音声言語情報処理 11－3
（1996．5．27）

パネル討論音声の話者と話題に関する自動インデキシングの検討

三村 正人　河原 達也　堂下 修司
京都大学 工学部 情報工学教室
〒606-01 京都市 左京区 吉田本町

あらまし

本研究ではパネル討論音声の自動インデキシングについて検討する。一般にパネル討論音声は複数の話者の複数の話題区間からなるものと考えられる。機械により入力音声を同一話者(話題)区間毎に区切り、かつ各区間に対して話者(話題)のラベルが与えられれば検索上、有益である。話者インデキシングについては、テキスト独立話者識別の技術に基づく手法を提案し、実験を行った。また話題インデキシングについては各小区間に対する話題同定の結果に基づいて行うが、その際、予稿からドメイン知識を取得することを考え、実験を行って手法の有効性を検討した。

和文キーワード　自動インデキシング、話者識別、話題同定

Automatic Indexing of Speakers and Topics for Panel Discussion Speech

Masato Mimura　Tatsuya Kawahara　Shuji Doshita
Department of Information Science, Kyoto University
Sakyo-ku, Kyoto 606-01, Japan
e-mail: mimura@kuis.kyoto-u.ac.jp

Abstract

In this report, we study automatic indexing for panel discussion speech. Panel discussion speech consists of different speakers and topics. It is useful to devide input speech by speaker(topic) boundary and give a speaker(topic)label to each section. About speaker indexing, we propose a method based on text-independent speaker recognition. Topic indexing is performed using results of topic identification for each small section. For topic identification, we propose to learn domain knowledge from manuscript for discussion.

英文 keywords　automatic indexing, speaker verification, topic identification

1 緒論

膨大な音声データの中から望む情報に効率的にアクセスするには、機械によるインデキシング、自動検索技術の支援が必須となる。最近では入力音声のおおまかな内容を同定する話題同定の研究が盛んに行われている [1] [2]。これらの研究では入力音声全体が同一の話題からなるものと見なしており、タスクもニュース音声などドメイン知識の不要な一般性の高いものである。またこれとは別に、異なる話題区間からなる音声を同一話題区間毎に切り分ける音声要約の研究もなされている [3]。これは音声認識に基づかないという利点があるが、逆に切り出された区間に対してラベルを与えないので、検索には適さない。

本研究ではパネル討論音声をタスクとする自動インデキシングについて検討する。一般にパネル討論音声は複数の話者による複数の話題区間からなる。討論を通しての連続した音声を入力として、それを話者、話題の境界で区切り、かつ各区間に対して話者ラベル、話題ラベルが与えられれば、検索上、有益であろう。従って本研究では、パネル討論音声に対するインデキシング処理を、話者インデキシングと話題インデキシングのふたつの要素に分けてとらえる。

2章では全体の処理の概要および仕様について述べる。3章ではテキスト独立話者識別に基づく話者インデキシングについて述べる。4章では予稿を用いた話題同定の手法を述べ、またそれに基づく話題インデキシングについて述べる。

2 インデキシング処理の概要

本研究でインデキシングとは、複数の話者、複数の話題からなるパネル討論音声を、それらの境界によりセグメンテーションを行い、同時に各区間に対して話者ラベル、話題ラベルのインデックスを与えることを言う。このインデキシング処理の概要を図1に示す。

まず入力音声に対して話者に関するインデキシング処理を行う。これはテキスト独立話者識別の結果に基づいて行われ、従ってセグメンテーションとラベル付けとは平行して行われることになる。

各参加話者のテンプレート音声は人手で予め登録するものとする。またパネル討論にはインデキシングの対象とはならないような話者も参加することが考えられるが、それらの話者の棄却は今後の課題とする。

次に、ここで得られた各同一話者区間に対して、話題に関するインデキシングを行う。同一話者の発言内でも話題は変化することが考えられるので、話者によるセグメンテーションと同様に、話題の境界も抽出

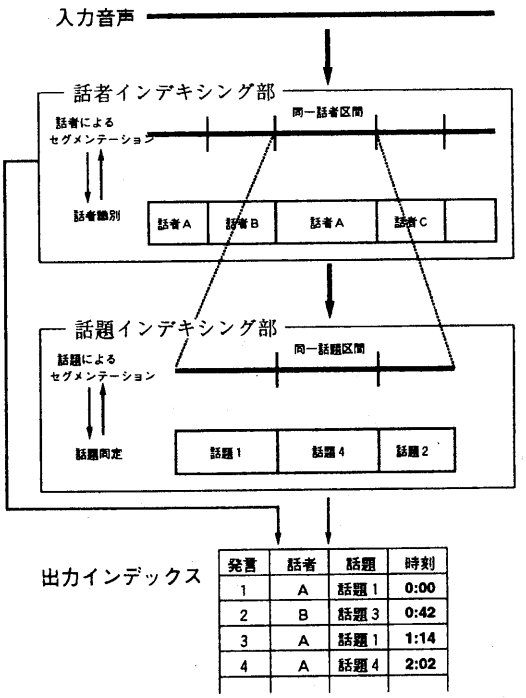

図 1: パネル討論音声のインデキシングの概要

しなければならない。このセグメンテーションは小区間毎に対する話題同定の結果に基づいて行う。話題同定のためのドメイン知識は、パネル討論に際して用意される予稿から取得する。予稿は容易に入手でき、また各話題とそれについてのテキストとの対応が明確である。予稿は通常話題毎に章分けされ、各章のテキストの内容は極めて話題依存性が高いと考えられる。

3 話者インデキシング

3.1 処理の概要

話者インデキシングは、

- テキスト独立に、話者を識別する。
- 話者の境界を抽出する。

の二つの処理からなる。この二つは互いに独立な処理ではなく、実際には入力音声の先頭から識別を行いながら、その結果に基づき、並行して話者境界の抽出も行なわれる。

話者識別には様々な手法が提案されているが、本研究ではベクトル量子化歪に基づく方式を用いる [4]。

この識別器の単独での性能を調べるために、登録話者44名で識別実験を行ったところ、学習用音声が

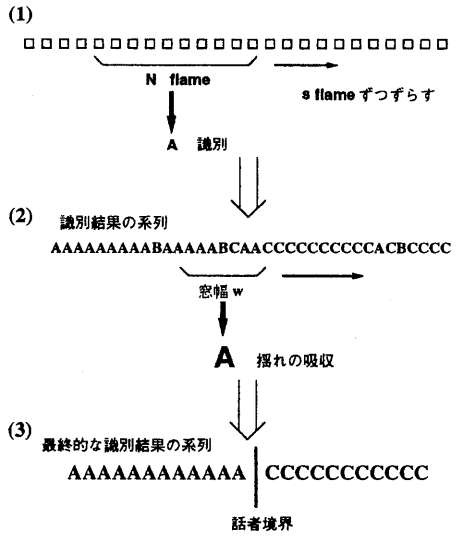

図 2: 識別結果に基づく話者境界の抽出

十分大きければきわめて高い性能 (最大で 96 %) を示すことが判った。

3.2 セグメンテーションの手法

前節で述べた識別器を用いて話者によるセグメンテーションを行う。入力音声の先頭から一定サイズの小区間毎に識別を行っていき、その結果の変化する箇所を境界として抽出する。ところが、入力音声は (とりわけ話者境界の付近では) 必ずしも安定な状態で取り出せるとは限らず、そのような箇所では識別器の性能が崩れ、識別結果の系列には揺れが生じることがあり得る。これをそのまま話者境界として抽出すれば誤りになるが、話者が数フレーム程度の短時間に入れ替わるということは現実にあり得ないので、周囲いくつかの結果を見れば、この揺れは吸収できるであろう。

本研究での手続きは、以下のようになる。

1. 入力音声の先頭から N フレームを 1 単位に、これを s フレームずつずらしながら識別を行っていく。
2. 周囲 w 個の識別結果を見て、そのうちで最も多く現れる結果のみを実際の話者と見なすことによって、揺れを識別誤りとして棄却する。
3. 2 で得られた最終的な識別結果の系列を見て、そこで話者の変化している箇所を実際の境界として抽出する。

この処理の様子を図 3 に示す。

この手続きでは、話者の境界にはポーズが挟まるという仮定は行っていない。

表 1: セグメンテーションのための識別結果

総区間数	16973
正解区間数	16463
不正解区間数	510
識別率	97.0
最大連続不正解区間数	9

(登録話者 5 名、 codebook size 64)

3.3 話者インデキシング実験

以上に述べた手法に基づき、話者境界を抽出する実験を行った。

実験データとして、実際に行われたパネル討論の書き起しテキストをもとにした朗読音声を用意した。このパネル討論は 95 年 5 月に行われた「話し言葉の文法構築は可能か」(情報処理学会音声言語処理研究会・自然言語処理研究会合同) ([5]) であり、予稿を用意したパネリストは 5 名である。実際に実験に用いるのは、この 5 名のみが参加する討論の前半部分とする。

前節で述べたように、まず入力音声の先頭から N フレームを単位に話者識別を行っていく。この N が小さ過ぎると識別の精度が下がり、また大き過ぎると (実際の境界フレームは N フレームの中のいずれかに埋もれることになるので) 精度良く境界を求めることができなくなる。従って N は識別精度を高く保ちながら、粒度細かく境界を調べられるように選ばなければならない。本研究では、識別精度の大きく崩れない最小のフレーム数として、 N=100 を求めた。

次にこの識別結果の系列から揺れを吸収し、話者境界を求めることになるが (図 3 の手続き 2)、この際に見る窓の幅は 20 とした。また用いた識別器のコードブックサイズは 64 である。

3.3.1 実験結果と考察

前節の条件のもとにセグメンテーションの実験を行った。まず参考のため、話者識別単独での結果を表 1 に示す。これは実験データの先頭から取っていった 1 つ 100 フレームの単位区間に対する識別結果 (すなわち、図 3 における手続き 2 での識別結果の系列に対応する) である。

表 1 から判るように、登録話者が 5 名と少ないにも関わらず、 3.1 節に示した登録話者 44 名の場合に比べて、識別精度はそれほど高くはなっていない。これは識別を行う単位が 100 フレームと短いため、ごく短い不安定な箇所にも影響を受けやすいためであ

る。しかしながら、識別性能は極めて局所的にしか崩れず、長い区間に渡って識別誤りを起こすといったことはない (表中、最大連続不正解区間数の項)。

連続して識別を誤る場合は、識別された結果が全て同一話者であるという傾向があり、これが $w/2$ 個以上続けば、誤りは揺れとして吸収されず、この区間もひとつの話者区間として抽出されてしまう。

今回の実験では、このような識別誤り区間の最大長は 9 であり (表 3 参照)、$w/2 = 10$ よりも小さいので、そのような誤りはなかった。

以上のように、識別結果に基づくセグメンテーションは、ほぼ適切に行われた。すなわち、

- 実際には境界ではない箇所 (境界とは大きく離れた箇所) に境界を抽出することはなかった。
- すべての話者境界を抽出した。
- 抽出された各話者区間に対する話者名のインデックスは正しく行われた。

本手法は話者境界にポーズなどを一切仮定しておらず、従ってさまざまなタスクに適用可能と考えられる。これにより適切に分けられた各話者区間を、それぞれ後段の話題同定の入力とする。

4 話題インデキシング

4.1 処理の概要

2 章で示した通り、話題インデキシングは話題境界によるセグメンテーションと、各区間に話題ラベルを与えることの二つの処理からなる。すなわち

1. 同一話者の音声区間を一定サイズの (その内部で話題が変化しないと考えられる程度の) 小区間毎に分割する。
2. 各小区間に対して話題同定を行う。
3. この同定結果の系列を見て、話題の境界を抽出する。

という手続きになる。

4.2 予稿を用いた話題同定

話題同定とは、入力音声の内容 (ひとつの話題からなると仮定) を、既に判っているいくつかの話題のひとつに対応づけることを言う。

話題同定を行うためには、各話題候補の取得、および各話題に対する話題モデルの学習が必要となる。本研究ではそれらドメイン知識の学習に、討論参加者の用意する予稿を用いる。この枠組みでは、予稿の各章題を話題名として、入力音声がどの章の内容を話しているかを同定することになる。話題同定のために

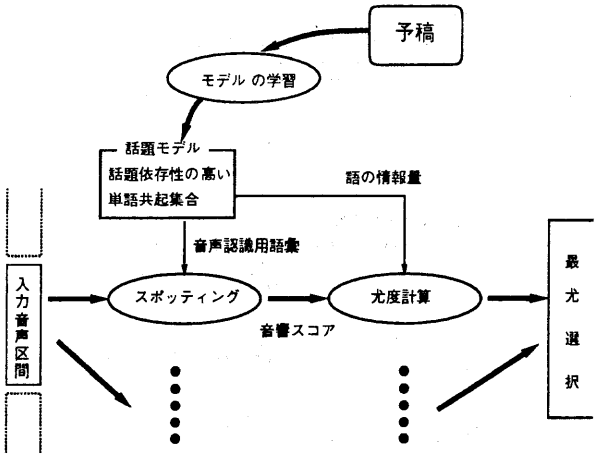

図 3: 話題同定の概要

は、まず各話題のモデル化を行わなければならない。本研究ではキーワード集合に加えて単語共起の利用も考える。同定はこうして作られた各話題モデルにより尤度を算出し、そのなかで最もスコアの高い話題を選ぶことで行われる。この様子を図 4 に示す。

図中に示したように、音声ベースの話題同定では、各語の持つ話題依存性 (相互情報量) の評価に加え、認識の際の音響スコアも統合するのが適当である。

理想的には、各小区間毎の同定結果の系列をもとに話題の纏まりが抽出され、それにより話題境界が求まることになる。

4.3 単語共起による話題モデル

予稿は入手しやすく、かなりの程度ドメイン知識を反映したものであると考えられる。従ってこれを話題モデルの学習に用いるのは有効であると考えられる。しかし一方、予稿は学習データとしては小規模であるという問題がある。従来のキーワード集合のみによるモデルでは、得られたキーワードが妥当なものである統計的な保証がない。そこで語と語との組み合わせ、一文中の共起をも考慮した、より強力なモデルを利用することを考える。ある語は文中に単独で出現したときよりも、他の語との組み合わせで出現したときの方が、より強くその文脈上の特徴を表すものと考えられる。予稿テキスト中から各話題毎に重要な単語共起を選定し、入力音声中にはその共起の組の両方の語が短時間中にともに現れたとき、スコアに加えることとする。

具体的には、まず予稿テキストの各話題 (章) 毎

に、一文中に共起する名詞の組で、かつその相互情報量の高いものをいくつかずつ選定する。相互情報量は次式で評価する。

$$I(w_1,w_2;T) = p(w_1,w_2|T)\cdot (\log\frac{1}{p(w_1,w_2)} - \log\frac{1}{p(w_1,w_2|T)})$$

このように重要な単語共起が求められると、次に音声中からそれらの語を抽出する。これにはヒューリスティックスポッティングの手法を用いる [6]。本研究では話題の境界を求めるため、一定サイズの小区間毎に話題同定を行っていく。この小区間中に共起の両方の語が現れたときのみ、スコアに加算する。ただし、共起のスコア (相互情報量) に統合する音響スコアは二つの語のうち低い方の値とする。すなわちある話題 T_i の尤度は次式で評価される。

$$P(T_i) = \sum_{w_1,w_2} I(w_1,w_2;T_i)\cdot \min\{f(w_1,t_1,t_2), f(w_2,t_1,t_2)\}$$

ここで例えば $f(w_1,t_1,t_2)$ とは、共起の一方の語 w_1 が小区間 t_1~t_2 に現れるヒューリスティックスコアである。

4.4 テキストベース話題同定実験

予稿テキストから各話題の単語共起集合が正しく得られたかを調べるために、まずテキストベースでの話題同定実験を行う。単語共起は、各話題毎に相互情報量の高いものから 100 ずつ選定して用いる。また今回実験に用いた予稿では、章 (=話題) の総数は 25 であった。

まずパネル討論音声を 10 秒毎の小区間に分割する。本実験ではこの小区間は 451 あった。テキストベースの実験には、この各小区間に対応する書き起しテキストを用意して用いる。

話題の尤度は前節で示したものとほぼ同様に評価する。すなわち小区間中に共起の両方の語がともに現れたときのみ、その共起の情報量をスコアに加算していく。表 2 にテキストベースでの同定結果を示す。

結果に示したように共起が一つでも現れている区間では、人間が判断して正しい同定結果が得られた (すなわち同定された章のテキスト内容にほぼ合致した)。従って予稿から学習された共起集合は、各章の内容を十分に特徴づけるものであったと考えられる。

しかし一方、10 秒程度の小区間では、特徴的なキーワードや共起のほとんど現れない、話題を持たない区間も多く、そのような区間では正しい結果が出なかった。そこでテキストベースでは、人手で話題の纏まりを抽出したある程度長い区間を入力として再び話題同定実験を行ってみた。このような話題区間は 21 あった。その結果、第一候補までで 14 の区間、第二候補までで 17 の区間で正しい同定結果が得られた。誤った区間では、単語共起が正しく学習されなかったというよりも、人手で見てもどの話題かが判断しにくいもの、あるいは予稿から得られる限られた話題候補中に近いものが見いだせなかったものであった。従って、予稿の少ないテキストから、単語共起を利用すればほぼ妥当な話題モデルが得られることが判った。

表 2: テキストベース話題同定結果

入力小区間のテキスト内容:
ポーズ情報というは、なんかポーズとかプロソディーとか色々ありますけども、一番ポーズ情報というのが取っつきやすい
同定結果: 「ポーズの調査」 ○
出現単語共起数:3
入力小区間のテキスト内容:
規則自体は，ある程度あちこちで使い回しする事はできる。統計コーパスからの統計でスコア付ける事はできる。
同定結果: 「N-gram ベースと文法ベース」 ○
出現単語共起数:1

4.5 音声ベース話題同定実験

次にヒューリスティックワードスポッティングを用いた音声ベースの同定実験を行った。この入力は前節でも示した 10 秒の長さを持つ 412 の小音声区間である。スポッティングの結果、実際に抽出するのは同一単語で上位 2 位までとする。この結果を表 3 に示す (テキストベースの実験と同じ区間について示す)。

図中、同定結果の欄に○で示されたのが正解である。この結果からある程度多くの共起の現れている区間では正しい同定が行えたことが判る。これは、スポッティングに伴う (話題に依存しない) 沸き出し誤りを抑えるためには、少なくとも 3 つ程度の共起が実際に含まれていることが必要であるためと考えられる。

また沸き出し誤りが多いのは、予稿から得られる語彙はもともと小規模であり、それをもとに共起を学習するため、各語の音響距離を大きくするなどの音声認識レベルでの語の選択は行っていないためであるとも考えられる。すなわち各共起集合は話題をよく特徴

表 3: 音声ベース話題同定結果

入力小区間のテキスト内容:
ポーズ情報というは、……
同定結果: 「ポーズの調査」 ○
出現単語共起数:3
入力小区間のテキスト内容:
規則自体は、……
同定結果: 「今後の展望」
出現単語共起数:1

づけているが、語彙のサイズがそもそも小さく、また予稿のみから統計的な処理で人手を介さず学習しているので、実際の音声に頻繁に現れる語を十分カバーしきれていないという問題がある。

ヒューリスティックスポッティングを適切に行うため、また話題の変化しない程度の長さという基準で、10秒というサイズを設けたが、この程度の長さでは話題を持たない区間が続出した。このような区間に対しては正しい同定は行えないので、このままでは話題の纏まりを抽出するのは難しいと考えられる。

4.6 まとめ

予稿を用いて話題候補と話題モデルを取得し、話題の明確な区間に対してはほぼ妥当な同定結果が得られた。しかしながら、パネル討論音声には、話題を持たない区間、また予稿から大きく外れた話題の区間が多く含まれるため、全ての音声区間に対して話題ラベルを与えようとするのは難しいと考えられる。従って本研究では話題境界の抽出まではできなかった。

一方ドメイン毎に人手で大量の学習データを与えてやるのは現実的でない。検索上の都合からも、予稿の内容との対応でラベルを与えるのが適当であろう。従って、まずユーザが予稿上で検索したい区間を示し、それに対応する話題区間を音声中からスポッティングするといったようなアプローチがより適当であると考えられる。また予稿を用いるにしても、語彙サイズが小さいなどの問題を補うために、ある程度人手で操作する支援システムという方向も検討しなければならないだろう。

5 結論

パネル討論音声をタスクとした話者と話題に関する自動インデキシングについて検討した。

話者インデキシングについてはテキスト独立話者識別の手法を用いて、ほぼ適切に話者によるセグメンテーション、および各区間へのラベル付けがなされた。本手法は話者境界にポーズ等の仮定を一切行っておらず、さまざまなタスクに適用可能と考えられる。今後はテンプレート用音声の自動登録、不要話者の棄却などを検討したい。

話題インデキシングについては、予稿からドメイン知識を獲得し、それを用いた話題同定についてまず検討した。話題によるセグメンテーションのためには小区間毎に話題同定を行っていくことが考えられるが、話題を持たない区間や予稿から外れた話題区間をどう棄却するかの問題が残り、入力音声の全区間に対してセグメンテーションを行うのは無理がある。前章で述べたようにより検索に適した仕様を検討する必要があるだろう。しかしそれ以外の区間では、正しい同定が行われ、単語共起を用いれば予稿からほぼ妥当な話題モデルが学習されることが判った。今回は話題のまとまりを抽出することができなかったが、音声内容に関するテキストデータからドメイン知識を学習するための基本的な枠組みが得られた。ニュース音声と新聞記事、講義とシラバスなど、この枠組みはさまざまなタスクに適用され得る。

参考文献

[1] 横井謙太郎、河原達也、堂下修司: キーワードスポッティングに基づくニュース音声の話題同定, 情報処理学会研究会報告, Vol. SLP-6, pp. 15–20 (1995).

[2] 伊藤慶明、木山次郎、岡隆一: スポッティングに基づくTVニュース番組の音声対話理解と音声検索, 音講論, pp. 3–P–22 (1995).

[3] 木山次郎、伊藤慶明、岡隆一: Incremental Reference Interval-free 連続 DP を用いた任意話題音声の要約, 信学技報, Vol. SP-95, pp. 81–88 (1995).

[4] F.K.Soong、A.E.Rosenberg: On the Use of Instantaneous amd Transitional Spectral Information in Speaker Recognition, *ICASSP*, pp. pp.877–880 (1986).

[5] 中川聖一、他: 話し言葉の文法構築は可能か, 情報処理学会, Vol. SLP-6, pp. pp.51–60 (1995).

[6] 河原達也、宗続敏彦、堂下修司: ヒューリスティックな言語モデルを用いた会話音声中の単語スポッティング, 電子情報通信学会論文誌, Vol. Vol.J78-D-2 No.7, pp. pp.1013–1020 (1995).