人工知能学会研究会資料
SIG-SLUD-B501-07

説明行為の質の推定に向けた会話者のマルチモーダル情報モデリング

A multimodal modeling for predicting the performance of storytelling

岡田 将吾 [1*]　米 航 [1]　新田 克己 [1]
Shogo Okada[1]　Mi Hang[1]　Katsumi Nitta[1]

[1] 東京工業大学大学院総合理工学研究科　知能システム科学専攻
[1] Dept. of Computational Intelligence and Systems Science, Tokyo Institute of Technology

Abstract: We present a multimodal analysis of storytelling performance in group conversation as evaluated by external observers. A new multimodal data corpus, including the performance score of participants, is collected through group storytelling task. We extract multimodal features regarding explanators and listener from a manual description of spoken dialog and from various nonverbal patterns. We also extract multimodal co-occurrence features, such as utterance of explanator overlapped with listener's back channel. In the experiment, we modeled the relationship between the performance indices and the multimodal features using machine learning techniques. Experimental results show that the highest accuracy is 82% for the total storytelling performance (sum of score of indices) obtained with a combination of verbal and nonverbal features in a binary classification task.

1　はじめに

対面会話において説明行為やストーリーテリングは情報共有，他者への教示，などのために重要な役割を担う．社会言語学の研究分野では，説明における説得力を上げるためには，説明中に用いられた言語情報だけでなく動作，韻律，表情といった非言語情報が重要であると報告されている [1]．本研究は，説明会話中の説明の質を，会話中に交わされた言語・非言語情報から推定するモデルの構築に焦点を当てる．外部観測可能な言語情報・非言語情報から会話中の説得力やストーリーテリングを推定する計算機モデルを構築出来れば，説明会話の質を評価する技術や，大規模な会話データから良質な説明シーンを検索するシステムに応用できる．またモデル構築・評価を通じて質の高い説明において観測されるマルチモーダル情報を分析することも可能となり，説明行為の理解にも役立つ．

この目的に向けて，複数のモーションセンサ・マイクを用いて説明会話を計測し，説明会話において交わされた言語・非言語情報を含むマルチモーダルデータセットを新規に収集する．マルチモーダルデータセットは音声データ，手・頭部の動作データといった複数の非言語データと，発話内容を書き起こした言語データを含む．また，収集した会話データを 3 名の観察者が閲覧し，説明内容の基となった動画と見比べながら，説明の円滑さ，情報の正確さといった複数の観点から説明の質を評価する．

次に，発話区間，韻律情報，手のジェスチャ，頭部のジェスチャ，顔向け方向といった非言語特徴量を抽出し，書き起こした言語データから発話内容中の単語特徴量を抽出する．さらに,「共同注視をしながらの発話」といった会話における説明者と聞き手のインタラクションや,「ジェスチャを伴いながらの発話」,「聞き手を見ながらの発話」といったマルチモーダルパターンを共起パターンマイニングにより抽出する．抽出した多様なマルチモーダル特徴量と，アノテーションされた説明行為の評価値の間の関係を分析，モデル化するために，多変量解析，機械学習を行い，説明行為の質の推定精度について議論する．マルチモーダルモデリングの概要を図 1 に示す．

2　関連研究

対面会話において表出される参加者の音声区間，韻律，体・頭部の動作，視線状態などの非言語情報に基づき，参加者の役割 [2] [3]，性格特性 [4]，リーダシップを有する人物 [5] といった参加者の高次特性を推定する研究が行われている．ただし，グループ会話における説明の質の推定に焦点を当てた研究は我々の知る限り存在しない．[6] はユーザが商品をのレビューを行っているビデオデータを収集し説得力の度合いをレビュアーの言語・非言語情報から推定する研究を行った．本研究では説明者だけでなく，聞き手とのインタラクションの過程で表出される言語・非言語情報を手掛かりに，説明会話における説明の質を推定する．

本研究の貢献は以下の二つである．説明の質を複数の

*連絡先：東京工業大学 大学院総合理工学研究科
〒 226-8502 横浜市緑区長津田町 4259
E-mail:okada@dis.titech.ac.jp

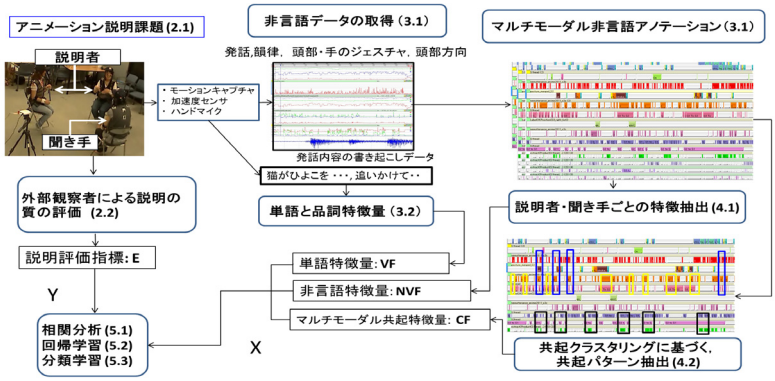

図 1: 説明評価値の推定のためのマルチモーダルモデリングの概要 (() 内の数字は論文中の節番号に対応している．)

外部観察者による説明への評価値として定義し，この評価値と，説明者・聞き手双方の言語・非言語情報とを関連付けることでグループ会話における説明の質を推定するモデルを構築・評価する．[2, 3, 4, 5, 6] では発話区間，ジェスチャ，言語といった個々の特徴量を 1 セッション中から抽出し，機械学習を行う際に統合していたが，複数のモダリティの時間共起特徴を考慮していなかった．本研究ではマルチモーダル共起パターンをデータマイニングのアプローチにより抽出し，モデルの推定に用いる．実験を通じて，説明の質を推定するために上記の共起特徴を含む多様な特徴量が説明の質を推定するために寄与することを示す．

3 データコーパス

3 人のグループ対話タスクとしてアニメーションの内容・各シーンの状況を説明する課題を設定した [7].

3.1 会話タスクの概要

本研究では McNeill [8] により考案された，動画を事前に観察した参加者（以後，説明者と呼称）がその動画を見ていない参加者（以後，聞き手と呼称）に動画内容を説明するタスクを選定した．3 人は着座状態で対話を行う．人材派遣会社を通じて計 30 名の実験協力者を募集した．募集した 30 名はいずれも初対面の 20 代前半の女性であり，3 人同士を 1 グループとして説明タスクを 10 セッション行い，対話データを収集した．この内，充分に説明を行わなかった 1 グループと，センサデータの欠損が著しかった 1 グループのデータを除外し，計 8 セッションのデータを本研究に使用した．各セッションの平均対話時間は約 11 分（合計で約 700 分）であった．

アニメーションは 8 つのエピソードで構成されており，基本的には全てのセッションにおいて，この 8 つのエピソードに関する説明が行われる．また全てのセッションの説明者はアニメーションの状況設定，登場人物等の説明を事前に行ったため，事前説明を含め計 9 つのエピソードで構成される．本研究では説明会話をエピソードごとに予め分割し，8 セッションで計 67 エピソードの会話データセットを構築した．

3.2 外部観察者による説明の質の評価

本研究では，外部観察者により質の高い説明シーンを評価する．説明の理解度を基準とする場合，聞き手のインタビューに基づき説明の質を評価する必要がある．しかしながら，セッションごとに異なる聞き手が説明を受けており，インタビューの結果，聞き手の記憶量が一定でないことを確認したため，本研究では複数の第三者による評価値を採用した．説明の質の評価には，図 1 に示す 10 個の指標（E_{1-10}）を用いた．E_{1-10} は「雄弁性」,「熱意」,「円滑さ」,「機知に富んだ説明内容」,「情報の正確さ」,「端的な説明」,「説明内容の要約」,「活発な説明態度」,「協調的な説明」,「質疑応答の適切さ」に関する評価値をそれぞれ示す．最後に E_{Total} は説明評価値 E_{1-10} の総和を示しており，本研究ではこの値を総合的な説明評価値と定義する．スケールは 10 段階 (最少 1，最大 10) とした．

上記の評価指標が妥当であるかどうかを判断するため 3 人の外部観察者による評価値の一致率をクローンバックの α 値により計算した結果，全て 0.7 を超えており，充分な一致率を得ることが出来た．各セッションの説明者は同一の動画内容に関する説明を行う．外部観察者はこの動画内容と，各セッションにおける説明内容を比較しながら説明の質を評価することができるため，充分な一致率を得ることが出来た考える．以降の実験では 3 人による評価値の平均値を，各指標の評価値として用いる．

4 マルチモーダルアノテーション

マルチモーダル情報の抽出のために，ハンドマイク，光学式モーションキャプチャMac3D，加速度・ジャイロセンサ (6 軸センサ) を用いた [7].

4.1 非言語情報のアノテーション

音声区間検出

各話者の音声区間の検出を行う．ここでは零点交差法により音声区間の候補を抽出し，事前に音声区間を学習しておいた混合ガウシアンモデルを用いて音声区間を検

出した．上記の実行にはソフトウェア Julius[1] を用いた．音声区間として抽出された発話断片ラベル sp の集合を SP と定義する．

ジェスチャ区間検出

動作状態・無動作状態の 2 クラスに分類した．ジェスチャ区間の検出には正規分布を出力確率分布にもつエルゴディック型の HMM を用いた．手の動作区間として抽出されたハンドジェスチャラベル hg の集合を HG と定義する．

頭部ジェスチャアノテーション

後頭部に装着した 6 軸センサより取得される時系列データから頭部方向の動きを検出する．ここでは首を縦方向に動かしたか否かの 2 クラスに分類する．[9] で利用された，離散ウェーブレット変換により算出される特徴量を用い，識別器にはガウシアンカーネルに基づく SVM を用いてアノテーションを行う．頭部の動作区間として抽出されたジェスチャラベル he の集合を HE と定義する．

頭部方向アノテーション

本研究では顔向きから視線方向を近似する．三話者会話タスクにおける顔向きを分類するために，ある参加者が，参加者の左または右の参加者を見ているかどうかの 2 クラスに分類する．訓練データセットから前頭部・頭頂部マーカの位置座標の差分ベクトル集合を算出しておき，これらのベクトル集合と最近傍識別を行い，テストデータのラベルを分類した．参加者 i が参加者 l に視線を向けていると推定された区間の視線方向ラベル $g_{i,l}$ の集合を $G_{i,l}$ と定義する．

4.2 発話内容の言語アノテーション

説明の質は説明者のボキャブラリーに関連していると考えられる．各発言内容を日本語話し言葉コーパス（CSJ）の転記フォーマットに従って書き起こし，書き起こしたデータに対して形態素解析を行う事で発話内容を単語に分割し，品詞のアノテーションを行った．品詞の特徴量セットを W と定義し，各品詞を w と定義する．形態素解析には茶筅[2]を用いた．今回研究に用いた品詞セットは，w_1：名詞，w_2：接続詞，w_3：動詞，w_4：形容詞，w_5：副詞，w_6：前置詞，w_7：助動詞，w_8：感嘆詞の計 8 種類と w_9：フィラー，w_{10}：笑い，のアノテーションを含め計 10 種類の特徴を用意した．

5 マルチモーダル特徴量の抽出

4 章でアノテーションしたマルチモーダルラベルセットから，説明の質を推定するために用いる特徴量の抽出を行う．まず説明者・聞き手ごとに発話量，発話内容に含まれる単語頻度といった個別の特徴量を抽出する．次に,「ジェスチャを伴いながらの発言」といった各ラベルを組み合わせた特徴量を抽出する．以下で述べる，マルチモーダル特徴量を表 2 にまとめる．全ての特徴量を，エピソードの時間長：EPL で割ることで正規化した．

表 1: 説明評価指標

説明評価指標	略記	質問事項
$E_1:(Eloquence)$	*Eloquent.*	雄弁に説明していたか?
$E_2:(Enthusiasm)$	*Enthus.*	説明者は熱意をもって説明していたか?
$E_3:(Fluentness)$	*Fluent.*	円滑な説明を行えていたか?
$E_4:(Wittiness)$	*Witt.*	説明内容は機知に富んでいたか?
$E_5:(Preciseness)$	*Precise.*	正確な情報を伝えていたか?
$E_6:(Compactness)$	*Compact.*	簡潔に説明を行っていたか?
$E_7:(Summarization)$	*Summary*	主要な点を要約して説明できていたか?
$E_8:(Liveliness)$	*Live*	活気のある説明を行えていたか?
$E_9:(Cooperation)$	*Cooperate.*	協調して説明を行えていたか?
$E_{10}:(Q\&A)$	*Q&A*	質疑応答は適切であったか?
E_{Total}	Total	E_1-E_{10} の値の総和

5.1 各モダリティからの特徴量

ラベルセット SP, HG, HE, G に含まれる個々のラベルセグメントは始点と終点を有しており固有の時間長を有している．j 番目のエピソード SE_j にお

いて各参加者 i ($i=\{E1, E2, L\}$) から観測されるラベル系列から特徴量 $F_{i,j}$ を計算する．ここで $E1, E2$ は説明者，L は聞き手のインデックスを示す．各セッションの説明者は 2 名であるため，この 2 名から抽出した特徴量の和を説明者の特徴量として定義し，特徴量は各エピソード毎に計算される．

発話中の単語特徴量

j 番目のエピソード SE_j において各参加者 i の発話内容に含まれる品詞の回数 $W_{i,j}$（4.2 節 ）を計算する．説明者の特徴量は $EW_j = W_{E1,j} + W_{E2,j}$ ，聞き手の特徴量は $LW_j = W_{L,j}$ として計算される．$W_{i,j}$ は 10 次元のベクトルであるため，EW_j, LW_j を合わせて 20 次元のベクトルが構成される．

発話ターンに関する特徴量

充分な情報量を相手に伝えるためには，一定の説明時間を確保する必要があり，発話の回数や発話量は説明の質に関連すると考えられる．j 番目のエピソード SE_j で参加者 i から N_i 個の発話断片が得られた場合の各発話断片を $sp_{i,k}$, $(1 \le k \le N_i)$ として，各参加者の発話回数 SL_i，発話長 ST_i を計算し，説明者の発話長 ESL，発話回数 EST を以下のように計算する．

$$SL_i = \sum_{k}^{N_i} SL_{i,k},\ ST_i = N_i$$
$$ESL = SL_{E1} + SL_{E2},\ EST = ST_{E1} + ST_{E2} \quad (1)$$

聞き手に関するそれらの特徴量は $LSL_{=}SL_L, LST = ST_L$ とする．次に，会話中にターンテイキングが起きた回数を計算し，これを特徴量 TT として計算する．異なる話者の発話断片が順番に観測された場合に一回のターンテイキングとして認定する．この処理により発話ターンに関して 5 次元の特徴量を抽出した．

韻律に関する特徴量

発話中の声の抑揚は説得力を上げるために効果的であることが知られている [10]．このため本研究では発話

[1] Julius: http://julius.sourceforge.jp
[2] Chasen: http://chasen.naist.jp/hiki/ChaSen/

断片中の音声データからエネルギーとピッチ（基本周波数）を計算する．ここでは発話者のみの特徴量を計算した．韻律情報の特徴抽出には Speech feature extraction code[3]を用いた．各セッション中に観測された 2 名の説明者の発話断片 $sp_{i,k}$ 群から計算したそれぞれのエネルギー E，ピッチ P の値について最大，最少，平均，標準偏差をそれぞれ計算し，計 8 次元の特徴量を抽出した．

視線の特徴量

視線方向について，各参加者から見て左右の参加者のどちらを見ていたかのラベルが付与されている．k 番目の視線ラベルについて，参加者 l が参加者 i に視線を向けている時間を $GL_{i,l,k}$ と定義すると，説明者，聞き手が視線を向けられていた時間の総和 EGF，LGF は．$GF_i = \sum_{l,k} GL_{l,i,k}$, $EGF = GF_{E1} + GF_{E2}$, $LGF = GF_L$ と計算される．次に，参加者 i,l の間で共同注視が観測された時間長を $MG_{i,l}$ と定義し，説明者同士の間での共同注視時間：$EMG = MG_{E1,E2}$ と説明者と聞き手の間での共同注視時間：$LMG = MG_{E1,L} + MG_{E2,L}$ を計算する．また，視線方向が変化した回数を TT と同様に計算し，説明者の視線変化回数：$EGT = GT_{E1} + GT_{E2}$ と聞き手の視線変化回数：$LGT = GT_L$ を計算した．視線について合計で 6 次元の特徴量を抽出した．

5.1.1 頭部と手のジェスチャの特徴量

頭部・手のジェスチャについても発話ターンと同様に，ジェスチャの持続長 HL_i, GL_i，回数 HT_i, GT_i を各参加者 i ごとに計算し，説明者の持続長 EHL, EGL，回数の総和 EHT, EGT を計算する．聞き手に関しても同様に計算し，頭部, 手のジェスチャに関して $HEF = \{EHL, LHL, EHT, LHT\}$ の 4 次元，$HGF = \{EGL, LGL, EGT, LGT\}$ の 4 次元の特徴量をそれぞれ抽出した．

5.2 非言語マルチモーダル共起特徴量の抽出

「視線を向けながらの発話」,「ジェスチャを伴いながらの発話」といった，複数のモダリティの共起パターンは単一の特徴量よりも，詳細な会話者の会話態度や状態を示している．また,「説明者の発言時に聞き手がうなづいている」というような 2 者以上の間のインタラクションパターンも重要な会話のコンテキストを示していると考えられる．

本研究では，各参加者 i から取得したラベルセット SP_i, HG_i, HE_i,$G_{i,l}$ に含まれるセグメントの時間共起度に基づき，上記で述べるマルチモーダル共起特徴やインタラクションパターンを抽出する．本研究では効率良く時間共起度（同時に観測される割合）の高い共起パターンのみを抽出するために共起クラスタリングを利用する．本研究では [11] で提案されている共起クラスタリングを拡張して，この目的に利用する．

[11] のアルゴリズムでは，段階的に各パターン p（本研究では各ラベルに対応）をマージしていく戦略で，マージされた共起パターン $CP_x = \{p_1, p_2, \ldots p_{N_x}\}$ のみを候補として抽出していた．本研究では，各段階でマージされたパターンを全て列挙する．これは例えば,「ジェスチャ(p_1)を伴いながらの発言 (p_2)」を示す共起パターン $\{p_1, p_2\}$ と「ジェスチャ(p_1) を伴いながらの発言 (p_2) 中に聞き手がうなづいた (p_3)」を示す共起パターン $\{p_1, p_2, p_3\}$ とを別個のパターンとして抽出するためである．また，[11] のアルゴリズムではパターンの総数の多い順に共起パターン抽出を行っていたが，この場合，出現頻度の低いパターンを列挙できないため，本研究ではパターン総数の昇順，降順，両方の場合でアルゴリズムを実行し，列挙された共起パターン CE_x を特徴量として抽出する．この結果として計 195 の共起パターンが列挙された．共起パターン CE_x が各エピソード内で観測される回数を計算し，この回数を特徴量として定義する．紙面の都合でアルゴリズムの詳細については [11] に譲る．

表 2: マルチモーダル特徴量セット

	変数	特徴量の概要
VF	単語特徴量 (4.2 節)	
	EW_{1-10}	説明者の発話内容に含まれる品詞の頻度
	LW_{1-10}	聞き手の発話内容に含まれる品詞の頻度
NVF	EPL	エピソードの時間長 (説明時間)
	音声特徴量	
	ESL	説明者の発話長
	LSL	聞き手の発話長
	EST	説明者の発話回数
	LST	聞き手の発話回数
	TT	ターンテイキングの回数
	ESE	説明者の音声中のエネルギー特徴
	ESP	説明者の音声中のピッチ特徴
	視線特徴量	
	EGF	説明者が視線を向けられた時間長
	LGF	聞き手が視線を向けられた時間長
	EMG	説明者同士が共同注視した時間長
	LMG	説明者と聞き手が共同注視した時間長
	EGT	説明者が視線方向を変化させた回数
	LGT	聞き手が視線方向を変化させた回数
	ハンドジェスチャの特徴量	
	EGL	説明者がジェスチャを行った時間
	LGL	聞き手がジェスチャを行った時間
	EGT	説明者がジェスチャを行った回数
	LGT	聞き手がジェスチャを行った回数
	頭部ジェスチャの特徴量	
	EHL	説明者がジェスチャを行った時間
	LHL	聞き手がジェスチャを行った時間
	EHT	説明者がジェスチャを行った回数
	LHT	聞き手がジェスチャを行った回数
CF	マルチモーダル共起特徴量	
	CE_{1-195}	列挙された共起パターンの頻度

6 マルチモーダル分析・モデリング

本章では，まず 6.1 節においてマルチモーダル特徴量と説明評価値との間の関係について相関分析し，説明評価値に対して有意に相関の高い特徴量を特定する．次に，6.2，6.3 節では説明評価値の連続値を予測する回帰学習タスクと，評価値の高・低の 2 クラスを予測する分類学習タスクの 2 種類の機械学習タスクを行い，説明評価値の推定精度を検証する．

[3]Speech feature extraction code, `http://groupmedia.mit.edu/data.php`.

表 3: 説明評価値と有意な相関係数が得られたマルチモーダル特徴量 ($p < 0.05$: $*$, $p < 0.01$: $**$)

特徴量の変数	最大の相関係数		有意な相関が認められた評価指標
	評価指標	r	
EPL	E_8 (*Live.*)	$+0.41^{**}$	E_{1-10}
単語特徴量			
$EW(w_6)$	E_8 (*Live.*)	-0.41^{**}	E_{2-10}
$EW(w_7)$	E_1 (*Eloquent*)	-0.49^{**}	E_{1-10}
$EW(w_8)$	E_1 (*Eloquent*)	$+0.53^{**}$	E_{1-10}
$EW(w_9)$	E_{10} (*Q&A*)	-0.40^{**}	E_{1-10}
$EW(w_{10})$	E_9 (*Cooperate*)	$+0.27^{*}$	$E_{5,7,9}$
$LW(w_3)$	E_8 (*Live.*)	$+0.30^{*}$	$E_{2,3,7,8}$
$LW(w_4)$	E_1 (*Eloquent*)	$+0.31^{**}$	$E_{8,10}$
$LW(w_7)$	E_2 (*Enthus*)	-0.46^{**}	E_{1-10}
$LW(w_8)$	E_2 (*Enthus*)	$+0.32^{**}$	$E_{1-4,6,7,9,10}$
$LW(w_9)$	E_2 (*Enthus*)	$+0.38^{**}$	E_{1-10}
音声特徴量			
ESL	E_1 (*Eloquent*)	$+0.27^{*}$	$E_{1,3,4,8}$
LST	E_1 (*Eloquent*)	-0.27^{*}	E_1
ESE_{min}	E_{10} (*Q&A*)	-0.49^{**}	E_{1-10}
ESE_{std}	E_{10} (*Q&A*)	-0.28^{*}	E_1
ESP_{min}	E_6 (*Compact.*)	-0.30^{*}	E_{3-10}
ESP_{mean}	E_1 (*Eloquent*)	-0.30^{*}	$E_{1-4,6,7}$
ESP_{std}	E_5 (*Precise.*)	$+0.33^{**}$	$E_{2,4-7,9,10}$
視線特徴量			
EGF	E_5 (*Precise.*)	-0.27^{*}	$E_{1,5,7}$
EMG	E_1 (*Eloquent*)	-0.35^{**}	E_{1-10}
EGT	E_9 (*Cooperate*)	$+0.45^{**}$	E_{1-10}

6.1 相関分析結果

本節では，表 2 に記載した全ての特徴量と 10 個の説明評価値の間でピアソン積率相関係数を計算し，$p < 0.05$ の有意な相関値が得られた特徴量を報告する．表 3 は言語・非言語特徴量に関する相関分析の結果を示している．r は相関係数であり，$p < 0.05$, $p < 0.01$ の有意な相関が認められた場合，相関値の横に $*$, $**$ をそれぞれ記載した．表 3 において，二，三列目には，一列目の特徴量との相関係数の絶対値が最大である評価値とその係数を，四列目には $p < 0.05$ で有意な相関を有する説明評価指標を記載する．

6.1.1 発話単語特徴に関する相関分析結果

表 3 より，説明者の発話内容に含まれる接続詞の頻度数 w_6，参加者全員の助動詞の頻度数 w_7　は多くの説明評価指標と負の有意な相関を有している．この結果は説明の質が高いとされた説明シーン（エピソード）における説明者は接続詞や助動詞を多く使わないことを示す．一方で，簡単詞やあいづちが多く使われたエピソードの説明評価は高い傾向にあることがわかる．基本的にあいづちは重要な聞き手の行為であることが知られており，この結果は自然な結果であると考えられる．また説明者のフィラーの頻度 w_9 は説明評価値全てと負の相関が認められており，E_1 と-0.40 の相関値を有することが示された．

6.1.2 非言語特徴に関する相関分析結果

表 3 より，エピソードの時間長 EPL は各説明評価指標と正の相関が認められることがわかる E_{1-10}（最大値は 0.41）．この結果は，質の高い説明をするためにはある程度の時間をかける必要があることを示している．また，それと同時に，説明者の発話時間 (ESL) も説明指標と正の相関を有している．一方で聞き手の発話時間 (LSL) は，説明指標と負の相関を有している．聞き手がターンを長く保持するエピソードでは，説明の質が低い傾向にあることが示された．視線の特徴量において説明者に視線が集まる場合，また説明者同士が共同注視する時間 (EMG) が長い場合，説明の評価値が下がる（説明評価指標と負の相関にある）ことが示された．この原因をビデオを観察し分析したところ，説明内容に自信がない場面で，説明者同士が見合ってしまうケースが多いことがわかった．

6.2 回帰学習に基づく予測実験結果

本節では，抽出した特徴量セットが説明評価値の予測に有効かどうかを検証するため，回帰学習・モデル評価を行う．回帰モデルには L2 ノルムを正規化項に用いる，リッジ回帰モデルを採用した．交差検定法によりモデル評価を行う．ここでは 1 エピソードをテストとして用い，残りの 66 エピソードをモデルの学習に用いた．表 2 から 6 種類の特徴量セット (1)NVF, (2)VF, (3)CF, (4)$VF + NVF$, (5)$VF + CF$, (6)$VF + NVF + CF$ (All) を用いて上記の評価実験を行い，有効な特徴量を検証した．マルチモーダル共起特徴とインタラクション特徴量 CE について，主成分分析を行う事により累積寄与率が 0.999 を超える最少数の主成分を特徴量として採用した．この結果，CE の次元数は 195 から 30 に削減された．交差検定法を行い，各試行でテストデータの予測誤差を算出し，全試行の予測誤差から決定係数 R^2 を計算した．リッジ回帰モデルの正則化項の重みパラメータを $[1-150]$ と変化させ，訓練データ集合を使って最適なパラメータを求め，これをテスト時に使用した．図 2 は各特徴量セットを用いて学習した回帰モデルのテストデータに対する R^2 を各指標ごとに示している．

言語特徴 VF, 非言語特徴 NVF，マルチモーダル共起特徴 CF で学習したモデルをそれぞれ比較すると，$E_{1,5-8,10,Total}$ については言語特徴，$E_{2-4,9}$ についてはマルチモーダル共起特徴を用いたモデルの精度が最も高かった．次に，全ての特徴量セットで学習したモデルの精度を比較すると，E_7 については言語特徴が最大，$E_{1,3-5,8,Total}$ については $VF + CF$: 言語とマルチモーダル共起特徴量で学習したモデルが最大，その他は全ての特徴量を用いて学習したモデルの精度が最大であった．

特に総合評価値 E_{Total} の精度は 0.24，精度の最大値は E_1 の評価値を予測した場合の 0.27 であった．テストデータに対する R^2 は最大で 0.3 未満とまだ十分ではない．一方で 67 エピソード全てのデータセットを使って，変数を p 値に従い追加しながら回帰モデルを学習し，E_{Total} について自由度調整済みの R^2 を計算したところ 0.72 であった．この結果から，予測精度は充分ではないものの，提案した特徴量から全データセットを高

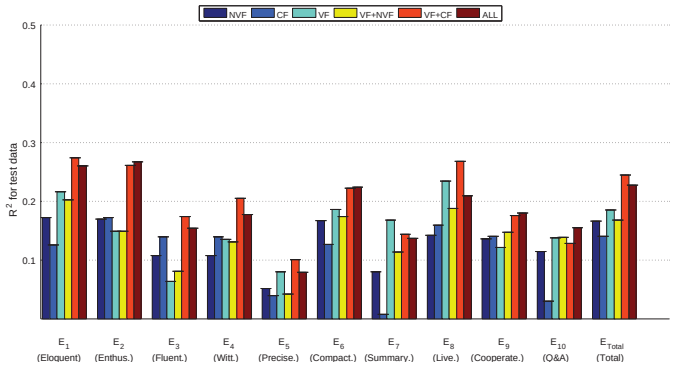

図 2: 説明評価値の回帰学習の結果

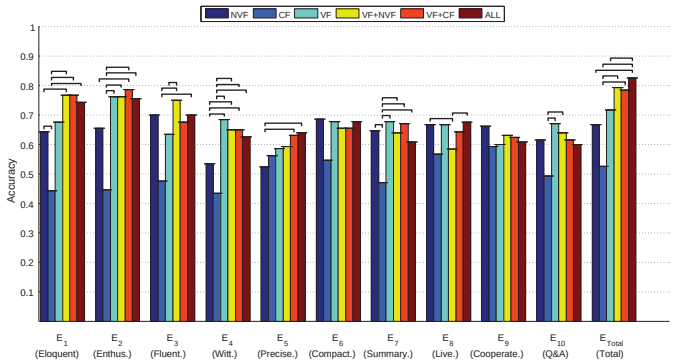

図 3: 説明評価値の高・低クラスを分類した結果 (⊓ は t 検定の結果，有意差 ($p < 0.05$) を認められたペアを示す．)

精度に説明するモデルを獲得できることが示された．

6.3 分類学習に基づく予測実験結果

中間値を閾値として用い，説明評価値を高レベルと低レベルの 2 クラスにカテゴリ化し，この高低レベルの分類タスクを行った．本タスクでは，訓練事例を増やすために，各エピソードをエピソード中の会話時間の中点を境に二分割した．この 2 分割されたサブエピソードには分割元のエピソードの評価値を割り当てた．言語・非言語・マルチモーダル共起特徴についても，サブエピソードごとに計算した．この操作により 134 個のサブエピソードデータセットを構築した．線形 SVM を用いて，10 分割交差検定を行った．訓練データセット内で交差検定を行い，SVM の C パラメータについて [0, 0.01, 0.1, 1] から最適な値を選択し，テストに用いた．

6.2 節と同様に 6 種類の特徴量セットを用意し，上記の評価実験を行い，分類精度を比較した．図 3 は各説明評価値の高・低レベルを分類した結果を示す．縦軸はテストデータの平均正答率である．図 3 における ⊓ は各モデル間の精度に対して t 検定を行った結果 $p < 0.05$ で有意差が認められたペアを示している．言語特徴 VF, 非言語特徴 NVF，マルチモーダル共起特徴 CF で学習したモデルを比較すると，$E_{2,4,7,10}$ について言語特徴で学習したモデルが有意に高い精度を得た．$E_{1-3,Total}$ について，言語特徴 VF, 非言語特徴 NVF，マルチモーダル共起特徴 CF いずれかを用いた場合よりも，特徴量を統合した場合の方が高い精度を得た．特に全ての特徴量で学習したモデルは，総合的な説明評価指標 E_{Total} のレベルを 82.7% で分類出来ており，言語特徴を用いたモデルの 71.7%，非言語特徴量を用いたモデルの 66.6%，マルチモーダル共起特量を用いた 52.5% を上回った．この結果より，抽出した多様なマルチモーダル特徴量は外部観察者により評価された説明の質に関する指標を高精度に推定可能であることが示された．

7 結論

本研究は説明の質を外部観測可能な言語・非言語情報より推定するモデルの構築に焦点を当てた．この目的のために，グループ説明タスクを収録し，複数の外部観察者による説明の質の評価を行い，説明評価付きマルチモーダルコーパスを新規に構築した．発話内容から取得される単語の品詞特徴，発話区間，韻律，手・頭部のジェスチャ，視線方向といった多様な非言語特徴，マルチモーダル共起特徴・インタラクション特徴を加えたマルチモーダル特徴量を抽出することによって，説明評価値の推定を試みた．

実験の結果，各種推定精度については向上する余地はあるものの，多様なマルチモーダル情報から説明の質を推定することの可能性を示すことが出来た．今後の課題は，ビジネス，教育における説明場面など他の性質を持つ説明タスクを設定し，同様の実験を行い，説明の質を推定する上で，重要な不変特徴量とタスクに依存した特徴量を特定することである．

謝辞 本研究は科研費若手研究（B）25730132，基盤研究 (B) 25280076,（C）15K00300 の助成による．

参考文献

[1] J. K. Burgoon, T. Birk, and M. Pfau, “Nonverbal behaviors, persuasion, and credibility,” *Human Communication Research*, vol. 17, no. 1, pp. 140–169, 1990.

[2] A. Vinciarelli, “Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling,” *IEEE Trans. on Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

[3] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, “Rolenet: Movie analysis from the perspective of social networks,” *IEEE Trans. on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[4] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, “Multimodal recognition of personality traits in social interactions,” in *Proc. of ICMI*, New York, NY, USA, 2008, pp. 53–60.

[5] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, “A nonverbal behavior approach to identify emergent leaders in small groups,” *IEEE Trans. on Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.

[6] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, “Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach,” in *Proc of ACM ICMI*, ser. ICMI ’14, New York, NY, USA, 2014, pp. 50–57.

[7] S. Okada, M. Bono, K. Takanashi, Y. Sumi, and K. Nitta, “Context-based conversational hand gesture classification in narrative interaction,” in *Proc of ACM ICMI*, 2013, pp. 303–310.

[8] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, ser. Psychology/cognitive science. University of Chicago Press, 1996.

[9] K. Otsuka, H. Sawada, and J. Yamato, “Automatic inference of cross-modal nonverbal interactions in multiparty conversations: ”who responds to whom, when, and how?” from gaze, head gestures, and utterances,” in *Proc. of ACM ICMI*, 2007, pp. 255–262.

[10] A. Tabensky, “Non-verbal resources and storytelling in second language classroom interaction,” *Journal of Applied Linguistics*, vol. 5, no. 3, pp. 321–348, 2012.

[11] A. Vahdatpour, N. Amini, and M. Sarrafzadeh, “Toward unsupervised activity discovery using multi-dimensional motif detection in time series,” in *Proc. of IJCAI*, ser. IJCAI’09, 2009, pp. 1261–1266.