一人称・三人称視点対話収録システムと エゴセントリック津軽弁音声対話コーパスの構築

阪井 瞭介 [1,3,*]　江 舒婷 [2,3,*]
郭 傲 [2]　高道 慎之介 [1,4]　小川 哲司 [5]　東中 竜一郎 [2,3]
[1] 慶應義塾大学　[2] 名古屋大学　[3]NII LLMC　[4] 東京大学　[5] 早稲田大学
r73ryo.s@keio.jp　jiang.shuting.g7@s.mail.nagoya-u.ac.jp
shinnosuke_takamichi@keio.jp　higashinaka@i.nagoya-u.ac.jp

概要

話者と対話システムが同時に発話・聴取を行う full-duplex 型音声対話システムの登場に伴い，コミュニケーションに関する複数モーダルの統合や，標準語以外の言語への拡張が，次なる課題として見据えられる．そこで我々は，一人称・三人称視点から音声・映像を同期収録可能な対話収録システムを構築し，そのシステムを用いて津軽弁音声対話コーパスを構築した．本論文では収録システムの設計とコーパスの分析結果を示す．

1　はじめに

大規模言語モデルの登場により，人間と対話システムとの言語的インタラクションの自然性は飛躍的に向上した [1, 2, 3]．これに追従する形で，音声対話においても，よりインタラクティブな対話を志向した研究が進展している．近年では，話者とシステムが同時に発話・聴取を行う full-duplex 音声対話モデルが提案されており [4, 5, 6, 7, 8]，音声対話研究のさらなる高度化が期待される．

このような full-duplex 音声対話モデルの発展において，特に重要な課題として二点が挙げられる．一つは**マルチモーダル化**である．人間同士の対話では，音声，表情，身振り手振りなど複数モダリティが相互作用する [9, 10, 11]．同様に，full-duplex 音声対話モデルにも複数モダリティの統合が求められるが，その学習と評価に適した，時刻同期された音声・映像日本語対話データは十分に整備されていない．また，対話システムは人間と同様に一人称視点で環境を知覚し応答すべきであり，エゴセントリック（一人称視点）での収録が望ましい．

もう一つは**言語変種，特に地域方言への対応**である．地域方言は，音韻・韻律といった音声学的特徴から語彙・文法・談話構造に至るまで多層的に変異し，自然対話の成立に大きく関与する．しかし，日本語の既存音声対話コーパス [12, 13][1)] や full-duplex 音声対話モデル [5][2)] は標準語中心であり，地域方言の多様性を十分に扱えていない．

本研究では，上記の課題に対処するため，**(i)** 一人称・三人称の両視点から音声・映像を同期収録できるマルチモーダル対話収録システムを設計・構築し，**(ii)** そのシステムを用いて，消滅が危惧されている方言の一つである津軽弁[3)] の音声対話コーパスを構築する．提案システムは，二者対話の音声・映像を多視点から同時に収録できる．本稿では，まず収録システムの設計を述べ，次に 12 名の話者による 30 セッション（総収録時間約 16.8 時間）からなる津軽弁音声対話コーパスの概要と分析結果を報告する．本コーパスは，国立情報学研究所[4)] が管理するリポジトリでの公開を予定している．

2　一人称・三人称視点対話収録システムの構築

二者対話を対象としたマルチモーダル対話収録システムの概要を図 1 に示す．以降では，その設計と運用について述べる．

2.1　概要

システムは主に以下の観測デバイスから成る．一人称ウェアラブルカメラ・マイクと三人称話者側面

[*] 同等の貢献

[1)] https://www.nii.ac.jp/dsc/idr/rdata/Hazumi/
[2)] https://huggingface.co/nu-dialogue/j-moshi
[3)] https://www.bunka.go.jp/seisaku/kokugo_nihongo/kokugo_shisaku/kikigengo/shinsai_jittaichosa/pdf/aomori_01.pdf
[4)] https://www.nii.ac.jp/

図 1　対話収録システムの構成と収録映像の例．全ての観測デバイスは時刻同期して映像・音声を保存する．各デバイスの録画／録音の開始・終了は単一の PC から一括制御する．映像例では，身振り手振りを一人称と三人称の両視点で取得している．

カメラ・マイクは収録用 PC から録画／録音を一括制御し，時間同期された一人称と三人称視点の映像と音声を保存する．最後の三人称話者正面カメラ・マイクは予備系として独立動作させる．

一人称ウェアラブルカメラ・マイク．デバイス装着者（各話者）視点から映像と音声を取得する．主に，デバイス装着者の発話音声と手振り，相手話者の表情や身振り手振りの観測を目的とし，将来的な一人称視点音声情報処理および full-duplex モデルでの利用を見据える．

三人称話者側面カメラ・マイク．固定視点でステレオ音声と話者側面映像を取得する．全話者の音声，表情，身振り手振りを定点観測し，多チャネル計測に基づく話者と環境の三次元的な解析を見据える．

三人称話者正面カメラ・マイク（予備）．各話者の正面から音声，表情，身振り手振りを収録する．前述の二系統の不動作時に備えた冗長系として運用する．

2.2　デバイス構成

一人称ウェアラブルカメラ・マイクとして Thinklet[5] を装着し，三人称話者側面および三人称話者正面カメラ・マイクとして GoPro HERO 13[6] を設置した．各デバイスの外観および装着例は図 2 に示すとおりである．

図 2　一人称ウェアラブルカメラ・マイクとして使用する Thinklet の外観（左）とその装着の様子（右）．赤丸部がカメラとして機能する．

表 1　収録データの保存形式

Device	Modality	Format
Thinklet	Audio	48 kHz・16 bit・5ch WAV
	Video	1920 × 1080 MP4（≈ 30.03 fps）
GoPro	Audio	48 kHz・2ch AAC（≈ 189 kbps）
	Video	1920 × 1080 MP4（29.97 fps）

接続形態は，Thinklet を Bluetooth で収録用 PC に無線接続し，三人称話者側面用 GoPro を USB type-A-to-C で有線接続した．各デバイスの録画／録音の開始・終了は，収録用 PC 上のソフトウェアで一括制御し，同ソフトウェアは GitHub[7] にて公開した．CUI から一括操作が可能で，GUI では Thinklet の状態（録画中／待機）やプレビュー映像をリアルタイムに確認できる．デバイス間の録画／録音の完全同時開始はデバイス仕様上保証できないため，開始直後に PC からベル音を再生し，各デバイスで検出したベル音の到来時刻を基準に後処理で時刻同期した．収録データの保存形式を表 1 に示す．

電源運用は，Thinklet と三人称話者正面 GoPro をモバイルバッテリで常時給電し，三人称話者側面 GoPro は USB ポートを収録用 PC との接続に使用したため，対話の区切りごとに残量を目視確認し，必要に応じて (残量 20%を目安に) 充電・交換した．

3　エゴセントリック津軽弁音声対話コーパスの構築

2 節にて構築したシステムを用いて，津軽弁音声対話コーパスを構築した．

3.1　収録

収録時期と収録量．2025 年 9 月に青森市内の静音な室内にて収録を行った．一セッションの目安は

[5] https://mimi.fairydevices.jp/technology/device/thinklet/

[6] https://gopro.com/ja/jp/shop/cameras/learn/hero13black/CHDHX-131-master.html

[7] https://github.com/takamichi-lab/llm-jp-thinklet-public

30 分とし，合計 30 セッションを収録した．なお，同一話者ペアは最大 2 回まで参加可能とし，作業量に応じて謝金を支払った．最終的な総収録時間は約 16.8 時間である．なお，本収録とアノテーションに関する一連の流れは、この後あたりに，NII における倫理審査を経て実施している．

参加者． 青森県在住の 12 名（男性 7 名，女性 5 名）が参加した．年齢層は 20 代と 30 代が各 2 名，40 代が 3 名，50 代が 4 名である．

対話トピック． 参加者はペアとなり，事前に指定されたトピックに沿って自由に対話した．各ペアには，(1) 話題に制限のない自由対話，(2) 所定の話題に関する自由対話，(3) 思わず人に話したくなる自分のとっておきのエピソードのいずれか 1 つのトピックを割り当てた．

アンケート． 各参加者は，以下の 3 種類のアンケートに回答した．

- **事前アンケート**：年齢，性別，職業，性格特性に関する調査．性格特性は既存研究 [14] の pre-questionnaire に準拠．
- **対話後アンケート**：各セッション終了後，自他の振る舞いを 5 段階で評価．既存研究 [14] の follow-up questionnaire に準拠．
- **事後アンケート**：全セッション終了後に，(1) 印象に残った対話と理由，(2) 対話で意識した点を自由記述で回答．

3.2 アノテーション

一人称ウェアラブルマイクの収録音をデバイスごとにモノラル化し，2 デバイスのモノラル音を左右チャネルに割り当てたステレオ音を作成した[8]．このステレオ音を ELAN[9] 形式にて保存し，ELAN を用いたアノテーションを津軽弁話者に依頼した．アノテーション時の規則の抜粋を以降に示す．

発話区間と時間内容． 各話者の各発話について，時間区間と発話内容をアノテーションする．発話区間が重複する場合も別発話として扱う．

音声に忠実な書き起こし． フィラーは除去せずに音に忠実に書き起こす．笑い声や喉払い等は専用記号で表記する．

区間の分割． 発話区間の間が 0.2 秒以下であれば，発話内容を結合して 1 つの区間に，0.2–0.4 秒であれば発話内容を読点を挟んで結合，0.5 秒以上は別の発話区間として扱う．

[8] 一部セッションでは一人称ウェアラブルマイクのファイル欠損が認められたため，三人称話者側面マイクのファイルで代用した．

[9] https://archive.mpi.nl/tla/elan

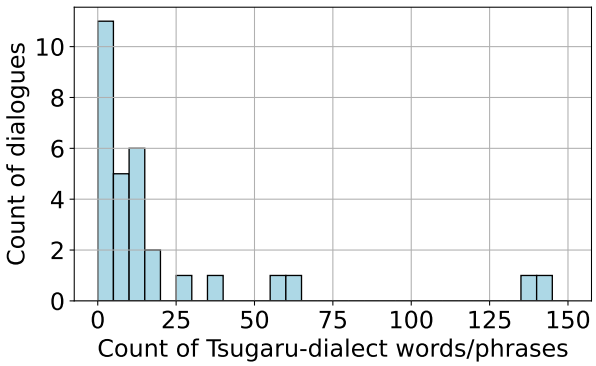

図 3 対話中に出現した津軽弁語句の回数．

津軽弁特有の語句． 方言特有と判断される語句について，時間区間と語句，意味を別ファイルに列挙する．方言性の判定は各アノテータに委ねた．

4 コーパス分析

4.1 津軽弁特有の語句

全 30 セッションで計 102 語の津軽弁特有語句を抽出した．その例を表 2 に示す．表より，接続詞や名詞など，日本標準語から表記が大きく変化した語句が多数確認された．

セッションごとの津軽弁特有語句の出現状況を図 3 に示す．30 セッション中 6 セッションでは，30 分間で 25 回以上（約 1 分に 1 回弱）の頻度で方言語句が出現した．一方で，方言語句が一度も出現しないセッションも 11 存在した．したがって，本コーパスを用いて津軽弁語句を含む対話モデルの学習・評価を行う際には，この出現頻度のアンバランスに留意する必要がある．

4.2 オーバーラップ率

対話の同時双方向性を定量的に評価するため，Nguyen ら [15] および Ohashi ら [5] の手法に基づき，1 分あたりの累積オーバーラップ時間（秒）を算出した．

分析の結果，本データセットの平均オーバーラップは 14.82 秒/分であった．この値は，日本語雑談対話で学習した J-Moshi モデルの報告値 (約 5.0 秒/分) や，その学習データである Ground-truth (約 8.0 秒/分) より顕著に高い [5]．

通常，高いオーバーラップ率は発話衝突として否

表 2 コーパスに含まれる津軽弁特有の語句．全ての項目は津軽弁話者アノテータによるもの．

語句	意味	由来
だはんで	なので、だから	語源は、室町時代の京都で使われていた「ほどに」が変化した「ほでえ」「ほで」「はで」に由来すると考えられています
ささってったのに	してあったのに	「ささる」は、主に北海道や東北地方で使われ、「意図せず～してしまう」「自然と～になってしまう」といったニュアンスを持つ方言です。
じゃっぱじる	じゃっぱ汁	「じゃっぱ汁」の語源は、津軽弁で「雑把（ざっぱ）」、つまり「魚のあら」を意味する言葉に由来します。この「じゃっぱ」を、大根やネギなどの野菜と一緒に煮込んだ汁物のため「じゃっぱ汁」と呼ばれます

定的に捉えられることがある．しかし本データセットでは，定性的な分析から，(i)「食い気味の相槌」，(ii)「共起する笑い」，(iii) 互いに文を補完し合う「発話の並走」など，親密性に基づく協調的な振る舞いが頻繁に観測され，これが高い値の主要因となっている．つまり，本コーパスは極めて活発で感情豊かなインタラクションを豊富に含むことが，オーバーラップ率の高さからも裏付けられる．

5 おわりに

本論文では，full-duplex 音声対話システムの構築に資する，一人称・三人称視点対話収録システムと津軽弁音声対話コーパスについて述べた．今後は，津軽弁 full-duplex 音声対話システムを試作する．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです．

参考文献

[1] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. **arXiv 2309.07864**, 2023.

[2] Ryuichiro Higashinaka, Michimasa Inaba, Zhiyang Qi, Yuta Sasaki, Kotrao Funakoshi, Shoji Moriya, Shiki Sato, Takashi Minato, Kurima Sakai, Tomo Funayama, Masato Komuro, Hiroyuki Nishikawa, Ryosaku Makino, Hirofumi Kikuchi, and Mayumi Usami. Dialogue system live competition goes multimodal: Analyzing the effects of multimodal information in situated dialogue systems. In **The 14th International Workshop on Spoken Dialogue Systems Technology**, 2024.

[3] Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. **ACM Comput. Surv.**, Vol. 58, No. 6, December 2025.

[4] Chen Chen Kevin Hu Ankita Pasad Elena Rastorgueva Seelan Lakshmi Narasimhan Slyne Deng Ehsan Hosseini Asl Piotr Zelasko Valentin Mendelev Subhankar Ghosh Yifan Peng Jason Li Jagadeesh Balam Vitaly Lavrukhin Boris Ginsburg Zhehuai Chen, Edresson Casanova. Open full-duplex voice agent with speech-to-speech language model. In **Proc. IEEE ASRU**, 2025.

[5] Atsumoto Ohashi, Shinya Iizuka, Jingjing Jiang, and Ryuichiro Higashinaka. Towards a Japanese Full-duplex Spoken Dialogue System. In **Proc. Interspeech 2025**, pp. 1783–1787, 2025.

[6] Ju Lin, Yiteng Huang, Ming Sun, Frank Seide, and Florian Metze. Directional Speech Recognition with Full-Duplex Capability. In **Proc. Interspeech 2025**, pp. 2570–2574, 2025.

[7] Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model. In **Proc. Interspeech 2025**, pp. 2715–2719, 2025.

[8] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. **arXiv 2410.00037**, 2024.

[9] Meng-Chen Lee and Zhigang Deng. Learning multimodal motion cues for online end-of-turn prediction in multiparty dialogue. In **Proceedings of the 27th International Conference on Multimodal Interaction**, p. 265–274, 2025.

[10] Jian Ding, Bo Zhang, Dailin Li, Jian Wang, and Hongfei Lin. Disentangling cross-modal interactions for enhanced multimodal emotion recognition in conversation. ICMI '25, p. 344–353, New York, NY, USA, 2025. Association for Computing Machinery.

[11] Kazushi Kato, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara. Real-time generation of various types of nodding for avatar attentive listening system. In **Proceedings of the 27th International Conference on Multimodal Interaction**, ICMI '25, p. 209–217, New York, NY, USA, 2025. Association for Computing Ma-

chinery.

[12] Yuki Saito, Eiji Iimori, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari. Calls: Japanese empathetic dialogue speech corpus of complaint handling and attentive listening in customer center. In **Proc. Interspeech 2023**, pp. 5561–5565, 2023.

[13] Wataru Nakata, Kentaro Seki, Hitomi Yanaka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. J-chat: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling. 2024.

[14] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In Chu-Ren Huang, Yasunari Harada, Jong-Bok Kim, Si Chen, Yu-Yin Hsu, Emmanuele Chersoni, Pranav A, Winnie Huiheng Zeng, Bo Peng, Yuxi Li, and Junlin Li, editors, **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 852–861, Hong Kong, China, December 2023. Association for Computational Linguistics.

[15] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 250–266, 2023.