The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

自己評定と他者評定のアノテーションを用いたインタビューにおける発話意欲レベルの推定

Attitude recognition of interviewee in interview on human robot interaction using both self and impression annotation

石原 卓弥 *1 Takuya Ishihara　長澤 史記 *1 Fuminori Nagasawa　岡田 将吾 *2 Shogo Okada　新田克己 *1 Katsumi Nitta

*1東京工業大学　情報理工学院
Tokyo Institute of Technology, School of Computing

*2北陸先端科学技術大学院大学　情報科学系
Japan Advanced Institute of Science and Technology, School of Computer Science

The goal in this research is to summarize the dialog contents which are collected through interaction with a robot interviewer by combining speech recognition and multimodal perception techniques. The key technique is to identify important statements in the dialog contents. We focus on using nonverbal behaviors including prosody, gesture and posture for the prediction of the important statements that user would like to emphasize. In this study, we collected two types of the annotation data set for important statements. One data is annotated by third party annotator (impression by outsider) and another data is annotated by interviewee (self answering). We investigated the difference of recognition accuracy in these two tasks using these annotations.

1. はじめに

本研究は対話システムを対話コンテンツ・知識を収集するために用いることで，個人の発信したい内容，聴衆が聞きたい内容を，対話を通じて獲得し，対話内容を抄録し，それを提示する技術に焦点を当てる．対話内容の収集・活用が行われる事例の一つとして，インタビューが挙げられる．本研究ではヒューマノイドロボットを用いたインタビューシステムを構築し，インタビュー対話内容を音声認識・自然言語処理により獲得し，重要なインタビュー内容を推定することを目的とする．提案するロボットシステムの要素技術は，(１）インタビュー相手（ユーザ）の発話態度・意欲に応じて，インタビュー方法を変えることで，ユーザに多くの事を語らせるインタビュー戦略の実装と（２）インタビューで得られた対話コンテンツにおける重要箇所を推定することである．本研究では（２）の対話コンテンツの抄録を作成するための検討を行い，その結果を報告する．本研究では，ロボットとユーザの対面対話を対象としているため，音声認識結果より得られたテキストデータだけでなく，対話時のユーザの韻律情報や上半身の動作情報を含め，重要箇所の推定に有効な特徴量の分析を行う．そこから，分析によって判明した有効な特徴量により機械学習を行い，重要箇所推定のためのモデルの構築を行う．

[8] では，重要箇所の推定に有効な特徴量を分類モデルのパラメータより分析し，重要箇所における人間の動作について明らかにした．本研究では，汎用的な重要箇所の推定を実現するために，複数のトピックを扱うインタビュー実験を行い，15 セッションのデータセットを収集し，インタビュー対象者の発話意欲の高い重要箇所の推定モデルを構築・評価する．特に，本研究では他者による発話意欲のアノテーションと，インタビュー当事者による自己アノテーションを用意して，その推定タスクにおける精度の差異と，有用な特徴量の比較を行う．

連絡先: 岡田将吾，北陸先端科学技術大学院大学，〒 923-1211 石川県能美市旭台 1 丁目 1，0761511201，okada-s@jaist.ac.jp

2. 関連研究

[1] では，議論参加者の注視行動，頭部動作，韻律情報といった非言語情報に着目しており，発言の重要度を F 値 0.7, 再現率約 0.7 の性能で検出できることを報告している．また，議論データについて，時間を 45 %に短縮した要約の生成が可能になったことを報告している．これらの研究に用いられている情報は非言語情報のみであり，言語情報は用いられていない．

Berna ら [2] は，テキストの分析に加え，音声分析と視覚的な動作の解析法を組み合わせることで，ミーティングのビデオの要約を作成する方法を提案した．重要なイベントの検出をビデオの局所的な輝度変化を分析することで行っている．この研究では TF-IDF 法に基づいたテキスト解析も行っている．この研究から得られたビデオ要約の作成法は，均一的にシーンを結合して得られる要約よりも重要なイベントをより検出することがわかった．しかしながら，この研究における題材は多人数が参加するミーティングであり，本研究とはシチュエーションが異なる．McCowan ら [3] は，韻律情報，画像特徴量を利用し，ミーティングにおけるシーンを発話や議論などのイベントから分割することに取り組んだ．音声情報が主となるようなミーティングであっても，視覚情報にはミーティングの分析に対する重要な情報が含まれていることを示した．本研究とは用いている特徴量のモーダル，また研究目的のシチュエーションが異なり，差異が存在する．

[4] では交渉対話において，対話相手をどの程度信頼できるかをマルチモーダル情報から推定している．この研究では，本研究と同様，対話参加者によるアノテーションラベルと，第三者によるアノテーションラベルの 2 種類を用意し，各アノテーションラベルを推定するために有効な特徴量が異なることを示している．インタビューにおける意欲の高い発言は，第三者の印象評価よりも，インタビューを受けた本人の方が正確に評価できる可能性があることから，本研究でも第三者・インタビューを受けた本人の 2 種類のアノテーションデータを用意し，これを推定する実験を行い，比較する．

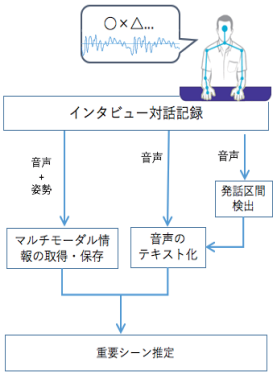

図 1: インタビューから得られたデータを処理し，重要シーン推定を行うための処理フロー

3. インタビューにおける発話意欲の推定

3.1 インタビュー環境

インタビュアーとして，アルデバランロボティクス社が開発した人型ロボットの Pepper[*1] を用いた．インタビュー対象者 (以下，対象者と呼ぶ) は Pepper との一対一の環境でインタビューに答えた．Pepper が質問を行い，対象者がその質問に答える形式でインタビューを進めた．なお，対象者が Pepper に対して質問を行うことは出来ない設定とした．インタビュー中は指向性マイクを対象者の頭部に取り付けることで，音声の記録を行った．また，対象者の前方に Kinect v2，Web カメラを設置した．Kinect v2 により実験中の対象者の動きをセンシングした．更に，Web カメラにより実験中の対象者の様子を撮影すると共に，俯瞰音声を取得した．本研究ではインタビューデータのアブストラクト作成に焦点を当てるため，インタビュー対話を正確に行う必要がある．よって，本実験では質問タイミングは実験管理者 (以下，管理者と呼ぶ) の遠隔操作によって制御した（WoZ 法を用いた）．インタビューから取得したデータの処理と，そこからの重要シーン推定を実現する概略図を図 1 に示す．

インタビューにより，Kinect v2 から対象者の行動データ，指向性マイクからの音声データ，ウェブカメラからの正面映像・俯瞰音声データが取得される．更に，音声データから音声認識によって対象者の発言内容を得る．取得したマルチモーダル情報 (韻律情報，動作情報，言語情報) から，インタビュー中の重要箇所の推定を行う．20〜60 歳の男女を対象としてインタビュー実験が行われ，収録件数は 30 件であった．世代と性別が均等になるように被験者を募集した．

3.2 マルチモーダルインタビューデータコーパス

本実験は一般人を対象としているため，対象者毎に答えられるトピックの知識に差がある．従って，質問セットは，[趣味，スポーツ，子育て，勉強，研究] のトピックから 2 つを選択し，それに合わせて質問セットを切り替える形とした．

Pepper の制御に [6] で開発された質問毎の発話意欲推定器を用いた．発話意欲の推定には韻律特徴量，動作特徴量が用いられている．質問に対する回答発話中のそれらの特徴量を分析することで，対象者がその質問に対して意欲を持った回答を行っていたかを推定する．

[8] で収集したデータセットに比べ，自由度の大きい実験環境でデータを収集している．対象者の属性を限定しないことで，対象者毎に挙動の大きさやその動作の分布は大きく異なることが予想される．加えて話題を 5 つから 2 つを選択する形にしたことで，対象者毎に話す内容が統一されない．更に，意欲推定器によって同じトピックを選んだ場合でも対象者に行われた質問は異なる．それらの要因が重要シーン推定にどのように影響するのか確認する．計測機器の不調により，ファイルの欠損が見られるインタビューもあったため，結果としてモデル構築の際には 30 件中，15 件分のデータを使用した．以後，15 件のデータをそれぞれ 2-1〜2-15 実験と呼ぶ．対象者の発話意欲が高いシーンを重要シーンとして取り扱う．任意のシーンの開始時刻，終了時刻をアノテーションし，それに該当する箇所を重要シーン，それ以外のシーンを非重要シーンとして扱う．第三者のアノテータに加えて，対象者本人にもアノテーションを行うように指示した．これを行うことで対象者本人が考える発話意欲の高いシーンと，第三者が考えるそれとの比較を行うことができ，発話意欲の包括的な理解が進むと考えられる．

アノテーションの際には，第三者のアノテータ，対象者本人共に，インタビュー中の全ての対象者の様子から相対的に見て，より話したがっている，もしくは聞いてほしいと対象者が考えながら話していると思われるシーン，つまり発話意欲が高いと考えられるシーンを記録するように指示した．

4. マルチモーダル特徴量

本研究では，Julius[5] を用いて対象者の発話ターン内の発話区間を検出し，それに従って指向性マイクから取得した音声データを分割した．Julius によって発話断片が生成された際の各発話断片の開始秒と終了秒をもとに，付与されたアノテーション区間と比較を行い，時間単位で 50 %以上重複する発話断片を重要発話断片，それ以外を非重要発話断片とする．図?? における赤枠の中が重要発話断片であり，それ以外が非重要発話断片である．なお，アノテーション区間が長く，複数の発話断片が 1 つのアノテーションに対応している場合，その全ての発話断片を重要発話断片とみなす．マルチモーダル特徴量は，いずれも発話断片単位で抽出した．なお，マルチモーダル特徴量は対象者毎に値の取るレンジが異なり，そのままでは分析の際に支障をきたす．従って，それぞれの対象者の特徴量毎に最大値を 1，最小値を 0 とする正規化，もしくは特徴量毎に標準化，つまり z 値 (標準化得点) を求めた．これにより，異なる人を同じ特徴量を基準として比較することが可能になる．

4.1 言語特徴量への変換

各発話断片について，Bing Speech API [*2] を用いて音声認識を行い，テキスト化した．そこから，Mecab を用いて形態素解析を行い．各発話断片に含まれる名詞数，動詞数，感動詞数，フィラー数，品詞数を抽出した．なお，実験 1 において取得された 4 件のデータを BingSpeechAPI によってテキスト化したものと，人間によって書き起こしされた文章の間の単語一致率を計測したところ，平均して 55.2 %の一致率となった．

品詞数の出現回数を記録した特徴量とは別に，音声認識結果から全実験で行われた品詞を抜き出し，出現頻度が高かったものを発話断片毎の出現回数を要素とする単語ベクトルに変換し，特徴量として用いた．なお，本研究ではインタビューという特性上，実験毎に話す内容が全く異なることから，単語ベク

*1 Pepper, https://www.softbank.jp/robot/consumer/products/

*2 Bing Speech API, https://azure.microsoft.com/ja-jp/services/cognitive-services/speech/

トルとする品詞はある程度汎用的に用いると予想される形容詞に限定した．

4.2 動作特徴量への変換

各発話断片について，動作特徴量として，頭部，肩部，肘頭部，手首，指先，手の動作の変化量を使用した．対象者の動作のセンシングには Kinect v2 を用いており，インタビュー中の対象者の身体動作が記録されている．

各発話断片は開始時刻，終了時刻を持つため，その時刻と Kinect v2 で記録されたデータとの対応を求め，各発話断片に対応する区間における Kinect v2 のデータを得た．その区間中の，フレーム間での動作の変化量の累計を求めた．その累計から各発話断片の各部位における変化量の平均値を求め，更にその値を用いて分散値を求めた．なお，これらの値はそれぞれ X 軸，Y 軸，Z 軸ごとのデータとなっており，更に頭部以外のデータは右腕，左腕の値をもつ．結果として，72 種類の特徴量を得た．

4.3 韻律特徴量への変換

対象者の口元に装着された指向性マイクから記録された音声データについて，Julius により発話断片単位に分割した．Julius によって検出された発話断片それぞれについて，Speech feature extraction code [*3] を用いてピッチ，エネルギー，メル周波数ケプストラム係数 (MFCC) を求めた．導出したそれらの値について，それぞれの最大値，最小値，平均値，標準偏差を特徴量として用いた．1 つの発話断片につき，ピッチ，エネルギー，MFCC 関連の最大値，最小値，平均値，標準偏差がそれぞれ求まり，特徴量とした．

5. 重要シーン推定モデルの構築

本研究では，マルチモーダル特徴量を用いることで，機械学習により重要シーン推定モデルの構築を行う．t 検定を行い，5 %水準で有意な特徴量を用いて機械学習を行い，分類モデルを構築した．

5.1 モデルと機械学習の概要

15 名のデータから，機械学習によりモデルを構築した．事前に SVM，Random Forest，XGBoost，RNN で分類実験を行った結果，XGBoost が最良の結果であったため，XGBoost の結果を記載する．XGBoost は kaggle をはじめとするデータ解析を競うサイトで最も使用されるアルゴリズムの 1 つであり，オーバーフィッティングを防ぐためにサブサンプリングや縮退化の仕組みも導入されている．

分類実験では，得られた特徴量をそのまま学習に用いる場合と，時刻 t の学習/テストを行う際に，時刻 $t-1$ における全ての特徴量，また時刻 t における全ての特徴量と時刻 $t-1$ でのそれとの差，変化の絶対値，変化率を特徴量として追加する場合の 2 つを試みた．時系列性を反映する特徴量を追加した場合，分類に用いる特徴量の数は通常の 5 倍となっている．以降，それぞれを Pre, Diff, Abs, Rate と呼称する．

実験 2 では，2-1〜2-15 の 15 回分のインタビューデータを用いるため，14 回の実験を学習データ，1 回の実験をテストデータとして，計 15 パターンの学習，テストを行った．学習の際には，分類正解率 （Accuracy）を基準として，XGBoost のハイパパラメータをチューニングした．言語・動作・韻律特徴量の寄与を分析するために，3 種類の各特徴量を組み合わせて，特徴量セットを 7 種類用意した．

*3 Speech feature extraction code, http://groupmedia.mit.edu/data.php

表 1: 対象者本人評価を用いた際の分類正解率 （モデルに時系列性を付与するための特徴量である Pre, Diff, Abs, Rate を学習に使用した場合と使用しない場合に分けて学習を行った.）

時系列特徴量あり	平均 Accuracy	時系列特徴量なし	平均 Accuracy
韻律＋動作＋言語	60.0 %	韻律＋動作＋言語	60.0 %
韻律＋動作	58.7 %	韻律＋動作	60.0 %
韻律＋言語	58.6 %	韻律＋言語	58.2 %
動作＋言語	59.3 %	動作＋言語	59.0 %
韻律のみ	58.3 %	韻律のみ	58.2 %
動作のみ	60.0 %	動作のみ	58.0 %
言語のみ	56.1 %	言語のみ	55.4 %

- 韻律特徴量＋動作特徴量＋言語特徴量
- 韻律特徴量＋動作特徴量
- 韻律特徴量＋言語特徴量
- 動作特徴量＋言語特徴量
- 韻律特徴量のみ
- 動作特徴量のみ
- 言語特徴量のみ

5.2 分類結果と考察

対象者本人評価による機械学習第 5.1 章で述べた機械学習の概要に基づき，発話断片の分類モデルを構築した．分類正解率を表 1 に示す．表 1 を見ると，実験 1 の際に確認された時系列を反映する特徴量を入れた場合のような大幅な分類正解率の向上は見られず，最高分類正解率は時系列の特徴量を入れた場合でも，入れない場合でも変わらないことがわかる．この理由として，アノテータが対象者本人であったことが影響を及ぼしている可能性が考えられる．つまり，第三者がアノテーションを行う場合は，対象者の動作や発話内容の変化などの行動・発言自体に注視する可能性があるが，対象者本人はアノテーションを行う時点である程度自分のインタビュー中の意欲の増減を把握しているため，そこまで行動や発言内容自体に注視する必要はない．更に，そういったアノテーションの特性上，外面には出ていないが内面では意欲があった箇所もアノテートされている可能性があり，本研究で用いたような外面の情報を強く反映する特徴量ではそれを捉えられなかった可能性がある．

更に，表 1 より，韻律モーダル単体で学習を行った場合をみると，発話断片間の特徴量の変化を反映する時系列の特徴量を加えたことでも分類正解率は 0.1 %しか向上していないことがわかった．3 つのモーダルの中でも特に，発話毎の行動の変化について考えられずにアノテーションされていた可能性が高い．

第三者評価による機械学習第 5.1 章で述べた機械学習の概要に基づき，発話断片の分類モデルを構築した．分類正解率を表 2 に示す．表 2 から判明したこととして 1 つ目に，実験 1 の際に比べて時系列性を反映する特徴量が強い影響を持っていないことがわかる．しかしながら，第 5.2 の結果に比べれば平均 Accuracy の変化は大きく，最大 Accuracy でも時系列性を反映する特徴量が最も良い結果を生んでいることがわかる．本研究における時系列的な特徴量をもとにした判断を行っている可能性が第 5.2 の結果に比べて多いことから，第 5.2 の結果のほうが時系列を反映する特徴量の影響を受けたと考えられる．

2 つ目に，表 2 の結果は，僅かではあるが表 1 の結果と比べて分類正解率が向上していることが判明した．この原因として

表 2: 第三者評価を用いた際の分類正解率 11（モデルに時系列性を付与するための特徴量である Pre, Diff, Abs, Rate を学習に使用した場合と使用しない場合に分けて学習を行った.）

時系列特徴量あり	平均 Accuracy	時系列特徴量なし	平均 Accuracy
韻律＋動作＋言語	62.1 %	韻律＋動作＋言語	61.9 %
韻律＋動作	62.0 %	韻律＋動作	61.9 %
韻律＋言語	62.4 %	韻律＋言語	62.0 %
動作＋言語	59.1 %	動作＋言語	58.8 %
韻律のみ	62.4 %	韻律のみ	62.0 %
動作のみ	55.6 %	動作のみ	56.4 %
言語のみ	59.4 %	言語のみ	58.8 %

は上記の時系列性の特徴量の有効性に加えて，アノテータの人数の影響が考えられる．表 1 の結果は対象者本人が行っていることから，アノテータ数としては 15 名となる．対して，表 2 の結果は第三者のアノテータ 1 名によるものである．本研究におけるアノテーションタスクは主観性が高いものであるので，アノテータによって評価尺度のばらつきが起きている可能性は十分考えられる．従って，1 名でのアノテーションが行われている表 2 のほうが，15 人の発話意欲を相対的に判定で来ており，結果として表 1 の結果に比べて良い結果を生んでいる可能性がある．

3 つ目に，実験 2 の第三者評価を用いた際のモデル構築では，韻律と言語を組み合わせたものが最も良い結果が得られることがわかった．単一の動作モーダルの特徴量による結果を見てみると，他の単一モダリティの結果に比べて精度が低下していることがわかる．つまり，動作特徴量が加わったことによって正解率が落ちてしまったと推測される．すなわち，第三者が他人のインタビュー動画を発話意欲という観点で評価する際には，その対象者の動作はほとんど確認せず，韻律によって発話意欲が判断されている可能性がある．

4 つ目に，表 2 より，単一モダリティでの学習では韻律情報が良い結果を得た．これは 2 つのモダリティを選択して学習した際にも同様の現象が確認できる．韻律を選択したものはそうでないものに比べて僅かに分類正解率が向上していることがわかる．第三者が他人の発話意欲を判断する際には，韻律特徴量に強く影響を及ぼすことが推察される．

5.3 対象者本人・第三者のアノテーションの比較

対象者本人評価と第三者評価のアノテーション結果を比較する．対象者本人がアノテーションしたデータと，第三者が行ったそれとの一致度を示すため，Cohen の κ 値を用いた．表 3 に，実験 2 の各セッションで計算した Cohen の κ 値を示す．[7] によると，Cohen の κ 値は 0.0〜0.2 で僅かに一致，0.21〜0.40 でまずまずの一致，0.41〜0.60 で中程度の一致，0.61〜0.80 でかなりの一致，0.81 以上で完全一致とされている．表 3 より，いくつかの実験においてはかなりの一致が見られており，平均としても 0.45 のスコアを取ることがわかった．つまり，発話意欲という尺度でみた場合，本人による評価と第三者によるそれは中程度に一致していると見なせる．つまり，評価者が本人，第三者を問わず，発話意欲を測る際にはエネルギーの変化やその大きさなどの指標が存在しており，それを感じたときに人は発話意欲を感じると推察される．

6. 結論

本研究では，重要シーンを (1) 発話意欲が高いと判断されるシーンとし，それぞれに対して言語・韻律・動作の 3 モーダルの特徴量を分析し，それを用いてモデルの構築を試みた．(1) の発話意欲について，対象者本人が考える発話意欲について最大 60.0 %，第三者が考えるそれについて最大 62.4 %の分類正解率を得た．時系列性を含む特徴量を用いることで，僅かながら分類正解率が向上したことから，時系列学習の有用性が示唆された．推定に重要な特徴量の分析結果は紙面の都合上除外したため，発表において詳細を述べる．

表 3: 実験 2 における対象者本人・第三者評価間の一致率.（[7] による基準で見ると，かなりの一致が 4 件，中程度の一致が 5 件，まずまずの一致が 4 件，僅かな一致が 2 件となっている.）

実験 ID	Cohen の κ 値
2-1	0.39
2-2	0.58
2-3	0.11
2-4	0.64
2-5	0.48
2-6	0.63
2-7	0.21
2-8	0.53
2-9	0.72
2-10	0.69
2-11	0.45
2-12	0.44
2-13	0.38
2-14	0.16
2-15	0.27
平均	0.45

参考文献

[1] 二瓶芙巳雄, 高瀬裕, 中野有紀子, ”非言語情報に基づくグループ議論における重要発言の推定 -グループ議論の要約生成に向けて-”, 電子情報通信学会論文誌 A, pp34-44, 2017.

[2] Berna Erol, Dar-Shyang Lee, Jonathan Hull, ”Multimodal Summarization of Meeting Recordings”, In proc. of IEEE ICME, 2003

[3] McCowan I, Gatica-Perez D, Bengio S, Lathoud G, Barnard M, Zhang D., ”Automatic analysis of multimodal group actions in meetings”, IEEE Transactions on Pattern Analysis Machine Intelligence, Vol. 27, No. 3, pp305-317, 2005

[4] Gale Lucas, Giota Stratou, Shari Lieblich, and Jonathan Gratch. 2016. Trust me: multimodal signals of trustworthiness. In proc. of ACM ICMI, 2016

[5] 河原達也, 李晃伸,”連続音声認識ソフトウエア Julius”, 人工知能学会誌, Vol.20, No.1, pp.41–49, 2005.

[6] 長澤史記, 石原卓弥, 岡田将吾, 新田克己, ”ユーザーの態度推定に基づき適応的なインタビューを行うロボット対話システムの開発”, 対話システムシンポジウム, 2017

[7] J. Richard Landis, Gary G. Koch, ”The Measurement of Observer Agreement for Categorical Data”, Biometrics Vol.33, No.1, pp159-174, 1977

[8] 石原卓弥, 長澤史記, 岡田将吾, 新田克己, ”マルチモーダル情報に基づくインタビューにおける重要シーンの推定”, Human Communication Group(HCG) シンポジウム, 2017