにより，各センサのタイムスタンプのずれが生じるのを防ぎ，360 度

カメラの動画に対してのラベリングデータを慣性センサや Pupil の解析に用いることが可能になる．

本研究では，慣性センサの加速度・角加速度データと 360 度カメラの動画を用いてマイクロ行動の認識を行い比較および考察を行う．

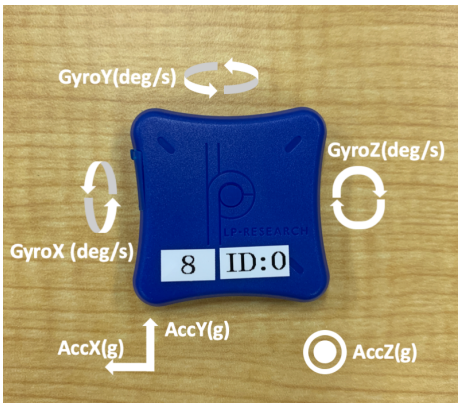


図 2 慣性センサの加速度・角加速度の出力

図 3 Pupil から取得される映像

図 4 Pupil を装着した様子

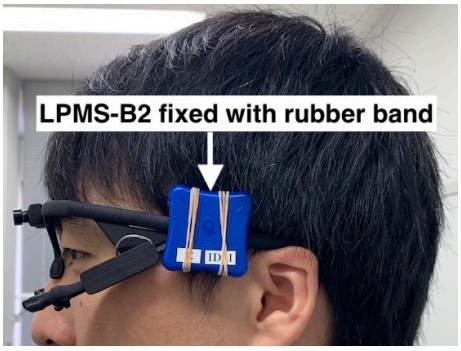


図 5 慣性センサを Pupil に装着した様子

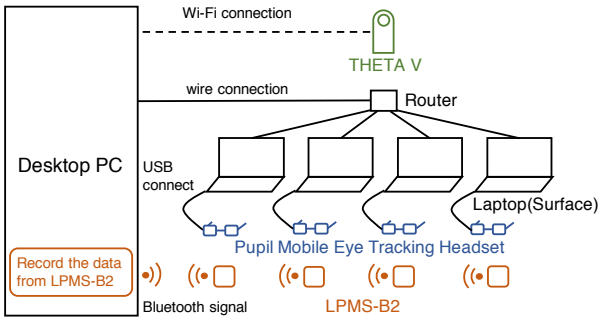


図 6 複数センサを用いたセンシングシステム

3.2 マイクロ行動の認識手法

歩行や走行，階段を登るといった動作を加速度や動画から認識を行う研究は数多くの研究が報告され [29][30]，行動認識に関するデータセットについても複数公開されている [31][32]．一方，実際のミーティングにおける頷きといった行動は非常に短く素早い行動な為，歩行や走行といった継続時間が長い動作と比べると認識が困難である．恣意的に発生させられた頷きでは，本来の無意識に発生する頷きに比べて深く継続時間の長い頷きが発生し，実際のミーティングでの頷きとは顔の角度やテンポが異なる可能性が考えられる．したがって，本研究では実際のミーティング中に発生するマイクロ行動の認識を行う．

本研究では，行動認識のアルゴリズムにスライディングウィンドウ法を用いた手法を採用した．図 7 は慣性センサの加速度と角加速度のデータを時系列上にプロットしたものである．図 7 に示すように，センサデータをある時間幅を持った部分系列に分割する．その部分系列から，センサデータの最大値 (Max) や歪度 (Skewness) といった特徴量を説明変数，ラベリングデータを目的変数とし機械学習に基づく認識を行う．スライディングウィンドウの正解ラベルには人手によるラベリングデータを使用するが，ウィンドウの正解ラベルの決定方法には次の 2 通りの定義が考えられる．

- 定義 1: 行動変化点を含む場合はそのウィンドウデータを無視する．
- 定義 2: ウィンドウに含まれるラベリングデータのうち最も多いものをそのウィンドウの正解データとする．

実際のミーティングのマイクロ行動の認識を行う際に，図 8 に示すようにウィンドウ内に行動変化点が含まれる場合が発生する．定義 1 では，このようなスライディングウィンドウのデータは扱わないものとする．この方法で抽出されたウィンドウのデータには 1 つの行動に関するセンサデータのみが含まれるため，各マイクロ行動の認識が比較的容易になる．したがって，5 章では慣性センサと動画のデータから定義 1 の手法によって抽出されたウィンドウデータを用いてマイクロ行動の認識を行い，どちらのセンサが有用であるかを評価する．

しかし，定義 1 で抽出されるスライディングウィンドウのデータは正解データがラベリングされていることが前提であり，正解データが未知のデータに対して認識を行うときには，スライディングウィンドウ内に行動変化点を含む場合を考慮する必要がある．しかし，本研究で取り扱うような微小な動きについては行動変化点の検知は非常に困難である．そこで，定義 2 ではウィンドウ内に含まれている行動ラベルで最も時間が長い行動ラベルをそのウィンドウの正解の行動ラベルとする．第 6 章では実際の予測に使われるデータを想定し，第 5 章で有効と判断したデバイスを用いて定義 2 の手法によって抽出されたウィンドウデータからマイクロ行動の予測を行う．

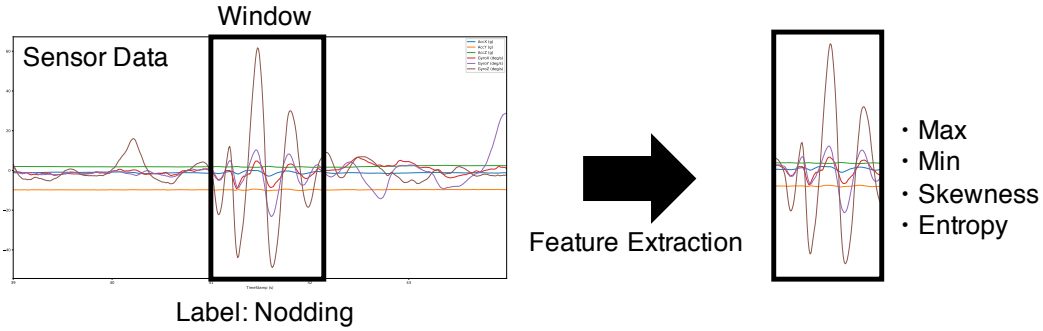
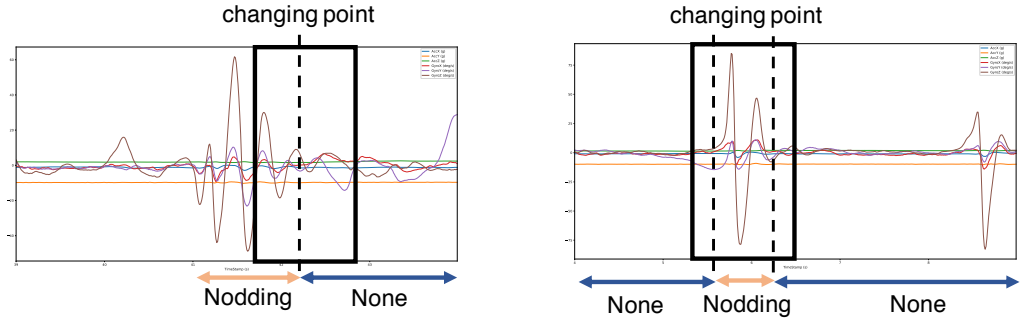


図 7 センサデータから特徴量を抽出する様子



(a) 行動変化点がウィンドウ内に含まれる場合　(b) 行動が短いため他の行動が含まれる場合

図 8 行動変化点がウィンドウ内に含まれる場合

3.3 認識結果の評価方法

認識精度の評価方法には適合率・再現率・F 値を用いる．適合率は各マイクロ行動を識別したときに，識別したマイクロ行動が実際の行動と一致した割合であり，再現率は識別するマイクロ行動が正しくその行動であると識別された割合である．F 値は適合率と再現率の調和平均から算出されるものであり，式 1 で表される．

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{1}$$

3.4 慣性センサを用いたマイクロ行動の認識

3.4.1 信号処理および特徴量抽出

慣性センサから取得される加速度，角加速度に対し，メディアンフィルタと 20Hz の 3 次バタワース・ローパスフィルタをかけてスパイクノイズなどのノイズ除去を行う．この周波数は，身体動作の 99 %が 15Hz 以下に含まれているため，動作の捕捉を行うために十分である [33]．ノイズ除去された加速度信号には重力および体動成分が混在しているため，0.3Hz のバタワース・ローパスフィルタをかけ，重力成分 (GravityAcc.XYZ) と体動成分 (HeadAcc.XYZ) に分離させる．さらに，時間微分を計算することによって取得されるジャーク信号 (HeadAccJerk.XYZ) と 3 軸信号から式 2 を用いてユークリッド距離を算出することで取得される合成信号 (GravityAccMag，HeadAccMag，HeadAccJerkMag) を追加する．同様に，ジャイロ信号についても，頭部のジャイロ信号 (HeadGyro.XYZ) に加えて，ジャーク信号 (HeadGyroJerk.XYZ) および合成信号 (HeadGyroMag，HeadGyroJerkMag) を追加する．

$$Magnitude(Mag) = \sqrt{(x^2 + y^2 + z^2)} \tag{2}$$

次に，スライディングウィンドウ法を用いて各ウィンドウから特徴量を抽出す

る．このときのウィンドウ幅は 1.28 秒 (128 サンプル)，オーバーラップは 50%とする．第 4 章の表 4 で後述する通り，特に Nodding の継続時間は平均で 1.44 秒と極めて短い動作である．従って，図 8 で説明した通り，ウィンドウ幅を大きくしてしまうと行動変化点が含まれてしまうウィンドウが増加してしまい，抽出されるマイクロ行動のサンプル数が大きく減少してしまうため，ウィンドウ幅を上記のように設定した．ウィンドウで区切られた信号から表 1 に示す，時間ドメイン特徴量 (Time domain features) と周波数ドメイン特徴量 (Frequency domain features) を抽出する．これらの特徴量を選択する理由として，主に慣性データを用いた運動認識に関する先行研究から有効性が示されているためである [29][34][35][36].

3.5 顔画像を用いたマイクロ行動の認識

3.5.1 Smile の認識について

Talk と Nodding については，時系列に依存するマイクロ行動のため 1 フレームからの認識が不可能であるが，Smile ついては 1 フレームの画像から認識が可能である．近年では深層学習の発展による影響もあり，笑顔の認識は非常に盛んに研究され精度も大きく向上している [37][38]．したがって，本研究では既存の手法による画像から Smile の認識を前提とし，動画からの Smile の認識は行わないものとする．

3.5.2 OpenFace を用いた顔画像の変換

本研究では，顔の特徴点の座標や角度を求めるためにオープンソースのソフトウェアである OpenFace[39] を使用した．OpenFace を顔画像に適用させた時の画像を図 9 に示す．OpenFace は顔画像を図 10 に示す 68 個の特徴点，および顔の角度をピッチ方向 (pose_Rx)，ヨー方向 (pose_Ry)，ロール方向 (pose_Rz) の 3 成分に変換する．

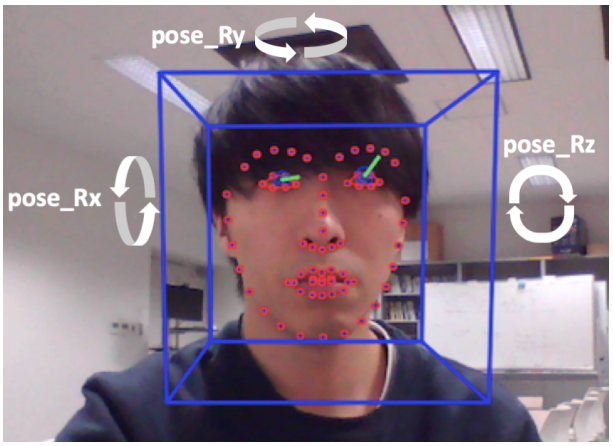


図 9 顔画像を OpenFace を用いて点群データに変換した様子

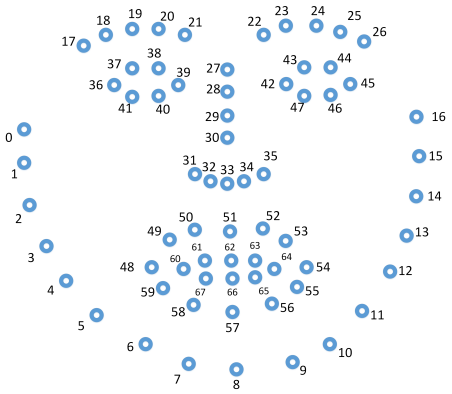


図 10 OpenFace が出力する特徴点

3.5.3 Nodding の認識に使用する特徴量抽出

Nodding は顔の垂直方向の角度が変化するため，図 9 に示す pose_Rx 成分についてスライディングウィンドウ法を用いて特徴量を抽出する．本実験で使用するカメラのフレームレートは 29.97fps のため，Nodding の継続時間に注意しウィンドウ幅は 1.06 秒 (32 フレーム)，オーバーラップは 50%とする．慣性センサによる行動認識と同様に，表 1 に示す特徴量を抽出する．このとき pose_Rx のみを信号と

して使用しているため，複数の信号を組み合わせて生成される sma，correlation，angle の 3 つの特徴量を除外した．

3.5.4 Talk の認識に使用する特徴量抽出

Talk は上唇と下唇が最も動くため，図 10 に示す特徴点の中で 62 番と 66 番の特徴点の距離成分からスライディングウィンドウ法を用いて特徴量を抽出した．ウィンドウ幅は Nodding と同様にウィンドウ幅は 32 フレーム，オーバーラップは 50%とし，表 1 に示す特徴量を抽出する．

表 1 特徴量リスト

Function	Description	Formulation	Type
mean（s）	Arithmetic mean	$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$	T，F
std（s）	Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2}$	T，F
mad（s）	Median absolute deviation	$median_i(\mid s_i - median_j(s_j) \mid)$	T，F
max（s）	Largest values in array	$max_i(s_i)$	T，F
min（s）	Smallest value in array	$min_i(s_i)$	T，F
energy（s）	Average sum of the square	$\frac{1}{N}\sum_{i=1}^{N} s_i^2$	T，F
sma（s_1，s_2，s_3）	Signal magnitude area	$\frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{N}\lvert s_{i,j}\rvert$	T，F
entropy（s）	Signal Entropy	$\sum_{i=1}^{N}(c_i \log(c_i)), c_i = s_i / \sum_{j=1}^{N} s_j$	T，F
iqr（s）	Interquartile range	$Q3(s) - Q1(s)$	T，F
autorregresion（s）	4th order Burg Autoregression coefficients	$a = arburg(s,4), a \in \mathbb{R}^4$	T
correlation（s_1，s_2）	Pearson Correlation coefficient	$C_{1,2}/\sqrt{C_{1,1}C_{2,2}}, C = cov(s_1, s_2)$	T
angle（s_1，s_2，s_3，v）	Angle between signal mean and vector	$\tan^{-1}(\parallel [\bar{s}_1,\bar{s}_2,\bar{s}_3] \times \upsilon \parallel, [\bar{s}_1,\bar{s}_2,\bar{s}_3] \cdot \upsilon)$	T
range（s）	Range of smallest value and Largest value	$max_i(s_i) - mix_i(s_i)$	T
rms（s）	Root square means	$\sqrt{\frac{1}{N}(s_1^2 + s_2^2 + \cdots + s_N^2)}$	T
skewness（s）	Frequency signal Skewness	$E[(\frac{s-\bar{s}}{\sigma})^3]$	F
kurtosis（s）	Frequency signal Kurtosis	$E[(s-\bar{s})^4]/E[(s-\bar{s})^2]^2$	F
maxFreqInd（s）	Largest frequency component	$argmax_i(s_i)$	F
meanFreq（s）	Frequency signal weighted average	$\sum_{i=1}^{N}(is_i)/\sum_{j=1}^{N} s_j$	F
energyBand（s，a，b）	Spectral energy of a frequency band [a, b]	$\frac{1}{a-b+1}\sum_{i=a}^{b} s_i^2$	F
psd（s）	Power spectral density	$\frac{1}{Freq}\sum_{i=1}^{N} s_i^2$	F

N : signal vector length，Q : Quartile，T : Time domain features, F : Frequency domain features.

4. マイクロ行動認識のためのデータ収集実験

4.1 実験概要

4.1.1 実験手順

4 人の被験者は指定された内容について 5 分間ミーティングを行った．このときのミーティングの内容は，被験者の事前知識に依存しないような簡単なものであり，発言が発散してしまうのを防ぐため回答が 2 択に絞られるようなトピックを作成した（例 :「犬と猫どちらが好きか？」「行けるのであれば過去と未来どちらに行きたいか？」）．

ミーティングが終了したのち，被験者はミーティングについての理解度や満足度に関するアンケートに回答を行う．

ミーティングとアンケートを 2 回ずつ行なった後，各被験者は自分自身のマイクロ行動について 360 度カメラによって撮影された動画から確認を行いラベリングを行う．

4.1.2 アンケート項目

アンケートの項目を表 2 に示す．A1，A2 については，意見が前者の場合であれば 1 と回答し，後者の場合であれば 5 と回答する．B1 から B8 については，否定であれば 1，肯定であれば 5 と回答する．

表 2 アンケート項目

Survey ID	アンケート内容
A1	ミーティングを行う前，あなたの意見は前者でしたか後者でしたか？
A2	ミーティングを行なった後，あなたの意見は前者でしたか後者でしたか？
B1	ミーティングに満足であった．
B2	自分の意見を発言出来た．
B3	同じ意見を持った人物の発言を聞くことが出来た．
B4	異なった意見を持った人物の意見を聞くことが出来た．
B5	ミーティングは楽しいものであった．
B6	自分の発言を遮られたと感じた．
B7	もう一度このグループでミーティングを行いたいと感じた．
B8	360 度カメラが気になった．

4.1.3 ラベリング項目

各被験者は，ELAN[40] という時系列データに対してラベリングを行うソフトを使用して，THETA V の動画を参照し自分自身のマイクロ行動についてラベリングを行う．実際にラベリングを行なっている画面を図 11 に示す．ラベリングの項目については表 3 に示す．このときに，3 つのラベリング項目に付随して，マイクロ行動を行なった理由についてもラベリングを行なった．この行動を行なった理由については Duncan ら [19][20][21] の文献を参照し決定した．

表 3 ラベリング項目

ラベル名	詳細なカテゴリ	詳細
smile	response	特別な意味を持たない笑い
	agree	同意の時に発生する笑い
	interesting	会話が面白い時に発生する笑い
	sympathy	同意を求める時に発生する笑い
nodding	response	特別な意味を持たない頷き
	agree	同意の時に発生する頷き
talk	description	意見を説明するときの発言
	objection	他者の意見を否定するときの発言
	agree	他人の意見に同意するときの発言
	say	他人に発言権を譲渡するときの発言

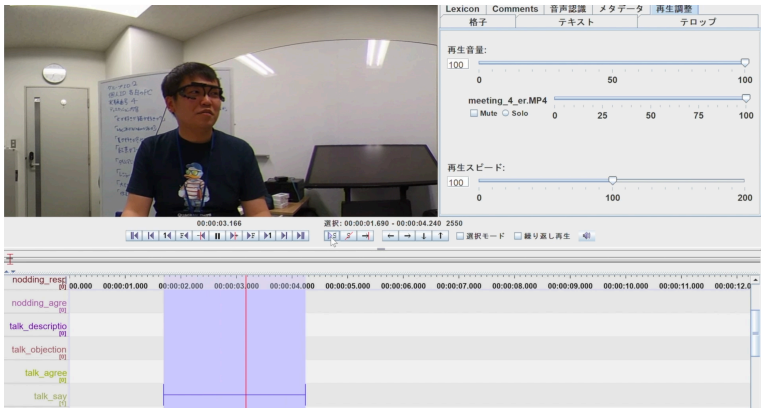


図 11 ELAN を用いてラベリングを行う様子

4.2 データセット

2019 年 5 月 7 日から 6 月 5 日の期間で実験を実施した．合計で 16 回のミーティングに関するデータを取得した．奈良先端科学技術大学院大学ユビキタスコンピューティングシステム研究室の学生 21 名が被験者として参加した．そのうち

男性は 17 名 (うち 9 名は 2 回参加) で女性は 4 名 (うち 2 名は 2 回参加) であった．データセットを作成するにあたって合計でのべ 130 時間要した．その多くの割合はラベリングの作業に費やされた．

4.2.1 各マイクロ行動に関しての統計量

図 12 および表 4 は各マイクロ行動の継続時間と統計量について，図 13 および表 5 は被験者一人に関するマイクロ行動の合計時間の分布と統計量を示す．

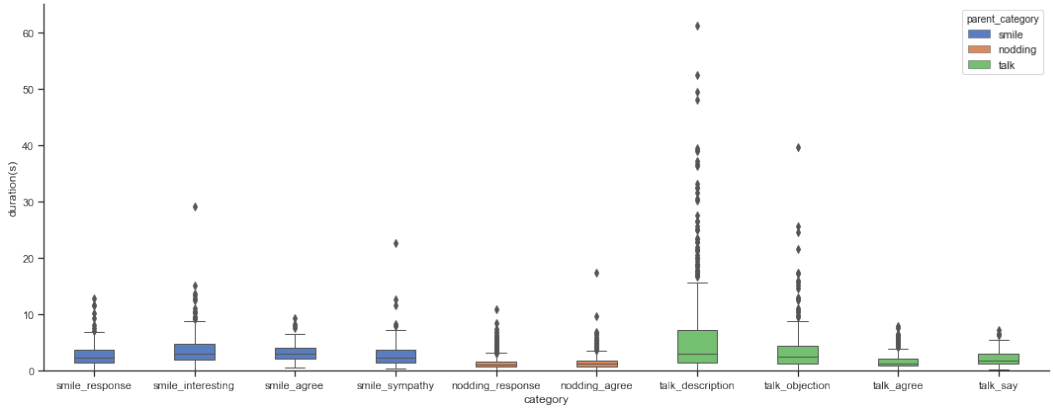


図 12 マイクロ行動の継続時間の分布

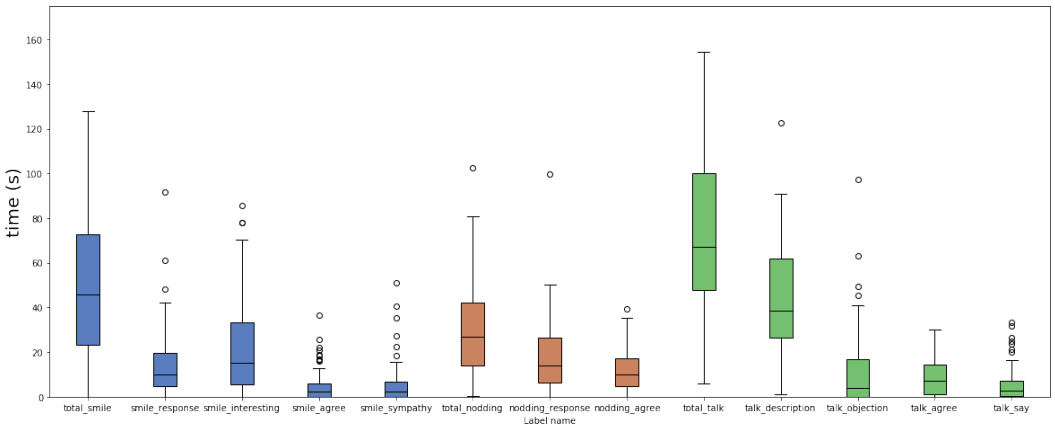


図 13 5 分間のミーティング中に被験者 1 人あたりが行うマイクロ行動の合計時間の分布

表 4 マイクロ行動の回数と継続時間に関する統計量

statistics	s_total	s_response	s_interesting	s_agree	s_sympathy	n_total	n_response	n_agree	t_total	t_description	t_objection	t_agree	t_say
count(time)	906	324	352	112	118	1404	859	545	1143	481	181	315	166
mean(s)	3.25	2.78	3.77	3.23	3.06	1.44	1.39	1.53	4.01	6.07	4.09	1.74	2.28
median(s)	2.75	2.36	3.06	2.96	2.36	1.11	1.07	1.2	1.96	2.98	2.44	1.32	1.73
SD(s)	2.4	1.85	2.79	1.72	2.76	1.19	1.12	1.29	6.14	8.35	5.08	1.29	1.47
max(s)	29.07	12.76	29.08	9.39	22.72	17.37	10.97	17.37	61.25	61.25	39.63	7.99	7.2

表 5 5 分間のミーティング中に被験者が行うマイクロ行動の統計量

statistics	s_total	s_response	s_interesting	s_agree	s_sympathy	n_total	n_response	n_agree	t_total	t_description	t_objection	t_agree	t_say
mean(s)	49.30	15.15	22.03	5.67	6.43	31.06	18.40	12.66	72.63	45.07	12.55	8.78	6.21
median(s)	45.67	10.22	15.41	2.56	2.49	27.08	14.11	10.11	67.04	38.54	4.01	7.30	2.63
SD(s)	31.94	16.82	21.82	8.06	10.65	21.41	16.65	10.23	35.31	24.31	19.21	7.94	8.53
max(s)	128.06	91.83	85.82	36.73	51	102.70	99.93	39.58	154.56	122.59	97.51	30.17	33.35

A character s means “smile”, n means “nodding”, t means “talk”. i.e. s_response means that smile for response.

4.2.2 アンケート結果

ミーティングを行う前後の意見の偏りについて図 14，図 15 に示す．ミーティングを通して全体的に意見が中立に変化したことが分かる．また，B1-B8 に関する回答の平均値を図 16 に示す．特に，質問番号 B8 の「360 度カメラが気になった」という項目について，平均 1.375 と 360 度カメラは気にならなかったという意見が強かったことから，机の中央に 360 度カメラを置くことは心理的な影響を与えにくいと考えられる．

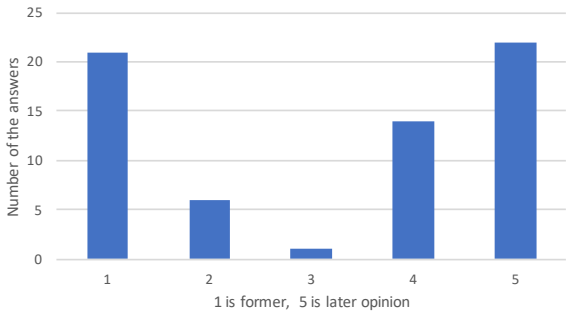


図 14 ミーティング前の意見の分布

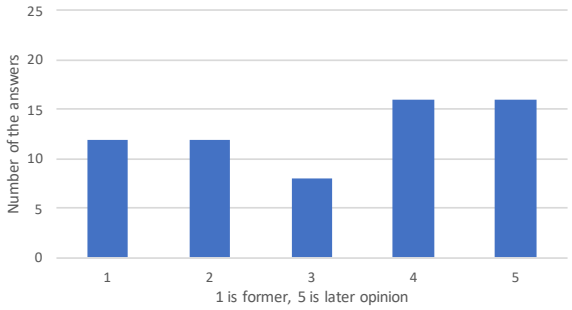


図 15 ミーティング後の意見の分布

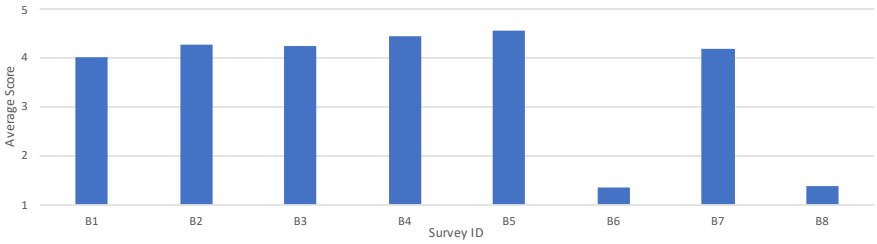


図 16 質問番号 B1-B8 についての回答の平均値

4.2.3 外部公開について

本研究で収集されたデータセットは，M3B(Multi Modal Meeting Behavior) Corpus と命名し，研究を目的とした者を対象に配布している [41]．M3B Corpus は顔画像や音声といった個人情報を排除している．顔の画像については，OpenFace を用いて点群に変換し，背景や顔の情報については黒背景にすることによって個人情報を排除する処理を行なった．

5. 慣性センサと動画によるマイクロ行動認識の結果と比較

慣性センサと動画によるマイクロ行動認識を行い，それぞれの結果について比較を行う．このとき，ウィンドウデータの正解ラベルの決定方法には第 3 章で定義 1 とした手法を用いる．

スライディングウィンドウ法で抽出されたデータは，被験者に関する情報や時系列に関するデータを無視し一つのデータセットとする．このデータセットを K 個のサブセットに分割し，1 つのサブセットに対して残りの K-1 個のサブセットから学習・予測を行う．これを全てのサブセットが試験データとなるように検証を繰り返す手法を本章で用いる (K-Fold Cross-Validation)．このときの K の値は 5 とする．

5.1 慣性センサを用いたマイクロ行動の認識

5.1.1 t-SNE を用いた特徴量の削減及び可視化

慣性センサから取得される加速度，角加速度から表 1 に示す特徴量を抽出した．教師なし機械学習アルゴリズムの 1 つである t-SNE[42] を用いて，抽出された特徴量を 2 次元にまで圧縮し可視化を行い，静止状態 (Idle) と Talk，Smile，Nodding のいずれかの動作を行なっている状態 (Move) がどのように分布しているかを確認する．

図 17 と図 18 に t-SNE を用いて特徴量を圧縮し，Idle(橙色の点) と Move(青色の点) を 2 次元空間上にプロットし可視化を行なったグラフを示す．それぞれのグラフは異なる人物の 1 回分のミーティングデータを使用した．プロットされた点群は曲線を描いているが，この曲線自体には特別な意味を持たない．図 17 は，比較的 Idle と Move が分離している場合を示している．一方，図 18 は Idle と Move が混在し線形的な分離が困難な場合を示している．後者は前者に比べて，センサデータから抽出した特徴量を用いて Idle と Move を認識するのが困難であることを意味している．

図 19 及び図 20 はそれぞれ図 17 と図 18 の元となるセンサデータの一部を時系列にプロットしたものである．上部のグラフは加速度センサの値，下部のグラフは各加速度センサの値を表す．被験者がラベリングしたデータをセンサデータのグラフの上に表示している．図 19 を確認すると，加速度及び角加速度のどちらとも，Talk，Smile，Nodding が発生した時に静止状態に比べて大きく変化していることが確認できる．一方，図 20 を確認すると，ラベリングが無い状態にも関わらず加速度，角加速度のどちらとも出力が変化していることが確認できる．これは，Talk，Smile，Nodding のいずれの動作を行なっていないにも関わらず，頭部が常に大きく動いているために加速度と角加速度を出力してしまう．このような場合，Idle と Move の識別が困難となり，t-SNE による可視化を行うと図 20 のように線形分離が困難なグラフがプロットされる．

5.1.2 Idle と Move の二値分類

慣性センサから抽出された特徴量から機械学習アルゴリズムを用いて Idle と Move の識別を行う．このときの機械学習アルゴリズムは Random Forest を使用する．以降の認識においても，機械学習アルゴリズムには Random Forest を使用する．パラメータは scikit-learn のデフォルト値である．以降の機械学習によるモデル構築時においてもパラメーターは常にデフォルト値を使用する．

認識で扱うデータセットは (a)「全データを用いるとき」と (b)「t-SNE による可視化を行った時に，Idle と Move の分離が困難と判断したデータを除去したとき」の 2 つの条件で認識を行なった．それぞれの条件での混同行列を図 21，図 22 に示す．また，それぞれの混同行列から Idle と Move の適合率，再現率，F 値を表 6 と表 7 に示す．

これらの混同行列および適合率，再現率，F 値を比較すると，いずれも t-SNE による可視化から識別が困難と判断した一部のデータを除去した場合の方が認識の精度が高かった．つまり，慣性センサを用いた認識を行う場合には常に頭部を動かすような人物については特に認識が困難であることが分かった．

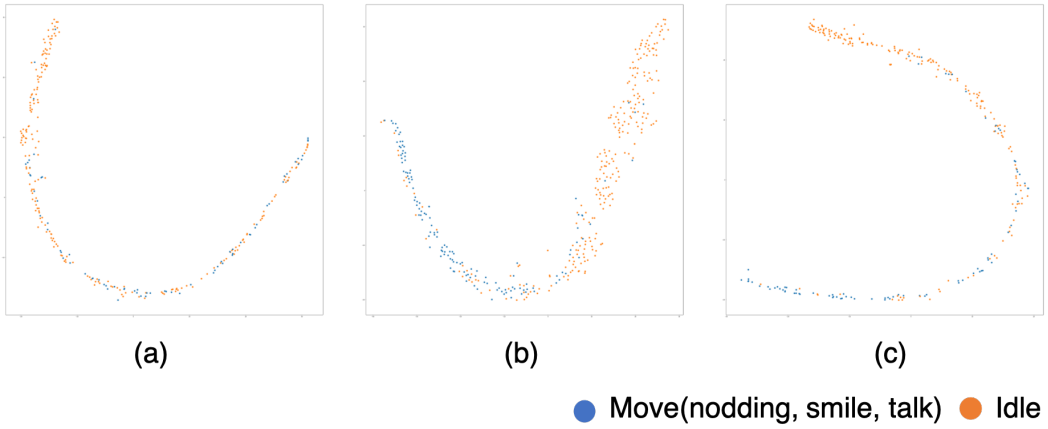


図 17 t-SNE による慣性センサデータの可視化：Idle と Move が分離している場合

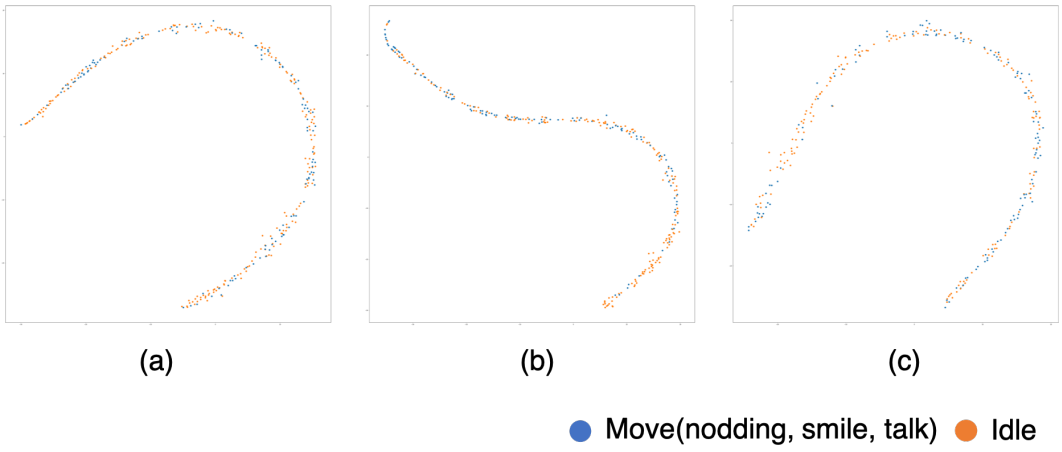


図 18 t-SNE による慣性センサデータの可視化：Idle と Move が混在している場合

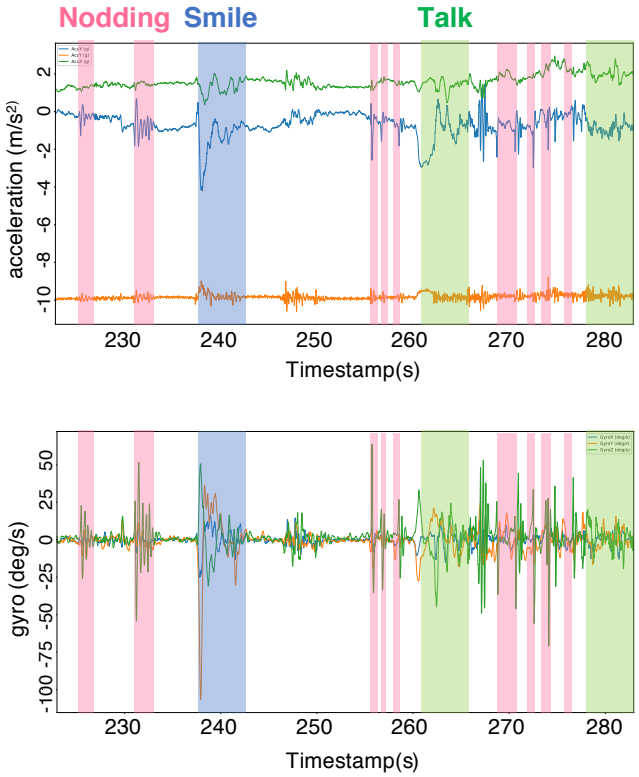


図 19 Move 時に加速度・角加速度が応答している様子

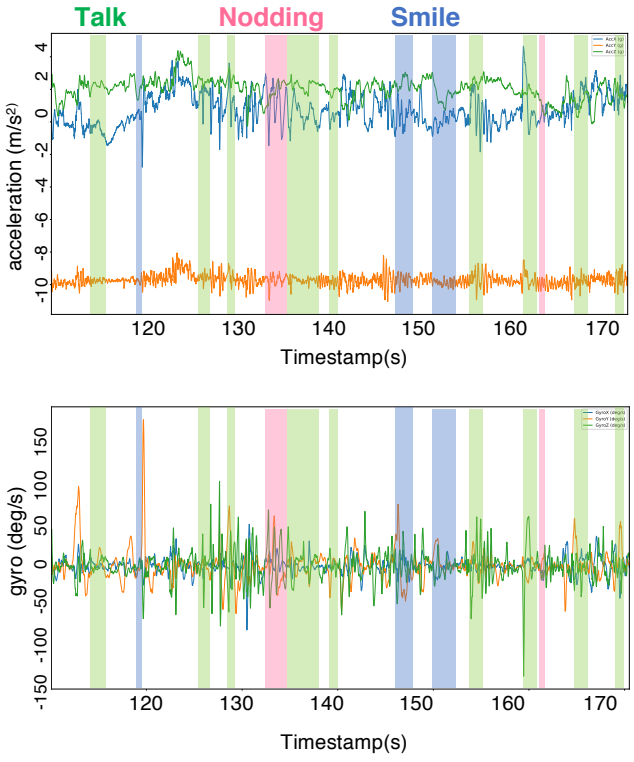


図 20 Move 時と Idle 時に加速度・角加速度が応答している様子

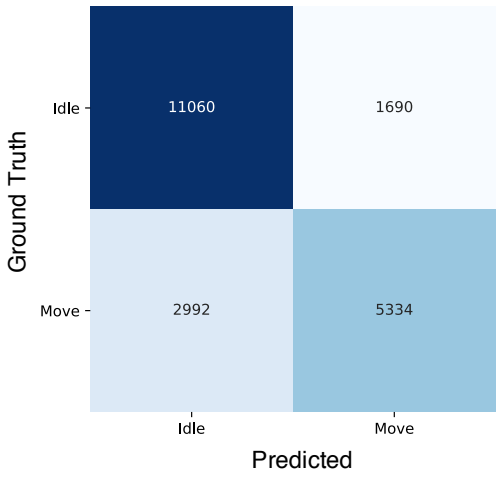


図 21 (a) 全てのデータを用いた Idle と Move 識別結果の混同行列

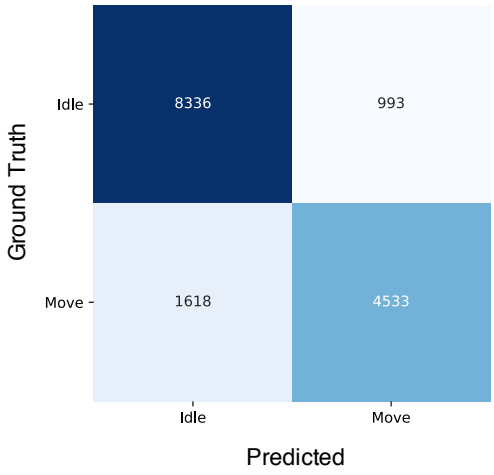


図 22 (b) 一部のデータを除去した場合の Idle と Move の識別結果の混同行列

表 6 (a) 全データを用いた識別の適合率・再現率・F 値のマクロ平均

	Precision	Recall	F-measure
Idle	0.867	0.787	0.825
Move	0.641	0.795	0.695
Macro Ave	0.754	0.773	0.760

表 7 (b) 一部のデータを除去した場合の識別の適合率・再現率・F 値

	Precision	Recall	F-measure
Idle	0.894	0.837	0.865
Move	0.738	0.821	0.777
Macro Ave	0.816	0.829	0.821

5.1.3 慣性センサによるマイクロ行動認識

慣性センサから取得されたセンサデータから Idle，Talk，Smile，Nodding の 4 クラスの行動認識を行なった．このとき，特に Nodding については他の動作と比べて極端に短く素早い動作であるため他の動作に比べてサンプル数が非常に少なく，クラス間のサンプル数に大きな偏りが発生してしまう．一般的に，不均衡なデータセットから機械学習によるクラス分類を行うと，サンプル数が多いクラスに学習が偏ってしまうため各クラスのサンプル数をバランシングを行う必要がある．したがって，ここでは慣性センサから得られたデータセットについて，Nodding のサンプル数にまで他クラスのサンプルをアンダーサンプリングする．認識を行なった結果に関する混同行列を図 23 に示し，各マイクロ行動の適合率，再現率，F 値を算出し表 8 にその結果を示す．

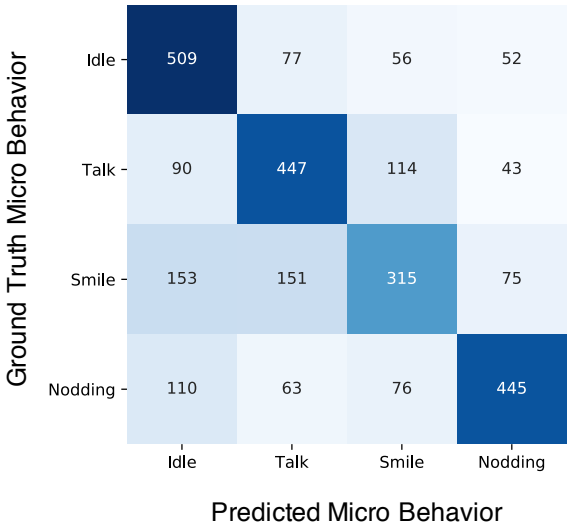


図 23 慣性センサを用いたマイクロ行動の認識結果の混同行列

表 8 慣性センサを用いたマイクロ行動の認識結果の適合率・再現率・F 値

	Precision	Recall	F-measure
Idle	0.733	0.590	0.654
Talk	0.644	0.606	0.624
Smile	0.454	0.561	0.502
Nodding	0.641	0.724	0.680

5.2 動画を用いたマイクロ行動の認識

5.2.1 Talk の認識

図 10 に示す OpenFace が出力する特徴点から口の開きを算出し，スライディングウィンドウで抽出した特徴量から Idle，Smile，Talk，Nodding の 4 種類の動作の認識を行なった．動画からウィンドウデータを抽出したときも慣性センサの場合と同様に Nodding のサンプル数が極端に少ないため，Nodding のサンプル数に

まで多クラスのサンプルについてアンダーサンプリングを行う．認識結果の混同行列を図 24 に示す．Talk の認識が目標のため，Talk についての適合率，再現率，F 値を算出し，その結果について表 9 に示す．

Talk の F 値は 68.4%と比較的高い値であることが確認できたが，Smile の F 値についても 67.9%と Talk とほぼ同じ数値であることが確認された．これは，Smile が発生する際にも口を開けるため，口の開きから特徴量に反映したと考えられる．しかし，アンダーサンプリングにより意図せずに Smile の認識率が向上した可能性が考えられる．そこで，アンダーサンプリングを行わずに認識を行なった結果の混同行列を図 25 に示す．図 24 と比較すると Talk の F 値は 69.4%である一方，Smile の F 値は 52.7%と低下し，Idle と誤認識している場合が増加していることから，口の開きのみで Smile を認識するのは困難であると考えられる．

5.2.2 Nodding の認識

図 9 に示す pose_Rx 成分 (顔の垂直方向の角度) からスライディングウィンドウ法で抽出した特徴量を用いて機械学習を行い，生成されたモデルから Idle，Smile，Talk，Nodding の 4 種類の動作の認識を行なった．Nodding のサンプル数にまで多クラスのサンプルに対してアンダーサンプリングを行う．認識結果の混同行列を図 26 に示す．Nodding の認識が目標のため，Nodding 適合率，再現率，F 値を算出し，その結果について表 10 に示す．

Nodding の F 値は 50.9%という結果になった．しかし，正解データが Nodding の場合に対して Talk と誤認識，もしくは正解データが Talk の場合に対してを Nodding と誤認識している場合も多く確認される．図 27 は，被験者の顔の垂直方向の角度とラベルデータの一部を表示したグラフである．このグラフから，Nodding 時には顔の垂直方向の角度が周期的な波形を描いている様子が確認できるが，Talk 時にも Nodding と似たような波形が発生している．これは，Talk 時にも顔が小さく上下に揺れていることが原因と考えられる．その結果，スライディングウィンドウによって抽出される特徴量が Nodding と類似し，誤認識が多く発生したと考えられる．

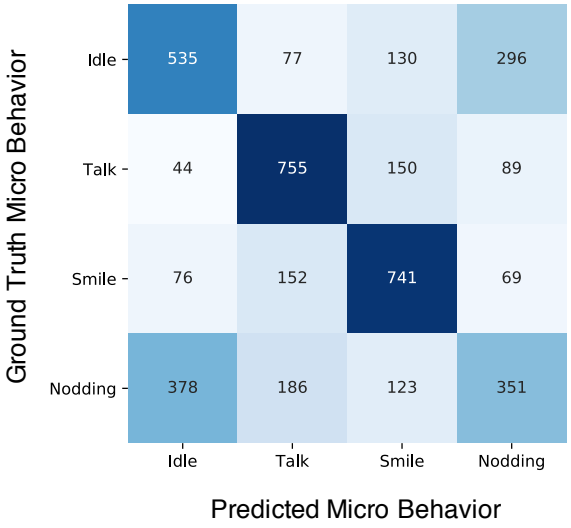


図 24 口の開きから Talk を対象に認識した結果の混同行列

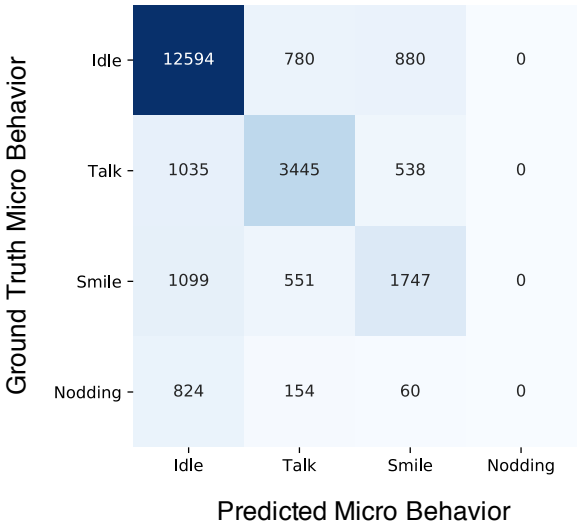


図 25 口の開きから Talk を対象に認識した結果の混同行列 (アンダーサンプリングなし)

表 9 口の開きから認識を行なった結果の Talk の適合率・再現率・F 値

	Precision	Recall	F-measure
Talk	0.727	0.645	0.684

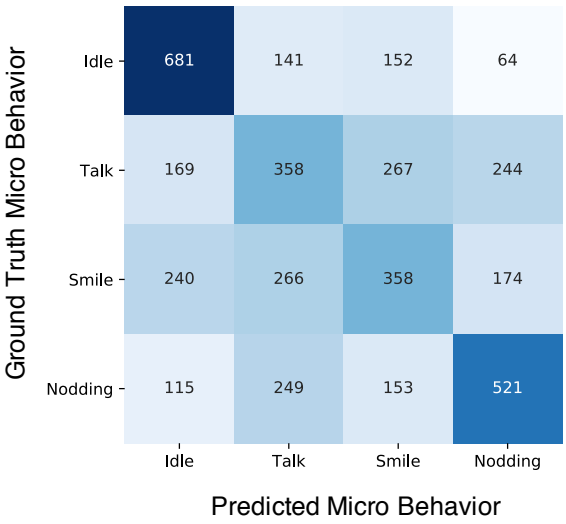


図 26 顔の角度から Nodding を対象に認識した結果の混同行列

表 10 顔の角度から認識を行なった結果の Nodding の適合率・再現率・F 値

	Precision	Recall	F-measure
Nodding	0.504	0.515	0.509

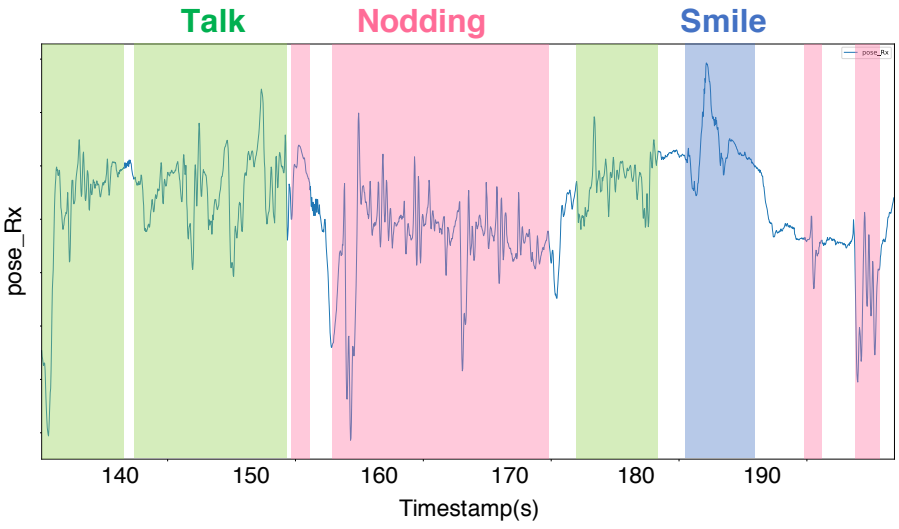


図 27 顔の垂直方向の角度とラベルデータ

5.3 Talk と Nodding を同時に認識

Talk の認識と Nodding の認識に用いられた特徴量を合わせたデータを使用し，これまでと同様の手法で Talk と Nodding の認識を同時に行なった．認識結果の混同行列を図 28 に示す．また，Idle，Talk，Nodding について適合率，再現率，F 値を算出し，その結果について表 11 に示す．図 26 と図 28 を比較すると，Talk を Nodding と誤認識または Nodding を Talk と誤認識している場合が大きく減少したことが確認できる．これは，特徴量を組み合わせたことにより顔の上下の動きより口の開き具合の特徴量が Talk の認識に強く影響したためと考えられる．

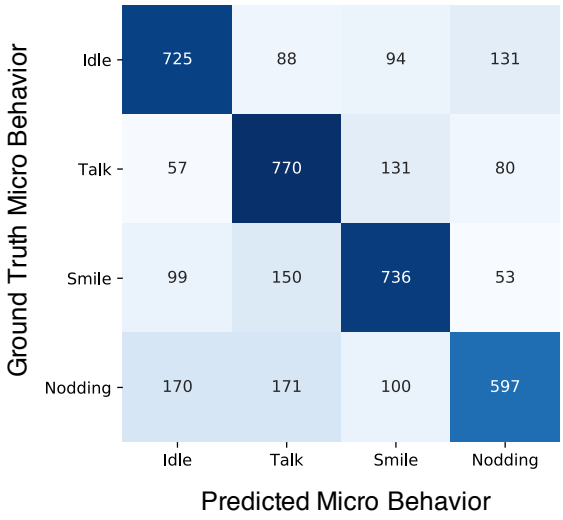


図 28 動画から Talk と Nodding を対象に認識を行なった結果の混同行列

表 11 動画から認識を行なった結果の Idle，Talk，Nodding の適合率・再現率・F 値

	Precision	Recall	F-measure
Idle	0.698	0.690	0.694
Talk	0.742	0.653	0.695
Nodding	0.575	0.693	0.629

5.4 慣性センサと動画からの認識の比較

本章では，ミーティング中に発生するマイクロ行動の認識を慣性センサと動画からの 2 種類のセンサを用いて行なった．慣性センサを用いた認識は，図 20 に示したように人物によっては Idle 時にも大きく頭部を動かしてしまい，誤認識が増加してしまう場合が確認された．実際のミーティングの中で使用する場合を考慮すると，汎用的にどのような人物についてもマイクロ行動を認識できることが好ましい．一方，動画を用いた認識手法については，顔の垂直方向の角度と口の開きのみを抽出しているため，頭の揺れの影響は小さく汎用的に認識が行えると考えられる．また，Smile の認識について，慣性センサからの認識精度は Talk と Nodding と比べて低い．動画からの Smile の認識は本研究では行なわれていないが，先行研究でも数多く笑顔の認識が行われ，その精度は高いことから，動画を用いた認識の方が慣性センサによる Smile の認識精度が高いと期待される．さらに，360 度カメラが撮影した動画を用いているため，1 つのカメラから複数人のマイクロ行動を認識することが可能である．一方，慣性センサの場合ではミーティングに参加する人数分の慣性センサを用意する必要がある．

以上より，Smile の認識精度が向上する可能性と導入が容易であることから，360 度カメラによって撮影された動画を用いた認識の方が慣性センサによる認識より有用であると考えられる．そこで，第 6 章では動画を用いた認識手法から実際のミーティングを想定したマイクロ行動の予測を行う．

6. 行動変化点を考慮したマイクロ行動の認識

5 章では，データに対して正解ラベルがある前提でスライディングウィンドウ法によるデータの抽出及び行動の認識を行なった．そのため，行動変化点を含むウィンドウについては無視することが可能であった．しかし，実際のミーティング中に発生するマイクロ行動の認識を行うとき，ウィンドウ内に行動変化点があることを検知することは困難であるため，行動変化点を含むウィンドウデータを無視しデータを抽出する手法は不適切である．そこで，本章では第 3 章で定義 2 とした手法を用いて，動画からスライディングウィンドウ法により再度データを抽出しマイクロ行動の認識を行なった．

ある人物が 5 分間のミーティングに参加している時に取得されたデータを 1 セッションと定義する．つまり，本実験で実施されたミーティングは全 16 回，1 回のミーティングにつき 4 人が参加しているため，合計で 64 セッションのデータが本研究では取得された．このうち 1 セッションを試験データ，残りの 63 セッションを教師データとし学習・予測を行う．これを全てのセッションが試験データとなるように検証を繰り返す手法を本章で用いる (Leave-One-Out-Cross-Validation, LOOCV).

6.1 Talk の認識

正解データは Talk 以外のラベルを Other(Idle，Smile，Nodding) と丸めこみ，Talk と Other の二値分類を行なった．図 29 に予測の一例を示す．上のグラフが正解ラベル，下のグラフが予測ラベルを表し，橙色の帯は Talk ラベル，空白の箇所は Other ラベルを表している．Talk の認識は概ね認識が行えている一方，Talk を行なっていないにも関わらず Talk と予測されていることが確認できる．つまり，この例については Talk の Recall は高いが Precision が低い．この傾向はほとんどのセッションでも確認された．本研究の目的はマイクロ行動の自動認識である為，Precision よりも Recall の値の方が高い場合の方が好ましいと考えられる．予測値を平滑化したグラフを図 30 に示す．極端に短い Talk の予測が丸め込まれることによってより正解ラベルに近い認識結果になった．

次に，Talk の予測精度が低い結果を図 31 に示す．このグラフは平滑化済である．Talk を行なっていないにも関わらず，Talk と誤った認識を頻発に行なっていることが確認できる．実際に動画を参照し，どのような場合に誤認識が発生しているかを確認した．図 31 の被験者の場合，Talk を行なっていない状態に唇を巻き込む，舌で唇を舐めるといった動作を行なっていることが確認された．このような口周辺に発生する癖が誤認識の原因であると考えられる．

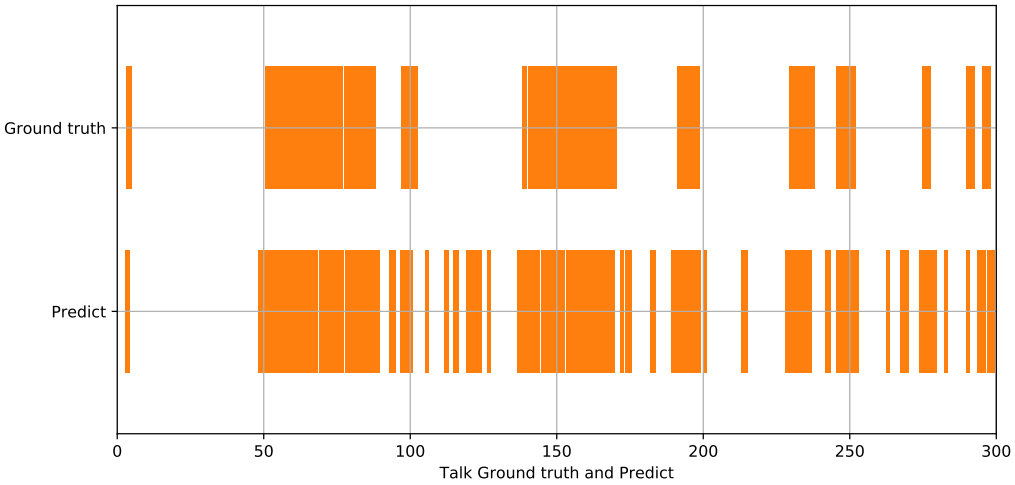


図 29 Talk の正解ラベルと予測ラベル (精度の高い例)

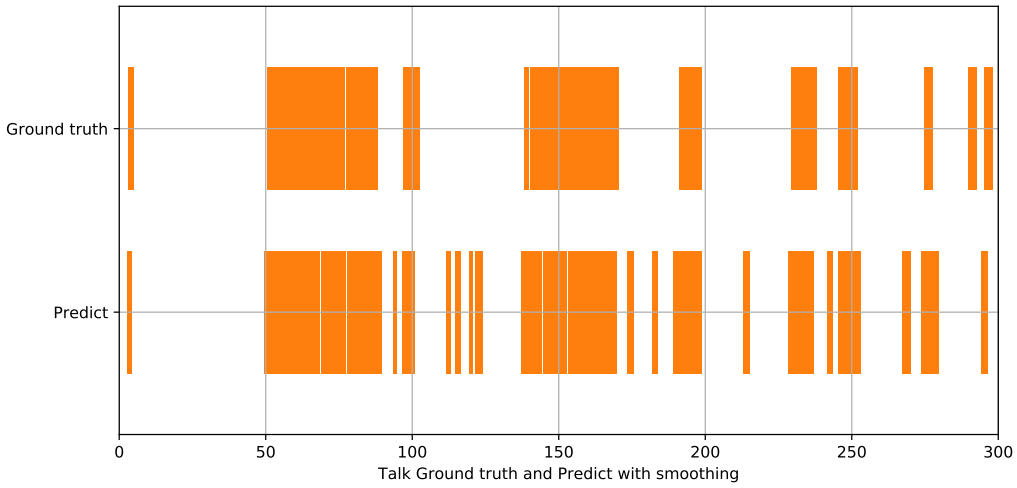


図 30 図 29 の予測値を平滑化した場合

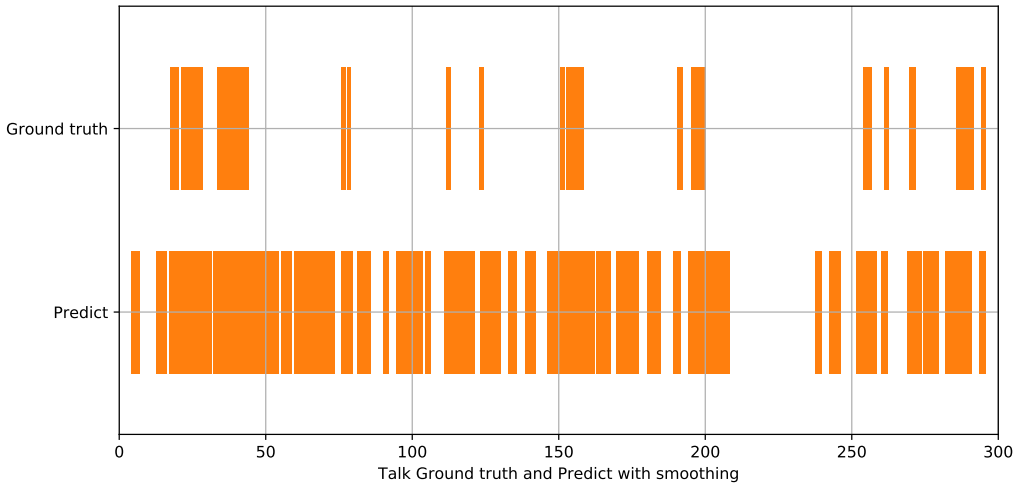


図 31 Talk の正解ラベルと予測ラベル (精度の低い例)

6.2 Nodding の認識

Talk の認識と同様に，正解データは Nodding 以外のラベルを Other(Idle，Talk，Smile) と丸めこみ，Nodding と Other の二値分類を行なった．図 31 にあるセッションの予測結果を示す．青色の帯は Nodding ラベル，空白箇所は Other ラベルを表している．Nodding が発生しているにも関わらず，予測値は Other と認識されている場合が非常に多いことが確認される．これはほとんどのセッションでも同じ傾向であった．原因として，Nodding は他のマイクロ行動 (Other) に比べてサンプル数が極端に少なく，前述したように不均衡データの学習は誤った認識を引き起こす可能性が高いためと考えられる．従って，Nodding の認識精度を向上させるにはバランシングを行う必要があり，バランシングには次の 2 通りの手法が考えられる．

1. データ数をアンダーサンプリングし機械学習モデルを作成する
2. データ数をオーバーサンプリングし機械学習モデルを作成する

それぞれの手法から Nodding の予測を行なった結果について記述する．

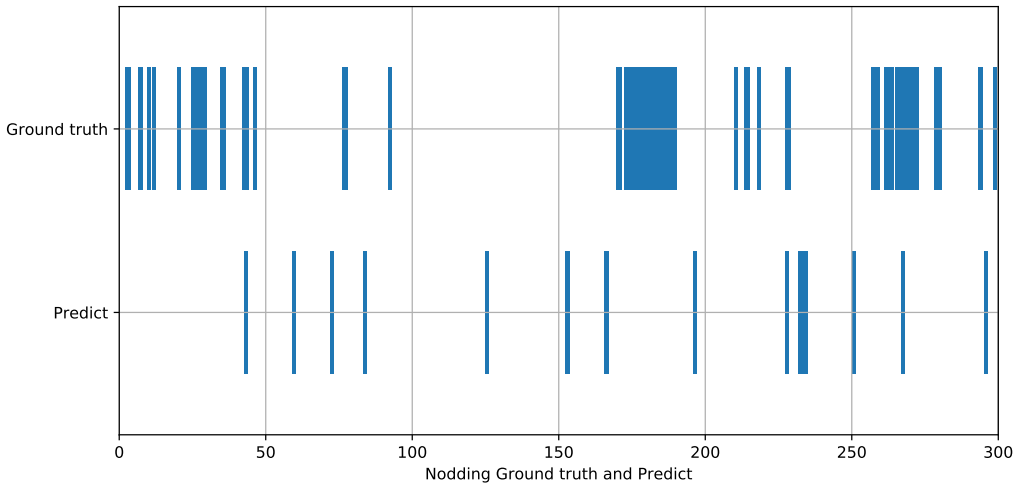


図 32 Nodding の正解ラベルと予測ラベル (バランシングなし)

Nodding は Other に比べてサンプル数が極端に少ないことから，Other のサンプル数を Nodding のサンプル数にまでアンダーサンプリングし，機械学習による認識モデルの作成を行った．アンダーサンプリングによってバランシングされたデータから作成された機械学習モデルによる予測結果を図 33 に示す．図 32 とは対照的に，Nodding が発生していない状況に関わらず，Nodding と予測を多く行なってしまう傾向が確認された．

Nodding のサンプル数が少ない状態を Other のサンプル数にまで増加させることによってバランシングを行い，機械学習による認識モデルの作成を行なった．このように擬似的にサンプル数の少ないクラスのデータを増加させることをオーバーサンプリングと呼ぶ．この時のオーバーサンプリングのアルゴリズムには ADASYN[43] を使用した．オーバーサンプリングによってバランシングされたデータから作成された機械学習モデルによる予測結果を図 34 に示す．図 32，図 33 と比較すると ADASYN を用いたオーバーサンプリングによるデータを用いた機械学習モデルの方が誤認識が減少し，より精度の高い認識が行えていることが確認された．これは他のセッションに関しても同様であった．

以上より，バランシングの手法について，オーバーサンプリングを用いた手法の方がより認識精度を向上することが確認された．

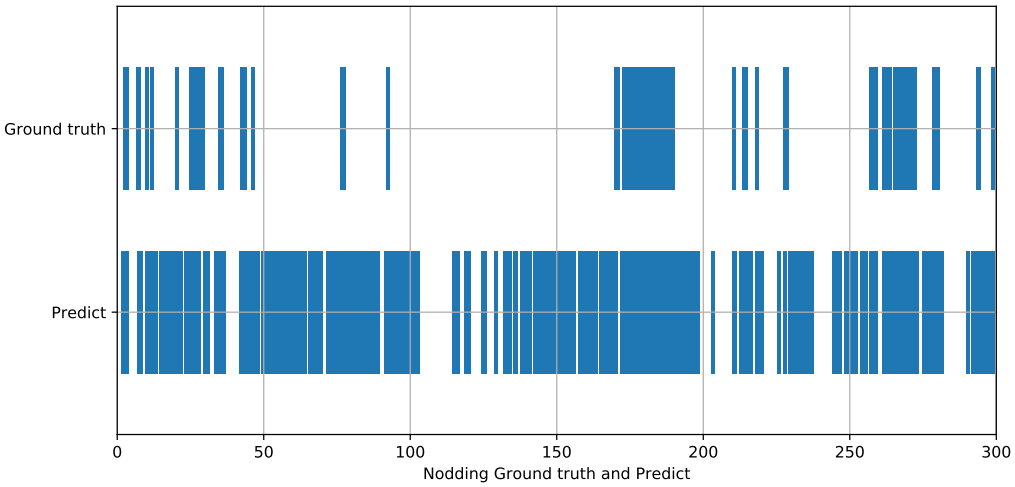


図 33 Nodding の正解ラベルと予測ラベル (アンダーサンプリング)

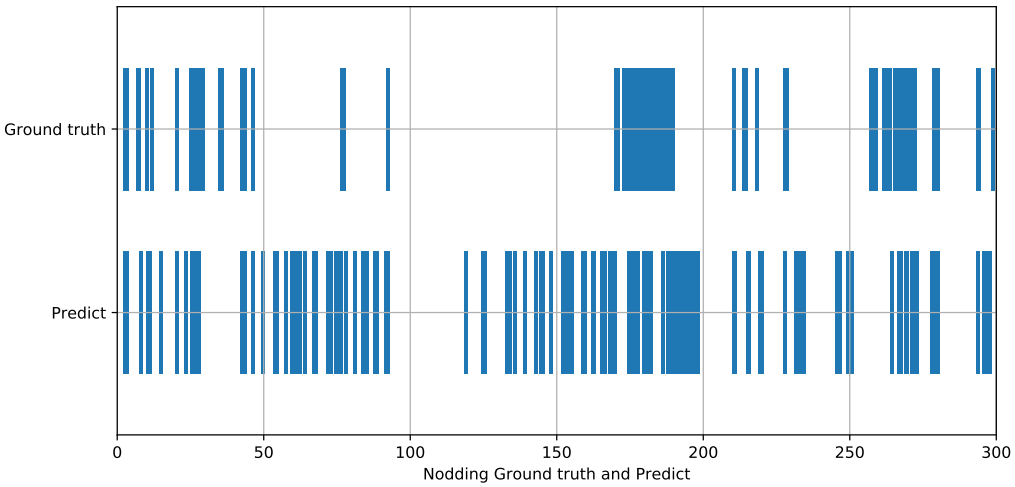


図 34 Nodding の正解ラベルと予測ラベル (オーバーサンプリング)

6.3 Talk と Nodding の認識

6.1 節と 6.2 節では，それぞれのマイクロ行動と Other ラベルの 2 値分類を行なった．本節では，Talk，Nodding，Other(Idle，Smile) の 3 値分類を行った．また，Nodding の認識精度を向上させるために，ADASYN アルゴリズムによるオーバーサンプリングを行なった．複数のセッションについて予測した結果を図 35〜図 37 に示す．このときグラフは平滑化済みのものである．図 35 のグラフは図 32〜図 34 と同じセッションのデータである．それぞれの予測結果を参照すると，Talk と Nodding の認識を同時に行うことによって，独立して認識を行う場合よりも精度が大きく向上していることが確認できる．また，図 34 で Nodding と誤認識を行なった箇所が Talk と正しく認識されていることが確認できる．

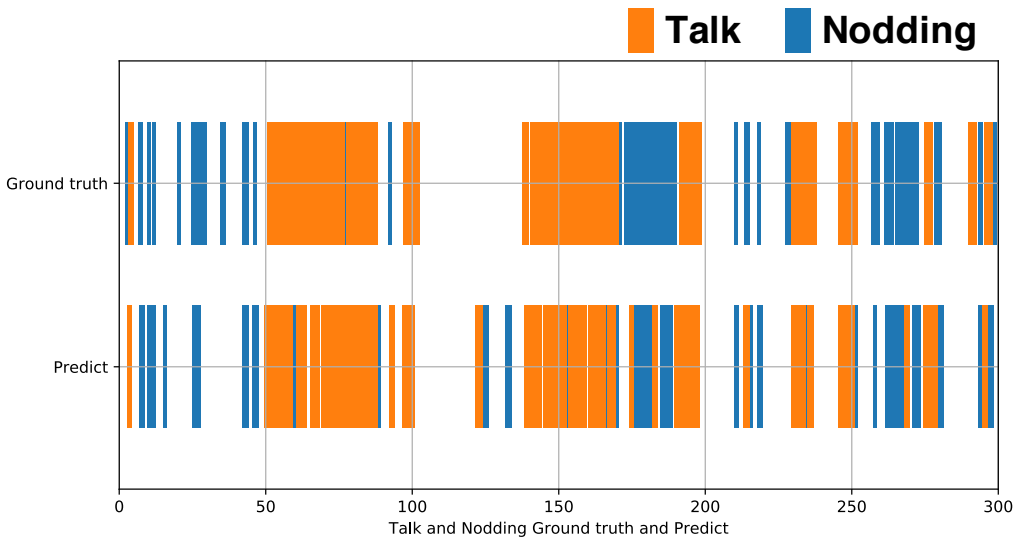


図 35 例 1：Talk と Nodding の認識結果

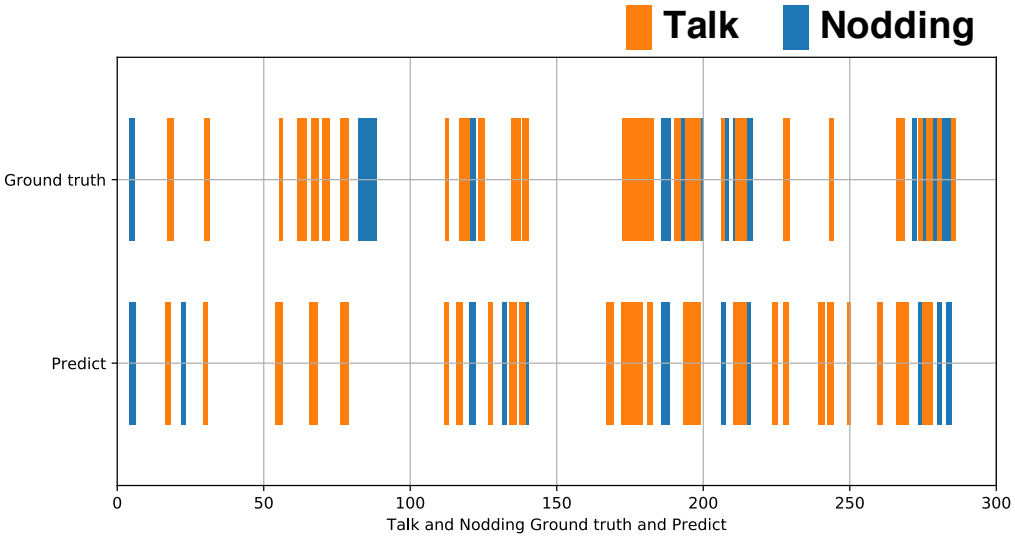


図 36　例 2：Talk と Nodding の認識結果

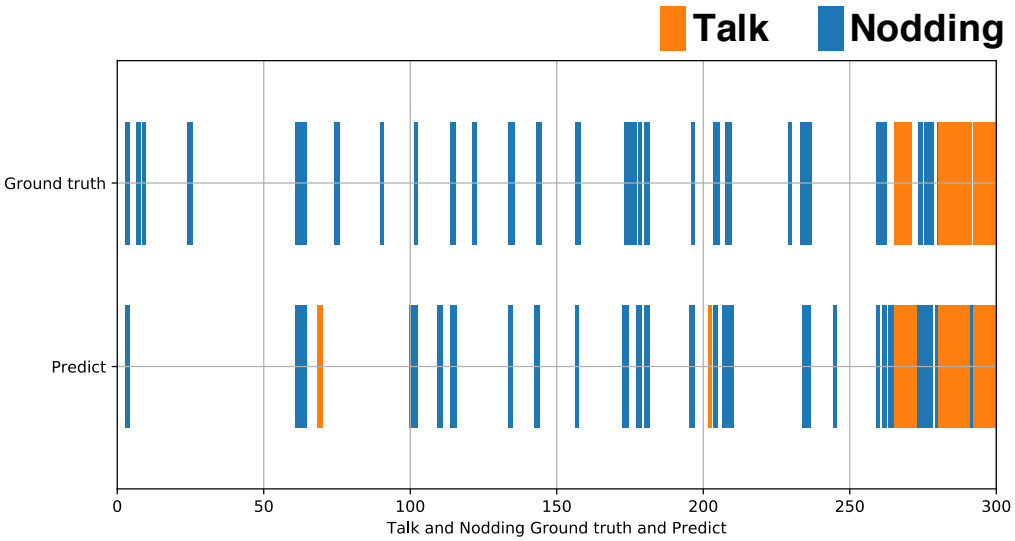


図 37　例 3：Talk と Nodding の認識結果

6.4 アプリケーションの開発

本研究で提案した手法を組み込まれたアプリケーションの開発を行った．アプリケーションの様子は図 38 に示す．このアプリケーションにより，ユーザーは THETA V によって撮影されたミーティングの動画を入力することによって，ワンストップで Talk と Nodding を認識することが可能となる．

図 38 開発したアプリケーション

7. 考察

7.1 各マイクロ行動の認識の考察

7.1.1 Talk の認識について

表 8 と表 11 から Talk の認識に関する F 値を確認すると，動画からの認識の方が慣性センサに比べ 7.1%高いことが確認された．これは，動画からの認識は口の開きを直接特徴量に反映させることが出来るためと考えられる．

Talk について，表 4 と表 5 を参照すると平均継続時間は 4.01 秒，５分間のディスカッションに対して 1 人が Talk を行う時間は平均で 72.63 秒と Talk のサンプル数が多くなり，また継続時間が Smile や Nodding と比べて長いことからラベリングが容易でかつ正確に行えるためラベリングの質が高いと考えられる．ラベリングの質は予測精度に大きく影響するため，Talk の認識精度はどちらのデバイスの場合でも比較的高かったと考えられる．

動画からの Talk の認識について，本研究では上唇と下唇のユークリッド距離を信号として用いている．しかし，この信号に対して正規化等の処理を行なっていない．カメラからの距離や他の特徴点からの距離を用いて正規化を行うことによって，ロバストな認識モデルを作成することが可能と期待される．

7.1.2 Smile の認識について

慣性センサからの Smile の認識について，図 23 の混同行列と表 8 から Smile の認識精度は他のマイクロ行動に比べて低いことが確認される．これは，笑い方には個人による癖や特徴が大きく発生し，また笑い方自体に規則性が低いことが原因と考えられる．

動画からの認識について，図 24 と図 28 に示す混同行列から，口の開きの特徴量から Smile の認識率が高いことが確認できる．これについては図 25 に示すアンダーサンプリングを行わずマイクロ行動の認識を行なったときの混同行列から，Smile の認識率が低下したことが確認され，口の開きからでは Smile の認識は困難であると考察した．

さらに，第 6 章の Talk と Nodding の予測に加えて Smile の認識を追加した場合の結果について図 39 に示す．緑色の帯が Smile のラベルを表す．この結果から，正解ラベルが Smile ではない状態に対して Smile と予測されていることが非常に多く，つまり適合率が極端に低いことが確認された．これは，Smile には口を閉じながら口角を上げる場合や口を開いて笑う場合など複数のパターンが存在するためと考えられる．したがって，口の開きを特徴量として用いた機械学習では，不適切な学習が行われてしまうため Smile でないサンプルに対して誤認識を行なってしまうと考えられる．

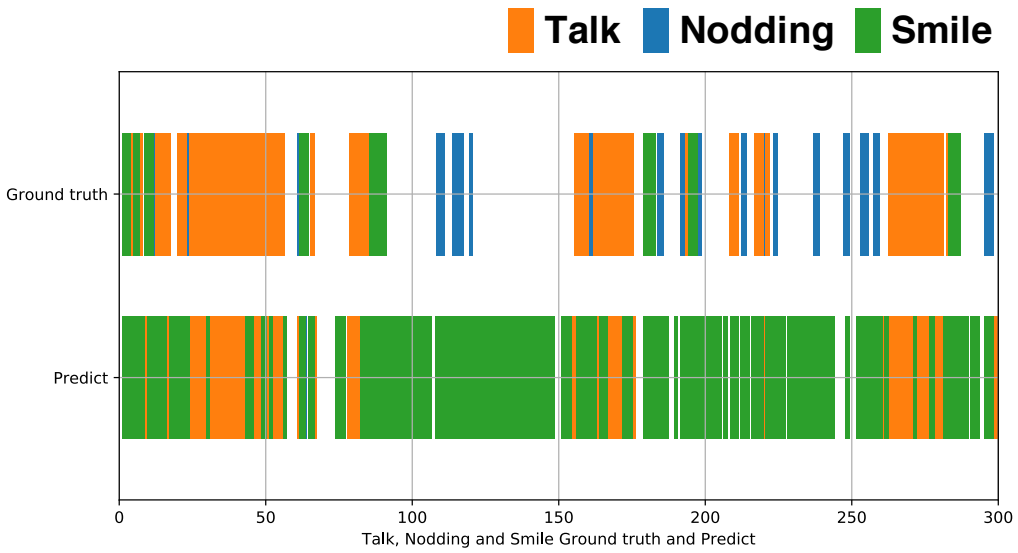


図 39 Smile の認識を追加した予測結果

7.1.3 Nodding の認識について

表 8 と表 11 から Nodding の認識に関する F 値を確認すると，慣性センサからの認識の方が動画からの認識に比べて 5.1%高いことが確認された．これは，Nodding の特徴的な周期的な揺れを慣性センサの方が動画よりも正確に捕捉出来たためと考えられる．また，表 8 から Nodding の F 値は Talk と Smile と比較しても高いこ

とからも，Nodding は他のマイクロ行動と比べて特徴的であると考えられる．

第 6 章での認識結果では，正解ラベルでは Nodding である場合を予測ラベルでは正しく認識できない場合が確認された．例えば，図 36 のグラフでは 80 秒付近で比較的継続時間の長い Nodding が発生しているが，予測では Nodding と認識されていない．動画から目視でこの Nodding について確認を行うと，ゆっくり深い特徴的な Nodding であったことが確認された．また，角度が他の場合と比べて大きく変化しない Nodding についても認識されていない傾向が確認された．このように，Nodding についても個人による特性の違いや性別による違い，Nodding を行なった理由に着目し，さらに詳細な解析を行うことにより Nodding の認識精度が向上すると期待される．

また，第 6 章では，Nodding の認識において「Nodding と Other(Idle，Talk，Smile)」の 2 値分類と「Nodding と Talk と Other(Idle，Smile)」の 3 値分類の 2 通りの方法から検証を行なった．このとき，不均衡状態の解消に ADASYN によるオーバーサンプリングが行われている．図 40 に 2 値分類と 3 値分類による Nodding の認識結果を比較したグラフを示す．上段のラベルは正解ラベル，中段は 3 値分類による予測ラベル，下段は 2 値分類による予測ラベルである．3 値分類の場合は，Nodding の正解ラベルに対して認識がなされていない (Precision が高い)．一方，2 値分類の場合は，3 値分類の場合に比べて Nodding の正解ラベルを捕捉しているが，過剰に Nodding と予測している (Recall が高い)．つまり，Nodding について 3 値分類では Precision が高く，2 値分類では Recall が高い．このような傾向を持つセッションは複数確認された．目視で確認を行うと，頷きが浅い人物についてこのような傾向が確認された．多くのセッションについては 3 値分類による予測の方が正確に Nodding を捕捉できる傾向にあるが，このような場合を加味し Nodding の 2 値分類と 3 値分類を組み合わせるようなアルゴリズムを考案することによりさらに認識精度が向上すると期待される．

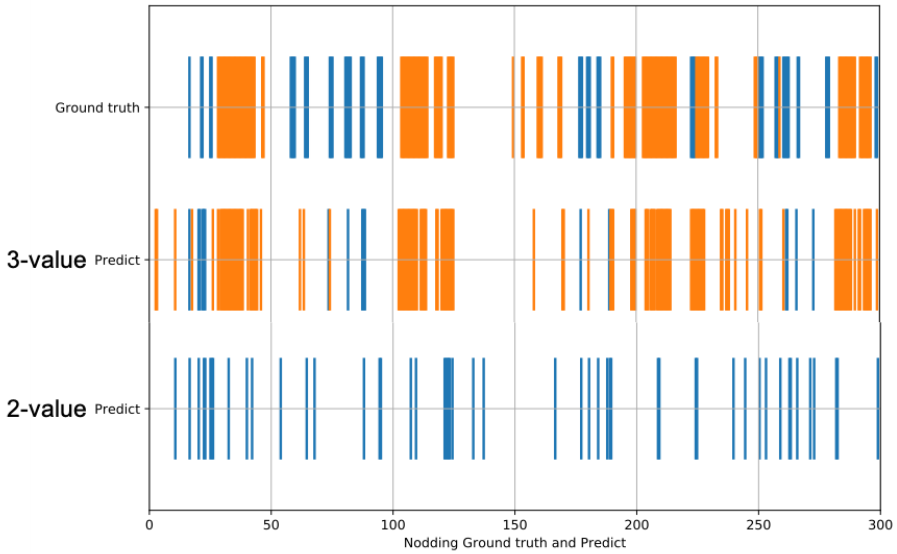


図 40　2 値分類と 3 値分類による認識結果の違い

7.2　オーバーサンプリングによる Nodding の認識精度の向上について

6.2 節では，Nodding の認識精度の向上のためオーバーサンプリングを行なった．その結果，不均衡データで学習する場合とアンダーサンプリングによるバランシングに比べて大きく認識精度が向上した．アンダーサンプリングを用いた場合の精度が悪い原因には

1. サンプル数が少ないため機械学習によって生成されたモデルの精度が低い
2. 他の行動ラベルが混在しているウィンドウデータを使用している

の 2 つが考えられる．1 つ目の原因については，機械学習による精度の高いモデルを生成するためには十分な数のラベル付きデータが必要である．しかし，Nodding は継続時間が短く発生回数も少ないマイクロ行動のため，アンダーサンプリングを行うと精度の低いモデルが生成されたと考えられる．2 つ目の原因に

ついては，定義 2 による正解ラベルの決定を行なっているため，複数の行動ラベルが含まれるウィンドウデータが生成される．したがって，他のクラスのマイクロ行動と近似した特徴量を持つサンプルが定義 1 の場合よりも増加する．本研究で使用した ADASYN と呼ばれるオーバーサンプリングは，サンプル数が少ないクラスと他クラスとの境界に対して重点的に擬似データを増加させるアルゴリズムである．つまり，サンプル数の多いクラスに誤って認識されてしまう可能性の高い領域について重点的に擬似データを増加させているため，より精度の高い機械学習モデルが生成されたと考えられる．

8. 結論

本研究では，ミーティングの定量的な評価と支援に向けて，ミーティング中に発生するマイクロ行動 (Talk，Smile，Nodding) の認識手法を提案した．認識には，(1)：頭部に慣性センサを装着し解析を行う手法と，(2)：動画から解析を行う手法の 2 つの手法を用い，Random Forest による認識結果を比較した．(1) の手法による認識結果から F 値を算出すると，Talk は 62.4%，Smile は 50.2%，Nodding は 68.0%，Idle は 65.4%となった．(2) の手法による認識結果から F 値を算出すると，Talk は 68.5%，Nodding は 62.9%，Idle は 69.4%となった．(2) からの手法では，画像から Smile の認識を認識する先行研究が多数報告されているため本研究では行わなかった．(1) と (2) の認識結果を比較すると，Idle，Talk，Nodding の F 値には大きな違いは見られなかった．しかし，(1) からの手法では Smile の認識率が低く，画像から Smile の認識は先行研究で高い認識精度が報告されているという背景から，動画によるマイクロ行動の認識手法の方がより精度の高い予測が行えると考えられる．また，慣性センサを用いた認識の場合では人数分の慣性センサを用意する必要があるが，360 度カメラを用いた認識の場合は 1 つのカメラから複数人のマイクロ行動の認識を行える．これらの違いから，動画を用いた認識手法の方が慣性センサを用いた認識手法よりも有用と考えた．

次に，実際のミーティングに対してマイクロ行動の認識を行うことを想定し，行動変化点を考慮したマイクロ行動の認識を行なった．Talk の認識については，サンプル数が他のクラスに比べて十分に多く精度の良い機械学習モデルを生成することが可能となり，精度の高い認識が行えた．しかし，Nodding については，Talk に比べてサンプル数が不十分であり，他クラスのラベルを含んだウィンドウデータを使用しているため誤った認識が多く発生したと考えられる．この状態を，オーバーサンプリングの手法である ADASYN を使用し，Nodding のサンプル数を擬似的に増加させることにより認識精度が向上した．

今後の展望としては，現時点では Smile の認識を動画からは行なわれていないが，先行研究を参考にフレーム毎に深層学習等を用いた Smile の認識を提案手法と組み合わせることが考えられる．また，現段階では頭部に発生するマイクロ行動の認識を行なったが，さらにミーティングに対する満足度や理解度とマイクロ

行動がどのような関係性があるのかについて解析を行うことによって，ミーティングの定量的な評価や参加者への支援を行えると期待される．本研究では，発言内容には個人情報が含まれる可能性があるため解析を行わなかったが，声の大きさやトーンといった個人情報を含まない言語コミュニケーションと提案手法を組み合わせることが期待される．さらに本研究では解析を行わなかった視線計のデータについても，ミーティング中の視線がコミュニケーションにおいてどのように影響するのか，マイクロ行動とどのような関係があるのかについて調査するなど様々な側面から解析を行えることが考えられる．

図 38 に示したアプリケーションを用いて，実際のミーティングに関するデータを容易に収集できることが期待される．簡単なコンピュータの扱いが可能な人物であれば使用出来る設計となっているため，様々な状況でのミーティングに関するデータを収集可能である．これらのデータは，ミーティングの理解度や集中度を推定する上で重要なデータとなることが期待される．

謝辞

本研究を進めるにあたり，安本慶一 教授には，研究全般に関し，多大なるご指導・ご助言を賜りました．また，充実した研究環境の整備など，研究活動を手厚くご支援いただきました．感謝の意を表すとともに，心より厚く御礼申しあげます．

中村哲 教授には，ご多忙の中，論文審査委員を引き受けてくださった上で，副指導教官として様々なご助言をいただきました．本研究関連のプロジェクトでも，ご指導いただき，的確なアドバイスを賜りました．感謝の意を表すとともに，心より厚く御礼申しあげます．

荒川豊 准教授には，研究の指導教官として大変お世話になりました．自分一人では思いつかないようなアイデアや的確な助言によって無事に研究を進めることができました．また，巨額の研究費によって実験や解析がスムーズに行えました．感謝の意を表すとともに，心より厚く御礼申しあげます．

松田裕貴 助教授には，ドイツから帰ってきて間も無く多大なる補助や助言を賜りました．国際論文の提出の際には私の拙い英語論文を丁寧に添削していただきました．ここまで充実した研究生活を送れましたのは松田先生の援助が非常に大きいです．感謝の意を表すとともに，心より厚く御礼申し上げます．

諏訪博彦 特任准教授には，研究に関する相談にも，丁寧に回答していただきました．また，疲れている時にも優しく話しかけていただき，とても学生のことをみていただいているという精神的安心感を多くの学生も感じていたと思います．感謝の意を表すとともに，心より厚く御礼申しあげます．

藤本まなと 助教には，夜遅くにも関わらず学生と近い距離でコミュニケーションを取っていただきました．持ち前の明るさでこちらまで前向きに頑張れる場面も多々ありました．感謝の意を表すとともに，心より厚く御礼申しあげます．

金岡恵 事務補佐員，山内奈緒 事務補佐員には，学会や出張に関する事務処理を始め，研究生活の様々な場面でご支援いただきましたこと，謹んで感謝申し上げます．

また研究全般において，的確なアドバイスをくださった中村優吾 先輩，河中祥吾 先輩を始めとする先輩の皆様，共に研究生活を過ごし大変なラベリング作業を

非常に正確な精度で行ってもらったユビキタスコンピューティングシステム研究室の同輩，後輩には頭が上がりません．心より感謝申しあげます．

最後に，今日まで学生生活を様々な面から支えてくださった両親，家族に心より感謝申し上げます．

参考文献

[1] 株式会社ジェイアール東海エージェンシー. ビジネスパーソンの「社内会議」に関する調査. ビジネスパーソンウォッチング調査, Vol. 14, , 2017.

[2] Steven G. Rogelberg, Clifton W. Scott, Brett Agypt, Jason Williams, John E. Kello, Tracy McCausland, and Jessie L. Olien. Lateness to meetings: Examination of an unexplored temporal phenomenon. *European Journal of Work and Organizational Psychology*, Vol. 23, No. 3, pp. 323–341, 2014.

[3] Arpan. C. Active learning through discussion. pp. 1–4, 2017.

[4] 佑磨林. 発話間関係の構造化による会議録からの議論マップ自動生成システム, feb 2016.

[5] Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. Long story short – global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, Vol. 52, No. 10, pp. 801 – 815, 2010.

[6] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pp. 97–100, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics.

[7] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pp. 28–39, Berlin, Heidelberg, 2006. Springer-Verlag.

[8] P.Ekman and L. W. Friesen. The repertoire of nonverbal behavior. *Semiotica*, pp. 49–98, 1969.

[9] Peter.E Bull. *Posture and Gesture*. Pergamon Press, 1987.

[10] Herbert H. Clark. *Using Language*. 'Using' Linguistic Books. Cambridge University Press, 1996.

[11] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech Language*, Vol. 33, No. 1, pp. 1 – 24, 2015.

[12] Kristiina Jokinen. Turn taking , utterance density , and gaze patterns as cues to conversational activity. 2011.

[13] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[14] Y. . Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 97–115, Feb 2001.

[15] Andrés Romero, Juán León, and Pablo Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 2018.

[16] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, p. 65. Paris, France, 2010.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and

emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, June 2010.

[18] Maia Garau, Mel Slater, Simon Bee, and Martina Angela Sasse. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pp. 309–316, New York, NY, USA, 2001. ACM.

[19] Starkey Duncan. Some signals and rules for taking speaking turns in coversations. *Journal of Personality and Social Psychology*, Vol. 23, No. 2, pp. 283 – 292, 1972.

[20] Starkey Duncan. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, Vol. 3, No. 2, pp. 161–180, 1974.

[21] Starkey Duncan and George Niederehe. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, Vol. 10, No. 3, pp. 234 – 247, 1974.

[22] Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed Ehsan Hoque. Coco: Collaboration coach for understanding team dynamics during video conferencing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 1, No. 4, January 2018.

[23] K. Murano A. Onishi and T. Terada. A method for structuring meeting logs using wearable sensors. *Internet of Things*, pp. 140–152, 2019.

[24] Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, Vol. 171, No. 8, pp. 568 – 585, 2007.

[25] Zhiwen Yu, Zhiyong Yu, Hideki Aoyama, Motoyuki Ozeki, and Yuichi Nakamura. Capture, recognition, and visualization of human semantic interactions in meetings. In *2010 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 107–115. IEEE, 2010.

[26] RICOH. Theta v, 2017.

[27] Pupil Labs. Pupil mobile eye tracking headset, 2017.

[28] LP-RESEARCH Inc. Lpms-b2, 2019.

[29] Masashi Takata, Manato Fujimoto, Keiichi Yasumoto, Yugo Nakamura, and Yutaka Arakawa. Investigating the capitalize effect of sensor position for training type recognition in a body weight training support system. UbiComp/ISWC 2018 - Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers, pp. 1404–1408. Association for Computing Machinery, Inc, 10 2018.

[30] X. Wang, L. Gao, J. Song, and H. Shen. Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition. *IEEE Signal Processing Letters*, Vol. 24, No. 4, pp. 510–514, April 2017.

[31] Nobuo Kawaguchi, Nobuhiro Ogawa, Yohei Iwasaki, Katsuhiko Kaji, Tsutomu Terada, Kazuya Murao, Sozo Inoue, Yoshihiro Kawahara, Yasuyuki Sumi, and Nobuhiko Nishio. Hasc challenge: Gathering large scale human activity corpus for the real-world activity understandings. In *Proceedings of the 2Nd Augmented Human International Conference*, AH '11, pp. 27:1–27:5, New York, NY, USA, 2011. ACM.

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[33] Dean Karantonis, Michael Narayanan, Merryn Mathie, Nigel Lovell, and B.G. Celler. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *Information Technology in Biomedicine, IEEE Transactions on*, Vol. 10, pp. 156 – 167, 02 2006.

[34] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Esann*, 2013.

[35] Janelle Mason, Christopher Kelley, Bisoye Olaleye, Albert Esterline, and Kaushik Roy. An evaluation of user movement data. In Malek Mouhoub, Samira Sadaoui, Otmane Ait Mohamed, and Moonis Ali, editors, *Recent Trends and Future Technology in Applied Intelligence*, pp. 729–735, Cham, 2018. Springer International Publishing.

[36] Igor Pernek, Gregorij Kurillo, Gregor Stiglic, and Ruzena Bajcsy. Recognizing the intensity of strength training exercises with wearable sensors. *Journal of Biomedical Informatics*, Vol. 58, pp. 145 – 155, 2015.

[37] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 17–24, May 2017.

[38] Michal Uricár, Radu Timofte, Rasmus Rothe, Jiri Matas, and Luc Van Gool. Structured output svm prediction of apparent age, gender and smile from deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 25–33, 2016.

[39] Peter Robinson Tadas Baltrušaitis and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[40] The Language Archive. Elan, 2002.

[41] Yusuke Soneda, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. M3b corpus: Multi-modal meeting behavior corpus for group meeting assessment. UbiComp/ISWC 2019- - Adjunct Proceedings of the 2019 ACM International

Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp. 825–834. Association for Computing Machinery, Inc, 9 2019.

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605, 2008.

[43] Haibo He, Yang Bai, Edwardo Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. pp. 1322 – 1328, 07 2008.

研究業績

国際会議

1. Yusuke Soneda, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. Multimodal Recording System for Collecting Facial and Postural Data in a Group Meeting. The 27th International Conference on Computers in Education (ICCE 2019) , Pingtung, Taiwan, December 2-6, 2019.

2. Yusuke Soneda, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. M3B Corpus: Multi-Modal Meeting Behavior Corpus for Group Meeting Assessment. ACM UbiComp/ISWC ’19 Adjunct (HASCA) , London, United Kingdom, September 9–13, 2019.

国内会議

1. 曽根田 悠介, 荒川 豊, 安本 慶一：オンラインミーティングを対象とした参加者のエンゲージメント計測システムの検討, 2018 年度 情報処理学会関西支部 支部大会， 大阪大学中之島センター， 大阪府， 2018 年 9 月.

2. 曽根田悠介，荒川 豊，安本 慶一：オンラインミーティングを対象とした会議の質評価システムの設計と構築，情報処理学会 第 81 回全国大会，福岡大学 七隈キャンパス，福岡県，2019 年 3 月.

3. Yusuke Soneda, Stirapongsasuti Sopicha, Yutaka Arakawa, and Keiichi Yasumoto: Observing Semantic Human Expression in a Group Meeting by using an Omnidirectional Camera，インタラクション 2019，学術総合センター／一橋大学 一橋講堂，東京都，2019 年 3 月.

4. 徳原 耕亮, Billy Dawton, 荒川 豊, 石田 繁巳, 曽根田 悠介, 松田 裕貴：グループミーティング動画からの発話量抽出手法の検討，情報処理学会 第 82 回全国大会，金沢工業大学 扇が丘キャンパス，石川県，2020 年 3 月.

5. 曽根田 悠介，中村 優吾，松田 裕貴，荒川 豊，安本 慶一 ：ミーティング映像からの発話およびマイクロ動作識別手法，情報処理学会 第 65 回ユビキタスコンピューティングシステム (UBI) 研究発表会，名古屋大学 IB 電子情報館，愛知県，2020 年 3 月．