政治家の YouTube を対象とした話者識別の定量的評価の試み

渡邉隆誠 [1]　木村泰知 [1]　森浩太 [2]　加藤賢 [2] 御器谷裕樹 [4]　吉田光男 [5]　粕谷祐子 [4]
[1] 小樽商科大学 [2] 株式会社 JDSC [3]VETA 株式会社 [4] 慶應義塾大学 [5] 筑波大学
kimura@res.otaru-uc.ac.jp

概要

本研究では，政治家の発言が一貫しているかを分析するための準備段階として，YouTube 動画に含まれる複数話者の中から対象となる政治家の発言のみを自動的に抽出するための実用可能性を検証するために，既存ツールによる話者識別精度を定量的に評価した．実験では，議会・対談・ショート動画などを対象に，テキスト解析への応用を考慮して文字単位のマクロ F 値による評価を行った．結果として，演説や対談では実用的な精度が確認された一方，ショート動画では発話量の不足に起因する話者数の誤りが発生することが明らかとなった．本稿では，これらの誤り分析を通じ，政治的発言の自動分析に向けた技術的課題について議論する．

1　はじめに

近年，政治家による情報発信の手段として，YouTube や X（Twitter）などのオンラインメディアが定着し，議会会議録に加えて，動画や短文投稿を通じた発言が日常的に公開されている．これら複数の媒体における発言を横断的に分析することで，政治家の主張に一貫性があるのか，あるいは時期や媒体によって変化しているのかを検証することが可能となる．実際に，国会会議録や SNS を対象として，政治家の発言内容や立場を定量的に分析する研究が蓄積されつつある．

例えば，松本らは国会議員個人に着目し，委員会における発言内容を体系的に分析することで，政策関心や発言傾向の差異を明らかにしている [1]．また，SNS を対象とした研究として，三輪は政治家間のフォロー関係に着目し，ネットワーク構造を用いて政治家のイデオロギー的位置を推定する手法を提案している [2]．これらの研究は，テキストデータやネットワーク情報を用いて政治的立場や発言特性を分析する点で重要な知見を提供しているが，主としてテキストベースのデータを対象としており，動画に含まれる発言そのものを直接扱うものではない．

一方で，YouTube に代表される動画コンテンツは，テキストデータと比べて検索や参照が困難であり，人手による書き起こしには多大なコストがかかる．この課題に対して，自動音声認識（ASR）技術を用いた発話のテキスト化は有効な手段であり，近年では Whisper に代表される高精度なモデルが提案されている [3]．しかし，政治的発言の分析においては，「何が話されたか」だけでなく，「誰が話したか」を正確に特定することが不可欠であり，話者識別（Speaker Diarization）は重要な前処理となる [4]．話者識別については，阿坂らは半教師あり学習を用いた対話ダイアライゼーション手法を提案し，少量のラベル情報による話者分離精度の向上を示している [5]．しかし，政治家個人の発言抽出における実用性は検証されていない．

そこで，本研究では，政治家が異なる時点や場面において行った発言内容の一貫性を分析することを最終目標とし，YouTube 動画に含まれる複数話者の中から対象となる政治家の発言のみを自動的に抽出できるかどうかという実用的観点から検証を行う．具体的には，WhisperX[6] を用いて動画中の発話を文字起こしおよび話者分離し，対象政治家の発言テキストを自動抽出できるかを評価する．評価指標としては，文字数単位の再現率，適合率，および F 値を用い，政治に関する動画分析における既存の話者識別手法の適用可能性を明らかにする．

2　話者識別実験の概要

本実験では政治家発言の定量的分析を行うための基盤技術として，自動音声認識とその付随機能である話者識別の精度検証を目的とする．図 1 に話者識別実験の概要を示す．

2.1　対象動画の選定

表 1 に，本調査で対象とした動画を示す．本研究では，日本の代表的な政党を幅広く網羅することを

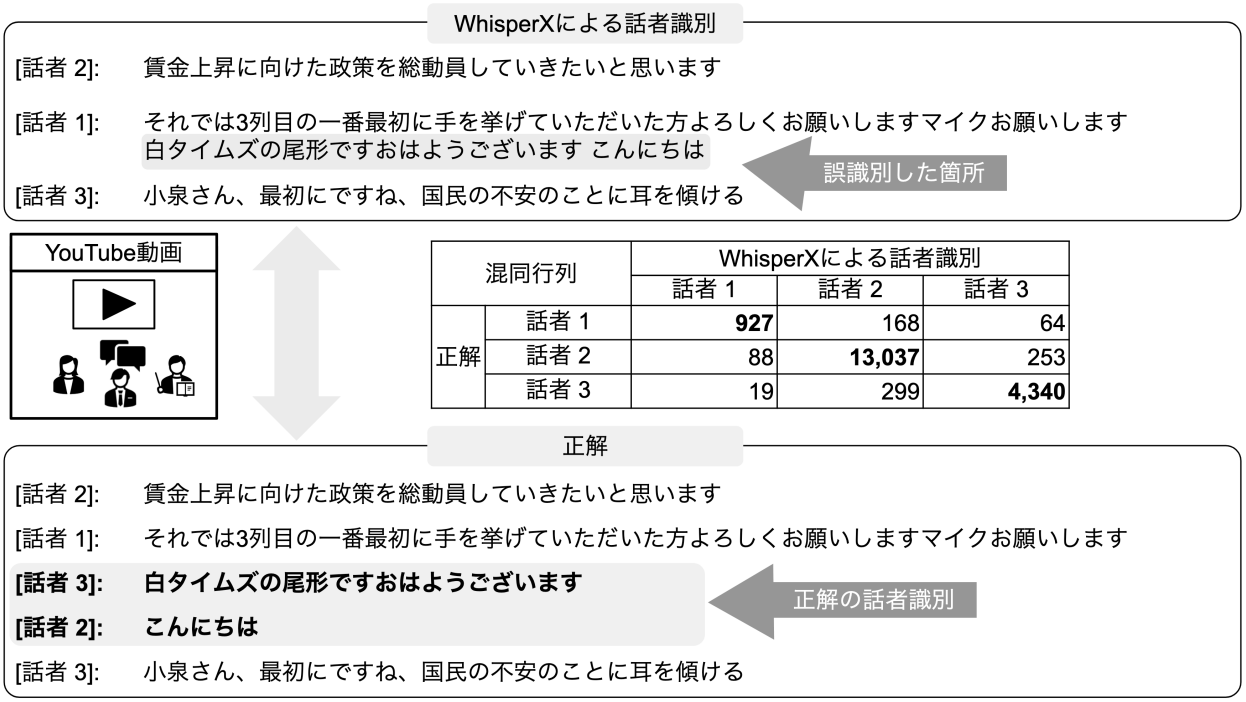

図 1　話者識別実験の概要

目的として，自由民主党の小泉進次郎，公明党の岡本三成，日本維新の会の前原誠司，国民民主党の玉木雄一郎，日本共産党の田村智子，参政党の神谷宗幣，ならびに「チームみらい」を率いる安野たかひろを分析対象とした．いずれも国政レベルでの知名度が高く，党の中核的立場や継続的な情報発信を通じて，各政党の発信傾向を一定程度代表していると考えられる．動画の選定にあたっては，研究者の恣意的判断や内容的偏りが生じないように，各政治家の公式 YouTube チャンネルからランダムに 3 本を選定した．

政治家の YouTube 活動は多岐にわたり，動画の種類によって収録環境や発話スタイルが大きく異なる．WhisperX による話者識別の精度が，どのような要因によって変動するかを詳細に分析するため，本研究では収集した動画を 7 つのカテゴリ (演説，ショート，対談，議会，密着，解説，記者会見) に分類した．

2.2　正解データの作成

正解データセットは，whisperX による話者識別精度を評価するために作成したものであり，システムが出力した SRT 形式の文字起こしデータを基に，人手によって構築した．具体的には，whisperX が生成した文字起こしテキストに対して実際の動画音声を

表 1　対象とした YouTube 動画の概要

政党	動画 ID	動画長	話者数	カテゴリ
共産	田村-1	1:55	2	演説
共産	田村-2	0:59	2	ショート
共産	田村-3	0:53	3	ショート
みらい	安野-1	0:42	3	ショート
みらい	安野-2	0:20	1	ショート
みらい	安野-3	0:57	2	ショート
維新	前原-1	16:39	2	対談
維新	前原-2	0:52	1	ショート
維新	前原-3	27:36	5	議会
公明	岡本-1	0:45	1	演説
公明	岡本-2	6:08	1	演説
公明	岡本-3	6:09	2	議会
国民	玉木-1	29:00	2	対談
国民	玉木-2	9:59	3	密着
国民	玉木-3	8:42	1	解説
参政	神谷-1	10:50	4	議会
参政	神谷-2	10:36	2	解説
参政	神谷-3	-	-	-
自民	小泉-1	1:37	2	ショート
自民	小泉-2	7:58	4	議会
自民	小泉-3	65:27	3	記者会見

聴取しながら，正しい話者ラベルを人手で確認および修正をした．これにより，システムの話者識別結果の妥当性を，文字起こしテキストの文字単位で定量的に評価することが可能となる．なお，本評価は話者識別の精度検証を主眼とするため，音声認識自体の誤り（誤字脱字）の修正は行わず，システムが出力したテキストに対する話者ラベルの正誤のみを修正対象とした．

2.3 評価方法

本研究では，文字起こしされた（音声認識誤りを含む）テキストを対象に文字数を単位とした話者識別の評価を以下の３つの評価で行った．

再現率 (Recall)

$$\text{再現率} = \frac{\text{正しく識別された話者の文字数}}{\text{正解の特定話者の総文字数}}$$

適合率 (Precision)

$$\text{適合率} = \frac{\text{正しく識別された話者の文字数}}{\text{システムが特定話者と識別した総文字数}}$$

F 値 (F-measure)

$$\text{F} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

一般的に話者識別の評価には時間ベースの指標（DER 等）が用いられるが，本研究は発言内容（テキスト）で評価を行う．評価は「どの発言（テキスト）が誰に紐づくか」という点を重視し，文字単位で算出した各話者のスコアを単純平均するマクロ平均（Macro Average）により評価する．

2.4 実験設定

本実験では，Linux 環境において whisperX を用いて音声の文字起こしおよび話者識別を行った．音声認識モデルには large モデルを採用し，計算精度（compute type）は float16 に設定した．また，話者識別（speaker diarization）機能を有効化し，出力結果として SRT 形式の字幕ファイルを生成し，後段の評価に用いた．なお，話者数については事前に指定せず，ツールが自動的に推定した値を用いた．

WhisperX が出力する話者ラベル（例: `SPEAKER_00`）は，システムが独自に付与した匿名 ID であり，正解データの人物名とは直接対応していない．そのため，評価に先立ち，システム出力と正解データの間で話者ラベルの対応付け（Mapping）を行う必要がある．本研究では，システムが出力した各話者クラスタの発話区間と，正解データの各話者の発話区間を比較し，**両者の共通部分（文字数の一致）が最大となる組み合わせ**を一対一で割り当てる「最適マッピング」を行った．なお，システムが推定した話者数が正解話者数より少ない場合（過少推定）は，最も重複度の高い話者のみをマッピングし，割り当てられなかった正解話者については「検出なし（Missed）」として扱った．

表 2 話者識別性能（動画単位のマクロ平均）

動画 ID	識別 話者数	正解 話者数	マクロ 再現率	マクロ 適合率	マクロ F 値
田村-1	2	2	0.8696	0.8702	0.8682
田村-2	2	6	0.2522	0.1343	0.1642
田村-3	2	3	0.5374	0.4004	0.3649
安野-1	3	3	1	1	1
安野-2	1	1	1	1	1
安野-3	1	2	0.5	0.4366	0.4662
前原-1	2	2	0.8487	0.8983	0.8696
前原-2	1	1	1	1	1
前原-3	5	5	0.8161	0.8803	0.8307
岡本-1	1	1	1	1	1
岡本-2	1	1	1	1	1
岡本-3	2	2	1	1	1
玉木-1	2	2	0.882	0.8850	0.8832
玉木-2	3	3	0.6004	0.8144	0.5610
玉木-3	1	1	1	1	1
神谷-1	3	4	0.75	0.7394	0.7446
神谷-2	2	2	0.886	0.8567	0.8703
神谷-3	-	-	-	-	-
小泉-1	1	2	0.5	0.384	0.4344
小泉-2	3	4	0.7423	0.7068	0.7233
小泉-3	3	3	0.902	0.9313	0.9157

3 話者識別の結果

表 2 は，動画単位で話者識別性能をマクロ平均により評価した結果を示している．話者数が正解話者数と一致する動画では F 値が高く，特に単一話者や話者構成が単純な動画では完全一致（F 値=1）が多く見られた．一方，話者数の不一致や話者構成が複雑な動画では再現率・適合率が低下し，識別性能にばらつきが生じる傾向が確認された．

(a) 例 1：馬渕磨理子氏の発言区間への相槌の混入

認識結果: [SPEAKER_01]: トラス政権のことをいまだに事例として挙げることがあるんですけど「いっぱい言うよね石破内閣の時とか」でもあの時と日本を比較すると...
↗ 下線部は本来 **玉木雄一郎氏** の発言

図 2 誤識別の例：短い発話（下線を付したカギカッコ部分）の話者交代を正確に検出できない

話者識別の結果には，特定の話者については全発話が他話者に誤って割り当てられ，再現率が 0 となるケースが確認された．この場合，当該話者に対する予測発話が存在しないため適合率は不定となるが，本研究では評価の一貫性を保つため，適合率および F 値を 0 と定義した．これらの話者も含めてマクロ平均を算出することで，話者識別性能の失敗事例を含めた公平な評価を行った．

4 考察

本実験の結果に基づき，動画のスタイルや発話特性が話者識別精度に与える影響について考察する．

4.1 話者識別は実用的な精度なのか

考察対象の動画として，国民民主党・玉木雄一郎氏の YouTube チャンネルに公開されている対談動画『**【対談】経済アナリスト 馬渕磨理子さん＆玉木雄一郎　高市新総裁で日本経済はどうなる 日本経済復活の切り札とは？**』(玉木-1) を選定した．本動画は，約 29 分間の二者対談形式で構成されており，話者ごとの発言量および識別精度を評価するのに適したデータである．

表 3 に WhisperX による話者識別結果と正解データとの混同行列を示し，その混同行列から算出した話者別の評価指標を表 4 に示す．その結果，両話者ともに F 値が 0.87 以上となり，本動画においては**実用的な話者識別精度が得られている**ことを確認した．

誤識別の傾向としては，図 2 に示すように，短い相槌など両者が同時に発話する区間において，主要話者の発言として一括して識別される事例が見られた．ただし，**これらの誤認識された区間は意見表明などの重要な内容を含む箇所ではなく，発言内容分析への影響は限定的である**と考えられる．

4.2 ショート動画における精度の低下

誤り分析として，動画カテゴリにおいて「ショート」と分類された動画に焦点を当てる．ショート動画では，全体としてマクロ F 値が低い傾向が確認された．とりわけ，6 名の話者が存在する田村氏の動画（田村-2）では，システムが話者数を「2 名」と著しく過少に推定された事例が観察された．このような精度低下は，主に以下の 2 点に起因すると考えられる．

表 3 話者識別の混同行列（単位：文字数）

		システムによる識別	
		玉木	馬渕
正解	玉木	**4120**	709
	馬渕	508	**5185**

表 4 話者識別精度の評価結果

	再現率	適合率	F 値
玉木雄一郎	0.8532	0.8902	0.8713
馬渕磨理子	0.9108	0.8797	0.8950

第一に，動画が短時間であるため話者ごとの発話量が不足し，話者モデルに必要な統計量が十分に得られず，話者クラスタの分離が困難となった点である．第二に，ショート動画特有の高速なカット割りにより無音区間が短縮され，話者境界の検出やアライメントが不安定となった点である．これらの結果は，現行の WhisperX では「超短尺かつ多人数」の動画が適用範囲外となる可能性を示しており，政治家によるショート動画活用が進む中で解決すべき課題である．

5 おわりに

本研究では，政治家の発言内容を網羅的に分析するための基礎検討として，WhisperX を用いた話者識別精度を，多様なカテゴリの YouTube 動画を対象に定量的に評価した．テキスト解析への応用を想定し，文字単位のマクロ F 値を用いて評価した結果，演説や対談など一定の発話長が確保された動画では，実用的な識別精度が得られることが確認された．一方，ショート動画では，話者識別精度が大きく低下することが明らかとなった．

謝辞

本研究は, JST, RISTEX, (課題番号:JPMJRS25L2), および, JSPS 学術知共創プログラム（課題番号：JPJS00123811919）の支援を受けて実施された.

参考文献

[1] 松本俊太, 松尾晃孝. 国会議員はなぜ委員会で発言するのか？選挙研究, Vol. 26, No. 2, pp. 84–103, 2011.

[2] 三輪洋文. Twitter データによる日本の政治家・言論人・政党・メディアのイデオロギー位置の推定. 選挙研究, Vol. 33, No. 1, pp. 41–56, 2017.

[3] 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.

[4] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. **IEEE Transactions on Audio, Speech, and Language Processing**, Vol. 20, No. 2, pp. 356–370, 2012.

[5] 阿坂脩平, Yen Benjamin, 糸山克寿, 中臺一博. 話者情報の半教師あり学習を用いたオフライン話者ダイアライゼーション. 人工知能学会第二種研究会資料, Vol. 2024, No. Challenge-066, p. 04, 2024.

[6] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. **arXiv preprint arXiv:2303.00747**, 2023.