Vol.2017-SLP-117 No.11
2017/7/28

日本語におけるG2Pによる統計的学習を用いた話し言葉に頑健な発音辞書の自動構築

寺田 卓矢[1,a)]　李 晃伸[1,b)]

概要：本研究では話し言葉における発音の変化に着目し，G2P による統計的学習を用いて作成した発音辞書によって対応することを提案する．G2P の学習データとして辞書の標準読みに加えて誤りを含む認識結果を与えることにより，崩れた発音による認識誤りをカバーできる発音辞書を構築した．この発音辞書を用いた認識では，従来の辞書と比べ日本語話し言葉コーパスにおける認識率が最大で 1%弱向上した．

キーワード：音声認識，話し言葉，発音辞書，Grapheme-to-Phoneme，日本語話し言葉コーパス

1. はじめに

近年では音声認識技術の発展に伴い，音声検索はもちろん音声対話システムや議事録自動作成など様々な用途に音声入力が用いられるようになった．これら音声インターフェースは人が直接手で操作せずにすむため手間が少なく，益々の普及が期待されている．

しかし音声インターフェースの応用にあたって，音声認識における認識誤りが問題となる．認識誤りは雑音下での認識や，想定しているタスクドメインの外の発話など様々な原因が存在するが，その一つとして話し言葉が挙げられる [1]．Apple の Siri や Google 音声検索は実用に耐える認識率を示すとされているが，これは基本的に機械を意識して発話しているため，概念的，言語的，音響的に制約が存在する条件での音声認識であるといえ，この発話スタイルは読み上げ言葉に近い．一方話し言葉においては，これらの制約が存在せず，より自由度の高い発話スタイルとなる．このため話し言葉は読み上げ言葉に対し発音がくずれて変化しやすい．この発音の変化を柔軟に捉えることができれば，認識精度の向上が期待できる．

そこで本研究では，他言語で用いられている G2P (Grapheme-to-Phoneme) による統計的学習を日本語に適用し話し言葉に頑健な発音辞書を構築する．G2P とは書記素（文字列）からその読みである音素を推定する技術で，外国語で主にテキスト読み上げや音声認識で利用されている．これはアルファベット表記の言語において，書記素は音素と一対一対応しないという問題が存在するためである．一方日本語は英語のようにスペースで単語が分かち書きされておらず，単語の区切りとその読みや漢字を決定するために形態素解析の研究が進み，ここから得られるひらがなの読みを一対一のルールに基づき音素に変換していたため，G2P はほとんど重要視されてこなかった．しかし話し言葉のように発音がくずれ意図した語ではなくなる場合，ひらがなと音素は一対一対応しないため G2P を適用する意義があるといえる．よって音声認識結果の認識誤りと書き起こし（正解文）の対応をとり G2P による統計的学習を行い学習モデルに基づいた発音の変化を取り入れた辞書を構築することで，話し言葉に対して頑健な認識を目指す．

2. 話し言葉の音声認識

2.1 読み上げ言葉と話し言葉認識

音声認識はカーナビやパソコンソフト等で採用され，近年では Apple の Siri や Google 音声検索などに代表されるようにシステムの認識精度も向上し，実用的になってきている．ただしこれらはあくまで，人間が機械に向かって話すヒューマン・マシン・インタフェースとしての位置づけである．これらシステムを利用する際，意識せずともその会話スタイルは人間同士の会話や講義といった話し言葉とは異なったものになる．機械相手の発話では，話し言葉にはない以下のような制約が存在する [2]．

概念的制約

ユーザは動機を持ってシステムを利用するので，事前に内容を考えてから発話する．話し言葉では喋りなが

[1] 名古屋工業大学大学院 工学研究科
Nagoya Institute of Technology
[a)] terada@slp.nitech.ac.jp
[b)] ri@nitech.ac.jp

ら次の内容を考えている．

言語的制約

システムができる動作に制約があるため，コマンドのような文法的で単純な文を発話する．話し言葉には文法的誤りが存在し，また数語で終わるような短い文章はほとんど存在せず，数十語の長い文章になる．

音響的制約

目的が明確で発話も短いため，発音が明瞭である．話し言葉はそこまで発音が明瞭でなく，そもそも言い淀みや冗長な語が含まれ，発話した語自体が人手でも判断できず正解の定義が不明である発話も存在する．

上記のように機械相手の発話は制約が存在し，例えば外国に旅行に行った際，ホテルやレストランでの要求や情報提示などのコミュニケーションを行っている話し方に近い．しかし話し言葉にはこれらの制約は存在しないため，より認識が難しいと言える．

2.2 日本語話し言葉認識の研究

話し言葉の多様性の一つとして，言語的に同一の単語が異なって発音される，発音変動あるいは言語変動と呼ばれる現象がある [3][4]．こうした発音の変化には，音節や音素と言った音響モデルの個々の単位内部で発生する変動と，音素の置換等の単位間の変動が考えられる．前者の単位内部での変動については，音節や音素の音響モデルにおいて解決が図られ，発音の変化を考慮して HMM における状態・共有ガウス分布を修正する手法 [5][6][7] などが検討されている．一方，後者の単位間の変化については，単語の音素・音素列を規定する発音辞書において，標準的な発音に加えて実際にありえる発音を登録する手法 [8][9][10] が用いられる．

日本語では，話し言葉音声の諸相を包含した日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ)[11][12] が構築され，発音変化に関する研究でよく用いられている．音響モデルに関しては，実際に話し言葉の音声データを用いて学習することの効果が示されている [13][14]．また発話音素の単語内位置情報を用いた音響モデル [15] や近年では音響モデルに話者適応を含む DNN-HMM を用いることで，単語認識精度が更に改善した [16]．一方発音辞書についても先行研究 [13][17] が行われているが，CSJ の語彙に特化したモデル化になっており，CSJ のテストセットにしか事実上適応できない．ここから語彙に依存しない統計的な発音変化のモデル化を行った研究 [18] もされている．しかしこのモデルの学習は CSJ の音声に対する「ケイタイテキナブンセキヲモウシアゲマス」のような基本形と「ケータイテキナブンセキオモーシアゲマス」のような発音形とを対応させるもので，実際に起きているが人間が書き下した時に捉えられていない発音変化は含まれない．

表 1 読み上げ言葉と話し言葉の認識率 （文献 [1] より）

認識データ	単語認識精度
読み上げ言葉（JNAS）	96.2%
話し言葉（CSJ）	82.5%

現在の読み上げ言葉と話し言葉における認識率は表 1 である [1]．話し言葉は読み上げ言葉と比べ，認識率が 10%以上低下するが，この差は音響モデル内では対応できない発音変化によるものだと考えられる．

3. G2P による話し言葉に頑健な発音辞書の自動構築

3.1 G2P

G2P とは Grapheme-to-Phoneme Conversion の略であり，書記素（文字）からその読みである音素を推定する技術である．例えば英単語の「Apple」と「April」では同じ書記素「A」を持つが，音素は「AE P AH L」と「EY P R AH L」であり異なった発音になる．G2P は統計的学習により書記素と音素の関係を学習し，ある書記素に対する尤もらしい読みを出力できる．このため G2P は音声認識や音声合成の分野で利用されている．書記素の集合を G，音素の集合を Φ，$*$ をクリーネ閉包とすると，$\boldsymbol{g}$ はある書記素，$\boldsymbol{\varphi}'$ は考えられるある音素の組み合わせ列であり，ある書記素に対する最も尤もらしい読みの推定は式 (1) のように計算できる．

$$\boldsymbol{\varphi}(\boldsymbol{g}) = \arg\max_{\boldsymbol{\varphi}' \in \Phi^*} p(\boldsymbol{g}, \boldsymbol{\varphi}') \tag{1}$$

3.2 日本語と G2P

日本語はアルファベット表記の言語と異なるシステムであるため，日本語では G2P はほとんど利用されていない．英語に代表される外国語では，単語がスペースで分かち書きされているため，解析は比較的容易である．一方日本語は漢字かな交じりの表記をし，単語の区切れを分かち書きのように明示しない．よって単語の区切りを決定するために，形態素解析の研究がなされてきた．形態素解析では語の区切りとその読みが付与される．例えばある文脈で出現した「今日」という語の区切りや読み方は複数考えられるが，「コンニチ」,「イマ・ヒ」ではなく「キョウ」という読みだと判定できる．一度読みが付与されると，音素列は図 1 のように変換ルールで得ることができる．G2P が日本語に用いられないのは，形態素解析によってカナが実用的なレベルで得られ，かつカナが音素と一対一対応するされているからだと考えられる．また実際の話し言葉に忠実で多様な書き起こしが多く存在せず，G2P の学習データが揃えられないということも言える．話し言葉をリアルに捉えていない綺麗な書き起こしとその音素列を G2P の学習データとしても，カナと音素の一対一変換ルールと同じものが学習されるだけだからである．

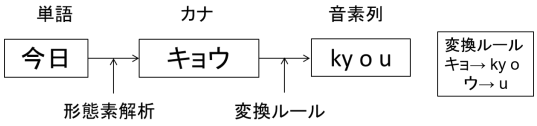

図 1 変換ルールによる音素列の取得

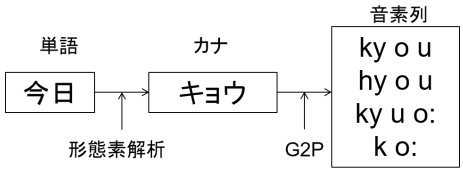

図 2 G2P による発音の変化を考慮した音素列の取得

3.3 認識誤りを利用した G2P の学習

本論文では認識誤りを実際の発音に近いデータとみなし，書き起こしと認識誤りを対応させて発音の変化を G2P で学習する枠組みを提案する．実際の発音の変化を音素単位で記述したデータを学習に利用するのが理想的であるが，そのような書き起こしの作成は大変な労力がかかり，大規模なものは存在しない．人が解釈し書き起こした CSJ の発音形よりも，認識誤りはリアルな発音に近いデータと考えられるのでこれを利用する．G2P によって発音の変化を考慮した音素列の取得は図 2 のようになる．以下，学習の手順について述べる．

図 3 は，G2P の学習に必要なデータを取り出すための認識結果と書き起こしの対応付けの手順である．まず認識結果は単語単位で分かち書きされているので，そこから DP マッチで単語ごとのアラインメントを行う．REF は書き起こし，HYP は認識結果，EVAL は DP マッチによる評価を表す．通常の発音を学習するために正解語部分を取り出し，発音の変化を学習するために置換エラー部分を取り出す．ここで図 4 のように認識結果と書き起こしで区切りが異なり，単純には取り出せない場合も存在する．この例で単純な置換エラーの取り出しを行うと「四」と「音声」，「正」と「生徒」，「と」と「不」という対応付けがなされるが，これは発音の変化によって生じたものではない．確実に語の区切りが正しいと判断できる正解語に挟まれた置換エラーのみを取り出す．こうして対応付けられた語の認識結果側を一般的な日本語のカナと音素の変換ルールによって変換し，この書記素と音素のペアを G2P の学習データとする．

漢字かな混じりのデータを用いた場合，同音異義語の誤りは認識誤りとみなされるが，これは発音の変化によって生じたエラーではない．よって今回の学習データは漢字かな混じりのデータでなく，カナのデータを用いる．

ここで認識誤りを学習データとするため，表記は音素単位であるが，従来の日本語のモデルから大きく逸脱するような音素の変化は捉えられず，音節単位で変化を捉えることには留意が必要である．これは「キョウ」が実際には「ky r:」という発音をされていても，日本語には「ky」や「r」に対応するカナが存在しないため，学習データに表されないからである．

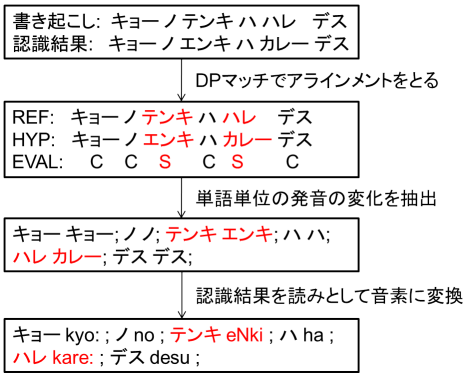

図 3 認識結果と書き起こしの対応付け

REF: 第 四 声 に つ い て は 正 と 負 の 指 令 で
HYP: 大 音声 に つ い て は 生徒 不 の 指 令 で
EVAL:S S D C C C C S S D C C C

図 4 認識結果と書き起こしで文節が異なる場合

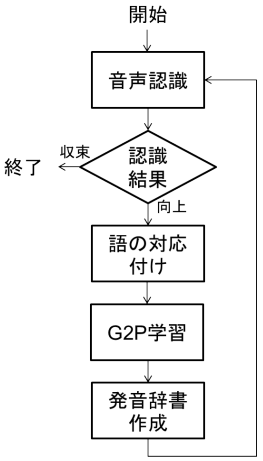

図 5 G2P の繰り返し学習（システム構成）

3.4 G2P の繰り返し学習と発音辞書の更新

少量の書き起こしデータから G2P によって自動的に音声認識用の大規模な発音辞書を作成する研究がなされている [19]．この論文では G2P のモデルを繰り返し学習することにより更新を行っている．本研究においても学習と音声認識は交互に行え，繰り返すごとに抽出できる語が増えるため，G2P の繰り返し学習と発音辞書の更新を提案する．G2P の繰り返し学習の過程を図 5 に示す．

まず音声認識を行い，認識率を算出する．ここで認識率が前回と比べて向上すれば，さらに上がる可能性があるため語の対応付けに進む．一方認識率の向上が見られず収束した場合は，このシステムで捉えられる発音の変化をすべて辞書に取り込んだとみなし，終了する．語の対応付けを

単語	候補	確率	読み
あ	0	0.837369	a
あ	1	0.095211	a:
あ	2	0.028033	a
あ	3	0.004017	w a

→

単語	候補	確率	読み
あ	0	0.865402	a
あ	1	0.095211	a:

図 6 読みの候補のマージと足切り

```
あ+ア+動詞 @-0.0628 [あ] a_S
あ+ア+動詞 @-1.0213 [あ] a:_S
```

図 7 発音辞書の例

行ったらそれを学習データとして G2P の学習を行う．ここで作成した G2P のモデルに元々の発音辞書のエントリを入力し，その語の読みを複数推定する．ここで得られる読みの候補に，発音の変化を考慮したものが出現する．これを元に G2P による新規の発音辞書を作成し，元々の辞書と差し替え再び音声認識を行う．このサイクルを認識率が収束するまで続ける．これにより一回では捉えきれなかった発音の変化も辞書に取り込むことを目指す．

3.5 発音辞書作成の実装

実際の G2P の学習・推定はジョイントシーケンスモデル [20] に基づき計算を行う Sequitur G2P[21] を用いる．この時得られる読みの候補はオプションで指定した数だけ，確率が高い順に出力される．この読みの候補を N-best と呼ぶ．N には任意の値が指定できるが，実際に動作させたところ，N-best に全く同じ読みが複数出現することが観測された．また出力される候補の確率が低いものも多いため，Sequitur G2P から得られた読みの候補の中から同じ読みのものをマージし，その後一定の足切り確率以下の候補を辞書に追加せず打ち切ることとした．この足切り確率は $prob$ と定義し任意の値を割り当てる．図 6 は単語「あ」に対して $N = 4$, $prob = 0.01$ でマージと足切りを行った例である．足切りを行わない場合，考えられる読みの変化に引きずられすぎて認識率が低下することが考えられる．こうして得た読みの候補を元に，確率の常用対数をとり図 7 のような形式に変換する．この発音辞書はそれぞれ言語モデル用エントリ，各読みの出現確率，認識結果用表記，音素列からなる．

G2P の学習に用いたデータの詳細を図 8 に示す．認識結果から抽出したデータのみでは元々の発音辞書のエントリに含まれている書記素を網羅できず，すべての読みを出力することができないため，元々の辞書データも学習データとして追加している．また辞書のデータをひらがなに直した時，同音異義語は同じエントリとして出現回数に差が生じるため，統合を行った．

4. 評価実験

4.1 実験条件

提案手法の有効性を検証するために，CSJ の評価

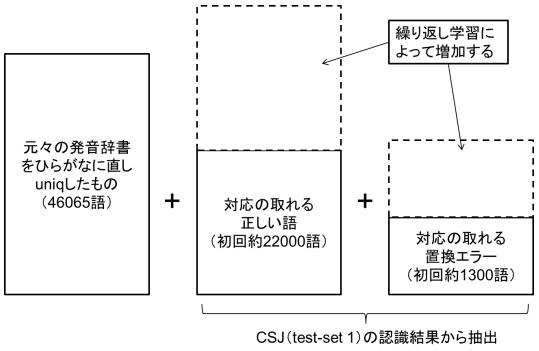

図 8 G2P の学習データ

表 2 評価実験の実験条件

評価基準	単語認識精度 WAcc（第 2 パス）
テストセット	CSJ test-set1 男性話者 10 講演 2907 ファイル
音響モデル	CSJ 講演 233 時間から DNN を学習 入力層：20×11=1320 ノード 中間層：2048 ノード ×7 層 出力層：HMM の状態数 4829 ノード
言語モデル	CSJ 学習データ 7315487 語から学習 第 1 パス 前向き 2-gram 第 2 パス 後ろ向き 3-gram
初期発音辞書	57553 語
デフォルト N-best	5
デフォルト $prob$	0.10

表 3 Julius の実験条件

バージョン	Julius rev.4.4.2(standard) –enabele-words-int
ビーム幅（第 1/2 パス）	2000/150
言語重みスコア（第 1/2 パス）	9/12
挿入ペナルティ（第 1/2 パス）	-3/-3
探索仮設文数	10
スタックサイズ	20000
第 2 パススコア幅	80
第 2 パス仮説展開しきい値	20000
第 2 パスフレーム範囲	5
その他オプション	-fallback1pass/-nostrip

セットを用いて評価実験を行った．G2P の学習には SequiturG2P[20][21], 音声認識デコーダとして Julius[22][23], を用いた．その他詳細な実験条件は表 2, 表 3 の通りである．この条件のもとで G2P の適用の有無，繰り返し回数，N-best 数，足切り確率 $prob$ についてパラメータを変化させ，認識率の変化を検証した．ここで用いる認識率とは，第 2 パスの単語認識精度 (Word Accuracy; WAcc) のことを指す．出現する単語の総数を N，置換エラーを S，削除エラーを D，挿入エラーを I，単語誤り率を WER(Word Error Rate) とすると WAcc は以下のように計算できる．

$$WAcc = 1 - WER = \frac{N - S - D - I}{N}$$

表 4 G2P の繰り返し学習による認識率 (%) 比較

G2P 繰り返し学習回数	認識率	置換エラー	辞書エントリ
ベースライン	82.88	10.36	57553
1	83.15	10.08	58075
2	83.01	10.11	58850
3	82.98	10.16	59177

表 5 N-best 数による認識率 (%) 比較

N-best	認識率	置換エラー	辞書エントリ
3	83.15	10.08	58066
5	83.15	10.08	58075
10	83.15	10.08	58075

4.2 実験結果

提案する認識誤りを用いた G2P の繰り返し学習を行った結果を表 4 に示す．この時 $N = 5, prob = 0.10$ で実験を行っている．G2P を適用しない通常の場合では認識率は 82.88%であるが，G2P を適用した場合は 83.15%と 0.27%上昇している．エラーを種類別に見ると置換エラーが顕著に減少した．これは増加した約 500 の辞書エントリが，起こりうる発音の変化を捉えたと言える．しかし認識結果を用い G2P の学習を繰り返すと，認識率はベースラインよりも高いものの，1 回目よりも低下した．これは 1 回目の学習で十分に実際に起こる発音の多くを拾い上げたということ，その後の学習では回復できなかった発音の変化に引きずられすぎて既存の発音を不適切に更新してしまうものがあるということが考えられる．この結果から今回の学習データでは G2P の繰り返し学習は 1 回で十分だとして，パラメータによる認識率の比較を行った．

次に G2P から出力する読みの候補 N-best 数をデフォルトの 5 から上下させて実験した結果を表 5 に示す．この時 $prob = 0.10$ で固定して実験を行っている．結果を見ると，どの場合でも認識率は変化しなかった．これは N-best が 3 つもあれば多くの場合であり得る発音の変化は出力しきり，残りは非常に確率の低い読み方ばかりが出力されるため，足切り操作と合わせてほぼ辞書に影響を及ぼさないからだと言える．

最後に足切り確率 $prob$ をデフォルトの 0.10 から上下させて実験した結果を表 6 に示す．この時 N-best=5 で固定して実験を行っている．$prob = 0.01$ までは足切り確率を下げることで認識率は向上している．しかしこれより足切り確率を下げても，辞書エントリは大幅に増加するものの認識率はほとんど変化していない．足切りはどこかで考えられうる読みの変化に引きづられすぎて認識率が悪化することを懸念し設定したが，足切りを行わずとも認識率は悪化しなかった．一方 $prob = 0.01$ 以下のエントリは多数存在するものの認識率の改善にも大きく寄与しなかった．

ここで G2P の適用によって具体的にどのように認識が

表 6 足切り確率 $prob$ による認識率 (%) 比較

$prob$	認識率	置換エラー	辞書エントリ
なし	83.69	9.68	281246
0.0001	83.69	9.68	234530
0.001	83.68	9.70	112076
0.005	83.63	9.75	73286
0.01	83.64	9.77	65647
0.05	83.29	9.98	58964
0.10	83.15	10.08	58075
0.20	83.03	10.17	57747

表 7 G2P 適用により改善した認識誤りとその回数の例

誤りパターン	G2P なし	G2P あり
コノ → ノ	32	13
マ → ア	33	14
コンカイ → オンカイ	4	0
モデル → デル	2	0
デ → エ	28	24
ンー → ン	13	8
コレ → オレ	11	0
タカイ → アカイ	4	2

表 8 G2P 適用により悪化した認識誤りとその回数の例

誤りパターン	G2P なし	G2P あり
エ → エー	45	47
エー → エ	38	39
マー → マ	32	36
ワ → ガ	20	23
ノ → ナ	3	4
マトメ → アトメ	0	1
ケレド → ケド	14	17

どのように改善したかを考察する．表 7，8 はベースラインと，N-best=5, 足切りなしの場合との具体的な認識誤りの比較である．改善した認識誤りでは特に語頭の音素が抜け落ちてしまっていた単語が正しく認識されるようになったのが目立つ．例えば「今回」は音素表記では「k o n k a i」であるが，実際は語頭の「k」がほとんど発音されず「o n k a i」に近い発音だったとしてもその変化を拾い上げ，「音階」ではなく「今回」と正しく認識されるようになったのが 4 回存在する．発音辞書のエントリは図 9 のようになった．

ただ認識結果が悪化した例も存在する．極端に認識誤りが増える語はほとんどなく，全体的に少量ずつ認識誤りが広がっているように見える．総数として新たな認識誤りより，回復した語が多いため認識率が向上したと言える．

5. むすび

本研究では，話し言葉で発生しやすい発音の変化を捉えるために認識誤りを利用した G2P の繰り返し学習を提案し，これにより作成した発音辞書を用いることにより認識率の向上を試みた．評価実験の結果，従来の辞書と比較し

- G2P なし
  今回+コンカイ+名詞 [今回] k_B o_I N_I k_I a_I i_E
- G2P あり
  今回+コンカイ+名詞 @-0.0629 [今回] k_B o_I N_I k_I a_I i_E
  今回+コンカイ+名詞 @-0.8726 [今回] o_B N_I k_I a_I i_E

図 9 G2P の適用の有無による発音辞書の変化の例

て最高で 1%弱の認識率の向上を達成した．日本語はひらがなから単純な変換ルールで音素列を得るだけで十分だとされていたが，発音の変化を考慮した場合，G2P を適用する意義があると言える．また認識誤りを発音の変化を捉えたデータと見なして用いることの有効性も確認された．

ただしこの手法は音響モデルを信頼しており，音響モデルがくずれた発音通りの出力をするという前提である．実際は話し言葉による認識誤りの他に音声認識システムの誤り傾向を G2P のモデルに含んでおり，発音の変化と音響モデルの誤りが区別されていない点は議論の余地がある．

今後の課題としては，より大量の話し言葉を含むデータセットでのオープンな認識実験や，パラメータと辞書作成の手法の吟味などが挙げられる．近年では，G2P の学習に Sequitur G2P で用いているジョイントシーケンスモデルでなく，重み付き有限状態トランスデューサとリカレントニューラルネットワーク言語モデルを G2P 学習・推定に用いる研究 [24] が存在するため，こうしたモデルに基づく G2P の学習も検討したい．さらに今回の提案手法では発音の変化を単語単位で処理しているが，実際の発音の変化は前後の単語や文章全体といったより大きな単位で影響を受けていると考えられる．よって単語の単位を超えた発音の変化のモデル化を考えていくのも課題である．

参考文献

[1] 河原達也：音声認識の方法論に関する考察ー世代交代に向けてー, 情報処理学会研究報告 IPSJ SIG Technical Report, Vol.2014-SLP-100, No.3, 2014-1-31
[2] 河原達也：話し言葉の音声認識の進展ー議会の会議録作成から講演・講義の字幕付与へー, メディア教育研究, Vol.9, No.1, pp.1-8, 2012
[3] 前川喜久雄, 小磯花絵, 菊池英明, 間淵陽子, 斎藤美紀：「日本語話し言葉コーパス」に捉えられた言語変異現象, 国立国語研究所公開研究発表資料, pp.41-42, 2003
[4] H. Strik and C. Cucchiarini: Modeling pronunciation variation for ASR, A survey of the literature, Speech Commun., Vol.29, pp.225-246, 1999
[5] M. Saraclar, H.Nock, and S. Khudanqur: Pronunciation modeling by sharing Gaussian Densities across phonetic models, Comput. Speech Lang., Vol.14, pp.137-160, 2000
[6] P. Kam, T. Lee, F.K. Soong: Modeling cantonese pronunciation variation by acoustic model refinement, Proc. Eurospeech, Vol.2, pp.1477-1480, 2003
[7] H. Yu and T. Schultz, Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition, Proc. Eurospeech, Vol.3, pp.1869-1872, 2003
[8] L. Lamel and G. Adda: On designing pronunciation lexicons for large vocabulary, continuous speech recognition, Proc. ICSLP, Vol.1, pp.6-9, 1996
[9] T. Sloboda and A. Waibel: Dictionary learning for spontaneous speech recognition, Proc. ICSLP, Vol.4, pp.2328-2331, 1996
[10] T. Imai, A. Ando, and E. Miyasaka: A new method for automatic generation of speaker-dependent phonological rules, Proc. ICASSP, Vol.1, pp.864-867, 1995
[11] 前川喜久雄, 籠宮隆之, 小磯花絵, 小椋秀樹, 菊池英明：日本語話し言葉コーパスの設計, 日本音声学会 音声研究, Vol.4, No.2, pp.51-61, 2000-8
[12] 日本語話し言葉コーパス (CSJ) http://pj.ninjal.ac.jp/corpus_center/csj/
[13] H. Nanjo and T. Kawahara: Language model and speaking rate adaptation for spontaneous presentation speech recognition, IEEE Trans. Speech Audio Process., Vol.12, No.4, pp.391-400, 2004
[14] 篠崎隆宏, 古井貞熙：日本語話し言葉コーパスを用いた講演音声認識, 情報処理学会論文誌, Vol.43, No.7, pp.2098-2107, 2002
[15] 五十川賢造, 篠田浩一, 嵯峨山茂樹：形態素情報と単語内位置情報を用いた話し言葉音声認識のための音響モデル, 電子情報通信学会技術研究報告, Vol.102, No.529, pp.111-116, 2002-12
[16] 三村正人, 河原達也：CSJ を用いた日本語講演音声認識への DNN-HMM の適用と話者適応の検討情報処理学会研究報告 Vol.2013-SLP-97, No.9, pp.1-6, 2013-7
[17] 堤怜介, 加藤正治, 小坂哲夫, 好田正紀：発音変形依存と教師なし適応による講演音声認識の性能改善, 話し言葉の科学と工学ワークショップ講演予稿集, pp93-98, 2004
[18] 秋田祐哉, 河原達也：話し言葉のための汎用的な統計的発音変動モデル電子情報通信学会論文誌, Vol.J88-D-II, No.9, pp.1780-1789, 2005
[19] Loang Lu, Arnab Ghoshal, and Steve Renals: Acoustic Data-Driven Pronunciation Lexicon For Large Vocabulary Speech Recognition, Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on 2013
[20] Maximilian Bisani and Hermann Ney: Joint-Sequence Models for Grapheme-to-Phoneme Conversion, Speech Communication, Vol.50, No.5, pp.434-451, 2008-5
[21] Sequitur G2P - A trainable Grapheme-to-Phoneme converter https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html
[22] 河原達也, 李晃伸：連続音声認識ソフトウェア Julius, 人工知能学会誌, Vol.20, No.1, pp.41-49, 2005
[23] Julius - Open-Source Large Vocabulary Continuous Speech Recognition Engine https://github.com/julius-speech/julius
[24] Josef R. Novak, Paul R. Dixon, Nobuaki Minematsu, Keikichi Hirose, Chiori Hori, and Hideki Kashioka: Improving WFST-based G2P with Alignment Constraints and RNNLM N-best Rescoring, Proc. Interspeech Conference of the International Speech Communication Association, Portland, Oregon, USA, pp.2526-2529, 2012-9