Vol.2010-NL-196 No.12
Vol.2010-SLP-81 No.12
2010/5/28

広域文書類似度と局所文書類似度を用いた講演音声ドキュメント検索

南條浩輝[†1]　弥永裕介[†1]　吉見毅彦[†1]

講演音声集合から探したい内容を表す 1 分程度の箇所を検索する方法について述べる．本論文では，講演には階層構造があること，すなわち講演の各部分には講演全体が扱うトピックのサブトピックがあることに着目し，講演単位などの大きなまとまり（広域文書）での粗い検索と，1 分程度の単位（局所文書）の検出したい単位での詳細な検索とを段階的に統合する方法を提案する．具体的には，クエリと広域文書との類似度を局所文書との類似度に統合して，局所文書を検索する方法を提案する．CSJ 音声ドキュメントコレクションの 15 発話単位の検索タスクにおいて，検索精度（11 点平均精度）の有意な向上（0.172→0.220）が確認できた．

Spoken Document Retrieval for Oral Presentations Integrating Global Document Similarities into Local Document Similarities

Hiroaki NANJO,[†1] Yusuke IYONAGA[†1]
and Takehiko YOSHIMI [†1]

A spoken document retrieval (SDR) method for oral presentations is addressed. We propose an integration method of global information and local information based on a topic hierarchy of presentations. Specifically, for detecting a part of oral presentations about 1 minute (local document), we integrate a similarity between a given query and a whole presentation (global document) into a similarity between the given query and a local document contained in the global document. For the 15-utterances-based pseudo-passage retrieval task of CSJ SDR, we confirmed a statistical improvement of information retrieval performance (11 point average precision) from 0.172 to 0.220.

1. はじめに

講演音声集合から探したい内容を表す箇所を見つける「講演音声ドキュメント検索」の研究を行う．これまで音声ドキュメント検索の主な研究対象は TREC SDR に代表されるようにニュースであった．ニュースは通常 1 分程度で自己完結的に作られているため，各ニュース音声を検索単位とするのが自然であり，ドキュメント拡張などを用いて音声認識誤りに対して頑健にインデキシングを行う方法や，クエリ拡張などが主な研究テーマであった．

これに対し，本研究ではある程度の長さを持つ講演や講義の音声を対象とした音声ドキュメント検索手法を研究する．講演や講義は通常短くても 10 分以上の長さを持っており，従来のようにこのような音声ドキュメントそのものを検索単位とするのは実用上問題がある．すなわち，適切な検索結果が得られても目的とする情報にダイレクトにアクセスできないという問題がある．例えばユーザが知りたい内容が講演や講義の一部である場合に，検索結果の講演や講義の中から該当個所を自ら探さなくてはならないという問題がある．そのため，講演や講義を話題のまとまり（サブトピック）ごとに分割し，それらを検索対象とすることが望ましい．ただし，このような講演や講義のトピック分割自体が難しい問題であるため，本研究では講演音声をあらかじめ 30 秒から 1 分程度の単位に区切って各区間を検索対象とすることで，欲しい情報の近くにダイレクトにアクセスできるようにすることを考える．

このように 1 分程度の短いドキュメントを検索単位とした場合，ドキュメントに含まれる単語の数が少ないことや，音声認識誤りが存在した場合にその影響が大きくなるため，一般的に高い検索精度を得るのが難しい．本研究では，講演にはそれ自体が扱う大きなトピックがあり，講演の各部分には講演内容をより詳細に表すサブトピックがいくつかあり，各サブトピック自体にもそれを詳細に表すより小さなサブトピックが存在することがあるという構造があることに着目し，講演単位などの大きなまとまり（広域文書）での粗い検索と，1 分程度の単位（局所文書）の検出したい単位での詳細な検索とを統合することを考える．具体的には，クエリと広域文書との類似度を局所文書との類似度に統合して，局所文書を検索する手法を提案する．

提案手法は，XML に代表されるような構造化文書の検索手法[1)2)] とみることもできる．ただし，音声ドキュメントではタイトルや段落区切りなどの構造が明示されないため[3)]，構

†1 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University

Vol.2010-NL-196 No.12
Vol.2010-SLP-81 No.12
2010/5/28

造化テキストに対する従来の検索手法は音声ドキュメント検索に利用できない．提案手法は，講演のような長い音声ドキュメントに対して，自動的に階層構造を持った局所文書および広域文書群を生成し，それらを構造化テキストのように用いる点において新規性を有する．

2. 講演音声ドキュメント検索

本研究では，講演音声を対象として検索を行う．講演を記録したもの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ことがある．スライドの文字を解析して，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．

2.1 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，クエリ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である．テストコレクションとは，文書集合，クエリ集合，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[4]を用いて研究を行った．これは,「日本語話し言葉コーパス」（以後，CSJ と略す）[5] の学会講演 987 件と模擬講演 1715 件の合計 2702 件の講演を検索対象とするものである．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．本研究は，この音声認識結果を使って行う．

クエリは自然言語文で記述された 39 件のテキストであり，各クエリに対する答えとしての適合情報が，どの講演のどの発話からどの発話までという単位で付与されている．なお適合度として適合（R）と部分適合（P）のラベルが存在する．本研究では適合（R）ラベルが付与された区間をクエリに対する正解として扱った．

2.2 評 価 尺 度

情報検索システムの検索性能の評価は，完全性と正確性の観点から再現率（recall）と適合率（precision）を用いるのが一般的である．本研究では，評価尺度として式（1）で示すこれらを組み合わせた評価尺度である補間 11 点平均精度（Interpolated 11-points Average Precision,"11ptAP"と記す）を用いる．これは各検索クエリ Q_k に対して 0.0 から 1.0 まで 0.1 刻みでの各再現率レベル x における補間精度 $IP_{Q_k}(x)$（式（3））を求め，それらの平均 $AP(Q_k)$（式（2））を全検索クエリで平均をとったものである．

$$11ptAP = \frac{1}{N}\sum_{k=1}^{N} AP(Q_k) \tag{1}$$

$$AP(Q_k) = \frac{1}{11}\sum_{i=0}^{10} IP_{Q_k}(\frac{i}{10}) \tag{2}$$

$$IP_{Q_k}(x) = \max_{x \le R_{Q_k}(T)} P_{Q_k}(T) \tag{3}$$

ここで $R_{Q_k}(T)$ と $P_{Q_k}(T)$ は，それぞれクエリ Q_k を用いて上位 T 番目まで検索したときの再現率と適合率である．

今回は，1 つのクエリに対して類似度が 0 でないものを全件検索し，検索結果全体での再現率よりも高い再現率レベル x の補間精度 $IP_{Q_k}(x)$ は 0 とした．

3. 検索システム

本研究ではベクトル空間モデル[6] に基づく文書検索システムを用いる．ベクトル空間モデルは，文書とクエリとをベクトルで表現し，ベクトル間の距離により検索を実現するモデルである．本研究では，ベクトル間の類似度に SMART[7] を用いる．すなわち，あるクエリ Q と文書 $D_i(1 \le i \le N)$ の類似度を，索引語を $t_k(1 \le k \le m)$ として，式（4）で与えるものである．

$$\mathrm{SMART}(Q, D_i) = \sum_{k=1}^{m} (q_{t_k} \cdot d_{i,t_k}) \tag{4}$$

ただし，

$$d_{i,t_k} = \begin{cases} \dfrac{1+\log(\mathrm{tf}_{i,t_k})}{1+\log(\mathrm{avtf})} & \text{if } t_k > 0 \\ 0 & \text{otherwise} \end{cases}$$

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

ここで，tf_{i,t_k} は D_i 中での t_k の出現数，avtf は D_i における単語の出現数の平均を表す．pivot は 1 文書中の異なり単語数の平均，utf_i は D_i 中の異なり単語数を表す．slope は補

Vol.2010-NL-196 No.12
Vol.2010-SLP-81 No.12
2010/5/28

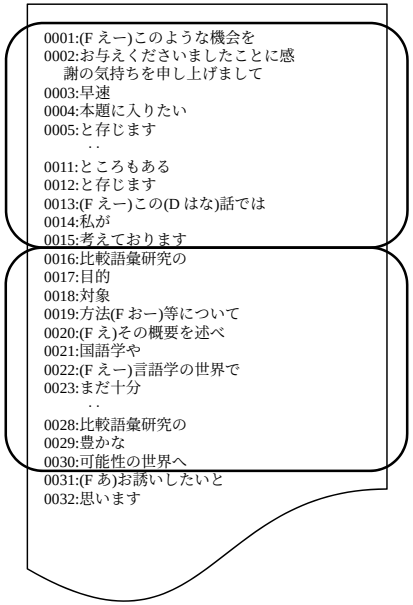

図 1 15 発話区間
Fig. 1 15-utterances-based-pseudo-passage unit

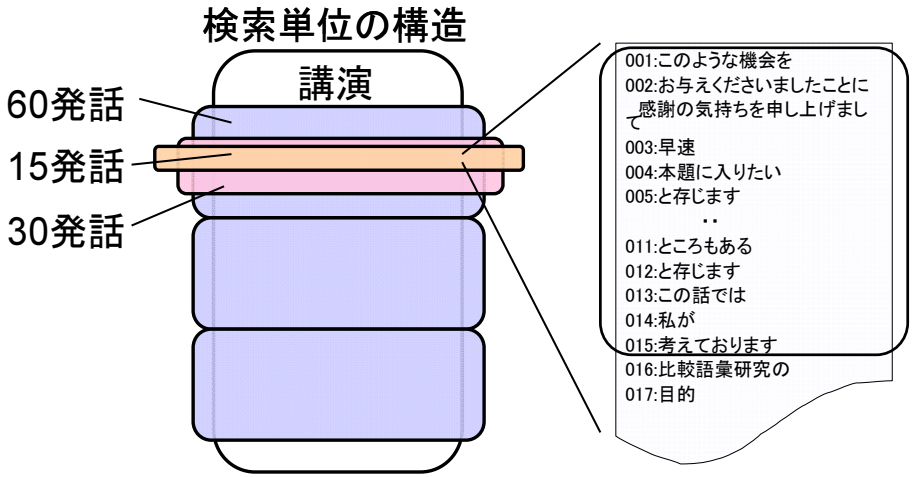

図 2 局所文書（15 発話単位）と広域文書（30 発話単位，60 発話単位，講演単位）の関係図
Fig. 2 Local document (15-utterances) and global document (30-utterances, 60-utterances, and a whole presentation)

表 1 CSJ 講演音声ドキュメント検索テストコレクションでの局所文書（15 発話単位）および広域文書（30 発話単位，60 発話単位，講演単位）の文書数
Table 1 Numbers of local document (15-utterances) and global documents (30-utterances, 60-utterances, and a whole presentation) on the spoken document retrieval test collection of CSJ

	15 発話単位	30 発話単位	60 発話単位	講演単位
文書数	60202	30762	16060	2702

間係数（0.2）である．qtf_{t_k} は，Q 中での t_k の出現数，avqtf は Q に含まれる単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．

本研究では，クエリ Q が与えられたとき，全ての文書 D_i について Q との類似度 SMART(Q, D_i) を算出し，類似度が 0 より大きいものを高い順に全件出力する．

4. 講演の構造を利用した講演音声ドキュメント検索

本研究では，講演には構造があり，探したい部分が講演のトピックの一部のサブトピックとなっている点に着目し，講演音声を検索する方法を提案する．具体的には，大きなトピック単位を広域文書，探し出したいサブトピック単位を局所文書と定義し，クエリと広域文書との類似度を局所文書との類似度に統合して，局所文書を検索する手法を提案する．局所文書として，1 分程度の長さを持ち，秋葉ら[4] の検索性能評価でも利用されていた 15 発話を採用する．具体的には図 1 に示すように講演の先頭から順に 15 発話ごとに区切り，各区間を 1 つの局所文書とする．その際の広域文書としては 15 発話を包含する 30 発話，60 発話，講演全体を採用する（図 2）．CSJ 講演音声ドキュメントテストコレクションでのこれらの統計量を表 1 に示す．

本論文では，広域文書と局所文書類似度の統合を行わず，局所文書をドキュメントとみなして検索する方法をベースラインの手法とし，広域文書類似度と局所文書類似度の統合手法（提案手法）の有効性を示す．また，局所文書がある程度の長さを持つ場合，具体的には局所文書を 30 発話単位とした場合および 60 発話単位とした場合に，広域文書情報を利用す

Vol.2010-NL-196 No.12
Vol.2010-SLP-81 No.12
2010/5/28

表 2 15 発話単位での検索性能
Table 2 Information retrieval performance for 15-utterances retrieval task

索引語	11 点平均精度
形態素の出現形	0.142
形態素の基本形 (名詞，動詞のみ)	0.172

る効果についても考察を行う．

4.1 ベースラインシステムによる検索結果

CSJ の 2702 講演を 15 発話ごとに区切り各区間を検索対象の文書として検索を行った．表 1 に示すとおり，文書数は 60202 である．索引語には先行研究[8)9)] を参考に形態素の出現形，形態素の基本形（名詞，動詞のみ）の 2 種類，検索システムには汎用連想計算エンジン GETA[10)] を用いた．

実験結果を表 2 に示す．索引語に形態素出現形を用いた場合の 11 点平均精度は 0.142 であった．索引語に形態素基本形（名詞，動詞のみ）を用いた場合に 11 点平均精度は 0.172 であった．この結果は，索引語として形態素基本形（名詞，動詞のみ）を用いることが有効であることを示しており，この結果は先行研究[8)9)] の傾向と一致する．以後の実験においては形態素基本形（名詞，動詞のみ）を索引語とする．

4.2 広域文書類似度と局所文書類似度の統合

次に，提案手法について述べる．具体的には，クエリと広域文書との類似度をクエリと局所文書との類似度に統合し，局所文書を検索する方法について述べる．この様子を図 3 に示す．

4.2.1 1 種類の広域文書類似度の統合

はじめに，クエリと 1 種類の広域文書の類似度を，クエリと局所文書の類似度に統合する方法を述べる．ここでは局所文書を 15 発話単位として説明を行う．

講演 D_i の l から $l+n$ までの発話区間を ${D_{i,l}}^{l+n}$ と表すことにすると，クエリ Q と局所文書（15 発話単位）との類似度は $\mathrm{SMART}(Q, {D_{i,k}}^{k+15})$ と表現できる．局所文書を包含する広域文書とクエリとの類似度を統合した後の，局所文書とクエリの類似度 $\mathrm{S}(Q, {D_{i,k}}^{k+15})$ は，Q と広域文書 ${D_{i,l}}^{l+n}$ の類似度 $\mathrm{SMART}(Q, {D_{i,l}}^{l+n})$ と，Q と ${D_{i,k}}^{k+15}$ との類似度 $\mathrm{SMART}(Q, {D_{i,k}}^{k+15})$ を対数線形補間したものと定義する（式 (5)）．なお，対数線形補間では，一方の文書類似度が高くても他方の文書類似度が低い場合は最終的な文書類似度が低くなるため，両類似度が高い文書のみが選択される．

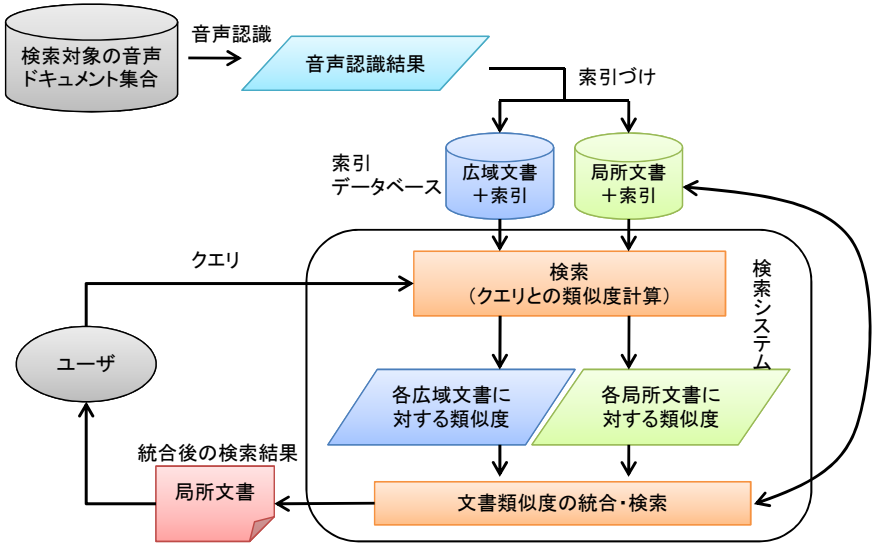

図 3 文書類似度の統合による音声ドキュメント検索
Fig. 3 Spoken document retrieval for oral presentations based on local and global documents

表 3 1 種類の広域文書類似度の統合
Table 3 Integration of one global document similarity into local document similarity

局所文書	広域文書	11 点平均精度
15 発話	なし（ベースライン）	0.172
	30 発話	0.203
	60 発話	0.196
	講演	0.211

$$\mathrm{S}(Q, {D_{i,k}}^{k+15}) = (1-\lambda)\log \mathrm{SMART}(Q, {D_{i,l}}^{l+n}) + \lambda \log \mathrm{SMART}(Q, {D_{i,k}}^{k+15}) \quad (5)$$

ここで，${D_{i,l}}^{l+n}$ は ${D_{i,k}}^{k+15}$ を包含しているため，$l \le k$，$k+15 \le l+n$ の条件を満たす．λ は各文書類似度に対する重み係数であり，0 から 1 までの値をとるものとする．

CSJ の講演音声ドキュメント検索タスクで用意された 39 件のクエリを用いて検索を行った．結果を表 3 に示す．ここでは，39 件のクエリを 3 つのセットにわけ，交差検定を行っ

Vol.2010-NL-196 No.12
Vol.2010-SLP-81 No.12
2010/5/28

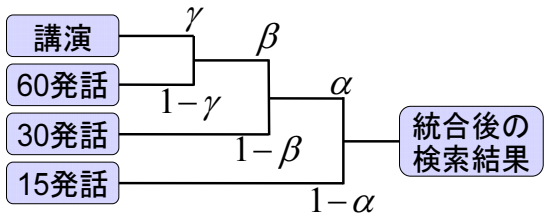

図 4 複数の広域文書類似度の階層的統合による講演音声ドキュメント検索
Fig. 4 Spoken document retrieval for oral presentations based on hierarchical incorporation of global document similarities

た．具体的には，26 件のクエリを用いて検索精度が最大となるように統合重み λ の推定を行い，残りの 13 件の検索実験にその推定重みを使用した．これを 3 つの分割セットそれぞれに対して行った．なお，講演単位を広域文書とした場合の統合重み λ は，0.42，0.48，0.46 であり，分割セット間での大きな違いは見られなかった．

広域文書の情報を用いることで，広域文書を用いない検索精度（0.172）よりも高い精度が得られた．広域文書の大きさ（長さ）による影響は確認できなかったものの，広域文書を講演単位としたときに，検索精度が最も高くなり，0.211 が得られた．

4.2.2 複数の広域文書類似度の統合

次に，複数の広域文書，具体的にはクエリと講演単位，60 発話単位，30 発話単位の 3 種類の広域文書の類似度を，15 発話単位の文書類似度に統合する方法について述べる．

本研究では，包含関係にある各広域文書を大きさの順に並べ，段階的に統合を行った．具体的には，式（6）に示すように，まず，講演単位の文書との類似度と 60 発話単位の文書との類似度を重み γ を用いて統合し，次に，その類似度と 30 発話単位の文書との類似度を重み β を用いて統合する．最後に，15 発話単位の文書との類似度を重み α を用いて統合し，最終的な結果とする．この様子は，図 4 にも示されている．

$$\begin{aligned}\mathrm{S}(Q, D_{i,k}^{\ k+l5}) = &\left(1-\alpha\right)\log \mathrm{SMART}(Q, D_{i,k}^{\ k+l5}) \\ &+\alpha\Big(\Big(1-\beta\Big)\log \mathrm{SMART}(Q, D_{i,l}^{\ l+30}) \\ &+\beta\Big((1-\gamma)\log \mathrm{SMART}(Q, D_{i,m}^{\ m+60}) + \gamma \log \mathrm{SMART}(Q, D_i)\Big)\Big) \quad (6)\end{aligned}$$

表 4 複数の広域文書類似度の階層的統合の効果
Table 4 Effect of hierarchical incorporation of global document similarities

局所文書	広域文書	11 点平均精度
15 発話	なし（ベースライン）	0.172
	講演+60 発話+30 発話	0.220

ただし，各広域文書はそれよりも小さい単位の文書を包含しているため，それぞれ $m \leq l \leq k$，$k+15 \leq l+30 \leq m+60$ を満たすものである．また，統合重み α，β，γ はそれぞれ 0 から 1 の間の値をとるものとする．

結果を表 4 に示す．広域文書を一つだけ使った場合（表 3）よりも，高い検索精度 0.220 が得られた．なお統合重みは 3 つのセットそれぞれで，$\gamma = 0.7, 0.76, 0.67$，$\beta = 0.35, 0.28, 0.4$，$\alpha = 0.1, 0.03, 0.16$ であり，各分割セット間での大きな違いは見られなかった．

ベースラインの結果に比べてこの結果が有意に高いかについて，対応のある 2 群の差の検定（T 検定）を行ったところ，有意水準 1%で有意差が認められた．このことは，クエリと広域文書との類似度を局所文書との類似度に統合する提案手法は有効であることを示している．

4.3 局所文書の長さと広域文書利用の効果

次に，局所文書を長い場合の広域文書を利用した検索結果について述べる．

はじめに，局所文書を 30 発話単位とした場合について述べる．30 発話単位を一つの文書とみなした場合，CSJ 講演音声ドキュメント検索テストコレクションでは 30762 文書の検索タスクとなる（表 1）．名詞および動詞の基本形の語を索引語として，ベースラインシステムで検索を行ったところ，検索精度（11 点平均精度）は 0.235 であった．

広域文書としてこの 30 発話単位文書を包含する 60 発話単位文書と講演単位文書を用いて，クエリと広域文書の類似度を局所文書との類似度に統合して検索を行った．結果を表 5 に示す．広域文書として講演単位文書を利用したときに，検索精度の大きな向上（0.235 → 0.261）が確認できた．60 発話単位文書を広域文書としても大きな改善は見られなかった．また，60 発話単位と講演単位の文書とを同時に利用する効果もみられなかった．これらの理由として，30 発話単位の音声ドキュメントは 2 分程度とやや長く，その発話前後の情報を十分に含んでおり，60 発話単位の文書を用いる効果が低かったことが考えられる．講演単位文書を用いた場合では，近傍以外の部分での情報をうまく取り入れられたと考えられる．なお，講演単位文書と 60 発話単位文書を用いたときの検索精度の改善（0.235 → 0.262）につい

Vol.2010-NL-196 No.12
Vol.2010-SLP-81 No.12
2010/5/28

表 5 局所文書を 30 発話とした場合の検索性能
Table 5 Information retrieval performance for 30-utterances retrieval task

局所文書	広域文書	11 点平均精度
30 発話	なし（ベースライン）	0.235
	60 発話	0.240
	講演	0.261
	60 発話+講演	0.262

表 6 局所文書を 60 発話とした場合の検索性能
Table 6 Information retrieval performance for 60-utterances retrieval task

局所文書	広域文書	11 点平均精度
60 発話	なし（ベースライン）	0.281
	講演	0.296

て，対応のある 2 群の差の検定（T 検定）を行ったところ，有意水準 5%で有意差が認められた．

次に，広域文書を 60 発話単位文書とした場合の結果について述べる．60 発話単位を一つの文書とみなした場合，テストコレクションでは 16060 文書の検索タスクとなる（表 1）．名詞および動詞の基本形の語を索引語として，ベースラインシステムで検索を行ったところ，検索精度（11 点平均精度）は 0.281 であった．広域文書として，この 60 発話単位文書を包含する講演単位文書を利用した．結果を表 6 に示す．検索性能の向上は見られたものの，局所文書が短い場合（15 発話や 30 発話）に比べた場合，向上する割合が小さくなっていることがわかる．局所文書が長くなるにしたがって，検索に必要な索引語が多く得られていく結果と考えられる．

これらのことより，局所文書を長くするにしたがって広域文書類似度を統合する効果は次第に小さくなるものの，広域文書の利用は局所文書の検索結果に悪影響は及ぼさず，効果があることがわかった．講演のような，長くかつ話題に階層構造をもつ音声ドキュメントに対して，その一部の短い区間を検索するタスクにおいて，提案手法が有効であることを示した．

5. おわりに

講演音声ドキュメント検索において，短い発話区間（局所文書）と，局所文書を包含する広域文書を用いて局所文書を検索する手法を提案し，CSJ を対象とした SDR テストコレクションで評価実験を行った．局所文書を 1 分程度の 15 発話単位の文書とした場合は，文書類似度の統合を行わない従来の検索方法での精度が 0.172 であったのに対し，提案手法を用いて文書類似度の統合を行うことで有意な精度の改善が得られ，検索精度 0.211 が得られた．また，局所文書の長さを 30 発話，60 発話と長くしたときにも，局所発話が短い場合（15 発話）に比べると広域文書類似度を統合する効果は小さくなるものの，局所文書の検索結果に悪影響は及ぼさず，効果があることがわかった．クエリと広域文書との類似度を局所文書との類似度に統合する提案手法の有効性を示した．

参 考 文 献

1) 江口浩二：情報検索のための確率的言語モデルに関する動向と課題，電子情報通信学会論文誌， Vol.J93-D, No.3, pp.157–169 (2010).
2) Ogilvie, P. and Callan, J.: Combining document representations for known-item search, *Annual ACM Conference on Research and Development in Information Retrieval*, pp.143–150 (2003).
3) Kawahara, T., Hasegawa, M., Shitaoka, K., Kitade, T. and Nanjo, H.: Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers, *IEEE Trans. Speech & Audio Process*, Vol.12, No.4, pp.409–419 (2004).
4) Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y. and Itou, K.: Construction of a test collection for spoken document retrieval from lecture audio data, *IPSJ-Journal*, Vol.50, No.2, pp.501–513 (2009).
5) 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).
6) 北 研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版株式会社，ISBN4-320-12036-1 (2002).
7) 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).
8) 重安幸治，南條浩輝，吉見毅彦：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，情報処理学会研究報告，SLP-76-8 (2009).
9) Shigeyasu, K., Nanjo, H. and Yoshimi, T.: A Study of Indexing Units for Japanese Spoken Document Retrieval, *10th Western Pacific Acoustics Conference (WESPAC X)* (2009).
10) 汎用連想計算エンジン GETA：`http://geta.ex.nii.ac.jp`.