Web 会議における予備動作を用いた発話欲求推定手法の提案

山田楓也 [1]　　白石陽 [2]　　石田繁巳 [2,3]

概要：近年，テレワークの普及に伴い，Web 会議サービスの利用が増加している．Web 会議では，画面構成や配置によって会議参加者の状況が把握しにくくなり，発話するタイミングが掴めないという問題がある．相手の状況把握が不完全なまま発話しようとすることで発話衝突が発生し，会話が中断される．そのため，円滑な話者交替を行いながら，会議進行することが困難である．しかし，会議参加者が発話する際に発話欲求を推定し提示できれば，会議参加者の円滑な話者交替が実現できると考えられる．そこで本研究では，会議参加者の発話欲求を推定する手法の実現を目指す．実現に向けた最初の段階として，会議参加者の数秒後に発話するか否かの予測（発話予測）を行う．発話予測を行う上で，発話前に行う予備動作に着目する．発話予測に有効な予備動作を明らかにすることで，その予備動作が現れた際に発話欲求の推定が可能になると考える．これらを踏まえ，本研究では Web 会議における予備動作を用いた発話予測に取り組む．発話予測の流れとして，まず Web 会議中の映像データから会議参加者の予備動作に関するデータの収集を行う．発話・非発話区間のアノテーションを行い，その中で予備動作が現れる区間のデータを切り出す．切り出したデータから特徴量を抽出し，抽出した特徴量を用いて機械学習モデルの構築を行う．構築したモデルの性能を検証するため，Web 会議における被験者 3 名の発話予測の精度評価を行った．10-分割交差検証を行い，F 値による評価を行った結果，7 割から 9 割の精度が得られた．また，精度評価と特徴量の重要度の結果から，発話予測には口の開きと視線移動が予備動作として関与していることが示唆された．

A Method for Estimating Utterance Desire Using Pre-Utterance Behavior in Web Meetings

FUYA YAMADA[1]　YOH SHIRAISHI[2]　SHIGEMI ISHIDA[2,3]

1. はじめに

近年，働き方改革による働き方の多様化や新しい生活様式の変化によりテレワークが普及している[1]．これに伴って遠隔でのコミュニケーションツールの一つとして，Web 会議サービスの利用が増加している．多種多様な Web 会議サービスが存在しており，同時に多数の参加者と接続できる機能やリアルタイムで同じ画面を見ながら議論し合える機能などが備わっている．しかし，画面構成や配置によって会議参加者の状況把握しにくくなり，発話するタイミングが掴めないという問題がある．相手の状況把握が不完全なまま発話しようとすることで発話衝突が発生し，会話が中断される．そのため，円滑な話者交替を行いながら，会議進行することが困難である．

これらの問題に対して，発話欲求に関する研究が行われている[2], [3], [4], [5], [6]．文献[2]では，Web 会議を対象として，発話前に行う意識的な予備動作を用いて発話欲求伝達システムの構築を行っている．意識的な予備動作として，「頷き」，「挙手」，「手を顔周辺へ動かす動作」の 3 つを定義している．しかし，この研究で定義している予備動作は，すべての会議参加者に対して同様に現れる動作とは限らない．そのため，発話欲求が低い場合でも，高いと判断してしまう問題がある．また，文献[3]では，対面会議を対象として，韻律・身体動作・言語情報を用いて発話欲求が高いと判断される区間を推定している．この研究では，発話中の発話欲求を対象にしており，数秒後の発話予測に有効な発話前の予備動作を抽出できない．

一方で，対面会議における予備動作を用いた発話予測の研究[7], [8], [9], [10], [11]がある．顔特徴を用いた研究[7], [8], [9]では，対面会議における頭部運動，視線移動，口の開き具合に着目している．また，韻律情報を用いた研究[10], [11]では，基本周波数（F0），発話のピッチ，音量レベルに着目している．しかし，これらの研究では，対面会議で現れる動作を対象にしており，Web 会議を想定した場合，発話前の予備動作が対面会議とは異なる可能性がある．そのため，Web 会議における発話予測に有効な予備動作を明らかにする必要がある．

そこで本研究では，Web 会議における予備動作を用いた会議参加者の発話欲求を推定する手法を提案する．具体的には，発話欲求に合わせて，Web 会議ツール中の会議参加者の映像画面の色を変化させるシステムを想定する．発話欲求の度合いを画面上に可視化して提示することで，会議

1 公立はこだて未来大学大学院　システム情報科学研究科
Graduate School of Systems Information Science, Future University Hakodate
2 公立はこだて未来大学　システム情報科学部
School of Systems Information Science, Future University Hakodate
3 九州大学　システム LSI 研究センター
SLRC, Kyushu University

参加者は相手の状況を把握することが可能になる．他の参加者の発話欲求を確認することで発話タイミングを掴むことが可能になり，発話衝突を抑制できると考える．そのため，本研究では Web 会議における発話欲求推定システムの実現を目指す．

本システムの実現に向けた最初の段階として，発話予測を行う．発話予測において，発話前に行う予備動作に着目する．マジョリーは，会議参加者が発話する際に非言語情報を活用すると述べている[12]．そのため，予備動作の一つに非言語情報が関連していると考える．発話予測に有効な予備動作を明確にすることで，その予備動作が発生した際に，発話欲求の推定が可能になると考える．以上を踏まえて，本研究では，Web 会議における予備動作を用いた発話予測に取り組む．

本稿では，Web 会議時のデータ収集を行い，収集データに基づいて発話予測モデルの構築を行い，各被験者の予備動作を明らかにするための分析を行った．まず，Web 会議中の映像と音声のデータを収集し，発話・非発話区間のアノテーションを行う．非発話区間のデータから，発話予測に用いるデータを切り出し，特徴量を抽出する．抽出した特徴量を用いて被験者ごとに学習モデルを構築する．構築した学習モデルを用いて，数秒後に発話するか否かの予測を行い，予測精度の評価を行う．精度評価と特徴量の重要度をもとに発話予測に有効な予備動作を明らかにする．

2. 関連研究

本章では，まず 2.1 節では発話欲求の研究について述べる．次に，2.2 節では予備動作を用いた発話予測の研究について述べる．

2.1 発話欲求の研究

Web 会議における意識的な予備動作を用いた発話欲求伝達手法の研究[2]がある．玉木らは，3 名による Web 会議を想定して，予備動作を用いた発話欲求伝達システムの構築を行っている．この研究では意識的な予備動作に着目し，「頷き」，「挙手」，「手を顔周辺へ動かす動作」の 3 つを定義している．これらの予備動作を検出し，発話欲求度合いを参加者に提示することで，発話衝突抑制の有効性を示している．発話欲求度合いは，Kinect センサを用いて予備動作を検出した後，検出するごとにスコアを加算し，スコアの総和としたものである．この研究で定義している予備動作は，すべての会議参加者に対して同様に現れる動作とは限らない．そのため，個人ごとに発話欲求の判定精度が異なり，正確に判定できない問題がある．また，Kinect センサでは，頭部，首，肩，関節，肘関節，手首などを 3 次元座標として取得している．しかし，顔の詳細な情報までは対象になっていないため，本研究の目的である発話欲求の推定に応用することは困難である．

また，1 対 1 の対面会議における韻律・身体動作・言語情報を用いた研究[3]がある．千葉らは，韻律・身体動作・言語情報を用いて発話欲求を推定している．発話欲求の指標は，話題に対する興味度合いに対して「発話したい」，「発話したくない」と定義している．しかし，「発話したい」の際に，実際に発話した否かの分析は行われていない．また，対面によるインタビューを想定しているが，3 人以上の会議の場合，頻繁に話者交替が行われることから環境が異なると考える．そのため，本研究の本研究で対象とする環境に応用することは困難である．

頷きや笑いなどの非言語コミュニケーションと会話との関係性の分析に向けて，360 度カメラで撮影した会議映像から非言語コミュニケーション・発話を自動的に検出する研究がある[4], [5]．しかし，データの分析を目的としており，会話中におけるリアルタイムの発話予測は行っていない．そのため，本研究の目的である発話欲求の推定に応用することは困難である．

2.2 予備動作を用いた発話予測の研究

顔特徴を用いた研究[7], [8], [9]，韻律情報を用いた研究[10], [11]がある．

まず，顔特徴を用いた研究[7], [8], [9]では，対面会議における頭部運動，視線移動，口の開き具合などを用いて発話予測を行っている．文献[7]では，話者継続時の非話者，話者交替時の非話者と次話者の 3 者間の頭部運動の特徴を分析している．分析結果として，話者継続と話者交替を比較した際，現話者の発話末の頭部位置・回転角の変化量と振幅のパラメータが異なると述べている．このパラメータを用いて，予測モデルを構築し，次に発話する会議参加者の予測を行っている．文献[8]では，現話者と非話者の視線交差が起きた時に，次話者が現話者である時，視線交差していない非話者である時，視線交差した非話者である時の 3 つの状況下で，視線交差の開始と終了時間，発話末と発話間隔時間の関係を分析している．分析結果として，話者継続時には非話者が先に現話者に視線を向ける確率が高いと述べ，この視線の特徴を用いて発話間隔の予測を行っている．文献[9]では，口の開き具合の違いを分析している．口の開き具合として，口を閉じている，狭く開いている，広く開いている，の 3 つを定義している．分析結果として，現話者は話者継続時に口を狭く開いたままにしていることが多い傾向があると述べている．この口の開き具合の特徴を用いて次に発話する会議参加者の予測，発話間隔の予測を行っている．これらの研究は，顔特徴を用いた発話予測の有効性を示している．しかし，これらの研究は全て対面会議を想定しているため，Web 会議に適用した場合に，同様の特徴が現れない可能性がある．そのため，Web 会議特有の予備動作を抽出する必要がある．

次に，韻律情報を用いた研究[10], [11]では，音響的特徴，相槌やフィラーを用いて発話予測を行っている．文献[10]では，話す速さ，基本周波数 (F0)，パワーなどの音響的特徴を用いて話者交替の予測を 2 段階で行っている．パワーとは，音量レベルの大きさである．1 段階目で，発話中か否かの判別を行い，2 段階目で，1 段階目の発話中から非発話に切り替わった際に話者交替及び継続の予測を行う．しかし，この研究は発話の状態を推定しており，数秒先の発話予測は行われていない．文献[11]では，相槌とフィラー予測を用いて発話予測を行っている．フィラーとは，「えー」や「あのー」といった言い淀み時に現れる場繋ぎに行う表現である．会話データは自律型ロボットと人間の会話データを用いている．しかし，ロボットとの会話の知見を人間同士の複数人会話に適用できるとは限らない．これらの発話予測の研究では，話者交替を対象にしており，数秒先の発話の予測をしていない．発話しない状況を考慮していないため，本研究の発話欲求の推定に応用することは困難である．

3. 発話欲求推定システム

本章では，3.1 節で本研究において実現したいシステムの概要について述べる．次に 3.2 節でシステムの実現に向けた研究課題について述べる．

3.1 システム概要

提案システムの全体像を図 1 に示す．

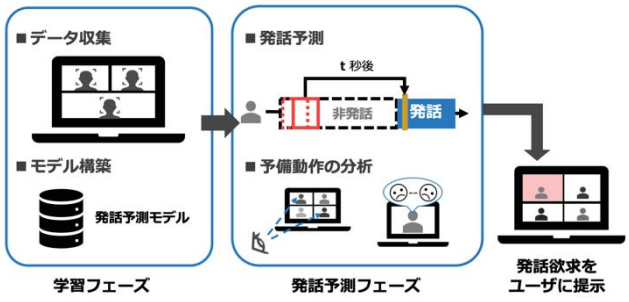


図 1　提案システムの全体像

提案システムは，Web 会議上でフィードバックを行うシステムを想定する．フィードバック方法として，会議参加者の発話欲求に合わせて，Web 会議ツール中の会議参加者の映像画面の色を変化させる提示方法を考えている．これにより，会議参加者が他の参加者の発話欲求を把握することが可能になることから，発話タイミングを掴むことができ，発話衝突を回避することができると考える．

本研究では，提案システムの実現に向けた最初の段階として，会議参加者の数秒先の発話を予測する手法の実現を目指す．発話予測を行う上で，発話前に行う予備動作に着目する．マジョリーは，対面会議において，会議参加者が発話前に非言語情報を活用すると述べている[12]．例えば，組んでいた腕や足を解いて身を前に乗り出す動作，体の向きを変えて相手を見る動作などを挙げている．このように予備動作は非言語情報の一つになると考える．そのため，予備動作を用いて発話予測を行う．発話予測では，まず Web 会議時のデータ収集を行い，収集データに基づいて発話予測モデルの構築を行う．発話予測モデルの精度結果に基づき，発話予測に有効な予備動作の分析を行う．予備動作を分析することで，予備動作が発生した際に発話欲求を推定できると考える．

3.2 研究課題

発話予測における研究課題を以下に示す．

課題 1．対象とする予備動作の検討
課題 2．発話予測に有効な特徴量の検討
課題 3．予備動作区間の検討

課題 1 について，個人ごとに予備動作が異なるため，会議参加者それぞれで調査する必要がある．また，Web 会議で予備動作を収集する際，自然な会話を阻害しない方法で収集する必要があると考える．そこで非接触型のデバイスで会議参加者の予備動作を収集することが望ましいと考える．そのため，非接触型デバイスによって検出可能な予備動作の検討を行う．

課題 2 について，発話予測に有効な特徴量を検討する必要がある．発話予測は会議参加者の情報から特徴量を抽出して機械学習を行う．発話予測に有効な特徴量を明らかにすることで，個人ごとの予備動作を検出することが可能であると考える．そのため，発話予測に有効な特徴量の検討を行う．

課題 3 について，予備動作を明らかにする上，発話予測に用いるデータの切り出し範囲（予備動作区間）を設定する必要がある．個人ごとにいつ予備動作が現れるか明らかになっていない．1 つの動作で現れる場合は，短い時間の範囲を設定する必要がある．また，複数の動作の組み合せで現れる場合は，長い時間の範囲の設定が必要である．そのため，予備動作区間としてデータを切り出す範囲の検討を行う．

4. 発話予測手法

本章では，まず 4.1 節では課題に対するアプローチを述べる．次に 4.2 節でシステム構成を示し，4.3 節で対象とする予備動作の検討について述べる．4.4 節以降では発話・非発話のアノテーション，特徴量抽出，モデル構築について述べる．

4.1 アプローチ

本研究では，発話予測の研究課題に対するアプローチとして次に述べる．

課題 1 に対するアプローチとして，会議参加者の顔特徴に着目する．文献[7], [8], [9]より，顔特徴が発話予測に有効であることを示している．さらに，Web 会議では PC 本体に内蔵している Web カメラを用いて自身の顔から上半身までを映すことが多いと考えられる．そのため，詳細な顔特徴を対象にすることで，予備動作を明らかにすることが可能になる．具体的な顔特徴として，頭部運動，視線移動，口の開きを用いる．会議参加者の顔映像から画像処理ツールを用いて顔特徴点を追跡する．顔特徴点の変化量を算出して，顔特徴における予備動作を抽出する．

課題 2 に対するアプローチとして，時系列データからスライディングウィンドウによって特徴量を算出する．まず，Web 会議中の会議参加者の顔映像を収録する．顔映像から顔特徴点を追跡した顔特徴データを収集する．収集したデータからスライディングウィンドウによりデータの切り出しを行う．スライディングウィンドウとは，時系列データに対してウィンドウサイズを設定し，前のウィンドウに対してオーバーラップさせながらウィンドウをスライドさせる手法である．ウィンドウごとに 7 種類 (平均, 標準偏差, 最小値，最大値，中央値，歪度，尖度) の基本統計量を算出する．基本統計量を用いることで，データの大まかな特徴をつかむことができると考える．

課題 3 に対するアプローチとして，予備動作区間を 0.5 秒，1.0 秒，2.0 秒，3.0 秒，5.0 秒で設定し，それぞれで予測精度を比較する．予備動作区間は課題 2 で述べた 1 つのウィンドウサイズのことを指す．予備動作区間の具体的な数値は，個人によって異なると考える．例えば，1 つの動作が単体で現れる場合，複数の動作の組み合わせで現れる場合があると考える．そのため，それぞれの被験者で検証し，個人に最適な予備動作区間を検討する．

4.2 システム構成図

発話予測のプロセスを図 2 に示す．

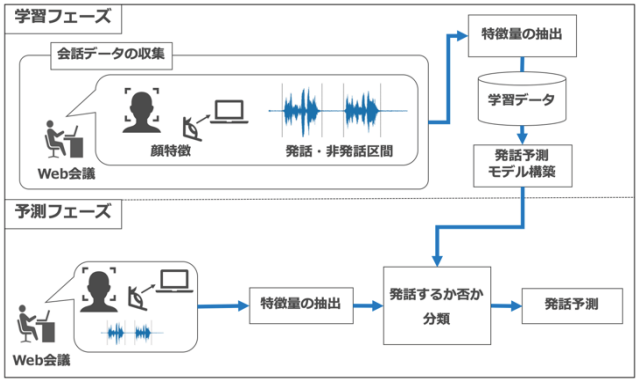


図 2　発話予測のプロセス

発話予測手法は，学習フェーズと発話予測フェーズから構成される．学習フェーズでは，Web 会議中の映像と音声のデータを収集する．収集したデータに対して，被験者ごとに発話・非発話区間のアノテーションを行う．非発話区間のデータからスライディングウィンドウによって特徴量を抽出する．抽出した特徴量を正解ラベルとともに学習データとする．この学習データを用いて個人ごとに発話予測モデルを構築する．

推定フェーズでは，学習フェーズと同様に映像と音声のデータを収集し，特徴量を抽出する．抽出した特徴量と発話予測モデルを用いて，会議参加者の発話予測を行う．発話欲求推定に向けて，予備動作として有効な特徴量を明らかにする．

4.3 対象とする予備動作の検討

発話予測に用いる予備動作として顔特徴データに着目する．Web 会議中の映像データから顔画像処理ツールである OpenFace[13]を用いて顔特徴データを収集する．OpenFace を用いることで，顔の特徴点を追跡し，その変化量を CSV データとして出力することが可能である．OpenFace を顔画像に適用させた例と顔の特徴点を図 3，本研究で対象にする顔特徴を表 1 に示す．

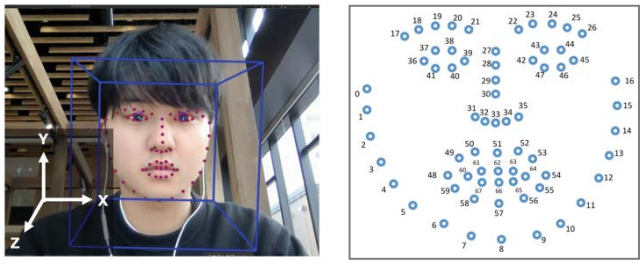


図 3　OpenFace の出力例

表 1　対象とする顔特徴

顔特徴	詳細
頭部運動	水平方向 (pose_Tx) 垂直方向 (pose_Ty) 手前から奥方向 (pose_Tz) 回転角 (pose_Rx，pose_Ry，pose_Rz)
視線移動	水平方向 (gaze_angle_x) 垂直方向 (gaze_angle_y)
口の開き	上唇と下唇に対する特徴点 (y_62，y_66) の差分 (mouth)

図 3 に示すように，画像の水平方向を x 軸，垂直方向を y 軸，手前から奥方向を z 軸とする．本研究で扱う顔特徴は，頭部運動，視線，口の開きである．頭部運動には，頭部の位置 (pose_Tx，pose_Ty，pose_Tz) の 3 成分，回転角

(pose_Rx，pose_Ry，pose_Rz) の 3 成分を使用する．頭部の位置の特徴点を用いることで，カメラに顔を近づける動作や頭を横に傾ける動作を把握できると考える.視線には，視線の水平方向 (gaze_angle_x) と垂直方向 (gaze_angle_y) の角度を使用する．視線の方向を用いることで，画面外か画面内かどうか，画面内でどの領域を注視しているかを把握できると考える．OpenFace で抽出した顔特徴点のうち，上唇と下唇に対する特徴点 (y_62，y_66) の y 座標の差分を算出した値 (mouth) とする．口の開きを用いることで，発話前に口が開く特徴を把握できると考える.

4.4 発話・非発話区間のアノテーション

発話・非発話区間のアノテーションを行う．まず会話注釈ツール ELAN[14]を用いる.ELAN は映像データや会話の音声波形を観察しながら，注釈を付与するツールである．区間ごとに注釈を付与することができ，例えば，発話の開始時間や終了時間，発話間隔時間を抽出できる．そこで無音区間自動認識の機能を用いて，無音区間と有音区間の自動判別を行った上でアノテーションを行う．自動判別の際に設定するパラメータを表 2 に示す.

表 2　自動アノテーションの設定値

パラメータ名	設定値
無音音量レベル (RMS)	-30dB
最短無音時間	200ms
最短有音時間	500ms

自動アノテーションを実行した後，発話区間と非発話区間であるか否かを確認し，自動判別できていない箇所は手動で修正を行う．今回は笑い声や咳，相槌などの言語内容として意味を持たない発話は発話区間の対象外としている.

4.5 特徴量抽出

4.4 節の手順により収集した非発話区間のデータから，スライディングウィンドウによって特徴量を抽出する
発話予測に用いる特徴量を表 3 に示す.

表 3　使用する特徴量

特徴量 (63 次元)	
基本統計量	平均 (mean) 標準偏差 (std) 最大値 (max) 最小値 (min) 中央値 (median) 尖度 (kurtosis) 歪度 (skewness)

4.4 節の手順により収集した非発話区間のデータから．スライディングウィンドウによって特徴量を抽出する．オーバーラップさせる割合を 50%とし，ウィンドウごとに基本統計量を算出する．基本統計量は 7 種類を用いる．頭部運動 6 種類，視線移動 2 種類，口の開き 1 種類であり，次元数は 63 次元とする.

4.6 モデル構築

4.2 節で述べた顔特徴データから 4.5 節で挙げた特徴量を抽出し，これらの特徴量を用いて数秒後に発話するか否かの予測する発話予測モデルを構築する．発話予測から会議参加者の予備動作を分析するために分類器はランダムフォレスト[15]を用いる．ランダムフォレストは弱学習器である決定木を組み合わせてアンサンブル学習を行う機械学習のアルゴリズムである.ランダムフォレストの特徴として，次元数が高い場合でも，過学習が発生しにくく，推定精度が低下しにくいことが挙げられる．予備動作を分析する際に,特徴量を増やすことで次元数が高くなると考えられる．また，変数重要度から発話予測に有効な特徴量が明らかになり，分岐の閾値から具体的な予備動作を特定できると考える．以上を踏まえて，本実験の発話予測モデルにはランダムフォレストを用いる.

5. 実験および考察

本稿では，発話欲求推定に向けて，有効な予備動作を把握するために発話予測に関する評価実験を行った．本章では，まず 5.1 節で実験環境について述べる．次に 5.2 節で実験において扱うデータの切り出しについて述べる．5.3 節以降で研究課題に関する発話予測モデルの評価実験の結果と考察について述べる.

5.1 実験環境

本実験における実験環境について表 4 に示す．また，Web 会議ツールとして活用した Zoom の画面レイアウトの構成を図 4 に示す.

表 4　実験環境

項目	詳細
Web 会議ツール	Zoom 5.4.6
被験者	3 名 (A, B, C)
会議シナリオ	10 分間のアイディア出し (ファシリテーションなし)
アノテーションツール	ELAN 6.0
動画収集ツール	QuickTime Player 10.5
フレームレート	30fps
PC の画面サイズ	2560×1600

図 4　Zoom による画面レイアウト

被験者は 20 代の男子大学生 3 名 (A, B, C) であり，Web 会議サービスの Zoom を使用した．事前知識を必要としない簡単なテーマを設定して，10 分間の会議を行った．データ収集の際，Zoom の記録機能を用いて，全体の画面の映像，全体の合成音声，各参加者の音声のデータを対象とした．Zoom の画面はフルスクリーンに設定し，被験者の画面を①，②，③の位置に配置し，チャット機能は④の位置に配置した．また同時に，参加者のみの映像データを記録するために，QuickTime Player を用いて 1280×720 のサイズ，30fps で被験者全員の顔映像データを収集した．

5.2 予備動作区間の顔特徴データ切り出し

会議映像データから発話予測に用いる顔特徴データを収集する．予備動作区間における顔特徴データの切り出しの一例として発話前を図 5，非発話時を図 6 に示す．縦軸は OpenFace による出力値，横軸は時間軸上のサンプル番号を表す．図 5 と図 6 を比較すると，発話前と非発話時のデータを比較すると，顔特徴データの変化が異なることがわかる．

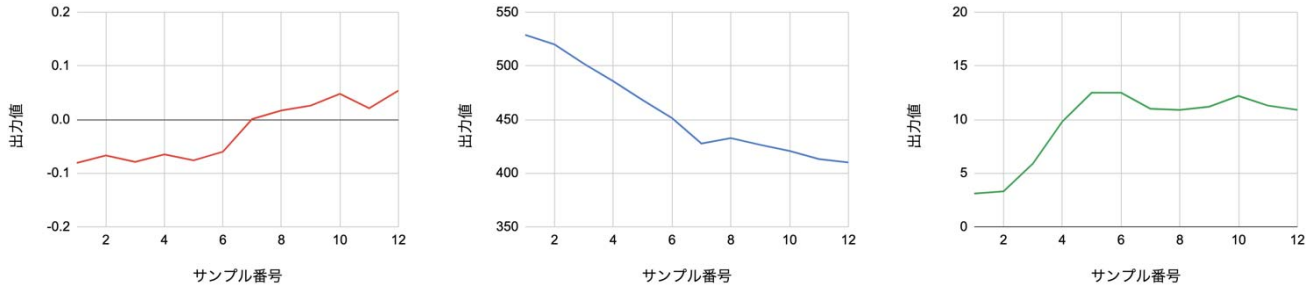
5.3 発話予測モデルの実験結果

5.3.1 モデルの精度評価

発話予測モデルを用いて発話するか否かの 2 値分類を行った．10-分割交差検証を行い，F-measure による精度評価を行った．また，3.2 節の研究課題を検討するため，パラメータを変えて実験を行った．各パラメータは，ウィンドウサイズと予測時間である．被験者 A，B，C の結果をそれぞれ図 7，図 8，図 9 に示す．縦軸は発話を予測するまでの時間（予測時間），横軸はウィンドウサイズを表す．

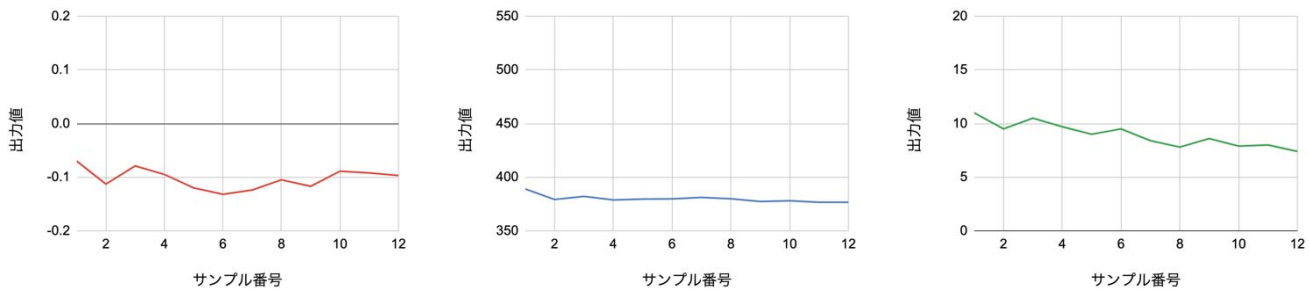
図 7 の被験者 A の結果から，ウィンドウサイズが大きくなるにつれ予測精度が向上していることがわかる．そのため，被験者 A は予備動作区間を長く設定することで予備動作を抽出できている可能性が高いと考える．また，予測時間が長くなるにつれ予測精度が低下している．そのため，発話直前に予備動作が現れると考える．予測時間が 0.5，1.0 秒間で，ウィンドウサイズが 1 秒以上の場合，7 割から 8 割の精度であることから，発話予測に有効な予備動作が抽出できている可能性がある．

図 8 の被験者 B の結果から，全体的に 8 割から 9 割の精度となった．特にウィンドウサイズを大きくすることで，精度が向上していることがわかる．他の被験者と比べ，ウィンドウサイズと予測時間のパラメータを変更した際の精



(a) 視線移動の水平方向　(b) 頭部運動の前後方向　(c) 口の開き

図 5　予備動作区間における発話前の顔特徴データ



(a) 視線移動の水平方向　(b) 頭部運動の前後方向　(c) 口の開き

図 6　予備動作区間における非発話時の顔特徴データ

度の変動が少ないことがわかる．被験者 B では，発話時と非発話時の動作間で大きく違いが出ている可能性がある．

図 9 の被験者 C の結果から，ウィンドウサイズが大きくなるにつれ予測精度が向上しており，予測時間が長くなるにつれ予測精度が低下している．予測時間が 0.5 秒の時の精度が 9 割を超えていることがわかる．そのため，被験者 A と同様に，発話直前に予備動作が現れると考える．

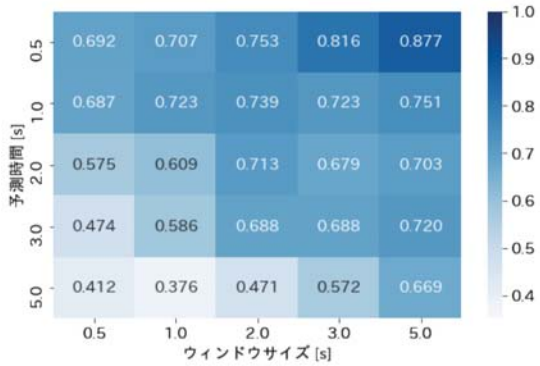


図 7　被験者 A の発話予測モデルの精度変化

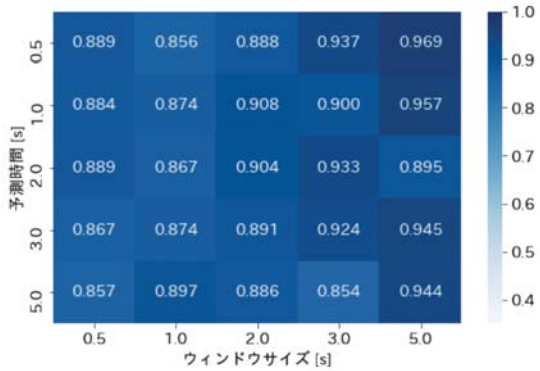


図 8　被験者 B の発話予測モデルの精度変化

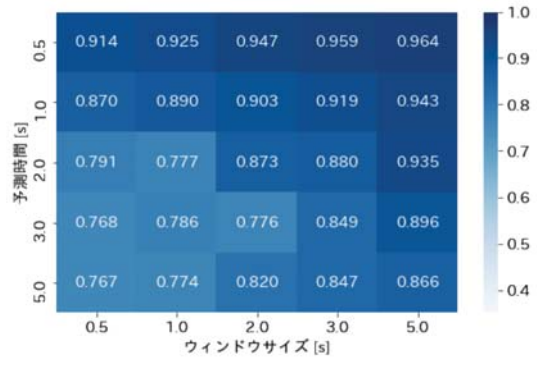


図 9　被験者 C の発話予測モデルの精度変化

5.3.2 特徴量評価

各被験者の予備動作の分析を行うため，特徴量の重要度を算出した．特徴量の重要度はランダムフォレストによる変数重要度を用いた．被験者 A，B，C の特徴量の重要度の上位 10 個をそれぞれ図 10，図 11，図 12 に示す．このときの，パラメータは，ウィンドウサイズ 1 秒間，予測時間 1 秒間である．

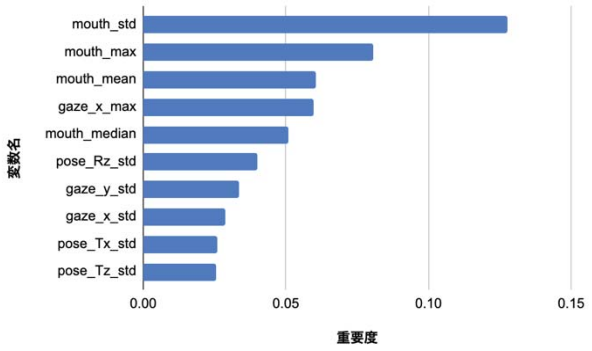


図 10　被験者 A の特徴量の重要度 (上位 10 個)

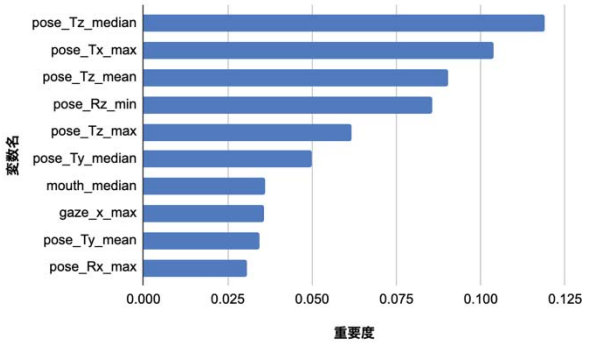


図 11　被験者 B の特徴量の重要度 (上位 10 個)

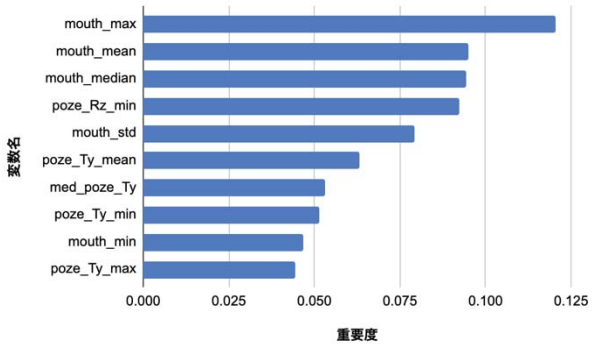


図 12　被験者 C の特徴量の重要度 (上位 10 個)

図 10 から被験者 A は，特に口の開きと視線が有効であることがわかる．口の開きは標準偏差，最大値，平均，中央値の順で高いことがわかる．特に口の開きの重要度が高いことから，被験者 A は口の開きの予備動作が現れる可能性がある．また，視線の水平方向の最大値が口の開きの一部の重要度と同等の数値であるため，視線移動も関与していると考える．

図 11 から被験者 B は，頭部運動が有効であることがわかる．特に，頭部の x 軸，z 軸，回転角が高いことがわか

る．また，1 つの動作単体で現れるのではなく，それぞれの動作の組み合わせることで現れる可能性もある．

図 12 から被験者 C は，特に口の開きと視線が有効であることがわかる．口の開きは最大値，平均，中央値，標準偏差の順で重要度が高いことがわかる．被験者 C は A と同様に口の開きの特徴が予備動作になる可能性が高い．また，口の開きの特徴量が A と同様であるため，口の開きに有効な特徴量であると考える．

5.4 考察

各被験者の予備動作を分析するために 5.3 節の結果をもとに 2 つの観点で考察を行う．5.4.1 項では，モデルの精度評価の結果から予備動作を現れるかの考察を行う．5.4.2 項では，特徴量の重要度から予備動作になる可能性のある動作について考察する．

5.4.1 精度評価の結果による考察

数秒先に発話するか否かの各クラスのどちらかに特徴的な動作が現れると，分類精度が高くなると考える．発話前に特徴的な動作が現れていた場合，その動作が予備動作になると考えられる．個人を対象にしても会議シナリオの違いにより現れる動作に差があると考えられるため，7 割以上の精度が妥当であると考える．以上を踏まえると，被験者 B，C の精度が 7 割を超えていることから，予備動作が現れると考える．被験者 A は予測時間が 0.5，1.0 秒の時に限り予備動作が現れると考える．

ウィンドウサイズ 1 秒間，予測時間 1.0 秒の被験者 A の混同行列を表 5 に示す．

表 5　被験者 A の混同行列

		予測結果		再現率
		発話	非発話	
正解	発話	969	231	.808
	非発話	428	772	.643
適合率		.694	.770	F 値/ .746

表 6 より 1.0 秒後に発話したクラスの再現率が 8 割を超えていることがわかる．将来的なシステムを想定した際に，数秒後に発話したクラスの再現率が高い方が，発話衝突の現象を抑制しやすいと考える．そのため，発話予測において数秒後に発話したクラスの再現率も評価指標の 1 つとして採用することができると考える．今回の被験者はモデルの精度評価の結果から，予備動作を明らかにすることが可能であると考える．

5.4.2 特徴量の重要度による考察

各被験者の予備動作を明らかにするために，特徴量の重要度を元にデータ収集で収録した各被験者の動画の観察を行った．

被験者 A では，発話前にフィラーを行う傾向があることがわかった．一方で，非発話時は頷きや口を小さく開けることがわかった．発話と非発話で口の開きに違いが出ているため，口の開きの特徴量の重要度が一番高い要因だと考える．また，発話前に視線移動の水平方向の動きを行うことがわかった．これは，画面内に映っている会議参加者を注視する動きであると考える．これを踏まえ，被験者 A の予備動作はフィラー，視線移動の水平方向であることが示唆された．

被験者 B では，非発話時に頭部運動の変化が少ないことがわかった．発話前の予備動作ではなく，非発話時の特徴が重要度の高い要因となると考えられる．被験者 B では，発話前の予備動作ではなく，非発話時の特徴が重要であることが示唆された．

被験者 C では，発話前には口を開けていることが多く，非発話時には口を閉じている傾向であることがわかった．口を開けている動作は，フィラーではないため，被験者 C 特有の癖である可能性も考えられる．非発話時には口を閉じているため，発話予測の際に明確に差が出ていたと考えられる．そのため，被験者 C の予備動作は口の開きであることが示唆された．

以上を踏まえ，被験者 A と C は予備動作が現れることが明らかになった．特に，予備動作として口の開き，視線移動が関与していることが示唆された．また，被験者ごとに有効な特徴量が異なることも明らかになった．一方で，被験者 B は非発話時の特徴が特徴量の重要度に関連していたため，予備動作かどうかの判断は困難であった．しかし，非発話時の特徴が明確になることで，数秒後に発話しない動作が明らかになると考えられる．

本実験において予備動作区間の設定方法に改善の余地がある．例えば，ウィンドウサイズが 5 秒の時に発話と発話の間隔が短い場合，予測対象の発話だけでなく前の発話の予備動作を抽出することになる．短い間隔で発話を繰り返す被験者が存在する場合，精度に影響が出てしまう可能性がある．しかし，連続した発話の特徴も発話予測に重要であると考える．そのため，今後は発話時間や回数などの別の特徴量を追加する必要があると考えられる．

また，被験者数を増やし，今回現れた予備動作の妥当性を確認する必要がある．会議シナリオを変えることで，今回明らかになっていない被験者 B の予備動作が現れる可能性がある．そのため，会議内容や会議参加者，会議回数などを変えて，精度評価や予備動作の分析を行う必要があると考えられる．

6. おわりに

本研究では，Web 会議における円滑な話者交替のために発話欲求推定システムの提案を行った．本研究では，発話欲求推定システムの実現に向けて，数秒先の発話を予測する機械学習モデルの構築を行った．発話前に行う予備動作に着目し，この予備動作を明らかにすることで発話欲求の推定に用いることができると考える．

本稿では，Web 会議中の映像データから会議参加者の顔特徴の収集を行った．さらに，映像データから現在のフレームに合わせて発話・非発話区間のアノテーションを行った．非発話区間のデータからスライディングウィンドウを行い，ウィンドウごとに特徴量を抽出した．抽出した特徴量を用いてランダムフォレストによる学習モデルの構築を行った．本実験では，発話予測の有効性の検証と予備動作を明らかにすることを目的に，被験者 3 名による発話予測モデルの精度評価を行った．10-分割交差検証を行い，F 値による評価の結果，7 割から 9 割の精度で発話を予測することが確認できた．予備動作の分析では，今回の実験で口の開きや視線移動が予備動作になることが示唆された．

今後の課題として，会議シナリオが予測精度に与える影響を調査する必要がある．本稿では，1 つの会議シナリオのみで実験を行っている．しかし，会議シナリオが異なることで，現れる予備動作にも影響が出ると考えられる．具体的な会議パラメータとして,マイク出力切り替えの有無,画面共有の有無，会議の目的，会議に参加する被験者の特性などが挙げられる.これらを変更して,データを収集し，精度比較を行う必要がある．また，予備動作を分析する被験者数を増やす予定である．今回は 3 名のみで分析を行ったが，顔特徴に現れない被験者も存在する可能性がある．そのため，他の予備動作の分析も行い，Web 会議に有効な予備動作を検討していく．

参考文献

[1] 東京都公式ホームページ，テレワーク導入実態調査結果，https://www.metro.tokyo.lg.jp/ (最終アクセス日：2021/4/20).

[2] 玉木 秀和，東野 豪，小林 稔，井原 雅行，“発話がぶつからない Web 会議を実現するための発話欲求伝達手法”，情報処理学会論文誌，Vol.54，No.1，pp.275-283 (2013).

[3] 千葉祐弥，伊藤彰則，“ユーザの対話意欲推定のための人対人対話データの分析と WOZ システムの検討”，情報処理学会研究報告 音声言語情報処理 (SLP)，Vol.2015-SLP-109，No.22，pp.1-6 (2015).

[4] 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”，情報処理学会研究報告 ユビキタスコンピューティングシステム (UBI), Vol. 2020-UBI-65, No.40, pp.1-8 (2020).

[5] 徳原 耕亮，ビリー ドートン，石田 繁巳，荒川 豊，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 回全国大会講演論文集，Vol.2020，No.1，pp.311-312 (2020).

[6] 二瓶 芙巳雄，田口 和佳奈，中野 有紀子，深澤 伸一，赤津 裕子，“多人数遠隔コミュニケーションにおける 肯定的感情表出支援の効果と支援適用タイミングの決定”，情報処理学会論文誌，Vol.62，No.2，pp.761-771 (2021).

[7] 石井 亮，大塚 和弘，熊野 史朗，大和 淳司，“複数人対話における頭部運動に基づく次話者の予測”，情報処理学会論文誌，Vol.57，pp. 1116-1127 (2016).

[8] 石井 亮，大塚 和弘，熊野 史朗，大和 淳司，“複数人対話における視線交差のタイミング構造に基づく次話者と発話開始タイミングの予測”，人工知能学会 全国大会論文集，Vol.29，pp.1-4 (2015).

[9] R.Ishii，K.Otsuka，S.Kumano，R.Higashinaka，J.Tomita，“Prediction of Who Will Be Next Speaker and When Using Mouth-Opening Pattern in Multi-Party Conversation”，Multimodal Technologies and Interaction. Vol.3，No.4 (2019)

[10] 小川 翼，伊藤 敏彦，“リアルタイム発話継続/交替予測システムの構築”, HAI シンポジウム 2014, Vol.43, pp.192-198 (2014).

[11] 原 康平，井上 昂治，高梨 克也，河原 達也，“相槌・フィラー予測とのマルチタスク学習によるターンテイキング予測”，情報処理学会 第 80 回全国大会講演論文集，Vol.2018，No.1，pp.409-410 (2018).

[12] マジョリー・F・ウォーガズ:非言語コミュニケーション，新潮社 (1987).

[13] T.Baltrusaitis，P.Robinson，and L-P.Morency，“OpenFace: An Open Source Facial Behavior Analysis toolkit”，Proc. of the 2016 IEEE Winter C onference on Applications of Computer Vision，pp.1-10 (2016).

[14] ELAN，https://archive.mpi.nl/tla/elan (最終アクセス日：2021/4/20).

[15] L. Breiman，“RandomForest”，Machine Learning，Vol.45，pp.5-32 (2001).