音声ドキュメント検索用テストコレクションにおける音節インデックスの効果

前沢慎吾† 杉本樹世貴‡ 西崎博光†† 関口芳廣††
† 山梨大学・大学院医学工学総合教育部 ‡ 工学部 †† 大学院医学工学総合研究部

1 はじめに

音声ドキュメント検索を簡便に行うためには, 音声認識システムから得た認識結果をテキスト検索と同様に処理するのが良い. しかしそのような場合, 音声認識誤りや未知語が問題となり検索精度の低下が起こる.

情報処理学会音声言語処理研究会のワークショップである「音声ドキュメント処理ワークグループ」が『日本語話し言葉コーパス』[1](以下 CSJ) を用いて音声ドキュメント検索用のクエリと適合文書のテストセットを構築した [2]. この文献 [2] では, テストコレクションを利用し, 音声認識結果を用いた音声ドキュメント検索の基本的な検索精度が述べられている.

また, 秋葉らはテキスト翻訳の技術を利用した語彙拡張を音声ドキュメント検索に応用し, テストコレクションでの有効性を示した [3]. さらに, 胡らは音声認識結果のコンフュージョンネットワークを利用することで検索精度が改善できることを示した [4].

本研究は, このテストコレクションにおいて音節インデックスを利用することで検索精度が改善できるかどうかの検証を目的とする. 過去に,TREC をはじめとする大規模な情報検索テストコレクションを用いた音声ドキュメント検索の研究が行われてきた. それらの研究の中で, 音節等のサブワード単位のインデックスが未知語や音声認識誤りに効果があることが示された.

本稿では日本語のテストコレクションにおけるサブワード (音節) 単位の有効性を調査する. さらに, 音声ドキュメントを認識するときの認識精度, クエリに対する未知語率の違いによる検索精度の変化と, 音節インデックスがどの程度有効なのか, あるいは有効でないのかを明らかにする.

2 検索テストコレクション

検索テストコレクションには,CSJ の 2745 講演を対象にした 39 個の検索クエリとそれに対する適合文書のセット, および講演の音声認識結果が含まれている.

本研究では, 音声認識精度の違いによる検索精度の変化を調査するために, テストコレクションに含まれるものとは別に 3 種類の言語モデルを使った認識精度別の音声認識結果を用意した. 言語モデル (認識辞書) は CSJ 学会講演 970 講演から学習し, クエリに対する未知語を増やすためにクエリ中に出現する単語のうち数単語を言語モデルから削除した. 音声認識システムには Julius[5] を使用した. 表 1 に各言語モデルの語彙サイズと認識率を示す. 表 1 における未知語率とは, 検索クエリに対する未知語率 (検索クエリに含まれる単語のうちの何%が辞書に登録されていないのか) である.

表 1: 言語モデルの語彙サイズと認識率

	語彙数	未知語率	Corr.	Acc.
5k	5000 語	11.0%	73.7%	66.6%
10k	10000 語	8.8%	75.6%	69.7%
17k	17000 語	7.0%	76.9%	71.6%

語彙数が小さいものほど未知語率, 認識精度が低くなった. この認識結果を利用して検索実験を行った.

3 検索システム

検索システムには汎用連想計算エンジン GETA[6] を使用した. GETA は, 検索対象のインデックスを作成し, それを利用して検索を行う.

インデックスの索引語重みには登場頻度を用いた. 文書と検索クエリの類似度である検索スコアの計算には GETA にあらかじめ用意されている尺度の中から SMART を選択した.

検索は講演単位 (1 講演 1 文書) で行う.

4 検索実験

検索システムによって得られた検索結果のうち, 上位 1000 件の補間 11 点平均精度で検索精度を評価した.

インデックスの索引語単位には以下のものを使用した.

- wd(n):単語単位 (認識結果の n-best までを使用)
- w2g(n):単語単位と文字 2-gram 単位の併用 (認識結果の n-best までを使用)
- syl(n):音節 2-gram,3-gram 単位 (認識結果の n-best までを使用)

今回は $n = 1$ と $n = 10$ について実験を行った. さらに,w2g(n) と syl(n) の検索結果のスコアを次に示す線形補間によって結合したときの検索精度 (w2g+syl) についても調査した.

$$w2g + syl = \alpha \times w2g(n) + (1 - \alpha) \times syl(n)$$
$$\alpha = \{0.1, 0.2, \ldots, 0.9\}$$

Effect of syllable-based indexing for the test collection of spoken document retrieval
† Shingo MAEZAWA, University of Yamanashi
‡ Kiyotaka SUGIMOTO
†† Hiromitsu NISHIZAKI,Yoshihiro SEKIGUCHI

表 2: 検索精度 (全てのクエリ)

言語モデルのサイズ	wd(1)	w2g(1)	syl(1)	w2g+syl(1)	w2g(10)	w2g+syl(10)
5k	0.3028	0.3402	0.3284	0.3466	0.3218	0.3391
10k	0.3136	0.3549	0.3376	0.3657	0.3396	0.3468
17k	0.3257	0.3611	0.3376	0.3669	0.3572	0.3572

表 3: 検索精度 (未知語クエリのみ)

言語モデルのサイズ	wd(1)	w2g(1)	syl(1)	w2g+syl(1)	w2g(10)	w2g+syl(10)
5k	0.1620	0.1699	0.2278	0.2291	0.1575	0.2062
10k	0.1328	0.1462	0.1994	0.2069	0.1263	0.1656
17k	0.1640	0.1660	0.2173	0.2272	0.1729	0.1999

表 4: 検索精度 (未知語クエリのみ音節利用)

サイズ	wd(1)	w2g(1)	w2g+syl(1)	w2g+syl(10)
5k	0.3348	0.3684	0.3684	0.3456
10k	0.3392	0.3753	0.3782	0.3547
17k	0.3421	0.3769	0.3799	0.3655

表 2 に各インデックスに対する検索精度を示す. 表 2 を見ると, 単語のみ利用するよりも, 単語と文字 2-gram を併用した方が検索精度が良い. また,w2g+syl* も w2g に比べて精度の改善が見られた. 検索精度は, 音声認識精度が下がるほど低下している.

次に, 未知語を含むクエリについてのみ評価を行った結果を表 3 に示す. wd,w2g に比べて syl の精度が高く, 音節の利用により大幅な改善が得られた. 未知語を含むクエリに対しては音節インデックスの効果が高いことが分かった.

未知語を含むクエリのみ音節インデックスを利用するようにした場合の検索精度を 表 4 に示す. 表 4 において,wd は未知語を含むクエリの時のみ音節インデックスを用い, 既知語しか含まないクエリの場合は単語インデックスを用いた場合の結果である.w2g は単語インデックスの代わりに単語と文字 2-gram インデックスを用いた結果である. また w2g+syl は, 未知語を含むクエリの時のみ音節インデックスを用いて線形結合し, 既知語しか含まないクエリの場合は単語と文字 2-gram を用いた結果である.

この結果では w2g+syl(1) が最も精度が高くなった. 表 2 の結果と比べても精度が大幅に改善されており, 改善の幅は認識精度が低いものほど大きくなった. 実験の結果から, 未知語を含むクエリに対して音節インデックスを利用することで検索精度を改善できることがわかった.

* 最適な α 使用時の結果を掲載

5 おわりに

本稿では, 音声ドキュメント検索のテストセットを用いて, 音声認識精度が検索精度に与える影響や音節をサブワードとして用いたときのの有効性を実験的に示した. 実験の結果, 単語 (+文字 2-gram) インデックスのみを利用した場合は音声認識精度が下がるにつれて検索精度も低下した. しかし, 音節 3-gram を組み合わせることで, 音声認識精度の影響を受けにくく安定した検索精度を得ることができた. このことから, 音声ドキュメント検索において音節が有効であることが分かった.

今後はコンフュージョンネットワーク等を用いて音節インデックスと組み合わせるなど, 音節インデックスを有効に利用する方法を検討する予定である.

参考文献

[1] 国立国語研究所:"日本語話し言葉コーパス",http://www.kokken.go.jp/katsudo/seikaslashcorpus/

[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, 安田宜仁, 山下洋一, 伊藤克亘:"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検索性能評価", 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, 2007.2

[3] 秋葉友良, 横田悠右:"認識候補から正解テキストへの翻訳モデルに基く講演音声ドキュメントのアドホック検索", 第 2 回音声ドキュメントワークショップ講演論文集,pp.79-84,2008.2

[4] 胡新輝, 呉友政, 柏岡秀紀:"Confusion Network を用いた音声ドキュメントの検索及び評価に関する研究", 第 2 回音声ドキュメント処理ワークショップ講演論文集,pp.85-90,2008.3

[5] Julius, http://julius.sourceforge.jp/

[6] 汎用連想計算エンジン (GETA), http://geta.ex.nii.ac.jp/