ユーザ発話中の未知語を自動補完する音声入力型検索システム

藤井 敦[†,†††] 伊藤克亘[††,†††] 石川徹也[†]
[†] 図書館情報大学
[††] 産業技術総合研究所
[†††] 科学技術振興事業団 CREST
fujii@ulis.ac.jp

1 はじめに

近年の音声認識技術は，ある程度内容が整理されている発話に対しては実用的な認識精度を達成できるようになっており，様々な応用が考えられる．情報検索の分野では音声認識を採り入れた研究も数多く行われている．これらの研究は目的に応じて「音声データの検索」と「音声による検索」の2つに大別される．前者は，TRECのSpoken Document Retrieval (SDR) トラック [4] で放送音声データを対象にしたテストコレクションが整備されていることを背景にして盛んに研究が行われ，既に実用レベルに達している [7]．

それに対して，音声による検索はカーナビゲーションシステムやコールセンターのようにキーボード入力を前提としないアプリケーションを支える重要な基盤技術であるにも拘らず，音声データ検索に比べて研究事例は少ない．また，従来の研究では既存の音声認識とテキスト検索システムが単純に接続されているだけであり，音声認識誤りによって検索精度が顕著に低下する [1, 2]．これに対して，筆者らは検索対象のコレクションを用いて音声認識用の言語モデルを作成し，音声認識と検索精度の両方を向上させる手法を提案した [3, 5]．

しかし，音声入力型の検索システムでは，未知語（システム辞書未登録語）の問題がある．近年の情報検索システムは，古典的な統制語彙型システムとは異なり，検索対象テキスト中の任意の語による検索を可能とする．索引のサイズが数100万のオーダーに達することは珍しくない．機能語などは不要語として索引から除外されるものの，これらが検索キーワードとして利用されることは稀であるため，事実上，語彙制限はないと考えてよい．

他方において，近年の音声認識システムでは語彙サイズ（辞書登録語数）が制限される．これはハードウェアに関する制約や統計モデルの学習効率が主な原因であるため [13]，登録語数を増やすという単純な方法では解決が困難である．多くの言語において，語彙サイズは高々数万語に制限されており [6, 9, 11]，実用的な検索システムの索引サイズに比べると極端に小さい．

また，統計的な音声認識では，機能語などの高頻出語ほど高い精度で認識されるのに対して，情報検索では特定の文書にしか出現しない低頻度語ほど効果的な索引語になりやすい．すなわち，ユーザ発話中の効果的な検索キーワードほど誤認識されやすいという矛盾が生じる．

以上まとめると，音声入力型の検索システムにおいて「未知語問題」は本質的に不可避であり，何らかの積極的な解決策が必要である．本研究では，音声認識でカバーできない単語を検索用の索引語によって自動的に補完する手法を提案する．また，評価実験によって実装したシステムの有効性を示す．

2 システム概要

本研究で提案する音声入力型テキスト検索システムの構成を図1に示す．本システムは，音声認識，テキスト検索，未知語補完の3つのモジュールで構成されている．現在は日本語を対象に実装されているものの，本研究で提案する手法は言語の種類を問わない．以下，図1に基づいて本システムの処理について説明する．

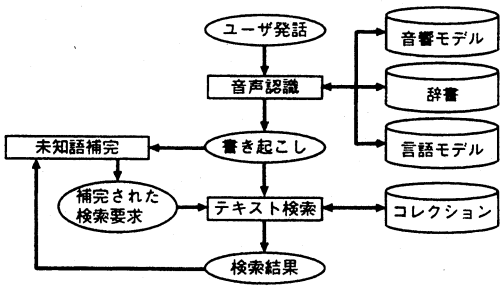

図 1: 音声入力型検索システムの構成

まず，ユーザが検索要求を発話すると，音声認識部が辞書，音響モデル，言語モデルを用いてユーザ発話の書き起こしを生成する．本システムでは，日本語ディクテーションツールキット [15] で提供されている音声認識エンジン（デコーダ）と音響モデルを利用した．しかし，ユーザ発話中に含まれる未知語を検出するために，辞書と言語モデルは独自に作成して利用した [14]．

具体的には，毎日新聞CD-ROM 10年分（1991–2000）の記事を「茶筌」[1]で形態素解析し，高頻度語20,000語

[1] http://chasen.aist-nara.ac.jp/

を抽出して辞書を構成した．通常は，辞書中の単語 N グラムなどによって言語モデルを作成する．しかし，これでは辞書未登録語は認識できない．そこで，辞書に登録されなかった約 30 万語（異なり数）を音節単位に分割し，単語と音節を併用してトライグラムを作成した．音節は異なりで 700 件あった．

すなわち，本システムの言語モデルにおいて，辞書未登録語は音節の組合せとしてモデル化されている．その結果，辞書未登録語は単語としては認識されないものの，音節単位でカタカナ列として書き起こしされる．また，当該言語モデルは通常の統計的 N グラムなので，既存のデコーダを拡張せずに利用できる．そこで，音韻系列の認識を別途必要とする手法 [12] とは異なる．

「オレンジやグレープフルーツなどの柑橘系果物の輸入に関する記事」という発話を例にとると，

> オレンジや /グ レ ー プ ラ チ ナ ガ ノ/ などの
> /カ ン キ ツ ケ イ/ 果物の輸入に関する記事

のように「グレープフルーツ」や「柑橘系」が未知語として検出される（ここでは未知語部分をスラッシュで括っている）．なお「柑橘系」のように未知語箇所の検出と音韻列の特定に成功する場合や「グレープフルーツ」のように音韻列の特定は不完全でも未知語箇所の検出に成功する場合がある．いずれの場合も，未知語に対する正しい語を推定することが出来れば，音声認識精度が向上し，結果として検索精度も向上する．

本システムのユーザは，検索対象のテキストコレクションから何らかの情報を引き出したいという意図を持って発話を行う．言い替えれば，ユーザの発話はコレクション中の情報に関連したものである可能性が高い．そこで，上記「グレープラチナガノ」や「カンキツケイ」に対応する正しい語がコレクション中に含まれていると考えることは自然な発想である．

直観的には，検索対象コレクションの索引語から，検出された未知語と音韻的に等価な語もしくは類似する語を探索してユーザ発話中の未知語を補完すればよい．しかし，音韻的に「類似する」語の探索（すなわち，音韻列の部分一致による探索）を大規模な索引に対して行うことは効率が悪く，実時間処理には耐えない．

そこで，まず，ユーザ発話中で単語として認識された部分だけを用いて初期検索を実行し，ユーザの検索要求に関連する文書を選択的に取得する．テキスト検索には確率型の「Okapi 法」[10] を用いた．当該手法は，与えられた検索要求に対するスコアを各文書に対して計算し，スコアが高い順番に文書を出力する．本システムでは，対象テキストを「茶筌」で形態素解析して名詞を索引語として抽出し，単語単位で索引付けを行って転置ファイルを事前に作成する．

次に，初期検索で得られた文書から，検出された未知語に対応する語を探索し，未知語と置き換えることで検索要求を補完する．具体的な方法については 3 章で説明する．最後に，補完された検索要求を用いて再検索を行い，最終的な検索結果が得られる．

上記の手法は，初期検索の結果を用いて最終的な検索精度を向上させるという点において，情報検索で用いられる検索要求の拡張（query expansion）やローカルフィードバックに類似している [8]．しかし，これらは検索精度を向上させることに主眼が置かれ，ユーザが意図しない索引語を追加する可能性がある．それに対して本手法は，ユーザの発話を正しく認識することを目的としている点が異なる．これは「自分が発話（意図）した通りに検索が行われている」という安心感をユーザに与える上で重要である（残念ながら，このような観点は情報検索の研究ではあまり考慮されていない）．

3 未知語の自動補完

3.1 方法論

本システムの特長は，音声認識で検出された未知語の音韻系列を，初期検索で取得された上位文書中の索引語に対応付けることによって単語として正しく認識する点にある．この処理を「未知語の補完」と呼ぶことにする．

同音意義語のために，一つの音韻系列が複数の単語に対応することがある（例えば「河川」と「架線」）．また，未知語の音韻系列は誤って検出されることがあるため，補完対象の音韻系列一つに対して，音韻的に類似する複数の索引語を考慮する必要がある．すなわち，未知語の自動補完では，複数の候補から適切な索引語を選択するための曖昧性解消が必要である．

そこで，選択されるべき索引語が満たす条件について検討し，以下に示す 3 つの基準を設定した．

- 補完対象の未知語との音韻的な類似度が高い（完全一致すれば類似度は最大となる）．
- 上位文書における出現頻度が高い．
- より上位の文書に出現する．

これらを確率論的な枠組で定式化すると，未知語補完は，式 (1) で計算されるスコアを最大化する t を選択することに相当する．

$$\sum_{d \in D_q} P(w|t) \cdot P(t|d) \cdot P(d|q) \qquad (1)$$

ここで，D_q は検索要求 q によって初期検索された上位文書の集合である．$P(w|t)$ は t が音韻的に w と等価である確率，$P(t|d)$ は上位文書の一つ d から索引語を無作為に選んだ場合に，それが t である確率，$P(d|q)$ は検索要求 q によって文書 d が検索される確率である．これらのパラメタは，上記 3 つの基準にそれぞれ対応している．

しかし，実際には $P(w|t)$ や $P(d|q)$ の確率値を正確に推定することは難しい．また，音韻的な類似度（上記，第1の基準）が他の基準よりもかなり強い制約になることが経験的に分かっている．そこで，予備実験の結果に基づいて，式 (1) を式 (2) のように近似する．

$$\sum_{d \in D_q} P(w|t) \cdot \log(P(t|d) \cdot P(d|q)) \qquad (2)$$

ここで，$P(w|t)$ は t と w が共有する音韻数と w に含まれる音韻総数の比率によって計算する．具体的には，DP マッチングによって t と w を音韻単位で比較し，両者に共通して含まれる音韻列を特定する．$P(t|d)$ は d における t の相対頻度で計算する．$P(d|q)$ として Okapi 法で計算される文書 d のスコアで代用する．また，$P(t|d)$ と $P(d|q)$ の log を用いることで，これら2つの影響力が相対的に小さくなるように制御している．

以上の方法は，索引付けの手法に依存しない点に注意が必要である．言い替えれば，索引語 t の単位として，文字，単語，複合語など文書中に現れる任意の文字列を対象とすることができる．

3.2 実装

初期検索によって文書数を制限しても，索引語数は膨大なものになる場合がある．特に，DP マッチングによる音韻単位の比較は実時間応答を低下させる要因となる．また，上位文書中の索引語の多くは，補完対象の未知語と音韻的に全く類似しないため，これらのノイズを早期に排除できれば，計算効率の向上が期待できる．

通常のテキスト検索に用いられる索引（本システムでは転置ファイル）は，入力されたキーワードとの完全一致によって，該当する項目を効率良く検索できる．しかし，未知語補完用の索引では，入力された音韻列に対して，部分一致を許容しながら，ある程度類似した項目だけを効率良く特定できなければならない．

本システムで用いる未知語検出の傾向を調査した結果，検出された未知語と，それに対応する正しい索引語は，前方もしくは後方で一致していることが多く，両端が一致せずに語中のみが一致することは少ない．そこで，未知語補完用の索引を以下の手順で事前に作成した．

まず，コレクション中の全文書を「茶筌」で形態素解析し，単語表記とカナ表記を抽出する．次に，カナ表記を規則によって音韻系列に変換する（規則数 143）．最後に，音韻系列の前方と後方から任意長の部分列を抽出して，前方/ 後方部分一致探索が可能な索引を編成する．

このとき，単語一つと単語バイグラムを併用して索引を作成することで「弥生/時代」や「オゾン/ホール」のように2単語で構成される複合語にも対応した．

原理的には，3単語以上で構成される長い複合語も扱うことができる．しかし，未知語の長さに比例して探索時間がかかるため，現在は2単語までとしている．また，現状の音声認識では機能語のような高頻度語は既知語として正しく認識されやすいため，長い単語列（例えば「情報検索の応用分野」）が一つにまとまった未知語として検出されることは稀である．

4 評価実験

本研究で実装したシステムを評価するために，IREX の日本語検索コレクション[2]を用いて実験を行った．当コレクションは，毎日新聞 1994–1995 年（記事総数 211,853 件）を対象にした検索課題 30 件と各課題に対する正解記事 ID で構成されている．検索課題の例を以下に示す．

```
<TOPIC><TOPIC-ID>1010</TOPIC-ID>
<DESCRIPTION>柑橘類の輸入</DESCRIPTION>
<NARRATIVE>オレンジ、レモン、グレープフルーツなどの柑橘系果物の日本への輸入の記事。政府の市場解放や輸入による日本生産地の影響、値段への影響や消費者の反応などの記事を含む。</NARRATIVE></TOPIC>
```

さらに，4名の話者（男女各2名）に<NARRATIVE>フィールドを読み上げてもらい，合計 120 件の音声発話データを作成して実験に利用した．初期検索，再検索ともに上位 300 件を出力した．

まず，未知語の検出と補完に関する評価を行った．30 件の検索要求（<NARRATIVE>のみ）に含まれる単語は，のべ数で約 400 語あり，14 単語（異なりで 13 単語）が音声認識用辞書に登録されていなかった．

未知語検出の再現率と精度はそれぞれ 71.4%と 22.6%であった．本システムは未知語を網羅的に特定する傾向があることが分かる．さらに，未知語の補完精度を調べた結果，36.2%であった．ここでは，辞書登録語が未知語として誤検出されても，補完処理によって正しい索引語に対応付けられた場合は正解と判定した．正しく補完された未知語と索引語の例を以下に示す．

グレープラチナガノ / グレープフルーツ
ヤヨイチタ / 弥生時代
ニククライス / ニックプライス
ベンピ / 便秘

次に，検索精度への影響を調べるために，以下の異なる検索手法（システム）を比較した．

1. テキスト入力型検索システム
2. 高頻度語 20,000 語のみを含む言語モデルを音声認識に使用した音声入力型検索システム
3. 本システム（検出した未知語は補完しない）
4. 本システム（未知語の検出・補完を併用）

[2] http://cs.nyu.edu/cs/projects/proteus/irex/

システム4が本研究で提案するシステムに相当する．システム2は未知語音節をモデル化していないため，未知語の検出と補完を行わない点を除けば，本システムと同じである．各システムの平均適合率（%）を以下に示す．

話者＼システム	1	2	3	4
男性#1	-	30.1	25.8	31.8
男性#2	-	27.9	26.8	29.6
女性#1	-	28.3	28.3	32.0
女性#2	-	27.5	24.4	28.5
平均	35.0	28.4	26.3	30.4

本システムの精度はテキスト検索には及ばないものの，約87%を再現している．また，全ての話者に対して，それ以外の音声入力型システム（2と3）の検索精度を向上させた．システム3と4を比較することで未知語補完の効果が分かり，システム2と4を比較することで，未知語の検出と補完を併用した提案手法の有効性が分かる．

しかし，精度の向上はそれほど大きくなかった．今回の実験では未知語が本質的に少なかったため，全体的な差異が大きくならなかった．また，未知語を人工的に作るような不自然な実験設定は避けた．未知語問題がより深刻な対象（例えば，技術文書やウェブページ）について，今後さらなる評価実験を行う予定である．

システム2に比べて，本システムの精度が顕著に低下した課題を分析した結果，初期検索の上位文書に正しい索引語が含まれているにも拘らず，式(2)のスコアで適切に選択されなかった事例が大半を占めた．例えば「制度」が未知語検出によって「セント」と誤認識されたために「鮮度」のように音韻的に等価な別の語が選択されてしまった．文書中の索引語頻度や文書順位などとのバランスについて今後検討が必要である．また，未知語音節をモデル化したために，辞書登録語を誤認識し，検索精度が低下した事例が若干あった．

最後に，オンライン処理のCPU時間を測定した．未知語の検出は通常の統計的音声認識の枠組内で行われるため，それに伴う付加的なCPU時間は発生しない．補完に要したCPU時間は未知語あたり平均3.5秒だった（AMD Athlon MP 1900+）．依然として改善の余地はあるものの，ほぼ実時間で動作すると考えてよい．

5 おわりに

音声入力型の検索システムでは，音声認識と検索における語彙サイズの不整合は不可避である．本研究は単語と音節を併用した言語モデルによってユーザ発話中の未知語を検出し，検索対象文書中の索引語によって適切に補完する手法を提案した．新聞記事を対象にした実験の結果，本手法は実時間で動作し，既存の手法を上回る検索精度を実現することができた．論文のような技術文書やウェブページの検索では，未知語問題がより深刻になる．今後は，これらを対象に研究を行う予定である．

参考文献

[1] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, pp. 1323–1326, 1997.

[2] Fabio Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pp. 267–281, 2000.

[3] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In *ACM SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications*, 2001.

[4] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pp. 83–91, 1997.

[5] Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

[6] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, and Kiyohiro Shikano. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustic Society of Japan*, Vol. 20, No. 3, pp. 199–206, 1999.

[7] Pierre Jourlin, Sue E. Johnson, Karen Spärck Jones, and Philip C. Woodland. Spoken document representations for probabilistic retrieval. *Speech Communication*, Vol. 32, pp. 21–36, 2000.

[8] K.L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 250–256, 1998.

[9] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of DARPA Speech & Natural Language Workshop*, pp. 357–362, 1992.

[10] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[11] Herman J. M. Steeneken and David A. van Leeuwen. Multilingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In *Proceedings of Eurospeech95*, pp. 1271–1274, 1995.

[12] Martin Wechsler, Eugen Munteanu, and Peter Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 20–27, 1998.

[13] Steve Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, pp. 45–57, September 1996.

[14] 伊藤克亘, 田中和世. 被覆率を重視した大語彙連続音声認識用統計的言語モデル. 日本音響学会講演論文集, pp. 65–66, March 1999.

[15] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. オーム社, 2001.