The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

ビジネスプレゼンテーションにおける言語・非言語的能力の自動推定

Estimating Verbal・Nonverbal Skills in Business Presentation

八木 悠太朗 *1 Yutaro Yagi　岡田 将吾 *1 Shogo Okada　塩原 翔太 *2 Shota Shiobara　杉村 聡太 *2 Sota Sugimura

*1北陸先端科学技術大学院大学 先端技術研究科
School of Information science, Japan Advanced Institute of Science and Technology (JAIST)

*2ソフトバンク株式会社
SoftBank Corp.

This paper focuses on developing a model for estimating presentation skills of each participant from multimodal (verbal and nonverbal) features. For this purpose, we use a multimodal presentation dataset including audio signal data and body motion sensor data, text data of speech contents of participants observed in 58 presentation sessions. The dataset also includes the presentation skills of each participant, which is assessed by two external observers of the Human Resources Department. We extracted various kinds of features such as spoken uttetances, acoustic features, and the amount of body motion to estimate the presentation skills. We created a regression model to infer the level of presentation skills from these features using support vector regression to evaluate the estimation accuracy of the presentation skills. Experiment results show that the multimodal model achieved 0.59 in R^2 as the regression accuracy of effective production elements.

1. はじめに

企業やビジネスの場において，プレゼンテーション能力は重要なスキルのひとつである．プレゼンテーション能力が必要とされる場面として，企画やアイデアを発表するプレゼンテーションや商品やサービス内容を説明する店頭販売，採用面接などが挙げられる．プレゼンテーション能力を数値化することが出来れば，プレゼンテーションの振り返りや練習などを行う自動評定システムの構築に役立つ．本研究では，「プレゼンテーション能力」に関する個人特性を計算モデルにより捉えることに焦点を当てる．

プレゼンテーション能力の向上支援に向け，本研究は，プレゼンテーションを通じて観測できる参加者個人の話し方，体の動作，発話内容から，プレゼンテーション能力値を推定するモデルを機械学習により構築し，評価することを目的とする．推定モデルの構築に用いるデータセットを，ソフトバンク株式会社で行われたプレゼンテーション研修にて収集した．

2. 関連研究

2.1 プレゼンテーションスキルの推定

Torsten ら [Torsten 15] は，Public speaking skill の評定値と相関関係にあるプレゼンテーションから得られる非言語情報を分析し，それらを推定するモデルを構築した．相関係数 0.75 の予測精度を得た．また，音声と動作の特徴量を両方用いた場合に予測精度が向上することも示した．Vikram ら [Vikram 15] は，プレゼンテーションの音声情報や話者の表情，動作などの時系列マルチモーダル情報から特徴量を抽出し，個人のプレゼンテーション能力の評定値を推定するモデルを構築した．ある動作の後に別の動作を行うなどの時間的な共起パターンを共起ヒストグラムによって特徴量として抽出し，その特徴量がプレゼンテーション能力の評価にどれだけ貢献するのか分析した．相関係数で最大 0.69 の予測精度を得た．前者の研究は，プレゼンテーション能力に関係する特徴量の分析を，後者の研究は，プレゼンテーション能力の向上支援のためのシステムを構築している．[Torsten 15, Vikram 15] の目的は本研究の目的に類似する点はあるが，いずれも非言語情報のみを用いており，本研究のようにプレゼンテーション中の言語内容が重要視されるビジネスプレゼンテーションを対象としていない．

福島ら [福島 16] は，TED のウェブサイトで公開されている 1646 本のプレゼンテーション動画に対して視聴者が抱く印象を予測するモデルを構築した．プレゼンテーション動画から得られた言語情報と音声情報を用いて，複数の視聴者がアノテーションした beautiful，confusing などの合計 14 の印象の中から特定の印象を抱くか否かの 2 クラス分類を行い印象の予測を行った．分類の最大精度は 97.0%であった．この研究はプレゼンテーションの印象を予測することが目的であり，本研究のようにプレゼンテーション能力の推定を目的としていない．

2.2 本研究の位置づけ

関連研究と本研究の相違点をまとめ，本研究の貢献を明らかにする．

言語・非言語両面を考慮した評価尺度の生成

本研究はビジネスプレゼンテーションの評価を対象としている．ビジネスプレゼンテーションでは，プレゼンテーションの内容が重要視される．したがって，ジェスチャーや音声を考慮した評価尺度に加えて，言語面を重視した新たな評価尺度を作成した．

言語・非言語情報の利用

[Torsten 15, Vikram 15] では主に非言語情報に基づき分析を行っている．一方で言語的な能力を推定するためには，プレゼンテーションにおける発話内容は重要である．本研究では非言語特徴量だけでなく，論理性やプレゼン内容の評定を予測するために，発話内容に含まれる名詞数などの言語特徴量も用いて，評定値を推定する機械学習モデルを構築する．

連絡先: 八木　悠太朗，北陸先端科学技術大学院大学，s1810184@jaist.ac.jp

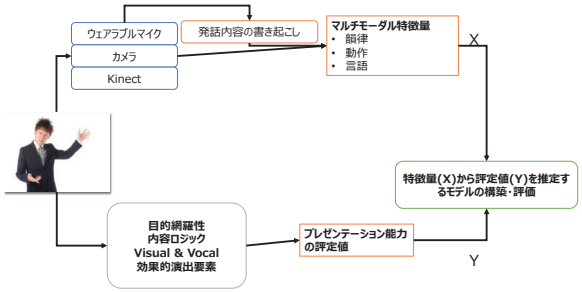

図 1: プレゼンテーション能力を推定する枠組み

表 1: 抽出する言語特徴量

特徴量	詳細
名詞数	発話内容に含まれる名詞数
動詞数	発話内容に含まれる動詞数
接続詞数	発話内容に含まれる接続詞数
フィラー数	発話内容に含まれるフィラー数
感動詞数	発話内容に含まれる感動詞数
同一名詞繰り返し最大数	同一の名詞を繰り返した最大数
同一動詞繰り返し最大数	同一の動詞を繰り返した最大数
名詞出現種類数	発話内容に含まれる名詞の種類数
動詞出現種類数	発話内容に含まれる動詞の種類数

3. 方法

3.1 概要

本研究の概要を図 1 にまとめる．収集したプレゼンテーションデータから音声・動作・言語に関する特徴量を抽出し，個人のプレゼンテーション能力を推定する機械学習モデルを構築し，評価する．推定モデルの構築に用いるデータセットは，ソフトバンク株式会社で行われているソフトバンクユニバーシティ「プレゼン話し方研修」で収集した．収集したデータセットに対して，ソフトバンク株式会社の人事部にプレゼンテーション能力の評価を依頼した．本章では，データセットの概要，抽出するマルチモーダル特徴量，および評価項目について述べる．

3.2 データセット

本実験で使用するデータセットの概要を説明する．データセットには，ウェアラブルマイクから収音した音声情報，Kinect から収音した音声情報，Kinect から得られた深度情報，Kinect から得られた各関節の位置情報，音声情報から文字起こしした発話内容が含まれる．このうち本研究では，ウェアラブルマイクから収音した音声情報，Kinect から得られた骨格の位置座標，音声情報から文字起こしした発話内容を用いてマルチモーダル特徴抽出を行う．実験には 58 人が参加し，一人一回 3 分間のプレゼンテーションを行う．実験で得られた合計 58 個のプレゼンテーションデータを実験に用いる．

3.3 特徴量

[岡田 16] を参考に韻律特徴量，言語特徴量，動作特徴量を抽出する．

3.3.1 韻律特徴量

ウェアラブルマイクから収音した音声情報から発話区間を推定し，OpenSMILE[*1] を用いて特徴抽出を行った．収音した音声のエネルギーの閾値処理に基づいた発話区間推定を用いた．The INTERSPEECH 2010 Paralinguistic Challenge feature set に含まれる 1582 次元の特徴量を抽出した．これらの特徴量には音の大きさ，0 次〜14 次メル周波数ケプストラム係数，0 次〜7 次メル周波数帯の対数パワー，8 つの LPC 係数から算出された 8 つの線スペクトラムペア周波数，平滑化基本周波数，最後の基本周波数の有声確率，及びそれらの一次微分が含まれる．

3.3.2 言語特徴量

音声情報から文字起こしした発話内容のテキストデータに対して形態素解析を行い，単語の品詞情報を取得した．形態素解析には MeCab[*2] を用いた．得られた単語の品詞情報をもとに，抽出した言語特徴量を表 1 にまとめる．

3.3.3 動作特徴量

Kinect から得られた各関節 (上半身 19 部位) の位置座標の時間変化を計算して，動作量と速度，加速度についてそれぞれ平均と標準偏差を求めた．また，発話中についても同様の特徴量を抽出した．

3.4 プレゼン研修講師による評定

プレゼンテーション評定の熟練者として，ソフトバンク株式会社のソフトバンクユニバーシティの講師二名が評定を行った．研修に参加した 58 名のプレゼンテーション映像を閲覧し，15 個の評価項目について，最低 0〜最高 2 の 3 段階で評定した．それらを 4 つの大項目に分類し，各項目の平均値を算出したものを評定値として用いた．評価項目とその定義を以下にまとめる．

目的網羅性

「誰に」,「何を」,「どうしてほしい」が伝えられている．

内容ロジック

「結論」,「根拠」,「相手の利益」が述べられている．

Visual & Vocal

「抑揚」,「声量」,「アイコンタクト」,「ジェスチャー」,「表情」が適切に用いられている．

効果的演出要素

「強調」,「繰り返し」,「具体表現」,「双方向性」が適切に用いられている．

各評定者による評定値の一致度を計算し，各項目に関する信頼性を確認した．一致度の計算には，カッパ値と重み付きカッパ値を算出した．ある項目に関してカッパ値が 0.4 以上であれば各評定者による参加者への評定値は概ね一致しており，その項目の評定値は信頼性が高いと言える [Landis 77]．

表 2 に項目ごとのカッパ値と重み付きカッパ値を示す．図中では，小数点以下第三位を四捨五入した値を記載している．全項目でカッパ値は 0.4 を上回っており，項目としての信頼性は高いと考える．

4. 評価実験

回帰予測タスクを行い，各参加者の評定値 (連続値) の推定精度を評価する．

*1 https://www.audeering.com/opensmile/

*2 http://taku910.github.io/mecab/

表 2: 各評定者間での評定値の一致度

能力項目	重み付きカッパ値	カッパ値
目的網羅性	0.99	0.81
内容ロジック	0.89	0.81
Visual & Vocal	0.91	0.54
効果的演出要素	0.99	0.93

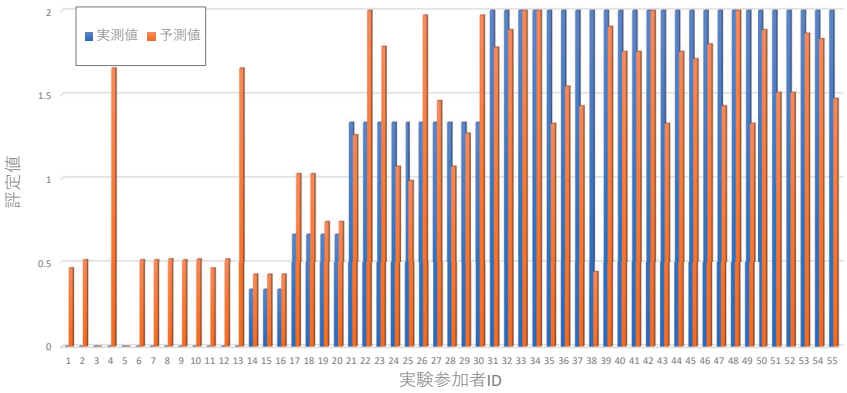

図 2: 「$All\,(p < 0.01)$」のモデルを用いた「目的網羅性」の評定値の予測値と実測値

4.1 推定モデルの構築手順

参加者の特徴量セットを入力 X，そのプレゼンテーションの評定値を出力 Y として，この入出力関係を学習する．合計で 58 個のデータを利用できる予定であったが，3 データの特徴量が機器の不具合などで一部欠損していたため，これらを除外し合計 55 個のデータを用いた．評価には一つ抜き (一人抜き) 交差検定法を利用した．ある参加者から得られたデータをテスト，それ以外の 54 個のデータを訓練に用いて回帰モデルの評価実験を行った．

回帰学習の概要

回帰学習には，RBF カーネルを用いた非線形サポートベクトル回帰モデル (Support Vector Regression:SVR) を用いた．SVR における誤差関数には，ϵ 許容誤差を用いた．誤差の許容度を調整する ϵ，損失関数とマージンの大きさの間のトレードオフを調整するパラメータである C，RBF カーネルの形状を調整するパラメータである γ に関して，それぞれ [0,0.01,0.1,1],[0.01,0.1,1,5,10]，[0.0001,0.001,0.01,0.1] の探索域を設け，訓練データを用いた交差検定を通じて，グリッドサーチにより最適パラメータを決定した．

回帰モデルの推定性能の評価指標にはテストデータに対する決定係数 R^2 を用いた．テストデータに対する評定値 (正解) と推定値の二乗誤差を，テストデータに対応する評定値の分散で割った値を E とすると，決定係数 R^2 は $1 - E$ で計算される．したがって，誤差が小さいほど R^2 は大きくなり，良い結果を示す．

4.1.1 比較対象の特徴量セット

各モダリティの推定性能への寄与を検証するため，以下の 9 種類の特徴量セットを準備し，推定精度の比較を行う．韻律，言語，動作の特徴量セットを A，L，M とそれぞれ略記する．

1. A：韻律特徴量
2. L：言語特徴量
3. M：動作特徴量
4. $A + L$：韻律と言語
5. $A + M$：韻律と動作
6. $L + M$：言語と動作
7. All：全ての特徴量 $(A + L + M)$
8. $All\,(p < 0.05)$：相関分析で有意な相関があった特徴量
9. $All\,(p < 0.01)$：相関分析で有意な相関があった特徴量

4.2 実験結果

プレゼンテーション能力を構成する 4 つの要素項目の評定値を推定する回帰タスクの結果を表 3 に示す．表 3 は各特徴量セットを用いて，各評価項目を推定した結果を表している．推定精度にはテストデータに対する決定係数 R^2 を用いた．

§1 単一モダリティ特徴量セットのモデルの精度

単一モダリティの特徴量セットを用いたモデルに関して「目的網羅性」,「内容ロジック」で「韻律特徴量 (A)」のモデルの精度が最大であった．「Visual & Vocal」では，「動作特徴量 (M)」のモデルの精度が最大であった．「効果的演出要素」では，「言語特徴量 (L)」のモデルの精度が最大であった．「目的網羅性」,「内容ロジック」,「Visual & Vocal」,「効果的演出要素」の最大推定精度はそれぞれ 0.32，0.07，0.05，0.00 であった．

§2 マルチモーダル特徴量セットのモデルの精度

マルチモーダル特徴量セットを用いたモデルに関して「目的網羅性」,「効果的演出要素」で「$All\,(p < 0.01)$」のモデルの精度が最大であった．「内容ロジック」で「$All\,(p < 0.10)$」のモデルの精度が最大であった．「Visual & Vocal」で「$All\,(p < 0.05)$」のモデルの精度が最大であった．「目的網羅性」,「内容ロジック」,「Visual & Vocal」,「効果的演出要素」の最大推定精度はそれぞれ 0.59，0.51，0.38，0.29 であった．

「目的網羅性」,「内容ロジック」,「Visual & Vocal」,「効果的演出要素」の要素項目に関して，マルチモーダル情報の統合は，単一モダリティ特徴量セットのモデルの精度を向上させるために役立つ．

推定精度が最大であった特徴量セットを以下にまとめる．「目的網羅性」,「効果的演出要素」の推定には $All\,(p < 0.01)$ が有効であった．「内容ロジック」の推定には $All\,(p < 0.10)$ が有効であった．「Visual & Vocal」の推定には $All\,(p < 0.05)$ が有効であった．

5. 考察

単一モダリティ特徴量セットの精度よりも，複数のモダリティを統合したマルチモーダル特徴量セットのモデルの精度のほうが高くなる傾向にあった．これは [Torsten 15] の結果とも一致している．

韻律特徴量について，今回は集団での研修の場においてデータ収集を行ったため，ウェアラブルマイクで収音した音声データにノイズが含まれている場合があった．実用化に向けては，音声認識を用いて自動的に発話内容から言語特徴量を抽出する必要がある．本研究で用いた音声データに音声認識を用いて得

表 3: SVM を用いた回帰タスクの結果：各特徴量セットを用いて各プレゼンテーション能力項目を推定した場合の精度を示している．精度はテストデータに対する決定係数 R^2 を用いた．

	目的網羅性		内容ロジック		Visual & Vocal		効果的演出要素	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
A	0.32	0.47	0.07	0.57	-0.11	0.20	-0.09	0.21
L	-0.69	1.18	-0.11	0.67	-0.21	0.22	0.00	0.19
M	-0.23	0.86	-0.36	0.82	0.05	0.17	-0.34	0.26
$A + L$	0.31	0.48	0.07	0.57	-0.12	0.20	-0.04	0.20
$A + M$	-0.16	0.81	0.22	0.47	0.16	0.15	-0.17	0.23
$L + M$	-0.25	0.87	-0.18	0.71	-0.31	0.24	-0.29	0.25
All	-0.17	0.81	0.21	0.48	0.17	0.15	-0.17	0.23
$All\,(p < 0.10)$	0.54	0.32	**0.51**	0.30	0.23	0.14	0.24	0.15
$All\,(p < 0.05)$	0.47	0.37	0.39	0.37	**0.38**	0.11	-0.14	0.21
$All\,(p < 0.01)$	**0.59**	0.28	-0.32	0.80	0.13	0.16	**0.29**	0.14

られた言語情報から言語特徴量を抽出した場合の推定精度についても調査する必要がある．

動作特徴量について，本研究では体の各部位の動作量や速度，加速度を特徴量として抽出した．しかし「Visual & Vocal」の評価項目では視線や表情も考慮されているため，視線や表情に関する特徴量を加えモデルを構築する必要がある．

言語特徴量について，本研究では音声情報から文字起こしした発話内容のテキストデータに対して形態素解析を行い，得られた単語の品詞情報から言語特徴量を抽出している．また学習の前に相関分析を行い，有意な相関がある特徴量のみから各評価項目の評定値を推定した場合に最も高い推定精度が得られた．言語面の評価項目である「目的網羅性」,「内容ロジック」に対して，名詞出現種類数，動詞出現種類数は $p < 0.10$ で有意な相関を示した.「効果的演出要素」では，名詞数，動詞数，名詞出現種類数，動詞出現種類数の特徴量が $p < 0.05$ で有意な相関を示した．単語レベルでなく，構文レベルの解析を行い,「目的網羅性」,「内容ロジック」の項目の推定に寄与する特徴量を抽出することが今後の課題の一つである．

図 2 は予測精度が最大であった「$All\,(p < 0.01)$」のモデルを用いた「目的網羅性」の評定値の予測値と実測値を比較している．評定値が 0 点〜1 点，1 点〜2 点の場合，推定誤差の平均がそれぞれ 0.16，0.35 であるのに対し，0 点の場合の推定誤差は 0.60 であった．この結果より，評定値が 0 のデータの推定精度が低いことがわかった．このようなデータに対しても高い精度で推定を行える機械学習モデルを構築する必要がある．

6. 結論

本研究では，プレゼンテーションデータから得られたマルチモーダル情報から「プレゼンテーション能力」を推定するモデルの構築・評価を行った．発話内容に含まれる品詞情報，韻律情報，動作量をマルチモーダル特徴量として抽出し，個人のプレゼンテーション能力値を推定するモデルを機械学習により構築した．評価実験の結果,「目的網羅性」,「内容ロジック」,「Visual & Vocal」,「効果的演出要素」に関する評定値に関して，回帰タスクでそれぞれ最大 0.59，0.51，0.38，0.29 の決定係数 R^2 を得た．

本研究では，プレゼンテーション中の言語・非言語特徴量の時系列変化を考慮していない．時系列マルチモーダル情報を用いたモデルの構築は今後の課題である．

7. 謝辞

ソフトバンク株式会社 人事本部の笠井 理恵様，高嶋 直人様，海上 博志様には，データセットの収集，および評定値の作成においてご支援いただきました．心より感謝申し上げます．

参考文献

[岡田 16] 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, 2016, 31 巻, 6 号, p. AI30-E_1-12

[福島 16] 福島 悠介, 山崎 俊彦, 相澤 清晴, 文書と音声解析に基づくプレゼンテーション動画の印象予測, 電子情報通信学会論文誌 D Vol.J99-D No.8 pp.699-708, 2016

[Landis 77] Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics. 33 (1): 159–174.

[Torsten 15] Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. 2015. Multimodal Public Speaking Performance Assessment. In Proceedings of the 2015 ACM on ICMI (ICMI '15).

[Mathieu 16] Mathieu Chollet, Helmut Prendinger, and Stefan Scherer. 2016. Native vs. non-native language fluency implications on multimodal interaction for interpersonal skills training. In Proceedings of the 18th ACM on ICMI (ICMI 2016).

[Vikram 15] Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. 2015. Evaluating Speech, Face, Emotion and Body Movement Time-series Features for Automated Multimodal Presentation Scoring. In Proceedings of the 2015 ACM on ICMI (ICMI '15).