音声クエリによる講演音声ドキュメント検索の基礎的評価

七里　崇[†1]　重安幸治[†1]
南條浩輝[†1]　吉見毅彦[†1]

音声ドキュメント検索ワーキンググループが作成中の CSJ 音声ドキュメント検索テストコレクションでは 39 件のクエリが用意されている．我々はこの 39 件のクエリの音声を 14 名分収録した．本発表では，はじめにこのクエリ音声を用いたテストコレクションの検索性能について述べ，次に，音声検索を指向したクエリ音声の認識方法について述べる．

A Study of CSJ Spoken Document Retrieval by Speech Queries

Takashi SHICHIRI,[†1] Koji SHIGEYASU,[†1]
Hiroaki NANJO[†1] and Takehiko YOSHIMI [†1]

CSJ spoken document retrieval using spoken queries is addressed. In the test collection of CSJ spoken document retrieval, query set which consists of 39 text queries is provided. we asked 14 people to read these queries and collected speech data of the query set. In this paper, first, we describe a baseline result of spoken document retrieval by spoken queries, and then, describe an ASR strategy of query speech oriented for speech-based information retrieval.

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

1. はじめに

現在，日本語の音声ドキュメント検索研究用プラットフォームとして，音声ドキュメント検索テストコレクションが構築されている[1)2)3)]．これは，検索対象の講演音声集合と 39 件の自然言語文で書かれた検索要求（クエリ），およびそれらに対する正解ラベルから構成されており，検索対象を音声とした情報検索の研究用のものである．WEB を介した動画検索などの応用が想定されており，テキスト入力型情報検索のタスク設定となっている．

しかし，近年普及しつつあるタッチパネル式の小型の携帯情報端末においては，検索用クエリのテキスト入力の負担は比較的大きく，音声入力が注目されている．したがって，音声入力型の音声ドキュメント検索に関しても技術を高める必要がある．このような研究を推進するためには，そのための共通データベースが重要である．

本研究では，CSJ の音声ドキュメント検索テストコレクションにおいて用意されている 39 件のクエリに対して，その音声データ（音声クエリ）を収集し，それを用いて検索性能を調べたのでその報告を行う．

2. 音声ドキュメント検索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である．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ワーキンググループによって作成された音声ドキュメント検索テストコレクション[1)] を用いて研究を行った．以下にこの詳細について述べる．

2.1 文書集合

検索対象の文書集合は，表 1 に示す日本語話し言葉コーパス（CSJ）[4)] の学会講演 987 件と模擬講演 1715 件の合計 2702 件である．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．

2.2 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．このテストコレクションの適合情報として，その適合度に応じて適合（R; Relevant），部分適合（P; Partially Relevant）の 2 種類が用意されている．なおこの適合情報は，講演の一部分（発話区間）に付与されている．

2

表 1　検索対象の CSJ の講演音声
Table 1　CSJ lecture speech

音声の種類	話者数	講演数	合計時間
学会講演	819	987	274.4
模擬講演	594	1715	329.9
合計	1413	2702	604.3

2.3　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[5)]．

- 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．

3.　クエリ読上げ音声の収録とその音声認識

39 件のクエリを男性 10 名，女性 4 名の合計 14 名に読み上げてもらった．言い間違いや言いよどみは含まれていない．

次にこのクエリ音声の音声認識について述べる．使用した言語モデルは CSRC 最終年度版に含まれる毎日新聞から学習した語彙サイズ 60K のモデルと WEB ページから学習した語彙サイズ 60K のモデル，および CSJ の講演から学習した語彙サイズ 20K のモデルの 3 種類である．それぞれの言語モデルの語彙サイズ，テストセット（クエリ 39 件）に対するカバレージ，パープレキシティを表 2 に示す．CSJ で学習した言語モデルが比較的カバレージが高いものの，未知語が多いことがわかる．語彙サイズが異なるためパープレキシティの比較は難しいが，どの言語モデルでも高めの値であることがわかる．毎日新聞言語モデルと WEB 言語モデルの比較では，WEB 言語モデルのほうがテストセットパープレキシティが小さく，39 件のクエリの認識には適していると考えられる．

表 2　クエリ音声認識用言語モデルの一覧
Table 2　Language models for ASR of queries

言語モデル学習データ	語彙サイズ	テストセットカバレージ	テストセットパープレキシティ
CSJ	20K	83.2%	132.6
WEB	60K	81.2%	160.6
毎日新聞	60K	80.1%	245.6

表 3　クエリ音声認識結果（WER（%））
Table 3　ASR results of spoken queries (WER (%))

	言語モデル学習データ		
話者 ID	CSJ	WEB	毎日新聞
F001	36.27	32.97	39.57
F002	19.39	17.00	23.24
F003	27.52	24.82	33.39
F004	34.35	30.63	38.38
M001	22.08	16.58	20.29
M002	19.39	20.65	26.17
M003	25.18	23.41	29.73
M004	14.93	11.33	18.91
M005	24.28	15.51	23.64
M006	20.47	16.45	22.36
M007	33.21	30.00	37.86
M008	32.25	29.22	38.99
M009	43.17	45.21	51.62
M010	48.65	44.36	49.01
AVG.	28.65	25.61	32.39

これらの言語モデルと読上げ音声（JNAS）で学習した状態数 2000 の triphone モデルを用いてクエリ音声を認識した．音声認識エンジンには Julius rev.4.1 を使用した．なお認識が実時間で行えるよう，ビーム幅は小さく設定した．39 件のクエリの認識率を表 3 に示す．WEB 言語モデルを用いることで，最も高い精度が得られた．

4. 音声クエリによる音声ドキュメント検索

このクエリの音声認識結果を用いて音声ドキュメント検索を行った結果について述べる．ここでは単純な bag-of-words 型のベクトル空間モデルに基づく検索を行う．

4.1 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 $D_i(1 \leq i \leq N)$ の類似度を，索引語を $t_k(1 \leq k \leq m)$ として，式（1）で与えるものである．

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

ただし，

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

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

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．

なお，索引語の単位は形態素に基づいている．形態素の出現形を索引語とし，ストップワード処理[8][9] は行っていない．

4.2 評価尺度

検索性能の評価尺度には，式（2）で示す補間 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 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 $IP_Q(x)$ を求め，それらの平均 $AP(Q)$ を全検索クエリで平均をとったものである．今回は，1 つのクエリに対して上位 1000 件を検索して出力し，検索結果 1000 件での再現率よりも高い再現率レベル x の補間精度 $IP_Q(x)$ は 0 とした．

$$\mathrm{11pt}AP = \frac{1}{N} \sum_{k=1}^{N} AP(Q) \tag{2}$$

$$AP(Q) = \frac{1}{11} \sum_{i=0}^{10} IP_Q(\frac{i}{10})$$

$$IP_Q(x) = \max_{x \leq R_{qi}} P_{qi}$$

ここで R_{qi} と P_{qi} は，それぞれクエリ Q に対する検索結果の上位 i 番目までの検索結果の再現率と適合率である．

4.3 検索結果

CSJ の講演書き起こしに対して検索を行ったときの検索性能を表 4 に示す．表 5 には，CSJ の音声認識結果（1-best）を対象に検索を行ったときの検索性能を示す．

毎日新聞記事で学習した言語モデルを用いてクエリ音声を認識した場合は認識率も低く，検索性能も最も低かった．CSJ 言語モデルを用いてクエリ音声の認識を行った場合と，WEB 言語モデルを用いてクエリ音声の認識を行った場合とを比較すると，音声認識に関しては WEB 言語モデルを用いた場合が精度が高いものの，検索精度は低かった．これについて調べたところ，音声認識における誤り単語数が WEB 言語モデルを用いて認識した場合の方が少なかった（CSJ：567 単語，WEB：519 単語）のに対して，検索にとって重要と考えられる名詞および動詞の誤り単語数は CSJ 言語モデルを用いた場合の方が少ない結果（CSJ：357 単語，WEB：373 単語）となった．WEB 言語モデルを用いた場合は，検索クエリ中の検索にとって重要な単語の認識精度が低かったと考えられる．

表 4　CSJ の書き起こしを検索対象としたときの平均検索性能
Table 4　Information Retrieval Performance for correctly written CSJ documents

正解判定	クエリ	クエリの音声認識に用いた言語モデル	音声認識精度 WER (%)	情報検索性能（11 点平均精度）
R	書き起こし	-	0	0.485
	音声	CSJ	28.7	0.352
		WEB	25.6	0.281
		毎日新聞	32.4	0.219
R+P	書き起こし	-	0	0.509
	音声	CSJ	28.7	0.359
		WEB	25.6	0.287
		毎日新聞	32.4	0.225

表 5　CSJ の音声認識結果を検索対象としたときの平均検索性能
Table 5　Information Retrieval Performance for ASR results of CSJ spoken documents

正解判定	クエリ	クエリの音声認識に用いた言語モデル	音声認識精度 WER (%)	情報検索性能（11 点平均精度）
R	書き起こし	-	0	0.455
	音声	CSJ	28.7	0.330
		毎日新聞	32.4	0.213
		WEB	25.6	0.269
R+P	書き起こし	-	0	0.470
	音声	CSJ	28.7	0.334
		毎日新聞	32.4	0.217
		WEB	25.6	0.274

クエリごとの情報検索性能を図 1 に示す．各クエリに対して 14 種類の音声認識結果で検索した結果を示している．なお，検索対象は講演の音声認識結果，クエリ音声の認識に使用した言語モデルは CSJ で学習したもの，正解判定は R+P である．

話者ごとの 39 クエリでの平均音声認識率と情報検索性能（11 点平均精度）を図 2 に示す．WER が高くなるにしたがって検索精度も低下しているが，必ずしも WER の増加と検索精度の低下が一致しないことがわかる．これは，たとえ 1 単語の音声認識誤りであってもそれが重要な単語であれば，検索にとって影響が大きいためと考えられる．

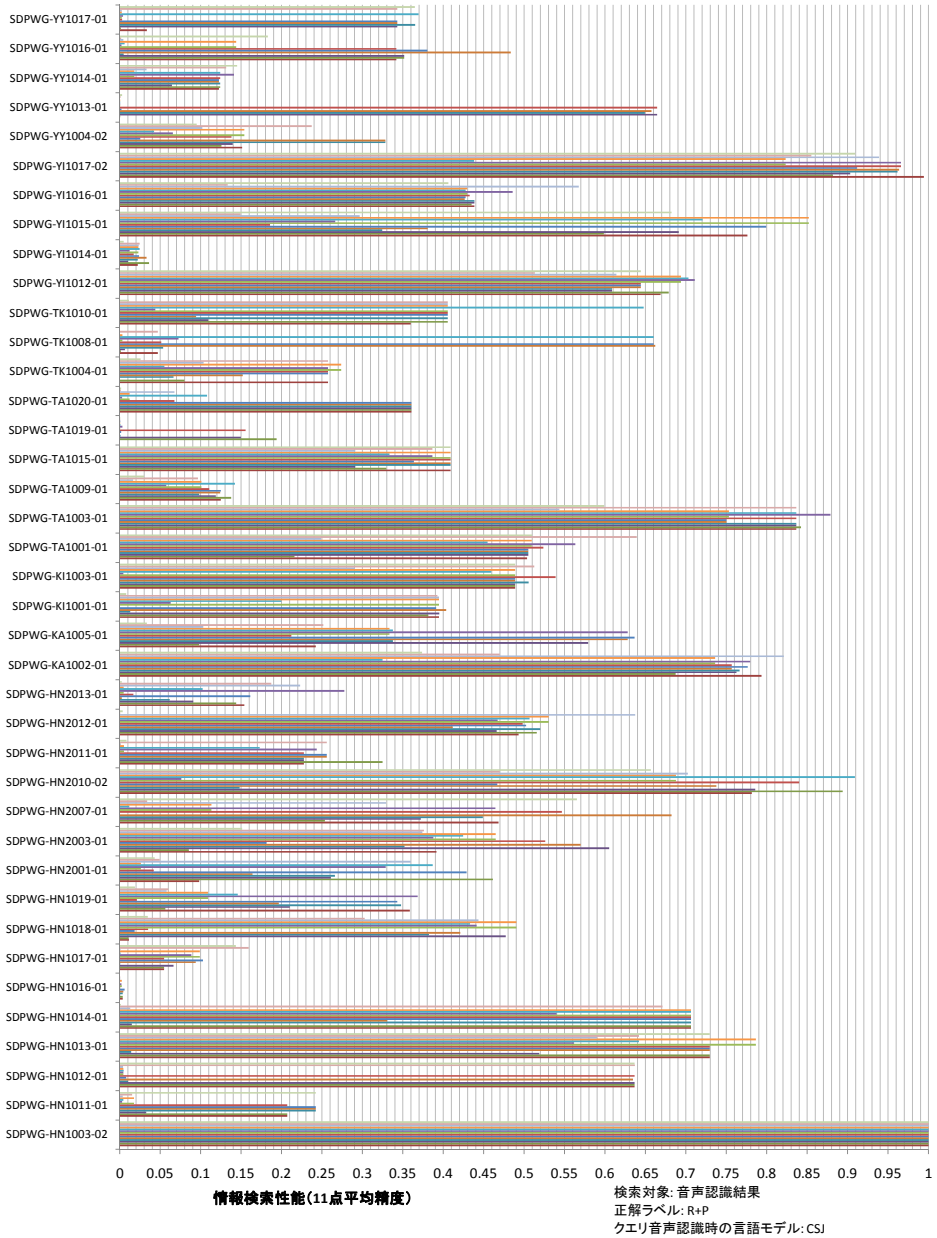

図 1　クエリごとの情報検索性能
Fig. 1　Information Retrieval Performance for each query

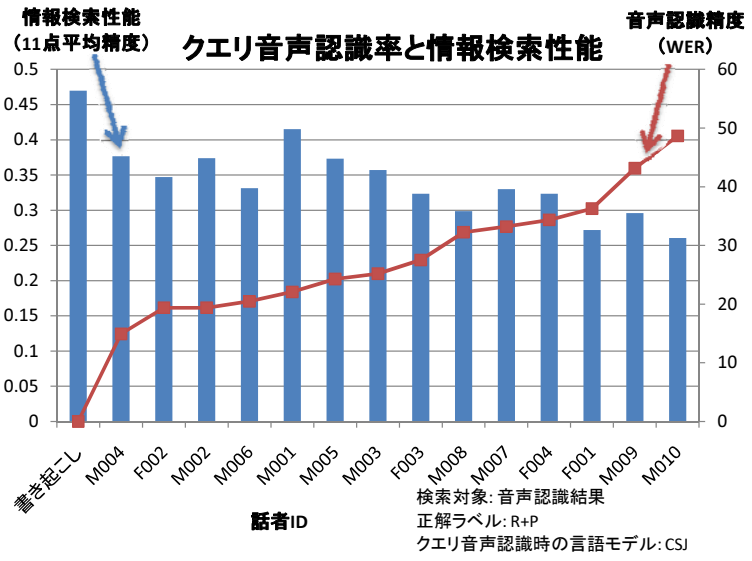

図 2　話者ごとのクエリの音声認識精度と情報検索性能
Fig. 2　ASR Accuracy and Information Retrieval Performance

4.4　情報検索に適したクエリ音声の認識手法

我々はこれまでに情報検索の音声入力のための音声認識手法の研究を行ってきている[10)11)]．これは，単語の重要度を考慮した音声認識の評価尺度として式（3）で定義する重み付き単語誤り率（Weighted Word Error Rate: WWER）を導入し，音声認識において WWER が小さくなるようにデコーディングする認識手法である．

$$\mathrm{WWER} = \frac{V_I + V_D + V_S}{V_N} * 100 \qquad (3)$$
$$V_N = \Sigma_{w_i}\ v_{w_i}, \quad V_I = \Sigma_{\hat{w}_i \in I}\ v_{\hat{w}_i}$$
$$V_D = \Sigma_{w_i \in D}\ v_{w_i}, \quad V_S = \Sigma_{seg_j \in S}\ v_{seg_j}$$
$$v_{seg_j} = \max(\Sigma_{\hat{w}_i \in seg_j} v_{\hat{w}_i}, \Sigma_{w_i \in seg_j} v_{w_i})$$

ただし，v_{w_i}，$v_{\hat{w}_i}$ はそれぞれ正解文と音声認識結果における単語の重みである．v_{seg_j} は置換誤り区間 seg_j の重みである．v_{seg_j} は，当該区間 seg_j に含まれる正解系列の単語の重みの合計値と認識結果の単語の重みの合計値のうち，大きい方とする．

ペット の 犬 の 名前 の リストアップ
⇒ペット:11.0 の:1.0 犬:11.0 の:1.0 名前:3.26 の:1.0 リストアップ:3.02

オークション における 自動 入札 戦略 を 知り たい
⇒オークション:11.0 における:4.14 自動:11.0 入札:1.0 戦略:1.0 を:1.0 知り:8.59 たい:1.0

図 3　単語重要度の推定結果例
Fig. 3　Sample of estimated word significance

表 6　音声認識の評価尺度と情報検索精度の相関
Table 6　Relationship between ASR evaluation measures and IR performance

	情報検索性能（11 点平均精度）との相関
WER	-0.386
WWER	-0.414

この認識のためには，各単語に対して適切にその重要度を付与することが必要である．実際に本検索タスクにおいて重要単語とその重要度の自動推定を行った[12)]．推定された単語重要度の例を図 3 に示す．検索にとって重要と考えられる単語「オークション」や「ペット」「犬」「名前」に対して高い重要度が推定できていることがわかる．

この重要度を用いて WWER を算出し，WWER と情報検索精度の関係を調べた．結果を表 6 に示す．WWER のほうが WER よりも情報検索精度との相関が大きいことがわかる．このことは，音声認識において WWER が小さくなるようにデコーディングを行うことで，情報検索精度の向上がより期待できることを示している．なおこのような音声認識は，ベイズリスク最小化の枠組みに基づいて行えばよい[13)14)]．

5. おわりに

CSJ の音声ドキュメント検索テストコレクションにおける 39 件のクエリに対して 14 名の音声クエリを用意し，音声による音声ドキュメント検索の基礎的評価を行った．このクエリ音声は研究用に公開する予定である．また，情報検索にとって重要な単語をきちんと認識することが重要であることもわかった．

参 考 文 献

1) 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).
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) 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，胡　新輝，中川聖一，伊藤克亘：SLP 音声ドキュメント処理ワーキンググループ活動報告，電子情報通信学会技術研究報告 SP2008-98，pp.115–120 (2008). NLC2008-43.
4) 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).
5) 伊藤克亘，相川清明，秋葉友良，伊藤慶明，河原達也，南條浩輝，西崎博光，安田宣仁，山下洋一：音声ドキュメント検索評価のためのテストコレクションの試作，情報処理学会研究報告，SLP-64-25 (2006).
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) 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).
9) 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，Vol.2009-SLP-76 No.8 (2009).
10) 南條浩輝，七里　崇，吉見毅彦：情報検索システムの音声インターフェースのための単語重要度の自動推定とそれに基づく音声認識，日本音響学会研究発表会講演論文集，1-10-14 (2008).
11) 七里　崇，南條浩輝，吉見毅彦：情報検索を指向した音声認識のための単語重要度の自動推定，日本音響学会研究発表会講演論文集，2-3-1 (2007).
12) 七里　崇，南條浩輝，吉見毅彦：音声入力型情報検索における単語重要度推定のための統計的機械翻訳を用いた音声認識シミュレート，日本音響学会研究発表会講演論文集，3-6-4 (2010).
13) 南條浩輝，河原達也，七里　崇：音声理解を指向したベイズリスク最小化枠組みに基づく音声認識，電子情報通信学会論文誌，Vol.J91-D, No.5, pp.1314-1324 (2008).
14) V.Goel and WJ.Byrne: Task dependent loss functions in speech recognition: Application to named entity extraction, *ESCA ETRW Workshop on Accessing Information from Spoken Audio*, pp.49–53 (1999).