情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2009-NL-191 No.8
Vol.2009-SLP-76 No.8
2009/5/21

日本語講演音声ドキュメント検索における索引付けの検討

重安幸治[†1]　南條浩輝[†1]　吉見毅彦[†1]

自然言語による講演音声ドキュメント検索について述べる．このような検索タスクでは適切な索引付けが重要であり，本研究ではこれに焦点をあてる．音声ドキュメント検索においては音声認識が行われるため，音声認識誤りに頑健な索引語の研究が必要である．さらに日本語では，語と語の間にスペースがおかれず語の区切りがあいまいである．したがって索引単位の研究も重要である．これらの背景に基づき，日本語話し言葉コーパスの音声ドキュメント検索評価用テストコレクションを用いて索引単位と索引語の研究を行った．ベクトル空間モデルに基づく音声ドキュメント検索システムを構築し，形態素，N 文字連鎖，それらの組み合わせの索引単位を研究した．

A Study of Indexing Units for Japanese Spoken Document Retrieval

Koji Shigeyasu,[†1] Hiroaki Nanjo[†1]
and Takehiko Yoshimi[†1]

Spoken document retrieval (SDR) from Japanese lectures is addressed. For SDR, appropriate indexing is significant. Automatic speech recognition (ASR) is performed to make index terms, and studies of indexing terms which are robust to ASR errors are necessary. In Japanese text, no space is put between words, and word unit is not obvious. Thus, studies of indexing unit are also important. Based on the background, indexing unit and index terms are investigated. We constructed SDR system based on the vector space model. As for indexing unit, morpheme, character N-gram, and combination of them were investigated.

†1 龍谷大学 理工学研究科
Graduate School of Science and Technology, Ryukoku University

1. はじめに

ネットワークの高速化とストレージの大容量化により，音声を含む動画データを容易に配信・保存できるようになった．過去の講義を学生に向けてネットワークを通じて配信している大学などもある．このような状況のもと動画データを処理するための様々な技術の需要は増加し，これらのデータを的確に検索する方法が求められている．

従来の検索では，Web 検索に挙げられるようにテキスト文書が主な検索対象であった．画像や音声，動画の検索も行われているものの，テキストによるメタデータの付与，すなわちデータ名を適切なものにしたり，検索用のテキストラベルを付与する必要があった．大量のデータに対して人手で検索用にリネームやラベル付けを行うにはコストがかかる．さらにこのようなラベルは動画のタイトルなど映像そのものを検索するために付与されることが多く，これのみでは，動画内の特定のシーンを探しだすことができないという問題があった．実際に講義映像などの検索を考えた場合は，DVD のチャプターのように話のまとまりごとに映像や音声を区切り，特定のシーンを直接視聴できることが望ましい．

このような背景に基づき，本研究では，講演で録音された音声から検索要求に合致する音声区間を検索する方法について研究を行う．具体的には，講演の音声に対して，音声認識を行って索引付けする方法を研究する．音声ドキュメント検索により，講演でわからなかった部分の復習が容易になることが期待できる．本稿では，講演単位での検索における索引語の検討を行った．

2. 講演音声ドキュメント検索

2.1 講演音声ドキュメント

本稿での研究対象は講演音声ドキュメントである．講演の映像には音声だけでなく話し手の身振り手振りや表情，スライドの画像などが含まれる[1)]．スライドを用いて行われる講演では，重要な用語を発話せずにスライドで指示するだけのこともある[2)]．この場合，従来の検索よりも難易度が上がる．スライドの文字などを解析して，索引語に追加することも考えられるが，本研究では，これは扱わず音声のみを検索対象として検索する方法を研究する．

2.2 情報検索のモデル

情報検索は，与えられた検索要求に適合するデータ（文書）を見つけ出す処理である．適合性の基準の観点から，全文検索と内容型検索の 2 つに大きく分けることができる[3)]．全文検索は，文書中から検索要求の文字列と完全に一致する部分を探し出すことを目的として

Vol.2009-NL-191 No.8
Vol.2009-SLP-76 No.8
2009/5/21

いる．内容型検索は，検索要求と意味的に類似した文書を探し出すことを目的としている．本研究では，内容型検索を行う．

内容型検索では，文書の内容を特徴付ける語（索引語）を抽出し，これらの索引語の出現頻度などを用いて文書および検索要求を表現することが一般的である．事前に索引データベースを作成しておき，文書本体ではなく索引データベースを参照して検索結果を出力することが多い．本研究では，このような内容型検索を実現するモデルとして，ベクトル空間モデルを用いる．これは，文書および検索要求を多次元のベクトルで表現し，文書と検索要求の類似度をベクトル間の類似度計算によって求めるものである．ベクトルの各要素は各索引語の重みであり，当該索引語の文書や検索要求での重要度である．したがって，ベクトル空間モデルでは，索引語の抽出と重み付けが重要である．

2.3 索引付け

索引語の抽出と各索引語の重み付けをする処理を索引付けと呼ぶ．人手による索引付けは作業コストが大きいため事実上不可能であり，自動索引付けの技術が求められている．

多くの場合，索引語として用いられるのは形態素である．日本語では形態素が空白で区切られていないため，その定義が問題となる．本研究でははじめに形態素解析システムを用いて形態素に分割して自動索引付けを行う．

本研究が対象とする講演音声の検索では，音声認識誤りを考慮した索引付けを考える必要がある．さらに形態素解析自体に誤りが含まれる可能性もある．この様な背景に基づき，形態素以外の索引付けとして文字連鎖を用いた索引付けも検討する．具体的には，文字列の先頭から 1 文字ずつずらしながら N 文字単位で索引付けを行う．ここでは，N 文字連鎖単位の索引付けと呼ぶことにする．

2.4 検索システム

本研究ではベクトル空間モデルに基づく文書検索システムを用いる．ベクトル間の類似度には SMART[4] を用いる．

ある質問 Q と文書 $D_i(1 \leq i \leq N)$ が与えられ，索引語を $t_k(1 \leq k \leq m)$ としたとき，質問 Q と文書 D_i のベクトル間の類似度 $\mathrm{SMART}(Q, D_i)$ は，式（1）で与えられる．

$$\mathrm{SMART}(Q, D_i) = \sum_{k=1}^{m} (q_{t_k} \times d_{i,t_k}) \tag{1}$$

$$d_{i,t_k} = \begin{cases} \dfrac{\frac{1+\log(\mathrm{tf}_{i,t_k})}{1+\log(\mathrm{avtf})}}{(1-\mathrm{slope})\times\mathrm{pivot}+\mathrm{slope}\times\mathrm{utf}_i} & \text{if} \quad \mathrm{tf}_{i,t_k} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$q_{t_k} = \begin{cases} \dfrac{1+\log(\mathrm{qtf}_{t_k})}{1+\log(\mathrm{avqtf})} \times \log \dfrac{N}{\mathrm{n}_{t_k}} & \text{if} \quad \mathrm{qtf}_{t_k} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

ここで，tf_{i,t_k} は文書 D_i に含まれる索引語 t_k の出現数を表す．avtf は各文書に含まれる語の出現数の平均を表す．pivot は各文書に含まれる異なり語数の平均を表す．utf_i は D_i 中の異なり語数を表す．slope は補間係数であり，本研究では 0.2 とした．

qtf_{t_k} は質問 Q における語 t_k の出現数を表す．avqtf は Q に含まれる語の出現数の平均を表す．N は検索対象の文書集合の全文書数を表す．n_{t_k} は t_k を含む文書の数を表す．

2.5 検索評価用テストコレクション

情報検索システムの評価を行う上で，検索質問文に対して文書集合中のどの文書が適合しているかという情報が必要になる．テストコレクションとは，文書集合，検索質問集合，適合情報を備えた情報検索システムの評価用データである．

本研究では，音声ドキュメント検索処理 WG によって作成されたテストコレクション[5] を用いて研究を行った．以下にこの詳細について述べる．

2.5.1 文書集合

検索対象の文書集合は，「日本語話し言葉コーパス」（以後，CSJ と略す）[6] である．

テストコレクションでは，CSJ の学会講演 987 件と模擬講演 1715 件の合わせて 2702 件の講演が検索対象となっている．学会講演の長さはほとんどが 10 分から 25 分程度であるが，なかには 1 時間を超えるものもある．模擬講演は，一般話者による日常的話題についての 12 分程度のスピーチである．

テストコレクションでは，この 2702 件の音声に対して，音声認識が行われており認識率は 65 %から 95 %である．本研究では，この音声認識結果を使って索引付けの研究を行う．

2.5.2 検索質問集合

テストコレクションにおける検索システムの評価用の検索質問集合は，検索対象文書の性質と講演音声を対象とすることを考慮して，以下の検索質問となっている[2]．

- 内容を問う質問

  「言い間違いを笑って取り繕う箇所を見つけたい」など従来の内容型検索で扱えないよ

Vol.2009-NL-191 No.8
Vol.2009-SLP-76 No.8
2009/5/21

うな検索質問はない．

- 10 件程度の適合情報が存在する
  適合箇所が CSJ 全体で一ヶ所となるような質問や，あらゆる講演に答えが見つかるような検索質問はない．
- 特定の分野に偏りがない検索質問
  様々な分野の講演が検索対象になるように偏りのない検索質問である．

このようにして合計 39 件の検索質問が作成されている．

2.5.3 適合情報

適合情報とは，検索質問集合の各検索質問文に対し，文書集合のどの文書が適合しているか，もしくは不適合であるかという情報である．本研究で扱うテストコレクションの適合情報には適合，部分適合，不適合の 3 段階が用意されており，本研究では適合判定のみを用いて評価する．

2.6 評価尺度

テストコレクションの検索対象と正解ファイルを用いて，各索引語単位の検索性能の評価を行う．今回は正解となる講演の一部を検索するのではなく，正解箇所を含む講演全体を検索対象として検索を行った．

評価尺度には式（4）に示す 11 点平均精度（11ptAP）を用いた[3]．再現率レベル L（ 0.0 から 1.0 まで 0.1 刻み）での補間精度 IP_L を平均した精度 11ptAP_k を各検索クエリ k に対して求め，各検索クエリの平均をとって評価を行った．

$$11\text{pt}AP = \frac{1}{N}\sum_{k=1}^{N} 11\text{pt}AP_k \tag{4}$$

$$11\text{pt}AP_k = \frac{1}{11}\sum_{l=0}^{10} IP_{\frac{l}{10}} \tag{5}$$

$$IP_L = \max_{L \le R_x} P_x \tag{6}$$

補間精度 IP_L は，再現率レベル L 以上の再現率 R_x を与える順位 x での適合率 P_x の最大値である．N は実験で用いた検索クエリの総数であり，$N = 39$ である．今回は，1 つのクエリに対して上位 1000 件まで検索し，1000 件のときの再現率 R_x よりも高い再現率レベル L の補間精度 IP_L は 0 とした．

3. 索引付けの実験と評価

音声ドキュメント検索に適した索引付けを行うために，種々の索引語の単位を検討する．索引語の単位には形態素と N 文字連鎖，それらの組み合わせを用いる．

検索システムには GETA[7] を使用し，全検索質問 39 件での 11ptAP を求めた．

3.1 形態素単位の索引付け

はじめに形態素単位で索引付けを行う．形態素解析には chasen ver2.2.1[8] を用いた．日本語には，漢字，ひらがな，片仮名，英数字などの文字種がある．単純なベクトル空間モデルでの検索のためには，検索質問文と索引語での語の一致が必要であり，表記の違いは大きな問題である．形態素単位の出現形，基本形，読みによる索引付けを検討する．

3.1.1 出現形の利用

はじめに最も単純な出現形について述べる．形態素が出現した形そのままで索引付けを行うものである．また出現形の読みを用いた検索も行う．表記を読みに統一することで表記の違いに対応することができる．例えば「煙草」と「たばこ」などであっても，読みの場合「タバコ」となり一致することが期待できる．

3.1.2 基本形の利用

次に基本形での索引付けについて述べる．動詞などの活用語に対して出現した形ではなく，語の基本的な形（終止形）で索引付けを行うものである．これにより索引語と検索質問の活用形の違いの影響をさけて一致をとることができる．

3.1.3 形態素単位の索引付けの評価結果

形態素ごとの索引付けの評価結果を**表 1** に示す．最も単純な形態素の出現形での索引付けでは，11ptAP 値 0.450 であった．読みでの索引付けでは，11ptAP 値 0.445 であった．形態素基本形で索引付けを行った場合は 0.466 であった．

形態素単位の索引付けにおいて，出現形と基本形では多くの検索質問に対して基本形の評価の方が高かった．読みによる索引付けでは漢字かな混じり表記による索引付けに比べて明らかに高い精度が得られた検索質問も見られた．例えば「煙草が体に及ぼす影響，有害性にはどのようなものがあるか？」（検索質問 ID:SDPWG-HN2010-02）という検索質問では，重要なキーワードである「煙草」が CSJ の音声認識結果では「たばこ」と表記されており，検索質問の漢字表記とマッチしなかった．読みでの索引付けでは，表記の違いの影響をさけることができ一致をとることができた．

Vol.2009-NL-191 No.8
Vol.2009-SLP-76 No.8
2009/5/21

表 1 形態素単位の索引付けの評価

索引語の単位	11ptAP
出現形 漢字かな混じり	0.450
出現形 読み	0.445
基本形 漢字かな混じり	0.466

表 2 形態素（出現形）と N 文字連鎖単位の索引語の例

索引語の単位	索引語
形態素（出現形）	世界 / 遺産 / に / は / どの / よう / な / ところ / が / ある / か
2 文字連鎖	世界 / 界遺 / 遺産 / 産に / には / はど / どの / のよ / よう / うな / なと / とこ / ころ/ ろが / があ / ある / るか
3 文字連鎖	世界遺 / 界遺産 / 遺産に / 産には / にはど / はどの / どのよ / のよう / ような / うなと / なとこ / ところ / ころが / ろがあ / がある /あるか
4 文字連鎖	世界遺産 / 界遺産に / 遺産には / 産にはど / にはどの / はどのよ / どのよう / のような / ようなと / うなとこ / なところ / ところが / ころがあ / ろがある / があるか

3.2 N 文字連鎖単位の索引付け

日本語では，形態素単位で索引付けを行うために形態素解析を行う必要がある．しかし，形態素解析は完全ではないうえに，日本語の正しい形態素区切りも明確ではない．例えば，「世界遺産」を 1 語にするか,「世界」と「遺産」で 2 語にするかがある．

そこで，形態素以外の索引付けの方法の 1 つとして N 文字連鎖単位に分割する方法を検討する[2)]．これは，文字列の先頭から 1 文字ずつずらしながら N 文字単位で索引語を抽出する方法である.「世界遺産にはどのようなところがあるか」を形態素の出現形と N 文字連鎖で分割した例を**表 2** に示す．N 文字連鎖単位では「世界遺産」のような複合語を扱うことを意識せず索引付けできる．

なお，文字連鎖で N を大きくした場合，索引語の種類は膨大な量になり，N が小さいとどの文書にも出現する不要な索引語が大量に作られてしまう．

3.3 N 文字連鎖単位の索引付けの評価結果

N 文字連鎖単位の索引付けの評価結果を**表 3** に示す．2 文字連鎖単位での精度は 0.420，3 文字連鎖単位では 0.346，4 文字連鎖単位では 0.302 となった．2 文字連鎖単位の索引付けでは平均的には形態素単位の索引付けよりも精度が低かったものの，一部の検索質問に対しては効果がみられた．実際に「世界遺産」を含む検索質問に対して，2 文字連鎖を用いた索引付けでは 11ptAP が 0.550 となった．形態素出現形単位の索引付けでは 0.361 であり，文字連鎖単位の索引付けの効果がみられた．このように 2 文字連鎖単位の索引付けで高い精度が得られた検索質問には，複合語が含まれているものがあり，これらに対しては形態素間の文字を補間する索引語，例えば「界遺」が使用され，その効果が高かったと考えられる．

表 3 N 文字連鎖単位の索引付けの評価

索引語の単位	11ptAP
2 文字連鎖	0.420
3 文字連鎖	0.346
4 文字連鎖	0.302

表 4 形態素と N 文字連鎖を組み合わせた索引付けの平均検索性能

索引語の単位	11ptAP
形態素基本形	0.466
2 文字連鎖	0.420
形態素基本形と 2 文字連鎖の組み合わせ	0.462

3.4 形態素と N 文字連鎖の並用

形態素と N 文字連鎖の 2 つの索引語の単位を組み合わせた索引付けも行った．索引語は形態素と文字連鎖の索引語を合わせたものになるため，総量は増大する．しかし，それぞれの単独では検索できなかったものに対して検索できるようになることが期待できる．

3.4.1 形態素と N 文字連鎖を並用した索引付けの評価結果

形態素と N 文字連鎖を並用した索引付けの評価結果を**表 4** に示す．2 文字連鎖の索引付けと形態素の基本形で索引付けを組み合わせて索引付けした結果は 0.462 であった．2 文字連鎖と形態素を並用した結果では双方の利点が活かされることを期待したが，今回は効果がみられなかった．

3.5 ストップワードの利用

3.5.1 品詞情報によるストップワードリストの作成

日本語の助詞（「は」,「が」など）はきわめて頻繁に使われる語であり，必要な文書を絞り込む能力が低い．このように高い頻度で検出される語は，索引語として適当ではない．こ

Vol.2009-NL-191 No.8
Vol.2009-SLP-76 No.8
2009/5/21

のような語を不要語（ストップワード）と呼ぶ．索引語の候補から不要語を除去することにより索引語の総数を減らすことができるため，検索システムの処理の効率化や高速化，高精度化を行うことができる．

どのような語を不要語と認定するかについては，さまざまな方法が考えられる[3]．自然言語の語は，大きく内容語と機能語の 2 つに分けることができる．内容語は，それ自体が意味を持った，ある特定の概念を表している語であり名詞や動詞がこれに相当する．機能語は，語と語の間の関係を表している語であり，日本語の場合には助詞や助動詞などが相当する．ここでは，品詞情報を用いたストップワードについて調査を行う．具体的には，助詞，助動詞を不要語としてストップワードリストを作成する方法と，名詞と動詞以外を不要語としてストップワードリストを作成する方法の 2 つを調べる．

3.5.2 DF 値によるストップワードリストの作成

品詞情報の利用は形態素解析システムの性能に依存する．また N 文字連鎖ではそもそも品詞の情報は使えない．したがって次に，品詞情報以外を用いてストップワードを設定することを考えた．はじめに DF 値（文書出現数）の高い形態素や N 文字連鎖を不要語としてストップワードリストを作成する方法を検討した．具体的には，文書集合の全文書数 2702 件の 1 割（270 件），2 割（540 件），3 割（810 件），…と閾値を設定し，閾値を超える DF 値を持つ形態素や N 文字連鎖をストップワードリストに登録した．

3.5.3 エントロピー値によるストップワードリストの作成

各索引語のエントロピー値を用いてストップワードリストを作成する方法も検討した．

索引語 w_i の文書 D_j での出現数 $tf_{i,j}$ と，文書集合全体での索引語 w_i の出現数 TF_i を用いると索引語 w_i のエントロピー H_i は式（7）で与えられる．

$$H_i = -\sum_{j=1}^{N} \frac{tf_{i,j}}{TF_i} \log \frac{tf_{i,j}}{TF_i} \quad (0 \leq H_i \leq \log N) \tag{7}$$

本研究では，式（7）を正規化した正規化エントロピー（式（8））を用いる．

$$\frac{H_i}{\log N} = -\frac{1}{\log N} \sum_{j=1}^{N} \frac{tf_{i,j}}{TF_i} \log \frac{tf_{i,j}}{TF_i} \quad (0 \leq \frac{H_i}{\log N} \leq 1) \tag{8}$$

索引語が各文書に等しく出現するほど正規化エントロピーは 1 に近い値をとり，少数の限られた文書でのみ出現する場合は 0 に近い値となる．本研究では，ある閾値を設定し，正規化エントロピー値がそれ以上の形態素や N 文字連鎖をストップワードとして検索を行った．

表 5 形態素単位の品詞情報に基づくストップワードを用いた索付けの評価

索引語の単位	11ptAP
形態素 基本形 助詞・助動詞をストップワード	0.462
形態素 基本形 名詞と動詞以外ストップワード	0.474

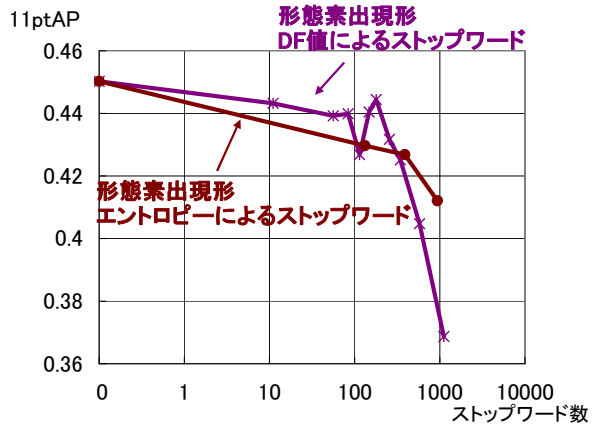

図 1 DF 値・エントロピー値に基づくストップワードの効果の比較

3.5.4 ストップワードを用いた索引付けの評価結果

品詞情報に基づくストップワードを用いた索引付けの評価結果を**表 5** に示す．ストップワードとして助詞・助動詞を設定し，これら以外の品詞の形態素基本形で索引付けを行った場合の 11ptAP は 0.462 であった．名詞と動詞以外の品詞の語をストップワードとした場合は 0.474 であり，本実験で行った索引付けで最も高い精度が得られた．このことはストップワードを利用することの有効性を示している．

次に品詞情報を用いないストップワードの設定の結果について述べる．はじめに DF 値とエントロピーに基づくストップワードの比較を行う．ここでは形態素出現形で実験を行った．**図 1** に結果を示す．ストップワードの効果はみられなかったものの DF 値に基づく手法の方がエントロピーに基づく手法より高い精度が得られることがわかった．次に，DF 値を用いたストップワードの設定についてさらに調査を行ったので，その結果について述べる．今回は，形態素基本形で索引付けの実験を行った．結果を**図 2** に示す．形態素を索引語の単位にしたときは，ここでもストップワードの効果はみられかった．2 文字連鎖単位で索引付

Vol.2009-NL-191 No.8
Vol.2009-SLP-76 No.8
2009/5/21

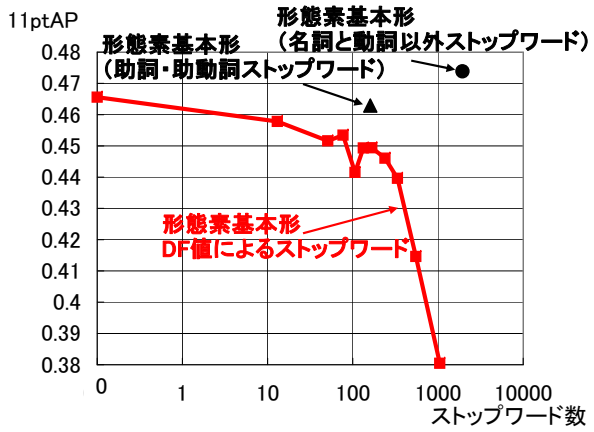

図 2 DF 値に基づくストップワードの効果

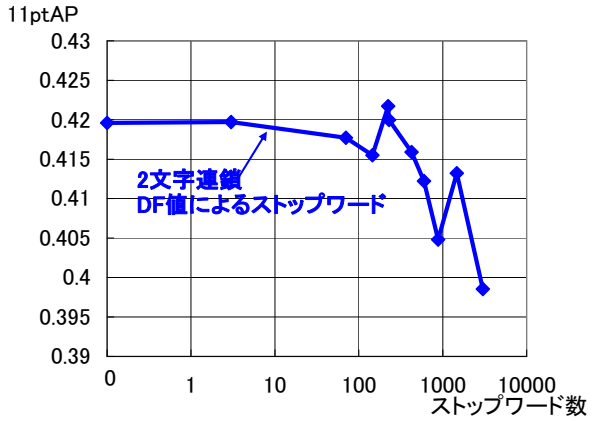

図 3 2 文字連鎖単位での DF 値に基づくストップワードの効果

けを行った場合の結果を**図 3** に示す．出現文書数が 7 割（1891 件）よりも多い索引語 223 語をストップワードとしたときに，ストップワードの効果がみられた．

4. 結　　論

音声ドキュメント検索のための索引語の検討を行った．形態素単位，文字連鎖単位での種々の索引付けを行い，検索質問に該当する講演全体を検索し評価した．名詞と動詞以外をストップワードにした形態素の基本形による索引付けで 11ptAP 値 0.474 が得られ，形態素の基本形を索引語の単位とすること，およびストップワードを用いることの有効性がわかった．

文字連鎖での索引付けでは品詞情報が使えないため，DF 値およびエントロピー値といった統計値に基づくストップワードの設定も検討した．2 文字連鎖単位の索引付けにおいて DF 値に基づくストップワードの効果がみられた．形態素単位の索引付けでは効果はみられなかった．今後も種々の統計値に基づくストップワードを研究していく予定である．

参　考　文　献

1) 岡本拓明，　仲野亘，小林隆志，　直井聡，横田治夫，岩野公司，古井貞熙：音声情報を統合したプレゼンテーションコンテンツ検索，電子情報学会論文誌 D Vol. J90-D No.2，pp.209–222 (2007).
2) 秋葉友良，相川清明，伊藤慶明，河原達也，南條浩輝，西崎博光，安田宣仁，山下洋一，伊藤克旦：音声ドキュメント検索テストコレクションの試作と基本性能評価，第 1 回音声ドキュメント処理ワークショップ講演論文集，pp.73–80 (2007).
3) 北　研二，津田和彦，獅子堀正幹：情報検索アルゴリズム，共立出版株式会社 (2002). ISBN4-320-12036-1.
4) 小作浩美，内山将夫，井佐原均，河野恭之，木戸出正継：WWW 検索における複数検索結果の結合処理とその評価，情報処理学会論文誌 Vol.44 No.SIG 8（TOD 18），pp. 78–91 (2003).
5) Tomoyosi Akiba, Kiyoaki Aikawa, Yoshiaki Itoh, Tatsuya Kawahara, Hiroaki Nanjo, Hiromitsu Nishizaki, Norihito Yasuda, Yoichi Yamashita, and Katunobu Itou: Construction of a test collection for spoken document retrieval from lecture audio data, *IPSJ-Journal, Vol.50. No.2*, pp.501–513 (2009).
6) 前川喜久雄：言語研究における自発音声，日本音響学会研究発表会講演論文集（春季），pp.19–22 (2001).
7) 汎用連想計算エンジン GETA：http://geta.ex.nii.ac.jp/.
8) 松本裕治，　北内啓，山下達雄，平野善隆，　松田寛，高岡一馬，浅原正幸：http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1-j.pdf (2000).