Vol.2020-MBL-97 No.8
Vol.2020-ITS-83 No.8
2020/11/24

Web 会議の非発話区間における 顔特徴量を用いた発話予測手法の検討

山田楓也 [1,a)]　白石陽 [1,b)]

概要：近年，在宅勤務やモバイルワークといったテレワークの推進，働き方改革宣言による働き方の多様化などから Web 会議サービスの利用が増加している．Web 会議サービスを用いた際，会議参加者が相手の状況を把握しづらいことがある．特に，他の参加者の発話の開始タイミングが把握できず，複数人による発話の重なり（発話衝突）が発生してしまう．発話衝突が発生することで，会話の中断や会議参加者の発言意欲の低下に繋がり，消極的な会議になることが考えられる．また，対面コミュニケーション時に相手の顔や体の動きなどの非言語情報を利用して円滑な会話を行うことから，発話と非言語情報の関係性が高いと考えられる．例えば，会議参加者の方向に視線を向ける，顔を傾けるなどといった動作が挙げられる．そこで，本研究では視線や頭部運動などの顔の特徴量を用いて発話予測を行う．Web カメラから参加者の顔の特徴を追跡し，非発話区間における顔の特徴を抽出する．発話予測に用いる特徴量は，会議参加者の顔から抽出した特徴点の変化量を用いることを検討している．非発話区間から抽出した特徴量から予測モデルを構築し，そのモデルを用いて予測を行う．

キーワード：Web 会議，発話衝突，発話予測，顔の特徴

1. はじめに

近年，在宅勤務やモバイルワークといったテレワークの推進，働き方改革宣言による働き方の多様化などから Web 会議サービスの利用が増加している[1]．遠隔コミュニケーションの増加とともに，Web 会議サービスや共同作業ツールの品質も向上している．Web 会議サービスは，テレワークだけでなく不特定多数が集まる各種イベントや教育現場などでも遠隔コミュニケーションの手段として活用されている．

Web 会議では，対面コミュニケーションと異なり，相手の顔の表情や頷きなどの非言語情報を把握しづらい．そのため，参加者が他者の発話の開始タイミングが把握できず，発話の重なり（発話衝突）が発生する．発話衝突に関する研究として，Sacks らは，発話衝突の発生時に，会話の流れを戻すために参加者のうち片方が発話を中断し，もう片方の参加者がもう一度発言の言い直しを行うと述べている[2]．また，玉木らは，Web 会議で発生する発話衝突は対面会議に比べて 30 倍増加すると述べている[3]．発話衝突の発生により，予期しない発言の阻止や会話の流れの中断し，会議進行を妨げると考えられる．さらに，会議進行が円滑でなければ，会議参加者の発言意欲の低下につながり，消極的な会議になることが考えられる．そこで，発話を中断させる発話衝突を解消し，Web 会議による円滑な会議を実現することが重要である．

発話衝突では，参加者が会話を終えてもう 1 度発話をした際に，他の参加者も発話するタイミングで発生すると考えられる．吉田は，発話衝突の分類として，声を発した瞬間に発話同士が重なるタイミングを挙げている[4]．このタイミングで発話衝突が発生すると，一度会話の流れが止まってしまう．このことから，発話する前の全体が無音になっている状態（非発話区間）では，発話衝突が発生しやすいと考えられる．そこで，非発話区間において参加者の発話を予測することで，会話の流れを中断させる発話衝突を解決できると考える．そこで，本研究では，Web 会議の非発話区間における発話予測を行う．

発話予測に関する研究として，韻律情報[5], [6]や発話前の非言語情報[7], [8], [9]を用いた研究がある．文献[7], [8]では，対面会話時の参加者の頭部運動や視線交差パターンを用いて発話予測を行っている．また，文献[9]では，Web 会議における発話前の頭の動きや頷きを検出して，発話衝突を低減するシステムが構築されている．これらの研究から，発話予測に頭部運動や視線などの顔情報を用いるのが有効であると考えられる．よって，本研究では発話予測を行うために，顔情報に着目する．

本研究では，Web 会議の非発話区間における顔特徴量を用いた発話予測手法を提案する．提案手法では，Web 会議サービスを用いて会話を行い，音声と映像データを収集する．収集した音声データの分析区間は非発話区間とする．発話予測に用いる特徴量は，発話前に行った動作の変化量を用いる．文献[7]の知見をもとに，発話予測では 3 段階で処理を行う予測モデルを構築する．1 段階目は，話者交替であるか話者継続であるかの予測である．2 段階目は，話者交替時に誰が次の発話者になるかの予測である．3 段階目は，特定された次の発話者が発話を開始するタイミングの予測である．それぞれの予測モデルを構築し，予測精度の検証を行う．本稿では，本研究で使用する顔特徴量の検

1 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.
a) b1017040@fun.ac.jp
b) siraisi@fun.ac.jp

討を行う．実際に Web 会議サービスを利用して会話を行い，参加者の顔の映像と音声データを収集し，発話に関係する特徴的な動作の分析を行った．

2. 関連研究

発話予測の研究として，韻律情報を用いた研究[5]，[6]，対面会話時の非言語情報を用いた研究[7]，[8]，Web 会議時の非言語情報を用いた研究[10]がある．

2.1 韻律情報を用いた発話予測の研究

韻律情報を用いた発話予測の研究として，原らは，対面の面接形式を想定して，相槌とフィラーを用いた話者交替の予測を行っている[5]．フィラーとは，「えー」や「あのー」といった言い淀み時に現れ，場繋ぎに行う表現である．相槌とフィラーそれぞれで予測し，話者交替の予測を行う．マルチタスク学習の仕組みを用いて，ニューラルネットワークで話者交替の予測モデルを構築している．しかし，Web 会議において相槌やフィラーは，他の参加者が同時に話した場合，聞き取ることが困難になる．そのため，参加者が相槌とフィラーを発する前に発話を予測する必要がある．小川らは，話速，基本周波数（F0），パワーの変化量を用いて 2 段階で話者交替の予測を行っている[6]．1 段階目は，発話中か無音かどうかを判別する．2 段階目は，1 段階目の発話中から無言に切り替わったところで話者交替及び継続かどうかの予測を行う．しかし，音声情報のみでは相手が発話する予兆を抽出することは困難である．

したがって，Web 会議において韻律情報を用いて発話予測を行うことは困難であると考えられる．

2.2 非言語情報を用いた発話予測の研究

2.2.1 対面会話時の非言語情報を用いた発話予測の研究

対面会話時の非言語情報を用いた発話予測の研究として，石井らは，対面での複数人会話を想定して，頭部運動[7]と視線交差のタイミング構造[8]を用いて次の話者の予測を行っている．文献[7]では，頭部の各座標位置と各回転角のデータから変化量，振幅と周波数の平均を算出する．算出したデータをもとに，話者継続時の非話者，話者交替時の非話者と次話者の 3 者間で頭部運動の特徴を抽出する．抽出した特徴から話者交替及び継続の予測を行っている．文献[8]では，視線交差が起きた際，話者と非話者どちらが先に視線を向けているかの情報から特徴を抽出し，話者交替及び継続の予測を行っている．また，話者と非話者の視線交差をした際，視線交差をしていない非話者が次話者になる時に，発話開始タイミングに差がある特徴を用いて発話末から次の発話開始までの発話間隔を予測している．しかし，これらの研究では，対面による複数人での対話を想定している．対面であれば参加者の方向を見る動作や頭を動かす動作を行う．しかし，Web 会議を想定した場合，これらの手法をそのまま適用することは困難であると考えられる．

2.2.2 Web 会議時の非言語情報を用いた発話予測の研究

Web 会議時の非言語情報を用いた発話予測の研究として，玉木らは，Web 会議の複数人会話を想定して，非言語情報を用いた発話予測を行っている[10]．具体的には，発話前の動作を検出し，動作を検出するごとに発話欲求度合いを表すインジケータを参加者に提示するシステムの構築を行っている．発話欲求度合いとは，Kinect センサを用いて「頷き」，「挙手」，「手を顔周辺へ動かす動作」を検出した後，それぞれにスコアの総和を可視化したものである．発話欲求度合いを用いて参加者にフィードバックする．Web 会議において上記の発話前の予備動作には，個人差が大きいと考えられる．Kinect センサでは，頭部，首，肩 関節，肘関節，手首などの 3 次元座標を取得する．手や顔，頭部などの位置関係が明らかになっても，顔の表情などの具体的な情報がわからないため，文献[10]で定義している予備動作を用いて発話予測を行うことは困難であると考えられる．

3. 提案手法

本章では，まず 3.1 節で本研究の目的について，3.3 節で本研究の提案システムについて，3.3 節で研究課題とアプローチについて述べる．3.4 節では発話予測に用いる特徴量について，3.5 節では発話予測の流れについて述べる．

3.1 研究目的

本研究の目的は，Web 会議における発話衝突を解消するために，顔の特徴量を用いて発話予測を行うことである．

3.2 提案システム

本研究の提案システムの全体像を図 1 に示す．

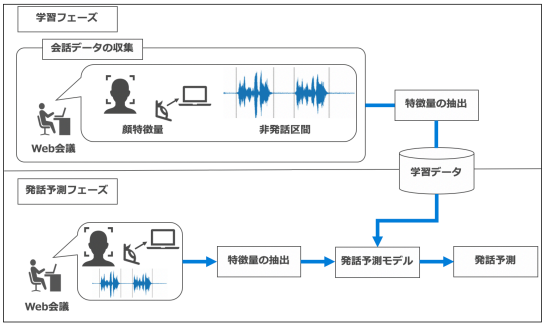


図 1 提案システムの全体像

本研究における提案システムは，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 会議サービスを用いた会話時の映像データから非発話区間にお

Vol.2020-MBL-97 No.8
Vol.2020-ITS-83 No.8
2020/11/24

ける参加者の顔の特徴を抽出し，学習モデルを作成する．発話予測フェーズでは， Web 会議サービスを用いた会話時の映像データから非発話区間における顔の特徴を抽出し，作成した学習モデルを用いて発話予測を行う．

3.3 研究課題とアプローチ

本研究では，以下の 3 つを研究課題とする．

a) 発話前に行う特徴的な動作の収集方法の検討
b) 発話予測に有効な特徴量の検討
c) 発話予測手法の検討

課題 a に対するアプローチとして，動画解析ツール ELAN[12]を用いて非発話区間における顔の特徴分析を行う．Web 会議サービスを用いた会話時の会話データを収集する．非発話区間を把握するために， ELAN を用いて音声波形の表示を行う．ELAN で映像を分析し，非発話区間に対して参加者の非言語情報をラベリングする．また，収集した映像データからオープンソースソフトウェアである OpenFace[11]を用いて顔の特徴を抽出する．石井らの研究[7], [8]では，発話予測を行うために頭部運動と視線交差パターンが有効であると述べている．さらに，玉木らの研究[9]では，Web 会議における発話前の特徴的な動作として頭部運動が有効であると述べている．Web 会議の特徴として，画面上でやりとりを行うため，画面を見ている視線や頭の動きから発話前の特徴的な動作が現れると考えられる．これらのことから，顔の特徴の中でも，頭部運動と視線を用いる．非発話区間における発話前の参加者の特徴を抽出する．

課題 b に対するアプローチとして，発話前に行った動作の変化量を特徴量に用いることを検討している．発話前に行う動作として，非発話区間における頭部運動と視線に着目する．頭部運動は，頭部の位置（3 軸），頭部の回転角（3 軸）から取得する．視線は正面を 0 度とした際に，X，Y 軸方向にどの程度向いているかの角度を用いる．これらの値から変化量を算出する．

課題 c に対するアプローチとして，話者交替または継続の予測，次話者と発話開始タイミングの予測することを検討している．発話予測では，3 段階で処理を行う予測モデルを構築することを検討している．1 段階目は，話者交替であるか話者継続であるかの予測を行う．2 段階目は，話者交替時に誰が次の発話者になるかの予測を行う．3 段階目は，特定された次の発話者が発話を開始するタイミングの予測を行う．それぞれの予測モデルを構築し，予測精度の検証を行う．学習モデルは SVM（Support vector machine）を用いることを検討している．

3.4 発話予測に用いる特徴量

Web 会議時の参加者の顔を収録した映像データから OpenFace を用いて顔の特徴の抽出を行う．OpenFace を顔画像に適用させた画像の例と顔の特徴点を図 2 に示す．本研究で扱う顔の特徴は，頭部運動と視線である．頭部運動には，頭部の位置（Pose_Tx，Pose_Ty，Pose_Tz）と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横に傾ける動作を把握することが可能である．視線には，視線の横方向（gaze_angle_x）と縦方向（gaze_angle_y）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することが可能である．これらの特徴点をトラッキングし，一定時間ごとの変化量を算出する．算出したデータを発話予測に用いる特徴量とする．

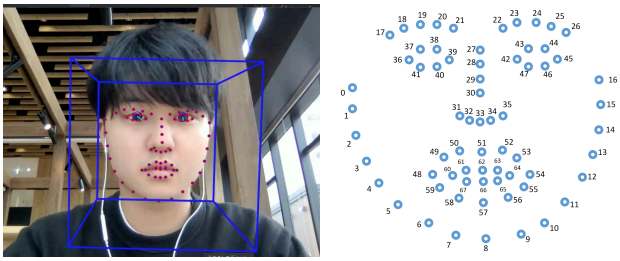


図 2 OpenFace の出力例

3.5 発話予測の流れ

発話予測の流れを以下に示す．

(1)会話データの収集
(2)分析結果から特徴の抽出
(3)発話予測

(1)として，発話予測に必要な顔の特徴を抽出するために，Web 会議サービスを用いた会話時のデータ収集を行う．複数人で行う Web 会議を想定し，会議シナリオを設定する．Web 会議中に映像と音声データを収集する．収集した映像データを用いて，顔の特徴を抽出する．映像データの分析区間として，発話者が発話する前を非発話区間とする．非発話区間は，ELAN を用いて発話終了から次の発話の直前にラベリングを行い，その区間の映像データを抽出する．非発話区間から顔の特徴を抽出して，1 秒あたりの変化量を算出する．

(2)として，(1)で算出した特徴量の分析を行うため，t 検定を行い，それぞれの特徴量で話者交替及び継続時に有意差があるかを検証する．有意差がある特徴量のみを用いて話者交替の予測を行う．

(3)として，発話予測を行うために，3 つの段階で発話予測モデルを構築する．1 段階目は，話者交替か話者継続のどちらが起こるかのモデルを構築する．2 段階目は，話者交替時に誰が発話者になるかのモデルを構築する．3 段階目は，特定された発話者が発話を開始するタイミングのモ

Vol.2020-MBL-97 No.8
Vol.2020-ITS-83 No.8
2020/11/24

デルを構築する．それぞれの予測モデルを構築し，予測精度の検証を行う．

4. 予備実験および考察

本章では，Web 会議における顔の特徴的な変化の調査を行った予備実験について述べる．4.1 節では Web 会議サービスを用いた会話データの収集について述べる．4.2 節では収集したデータの分析について述べる．最後に 4.3 節では考察を述べる．

4.1 会話データの収集

4.1.1 実験環境

予備実験の実験環境について表 1 に示す．

表 1 予備実験の実験環境

ハードウェア（PC）	
OS	Catalina 10.15.7
CPU	Intel Core i7 2.5GHz
RAM	16.00 GB
使用ツール	
顔解析	OpenFace2.2.0
動画解析	ELAN 5.9
動画収録	QuickTime Player 10.5
Web 会議	Zoom 5.3.1

今回の予備実験の目的は，Web 会議における発話前の顔の特徴的な変化を調査することである．被験者は，20 代の男子大学生 3 名とした．被験者全員が Web 会議サービスの Zoom を使用し，10 分程度のファシリテーションを設けないアイディア出し（創造会議[13]）を行った．アイディア出しをする際，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収集した．Zoom 画面内の構成を図 3 に示す．

図 3 Zoom による画面レイアウト

Zoom の画面はフルスクリーンに設定した．被験者の画面を①の位置に配置し，チャット機能は④の位置に配置を行った．また，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，約 30fps で被験者全員の顔の映像データを収集した．

4.1.2 ELAN を用いた会話データの解析

映像・音声データを収集した後，ELAN を用いて，非発話区間の抽出を行った．4.1.1 節で収集したデータを ELAN に表示した．ELAN を用いた分析の様子について図 4 に示す．

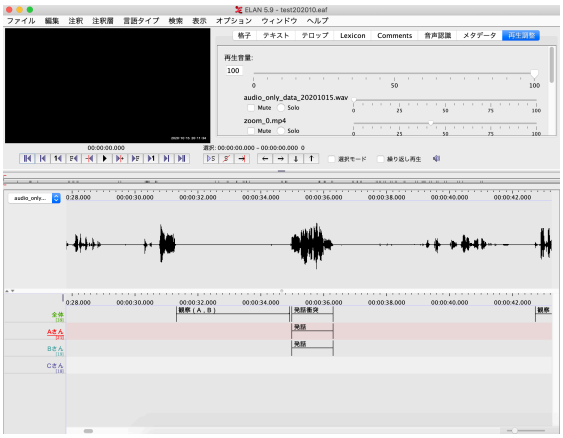


図 4 ELAN を用いた分析の様子

音声データと映像データを観察し，各参加者の様子をラベリングした．ラベリングの内容は，発話前の非発話区間における観察対象者（A，B，C），各参加者の発話開始タイミング，発話衝突とした．

4.2 収集したデータの分析

4.2.1 収集したデータ

非発話区間における顔特徴量を収集したデータの一例として，被験者 A の音声データ（図 5），視線方向（図 6），頭部の位置（図 7），頭部の回転角（図 8）のデータをグラフ化し，以下に示す．縦軸は出力値，横軸は時系列データのサンプル番号を示す．また，図 5 と図 6 のグラフにある青の領域は，非発話区間を指す．また，図 7 のグラフにある緑の領域と図 8 の黒の領域は，それぞれ被験者 A の音声データの別の非発話区間を指す．

OpenFace で顔認識した特徴量を出力した際，顔認識が失敗した時の欠損データが含まれていた．また，被験者の 1 名の映像データでは，人物のポスターを壁に貼っている影響により，誤認識しているデータも存在した．そのため，上記の欠損データを含めずに分析を行った．

Vol.2020-MBL-97 No.8
Vol.2020-ITS-83 No.8
2020/11/24

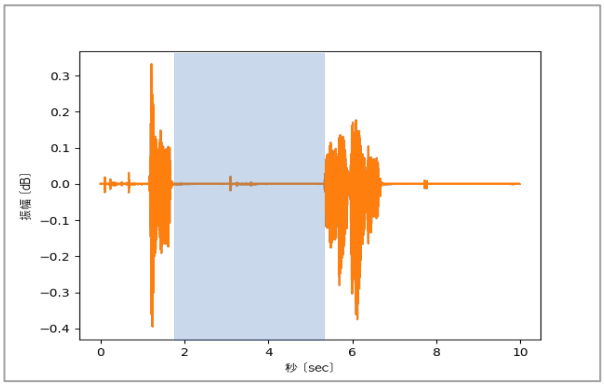
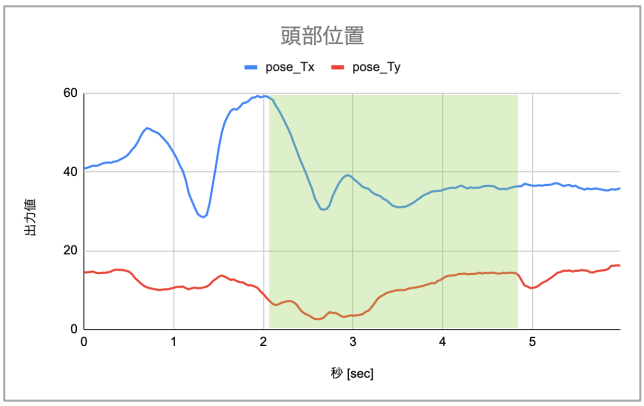


図 5 音声データの一例

図 6 視線データの一例



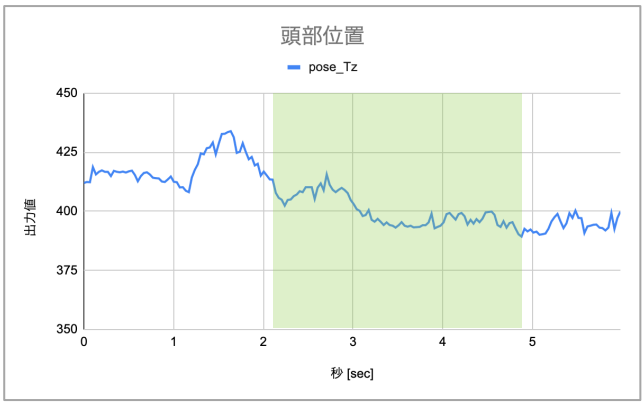


図 7 頭部位置のデータ一例

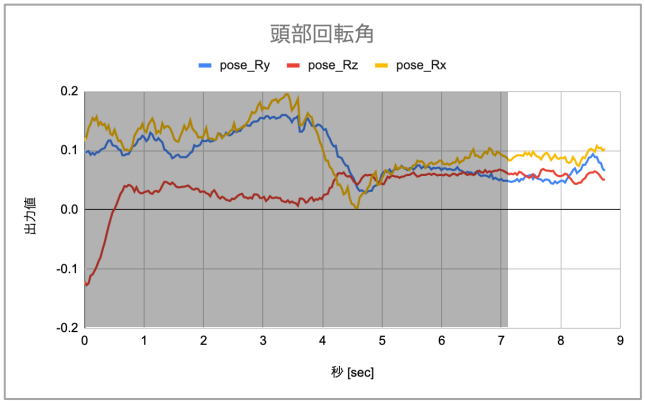


図 8 頭部の回転角データの一例

4.2.2 収集したデータの分析

図 6 の視線方向のデータでは, 非発話区間に gaze_angle_x が正の方向に変化しており，gaze_angle_y は負の方向に変化している．横方向に視線を動かしていることから，他の被験者が映っている画面に視線の方向を変えたことが考えられる．図 7 では，pose_Tx と pose_Tz は変化し，pose_Ty は大きな変化は見られなかった．Pose_Tz は徐々に減少しており，発話 ccc する前に顔をカメラに近づける可能性が考えられる．図 8 では，非発話区間内の 4 秒から 5 秒にかけて値に変化があった．動画と比較すると，顎に手をつけて頬杖をついているシーンであることがわかった．

4.3 考察

本稿では，Web 会議における発話前に行う顔の特徴的な変化を調査するため，Zoom を用いて被験者の映像と会話データを収集し，OpenFace を用いて非発話区間における特徴点を検出した．

4.2 節の結果から，視線方向と頭部運動は発話予測の特徴量として有効であると考える．視線方向に関する特徴量では，Web 会議の参加者が画面上に映っている参加者を見ることから，発話前の特徴に関連すると考える．しかし，Web 会議参加者が画面オフの場合，同じような視線データを収集できると限らないため，他の特徴の組み合わせについて検討を行う必要がある．頭部運動に関する特徴量では，顔をカメラに近づける動作が発話前の特徴に関連すると考える．また，4.1 節の予備実験の方法を改善する必要である．実験中では，普段の日常会話と比べて，発話する回数が少なかったと考える．原因として，アイディア出すために参加者思考する時間が長いことが考えられる．そのため，各参加者の発話数が多くなるように実験環境を構築する必要がある．

5. まとめ

本研究の目的は，Web 会議における発話衝突を解消する

ために，顔特徴量を用いて発話タイミングを予測することである．本稿では，研究課題の発話前に行う特徴的な動作の収集方法の検討に関する予備実験について述べた．結果として，非発話区間における視線方向と頭部運動は，発話予測に用いる特徴量として有効であると考える．

今後は，会話のデータ数を増やし，発話予測に有効な特徴量の検討を行う．本稿で行った予備実験の結果をもとに，変化量の平均を算出し，話者交替及び継続時のそれぞれに差があるのかを t 検定を用いて分析する．t 検定の結果から話者交替及び継続時に有意差が見られた場合，そのデータを特徴量とする．

参考文献

[1] 総務省，情報通信統計データベース，https://www.soumu.go.jp/johotsusintokei/statistics/data/2200521.pdf (最終アクセス日：2020/10/4).
[2] H.Sacks, A.E.Schegloff and G.Jefferson，“A Sim - plest Systematics for the Organization of Turn - Taking for Conversation”，Language，Vol.50，No.4，Pt 1，pp.696-735 (1974).
[3] 玉木 秀和，東野 豪，小林 稔，井原 雅行，“遠隔会議における発話の衝突と精神的ストレスの関係”，情報処理学会研究報告，Vol.2011-GN-79，No.10，pp.1-6 (2011)．
[4] 吉田智子, “発話の重なり現象の考察−電話の会話分析−”，日本語教育論集，No.6，pp.76-93（1989）．
[5] 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”, 第 80 回全国大会講演論文集, Vol.2018, No.1，pp.409-410 (2018).
[6] 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).
[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] 玉木 秀和，東野 豪，小林 稔，井原 雅行，岡田 謙一，“遠隔会議における発話衝突低減手法”，情報処理学会研究報告，Vol.2011-GN-79，No.10，pp.1-6 (2011).
[10] 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).
[11] T.Baltrusaitis, P.Robinson, and L-P.Morency, “OpenFace: An Open Source Facial Behavior Analysis oolkit,” Proc. of the 2016 IEEE Winter Conference on Applications of Computer Vision，pp.1-10 (2016).
[12] ELAN，https://archive.mpi.nl/tla/elan (最終アクセス日：2020/10/25).
[13] 高橋 誠, 会議の進め方, 日本経済新聞出版社 (1987).