自然言語処理 143－11
音声言語情報処理 36－6
（2001. 6. 1）

談話標識の抽出に基づいた講演音声の自動インデキシング

長谷川 将宏　秋田 祐哉　河原 達也

京都大学 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町
e-mail: hasegawa@kuis.kyoto-u.ac.jp

あらまし　パラグラフの先頭部分に頻出する特徴的な単語 (談話標識) を用いて講演音声に対して自動インデキシングを行う手法を提案する。本研究では、種々の講演のなかでも流れが比較的明確で共通性のある学会講演を対象とする。学習セットの講演の書き起こしからポーズ情報を用いてパラグラフ境界を検出し、統計的言語モデルを用いて句点を挿入して各パラグラフの先頭の一文を抽出する。その中に含まれる名詞から tf・idf に基づいて談話標識を選定する。評価データの各文について談話標識の tf・idf 値を計算し、その合計が閾値以上であればインデックスを付与する。実際の講演音声の書き起こしと認識結果に対して評価を行った結果、再現率は 90%程度 (適合率は 20%程度) となり、高精度にインデキシングできた。

キーワード　講演音声, 談話標識, 自動インデキシング, tf·idf

Automatic Indexing of Lecture Speech by Extracting Discourse Makers

Masahiro Hasegawa　Yuya Akita　Tatsuya Kawahara

Graduate School of Informatics, Kyoto University
Kyoto 606-8501, Japan
e-mail: hasegawa@kuis.kyoto-u.ac.jp

Abstract　We address a method of automatic indexing for lecture speech by suggestive words that frequently appear in the initial sentences in each paragraph, and we define such words as discourse markers. We deal with academic presentations because these presentations can be segmented into relatively distinct parts. At first, we segment transcriptions into paragraphs and sentences by using average length of pauses during the lecture as a threshold. Next, each paragraph is segmented into sentences by using a statistical language model. Then, discourse markers are selected from nouns based on tf and idf statistics. We evaluated these discourse markers with recall and precision rate on indexing task of the lecture speech. Sentences are indexed if sum of the tf·idf value of detected discourse markers exceeds a threshold. As a result, we achieved a recall rate of 90%.

keyword　lecture speech, discourse markers, automatic indexing, tf·idf

1 緒論

近年、計算機の性能が飛躍的に向上し、音声メディアをディジタルアーカイブとして保存できるようになった。しかし、音声アーカイブは一見して内容を把握することがテキストや映像以上に困難である。したがって、求める情報を効率よく検索するには、音声アーカイブにインデックスが付与されていることが必要であるが、インデックスを人手で付与することは手間と時間を要し、大量のデータに対して人手で行うことは困難である。そこで、音声アーカイブに対して音声認識を適用することを考える。

本研究では講演音声を対象として自動インデキシングの検討を行う。講演には予稿があることが多いが、実際に予稿を読み上げる人はほとんどおらず、かなり自由な発話が行われるので、話し言葉特有のくだけた言い回しや言い淀み、発話速度の変化や発声の怠けなどの種々の特徴が含まれる [1]。そのため、現在の音声認識技術では講演のような話し言葉を高精度で認識することができない。講演音声を自動認識した結果には誤りが多く含まれているため、そのまま講演録として使用できるレベルではなく、人手による修正や編集が必要である。

したがって音声メディアの形で保存しておいた上で、音声認識結果を利用してそのインデックスを自動的に付与することを考える。話題の転換点のインデックスとなるような単語を談話標識として定義し、講演の書き起こしテキストから自動抽出する。これをもとに講演音声の認識結果から自動的にインデキシングを行う。これによって、講演音声の高速な検索や内容の把握が可能になる。

2 講演音声の自動インデキシング

本研究では、要約を作成するのではなく、録音音声の参照を容易にするための自動インデキシングについて検討する。音声情報が適切にインデキシングされていれば、何がどの部分に述べられているかが直ちにわかり、参照する際にきわめて有効である。音声認識技術を用いて講演音声に対して自動でインデキシングすることを目標とする。

先行研究では、ニュース音声の話題同定 [2][3] や講義音声の文単位へのセグメンテーション [4][5]、会議音声のアーカイブ化 [6] などの研究がなされていたが、いずれの研究でも連続音声認識ではなく、単語スポッティングや韻律情報を併用している。これに対して本研究では、大語彙連続音声認識を用いる。

講演には、その分野や話題、長さ、スタイルなどによって様々な種類がある。使用する語彙や講演の流れはその種類によって異なるため、すべての種類の講演に対してインデキシングを行うことは難しい。そこで、本研究では講演の流れが比較的明確で共通性がある学会での研究発表を対象とする。種々の学会講演が融合研究プロジェクトで大規模に収録、データベース化されており、これを利用する [7]。

いくつかの学会での研究発表 (主に工学系) の書き起こしを分析したところ、各講演の構成には一定のパターンが存在することがわかった。そのモデルを図 1 に示す。多くの場合、導入・背景、目標、手法の概要とその説明、実験とその評価、まとめの 5 つの部分に大きくわけることができ、またこの順で述べられている。これらの各部分は一つのパラグラフからなっている場合もあれば多数のパラグラフからなっている場合もある。

パラグラフの先頭の一文は、そのパラグラフで述べようとしている内容を短く端的に表している。例えば、「本報告のアプローチについて説明します。」、「次にその実験結果を示します。」などといったものである。したがって、パラグラフの先頭部分に対して正しくインデキシングすることができれば、講演音声のどの区間にどのようなことが述べられているかを容易に知ることができる。図 1 のモデルの各部分の境界はパラグラフの境界の中に含まれているので、パラグラフの先頭部分を検出し、その部分に対してインデキシングすることにより講演音声のインデキシングとする。

学会講演の書き起こしを用いてパラグラフの先頭の一文を取り出してみたところ、この部分に頻出する特徴的な単語が見られた。例えば「実験」や「説明」、「結果」、「背景」、「今回」、「最後」などである。本研究では、このようなパラグラフや話題の転換点を示すような単語を談話標識とよぶ。この談話標識を検出することで、パラグラフの先頭に対してインデキシングができると期待できる。

本研究では、談話標識を講演の書き起こしテキストコーパスを利用して自動的に抽出し、講演音声の音声認識結果から、談話標識の出現する部分を検出することによってインデキシングを作成する。

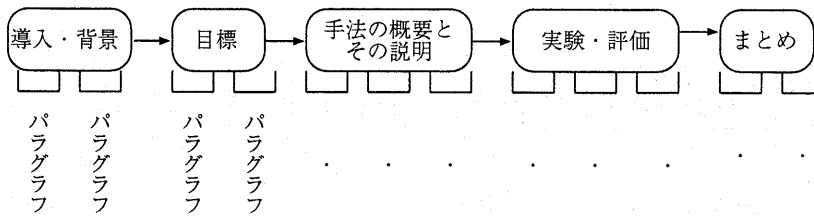

図 1: 学会での研究発表 (工学系) のモデル

3 インデキシングのための談話標識の抽出手法

3.1 抽出手法の概要

あらかじめ人手で選定した談話標識 (表 1) に基づいて予備実験を行ったところインデキシングが十分可能であることがわかった。そこで、ここでは談話標識を自動的に抽出するための手法について述べる。図 2 に処理の流れと談話標識の抽出に使用する情報を示す。

学習セットには、表 2 に示した融合研究コーパス中の講演の忠実な書き起こしとその形態素解析結果を使用する。形態素解析には ChaSen ver2.02 を用いている。

まず学習セットの講演をパラグラフ単位に区切る必要があるが、融合研究コーパスにはパラグラフのタグは付与されていない。パラグラフ境界は話題の転換点であるので、スライドを使う講演であればここでスライドを換える可能性が高い。また、スライドを使用しない講演であっても、話者はここで一呼吸おくことが多い。したがって、パラグラフ境界には通常の発話間より長いポーズが挿入されると考えられる。本研究では、適当なポーズ長の閾値を定め、その長さ以上のポーズが挿入されている部分をパラグラフ境界の候補とする。

表 1: 人手により抽出された談話標識 (16 個)

今日	研究	実験	目的	結果	我々
説明	背景	発表	今回	最後	今後
課題	まとめ	評価	検討		

表 2: 学習セットとして使用する講演

講演の種類	講演数
日本音響学会 春季&秋季研究発表会 (AS)	42
言語処理学会 年次大会 (NL)	7
国語学会 (JL)	5
音声学会 全国大会 (PS)	9
国立国語研究所内の種々の研究会 (KK)	6
融合研究の会合 (YG)	3
合計	72

次に、パラグラフの先頭の一文を抽出する。そのためには各パラグラフを文単位に区切る必要がある。文と文の間には通常ポーズが挿入されているので、ポーズによって文境界を抽出できる。しかし、言い淀みや言い間違いの際に生じるポーズも文境界として誤検出してしまう。

そこで、本研究では文境界の判定、すなわち句点の挿入に統計的言語モデルの単語 N-gram を利用する。文と文の間にはポーズが存在すると仮定する。ポーズの前後の単語列に対して、ポーズ部分に句点を挿入した単語列としない場合の単語列の言語的な尤度 (パープレキシティ) を計算し、比較することによって文境界か否か判定する。

最後に、パラグラフの先頭の一文から談話標識を抽出する。学習セットからは、パラグラフの先頭として抽出された文に出現している単語を抽出し、tf(単語頻度) と idf(文書頻度の逆数) の統計量を利用して談話標識を選択する。評価音声に対しては、この談話標識を用いてパラグラフ境界を絞り込みインデックスとする。

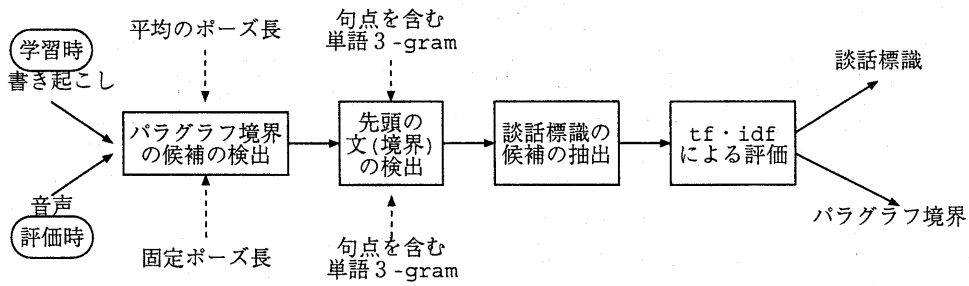

図 2: インデキシングの処理 (学習時と評価時) の概要

3.2 ポーズ情報を用いたパラグラフ境界候補の検出

話題の転換点であるパラグラフ境界に挿入されるポーズ (ロングポーズ) は、パラグラフ内の文の間に挿入されるポーズ (ショートポーズ) よりも、長いポーズになっていると考えられる。

そこで、適切な閾値を決定することができれば、閾値よりも長いポーズが挿入されている部分をパラグラフの境界の候補として抽出する。

ここでは、後で談話標識の情報を用いて絞り込みを行うので、できるだけ正しい境界がもれないように候補を抽出することが望ましい。すなわち、多少適合率が悪くとも再現率が十分に高いことが必要である。ただし最適な値は講演によって異なると考えられる。これは、講演者によって発話速度が異なっているためである。発話速度が遅い人であれば文と文との間に挿入するポーズが長く、速度が速い人であれば短くなる。したがって、すべての講演に対して同じポーズ長を閾値にして区分化するのではなく、講演者の発話速度に応じたポーズ長を用いる必要がある。

そこで、書き起こしにあるポーズ情報をもとに、各講演毎にポーズの平均の長さを求めた。

3.3 単語3-gramモデルを用いた文境界の検出

インデキシングすべき部分は各パラグラフの先頭の一文とし、各講演のパラグラフ境界候補を検出した後に、パラグラフの先頭の一文を取り出す。しかし、学習コーパスは句点などによって文単位に区切られておらず、また同コーパスによって構築された言語モデルを用いた音声認識結果にも句点は挿入されないので、これらを文単位に区切る必要がある。

本研究では、文境界の候補を検出するために、句点が含まれる Web 講演録から学習された単語 3-gram モデルを利用する。学習に用いた講演数は 81 個、学習テキストサイズは 1692802 語、総語彙数は 37462 語である。これらからカットオフを 1 にして単語 3-gram モデルを作成した。

文と文の間にはポーズが挿入されると仮定し、ポーズ部分での句点の有無による言語モデル尤度の差異に基づいて判定する。具体的にはポーズのその前の 2 つの単語 w_1、w_2と、後ろの 2 つの単語 w_3、w_4 を取り出す。そして、その 4 つの単語をそのまま並べた単語列 "w_1,w_2,w_3,w_4" の尤度 $P(w_1,w_2,w_3,w_4)$ と、句点を挿入した単語列 "w_1,w_2, 句点,w_3,w_4" の尤度 $P(w_1,w_2,$句点$,w_3,w_4)$ を計算しそれらを比較する [8]。そして、句点を挿入した場合の尤度の方が大きい場合は、その部分に句点を挿入し文境界と見なす。ただし、本実験では尤度のかわりにテストセットパープレキシティ $-\frac{1}{n}\log P(w_1 \cdots w_n)$ を用いた。

しかし、この手法では文境界でない部分の誤検出が多く見られた。その多くが未知語列を含み、パープレキシティが非常に大きな値になってしまう場合である。また、句点を挿入した場合と句点を挿入しなかった場合のパープレキシティの値があまり変わらない場合にも誤検出が散見された。

したがって、単純に尤度を比較するのではなく判定にマージンをとることとした。具体的には、句点を挿入しなかった場合のパープレキシティが句点を

表 3: 単語 3-gram モデルを用いた区点挿入の結果

講演	再現率	適合率
AS00MAR011	61/61(100%)	61/78(78.2%)
AS00MAR015	31/32(96.9%)	31/52(59.6%)
AS00MAR020	69/69(100%)	69/81(85.2%)
AS00MAR026	48/51(94.1%)	48/67(71.6%)
total	209/213(98.1%)	209/278(75.2%)

挿入した場合のパープレキシティの 3 倍以内の値であった場合は、句点を挿入しないこととした。また、パープレキシティは未知語に対して高い値を示すが、文末に現れるような単語が未知語であるとは考えにくいので、パープレキシティが 1000 以上の値になっていた場合にもその部分には句点を挿入しないこととした。本研究では、誤って短く区切られるよりはある程度以上長い方がインデックスとして意味があり、ここでも適合率よりも再現率を優先することとした。

この手法を、AS00MAR011, AS00MAR015, AS00MAR020, AS00MAR026 の 4 つの講演を用いて評価した。実際にこの 4 つの講演の書き起こしから人手で文境界を判断し、正解を設定した。再現率と適合率の結果を表 3 に示す。

再現率はかなりよく、句点を挿入するべき部分には 90%以上の割合で句点を挿入できた。適合率は 60%程度から 80%程度とばらつきがあり平均すると 75%程度である。

なお、文境界でないのに誤って句点を挿入された部分として、文中に文末表現になりうる表現が使われていて、その直後にポーズが入っているものが多く見られた。例えば「後ろにですね＜ポーズ＞えっとどれだけの」、「使用しました＜ポーズ＞装置について」などである。また、「…のようなもの＜ポーズ＞…のようなもの＜ポーズ＞…のようなものについて」といったように、評価する対象や実験条件などを並べて述べている部分も文境界として誤検出することが多かった。他に「言いました＜ポーズ＞述べましたが」のような言い直しにおいても、句点の挿入誤りが見られた。

3.4 tf・idf 値を用いた談話標識の選択

学習セットに対して 3.3 節と 3.4 節で述べた手法により、各講演をパラグラフ単位に区切りパラグラフの先頭の一文を取り出した。次に、これらの文から談話標識となる単語集合を選択する。

談話標識として抽出する単語は多くの講演に共通して出現する単語なので、話題独立な単語であることが望ましい。ただし、話題独立な単語といっても、動詞や助詞は一つの講演内でパラグラフの先頭以外の部分にも多く出現するので、本研究では名詞に注目した。また、名詞の中でも固有名詞や数詞は話題独立な単語であるといえないので、これらを除いた名詞から談話標識を選択する。

ここでは、tf・idf 値を利用することを考えた [9]。N 個の文書からなる文書集合が与えられたときに、単語 T_iの単語頻度 tf_iは、全文書における単語 T_iの出現回数と定義される。また、文書頻度 df_iは全文書中の単語 T_iを含む文書数で、idf は N/df_iで定義される。本研究においては、名詞 w_iの単語頻度 tf_iはパラグラフの先頭の一文として抽出された文の集まりの中に名詞 w_iが出現する回数であり、df_iは学習セットの全講演中で対象とする単語が出現する文の数となる。ある名詞 w_iについて、tf_iの値が大きいとパラグラフの先頭の文によく出現していることを示し、df_iの値が大きいと多くの講演中にまんべんなく出現していることを示す。したがって tf_iの値は大きく df_iの値が小さいものを談話標識として選択する。

パラグラフの先頭の一文として抽出された文の形態素解析結果から、固有名詞、数詞を除いて全講演で名詞を抽出した結果、3000 個程度の名詞が得られた。このように多数になった理由としては、パラグラフ単位に区切る際に、適合率よりも再現率を重視し、誤った検出をかなり許したためである。

ここで、tf と idf を統合した評価尺度を導入した。各単語 w_iについて式 (1) からスコア S_{w_i}を求める。

$$S_{w_i} = \mathrm{tf}_i^a * log(b * \mathrm{idf}_i) \quad (1)$$

ここで、式 (1) の重み重み a と b の値を変化させることを検討したが、有意な差は見られなかったので、重みはどちらも 1 とした。ここで、これらの重みを決める評価値として F-measure を用いた。F-measure は式 (2) で表される。

$$F\text{-}measure = \frac{(1+\alpha) * recall * precision}{\alpha * recall + precision} \quad (2)$$

適合率よりも再現率を重視するので再現率の重みαとして1以外に$\frac{1}{10}$を用いた。

3.5 インデキシングの手法

新しい講演に対して、談話標識を利用して自動インデキシングを行う手法について述べる。インデキシングは講演音声を音声認識した結果に対して行う。

まず、テストセットの講演音声のパラグラフ境界の候補を検出する。音声認識に際しては、適当な閾値以上のポーズで音声を分割するので、この際に適切なポーズの閾値を用いればパラグラフ境界の候補で区切ることができる。談話標識の抽出時には書き起こしに記録されているポーズ長の平均値を用いてパラグラフ境界の候補を検出したが、講演音声に挿入されているポーズ長の平均値を自動的に求めることは容易ではない。そのため固定のポーズ長を用いることとした。

インデキシングは文の先頭に対して行うので、各文は音声区間の先頭から始まっている必要がある。談話標識が出現している文の先頭が音声区間の途中にあると、よいインデキシングになっているとはいえない。そこで、AS99SEP097とPS99SEP025の2つの講演に対して、平均のポーズ長として最低の値であった500msを閾値として各講演音声を分割し、どの程度の割合で文境界がファイルの切れ目になっているかを調べた。その結果、AS99SEP097では全63文中46文(73.0%)で、PS99SEP025では全125文中99文(79.2%)で音声区間の境界と文境界が一致した。これらを平均すると、77%程度の文境界が音声区間の境界と一致しており、固定のポーズ長として500msのポーズを使用して問題ないと考えた。

次に文の終端の検出を行う。音声認識の言語モデルに句読点が含まれない場合は、3.4節と同じ手法を適用して句点を自動挿入し、文境界を検出する。ショートポーズが句読点に対応づけられて言語モデルに含まれている場合は、認識時に句点が挿入され、文の境界が検出される。

最後に3.4節で決定した式に基づいて抽出した談話標識を用いて判定を行う。3.4節で抽出した談話標識が音声認識結果の中に出現した場合に、一文中に出現する談話標識のtf・idf値の合計が一定の閾値を上回った場合にインデキシングを行う。

4 評価実験

4.1 評価データの仕様

評価実験には表4にあるデータを用いた。これらの大半が日本音響学会での研究発表であり、図1のモデルに沿っていると考えられる。これらは学習セットには含まれていない。正解のインデックスは、パラグラフの境界を人手で付与した。

表4: 評価データの概要

	講演時間		正解
	時間	単語総数	インデックス数
AS99SEP097	12分	2508語	12
PS99SEP025	27分	5372語	11
AS99SEP008	12分	1943語	8
AS99SEP009	11分	1680語	9
AS99SEP011	13分	2541語	7
AS99SEP014	11分	2067語	8
AS99SEP015	12分	1871語	12
AS99SEP018	13分	1628語	11
AS99SEP019	14分	1926語	16
AS99SEP027	12分	1460語	18

4.2 談話標識の効果

まず、談話標識を用いてインデキシングを行うことの基本的な有効性を確認した。談話標識を用いない場合として、100msから5000msまで50msごとにポーズの閾値を設定し、閾値以上の長さのポーズ部分で書き起こしを分割して、それらの部分に対してインデキシングを行った。談話標識の数は75個である。談話標識を用いる場合は、500msのポーズで講演の書き起こしを分割し、その先頭の一文に含まれる談話標識のtf・idf値の合計を求めて、その合計値が閾値以上となった文にインデキシングを行った。この結果を図3、図4に示す。

これから談話標識を用いる手法(図4)の方がF-measureが大きな値になっている。特に、再現率(recall)の高い部分においては談話標識を用いたインデキシングの方が適合率(prescision)も高い値を示している。実際に再現率に重みをおいたF-measure(10)で両手法の差は顕著である。本研究では適合率よりも再現率を重視するので、インデキシングに談話標識を用いることは有効であるといえる。

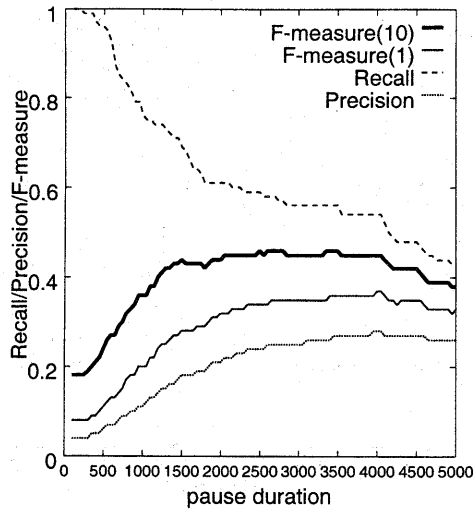

図 3: ポーズ長のみによるインデキシング結果

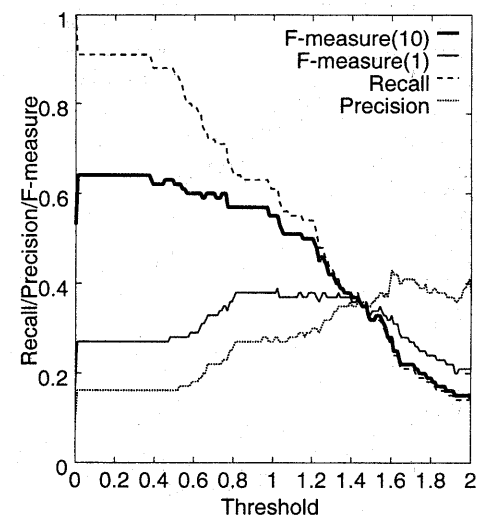

図 4: 談話標識を用いたインデキシング結果

4.3 談話標識の数の影響

次に談話標識の数による影響を調べた。談話標識の数として 25 個、75 個、125 個の値を用いてインデキシングを行った。これらは tf・idf 値 (S_{w_i}) の大きい順に抽出している。その結果を図 5 に示す。

談話標識の数が 75 個の場合が最も F-measure の値がよい。談話標識の数が少なすぎると、インデキシングすべき部分を抜き出せないことが多くなり、また、多すぎると無関係の部分を多く抜き出してしまうためである。したがって、談話標識の数は 75 個とする。

4.4 音声認識結果に対するインデキシング

次に、実際の講演音声を認識した結果にインデキシングし、手法の評価を行った。評価データとして表 4 の中の AS99SEP097 と PS99SEP025 の 2 つの講演を用いた。また、音声認識の際の音響モデルとして 2000 状態 16 混合の triphone を使用し、言語モデルとして融合研究コーパスと Web 講演録を用いて作成した単語 3-gram モデルを使用した [10]。単語認識精度は AS99SEP097 が 66.9%、PS99SEP025 が 58.4%である。

自動インデキシングの結果を図 6 に示す。比較として評価データの書き起こしに対するインデキシングの結果も示している。

これからわかるように認識結果に対して本手法を用いても書き起こしに対して行った場合とほとんど差がない結果が得られた。

5 結論と今後の課題

講演音声に対して談話標識を用いて自動的にインデキシングを行う手法について検討した。パラグラフの先頭の一文は、そのパラグラフで述べる内容を端的に表しており、これをインデキシングすることによって内容把握や参照が容易になると考えた。また、パラグラフの先頭部分には頻出する特徴的な単語 (談話標識) があるので、これに着目してインデキシングを行うことを考えた。そこで、学習セットの講演の書き起こしを用いて、談話標識を自動的に抽出する方法を提案した。

抽出された談話標識を用いてインデキシングを行い評価した。書き起こしテキストに対しては、再現率は 90%、適合率は 20%程度であり、談話標識を用いることによって高精度にインデキシングすることができた。

しかし、問題点として適合率がまだ十分でないこ

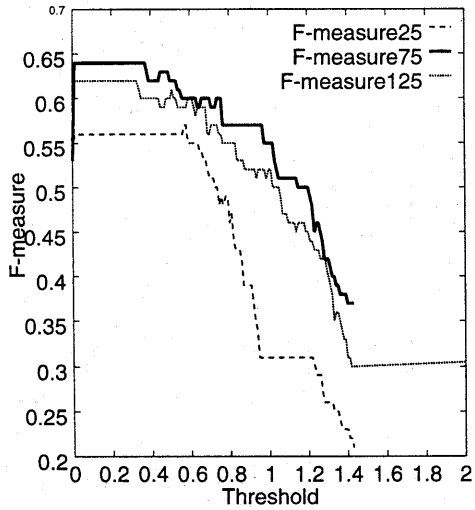

図 5: 談話標識の数のインデキシングへの影響

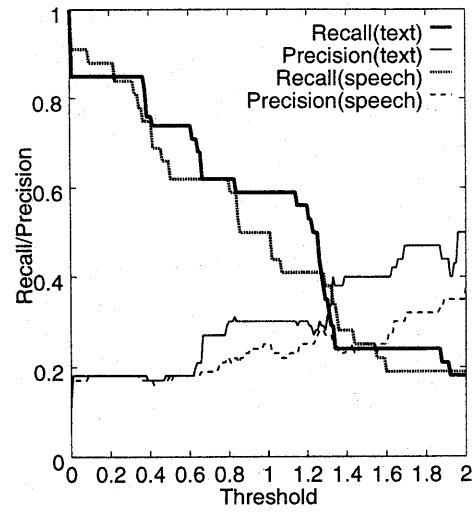

図 6: 音声認識結果に対するインデキシング

とが挙げられる。パラグラフの先頭以外の部分も多数インデキシングしてしまっていた。談話標識の数をさらに絞っても適合率はそれほどよくならないことから、今後、談話標識によってインデキシングするべき部分を検出し選択した後に、別の方法によってさらに絞り込むことが考えられる。また、本研究では講演の全体に対して同様の処理を行い談話標識を抽出したが、講演を図 1 に示したモデルの各部分ごとに分割して、談話標識抽出の処理を行うことで、抽出される談話標識をより精密なものにできると考えられる。

参考文献

[1] 峯松信明, 片岡嘉孝, 中川聖一. 講演調の話し言葉に対する言語的解析. 情報処理学会研究報告, 95-SLP-8-7, 1995.

[2] 横井謙太朗, 河原達也, 堂下修司. 単語の共起情報を用いたニュース朗読音声の話題同定機構. 電子情報通信学会技術研究報告, SP96-105, 1997.

[3] 横井謙太朗, 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. 情報処理学会研究報告, 95-SLP-6-3, 1995.

[4] 野村和弘, 河原達也, 堂下修司. 講義の自動アーカイブ化のための韻律情報を用いた講義音声の文境界の抽出. 電子情報通信学会技術研究報告, SP98-80, 1998.

[5] 野村和弘, 河原達也, 堂下修司. F0 パターンに基づく講義音声の文単位へのセグメンテーション. 電子情報通信学会技術研究報告, SP99-13, 1999.

[6] 秋田祐哉, 河原達也. 会議音声の自動アーカイブ化システム. 情報処理学会研究報告, 2000-SLP-34-11, 2000.

[7] 小磯花絵, 前川喜久雄. 『日本語話し言葉コーパス』の概要と書き起こし基準について. 情処学研報, 2001-SLP-36-1, 2001.

[8] 西村雅史, 伊東伸泰, 山崎一孝. 単語を認識単位とした日本語の大語彙連続音声認識. 情処学論, Vol. 40, No. 4, pp. 1395–1403, 1999.

[9] 長尾真（編）. 自然言語処理. 岩波講座ソフトウェア科学, 1996.

[10] 加藤一臣, 南條浩輝, 河原達也. 講演音声認識のための音響・言語モデルの検討. 情報処理学会研究報告, 2000-SLP-34-23, 2000.