Vol.2022-GN-116 No.19
2022/3/15

Web会議におけるミクロな顔特徴に着目した発話予測手法の検討

山田 楓也[1,a)]　石田 繁巳[2,b)]　白石 陽[2,c)]

概要：近年，遠隔で会議を行う Web 会議の利用が広がっている．Web 会議では画面構成やサイズの制約により他の参加者の様子を読み取りにくく，誰がいつ発話し始めるのかを予測しづらいことから発話衝突が発生し，会議の進行が妨げられるという問題がある．これに対して，著者らは発話前の予備動作を用いて数秒後の発話状態を予測する手法を提案している．先行研究では，頭部運動，視線移動，口の開きに関する特徴量を Web 会議映像から抽出して被験者ごとに予備動作を認識することで，会議参加者の数秒後の発話予測を行う．初期的評価として被験者 3 名の発話予測を行った結果，F-maesure で 7 割から 9 割の精度で予測が可能であることを確認した．しかし，被験者を増やして先行研究の手法を適用した結果，被験者ごとの精度のばらつきが発生した．より多くの人に対応するため，本稿では先行研究を拡張し，ミクロな顔特徴及び他者の発話状態を考慮することで予測精度の向上を図る．被験者 9 名について Web 会議の映像を記録し，記録した映像から個人ごとに有効な特徴量を抽出して発話予測モデルを構築した．構築した発話予測モデルの精度を評価した結果，予測精度は平均 F-measure 0.694 となり，ミクロな顔特徴を用いることで予測精度が向上することを確認した．特徴量重要度の分析結果から，発話予測には口の開きという共通の特徴量と個人ごとに異なる特徴量を用いる必要があることを示した．

キーワード：Web 会議支援，会議センシング，発話予測，予備動作，機械学習

1. はじめに

IT 技術の発展により，遠隔で会議を行う Web 会議の利用が広がっている．テレワークの普及による後押しを踏まえると，参加する場所の制約の少ない Web 会議は今後ますます需要が高まると予想される．

しかしながら，Web 会議では誰がいつ発話し始めるのかを予測しづらいことから発話衝突が発生し，発話の繰り返しによって会議の進行が妨げられるという問題がある．対面会議とは異なり，Web 会議では他の参加者の様子を読み取りにくいために発話するタイミングの把握が難しい．そのため，円滑な Web 会議の進行に向けては他の参加者の様子を共有し，会議参加者が他者の発話を予測できるようにすることが重要である．

円滑な会議進行の実現に向けて，Web 会議，対面会議の発話交替を補助する手法がこれまでにも提案されている [1–4]．文献 [1] では，発話前に行われる予備動作を用いて Web 会議の参加者に発話欲求を伝達する手法を示している．この手法では，ヴァーガスら [5] の知見に基づいて頷き，挙手，手を顔周辺に近づけるといった予備動作を定義して発話欲求を推定している．文献 [2–4] では，頭部運動，視線移動，口の開きを用いて，発話交替及び次の発話者の予測を行う手法を示している．これらの手法は対面会議を前提とした予備動作に基づいて推定や予測を行っており，Web 会議に応用可能であるかは検証されていない．

著者らは，発話前に行われる予備動作を前提とした Web 会議向けの発話予測手法を先行研究 [6] において提案した．先行研究では，頭部運動，視線移動，口の開き（以降，マクロな顔特徴）の変化に基づき予備動作を認識することで，会議参加者の数秒後の発話予測を行う．初期的評価として，被験者 3 名の発話予測を行った結果，F-measure で 7 割から 9 割の精度で予測が可能であることを確認した．

本稿では先行研究を拡張し，より多くの人に対応可能な発話予測手法を提案する．被験者数を増やして先行研究の手法を適用した結果，被験者ごとに精度に大きくばらつくことが分かった．精度のばらつきは特徴量の不足が原因で

[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, Japan
[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, Japan
[a)] g2121057@fun.ac.jp
[b)] ish@fun.ac.jp
[c)] siraisi@fun.ac.jp

あると考え，本稿ではミクロな顔特徴及び他者の発話状態を考慮することで予測精度の向上を図る．例えば，顔を上げる，眉を下げるなどのミクロな顔特徴は，直後の発話に関連すると考えられる．対面会話において発話する際には他者の発話状態を考慮することから，ミクロな顔特徴の変化に加えて他者の発話状態を用いて教師あり学習により発話予測を行う．ミクロな顔特徴の変化は，OpenFace [7] の Action Unit から抽出する．対面会議における発話交替予測に関する研究 [4,8] では 7 割から 8 割以上の予測精度が必要であると報告されていることから，本研究では，予測精度が F-measure で 7 割以上，Recall で 8 割以上を目指す．

提案手法の有効性を検証するため，実際の Web 会議の映像データを用いて発話予測精度を評価した．相異なる 3 名が参加する Web 会議を 3 回行い，その映像を記録した．その映像データに提案手法を適用し，被験者ごとに発話予測モデルを構築して予測精度を評価した．その結果，平均 F-measure 0.694 という結果が得られた．また，予測モデルにおける各特徴量の重要度を分析した結果，口の開きに関する特徴量は多くの被験者で共通して高い重要度を持つものの，予測に大きく寄与する特徴量の多くは被験者ごとに異なることが分かった．

本稿の構成は以下の通りである．2 章では，Web 会議と対面会議における関連研究について示す．3 章では，ミクロな顔特徴と他者の発話状態を追加した発話予測手法を示す．4 章では，拡張した発話予測モデルの評価を行い，結果に基づき予備動作について考察する．最後に 5 章でまとめとする．

2. 関連研究

2.1 対面会議における発話予測に関する研究

対面会議における発話予測に関する研究として，顔特徴を用いた研究 [2–4]，韻律情報を用いた研究 [8], 音声対話システムの発話欲求を用いた研究 [9] が報告されている．

石井ら [2–4] は，顔特徴を用いて発話交替/継続を予測し，次の発話者を予測する手法を提案している．顔特徴は頭部運動，視線移動，口の開きの 3 つの動作を対象にしている．文献 [2] では，発話継続時の非発話者，発話交替時の非発話者，次の発話者における 3 者の頭部運動の特徴を分析している．分析結果により，会議参加者の頭部位置と回転角の特徴の違いから，発話予測に用いる特徴量を抽出している．文献 [3] では，会議参加者同士による視線移動の特徴を分析している．分析結果により，発話継続時に非発話者が先に発話者に視線を向けるという特徴から，発話予測に用いる特徴量を抽出している．文献 [4] では，口の開きの違いを分析している．口の開きとして，「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，現話者は発話継続時に口を狭く開いたままにしていることが多い特徴を分析している．しかし，これらの手法は，全て対面会議における特徴に基づき予測を行うため，Web 会議の環境に応用できるとは限らない．

大須賀ら [8] は，韻律情報を用いて発話交替/継続を予測する手法を提案している．韻律情報は基本周波数 (F0)，パワー，時間長を用いている．相手の発話の音声に基づいて韻律情報を抽出して，発話交替/継続の予測を可能としている．しかし，数秒間の無音状態が続いた際，音声を用いた発話の予測が困難である．

藤江ら [9] は，音声対話システムの発話欲求モデルに基づいて発話予測を行う手法を提案している．この発話欲求モデルは，音声対話システムがユーザに向けて発話しようとする欲求をモデル化したものを指している．1 対 1 の会議を想定したデータセットを用いているため，3 名以上の Web 会議を対象にする本研究の想定と異なる．

2.2 Web 会議における発話欲求に関する研究

Web 会議における発話欲求に関連する研究として，予備動作を用いた研究 [1,6] と呼吸動作を用いた研究 [10] が報告されている．

玉木ら [1] は，予備動作を用いて発話欲求を伝達する手法を提案している．予備動作として頷き，挙手，手を顔周辺に近づけるといった動作を検出し，会議参加者にインジケータを通じて発話欲求を提示する．評価実験では，通常通り Web 会議を利用した時の条件で比較すると，提案システムを利用することで，発話衝突の回数が減少することを報告している．この手法では，ヴァーガスら [5] の知見に基づき対面会議を前提とした予備動作を扱っている．しかし，対面会議と Web 会議では会話を行う環境が異なるため，Web 会議の環境に応用できるとは限らない．例えば，会議参加者が常にビデオをオフにしていると，手を挙げても他の会議参加者には伝わらないため,「挙手」は予備動作として現れないと考える．また,「頷き」は相手の発話内容への理解を示すために用いる可能性もあり，必ずしも発話すると限らない．そのため，個人ごとに発話予測を行うことは困難であると考えられる．

伊藤ら [10] は，呼吸動作を可視化して発話のタイミングを把握する手法を提案している．呼吸動作に関する時間（呼吸時間）として，発話直前の息を吸う時間と発話中の息を吐く時間を計測し，これらの時間情報を用いて発話のタイミングを把握している．しかし，呼吸動作は発話以外でも行うものであり，同じ呼吸時間でも発話の場合と非発話の場合があるため，呼吸時間のみを用いて呼吸直後の発話を予測することは困難である．

著者ら [6] は，発話直前の予備動作を用いた発話予測手法を提案している．予備動作をマクロな顔特徴の変化に基づいた特徴を取得することで，会議参加者の数秒後の発話予測を行う．初期的評価として，被験者 3 名の発話予測を

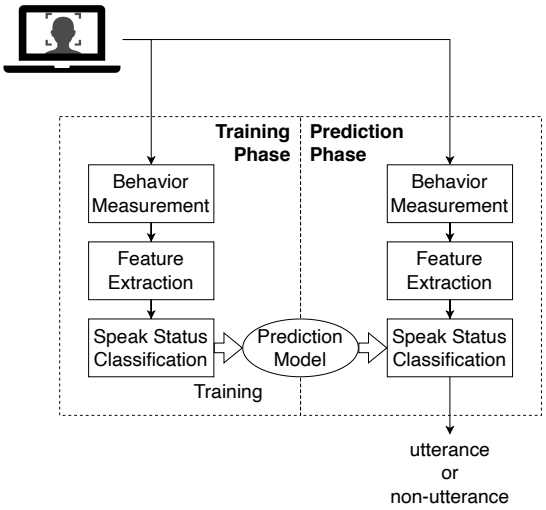


図 1 発話予測手法の概要

行った結果，F-measure で 7 割から 9 割の精度で予測が可能であることを確認している．被験者数を増やして実験を行い，この手法を評価した結果，精度が大きくばらつくという問題が発生した．本稿では，この手法を拡張し，より多くの人に適用可能な発話予測手法を提案する．

3. 発話予測手法

3.1 概要

図 1 に提案する発話予測手法の概要を示す．提案手法は，発話行動計測，特徴量抽出，発話予測の 3 つのステップで構成され，教師あり学習を用いて Web 会議映像から抽出した特徴量を用いて発話を予測する．次節以降では各ステップについて詳述する．

3.2 発話行動計測

発話行動計測では，顔特徴データ及び発話状態データを計測する．

まず，顔特徴データとして，映像データから顔特徴点の時系列変化をマクロな顔特徴，ミクロな顔特徴のそれぞれについて計測する．顔特徴点は，顔動作解析ツールである OpenFace [7] を用いて映像データの各フレームで計測する．

マクロな顔特徴では，頭部運動，視線移動，口の開きの 3 種類を計測する．マクロな視点で顔の大まかな動作を追跡し，予備動作の特徴を把握できると考える．マクロな顔特徴の一覧について**表 1** に示す．頭部運動に関しては，映像内の会議参加者から見て，左から右の水平方向を x 軸 (pose_Tx)，下から上の垂直方向を y 軸 (pose_Ty)，手前から奥方向の前後方向を z 軸 (pose_Tz) とする．さらに，頭部の x 軸，y 軸，z 軸回りの回転角度 (pose_Rx, pose_Ry, pose_Rz) も計測する．視線移動に関しては，映像内の会議参加者から見て，左から右の水平方向を x 軸方向の角度 (gaze_x)，下から上の垂直方向を y 軸方向の角度 (pose_Ty) とする．口の開きに関しては，OpenFace で出力した顔特徴点の 68 点のうち，上唇と下唇に対する特徴点 (y_62, y_66) の y 軸方向の差分を算出した値 (mouth) とする．

表 1 マクロな顔特徴の一覧

顔特徴	詳細
頭部	水平方向 (pose_Tx) 垂直方向 (pose_Ty) 前後方向 (pose_Tz)
視線	水平方向 (gaze_x) 垂直方向 (gaze_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)

表 2 Action Unit の一覧

AU	内容	AU	内容
01	眉の内側を上げる	14	笑窪を作る
02	眉の外側を上げる	15	唇の両端を下げる
04	眉を下げる	17	顎を上げる
05	上瞼を上げる	20	唇の両端を横に引く
06	頬を持ち上げる	23	唇を固く閉じる
07	瞼を緊張させる	25	顎を下げずに唇を開く
09	鼻に皺を寄せる	26	顎を下げて唇を開く
10	上唇を上げる	45	瞬きをする
12	唇の両端を引き上げる		

ミクロな顔特徴では，表情変化を計測する．表情変化の中でも顔の細かい動作に着目することで，発話前の予備動作の特徴を把握できると考える．例えば，感情の抑制をしていようとも表出してしまう微表情が存在する．微表情は 500 ms しか現れない瞬間的な動作である [11]．発話前に瞬間的に微細な表情変化を捉えることができれば，予備動作を抽出できると考える．そこで微細な表情変化を計測するため，Action Unit を用いる．Action Unit とは，顔の筋肉群の基本的な動作の単位のことである．Action Unit の中でも**表 2** に示している 17 種類の強度変化（0 から 5 の連続値）を用いる．

発話状態データは，Web 会議映像データの各フレームにおいて，各会議参加者が発話しているか否かを時系列データとして音声に基づいて計測する．

3.3 特徴量抽出

3.2 節で示した顔特徴データ及び発話状態データに対してスライディングウィンドウを適用し，各ウィンドウで特徴量を抽出する．

まず，時系列の顔特徴データ及び他の会議参加者の発話状態データをスライディングウィンドウを用いて切り出す．**図 2** にスライディングウィンドウを用いたデータ切り出しの概要を示す．図に示すように，幅 w のウィンドウを 50% ずつ重ねてずらしながらデータを切り出し，各ウィンドウのデータを用いて，予測先時間 T だけ未来の発話予測を行

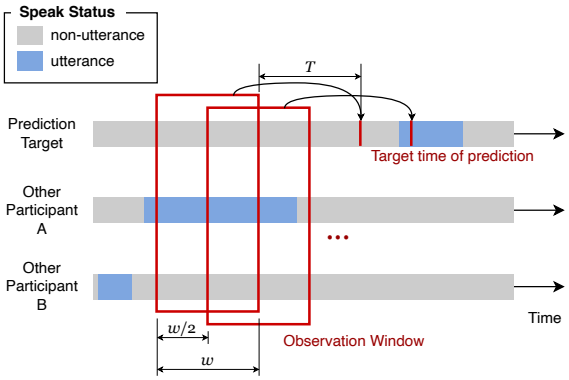


図 2 データ切り出しの概要

う．本稿では試験的に $T = 1$ 秒，$w = 1$ 秒と設定する．

次に，各ウィンドウのデータから特徴量を抽出する．マクロ・ミクロな顔特徴点データからは，基本統計量を計算して特徴量とする．マクロな顔特徴に対しては 7 種類の基本統計量（平均，標準偏差，最大値，最小値，中央値，尖度，歪度）を，ミクロ特徴に対しては 3 種類の基本統計量（標準偏差，最大値，最小値）をそれぞれ計算し，114 次元の特徴量を得る．

他者の発話状態については，発話状態データから予測対象者以外の会議参加者の発話状態を特徴量として抽出する．各ウィンドウの最後の時点のみを参照し，予測対象者以外の会議参加者が 1 人でも発話していれば他者が発話状態であるとする．

3.4 発話予測

3.3 節で述べた特徴量を用いて，教師あり機械学習により会議参加者の数秒後の発話予測を行う機械学習モデルを構築する．本提案手法では，使用する機械学習アルゴリズムは限定しない．発話，非発話の 2 クラスの分類を可能な教師あり学習アルゴリズムであればどのようなものでも適用可能である．本稿では，特徴量の重要度を評価することが可能である Random Forest を用いる．予備動作の候補となる特徴を確認にするために次元の大きい特徴量を用いるため，Random Forest を用いて過学習を防ぐ．

学習に用いる正解データは，半手動でアノテーションした．アノテーションの手順は 4.2 節で詳述する．

4. 評価実験

ミクロな顔特徴と他者の発話状態の特徴量の追加による発話予測モデルの予測精度の変化を調査するため，Web 会議の映像データを取得して評価を行った．

4.1 実験環境

実験は Web 会議ツール Zoom を用いて行った．被験者は 20 代の 9 名であり，9 名を 3 つの 3 名グループに分けて Web 会議による議論を行わせ，その映像を記録した．映像データのフレームレートは 25 Hz である．

表 3 会議条件

項目	詳細
会議人数	1 グループ 3 名
会議シナリオ	10 分間程度の議論
画面共有の有無	なし
ミュート機能の使用	なし
別画面の閲覧	なし
カメラのオンオフ	常にオン

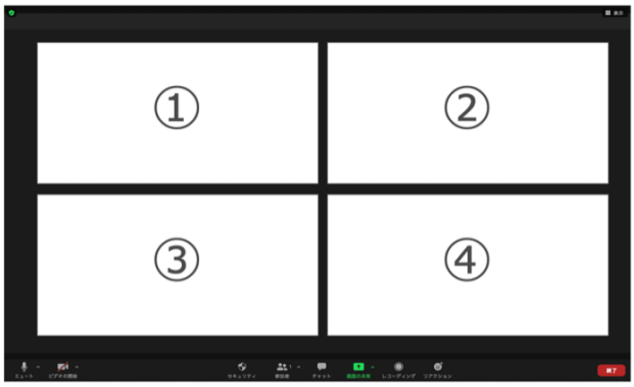


図 3 Zoom の画面構成

表 4 評価モデルの構築に使用する特徴量

モデル	使用する特徴量
C1	マクロな顔特徴
C2	ミクロな顔特徴
C3	マクロ・ミクロな顔特徴
C4	マクロ・他者の発話状態
C5	ミクロ・他者の発話状態
C6	全ての特徴量

表 3 に会議条件を示す．発話の予備動作は会議環境の影響を受けると考えられるため，本実験では表 3 に示す制限を設けた会議環境で実施した．画面共有，ミュート機能，別画面の閲覧はなしとした．また，カメラを常時オンして顔を見せて議論するように被験者に指示を行った．会議中の役割は事前に決めずに，被験者に委譲した．

Zoom の画面構成を**図 3** に示す．Zoom の画面はフルスクリーンに設定し，被験者を①, ②, ③の位置に，実験実施者を④の位置に配置した．実験実施者は実験を記録するために参加している．実験への影響を避けるために実験実施者のカメラは常時オフ，マイクをミュートに設定し，会議の議論には参加しなかった．

評価では，**表 4** に示す 6 種類の特徴量の組み合わせで発話予測モデル C1 から C6 を構築し，それぞれの予測精度を比較した．

4.2 発話・非発話のアノテーション

特徴量の抽出で用いる発話・非発話区間を抽出するため，映像・音声データに対して発話・非発話のアノテーション

表 5　各被験者の C1〜C6 における F-measure

被験者	C1	C2	C3	C4	C5	C6	平均
A	0.703	0.672	0.696	0.682	0.670	0.696	0.687
B	0.614	0.597	0.612	0.610	0.602	0.619	0.609
C	0.611	0.587	0.594	0.616	0.592	0.599	0.600
D	0.611	0.721	0.659	0.600	0.701	0.647	0.657
E	0.655	0.666	0.661	0.675	0.672	0.663	0.665
F	0.652	0.665	0.647	0.638	0.673	0.653	0.655
G	0.820	0.756	0.773	0.810	0.758	0.781	0.783
H	0.702	0.744	0.724	0.717	0.752	0.720	0.727
I	0.652	0.651	0.652	0.662	0.649	0.658	0.654
平均	0.669	0.673	0.669	0.668	0.674	0.671	0.671

を行った．アノテーションツールである ELAN[*1] [12] を用いた．この時，発話区間は，非発話区間が 700 ms 未満の連続した音声区間と定義した．

機械的に判別するため，アノテーションの半自動化を行った．発話区間検出ライブラリである inaSpeechSegmenter[*2] [13] を用いて，発話区間の自動アノテーションを行った．自動アノテーションが終わった後，誤検出を探し，目視で手動アノテーションを行った．この手動アノテーションの際，判断基準が揺るがないよう，1 名のアノテータのみが修正を行った．今回は笑い声や相槌などの言語内容として解釈不可能な発話区間は対象外とした．

4.3　評価指標

本研究では，F-measure，Precision，Recall を用いて発話予測モデルの評価を行う．Precision は発話と予測されたうち，実際に発話である割合，Recall は実際に発話であるうち，正しく発話と予測できた割合である．F-measure は Precision と Recall の調和平均である．各被験者それぞれで評価値を算出し，それら全てを平均したものを実験結果に示す．

4.4　評価結果

表 5 に，各被験者の予測精度を示す．表は，C1 から C6 のそれぞれのモデルを用いた場合の F-measure を示しており，被験者ごと，モデルごとの平均値も併せて示している．色付きのセルは，各被験者で最も精度が高かった場合を示している．

各特徴量群の平均 F-measure について確認する．C1 のモデルは，先行研究の手法で構築したモデルである．C1 に比べて，C2，C5，C6 のモデルの方が高い F-measure を示している．C5 の F-measure は最も高い 0.674，C4 の F-measure は最も低い 0.668 である．この結果から，ミクロな顔特徴と他者の発話状態の特徴量を追加することで予測精度が向上することを確認した．

各モデルで最大の精度を示した被験者の数は，C1：2 名，C2：1 名，C4：3 名，C5：2 名，C6：1 名である．全被験者の平均 F-measure ではモデル C5 の F-measure が最も高い一方で，C4 が最大 F-measure となる被験者数が最も多い．被験者ごとに最大精度となった F-measure の平均値は 0.694 である．被験者に最大 F-measure となるモデルが異なることから，各被験者で異なる特徴量を用いる方が良いと言える．

[*1] https://archive.mpi.nl/tla/elan
[*2] https://github.com/ina-foss/inaSpeechSegmenter

表 6　被験者ごとに F-measure が最大となるモデルを用いた場合の Precision，Recall

被験者	モデル	Precision	Recall
A	C1	0.668	0.749
B	C6	0.524	0.764
C	C4	0.541	0.725
D	C2	0.660	0.806
E	C4	0.560	0.814
F	C5	0.555	0.864
G	C1	0.761	0.914
H	C5	0.816	0.705
I	C4	0.565	0.807
平均	—	0.628	0.794

続いて，各被験者の F-measure が最大となるモデルを用いた場合の Precision, Recall について確認する．**表 6** に，被験者ごとに F-measure が最大となるモデルを用いた場合の Precision，Recall を示す．Recall の平均は 0.794 であり，1 で示した目標精度の値に概ね近いことが確認できる．一方で，Recall と比べてが大幅に低い Precision となる被験者がいることが分かる．Precision 平均は 0.628 であり，Recall の平均と比べても低い．これは，発話を示す予備動作が非発話時にも観測されるためと考えられる．実際，実験の映像データを確認すると「口を開いたのに発話しない」などの状況があることが確認できた．

最後に特徴量の重要度を確認する．**図 4** に特徴量の重要度上位 10%の出現回数を示す．図より，口の開きの最大値（mouth_max），AU25 の最大値（AU25_max）は数人の被験者で高い重要度となっていることが分かる．一方，他の特徴量は 1 回ずつ出現しており，各被験者で有効な特徴量が異なることが確認できた．

4.5　考察

図 4 の特徴量の重要度に基づいて会議の映像データを観察し，それぞれの重要度が発話予測にどのように影響しているかを確認し，予備動作になるのかを考察する．

4.5.1　マクロな顔特徴の予備動作に関する考察

特徴量の重要度に基づきマクロな顔特徴である口の開き (mouth)，頭部運動の水平方向 (pose_T)，前後方向 (pose_Tz) の順で考察する．

会議中のマクロな顔特徴の変化の一例について**図 5** に示す．縦軸は顔特徴点の出力値，横軸は会議時間である．青の領域は発話区間を表している．

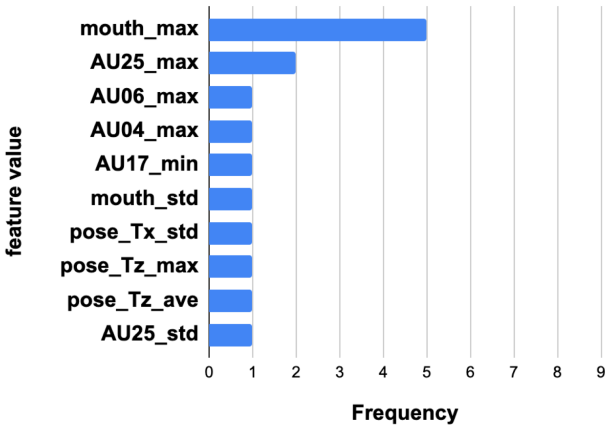
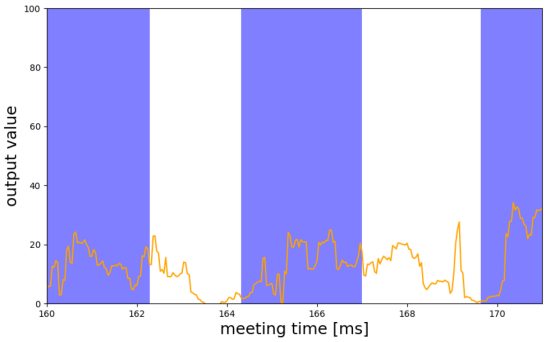
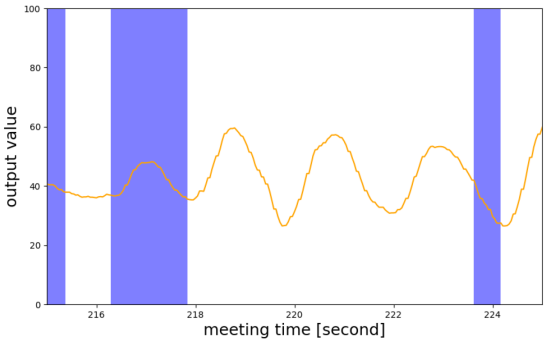


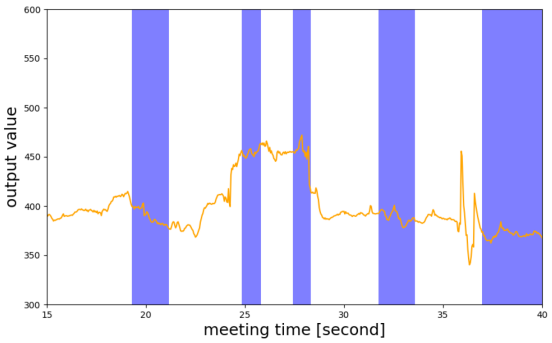
図 4 特徴量の重要度上位 10%の出現回数



(a) 被験者 A の口の開きデータ

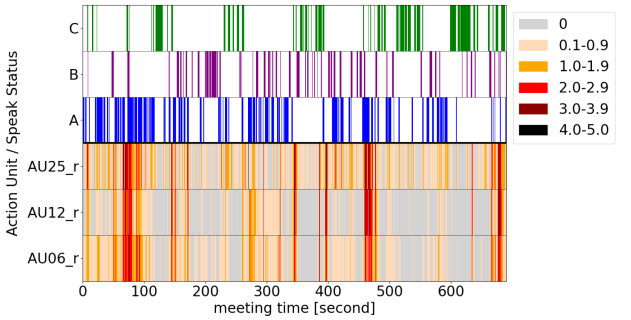


(b) 被験者 E の水平方向の頭部運動データ

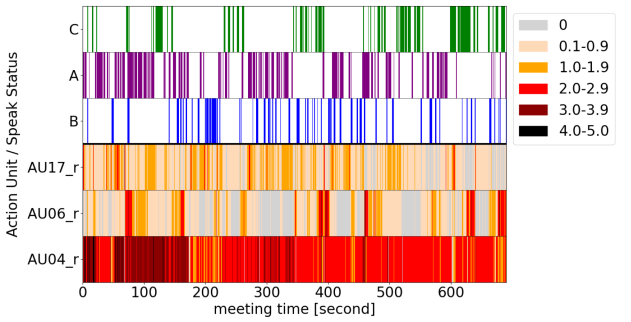


(c) 被験者 F の前後方向の頭部運動データ

図 5 会議中のマクロな顔特徴の変化の一例



(a) 被験者 A



(b) 被験者 B

図 6 会議中の AU 強度変化の一例

mouth では，発話直前に小さく口を開いて発話しようとする特徴が現れていると考える．図 5 (a) から発話直前に小さく口を開く特徴が確認できる．これにより，発話直前に口を開く動作を捉えることが可能だと考える．対面会議においても発話する直前に会議参加者は口を小さく開ける特徴が現れることを報告している [5]．このことから，Web 会議でも同様の予備動作が現れると考えられる．

pose_Tx では，顔を左右に傾ける特徴が現れていると考える．図 5 (b) から連続的に顔を左右に傾ける特徴が確認できる．被験者 E は 10 分間の議論中，頭部の水平方向動作が 8 回現れ，そのうち 6 回は発話直前に現れることが確認されている．他の被験者には同様の動作が見られず，被験者特有の癖である可能性がある．

pose_Tz では，発話直前に瞬間的に前傾姿勢，もしくは一定時間静止する特徴が現れていると考える．図 5 (c) から発話直前に値の上昇，値が一定になっていることが確認できる．会話中の前傾姿勢は，話に興味を持っている，もしくは親しい人との会話で現れる特徴として現れている可能性がある．

4.5.2 ミクロな顔特徴の予備動作に関する考察

特徴量の重要度に基づきミクロな顔特徴である AU25, AU06, AU04, AU17 の順で考察する．

被験者 A と B の AU 表出タイミングを**図 6** に示す．縦軸は各 Action Unit の強度変化，被験者 A, B, C の会話状態を可視化して対応づけており，横軸は会議時間である．

まず AU25（顎を下げずに唇を開く動作）は，マクロな顔特徴の口の開きと同様に重要度が高かった．このことか

ら Action Unit からでも口の開きの特徴が発話予測に重要であることが考えられる．

AU06（頬を持ち上げる）は被験者 A の重要度が高い．図 6 (a) から AU06 と同じタイミングで AU12，AU25 の強度が上がっていることが確認できる．AU06，AU12，AU25 の組み合わせは，幸福の感情の表情であると示されている [14]．また，発話交替よりも，発話継続に現れていることも確認できる．このことから，発話継続時に笑顔，もしくは笑う仕草が現れる予備動作となることが考えられる．

AU04（眉を下げる），AU17（顎を上げる）は，被験者 B の重要度が高い．図 6 (b) から AU06 の値は常に高いことが確認できる．映像データからは AU04 の動作が確認できないため，誤検出の可能性が高いと考える．AU17 は被験者 A の発話中に強度が上昇することを確認した．映像データと照合すると，顎を上げて，斜め上を見ている様子が見られた．これは考える動作になり，被験者 A の発話に対して，考える動作が現れていると考える．

以上の重要度の高い特徴量に関する考察を踏まえ，発話予測において口の開きと個人ごとに異なる特徴量が重要であることを示した．

4.5.3 他者の発話状態の考慮に関する考察

他者の発話状態の特徴量は，被験者 9 名の特徴量の重要度上位 10%以内に含まれず，重要ではないことを確認した．今回追加した発話の有無のみでは，発話予測に必要な特徴を捉えきれていない可能性がある．そこで今後は発話時間などの特徴量の抽出を検討する．音声特徴量，発話継続時間，無音継続時間などが考えられる．質問や確認などの相手に問いかける発話の場合，語尾にアクセントをつけることが考えられ，音声特徴量が有効であると考える．また，他者の発話時間が継続すると，発話欲求の低下，もしくは発話の割り込みが発生する．反対に，無音時間が継続すると，発話欲求が高くなる．そのため，発話継続時間，無音継続時間も検討する余地があると考える．

5. おわりに

本稿では，Web 会議における円滑に会議進行に向けて，会議参加者の数秒後の発話を予測する手法を提案した．先行研究では，発話前に行う予備動作としてマクロな顔特徴を用いていたが，新たにミクロな顔特徴と他者の発話状態の特徴量を追加し，発話予測モデルの評価を行った．

評価実験では，個人ごとに有効な特徴量を用いた発話予測モデルを用いた場合，被験者ごとに最大精度となった F-measure の平均値は 0.694 という結果が得られた．また，発話予測モデルにおける各特徴量の重要度を分析した結果，口の開きに関する特徴量は多くの被験者で共通して高い重要度を持つものの，発話予測に重要な特徴量の多くは被験者ごとに異なることを示した．

今後の課題として，会議環境に依存しない発話予測モデルの構築が挙げられる．本稿では会議条件を固定しているため，会議参加者の人数の変化に対応した予測は困難である．そこで会議の条件や会議参加者の人数を変えて実験を行い，発話予測モデルを改善する．加えて，他の予備動作の検討，学習器の検討なども行う予定である．

参考文献

[1] 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).
[2] 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, No. 4, pp. 1116–1127 (2016).
[3] 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. JSAI2015, pp. 2L32in–2L32in (2015).
[4] Ishii, R., Otsuka, K., Kumano, S., Higashinaka, R. and Tomita, J.: 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, p. 70 (2019).
[5] ヴァーガス M.F.：非言語 (ノンバーバル) コミュニケーション，新潮社 (1987).
[6] 山田楓也，白石　陽，石田繁巳：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，情報処理学会マルチメディア, 分散, 協調とモバイルシンポジウム（DICOMO）(2021).
[7] Baltrušaitis, T., Robinson, P. and Morency, L.-P.: OpenFace: An Open Source Facial Behavior Analysis Toolkit, *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10 (2016).
[8] 大須賀智子，堀内靖雄，西田昌史，市川　熹：音声対話での話者交替／継続の予測における韻律情報の有効性，人工知能学会論文誌，Vol. 21, No. 1, pp. 1–8 (2006).
[9] 藤江真也，片山颯人，小林哲則：音声対話システムにおける発話期待度の逐次推定に基づくターンテイキングタイミングの予測, 人工知能学会全国大会論文集, Vol. JSAI2020, pp. 2Q1GS1003–2Q1GS1003 (2020).
[10] 伊藤淳子，永吉亜優：Web 会議における話者交代円滑化のためのアバターによる呼吸の視覚化，情報処理学会研究報告デジタルコンテンツクリエーション（DCC），Vol. 2022-DCC-30, No. 17, pp. 1–6 (2022).
[11] Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H. and Fu, X.: How Fast Are the Leaked Facial Expressions: The Duration of Micro-Expressions, *J Nonverbal Behav*, Vol. 37, No. 4, pp. 217–230 (2013).
[12] Nijmegen, Max Planck Institute for Psycholinguistics: ELAN.
[13] Doukhan, D., Carrive, J., Vallet, F., Larcher, A. and Meignier, S.: An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5214–5218 (2018).
[14] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. and Pollak, S. D.: Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements, *Psychological Science in the Public Interest*, Vol. 20, No. 1, pp. 1–68 (2019).