マルチモーダル情報を用いたソーシャルスキルの客観的推定

佐賀　健志† 　田中　宏季† 　岩坂　英巳†† 　中村　哲†

† 奈良先端科学技術大学院大学 情報科学研究科　〒 630–0192 奈良県生駒市高山町 8916–5

†† 奈良県立医科大学 精神医学講座　〒 634-8521 奈良県橿原市四条町 840 番地

E-mail: †{saga.takeshi.sn0,hiroki-tan,s-nakamura}@is.naist.jp

あらまし　ソーシャルスキルトレーニング（SST）は、日常のコミュニケーション中に適切なソーシャルスキルを習得するために使われる認知行動療法のひとつである。しかし、SST のアクセス性にはいくつかの課題が存在しているため、これを解決するために SST の自動化研究が行われている。本稿では SST の自動訓練システムの実現にむけて、マルチモーダル情報を利用したソーシャルスキルの客観的評価手法を提案する。客観的評価指標として対人応答性尺度（SRS-2）を利用した。モデルには線形回帰を使用し、マルチモーダル情報を入力として学習をさせることで自動評価を試みた。28 種類の音声特徴量と BERT 埋め込みに基づく発話系列類似性および Facial Action Units 等の特徴量を新たに採用することで、SRS-2 の全体スコア予測では相関係数 0.52、SRS-2 の治療下位尺度である社会的コミュニケーションスコア予測では相関係数 0.65 を達成した。

キーワード　ソーシャルスキルトレーニング、社会的信号処理、マルチモーダル分析、線形回帰、Facial Action Coding System、対人応答性尺度

Takeshi SAGA†, Hiroki TANAKA†, Hidemo IWASAKA††, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology, Facility of Information Science,Takayama-cho 8916-5, Ikoma-shi, Nara, 630–0192 Japan

†† Nara Medical University, Department of Psychiatry, Shijoh-cho 840, Kashihara-shi, Nara, 634-8521 Japan

E-mail: †{saga.takeshi.sn0,hiroki-tan,s-nakamura}@is.naist.jp

Abstract　Social Skills Training (SST) is an effective method to acquire appropriate social skills in daily communications. In this paper, we present objective social skills level estimation method for SST automation by using linear regression with multimodal information. By adding new feature set including 28 audio features and BERT-based sequential similarity, we achieved correlation coefficient of 0.52 for SRS-2 overall score, and 0.65 for SRS-2 social communication score which is a treatment subscale of SRS-2.

1. はじめに

ソーシャルスキルは他者とのコミュニケーションにおいて重要な要素の 1 つである。 Bellack らの定義では、ソーシャルスキルは主に「アイコンタクトや姿勢による表現行動」、「感情認識に関連する手がかりに注意を払い、その解釈を行う受容行動」、「応答のタイミングや話者交替などのインタラクティブ動作」の 3 つの要素によって構成されている [1]。通常、私たちは幼少期にこれらのスキルを習得するが、自閉症スペクトラム症候群（ASD）などの理由でうまく学習できない人もいる [2]。ほとんどの日常生活はいくつかのソーシャルスキルの複合的な応用を必要とするため、そのようなスキルがなければ日常生活は非常に難しいものとなる。このように日常生活において非常に重要なソーシャルスキルだが、幼少期で学んできたことを前提に社会活動は営まれているため、大人になってから学ぼうと思っても簡単には身に付けることはできない。

この難しさを軽減する方法の 1 つとして、認知行動療法の一種、ソーシャルスキルトレーニング（SST）がある。Bellack らによると SST は「条件付き反射療法」と「相互抑制による心理療法」、「社会的学習理論」に基づいており、様々な分野で広く使用されてきた [1], [3]〜[6]。また、SST は特定の疾患だけを対象としているわけではないため、ソーシャルスキルに悩みを抱える人なら誰でもソーシャルスキル向上の効果が期待できる点も特徴のひとつである。

SST はリーダー 1 人とメンバー 1 人を最小構成としているが、複数のメンバーによって構成されたグループで行うこともできる [1]。その場合は、サブリーダーがリーダーを助けるために設けられることが一般的である。SST の基本的な流れは以下の通りである。まず、リーダーとメンバーが相談しながら SST の目標とそのスキルが必要となる状況を設定する。次に、

リーダーが状況設定に従い演技することでその状況に適切な立ち振る舞い方を示す。その後、メンバーは示された手本を基に自分自身で立ち振る舞い方を模倣する。その演技に対してリーダーはうまくできていた点とできていなかった点についてのフィードバックを与え、それに基づき、メンバーは繰り返し演技を行うことでソーシャルスキルの習得を目指す。また、SSTによって得られたソーシャルスキルをより汎化させるために、リーダーの判断で次回 SST 実施時までの宿題をメンバーに与える場合もある。

SST は一般的な手法になりつつあるが、誰もがアクセスできるわけではないのが現状である。理由の 1 つとして、精神疾患を抱えた人々に対する社会的な偏見が挙げられる [7], [8]。うつ病などの精神疾患が社会に正しく認知され状況は変わりつつあるが、いまだに精神疾患患者に対する社会的偏見は存在する。そのため、患者は精神疾患を抱えていることが隣人に知られないように気を配らなくてはならない場合がしばしば存在する。もう 1 つの理由として、SST リーダーになるためのトレーニングのハードルの高さが挙げられる。SST を効果的に実施するには、SST の仕組みを正しく理解し、適切なフィードバックを行う方法を身に付けなくてはならない。一般にこの技術習得には長期間が必要とされ、SST 普及協会の SST 認定講師になるためには最低でも 90 時間のリーダー経験が必要である [9]。

この問題を解決するために先行研究において SST の自動化が提案されてきた。本稿では、そのような自動 SST システムでの使用を想定し、ソーシャルスキルレベルの客観的推定手法について提案する。客観的評価指標には社会的応答性尺度第 2 版（SRS-2）を利用して、線形回帰モデルでの推定を行った [10]。この際、より正確な SRS-2 推定モデルを構築するために、BERT 埋め込みに基づく発話系列類似性（以下、Seq-similarity と呼称する）を含むいくつかの新しい音声およびテキスト特徴量を使用している。

本研究はまだ実現可能性の検証段階にあるため、本稿では最初のステップとして健康な被験者のみを対象として提案手法の検証を行った。本手法の実際の精神疾患患者を対象とした有効性については、さらなる検証が必要である。

2. 先 行 研 究

先行研究において、誰でも SST を利用できるような環境実現を目指し、いくつかの仮想エージェントによる自動化が提案されている。SimSensei は、PTSD 患者向けのセラピーを実施する仮想心理療法エージェントであり、人間による対面セラピーに近い効果が報告されている [11]。しかし、本論文執筆時で SimSensei を SST のような行動療法の自動化に応用した研究は報告されていない。田中らは仮想エージェントによるマルチモーダル対話システムを使用して、SST の自動化を試みた [12]。ユーザーのスピーキングスキルを向上させることを目的とした彼らの研究では、ユーザーのスピーキングレベルを測定するための評価指標としてスピーキングスキルスコアを用いている。このスコアは経験豊富な数人のセラピストによってつけられた点数の平均値で、ユーザーのソーシャルスキルレベルの推定モデルを学習させる際の正解ラベルとして用いられた。しかし、このスコアはセラピストの主観に基づいているため、個人のバイアスがかかっている可能性がある。

この問題を解決するために本研究では客観的指標に基づいたモデルの実現に取り組んだ。

3. 対人応答性尺度（SRS-2）

対人応答性尺度（SRS-2）は、65 項目の質問で構成される評価指標である。SRS-2 は元々、自閉症スペクトラム障害（ASD）を潜在的に抱えている人を評価するために設計されたが、統合失調症などの精神疾患を区別することもできることが報告されている [10]。また、健康な人々を対象とした実験においても被験者の ASD 傾向を判別できることが確認されているため、総合的かつ客観的な評価指標として使用できると期待できる。

本論文では、SRS-2 の「総合スコア」とその治療下位尺度である「社会的コミュニケーション（65 項目のうち 22 項目）」を学習時の目的変数として利用した。SRS-2 の総合スコアには、純粋なコミュニケーションスキルだけではなく生活スタイルなどに関する項目も含まれるため、ユーザーのソーシャルコミュニケーションスキルをより明確に示している社会的コミュニケーションスコアも使用することとした。社会的コミュニケーションスコアは社会的相互作用の物理的側面を示しており、直感的に理解でき、客観的に観察できるため、今後の SST 自動化にも適していると考えられる。

社会的コミュニケーションスコアと総合スコアは、相関係数 0.92 と高い相関であった。その一方で、先行研究で使用されていた主観的評価指標であるスピーキングスコアとの相関係数は、それぞれ-0.19 と-0.29 という弱い負の相関であった [13]。

4. マルチモーダル特徴量

ソーシャルスキルは音声や視覚といった複数のコミュニケーション手段によって構成されているため、総合的なソーシャルスキルの推定を行うためにはマルチモーダル特徴量を使用する必要がある。このセクションでは、各モダリティで使用した特徴量と基本的な考え方を紹介する。提案モデルで使用した特徴量を表 1 に示す。

4.1 音声特徴量

先行研究において、スピーキングスキルを推定するための音声情報の重要性が示された [12], [14]。これらの研究では 4 つの異なる音声特徴量を使用していたが、本研究ではその数を 28 に増加させることとした。今回追加した特徴量は就職面接におけるスコアの推定システムに使用されていたものである [15]。就職面接は SST 同様にマルチモーダル情報を統合的に使用していることから、ソーシャルスキルスコアの推定に有効であると容易に予測できる。

特徴量の抽出には韻律分析用オープンソースソフトウェアである Praat を使用した [16]。

4.2 テキスト特徴量

田中らの研究で使用されたテキスト特徴量を基本特徴量として使用した [12]。1 分あたりの単語数（WPM）はユーザーの

表 1　マルチモーダル特徴量（上から音声、テキスト、画像）

Feature name	Description
Energy	平均スペクトルエネルギー
F0, F1, F2, F3 Mean	F0,F1,F2,F3 の平均周波数
F0, F1, F2, F3 SD	F0,F1,F2,F3 の標準偏差
F0 Min	F0 周波数の最小値
F0 Max	F0 周波数の最大値
F0 range	F0 周波数の最小値と最大値の差
F1, F2, F3 BW	F1,F2,F3 の平均バンド幅
F2/F1, F3/F1 Mean	F2-F1, F3-F1 の平均比率
F2/F1, F3/F1 SD	F2-F1, F3-F1 の比率の標準偏差
Int mean	平均音声強度
Int Min	最小音声強度
Int Max	最大音声強度
Int range	最小音声強度と最大音声強度の差
Int SD	音声強度の標準偏差
Jitter	F0 周波数のゆらぎ
Shimmer	音声強度のゆらぎ
Unvoiced %	無声区間の比率
Breaks %	短時間無声区間の比率
WPM	1 分間の発話単語数
Six plus	6 文字以上の単語の数
Fillers	フィラーの数
Vocabulary size	語彙数
Seq-similarity	BERT 埋め込みに基づく発話系列類似性
AU01	眉の内側を上げる
AU02	眉の外側を上げる
AU04	眉を下げる
AU05	上瞼を上げる
AU06	頬を持ち上げる
AU07	瞼を緊張させる
AU09	鼻にしわを寄せる
AU10	上唇を上げる
AU12	鼻唇溝を深める
AU14	えくぼを作る
AU15	唇両端を下げる
AU17	オトガイを上げる
AU20	唇両端を横に引く
AU23	唇を固く閉じる
AU25	顎を下げずに唇を開く
AU26	顎を下げて唇を開く
AU28	唇を吸い込む
AU45	まばたく
AU06+12	幸せな感情と相関する AU の組み合わせ
Pitch	頭のピッチ方向の角度
Yaw	頭のヨー方向の角度
Roll	頭のロール方向の角度

発話速度を示している。先行研究において、この特徴量が面接における総合スコアと相関していると報告されていたためソーシャルスキルの推定にも応用できると期待される [17]。 6 文字以上で構成される単語数（Six Plus）は、ユーザーが複雑すぎる不自然な単語をどれくらい使用しているかを示す特徴量である。ASD 患者は健常者に比べて複雑すぎる不自然な単語を頻繁に使用する傾向があると報告されているため、特徴量として加えている [18]。6 文字という具体的な数字は先行研究における提案に基づいている [19]。Fillers は発話中のフィラーの出現回数を示している。フィラーが多すぎると対話相手は内容に集中できなくなってしまう。

本論文では、次の 2 つの仮説に基づいて、「語彙数」「Seq-similarity」という新しいテキスト特徴量を追加した。

（1） ソーシャルスキルの高い人はより多くの語彙を使う

（2） ソーシャルスキルの高い人の話には一貫性がある

語彙数を計算する際に、単語の活用による影響をなくすために日本語形態素解析器 MeCab と追加辞書 ipadic-NEologd を使用して、全ての単語を原型に戻している [20]〜[23]。

また、先行研究で使用されていた手法を参考に、BERT を利用した Seq-similarity を新たに特徴量として使用した [24]。まず、各単語に対応する埋め込みを BERT によって抽出する。その後、それら埋め込みを利用して、1 番目と 2 番目の単語、2 番目と 3 番目の単語の間でコサイン類似度を算出する。最終的に、それらを平均化することによって Seq-similarity は計算される。先行研究では tf-idf に基づいた埋め込みが使用されていたが、本研究では以下のような理由で BERT による埋め込みを使用した。第一の理由はデータによる制限である。今回使用したデータは発話者ごとの発話内容がまったく異なっている上に、データの絶対量が不足していた。そのため、出現頻度に基づく tf-idf がうまく機能しないと考えられた。第二の理由は BERT の汎用性の高さである。BERT は、Transformer 構造に基づくニューラルネットワークの一種であり、隣接するコンテキストを考慮して単語の埋め込みを出力できるとされる [25]。そのため、これまでの Word2Vec よりも汎用性が高く、様々な研究トピックにおいてうまく機能することが報告されている。本研究では Transformer に基づいたモデルを複数含むライブラリ Transformers の中から、Whole-word-masking で事前学習された日本語 BERT モデルを使用した [26]。

4.3　画像特徴量

本研究では各映像フレームにおける表情に着目し、「全フレーム数に対する笑顔フレーム数の割合」と「頭の向き」および「Action Unit（AU）」を画像特徴量として使用した。笑顔検出には画像処理ライブラリ OpenCV が提供する Haar-like 特徴量に基づいた学習済みの笑顔検出器を使用している [27]。頭の向きおよび Action Unit の検出には顔解析ライブラリ OpenFace を使用した [28]。Ekman らによる Facial Action Coding System によると、表情は表情筋の収縮や弛緩によって生じる AU の組み合わせによって構成されている [29]。この理論に基づき、本研究では表情を AU レベルに分解し、モデルに入力することでスコア推定性能向上を試みた。各フレームに対して AU が出現しているかどうかは 0 または 1 の二値で表現され、全フレームの平均値を特徴量として用いた。また、EMFACS によると AU06 と AU12 が同時に検出された場合は「幸せ」な感情を意味しているとされるため、AU06 と AU12 が同時に出現しているかどうかも特徴量として加えた [30]。

5. 実　　験

5.1 マルチモーダル特徴量を用いた SRS-2 スコア予測

Table 1 に示すマルチモーダル特徴量を使用して線形回帰モデルを学習させた。今回は、予備実験において学習結果が悪化したことから正則化は適用しなかった。また、先行研究のモデルと比較するために、田中らのモデルを同じデータで学習させて基準モデルとして比較を行った [12]。ベースラインモデルの入力特徴量には、1 分あたりの単語数、6 文字を超える単語数、フィラーの数、音声振幅の平均、F0 変動係数、発話中の一時停止率、F1 と F3 の間のスペクトル傾斜（H1A3）、全フレームに対する笑顔フレームの割合、頭の向き（ピッチ、ヨー、ロール）が使用されている。

5.1.1 学習データ

モデルの学習と評価には、先行研究によって得られた日本語の 1 分間スピーキングビデオを使用した [13]。まず、被験者はノートパソコンの画面を介して MMDAgent（`http://www.mmdagent.jp/`）で作成された仮想エージェントに対して、最近あったうれしかった出来事について 1 分間話す。被験者が話している間、被験者の顔の映像と声が記録され、後日、注釈者によってテキストに書き起こされた。 1 分間の発話終了後、65 項目の SRS-2 質問紙を用いて、被験者の生活スタイルやコミュニケーションスタイルなどについて記入してもらった。以上のような手順によって 27 人の健康な被験者のデータセットは作成された。

5.1.2 外れ値の除去

基準モデルの画像特徴量を抽出する際に、一部のデータにおいて頭の向き推定が失敗していたことが原因で特徴量に大きな外れ値が発覚した。そのため、このデータをデータセットから削除し、残りの 26 人のデータで学習を行った。

5.1.3 実 験 詳 細

特徴量ごとの偏りを是正するために、学習前の全特徴量に対し、平均が 0 、分散が 1 になるように標準化した。また、他の研究よりも比較的小さいデータセットサイズの問題に対処するために、Leave-one-subject-out 交差検証を適用して学習と評価を行っている。

5.2 結　　果

総合スコア予測では基準モデルと提案モデルについてそれぞれ、相関係数 0.28 と 0.52、 RMSE32.13 と 17.44、絶対誤差の標準偏差 24.08 と 10.69 となった。図 1 にモデルの SRS-2 全体スコアの正解値と予測値の散布図を示す。

社会的コミュニケーションスコア予測では基準モデルと提案モデルについてそれぞれ、相関係数 0.45 と 0.65、RMSE11.33 と 7.43、絶対誤差の標準偏差 8.17 と 4.24 となった。社会的コミュニケーションスコア予測では、基準モデルと提案モデルの両方が無相関検定において有意に相関していることが確認された（$p < 0.05$）。さらに、提案モデルの結果は p 値 0.01 のしきい値でも有意に相関していた（$p = 0.006$）。また、提案モデルの RMSE は基準モデルより 3.90 小さくなっている。図 2 は、提案モデルの社会的コミュニケーションスコアの正解値と予測を使用した散布図を示している。

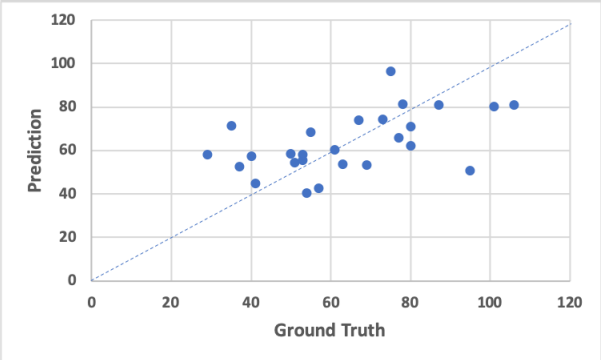

図 1　提案モデルによる SRS-2 総合スコア予測（r：0.52）

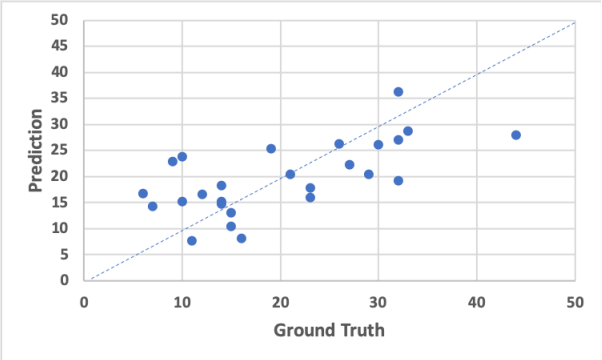

図 2　提案モデルによる SRS-2 社会的コミュニケーションスキルスコア予測（r：0.65）

表 2 の左側は、SRS-2 の総合スコア予測と社会的コミュニケーションスコア予測におけるモデル係数の上位 5 つを示している。これらの係数は 26 回の交差検証結果を平均した値となっている。また、全試行において係数が安定しているもしくは結果に影響を与えるには小さすぎたため、本実験において重大な多重共線性は存在していないと考えられる。

6. 考　　察

SRS-2 総合スコア予測の結果から、提案モデルは SRS-2 総合スコアと社会的コミュニケーションスキルスコアの双方においての基準モデルの結果を上回っていることが確認された。特に社会的コミュニケーションスキルスコアの予測では、相関係数、RMSE、絶対誤差の標準偏差がそれぞれ 0.65、7.43、4.24 であったため、提案モデルの予測はより正確でより精度の高いものであった。さらに、スコア予測に最も影響を与えた特徴量を特定するために、学習済みモデルの係数についての解析を行った。本実験において、全ての入力特徴は-1 から 1 の値範囲に正規化されているため、各特徴量に対応するモデル係数を特徴量の重要度として取り扱った。表 2 は提案モデルにおけるモデル係数（絶対値）の大きい上位 20 の特徴量の名称とそのモデル係数を、総合スコア予測（左側）と社会的コミュニケーションスキルスコア予測（右側）に分けて示したものである。総合スコア予測の結果に着目すると、「AU28」と「BERT を用いて計算された発話系列類似性（Seq-similarity）」が他のものと比較してはるかに支配的であることがわかる。同様の傾向が社会

表 2　提案手法におけるモデル係数（左）SRS-2 総合スコア予測
（右）SRS-2 社会的コミュニケーションスコア予測

Name	Coefficients
AU28	-409.6
Seq-similarity	-115.6
Pitch	45.9
AU25	-42.7
AU15	-39.0
AU01	35.5
AU26	-33.6
AU45	31.9
AU17	-29.3
AU09	27.3
Yaw	-25.7
AU02	20.8
Smile Ratio	-12.7
Energy	-12.4
AU07	12.2
AU12	-12.2
F2/F1 Mean	8.0
AU05	6.1
AU06+12	-6.0
AU04	4.9

Name	Coefficients
AU28	-373.6
Seq-similarity	-78.4
AU45	17.1
Pitch	14.3
AU15	-13.7
AU01	12.0
AU17	-11.9
AU26	-11.8
AU09	9.1
AU25	-8.3
Energy	-7.6
AU02	7.5
AU20	-7.4
AU07	7.1
Yaw	-6.2
F2/F1 Mean	6.0
Smile Ratio	-4.7
AU12	-4.7
AU23	-3.4
AU06+12	-3.3

的コミュニケーションスキルスコア予測の結果においても確認できた。したがって、私たちの 2 番目の仮説「高い社会的スキルを持つ人々による話は意味的に一貫しているべきである」を補強する結果が確認できた。その一方で、上位に語彙数が入っていないことから、最初の「社会的スキルの高い人は語彙数が多い」に対しては否定的な結果となった。

より深く実験結果について理解するために、これらの係数が示す結果についての解釈を試みた。ここで、SRS-2 はもともと精神疾患の重症度を評価するために開発されたものであるため、SRS-2 スコアが低いほど社会的スキルが優れていることを示しているため注意が必要である。モデル係数上位 20 の中で 17 の特徴量が画像由来、2 つの特徴量が音声由来、1 つの特徴量がテキスト由来であった。それら特徴量の中で、感覚的に理解しにくい「F2/F1 の平均」について詳細に説明する。

籠宮は、/i/、/e/、/u/の長い母音は短い母音よりも F2/F1 が低いことを報告している [31]。その反面、/a/や/o/の長い母音は短い母音よりも F2/F1 が高い傾向があるが、前者が数的に支配的であるため（5 母音のうち 3 つ）、後者は無視できると考えられる。したがって、短母音・長母音と F2/F1 の関係についての籠宮の理論と、私たちの結果より、ソーシャルスキルの高い人は発話中に長い母音を単母音よりも多く使用する可能性が示唆された。この結果は「日本語の会話において長母音をあまり使わないで話す場合、事務的で友好的ではないドライな印象を感じることがある」という経験的な直感に一致していると考えられる。

以上のような考察から、社会的スキルを向上させるためには、以下の点が重要であるとの解釈を得た。

- 意味的一貫性を保って話す (Seq-similarity)
- 大きい声で話しすぎない (Energy)
- 長母音をより多く使う (F2/F1 Mean)

総合スコア推定と類似の傾向が社会的コミュニケーションスコア推定時にも確認された（表 2 の右側）。これは、社会的コミュニケーションスコアが SRS-2 総合スコアの治療下位尺度であり、相関係数 0.92 と互いに高い相関があったことに起因するものであると考えられる。

また、今回の実験では社会的コミュニケーションスキルスコア予測時よりも総合スコア予測時のほうが総じて悪い結果となった。これは総合スコアが社会的コミュニケーションスコアより生活習慣などの複雑なユーザー特性を含むため、社会的コミュニケーションより予測が難しいことに起因する。全体的なスコアをより正確に予測するには、ソーシャルスキルに関連した重要な特徴量を今よりも多く加える必要があると考えられる。

7. 結　　論

本研究では、新しい組み合わせのマルチモーダル特徴量を使用し、SRS-2 総合スコアとその治療下位尺度である社会的コミュニケーションスコアを予測するモデルを提案した。提案モデルでは、総合スコア予測で 0.52、社会的コミュニケーションスコア予測で 0.65 の相関係数を達成した。

今回の実験によってソーシャルスキルのレベルの予測におけるマルチモーダル特徴量の有効性が確認されたが、いくつか改善すべき点を残すこととなった。第一に、データ収集の観点から、仮想エージェントの効果を調査する必要がある。今回の実験で用いたデータは仮想エージェントをセラピストに見立ててデータ収集を行ったものであるが、人間のセラピストが代わりにセッションを行った場合に結果が変わる可能性がある。第二に、この実験の参加者はすべて健康な被験者であったため、実際に精神疾患を抱えている人々を対象として提案手法をテストする必要がある。第三に、結果に対する各特徴量の重要度および解釈についてさらなる検討が必要である。

8. 謝　　辞

本研究は、CREST 戦略的創造研究推進事業 (JPMJCR19A5) および JSPS 科研費 (JP17H06101 および JP18K11437) の支援を受けて行われたものである。

文　　献

[1] A.S. Bellack, K.T. Mueser, S. Gingerich, and J. Agresta, Social Skills Training for Schizophrenia: A Step-by-Step Guide, 2 edition, Guilford Press, 2004.
[2] “Relationships & social skills”.
[3] A. Salter, Conditioned reflex therapy, Creative Age Press, 1949.
[4] J. Wolpe, Psychotherapy by reciprocal inhibition, Stanford University Press, 1958.
[5] A. Bandura, Principles of behavior modification, Holt, Rinehart and Winston, 1969.
[6] K.T. Mueser and A.S. Bellack, “Social skills training: Alive and well?,” Journal of Mental Health, pp.549–552, 2007.
[7] J. Hunt and D. Eisenberg, “Mental health problems and help-seeking behavior among college students,” Journal of Adolescent Health, vol.46, no.1, pp.3–10, 2010.

[8] B. etal, “The mental health and well-being of ontario students, 1991-2015: Detailed osduhs findings,” Technical Report 43, Toronto: Centre for Addiction and Mental Health, 2016.

[9] 一般社団法人 SST 普及協会，“認定講師”． Accessed October 5, 2020)． http://www.jasst.net/top/lecturer

[10] M. John N. Constantino and P. Christian P. Gruber, Social Responsiveness Scale, Second Edition (SRS-2)Back, Western Psychological Services, 2012.

[11] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, and L.-P. Morency, “Simsensei kiosk: A virtual human interviewer for healthcare decision support,” 13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014, vol.2, pp.1061–1068, 01 2014.

[12] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, “Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders,” PLOS ONE, vol.12, no.8, pp.1–15, 08 2017. https://doi.org/10.1371/journal.pone.0182151

[13] H. Tanaka, H. Iwasaka, H. Negoro, and S. Nakamura, “Analysis of conversational listening skills toward agent-based social skills training,” Journal on Multimodal User Interfaces, vol.14, no.1, pp.73–82, March 2020. https://doi.org/10.1007/s12193-019-00313-y

[14] H. Tanaka, S. Sakriani, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura, “Teaching social communication skills through human-agent interaction,” ACM Trans. Interact. Intell. Syst., vol.6, no.2, p.18 with 26 pages, Aug. 2016. https://doi.org/10.1145/2937757

[15] I. Naim, M.I. Tanveer, D. Gildea, and M.E. Hoque, “Automated analysis and prediction of job interview performance,” IEEE Transactions on Affective Computing, vol.9, no.2, pp.191–204, April 2018.

[16] V. vanHeuven, “Praat, a system for doing phonetics by computer,” Glot International, vol.5, no.9/10, pp.341–345, 2001.

[17] M. Courgeon, J.-C. Martin, B. Mutlu, R. Picard, and M. Hoque, “Mach: My automated conversation coach,” UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp.697–706, Dec. 2014.

[18] M. Rouhizadeh, E. Prud’hommeaux, B. Roark, and J. vanSanten, “Distributional semantic models for the evaluation of disordered language,” Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.709–714, Association for Computational Linguistics, Atlanta, Georgia, June 2013. https://www.aclweb.org/anthology/N13-1084

[19] J. Pennebaker, M. Francis, and R. Booth, “Linguistic inquiry and word count:liwc2007,” 01 2007. Available at https://liwc.wpengine.com/.

[20] T.H. Toshinori Sato and M. Okumura, “Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese),” Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing, pp.NLP2017–B6–1, The Association for Natural Language Processing, 2017.

[21] T.H. Toshinori Sato and M. Okumura, “Operation of a word segmentation dictionary generation system called neologd (in japanese),” Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL), pp.NL–229–15, Information Processing Society of Japan, 2016.

[22] S. Toshinori, “Neologism dictionary based on the language resources on the web for mecab,” 2015. https://github.com/neologd/mecab-ipadic-neologd

[23] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230–237, Association for Computational Linguistics, Barcelona, Spain, July 2004. https://www.aclweb.org/anthology/W04-3230

[24] L. Bertola, N.B. Mota, M. Copelli, T. Rivero, B.S. Diniz, M.A. Romano-Silva, S. Ribeiro, and L.F. Malloy-Diniz, “Graph analysis of verbal fluency test discriminate between patients with alzheimer’s disease, mild cognitive impairment and normal elderly controls,” Frontiers in aging neuroscience, vol.6, pp.185–185, July 2014. 25120480[pmid]. https://pubmed.ncbi.nlm.nih.gov/25120480

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp.4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://www.aclweb.org/anthology/N19-1423

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, “Huggingface’s transformers: State-of-the-art natural language processing,” 2019.

[27] G. Bradski, “The OpenCV Library,” 2000.

[28] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L. Morency, “Openface 2.0: Facial behavior analysis toolkit,” 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp.59–66, 2018.

[29] P. Ekman and W. Friesen, Facial action coding system: A technique for the measurement of facial movement, Consulting Psychologists Press, 1978.

[30] P. Ekman and W. Friesen, “Rationale and reliability for emfacs coders” unpublished.

[31] T. Kagomiya, “Articulatory positions of japanese vowels as a function of duration computed from a large-scale spontaneous speech corpus,” 2015.