# YouTubeを利用した大規模会話コーパス構築のための複数話者動画抽出

☆米山蒼祐，森大毅（宇都宮大）

## 1　はじめに

相槌や笑いに代表される聞き手反応、あるいは話者交替など、音声対話に見られる実時間の現象をモデル化するには、対話コーパスが必要不可欠である。日本語日常会話コーパス (CEJC) [1]は約200時間の会話データを含んでいるが、深層学習のためのデータとしてはこれでも規模が大きいとは言えない。

本研究では、YouTube 上の動画を使った大規模会話コーパス構築を目的として、クロールした動画から話者が複数人登場する動画を取り出すことを試みる。これは、YouTube 動画から音声認識や話者照合用のコーパスの構築を行った JTubeSpeech [2]の会話版と位置付けられる。JTubeSpeech では、話者表現ベクトルの分散を用いて合成音声、独話、複数話者動画の分類を行っている。しかし、複数話者動画に対しては適合率が悪くなってしまう傾向があった [2]。

そこで本研究では、動画の音響的特徴だけでなく、動画のサムネイルや動画シーン内の人物の数 (視覚的特徴) 、および自動字幕から得られる発話内容の情報 (言語的特徴) をも併用することで、複数話者動画の抽出精度の向上を試みる。

## 2　複数話者動画に関連する特徴

### 2.1 視覚的特徴

動画のサムネイル、および動画シーンの最初の時点、1/4 の時点、1/2 の時点、3/4 の時点、最後の時点の合計 6 枚の画像に対し、YOLOv8 [3]を使って人物の検出を行い、その数を特徴量とする。Fig. 1 に 1/4 の時点における検出人数の分布を示す。単一話者動画に比べ複数話者動画の方が多くの人数が検出されることがわかる。

### 2.2 言語的特徴

対話における発話内容は、独話と大きく異なる。例えば、発話内容の中に「なるほど」のような同意表現があれば、その動画は複数話者によるものである可能性が高い。

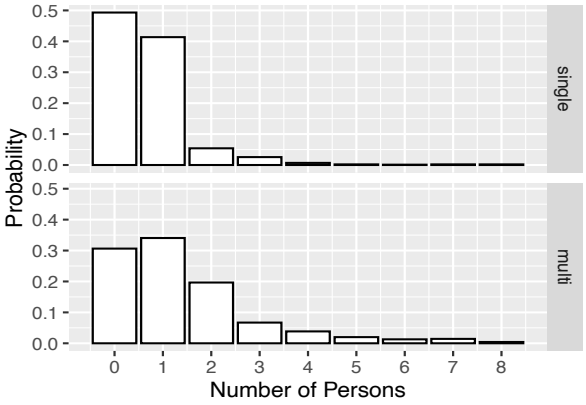


Fig. 1 Distribution of the number of persons detected by YOLO at 1/4 time point.

Classify the transcript into one of the following categories:
single speaker, multi speaker. Return only single or multi, nothing else. It should be all lowercase.
Video Transcription: {冒頭 20 発話}
Classify:

Fig. 2 Prompt given to GPT-3.5 Turbo.

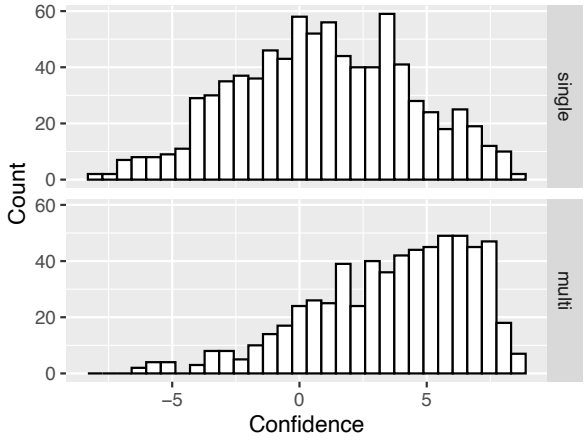


Fig. 3 Distribution of confidence of multi-speaker video estimated by GPT.

今回は、自動字幕の冒頭 20 行、およびその話者が単一 (single) か複数 (multi) かを問う質問文をプロンプトとして GPT-3.5 Turbo に与える (Fig.2) 。この際、GPT が推定した「multi」という語が後続する確率をロジット

---

* Extracting multi-speaker videos from YouTube for building a large-scale conversation corpus. By YONEYAMA, Sousuke, MORI, Hiroki (Utsunomiya University).

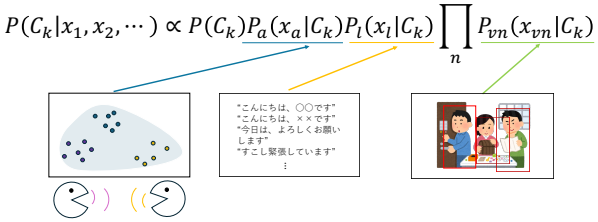


Fig. 4 Integration of acoustic, linguistic, and visual features using naive Bayes.

変換したものを複数話者である確信度と考え、これを言語的特徴量とする。Fig. 3 に確信度の分布を示す。単一話者動画と複数話者で分布が異なる事がわかる。

### 2.3 音響的特徴

音響的特徴として、話者表現ベクトル x-vector を使う。話者が複数人いれば話者表現ベクトルの分散が大きくなることが期待される [2]。今回は、[4] で公開されている学習済み x-vector 抽出器を使用した。各発話区間に対して計算された x-vector を t-SNE で 2 次元に削減し、その共分散行列の行列式の値を特徴量とする [2]。

## 3 ナイーブベイズ分類器による統合

視覚的、言語的、音響的特徴量をナイーブベイズ分類器で統合する (Fig. 4)。

言語的特徴量と音響的特徴量についてはガウス分布を仮定する。また、視覚的特徴量についてはカテゴリカル分布を仮定し、式(1)でモデル化する。

$$P_v(x=i|C_k) = \max\left(\frac{N_i^{(k)}}{\sum_i N_i^{(k)}}, 10^{-9}\right) \quad (1)$$
$$k \in (\text{single}, \text{multi})$$

ただし、$N_i^{(k)}$は訓練データにおいて$i$人が検出された画像の数を表す。

## 4 実験的評価

### 4.1 データセット

データセット作成のため、1467 個の動画を取得し、人手で複数話者動画またはそれ以外に分類した。その内訳は、複数話者動画 635 個、それ以外が 832 個となった。

これらのデータをテストデータ 2 割、訓練データ 8 割に分けて実験を行う。

### 4.2 結果

提案法の組み合わせによる複数話者動画検出の精度、適合率、再現率、F 値を Table 1 に示す。

Table 1 Accuracy of multi-speaker video detection

| | 精度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|
| – | 0.567 | N/A | 0 | N/A |
| 視覚 | 0.719 | 0.711 | 0.540 | 0.614 |
| 言語 | 0.701 | 0.664 | 0.633 | 0.648 |
| 音響 | 0.605 | 0.562 | 0.422 | 0.482 |
| 視覚&言語 | 0.741 | 0.750 | 0.609 | 0.672 |
| 言語&音響 | 0.721 | 0.702 | 0.625 | 0.661 |
| 音響&視覚 | 0.731 | 0.738 | 0.594 | 0.658 |
| 視覚&言語&音響 | 0.759 | 0.777 | 0.625 | 0.693 |

Table 1 より、すべての特徴を組み合わせたものが最も高い F 値となった。各特徴を単体で見ると視覚的特徴が最も良い結果となった。このことから、複数話者動画抽出において視覚情報が大きな手がかりとなることがわかる。一方で音響的特徴は視覚的特徴や言語的特徴に比べると大きく劣る結果となった。これは、データ中に音質が悪いものが含まれたことが原因だと考えられる。

## 5 おわりに

本研究では、YouTube 上の動画から複数話者動画を抽出するために、サムネイルなどの視覚的情報、発話内容の言語的情報、話者表現ベクトルを使った音響的情報をもとに分類を行った。

YouTube 動画 1467 個を使ってデータセットを作成し検証を行った結果、視覚的特徴、言語的特徴、音響的特徴を組み合わせたものが最も良い結果となった。

今後の研究では言語的、視覚的特徴による分類が難しいミュージックビデオなどの動画に対しての分類精度を上げるために、音響的特徴の抽出法の洗練を行っていく。

### 参考文献

[1] 小磯他, 国立国語研究所論集, 24, 153–168, 2023.
[2] Takamichi et al., arXiv, vol. abs/2112.09323, 2021.
[3] Redmon et al., CVPR, 779–788, 2016.
[4] https://github.com/sarulab-speech/xvector_jtubespeech