大阪大学マルチモーダル対話コーパス Hazumi 概要（対面収録版）

駒谷 和範
大阪大学 産業科学研究所

岡田 将吾
北陸先端科学技術大学院大学

2023/1/17　1902F4007 のビデオについて追記
2022/7/7　対面収録版である旨追記，表 1 追加，表 2 追記，等
2021/3/8　Hazumi1911 追加
2020/11/25　一部改訂
2020/8/9　初版

1　はじめに

Hazumi は大阪大学産業科学研究所で収集されたマルチモーダル対話コーパスである [1, 2]．一般から募集した人（実験参加者）と，Wizard-of-Oz (WoZ) 方式で動作するエージェントとが，いくつかの話題について目的を定めずに対話する様子が収録されている．

マルチモーダル対話システムの研究において，システムに対して一般のユーザがどのようにふるまうかを知るのは極めて重要である．しかしながら，このようなデータの収集にはまずシステムが必要であり，また個人情報を含むデータを収集することから各種の倫理的配慮が必要となるなど，研究を始める時点で既に多くの障壁がある．この障壁を緩和し，様々な分野の研究者がマルチモーダル対話システムの研究に参入できるようにしたいという願いから，本データを公開する．

これまでにも人対人の対話データはいくつか公開されているものの，人対システムの対話データは少ない．一般ユーザのふるまいは，人に対する場合とシステムに対する場合とで大きく異なり，一般ユーザがシステム開発者の期待通りにふるまうことはない．したがって，人対人ではなく，人対システムの対話に基づき，対話システムを設計することが必須である．

Hazumi には収集時期に対応したバージョン名を付与している．2017 年度（2017 年 12 月）に収集したコーパスは Hazumi1712 [3, 4]，2018 年度（2019 年 2 月開始）に収集したコーパスは Hazumi1902 [5]，2019 年度（2019 年 11 月開始）に収集したコーパスは Hazumi1911 [6] である．この文書では，対面で収録を実施したこの 3 バージョンについて説明する．オンラインで収録したバージョン（Hazumi2010, Hazumi2012, Hazumi2105）については，別文書[*1]を参照されたい．対面収録版の 3 バージョンでは全て，実験参加者が概ね 15 分間話す様子を，ビデオと Microsoft Kinect を用いて収録した．さらに Hazumi1911 では生体信号センサも使用した．本コーパスには，これらのセンサを用いて収録したデータとともに，これらに対して付与した各種のアノテーションや，そこから抽出した特徴量ファイル，事前と事後に行ったアンケートの結果が含まれている．

なお，Hazumi という名前は，話を弾ませることができるような対話システムを作りたいという願いから名付けたものである．

[*1] https://www.nii.ac.jp/dsc/idr/rdata/Hazumi/documents/HazumiOverviewOnline.pdf

表 1　収集データの概要（対面収録 3 バージョン）

	Hazumi1712	Hazumi1902	Hazumi1911
収集時期	2017 年 12 月	2019 年 2 月開始	2019 年 11 月開始
概要	人間とシステム（Wizard-of-Oz 方式）との対話 1 対話あたり 15 分〜20 分		
Wizard への指示	雑談 興味あり 3 話題となし 3 話題	雑談 「実験参加者が対話を楽しむ時間が長くなるように」	
参加者	29 名（男性 14 女性 15）	30 名（男性 10 女性 20）	30 名（男性 15 女性 15）
	20 代〜50 代	20 代〜70 代	
交換数	2,422	2,514	2,859
センサ	実験参加者のビデオ（顔＋上半身） Microsoft Kinect 収録データ（深度，姿勢，音声）		
	-	エージェントのビデオ	
	-		生体センサ（Empatica E4）
交換ごとのアノテーション	興味度（3 段階；3 or 6 名）	-	
	第三者による心象評定（7 段階；5 名） 第三者による話題継続（7 段階；5 名）		
	-	実験参加者本人による心象評定（7 段階）	
	システム発話，システム発話の対話行為，実験参加者の発話の書き起こし		
対話全体のアンケート	-	実験参加者本人による事前・事後アンケート（8 段階；18 項目） Wizard による事前・事後アンケート（8 段階；3 項目）	
	- -		性格特性（TIPI-J） 人間による操作に気づいたか

2　収集環境

実験参加者と対話するシステムは，Wizard-of-Oz (WoZ) 方式で動作させた．つまり人間のオペレータが，遠隔から，実験参加者の様子を見ながら専用のインタフェースを通じて，システムの応答を選択した．システムを遠隔から人間が操作していたことは，実験前や実験中には実験参加者に開示せず，実験終了後に開示した．

各バージョンごとの差異を含む，収集データの概要を表 1 に示している．

Wizard への指示は，Hazumi1712 と Hazumi1902 以降とで異なる．Hazumi1712 では，話題に対する興味の有無推定のためのデータを得ることを目的としていたため，興味のある話題とない話題が半数ずつになるように設定した．このうえで，実験参加者が興味がなさそうにしている場合でも，Wizard 役はその話題を継続するようにした [3, 4]．これに対して Hazumi1902 や Hazumi1911 では，実験参加者が対話を楽しんでいる時間が長くなるように，Wizard は発話を選択した．具体的には，実験参加者が興味がなさそうな場合に話題を変えたり，実験参加者が興味を持った様子で積極的に話している際には Wizard は聞き役に回るようにした [5]．

実験参加者は，一般から報酬付きで公募された．Hazumi1712 では 20 代から 50 代の男女 29 名（男性 14 名，女性 15 名）[*2]，Hazumi1902 では 20 代から 70 代の男女 30 名（男性 10 名，女性 20 名），Hazumi1911 では 20 代から 70 代の男女 30 名（男性 15 名，女性 15 名）である．研究の意義や実験参加者の権利（いつでも実験参加を撤回できることなど）を説明した同意書に同意した者のみからデータを収録した．同意書には，データ利用に関する契約が交わされることを前提として，研究者に対して研究開発目的でデータを配布できることが明記されている．顔映像などの学会発表での表示については，同意された部分のみ利用可能である．

収録されたデータの中には，住所などの個人の属性が特定される可能性がある表現を含む発話や，個人の信条嗜好を表明している発話も一部存在した．このような部分については，配布に際して，音声を無音化したり，書き起こしテキストを伏字にするなどした．

[*2] 30 名からデータを収集したが，1 名分は機材トラブルのため収録できていなかった．

YYMM	コーパスのバージョン．1712，1902，1911 のいずれか．
G	実験参加者の性別．M（男性）もしくは F（女性）．
AA	実験参加者の年代．20 から 70 まで．
NN	上記 7 文字と合わせて ID となるように付番．

図 1　ファイルの命名規則

3　データの仕様

データはその入手方法において，以下の 2 種類に大きく分けられる．

- NII IDR（国立情報学研究所 情報学研究データリポジトリ）[*3]から配布されるもの．データ利用に関する契約を交わした利用者のみに対して配布される．
  - ビデオで収録したデータ（3.1 節）
  - MS Kinect で取得したデータ（3.2 節）
- Github からダウンロードできるもの．
  - 生体信号データ（3.3 節）
  - 閲覧用 ELAN ファイル（3.4 節）
  - 実験用ダンプファイル（3.5 節）
  - アンケートデータ（3.6 節）

  URL は以下のとおりである．

  `https://github.com/ouktlab/Hazumi1712/`

  `https://github.com/ouktlab/Hazumi1902/`

  `https://github.com/ouktlab/Hazumi1911/`

ファイルは実験参加者ごとに分かれている．ファイル名には実験参加者 ID が使用されており，図 1 に示される YYMMGAANN の 9 文字（例： 1712F2006）で構成される．例えば 1712F2006 は，「Hazumi1712 の，ある 20 歳代女性のデータ」を意味する．

3.1　ビデオデータ

1. 実験参加者ビデオ
   エージェントと対話する実験参加者を，正面から 30fps で撮影したビデオである．mp4 形式である．対話開始前や対話開始後に数分間程度の余計な区間が含まれるが，これは 3.4 節で説明する ELAN 用ファイルとの時間的な整合を取るためである[*4]．
   画像や映像を，論文や学会発表で使用するのは，同意がなされた項目のみ可能である．具体的には，ビデオデータに同梱される同意情報（Hazumi{1712,1902,1911}_consent.pdf）を参照のこと．
2. エージェントビデオ
   Hazumi1902 と Hazumi1911 では，対話中に表示されていたエージェントを，記録のために撮影した．mp4 形式である．実験参加者ビデオや Kinect データとの間で，時間同期は特に取られていない．

[*3] `https://www.nii.ac.jp/dsc/idr/`

[*4] 1902F4007 のビデオは，対話のクロージング部分が収録されていない．

SpineBase, SpineMid, Neck, Head,
ShoulderLeft, ElbowLeft, WristLeft, HandLeft, ShoulderRight, ElbowRight, WristRight, HandRight,
HipLeft, KneeLeft, AnkleLeft, FootLeft, HipRight, KneeRight, AnkleRight, FootRight,
SpineShoulder, HandTipLeft, ThumbLeft, HandTipRight, ThumbRight

図 2　Kinect における関節部位の変数名

3.2　MS Kinect データ

実験参加者の正面に Microsoft 社製の Kinect V2 (以降 Kinect) を配置し，Kinect 内蔵の深度センサと Microphone アレイにより，実験参加者の音声と動作を収録した．音声データ，深度画像データ，姿勢データが含まれる．各データの取得には Kinect V2 SDK[*5] を使用した．

音声データ　Audio/Audio1.wav（16kHz，32 ビット）として録音されている．さらに，おおよそ 256 サンプルを 1 フレームとして，水平面音源方向とその信頼度が Audio/AudioTimeStamp.csv に記録されている．このファイルは，1 列目が時刻，2 列目が音源方向，3 列目がその信頼度である[*6]．音源方向は，Kinect v2 の中心正面が 0° であり，単位はラジアン（弧度法）である．水平面方向左右に ±50° の範囲（対応するラジアンの範囲は ±0.87）で取得できる．信頼度は，0.0 と 1.0 の間の値を取り，数値が大きいほど音源方向の推定結果が信頼できることを表す．

深度画像データ　10〜30 フレーム毎秒で，Depth/DepthX.png (X はフレーム番号) として保存されている．Depth/DepthData.csv の 1 列目は各深度画像の取得時刻，2 列目はファイル名に対応したフレーム番号である．

姿勢データ　Body/BodyData.csv に格納されている．時刻情報はこの中の SysTime 列にある．この各時刻における，実験参加者の関節の位置座標の推定値が出力される．学習済みの姿勢推定モデルによる，全身の 25 個の関節の 3 次元位置座標データが保存されている．つまり，フレームあたり計 75 個の値がある．図 2 に各関節の部位の変数名を示す[*7]．なお，Kinect は机上に設置されていて実験参加者の下半身は計測可能範囲に入っていないため，下半身の関節位置，例えば足（Foot）や臀部（Hip）などの位置の推定結果は無視してよい．

ここで本データセットで使用されている時刻について説明する．以下の (A) から (C) の 3 種類がある．単位はいずれもミリ秒である．

(A) Kinect デバイス上の時刻： Kinect 収録日の午前 0 時が 0
(B) Kinect 収録時の時刻： Kinect 収録開始時刻（wav ファイルの開始時刻）が 0
(C) 実験参加者ビデオ上の時刻： ビデオの録画開始時刻が 0

上述した Kinect データの各 csv ファイル中の時刻は (A) で表記されている．なお Kinect V2 SDK では，収録時の PC の負荷などによりフレームレートが変わるため，各 csv ファイルの間で，時刻の値は必ずしも一致しない．また AudioTimeStamp.csv では同じ時刻が複数回出力されていることもある．

ビデオと Kinect は異なるデバイスで収録したため，収録時点では時刻の同期は取れていない．このため，事後的に人手で，これらの間の同期を取った．この情報は，実験用ダンプファイル（3.5 節）に含まれている．

[*5] https://www.microsoft.com/en-us/download/details.aspx?id=44561
[*6] ただし，本データでは発話者は 1 名で音源は 1 つであるため，これらの情報は無視してよい．
[*7] 全身に各関節部位の変数名をマップした模式図:
https://social.msdn.microsoft.com/Forums/sqlserver/ja-JP/c818387d-8717-48a9-b562-738e9e0b69e5/joints-in-kinect-v2?forum=kinectv2sdk

実験用ダンプファイル中の kinectstart(exchange), kinectend(system), kinectend(exchange) という列は，上記 (B) の時刻である．また start(exchange)，end(system)，end(exchange) は上記 (C) の時刻である．これらの時刻情報については 3.5.8 節で詳述する．

AudioTimeStamp.csv の 1 行目にある (A) での時刻は，(B) の時刻での 0 秒に概ね対応する．このため，この時点からの差分を計算することで，実験ダンプファイル内の時刻と，Kinect の各 csv ファイル内の時刻との対応を得ることができる．このようにして，実験用ダンプファイル内のある時点に最も近い，AudioTimeStamp.csv，DepthData.csv，BodyData.csv 内の時刻を同定できる．

なお Hazumi1902 では，閲覧用 ELAN ファイルの閲覧時のズレを解消するために，上述の同期を取った後に，いくつかのビデオファイルを約 1 秒前後短くした．このため，新たにビデオデータから特徴量を抽出する際には，厳密には新たに同期を取り直す必要がある．

3.3 生体信号データ

Hazumi1911 でのみ収録されている．

生体信号は人の感情状態を反映しているという知見が多くの研究で得られている．システムとの対話におけるユーザの心象状態を推定するうえで，生体信号データから抽出した特徴量は有効である可能性がある．そこで，Empatica 社製の E4 wristband（以降 E4）を用いて，対話中の生体信号データを取得した．E4 はワイヤレスのリストバンド型生体センサであり、実験参加者への負担なく生体信号データを取得可能である．皮膚に接触する 2 つの電極および赤色／緑色 LED から構成され，皮膚電位および心拍などを測定できる．具体的には以下の各データが格納されている．

- 皮膚電位データ（EDA）
  4Hz で計測され，単位は μSiemens である．
- 容積脈波データ（BVP）
  64Hz で計測される．2 つの LED から得られた光学的情報に基づき，E4 独自のアルゴリズムにより出力される．
- 心拍データ（HR）
  容積脈波データから算出され，1Hz で出力される．
- 皮膚温度データ（TEMP）
  4Hz で計測され，単位は °C である．
- 加速度データ（ACC）
  32Hz で計測され，重力加速度 (1/64g) を単位として得られる．csv ファイルの 1，2，3 列目が x，y，z 軸に対応する．

全て csv ファイルとして出力され，1 行目に E4 データ測定開始時間（UNIX 時間），2 行目に各データのサンプリング周波数，3 行目以降に測定値が記録される．なお，静的条件下であれば容積脈波データから拍動間隔データ（inter-beat interval, IBI）も算出されるが，今回は実験参加者の動作を含む自然な対話条件下での測定であったため，拍動間隔データの多くは欠損していた．

生体信号データに含まれる時刻は Unix 時間で記載されており，Kinect で収録された時刻と対応している．

3.4 閲覧用 ELAN ファイル

アノテーションや書き起こしを全て含んだ eaf（ELAN annotation fomat）ファイルである．アノテーションツール ELAN[*8] 上で，3.1 節で説明した実験参加者ビデオを読み込んで使用する．ELAN 5.9 で動作を確認

[*8] https://archive.mpi.nl/tla/elan

表 2　システム発話の対話行為タグ（二重線以下の 3 つは Hazumi1911 以降で追加）

記号	内容	例文
qy	Yes-No 疑問文	スポーツはよくするんですか？
qw	Wh 疑問文	どこで食べられたんですか？
pa	肯定的回答	そうなんですか．一度乗ってみたいものですね！
na	否定的回答	すみません，知らないです．
oa	その他の回答	そうですか，いいものが見つかるといいですね．
op	開始	これから，旅行について話しましょう！
io	情報提供	人気漫画「ナルト」をテーマにした作品で，来年の夏頃に公演されるそうです．
su	提案	写真を撮る場所として，大阪の箕面の滝や和歌山の白浜などがオススメですよ！
th	感謝	きょうは実験に参加いただいてありがとうございます．
fu	了解	沖縄ですか （働きかけと応答の対の後に続く了解部分 [*10]）
no	エラー	（交換への分割が正しく行われていない箇所など）

している．

付与単位として，システムの発話と実験参加者の発話の対である交換 (exchange) を設定している．具体的には，システム発話開始時刻から，次のシステム発話開始時刻までを一交換としている．WoZ 方式での収集であるため，システムの発話開始時刻（つまり，操作役の Wizard が発話開始ボタンを押した時刻）がログに記録されている．交換の終了時刻は次のシステムの発話開始時刻とした．このように機械的に交換を認定した．

この付与単位に対して，eaf ファイルの各層に，以下の各節で説明するアノテーションがなされている．

なお，一部の交換への心象アノテーションや話題継続アノテーション，興味度アノテーションの値として，e や E が付与されている場合がある．これは error を意味する．これには，データ収集時のシステムの不具合などにより，発話開始ボタンを押したにも関わらずシステムが発話しなかった箇所などが含まれる．この場合ユーザは応答していないため，7 段階の数値や o/t/x に代えて，上記の記号を付与した．

3.4.1　実験参加者の発話の書き起こし

その交換に含まれる実験参加者の発話を，人間が聞いて書き起こしたテキストである．`user_utterance` というレイヤに記入されている．各交換内では，概ね以下の基準に基づいて書き起こした．

- 句読点は用いない
- フィラーは (F) で囲む（例：(F えーと)）
- 言い誤り・言い直しなどは (D) で囲む（例：(D ちょ）丁度いいです）
- 単位区切り記号は | (半角縦棒)
- 単位区切りの基準は 0.2 秒以上の無音（直観的には「息継ぎしたところ」）
- 明らかな文の終わりと思われるところは無音が短くても区切る

これは日本語話し言葉コーパス（CSJ）[*9]の基準のうち，基本的な部分を参考にしたものである．

3.4.2　システム発話とその対話行為

その交換のシステム発話である．これは，WoZ のインタフェース上で，Wizard が選んだものを記録したものである．`sys_utterance` というレイヤに記載されている．

さらに，システム発話の対話行為が `dialogue_act` というレイヤに付与されている．対話行為は参考文

[*9] `https://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/CSJ_rep.pdf`

[*10] 記号 fu は Follow-Up の頭文字である．例えば，システム：「どこか旅行に行きましたか？」，ユーザ：「沖縄に行きました」，システム：「沖縄ですか」の 3 発話目がこれにあたる [7]．表層的には qy にも見えるが，機能としてはユーザに Yes/No の回答を求めてはいないため，qy とはしない．

献[*11] をベースに，これを簡略化した 8 種類を基本としている（表 2）．ただし Hazumi1911 では，表 2 の二重線より下の 3 種類が追加され，合計 11 種類となっている．対話行為は作業者 1 名で簡易的に付与し，別アノテータとの一致などは検証していない．発話中に二つ以上の機能がある場合は，より後に出てくる機能を，その発話の機能とみなした．例えば，「今現在，「ダフィン」が流行っているようですよ．ご存知ですか？」には情報提供と Yes-No 疑問文の 2 つの機能が含まれるが，この例の発話の対話行為は，後に出てくる Yes-No 疑問文 qy とした．

3.4.3 心象アノテーション

実験参加者がシステム発話を受けてどのように感じているかを，人手で付与したものである．各交換に対して 7 段階で付与した．1 をネガティブ，7 をポジティブとした．ポジティブの例として，楽しい，話し続けたい，満足などを示し，ネガティブの例として，楽しくない，話し続けたくない，不満，困惑などを示した．付与者は，対話のビデオを前から順番に見ながら，各交換に対して付与した．

付与主体によって，以下の 2 種類のアノテーションがある．

第三者心象 アノテータ 5 名が，事後に第三者視点から付与した結果である．これらは `UI_XX` というレイヤに記入されている[*12]．`XX` は 2 文字のアノテータ ID である．

振り返りアノテーション Hazumi1902 と Hazumi1911 で付与されている．対話終了直後に，実験参加者本人に対して，各交換ごとに，撮影したビデオを見ながら逐一印象を尋ねることで付与した．`UI_self` というレイヤに記入されている．

3.4.4 話題継続アノテーション

自分がシステム役だったとした場合に，実験参加者の回答を聞いた後，次に話題を変えようと思うかどうかを人手で付与したものである．各交換に対して 7 段階で付与した．1 が「話題を変える」，7 が「この話題を続ける」を表すとした．5 名のアノテータにより付与し，`TC_XX` というレイヤに記入されている[*13]．`XX` は 2 文字のアノテータ ID である．

さらに，自分の付与結果（話題を変えるか否か）と，収録されているシステムの実際の次発話に乖離があった場合に，自分がシステム役ならどのような発話をするのかを，自由に記述した．これは 1 対話あたり 5 箇所を目処に記述されており，全ての交換に記述されてはいない．5 名のアノテータにより付与し，`TC_XX_`**（自由記述）**というレイヤに記入されている．`XX` は 2 文字のアノテータ ID である．

3.4.5 興味度アノテーション

実験参加者が，現在の話題に興味を持っているか否かを，第三者が人手で付与したものである．各交換に対して，興味あり（o），不明（t），興味なし（x）の 3 段階で付与した．

Hazumi1712 のみで付与されている．アノテータは 6 名または 3 名である．`Int_XXX` というレイヤに記入されており[*14]，`XXX` は，アノテータ ID を表す 3 から 5 文字の記号列である．

3.5 実験用ダンプファイル

実験用ダンプファイルは，前節までで説明した情報から特徴量を抽出し，簡便に機械学習の実験が行えるようにしたファイルである．実験参加者ごとに一つの CSV ファイルがあり，ファイル名は「実験参加者 ID.csv」である．各行が 3.4 節で述べた交換，各列がアノテーションや特徴量に対応する．また，閲覧用 ELAN ファイ

[*11] `https://web.stanford.edu/~jurafsky/ws97/manual.august1.html`
[*12] UI は User Impression の頭文字である．
[*13] TC は Topic Continuance の頭文字である．
[*14] Int は Interest を表す．

ル（ビデオファイル）と，MS Kinect データとの，時間同期を示すための列も含まれている（3.5.8 節）．

特徴量は，各センサから得た，ユーザの音声，映像，言語情報（音声認識結果）などから抽出されたものである．特徴量セットは，論文 [8] に記載されている方法で抽出されたものである．これを用いて，例えば，上記の各センサから得られた実数値の特徴量を入力とし，興味度・心象・話題継続アノテーションの値（連続値とカテゴリ値）を出力とした機械学習実験が行える．

なお，7 名分（1712F4002, 1712M5003, 1902F4007, 1902M4003, 1911F7001, 1911M3001, 1911M3002, 1911M5003）については，ダンプファイルは存在しない．したがってダンプファイルは，Hazumi1712 で 27 人分，Hazumi1902 で 28 人分，Hazumi1911 で 26 人分である．1712F4002 では，Kinect で取得した Audio 関係のファイルの中身が消失していた．1712M5003 と 1902F4007，1902M4003 については，作成時にシステムログと他のファイルとの間で不整合があった．1911F7001，1911M3001，1911M3002，1911M5003 については顔表情特徴量データが一部欠損したためダンプファイルを生成しなかった．

3.5.1 韻律特徴量

実験参加者の発話から openSMILE[*15]を使用して韻律特徴量を抽出した．INTERSPEECH 2009 Emotion Challenge feature set (IS09) [9] で使用された特徴量を用いた．計 384 次元の特徴量で構成されている．

3.5.2 顔表情特徴量

ビデオカメラの映像から OpenFace [10] を使用して，ランドマーク特徴量と Action Unit 特徴量の 2 種類の顔表情特徴量，計 66 次元を得た．

ランドマーク特徴量 48 次元は以下の手順で得た．まず OpenFace によりランドマーク（顔の特徴点）の 2 次元座標を求めた．具体的には，目の周りの 4 点，口の周りの 4 点，眉の周りの 4 点の合計 12 点である．これら 12 点それぞれについて 4 種類の統計量を求めた．具体的には，速度の絶対値の最大値，平均値，標準偏差と，加速度の絶対値の最大値である．速度は，フレーム t における座標データを $c(t)$，フレームインターバルを定数 I（1/30 秒）として，フレーム間速度の絶対値 $v(t) = \frac{||c(t+1)-c(t)||}{I}$ として求めた．同様にフレーム間加速度の絶対値は，$a(t) = \frac{|v(t+1)-v(t)|}{I}$ で求めた．

Action Unit 特徴量 18 次元として，顔表情を記述する動作単位である Action Unit (AU) の各 Unit を用いて特徴量とした．OpenFace には，18 種類の AU の有無をフレームごとに検出する学習済みモデルがある．このモデルによる検出結果を用い，交換内でそれぞれの Action Unit が検出されたフレームの割合を特徴量とした．

3.5.3 動作特徴量

Kinect V2 から得られる上半身の関節の 3 次元座標データをもとに，動作特徴量計 20 次元を抽出した．ここでは頭部と両肩と両手の計 5 箇所の座標データを用いた．このそれぞれについて，顔表情特徴量におけるランドマーク特徴量と同様の 4 種類の統計量，つまり速度の絶対値の最大値，平均値，標準偏差と，加速度の絶対値の最大値を計算して用いた．

3.5.4 生体特徴量

Hazumi1911 では収録した生体信号データを用いて，生体特徴量計 4 次元をダンプファイルに含めた．具体的には，交換内での，皮膚電位（μSiemens）の平均値および標準偏差，および，心拍の平均値および標準偏差である．この特徴量は論文 [11] に記載されている方法で抽出されたものの一部である．

[*15] https://www.audeering.com/opensmile/

3.5.5 言語特徴量

ユーザの発話内容から言語特徴量を抽出した．

Hazumi1712，Hazumi1902 ではユーザ発話の音声認識結果として得られたテキストから特徴量を抽出した．音声認識には Google Speech API を用いた．Hazumi1911 では，ユーザ発話の書き起こしテキスト（3.4.1 節）から特徴量を抽出した．

各ユーザ発話に対し，日本語形態素解析器 MeCab [12] を使用して形態素解析を行い，1 発話内に含まれる名詞，感動詞，形容詞，副詞の数，および各単語の出現回数を表す Bag-of-Words（BoW）を言語特徴量とした．Hazumi1712 では 951 次元，Hazumi1902 では 972 次元，Hazumi1911 では 2601 次元である．

3.5.6 対話的特徴量

対話という状況から得られる特徴量を，対話的特徴量計 13 次元として抽出した．まずシステムログをもとに，システム発話の単語数（lenS），ユーザ発話の単語数（lenU），システム発話とユーザ発話との単語数の差（lenS−lenU）を特徴量とした．

また Hazumi1712，Hazumi1902 では，システム発話終了からユーザ発話開始までの時間（reaction），その交換の時間長（ミリ秒）も特徴量に加えた．さらに，3.4.2 節で述べたシステム発話の対話行為 8 種類を，8 次元の one-hot ベクトルで表現して特徴量に加えた．

3.5.7 アノテーション値

4 つのアノテーション結果が含まれている．つまり，ユーザ本人による心象 (self sentiment: SS)，第 3 者によるユーザ心象 (third sentiment: TS)，興味度 (interest: IN), 話題継続 (topic continuance: TC) である．また，それらの値を三段階でラベル化した離散値（2 が高群, 1 がニュートラル, 0 が低群）が，それぞれのラベルに `_ternary` をつけた列に格納されている．アノテーション内容の詳細については 3.4 節を参照のこと．

Hazumi1712 には上記のうち TS，IN，TC が，Hazumi1902 と Hazumi1911 には SS，TS，TC が，それぞれ含まれている．なお，Hazumi1712 での IN のうち，3 名でアノテーションを行ったユーザのデータについては，4 人目から 6 人目の列（IN4 から IN6）には no という値を入れている．

3.5.8 時刻情報

実験参加者ビデオデータの時刻表現（3.2 節での (C) の時刻）と，Kinect データの時刻表現（3.2 節での (B) の時刻）との両方で，各行に対応する交換の時刻が記載されている．単位はミリ秒である．

ファイル中の各列，start(exchange), end(system), end(exchange) はそれぞれ，実験参加者ビデオデータの時刻表現における，システム発話の開始時刻，システム発話の終了時刻，次のシステム発話の開始時刻である．これらの値は，実験参加者ビデオの録画を開始した時点を 0 とした場合の時刻であり，システムログとビデオとの対応を人手で与えて算出した．システム発話の開始時刻は，交換の認定と同様，Wizard が発話開始ボタンを押したタイミングである．システム発話の終了時刻はシステムログから取得した．

同様に，kinectstart(exchange)，kinectend(system)，kinectend(exchange) は，Kinect データの時刻表現でそれぞれ，システム発話の開始時刻，システム発話の終了時刻，次のシステム発話の開始時刻である．これらの値は，Kinect の収録を開始した時点を 0 とした場合の時刻である．

3.6 アンケートデータ

実験参加者（一部 Wizard）から取得したアンケートデータが，Hazumi1902 と Hazumi1911 のみに存在する．Hazumi1712 には存在しない．

3.6.1 Hazumi1902

実験参加者と Wizard の双方に対して，実験開始前と実験終了後にそれぞれ，会話者の対人コミュニケーション認知に関する測定項目 18 項目 [13] に関するアンケートを実施した．各項目は 8 段階である．

この結果は 1902questionnaire.xlsx に保存されている．このファイル内には，実験参加者（実験前），実験参加者（実験後），Wizard（実験前），Wizard（実験後）の 4 つのタブがある．このそれぞれにおいて，実験参加者は上記の 18 項目，Wizard はこれを簡略化した 3 項目に対して，8 段階で回答した結果が記録されている．この際に使用した質問紙が 1902questionnaire-items.pdf である．

3.6.2 Hazumi1911

Hazumi1902 でのアンケートに加え，実験参加者に対して対話後に，以下の 3 つの内容を新たに尋ねた [6]．

1 点目として，「近い将来，今回の収録で用いたような，雑談ができる AI を使ってみたいですか？」という質問に対して，7 段階の評点で回答してもらった．評点は，1 が「使ってみたい」，7 が「使いたくない」とした．さらに理由があれば自由に記述してもらった．

2 点目として「今回使用したメイちゃんは，実は AI ではなく，別室にいる人間が操作していました．そのことに気づいていましたか？」という質問に対して，7 段階の評点で回答してもらった．評点は，1 が「気づいていた」，7 が「気づかなかった」とした．またその補足や実験全体に対する意見や感想についても記述がある．

3 点目として，各実験参加者のパーソナリティ特性を調査した．具体的には，ビッグファイブ [14] の 5 特性を 10 項目で測定する Ten Item Personality Inventory (TIPI) の日本語版 TIP-J[15] の質問文を利用して，実験参加者に自分自身の性格について尋ねた．

Hazumi1902 と同じ項目のアンケート結果に加え，上記 3 点の結果を加えたものが，1911questionnaire.xlsx に保存されている．このファイル内には，Hazumi1902 と同様の 4 つのタブに加え，「記述式」というタブに上記の 1 点目と 2 点目が，「性格特性」というタブに上記の 3 点目の結果が記録されている．使用した質問紙は 1911questionnaire-items.pdf である．

4 問い合わせ先

ご意見・ご質問などありましたら下記までお知らせください．

〒567-0047 大阪府茨木市美穂ヶ丘 8-1
大阪大学産業科学研究所
駒谷 和範
Email komatani@sanken.osaka-u.ac.jp

謝辞

本コーパス作成の一部は，物質・デバイス領域共同研究拠点における「人・環境と物質をつなぐイノベーション創出ダイナミック・アライアンス」共同研究プログラムの助成を受けました．本コーパス作成の始点となった，人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD)「人システム間マルチモーダル対話共有コーパス構築グループ」のメンバー各位に感謝します．また，データ収集に協力いただいた方々や，データを提供くださった実験参加者の皆様に感謝します．

参考文献

[1] 駒谷和範. マルチモーダル対話コーパスの設計と公開. 日本音響学会誌, Vol. 78, No. 5, pp. 265–270, 2022.

[2] Kazunori Komatani and Shogo Okada. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, 2021.

[3] 駒谷和範, 岡田将吾, 西本遥人, 荒木雅弘, 中野幹生. 配布可能なマルチモーダル対話データの収集とアノテーション不一致傾向の分析. 人工知能学会研究会資料, SIG-SLUD-B802-08, pp. 45–50, 2018.

[4] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. Multimodal dialogue data collection and analysis of annotation disagreement. In *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*, 2019.

[5] 駒谷和範, 岡田将吾. 複数の主観評定を付与した人システム間マルチモーダル対話データの収集と分析. 電子情報通信学会技術報告, Vol. 119, No. 179, HCS2019-33, pp. 21–26, 2019.

[6] 駒谷和範, 岡田将吾, 堅田俊. マルチモーダル対話コーパス hazumi 公開と生体信号を含む新規データ収集. 人工知能学会研究会資料, SIG-SLUD-C002-35, pp. 170–177, 2020.

[7] 荒木雅弘, 伊藤敏彦, 熊谷智子, 石崎雅人. 発話単位タグ標準化案の作成. 人工知能学会論文誌, Vol. 14, No. 2, pp. 251–260, 1999.

[8] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2019.

[9] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 312–315. ISCA, 2009.

[10] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proc. International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 59–66. IEEE Computer Society, 2018.

[11] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. Is she truly enjoying the conversation?: Analysis of physiological signals toward adaptive dialogue systems. In *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2020.

[12] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 230–237, 2004.

[13] 木村昌紀, 余語真夫, 大坊郁夫. 感情エピソードの会話場面における表出性ハロー効果の検討. 感情心理学研究, Vol. 12, No. 1, pp. 12–23, 2005.

[14] R. Lewis Goldberg. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, pp. 1216–1229, 1990.

[15] 小塩真司, 阿部晋吾, Pino Cutrone. 日本語版 Ten Item Personality Inventory（TIPI-J）作成の試み. パーソナリティ研究, Vol. 21, No. 1, pp. 40–52, 2012.