社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

マルチモーダル情報を伴った自動対話訓練システム

田中　宏季† 　サクリアニ　サクティ† 　グラム　ニュービッグ† 　根來　秀樹†† 　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
†† 奈良教育大学　特別支援教育研究センター　〒 630–8528 奈良県奈良市高畑町
E-mail: †{hiroki-tan,ssakti,neubig,s-nakamura}@is.naist.jp, ††{gorosan,hiwasaka}@nara-edu.ac.jp

あらまし　社会的コミュニケーションを苦手とする人々は，一方でコンピュータなどに優れた能力を発揮する傾向があり，コンピュータを使用したソーシャルスキル訓練が注目されている．我々は自動ソーシャルスキルトレーナと題して，ソーシャルスキルトレーニングの過程を人間と対話エージェントの会話によって自動化する研究を進めている．これまでの自動ソーシャルスキルトレーナは音声および言語情報のみしか考慮していないという問題があった．実際のソーシャルスキルトレーニングでは，画像や姿勢など視覚情報も含めたフィードバックを行っていることから，これらのマルチモーダルな情報も組み込む必要があると考えられる．本稿では既存の基本訓練モデルの枠組みに従い，マルチモーダル情報を含めたシステムの構築を目指す．具体的には，ユーザへのフィードバックに，「笑顔の頻度」と「顔の向き」の特徴量を追加した．ユーザがシステムに向かって話を伝える評価実験によって，ソーシャルスキルと画像情報との関連性が確認された．
キーワード　音声対話システム，マルチモーダル，社会性

Automated Commucation Training System with Multimodal Information

Hiroki TANAKA†, Sakti SAKRIANI†, Graham NEUBIG†, Hideki NEGORO††, Hidemi IWASAKA††, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology　Takayama-cho 8916–5, Ikoma-shi, Nara, 630–0192 Japan
†† Center for Special Needs Education, Nara University of Education　Takabatake-cho, Nara-shi, 630–8528 Japan
E-mail: †{hiroki-tan,ssakti,neubig,s-nakamura}@is.naist.jp, ††{gorosan,hiwasaka}@nara-edu.ac.jp

Abstract　People with social communication difficulties tend to have superior skills using computers, and computer-based social skills training systems are flourishing. Social skills training is a well-established method to decrease human anxiety and discomfort in social interaction, and obtain appropriate skills. We have attempted to automate the process of social skills training by developing a spoken dialogue system named automated social skills trainer, which provides the social skills training through human computer interaction. While the previous system considered only acoustic and linguistic features, human social skills trainers take into account visual features (e.g. facial expression, posture). In this paper, we extend the automated social skills trainer by adding multimodal information. Specifically, we extracted features regarding the ratio of smiling, yaw, and pitch. An experimental evaluation measured the relationship between social skills and extracted visual features.
Key words　Spoken Dialogue System, Multimodal, Social Interaction

1. はじめに

多くの人々が対人関係、面接、プレゼンテーションなどの社会的コミュニケーションに対して不安や困難を抱えている [1]．一方で，これらのスキルを苦手とする人は，コンピュータなど非社会的なものに対して高い能力と興味を示し，人間と直接関

わるよりもコンピュータの方が不安が少ないことが多い [2]．これらのことからコンピュータアプリケーションやロボットを，社会的コミュニケーションの訓練に利用する研究が注目されてきている [3], [4]．

ソーシャルスキルトレーニング（SST）は幅広く社会的コミュニケーションを苦手としている人々に適用されている方法であり， 医療機関や各種の社会復帰施設，作業所，矯正施設，学校，職場などで人間のトレーナにより実施されている [5]．SST の全体もしくは一部分をコンピュータで自動化することができると，希望者がいつでもどこでも、SST を受けることができると考えられる．そこで，我々は音声対話システムを用いて SST の自動化を行う研究を進めており， コンピュータを用いた従来の SST を模倣した「自動ソーシャルスキルトレーナ」を提案した [6]（図 1）．システムは人間の行動のビデオモデリング，リアルタイム行動認識およびフィードバックを含んでおり，具体的に，1) 現実的な状況において実ユーザのソーシャルスキルトレーニングを目指した，2) 改善すべき行動について視覚的にフィードバックを行った，3) システム自身がユーザの発話や行動を認識し，フィードバックや頷きを行った．また実験により，提案システムが従来の本を読むトレーニングと比較して，ソーシャルスキルの学習に有効であることを報告した．

一方で，これまでの自動ソーシャルスキルトレーナは音声および言語情報のみしか考慮していないという問題があった．実際の SST では，画像や姿勢など視覚情報も含めたフィードバックを行っていることから [7]，これらのマルチモーダルな情報も組み込む必要があると考えられる．また，画像を中心とした自動フィードバックの研究も存在していることから [3]，本稿では既存の SST の枠組みに従い，さらにマルチモーダル情報を含めたシステムの構築を目指す．具体的には，ユーザへのフィードバックに,「笑顔の頻度」と「顔の向き」の特徴量を追加する．実験により，これらの画像特徴量がどの程度ソーシャルスキルと関連するのかについて調査を行う．

2. SST と自動ソーシャルスキルトレーナ

2.1 SST の基本訓練モデル

ここで SST [8] について説明を行う．一般社団法人 SST 普及協会によると，“SST とは，Social Skills Training の略で,「社会生活技能訓練」や「生活技能訓練」などと呼ばれていおり，小児の分野では「社会的スキル訓練」，教育の分野では「スキル教育」とも呼ばれている．現在では、医療機関や各種の社会復帰施設，作業所，矯正施設，学校，職場などさまざまな施設や場面で実践されている．家庭や職場訪問など地域生活の現場での支援も行われている．社会生活の上で様々な困難を抱えるたくさんの人たちの自己対処能力を高め（エンパワメント），自立を支援するためにこの方法が広く活用されることが期待されている” と説明されている (注1)．

SST ではリバーマンがまとめた基本訓練モデル（ステップ方式：図 1 参照）が，構造化された手順として利用されることが多い [7]．SST の基本訓練モデルは，課題設定，モデリング，ロールプレイ，フィードバック，正の強化，宿題により構成される．以下，基本訓練モデルの各ステップについて簡単に述べる．

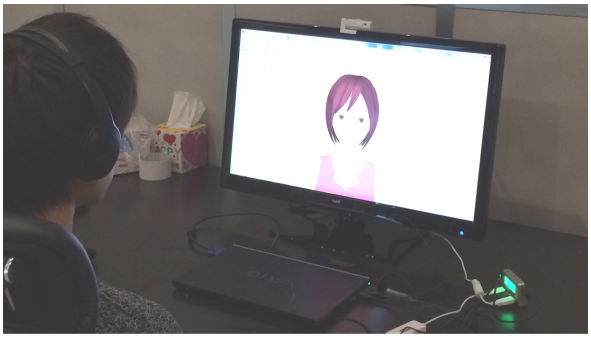

図 1　自動ソーシャルスキルトレーナによる SST.

表 1　SST の基本訓練モデル

① 前回の宿題の報告
② 課題の設定
③ モデリング
④ ロールプレイ
⑤ 正のフィードバック
⑥ 宿題の設定

まず初めに課題を設定する．課題を決めるために，トレーナと参加者がボトムアップ式で話し合う場合と，トレーナが決定する場合がある．課題が決まると，それに伴うゴールがトレーナによって設定される．課題の例としては，プレゼンテーション，自己紹介，要求の断り方，気持ちの表現など，実際の場面での課題が挙げられる．

次に参加者が課題を行う前に，トレーナがモデルとして対象行動の見本を示す．参加者はそれを観察することによって，対象のスキルについてどのように行動すれば良いのかを学習する．例としては，トレーナが適切な言語および非言語情報を使用して上手に話をする見本を提示することなどが挙げられる．この時，モデルのどの点を見るかについて参加者に注意を与えたり，モデルの観察によって得たものを言語化したりイメージ化することによって保持することも行われる．

次に参加者が課題のロールプレイを行う．例として，参加者がトレーナに向かって，上手に話を伝えるロールプレイを行う．その際，トレーナは参加者のスキルを主観的に観察する．具体的には，ことばの内容，声の大きさ，表情，姿勢，ジェスチャーなど言語情報および非言語情報などに着目する．

ロールプレイの後にトレーナは参加者にフィードバックを行う．フィードバックは参加者が自身の強みを知り，モチベーションを強化するのに有効だと言える．例えば先ほどの例だと，トレーナは参加者に「適切な声の大きさでとても良く伝わりました」と伝える．フィードバックの際，トレーナは参加者に正の強化を与える必要がある．ASD などの発達障害をもつ人々は社会的インタラクションに対する成功体験が少ないため,「ほめる」ということは重要である．またこの際に修正のフィードバックも与える．

最後にトレーナは参加者に対して，学んだスキルを実環境で

(注1)：http://www.jasst.net/

使うことを宿題として与える．これによって学んだスキルを普段の生活で自然に汎化できるようになることを目指す．例えば，「今日学んだスキルで友達や親に話を伝えてみてください」と宿題を出す．宿題の確認は次回のセッションのはじめに行われることが一般的である．

以上の過程を繰り返すことによって，SST は適切なソーシャルスキルを獲得するのに有用なプログラムとなる．

2.2 自動ソーシャルスキルトレーナ

ここでは，SST の基本訓練モデルに従った自動ソーシャルスキルトレーナの概要について述べる．実際に人間が行う SST は各ステップに対しての拡張性が高いが，今回のシステムでは以下にまとめるような制約を加えてシステムを構築している．各ステップの改良については今後の課題としていく．

- **課題設定:** 課題設定では，上手な話しの伝え方（ナラティブ）に焦点を当てる．話しの伝え方は他の課題の基本となるスキルであり，プレゼンテーションや面接などの際にも特に役立つスキルである [9]．また，ナラティブは ASD 児と定型発達児を特定することにも役立つことが知られており [10]，自分の体験を話すことによる種々の効果についても近年注目されてきている (注2)．課題の目標とする所として，システムが「このアプリケーションは，上手にお話する練習をするものです．トレーニングをした後には，他の人に自分の体験を伝えるのがもっと楽しくなります」と伝える．
- **モデリング:** 個別 SST の欠点の 1 つとして，他の参加者をモデルとすることができないという点が挙げられるが，自動ソーシャルスキルトレーナで他者のビデオモデリングを代替案とする [11]．ユーザが「上手なお手本を見せてください」とシステムに発話すると，ユーザは収録されたビデオモデルを見ることができる．モデルとなる人物は，他者と比較して高いスキルがある人物としている．ユーザは動画を視聴し，それを手本とする．
- **ロールプレイ:** ロールプレイではユーザがアバターに向かって「最近あった楽しかった出来事」を伝えることによって行われる．話は 1 分間行われ，アバターはユーザの発話に対して頷き，音声言語特徴量を抽出する．本研究では，ASD 児と定型発達児で差が生じた音声言語特徴量（F0 の変動係数，パワー，声質，ポーズ，1 分間の単語数（WPM），6 文字以上の単語の割合，フィラーの割合 [10]）に加え，新たに「笑顔の頻度」「顔の向き」の画像特徴量を抽出する．
- **正のフィードバック:** ロールプレイが終わると，システムは抽出した特徴量に従ってフィードバックを表示する．フィードバックは，コメント，ユーザのビデオ，モデルとの比較，総合スコアという 4 つの項目を含む．これによりユーザは客観的に自身の強みを確認することができる．フィードバックは単純にスコアを出すだけではなく，話しの伝え方で良かった点、修正点およびそのコメントを提示する．
- **宿題:** 宿題として，アバターはユーザに「まわりの人にお話を伝えてみてください．それで，どうだったか教えてくださいね」と伝える．本研究では複数セッションの SST を実施していないため，宿題のフォロー・確認に関しては今後の課題となる．

(注2)：http://otoemojite.com/

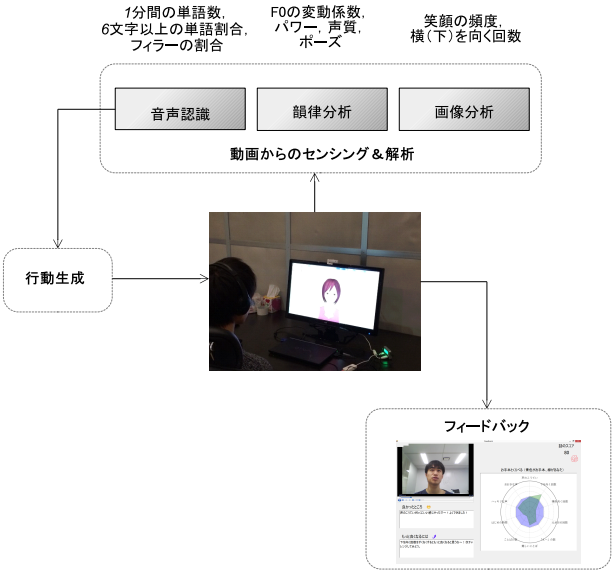

図 2　ロールプレイとフィードバックの要素．

3. マルチモーダルシステムの詳細

本節では自動ソーシャルスキルトレーナのマルチモーダル化の詳細について，ロールプレイとフィードバックに焦点を当てて述べる．システムはロールプレイ中にユーザの言語・音声・画像情報を収録する．ロールプレイ中に収録したこれらの情報に基づいてフィードバックの生成を行う．図 2 に示す通り，本研究で技術的な軸となるロールプレイ，フィードバックに関しては次の 3 つの要素から成り立つ：行動生成，動画からのセンシングと解析，フィードバック．

3.1 行 動 生 成

自動ソーシャルスキルトレーナは音声認識，対話制御，音声合成および行動生成を行う MMDAgent (注3) を用いる．ユーザは全て音声のみでアバターと対話していくことが可能である．アバターはユーザの興味を持続させるために人間らしい振る舞いをする．アバターは 3 秒毎にまばたきを行い，またユーザの発話を認識した数秒後，頷きの動作を行う．

3.2 動画からの各種センシングと画像解析

言語に関連する特徴量を計算するため，Julius (注4) ディクテーションキットによる音声認識を使用した．日本語の品詞解析には，Mecab (注5) を使用した．音声に関連する特徴量に関しては，Snack sound toolkit (注6) を用いて抽出した．

ここで画像解析について述べる．まずユーザが話している様子を内蔵カメラを用いて撮影した．その撮影した動画に対して

(注3)：http://www.mmdagent.jp/
(注4)：http://julius.sourceforge.jp/index.php
(注5)：https://code.google.com/p/mecab/
(注6)：http://www.speech.kth.se/snack

リアルタイムで face tracker(注7) を使用して顔の特徴点抽出を行った（図 3）．全 66 の特徴点から，目の外側と眉の外側の間の縦の距離， 目の内側と眉の内側の間の縦の距離, 唇の外側の縦の距離, 唇の内側の縦の距離, 目の開く量, 唇の両端の横の距離，を特徴量として算出した．これらの特徴量を使用して The Japanese Female Facial Expression (JAFFE) データベース [12] により笑顔の学習を行った．JAFFE データベースは日本人女性 10 名の各種感情に対しての表情データであり，中立，幸福，悲しみ，怒り，嫌悪の表情ラベルがそれぞれ付与されている．このうち，笑顔（幸福）のサンプルとして 31 個，中立のサンプルとして 30 個の表情が登録されている．線形カーネルの SVM により，笑顔と中立の表情のモデル学習を行った．JAFFE データベースから leave-one-out の交差検証を行った結果，笑顔の判定について適合率，再現率，F 値はそれぞれ .91, .97, .94 となった．本稿では上記のモデルを笑顔判定に使用している．動画に対しては，各フレームが笑顔に属するか中立に属するかを予測し，全フレームの内で笑顔と判定された割合を笑顔の頻度としている．検証として，NOCOA+のデータベース [13] の 4 名の男性における冷笑的な動画と友好的な動画に対して笑顔の頻度を抽出し，Student の t 検定を行ったところ，友好的な動画では冷笑的な動画よりも有意傾向をもって笑顔の頻度が多いことが確認された．

図 3 に示す特徴に加え，我々は顔の向き（頭部姿勢）の特徴量（ヨーおよびピッチ）を回転行列の対応する箇所により抽出した．ここで，ヨーは横方向，ピッチは縦方向の頭部姿勢に対応する特徴量である．ヨーは 0（正面）からどの程度離れたかをヨーの絶対値により算出し，その全フレームの平均値をとった．ピッチに関しては，下を向くか（正の値）上を向くか（負の値）が重要であると考えるため，全フレームでの平均値を算出した．なお，本課題においては，胸部より下の画像収録は行っていないため，ジェスチャーなどは考慮していない．

以上により抽出された特徴量について以下にまとめる：1) F0 の変動係数：100Hz 以上の F0 に関する変動係数を抽出した．個人差や性別があるため F0 に関しての最大値，最小値，平均値などの統計量の抽出は行わなかった，2) パワー：パワー値の平均を抽出した，3) 声質：スペクトル傾斜について，第一倍音と第三フォルマントの差の特徴量を抽出した，4) ポーズ：アバターの発話終了からユーザの発話開始までの時間を抽出した，5) WPM：ユーザが 1 分間発話をするため，その間の単語数を抽出した，6) 6 文字以上の単語割合：全発話から 6 文字以上の単語を使用していた割合を抽出した，7) フィラーの割合：Mecab の出力により,「えー」や「ああ」などのフィラーの割合を抽出した，8) 笑顔の頻度：全フレームに対する笑顔の割合を抽出した，9) 横を向く回数：ヨーの絶対値の平均を抽出した，10) 下を向く頻度：ピッチの平均値を抽出した．

3.3 フィードバック

抽出した特長量により，ユーザのナラティブスキルに関してフィードバックを行う（図 4）．フィードバックは以下のものを含んでいる．

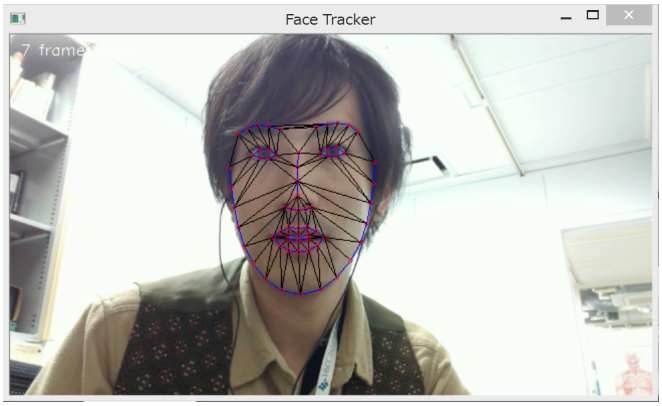

図 3　face tracker による画像からの顔特徴点抽出．

- **ユーザの動画:** ユーザは録画された自身の音声と動画を視聴することができる．これにより，ユーザはフィードバックとともに自身の様子を繰り返し確認することができる．
- **総合スコア:** システムは予測した総合スコアを表示する．これはユーザがより高得点を目指す動機付けとなる．総合スコアは 0 から 100 の間の値で，特徴量から重回帰モデルで予測する．
- **モデルとの比較:** Z 値によって，現在のナラティブにより抽出された特徴量が，モデルの平均的な値と，どの程度ずれているのかを表示する．レーダーチャートの形式でモデルとのズレを提示することにより，ユーザが視覚的に理解し易いようにしている．ユーザはモデルを手本としてナラティブを行うように指示される．
- **良かった点および修正点:** ユーザの動機付けとなるように，システムは正のコメントおよび修正のコメントを生成する．正のコメントはモデルと最も近い値の特徴量を元に生成される．修正点はモデルから中央値程度で離れている特徴量を元に生成される．

4. 実験：画像情報とナラティブスキルの関係性

画像情報を伴った自動ソーシャルスキルトレーナの有効性を確認するため，我々は 2 つの実験を行った．実験 1 においては，以下の質問について調査した．

1) ナラティブスキルは画像情報と関係があるか？
2) 画像特徴量はナラティブスキルを予測するのに有効か？

4.1 手　続　き

我々は 19 名の大学院生（男性 16 名，女性 3 名）によるデータ [6] を用いて実験を行った．このデータは自動ソーシャルスキルトレーナを使用し，ウェブカメラとヘッドセットにより動画と音声をそれぞれ収録している．ソーシャルスキルの高い 2 名（男性と女性）の評価者が動画を見て各種質問に答えている．質問は話の全体的なスキル，関連要因，および音声言語情報について尋ねる項目で構成されている．1（とても良くない）から 7（とても良い）までの範囲で評価されている．本稿ではこの内，話の全体的なスキルの評価結果を使用した．なお，話の

(注7)：https://github.com/kylemcdonald/FaceTracker

図 4　自動ソーシャルスキルトレーナのフィードバック画面.

全体的なスキルに関してのカッパ係数は 0.58 である.

4.2　抽出した特徴量とナラティブスキル

我々は，評価値から上位 5 名を選びモデルと設定し（このモデルはビデオモデリングの動画に組み込まれている），モデルとモデル以外の被験者において有意な差がある特徴量を Student の t 検定により分析した．結果として，パワー（モデルの方が有意に大きい），WPM（モデルの方が有意に多い），6 文字以上の単語割合（モデルの方が有意に少ない）について差が見られた ($p < .05$)．今回抽出した笑顔の頻度は，図 5 に示すように，モデルの方でより笑顔を表出することが確認された．なお顔の横向き下向きに関しては，図 6 に示すように，有意な差は生じなかった．我々はこれらの有意差の生じた特徴量を用いて，全体的なスコアを予測する重回帰モデルを作成した．重回帰モデルによる Leave-one-speaker-out 交差検証の予測値と，実際の観測値との相関係数は，音声と言語特徴量のみで 0.51 であり，笑顔の頻度特徴量を追加すると 0.54 （$p < .05$）に向上した (この重回帰モデルはフィードバックの総合スコアに組み込まれている)．

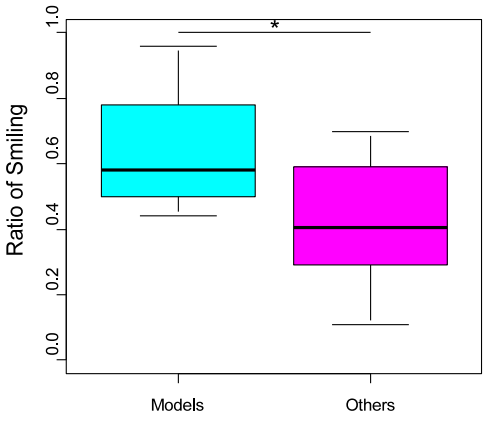

図 5　笑顔の頻度とナラティブスキルの関係 (*: $p < .05$).

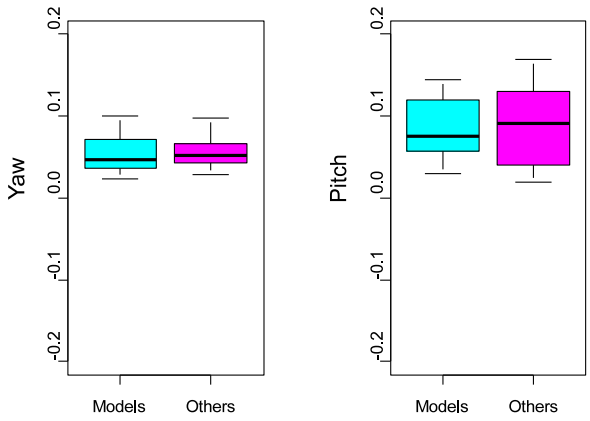

図 6　顔の横（縦）向きとナラティブスキルの関係.

5.　ま　と　め

我々は対話システムによって従来の SST を模倣する自動ソーシャルスキルトレーナを開発した．自動ソーシャルスキルトレーナは課題設定，モデリング，ロールプレイ，フィードバック，正の強化，宿題を含んでいる．本研究では，従来の音声言語のみに基づいてフィードバックを行うものから，画像情報も含めたシステムを構築した．我々は評価実験を通して，ナラティブスキルと画像特徴量との関連性を確認した．

今後は，提案システムを使用した SST の実験を行い，音声言語情報と画像を含めたトレーニング効果の違いを検証していく．また，SST の基本訓練モデルの各モジュールについて検討し，システムの改良を行う．

6. 謝　　辞

本研究は，JSPS 科研費 26540117 の助成を受けて行われたものである．

文　　献

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC, 2013.

[2] Bishop, J. The Internet for educating individuals with social impairments. *J. of Computer Assisted Learning 19*, 546–556, 2003.

[3] Hoque, E., Courgeon, M., Mutlu, B., Martin, C., Picard W. MACH: my automated conversation coach. *Proc. 15th Conf. on UbiComp*, 697-706, 2013.

[4] Ricks, J., Colton, B. Trends and considerations in robot-assisted autism therapy. *IEEE International Conference on Robotics and Automation (ICRA)*, 4354-4359, 2010.

[5] Bauminger, N. The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: Intervention outcomes. *J. Autism and Developmental Disorders 32*, 283–298, 2002.

[6] Tanaka, H., Sakriani, S., Neubig, G., Negoro, H., Iwasaka, H., Toda, T., Nakamura, S. Automated Social Skills Trainer. *International Conference on Intelligent User Interfaces*, 17-27, 2015.

[7] Bellack, A. S. Social skills training for schizophrenia: A step-by-step guide. Guilford Press. 2004.

[8] Wallace, C. J., Nelson, C. J., Liberman, R. P., Aitchison, R. A., Lukoff, D., Elder, J. P., Ferris, C. A review and critique of social skills training with schizophrenic patients. *Schizophr Bull 6*, 42-63, 1980.

[9] Davis, M., Dautenhahn, K., Nehaniv, C., Powell, S. Towards an Interactive System Facilitating Therapeutic Narrative Elicitation. *Proc. 3rd Conf. on NILE*, 2004.

[10] Tanaka, H., Sakriani, S., Neubig, G., Toda, T., Nakamura, S. Linguistic and Acoustic Features for Automatic Identification of Autism Spectrum Disorders in Children's Narrative. *ACL2014 Workshop on Computational Linguistics and Clinical Psychology*, 88-96, 2014.

[11] Essau, C. A., Olaya, B., Sasagawa, S., Pithia, J., Bray, D., Ollendick, T. H. Integrating video-feedback and cognitive preparation, social skills training and behavioural activation in a cognitive behavioural therapy in the treatment of childhood anxiety. *J. Affect Disorders 167*, 261-267, 2014.

[12] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J. Coding facial expressions with Gabor wavelets. *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 200-205, 1998.

[13] Tanaka, H., Sakriani, S., Neubig, G., Toda, T., Nakamura, S. NOCOA+: Multimodal Computer-Based Training for Social and Communication Skills. *IEICE Transactions on Information and Systems*, E98.D(8), 1536-1544, 2015