# イベントカメラを用いた マルチモーダル対話コーパスの構築と分析

行旨王我 [1]　水谷航太 [1]　延原章平 [1]　河野誠也 [1,2]
[1] 京都工芸繊維大学　[2]RIKEN GRP
{oga.yukimune,kota.mizutani}@vision.is.kit.ac.jp　{nob,kawano}@kit.ac.jp

## 概要

対話における音声活動予測や感情認識などのタスクでは，発話に先行・随伴する微細な表情変化や身体動作が重要な手掛かりとなる．しかし，既存の対話コーパスで主に用いられている RGB カメラは時間分解能に制約があり，これらの高速な変化を十分に記録することが難しい．本研究では，高時間分解能および高ダイナミックレンジを特徴とするイベントカメラに着目し，雑談対話を対象として音声，RGB 映像，イベントストリームを同期収録したマルチモーダル対話コーパスを構築した．本コーパスは 25 名・26 セッション・総対話時間約 3.5 時間から構成される．構築したコーパスの有用性を検証するため，対話データを用いた事例解析を行った結果，音声活動に先行する微細な調音動作が捉えられていることが確認された．

## 1　はじめに

対話システム研究では言語情報の処理が主要な研究対象となってきた．しかし実際の対話では，発話の重なりや言い淀み，省略表現が頻繁に生じ，言語情報のみでは話者の意図を理解することは困難である．これらの現象を理解するには表情や身体動作，韻律情報などの非言語情報が不可欠である．特に雑談のような社会的対話では，情報伝達よりも関係性の構築や維持が重視され，表情や視線，韻律といった非言語的手がかりが相手の感情状態や発話意図の理解に大きく寄与する [1]．このような背景から，音声・映像など複数のモダリティを統合的に扱うマルチモーダル対話システムに関する研究が重要性を増している [2, 3]．

マルチモーダル対話研究の推進には，複数のモダリティを同期的に記録し，各モダリティの寄与を分析可能なコーパスの構築が不可欠である．しかし，従来のマルチモーダル対話コーパスで用いられる RGB カメラは 30fps 程度のフレームレートに制限され，発話開始の予兆となる口唇の微動など，フレーム間で生じる微細な変化を捉えることが困難である．

そこで本研究では，マイクロ秒単位の高時間分解能と高ダイナミックレンジを有するイベントカメラを導入したマルチモーダル対話コーパスの構築に取り組む．イベントカメラは輝度変化を非同期的に検出するため，従来のカメラでは捉えきれなかった対話中の瞬間的な非言語現象の観測が可能になると期待される．しかし，イベントカメラを用いた対話コーパスはほとんど存在しない [4]．本研究によるコーパスの構築により，音声活動の検出・予測，ターンテイキングや聞き手反応の詳細な時間構造の解明，数十ミリ秒で生起・消失する微細な表情変化による感情認識の精緻化など，対話ダイナミクスの理解に向けた新たな研究展開が期待される．

本論文では，マルチチャネル音声，RGB 映像，イベントストリームを同期させたコーパス構築について報告する．26 セッション・約 3.5 時間のデータを収集し，音声波形とイベント発生パターンの定量的・定性的分析により，本コーパスが微細な非言語現象の解析に有用であることを確認した．

## 2　関連研究

テキスト・音声・表情情報を統合的に活用したマルチモーダル対話コーパス [5, 6, 7] に加え，近年では生理信号や視線データなどの異種センサーを統合した日本語対話データセットも報告されている [8, 9, 10]．しかし，これらで用いられる RGB カメラは一般に 30fps 程度のフレームレートであり，フレーム間で生じる瞬間的な口唇動作，視線変化，身体動作の開始タイミングなどを十分に捉えることは困難である．



This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

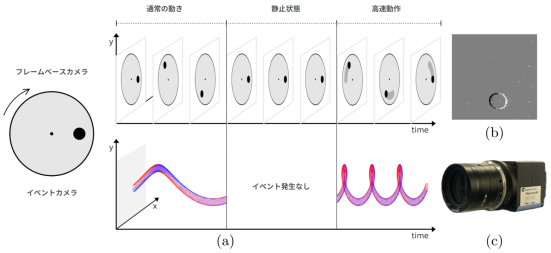


図 1: イベントカメラと従来のカメラの撮像原理の比較（文献 [13] の図を基に一部改変）．従来のカメラは一定間隔で全画素を同時に露光するのに対し，イベントカメラは画素ごとの輝度変化を非同期に検出し，変化のある部分のみを高時間分解能で記録する．

そこで本研究では，イベントカメラに着目した．イベントカメラは生体の視覚神経系を模倣したセンサーであり [11, 12]，図 1 に示すように，従来のフレームベースカメラが一定間隔で全画素を同時に露光するのに対し，各画素が独立して輝度変化を検出し，変化があった時のみ非同期的にイベント $(x, y, t, p)$ を生成する．ここで，$(x, y)$ は座標，$t$ は時刻，$p \in \{+1, -1\}$ は輝度の増減を表す．静止領域ではイベントが発生しないため，対話のように動きが局所的な場面ではデータ量を大幅に削減できる．この動作原理により，マイクロ秒単位の高時間分解能と 120dB 以上の高ダイナミックレンジを実現している．

イベントカメラはこれまで，オプティカルフロー推定や物体追跡などの視覚タスクを中心に研究が進められてきた [14, 15]．近年では人間の行動・感情認識への応用が注目されており，微表情認識データセット NEFER[16] でも，フレームベースの手法と比較して瞬間的な表情変化の検出において優れた性能が報告されている．また，音声とイベントストリームを統合した発話区間検出システムでは，発話に伴う顔面動作を高時間分解能で捉えることで，電力消費と計算量の両面で効率的な処理が可能となることが示されている [17]．

これらの研究は，イベントカメラが対話中の非言語的手がかりの理解・分析に有用であることを示唆している．しかし，いずれも特定タスクに焦点を当てた研究であり，人間同士の自然な対話における多様な非言語現象を包括的に記録したイベントカメラベースのコーパスは存在しない．

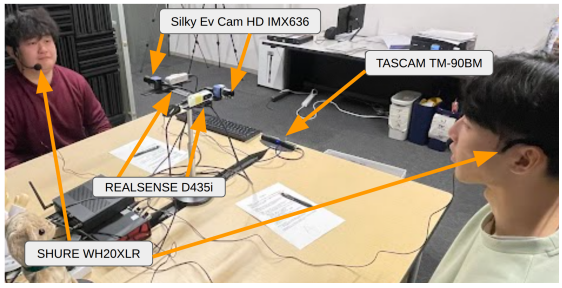


図 2: 対話収録環境の構成．被験者が対面で着席し，それぞれヘッドセットマイク（SHURE WH20XLR）と卓上マイク（TASCAM TM-90BM）で音声を収録している．表情および動作の記録には，イベントカメラ（SilkyEvCam HD IMX636）と RGB カメラ（Intel RealSense D435i）を併用した．

# 3 マルチモーダル対話コーパス

本章では，構築したマルチモーダル対話コーパスのデータ収集方法と仕様について述べる．

## 3.1 対話タスクと収録環境

1 対 1 形式の自由対話（雑談）を収録対象とし，大学内で募集した被験者がペアで約 5〜8 分間の対話を行った．特定のトピックは設定せず，必要に応じて j-tocc の話題リスト [18] を参考にした．収録環境は図 2 に示すように，被験者が互いに向かい合って着席し，各被験者の前方に RGB カメラとイベントカメラを配置した．

## 3.2 使用センサー

**イベントカメラ** 各被験者正面に CenturyArks 社製 SilkyEvCam HD（IMX636）を配置した．解像度は 1280 × 720，マイクロ秒単位の時間分解能と高ダイナミックレンジにより，発話に先行する口唇の微動や瞬間的な表情・視線変化を高精度に記録できる．

**RGB カメラ・音声** Intel RealSense D435i を各被験者に 1 台ずつ配置し，RGB 映像（1280 × 720, 30fps）を取得した．音声は各被験者にヘッドセット型マイク（SHURE WH20XLR）を装着させ，さらに卓上マイク（TASCAM TM-90BM）を設置して計 3 チャンネル（48kHz ／ 16bit）で収録した．

## 3.3 アンケート

各セッションの前後に，被験者の個人特性および対話体験に関するアンケートを実施した．

事前アンケートでは，基本属性（年齢・性別），対



This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 1: 収集したマルチモーダル対話データの概要

| 項目 | 値 |
|---|---|
| 被験者数 | 25 名（男性 18，女性 7） |
| セッション数 | 26 |
| 総対話時間 | 約 3.52 時間（12,659 秒） |
| 平均対話時間 | 486.9 ± 103.7 秒 |
| 総発話数（VAD 検出） | 3,790 |
| 平均発話数/セッション | 338.6 ± 102.3 |
| 総音素数 | 191,887 |
| 音素アライメント成功率 | 90.2%（7,937/8,803） |
| 平均音素数/発話 | 24.2 |
| 平均イベント数 [$\times 10^6$] | 829.9 ± 635.1 |

表 2: バイアス設定別のイベント総数統計

| 条件 | セッション数 | 平均 [$\times 10^6$] | 標準偏差 |
|---|---|---|---|
| バイアスなし | 8 | 1,701.3 | 315.2 |
| 中バイアス設定 | 8 | 562.7 | 203.7 |
| 高バイアス設定 | 10 | 346.4 | 126.7 |

話相手への初期印象，気分状態（PANAS），および性格特性（Big Five 尺度短縮版）を収集した．事後アンケートでは，PANAS による気分変化の測定に加え，対話品質の主観評価，対話相手への印象の再評価，カメラの存在や室内環境が対話に与えた影響についての回答を得た．

これらのアンケートデータは，被験者ごとの個人差や環境要因がイベント発生パターンや対話行動に与える影響を分析する際の補助指標として活用する．

### 3.4 データ概観

収集データの概要を表 1 に示す．本稿執筆時点で，参加者 25 名（男性 18，女性 7），セッション数 26，総対話時間約 3.52 時間を収録した．

音声活動検出には pyannote.audio 3.1 を使用し発話区間を抽出した．発話の書きおこしには Whisper large-v3 を使用し，認識結果の誤りについては人手での修正を進めた．さらに，pyopenjtalk による Grapheme-to-Phoneme（G2P）変換で音素列を生成し，発話区間内の各音素に時間情報を付与した．

イベントカメラの記録設定については，バイアスパラメータ[1)]を変えた 3 条件で収録を行った（表 2）．

1） bias-diff-on/off は輝度変化の検出感度を調整するパラメータであり，値が小さいほど微小な輝度変化を検出できる．

表 3: バイアス設定別の音声-イベント相関分析結果

| 条件 | 最大相関 $r_{\max}$ | 最適ラグ [ms] | 平均相関 |
|---|---|---|---|
| バイアスなし | 0.223 | +3 | 0.127 |
| 中バイアス | 0.479 | −96 | 0.346 |
| 高バイアス | 0.357 | −94 | 0.059 |

※ Smoothed ROI Ratio と音声 RMS の相互相関に基づく．最大相関は相互相関関数の最大値，最適ラグはその最大値が得られたラグ，平均相関は全ラグ範囲における相関係数の平均を示す．

## 4 コーパスの分析

本章では，収集したコーパスの統計的特性を示すとともに，イベントカメラによる記録データの特性を分析する．

### 4.1 音声活動とイベントの時間的関係

イベントカメラが発話に伴う顔面動作を捉えられているかを検証するため，音声波形とイベント発生量の相関分析を実施した．相互相関関数を用いて，音声 RMS とイベント指標との相関を評価した．具体的には，±150ms の範囲でラグを変化させながら相関係数を算出した．イベント指標には，時間ビン内における顔領域内イベント数を全画面イベント数で除し平滑化処理を施した Smoothed ROI Ratio を用いた．この指標は，照明変化や被験者の体動など顔面動作以外の要因による全体的な変動の影響を緩和し，顔領域特有の動作（口唇動作など）の寄与を抽出することを意図している．

表 3 に各バイアス設定における相関分析結果を示す．中バイアス設定および高バイアス設定では，相互相関のピークが負のラグ側（それぞれ −96ms，−94ms）に現れており，イベント発生が音声出力に先行する傾向が確認された．これは，発話開始前の口唇準備動作や表情変化がイベントとして検出されていることを示唆する．

バイアス設定がデータ特性に与える影響について，イベント発生量（表 2）と音声波形との相関強度（表 3）の両面から分析した．S/N 比の観点から以下のように解釈できる．バイアスなし条件では多数のイベントを記録できるが，ノイズイベントも増加し相関が低下した．高バイアス設定ではノイズは抑制されるが，発話関連イベントの検出感度も低下し相関は中程度にとどまった．中バイアス設定ではノイズ抑制と信号検出のバランスが取れ，最も高い相関が得られた．これらの結果は，Delbruck ら [19] の知見と整合しており，適切なバイアス設定により



This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

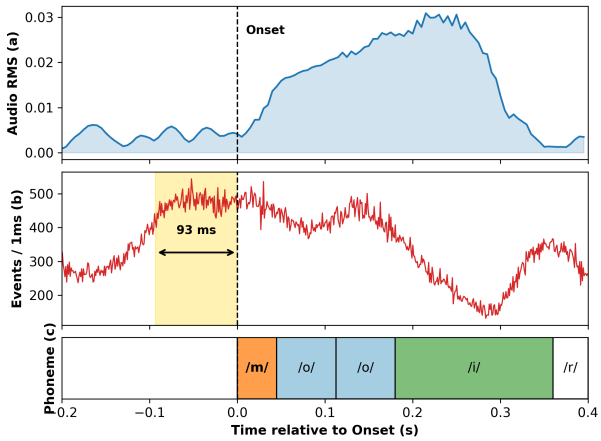


図 3: 対話例の時系列可視化（発話内容：「もういらんか」）．(a) 音声 RMS エンベロープ，(b) イベント発生率（1ms ビン），(c) 音素アライメント．音声開始（$t = 0$）に対し，両唇音/m/の閉鎖動作や準備動作に伴うイベント発生が約 93 ms 先行して観測されている（黄色領域）．

有効な記録が得られることを示している．

このような特性は，従来のフレームベースカメラでは困難であった発話予兆動作の検出やターンテイキングの時間的ダイナミクスの解析など，音声活動予測タスク [20] への応用が期待される．

### 4.2 発話例とイベントの先行性

本コーパスの最大の特徴である，音声活動に先行する非言語動作の記録を確認するため，実際の対話データから抽出した事例を図 3 に示す．紹介する例は，話者 A が発話権を取得し，「もういらんか（/m/o/o/i/r/...）」と発話を開始する局面である．なお，可視化にあたっては音素アライメントと音声波形の間に生じていた体系的なタイムラグを補正し，音声の立ち上がり（Audio Onset）を $t = 0$ として整列させている．

図 3(a) の音声 RMS エンベロープ（青線）において，振幅が立ち上がる時刻は $t = 0$ である．これに対し，図 3(b) のイベント発生率（赤線）を確認すると，音声開始の約 93 ms 前から，イベント数が急激に上昇していることが確認できる（図中黄色領域）．この区間では音声信号は依然としてノイズレベルであるが，イベントカメラは発話に向けた予備動作を高感度に捉えている．

図 3(c) の音素アライメントとの対応を見ると，このイベント先行区間および直後のピークは，先頭音素である両唇鼻音 /m/ の生成区間および続く母音 /o/ への遷移と時間的に対応している．「もう（/m/o/o/）」の発音において不可欠な口唇閉鎖（閉じる動作）や，続く円唇動作（突き出し）といった調音運動が，音声信号の発生に先立って開始されている様子が，イベント発生量の変化として明確に記録されている．

以上の結果は，本コーパスが音声活動検出のみでは捉えきれない発話の予兆動作を可視化・分析可能なマルチモーダルリソースであることを示している．

## 5 おわりに

本研究では，対話における微細な非言語情報を捉えるため，マイクロ秒単位の高時間分解能を有するイベントカメラを導入した新たなマルチモーダル対話コーパスを構築した．本コーパスは，25 名の被験者による 26 セッション，総対話時間約 3.5 時間の雑談対話から構成され，マルチチャネル音声，RGB 映像，イベントストリームを高精度に同期させて記録している．加えて，被験者の属性や心理状態，対話体験に関する主観評価を把握するため，事前・事後アンケートを実施し，行動データと主観指標を対応付けて分析可能な設計とした．

収録データに対しては，オープンソースライブラリを活用した音声書きおこしと，音素・韻律情報を含む時間情報の付与を進めた．これにより，音声，視覚的非言語行動，および言語単位を時間軸上で統合的に扱うための，半自動アノテーション基盤を整備した．

本コーパスは，高時間分解能モダリティを活かして対話中の微細な非言語現象を記述可能な，新たなマルチモーダル対話解析の基盤を提供するものである．本コーパスの活用により，音素レベルでの調音動作とイベント発生の対応分析をはじめ，より高精度な音声活動予測や感情認識モデルの構築，対話における協調的振る舞いのパターンの解明 [21, 22] など，対話研究のさらなる発展が期待される．今後は，コーパス規模の拡大やアノテーションの精緻化に取り組み，広く研究コミュニティへ貢献することを目指す．



This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

## 謝辞

本研究は, JSPS 科研費 25K21287 および 24H00742 の支援を受けた．また，実験にご協力いただいた被験者の皆様に深く感謝いたします.

## 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**, 78(5):265–270, 2022.

[2] 金崎翔大, 渡邉寛大, 河野誠也, 湯口彰重, 桂井麻里衣, and 吉野幸一郎. 対話行為予測とエントレインメント予測に基づいたマルチモーダル対話システム. In **人工知能学会研究会資料 言語・音声理解と対話処理研究会 96 回 (2022/12)**, page 20. 一般社団法人人工知能学会, 2022.

[3] Ryuichiro Higashinaka, Tetsuro Takahashi, Michimasa Inaba, Zhiyang Qi, Yuta Sasaki, Kotrao Funakoshi, Shoji Moriya, Shiki Sato, Takashi Minato, Kurima Sakai, et al. Dialogue system live competition goes multimodal: Analyzing the effects of multimodal information in situated dialogue systems. In **Proc. IWSDS**, 2024.

[4] 行旨王我, 水谷航太, 延原章平, and 河野誠也. イベントカメラを用いたマルチモーダル対話コーパスの構築に向けて. Technical Report 2025-SLP-158(14), 情報処理学会 音声言語情報処理研究会 (SLP), December 2025.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. **Language Resources and Evaluation**, 42(4):335–359, 2008.

[6] Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. **Lang Resources & Evaluation**, 41:181–190, 2007.

[7] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In **Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Jingjing Jiang, Ao Guo, and Ryuichiro Higashinaka. Estimating the emotional valence of interlocutors using heterogeneous sensors in human-human dialogue. In **Proc. of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pages 718–727, September 2024.

[9] Jingjing Jiang, Ao Guo, and Ryuichiro Higashinaka. Integrating physiological, speech, and textual information toward real-time recognition of emotional valence in dialogue. In **Proc. of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pages 591–600, August 2025.

[10] 船越 孝太郎 and 小尾 賢生. 呼吸信号付きマルチモーダル対話コーパス bind. **人工知能学会研究会資料 言語・音声理解と対話処理研究会**, 104:01–08, 2025.

[11] Tobi Delbruckl. Neuromorophic vision sensing and processing. In **ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference**, pages 7–14, 2016.

[12] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. **Proc. of the IEEE**, 102(10):1470–1484, 2014.

[13] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3D reconstruction and 6-DoF tracking with an event camera. In **Proc. of ECCV**, pages 349–364, 2016.

[14] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. **IEEE TPAMI**, 44(1):154–180, January 2022.

[15] Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Graph-based asynchronous event processing for rapid object recognition. arXiv:2308.14419, 2023.

[16] Lorenzo Berlincioni, Luca Cultrera, Chiara Albisani, Lisa Cresti, Andrea Leonardo, Sara Picchioni, Federico Becattini, and Alberto Del Bimbo. Neuromorphic event-based facial expression recognition. arXiv:2304.06351, 2023.

[17] Arman Savran, Raffaele Tavarone, Bertrand Higy, Leonardo Badino, and Chiara Bartolozzi. Energy and computation efficient audio-visual voice activity detection driven by event-cameras. In **Proc. of FG**, pages 333–340, 2018.

[18] 中俣 尚己, 太田 陽子, 加藤 恵梨, 澤田 浩子, 清水 由貴子, and 森 篤嗣. 『日本語話題別会話コーパス：j-tocc』. **計量国語学**, 33(1):11–21, 2021.

[19] Tobi Delbruck, Rui Graca, and Marcin Paluch. Feedback control of event cameras. arXiv:2105.00409, 2021.

[20] Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In **Proc. of HAI**, page 13–21, 2023.

[21] Seiya Kawano, Masahiro Mizukami, Koichiro Yoshino, and Satoshi Nakamura. Entrainable neural conversation model based on reinforcement learning. **IEEE Access**, 8:178283–178294, 2020.

[22] Seiya Kawano, Shota Kanezaki, Angel Fernando Garcia Contreras, Akishige Yuguchi, Marie Katsurai, and Koichiro Yoshino. Analysis of style-shifting on social media: Using neural language model conditioned by social meanings. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pages 7911–7921, 2023.



This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).