プロジェクト名：デバイス１つですべての会話を記録し可視化するシステムの開発
申請者名：藤巻晴葵

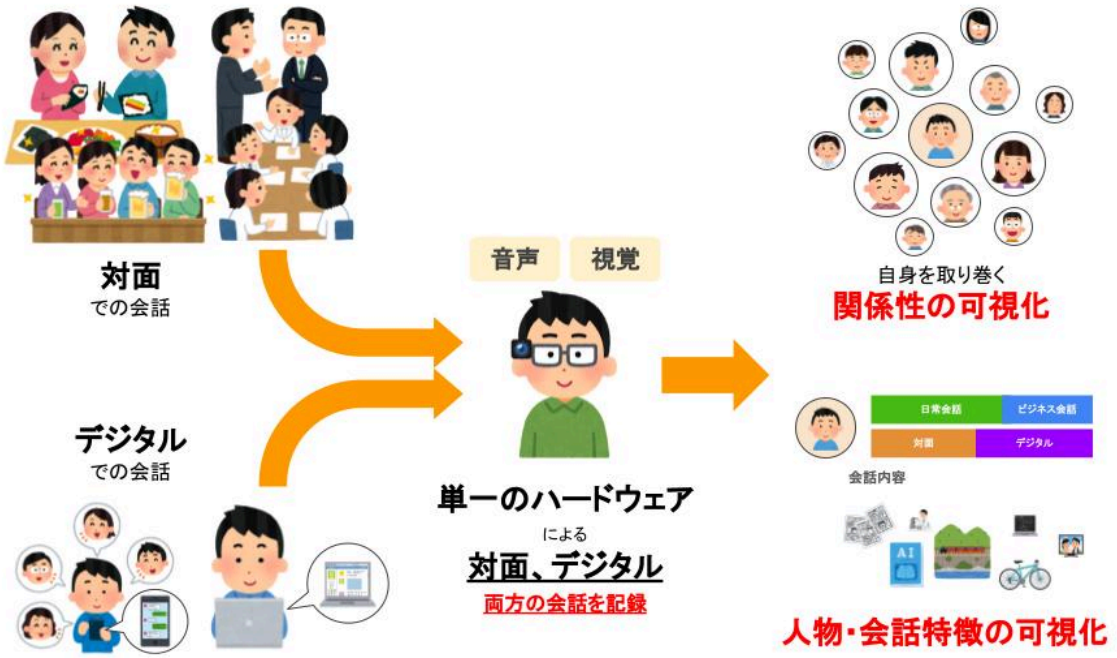

図1 提案するシステムで実現すること

概要

自身を取り巻く人たちとの関係性や会話特徴を可視化し、自身の社会活動を高度に理解できるシステムを実現する。ユーザーは一人称カメラとマイクを搭載した単一のハードウェアを装着するだけで、あらゆる場面での会話（対面および、PCやスマホを利用したデジタルの会話）を記録することができる。加えて会話内容を分析し、ユーザーの社会活動の特徴やユーザーを取り巻く人物特徴を詳細にレポートする。私の目標は、自身の社会活動を理解し、改善を行えるシステムの構築を目指すことである。

1. 何を作るのか

1.1. 背景

現代社会において、人々の健康的な生活を維持・向上させることは極めて重要な課題となっている。近年、ウェアラブルデバイスの普及により、個人の身体的健康状態を常時モニタリングし、可視化することが可能となった。Apple WatchやFitbitに代表されるヘルスケアデバイスは、活動量や心拍数、睡眠の質など、様々な身体指標をトラッキングし、ユーザーの健康管理をサポートしている。

しかし、人間の健康は身体的側面だけでなく、社会的側面も含めて捉える必要がある。私たちは社会的な存在であり、家族や友人、同僚など、他者との関わりの中で生きている。対面でのコミュニケーションはもちろん、デジタル技術の発展に伴い、オンライン上でのコミュニケーションも日常的なものとなった。こうした社会的つながりは、個人の精神的健康や幸福感に直結する重要な要素である。

一方で、現代人の社会的つながりは複雑化・多様化しており、自身の社会的健康状態を把握することは容易ではない。比較的内向的な私ですら、対面では大学のサークル、インターン先の上司、研究室の先輩など、様々な場面で日常的に会話を行っている。加えて、デジタル上でもプライベートの会話ではLINE、Instagram、Facebook、Xを、仕事上の会話ではSlack、Teams、メールを、趣味関係の会話ではDiscordや専門サービスを用いて、多様な人物と会話している。このように、仕事とプライベート、対面とデジタルなど、様々な場面でのコミュニケーションが絡み合う中で、バランスの取れた良好な人間関係を構築・維持するためには、一部分の会話だけでなく、あらゆる場面の会話を網羅的に分析し、客観的な指標に基づいた評価と改善が不可欠である。

そこで私は、ウェアラブルデバイスとAI技術を活用し、個人の社会的健康状態を可視化するシステムの開発を提案する。このシステムは、業務上の会話など一場面のみの会話だけでなく、あらゆる場面での会話を記録し、分析することを目的としており、一人称視点カメラとマイクを搭載したウェアラブルデバイスを用いて、日常生活における対面とデジタルの両方の会話を記録し、その特徴を分析する（図1）。**本システムは一つのウェアラブルデバイスで対面とデジタルの会話を記録するため、ユーザーは自身が利用する各デバイスに対してソフトウェアをインストールする必要がなく、特定のプラットフォームに制限されない会話の記録ができる。**そして、記録した会話データを元に会話の量や頻度だけでなく、会話内容の特徴、相手の特徴や関係性などの様々な指標を可視化することで、自身の社会との関わり方やコミュニケーションの特徴、課題を明らかにする。このシステムでは、NEC 協働支援サービス[1]のような主要なチャットツールと連携し、コミュニケーションを可視化するto B向けサービスとは異なり、日常のあらゆる場面での会話、あらゆるコミュニケーションサービスによる会話を網羅的に記録し、ユーザーの多様な人間関係全てを分析の対象とする。

本システムにより、ユーザーは自身の社会的つながりを客観的に把握し、適切なコミュニケーションを図ることが可能となる。分析結果を時系列でモニタリングすることで、環境の変化に伴う人間関係の変化も追跡できる。これは、孤独の解消や生きがいの創出など、個人のウェルビーイング向上に直結する取り組みである。

また、本システムで得られる知見は、社会全体の健康増進にも寄与すると期待される。個人の社会的健康状態を集約・分析することで、コミュニティレベルでの課題や傾向を明らかにできる。以上のように、本プロジェクトは、ウェアラブルデバイスとAI技術を活用し、個人の社会的健康状態を可視化することで、ウェルビーイングな社会の実現に寄与することを目的とする。

1.2. 提案するもの

本システムは、ユーザーの対面、デジタル上でのやり取りを把握、記録し、自身を取り巻く人との関係性を可視化する。提案するシステムの大きな機能は以下の2つである。

- **ウェアラブルデバイスによるユーザーの対面、デジタル上の会話の記録**
- **会話情報から特徴を抽出し、ユーザーを取り巻く人たちとの関係性を可視化**

上記の機能について順に説明する。

1.2.1. ウェアラブルデバイスによるユーザーの対面、デジタル上の会話の記録

本システムの実現にあたり、ラズパイなどのマイコンを使用したウェアラブルデバイスを開発する。このデバイスはカメラとマイクを搭載し、ユーザーは対面とデジタルの会話両方をデバイスの装着のみで簡単に記録を開始できる。そして、記録されたデータはサーバーにアップロードされ、データを解析し、保存する。

デジタルの会話を記録するためには、デバイスで画面の文字が認識できる程度の解像度で記録する必要がある。私が胸元にカメラを装着して検証したところ、Full HDの解像度では文字を認識するのは難しく、4Kレベルの解像度では文字を認識できることを確認した。しかし、4Kの解像度での長時間の記録は膨大な容量を必要とし、加えて、常時記録した視覚、音声データを全てサーバで処理を行うのは現実的ではない。そこで、図2に示すようにデバイス側で必要な部分だけを記録する仕組みを適用し、この問題を解決する。検出する際には解像度の低いカメラを用いて行い、記録の際には高解像度カメラを用いる。これにより、関連する情報のみを効率的に記録することが可能となる。

デバイスに記録されたデータをどのように処理し、保存するのか対面での会話、デジタル上での会話の二つのケースに基づいて処理方法を説明する。

対面での会話の記録

デバイスから記録された対面での会話が行われている可能性のある視覚情報（動画）と音声情報を用いて、以下の順で処理を行い記録する。

1. **視覚情報の処理（A1）：カメラから得られる画像から顔を検出し、識別する。**
2. **音声情報の処理（A2）：マイクから得られる音声を話者ごとに分離し、識別する。さらに、音声を文字情報に変換する。**
3. **顔画像と音声の紐付け（A3）：検出された顔画像と音声を紐付けし、同一人物の会話情報を蓄積する。**

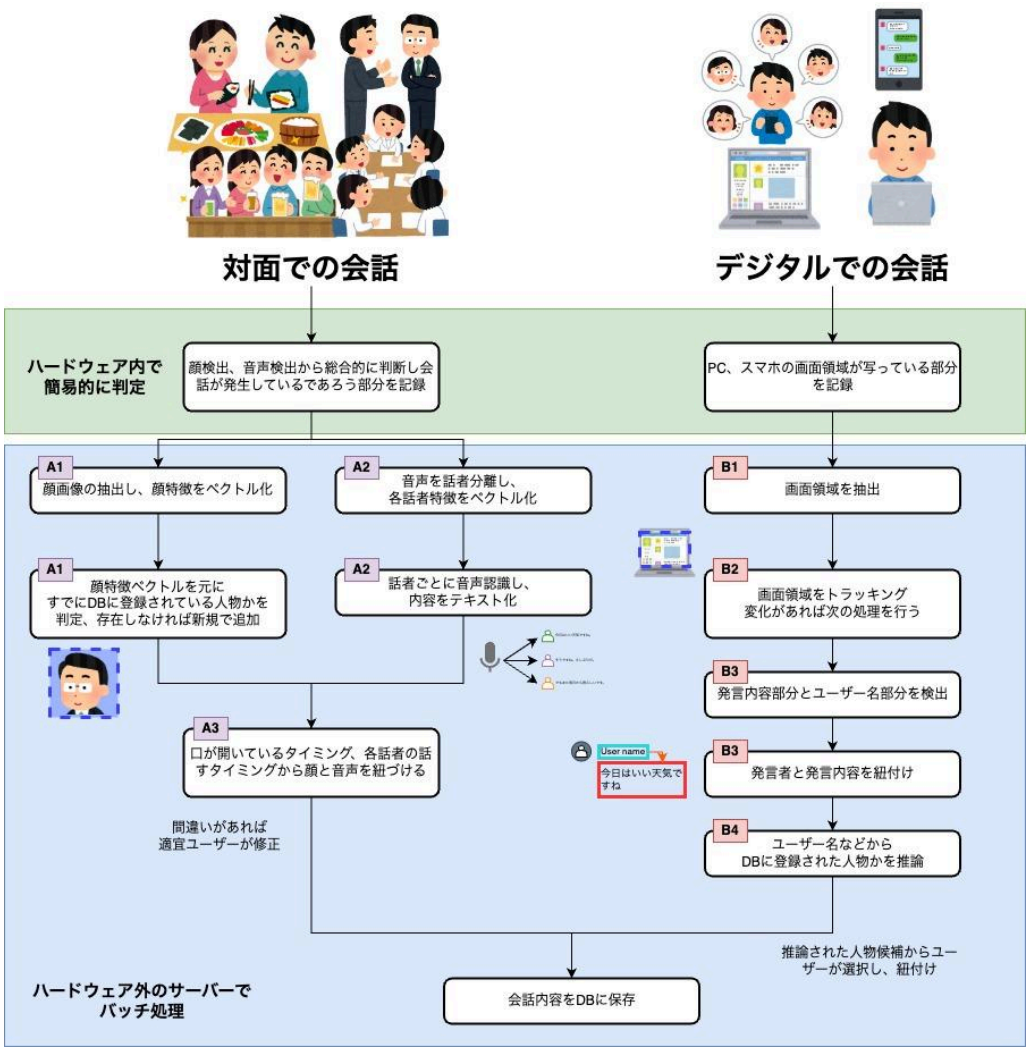

図2 対面、デジタルそれぞれの会話の記録フロー

まず、**1. 視覚情報の処理**についてだが、これは顔を元に今まで記録された人物と紐づけるために行う。記録された動画をYOLO v8[2]などの軽量な顔検出AIモデルを利用し顔を検出する。検出した顔画像はArcFace[3]などの顔識別モデルを用いることで特徴ベクトルに変換する。これにより、今まで記録された人物の顔特徴ベクトルと照合し、以前記録された人物であれば、その人物に紐付けて記録し、そうでなければ新たな人物として記録する。

次に、**2. 音声情報の処理**について、これは自身を含む複数人物の会話において、人物ごとに発言を記録するために行う。記録された音声をPyannote-Audio[4]という音声処理ツールと学習された話者分離モデルを利用し、音声を話者ごとに分離する。分離された音声はWhisper[5]やReazonSpeech v2[6]といった音声認識モデルを利用し、音声認識を行う。また、Jungらの音声識別の研究[7]などから話者分離された音声を識別し、人物と紐づけることを検討しており、これにより正確な会話の記録が行えると考えている。

最後に、**3. 顔画像と音声の紐付け**についてだが、会話内容を人物に紐づいて行うために行う。動画に映る顔の口の動きを検出し、そのタイミングを音声情報と照合し、どの人物が何を喋っているのかを紐づける。

デジタル上での会話の記録

デバイスから記録されたPC、スマホの画面が映っている可能性のある視覚情報（動画）を用いてデジタル上での会話の記録を行う。ここでのデジタル上での会話とは、任意のコミュニケーションプラットフォームにおけるテキストを用いた会話のことを指し、ビデオ通話などのオンラインの会話については、前述の対面での会話の記録と同様の処理によって記録を行う。デジタル上での会話は以下の順で処理を行い記録する（図3）。

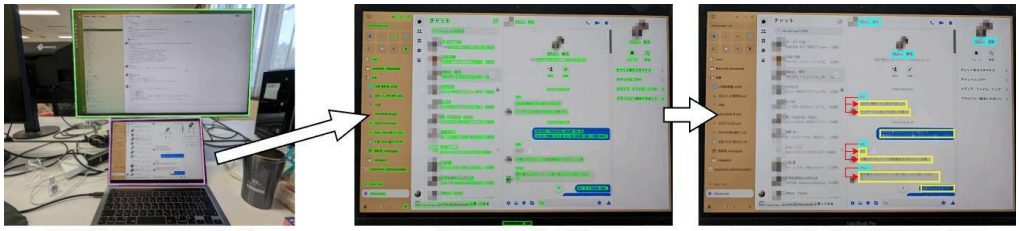

図3 会話情報抽出プロセス

1. **PC・スマホの画面領域の検出（B1）**
2. **検出した画面領域のトラッキング（B2）**
3. **抽出された画面領域の解析（B3）**

まず、**1. PC・スマホの画面領域の検出**についてだが、これは動画全体ではなく、画面領域部分に分離し、次の処理に渡すことで高精度化と処理コストの低減を目的としている。この検出については、検証を行っている。まず、すでに公開されているデータセットから学習することで検出モデルを学習できないかを考えたが、端末の画面の領域のみセグメンテーションされたデータセットはあまり存在しなかったため、今回はSegment Anything[8]という汎用セグメンテーションモデルを利用し画面部分のセグメンテーションと処理を行って4点の座標からなる画面領域検出が行えるかの検証を行った。検証の結果は参考資料に動画を添付している。この検証から分かったことは、Segment Anythingでの処理では手法の都合上、安定した検出を行うのが難しく、処理コストについても理想的ではないことがわかった。このことから、未踏期間では、画面領域検出に向けたデータセットの作成から行い、そのデータセットと物体検出モデルのYOLO v8を用いて学習させ、実現する。

次に、**2. 検出した画面領域のトラッキング**についてだが、これは画面をトラッキングすることで、画面内の変化を追跡できるようにする。これにより、適切な画面処理を行うことができ、重複した無駄な処理を減らし、処理コストを下げ、より高度な会話記録を可能にする。この処理について、一般的なトラッキングアルゴリズムで検証を行った。結果としてはあまり期待した精度のトラッキングを行うことができなかった。これは、前述の画面検出が不安定であることと、アルゴリズムの仕様上、画面検出の結果で逐次補正できないためだと考えられる。そのため、検出された物体のトラッキングについては、Zhangらによって提案されているByteTrack[9]やAharonらによって提案されているBoT-SORT[10]といった画面検出の結果を利用してトラッキングする手法を利用し、PC・スマホの画面領域の検出情報を元にトラッキングを行う。

次に、**3. 抽出された画面領域の解析**についてだが、これは複数の文字が配列された画面画像から人物の発言のみを抽出し、だれが何を発言したかを解析する。具体的には、まず画面画像をTesseract[11]やGCP Cloud Vision API[12]などを用いてOCRを行い、その後、発言者と発言内容の領域をラベリングする。ラベリングの手法として2つ考えられる。まずはYOLO v8、DETR[13]といった視覚情報のみを活用し、発言者、発言内容を示しているであろう領域を検出する方法が考えられる。二つ目の手法としては、Huangらによって提案されている画像とテキスト、レイアウト情報を用いるLayoutLMv3[14]のようなマルチモーダル言語モデルを利用した手法が考えられる。人の名前、ユーザー名というのはテキスト情報だけでもある程度判断することができると考えられる。そのため、2つ目の手法というのは画像とテキスト情報を入力とするので、前者よりも精度の高いラベリングが行えることが期待できる。しかし、公開されているLayoutLMv3モデルは文書画像で事前学習されており、加えて日本語に対応していない。そのため、初めは1つ目の手法によるラベリングを行い、その後、端末画面特化の言語モデルの事前学習、またはマルチモーダルなラベリングモデルの開発を検討することとする。検出した会話情報は過去の情報からどの人物の発言かを予測し、ユーザーは候補から選択することで、人物ごとに会話を記録する。

デジタル上の会話の記録について、未踏期間では以上3点を開発する。未踏期間終了後には、より正確にやりとりの流れを保持したデジタル上での会話の記録を実現したいと考えている。

1.2.2. 会話情報から特徴を抽出し、ユーザーを取り巻く人たちとの関係性を可視化

本システムは、ユーザーの会話内容を深く分析することで、ユーザーの社会的な関わりを詳細に理解

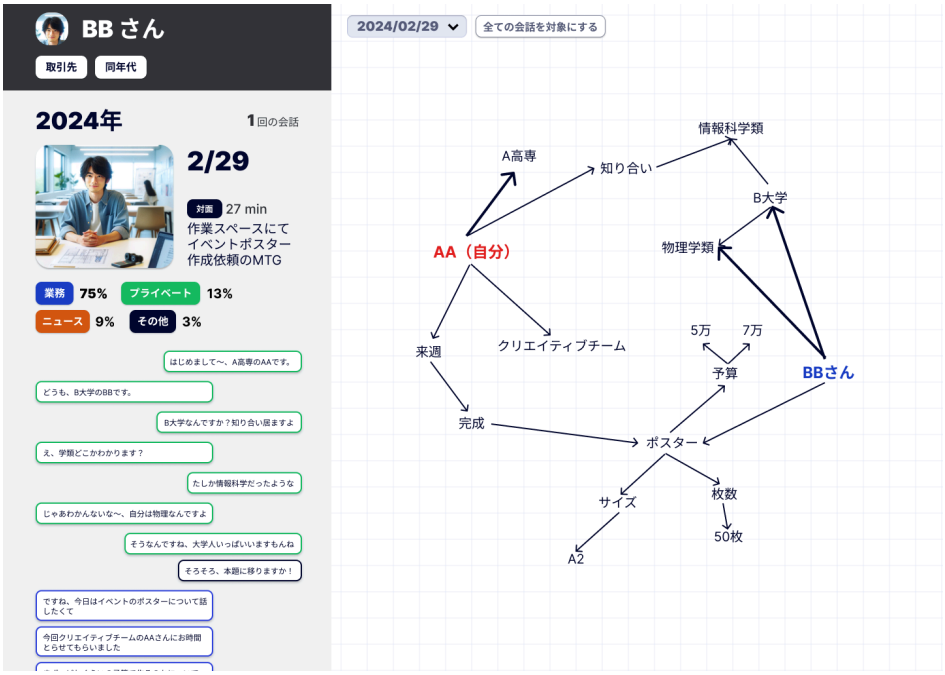

図4 会話、人物特徴可視化イメージ

することを目指している。具体的には、以下の3つのアプローチによってユーザーを取り巻く人々との関係性、会話特徴を可視化する。図4に示すような可視化を想定している。

1. **固定指標による会話、人物、関係性の分析**
2. **会話、人物についての動的特徴の抽出**
3. **視覚的な状況特徴の記録**

まず、**1. 固定指標による会話、人物、関係性の分析**について、これは会話情報を会話のカテゴリーなどの予め定義された項目で分類することを指している。具体的には、会話内容の分類（10項目）、会話相手との関係性分類（15項目）などが挙げられる。この分類では、Distil-BERT[15]などの比較的軽量な言語モデルを利用し、ファインチューニングを行うことでそれぞれの会話分類と関係性分類を行う。

次に、**2. 会話、人物についての動的特徴の抽出**について、これは会話の内容から個々の人物に関する具体的な情報や関係性を抽出することを指している。固定指標の分類では捉えきれない、より詳細で個別的な情報を見出すことが目的である。例えば、会話の中で「私は、イタリア料理が好きで、毎年イタリアに旅行に行っている」といった発言があった場合、この情報からその人物の好みや趣味、行動パターンなどの特徴を抽出できる。この特徴を可視化するにあたり、知識グラフとして可視化したいと考えている。Oshin Agarwalらは、T5を用いたSeq2Seqの知識グラフの作成手法[16]、Melnykらは、言語モデル等を用いたノード生成、エッジ生成の2段階式によるテキストから知識グラフを構築する手法[17]を提案しており、これらを参考に会話文に対して知識グラフを構築する。また、実装するにあたって、独自の手法を含めたいくつかのパターンで比較検証を行い、最適な手法を選択する。

最後に、**3. 視覚的な状況特徴の記録**について、これは会話が行われた状況を画像と文章で記録することを指している。例えば、「同級生と居酒屋での会話」といった状況の情報も記録したい。まず、会話が行われた画像をCLIP[18]を用いて特徴ベクトルに変換し、記録することで、テキストから該当する状況の画像、つまり会話を検索することが可能になる。また、Mokadyらの提案したClipCap[19]と言われる画像キャプションの研究から画像からキャプションを生成し、会話特徴の知識グラフの構築の際に用いることができる。

これらの機能を実装するにあたり、各深層学習モデルを学習させるためのデータセットは本システムで記録した会話データからデータセットを作成することに加えて、国立国語研究所が有償で公開している日本語の日常会話コーパス[20]などの会話コーパスを用いて開発・検証を行うことを考えている。

2. どんな出し方を考えているか

私は、このシステムを最終的にはスマートフォンのアプリケーションで利用できる形として公開した

表1 既存サービス、システムとの比較

	提案システム	NEC 協働支援サービス	Okunoらの開発したシステム	内橋らの開発したシステム
対象	個人（to C）	企業（to B）	個人（to C）	個人（to C）
対面のコミュニケーション記録	◎	×	△	○
デジタル上のコミュニケーション記録	◎	○ （Slack、Teams）	×	×
会話の分析	◎	○	×	○

いと考えている。このシステムは、ハードウェア1つで対面とデジタル上のやりとりを記録することによってユーザーの記録するための煩雑さをなくしている。そのため、システムもPCのみならず、スマートフォンで利用できるような形にし、ユーザーの利用ハードルを低くしたい。また、私は今後ヘッドセットなどを利用したMRシステムの普及が加速すると考えている。このシステムはカメラとマイクによって会話を記録するため一般的なヘッドセットに移植可能であり、Oculus Quest 3やXREAL Air 2 Ultra, Apple Vision Proなどの利用を想定した、専用のウェアラブルデバイスを利用せずともコミュニケーションを記録できるシステムとしての提供も行いたいと考えている。また、その際にはヘッドセットのディスプレイ情報をもとに本システムと同様な解析を行うことでデジタルの会話の記録を行えると考えている。

3. 新鮮さの主張、期待される効果など

3.1. 既存サービス、システムとの比較

社会活動やコミュニケーションを測定するシステムはいくつか挙げられる。例えば、NEC社はSlackやTeamsの会話履歴を取得し、従業員のコミュニケーション量などを測定するNEC 協働支援サービスを提供している。Okunoらはスマートウォッチと胸につける一人称ライフログ映像から日常活動を分類し、社会的・身体的活動量の計測を行うシステムを開発した[21]。内橋らはAndroidの視覚、音声センサを用いた個人に紐づいたコミュニケーション可視化システムSignaLogを開発した[22]。しかし、これらのシステムは対面またはデジタルのいずれか一方のコミュニケーションのみを対象としており、両方を統合的に扱うものではない。比較した結果を表1に示す。NEC 協働支援サービスは特定のサービスに限られたデジタル上での会話の分析に特化しており、対面での会話は考慮されていない。一方、Okunoらや内橋らのシステムは対面コミュニケーションに焦点を当てているが、デジタルでの会話は扱っていない。本システムの**対面とデジタルの両方の会話を単一のウェアラブルデバイスで記録・分析できる**点は特筆すべき点であると言える。また、あらゆる場面の会話を記録するということは、**自身の社会的な関わりを総合的に評価することができ、社会的健康状態を多角的に把握することが可能となる。**

3.2. 技術面での斬新さ

本システムは以下の3点で技術的に優位性を持っていると考えられる。

単一のハードウェアで対面、デジタルの会話両方を記録できる

本システムはカメラとマイクを搭載した単一のハードウェアで対面の会話とデジタルの会話を記録することができる。通常であればデジタルの会話を記録しようと考えた時、各コミュニケーションサービスに対してAPIなどを利用して収集する方法があるが、APIが存在しないことも考えられ、加えて様々なコミュニケーションサービスを一つ一つ対応するのは現実的ではない。そこで、人は五感を元にコミュニケーションを行っていることに着目し、**本システムではウェアラブルデバイスからデジタルの会話を記録するという新しいアプローチで解決する。**これにより、ユーザー目線ではデジタルの会話を記録するために専用のソフトウェアをインストールする必要がなく、一つのハードウェアを装着するだけで対面とデジタルの両方の会話を記録できるため、ユーザー体験の向上が期待できる。

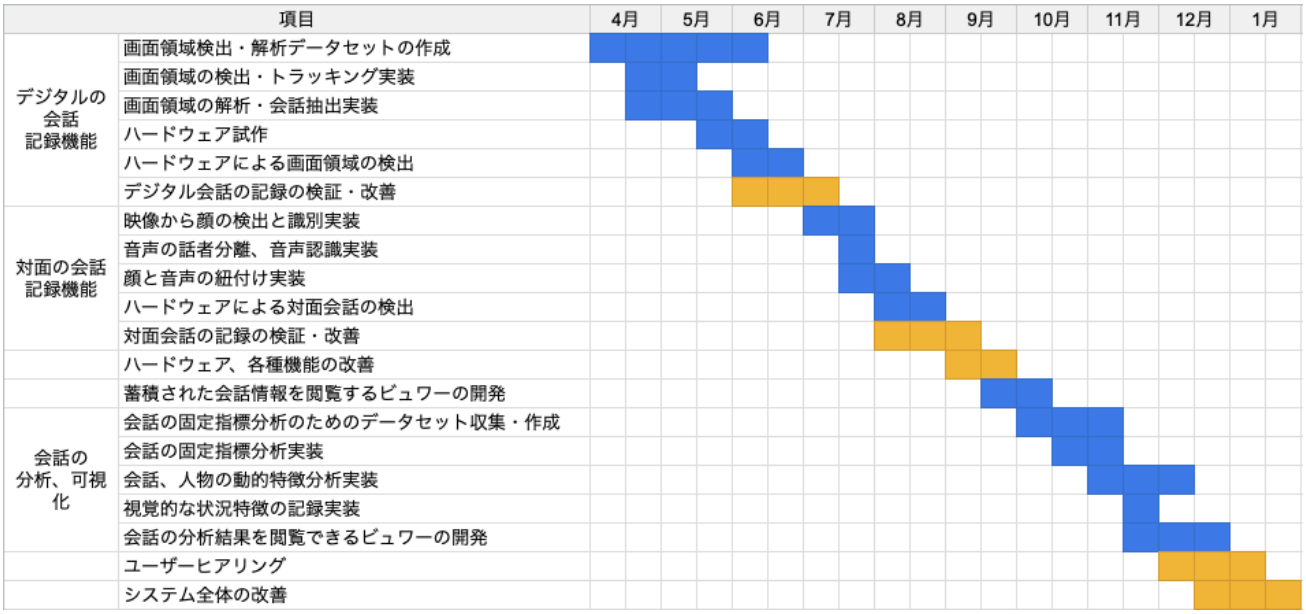

図5 開発線表

また、あらゆる会話を人物に紐づいて記録が行えるようになっており、継続的に利用することで蓄積されるデータはより価値の高いものと考える。

記録した会話から関係性や相手の特徴、関連情報を抽出する

本システムは、記録された対面とデジタルでのあらゆる会話情報を元に自身がどのような会話をするのかを多角的に分析し、様々な指標で可視化することができる。**あらゆる場面での会話の記録が想定され、パターン別で分析方法を変えるなどの分析の工夫を行うことでユーザーの社会活動をより高度に言語化・可視化することができる。**

活用しやすい会話データの管理

本システムは人物に紐づいた会話データの記録に加えて、会話特徴を抽出し、グラフ構造として記録しているため、特定のトピックの情報抽出を容易に行うことができ、ただ会話内容をテキストで保存するのに比べ、検索などのユーザー体験が高くなることが期待できる。そして、**分析されたデータをシステムで活用することで新たな価値の創出も期待できる。**

3.3. 期待される効果など

本システムを利用することで以下の3点が効果として期待できる。

自身と他人との関係性や会話特徴を可視化することによって、自身の社会活動量を俯瞰してふりかえることができ、自身の社会的健康状態を理解することができる

本システムでは、対面、デジタルのあらゆる会話を網羅的に記録する特徴を活かし、様々な視点、指標で分析を行いたいと考えている。これにより、自身が今どんな社会的な関わり方をしているのか、ワークライフバランスは理想的な状態であるか、どういった話題に偏りを持っているのかなど、**細かく自身の社会的健康状態を総合的に理解し、早期に問題を改善することが可能になる。**そして、様々な指標を総合的に見ることで、今までにない人間に関する示唆を得ることを期待できる。

会話の内容まで入り込んだ記録を可能にすることによって、より効果的で具体的な行動改善、推薦が可能になる

自身と相手の会話内容を記録・分析することによって、**相手について詳しく理解することができ、相手との関わり方を見直し、より良い関係性を構築するための示唆を得ることができる。**また、システム側からの具体的な行動改善のアドバイスを行うこともできると考えている。

高度にパーソナライズされたアシスタントAIの実現

このシステムはユーザーに関わる人物との会話を記録する。そのため、自身の発言が一番多く記録される。その自身の会話情報を分析し、言語化・可視化することは、コンピュータに自身の考えを伝える大きな手助けとなると考えられる。つまり、自身の意図を高度に汲み取ったコンピュータの応答や行動の実現が可能であり、**高度にパーソナライズされたAIアシスタントの実現にむけた基盤となることが期待できる。**

4. 具体的な進め方と予算

4.1. 主な開発場所

筑波大学春日エリア 加藤研究室

4.2. 使用する計算機環境

MacBook Pro, さくらのクラウド, GCP, 任意のGPU（VRAM 24GB以上）を搭載したデスクトップPC

4.3. 使用する言語・ツール

Python, JavaScript, Flutter

4.4. 開発計画

プロジェクトの進め方については図5に示すような流れで開発する。

4.5. 予算内訳

表2 予算内訳

用途	金額
人件費	2,880,000 円
必要経費	79,050 円
合計	2,955,050 円

人件費

週40時間（月160時間）ペースでプロジェクトを推進する。このことから、160 [時] x 9 [月] x 2000 [円] = 2,880,000 [円]。よって、2,880,000円を申請する。

必要経費

本システムを開発、運用するにあたり、多くの深層学習モデルを学習、推論させる必要がある。初期運用での推論は基本的にはローカルのGPUを搭載したPCで行うようにし、モデルの学習に関しては規模に応じてクラウドのGPUサーバーを用いて学習を行いたいと考えており、NVIDIA V100を50時間利用すると見積もっている。さくらのGPUサーバーでNVIDIA V100を50時間分利用するとV100は１時間481円であるため、481 [円] x 50 [時] = 24,050 [円]である。加えて、ウェアラブルデバイスの開発に関しては試作などを含め最低でも30,000円必要と考えている。また、会話分析機能の開発にあたり、有償のデータセットを利用し検証を行いたいと考えており、このデータセットの利用に25,000円支払う必要がある。これらの経費については自費で負担する。

5. 提案者の腕前を証明できるもの

藤巻晴葵

現在に至るまで、授業、研究、ハッカソン・コンテスト、個人開発、会社で数十のシステムの開発を行なってきた。基本的にはソフトウェアの開発をメインとして行なってきたが、ハードウェアの開発に関してもいくつか作成してきた。また、高専4年次に会社を創業し、その会社での活動がいくつかのメディアで取り上げられている。以下にこれまでの成果を示す。

2019年度 全国高専プログラミングコンテスト 課題部門（最優秀賞・文部科学大臣賞）

「:::doc（てんどっく）」という墨字の文書と点字の文書を相互に変換するシステムを開発した。

2020年度 全国高専プログラミングコンテスト 課題部門（最優秀賞・文部科学大臣賞）

「ぷらんとこれくしょん」という小学生の観察学習での利用を目的としたアプリケーションを開発した。児童たちが植物や昆虫を撮影し、名前を自動で判別しマップにプロットされ、季節ごとや他学校とのデータを利用し、植物や昆虫の生息場所などを比較しながら学習することができる。

図6 藤巻の実績:左から:::doc, ぷらんとこれくしょん, DCON

2020年度 高専ディープラーニングコンテスト（DCON）（最優秀賞）

前述の「:::doc（てんどっく）」をtoB向けに改良した自動点字翻訳システムを開発した。情報の発信者が手軽に文書を点字などの視覚障害者が読める形に変換し提供することができる。

2022年度 情報処理学会 コンシューマ・デバイス＆システム（CDS）研究会 CDS36（優秀発表賞）

レイアウト付き文書に対応したクラウド型点字翻訳システムの実用化と深層学習による半自動化

2022年度 総務省 異能vation

異能ジェネレーションアワード受賞

その他の制作物

- ワンタイムQRコードによる不正防止モバイルスタンプラリーシステム
- 画像認識による三次元的な駅構内の位置特定とARによる案内システム
- 紙媒体、デジタルカレンダーの情報を相互に共有できるシステム
- 画像認識を用いた、簡単に利用できる廃棄野菜のオークションシステム
- BLEビーコンを用いた、視覚障害者への駅構内に関するリアルタイム情報提供システム

6. プロジェクト遂行にあたっての特記事項

現在、私は筑波大学情報学群知識情報・図書館学類3年に所属しており、来年度4年に進学する。私は大学の授業をほとんど取り切っており、4年では卒業研究がメインとなる。研究室の指導教員から、未踏の応募について容認されており、プロジェクト遂行に当たって問題になる可能性は極めて低い。

7. ソフトウェア作成以外の勉強、特技、生活、趣味など

藤巻晴葵

趣味はいくつかあるが、主に二つある。一つ目はサイクリングである。筑波大学へ編入する前の高専に在学していた際は、家から高専まで、アップダウンのある片道10kmをクロスバイクで登校していた。そのため、自身の中でのクロスバイクでの移動は日常であり、体を動かしたくなった際には100km未満の距離のサイクリングをよく行なっている。二つ目はアニメ、ドラマ、映画などの動画コンテンツの視聴である。特にアニメは毎シーズン必ず何本かのアニメを見るようにしている。好きな内容の傾向としては、バトル系というよりも、心理描写がしっかりと描かれているヒューマンドラマなどを好みとしている。

また、大学のサークルなどを通してイベントの運営などを行なっている。TEDxUTsukubaという学生団体では、毎年100名規模のイベントを開催している。このイベントではワークショップや6名のスピーカーによる登壇セッションなどがあり、自身は協賛していただけるパートナーとの交渉や、翻訳システムの開発に携わった。また、エンジニアを対象としたミートアップ運営に携わっている。イベントを通じて、様々な背景を持った人たちと関わることで日々新しい気づきを得ている。

8. 将来のITについて思うこと・期すること

近年、LLMが話題になったことによって、エンジニアの業務を改善するためのツールが多くリリースされ、普及も進んでいる。例として、Cursorや

Github Copilotがあげられる。どちらもコーディングを支援するツールである。今までコードエディターに搭載されていたルールベースのコード補完とは違い、先ほどのツールはプログラマーの意図を高度に先読みし、ほとんどのコードを補完してくれる。私はこのツールの中でもGithub Copilotを利用しているが、体感として、自分がコードを書くという作業が1/2になったのではないかと感じている。それほどに強力なツールである。ここで一つ思うことがあるかもしれない。「このツールが発展していく、AIが発展していくことでエンジニアがいらなくなるのではないか」と。私は、エンジニアという存在がなくなることはないと感じている。一方、AIの進化によって、より開発することのコストは小さくなっていくと容易に想像がつく。今、あるサービスを開発するのに1年かかったとして、10年後には1週間、1日、1時間で作れるようになると考えられる。だが、考えてみてほしい。プログラミングに関しては、過去10年間の間で新しいプログラミング言語やフレームワーク、ライブラリの登場によってどんどんサービス開発等のコストが下がっていることに。しかし、今もプログラマーという需要は大きい。結論として、技術的なパラダイムシフトがあっても本質的に人間が行う活動というのはなくならないのだ。確かにプログラマーという存在はなくなるかもしれない。だが、人間が思考するというのは自分のクローンが作られるまで続く。エンジニアという存在も時の流れによって、当時やっていたことをしなくなり、また違う部分に焦点を当てて活動するようになる。私は将来のITは日々新しいことに取り組み、それぞれが常に誰かに影響を及ぼすよりカオスな状態となると考える。

参考文献

[1] NECソリューションイノベータ「NEC 協働支援サービス」, https://www.nec-solutioninnovators.co.jp/sl/wcs/index.html
[2] Jocher et al. (2023). Ultralytics YOLO (Version 8.0.0), https://github.com/ultralytics/ultralytics
[3] J. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in IEEE, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.
[4] Hervé Bredin (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In Proc. INTERSPEECH 2023.
[5] Radford et al. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.
[6] Reazon Human Interaction Laboratory「ReazonSpeech」 https://research.reazon.jp/projects/ReazonSpeech/
[7] Jee-weon Jung et al.. (2024). ESPnet-SPK: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models.
[8] Alexander Kirillov et al. (2023). Segment Anything.
[9] Yifu Zhang et al. (2022). ByteTrack: Multi-Object Tracking by Associating Every Detection Box.
[10] Nir Aharon et al. (2022). BoT-SORT: Robust Associations Multi-Pedestrian Tracking.
[11] Kay, A. (2007). Tesseract: an open-source optical character recognition engine. Linux J., 2007(159), 2.
[12] Google Cloud「Vision AI | Google Cloud」 https://cloud.google.com/vision
[13] Nicolas Carion et al. (2020). End-to-End Object Detection with Transformers.
[14] Yupan Huang et al. (2022). LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking.
[15] Victor Sanh et al. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
[16] Agarwal et al. (2020). Machine Translation Aided Bilingual Data-to-Text Generation and Semantic Parsing.
[17] Igor Melnyk et al.(2022). Knowledge Graph Generation From Text.
[18] Alec Radford et al. (2021). Learning Transferable Visual Models From Natural Language Supervision.
[19] Ron Mokady et al. (2021). ClipCap: CLIP Prefix for Image Captioning.
[20] 国立国語研究所「日本語日常会話コーパス」, https://www2.ninjal.ac.jp/conversation/cejc.html
[21] Okuno et al. (2020). Lifelog visualization based on social and physical activities. Association for Computing Machinery.
[22] 未踏iPedia「個人に紐付くメディア情報を用いたコミュニケーション可視化ツールの開発」, https://jinzaiipedia.ipa.go.jp/mitou_ipedia/development_result/post/個人に紐付くメディア情報を用いたコミュニケー