Vol.2020-MBL-94 No.40
Vol.2020-UBI-65 No.40
2020/3/3

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

曽根田 悠介[1,a)]　中村 優吾[1]　松田 裕貴[1]　荒川 豊[2]　安本 慶一[1]

概要：近年，働き方改革と叫ばれるように日本の働き方の改善が注目されている．特にオフィスワーカに関して，ミーティングの時間は 1 日の労働時間のうち平均で 68.2 分と報告され，労働時間の多くの時間がミーティングに費やされている．また，若手の社員は長時間のミーティングや発言権のないミーティングに対して苦手意識を持つ傾向にあり，無駄なミーティングは経済的損失を引き起こすという報告も存在する．このように，ミーティングはオフィスワークにとって重要であり，ミーティングの質の向上が働き方改革において重要な事項である．ミーティングや会話といったコミュニケーションは言葉を用いた言語コミュニケーションと視線や頷きといった非言語コミュニケーションに分けられる．本研究では，発話内容に個人情報が含まれる可能性があるという観点から非言語コミュニケーションに着目し，ミーティングの質の定量化に向けて，映像情報のみからミーティング中に発生するマイクロ行動（Nodding，Talk）を認識する手法を提案する．K-分割交差検証法によって提案手法を評価した結果，Nodding の F 値は 62.9%，Talk の F 値は 69.5%であった．次に，時系列上で認識結果を可視化し，正解と認識結果の比較から考察した結果について報告する．

1. はじめに

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，オフィスワーカの QOL の推定 [1] や就業中における姿勢の推定 [2][3] といった，職場でのセンシングに関する研究が数多く報告されている．一方，ある企業の調査によると [4]，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が，定量的にミーティングの質の評価を行う研究は確認されない．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，働き方改革を推進する上で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．

ミーティングを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報告されている [5]．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を解析はプライバシー保護の観点からリスクが発生する．

[1] 奈良先端科学技術大学院大学
NAIST, Takayamacho, Nara 630–0192, Japan
[2] 九州大学
Kyushu University, Motooka, Fukuoka 819–0395, Japan
[a)] soneda.yusuke.su2@is.naist.jp

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 [6] や Peter [7] 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．しかし，自動で非言語コミュニケーションを認識する手法は確立されておらず，人手によるラベリング作業が必須なため，非言語コミュニケーションが人対人のコミュニケーションにどのような影響を与えるかについて解析を行うには膨大な時間と人手を要する．

本論文では，360 度カメラによって撮影された動画からミーティング中の頷き，発言，笑いといったマイクロ行動を認識する手法を提案する．具体的には、動画から対象者の頭部角度や口の開き具合を表す時系列データを生成し、これをセンサデータとみなして、機械学習ベースのアプローチによってマイクロ行動の認識を行う．これまで，マイクロ行動がミーティングに及ぼす影響を調査するためには全て手動でラベル付けする必要があったが，本手法は，ミーティング中に撮影された動画から、参加者のマイクロ行動を自動で記録することが可能であることから高い有用性が期待される．

本論文の構成は以下の通りである．第 2 章では，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

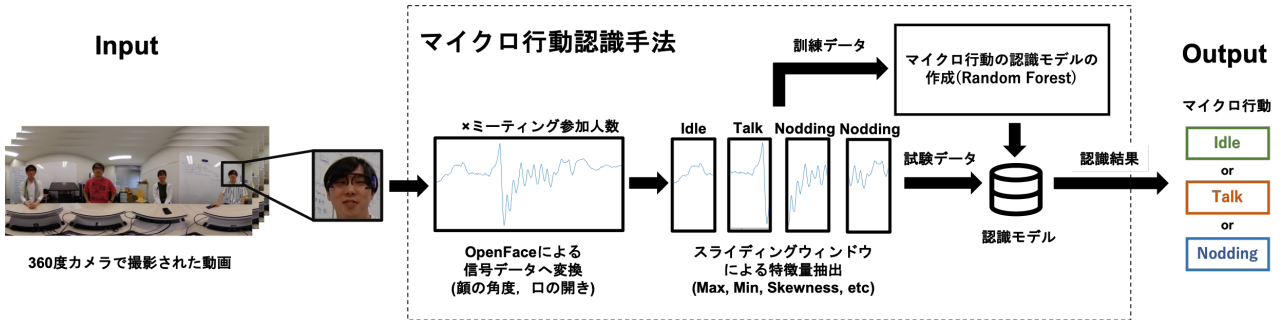

図 1: マイクロ行動認識の提案手法の全体図

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，使用したデバイス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について記述する．第 5 章では，実験から取得されたデータセットからマイクロ行動の認識した結果について記述する．第 6 章では，行動変化点を含むウィンドウを考慮した認識を行い，正解と認識結果を時系列上で比較する．第 7 章では，認識結果に関する考察について記述する．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

大西らは [8] 頭頂部に慣性計を装着し，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 [9] 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の認識は行われていない．また，発言内容を解析する必要があるため，個人情報を扱うリスクが発生する．

Yu らは [10] 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，ミーティング中に発生するマイクロ行動のセンシングに関する先行研究は多く存在する．特に，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といった課題を抱える研究も確認された．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の向上を試みられているが，実際のミーティングについて解析を行うためにはより汎用的な状況で導入できるシステムが求められる．本研究の提案手法では，このような背景から実際のミーティングを想定し，比較的安価な価格で手に入る 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う．

3. ミーティング中に発生するマイクロ行動の認識手法

本研究では，ミーティング中に発生するいくつかの仕草や行動の中で，参加者に共通して観測されるマイクロ行動である「Talk（発言）」「Smile（笑い）」「Nodding（頷き）」を認識対象とする．本手法は，対象とするマイクロ行動が，ミーティング参加者の頭部の動きに違いがあることに着目し，ミーティング映像から頭部の動きに関する特徴量を抽出し，機械学習によって認識するアプローチを採用する．本章では，提案手法で使用するデバイスとマイクロ行動の認識方法について記述する．

提案手法の全体の流れを図 1 に示す．入力は 360 度カメラによって撮影された動画，出力は予測されたマイクロ行動である．本研究では，動画から対象者の頭部角度や口の開き具合を表す時系列データを生成し、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後，機械学習ベースのアプローチによってマイクロ行動の認識を行う．機械学習アルゴリズムには Random Forest を採用する．

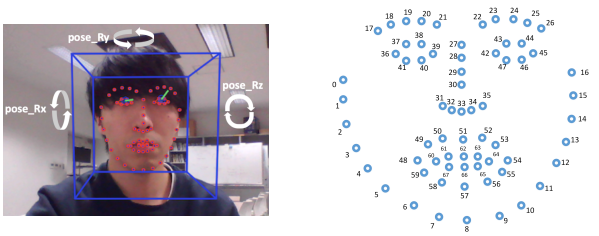

図 2: OpenFace の出力

3.1 入力デバイス

提案手法では，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の入力デバイスとして，RICOH THETA V [11] を採用する．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．THETA V を机の中央に設置し，ミーティングの参加者を撮影する．

3.2 OpenFace を用いた信号データへの変換

そして，撮影された動画をオープンソースのソフトウェアである OpenFace [12] を使用し，顔の特徴点の座標や角度を求める．OpenFace を顔画像に適用させた時の画像とその特徴点について図 2 に示す．OpenFace は顔画像を図 2 に示す様に 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．

3.3 スライディングウィンドウによる特徴量抽出

これらのデータの一部（顔の垂直の角度，口の開き）をセンサデータとして時間幅を持った部分系列に分割する．その部分系列から，表 1 に示した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）を抽出したものを説明変数，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．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 [13][14][15]．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，マイクロ行動の継続時間が短いことが原因となり，ウィンドウ内に行動変化点が含まれる場合が多く発生する．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．5 章では動画から定義 1 の手法によって抽出されたウィンドウデータからマイクロ行動の認識を行う．

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．6 章では実際の予測に使われるデータを想定し，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．

3.3.1 Nodding の認識に使用する特徴量抽出

Nodding は顔の垂直角度が変化するため，図 2 に示す pose_Rx 成分を Nodding の認識するためのセンサデータとする．本研究で取得したデータセットから，特に継続時間の短い Nodding の平均継続時間が 1.44 秒であること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．このときウィンドウ幅が大きすぎると，特に Nodding のように他のマイクロ行動よりも継続時間が短い行動について，行動変化点を含むウィンドウが増加し抽出されるサンプル数が減少してしまう．したがって，ウィンドウ幅を上記のように設定した．

3.3.2 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 2 に示す特徴点の中で 62 番と 66 番の特徴点の距離を Talk を認識するためのセンサデータとする．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%とする．

3.3.3 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は不適切であるが，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 [16]．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．

3.4 マイクロ行動認識モデルの作成

機械学習アルゴリズムには Random Forest を採用した．学習モデルの構築には，Python の機械学習ライブラリである scikit-learn [17] を用いており，パラメータは scikit-learn のデフォルト設定である．機械学習アルゴリズムに Random Forest を選択する理由として，行動認識に関する先行研

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
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
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.

究 [13][14][15] において有効性が示されているためである．

4. 実験内容

4.1 実験手順

4 人の被験者は指定された内容について 5 分間ディスカッションを行う．このときのディスカッションの内容は，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？」など）．ディスカッションを 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．

各被験者は，ELAN [18] 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に対して 3 つのマイクロ行動「Talk」「Smile」「Nodding」についてラベリングを行う．

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 回のディスカッション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．

4.3 認識結果の評価方法

機械学習のモデルの汎化性能を評価するために交差検証法を用いる．第 5 章では，各被験者の全てのデータを 1 つのデータセットと捉え，K-分割交差検証法（K-Fold Cross Validation）による評価を行う．K-分割交差検証法とは，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 個のサブセットを用いて認識を行い評価する方法である．本研究では，K の値を 5 と設定した．

第 6 章では，各被験者の 1 回のミーティングに関するデータを 1 セッションと捉え，一個抜き交差検証法（Leave-One-Out Cross Validation）による評価と時系列上での予測結果の可視化を行う．一個抜き交差検証法とは，ある 1 セッションに対して残りの全てのセッションを用いて認識を行い評価する方法である．この方法では，時系列に続いているデータに対して予測を行うため，予測結果を時系列上で確認することが可能となる．

5. マイクロ行動の認識結果

5.1 Nodding の認識

図 2 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によって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（何もしていない状態），Smile，Talk，Nodding の 4 種類の動作の認識を行なった．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．不均衡なデータセットから機

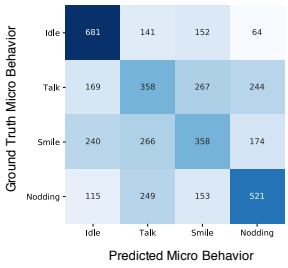

図 3: Nodding を対象に認識した結果の混同行列

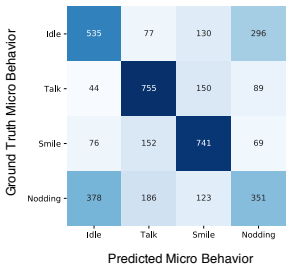

図 4: Talk を対象に認識した結果の混同行列

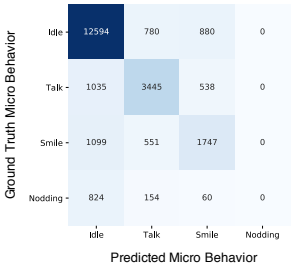

図 5: Talk を対象に認識した結果の混同行列（アンダーサンプリングなし）

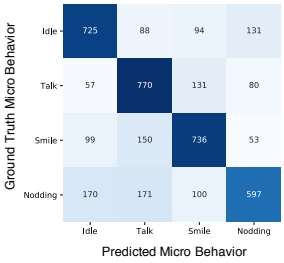

図 6: Talk と Nodding を対象に認識した結果の混同行列

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，Nodding のサンプル数にまで他クラスのサンプル数を減少させる．この手法はアンダーサンプリングと呼ばれる．認識結果の混同行列を図 3 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 2 に示す．

Nodding の F 値は 50.9%であった．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．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法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．

5.2 Talk の認識

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法によって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．このとき，Nodding の認識と同様に，Nodding のサンプル数にまで他のマイクロ行動についてアンダーサンプリングする．認識結果の混同行列を図 4 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 3 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 5 に示す．Talk の F 値は 69.4%である一方，Smile の F 値は 52.7%と低下していることから，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．

表 2: 図 3 から算出した Nodding の適合率・再現率・F 値

Behavior	Precision	Recall	F-measure
Nodding	0.504	0.515	0.509

表 3: 図 4 から算出した Talk の適合率・再現率・F 値

Behavior	Precision	Recall	F-measure
Talk	0.727	0.645	0.684

表 4: 図 6 から算出した Idle，Talk，Nodding の適合率・再現率・F 値

Behavior	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

5.3 Nodding と Talk の同時認識

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 6 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 4 に示す．図 3 と図 6 の混同行列を比較すると，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．

6. 行動変化点を考慮したマイクロ行動の認識

実際のミーティングに対するマイクロ行動の予測を想定し，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データを再度抽出しマイクロ行動の認識を行なった．

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 7 に予測の一例を示す．上のグラフが正解，下のグ

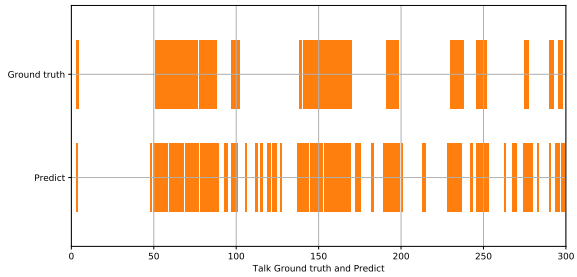

図 7: Talk の正解と認識結果（精度の高い例）

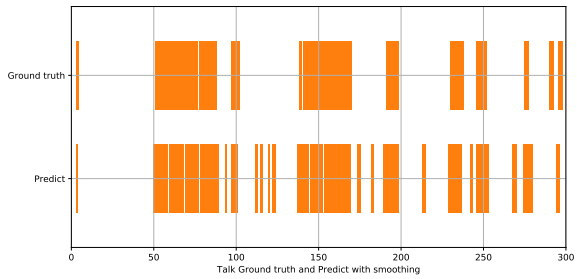

図 8: 図 7 の予測値を平滑化した場合

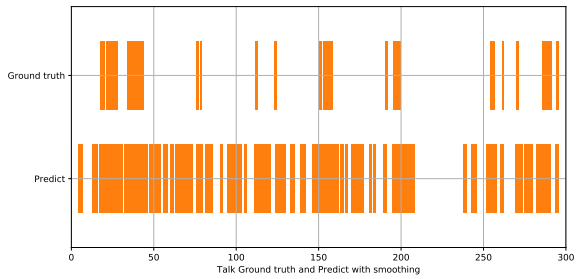

図 9: Talk の正解と認識結果（精度の低い例）

ラフが認識結果を表し，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．予測値を平滑化したグラフを図 8 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 9 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 9 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．

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．しかし，Nodding は他のマイクロ行動と比べてサンプル数が少ないため，各マイクロ行動のサンプル数とバランシングを行う必要がある．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

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 10 に示す．青色の帯は Nodding ラベル，空白箇所は Other ラベルを表している．正解と認識結果を比較すると，Other に対して Nodding と予測を多くされている．これは他のセッションについても同様に確認された．

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 [19] 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 11 に示す．図 10 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ついても同様に確認された．

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機械学習による認識精度を向上することが確認された．

6.3 Talk と Nodding の認識

次に，Talk と Nodding の認識で用いられた特徴量を組み合わせ，Talk，Nodding，Other（Idle，Smile）の 3 値分類を行う機械学習モデルの作成を行なった．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 12 と図 13 に示す．このときグラフは平滑化済みのものである．図 12 のグラフは図 10 と 11 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 11 で Nodding と誤認識を行なった箇所が図 12 では Talk と正しく認識されていることが確認できる．

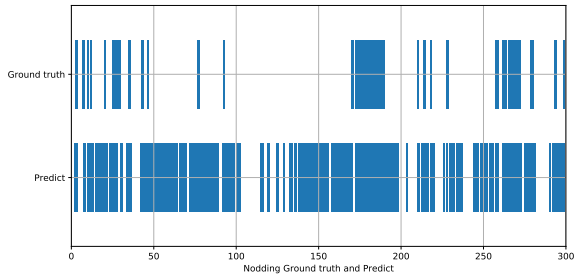

図 10: Nodding の正解と認識結果（アンダーサンプリング）

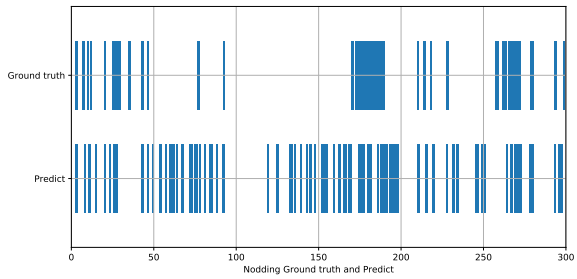

図 11: Nodding の正解と認識結果（オーバーサンプリング）

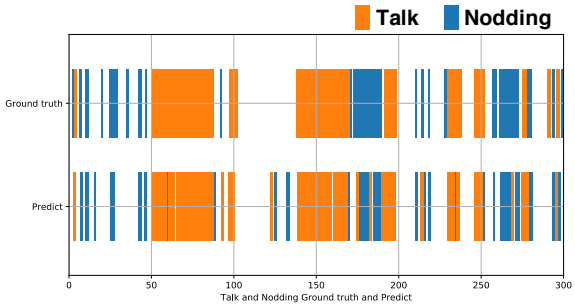

図 12: Talk と Nodding の認識結果（例 1）

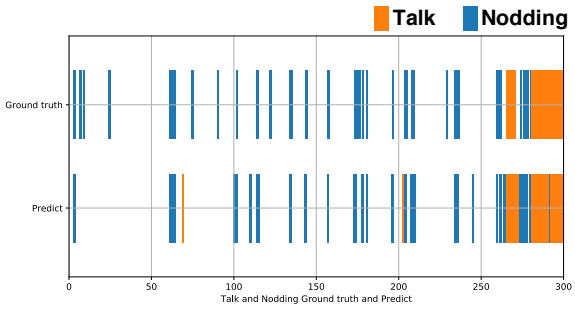

図 13: Talk と Nodding の認識結果（例 2）

7. 考察

7.1 Smile の認識について

第 5 章で，Smile の認識は口の開きからでは困難であると考察を行なった．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 14 に示す．緑色の帯が Smile のラベルを表す．この結果から，正解が 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が発生したと考えられる．

7.2 オーバーサンプリングによる Nodding の認識精度向上について

第 6 章では，Nodding の認識精度の向上のためにオーバーサンプリングを行なった．その結果，アンダーサンプリングによるバランシングに比べて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ことが挙げられる．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．本研究で使用した ADASYN 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. おわりに

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を動画から認識する手法を提案した．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．

次に，行動変化点がウィンドウ内に含まれる場合を考慮したマイクロ行動の認識を行い，正解と認識結果の比較を行なった．特に Nodding の認識精度が低い状態を，オーバーサンプリングの手法である ADASYN により，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．最後に，Talk と Nodding の認識を同時に行うことによりさらに誤認識が減少し，より正確な認識が可能となった．

今後の展望としては，現時点では Smile の認識を行なっていないが，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．本研究で提案した手法を用いて，実際のオフィスや教育現場のデータを収集し，満足度や理解度とマイクロ行動がどのように関係しているかについて解析し，ミーティングの定量的な評価を行えると期待される．

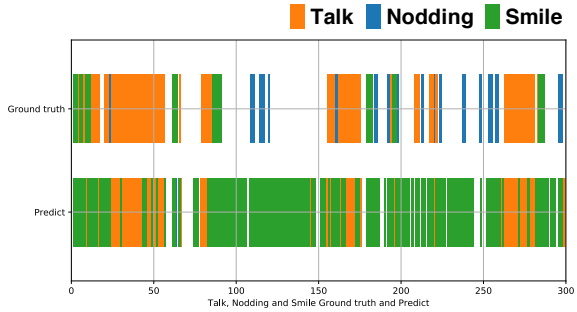

図 14: Smile の認識を追加した予測結果

データセットの公開

本研究では，THETA を用いた撮影以外に，慣性計と装着型視線計による計測が行われた．慣性計は各被験者の頭部の動きを計測，視線計は各被験者の注視点を計測している．これらのデータセットは，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 [20][21]．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．

謝辞 本研究は，JST さきがけ，および，Society 5.0 実現化研究拠点支援事業，の支援のもと実施されている．

参考文献

[1] Arakawa, Y.: Sensing and Changing Human Behavior for Workplace Wellness, *Journal of Information Processing*, Vol. 27, pp. 614–623 (online), DOI: 10.2197/ipsjjip.27.614 (2019).

[2] Nakamura, Y., Matsuda, Y., Arakawa, Y. and Yasumoto, K.: WaistonBelt X: A Belt-Type Wearable Device with Sensing and Intervention Toward Health Behavior Change, *Sensors*, Vol. 19, No. 20, p. 4600 (2019).

[3] Otoda, Y., Mizumoto, T., Arakawa, Y., Nakajima, C., Kohana, M., Uenishi, M. and Yasumoto, K.: Census: Continuous posture sensing chair for office workers, *2018 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–2 (online), DOI: 10.1109/ICCE.2018.8326275 (2018).

[4] 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).

[5] Riedhammer, K., Favre, B. and Hakkani-Tür, D.: Long story short – Global unsupervised models for keyphrase based meeting summarization, *Speech Communication*, Vol. 52, No. 10, pp. 801 – 815 (online), DOI: https://doi.org/10.1016/j.specom.2010.06.002 (2010).

[6] P.Ekman and Friesen, L. W.: The repertoire of nonverbal behavior, *Semiotica*, pp. 49–98 (online), DOI: 10.1515/semi.1969.1.1.49 (1969).

[7] Bull, P.: *Posture and Gesture*, Pergamon Press (1987).

[8] A. Onishi, K. M. and Terada, T.: A method for structuring meeting logs using wearable sensors, *Internet of Things*, pp. 140–152 (online), DOI: 10.1016/j.iot.2019.01.005 (2019).

[9] Morency, L.-P., Sidner, C., Lee, C. and Darrell, T.: Head gestures for perceptual interfaces: The role of context in improving recognition, *Artificial Intelligence*, Vol. 171, No. 8, pp. 568 – 585 (online), DOI: https://doi.org/10.1016/j.artint.2007.04.003 (2007).

[10] Yu, Z., Yu, Z., Aoyama, H., Ozeki, M. and Nakamura, Y.: Capture, recognition, and visualization of human semantic interactions in meetings, *2010 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, IEEE, pp. 107–115 (2010).

[11] RICOH: THETA V (2017).

[12] Tadas Baltrušaitis, P. R. and Morency, L.-P.: OpenFace: an open source facial behavior analysis toolkit, *IEEE Winter Conference on Applications of Computer Vision*, (online), DOI: 10.1109/WACV.2016.7477553 (2016).

[13] Takata, M., Fujimoto, M., Yasumoto, K., Nakamura, Y. and Arakawa, Y.: 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, Association for Computing Machinery, Inc, pp. 1404–1408 (online), DOI: 10.1145/3267305.3267504 (2018).

[14] Anguita, D., Ghio, A., Oneto, L., Parra, X. and Reyes-Ortiz, J. L.: A public domain dataset for human activity recognition using smartphones., *Esann* (2013).

[15] 中村優吾， 荒川豊，安本慶一ほか：ウェアラブルセンサ装着位置/向きの違いにロバストな行動認識システムの実現に向けたデータ変換手法の検討，マルチメディア, 分散協調とモバイルシンポジウム 2019 論文集， Vol. 2019, pp. 135–146 (2019).

[16] Ranjan, R., Sankaranarayanan, S., Castillo, C. D. and Chellappa, R.: An All-In-One Convolutional Neural Network for Face Analysis, *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 17–24 (online), DOI: 10.1109/FG.2017.137 (2017).

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).

[18] Archive, T. L.: ELAN (2002).

[19] He, H., Bai, Y., Garcia, E. and Li, S.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, pp. 1322 – 1328 (online), DOI: 10.1109/IJCNN.2008.4633969 (2008).

[20] Soneda, Y., Matsuda, Y., Arakawa, Y. and Yasumoto, K.: 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, Association for Computing Machinery, Inc, pp. 825–834 (online), DOI: 10.1145/3341162.3345588 (2019).

[21] Soneda, Y., Matsuda, Y., Arakawa, Y. and Yasumoto, K.: Multimodal Recording System for Collecting Facial and Postural Data in a Group Meeting, *The 27th International Conference on Computers in Education (ICCE 2019)*, No. 153 (2019).