表情に応じた字幕映像を出力する仮想カメラ EmoScribe Cameraのオンライン会議への応用

荒巻 美南海[1,a)]　Hautasaari　Ari[2,b)]　苗村 健[2,c)]

概要：
オンライン会議中にウェブカメラとマイクをオンにできない場合，代替手段としてチャットが用いられることが多い．しかし，チャットにはタイムプレッシャーが大きかったり，感情のニュアンスが伝わりにくいという課題がある．この課題に対処し，オンライン会議をより活発化するため，音声から自動生成した字幕映像を出力する仮想カメラシステムである EmoScribe Camera を提案した．仮想カメラとして実装することで，通常のカメラ入力と同様にさまざまなアプリケーションに字幕映像を入力することが可能となる．さらに，自動字幕の表示に使用するフォントは，自動的に推定された発話者の表情に応じて変化する．提案手法とテキストチャットを Zoom 上のグループワークで用いた際の比較を行ったところ，提案手法を用いた際にユーザのワークロードが軽減し，発話量の増加が見られた．

1. はじめに

オンライン会議で最大限対面での自然なコミュニケーションに近づけ，議論を活発に進めるためには，カメラとマイクの両方をオンにすることが理想である．しかし，身だしなみを整えるのが間に合わなかったり，生活音まで配信することがためらわれたりといった理由でカメラやマイクをオフにして参加するという状況も多々ある．そのような場合，キーボードで会議ツールのチャットにテキスト入力したり，リアクションボタンで反応したりという対応が考えられるが，このような方法ではコミュニケーションのたびに操作が必要となる上，文章や絵文字だけでは細かい感情のニュアンスが伝わりにくい．そのような問題に対処するため，ビデオ通話とテキストチャットの中間に位置するコミュニケーション手段として EmoScribe Camera という仮想カメラシステムを提案した [1]．本システムは，発話内容を自動字幕の画像としてリアルタイムで生成し，通常のカメラ入力と同様にさまざまなアプリケーションに字幕映像を入力することができる．さらに，自動字幕の見た目は発話者の感情に合わせて変化するため，通常の自動字幕では伝えることのできない感情のニュアンスを表現できる．この手法を用いることによって，オンラインコミュニケーションにおいて直接顔の映像や音声を伝えずともより活発な議論を行うことができるようになる．本研究では，実際のオンライン会議を想定した場面で提案手法の評価実験を行うことにより，その有効性を検証した．

1 東京大学大学院学際情報学府
2 東京大学大学院情報学環
a) aramaki@nae-lab.org
b) ari@nae-lab.org
c) naemura@nae-lab.org

2. 関連研究

先行事例として，DNP が開発した感情表現字幕システム [2] がある．このシステムは提案手法と同様に発話者の感情に合わせて自動字幕のフォントや大きさを変化させているが，異なる点もいくつかある．まず，感情表現字幕システムはテレビ番組向けのもので，仮想カメラとしての出力は行なっていないため，Web 会議ツールで利用したい場合には画面共有をする必要がある．したがって使用に手間がかかる上，ツールによっては複数人で同時に画面共有ができない場合もあるため，オンラインコミュニケーションで使用するには不便である．提案手法は仮想カメラであるため，**図 1** のように Zoom や Microsoft Teams などのビデオ会議ツールで使えるのはもちろん，カメラ入力を扱う様々なソフトウェアで活用することができるように設計した．加えて，感情表現字幕システムはカメラ映像に重ねて字幕を表示しているが，提案手法では字幕の画像のみを表示することもできる．したがって，Web カメラの映像を対話相手に見せたくない時でもこの手法を使える．さらに，複数の会議参加者が EmoScribe Camera を使うことで，画面に並んだ字幕映像の間でテキストによる対話を行うこともできる．

図 1 提案手法をビデオ会議ツールで使っている様子 (左: Zoom，右: Microsoft Teams)

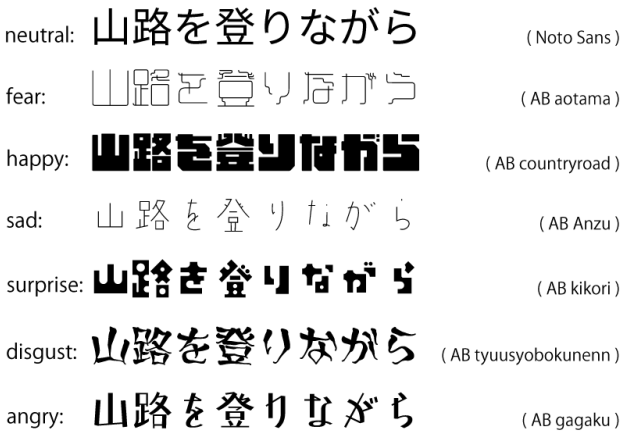

図 2 Chujo らの研究によるデータセット上で最も強く各感情を表現しているとされたフォント

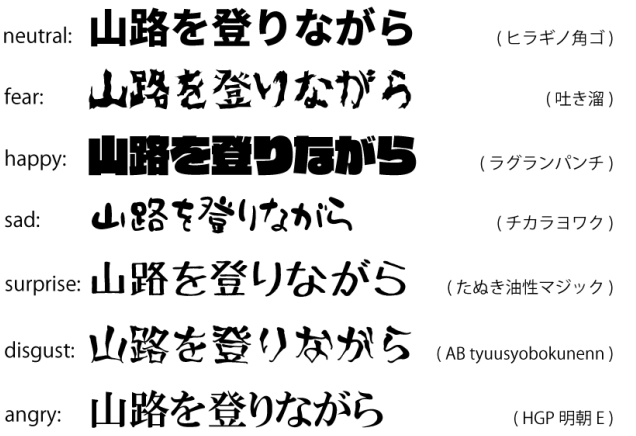

図 3 EmoScribe Camera に使用しているフォント

Chujo らは，日本語のフォントが人に与える印象に関して調査を行った [3]．さらに，その結果に基づき，フォントとそれが与える感情を対応付けたデータセットを作成した．このデータセット上では，フォントを見た時に与える印象が neutral, angry, disgust, fear, happy, sad, surprise の計 7 種類の感情のどれに最も近く感じられたがどの程度いたかが割合で示されている．これらの感情が選ばれているのは，Ekman による普遍的感情の研究 [4] が理由である．ただし，感情を強く表現できるフォントは**図 2** のように可読性に乏しい傾向があるため，提案手法を用いて行われる字幕によるコミュニケーションには適さない．そこで，現時点の提案手法では，**図 3** のような可読性の問題が少ないと思われるフリーフォントを用いている．将来的には，このデータセットに含まれるフォントに対して可読性の評価を行い，一定以上の可読性を持つものを採用する予定である．

3. 提案手法

EmoScribe Camera は, Swift を用いて実装された macOS 上で動作する仮想カメラアプリケーションである [1].

3.1 音声認識の手法

ユーザが発話すると，その内容が自動で文字起こしされる．この際の speech to text には Speech フレームワーク [5] を用いた．

3.2 感情認識の手法

本アプリケーションでは，ユーザが選択したカメラの情報をもとに表情認識を行うことによってユーザの感情の推定を行なっている．ある発話をした時点での感情を推定する方法としては，表情のほかに文章の内容や声色の分析など，様々なものが考えられる．しかし，これらの方法では文頭から感情推定を行うということができないため，提案手法に求められるリアルタイム性を達成することが難しい．加えて，文章の内容から推定を行う方法では，発話者の実際の感情と発話内容が乖離している場合にその情報を取りこぼしてしまう可能性がある．したがって，提案手法では表情から感情を推定する方法を取っている．

3.3 感情認識の実装

ユーザの顔認識と表情認識のため，CoreML[6] という Swift 上で動作する機械学習を扱うためのライブラリを用

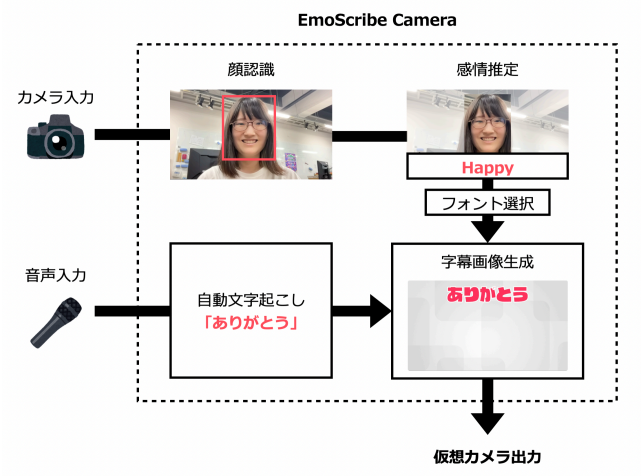

図 4 システムの概略を示すフローチャート

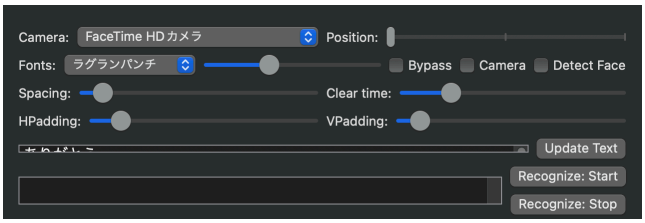

図 5 提案手法のコントロールパネル

いた．CoreML にデフォルトで含まれているモデル [7] を用いて顔部分の画像を切り出し，表情認識にかけるという手順で感情の推定を行なった．感情推定には CNNEmotion[8] という CoreML 用のモデルを用いた．推定した感情は Ekman の研究 [4] に基づいた 7 つの基本感情 (neutral, sad, happy, angry, fear, disgust, surprise) に分類され，そこで最も強く検知された感情に基づいて字幕に用いられるフォントが選定される．

3.4 処理の流れ

ユーザが発話すると，選ばれたフォントでリアルタイムに字幕の画像が生成される．字幕の背景にはカメラ映像をそのまま使うこともできるし，ユーザが選んだ背景画像を用いることもできる．こうして生成された映像は，仮想カメラとして出力され，様々なツールやデバイス上でウェブカメラによる入力と同様に扱われることができる．**図 4** のフローチャートは，以上の流れをまとめたものである．

また，ユーザはコントロールパネルを用いることによって手動でも自動字幕の表示形式を変更することができる．コントロールパネルのレイアウトは**図 5** の通りである．

4. オンライン会議への応用

提案手法の使用場面としては，カメラがオン・オフの場合，マイクがオン・オフの場合，合わせて 4 通りの状況が考えられる．カメラがオフの場合はあらかじめ設定された背景画像の上に字幕が表示され，オンの場合は発話者の映像に重ねて字幕が表示される．マイクがオフの場合は字幕のみで発話を解釈することになり，オンの場合は音声を聞き取る補助として字幕を用いることになる．

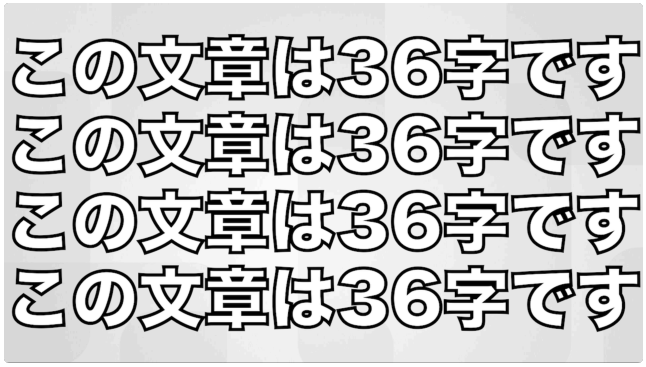

図 6 提案手法の画面いっぱいに 200pt の文字を表示した様子

本研究では，これらのうちカメラもマイクもオフである状況についての評価実験を行った．今回のデモでは，この実験と同様にカメラ・マイクをオフにしながら，より少ない人数でのオンライン会議において実際に提案手法を体験できる．

5. 評価実験

本実験では従来手法として Zoom のチャット機能を用いることとした．従来手法と提案手法のユーザビリティの違いや，提案手法を用いてビデオ会議をすることによって議論の活発さがどのように変化するかを確かめるための比較実験を行なった．

5.1 実験概要

カメラ・マイクオフで Zoom 上でグループワークを行う場面を想定し，それぞれ従来手法 (テキストチャット) と提案手法を用いて参加者間でグループワークを行うことによって参加者内比較を行った．

グループワークのテーマは「東京大学生活で一番驚いたこと」「東京大学生活で一番腹が立ったこと」の 2 種類で，各参加者がテーマに沿って自分のエピソードを話した後，その中で一番を決めるという流れで行った．従来手法と提案手法，2 つの手法を使う順序と共にテーマ 2 種類を入れ替え，計 4 条件で実施した．

参加者の総数は 24 名 (女性 12 名，男性 12 名) であった．用意した話題についてどの参加者も最低限発言することができるように，参加者は全て東京大学に在学中であることを条件に募集した．参加者を 3 名ずつの 4 グループに分け，順序効果を考慮し条件を入れ替えた上でそれぞれで実験を行なった．このとき，グループ内の性別が混在していると少数派の性別の参加者が萎縮してしまう可能性があるため，各グループは全て男性のみ，女性のみのいずれかで構成されるように調整した．

提案手法には字幕の表示に関わるさまざまなパラメータが存在するが，その中でも特に重要な字幕のフォントサイズや字幕の表示時間に関しては事前に実験で適切な値を決定した [1]．具体的には，フォントサイズは 200pt, 字幕の表示時間は 4 秒として実験を行った．このときの字幕の見た目は**図 6** のようであった．

参加者が Zoom 上で話す際は，直接お互いの声が聞こえないように仕切り越しに離れて着席し，ヘッドフォンを着用した．

5.2 実験の流れ

まず，各参加者に提案手法がインストールされた MacBook Pro(macOS Ventura) を貸与した．最初にアイスブレイクとグループワークの練習を兼ね，参加者は Zoom 上で順に自己紹介した後，その内容について相互に質疑応答を行なった．なお，参加者 3 名の Zoom 上の名前はそれぞれ「参加者 1」「参加者 2」「参加者 3」であった．

次に，参加者に提案手法の使用方法を説明した上で，自由に発言させて文字起こしを体験させたり，表情を変えて発言をさせることで提案手法の練習を行なった．また，従来手法を問題なく扱えるようにするため，タイピングの練習を行わせた．

次に休憩を挟み，グループワークを 10 分間実施した後にアンケートへの回答を行わせた．その後，再び休憩を挟んだ上で，使用するツール・話題を入れ替えて同様にグループワーク・アンケートへの回答を行わせた．

最後に，参加者それぞれにインタビューを行った．

5.3 評価方法

グループワーク中の議論の活発さを評価するため，参加者ごとの発話量を比較した．ここで言う発話量の定義は，従来手法を使った際は送信された文章をひらがなに直した際の文字数，提案手法を使った際は発話の自動文字起こしをひらがなに直した際の文字数とした．

また, グループワーク後のアンケート上では, SUS(System Usability Scale) を用いてユーザビリティの評価を，NASA-TLX を用いてワークロードの評価を行った．SUS に関しては，佐藤らによる日本語訳 [9] を使用した．NASA-TLX に関しては，**表 1** に示した通りの三宅らによる和訳 [10] を使用した．

加えて，「今回のグループワークで使ったツールについて，良い点と悪い点を挙げながら感想を述べてください」という設問に自由記述で答えさせた．

実験の最後では，以下のような質問を基本とした半構造化インタビューを行った．

- アンケート上の回答について，なぜそのように回答したか
- 従来手法と提案手法ではどちらのほうが議論がスムーズに進んだか
- 従来手法と提案手法ではどちらのほうが議論が盛り上がったか

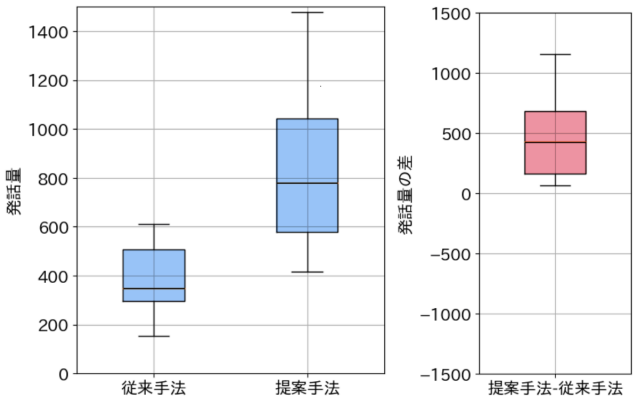

図 7 左: 従来手法と提案手法の発話量, 右: 提案手法の発話量から従来手法の発話量を引いたもの

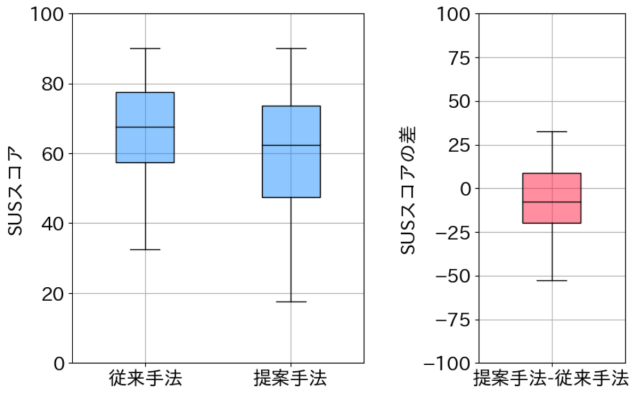

図 8 左: 従来手法と提案手法の SUS スコア, 右: 提案手法の SUS スコアから従来手法の SUS スコアを引いたもの

6. 結果

発話量を**図 7** に示した．これを対応のある t 検定によって調べたところ，提案手法の方が有意に発話量が多かった ($p < .05$)．

SUS スコアを**図 8** に示した．このスコアを対応のある t 検定によって調べたところ，統計的な有意差があるとはいえなかった ($p > .05$)．

NASA-TLX スコアを**図 9** に示した．このスコアを対応のある t 検定によって調べたところ，提案手法の方が有意に低かった ($p < .05$)．言い換えると，提案手法の方がワークロードが小さかった．

ツールに対する感想を問う自由記述の設問では，24 名中 8 名が「感情のニュアンスが伝わることによって，比較的コミュニケーションが取りやすかった」という旨の回答を行った．

提案手法の方が有意にワークロードが低かった理由について半構造化インタビューでより深掘りした結果，従来手法に対して提案手法はタイピング等の身体的負荷が少ない

表 1 NASA-TLX の質問内容を和訳したもの [10]

	質問内容
知的・ 知覚的要求	どの程度の知的・知覚的活動（考える，決める，計算する，記憶する，見るなど）を必要としましたか．課題は やさしかったですか難しかったですか，単純でしたか，複雑でしたか，正確さが求められましたか大雑把でよかったですか
身体的要求	どの程度の身体的活動（押す，引く，回す，制御する，動き回るなど）を必要としましたか，作業はラクでしたかキツかったですか，ゆっくりできましたかキビキビやらなければなりませんでしたか，休み休みできましたか働きづめでしたか
タイム プレッシャー	仕事のペースや課題が発生する頻度のために感じる時間的切迫感はどの程度でしたか，ペースはゆっくりとして余裕のあるものでしたか，それとも早くて余裕のないものでしたか
作業成績	作業指示者 (またはあなた自身) によって設定された課題の目標をどの程度達成できたと思いますか．目標の達成に関して自分が製作しようと思ったものができた度合いにどの程度満足していますか
努力	製作物を完成させたり製作を進めたりするために，精神的・身体的にどの程度いっしょうけんめいに作業しなければ なりませんでしたか
ストレス	作業中に，不安感，落胆，いらいら，ストレス，悩みをどの程度感じましたか．あるいは逆に，安心感，満足感，充足感， 楽しさ，リラックスをどの程度感じましたか

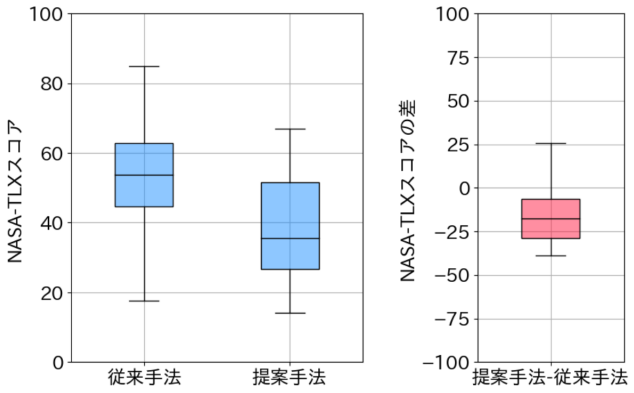

図 9 左: 従来手法と提案手法の NASA-TLX スコア, 右: 提案手法の NASA-TLX スコアから従来手法の NASA-TLX スコアを引いたもの

ことや，提案手法ではコミュニケーションがよりリアルタイムに行われるため相手が発話中かどうかなどの様子がわかりやすく，タイムプレッシャーをはじめとした精神的負荷が少なかったことが挙げられた．

また，インタビューの結果，参加者 24 名中 16 名が提案手法を用いた際の方がスムーズに議論が進められたと回答した．参加者 24 名中 21 名が提案手法を用いた際の方が議論が盛り上がったと回答した．

加えて，提案手法を使っての感想をインタビューで質問した結果，24 名中 8 名が，表情と同期してフォントが変わることでより話しやすくなったと回答した．具体的な意見としては，自分の感情や相手の感情のニュアンスが相互に伝わることで安心感が生まれた，ぱっと画面を見てどういう雰囲気の会話をしているのかわかりやすい，といったものがあった．

7. 考察

インタビューの結果，リアルタイム性という意味で対面でのコミュニケーションにより近い提案手法の方が従来手法より身体的・精神的負荷が少なく，その結果としてより積極的なコミュニケーションが誘発され，参加者の発話量が増え，参加者の主観的としてもより円滑かつ活発なグループワークを行うことができたと考えられる．

さらに，表情に応じてフォントが変化することについてポジティブな意見が多くあったことから，単なる文字起こしによる字幕ではなく感情を反映した字幕を用いることの意義は十分にあると考えられる．

8. おわりに

本研究では，ビデオ会議ツール上でカメラとマイクをオフにした状態でも円滑なコミュニケーションが行えるようにするため，参加者の感情に合わせて変化する自動字幕を仮想カメラとして出力できるアプリケーション，EmoScribe Camera を提案した．

今回の評価実験では，感情の伝達が議論の盛り上がりにどの程度寄与するかの定量的な比較はしていない．そこで今後は，通常の提案手法と，推定した感情による字幕の変調を行わないよう設定を変更した提案手法とで比較実験を行うことによって，その点に関する評価も行う．

また，提案手法は，カメラ映像に重ねて字幕を表示することもできる．今後はそのような使い方をした場合についても実験を行い，検討を行っていく予定である．

さらに，現在 5-10 名程度がオンラインで参加する定期的なミーティングで提案手法を使用してもらい，フィードバックを収集するユーザスタディを行っている．今後も引き続きこれを継続し，実用性の高いシステムを目指していく．

謝辞 本研究は，株式会社メルカリ R4D とインクルーシブ工学連携研究機構との共同研究である価値交換工学の成果の一部である．

提案手法のアプリケーション開発は，服部智氏の協力のもとで行われた．

参考文献

[1] 荒巻美南海，ハウタサーリアリ，苗村　健：EmoScribe Camera: 音声を感情豊かな字幕にリアルタイム変換する仮想カメラ，HCG シンポジウム，pp. A–6–5 (2023).
[2] DNP: 映像を AI で解析し臨場感を伝える「感情表現字幕システム」を開発しました, https://shueitai.dnp.co.jp/news/detail/10161885_3733.html. 参照 2023/7/9.
[3] Chujo, R., Suzuki, A. and Hautasaari, A.: Exploring the Effects of Japanese Font Designs on Impression Formation and Decision-Making in Text-Based Communication (2023). arXiv:2309.06743 [cs].
[4] Ekman, P.: Are There Basic Emotions?, *Psychological Review*, Vol. 99, No. 3, pp. 550–553 (online), DOI: 10.1037/0033-295x.99.3.550 (1992).
[5] Apple: Speech, https://developer.apple.com/documentation/speech/. 参照 2023/7/10.
[6] Apple: Core ML, https://developer.apple.com/jp/machine-learning/core-ml/. 参照 2023/7/10.
[7] Apple: VNDetectFaceRectanglesRequest, https://developer.apple.com/documentation/vision/vndetectfacerectanglesrequest. 参照 2023/7/10.
[8] Levi, G. and Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 503–510 (2015).
[9] Sato, K., Mitomi, N., Kon, K. and Haruna, H.: 義肢装具領域における System Usability Scale (SUS) の信頼性の検討，*The Journal of the Japanese Academy of Prosthetists and Orthotists*, Vol. 30, No. 1, pp. 32–37 (2022).
[10] 三宅晋司，神代雅晴：メンタルワークロードの主観的評価法 NASA-TLX と SWAT の紹介および簡便法の提案，人間工学，Vol. 29, No. 6, pp. 399–408 (1993).