G-002

音声文書インデキシングのためのWEB文書を利用した自動誤り訂正
Autmatic Error Correction using WEB Documents for Spoken Document Indexing

伊藤 友裕[†]　西崎 博光[‡]　関口 芳廣[‡]　中川 聖一[§]

Yusuke Itoh　Hiromitsu Nishizaki　Yoshihiro Sekiguchi　Seiichi Nakagawa

1. はじめに

近年，情報通信技術の発達に伴ない，大量の電子化されたマルチメディアデータを誰もが容易に送受信できるようになり，それらのデータを検索する技術の開発が行われている．一般的にマルチメディアデータの検索を行うためには，データに対し何らかのインデキシングを行わなければならない．マルチメディアデータ，特に音声文書を検索するために必要な自動インデキシングを行う方法として音声認識結果を利用することが考えられる．しかし，自動でインデックスを付与するために音声認識技術を用いると，未知語（音声認識辞書に含まれない単語）や誤認識の問題に直面する．名詞，特に固有名詞など文書の特異性を表す重要単語は未知語になりやすく，音声認識での正確な書き起こしが難しいとされている．特にニュースなどの時事文書を扱う場合はなおさらである．

そこで，本稿では音声認識結果中の認識誤りを自動的に特定し，マルチメディアデータをインデキシングする際に特に重要で，誤認識されやすい（未知語になりやすい）固有名詞に着目し，固有名詞の認識誤りを訂正する方法を提案する．音声誤り個所の検出には，2種類の音声認識システムを用い，2つのシステムで共通して同じ単語に認識された部分は正解と見なし，一致しない部分を訂正必要箇所と判断し，認識誤り箇所検出を行なっている．また，本システムの訂正候補の単語はインターネット上のWEB文書中から選択する．

2. 認識誤り訂正処理

2.1 処理の概要

処理の概要を図1に示す．

本稿では，認識結果1文ずつ別々に処理をするのではなく，ある一つの塊り（今回はニュース文書を扱うのでその文書）を単位として訂正を行う．まず，2つの音声認識システムからの認識結果に対して，誤り訂正個所の検出を行い，認識結果に対して訂正必要箇所を判別しておく．同時に，正しく書き起こされたと判断した単語から，WEB文書を検索するのに用いる単語を選択する．その単語集合を利用してWEB文書を検索し，得られたWEB文書集合から誤認識箇所に対する代替単語の候補を選択し，訂正必要箇所と音韻的なマッチングを行い，正しくマッチングした代替単語を訂正必要箇所に置換する．
なお，置換単語は最大5つまで許している．

2.2 誤認識箇所の検出処理

誤認識箇所の検出には，2種類の音声認識システム（Julius[1]，SPOJUS[2]）を用いる．ニュース音声データに対して，2つの音声認識システムで共通して同じ単語に認識されると，高確率（約95%）で正解であると判定できる[3]ことが実験的に実証されており，2種類の音声認識結果が一致しない単語を訂正必要箇所と判断し，音節認識（言語モデルを用いない認識結果）のデータと比較し訂正必要箇所に相当する音節認識データを比較対象データとする．

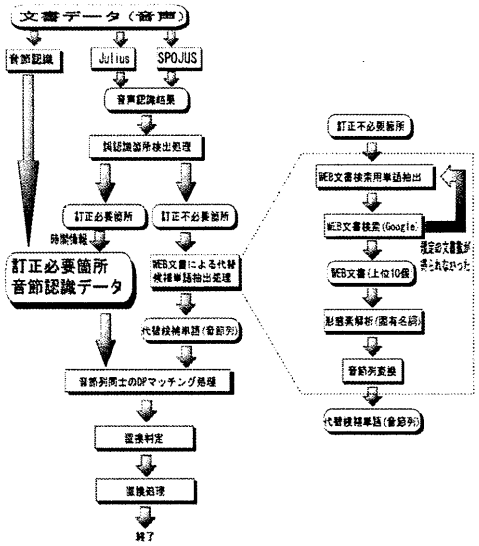

図1: 認識誤り訂正処理の流れ

2.3 代替候補単語の抽出

認識誤り箇所の代替候補単語の探索に用いる範囲として，インターネット上のWEB文書を用いる．認識結果文と類似した内容の文書集合をインターネット検索エンジンGoogle用いて検索する．

Googleで検索するときに用いる検索キーワードとして，訂正必要箇所を含む同じ記事（文書）内から選択した固有名詞を用いる．その理由としては，固有名詞は検索に重要な単語であるためである．当然，検索キーワードに使う固有名詞は2つの認識システムで共通で認識した正解（らしき）単語である．Googleで得られた上位10文書を形態素解析器ChaSenを使って形態素解析を行い，固有名詞のみを音節列に変換し，代替候補の単語集合を作成する．もし，WEB文書が1つも検索されてこなければ，キーワード集合から固有名詞を1つ取り除き検索を行う．また，10文書未満であれば，検索された文書だけを用いることにする．

2.4 正解代替候補との置換判定処理

2.4.1 訂正箇所と代替候補とのマッチング

認識誤り箇所検出部で述べた処理により検出された誤認識箇所に対して，代替候補単語と音節認識での比較対象データとの音韻的な距離尺度に基づきマッチングを行い，置換の妥当性を判断する．つまり，訂正必要箇所に対して正解代替候補が含まれているかどうかワードスポッティングする[4]．

[†]山梨大学大学院医学工学総合教育部
[‡]山梨大学大学院医学工学総合研究部
[§]豊橋技術科学大学情報工学系

2.4.2 置換判定

訂正必要箇所における音節認識での比較対象データに対して代替候補でワードスポッティング処理を行い，その結果ある設定閾値以上の値の代替候補のうち，上位3単語を最適候補と呼び，置換単語とし，ある設定の閾値以上の代替候補が存在しなかった場合は，その訂正必要箇所に入る単語は存在しないと判断する．また，Juliusで正解に認識されていてSPOJUSで誤認識してしまった単語，また，逆の場合すなわちSPOJUSで正解に認識されていてJuliusで誤認識してしまった単語も本稿では訂正必要箇所としているため，最適候補3単語に更にJuliusでの認識結果，SPOJUSでの認識結果，(固有名詞限定）というものも置換単語として追加する．この処理を訂正必要箇所すべてに行う．よって，1つの訂正必要区間に対し最大5単語の置換単語が入ることになる．

3. 訂正実験

3.1 評価用音声データ

評価用音声データは，NHKニュース音声データベースから，1996年6月1日,2日の「ニュース7」と「おはよう日本」の収録音声を用いた．記事（文書）数は33文書，10429単語である．固有名詞は630単語である．

本実験で用いる音声データにおいて，各認識システム毎の固有名詞の認識率を表1に示す．

表1: 各認識システムの固有名詞認識率 [%]

LVCSR	固有名詞正解率	正解固有名詞数
Julius	66.8%	421
SPOJUS	59.5%	375

3.2 訂正実験

提案した一連の動作を行なう実験システムを作成し，訂正実験を行なった．

音声認識結果中の正解（共通で認識された）単語の固有名詞を検索キーワードとし，それによって検索されたWEB文書データに含まれる固有名詞（代替候補）と訂正必要箇所に当たる音節認識の結果を音節列レベルでのDPマッチング（ワードスポッティング）を行ない．マッチングによって得られた一定の閾値以上でスコア上位3つの代替候補単語を置換単語とする．また，片方の認識システムで正しく認識された固有名詞も存在するが，本実験では誤認識として扱っているため，更にJuliusでの認識単語，SPOJUSでの認識単語（固有名詞限定）も正解置換候補単語として追加し，最大5単語を置換結果とする．なお，1音素の固有名詞で重要なものはほとんど無いため，音素数1の訂正必要箇所は訂正必要箇所に含めていない．

表2〜表4に訂正実験結果を示す．2つの音声認識システムを使った誤認識の検出実験では，誤認識した固有名詞の約90%(232/259 = 0.896)を誤りであると検出できる．なお，そのうちの246単語，約95%(246/259 = 0.951)の単語に対して，2つの認識システムで共通に認識された名詞から，人間の主観によって検索キーワードを選び検索された（手動で検索した）WEB文書中に存在することが調査済みである．

このシステムにより，自動で検索語を選定し検索を行うと，訂正対象の固有名詞の143単語，約55%(143/259 = 0.552)がインターネット上から自動で検索できる．DPマッチングのスコア上位3ワードかつ一定の閾値以上の代替候補（置換単語）において，正しく置換された単語数は62単語，約24%(62/259 = 0.239)が訂正できる．このうち，2つの認識システムが共に認識出来なかった単語は32単語であった．すなわち，30単語は一方の認識システムでは正しく認識している．

表2: 誤り箇所検出

共通部分の固有名詞正解率	92.7%
共通に認識した固有名詞	371 単語
共通で誤認識された固有名詞	27 単語
自動検出した訂正区間	1025 箇所
訂正必要固有名詞を含む区間	239 箇所
訂正区間に含まれる誤認識固有名詞	259 単語

表3: ワードスポッティング処理

訂正単語	259 単語
自動取得 WEB 文書に含まれる正解単語	143 単語
DP 上位 3 ワードで正しく置換できた単語	62 単語
Julius では正確に認識されていた単語	21 単語
SPOJUS では正確に認識されていた単語	9 単語
両システムの誤認識を正しく置換できた単語	32 単語

この上位3ワードに2つの認識システムの結果を加えて置換した結果，訂正区間に含まれる誤認識した固有名詞140単語を自動で置換できた．すなわち，このシステムによって，約54%(140/259 = 0.541)のインデキシングに必要な固有名詞を訂正出来たことになる．

表4: 置換結果

訂正単語	259 単語
このシステムにより正しく置換できた単語	140 単語
Julius では正確に認識されていた単語	77 単語
SPOJUS では正確に認識されていた単語	31 単語
両システムの誤認識を正しく置換できた単語	32 単語

この処理によって，音声データをインデキシングする際必要な固有名詞の再現率をJulius 66.8% (421/630 = 0.668)，SPOJUS 59.5% (375/630 = 0.595) から76.8%(484/630 = 0.768)に上げることができた．

4. まとめ

本稿では，インターネット上のWEB文書を利用して音声認識の誤り訂正を行ない，特にニュース音声を対象とし，検索に重要である固有名詞に限った誤り訂正のシステムを作成し報告をした．この手法によって，音声データをインデキシングする際必要な固有名詞の再現率を(Julius,SPOJUSそれぞれでは67%,60%から）約77%に上げることができた．

本稿では，訂正必要区間において最大5単語を置換対象としているため，音声データのインデキシングを行なう際に曖昧性が生じる．よって今後の課題としては，置換する単語の候補を更に減らす処理が必要がある．また，人間の主観による検索キーワードを選んだ結果，約95%は正解単語を含んだページを検索出来ることが分かっているため，WEB用の検索キーワードの抽出方法を改善することにより正解単語を持ってくることができると考えられる．更に，意味的な情報，例えば,『○○ 総理』という単語列では，○にはほぼ必ず人名が入るといった情報を利用することで，さらに訂正精度の向上を図ることが可能であると考えられる．

参考文献

[1] 河原他，日本語ディクテーション基本ソフトウェア (99年度版) の性能評価，情報研報,2000-SLP-31-1, pp.9–16, 2000.
[2] 甲斐他，単語n-gram言語モデルを用いた音声認識システムにおける未知語・冗長語の処理，情処論文誌, Vol.40, No.4, pp.1383–1394, 2000.
[3] 宇津呂他，複数の大語彙連続音声認識モデルの出力の共通部分を用いた高信頼度部分の推定，信学論誌 D-II, Vol.J86-D-II,No.7, pp.974–987, 2003.
[4] 西崎他，音声認識誤りと未知語に頑健な音声文書検索手法”，通信学会論文誌 D-II, Vol.J86-D-II, No.10, pp.1369–1381, 2003.