(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2019-12095
(P2019-12095A)
(43)公開日　平成31年1月24日(2019.1.24)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/06	(2013.01)	G10L	15/06	300C
G10L 15/18	(2013.01)	G10L	15/18	300H
G10L 15/187	(2013.01)	G10L	15/187	

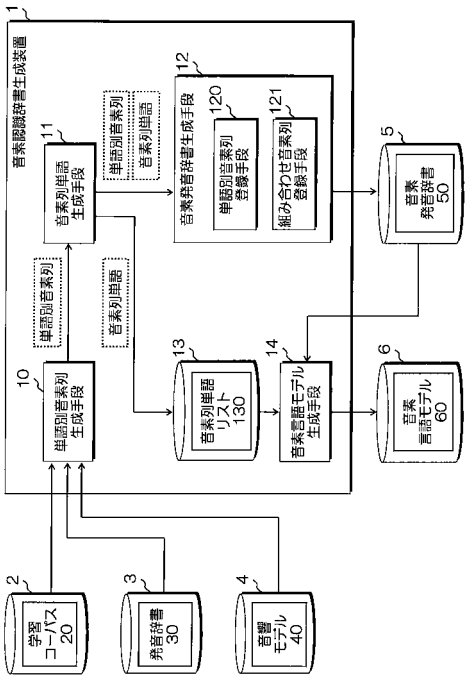
審査請求 未請求　請求項の数 7　OL　（全 13 頁）

(21)出願番号　特願2017-126929 (P2017-126929)
(22)出願日　平成29年6月29日 (2017.6.29)

(71)出願人　000004352
日本放送協会
東京都渋谷区神南２丁目２番１号
(74)代理人　110001807
特許業務法人磯野国際特許商標事務所
(72)発明者　一木　麻乃
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内
(72)発明者　尾上　和穂
北海道室蘭市山手町１－３－５０　日本放送協会　室蘭放送局内

(54)【発明の名称】音素認識辞書生成装置および音素認識装置ならびにそれらのプログラム

(57)【要約】
【課題】音素認識に用いる音素発音辞書および音素言語モデルを生成する音素認識辞書生成装置を提供する。
【解決手段】音素認識辞書生成装置１は、学習コーパスから、強制アライメントにより単語別音素列を生成する単語別音素列生成手段１０と、単語別音素列を１単語のテキストデータ形式に変換して音素列単語を生成する音素列単語生成手段１１と、音素列単語を見出し語とし、当該音素列単語に対応する単語別音素列を発音表記とすることで、音素発音辞書５０を生成する音素発音辞書生成手段１２と、音素列単語のリストから、Ｎ－ｇｒａｍ言語モデルを学習し、音素言語モデル６０を生成する音素言語モデル生成手段１４と、を備える。
【選択図】図１



【特許請求の範囲】
【請求項１】
音響モデルと発音辞書と学習コーパスとを用いて、音素認識に用いる音素発音辞書および音素言語モデルを生成する音素認識辞書生成装置であって、
前記学習コーパスの音声を、前記音響モデルと前記発音辞書とに基づいて音声認識し、前記発音辞書に登録されている見出し語に対応する単語ごとの音素列である単語別音素列を生成する単語別音素列生成手段と、
前記単語別音素列を１単語のテキストデータ形式に変換して音素列単語を生成する音素列単語生成手段と、
前記音素列単語を見出し語とし、当該音素列単語に対応する前記単語別音素列を発音表記とすることで、前記音素発音辞書を生成する音素発音辞書生成手段と、
前記音素列単語生成手段で生成される前記音素列単語のリストから前記音素列単語の連鎖としてＮ－ｇｒａｍ言語モデルを学習することにより、前記音素言語モデルを生成する音素言語モデル生成手段と、
を備えることを特徴とする音素認識辞書生成装置。
【請求項２】
前記単語別音素列生成手段は、前記単語別音素列の音素間に音素以外の予め定めた文字を挿入することで、前記音素列単語を生成することを特徴とする請求項１に記載の音素認識辞書生成装置。
【請求項３】
前記音素発音辞書生成手段は、予め定めた数の音素を組み合わせた音素列を前記テキストデータ形式に変換した見出し語とし、当該見出し語に対応する音素列を発音表記として前記音素発音辞書に登録することを特徴とする請求項１または請求項２に記載の音素認識辞書生成装置。
【請求項４】
前記音素言語モデル生成手段は、前記音素列単語生成手段で生成される前記音素列単語のリストに存在しない音素列単語の連鎖に対して、スムージングによりＮグラム確率を与えることを特徴とする請求項１から請求項３のいずれか一項に記載の音素認識辞書生成装置。
【請求項５】
コンピュータを、請求項１から請求項４のいずれか一項に記載の音素認識辞書生成装置として機能させるための音素認識辞書生成プログラム。
【請求項６】
音響モデルと、請求項１から請求項４のいずれか一項に記載の音素認識辞書生成装置により生成された音素発音辞書および音素言語モデルとを用いて、音声の音素を認識する音素認識装置であって、
前記音響モデルと前記音素発音辞書と前記音素言語モデルとにより、前記音声を音素列単語単位で認識する認識手段と、
この認識手段で認識された１単語のテキストデータ形式である音素列単語を、個々の音素に分離して音素列を生成する音素列生成手段と、
を備えることを特徴とする音素認識装置。
【請求項７】
コンピュータを、請求項６に記載の音素認識装置として機能させるための音素認識プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、発話音声の音素認識に用いる音素発音辞書および音素言語モデルを生成する音素認識辞書生成装置およびそのプログラム、ならびに、音素発音辞書および音素言語モデルを用いた音素認識装置およびそのプログラムに関する。

【背景技術】
【０００２】
通常、音声認識では、単語と当該単語の発音系列（音素列）とを対応付けた発音辞書を用いている。この発音辞書には、一般的な辞書に記載されているような読みが発音として登録されている。
しかし、表記上の読みと実際に発話された発音とでは異なることが多い。例えば、放送番組では、ニュース番組のアナウンサの正確な（発音辞書の発音と近い）発音に比べ、情報番組の出演者の発話は曖昧な発音が多い。
そこで、統計的機械翻訳モデルを利用して、アナウンサ等の正確な発音を前提とした音素列から、発音が不明瞭な発話の音素列の単語を推定して、発音辞書を拡張する技術が開示されている（特許文献１参照）。
【０００３】
特許文献１の技術（以下、従来技術という）では、認識対象音素の前後の音素に対する依存性（環境依存）を考慮して音素認識を行う。
この従来技術は、学習コーパスから、トライフォンを１つの単語として発音辞書を学習するとともに、トライフォンの連接確率を与える言語モデルを学習する。ここで、トライフォンは、例えば、「警察」の発音では、「（けー）ｋ－ｅ：＋ｓ」，「（さ）ｅ：－ｓ＋ａ」，「（つ）ｓ－ａ＋ｔｓ」のように、中心音素を含めた前後の発音を含めて表現したものである。
【０００４】
そして、従来技術は、音声と書き起こしテキストとを対応付けた学習コーパスから強制音素アライメントを行った音素列（標準音素列）と、音素のトライフォンの言語モデルおよび発音辞書を用いて学習コーパスの音声を音素認識した音素列（実発話音素列）とを用いて、統計的機械翻訳モデルを学習する。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２０１６－１６１７６５号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
前記した従来技術は、統計的機械翻訳モデルを学習するために、強制音素アライメントを行った音素列（標準音素列）と、音素認識した音素列（実発話音素列）とを用いる。この統計的機械翻訳モデルの精度を高めるには、標準音素列と実発話音素列の質が重要になる。
従来技術で、アナウンサ等の正確な発音の音声とその書き起こしテキストとを学習コーパスとして用いて標準音素列と実発話音素列とを生成した場合、理想的には、それぞれの音素列がほぼ同じであることが望ましい。
しかし、従来技術では、標準音素列と実発話音素列とをＤＰ（Dynamic Programming）マッチングした結果、音素が異なる割合（音素異なり率）が、２２．８％あり、さらなる音素認識の精度改善が望まれている。
【０００７】
本発明は、このような問題に鑑みてなされたものであり、音素認識の精度を高める音素認識辞書（音素発音辞書および音素言語モデル）を生成する音素認識辞書生成装置およびそのプログラム、ならびに、音素発音辞書および音素言語モデルを用いた音素認識装置およびそのプログラムを提供することを課題とする。
【課題を解決するための手段】
【０００８】
前記課題を解決するため、本発明に係る音素認識辞書生成装置は、音響モデルと発音辞書と学習コーパスとを用いて、音素認識に用いる音素発音辞書および音素言語モデルを生

成する音素認識辞書生成装置であって、単語別音素列生成手段と、音素列単語生成手段と、音素発音辞書生成手段と、音素言語モデル生成手段と、を備える。

【０００９】

かかる構成において、音素認識辞書生成装置は、単語別音素列生成手段によって、学習コーパスの音声を音響モデルと発音辞書とに基づいて音声認識し、発音辞書に登録されている見出し語に対応する単語ごとの音素列である単語別音素列を生成する。

そして、音素認識辞書生成装置は、音素列単語生成手段によって、単語別音素列を１単語のテキストデータ形式に変換して音素列単語を生成する。例えば、音素列単語生成手段は、単語別音素列の音素間のスペースに音素以外の予め定めた文字（例えば、“＋”）を挿入することで、音素列単語を生成する。これによって、音素認識辞書生成装置は、音素列単語を１単語として扱うことが可能になる。

【００１０】

そして、音素認識辞書生成装置は、音素発音辞書生成手段によって、音素列単語を見出し語とし、当該音素列単語に対応する単語別音素列を発音表記とすることで、音素発音辞書を生成する。これによって、音素発音辞書生成手段は、単語単位で音素列の発音を音素発音辞書に登録する。

さらに、音素認識辞書生成装置は、音素言語モデル生成手段によって、音素列単語生成手段で生成される音素列単語のリストから音素列単語の連鎖としてＮ－ｇｒａｍ言語モデルを学習することにより、音素言語モデルを生成する。これによって、音素言語モデル生成手段は、音素認識を行う際の音素列単語の接続確率を計算するため音素列単語の出現確率をモデル化する。

【００１１】

なお、音素認識辞書生成装置は、コンピュータを、単語別音素列生成手段、音素列単語生成手段、音素発音辞書生成手段、音素言語モデル生成手段として機能させるための音素認識辞書生成プログラムで動作させることができる。

【００１２】

また、前記課題を解決するため、本発明に係る音素認識装置は、音響モデルと、音素認識辞書生成装置により生成された音素発音辞書および音素言語モデルとを用いて、音声の音素を認識する音素認識装置であって、認識手段と、音素列生成手段と、を備える。

【００１３】

かかる構成において、音素認識装置は、認識手段によって、音響モデルと音素発音辞書と音素言語モデルとにより、音声を音素列単語単位で認識する。これによって、認識手段は、単語の繋がりに依存した音素列を認識することが可能になる。

そして、音素認識装置は、音素列生成手段によって、認識手段で認識された１単語のテキストデータ形式である音素列単語を、個々の音素に分離して音素列を生成する。例えば、音素列生成手段は、単語別音素列の音素間に挿入されている予め定めた文字（例えば、“＋”）をスペースに置き換えることで、個々の音素に分離する。

【００１４】

なお、音素認識装置は、コンピュータを、認識手段、音素列生成手段として機能させるための音素認識プログラムで動作させることができる。

【発明の効果】

【００１５】

本発明は、以下に示す優れた効果を奏するものである。

本発明によれば、音素列を単語単位とした音素発音辞書および音素言語モデルを生成することができる。

この音素発音辞書および音素言語モデルを用いることで、音素認識する際の音素の連結確率を、単に音素の前後の依存性だけではなく、音素の単語内および単語間における依存性も加味して算出することが可能になり、音声から音素を認識する際の認識精度を高めることができる。

【図面の簡単な説明】

【００１６】
【図１】本発明の第１実施形態に係る音素認識辞書生成装置の構成を示すブロック構成図である。
【図２】図１の単語別音素列生成手段における単語別音素列の生成例を説明するための説明図であって、（ａ）は学習コーパスの音声の書き起こし例、（ｂ）は発音辞書の一部、（ｃ）は生成した単語別音素列の例を示す。
【図３】音素の表記例を示す図である。
【図４】図１の音素発音辞書生成手段が生成する音素発音辞書の例を示す図である。
【図５】図１の音素列単語生成手段が生成する音素列単語リストの例を示す図である。
【図６】図１の音素言語モデル生成手段が生成する音素言語モデルの例を示す図である。
【図７】本発明の第１実施形態に係る音素認識辞書生成装置の動作を示すフローチャートである。
【図８】本発明の第２実施形態に係る音素認識装置の構成を示すブロック構成図である。
【発明を実施するための形態】
【００１７】
以下、本発明の実施形態について図面を参照して説明する。
＜第１実施形態＞
〔音素認識辞書生成装置の構成〕
まず、図１を参照して、本発明の第１実施形態に係る音素認識辞書生成装置１の構成について説明する。
【００１８】
音素認識辞書生成装置１は、音声データから音素を認識するための辞書として、音素発音辞書および音素言語モデルを生成するものである。この音素認識辞書生成装置１は、学習コーパス記憶装置２、発音辞書記憶装置３および音響モデル記憶装置４にそれぞれ記憶されている学習コーパス２０、発音辞書３０および音響モデル４０から、音素発音辞書５０および音素言語モデル６０を生成する。
【００１９】
具体的には、音素認識辞書生成装置１は、学習コーパスから強制アライメントにより単語別音素列を生成し、生成した単語別音素列を１単語のテキストデータ形式に変換して音素列単語を生成する。そして、音素認識辞書生成装置１は、生成した音素列単語を見出し語とし、当該音素列単語に対応する単語別音素列を発音表記とすることで、音素発音辞書５０を生成する。さらに、音素認識辞書生成装置１は、生成した音素列単語のリストから、Ｎ－ｇｒａｍ言語モデルを学習し、音素言語モデル６０を生成する。
【００２０】
学習コーパス２０は、予め大量の音声データ（音声コーパス）と、音声データの書き起こしテキスト（テキストコーパス）とを対応付けたデータである。この学習コーパス２０は、例えば、ニュース番組、情報番組等におけるアナウンサ、リポータ等の約１０００時間程度の音声（音声コーパス）と、その音声を書き起こしたテキスト（テキストコーパス）である。
【００２１】
発音辞書３０は、所定の文字列である見出し語（ここでは、単語とする）ごとに、その発音表記（音素列）を示した辞書である。
この発音辞書３０は、一般的な発音辞書であって、例えば、人手を介して見出し語（単語）とその発音表記（音素列）とを対応付けた辞書である。
【００２２】
音響モデル４０は、大量の音声データから予め学習したディープニューラルネットワーク（ＤＮＮ：Deep Neural Network）音響モデルである。例えば、ＤＮＮの入力には、メルフィルタバンク対数パワーの４０次元に時間変化（Δ＋ΔΔ）を加えて１１フレーム分の特徴量を連結（スプライス）した特徴量を用い、ＤＮＮの隠れ層を８層とする。
なお、音響モデル４０における音響特徴量の尤度計算は、隠れマルコフモデル（ＨＭＭ

：Hidden Markov Model）や、ガウス混合モデル（ＧＭＭ：Gaussian mixture model）音響モデルであっても構わない。

以下、音素認識辞書生成装置１の構成について詳細に説明する。

【００２３】

音素認識辞書生成装置１は、図１に示すように、単語別音素列生成手段１０と、音素列単語生成手段１１と、音素発音辞書生成手段１２と、音素列単語リスト記憶手段１３と、音素言語モデル生成手段１４と、を備える。

また、音素認識辞書生成装置１は、生成した音素発音辞書５０を記憶する音素発音辞書記憶装置５と、生成した音素言語モデル６０を記憶する音素言語モデル記憶装置６と、を外部に接続している。もちろん、音素発音辞書記憶装置５および音素言語モデル記憶装置６は、音素認識辞書生成装置１の内部に備える構成としてもよい。また、音素発音辞書記憶装置５および音素言語モデル記憶装置６は、１つの記憶装置で構成してもよい。

【００２４】

単語別音素列生成手段１０は、発音辞書３０と音響モデル４０とに基づいて、学習コーパス２０の音声（音声コーパス）を強制アライメントすることで、発音辞書３０に登録されている見出し語に対応する単語ごとに、音声の音素列を切り分けて単語別音素列を生成するものである。

【００２５】

この単語別音素列生成手段１０は、学習コーパス２０の音声から、音響モデル４０に対応する音響特徴量（メル周波数ケプストラム係数等）を抽出する。そして、単語別音素列生成手段１０は、発音辞書３０と音響モデル４０とを用いて、音声の書き起こしテキスト（テキストコーパス）を事前知識とする音声認識を行い、発音辞書３０に登録されている文字列（見出し語）に対応して強制アライメントする。これにより、単語別音素列生成手段１０は、図２（ｂ）に示されているように、発音辞書３０に登録されている単語に複数存在する発音の音素列に対し、尤も音声に近い発音の音素列を選択し、単語別音素列を生成する。

【００２６】

図２は、単語別音素列生成手段１０における単語別音素列の生成例を示す。例えば、単語別音素列生成手段１０は、学習コーパス２０として、「世界一短い東京の橋でイベントが開かれました」の音声データを入力した場合、音響モデル４０に対応する音響特徴量を抽出する。

そして、単語別音素列生成手段１０は、音声データに対応する図２（ａ）に示す学習コーパス２０の書き起こしテキスト「世界一　短い　東京　…」を事前知識として、図２（ｂ）に示す発音辞書３０と、音響モデル４０と、を用いて音声認識を行う。

【００２７】

これによって、単語別音素列生成手段１０は、図２（ｃ）に示すように、単語ごとの音素列（単語別音素列）「ｓ△ｅ△ｋ△ａ△ｉ△ｉ△ｃｈ△ｉ／ｍ△ｉ△ｊ△ｉ△ｋ△ａ△ｉ／ｔ△ｏ：△ｋｙ△ｏ：／…」（ここで、“△”はスペースを示す）を生成する。

単語別音素列生成手段１０は、生成した単語別音素列を音素列単語生成手段１１に出力する。

【００２８】

音素列単語生成手段１１は、単語別音素列生成手段１０で生成された単語別音素列を、単語ごとに１単語のテキストデータ形式に変換した音素列単語を生成するものである。

この音素列単語生成手段１１は、単語別音素列の音素間に音素以外の予め定めた文字を挿入することで、個々に分離した音素列を、１単語のテキストデータ形式に変換する。

【００２９】

具体的には、音素列単語生成手段１１は、音素ごとにスペースを含んだ単語別音素列のスペースを、音素以外の予め定めた文字に置き換えて１つの単語テキストとする。例えば、音素列単語生成手段１１は、単語別音素列のスペースを“＋”に置き換え、“ｓ△ｅ△ｋ△ａ△ｉ△ｉ△ｃｈ△ｉ”を“ｓ＋ｅ＋ｋ＋ａ＋ｉ＋ｉ＋ｃｈ＋ｉ”等に変換する。

【００３０】
音素列単語生成手段１１は、スペースを含んだ単語別音素列と、テキスト置換した音素列単語とを対にして、順次、音素発音辞書生成手段１２に出力する。また、音素列単語生成手段１１は、テキスト置換した音素列単語のみを、順次、音素列単語リスト記憶手段１３に書き込む。
【００３１】
音素発音辞書生成手段１２は、音素列を単語とみなした音素列単語の発音辞書である音素発音辞書を生成するものである。音素発音辞書生成手段１２は、図１に示すように、単語別音素列登録手段１２０と、組み合わせ音素列登録手段１２１と、を備える。
【００３２】
単語別音素列登録手段１２０は、単語別音素列と音素列単語とを対として登録した音素発音辞書を生成するものである。単語別音素列登録手段１２０は、音素列単語生成手段１１で生成された音素列単語を見出し語とし、音素列単語と対となる単語別音素列をその見出し語の発音として、音素発音辞書記憶装置５の音素発音辞書５０に登録する。
【００３３】
なお、単語別音素列登録手段１２０は、同じ見出し語となる音素列単語に対して、異なる発音の単語別音素列が入力された場合、見出し語に複数の発音を登録する。また、単語別音素列登録手段１２０は、同じ見出し語となる音素列単語に対して、同じ発音の単語別音素列が入力された場合、登録を行わないこととする。
【００３４】
組み合わせ音素列登録手段１２１は、任意の音素の組み合わせで構成される音素列を単語とみなした見出し語と、その音素列とを対として、音素発音辞書に登録するものである。
具体的には、組み合わせ音素列登録手段１２１は、図３に示す音素の例において、すべての音素（図３の例では、４０音素）に対して、予め定めた最大音素数（ここでは、“４”とする）の音素の組み合わせ（$４０^{1}＋４０^{2}＋４０^{3}＋４０^{4}$通り）の音素列を、音素発音辞書記憶装置５の音素発音辞書５０に登録する
【００３５】
この組み合わせ音素列登録手段１２１は、音素列単語生成手段１１と同様に、音素を組み合わせた音素列を、１つのテキストデータ形式に変換する。具体的には、組み合わせ音素列登録手段１２１は、音素を組み合わせた音素列のスペースを音素以外の予め定めた１つのテキスト（ここでは、“＋”）に置き換えた単語に変換し、見出し語とする。
【００３６】
ここで、図４を参照して、音素発音辞書生成手段１２が音素発音辞書記憶装置５に登録する音素発音辞書５０の例について説明する。
図４に示すように、音素発音辞書５０は、単語別音素列登録手段１２０で登録される辞書Ａと、組み合わせ音素列登録手段１２１で登録される辞書Ｂとで構成される。
辞書Ａは、学習コーパス２０の書き起こしに含まれる単語の発音を示す単語音素列のスペース部分を“＋”に置き換えた単語別音素列を見出し語とし、スペースを含んだ音素列（単語別音素列）を見出し語に対応する発音表記とする。
【００３７】
辞書Ｂは、すべての音素の予め定めた最大音素数の組み合わせにおいて、音素列のスペース部分を“＋”に置き換えた組み合わせ音素列を見出し語とし、スペースを含んだ音素列を見出し語に対応する発音表記とする。これによって、学習コーパス２０に含まれていない音素の組み合わせであっても、音素発音辞書５０内に見出し語と発音表記とが登録される。
図１に戻って、音素認識辞書生成装置１の構成について説明を続ける。
【００３８】
音素列単語リスト記憶手段１３は、音素列単語生成手段１１で生成される音素列単語を、音素列単語リストとして記憶するものである。音素列単語リスト記憶手段１３は、半導

体メモリ、ハードディスク等の一般的な記憶装置で構成することができる。

【００３９】

図５に、音素列単語リスト記憶手段１３に記憶される音素列単語リスト１３０の例を示す。図５に示すように、音素列単語リスト１３０は、音素列単語生成手段１１で生成した単語別音素列のスペースを“＋”に置き換えた音素列単語を逐次記憶したものである。

この音素列単語リスト１３０には、学習コーパス２０の書き起こしに含まれる単語の音素列を１つの単語として順次書き込まれる。

【００４０】

音素言語モデル生成手段１４は、音素列単語リスト記憶手段１３に記憶されている音素列単語リスト１３０から、音素言語モデルを学習により生成するものである。

音素言語モデルは、任意の音素列単語の単語列において、それが文である確率（尤度）を付与する確率モデル（統計的言語モデル）である。この音素言語モデルは、例えば、Ｎ－ｇｒａｍ言語モデルであって、以下の式（１）に示すように、音素列単語の列$w_1 w_2 \cdots w_{i-1}$の後にｉ番目の音素列単語w_iが出現する条件付き確率（Ｎグラム確率）を与えるモデルである。なお、桁あふれを防止するため、式（１）の尤度を対数とし、対数尤度とすることが好ましい。

【００４１】

【数１】

$$P\left(w_i \middle| w_{i-N+1} \cdots w_{i-1}\right) \qquad \cdots式（1）$$

【００４２】

例えば、学習コーパスの書き起こしで「東京の橋で」という単語列が存在する場合、音素言語モデル生成手段１４は、音素列単語リスト１３０として生成される「ｔ＋ｏ：＋ｋｙ＋ｏ：」、「ｎ＋ｏ」、「ｈ＋ａ＋ｓｈ＋ｉ」、「ｄ＋ｅ」の音素列単語からなる「ｔ＋ｏ：＋ｋｙ＋ｏ：△ｎ＋ｏ△ｈ＋ａ＋ｓｈ＋ｉ△ｄ＋ｅ」という学習テキストでＮ－ｇｒａｍ言語モデルを学習する。

【００４３】

なお、音素言語モデル生成手段１４は、学習テキストとして音素列単語リスト１３０に現れない音素列単語の連鎖には、一般的なスムージング手法によってＮグラム確率を与える。音素言語モデル生成手段１４は、スムージング手法として、例えば、バックオフスムージング（back-off smoothing）を用いることができる。バックオフスムージングは、学習テキストに出現しない音素列単語の連鎖のＮグラム確率を、連鎖数の少ない音素列単語の連鎖に与えられているＮグラム確率から推定するものである。

【００４４】

これによって、音素言語モデル生成手段１４は、すべての音素の組み合わせを含んだ音素発音辞書５０に登録されている見出し語の音素列単語の連鎖に、Ｎグラム確率を付与することができる。

音素言語モデル生成手段１４は、生成した音素言語モデルを音素言語モデル記憶装置６に書き込み記憶する。

【００４５】

図６に、音素言語モデル記憶装置６に記憶される音素言語モデル６０の例を示す。ここでは、Ｎ－ｇｒａｍ言語モデルとして、２－ｇｒａｍ言語モデルの例を示す。

図６に示すように、音素言語モデル６０は、２つの音素列単語w_1，w_2に対して、Ｎグラム確率（ｌｏｇＰ（w_2｜w_1））を対応付けたものである。

【００４６】

以上説明したように音素認識辞書生成装置１を構成することで、音素認識辞書生成装置１は、発話音声から音素を認識するための辞書として、音素発音辞書および音素言語モデルを生成することができる。このように生成された音素発音辞書および音素言語モデルは、音素認識を行う際に、単に音素の前後の依存性だけではなく、音素の単語内および単語

間における音素列の依存性を加味して、音素認識の精度を高めることができる。
なお、音素認識辞書生成装置１は、図示を省略したコンピュータを、前記した各手段として機能させるプログラム（音素認識辞書生成プログラム）で動作させることができる。
【００４７】
〔音素認識辞書生成装置の動作〕
次に、図７を参照（構成については適宜図１参照）して、本発明の第１実施形態に係る音素認識辞書生成装置１の動作について説明する。
【００４８】
ステップＳ１において、単語別音素列生成手段１０は、学習コーパス２０の音声から音響特徴量を抽出し、発音辞書３０と音響モデル４０を用いて、学習コーパス２０の音声の書き起こしテキストを事前知識とする音声認識を行い、発音辞書３０に登録されている見出し語に対応して強制アライメントした単語別音素列を生成する。
【００４９】
ステップＳ２において、音素列単語生成手段１１は、ステップＳ１で生成した単語別音素列の音素間のスペースを音素以外の予め定めた１つのテキスト（例えば、“＋”）に置き換えて、音素列単語を生成する。これによって、以降の動作において、単語別音素列を、スペースのない、１つの単語テキストとして扱うことが可能になる。
【００５０】
ステップＳ３において、音素列単語生成手段１１は、ステップＳ２で生成した音素列単語を、順次、音素列単語リスト記憶手段１３に書き込み記憶する。これによって、音素列単語リスト記憶手段１３には、学習コーパス２０の音声に対応する音素列を単語ごとにテキスト化した音素列単語リスト１３０が記録される。
【００５１】
ステップＳ４において、音素発音辞書生成手段１２は、単語別音素列登録手段１２０によって、ステップＳ２で生成した音素列単語を見出し語とし、ステップＳ１で生成した単語別音素列をその見出し語に対応する発音表記として、音素発音辞書記憶装置５の音素発音辞書５０に登録する（図４の辞書Ａ参照）。
【００５２】
ステップＳ５において、単語別音素列生成手段１０は、学習コーパス２０の音声についてすべて入力が終了したか否かを判定する。ここで、学習コーパス２０の入力が終了していない場合（ステップＳ５でＮｏ）、音素認識辞書生成装置１は、ステップＳ１に動作を戻す。
一方、学習コーパス２０の入力が終了した場合（ステップＳ５でＹｅｓ）、音素認識辞書生成装置１は、ステップＳ６に動作を進める。
【００５３】
ステップＳ６において、音素発音辞書生成手段１２は、組み合わせ音素列登録手段１２１によって、任意の音素の組み合わせで構成される音素列を単語とみなした見出し語と、その音素列とを対として、音素発音辞書記憶装置５の音素発音辞書５０に登録する（図４の辞書Ｂ参照）。これによって、学習コーパス２０からは抽出することができない音素の並びに対して、見出し語と発音表記とを割り当てることができる。
【００５４】
ステップＳ７において、音素言語モデル生成手段１４は、ステップＳ３で順次、音素列単語リスト記憶手段１３に記憶された音素列単語リスト１３０から、Ｎ－ｇｒａｍ言語モデルの音素言語モデル６０を生成し、音素言語モデル記憶装置６に記憶する。
【００５５】
さらに、ステップＳ８において、音素言語モデル生成手段１４は、音素発音辞書５０に登録されている音素の組み合わせから生成された見出し語を含めて、学習コーパスとして音素列単語リスト１３０に現れない音素列単語の連鎖に対して、スムージング手法によってＮグラム確率を与える。これによって、音素言語モデル６０を用いて音素認識を行う際に、音素列単語の連結確率が“０”になることを防止することができる。

以上の動作によって、音素認識辞書生成装置１は、音声から音素を認識するための辞書として、音素発音辞書および音素言語モデルを生成する。

【００５６】

＜第２実施形態＞

〔音素認識装置〕

次に、図８を参照して、本発明の第２実施形態に係る音素認識装置２００について説明する。

【００５７】

音素認識装置２００は、音響モデルと、音素認識辞書生成装置１で生成した音素発音辞書および音素言語モデルとを用いて、音声データから音素を認識するものである。この音素認識装置２００は、音響モデル記憶装置４、音素発音辞書記憶装置５および音素言語モデル記憶装置６にそれぞれ記憶されている音響モデル４０、音素発音辞書５０および音素言語モデル６０を用いて、音声データから音素を認識する。

【００５８】

音響モデル４０は、図１で説明した音響モデルと同じであって、大量の音声データから予め学習した音素ごとの音響特徴量をディープニューラルネットワーク（ＤＮＮ）によってモデル化したものである。

【００５９】

音素発音辞書５０は、図１で説明した音素認識辞書生成装置１で生成されたものである（図４参照）。

音素言語モデル６０は、図１で説明した音素認識辞書生成装置１で生成されたものである（図６参照）。

【００６０】

音素認識装置２００は、図８に示すように、認識手段２０１と、音素列生成手段２０２と、を備える。

【００６１】

認識手段２０１は、音響モデル４０と、音素発音辞書５０と、音素言語モデル６０とを用いて、音声データから音素列を認識するものである。

この認識手段２０１は、外部から入力される音声データから音響特徴量を抽出し、音響モデル４０と音素発音辞書５０とから音素列単語の候補をリストアップする。そして、認識手段２０１は、その候補の中で、音素言語モデル６０に基づく接続確率が最大となる音素列単語を認識結果とする。

【００６２】

具体的には、認識手段２０１は、音素列単語列w_1，w_2，…，w_nで、以下の式（２）に示す、w_{n-1}の次にw_nが出現する確率（事後確率）Ｐ（$w_n \mid w_{n-1}$）の接続確率が最大となる音素列単語列を認識する。

【００６３】

【数２】

$$P(w_1, w_2, \cdots, w_n) = \prod_{k=1}^{n} P(w_k | w_{k-1}) \qquad \cdots式（2）$$

【００６４】

このように、認識手段２０１は、一般的な音声認識が発音辞書に登録されている単語単位で音声を認識するのに対し、音素発音辞書５０に登録されている単語とみなした音素列単語単位で音声を認識する。

認識手段２０１は、認識した音素列単語を、順次、音素列生成手段２０２に出力する。

【００６５】

音素列生成手段２０２は、認識手段２０１で認識された１単語のテキストデータ形式である音素列単語から音素列を生成するものである。

具体的には、音素列生成手段２０２は、音素列単語から、音素以外の予め定めた文字（ここでは、“＋”）をスペースに置き換えて、音素列を生成する。例えば、音素列生成手段２０２は、音素列単語“ｓ＋ｅ＋ｋ＋ａ＋ｉ＋ｉ＋ｃｈ＋ｉ”を音素列“ｓ△ｅ△ｋ△ａ△ｉ△ｉ△ｃｈ△ｉ”に変換して出力する。

この音素列生成手段２０２が行う変換処理は、図１で説明した音素列単語生成手段１１の変換処理の逆変換に相当する。

【００６６】

以上説明したように音素認識装置２００を構成することで、従来、音響モデルにおけるトライフォンＨＭＭにより文脈として前後の音素の依存性で認識をしていた音素認識に対し、音素認識装置２００は、単語の繋がりを用いた、より長い文脈の依存性を考慮して音素認識を行う。

【００６７】

これによって、音素認識装置２００は、従来よりも精度よく音素認識を行うことができる。具体的には、従来技術の課題で説明したように、従来の音素認識の音素異なり率が２２．８％であったのに対し、音素認識装置２００は、音素異なり率を１．２％に改善することができた。

なお、音素認識装置２００は、図示を省略したコンピュータを、前記した各手段として機能させるプログラム（音素認識プログラム）で動作させることができる。

【００６８】

以上、本発明の実施形態について説明したが、本発明は、これらの実施形態に限定されるものではない。

ここでは、音素発音辞書５０の見出し語と音素言語モデル６０の接続対象とを、音素列単語生成手段１１（図１参照）が生成した単語別音素列のスペースを“＋”とした音素列単語とすることで、１単語分の音素列を１つの単語として扱うこととした。

【００６９】

しかし、音素列を１単語とみなす手法は、これに限定されるものではない。例えば、１単語分の音素列の末尾に音素以外の予め定めた文字（例えば、“￥”）を付加することとしてもよいし、単語分の音素列の前後に音素以外の予め定めた文字（例えば、“＜”，“＞”）を付加することとしてもよい。

この場合も、音素列生成手段２０２（図８参照）は、音素列単語生成手段１１（図１参照）が行った処理の逆変換を行えばよい。

【００７０】

また、ここでは、音素言語モデル生成手段１４が生成する音素言語モデル６０として、２－ｇｒａｍ言語モデルを例示した。

しかし、音素言語モデル生成手段１４は、Ｎ－ｇｒａｍ言語モデルであれば、１－ｇｒａｍ言語モデル、３－ｇｒａｍ言語モデル等であっても構わない。

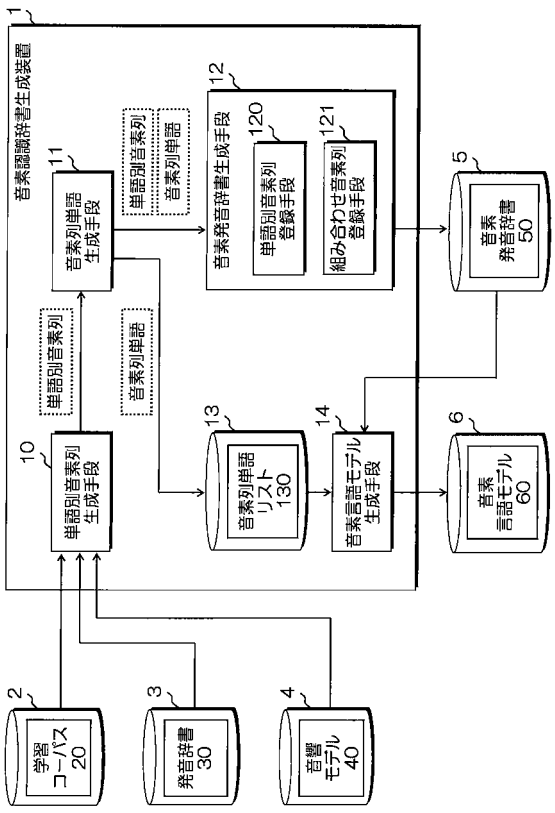
【符号の説明】

【００７１】

１　音素認識辞書生成装置
１０　単語別音素列生成手段
１１　音素列単語生成手段
１２　音素発音辞書生成手段
１２０　単語別音素列登録手段
１２１　組み合わせ音素列登録手段
１３　音素列単語リスト記憶手段
１３０　音素列単語リスト
１４　音素言語モデル生成手段
２　学習コーパス記憶装置
２０　学習コーパス
３　発音辞書記憶装置

３０　　発音辞書
４　　音響モデル記憶装置
４０　　音響モデル
５　　音素発音辞書記憶装置
５０　　音素発音辞書
６　　音素言語モデル記憶装置
６０　　音素言語モデル

【図１】



【図２】

(a)

20

世界一	短い	東京	の	橋	で	イベント	が	開か	れ	ました

(b)

30

見出し語（単語）	発音表記（音素列）
世界一	s△e△k△a△i△i△ch△i／s△e△k△a△i: △ch
短い	m△i△j△i△k△a△i
東京	t△o: △ky△o:／t△o△u△ky△o△u／t△o△o△ky△o△u
：	：

(c)

s△e△k△a△i△i△ch△i／m△i△j△i△k△a△i／t△o:△ky△o:／・・・

【図３】

a , a: , b , by , ch , d , dy , e , e: , f , g , gy ,
h , hy , i , i: , j , k , ky , m , my , n , ny ,
o , o: , p , py , r , ry , s , sh , t , ts , u , u: ,
w , y , z , N , Q

【図４】

50

見出し語（単語）	発音表記（音素列）
s+e+k+a+i+i+ch+i	$s_{△}e_{△}k_{△}a_{△}i_{△}i_{△}ch_{△}i$
m+i+j+i+k+a+i	$m_{△}i_{△}j_{△}i_{△}k_{△}a_{△}i$
t+o:+ky+o:	$t_{△}o:{}_{△}ky_{△}o:$
n+o	$n_{△}o$
h+a+sh+i	$h_{△}a_{△}sh_{△}i$
:	:
a	a
a:	a:
b	b
:	:
a+a	$a_{△}a$
a+a:	$a_{△}a:$
a+b	$a_{△}b$
:	:
a+a +a+a	$a_{△}a_{△}a_{△}a$
a+a +a+a:	$a_{△}a_{△}a_{△}a:$
a+a +a+b	$a_{△}a_{△}a_{△}b$
:	:

A

B

【図５】

130

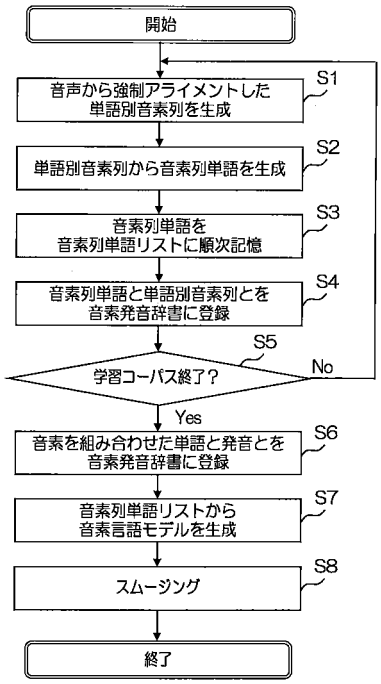
s+e+k+a+i+i+ch+i
m+i+j+i+k+a+i
t+o:+ky+o:
n+o
h+a+sh+i
d+e
i+b+e+N+t+o
g+a
h+i+r+a+k+a
r+e
m+a+sh+i
t+a
:

【図６】

60

$\log P(w_2 \mid w_1)$	w_1	w_2
-2.198956	s+o+r+e+d+e+w+a	h+i+r+o+sh+m+a
-1.394523	s+o+r+e+d+e+w+a	h+i+t+o+m+a+z+u
-8.217644	s+o+r+e+d+e+w+a	h+u+sh+i+N
:	:	:

【図７】



【図８】