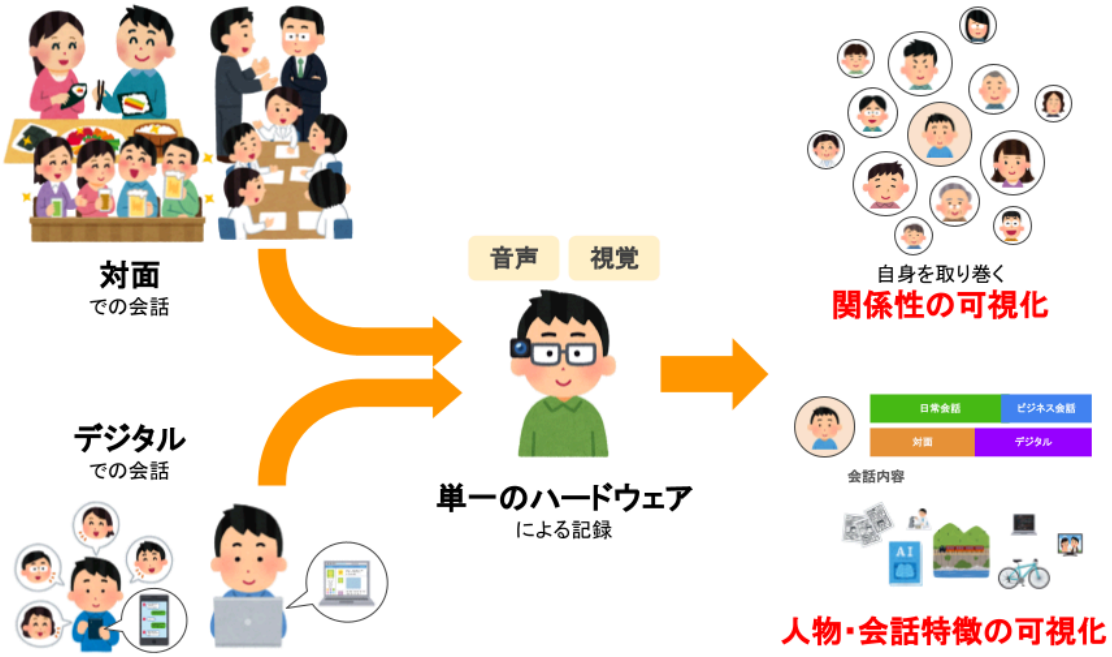
プロジェクト名：一人称カメラとマイクによる対面、デジタルの双方の会話を記録し、特徴を分析、可視化を行うシステムの開発
申請者名：藤巻晴葵



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、対面での会話、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。我々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。対面とデジタル、両方のコミュニケーションをバランス良く行い、良好な人間関係を構築・維持するためには、客観的な指標に基づいた評価と改善が不可欠である。
そこで本プロジェクトでは、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。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。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したり、各プラットフォームのデータを抽出する作業を行う必要がない。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。
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。

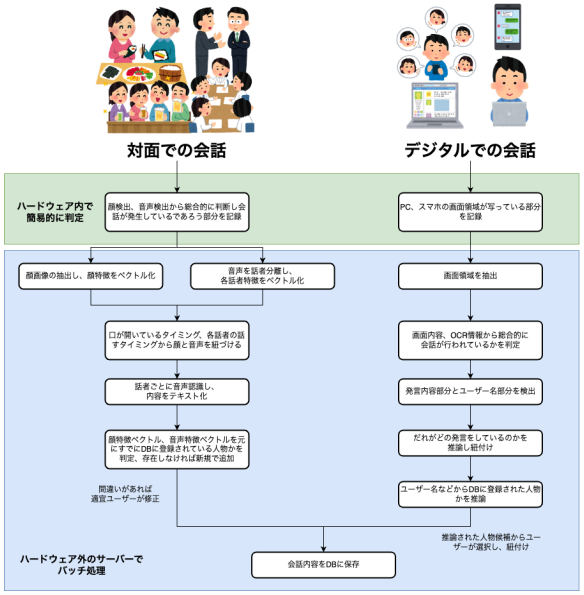
1.2. 提案するもの

本システムは、ユーザーの対面、電子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3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。私は胸元にカメラを装着し検証したところ、FullHD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、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。具体的には、顔検出や音声検出、PC、スマホの画面検出を、デバイス側にも簡易的に組み込み実装する。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: 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: 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: 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。**

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[1]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[2]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[3]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[4]やReazonSpeech v2[5]といった音声認識モデルを利用し、音声認識を行う。また、Youngmoon Jungらの音声識別の研究[6]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。

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

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

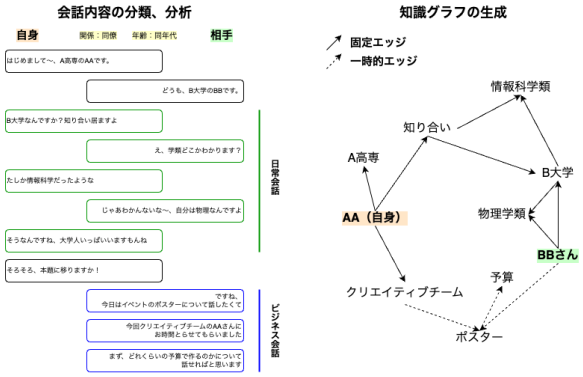
いる。この検出については、検証を行っている。まず、すでに公開されているデータセットから学習することで検出モデルを学習できないかを考えたが、端末の画面の領域のみセグメンテーションされたデータセットはあまり存在しなかったため、今回はSegment Anything[7]という汎用セグメンテーションモデルを利用し画面部分のセグメンテーションと処理を行って4点の座標からなる画面領域検出が行えるかの検証を行った。検証の結果は参考資料に動画を添付している。この検証から分かったことは、Segment Anythingでの処理では手法の都合上、安定した検出を行うのが難しく、処理コストについても理想的ではないことがわかった。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。検出された物体のトラッキングについては、Yifu Zhangらによって提案されているByteTrack[8]やNir Aharonらによって提案されているBoT-SORT[9]といった手法を利用し、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[10]やGCP Cloud Vision API[11]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二つ考えられる。まずはYOLO v8、DETR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、Yupan 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[12]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。二つ目の手法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、初めは一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。

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。

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[13]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[14]、Igor 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[15]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、「ユーザーが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[16]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、Ron Mokadyらの提案したClipCap[17]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。

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いと考えている。このシステムの売りは単一のハードウェアで対面、電子的な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おり、Oculus Quest 3やXREAL Air 2 Ultra, Apple Vision Proなどの利用を想定した、専用のハード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、このシステムは個人での利用のみならず、企業単位での利用にも発展できると考えている。例えば、接客業や営業においての客との関係性や特徴を記録することによるコンバージョン増加の期待や社内でのコミュニケーションを可視化することによる組織で発生しうる問題の早期発見と予防を期待することできる。本システムをそのままビジネス利用することは難しいが、基礎技術の確立として開発することはとても意義のあるものだと考えられる。

3. 新鮮さの主張、期待される効果など

3.1. 既存サービスとの比較

社会活動を測定するシステムはいくつか挙げられる。例えば、Laneらはスマートフォンに搭載されている加速度センサやマイクなどを用いて、身体的活動の記録や社会的活動の記録、睡眠の記録を行い、日常生活を可視化することで、ウェルビーイングの向上を促すBeWellアプリケーションを開発した。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するシステムを開発した。Jen-Anらは、胸につけた一人称カメラから得た社会活動を写した動画を顔の写り込みなどを元に要約し、短い動画に出力するシステムを開発した。いずれも本システムの目的や計測のアプローチが異なる。本システムは視覚情報と音声情報の両方を活用し、総合的な会話量を計測するだけでなく、**対面、デジタル両方の会話内容を人物別で記録することで詳細な社会活動量を可視化することができ、自身を取り巻く人物との関係を客観視し、具体的な改善行動を考えることができる。**

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。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、本システムでは一つのウェアラブルデバイスからデジタルの会話を記録できる仕組みを開発する。これにより、デジタルの会話の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、開発者目線では各プラットフォームに応じた開発を必要とせず、ユーザー目線では一つのハードウェアを装着するだけで両者の記録ができ、ユーザー体験が良い。また、人物に紐づいた会話の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会話情報を元に自身がどのような会話をするのかを分析し、色々な指標で可視化することができる。また、自身を取り巻く人物の特徴や情報を知識グラフとして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、会話に関する分析をより発展させることが可能である。

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、デジタル両方での会話を人物別で記録し、会話の特徴、相手との関係性を可視化できる。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理解し、早期に問題を改善することが可能になる。

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。

組織内で利用することによって組織の会話を可視化することができ、活動を活発化させる

組織で本システムを使用することで、組織状態を詳細に可視化することが可能である。どこまで情報を開示するのかに関しては別途考える必要があるが、組織内の部門間や人の間でどれくらい交流を持てているのか理解することや、各個人についての理解を深めること可能であり、より効果的に仕事を進めるための施策を考えることができ、組織内の生産性の工場に繋がることが期待される。

4. 具体的な進め方と予算

主な開発場所

筑波大学春日エリア 加藤研究室

使用する計算機環境

MacBook Pro, GCP, etc.

使用する言語・ツール

Python, JavaScript

開発計画

予算内訳

人件費

まず、提案者である藤巻は1440 [時間] x 2000 [円]=2,880,000 [円]

必要経費

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）
    - 自動点字墨字相互翻訳システム

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
    - 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
    - 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。

2020年度　高専ディープラーニングコンテスト 2020

- 最優秀賞
- :::doc（てんどっく）
    - toB向けに自動点字翻訳システムを開発し、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、藤巻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。藤巻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での在学中は家から高専まで、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ができることに喜びを感じている。

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] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

[2] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 5962-5979, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.

[3] Hervé Bredin (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In Proc. INTERSPEECH 2023.

[4] Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I.. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

[5] Reazon Human Interaction Laboratory「ReazonSpeech」https://research.reazon.jp/projects/ReazonSpeech/(2024、3月12日閲覧).

[6] Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Barry-John Theobald, Ahmed Hussen Abdelaziz, & Shinji Watanabe. (2024). ESPnet-SPK: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, & Ross Girshick. (2023). Segment Anything.

[8] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, & Xinggang Wang. (2022). ByteTrack: Multi-Object Tracking by Associating Every Detection Box.

[9] Nir Aharon, Roy Orfaig, & Ben-Zion Bobrovsky. (2022). BoT-SORT: Robust Associations Multi-Pedestrian Tracking.

[10] Kay, A. (2007). Tesseract: an open-source optical character recognition engine. Linux J., 2007(159), 2.

[11] Google Cloud「Vision AI | Google Cloud」https://cloud.google.com/vision(2024、3月12日閲覧).

[12] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, & Furu Wei. (2022). LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

[14] Agarwal, O., Kale, M., Ge, H., Shakeri, S., & Al-Rfou, R. (2020). Machine Translation Aided Bilingual Data-to-Text Generation and Semantic Parsing. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+) (pp. 125–130). Association for Computational Linguistics.

[15] Igor Melnyk, Pierre Dognin, & Payel Das. (2022). Knowledge Graph Generation From Text.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, & Ilya Sutskever. (2021). Learning Transferable Visual Models From Natural Language Supervision.

[17] Ron Mokady, Amir Hertz, & Amit H. Bermano. (2021). ClipCap: CLIP Prefix for Image Captioning.