漫才対話の収集及び自動アノテーションのパイプラインの検討

佐々木裕多 [1]　張建偉 [1]
[1] 岩手大学 理工学部
{s0619027,zhang}@iwate-u.ac.jp

概要

笑いは健康を促進することができ，誰かと一緒のときに起こることが多いため，対話システムへのユーモアの実装によりユーザの心身の満足度が向上すると考えられる．また，日本人の笑いは「会話型」コミュニケーションの笑いが多く，一般的にコミュニケーションにおいて非言語コミュニケーションの割合が高い．このことから，漫才対話に着目し，マルチモダリティを用いて，会話形式のユーモアを学習するためのデータセットの構築を行う．このデータセット構築におけるアノテーションは負荷が高く，再現性が低くなりやすい．アノテーションの負荷軽減と再現性向上のため，漫才対話の収集から自動アノテーションまでのパイプラインの構築を試みる．

1　はじめに

笑いは健康に良い影響を及ぼしている．漫才動画鑑賞を用いた実験から，笑うことによる認知機能改善とストレス応答抑制の有効性や [1]，他者と笑うことによる高齢期での機能不全のリスクの軽減など [2]，いくつもの医学研究が健康に対する笑いの影響を示している [3]．また，笑いは大抵誰かと一緒にいるときに起こると報告されており [4]，コミュニケーションの中で笑いが起きやすいと考えられる．これらのことから，対話システムに笑いを引き起こすユーモアを実装することで，ユーザの心身の満足度向上が期待される．

対話システムへのユーモアの実装を目的として，会話におけるユーモアの理解や生成を学習するためのデータセットの構築を試みる．荒木ら [5] は，ユーモアの面白さを評価する標準的なデータセットの構築の第一段階として，駄洒落データベースの構築を行なった．日本人の笑いは 2 人以上の時に多く成立し，「合いの手」が入ることで笑いが起こる「会話型」コミュニケーションにおける笑いが多い [6]．しかし，駄洒落はジョークの一種であり，対話へのユーモアの応用を考慮すると，文脈を必要とするユーモアに着目する必要がある．また，コミュニケーションにおける笑いを発話のみから判断するのは適切ではない． Mehrabian は 2 つの研究を基に，7%の言葉，38%の音声，55%の表情から感情を伝えると導いた [7, 8]．笑い声や笑い顔が人類に共通し，強い共鳴性を持つことからも，コミュニケーションにおける笑いに対して非言語コミュニケーションの要素を考慮することが重要だと考えられる．そこで，マルチモーダル漫才対話ユーモアデータセットの構築を試みる．Patro ら [9] は TV シリーズのシットコムを用いてラフトラックを予測するマルチモーダルデータセットを構築した．これを参考とし，話者情報，発話テキスト，発話開始／終了時間，面白さ，笑い声区間のアノテーションを行う．しかし，これら全てのアノテーションは負荷の高い作業であり，アノテータ毎に時間のずれや表記ゆれ等による不一致が起こりやすいため，再現性が低いと考えられる．アノテータが自動アノテーションの結果を修正することで，アノテーションの再現性を向上できると考え，本研究では発話テキスト，発話開始／終了時間，笑い声区間を対象とした自動アノテーション手法を検討し，データセット作成のためのパイプライン構築を試みる．

2　パイプライン構築

データセット構築のパイプラインを図 1 に示す．本研究では，漫才対話データ収集と自動アノテーションについて取り組む．

2.1　漫才対話データ収集

マルチモダリティを扱うことができ，漫才対話生成やユーモア検出，ラフトラック予測などの複数のタスクに対応できる抽象度の高いデータセットの構築を目指す．そこで，多くのモダリティ情報を獲得

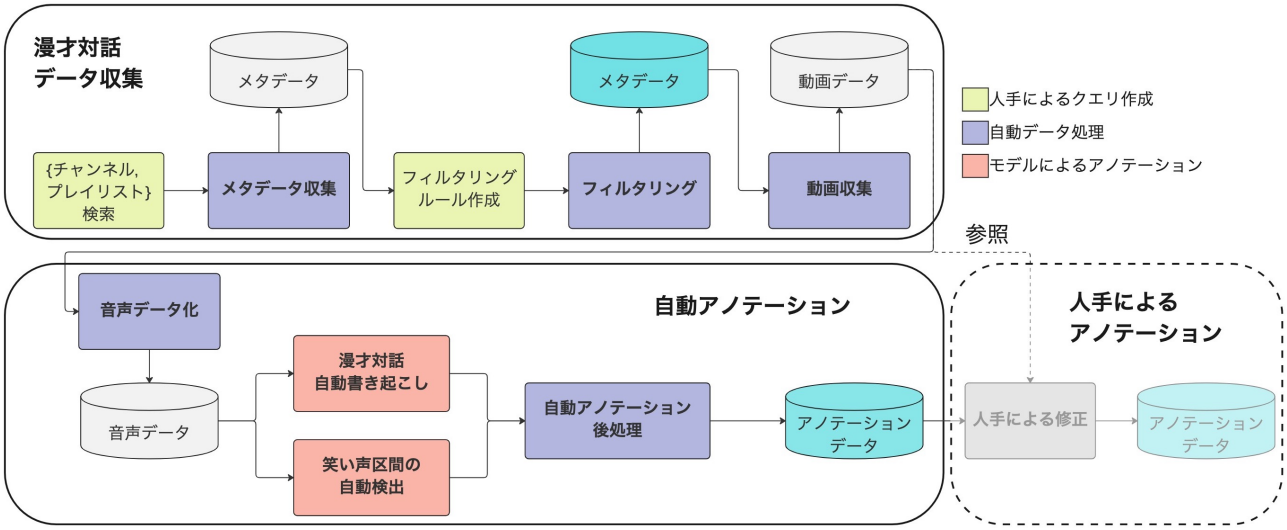

図 1　データセット構築のパイプライン

でき，漫才対話データも多い YouTube [1)] を対象としてデータの収集を行う．

2.1.1　収集方法

収集するデータの決定のため，YouTube Data API [2)] を用いて，動画のメタデータを収集する．メタデータには動画 ID やタイトル，チャンネル ID などが含まれる．メタデータの収集には次の 2 つの方法を採用する．

- M-1 グランプリ公式チャンネルを対象として，「M1 グランプリ」のクエリを用いて検索
- 芸能事務所公式チャンネルが公開しているプレイリストを対象として検索

近年では，一般の人も簡単に動画をアップロードすることができ，違法アップロードされた動画も多い．そのため，人手で API のパラメータを慎重に決定することで，確かに公式に公開された動画のメタデータを収集する．また，漫才師が公式に公開している動画も存在するが，漫才師の偏りを防ぐため，芸能事務所や大会が公開する漫才のみを対象とした．

2.1.2　フィルタリングによる後処理

コントはセットを用いた寸劇形式であり，会話形式のデータ収集には不適切である．したがって，コントやインタビューの動画を除くため，以下のルールに基づくフィルタリングを行う．

1) https://www.youtube.com/
2) https://developers.google.com/youtube/v3


```
{
  "漫才動画 ID": {
    "laughter": [
      "00:30.000 -> 00:30:960",
      ...,
    ],
    "uttr0": {
      "speaker": " ボケ",
      "utterance": " どーもー",
      "utterance_start": "00:00.000",
      "utterance_end": "00:03.000",
      "is_humor": false,
    },
    "uttr1": {...},
    ...
  }
}
```


図 2　アノテーションデータ形式例

- M-1 グランプリ公式チャンネルを対象として，タイトルに「【決勝ネタ】」，「【敗者復活戦ネタ】」，「【ナイスアマチュア賞】」のいずれかを含む
- プレイリストを対象として，タイトルに「漫才」を含む

このフィルタリングによって残ったメタデータを基に動画データを収集する．

2.2　自動アノテーション

漫才対話データのアノテーション結果のデータ形式を図 2 に示す．この形式を満たすためには，漫才を観た上で，発話テキストや発話時間，話者情報，面白さ，笑い声区間の全てのアノテーションを行う必要がある．これらのアノテーションは負荷が高く，テキストの表記ゆれや時間のずれ等が起こりうる．そこで，アノテータの負担を減らし，アノテーションの再現性を向上するため，次の自動アノテーションを行う．

2.2.1 漫才対話自動書き起こし

音声の書き起こしとテキストに対応する発話時間の推定を行う Whisper large-v2[3)] [10] を用いて，漫才対話の自動書き起こしと発話時間推定を行う．同 small モデルを用いた書き起こしも行なったが，漫才に対する音声認識性能が低く，誤字も多いことから，large-v2 モデルを採用する．

2.2.2 笑い声区間の自動検出

各発話の面白さの評価は，アノテータ毎にずれがあると考えられる．観客の笑い声をラフトラックとして扱うことで，各発話に対するユーモアのラベル付けが可能になると考え，笑い声区間の自動検出を行う．検出モデルには AudioSet-YouTube corpus [11] で訓練された MobileNet[4)] [12] を用いて，521 個の音声クラスのうち Laughter クラスの予測確率が 5%以上の区間を笑い声区間として検出する．このモデルは音声を 960ms のフレーム毎に分割し，各フレームに対してクラスを予測する．連続するフレームは 480ms の重複を持つため，笑い声区間と予測されたフレームが連続した場合，それらに対応する時間を結合する後処理を加える．

3 分析

収集されたデータと自動アノテーションの分析を行うことで，目的とした漫才対話が収集できているか，自動アノテーションが実用に足る精度であるかを調査し，課題を考察する．

3.1 収集データの分析

収集されたデータの統計情報を表 1，収集データ内における漫才師の被り数の分布を図 3 に示す．データ件数 280 件に対して漫才師が 201 組であったことや同じ漫才師の出現頻度が小さいことから，漫才師の偏りを防いでデータ収集できていることがわかる．今回収集された YouTube チャンネル数は 3 件であり，これらは大会か芸能事務所の公式チャンネルであった．したがって，違法アップロードされた動画は収集されていないことがわかる．しかし，いくつかの動画を確認した際，コントと思われるデータが確認された．メタデータを収集する API のパラメータの設定やフィルタリングによってコント動画を自動的に除くことは難しく，人手によるフィルタリングも行う必要性がある．

表 1 収集データの統計情報

データ件数	280 (件)
YouTube チャンネル数	3 (件)
漫才師数	201 (組)
平均動画時間	152.7 (秒)
最大動画時間	354.8 (秒)
最短動画時間	72.8 (秒)

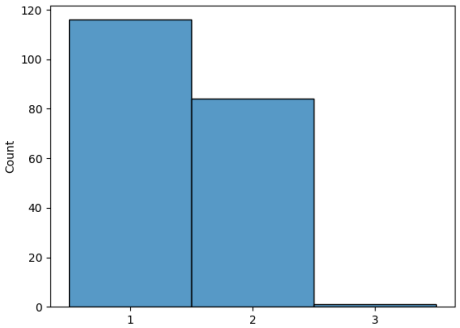

図 3 漫才師の被り数の分布

3.2 自動アノテーションの分析

自動アノテーションの分析のため，Web アプリケーションの構築を行なった．漫才動画，笑い声区間，書き起こし，発話時間をシングルページで参照でき，図 4 のようになっている．画面左では動画鑑賞と笑い声区間検出の確認を行い，画面右では書き起こしや発話時間の確認を行うことができる．このアプリケーションを用いて，20 件程度の漫才動画を鑑賞し，自動アノテーションの傾向の観察と課題の考察を行った．

3.2.1 漫才対話自動書き起こし

Whisper による自動書き起こしは，人手で修正することでアノテーションができるほど高い精度であったが，課題が多くあった．人名やコンビ名のよ

3) https://github.com/openai/whisper
4) YAMNet とも言われる．
https://tfhub.dev/google/yamnet/1

図 4　アプリケーション画面

うな固有名詞や言い間違えボケ，テンポの速すぎる会話などはうまく書き起こしができていなかった．例えば，「竹内豊」という俳優の名前に対して「竹ルーツ居たかったら」，「かわいい」を「ハワイ」と言い間違えるボケに対して「かわいい」と書き起こす誤りがあった．また，出囃子が流れている言葉の無い時間に発話時間を割り当て，ネタ前半部分の発話推定時間が大きくずれる現象も見られた．Whisper は同時に起こる発話の時間を適切に推定できず，重なった発話の片方のみを書き起こすケースも多かった．自動書き起こしと発話時間推定は，発話の重なりのない音声において高い精度であったが，テンポが速く発話の重なった音声において工夫すべき点が多く見られた．

3.2.2　笑い声区間の自動検出

笑い声区間の自動検出精度は，漫才動画の性質に大きく依存していた．観客の笑い声が鮮明に聞き取れる動画に対しては高精度で笑い声区間を検出できていたが，観客のリアクションが小さく収録された動画に対しては笑い声区間を検出できないケースが多かった．さらに，漫才師自身が笑う場合や漫才師が甲高い声または叫ぶような声で発声する場合において，笑い声区間の誤検出が多かった．観客だけのリアクションを考慮するため，観客のリアクションの抽出や増幅のような工夫が必要である．また，検出モデルが予測するフレームは 960ms と長いため，観客の笑い声が長く続くケースや笑いが起こるテンポが速いケースにおいて，検出される笑い声区間が極端に長くなってしまい，不適切であった．より短い時間のフレームに対する予測や音声イベント検出の組み合わせによって，笑い声区間の自動検出精度が向上すると考えられる．

4　まとめと今後の展望

笑いは健康を促進でき，日本人の笑いは「会話型」コミュニケーションの中で起こることが多いため，対話システムへの会話形式のユーモアの実装によりユーザの心身の満足度を向上できると考えられる．本研究では，マルチモーダル漫才対話ユーモアデータセットの構築のため，データ収集とアノテーションの負荷軽減と再現性向上を目的とした自動アノテーションのパイプラインを検討した．今回使用したデータ収集のロジックにより，人手でコントを除く処理は必要であるが，適切にアップロードされた動画を対象に収集できた．しかし，今回行った自動アノテーションの方法には多くの課題が見られ，工夫の余地があることがわかった．すでに検証した自動アノテーションの精度改善実験やデータ収集方法改善の結果を付録 A，B に示している．

今後はアノテーションのルールの細かな設定を決定し，自動アノテーションの手法を改善する．漫才は「会話型」コミュニケーションの形式でネタが行われており，意味のある言葉を含み笑いを狙ったツッコミだけでなく，相槌や合いの手も頻繁に行われている．相槌や合いの手をどの程度テキストに起こすかを決定する必要がある．また，笑い声はピークを持つが一定時間持続するものであり，笑い声が短時間に連続して発生した場合，笑い声区間のアノテーションは不適切になりうる．そのため，笑い声のピークや音声イベントを用いたルールの設定が必要である．これらの細かな設定をした上で，精度と再現性の高い自動アノテーション手法をさらに検討し，データセットの構築と公開を目指す．

参考文献

[1] 山越達矢, 阪本亮, 西垣翔梧, 田中爽太, 福田隆文, 金留理奈, 鈴木久仁厚, 梁弘一, 小山敦子, 阿野泰久. 笑いによるストレス応答抑制と認知機能改善効果. 日本健康心理学会大会発表論文集, Vol. 34, p. 87, 2021.

[2] Yudai Tamada, Chikae Yamaguchi, Masashige Saito, Tetsuya Ohira, Kokoro Shirai, Katsunori Kondo, and Kenji Takeuchi. Does laughing with others lower the risk of functional disability among older japanese adults? the jages prospective cohort study. **Preventive Medicine**, 2022.

[3] Rosemary Cogan, Dennis Cogan, William Waltz, and Melissa McCue. Effects of laughter and relaxation on discomfort thresholds. **Journal of behavioral medicine**, Vol. 10, No. 2, pp. 139–144, 1987.

[4] Rod A Martin and Nicholas A Kuiper. Daily occurrence of laughter: Relationships with age, gender, and type a personality. 1999.

[5] 荒木健治, 内田ゆず, 佐山公一, 谷津元樹, 北海道大学, 小樽商科大学. 駄洒落データベースの構築及び分析. 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B702-3, pp. 13–24, 2017.

[6] 大島希巳江. 日本の笑いと世界のユーモア: 異文化コミュニケーションの観点から. 世界思想社, 2006.

[7] Albert Mehrabian and Morton Wiener. Decoding of inconsistent communications. **Journal of personality and social psychology**, Vol. 6, No. 1, p. 109, 1967.

[8] Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. **Journal of consulting psychology**, Vol. 31, No. 3, p. 248, 1967.

[9] Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh Sarvesh, Hunar Singh, and Vinay P. Namboodiri. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In **2021 IEEE Winter Conference on Applications of Computer Vision (WACV)**, pp. 576–585, 2021.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. **arXiv preprint arXiv:2212.04356**, 2022.

[11] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In **Proc. IEEE ICASSP 2017**, New Orleans, LA, 2017.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. **arXiv preprint arXiv:1704.04861**, 2017.

[13] M Iftekhar Tanveer, Diego Casabuena, Jussi Karlgren, and Rosie Jones. Unsupervised speaker diarization that is agnostic to language, overlap-aware, and tuning free. **arXiv preprint arXiv:2207.12504**, 2022.

[14] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Frédéric Lepoutre, and François Grondin. Resource-efficient separation transformer. **arXiv preprint arXiv:2206.09507**, 2022.

[15] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In **Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)**, pp. 4211–4215, 2020.

A　自動アノテーションの精度を改善できなかった手法

A.1　ボケとツッコミの発話者推定

漫才では，ボケとツッコミの役割が重要であるため，発話者のアノテーションが必要である．Whisper は複数の話者をある程度分離して書き起こしを行うことができていたが，発話者を割り振ることはできないため，以下の発話者推定の手法を検討した．

1. PyAnnote[5] を用いた話者ダイアリゼーション
2. 事前学習済みモデルを用いた教師なしによる話者ダイアリゼーション [13]
3. RE-SepFormer[6] [14] を用いた話者分離

1 と 2 の手法は話者を適切に割り振ることができず，オーバラップした部分の検出もできていなかった．3 の手法において，話者ごとに分離された音声を自動書き起こしすることを目標としていたが，全く分離することができず，実用に足る精度が得られなかった．要因として，日本語の音声に対応したモデルを用いていないことが考えられる．日本語と英語は音声学的特徴が異なり，音声のスペクトログラムの傾向が異なる可能性が推測され，これが英語に対応したモデルが日本語の音声に対してうまく機能していない要因であると考えられる．3 において用いた RE-SepFormer を Common Voice[7] [15] の日本語データセットを用いてファインチューニングし，話者分離を試みたがうまく分離できなかった．しかし，日本語対応モデルの構築は一つの改善策である可能性があるため，引き続き検討していく．

A.2　発話開始／終了時間の推定精度向上

Whisper が推定するテキストに対応した発話時間には誤差があるため，その精度向上を試みた．A.1 で説明した 1 の手法を用いて，話者ダイアリゼーションの結果を発話開始／終了時間の誤差の軽減に使用する実験を行ったが，話者ダイアリゼーションの精度も高くないことから，良い結果が得られなかった．また，出囃子などの影響により Whisper が推定する発話時間が不適切な場合もあったため，ネタの前後の情報のない時間を削除することでノイズを軽減することが最優先の取り組みであると考えられる．まずはこの処理を今後行った上で追加実験を行っていく．

B　データ収集手法の改善

本研究においてデータ収集を行なった際（2022 年 11 月時点），M-1 グランプリ 2022 が終了していなかったが，2023 年 1 月 13 日（論文投稿締切）時点において M-1 グランプリ 2022 が終了していたため，データ収集とフィルタリングのパラメータの変更を行い，データ収集を試みた．主な変更点は以下である．

- M-1 グランプリ公式チャンネルを対象として，「 M-1 」のクエリを用いて検索（ M1 グランプリを M-1 に変更）
- M-1 グランプリ公式チャンネルを対象として，タイトルに【決勝ネタ】」,「【敗者復活戦ネタ】」,「【ナイスアマチュア賞】」,「【準々決勝ネタ】」,「【3 回戦全ネタ】」,「【1 回戦 TOP3】」のいずれかを含む（【準々決勝ネタ】」,「【3 回戦全ネタ】」,「【1 回戦 TOP3】」を追加）

これにより，合計 326 件の動画に関するメタデータを収集できた．また，「【3 回戦全ネタ】」か「【1 回戦 TOP3】」をタイトルに含む動画には，動画 1 本あたり 2 または 3 組の漫才師による漫才が収録されているため，収集できる漫才がさらに増加すると考えられる．

5） https://huggingface.co/pyannote/speaker-diarization
6） https://huggingface.co/speechbrain/resepformer-wsj02mix
7） https://commonvoice.mozilla.org/en