(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5327054号
(P5327054)

(45)発行日　平成25年10月30日(2013.10.30)　(24)登録日　平成25年8月2日(2013.8.2)

(51)Int.Cl.　　F I
G10L 15/18 (2013.01)　G10L 15/18 302H
G10L 15/10 (2006.01)　G10L 15/10 302G
G10L 15/065 (2013.01)　G10L 15/06 310Z

請求項の数 12　(全 16 頁)

(21)出願番号　特願2009-546202 (P2009-546202)
(86)(22)出願日　平成20年11月27日(2008.11.27)
(86)国際出願番号　PCT/JP2008/071500
(87)国際公開番号　WO2009/078256
(87)国際公開日　平成21年6月25日(2009.6.25)
審査請求日　平成23年9月8日(2011.9.8)
(31)優先権主張番号　特願2007-326132 (P2007-326132)
(32)優先日　平成19年12月18日(2007.12.18)
(33)優先権主張国　日本国(JP)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100102864
弁理士　工藤　実
(72)発明者　越仲　孝文
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　菊池　智紀

最終頁に続く

(54)【発明の名称】発音変動規則抽出装置、発音変動規則抽出方法、および発音変動規則抽出用プログラム

(57)【特許請求の範囲】
【請求項１】
音声データを記憶する音声データ記憶手段と、
前記音声データの標準形発音を表す標準形発音データを記憶する標準形発音記憶手段と、
前記標準形発音データからサブワード言語モデルを生成するサブワード言語モデル生成手段と、
前記サブワード言語モデルを用いて前記音声データを認識する音声認識手段と、
前記音声認識手段が出力する認識結果と、前記標準形発音データとを比較して、これらの差分を抽出する差分抽出手段と、
前記サブワード言語モデルの重み値を制御する言語モデル重み制御手段を
備え、
前記言語モデル重み制御手段は、
複数の重み値を出力し、
前記音声認識手段は、
前記複数の重み値のそれぞれについて、前記音声データを認識し、
前記言語モデル重み制御手段は、
前記重み値を制御する際、あらかじめ定められた値の組に基づいて、所定の回数だけ前記重み値を更新し、
前記言語モデル重み制御手段は、

前記差分に応じて、前記重み値を更新することの有無を随時決定する
発音変動規則抽出装置。
【請求項２】
前記言語モデル重み制御手段は、
前記差分が所定のしきい値よりも小さい場合に、前記重み値を減少させるように更新する
請求項１記載の発音変動規則抽出装置。
【請求項３】
前記言語モデル重み制御手段は、
前記差分が所定のしきい値よりも大きい場合に、前記重み値を増加させるように更新する
請求項２記載の発音変動規則抽出装置。
【請求項４】
前記差分抽出手段は、
前記差分を、前記認識結果と前記標準形発音データとの間の編集距離として計算する
請求項１乃至３のいずれかに記載の発音変動規則抽出装置。
【請求項５】
前記差分抽出手段は、
前記差分として、前記認識結果と前記標準形発音データとの相違箇所の文字列対と、前記認識結果が得られたときに前記音声認識手段が前記言語モデル重み制御手段から受け取ったサブワード言語モデルの重み値とを含む発音変動事例を抽出する
請求項１乃至４のいずれかに記載の発音変動規則抽出装置。
【請求項６】
前記発音変動事例から発音変動の確率的規則を生成する発音変動確率推定手段を更に備える
請求項５記載の発音変動規則抽出装置。
【請求項７】
前記発音変動確率推定手段は、
ある発音変動事例が観測されたときのサブワード言語モデルの重み値の大きさに応じて、前記ある発音変動事例の発現確率が高くなるように前記発音変動の確率的規則を生成する
請求項６記載の発音変動規則抽出装置。
【請求項８】
音声データの標準形発音を表す標準形発音データを記憶することと、
前記標準形発音データからサブワード言語モデルを生成することと、
前記サブワード言語モデルを用いて前記音声データを認識することと、
前記認識することによる認識結果と、前記標準形発音データとを比較して、これらの差分を抽出することと、
前記サブワード言語モデルの重み値を制御することと
を具備し、
前記制御することは、
複数の重み値を出力することを含み、
前記認識することは、
前記複数の重み値のそれぞれについて、前記音声データを認識することを含み
前記制御することは、
前記重み値を制御する際、あらかじめ定められた値の組に基づいて、所定の回数だけ前記重み値を更新することと、
前記差分に応じて、前記重み値を更新することの有無を随時決定することと、
前記差分が所定のしきい値よりも小さい場合に、前記重み値を減少させるように更新することと、

前記差分が所定のしきい値よりも大きい場合に、前記重み値を増加させるように更新することとを更に含む
発音変動規則抽出方法。
【請求項９】
前記抽出することは、
前記差分を、前記認識結果と前記標準形発音データとの間の編集距離として計算することと、
前記差分として、前記認識結果と前記標準形発音データとの相違箇所の文字列対と、前記認識結果が得られたときに受け取ったサブワード言語モデルの重み値とを含む発音変動事例を抽出することとを含む
請求項８記載の発音変動規則抽出方法。
【請求項１０】
前記発音変動事例から発音変動の確率的規則を生成することを更に具備し、
前記確率的規則を生成することは、
ある発音変動事例が観測されたときのサブワード言語モデルの重み値の大きさに応じて、前記ある発音変動事例の発現確率が高くなるように前記発音変動の確率的規則を生成することを含む
請求項９記載の発音変動規則抽出方法。
【請求項１１】
音声データを記憶する音声データ記憶手段と、
前記音声データの標準形発音を表す標準形発音データを記憶する標準形発音記憶手段と、
前記標準形発音データからサブワード言語モデルを生成するサブワード言語モデル生成手段と、
前記サブワード言語モデルを用いて前記音声データを認識する音声認識手段と、
前記音声認識手段が出力する認識結果と、前記標準形発音データとを比較して、これらの差分を抽出する差分抽出手段と、
前記サブワード言語モデルの重み値を制御する言語モデル重み制御手段と、
としてコンピュータを機能させるための
発音変動規則抽出用プログラムであって、
前記言語モデル重み制御手段は、
複数の重み値を出力し、
前記音声認識手段は、
前記複数の重み値のそれぞれについて、前記音声データを認識し、
前記言語モデル重み制御手段は、
前記重み値を制御する際、あらかじめ定められた値の組に基づいて、所定の回数だけ前記重み値を更新し、
前記差分に応じて、前記重み値を更新することの有無を随時決定し、
前記差分が所定のしきい値よりも小さい場合に、前記重み値を減少させるように更新し、
前記差分が所定のしきい値よりも大きい場合に、前記重み値を増加させるように更新する
発音変動規則抽出用プログラム。
【請求項１２】
前記差分抽出手段は、
前記差分を、前記認識結果と前記標準形発音データとの間の編集距離として計算し、
前記差分として、前記認識結果と前記標準形発音データとの相違箇所の文字列対と、前記認識結果が得られたときに前記音声認識手段が前記言語モデル重み制御手段から受け取ったサブワード言語モデルの重み値とを含む発音変動事例を抽出し、
前記発音変動事例から発音変動の確率的規則を生成する発音変動確率推定手段を更に備

え、
前記発音変動確率推定手段は、
ある発音変動事例が観測されたときのサブワード言語モデルの重み値の大きさに応じて、前記ある発音変動事例の発現確率が高くなるように前記発音変動の確率的規則を生成する
請求項１１記載の発音変動規則抽出用プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、発音変動規則抽出装置、発音変動規則抽出方法、および発音変動規則抽出用プログラムに関し、特に、対応する書き起こしテキストが付随した音声データ等から、自由な話し言葉によく現れる発音変動の規則を抽出することができる発音変動規則抽出装置、発音変動規則抽出方法、および発音変動規則抽出用プログラムに関する。
【背景技術】
【０００２】
『堤、加藤、小坂、好田著「発音変形依存モデルを用いた講演音声認識」電子情報通信学会論文誌、第Ｊ８９－Ｄ巻、２号、３０５～３１３頁、２００６年』、『秋田、河原著「話し言葉音声認識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、第Ｊ８８－Ｄ２巻、９号、１７８０～１７８９頁、２００５年』に、発音変動規則抽出装置の一例が記載されている。図１に示すように、この発音変動規則抽出装置２００は、標準形発音記憶手段２０１と、変形発音記憶手段２０２と、差分抽出手段２０３と、発音変動計数手段２０４とから構成されている。
【０００３】
このような構成を有する発音変動規則抽出装置２００は次のように動作する。すなわち、差分抽出部２０３は、標準形発音記憶部２０１および変形発音記憶部２０２からそれぞれ書き起こしテキストを抽出し、差分、すなわち相違箇所を抽出する。
【０００４】
ここで標準形発音記憶部２０１および変形発音記憶部２０２には、長時間の音声データの発音内容を書き起こした結果である書き起こしテキストが記憶されている。より具体的には、標準形発音記憶部２０１には、例えば以下のような書き起こしテキストが格納されている。
その　ような　しゅじゅつ　を　ほぼ　まいにち　おこない　まし　た
【０００５】
また、変形発音記憶部２０２には、標準形発音記憶部２０１に記憶された書き起こしテキストと対応する形で、例えば以下のような書き起こしテキストが格納されている。
その　ような　しじつ　を　ほぼ　まいんち　おこない　まし　た
【０００６】
標準形発音記憶部２０１には、元となった音声データの標準的な発音、つまり正しく発音された場合に観測されるべき本来の発音が書き起こしテキストとして記憶されている。一方、変形発音記憶部２０２には、音声データを実際に人が聞いて、聞こえるままの発音を忠実に書き起こした書き起こしテキストが記憶されている。上の例では、標準形発音「しゅじゅつ（手術）」、「まいにち（毎日）」に対して、それぞれ「しじつ」、「まいんち」という変形発音が記憶されている。
【０００７】
差分抽出部２０３は、標準形の書き起こしテキストと変形の書き起こしテキストとを比較し、相違する箇所の文字列対を抽出する。上の例では、「しゅじゅつ」と「しじつ」、「まいにち」と「まいんち」が抽出される。以下、これらの対を発音変動事例と呼ぶ。また、標準形発音と変形発音が等しい、すなわち変形がない場合の発音変動事例を特に、恒等発音変動と呼ぶことにする。
【０００８】

発音変動計数部２０４は、差分抽出部２０３から発音変動事例を受け取り、同じ標準形、同じ変形ごとに分類し、恒等発音変動も含めて観測回数を計数する。さらに、計数結果を正規化して確率値に変換する。例えば、上の例で標準形発音「まいにち」に対する変形発音として「まいにち（恒等変形）」、「まいんち」、「まいち」、「まんいち」があり、それぞれ９６６回、１１２回、１３回、２回観測されたとする。標準形発音「まいにち」の観測回数は９６６＋１１２＋１３＋２＝１０９３であるから、確率値に変換すると、

まいにち→まいにち　０．８８４（９６６／１０９３）
まいにち→まいんち　０．１０２（１１２／１０９３）
まいにち→まいち　０．０１２（１３／１０９３）
まいにち→まんいち　０．００２（２／１０９３）

となる。この結果は、標準形発音「まいにち」に対する変形発音の出現傾向に関する確率的な規則と解釈できる。発音変動計数部２０４は、上記結果を発音変動規則として出力する。

【０００９】

なお、上の例では、標準形発音や変形発音を単語単位で扱っているが、他の単位、例えば所定の長さの音素（母音、子音等、音声を構成する最小単位）の系列として扱うことも可能である。また、上記確率値を計算する際に、適当な平滑化操作、例えば観測回数が所定値に満たない特殊な発音変動規則を無視する等を行ってもよい。

【００１０】

『緒方、有木著「発音変形と音響的誤り傾向を考慮した話し言葉音声認識の検討」日本音響学会２００３年春季研究発表会講演論文集、９～１０頁、２００３年３月』、『緒方、後藤、浅野著「話し言葉音声認識のための動的発音モデリング法の検討」日本音響学会２００４年春季研究発表会講演論文集、２０３～２０４頁、２００４年３月』に、発音変動規則抽出装置の別の一例が記載されている。図２に示すように、この発音変動規則抽出装置３００は、音声データ記憶部３０１と、標準形発音記憶部３０２と、音節辞書記憶部３０３と、音響モデル記憶部３０４と、音声認識部３０５と、差分抽出部３０６と、発音変動計数部３０７とから構成されている。

【００１１】

このような構成を有する発音変動規則抽出装置３００は次のように動作する。すなわち、音声認識部３０５は、音節辞書記憶部３０３に記憶された辞書、および音響モデル記憶部３０４に記憶された音響モデルを用いて、音声データ記憶部３０１に記憶された音声データに対して、公知の連続音節認識処理を行い、認識結果の音節系列を出力する。

【００１２】

ここで、音節辞書記憶部３０３に記憶された辞書は、日本語の場合、あ、い、う、え、お、か、き、く、け、こ、…のようにあらゆる音節を記録したリストであり、各音節について、その音響的特徴が参照できるよう、音響モデルへのポインタが付与されている。他の言語の場合でも、その言語に即して適当な単位を定義し、辞書を構成することが可能である。また、音響モデル記憶部３０４に記憶された音響モデルは、所定の認識単位、すなわち音節、音素などに関する音響的特徴が、公知の隠れマルコフモデル等の手法に基づいて記述されたモデルである。

【００１３】

差分抽出部３０６は、音声認識部３０５から認識結果を、標準形発音記憶部３０２から書き起こしテキストをそれぞれ受け取り、両者の差分、すなわち相違箇所を抽出する。ここで、標準形発音記憶部３０２に記憶された書き起こしテキストは、図１の標準形発音記憶部２０１に記憶された書き起こしテキストと同様であるが、音声データ記憶部３０１に記憶された音声データと対応付いている、すなわち、音声データ記憶部３０１の音声データの内容が正しく発音された場合に観測されるべき本来の発音が書き起こしテキストとして記憶されている。発音変動計数部３０７は、図１の発音変動計数部２０４と同様の動作により、差分抽出部３０６から発音変動事例を受取り、発音変動規則を出力する。

【００１４】

『大西著「認識誤りの話者性を考慮した発声変形抽出と認識辞書拡張」日本音響学会２００７年春季研究発表会講演論文集、６５～６６頁、２００７年３月』に、発音変動規則抽出装置のさらに別の一例が記載されている。図３に示すように、この発音変動規則抽出装置４００は、音声データ記憶部４０１と、標準形発音記憶部４０２と、単語言語モデル・辞書記憶部４０３と、音響モデル記憶部４０４と、音声認識部４０５と、差分抽出部４０６と、発音変動計数部４０７とから構成されている。

【００１５】

このような構成を有する発音変動規則抽出装置４００は次のように動作する。すなわち、音声認識部４０５は、単語言語モデル・辞書記憶部４０３に記憶された言語モデルと辞書、および音響モデル記憶部４０４に記憶された音響モデルを用いて、音声データ記憶部４０１に記憶された音声データに対して公知の連続単語認識処理を行い、認識結果の単語系列を出力する。

【００１６】

ここで、単語言語モデル・辞書記憶部４０３に記憶された辞書および言語モデルは、一般的な大語彙連続音声認識システムが備える辞書および言語モデルと同様のものでよい。辞書は数万語の単語を含み、各単語について、その発音と、音響的特徴を参照するに足る音響モデルへのポインタが付与されている。言語モデルは、公知のｎ－ｇｒａｍモデルに基づき、ｎ－１個の単語並びを仮定した場合に、次にどのような単語が現れるかを確率の形で規定したモデルとなる。

【００１７】

また、音響モデル記憶部４０４に記憶された音響モデルは、図２の音響モデル記憶部３０４に記憶された音響モデルと同様、所定の認識単位、すなわち音節、音素などに関する音響的特徴が、公知の隠れマルコフモデル等の手法に基づいて記述されたモデルである。

【００１８】

差分抽出部４０６は、図２の差分抽出部３０６と同様の動作により、音声認識部４０５から認識結果を、標準形発音記憶部４０２から書き起こしテキストをそれぞれ受け取り、両者の差分、すなわち相違箇所を抽出する。ここで、標準形発音記憶部４０２に記憶された書き起こしテキストは、図２の標準形発音記憶部３０２と同様であり、音声データ記憶部４０１に記憶された音声データと対応付いていることが必要である。発音変動計数部４０７は、図１の発音変動計数部２０４や、図２の発音変動計数部３０７と同様の動作により、差分抽出部４０６から発音変動事例を受取り、発音変動規則を出力する。

【００１９】

これらの５つの文献に記載された発音変動規則抽出装置１００，２００，３００における第１の問題点は、発音変動規則やその元となる発音変動事例を得るために多大な労力を要するということである。その理由は、標準形発音とそれに対応する変形発音を大量に用意する必要があるためである。妥当性の高い発音変動規則を獲得するために、図１の発音変動規則抽出装置１００では、大量の音声データの書き起こしにより、標準形発音記憶部２０１に記憶される標準形発音、および変形発音記憶部２０２に記憶される変形発音をあらかじめ作成しておく必要がある。しかしながら、標準形発音と変形発音、特に後者の作成は、音声の聞き取りに習熟した作業者が注意深く音声を聞き、曖昧で判断がつきにくい変形発音を文字列として書き起こす作業となるため、とりわけ時間と労力がかかる。

【００２０】

第２の問題点は、汎化性の高い発音変動規則を得ることが難しいということである。その理由は、自由な話し言葉の音声データから正確な発音変動事例を得ることが難しいためである。例えば、図１の発音変動規則抽出装置１００では、人手により変形発音を書き起こすが、大量の書き起こしを得るためには、多数の作業者が分担して作業を行うのが普通である。しかしながら、話し言葉の発音は本質的に曖昧であるため、書き起こしには作業者の主観が多分に入り、作業結果にばらつきが生じる。また、図２の発音変動規則抽出装置２００では、音声認識部により統一的な基準で変形発音を自動的に取得することが可能である。しかしながら、現在の音声認識の技術水準では、言語的な事前知識のない状況で

音節の並びを求める連続音節認識処理を正確に行うことは極めて難しい。例えば、「ひろしま」という発声を連続音節認識すると、「けるせま」、「かるりか」というような、実際の発音の変動とは程遠い結果がしばしば得られる。すなわち、連続音節認識を適用しても、ランダムで有用性の乏しい文字列が得られるのみである。

【００２１】

図３の発音変動規則抽出装置３００でも、単語辞書と言語モデルという事前知識が利用可能とはいえ、図２の発音変動規則抽出装置２００と同様、音声認識の不正確さの問題がなお残る。さらに図３の発音変動規則抽出装置３００では、単語辞書と言語モデルが音声認識処理における言語的な制約として働くことから、得られる発音変動事例は単語辞書と言語モデルの影響を受ける。よって、実際に起こっている発音変動現象とは一般に異なる発音変動事例が得られる。例えば、「せんたくき（洗濯機）」が「せんたっき」に変わったり、「しょくぱん（食ぱん）」が「しょっぱん」に変わったりするような現象は一般的にみられるが、図３の発音変動規則抽出装置３００では、単語辞書に含まれる単語の組合せとしてしか音声認識結果が得られないため、「せんたっき」という発音と一致する認識結果が得られる保証はどこにもない。

【発明の開示】

【００２２】

本発明の目的は、少ない労力で、発音変動事例を頑健に検出し、汎化性の高い発音変動規則を獲得することにある。

【００２３】

本発明の一つ目のアスペクトによる発音変動規則抽出装置は、音声データ記憶部と、標準形発音記憶部と、サブワード言語モデル生成部と、音声認識部と、差分抽出部とを備える。音声データ記憶部は、音声データを記憶する。標準形発音記憶部は、音声データの標準形発音を表す標準形発音データを記憶する。サブワード言語モデル生成部は、標準形発音データからサブワード言語モデルを生成する。音声認識部は、サブワード言語モデルを用いて音声データを認識する。差分抽出部は、音声認識部が出力する認識結果と、標準形発音データとを比較して、これらの差分を抽出する。

【００２４】

本発明の二つ目のアスペクトによる発音変動規則抽出方法は、記憶することと、生成することと、認識することと、抽出することとを具備する。記憶することは、音声データの標準形発音を表す標準形発音データを記憶する。生成することは、標準形発音データからサブワード言語モデルを生成する。認識することは、サブワード言語モデルを用いて音声データを認識する。抽出することは、認識することによる認識結果と、標準形発音データとを比較して、これらの差分を抽出する。

【００２５】

本発明の三つ目のアスペクトによる発音変動規則抽出用プログラムは、コンピュータを、音声データ記憶部と、標準形発音記憶部と、サブワード言語モデル生成部と、音声認識部と、差分抽出部ととして機能させる。音声データ記憶部は、音声データを記憶する。標準形発音記憶部は、音声データの標準形発音を表す標準形発音データを記憶する。サブワード言語モデル生成部は、標準形発音データからサブワード言語モデルを生成する。音声認識部は、サブワード言語モデルを用いて音声データを認識する。差分抽出部は、音声認識部が出力する認識結果と、標準形発音データとを比較して、これらの差分を抽出する。このプログラムは、コンピュータ読み取り可能な記録媒体に格納でき、その記録媒体からコンピュータに読み込ませることができる。

【００２６】

本発明による効果は、正確で汎化性の高い発音変動規則を獲得できることにある。その理由は、制約のないサブワードを単位とした音声認識を基本として、音声データに対応する標準形発音という言語制約を任意の強さでかけながら音声認識を行うことにより、個々の音声データの違いに依存せず、多くの音声データに共通して現れる発音変動を抽出できるからである。また、人手作業で発生する、主観判断によるばらつきもないからである。

【図面の簡単な説明】
【００２７】
【図１】従来技術の一例を示すブロック図である。
【図２】従来技術の一例を示すブロック図である。
【図３】従来技術の一例を示すブロック図である。
【図４】本発明による第１の発明を実施するための最良の形態の構成を示すブロック図である。
【図５】第１の発明を実施するための最良の形態の動作の具体例を示す図である。
【図６】第１の発明を実施するための最良の形態の動作の具体例を示す図である。
【図７】第１の発明を実施するための最良の形態の動作の具体例を示す図である。
【図８】第１の発明を実施するための最良の形態の動作を示す流れ図である。
【図９】本発明による第２の発明を実施するための最良の形態の構成を説明するブロック図である。
【発明を実施するための最良の形態】
【００２８】
本発明を実施するための最良の形態の一つについて図面を参照して詳細に説明する。図４を参照すると、本発明の第１の実施の形態における発音変動規則抽出装置１００は、音声データ記憶部１０１と、標準形発音記憶部１０２と、サブワード言語モデル・辞書生成部１０３と、音響モデル記憶部１０４と、音声認識部１０５と、差分抽出部１０６と、発音変動確率推定部１０７と、言語モデル重み制御部１０８とを含む。
【００２９】
音声データ記憶部１０１は、発音変動事例が含まれると思われる多数の音声データを記憶する。標準形発音記憶部１０２は、音声データ記憶部１０１に記憶された音声データの書き起こしテキストを記憶する。ここに書き起こしテキストは、音声データの発音内容が、標準形で書き起こされたテキストデータであり、ひらがな、カタカナ、あるいは任意の発音記号の列で表される。ひらがなで表した書き起こしテキストの例を以下に示す。
みなさん　こんにちわ
発音を表す書き起こしであることから、「こんにちは」は「こんにちわ」と記述される。
【００３０】
音響モデル記憶部１０４は、後述する音声認識部１０５が音声認識処理を行う際に必要となる音響モデルを記憶する。音響モデルは、隠れマルコフモデルに基づいて個々の音素（日本語の場合は母音ａ，ｉ，ｕ，ｅ，ｏ，子音ｋ，ｓ，ｔ，ｎ，…）をモデル化したもの等を用いることができる。
【００３１】
サブワード言語モデル・辞書生成部１０３は、標準形発音記憶部１０２に記憶された書き起こしテキストを用いて、後述する音声認識部１０５が音声認識処理を行う際に必要となるサブワード言語モデル・辞書を生成する。ここに辞書は、例えばサブワードを音節とした場合、「あ，い，う，え，お，か，き，く，け，こ，…」の各音節を１単語として構成された辞書である。各単語、すなわち各音節についてその音響的特徴がわかるように、例えば「あ→ａ」、「か→ｋａ」、「さ→ｓａ」、…のように、単語から音響モデルへのポインタが付与されている。なお、音声データに一般に含まれ得るすべての音節を辞書が網羅しているなら、辞書は言語モデルと併せて生成する必要はない。この場合は、図示しない記憶部にそのような辞書を記憶しておいてもよい。
【００３２】
また、サブワード言語モデルは、サブワードを単語として、音声認識で広く用いられるｎ－ｇｒａｍモデルの考え方に基づき、各単語について、履歴ｈに続いて単語ｗが出現する確率Ｐ（ｗ｜ｈ）を規定したモデルである。具体的には、例えばｎ＝３のモデル（ｔｒｉｇｒａｍモデル）の場合、音節ｓｉ－２，ｓｉ－１がこの順に出現したとき、次に音節ｓｉが出現する確率Ｐ（ｓｉ｜ｓｉ－２，ｓｉ－１）が種々のｓｉ－２，ｓｉ－１，ｓｉについて規定されている。さらに、ここで生成されるサブワード言語モデルは、標準形発

音記憶部１０２に記憶された標準形の書き起こしテキストを学習データとして生成される。

【００３３】

例えば、上述の例の　みなさん　こんにちわ　という１発話を学習データに用いて生成されるサブワード言語モデルは、図５のように表される。なお、図５に示されていない履歴ｈを含む確率については等確率を与えるとすることができる。また、図５に示された履歴ｈを含むが図５に示されていない確率については０とすることができる。図５の”ｈ”欄において、φは空文字列であり、ここでは特に文頭を意味する。また、＃は単語間のポーズ（無音）を意味し、単語間にポーズが入る場合と入らない場合とで確率を二分している。上述のように、１発話のような短い単位の書き起こしテキストから学習されたサブワード言語モデルは、学習した書き起こしテキストを極めて高い確率（多くは確率１）で受理し、それ以外のほとんどすべてのテキストを拒否するサブワード言語モデルとなる。

【００３４】

なお、ここでは１発話を単位としてサブワード言語モデルを生成しているが、第１の実施の形態はこの単位の取り方を制限するものではなく、数個の発話を１単位とする、あるいは１個ないし数個の単語を単位とすることも可能である。また、サブワード言語モデル・辞書を構成する単語の単位を、ここでは音節としているが、一般にサブワードと呼ばれる単位、すなわち、音節、半音節、モーラ、音素等を単位としてサブワード言語モデル・辞書を生成することが可能である。

【００３５】

言語モデル重み制御部１０８は、サブワード言語モデルの重み値を少なくとも１回決定し、音声認識部１０５に送る。１回だけ決定する場合は、例えばあらかじめ実験的に定めた定数を使えばよい。また、複数回決定する場合は、同様にあらかじめ実験的に定めた複数個の定数を順に選択したり、あらかじめ定めた初期値から、あらかじめ定めた値を順次加算あるいは減算すればよい。ここにサブワード言語モデルの重み値とは、一般に正の値をとり、後述する音声認識部１０５が音響モデルやサブワード言語モデル・辞書を参照して音声認識処理を行う際に、サブワード言語モデルから計算されるスコアをどの程度重視するかを規定するパラメータである。

【００３６】

音声認識部１０５は、音響モデル記憶部１０４から音響モデルを、サブワード言語モデル・辞書生成部１０３から言語モデル・辞書をそれぞれ受け取り、また、言語モデル重み制御部１０８からサブワード言語モデルの重み値を少なくとも１回受け取る。そして、サブワード言語モデルの重み値ごとに、音声データ記憶部１０１に記憶された音声データに対して音声認識処理を行い、認識結果の音節列を求める。なお、音声認識処理は次の数式１により表すことができ、認識対象の音声データＯに対して、認識結果Ｗが得られる。

【００３７】

$$W = \arg\max_{W'} \left[\log P\left(O \middle| W', \theta_{AM}\right) + \lambda_{LM} \log P\left(W' \middle| \theta_{LM}\right) \right] \quad \cdots(1)$$

ここに、右辺ａｒｇｍａｘ関数内の第１項、第２項は、それぞれ音響スコア、言語スコアと呼ばれる。θＡＭは音響モデルであり、音響モデル記憶部１０４に記憶されている。θＬＭはサブワード言語モデル・辞書であり、サブワード言語モデル・辞書生成部１０３により生成される。λＬＭはサブワード言語モデルの重み値であり、言語モデル重み制御部１０８により決定される。Ｗ’は、認識結果Ｗになる候補であり、いずれかのＷ’が、Ｗとして算出される。ａｒｇｍａｘは、変数Ｗ’を動かしたときに、最大値を与えるＷ’を求める関数である。

【００３８】

サブワード言語モデルの重み値λＬＭが十分大きい場合、認識結果は、サブワード言語

モデルの学習データとなった書き起こしテキストと極めて高い確率で一致する。逆に、サブワード言語モデルの重み値λＬＭが小さい場合は、認識結果は先述の図２に示したような、連続音節認識の結果に近づく。なお、サブワード言語モデルの重み値を設定する代わりに、音響モデルの重み値を設定してもよい。すなわち、言語スコアの項に係数λＬＭをかける代わりに、音響スコアの項に同様の係数をかけても同じことである。サブワード言語モデルの重み値を大きくすることは、音響モデルの重み値を小さくすることと同値である。

【００３９】

差分抽出部１０６は、音声認識部１０５から少なくとも１つの認識結果を、また標準形発音記憶部１０２から標準形の書き起こしテキストをそれぞれ受け取り、両者の差分、すなわち相違箇所を抽出する。図６は、差分抽出部１０６が音声認識部１０５から受け取る認識結果の一例である。この例では、複数のサブワード言語モデルの重み値（１０．０～０．５）について、それぞれ得られた認識結果が示されている。差分抽出部１０６は、図６の認識結果を、標準形の書き起こしテキストと比較して、図７に示すように相違箇所をサブワード言語モデルの重み値とともに抽出する。図７の各行を、ここでは発音変動事例と呼ぶ。

【００４０】

なお、ここでは単語単位で差分すなわち発音変動事例を抽出しているが、第１の実施の形態は単位の取り方を単語に限定するものではなく、他の任意の単位でも実施可能である。例えば、前出の２つ目の文献では、所定の長さの音素系列を単位として差分抽出を行っているが、第１の実施の形態においても、このような形式による差分抽出を容易に適用することが可能である。

【００４１】

発音変動確率推定部１０７は、差分抽出部１０６から発音変動事例を受け取り、標準形発音、変形発音ごとに分類し、発音変動規則を得る。図７に示したように、発音変動事例を標準形発音、変形発音、サブワード言語モデルの重み値の組として、音声データ記憶部１０１に記憶された音声データから、数式２のようなＮ個の発音変動事例が得られたとする。

【００４２】

$$\left\{ w_i, \tilde{w}_i, \lambda_i \,\middle|\, i = 1,2,\cdots,N \right\} \qquad \cdots(2)$$

サブワード言語モデルの重み値が大きく、言語的制約が強い場合でも観測される変形発音は、一般的に発現しやすいであろうことを考慮すると、標準形発音ｗを所与とした発音変動規則が数式３のように確率論的に定義される。

【００４３】

$$P\left(\tilde{w} \middle| w\right) = \frac{\sum_i \lambda_i \delta_{w,wi} \delta_{\tilde{w},\tilde{w}i}}{\sum_i \lambda_i \delta_{w,wi}} \qquad \cdots(3)$$

ただし、δｉ，ｊはクロネッカのデルタ（ｉ＝ｊなら１、そうでなければ０）である。なお、数式３の変形例として、サブワード言語モデルの重み値λｉを考慮せず、λｉを１に置き換えて計算してもよい。また、数式３のλｉを、λｉを変数とする関数、例えばλｉの多項式関数などに置き換えてもよい。さらに、数式３の確率値を計算する際に、適当な平滑化操作を行ってもよい。適当な平滑化操作とは、例えば、サブワード言語モデルの重み値が小さい発音変動事例を無視する、観測回数が所定値に満たない発音変動事例を無視する等の操作に相当する。

【００４４】

次に、図４のブロック図および図８のフローチャートを参照して、第１の実施の形態における動作について詳細に説明する。まず、音声認識部１０５は、音響モデル記憶部１０４から音響モデルを読み込む（図８のステップＡ１）。次に、サブワード言語モデル・辞書生成部１０３は、標準形発音記憶部１０２に記憶された１発話分の書き起こしテキストを選択し（ステップＡ２）、読み込み（ステップＡ３）、これを学習データとしてサブワード言語モデルを生成し、および、必要に応じて辞書を生成する（ステップＡ４）。音声認識部１０５は、サブワード言語モデル・辞書生成部１０３が生成したサブワード言語モデル・辞書を読み込む（ステップＡ５）。次に、音声認識部１０５は、ステップＡ２で選択された書き起こしテキストに対応する音声データを、音声データ記憶部１０１から読み込む（ステップＡ６）。

【００４５】

言語モデル重み制御部１０８は、サブワード言語モデルの重み値として所定の値、例えば十分大きな値をセットし、音声認識部１０５に送る（ステップＡ７）。音声認識部１０５は、言語モデル重み制御部１０８がセットしたサブワード言語モデルの重み値に基づいて、音声認識処理を行い、音声認識結果すなわち音節列と、サブワード言語モデルの重み値を記憶する（ステップＡ８）。言語モデル重み制御部１０８は、サブワード言語モデルの重み値を一定量だけ増加又は減少させるなどして、サブワード言語モデルの重み値を更新する（ステップＡ９）。サブワード言語モデルの重み値の更新が所定回数Ｉを超えていれば次のステップに進み、そうでなければ、前述のステップＡ８、Ａ９を繰り返す（ステップＡ１０）。差分抽出部１０６は、音声認識部１０５が行った音声認識処理の結果を、図６にすでに示したような形式で受け取り、またステップＡ３でサブワード言語モデル・辞書生成部１０３が選択した標準形書き起こしテキストを受け取る。そして、図７や数式２ですでに示したような形式で、両者の相違箇所すなわち発音変動事例を抽出する（ステップＡ１１）。以上示したステップＡ２からＡ１１までの処理を、未処理の発話がなくなるまで繰り返す（ステップＡ１２）。最後に、発音変動確率推定部１０７は、差分抽出部１０６が求めたすべての発音変動事例を、数式３に従ってまとめ上げ、発音変動規則として出力する（ステップＡ１３）。

【００４６】

なお、第１の実施の形態における音声認識部１０５とわずかに異なる別の音声認識部を適用することが可能である。この別の音声認識部は、図示しない記憶部に無情報なサブワード言語モデルを記憶している。ここで無情報とは、任意の履歴において各単語が等確率で出現し得ることを意味する。無情報なサブワード言語モデルとは、サブワードを音節とした場合、任意の音節の並びであるｓｉ－２，ｓｉ－１，ｓｉについてＰ（ｓｉ｜ｓｉ－２，ｓｉ－１）＝ｃｏｎｓｔと表されるモデルのことである。無情報なサブワード言語モデルを用いた場合、音声認識処理は数式１に代わり、数式４となる。

【００４７】

$$W = \arg\max_{W'} \left[\log P\left(O \middle| W', \theta_{AM}\right) + K \log\left\{ \left(1-\lambda_{LM}\right) P\left(W' \middle| \theta_{LM0}\right) + \lambda_{LM} P\left(W' \middle| \theta_{LM}\right) \right\} \right] \quad \cdots(4)$$

ここにθＬＭ０は、無情報なサブワード言語モデルを表す。また、Ｋはあらかじめ定める定数である（なくてもよい）。数式４を用いた場合でも、言語モデル重み制御部１０８がサブワード言語モデルの重み値λＬＭを大小させることによって、第１の実施の形態における音声認識部１０５の場合と同様の結果が別の音声認識部から得られる。ただしこの場合、サブワード言語モデルの重み値には０≦λＬＭ≦１なる制約が生ずる。よって、言語モデル重み制御部１０８は、この制約の中でサブワード言語モデルの重み値を決定するよう動作する。

【００４８】

また、音声認識部１０５、差分抽出部１０６、および言語モデル重み制御部１０８の動

作についても、第１の実施の形態とわずかに異なる変形例を適用することが可能である。すなわち、第１の実施の形態では、図８のステップＡ８、Ａ９およびＡ１０を所定の回数だけ反復するとしているが、以下に述べるように、変形例における差分抽出部の抽出結果に応じて適応的に反復回数を決定することも可能である。

【００４９】

例えば、ステップＡ７にて、十分大きな値をサブワード言語モデルの重み値の初期値とし、ステップＡ９で順次サブワード言語モデルの重み値が減少するように動作させる場合は、差分抽出部にて標準形発音と認識結果の差分が所定のしきい値よりも大きくなった時点で反復を止めればよい。ここで、標準形発音と認識結果の差分を定量的に測るには、例えば文字列間の相違度合いの一般的尺度として知られている編集距離などが利用できよう。

【００５０】

あるいは逆に、ステップＡ７にて、十分小さな値をサブワード言語モデルの重み値の初期値とし、ステップＡ９で順次サブワード言語モデルの重み値が増加するように動作させる場合は、差分抽出部にて標準形発音と認識結果の差分が所定のしきい値よりも小さくなった時点、または標準形発音と認識結果が完全に一致した時点で反復を止めればよい。

【００５１】

次に、第１の実施の形態の効果について説明する。第１の実施の形態では、標準形発音のみ受理可能なサブワード言語モデルを生成するサブワード言語モデル・辞書生成部１０３と、サブワード言語モデルの重み、すなわちサブワード言語モデルの重み値を決定する言語モデル重み制御部１０８と、サブワード言語モデルおよびその重み値を用いて標準形発音に対応する音声データを認識する音声認識部１０５と、音声認識部１０５が出力する認識結果を標準形発音と比較して相違箇所を発音変動事例として抽出する差分抽出部１０６と、発音変動事例をまとめ上げて発音変動規則を出力する発音変動確率推定部１０７とを備える。そして、いくつかのサブワード言語モデルの重み値で音声認識処理を実行した結果をそれぞれ標準形発音と比較し、抽出される差分を発音変動事例とし、この発音変動事例をサブワード言語モデルの重み値を考慮してまとめ上げるようにしているため、正確で汎化性が高く、発現のしやすさに応じて確率値が付与された発音変動規則を獲得できる。

【００５２】

次に、本発明による第２の発明を実施するための最良の形態について図面を参照して詳細に説明する。第２の実施の形態は、第１の実施の形態を、プログラムを用いて実現するものである。このプログラムは、コンピュータを、第１の実施の形態における部１０１～１０８が結合されたものとして機能させる。図９を参照すると、発音変動規則抽出用プログラム９２は、コンピュータ読み取り可能な記録媒体９０に格納されていて、コンピュータ９１に読み込まれ、コンピュータ９１の動作を制御する。

【００５３】

発音変動規則抽出用プログラム９２は、コンピュータ９１に読み込まれた後、起動すると、記憶装置９４内の音声データ記憶部９４１を音声記憶部１０１として機能させ、標準形発音記憶部９４２を標準型発音記憶部１０２として機能させ、および音響モデル記憶部９４３を音響モデル記憶部１０４として機能させる。また、データ処理装置９３は発音変動規則抽出用プログラム９２の制御により、第１の実施の形態におけるサブワード言語モデル・辞書生成部１０３、音声認識部１０５、差分抽出部１０６、発音変動確率推定部１０７、および言語モデル重み制御部１０８として機能し、記憶装置９４内の音声データ記憶部９４１、標準形発音記憶部９４２、および音響モデル記憶部９４３に記憶されたデータを処理し、発音変動規則を出力する。

【００５４】

本発明によれば、大規模な音声データから発音変動規則を抽出する発音変動抽出装置や、発音変動規則抽出装置をコンピュータに実現するためのプログラムといった用途に適用できる。また、情報入力、情報検索、書き起こし支援、映像インデクシング等に広く用い

られる音声認識装置が知られるが、このような音声認識装置が使用する音響モデルや言語モデルを発音変動に対して頑健に作成するための音声認識用モデル作成装置、あるいは発音練習装置、語学学習装置、といった用途にも適用可能である。

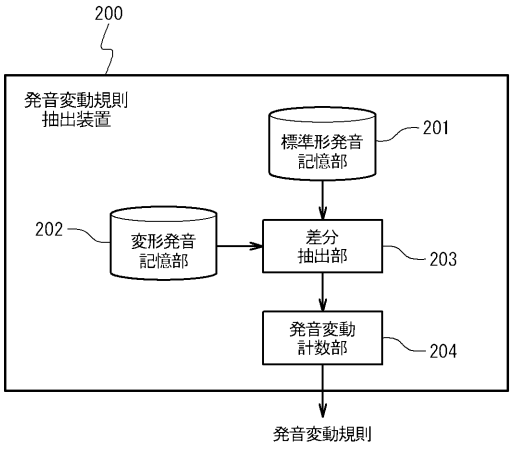
【００５５】

以上、実施の形態を参照して本願発明を説明したが、本願発明は上記実施の形態に限定されるものではない。本願発明の構成や詳細には、請求の範囲に記載された本願発明の技術的思想の範囲内において、当業者が適宜、様々な変形又は変更を加えることが可能である。

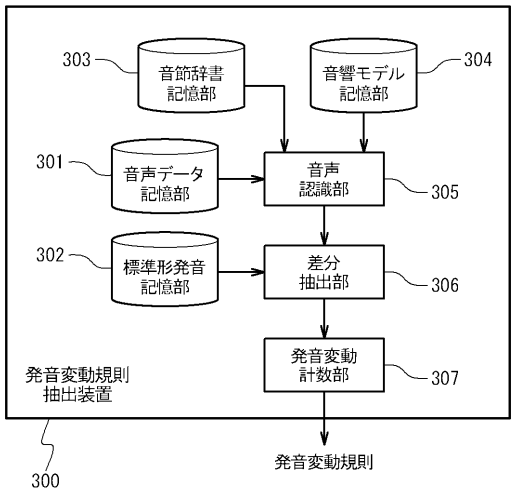
【００５６】

この出願は、２００７年１２月１８日に出願された特許出願番号２００７－３２６１３２号の日本特許出願に基づいている。本願は、この基礎出願により生じた優先権の利益を享受しており、この基礎出願における開示の内容の全てを、引用により、そっくりそのままここに取り込んでいる。

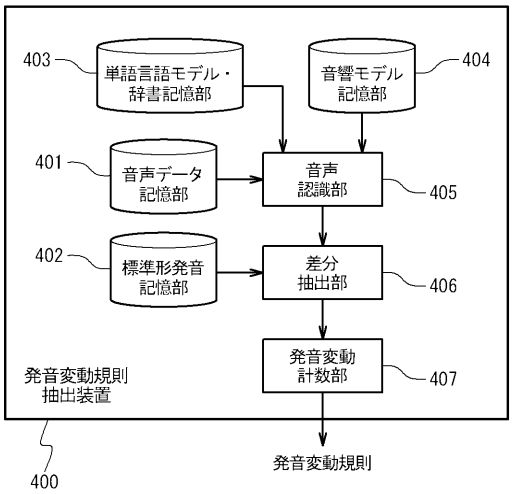
【図１】



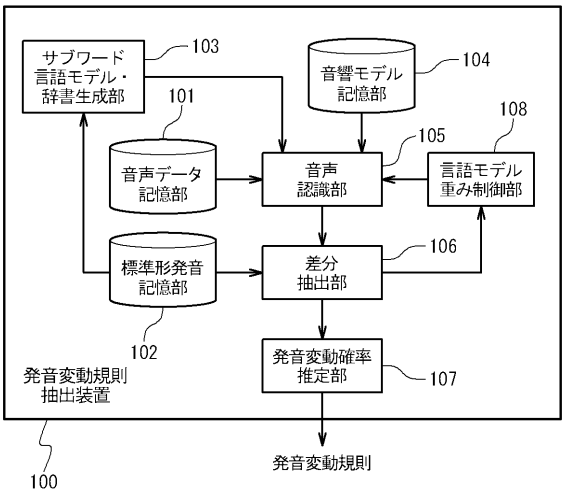
【図２】



【図３】



【図４】



【図５】

h	w	P(w|h)	備考
φ	み	1.0	文頭unigram
φみ	な	1.0	文頭bigram
みな	さ	1.0	
なさ	ん	1.0	
ん#	こ	1.0	
さん	こ	0.5	
さん	#	0.5	単語間ポーズの有無で分岐
んこ	ん	1.0	
#こ	ん	1.0	
こん	に	1.0	
んに	ち	1.0	
にち	わ	1.0	

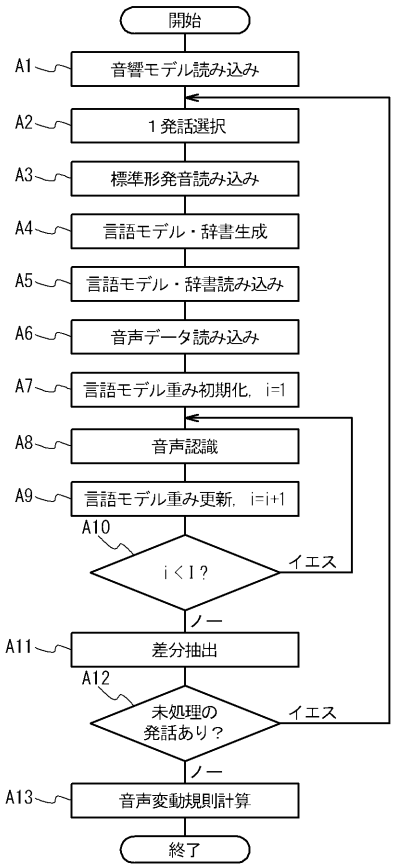
【図６】

言語モデル重み	認識結果
10.0	#みなさん#こんにちわ#
⋮	⋮
2.0	#みなさん#こにちわ#
1.0	#みんさん#こにつわ#
0.5	#みなさ#ごぬずば#

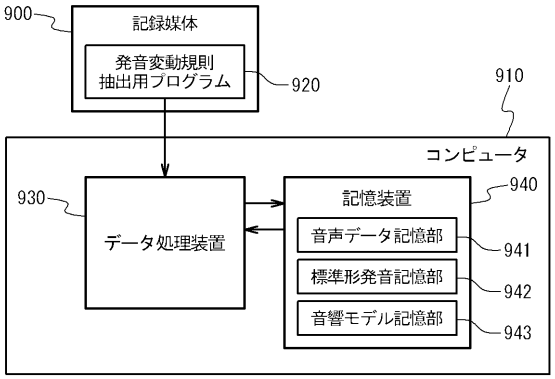
【図７】

標準形	変形	言語モデル重み	備考
みなさん	みなさん	10.0	恒等変形
こんにちわ	こんにちわ	10.0	恒等変形
みなさん	みなさん	2.0	恒等変形
こんにちわ	こにちわ	2.0	脱落”ん”
みなさん	みんさん	1.0	置換”な→ん”
こんにちわ	こにつわ	1.0	置換”ち→つ”
⋮	⋮	⋮	⋮

【図８】


開始
A1
音響モデル読み込み
A2
１発話選択
A3
標準形発音読み込み
A4
言語モデル・辞書生成
A5
言語モデル・辞書読み込み
A6
音声データ読み込み
A7
言語モデル重み初期化，i=1
A8
音声認識
A9
言語モデル重み更新，i=i+1
A10
i＜I？
イエス
ノー
A11
差分抽出
A12
未処理の
発話あり？
イエス
ノー
A13
音声変動規則計算
終了

【図９】


900
記録媒体
発音変動規則
抽出用プログラム
920
910
コンピュータ
930
データ処理装置
記憶装置
940
音声データ記憶部
941
標準形発音記憶部
942
音響モデル記憶部
943

フロントページの続き

(56)参考文献 特開平０８－１２３４７０（ＪＰ，Ａ）
特開平１０－３０８８８７（ＪＰ，Ａ）
秋田祐哉 他，"話し言葉音声認識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D-II，２００５年 9月 1日，Vol.J88-D-II，No.9，p.1780-1789
塚田元，"重み付き有限状態トランスデューサの自動学習と発音変形のモデル化"，電子情報通信学会論文誌D-II，２０００年１１月２５日，Vol.J83-D-II，No.11，p.2457-2464
深田俊明 他，"発音ネットワークに基づく発音辞書の自動生成"，電子情報通信学会論文誌D-II，１９９７年１０月２５日，Vol.J80-D-II，No.10，p.2626-2635

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １１／００－２５／９３
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩＩ）