未知語に頑健な音声ドキュメント検索手法

中川　聖一
情報工学系／センター長

1　はじめに

近年，インターネット利用者の増加や，インターネット通信の高速化に伴い，ウェブ上のコンテンツが豊富になってきている．中でも，音声ドキュメントに代表されるマルチメディアデータは，ウェブ上に多数存在している．ニュースや新聞記事のように，テキスト情報を含むものであれば，Google や Yahoo など種々のテキスト検索エンジンを用いることで，欲しい情報を検索することができる．しかし，現在音声ドキュメントに対する検索方法は確立されていない．特に音声ドキュメント検索において，大きな障害となるのが未知語や認識誤りといった問題である．これらの問題に対処できる音声検索手法の確立が望まれている．そこで，本研究では，音声ドキュメントに対する検索手法について検討した．

2　音声ドキュメント検索手法

2.1　大語彙連続音声認識の利用

従来の音声ドキュメント検索手法として，最も簡単な方法は，大語彙連続音声認識の書き起こしの結果に対して単語単位のテキスト検索を行う方法である．しかし，この方法では，未知語や，音声認識誤りの問題に対処することができない．

未知語とは，大語彙連続音声認識の辞書にない単語のことである．辞書にない単語は，認識結果に現れることがないため，単語単位のテキスト検索では，未知語を検出することは不可能である．

また，認識誤りの問題もある．認識誤りによって，辞書に登録されている単語であっても認識結果に現れない場合があり，その場合はテキスト検索を使用しても検索することができなくなってしまう．そのため，大語彙連続音声認識システムの性能によってテキスト検索の性能も決まってしまう．

2.2　サブワード単位認識の利用

未知語に対しては，サブワード列として音節単位で認識した結果を使用する．日本語の音節数は 100 余種類である．音節列として認識することによって，認識の際に単語辞書を使用しないので，文法の制約を無視でき，未知語の発音をそのまま認識できる可能性がある．そこで，音節単位で認識した音節ラティスをサブワード列として用意しておき，音節ラティスのトライグラムを用いる．

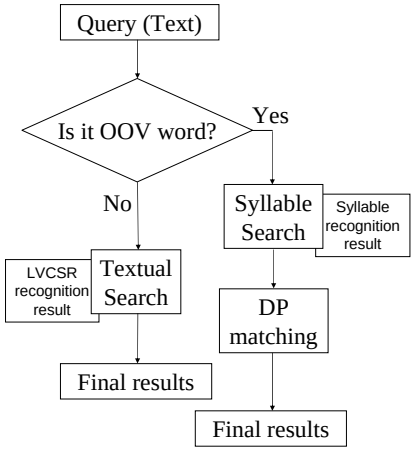

Fig. 1　提案手法のフローチャート

しかし，音声ドキュメントを事前にサブワード認識しておく必要がある．さらに，既知語に対しては，単語単位のテキスト検索よりも精度が低下することが知られている．そこで，本研究では，クエリが既知語の場合は，従来の大語彙連続音声認識の結果にテキスト検索を行い，クエリが未知語の場合のみ事前に用意しておいたサブワード（音節）の認識結果を用いて検索を行うこととする．

3　提案する未知語検索手法

本研究では，未知語に頑健な検索手法として，音節ラティスを使用して検索の際に認識誤りを考慮して検索を行う．本手法の概略を Fig. 1 に示す．検索対象の音声ドキュメントに対して大語彙連続音声認識と連続音節認識を行い，インデックス化する．

検索の最初に，まずクエリが与えられると，それが辞書に登録されているか，未知語であるかの判断をする．辞書にある場合，既知語に対しては，従来の手法を使用するため，大語彙連続音声認識の結果に対してテキスト検索を行う．未知語であれば，クエリを音節に変換して連続音節認識結果に対して検索を行う．

未知語を検索可能にするため，サブワード列として音節ラティスの上位 3 ベストを使用する．そして，音節ラティスのデータを保持させておくデータ構造としてトライグラムアレイを定義する．トライグラム

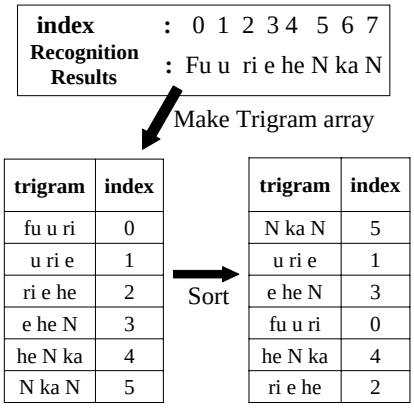

Fig. 2　トライグラムアレイ作成手順

アレイとは，サフィックスアレイを元に考えたデータ構造である．サフィックスアレイでは，音声ドキュメント内での出現位置情報のみを保持していたが，トライグラムアレイでは，出現位置とそこで出現するトライグラムの情報も保持させておく．トライグラムアレイの作成方法を Fig. 2 に示す．

まず，音声ドキュメント内での出現位置を割り振っておき，各出現位置での音節のトライグラム (3 つ組み) を作成する．そしてトライグラムを辞書順にソートしておく．それによって，二分探索を用いて高速に検索することが可能になる．検索の際は，クエリを音節に変換し，トライグラムを作成してトライグラムアレイに対して検索を行う．その際，3 音節のクエリであればそのままトライグラムになり，検索することができる．4,5,6 音節のクエリのときは，前半と後半の 2 つのトライグラムを作成し，2 回検索を行う．そして，検出結果の中で，前半と後半の出現位置のずれが ±1 以内であれば，連続していると考えて，4〜6 音節のクエリに対応させた．同様に 7〜9 音節のクエリは，音節列を 3 分割して 3 つのトライグラムを作り，検索を行う．出現位置のずれを考慮するのは，認識誤りに対処するためである．

また置換誤り対策は，音節ラティスの上位 3 候補までを使用し，トライグラムアレイを作成し対処した．挿入誤り対策は，トライグラムアレイを作成する際に，音節を 1 つ飛ばしたトライグラムを作ることで，1 文字挿入に対処した．最後に脱落誤りは，クエリを脱落させて検索することで対処した．検索の際は，クエリのトライグラムを作り，インデックス順にソートされたトライグラムアレイに対して検索を行う．検索結果は，未知語や認識誤りも検出できるように条件を緩めて検索を行うため，実際に発話した部分以外も検出される湧き出し誤りが生じる．そこで，その検索結果に対してクエリとの音節列同士の DP マッチングで検索結果のスコア付けを行い，ある閾値以下の検索結果は，除くことにした．

4　評価実験

4.1　実験データ

評価用データは，本学で作成された日本語講義音声コンテンツコーパスの 3 講義分約 210 分を SPOJUS でコンテキスト依存音節モデル (918 音節モデル) で音節認識した結果に対して検索，評価を行った．

音節認識率は，本講義音声に対しては，1 ベストのみ考慮した場合で約 62%，3 ベストまで考慮しても認識率は約 74%程度である．

4.2　検索結果

まず，講義音声ドキュメント内に存在していた既知語 (「対話」,「人間」,「形態素」,「意味解析」,「音声認識」など)31 種類 458 個に対して，単語認識率（正解率）が 50%〜60%の大語彙連続音声認識結果を用いて検索を行った結果，Recall は約 87%で，Precision は 100 %であった．つまり，検索語となりうるキーワードの音声認識率は 87%で挿入誤りはなかった．

次に，講義音声ドキュメント内の未知語 (「愛知県」,「河口湖」,「中川」,「酷使」,「望遠鏡」など)45 種類 150 個に対して，音節列認識結果を用いて検索，評価を行った．検索時間は，1 個のクエリに対して，平均約 3 秒程度であった．

各認識誤り対策を個々で行った結果，脱落と置換誤りの対処を行うと検出がかなり増加した．そこで，各認識誤り対策を組み合わせて評価を行った．組み合わせることによって Recall の改善がみられ，全ての認識誤り対策を行うことで，未知語 150 個のうち 71 個の検出に成功した．

5　おわりに

本研究では，音声ドキュメントに対して，音節ラティスを使用した未知語に頑健な検索手法について，評価した結果を報告した．音節認識率 62%（3 ベストで 74%）の講義ドキュメントに対して，未知語のうち約半分の未知語の検出に成功した．各認識誤りへの対処を行っていけば，未知語の検出率は増加するが，湧き出し誤りも増加してしまった．そこで，湧き出し誤りを減らすため，DP マッチングの挿入や脱落のコストを導入し，湧き出し誤りを大幅に減少できた．

発表論文

[1] 中川聖一，高橋将史, 藤井康寿，山本一公：『未知語に頑健な音声ドキュメント検索手法の検討』第 3 回音声ドキュメント処理ワークショップ講演論文集, pp. 7–14 (2009.2).