マルチモーダル特徴量を用いた教員志望者の授業スキル評価モデルの提案

Proposal of a Assessment Model of Student Teacher's Teaching Skill using Multimodal Features

福田 匡人 [1*] 黄 宏軒 [1,2,3] 桑原 和宏 [3] 西田 豊明 [1,2]

[1] 国立研究開発法人理化学研究所革新知能統合研究センター
[1] RIKEN Center for Advanced Intelligence Project
[2] 京都大学大学院情報学研究科
[2] Graduate School of Informatics, Kyoto University
[3] 立命館大学情報理工学部
[3] College of Information Science and Engineering, Ritsumeikan University

Abstract: 教員志望者が生徒としての CG キャラクタ (仮想生徒) とのインタラクションを通して，指導力を養う新たなプラットフォームの構築を進めている．自律システムによる授業のフィードバックの実現には，授業者の指導方法や生徒対応のリアルタイムな評価が求められる．本稿では，教員志望者の授業の意図，韻律情報およびジェスチャを含むマルチモーダル特徴量を用いた授業スキル評価モデルを提案する．また授業スキルの評価タスクにおける，有効なモダリティと学習アルゴリズムを検討する．

1 はじめに

教員志望者は，志望段階から実際の教育現場で発揮可能な授業スキル (Teaching Skill) の習得が求められている．授業スキルとは「言語的スキルと非言語的スキルを用いて，生徒集団に対して考えを伝える力」である [13]．一方で，教員志望者の指導訓練は教職課程における座学が主であり，生徒に対して指導訓練を実施する機会は多くない．ロールプレイ形式の授業シミュレーションには生徒役としての多くの協力者と，その指導訓練を評価する教育経験者が求められ，反復的な指導訓練は従来難しかった．

我々は，教員志望者 (授業者) の指導力を養う新たなプラットフォームとして仮想学級を用いた指導訓練システムを研究している [7] [6]．仮想学級とは生徒としての CG キャラクタ (仮想生徒) の集団で構築される学級である．教員志望者は実際の授業環境の代わりに，仮想生徒に対して指導を実施し，生徒とのインタラクションの中で自身の指導に関する改善点を検討する．

常に変化する指導状況において，システムが発話の意図や言動を認識し，それらの言動が指導として適切・効果的であるか評価することにより，仮想生徒はその状況ごとに適切な振る舞いを決定できる．仮想生徒とのインタラクションには，授業者の指導状況の把握は欠かせない．

システムによる指導評価の自動化には，センシング可能な授業者の言語・非言語情報を用いて授業スキルを評価することが求められる．一方，教員採用試験で利用される広義的な評価指標は存在するが，実際に評価する際の着眼点は豊富な経験を持つ教育経験者の暗黙知に依存する．

本研究では，模擬授業コーパスに対して教育専門家による評定を実施し，マルチモーダル情報を用いて，教員志望者の授業スキルを評価可能な授業スキル評価モデルを提案する．また授業スキルの評価というタスクにおいて，有効なモダリティとモデル構築のための学習アルゴリズムを検討する．予測モデルの学習アルゴリズムには Support Vector Machine (SVM) と Neural Network (NN) を採用する．先行研究で我々は，マルチモーダル特徴量を用いた SVM による授業者の授業意図の推定モデルを構築し，高い精度を示した [16]．よって，SVM が評価モデルに対しても有効であると考えられる．

一方で，授業の評価は個人の主観と経験に基づいており，授業評価に必要な特徴量を定義することは困難であると想定される．また，全てのセンサ情報を活用す

*連絡先：国立研究開発法人理化学研究所革新知能統合研究センター 〒 606-8501 京都市左京区吉田本町 京都大学構内総合研究 15 号館 108 号室 E-mail: hhhuang@acm.org

ることは，モデルの次元数とノイズの関係から現実的でない．各モダリティの潜在的な特徴を抽出することで，授業スキルの評価が可能であると考え，SVM の比較対象として LSTM 層を含む Deep Neural Network (DNN) を採用する．

2 関連研究

授業は複数の生徒とのインタラクションの場であり，起こりうるシチュエーションは多様である．その状況に応じた適切な振る舞いを自律エージェントとしての仮想生徒が実行することは困難である．よって，既存研究は各生徒が人間の外部入力により制御される Wizard of Oz (WOZ) 法の基盤としている．

Breaking Bad Behaviors[10][11] は VR 環境下で指導訓練が実施可能なシステムである．仮想学級には 24 人の仮想生徒が存在する．操作者は仮想生徒全体の Disruption レベルを制御し，仮想生徒に対して問題行為を実行させることができる．TeachLive[2] はディスプレイに投影される 5 人の仮想生徒とのインタラクションが可能な指導訓練システムである．

一方で，WOZ 法では，状況に応じた適切な仮想生徒の制御が可能だが，指導評価が可能な教育経験者や仮想生徒を制御する操作者が求められる．自律システムとしての仮想学級環境の構築には，WOZ 法において人手で実施されている指導評価の自動化は欠かせない．

インタラクションのリアルタイムな評価手法に関する研究は様々な分野で行われ，言語・非言語情報からなるマルチモーダルなセンシング情報がインタラクションの評価に有用であることが示されている．パブリックスピーキングの分野では顔情報，ジェスチャ，韻律を含むマルチモーダル情報を用いた話者の自動評価が提案されている [4][12]．栗原ら [14] はプレゼンテーションリハーサルの自動評価を提案した．指導訓練を行う発表者の音声および振る舞いを分析し，話速度，声の抑揚，聴衆とのアイコンタクトの度合いなどの指標をリアルタイムに発表者にフィードバックすることの有用性を示した．岡田ら [15] は対話に参加する人間のコミュニケーション能力を，参加者の言語と非言語行動の情報から推定できることを確認した．また，人事採用試験経験者による評定を基にコミュニケーション能力を付与し， 客観的かつ, 信頼性の高い評定値を得ている．

一方で，評定基準は個人の経験や主観によって異なる．インタラクションの自動評価には， 信頼性の高い評定値を獲得することに加えて，その評定値を出力可能な特徴量選択とアルゴリズムが求められる．岡田ら [15] はコミュニケーション能力の推定において，意味的解釈が可能なマルチモーダル特徴量を用いて線形の SVM による推定モデルを提案した．一方で Liu らら [9] は，CNN（畳み込みニューラルネットワーク）と RNN（リカレントニューラルネットワーク）を組み合わせ，対話行為を分類する際の文脈情報を自動的に抽出した．

3 模擬授業データコーパス

本研究は，仮想学級のプロトタイプシステムを用いた模擬授業形式の実験にて収録された 9 人の学生による模擬授業コーパスを使用する．実験協力者は教職課程の履修もしくは塾講師のいずれかの指導経験をもつ．授業科目は，板書が頻繁に観察されることに期待し，指導のテーマは円または三角形の面積の計算と設定した．実験では，各参加者は 5 〜 10 分の授業を実施した．各模擬授業に対して複数人の評価を獲得するため，それらの模擬授業の収録動画を前半と後半に分割し, 9 人 × 2 の計 18 本の動画とした．これらの模擬授業にたいして，授業スキルを評価可能な評価項目の定義に従って，教育経験者による授業スキルの評定を実施した．

3.1 授業スキル評価項目の定義

授業者を評価するための指標は自治体・評価者ごとに異なる．一方で，システムが識別可能なセンシング情報から評価値を予測するには，自治体ごとに具体化された評価指標でなく，一般化された授業評価の指標が求められる．

各自治体の教育委員会は, 教員採用試験において, 受験者を評価するにあたって, 理想とする教師像というものをあらかじめ提示している．全国 66 自治体 (47 都道府県 +19 市教育委員会) の理想とする教師像のうち過半数を超え採用されている項目は「教育の専門家としての確かな力量」「教職に対する強い情熱」「人間性や信頼」である [17]．これらの指標から，総合評価を含む以下の 6 つを評価項目と定義する．

Enthusiasm (Enth): 児童生徒に情熱, 責任感, 愛情が伝わるものであり, 生徒がこれからの授業展開に期待を寄せられるか

Attentiveness (Atte): 児童生徒一人一人の目線に立って, 理解しようとしているものであり, 生徒が授業者に安心感を抱くことができるか

Dominance (Domi): 学級という集団をまとめ, 教師としての威厳やリーダーシップが感じられるものであるか

Attractiveness (Attr): 生徒の興味関心を引き付け, 工夫されたものであり, 児童生徒が授業に興味を持ち, 授業の面白さを感じ取れるか

Knowledge Propagation (KnPr): 授業における知識伝達の手段として有効であり, 生徒が知識を身につけることができたと実感できるか

Overall (AlEv): 授業者の総合評価

3.2 教育経験者による評定

20 年以上の指導経験をもつ高等学校教諭 2 名がそれぞれ前半もしくは後半の 9 本の模擬授業コーパスに対し評定を行った．授業スキルの評定区間を評価者間で統一すべく，意図検出が可能であり，意味的な解釈が可能と推測される 200ms 以上の無音区間を区切り位置とした．全ての発話区間および非発話区間に対して最低を 1，最大を 10 とする 10 段階で評定した．

各評価者の評定結果を表 1 に示す．評価者をそれぞれ E1,E2 と略称し，評価項目ごとの平均得点と標準偏差を示す．

表 1: 教育経験者による評価結果の概要

	Enth	Atte	Domi	Attr	KnPr	AlEv
E1Ave	5.44	4.8	5.32	5.15	4.39	5.20
E2Ave	4.20	4.45	4.33	4.30	4.62	4.44
E1S.D.	2.16	1.98	2.20	2.15	2.37	2.13
E2S.D.	2.57	2.67	2.69	2.65	2.81	2.68

評定結果のうち，1 点から 3 点を Low (L)，4 点から 7 点を Medium (M), 8 点から 10 点を High (H) とし，授業スキルを評価する．評価対象区間は全てで 2394 個である．評価項目ごとの分布を表 2 に示す．

表 2: 評価クラスの分布

	Enth	Atte	Domi	Attr	KnPr	AlEv
L	783	817	824	795	904	784
M	1307	1324	1186	1282	1166	1279
H	304	253	384	317	324	331

各評価クラスに共通し，授業スキルが高いと評価された H クラスの度数が小さいことがわかる．また，M クラスが最も多い．

4 マルチモーダル特徴量の抽出

教育経験者によって評定されたラベルを授業者のマルチモーダル特徴量を用いて検出する．本研究では，ジェスチャ情報，韻律情報，顔表情に加えて先行研究にて推定可能な授業意図の 4 つのモダリティを特徴量として定義する．

一方，SVM と DNN では入力する特徴量の性質が大きく異なる．SVM の場合は一般的に，入力はセンサー情報等のローレベルシグナルではなく，意味的解釈が可能な人為的に選択された特徴量が有効である．DNN の場合は学習の過程で特徴を抽出することから人為的な特徴量の選択は必須ではない．本章では授業スキル評価モデルの構築に使用する特徴量を SVM と DNN の 2 つの観点から述べる．

4.1 ジェスチャ情報

本研究では単眼カメラによる骨格検出アルゴリズム OpenPose[1] を用いてビデオコーパスと同じフレームレート（29.97fps）で，14 所の関節位置 2 次元座標（X，Y）を抽出した．人ごとの座標系を統一すべく，右大腿骨の座標と左大腿骨の中心を原点とし，直立状態の頭部方向を Y 軸正方向，右手を X 軸正方向とする．

ジェスチャ情報は時系列データであり，時間軸方向に処理されることが望ましい．DNN には 14 所の関節位置 2 次元座標（X，Y）を 2 秒間の時系列データとして入力する．

一方，SVM において時系列データの処理は困難である．よって本研究では，OpenPose で書き出された関節位置座標のなかで，授業中の運動量が多いと考えられる左右の手首，肘，肩の座標を活用する．中でも手首の関節座標は，OpenPose で取得可能な関節のうち，最も体から離れた箇所であるため，変化量を顕著に取得可能だと考えられる．また，両手首の関節間の距離を扱うことにより，資料を持ち説明している状態や，板書中の状態の識別に有効だと考えられる．肩関節の距離は体の回転を示す特徴量として期待できる．OpenPose の出力は 2 次元座標であることから，関節の回転は取得できない．一方で，体の向きが変わると，腰を原点とした場合，両肩の X 座標の距離は小さくなる．よって肩関節の X 座標の距離関係は，授業意図の識別に有効であると考えられる．最大値に加え，標準偏差は発話中に向きを変えているかどうかの判断に寄与する．よって発話・非発話区間における両肩，肘，手首の距離の平均及び偏差，最大値の 18 個を特徴量とする．

4.2 韻律情報

一般的に話者の発話意図や感情状態は言語に加えて，声の調子に表出され，韻律情報がタスクの評価に有効であることは多くの研究で示されている．本研究では OpenSMILE [2] [5] を使用し，検出された発話に対する

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[2] https://www.audeering.com/opensmile

音響的特徴量を 100fps の粒度で抽出した．抽出する特徴量と素性値は Interspeech 2009 Emotion Challenge で用いられた特徴量セットを採用する．音量の 2 乗平均平方根値，1 次 - 12 次元のメル周波数ケプストラム係数 (MFCC)，基本周波数 (F0)，声である確率，音声波形のゼロ交差率に対して，それぞれ最大，最小，最大値と最小値の差分，最大値の位置，最小値の位置，平均，購買度，オフセット，2 乗誤差，標準偏差，歪度，尖度からなる 383 個の特徴量を発話区間ごとに抽出する．

DNN には発話区間ごとに抽出された全ての特徴量を使用する．一方，SVM では韻律情報としてすべての特徴量を使用した場合，特徴量間のバランスがとられない．学習結果は特徴量サイズの大きいモダリティに影響を受ける．よってジェスチャ情報の特徴量数 18 を目標値とし，目的変数と韻律情報の相関値上位 18 個を特徴量として選択し，他のモダリティとのバランスをとった．特徴量選択によってゼロ交差率と MFCC 及び MFCC の 1 次微分が抽出された．

4.3 顔表情

授業者の表情は，OpenFace[3]を用いて抽出される．OpenFace は，Ekman らの Facial Action Coding System（FACS）[1] に従い，頭部及び視線の方向，17 種のアクションユニット（AU），そして推定結果の信頼度を抽出できる．授業者が板書をしている間，授業者はホワイトボード (黒板) 方向，もしくは横を向くため，OpenFace による表情の信頼度は低下する．OpenFace の検出信頼度は，単眼カメラでは検出困難な顔向きを示す特徴量となりうる．よって本研究では 17 種の AU と，検出の信頼度を特徴量とする．

DNN には 17 の AU と信頼度を 2 秒間の時系列データとして入力する．SVM には発話，発話・非発話区間における 17 の AU と信頼度それぞれの平均値，標準偏差，最大値と最小値の差分から，目的変数との相関値を計算し韻律情報と同じく上位 18 個を特徴量として選択し，他のモダリティとのバランスをとった．選択された特徴量は主に AU4〜AU6，AU14, AU45 であり，それらは表情の中で目元と口元の動きを示す特徴量である．

4.4 授業意図

授業者の言動の意図は評価値の推定に有効であると考えらえれる．我々はすでに，マルチモーダル情報を活用した教育志望者の授業意図推定モデルを構築している [16]．授業意図とは授業者の発話が意味する目的である．授業意図推定モデルでは言語情報・韻律情報・ジェスチャ情報を用いて，発話区間ごとにその発話意図を検出する．推定可能な授業意図は Information Providing Oral, Information Providing Pictorial, Instruct, Auto Feedback, Communication Management の 6 項目である．

本研究における指導評価は発話区間だけでなく，非発話区間を含めた連続的な出力である．よって授業意図推定モデルで推定可能な 6 種の発話意図に加え,「非発話」という状態を加えた 7 つの状態を 1 つの特徴量とする．また，検出された発話・非発話区間に対する発話の 2-gram, 3-gram を特徴量とする．最終的に DNN 及び，SVM ともに授業スキル評価対象区間の発話状態及び，発話意図の 2-gram, 3-gram を特徴量とする．

5 授業スキル評価モデル

発話ごとに評定された評価値の分類タスクを行い，分類精度を評価する．また，授業スキルの評価タスクにおいて有用な学習アルゴリズムの検討を SVM と DNN の比較によって行う．それぞれのアルゴリズムにおいて，各モダリティ毎に学習を行い，授業スキルの評価に有用なモダリティの検討を行う．モダリティの組み合わせは以下である．本論文では授業意図，韻律，ジェスチャ，顔情報をそれぞれ I, S, P, F と表記する．

1. I: 授業意図
2. S: 韻律
3. P: ジェスチャ
4. F: 顔
5. I+: 授業意図と顔韻律
6. I+P: 授業意図とジェスチャ
7. I+F: 授業意図と顔
8. S+P: 韻律とジェスチャ
9. S+F: 韻律と顔
10. P+F: ジェスチャと顔
11. I+S+P: 授業意図と韻律とジェスチャ
12. I+P+F: 授業意図とジェスチャと顔
13. I+S+F: 授業意図と韻律と顔
14. S+P+F: 韻律とジェスチャと顔
15. I+S+P+F: 全ての特徴量

5.1 SVM による授業スキル評価モデル

評価ラベルの分布は表 2 に示される通り，不均衡である．よって SMOTE アルゴリズム [3] を適用し，重みを維持しながらバランスを調整した．全ての学習は Weka API [8] における SMO アルゴリズムを採用した SVM により実施された．カーネルは最も性能の高

[3]https://cmusatyalab.github.io/openface/

かった RBF カーネルを使用した．複雑度合いを示す Complexity Parameter である C は 1.0〜10.0 の間でグリッドサーチにより探索し，最も性能の高い C = 4.0 を採用した．

図 1 に授業意図，韻律，ジェスチャ，顔情報の 4 つのモダリティの全ての組み合わせにおける SVM モデルの総合評価に対する性能を示す．また表 3 に全モダリティを使用した場合の各評価項目の精度を示す．結果は 10 分割交差検証でバリデーションされ，F 値は全ての結果の加重平均である．結果から以下の特徴が明らかになった．

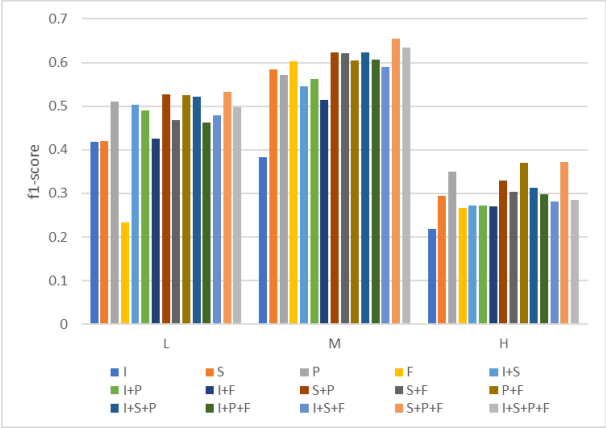

図 1: 特徴量セット毎の評価モデルの結果 (SVM)

表 3: SVM による評価モデルの精度 (I+S+P+F)

	Enth	Atte	Domi	Attr	KnPr	AlEv
L	0.50	0.50	0.51	0.51	0.53	0.50
M	0.64	0.67	0.62	0.63	0.59	0.63
H	0.24	0.36	0.30	0.25	0.37	0.28
Ave	0.55	0.58	0.53	0.54	0.54	0.54

- 複数のモダリティを併用することで，検出精度を向上させることができる．
- 授業スキルの低い L クラスの分類には顔表情はあまり効果的でない．顔表情は連続的に変化する時系列データであることから，発話区間内の平均，偏差及び最大値だけでは十分な情報量が得られなかったと考えられる．
- H クラスの検出精度が最も低く，他のクラスに比べその差は大きい．H クラスはインスタンス数が少なく，汎化性能を向上させるためには，より多くの訓練データが求められる．

5.2 DNN による授業スキル評価モデル

SVM は特徴量間のスケーリングの影響を強く受けるため，各モダリティそれぞれにおいて，意味的に重要な特徴量選択が求められる．授業スキルの評価は教育経験者の指導経験と主観に基づいて行われるように抽象度が高いため，センシング可能な情報から目的変数の予測に寄与する特徴量の選択は難しい．一方で，深層学習 (DNN) は学習の過程において，深い層で普遍的な特徴表現が獲得できることが知られている．

ネットワークの構造を図 2 に示す．入力層は 6 つのグループに分けられる．図中 $input_1$ から $input_3$ は授業意図の状態と N-gram の one hot vector，$input_4$ はジェスチャ情報，$input_5$ は顔表情，$input_6$ は韻律情報を入力とする．授業意図の one hot vector はそれぞれノード数 1 の Dense 層を通り，結合層にて意図の中間表現を抽出する．ジェスチャ情報と顔情報はそれぞれ，ユニット数 64 の LSTM(Long short-term memory) 層，韻律情報は Dense 層に入力される．各モダリティの特徴の統合を図り，中間層の潜在変数を共有することにより，異なるモダリティの情報から，共通な高次の特徴表現の獲得が見込まれる．出力層を除くすべての Dense 層において活性化関数はランプ関数 (Relu) を使用し，出力層は Softmax 関数により評価クラスの確率を出力する．誤差関数は多クラス交差エントロピーを使用する．

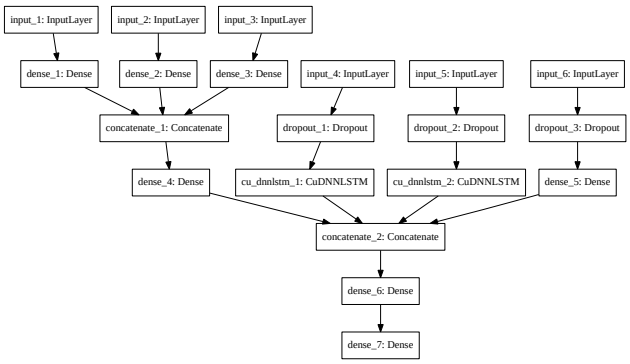

図 2: モデルのネットワーク構造

図 3 に授業意図，韻律，ジェスチャ，顔情報の 4 つのモダリティの全ての組み合わせにおける DNN モデルの総合評価に対する性能を示す．また表 4 に全モダリティを使用した場合の各評価項目の精度を示す．学習は Batch Size を 64, epoch 数を 100 とする．結果は SVM と同様に 10 分割交差検証で評価され，F 値は加重平均である．モダリティ別の学習を行う際は，各モダリティからのノードの統合を図る Concatnate より浅い層を切り離し，特徴量セットごとの学習を行った．その他のネットワーク構造及びパラメータは全ての学習で固定した．結果から以下の特徴が明らかになった．

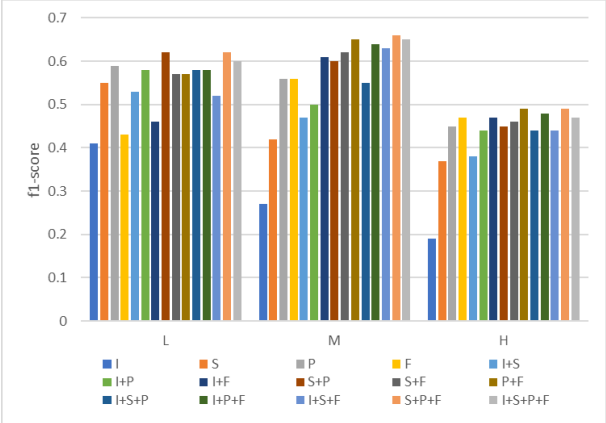

図 3: 特徴量セット毎の評価モデルの結果 (DNN)

表 4: DNN による評価モデルの精度 (I+S+P+F)

	Attr	Atte	KnPr	Domi	Enth	AlEv
L	0.62	0.61	0.64	0.59	0.60	0.60
M	0.67	0.65	0.60	0.59	0.66	0.65
H	0.55	0.47	0.47	0.51	0.50	0.47
Ave	0.64	0.62	0.60	0.58	0.62	0.61

- 授業スキルの評価にマルチモーダル情報は有効である．
- 授業スキルの低い L クラスの分類にはジェスチャ情報が有効である．また単一モダリティに顔表情，韻律情報を加えることで検出精度は向上する．
- 授業スキルの高い評価に値する H クラスでは顔表情が最も効果的である．
- 全ての授業スキル評価クラスの中で H クラスの検出精度が最も低い．
- 授業意図は他のモダリティほど効果的ではないが，授業意図と組み合わせることで，単一のモダリティを使用する場合よりも検出精度は向上する．
- 表 3 及び，表 4 から，全ての評価クラスにおいて，SVM よりも高い精度が示された．また DNN によって，ＳＶＭでは検出困難であった H クラスの検出が向上することが示された．

6 おわりに

仮想学級を用いた指導訓練システムにおける，マルチモーダル特徴量を用いた授業スキルの検出モデルを提案した．マルチモーダル特徴量にはジェスチャ，顔表情，授業意図及び韻律情報を採用した．授業スキルを High, Medium, Low の 3 クラスに分類するタスクを SVM と DNN の異なるアルゴリズムを用いて行った．結果，授業スキルの評価タスクにおいてはＳＶＭに比べ DNN が高い精度を示した．また単一のモダリティを使用する場合より，複数のモダリティを使用することで検出精度が向上することが明らかになった．

一方で，授業意図特徴量の精度への貢献は高くはない．これは，授業意図の検出に用いられる特徴量と，評価モデルにおける特徴量が重複することが問題としてあげられる．つまり，授業スキルの評価モデルにおいても，ジェスチャ情報や韻律情報等の複数モダリティの中間表現の組み合わせによって，授業意図のが抽出できている可能性が考えられる．よって授業スキルの評価モデルの特徴量として授業意図を使用するのではなく，授業意図の検出モデルと授業スキル評価モデルそれぞれに対し，マルチモーダル情報を入力することで，授業状況に関する情報を出力可能と考えられる．今後は，授業意図ごとに評価値を出力することで，より細分化された指導状況の把握が可能であると考えられる．

参考文献

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 59–66, May 2018.

[2] Roghayeh Barmaki and Charles E. Hughes. Providing Real-time Feedback for Student Teachers in a Virtual Rehearsal Environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pp. 531–537, New York, NY, USA, 2015. ACM.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, June 2002.

[4] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pp. 1143–1154, New York, NY, USA, 2015. ACM.

[5] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The Munich Versatile and

Fast Open-source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462, New York, NY, USA, 2010. ACM.

[6] Masato Fukuda, Hung-Hsuan Huang, Kazuhiro Kuwabara, and Toyoaki Nishida. Proposal of a Multi-purpose and Modular Virtual Classroom Framework for Teacher Training. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, pp. 355–356, New York, NY, USA, 2018. ACM.

[7] Masato Fukuda, Hung-Hsuan Huang, Naoki Ohta, and Kazuhiro Kuwabara. Proposal of a Parameterized Atmosphere Generation Model in a Virtual Classroom. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pp. 11–16, New York, NY, USA, 2017. ACM.

[8] Stephen R. Garner. Weka: The Waikato Environment for Knowledge Analysis. In *Proceedings of the New Zealand Computer Science Research Students Conference*, pp. 57–64, 1995.

[9] Yang Liu, Kun Han, Zhao Tan, and Yun Lei. Using Context Information for Dialog Act Classification in DNN Framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2170–2178, 2017.

[10] Jean-Luc Lugrin, Fred Charles, Michael Habel, Jamie Matthews, Henrik Dudaczy, Sebastian Oberdörfer, Alice Wittmann, Christian Seufert, Julie Porteous, Silke Grafe, and Marc Erich Latoschik. Benchmark Framework for Virtual Students' Behaviours. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 2236–2238, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[11] Jean-Luc Lugrin, Marc Erich Latoschik, Michael Habel, Daniel Roth, Christian Seufert, and Silke Grafe. Breaking Bad Behaviors: A New Tool for Learning Classroom Management Using Virtual Reality. *Frontiers in ICT*, Vol. 3, , 2016.

[12] Volha Petukhova and Harry Bunt. Incremental Dialogue Act Understanding. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pp. 235–244, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[13] 河野義章. パブリックスピーキング・スキルの研究-対話をイメージさせる要因. 昭和女子大学生活心理研究所紀要, Vol. 16, pp. 95–102, 2014.

[14] 栗原一貴, 後藤真孝, 緒方淳, 松坂要佐, 五十嵐健夫. プレゼン先生: 音声情報処理と画像情報処理を用いたプレゼンテーションのトレーニングシステム. WISS 第 14 回インタラクティブシステムとソフトウェアに関するワークショップ, pp. 59–64, 2006.

[15] 岡田将吾, 松儀良広, 中野有紀子, 林佑樹, 黄宏軒, 高瀬裕, 新田克己. マルチモーダル情報に基づくグループ会話におけるコミュニケーション能力の推定. 人工知能学会論文誌, Vol. 31, No. 6, pp. AI30–E_1–12, November 2016.

[16] 匡人福田, 宏軒黄, 豊明西田. 仮想学級におけるマルチモーダル特徴量を用いた教員志望者の授業意図の検出. HCG シンポジウム 2018, 2018.

[17] 文部科学省. これからの社会と教員に求められる資質能力. 2017.