音声とテキストを用いたニュース映像中の話者の推定

田路 茂　渡辺 靖彦　岡田 至弘
龍谷大学 理工学部 電子情報学科
touji@mail433.elec.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

ニュース映像には、アナウンサーやレポータだけでなくニュースの主題に関連する人物が話者としてあらわれることが多い。こうした話者の情報はニュース映像の内容検索に重要である [1]。例えば、図1は首相が国会で行った所信表明についてのニュース映像から取り出したものである。この映像の話者が首相自身であるのか、それとも話者はアナウンサーあるいはレポータで首相の発言を引用しているのかは重要な情報である。そこで、本研究では、ニュース映像中での話者が誰であるのか、特にニュースの主題に関連する人物があらわれた場合はその人物が誰であるかを、音響特徴とニュース映像の内容を説明するテキストを利用して推定する方法について述べる。

話者が誰であるのか、またその話者がなにを話しているのかを調べるのには一般に音声認識の技術が用いられている。これまでの話者認識の研究では、あらかじめ作成した話者モデルを利用して話者を認識していた。しかし、ニュース映像にあらわれる人物、特にニュースの主題に関連する人物を予想してそれらの話者モデルをあらかじめ用意することはむずかしい。このため、ニュース映像の話者を話者モデルを用いないで認識する方法について検討するのは重要である。西田らは部分空間法による話者照合に基づいて自動的に話者のインデキシングを行い、ニュース映像中の話者を推定する方法を提案している [2]。この研究では、あらかじめ話者モデルを作成せず、話者の発話区間を抽出して話者照合と学習を繰り返し、話者のインデキシングを行っている。そして、大語彙連続音声認識とワードスポッティングによりニュース音声中から人名を抽出して、話者の名前を自動的に付与している。これに対して本研究では、音声認識の結果を利用するのではなく、映像内容を説明するテキスト (ニュース原稿) を利用してニュース映像中の話者を推定する方法を提案する。

これまでにわれわれは、映像の内容を説明するテキスト (ニュース原稿) を利用してニュース映像の論理的構造を抽出する方法を提案している [3]。しかし、その研究で用いたテキスト (ニュース原稿) には映像中の話者が誰であるのかは記述されていない。そこで、本研究で提案する方法を利用して話者を推定し、ニュース映像の論理的構造の情報と一緒に話者情報を記述して内容検索や編集に利用できるようにすることをめざす。

2　音響特徴を用いた話者交替の検出と話者同定

話者交替の検出にはベクトル量子化歪みによる方法 [4] や、隠れマルコフモデルによる方法などが提案されている。

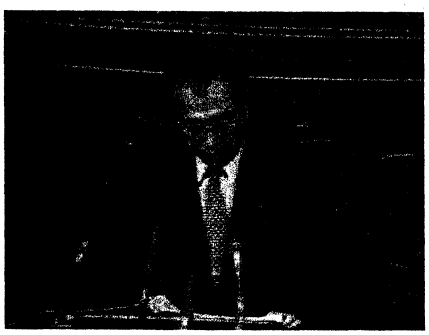

図 1: 国会で所信表明をする首相

これらの方法では、あらかじめ作成した話者モデルを利用することで話者の交替を高精度に検出している。しかし、ニュース映像の主題とそれに関連する人物をあらかじめ予想するのはむずかしいので、ニュース映像にあらわれる話者の話者モデルをあらかじめ作成しておくのはむずかしい。

そこで本研究では、話者モデルを利用せずに、話者が交替した可能性がある点の前後の音響特徴 (メルケプストラム) を比較して話者の交替を検出する。

最初にニュース映像中で話者が交替した可能性がある点の検出について述べる。対話が行われている場合をのぞいて、ニュース映像中では話者が交替するのはショットのかわり目 (カット点) である場合が多い。そこで以下の手順で話者が交替している可能性があるカット点を抽出する。

1. DCT 成分を手がかりに映像中のカット点を抽出する [5]。105 個のニュース映像を対象に行った実験では再現率 93%、適合率 79% の精度でカット点を検出することができた。

2. 検出したカット点のうち、発話の休止区間 (無音区間) に含まれるものを話者が交替している可能性があるカット点として抽出する。

次に、話者交替の検出と話者同定の方法について述べる。図2にその処理手順の概要と例を示す。

話者交替の検出と話者同定に用いる音響特徴としてメルケプストラム特徴を用いた。人間の聴覚特性は低い周波数域で細かい分解能を、高い周波数域で粗い分解能をもつことが知られている。メルケプストラムはこのような聴覚特性を考慮した非直線周波数軸上で定義されたケプストラムである。また、メルケプストラムは通常のケプストラムの半分程度の次数で表現でき、音声認識のパラメータとして用いたとき、高い認識率が得られるなどの利点をもってい

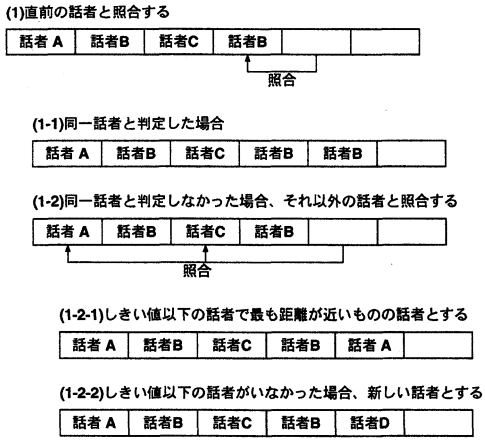

図 2: 話者交替の検出と話者同定の処理の手順と例

ると報告されている [6][7]。本研究では、メルケプストラム分析により抽出したメルケプストラム特徴を用いる。すなわち、話者が交替した可能性があるカット点の前後の映像断片に対してメルケプストラム分析を行う (図 2の (1))。抽出したメルケプストラム特徴間の距離がしきい値よりも小さければ同一話者と判定する (図 2の (1–1))。しきい値よりも大きければ話者の交替が行われたと判定する (図 2の (1–2))。ところでニュース映像中では、「アナウンサー → レポータ → アナウンサー」のように、同じ話者が二度以上あらわれることがある。そこで、話者の交替を検出すると、新しい話者がこれまでにあらわれた話者ではないか話者の同定を行う (図 2の (1–2))。話者同定でも、話者交替の検出と同様にメルケプストラムの距離尺度 [8] を用いる。すなわち、それまでにあらわれたすべての話者の中からメルケプストラム特徴による距離がもっとも小さいものを取り出し、その距離がしきい値以下の場合、交替した話者はその話者であると判定する (図 2の (1–2–1) では 5 番目の映像断片の話者を話者 A と判定している)。その距離がしきい値以上の場合、交替した話者は新しい話者であると判定する (図 2の (1–2–2) では話者 D)。

3 テキストを用いた話者の推定

本節では音響特徴 (メルケプストラム特徴) を利用して行った話者交替の検出・話者同定の結果と映像内容を説明するテキスト (ニュース原稿) を用いて、ニュース映像中にあらわれる話者が

- アナウンサー
- レポータ (記者など)
- ニュースの主題に関連する人物 (首相、大統領など)
- その他

のいずれであるか、推定する方法について述べる。

3.1 ニュース原稿

話者を推定するために利用する映像内容を説明するテキストにはニュース原稿を用いた。ニュース原稿とはアナウンサーやレポータがニュース中で読み上げる原稿である。このため、ニュース原稿の記述とアナウンサーおよびレポータの発話はほぼ一致する。一方、ニュースの主題に関連する人物の発話はニュース原稿には記述されていない。例えば、図 1に示す映像断片の話者は首相であるので、その発話に対応する記述はニュース原稿中にはない。なお本研究では、話者推定を行う映像断片の発話に対応する記述をニュース原稿から人手で取り出し、それぞれの映像断片と関係づけた。

3.2 テキストを利用した話者推定

話者交替の検出結果と話者同定の結果、そしてニュース原稿を用い、それぞれの映像断片の話者を以下の手順で推定する。図 3に処理手順の概要と例を示す。

手順 1 最初の話者をアナウンサーと判定する。また、話者同定の結果から最初の話者と同じ話者であると判定された話者もアナウンサーと判定する。

手順 2 手順 1で話者が推定できなかったもののうち、3 回以上あらわれた話者をレポータと判定する。

3 回以上としたのはニュースの主題に関連する人物が話者として 2 回あらわれることがあるからである。

手順 3 手順 1と 2で話者が推定できなかったものについて、ニュース原稿を利用して話者がニュースの主題に関連する人物であるものを検出する。

ニュース映像では、ニュースの主題に関連する人物が話者になる映像断片の 1 つ前の映像断片ではアナウンサーやレポータがその人物の名前を述べることが多い。このとき、アナウンサーやレポータが利用するニュース原稿にも、その人物の名前が記述されている。そこで、話者を推定する映像断片の 1 つ前の映像断片での発話に対応するニュース原稿中の記述に対して以下の規則を順に適用し、話者がニュースの主題に関連する人物であるかどうか判定する。

規則 1 ハ格の格要素が人間あるいは組織で、それが「述べました」「話しています」など、ある人物が話したことを明示的に示す用言[1] にかかるとき、その人物あるいは組織の構成員を話者として判定する。

(例文 1) 首相は 次のように 述べました。

規則 2 ハ格の格要素が人間あるいは組織である場合、その人物あるいは組織の構成員がニュースの主題と関連する人物であり話者として判定する。

[1]本研究では次の 13 種類の用言とする: 示しました, 指示しました, 述べました, 強調しました, 答えました, 挨拶しました, 明らかにしました, 非難しました, 質しました, 話しています, 話しました, 語りました, 話していました

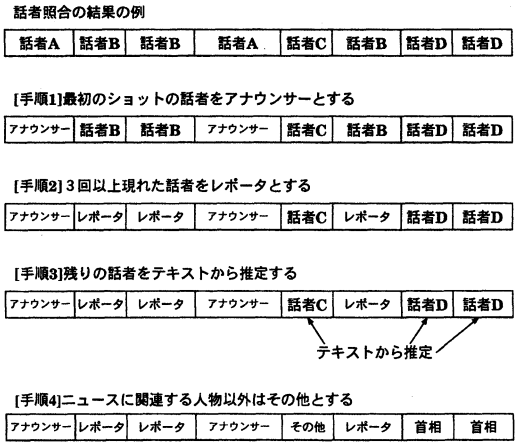

図 3: 話者推定の手順

(例文 2) 首相は大統領と会談を行いました。

なお、ハ格の格要素が人間あるいは組織であるかどうかはシソーラス [9] を用いて判定した。

手順 4 手順 1 ～ 3 で話者が推定できなかったものについて、手順 2 でレポータと判定された話者がいる場合はその話者をその他と判定する。手順 2 でレポータと判定された話者がいない場合はその話者をレポータと判定する。

4 実験と検討

4.1 話者交替の検出

メルケプストラム分析により抽出した音響特徴を用いて話者交替を検出する実験を行った。実験には 1998 年 8 月の「NHK 7 時のニュース」 1ヶ月分のニュース映像 67 個 (それぞれ 2 ～ 4 分程度の長さ) を使用した。この 67 個のニュース映像には話者が交替する可能性があるカット点が 241 個あった。音響特徴を抽出した際の実験条件を表 1に示す。話者交替の検出結果を表 2に示す。それぞれの正解率は次のように定義した。

$$\text{同じ話者であると判定した正解率} = \frac{\text{隣接する映像断片の話者が同じであると正しく判定したカット点の数}}{\text{隣接する映像断片の話者が同じであるカット点の数}}$$

$$\text{話者が交替したと判定した正解率} = \frac{\text{話者が交替したことを正しく判定したカット点の数}}{\text{隣接する映像断片の話者が違うカット点の数}}$$

なお、話者の交替を判定するしきい値を 7 に設定した。

同じ話者と正しく判定することができなかった原因として、映像断片に含まれる無音区間や、騒音などによる雑音が考えられる。映像断片に含まれる無音区間や雑音が多くなると、話者の特徴が変動するため同じ話者でも違う話者と判定してしまう場合があった。

表 1: 音響特徴を抽出する際の実験条件

サンプリング周波数	11kHz
フレーム長	30ms
フレーム周期	10ms
窓タイプ	ハミング窓
特徴パラメータ	メルケプストラム (26 次)

表 2: 話者交替の検出結果

	正解率
同じ話者であると判定	81% (76/93)
話者が交替したと判定	85% (126/148)

一方、話者交替を正しく判定することができないものとして、 1 つの映像断片に複数の話者が存在する場合がある。提案した手法では、 1 つの映像断片には 1 人の話者と仮定していて、複数の話者があらわれる場合を考慮していない。複数の話者があらわれる映像断片には、次々と話者が変わっていくものや、会話が行われているものがあった。

提案した手法は、特徴ベクトルを 1 つ (メルケプストラム分析により抽出した音響特徴の平均ベクトル) だけを用いて話者交替の検出を行っている。したがって、検出精度をあげるために、話者の特徴ベクトルを複数用いて話者交替を検出する方法について検討する必要がある。

4.2 話者推定の実験

話者交替を検出した結果をもとに話者を推定する実験を行った。話者交替の検出の実験 (4.1節) で用いたニュース映像のうち、 61 個のニュース映像には対応するニュース原稿があった。この 61 個のニュース映像には 269 個の映像断片があり、それらの話者を推定する実験を行った。この 269 個の映像断片の話者の内訳は、アナウンサーが話者であるものが 122 個、レポータが話者であるものが 85 個、ニュースの主題に関連する人物が話者であるもの 28 個、その他が話者であるものが 34 個であった。

手順 1 によって 114 個の映像断片の話者がアナウンサーと判定された。このうち 104 個が正解であった。手順 1 によるアナウンサーの検出結果は適合率が 91% (114 例中の 104 例が正解)、再現率が 85% (122 例中の 104 例を検出) であった。手順 2 と手順 4 によるレポータの検出結果は、適合率は 60% (96 例中の 58 例が正解)、再現率は 68% (85 例中の 58 例を検出) であった。手順 2 でレポータが話者であると判定した映像断片の数は 63 個で、そのうち 41 個が正しかった。手順 4 でレポータが話者であると判定した映像断片の数は 33 個で、そのうち 17 個が正しかった。手順 1 と手順 2 によるアナウンサーとレポータの判定の失敗は話者交替の検出の失敗によるものである。

次に、ニュースの主題に関連する人物が話者である場合の推定について検討する。ニュースの主題に関連する人物が話者である映像断片は 28 個あり、テキストの情報にもと

表 3: 話者はニュースに関連する人物ではないのにまちがってそうであると判定した原因 (適合率を悪くする原因)

原因	例
(音響特徴による) 話者交替の検出失敗によるまちがい	10
導入映像によるまちがい	5
話者の情報が述べられていない	2
合計	17

表 4: 話者はニュースに関連する人物であるのにまちがってそうではないと判定した原因 (再現率を悪くする原因)

原因	例
(音響特徴による) 話者交替の検出失敗によるまちがい	2
導入映像によるまちがい	4
対応する記述がニュース原稿にない	2
話者の情報が述べられていない	7
合計	15

づいて正しく話者を推定できたのはそのうち 14 個であった。映像断片中の話者がニュースの主題に関連する人物ではない (アナウンサー・レポータなど) のにまちがってそうであると判定したのが 17 例あった。それらの失敗の原因を表 3に示す。一方、映像断片の話者がニュースの主題に関連する人物であるのにまちがってそうではないと判定したものが 15 例あった。それらの失敗の原因を表 4に示す。以下では、これらのまちがいの原因について述べる。

導入映像の例を図 4に示す。図 4(b) は国会で所信表明をする首相の映像であるが、ニュースではこの映像の前に衆議院議長が首相の名前をよぶ図 4(a) の映像が挿入されていた。本研究では図 4(a) のような映像を導入映像とよぶことにする。この導入映像のため、話者推定は次のように失敗する。この映像の前にアナウンサーが「首相は次のように述べています」と述べた場合、図 4(a) の映像の話者を首相と誤って推定してしまう。さらに、図 4(a) の映像の発話に対応する記述はニュース原稿にはないので、図 4(b) の映像の話者を首相と推定することに失敗する。

表 3に示すように、話者の情報が述べられていないことによって判定に失敗した例が 2 つあった。この失敗は、話者推定を行っている映像断片の 1 つ前の映像断片での発話でハ格の格要素になった人物とは異なる人物が次の映像断片の話者になっていたために生じた。

表 4に示すように、対応する記述がニュース原稿にないため判定に失敗した例が 2 つあった。これは、話者推定を行っている映像断片の 1 つ前の映像断片での発話に対応する記述がニュース原稿にない場合である。ニュース原稿のかわりにクローズドキャプションを利用できれば、この原因で失敗することはなくなると考えている。

謝辞

本研究を進めるにあたり、さまざまな助言と援助をいただいた龍谷大学の西田昌史さんと京都大学大学院の秋田祐哉さんに感謝いたします。

(a) 話者: 衆議院議長

(b) 話者: 首相

図 4: 導入映像の例

参考文献

[1] 西尾 章治郎, 上原 邦昭, 有木 康雄, 加藤 俊一, 河野 浩之 : "情報の構造化と検索", 岩波講座、マルチメディア情報学 (2000)

[2] 西田 昌史, 緒方 淳, 有木 康雄 : "話者と発話内容の同時検索に関する検討", 人工知能学会研究資料, -SIG-CII-2000-NOV-12. (2000)

[3] Watanabe,Y., Okada,Y., Kurohashi,S., Iwanari,E.: "Discourse Structure Analysis for News Video", ICME2000, (2000).

[4] 森 一将, 山本 一公, 中川 聖一 : "発話間の VQ 歪みを用いたオンライン話者交替識別と話者クラスタリング", 信学技法, SP2000-18 (2000)

[5] 岩成, 有木: "DCT 成分を用いたシーンのクラスタリングとカット検出", 信学技報, PRU93-119 (1994)

[6] 徳田 恵一, 小林 隆夫, 深田 俊明, 斎藤 博徳, 今井 聖 : "メルケプストラムをパラメータとする音声のスペクトル推定", 電子情報通信学会論文誌. Vol.J74-A No.8, pp.1240-1248 (1991)

[7] 徳田 恵一, 小林 隆夫, 深田 俊明, 今井 聖 : "音声の適応メルケプストラム分析", 電子情報通信学会論文誌, Vol.J74-A No.8, pp.1249-1256 (1991)

[8] 今井 聖 : "音声認識", 共立出版 (1995)

[9] 国立国語研究所 : "分類語彙表", 秀英出版 (1964)