情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-NL-201 No.14
Vol.2011-SLP-86 No.14
2011/5/17

子音に注目した早口言葉の検索

鶴 巻 有 香[†1]　安川 美智子[†1]　横 尾 英 俊[†1]

滑舌訓練用の類似音の検索を行うことを目的として，日本語の子音の特徴に注目した早口言葉の検索方法を提案する．まず最初に，被験者実験を行い，言い間違いの具体例を調査分析した結果を報告する．次に，言い間違いは特に子音で生じやすいことから子音の特徴に注目した類似文字列の検索を提案する．漢字かな混じりの文字列を読み仮名に変換し，さらに母音の特徴を抽象化する記号体系に文字列を変換することにより，表記ではなく単語の読みの類似性で検索が行えることが期待できる．滑舌訓練用の例文を用いた評価実験により，提案法は従来法と比較して，類似音を持つ早口言葉の検索性能が高いことを確認した．

Tongue Twister Retrieval based on Japanese Consonants

Yuka Tsurumaki,[†1] Michiko Yasukawa[†1]
and Hidetoshi Yokoo[†1]

We propose a new method for Japanese tongue twister retrieval based on the manner and position of consonant articulation. In this paper, we first study the mechanism of mistakes in a set of difficult phrases to say correctly. Then, we investigate the requirements for the tongue twister retrieval and design two types of encoding tables to abstract the pronunciation of each tongue twister. According to the result of the evaluation experiment, our proposed method is effective to search similar tongue twisters that do not necessarily have common spellings but have common speech sound.

1. はじめに

俳優やアナウンサーなどが発音する際に，その舌の回りが滑らかなことを「滑舌[⋆1]」ということがある[1)]．滑舌はもともとは放送業界における専門用語であったが，最近では，専門家以外の人たちの間でも,「滑舌が良い（悪い)」といった表現が使われるようになっている．滑舌を良くするための訓練は「滑舌訓練」と呼ばれており，滑舌訓練の一つに,「早口言葉」を用いる方法がある．早口言葉とは，同音が重複するなどして，早く正確に言うことが難しい語句や台詞を早く唱える言語遊戯である．本研究では，滑舌訓練を目的とした早口言葉の検索を扱う．

本研究では滑舌の良い話し方を，以下の 2 つの条件を満たす話し方であると定義する．

- 言い間違いがない（明瞭，明確，正確）
- 発話の速度が一定で，かつ，遅くない（軽快，軽妙）

従来より，滑舌が良い話し方が求められる場面が存在した．たとえば，音声認識技術を用いた音声入力では，話者が適度な速度で正確な日本語を音声を話すことが求められ，言い間違いをしない方が効率的に入力できる．また俳優やアナウンサーではない一般の人たちの間でも，会社員であれば会議，教員であれば講義，学生であれば卒業研究の発表といった場面において，人前で話す機会がある．最近では，会議，講義，研究発表の模様がインターネットを通じて配信されることも増えており，字幕がなくても視聴者が理解できる滑舌が良い話し方は，俳優やアナウンサーに限らず，広く求められ，滑舌訓練の必要性が増している．

滑舌訓練において，もともとうまく言えている早口言葉を漫然と練習することは非効率であり，言い間違いやすい早口言葉を重点的に訓練することが効率的かつ効果的である．そこで，本研究では，言い間違えた部分と類似する音を含む早口言葉を検索し，滑舌訓練を効率的かつ効果的に行えるようにする早口言葉の検索方法を提案する．

2. 言い間違いに関連する従来研究

「言い間違い」とは,「X と言おうとしたが，Y と言ってしまった」ということである．外池[2)] は，文法の観点から，言い間違いを以下の 3 つに分類している．

(1) 音声的間違い ··· 誤：テンゴクゼンキガイヨウ　正：全国天気概要
(2) 統語的間違い ··· 誤：41 年ドンナ，トウジ　正：41 年当時，どんな
(3) 意味的間違い ··· 誤：覚えるのがトクイデス　正：覚えるのが苦手です (得意ではありません)

また寺尾[3)] は，言い間違いは音韻単位，形態素単位，語彙単位など様々な言語学的単位で生じることを指摘し，誤りのタイプとして付加，欠落，交換，混成などを挙げている．

†1 群馬大学
Gunma University
⋆1 活舌と表記されることもある．

Vol.2011-NL-201 No.14
Vol.2011-SLP-86 No.14
2011/5/17

(1)　付加　誤：草野仁さんシカイノ司会で　正：草野仁さん司会で　語彙単位
(2)　欠落　誤：青木とナカジの組　正：青木と中島の組　音韻単位
(3)　交換　誤：ナガせばハナい　正：話せば長い　形態素単位
(4)　混成　誤：ショッタイ　正：接待と招待　語彙単位

上に示した例のうち,「テン ゴク ゼン キガイヨウ（全国天気概況)」や「ナガ せば ハナ い（話せば長い)」のように 1 組の単語の先頭が入れ替わる言い間違いは，特に spoonerism[7] と呼ばれている．英語の spoonerism の例としては，the long river と言おうとして，the wrong liver と言ってしまう誤りが挙げられる[4]．ところで，この英語の spoonerism の例は，long/wrong と river/liver をもともと正確に発音できる英語話者の言い誤りである．英語の R/L を発音できない日本語話者の R/L の混同は,「発音の誤り」であると考える．また同様に，以下のような「日本語を母語としない話者の日本語の誤り」[5] も発音の誤りであると考える．

(1)　拗音　誤：ワイセツください．　正：ワイシャツください．
(2)　促音　誤：キタナイパンはありますか？　正：切ってないパンはありますか？
(3)　母音　誤：サワッテいいですか？　正：座っていいですか？

言い間違いに類似する概念として,「言い淀み」と「言い直し」がある．これらの概念を本研究では以下のように定義し，言い間違いとは区別する．

(1)　言い淀み：発話の途中に「あー」「えーと」「なんだっけ」などの不要な語句が挿入されること
(2)　言い直し：発話の途中で相手にさえぎられるなどして発話が中断され，反論や強調の目的で，既に発話した語句を別の語句で置き換えて発話が再開されること

土井[6] は会議中の自由発話において，言い直しが発生する箇所を分析し，言い直しは，話者の関心が高い単語と関係があることを報告している．

言い淀みと言い直しは，議論や説明を進める上で，円滑さや流暢さを欠く話し方であるという点で，言い間違いと共通点があるが,「X と言おうとしたが，Y と言ってしまった」という間違いではないため，言い淀みと言い直しは，言い間違いではないと考える．

Fromkin[7] は発話における話者の発話のメカニズムと実際の言い間違いの例を分析し，以下の点を指摘している．

- 有声音/無声音，鼻音，両唇音といった子音の特徴が話者の心内辞書に記憶されているので，音韻の転位が生じて言い間違いとなる．
- 音韻の転位は，発声のために筋肉を動かす前，または筋肉を動かしている最中に生じる．

本研究では，話者の自由な発話における言い間違いではなく，早口言葉を用いて滑舌訓練を行う際の言い間違いを対象としていることから，特に音声的間違いに注目し，子音の特徴という観点から言い間違いの分析を行う．

表 1　アナウンサーが最も言いづらい言葉 (2004 年)
Table 1　The 10 most difficult words for announcers to speak about in 2004.

順位	言いにくい言葉	読み	人数
第 1 位	高速増殖炉もんじゅ	こうそくぞうしょくろもんじゅ	75
第 2 位	手術中	しゅじゅつちゅう	47
第 3 位	貨客船万景峰号	かきゃくせんまんぎょんぼんごう	26
第 4 位	取りざたされる	とりざたされる	24
第 5 位	白装束集団	しろしょうぞくしゅうだん	21
第 6 位	出場	しゅつじょう	20
第 7 位	栃乃洋	とちのなだ	16
第 8 位	老若男女	ろうにゃくなんにょ	13
第 9 位	暖かく	あたたかく	12
第 10 位	火星探査車	かせいたんさしゃ	10

出展:フジテレビ系列 27 局のアナウンサー 336 人にアンケート調査した結果[8]

3. 言いにくい言葉を用いた言い間違いの調査

表 1 に示す「アナウンサーが最も言いづらい言葉」[8] を用いて被験者実験を行い，言い間違いの具体例を収集した．表 1 の人数は，文献 8) で報告されているアンケートにおいて，言いにくい言葉を回答したアナウンサーの人数であり，順位は言いにくいと答えたアナウンサーの人数の多い順を示している．本研究では表 1 の言いにくい言葉と読み (ひらがな) を，10 人の被験者（日本語話者 9 名＋中国語話者 1 名；男性 8 名＋女性 2 名）に配付し，各自で音読してもらい，言い間違いが発生した箇所に下線を付してもらった．被験者から収集した言い間違いを表 2 に示す．表中において被験者 (10) は中国語話者であり，日本語を母語としない話者にとって，これらの言葉を流暢に発音することは容易ではないことが推察される．被験者 (1)〜(9) は日本語を母語とする話者であり，被験者 (8) は間違いの箇所が最も少ない．被験者によって言い間違いの傾向は様々であり，拗音 (小文字の「ャ」「ュ」「ョ」)，濁音（「ガ行」「ダ行」)，撥音（「ン」)，カ行，サ行，タ行，ナ行，マ行で言い間違いが発生している．言い間違いをした被験者の人数が多い言葉は,「火星探査車」(8 人)，および,「手術中」,「貨客船万景峰号」,「白装束集団」,「出場」(それぞれ 7 人) であった.「取りざたされる」と「栃乃洋」を言い間違えた被験者は少なかった．アナウンサーは原稿の中で

Vol.2011-NL-201 No.14
Vol.2011-SLP-86 No.14
2011/5/17

表 2 被験者が言い間違えた箇所
Table 2 Mispronounced parts by subjects.

識別子	被験者 (1)	被験者 (2)	被験者 (3)	被験者 (4)	被験者 (5)
T01	しょく	n/a	n/a	しょくろ も	n/a
T02	n/a	しゅじゅつ	じゅつ ち	じ ゅつちゅう	しゅじゅつ
T03	ぎょん	ん ぎょん ぼ んごう	ぼ ん ごう	ん まんぎょん	n/a
T04	n/a	n/a	n/a	n/a	ざ た さ
T05	しょう	しょうぞく	n/a	う ぞく しゅ	ろ しょう
T06	しゅつ	ゅつじょう	しゅつ じょう	しゅつじょ	しゅつ じょ
T07	n/a	n/a	n/a	の な だ	n/a
T08	n/a	n/a	な んにょ	く な ん	ろう にゃく なん
T09	n/a	n/a	n/a	n/a	n/a
T10	さし ゃ	い たん さ しゃ	た んさ しゃ	ん さ しゃ	たん さしゃ
識別子	被験者 (6)	被験者 (7)	被験者 (8)	被験者 (9)	被験者 (10)
T01	n/a	う しょくろ も	く ろ も	う そく ぞ	う そくぞう しょく
T02	しゅじゅつ ちゅ	n/a	n/a	しゅじゅつ ちゅ	じ ゅつ ちゅ
T03	ぼん ごう	か きゃく せ	n/a	n/a	く せんまん ぎょん
T04	n/a	n/a	n/a	n/a	り ざた さ
T05	n/a	ろ しょう ぞ	う ぞく しゅ	n/a	しろしょうぞく しゅ
T06	n/a	n/a	n/a	し ゅつ じょ	し ゅつ じょ
T07	n/a	n/a	n/a	n/a	の なだ
T08	く なん にょ	n/a	n/a	n/a	ん にょ
T09	あ たたか く	na	na	あ たた か	た たか く
T10	ん さ しゃ	ん さしゃ	na	na	た んさしゃ

これらの言葉を読み上げる際に，前後につながる言葉の影響を受けるのに対して，被験者は各単語を単独で読み上げたため，言いやすくなったと考えられる⋆1．表 1 は 2004 年に実施されたアンケート調査であるが，2 年後の 2006 年に行われたアンケート調査[9] では，以下のような言葉と言い間違いの具体例が報告されている．

- 摘出手術 (てきしゅつしゅじゅつ) 誤：てきしゅつしゅじゅちゅ
- 腹腔鏡手術 (ふくくうきょうしゅじゅつ) 誤：ふくくう こう しゅじゅつ
- 低所得者層 (ていしょとくしゃそう) 誤：ていしょとくしゃしょう
- 六ヶ国協議 (ろっかこくきょうぎ) 誤：ろっか きょく きょうぎ
- 貨客船万景峰号 (かきゃくせんまんぎょんぼんごう) 誤：かきゃくせんまん ぼんぎょん ごう
- マサチューセッツ州 (まさちゅーせっつしゅう) 誤：まさちゅーせっちゅ しゅー
- 偽札造り (にせさつづくり) 誤：にせさつ じゅ くり
- 高速増殖炉もんじゅ(こうそくぞうしょくろもんじゅ) 誤：こうそく じょう しょく りょ もんじゅ

⋆1 たとえば,「栃乃洋」と単独で言うよりも「栃乃洋などの」と言う方が言いにくくなる．

上述のようなアナウンサーの言い間違いや，表 2 のような被験者の言い間違いが生じる原因について塩原[10] は，子音の調音位置と調音様式の観点から様々な分析を行っている．たとえば,「風邪・全然」を舌先の摩擦なしで読むと「カデ・デンデン」となり，前舌と硬口蓋を使う読み方をすると「カジェ・ジェンジェン」となる．また,「シュ・ジュ」の拗音を「シ・ジ」の直音へと変えてしまう東京方言の訛りの影響を受けて「静粛（セーシュク)」が「セーシク」,「千住（センジュ)」が「センジ」,「算術（サンジュツ)」が「サンジツ」と発音されることがある．さらに，近似音交錯型の難読文 (口構えや舌先のすばやい変化が必要) や異種音交錯型の難読文 (調音点の激しい移動をともなう) など多数の滑舌訓練用の例題を子音の特徴別に分類して紹介し，正確な発音を行うために，調音様式や調音位置を意識しながら同型の例文を用いた反復練習を行うことを塩原[10] は提案している．そこで，言い間違えた文と同型の例文，すなわち，言い間違いをしやすい調音パターンを含む例文を検索するシステムがあれば，滑舌訓練を行う際に役立つと考えられる．

```
There is truble.

1) trouble          6) tribal
2) treble           7) Trimble
3) triable          8) troubled
4) troubler         9) troubles
5) ruble            0) terrible
i) Ignore           I) Ignore all
r) Replace          R) Replace all
a) Add              l) Add Lower
b) Abort            x) Exit
?
```

図 1 英語の綴りの修正の例
Fig. 1 Example of spelling correction.

```
[michi@lemon NL201]$ simstring -u -d db/citrus.db
シークヮーサー
        シークヮーサー
        シィークヮーサー
        シイークヮーサー
3 strings retrieved (0 sec)
```

図 2 日本語の表記揺れの検出の例
Fig. 2 Example of orthographical variants.

4. 早口言葉の検索

本研究では「滑舌訓練を目的とした類似音を含む早口言葉の検索」を提案する．提案する検索は，単語の音に注目した検索を行うという点に特徴があり，この点において単語の意味

Vol.2011-NL-201 No.14
Vol.2011-SLP-86 No.14
2011/5/17

に注目した従来の情報検索とは異なっている．たとえば，Google 検索[⋆1]で「新宿」を検索すると「新宿」という地名に関連する Web ページや地図情報が検索結果として提示され，ユーザは新宿駅の電車の運行情報や新宿区のレストランの営業時刻などを調べることができる．一方，早口言葉の検索では,「新宿」という単語の意味ではなく,「新宿」の読み「しんじゅく」の音に注目した検索を行う．たとえば，ユーザが「新宿」という検索文字列を入力すると「学習塾 (がくしゅう じゅく)」「三重苦 (さ んじゅ うく)」「千手観音（せ んじゅ かんのん)」といった「しんじゅく」の類似音が検索される．

これまでに文字列編集距離[11] や N-gram を用いた類似文字列検索[12] のアルゴリズムが研究され，Aspell[⋆2]などのスペルチェッカ (図 1) や，表記揺れの検出 (図 2) といった用途に応用されている．そこで既存の類似文字列検索の手法を単語の読みに適用し，類似音の検索が行えるかを確認した．

具体的には，まず，形態素解析器 ChaSen 用の辞書 ipadic-2.7.0[⋆3]から名詞の読みを抽出し，重複を省いたものを入力文字列集合とし，類似文字列検索ライブラリ SimString[⋆4] のユーティリティを用いて，データベースを構築した．次に，構築したデータベースに対して表 1 の読みを検索文字列として検索を行った．

検索結果を調べたところ，いくつかの検索文字列に対しては，似て非なる類似文字列を検索することに成功したが，言いにくさという点で検索文字列とは共通の性質を持たないものがほとんどであった．たとえば検索文字列「しゅつじょう」に対して類似文字列「かつじょう」「けつじょう」「しゅつじん」「はつじょう」「ひつじょう」が検索されたが，拗音と「サ行音・ザ行音」の交錯という，検索文字列と共通の言いにくさを有する文字列は「しゅつじん」だけである．滑舌訓練に適した早口言葉を検索するためには,「検索文字列と同様の言いにくさ」という観点からの類似度計算用の文字列の特徴の取り方を工夫する必要がある．

以上のことをふまえて，文字列の読みを検索用の記号列に変換し，変換した文字列から類似度計算用の特徴を取る手法を考える．表 2 を見てみると，日本語の五十音の「ア行」の音では言い間違いは生じておらず，子音が言いにくさに関係していると考えられることから，特に子音の特徴に注目し，文字列の読みをローマ字綴りに変換し，変換した文字列から子音の特徴を取ることとする．文字列の特徴の取り方は，uni-gram，bi-gram，tri-gram など任意の長さの N-gram を作成することができるが，以下の説明では tri-gram を用いる．

⋆1 http://www.google.co.jp/
⋆2 http://aspell.net/
⋆3 http://sourceforge.jp/projects/ipadic/
⋆4 http://www.chokkan.org/software/simstring/

表 3 仮名から子音への変換
Table 3 Encoding table for consonants.

記	五十音	記	濁音	記	鼻・半濁音	記	特殊音
V	あいうえお (母音)					V	ー (長音)
k	かきくけこ	g	がぎぐげご	x	ガギグゲゴ		
s	さしすせそ	z	ざじずぜぞ				
t	たちつてと	d	だ でど			Q	っ(促音)
n	なにぬねの						
h	はひふへほ	b	ばびぶべぼ	p	ぱぴぷぺぽ		
m	まみむめも						
y	や ゆ よ					j	ゃゅょ(拗音)
r	らりるれろ						
w	わ					N	ん (撥音)

表 4 子音の調音様式
Table 4 Manner of articulation.

記	調音様式	子音
A	破裂音	p, b, t, d, k, g
B	鼻音など	m, n, x
C	はじき音	r
D	摩擦音など	h, s, z, (ち)(し)(じ)
E	接近音	y, w

表 5 子音の調音位置
Table 5 Place of articulation.

記	調音位置	子音と例外的な仮名
1	両唇音	p, b, m, w, (ふ)
2	歯音など	t, d, n, r, s, z
3	後部歯茎音など	(ち)(し)(じ)
4	硬口蓋音など	y, (に)(ひ)
5	軟口蓋音	k, g, x
6	声門音	h

4.1 提案法１：仮名から子音の記号列への変換

読み仮名をローマ字綴りに変換し，子音につく母音を削除し，子音の記号列で読みの特徴を表現する．ただし，(子音につかない) 単独の母音，長音，促音，拗音，撥音には子音とは別の文字列を割り当て区別できるようにする．具体的には，表 3 に示す変換表に従って，読み仮名を子音の記号列に変換する．表中の「記」は変換後の記号を表している．たとえば，漢字かな混じり文字列「高速増殖炉もんじゅ」の読み「こうそくぞうしょくろもんじゅ」は記号列「kVskzVsjkrmNzj」に変換される．変換された記号列から，文字 tri-gram を作成する．具体的には，記号列「kVskzVsjkrmNzj」から kVs, Vsk, skz, kzV, zVs, Vsj, sjk, jkr, krm, rmN, mNz, Nzj の 12 個の文字 tri-gram を作成する．記号列への変換と文字

Vol.2011-NL-201 No.14
Vol.2011-SLP-86 No.14
2011/5/17

tri-gram の作成をすべての入力文字列に対して行い，記号列を document，文字 tri-gram を索引語としてインデクシングを行う．検索を行う際は，検索文字列に対しても同様に記号列への変換と文字 tri-gram の作成を行い，検索文字列を query として，あらかじめインデクシングを行った document 集合 (文字列集合) に対して検索を行う．検索結果は，query に対する類似度の大きい document の順で順序付けして表示する．

4.2 提案法２：子音の特徴別分類

子音を調音様式と調音位置で分類し，分類に対応する記号で記号列を生成する．具体的には，まず，表 3 に示す変換表で仮名を子音の記号列に変換し，変換した記号列を入力として，さらに表 4 と表 5 に示す変換表に従って，2 種類の記号列に変換する．ただし，子音以外（母音，長音，促音，拗音，撥音）の記号には変換を行わず，そのまま変換後の記号列に含める．また，表 4 と表 5 において括弧書きした仮名に対しては例外的な処理を行うこととし，子音ではなく，仮名に対応する記号に変換する[⋆1]．たとえば，読み「こうそくぞうしょくろもんじゅ」を子音の記号列「kVskzVsjkrmNzj」に変換し，さらに調音様式を表す記号列「AVDADVDjACBNDj」と調音位置を表す記号列「5V252V3j521N3j」の 2 つの記号列に変換する．そして変換後の 2 種類の記号列からそれぞれ文字 tri-gram を作成する．具体的には，AVD, VDA, DAD, ADV, DVD, VDj, DjA, jAC, ACB, CBN, BND, NDj と 5V2, V25, 252, 52V, 2V3, V3j, 3j5, j52, 521, 21N, 1N3, N3j を作成し，これらを索引語としてインデクシングを行う．tri-gram を作成した後の提案法 2 の検索処理は提案法 1 の検索処理と同様である．

5. 評 価 実 験

提案法は，単語の意味ではなく，単語の読みの類似性を考慮して早口言葉の検索を行う点に特徴がある．提案法の有効性を検証するため，文献 10) に掲載されている滑舌訓練用の早口言葉を用いた評価実験を行った．文献 10) では，類似の音を含む早口言葉が子音のタイプ別に分類され，すべての早口言葉に読み仮名が振られている．各分類につき先頭から 5 個以上，最大 10 個までを，人手でテキストファイルに入力することとし，早口言葉 257 個を収集した．同じ分類の早口言葉は互いに類似性があるため，同じ分類の早口言葉を正解とし，1 個の早口言葉を query，残りの 256 個の早口言葉を検索対象の document 群として検索を行い，257 通りの検索を行った．検索評価には，検索システムの評価基盤としてよく用いられている Indri version 5.0[⋆2] を使用し，検索モデル Okapi BM25[⋆3]で検索を行うこととした．

提案法は，早口言葉の読みを入力とし，入力文字列を子音の特徴を考慮した特殊な記号体系に変換し，変換文字列から作成した N-gram を索引語とし，検索を行う．前節で述べた 2 種類の記号体系による変換を行い検索システムの性能を比較する．

- 提案法１：子音の記号列
- 提案法２：子音の特徴別分類

また，提案法による早口言葉検索の性能を比較評価するため，以下のような検索システムを作成し，提案法の比較対象とした．

- 従来法１：形態素インデックス
  漢字かな混じりの早口言葉の原文に対して，形態素解析器 ChaSen[⋆4]を用いた分かち書きを行い，形態素を索引語とした検索を行う．
- 従来法２：読みの類似文字列検索
  早口言葉の読み仮名から N-gram を作成し，作成した N-gram を索引語とした検索を行う．

検索結果の上位 10 件に着目し，再現率，精度，MAP 値 (Mean Average Precision)[11] を求めた．従来法の性能比較を表 6 に，提案法の性能比較を表 7 に示す．また，再現率・精度グラフを図 3 に示す．実験結果から，従来型の情報検索を早口言葉の検索にそのまま適用すると，検索性能が低くなることが分かる．漢字かな混じりの早口言葉の原文を検索するよりも，読み仮名に対して N-gram の文字列検索を行う方が，類似早口言葉の検索を検索しやすいと言える．N-gram の N を大きくすると，類似文字列が検索されにくくなり，MAP 値が低下する．提案法 2 の「子音の特徴別分類」は,「読み仮名の N-gram 検索」と比較して特に性能向上が見られなかった．この理由として，子音の調音様式と調音位置の 2 つの観点にわけて特徴を表現したため，子音の特徴が曖昧になってしまい，類似性の高い早口言葉が見分けられなくなってしまったということが考えられる．子音の記号列に対する 2-gram の検索は，従来法と比較して，類似早口言葉の検索性能向上に成功しているといえる．読

⋆1 具体的には調音様式については,「ち」「し」「じ」を記号 D，調音位置については,「ふ」を記号 1,「ち」「し」「じ」を記号 3,「に」「ひ」を記号 4 に変換する．

⋆2 http://www.lemurproject.org/
⋆3 Okapi BM25 のパラメータ設定はデフォルト値，すなわち，k1=1.2，b=0.75，k3=7 とした．
⋆4 http://chasen-legacy.sourceforge.jp/

Vol.2011-NL-201 No.14
Vol.2011-SLP-86 No.14
2011/5/17

み仮名で類似度を計算すると類似度が低くなる早口言葉であっても，子音以外の文字を抽象化することにより，子音の特徴が取りやすくなり，類似音を持つ早口言葉が検索されやすくなったと考えられる．長い早口言葉は，前半部分と後半部分で特徴が異なることもあるため，一定の長さになるように，検索前に分割し，長さを正規化しておくことを今後，検討する必要がある．

表 6 従来法の検索性能
Table 6 Evaluation of conventional methods.

索引語	MAP	精度	再現率
形態素	0.0593	0.0951	0.1016
読み (2-gram)	0.2409	0.2996	0.0573
読み (3-gram)	0.154	0.1927	0.1977
読み (4-gram)	0.1001	0.1183	0.1195

表 7 提案法の検索性能
Table 7 Evaluation of proposed methods.

索引語	MAP	精度	再現率
子音 (2-gram)	0.4397	0.4989	0.5262
子音 (3-gram)	0.3129	0.3835	0.4012
子音 (4-gram)	0.1786	0.2495	0.2693
調音 (2-gram)	0.1999	0.2629	0.2629
調音 (3-gram)	0.2265	0.3091	0.3234
調音 (4-gram)	0.1786	0.2495	0.2693

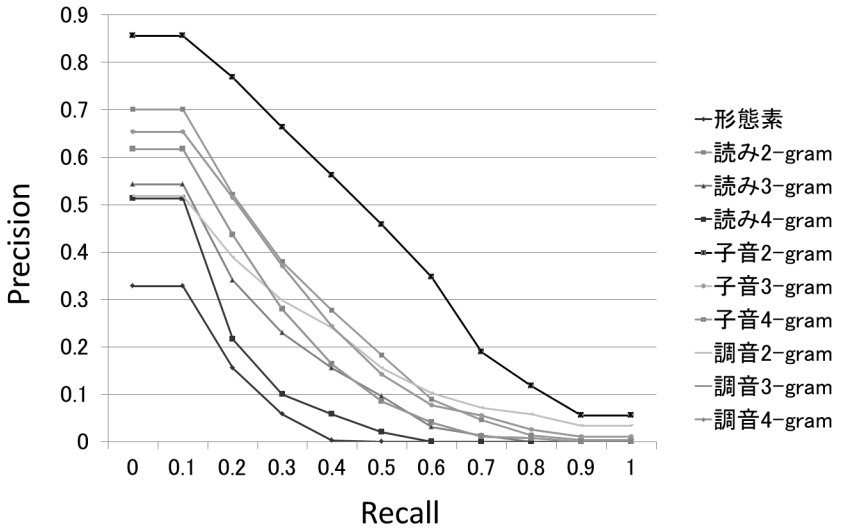

図 3 早口言葉検索システムの再現率・精度グラフ
Fig. 3 Recall-precision curves for tongue twister IR.

6. おわりに

滑舌訓練を支援することを目的として，多数の早口言葉の中から，類似の音を持つ早口言葉を検索する早口言葉検索の方法を提案した．漢字かな混じりの早口言葉の原文に対して，従来型の情報検索を適用すると，類似音を持つ早口言葉を見つけにくい．読み仮名に対して bi-gram で特徴を取り，類似文字列検索を行うことで，漢字かな混じりをそのまま検索する場合と比較して，早口言葉の検索性能が向上する．さらに，読み仮名をローマ字綴りに変換し，子音につく母音を削除して，子音の特徴を表す記号列に変換することで，類似音を持つ早口言葉の検索性能が大きく向上することが評価実験により確認された．長い早口言葉の特徴の取り方を工夫し，早口言葉の類似度計算を検索以外のアプリケーションにも適用することを今後検討していく予定である．

参 考 文 献

1) 橋本行洋：「カツゼツ (滑舌・活舌)」の語誌–近代の漢語受容と辞書，国語と国文学，Vol.82, No.12, pp.50–65 (2005).
2) 外池 滋生：言い間違いの言語学的意味，失語症研究，Vol.4, No.1, pp.537–541 (1984).
3) 寺尾 康：言語産出メカニズムの連続性について: 言い間違いからみた言語発達，ことばと文化，Vol.9, pp.115-131 (2006).
4) Hornby, A. S. et al.: Oxford Advanced Learner's Dictionary, Oxford University Press, 8th Ed. (2010).
5) トム・ディラン：「いいまちがい」大全集―外国人の日本語，IBC パブリッシング (2008).
6) 土井 晃一：自然な発話における言い間違いに関する考察，電子情報通信学会技術研究報告, 言語理解とコミュニケーション，Vo.95, No.169, pp.47–52 (1995).
7) Fromkin, V.A.: Slips of the tongue, Scientific American, Vol.226, No.6, pp.110–116 (1973).
8) フジテレビトリビア普及委員会：トリビアの泉 (6)，講談社, pp.131–132 (2004).
9) フジテレビトリビア普及委員会：トリビアの泉 (19)，講談社, pp.83–86 (2007).
10) 塩原 慎次郎：声を出して読む日本語の本―豊かな声をつくる早口ことばと滑舌例題集，創拓社 (1987).
11) Manning, C.D., Raghavan, P., and Schütze, H.: An Introduction to Information Retrieval, Cambridge University Press (2009).
12) 岡崎直観，辻井潤一：高速な類似文字列検索アルゴリズム，情報処理学会創立 50 周年記念全国大会，1C-1, (2010).