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

特許第5366050号
(P5366050)

(45)発行日　平成25年12月11日(2013.12.11)　(24)登録日　平成25年9月20日(2013.9.20)

(51)Int.Cl.		F I		
G10L 15/06	(2013.01)	G10L	15/06	300C
G10L 15/187	(2013.01)	G10L	15/18	200D
G10L 15/197	(2013.01)	G10L	15/06	300Y

請求項の数 6　(全 24 頁)

(21)出願番号　特願2009-94212(P2009-94212)
(22)出願日　平成21年4月8日(2009.4.8)
(65)公開番号　特開2010-243914(P2010-243914A)
(43)公開日　平成22年10月28日(2010.10.28)
審査請求日　平成24年4月6日(2012.4.6)

特許法第３０条第１項適用　日本音響学会２００９年春季研究発表会講演論文集（平成２１年３月１０日、社団法人日本音響学会発行）の第２５～２６頁に発表

(73)特許権者　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　三村　正人
京都府京都市左京区吉田本町　国立大学法人京都大学学術情報メディアセンター内
(72)発明者　河原　達也
京都府京都市左京区吉田本町　国立大学法人京都大学学術情報メディアセンター内

審査官　山下　剛史

最終頁に続く

(54)【発明の名称】音響モデル学習装置、音声認識装置、及び音響モデル学習のためのコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
音声データベースを人間が書き起こし、整形して得られた文書スタイルテキストにより学習した言語モデルから、実際の発言内容に忠実な話し言葉スタイル書き起こしの言語モデルを推定するための言語モデル推定手段と、
予め準備された初期音響モデルと、前記言語モデル推定手段により推定された話し言葉スタイル書き起こしの言語モデルとを用いた音声認識により、前記音声データベースに書き起こしとその音素ラベルとを付すための音素ラベリング手段と、
前記音素ラベリング手段により音素ラベルが付された前記音声データベースを学習データとして、音声認識用音響モデルの学習又は更新を行なうための音響モデル学習手段とを含む、音響モデル学習装置。
【請求項２】
前記言語モデル推定手段は、
前記音声データベースの発話のターンごとに対応した文書スタイルテキストから、ターンごとのＮ－グラム言語モデルを作成するためのＮ－グラム作成手段と、
前記Ｎ－グラム作成手段により作成されたターンごとのＮ－グラム言語モデルの各々から、前記話し言葉スタイル書き起こしの話し言葉用Ｎ－グラム言語モデルを推定するための手段とを含み、
前記音素ラベリング手段は、
前記音声データベースのターンごとに、前記話し言葉用Ｎ－グラム言語モデルのうち、

対応するＮ－グラム言語モデルを選択するための言語モデル選択手段と、
　前記音声データベースの発話のターンごとに、前記言語モデル選択手段により選択されたＮ－グラム言語モデルと、前記初期音響モデルとを用いて音声認識を行なって、前記音声データベースのターンごとに書き起こしとその音素ラベルとを付与するための音声認識手段とを含む、請求項１に記載の音響モデル学習装置。
【請求項３】
前記音声データベースの一部の話し言葉スタイル書き起こしと、前記文書スタイルテキストのうちで当該一部に対応する部分とに基づいて作成された対応付けコーパスに基づいて、前記文書スタイルテキスト内の表現から前記話し言葉スタイル書き起こしの表現への変換を統計的に示す変換モデルを学習するための変換モデル学習手段をさらに含み、
　前記言語モデル推定手段は、ターンごとのＮ－グラム言語モデルの各々に対し、前記変換モデルを適用することにより、前記話し言葉スタイル書き起こしのＮ－グラム言語モデルを推定するための手段を含む、請求項１に記載の音響モデル学習装置。
【請求項４】
前記音声データベースは何らかの会議の音声を収録した審議音声コーパスであり、
　前記文書スタイルテキストは、前記会議の会議録である、請求項１～請求項３のいずれかに記載の音響モデル学習装置。
【請求項５】
所定の音声データベースを学習データとして、請求項１～請求項４のいずれかに記載の音響モデル学習装置により学習が行なわれた前記音声認識用音響モデルを記憶するための音響モデル記憶手段と、
　前記音響モデル記憶手段に記憶された前記音声認識用音響モデルと、音声認識用言語モデルとを用いて、入力される発話データに対する音声認識を行なうための音声認識手段とを含む、音声認識装置。
【請求項６】
コンピュータを、
　音声データベースを人間が書き起こし、整形して得られた文書スタイルテキストにより学習した言語モデルから、実際の発言内容に忠実な話し言葉スタイル書き起こしの言語モデルを推定するための言語モデル推定手段と、
　予め準備された初期音響モデルと、前記言語モデル推定手段により推定された話し言葉スタイル書き起こしの言語モデルとを用いた音声認識により、前記音声データベースに書き起こしとその音素ラベルとを付すための音素ラベリング手段と、
　前記音素ラベリング手段により音素ラベルが付された前記音声データベースを学習データとして、音声認識用音響モデルの学習又は更新を行なうための音響モデル学習手段として機能させる、音響モデル学習のためのコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
　この発明は音声認識技術に関し、特に、話し言葉の音声を精度高く認識可能な音声認識装置、及びそのための音響モデルの学習技術に関する。
【背景技術】
【０００２】
　近年、大語彙連続音声認識の主要な対象は、音声認識用に丁寧に発音した音声（以下「読上音声」と呼ぶ。）から、講演及び会議などの話し言葉の音声（以下「話し言葉音声」と呼ぶ。）に移行しつつある。
【０００３】
　話し言葉音声は読上音声では見られないような流暢でない現象を伴う。これらの現象とは、例えば、言直し、言いよどみ、「あー」とか「うー」というようなフィラーと呼ばれる発声の挿入、日本語の場合の助詞の欠落、及び発音の怠けなどである。
【０００４】

一般に、音声を統計的音声認識技術を用いて音声認識するためには、音響モデルが必要である。音響モデルの学習には、音声とその忠実な書き起こしとの組である音声コーパスを準備しなければならない。音声認識の精度を高めるためには、音声コーパスの規模は大きい方が望ましい。通常、こうした音声コーパスの作成は人手で行なわれる。しかし話し言葉音声の場合、上記したような現象のために人手による書き起こしの作成には多大なコストがかかる。したがって、大規模なコーパスの構築は極めて困難である。その結果、音声認識に必要な音響モデルの学習のためのデータ量不足が問題となる。

【０００５】

この問題に対処するため、Ｌａｍｅｌらは、非特許文献１において、ｌｉｇｈｔｌｙ ｓｕｐｅｒｖｉｓｅｄ ｔｒａｉｎｉｎｇ（以下「準教師付学習」と呼ぶ。）と呼ばれるアプローチを提案している。このアプローチでは、発話の忠実な書き起こしの代わりに、低コストで利用できる整形済テキストデータから音響モデルの学習のための音素ラベルを作成する。非特許文献１では、ニュース音声を対象として以下のように音素ラベルを付与することが提案されている。

【０００６】

多くの放送には、字幕が付与される。この字幕を放送に対するテキストデータとして音素ラベルを作成することが考えられる。しかし、非特許文献１によれば、字幕は多くの誤りを含み、そのままでは音素ラベルとして利用できない。そこで、非特許文献１では、字幕のテキストデータから学習した言語モデルを用いて音声認識を行なうことで、放送音声に対する音素ラベルを作成している。非特許文献１によれば、ニュース音声には音楽及びいわゆるＣＭなどの非音声区間が多数存在するため、音声認識結果の信頼性は高くない。そこで非特許文献１は、音声認識の後、その結果と字幕とを再度マッチングさせ、合致した区間の音声認識結果のみを用いるのが効果的であると報告している。

【０００７】

非特許文献２は、同様に放送音声を対象としているが、字幕には現れない表現にも対応するために、字幕から構築した言語モデルと、別途構築したベースライン言語モデルとを、前者に大きな重みをかけて合成し、この言語モデルを用いて音声認識を行なっている。非特許文献２は、作成された音素ラベルを用いた学習データの追加により、通常のＭＬ（最尤）学習だけでなく、識別学習の一種である音素誤り最小（ＭＰＥ：Ｍｉｎｉｍｕｍ Ｐｈｏｎｅ Ｅｒｒｏｒ）学習においても認識精度が向上したと報告している。

【先行技術文献】

【非特許文献】

【０００８】

【非特許文献１】Ｌ．ラメルら、「準教師付音響モデル学習の研究」、ＩＣＡＳＳＰ，Ｖｏｌ．１、ｐｐ．４７７－４８０、２００１年（L. Lamel et al. “Investigating lightly supervised acoustic model training.” In ICASSP, Vol. 1, pp. 477-480, 2001）

【非特許文献２】Ｈ．Ｙ．チャンら、「準教師付識別学習による放送ニュース書き起こしの改良」、ＩＥＥＥ－ＩＣＡＳＳＰ，Ｖｏｌ． １，ｐｐ．７３７－７４０、２００４年（H.Y. Chan et al., “Improving broadcast news transcription by lightly supervised discriminative training.” In IEEE-ICASSP, Vol. 1, pp. 737-740, 2004）

【非特許文献３】Ｐ．モーリック他、「ＥＰＰＳ録音に対する準教師付音響モデル学習」ＩＮＴＥＲＳＰＥＥＣＨ、ｐｐ．２２４－２２７，２００８年（M. Paulik et al., “Lightly supervised acoustic model training on epps recordings” In INTERSPEECH pp. 224-227, 2008）

【非特許文献４】秋田祐哉他、「統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換」、電子情報通信学会技術研究報告、ＳＰ２００５－１０８、ＮＬＣ２００５－７５（ＳＬＰ－５９－１９）、２００５．

【発明の概要】

【発明が解決しようとする課題】
【０００９】
近年、国会、地方議会などにおいて、音声認識を用いて会議録を作成しようとする試みがされている。これは、公的機関の業務について効率化及び経費節減が求められていること、会議録作成を担ってきた熟練速記者の数が減少していること、速記者の養成が難しい社会情勢となっていること、などが理由である。もちろんその背景には、高性能なコンピュータの普及及び音声認識技術の発達など、必要なハードウエア及びソフトウエアの充実という事情もある。
【００１０】
しかし、国会、特に委員会の質疑応答などは典型的な話し言葉であるため、既に述べたように音声コーパスの作成が困難である。その結果、話し言葉音声のための音響モデルの精度を高めることができず、音声認識の結果も芳しくないという問題がある。
【００１１】
非特許文献１及び非特許文献２の報告から考えて、準教師付学習は放送についての話し言葉音声認識に有効な技術であると考えられる。国会の委員会などでの発話は典型的な話し言葉であるから、準教師付学習によって学習した音響モデルを使用して音声認識を行なうことで会議録の作成を行なうことができる可能性が高い。
【００１２】
既に、非特許文献３に、欧州議会音声を対象とした、準教師付学習を用いた会議録作成が報告されている。非特許文献３では、欧州議会の会議録のテキストを用いた準教師付学習を、音声データに対する音素ラベルの作成に使用している。具体的には、人手により作成された会議録をそのまま用いて言語モデルを構築し、この言語モデルを用いて会議録に対応する音声の音声認識を行なって音素ラベルを作成している。この音素ラベルが付された音声を用いて音響モデルを構築し、新たな会議音声の音声認識を行なって会議録を作成する。
【００１３】
非特許文献３ではさらに、特定会議のテキストに大きな重みをかけて言語モデルを学習してその会議の音声の音声認識をすることで、全ての会議の会議録を一様に用いて学習した言語モデルを使用したときよりも高い精度の音素ラベルが得られたことが報告されている。
【００１４】
非特許文献３で報告されているように、人手により作成された会議録そのものを言語モデルとして使用して音素ラベルを付与したときの精度が満足すべき値となれば問題はない。しかし、以下に述べるように、特に日本の国会、地方議会などの会議録を作成するためには、解決すべき問題がある。
【００１５】
欧州議会の場合、日本の国会の本会議での発言に相当するものが多いため、発言が比較的丁寧に行なわれ、話し言葉特有の問題がそれほど生じない。その結果、欧州議会では、会議録と実際の発話との相違が小さく、会議録のテキストデータをそのまま言語モデルの作成に使用しても、音素ラベル付与の精度はそれほど低下しない。
【００１６】
しかし、日本の国会での議論は、本会議ではなく委員会を中心になされている。委員会での議論は、本会議と比較してよりインタラクティブであり、自発的な発話が主となる。特に、委員会での質問者は、簡単なメモを手にして考えながら、かつ答弁の内容を考慮しながら発言を行なうので、発話中に頻繁に言直し、ポーズ、及びフィラーの挿入などが発生する。答弁者の場合は、質問者と比較してそうした問題は少ないが、それでも本会議での発言と比較して話し言葉特有の問題が多く発生する。
【００１７】
現在、会議録の作成は速記者によって行なわれている。そのため、上記したような無意味な音声、言直し、発音の怠けなどが訂正され、書き言葉に近い表現に整形される。こう

した作業は知的に高度な作業であって、機械で再現することは非常にむずかしい。しかしそれだけに、実際の発話内容と会議録との間の相違が大きくなり、音響モデル作成のための音声データへの音素ラベル付与に会議録をそのまま使用するのは無理である。

【００１８】

しかし、会議録を全く使用しないで会議音声に音素ラベル付けをしようとすれば、前述したとおり人手により新たに書き起こしを行なう必要が生じ、膨大なコストがかかってしまう。そこで、既存の会議録を有効に使用しながら、大量の音声に対する効率的な音素ラベル付けを可能とする技術が求められている。こうした問題は、会議録に限らず、例えば大学・高校などにおける講義録又は講演録の作成など、整形済の書き起こしテキストデータが存在している話し言葉音声データのテキスト化を自動化する場合に共通した問題である。さらに、例えば裁判などで、撮影済の画像を参照する際、画像内の主な発言内容を文書化した後に、再度画像内の関連する箇所を検索したい、というような要求が発生することが考えられる。そのような場合にも、音声に効率的に音素ラベルを付与することができれば便利である。

【００１９】

また、話し言葉の場合、話者、話題の内容、周囲の音響的環境などがときにより変化していく場合がある。例えば内閣改造があった場合、国会で答弁に立つ閣僚は変わる。政権交代があれば、それまでの与野党が逆転することがありえるが、立場の変化に応じて発話スタイルが変化する可能性が高い。そうした場合には、書き起こし作成のための音響モデルについても、環境の変化に追従できるように簡単に更新できることが望ましい。従来は、そのように簡便に大量の話し言葉音声データに効率的に音素ラベルを付与する技術は存在していなかった。

【００２０】

それゆえに本発明の目的は、整形済のテキストデータが存在している話し言葉音声データのテキスト化のための音響モデルを、効果的に作成することが可能な音響モデル学習装置を提供することである。

【００２１】

本発明の他の目的は、整形済のテキストデータが存在している話し言葉音声データのテキスト化のための音響モデルについて、環境の変化に応じて簡単に更新することが可能な音響モデル学習装置を提供することである。

【課題を解決するための手段】

【００２２】

本発明の第１の局面に係る音響モデル学習装置は、音声データベースを人間が書き起こし、整形して得られた文書スタイルテキストにより学習した言語モデルから、実際の発言内容に忠実な話し言葉スタイル書き起こしのための言語モデルを推定するための言語モデル推定手段と、予め準備された初期音響モデルと、言語モデル推定手段により推定された話し言葉スタイル書き起こしの言語モデルとを用いた音声認識により、音声データベースに書き起こしとその音素ラベルとを付すための音素ラベリング手段と、音素ラベリング手段により音素ラベルが付された音声データベースを学習データとして、音声認識用音響モデルの学習又は更新を行なうための音響モデル学習手段とを含む。

【００２３】

この音響モデル学習装置では、言語モデル推定手段が、文書スタイルテキストにより学習した言語モデルから、話し言葉スタイル書き起こしのための言語モデルを推定する。この言語モデルと、初期音響モデルとを用い、音素ラベリング手段が発話のもとになった音声データベースに書き起こしとその音素ラベルとを付与する。音素ラベルが付与された音声データベースを学習データとして、音響モデル学習手段が音声認識用音響モデルの学習を行なう。

【００２４】

文書スタイルテキストにより学習した言語モデルから、話し言葉スタイル書き起こしのための言語モデルが推定される。この言語モデルを用いることにより、発話スタイルテキ

ストのもとになった音声データベースに書き起こしと音素ラベルとが付されるため、音声データベースの発話内容に、話し言葉特有の現象（言い淀み、繰返し、フィラーの挿入など）があったとしても、精度高く、発話音声に忠実に音声認識を行なうことができる。このように、発話音声に忠実にラベリングがされた音声データを学習データとして音声認識用音響モデルの学習を行なうため、この音声認識用音響モデルを用いて新たな発話データの音声認識を行なうときの精度を高めることができる。

【００２５】

好ましくは、言語モデル推定手段は、音声データベースの発話のターンごとに対応した文書スタイルテキストから、ターンごとのＮ－グラム言語モデルを作成するためのＮ－グラム作成手段と、Ｎ－グラム作成手段により作成されたターンごとのＮ－グラム言語モデルの各々から、話し言葉スタイル書き起こしの話し言葉用Ｎ－グラム言語モデルを推定するための手段とを含む。音素ラベリング手段は、音声データベースのターンごとに、話し言葉用Ｎ－グラム言語モデルのうち、対応するＮ－グラム言語モデルを選択するための言語モデル選択手段と、音声データベースの発話のターンごとに、言語モデル選択手段により選択されたＮ－グラム言語モデルと、初期音響モデルとを用いて音声認識を行なって、音声データベースのターンごとに書き起こしとその音素ラベルとを付与するための音声認識手段とを含む。

【００２６】

音声データベース内の発話の発声のスタイルは、発話者及び話題などにより変化する。ターンごとに話し言葉スタイル書き起こしの話し言葉用Ｎ－グラムを作成し、ターンごとにそのターンから得られた話し言葉用Ｎ－グラムを用いて音声認識を行なうことで、ターンごとの音声データベースの音素ラベリングの精度を高めることができる。その結果、音声認識用音響モデルの学習効率を高めることが可能になり、音声認識用音響モデルを用いた音声認識の精度を高めることができる。

【００２７】

より好ましくは、音響モデル学習装置は、音声データベースの一部の話し言葉スタイル書き起こしと、文書スタイルテキストのうちで当該一部に対応する部分とに基づいて作成された対応付けコーパスに基づいて、文書スタイルテキスト内の表現から話し言葉スタイル書き起こしの表現への変換を統計的に示す変換モデルを学習するための変換モデル学習手段をさらに含む。言語モデル推定手段は、ターンごとのＮ－グラム言語モデルの各々に対し、変換モデルを適用することにより、話し言葉スタイル書き起こしのＮ－グラム言語モデルを推定するための手段を含む。

【００２８】

音声データベースの一部の話し言葉スタイル書き起こしと、文書スタイルテキストのうちで対応する一部とから対応付けコーパスを作成すると、その対応付けコーパスから変換モデル学習手段が変換モデルを学習する。この変換モデルは、文書スタイルテキスト内の表現から話し言葉スタイル書き起こし内の表現への変換を統計的に示すものである。言語モデル推定手段は、ターンごとのＮ－グラム言語モデルの各々に対してこの変換モデルを適用して、話し言葉スタイル書き起こしのＮ－グラム言語モデルを作成する。

【００２９】

対応付けコーパス自体は、人手により作成することが想定される。しかし、このようにして得られた言語モデルを使用すると、対応付けコーパスを作成するもとになった音声データベースの一部だけでなく、その一部の音声データベースを含むより大きな音声データベースの音素ラベリングを自動的に行なうことができる。音声データベース全体について対応付けコーパスを作成する場合と比較して、より少ない労力で大量の音声データベースの音素ラベリングを、高精度に、かつ効率よく行なうことができる。

【００３０】

より好ましくは、音声データベースは何らかの審議の音声を収録した審議音声コーパスであり、文書スタイルテキストは、その審議の会議録である。

【００３１】

国会などの審議の音声には、話し言葉特有の現象（フィラー、言い淀みなど）が頻繁に出現し、しかも大量に存在する。そのため、音声データベースの音素ラベリングを手作業で行なうのは困難である。しかし審議中の発言を文書スタイルに整形した会議録が完備している。そこで、この会議録を文書スタイルテキストとし、審議音声データベースを音声データベースとして上記したような音声認識用音響モデルの学習を行なうことで、審議の音声を、効率よく、精度高く音声認識することが可能になる。

【００３２】

本発明の第２の局面に係る音声認識装置は、所定の音声コーパスを学習データとして、上記のいずれかの音響モデル学習装置により学習が行なわれた音声認識用音響モデルを記憶するための音響モデル記憶手段と、音響モデル記憶手段に記憶された音声認識用音響モデルと、音声認識用言語モデルとを用いて、入力される発話データに対する音声認識を行なうための音声認識手段とを含む。

【００３３】

本発明の第３の局面に係るコンピュータプログラムは、コンピュータを、音声データベースを人間が書き起こし、整形して得られた文書スタイルテキストにより学習した言語モデルから、実際の発言内容に忠実な話し言葉スタイル書き起こしの言語モデルを推定するための言語モデル推定手段と、予め準備された初期音響モデルと、言語モデル推定手段により推定された話し言葉スタイル書き起こしの言語モデルとを用いた音声認識により、音声データベースに書き起こしとその音素ラベルとを付すための音素ラベリング手段と、音素ラベリング手段により音素ラベルが付された音声データベースを学習データとして、音声認識用音響モデルの学習又は更新を行なうための音響モデル学習手段として機能させる。

【図面の簡単な説明】

【００３４】

【図１】本発明の第１の実施の形態に係る会議録作成システム３０のブロック図である。

【図２】図１に示す審議音声コーパス４０と会議録４２との対応関係を模式的に示す図である。

【図３】図１に示す音素ラベリング処理部７８のブロック図である。

【図４】本発明の実施の形態で使用される対応付けコーパスの内容の一部を示す模式図である。

【図５】話し言葉／書き言葉の変換モデルを学習する処理部を実現するコンピュータプログラムの制御構造を示すフローチャートである。

【図６】ターンごとにＮ－グラムを作成する処理部及びＮ－グラムの書き言葉から話し言葉への変換を行なう処理部を実現するコンピュータプログラムの制御構造を示すフローチャートである。

【図７】第１の実施の形態に係る会議録作成システムを構成するコンピュータの関係を模式的に示す図である。

【図８】図７に示す会議録作成システムにおいて、音響モデル作成用のコンピュータの外観図である。

【図９】図８に示すコンピュータのハードウエア構成を示すブロック図である。

【図１０】図７に示す会議録作成システムにおいて、会議録作成用に使用されるコンピュータの外観図である。

【発明を実施するための形態】

【００３５】

以下の説明では、同一部品には同一の参照番号を付してある。それらの名称及び機能も同一である。したがって、それらについての詳細な説明は繰返さない。また、以下に述べる実施の形態では、Ｎ－グラムとしてユニグラム、バイグラム、及びトライグラムを用いている。

【００３６】

［実施の形態の原理］

本実施の形態では、以下の考え方によって、国会審議音声の自動書き起こしシステム（会議録作成システム）を構築している。日本の国会では、前述したとおり、欧州議会と異なり議論は主として委員会で行なわれる。そのため、欧州議会の審議よりもインタラクティブで自発的な発話が主となる。そうした発話には、多くのフィラー、言いよどみ、繰返しなどが含まれる。人手で作成された審議録では、そのような流暢でない発話も流暢な発話に「翻訳」されている。すなわち、日本では、実際の発話内容と会議録との相違が大きい。したがって、会議録をもとに音素ラベルを作成する処理はそのままでは難しく、話し言葉特有の現象にいかに適切に対応するかが問題となる。

【００３７】

国会審議音声における実際の発話と会議録との例を図２に示す。

【００３８】

図２には、実際の発話からなる審議音声コーパス４０と、対応する会議録４２とを対比して示してある。審議音声コーパス４０は、たとえば国会の審議の音声を収録したものであって、音声データベースを構成している。発話１００と、会議録１１０、発話１０２と会議録１１２、及び発話１０４と会議録１１４がそれぞれ対応している。

【００３９】

図２から分かるように、会議録では助詞「が」の挿入、並びに「いー」、「えー」、及び「あのー」などのフィラーの除去による整形が行なわれている。いわば話し言葉から書き言葉への変換が行なわれている。

【００４０】

このような話し言葉（発言の内容の忠実な書き起こし）と、整形済文書（会議録）との対応付けコーパスから、言語モデルのスタイル変換のための統計的モデルを構築する枠組みが、非特許文献４で提案されている。以下に述べる実施の形態では、この統計的な言語モデル変換を、個々の会議録に適用することにより、書き言葉の言語モデルから話し言葉の言語モデルを構築し、この言語モデルを用いて音声認識を行なうことにより、話し言葉に対する音素ラベルを作成する。

【００４１】

言語モデルの統計的スタイル変換では、統計的機械翻訳の枠組みに基づき、話し言葉スタイルＶと文書スタイルＷとの変換を行なう。この変換は双方向的である。すなわち、話し言葉の書き起こしから文書スタイルへ整形を行なう方向へも、文書スタイルのテキストから書き起こしを復元する方向へもこの変換モデルを適用することができる。

【００４２】

デコードは、統計的機械翻訳の枠組みにしたがい、次のベイズ則に基づいて行なわれる。

【００４３】

【数１】

$$p(W|V)=\frac{p(W)\cdot p(V|W)}{p(V)} \qquad (1)$$

$$p(V|W)=\frac{p(V)\cdot p(W|V)}{p(W)} \qquad (2)$$

【００４４】

この式において、ｐ（Ｗ）は文書スタイルのＮ－グラム確率、ｐ（Ｖ）は話し言葉スタイルのテキストＶのＮ－グラム確率、ｐ（Ｗ｜Ｖ）は話し言葉スタイルのテキストＶに対する文書スタイルのテキストＷの条件付確率、ｐ（Ｖ｜Ｗ）は文書スタイルのテキストＷに対する話し言葉スタイルのテキストＶの条件付確率を、それぞれ示す。各式の分母は通常は無視される。

【００４５】

ここで重要なのは、式（２）により話し言葉スタイルのテキストＶを一意に決定するのは、テキストＶが多様であり得るため、式（１）により整形を行なうプロセスよりもはるかに難しい点である。例えば、式（２）においてフィラーはランダムに挿入され得る（つまり、フィラーを含む話し言葉スタイルのテキストＶの形式が多様であり得る）が、式（１）においてはフィラーは確率１で除去される（すなわち、話し言葉スタイルのテキストＶ中のフィラーは文書スタイルのテキストＷへの変換の際に確実に除去される。）と考えてよい。したがって、話し言葉スタイルのテキストＶを一意に復元することよりも、次の式（３）のように話し言葉スタイルのテキストＶの統計的言語モデルを推定することの方が有意義である。

【００４６】

【数２】

$$p(V) = p(W) \cdot \frac{p(V|W)}{p(W|V)} \qquad (3)$$

【００４７】

重要な点は、文書スタイルのテキストＷは話し言葉を忠実に書き起こしたテキストＶよりも豊富に存在する点である。すなわち、式（３）にしたがえば、豊富な文書スタイルのテキストを用いて話し言葉音声認識のための言語モデルｐ（Ｖ）をロバストに推定できる。

【００４８】

実際の変換は、次式のようにＮ－グラム計数を操作することで行なわれる。

【００４９】

【数３】

$$N_{gram}(v_1^n) = N_{gram}(w_1^n) \cdot \frac{p(v|w)}{p(w|v)} \qquad (4)$$

【００５０】

ｖ及びｗは、各スタイルにおける変換パターンである。式（４）により、置換ｗ→ｖ、ｗの脱落、ｖの挿入を文脈を考慮してモデル化することができる。条件付確率ｐ（ｖ｜ｗ）及びｐ（ｗ｜ｖ）は、書き起こしと文書スタイルテキストとの対応付けコーパスから統計的に推定される。より具体的には、これら条件付確率条件付確率ｐ（ｖ｜ｗ）及びｐ（ｗ｜ｖ）は、コーパス中の各パターンの出現回数から推定される。

【００５１】

適切なモデルとなるように、パターンの隣接単語も考慮する。例えば、フィラー「あー」は、｛ｗ＝（w_{-1}、w_{+1}）→ｖ＝（w_{-1}，あー，w_{+1}）｝のようにモデル化される。品詞情報を用いたスムージングを行なうと、データのスパースネスに対応することができる。

【００５２】

［第１の実施の形態］

図１を参照して、本発明の第１の実施の形態に係る会議録作成システム３０は、一般的には音声認識システムであって、審議音声コーパス４０と、審議音声コーパス４０に対応する会議録４２とから、審議音声５４を音声認識することによって書き起こし５６を出力するためのものである。この実施の形態は、前記した言語モデルの統計的スタイル変換（書き言葉→話し言葉）を、音響モデルの準教師付学習に適用したものである。国会では、収録した音声データによる大規模なアーカイブが作成されている。これらの音声に対しては、人手による書き起こしは付与されていないが、整形済の会議録が利用可能である。したがって、会議録をもとに音素ラベルを自動で作成できれば、豊富な音声データがそのまま音響モデルの学習データとして利用できることになる。

【００５３】
図１を参照して、この目的のために、会議録作成システム３０においては、審議音声コーパス４０の一部である部分コーパス６８から作成した忠実な書き起こし７０と、会議録４２のうち部分コーパス６８に対応する部分会議録７２とから、手作業の対応付けコーパス作成処理７４により、最初に対応付けコーパス７６を作成する。部分コーパス６８と部分会議録７２とは互いに対応付けられている。すなわち、部分コーパス６８に含まれる音声に対し、部分会議録７２のテキストデータを構成する文字・記号が予め割当てられている。書き起こし７０により、部分コーパス６８に音素ラベルを付与できる。
【００５４】
会議録は、予算委員会、法務委員会などの会議毎に作成されるが、各発言には会議内の話者ＩＤが付与されており、それにしたがってターン毎のテキストが抽出できる。各会議はおよそ２時間から５時間の長さであり、各ターンは１０秒から３分程度（平均１分）の長さである。ここで「ターン」とは、ある話者がまとめて話したひとまとまりの発話のことをいう。例えば質問者が質問を発したときの発話で１ターン、答弁者がその質問に答弁して次の１ターン、などのように一連の発話が複数のターンに分割される。同一の話者による連続した発話でも、話題が異なれば別ターンとされている。図２に示す発話１００、１０２及び１０４はそれぞれ１ターンとなっている。それに対応する会議録１１０、１１２及び１１４もターンごとに読出すことができる。
【００５５】
本実施の形態では、音素ラベル付与のための音声認識の際に言語モデルとして使用されるＮ－グラムが、より強い制約となるように、多くの話者又は話題を含む会議全体ではなく、個々のターンごとにＮ－グラムを作成する。本実施の形態に係る手法では、個々のＮ－グラムのサイズが大きくならないので、ターンのような詳細な単位ごとにＮ－グラムを用意することが可能である。その上、ベースライン言語モデルを音声認識に使用する場合のように、余計な表現が混入する可能性が極めて低いという利点がある。
【００５６】
対応付けコーパス作成処理７４は、部分コーパス６８の書き起こし７０を作成した後、書き起こしの各単語を部分会議録７２の単語と対応付ける処理である。この処理は手作業である。しかし、対応付けコーパス７６は、審議音声コーパス４０の一部（部分コーパス６８）及び会議録４２の一部（部分会議録７２）のみに対応するものである。したがって、対応付けコーパス７６を作成するための作業量は、審議音声コーパス４０の全体を書き起こす場合と比較してはるかに小さくてよい。
【００５７】
なお、本実施の形態ではＮ－グラムを言語モデルとして使用するため、対応付けコーパス７６の作成において、ポーズの取扱いに注意する必要がある。音声データではポーズが挿入されていても、会議録ではポーズはそのままで挿入されているわけではなく、句読点の形で挿入されていることが多いためである。ポーズの取扱い方には種々あるが、本実施の形態では「、」はショートポーズ（＜ｓｐ＞）、「。」は無音区間（＜ｓｉｌ＞）として取扱っている。対応付けコーパス７６の作成時には、このようにしてポーズの標記を統一している。
【００５８】
会議録作成システム３０は、このようにして作成された対応付けコーパス７６を用い、式（４）によって書き言葉用の言語モデルを話し言葉用の言語モデルに変換する変換モデル１２２を推定するための話し言葉／書き言葉変換モデル学習部１２０と、この変換モデル１２２を使用して、審議音声コーパス４０から話し言葉の音声認識に対応した音響モデル４８の学習を行なうための音声認識用音響モデル学習部４４と、会議録４２の全体から音声認識用の統計的言語モデル５８の学習を行なうための言語モデル学習部４６と、変換モデル１２２を使用して、会議録４２から学習された書き言葉用の言語モデル５８を話し言葉用の言語モデル５０に変換するための言語モデル変換部６０と、各々話し言葉用に適応化された音響モデル４８及び言語モデル５０を用い、審議音声５４を音声認識して認識

結果を書き起こし５６として出力するための音声認識装置５２とを含む。

【００５９】

具体的には、話し言葉／書き言葉変換モデル学習部１２０は、部分会議録７２に出現するＮ－グラムの各々について、書き起こし７０内の対応する部分がどのように変化しているかを調べ、その結果を計数する。例えば部分会議録７２中にｗ＝「＜ｓｐ＞この　法案」（＜ｓｐ＞はショートポーズを表す。）が５００回出現し、書き起こし７０ではそのうち５０回がｖ＝「＜ｓｐ＞えー　この　法案」となっていた（フィラー「えー」が挿入された）とすれば、ｐ（ｖ｜ｗ）＝５０／５００となる。このような計数を、全てのＮ－グラムとその変化形とについて集計することで、式（４）にしたがった変換モデル１２２が得られる。この集計により得られるのは、どのような変化が何回あったかを示す計数である。この値は、文書スタイルの表現が話し言葉スタイルのどのような表現にどのような確率で変化するかを示す確率と同視することができる。

【００６０】

音声認識用音響モデル学習部４４は、審議音声コーパス４０、音素ラベル付部分コーパス６８、及び変換モデル１２２を用いた音声認識により審議音声コーパス４０の音声に対して音素ラベルを付す処理を行ない、音素ラベル付音声データベース８０を出力するための音素ラベリング処理部７８と、音素ラベル付音声データベース８０を学習データとして、通常の学習方法により話し言葉用の音響モデル４８の学習を行なうための音響モデル学習部８２とを含む。

【００６１】

図３を参照して、音素ラベリング処理部７８は、音素ラベル付部分コーパス６８から初期音響モデル１３２の学習を行なうための初期音響モデル学習部１３０と、会議録４２のターンごとに会議録４２のテキストデータからＮ－グラム統計データを作成することにより、ターンごとＮ－グラム１８６を作成するためのターンごとＮ－グラム作成部１８４と、ターンごとＮ－グラム１８６の各々に含まれるＮ－グラムの確率に対し、変換モデル１２２により定まる、式（４）により表現される変換を行なうことによって話し言葉用Ｎ－グラム１３６を出力するためのＮ－グラム変換部１８８とを含む。

【００６２】

ターンごとＮ－グラム作成部１８４は、各ターンの会議録のテキストからＮ－グラムエントリの抽出とそれらの出現回数との計数を行なう。この結果、ターンごとにターンごとＮ－グラム１８６が得られる。ターンごとＮ－グラム１８６内の各エントリについて、変換モデル１２２を適用することによって話し言葉用Ｎ－グラム１３６がターンごとに得られる。

【００６３】

音素ラベリング処理部７８はさらに、審議音声コーパス４０内の各ターンを順番に選択し、ターンを特定する情報と、選択されたターンの音声とを出力するためのターン・音声選択部１３８と、ターン・音声選択部１３８が選択したターンを示す情報を受け、話し言葉用Ｎ－グラム１３６の中から、そのターンに対応するＮ－グラム１４２を選択するためのＮ－グラム選択部１４０と、初期音響モデル１３２及びＮ－グラム１４２を用い、特にＮ－グラム１４２を言語モデルとして用いて、ターン・音声選択部１３８の出力した発話音声の音声認識を行なって、その音声に、単語レベル及び音素レベルの認識結果を付して音素ラベル付音声データベース８０に出力するための音声認識装置１４４とを含む。

【００６４】

音声認識装置１４４には、既存の統計的音声認識装置を用いることができる。ここでは単語レベル及び音素レベルの認識結果を出力するものを用いるが、音素レベルの結果のみを出力するものでもよい。音声認識装置１４４は、発話中のポーズにより、最長で３０秒程度の短い発話区間に分割した形で認識結果の付された音声データを出力する。以降の学習はこの区間を単位として行なう。

【００６５】

このようにして得られた音素ラベル付音声データベース８０の各音素ラベルは、話し言

葉には出現するが文書スタイルでは出現しないような音素列の出現確率を考慮して決定されている。しかもターンごとに、そのターンのみについて学習されたＮ－グラムを用いているため、音声認識の精度、すなわち付与される音素ラベルの精度は高くなる。その上、審議音声コーパス４０に大量の音声が存在する場合にも、その全てに対して、自動的に高精度で音素ラベルを付与することができる。

【００６６】

したがって、この音素ラベル付音声データベース８０から、図１に示す音響モデル学習部８２によって通常の方法で音響モデル４８を作成すると、音声認識装置５２による認識結果の精度が高くなることが十分に期待できる。

【００６７】

一方、音声認識装置５２が使用する言語モデル５０も、会議録４２中に出現するＮ－グラムについて、変換モデル１２２を適用して得られたものであり、話し言葉に特有の音素列の発生確率が算入されたものである。

【００６８】

このように、話し言葉特有の音素列の発生確率を考慮して得られた音響モデル４８及び言語モデル５０を使用するため、音声認識装置５２は、話し言葉においてよく発生する事象、すなわちフィラーの挿入、言い淀み、発音の怠けなどにもかかわらず、審議音声コーパス４０の高精度な書き起こしを出力することができる。

【００６９】

図４は、対応付けコーパス７６中の２つの文例を示す。図４において、審議音声コーパス４０では発話されているが会議録４２では削除されている音声を図４（Ａ）の発話１６０の先頭の「{えー}」のように中カッコ{　}で囲んで示してある。審議音声コーパス４０では発話されていないが会議録４２では挿入されている音声は、図４（Ｂ）の発話１６２内の「いただいて（い）るつもりで…」のようにカッコ（　）で囲んで示してある。審議音声コーパス４０の発話での表現が会議録４２では他の表現に変えられている部分は、発話１６０内の「{んで／ので}」のように、全体を中カッコで囲み、審議音声コーパス４０での表現を「／」の前に、会議録４２での表現を「／」の後に、それぞれ示してある。

【００７０】

この対応付けコーパスは、書き起こし７０と部分会議録７２とを別の言語によるものと考えたときの翻訳モデル作成のためのパラレルコーパスと考えることができる。通常、翻訳モデルでは、単語の挿入、削除、置換に加え、順序の入替えという編集を考えるが、ここでは言語自体は同一限度であるため、順序の入替えは考えていない。

【００７１】

［話し言葉／書き言葉変換モデル学習部１２０のプログラム構造］

図５を参照して、話し言葉／書き言葉変換モデル学習部１２０による変換モデル１２２の学習処理を実現するコンピュータプログラムは、利用者からの処理開始の指示に応答してプログラムの実行を開始し、記憶領域の確保、変数のクリアなどの初期設定を行なうステップ１９０と、対応付けコーパス７６のファイルをオープンするステップ１９２と、繰返し変数ｉに０を代入するステップ１９４とを含む。

【００７２】

繰返し変数ｉは、対応付けコーパス７６のうち、処理対象となっている単語の位置を示す変数であり、０から１ずつ増加する。以下、変数ｉによって示される位置の単語を「単語（ｉ）」と書く。

【００７３】

このプログラムはさらに、変数ｉの値が対応付けコーパス７６中の全単語の数より大きくなったか否かを判定し、判定結果に応じて制御の流れを分岐させるステップ１９６と、ステップ１９６の判定結果がＮＯのときに実行され、対応付けコーパス７６の中で、部分会議録７２の単語（ｉ）を先頭とするユニグラム、バイグラム、及びトライグラムの計数にそれぞれ１ずつ加算するステップ１９８と、変数ｉに１を加算して制御をステップ１９６に戻すステップ２００とを含む。ステップ１９６からステップ２００の処理を、対応付

けコーパス７６中の全単語に対して実行することにより、部分会議録７２のＮ－グラムモデルが作成される。

【００７４】

このプログラムは更に、ステップ１９６での判定結果がＹＥＳのときに実行され、対応付けコーパス７６の読出位置を先頭に再設定するステップ２０２と、ステップ２０２に続き、部分会議録７２で計算されたユニグラム、バイグラム、トライグラムの各々について、書き起こし７０ではどのように変化しているかを集計することにより、変換モデル１２２を計算するステップ２０４と、ステップ２０４で計算された変換モデル１２２をファイルとして出力し、処理を終了するステップ２０６とを含む。

【００７５】

［ターンごとＮ－グラム作成部１８４及びＮ－グラム変換部１８８のプログラム構造］

図６を参照して、ターンごとＮ－グラム作成部１８４及びＮ－グラム変換部１８８を実現するためのコンピュータプログラムは、プログラムの実行開始とともに、必要な記憶領域の確保及び初期化などの初期設定を行なうステップ２１０と、繰返し変数ｉに０を代入するステップ２１２と、繰返し変数ｉを処理対象の部分会議録７２に含まれるターン数と比較することにより、全ターンの処理が終了したか否かを判定し、判定結果により制御の流れを分岐させるステップ２１４とを含む。

【００７６】

このプログラムはさらに、ステップ２１４の判定結果がＮＯの場合に実行され、ターン（ｉ）の会議録を部分会議録７２から読出すステップ２１６と、ステップ２１６で読出されたターン（ｉ）の会議録のＮ－グラムを作成し、所定の記憶媒体に出力するステップ２１８と、ステップ２１８に続き、繰返し変数ｉの値に１を加算し、制御をステップ２１４に戻すステップ２２０とを含む。

【００７７】

このプログラムはさらに、ステップ２１４の判定結果がＹＥＳの場合に実行され、変換モデル１２２を外部記憶媒体から主記憶装置に読出すステップ２２２と、繰返し変数ｉに０を代入するステップ２２４と、繰返し変数ｉの値と部分会議録７２に含まれるターン数との比較により、部分会議録７２の内の全ターンの会議録についてＮ－グラムの変換（文書スタイル→話し言葉スタイルの変換）を行なったか否かを判定し、判定結果に応じて制御の流れを分岐させるステップ２２６と、ステップ２２６において、部分会議録７２の内の会議録についてのＮ－グラムの変換が完了していないと判定されたことに応答して実行され、ターン（ｉ）のＮ－グラムの全てについて、変換モデル１２２を適用することにより話し言葉スタイルにおける確率の推定値を再計算し更新するステップ２３０と、繰返し変数ｉに１を加算して制御をステップ２２６に戻すステップ２３２とを含む。

【００７８】

［コンピュータシステムによる実現］

上に構造を説明した会議録作成システム３０は、実質的にはコンピュータにより実現される。会議録作成システム３０の全体を１台のコンピュータ上に実装することも可能である。しかし、音響モデル４８及び言語モデル５０は大量の審議音声コーパス４０及び会議録４２を使用して学習するものであるのに対し、会議録作成には審議音声コーパス４０及び会議録４２は不要である。したがって、両者を分離する方がメンテナンス上都合がよい。また、変換モデルの学習及び音響モデルの学習は、システムの性能に大きな影響を及ぼすため、システムのユーザではなく、システムの管理者又は行なう方が好ましい。

【００７９】

したがって、本実施の形態に係る会議録作成システム３０は、図７に示されるように、音響モデル４８及び言語モデル変換部６０の学習を行なう学習用コンピュータシステム２５０と、コンピュータシステム２５０により学習が行なわれた音響モデル４８及び言語モデル５０を使用して、審議音声を音声認識し書き起こしを出力する処理を行なう会議録作成用コンピュータシステム３００とを含む。当業者には容易に分かるように、会議録作成用コンピュータシステム３００を複数使用すれば、共通の音響モデル４８及び言語モデル

５０を用いて、複数の委員会における審議の会議録を作成することができる。

【００８０】

図８を参照して、学習用コンピュータシステム２５０は、コンピュータ２６０と、いずれもコンピュータ２６０に接続されるモニタ２６２、キーボード２６６、マウス２６８、マイクロホン２９０及び一対のスピーカ２５８とを含む。コンピュータ２６０には、ＤＶＤ（Ｄｉｇｉｔａｌ　Ｖｅｒｓａｔｉｌｅ　Ｄｉｓｃ）の再生及び記録が可能なＤＶＤドライブ２７０と、所定の規格にしたがった半導体メモリ記憶装置が装着可能なメモリポート２７２とが備えられている。コンピュータ２６０の内部構成については図９を参照して後述する。

【００８１】

図９を参照して、コンピュータ２６０は、図８に示すＤＶＤドライブ２７０及びメモリポート２７２に加え、ＣＰＵ（中央演算処理装置）２７６と、ＣＰＵ２７６に接続されたバス２８６と、いずれもバス２８６に接続されたＲＯＭ（読出専用メモリ）２７８、ＲＡＭ（ランダムアクセスメモリ）２８０、大容量ハードディスク２７４、ネットワークインターフェイス２９６、及びサウンドボード２８８を含む。

【００８２】

ＤＶＤドライブ２７０には、ＤＶＤ２８２が装着される。メモリポート２７２には半導体メモリ２８４が装着される。ＣＰＵ２７６は、バス２８６並びにＤＶＤドライブ２７０及びメモリポート２７２をそれぞれ介して、ＤＶＤ２８２及び半導体メモリ２８４をアクセスし、データの読出及び書込を行なえる。

【００８３】

キーボード２６６、マウス２６８、モニタ２６２は、いずれも図示しないインターフェイスを介してコンピュータ２６０のバス２８６に接続される。スピーカ２５８及びマイクロホン２９０は、サウンドボード２８８に接続される。

【００８４】

上記実施の形態における審議音声コーパス４０、会議録４２、部分コーパス６８、書き起こし７０、部分会議録７２、対応付けコーパス７６、変換モデル１２２、音素ラベル付音声データベース８０、音響モデル４８、言語モデル５０及び５８等は、ＲＡＭ２８０、大容量ハードディスク２７４、ＤＶＤ２８２、半導体メモリ２８４のいずれでも実現できる。実際には、格納するデータの容量、読出し、書込みに要求される速度などによって、最も効率のよい記憶装置が各記憶部を実現するために選択される。本実施の形態では、これらは大容量ハードディスク２７４に記憶され、利用時にＲＡＭ２８０にロードされる。

【００８５】

図１０を参照して、本実施の形態に係る会議録作成システム３０で用いられる会議録作成用コンピュータシステム３００は、コンピュータ３１０と、いずれもコンピュータ３１０に接続された、モニタ３２０、キーボード３２２、マウス３２４、マイク３２８及び一対のスピーカ３２６とを含む。図示していないが、コンピュータ３１０にはヘッドホン接続端子が設けられており、ヘッドホンによる音声の再生を行なうこともできる。コンピュータ３１０には、図１に示す音声認識装置５２を実現するための音声認識プログラムと、この音声認識プログラムにより出力される審議録ファイルを編集するための編集プログラムとが予めインストールされている。さらに、コンピュータ３１０は、大容量のＨＤＤを持ち、コンピュータシステム２５０からネットワークを介して受信した音響モデル４８及び言語モデル５０をこのＨＤＤに記憶することができる。

【００８６】

会議録作成用コンピュータシステム３００のハードウエア構成は、図９に示すものと同様である。したがってここではその詳細については繰返さない。

【００８７】

［動作］

上に構成を説明した会議録作成システム３０は以下のように動作する。会議録作成システム３０の動作はいくつかのフェーズに分けられる。以下、それらフェーズを順番に説明

する。
【００８８】
－対応付けコーパス７６の作成－
図１を参照して、最初に、既存の審議音声コーパス４０及び会議録４２から、コンピュータシステム２５０において対応付けコーパス７６が作成される。手作業により、部分コーパス６８が審議音声コーパス４０から抽出され、対応する部分会議録７２が会議録４２から抽出される。部分コーパス６８を再生し、手作業により審議音声の忠実な書き起こし７０をターンごとに作成する。このようにして作成された書き起こし７０と部分会議録７２とから、これも人手による対応付けコーパス作成処理７４が行なわれ、対応付けコーパス７６が作成される。
【００８９】
ここでは、書き起こし７０を一旦作成してから対応付けコーパス７６を作成するが、部分コーパス６８を再生しながら、部分会議録７２を画面で直接編集することにより対応付けコーパス７６を作成してもよい。
【００９０】
完成した対応付けコーパス７６は大容量ハードディスク２７４に格納される。
【００９１】
－変換モデル１２２の作成－
対応付けコーパス７６は、話し言葉スタイルの部分コーパス６８の忠実な書き起こしと、整形済の（文書スタイルの）部分会議録７２とが対になったものであり、本実施の形態では図４に示すような形式となっている。話し言葉／書き言葉変換モデル学習部１２０は、この対応付けコーパス７６のうち、部分会議録７２の部分について通常のＮ－グラムを作成する（図５、ステップ１９６－２００）。さらに話し言葉／書き言葉変換モデル学習部１２０は、このＮ－グラムの各エントリについて、書き起こし７０内の対応部分を調べ、変化しているものがあればその数をそれぞれ計数し、全て計数した時点で、各エントリに対する変化形ごとにその割合を算出することで変換モデル１２２を得る（ステップ２０４）。
【００９２】
この処理は例えば以下のように行なう。部分会議録７２内に、Ｎ－グラムのトライグラムｗ＝「＜ｓｐ＞この　法案」が５００回出現し、書き起こし７０ではそのうち５０回がｖ＝「＜ｓｐ＞えー　この　法案」となっていたとする。この場合、ｐ（ｖ｜ｗ）＝５０／５００となる。話し言葉／書き言葉変換モデル学習部１２０はｖの生起回数（上の場合、５０）を計数する。他にトライグラムｗ＝「＜ｓｐ＞この　法案」の変形がなかったとすれば、文書スタイルのトライグラムｗ＝「＜ｓｐ＞この　法案」が全部で５００あれば、それに対応する話し言葉スタイルの表現の生起回数は、「＜ｓｐ＞えー　この　法案」が５０、「＜ｓｐ＞この　法案」が４５０（＝５００－５０）となる。
【００９３】
話し言葉／書き言葉変換モデル学習部１２０は、このようにして、対応付けコーパス７６から得られるＮ－グラムの各エントリに対し、その変形ごとに書き起こし７０内での発生回数を計数する。この計数結果に基づき、式（４）の変換係数が、書き起こし７０中に出現する話し言葉スタイルの各Ｎ－グラムについて算出される。これらにより変換モデル１２２が得られる。得られた変換モデル１２２はＨＤＤに出力され記憶される（図５、ステップ２０６）。
【００９４】
－審議音声コーパス４０の音素ラベリング処理－
以上のようにして変換モデル１２２が得られると、審議音声コーパス４０について以下のようにして音素ラベルが付与できる。
【００９５】
最初に、図３に示されるように部分コーパス６８及び部分会議録７２を用い、初期音響モデル学習部１３０によって、通常の方法で初期音響モデル１３２の学習が行なわれる。

次いで、会議録４２の各ターンに対し、ターンごとＮ－グラム１８６（図３参照）がターンごとＮ－グラム作成部１８４により得られる（図６、ステップ２１４－２２０）。得られたターンごとＮ－グラム１８６に対して、Ｎ－グラム変換部１８８が変換モデル１２２を適用することにより、各ターンについて話し言葉用Ｎ－グラム１３６が得られる。

【００９６】

ターン・音声選択部１３８は、審議音声コーパス４０の各ターンを順番に選択してターン情報をＮ－グラム選択部１４０に与える。Ｎ－グラム選択部１４０は、与えられたターン情報に応じ、話し言葉用Ｎ－グラム１３６の中で、選択されたターンから得られた話し言葉用Ｎ－グラムを選択し、Ｎ－グラム１４２として音声認識装置１４４に与える。一方、ターン・音声選択部１３８は、選択されたターン中の音声データを音声認識装置１４４に与える。

【００９７】

音声認識装置１４４は、Ｎ－グラム１４２を言語モデルとして用い、初期音響モデル１３２を使用して、審議音声コーパス４０から選択された音声に対する音声認識を行ない、音声認識結果を音素ラベルとして審議音声コーパス４０の音声データに付与する。音声認識装置１４４による音声認識では、ターンごとにそのターンから得られた話し言葉用に変換したＮ－グラム１４２が言語モデルとして使用される。そのため、審議音声コーパス４０の各ターンについて、話された際の音声に忠実な音声認識結果が得られる。すなわち、音素ラベリング処理部７８により音素ラベルが付与された音素ラベル付音声データベース８０は、話し言葉の発音に忠実な、精度の高い音素ラベルを有した音声コーパスとなる。しかも、審議音声コーパス４０に含まれる全ての音声に対し、このようにして自動的に音素ラベルを付与することができる。

【００９８】

－音響モデル４８の学習－

上記のように得られた音素ラベル付音声データベース８０は、話し言葉に忠実な音素ラベルが付与された音声コーパスである。したがってこの音素ラベル付音声データベース８０を使用した学習を行なうことにより、話し言葉を音声認識するのに適した音響モデル４８が得られる。音素ラベル付音声データベース８０が話し言葉に忠実な音素ラベルを有しているため、音響モデル学習部８２は通常の音響モデルの学習を行なうだけでよい。

【００９９】

－言語モデル５０の学習－

音響モデル４８の学習とは別に、言語モデル５０の学習も以下のようにして行なれる。言語モデル学習部４６は、通常の言語モデルの学習方法を用い、会議録４２を学習データとして言語モデル５８の学習を行なう。本実施の形態では、言語モデルとしてユニグラム、バイグラム及びトライグラムを用いる。

【０１００】

言語モデル変換部６０はさらに、言語モデル５８内の各Ｎ－グラムに対し、変換モデル１２２を適用することで、話し言葉に対応した言語モデル５０への変換を行なう。変換後の言語モデル５０においては、文書スタイルのＮ－グラムの生起確率の一部が、話し言葉特有のＮ－グラムの生起確率に割り振られ、その分だけ文書スタイルのＮ－グラムの生起確率がディスカウントされている。

【０１０１】

－新たな書き起こしの作成－

このようにしてコンピュータシステム２５０で得られた音響モデル４８及び言語モデル５０を、会議録作成用コンピュータシステム３００に送信し、会議録作成用コンピュータシステム３００に保存する。会議録作成用コンピュータシステム３００の音声認識装置５２は、新たに録音された審議音声５４を、これら音響モデル４８及び言語モデル５０を用いて音声認識し、音声認識結果を新たな書き起こし５６として出力する。

【０１０２】

音響モデル４８の学習のときに、審議音声コーパス４０の全体を学習データとすること

ができる。そのため、音響モデル４８は多様な話し言葉表現をカバーすることができる。さらに、言語モデル５０では、話し言葉特有の表現について、書き起こし７０及び部分会議録７２の比較結果に応じた生起確率が割当てられる。そのため、文書スタイルのみの言語モデル５８を用いた場合と比較して、話し言葉スタイルの発話の音声認識の精度を高めることができる。

【０１０３】

以上述べたように、この実施の形態に係る会議録作成システム３０によれば、審議音声コーパス４０の一部である部分コーパス６８から書き起こし７０を作成し、対応する部分会議録７２と結合して対応付けコーパス７６を作成する処理を行なえば、後は自動的に審議音声コーパス４０への音素ラベル付与、音響モデル４８の学習、及び言語モデル５０の学習が行なえる。例えば政権交代などがあり、審議音声の状況に相当大きな変化があったときにも、対応付けコーパス７６を作成する処理までを手操作で行なえば、後は自動的な処理で音響モデル４８及び言語モデル５０の再構築をすることができる。その結果、新たな状況で得られた審議音声５４でも、音声認識装置５２によって正確な書き起こしを作成することができる。

【０１０４】

上記した実施の形態に係る会議録作成システム３０を実現するためのコンピュータプログラムは、単一のプログラムでもよいし、複数のプログラムを組合せたものでもよい。ただし、上記した実施の形態のように、会議録作成システム３０を２系統のコンピュータシステムで分割して実現する場合には、それらプログラムも別々にする必要がある。上記した各部の機能のうち、図１に示す話し言葉／書き言葉変換モデル学習部１２０において行なわれるＮ－グラム作成、言語モデル学習部４６において行なわれる言語モデル作成、初期音響モデル学習部１３０及び音響モデル学習部８２が実行する音響モデルの学習処理、などの個々の機能については、既に広く流布しているプログラムをそのまま使用できる。もちろん、これらプログラムは汎用に作成されているため、適切な調整を行なうことは要求されるが、それらはこの技術分野における通常の知識を持つ者にとっては、目的に照らして容易に実現できる範囲に留まる。

【０１０５】

これらプログラムは、例えばＤＶＤ２８２等のような記憶媒体に記憶され、又はインターネット２５２等のネットワークを通じて流通し、通常は大容量ハードディスク２７４等の不揮発外部記憶装置に記憶される。そして実行時には大容量ハードディスク２７４からＲＡＭ２８０にコピーされ、ＣＰＵ２７６内の図示しないプログラムカウンタと呼ばれるレジスタにより指し示されるアドレスから読出された命令がＣＰＵ２７６により実行され、上記した所期の機能を実現する。コンピュータハードウエアそのものの動作形態については周知であるので、ここではこれ以上の詳細な説明は行なわない。

【０１０６】

［評価実験］

－実験条件－

上記実施の形態の考え方にしたがって構築した会議録作成システムの性能について、衆議院審議音声により評価した。

【０１０７】

ベースライン音響モデル及び統計的変換モデルは２００３年及び２００４年のデータを用いて学習した。これらのデータについては人手による書き起こしが存在し、予め会議録との対応付けを行なっておく。音声データのサイズは１３４時間であり、審議録のテキストサイズは１．８Ｍ単語である。

【０１０８】

音声認識の際の音響特徴量は、１２次元のＭＦＣＣ（Ｍｅｌ－Ｆｒｅｑｕｅｎｃｙ Ｃｅｐｓｔｒｕｍ Ｃｏｅｆｆｉｃｉｅｎｔ）、ΔＭＦＣＣ、ΔΔＭＦＣＣ，Δパワー、ΔΔパワーの計３８次元である。

【０１０９】

－音素ラベル作成実験－

２００６年及び２００７年の衆議院審議音声を対象に、音素ラベル作成の実験を行なった。会議数は２６、ターン数は５，１７０、データ量は９１時間である。音響モデルは２００３年及び２００４年のデータ（１３４時間）を用いて学習したＨＭＭ（隠れマルコフモデル）のベースラインモデルである。ＨＭＭの状態数は３０００、混合数は１６であり、ＭＰＥ学習済である。特徴量にはＣＭＮ（Ｃｅｐｓｔｒａｌ Ｍｅａｎ Ｎｏｒｍａｌｉｚａｔｉｏｎ）及びＣＶＮ（Ｃｅｐｓｔｒａｌ Ｖａｒｉａｎｃｅ Ｎｏｒｍａｌｉｚａｔｉｏｎ）を適用した。音声認識は、Ｊｕｌｉｕｓ（http://julius.sourceforge.jp/）を用いて行なうが、大量のデータを処理することを想定して、サーチパラメータは軽く設定している（リアルタイムの２倍程度の時間を許容）。

【０１１０】

比較のため、以下の種々のモデルで音素ラベル作成実験を行なった。言語モデルの単位としては、会議全体で１つのモデルを作成する条件と、ターン毎に個別のモデルを作成する条件とを比較した。手法としては、本実施の形態に係る手法（「会議録、話し言葉変換」と呼ぶ。）に加え、話し言葉用ベースラインモデル（「ベースライン」）、会議録のみから作成したモデル（「会議録」）、それらを会議録に１００倍の重みをかけて合成したｂｉａｓｅｄ ＬＭ（「ｂｉａｓｅｄ ＬＭ」）、及び会議録モデルのポーズ位置にフィラーのエントリのみを追加したモデル（「会議録、フィラー」）をそれぞれ用いた。ベースラインモデルは１９９９年から２００５年の７年分の会議録に話し言葉変換を適用して作成した。

【０１１１】

音声認識により得られた音素ラベルの精度をテーブル１に示す。テーブル１において、Ｃｏｒｒ．（単語正解率）及びＡｃｃ．（単語認識精度）は人手による書き起こしを正解として算出した値である。

【０１１２】

【表１】

テーブル1

	作成単位	Corr.	Acc.
ベースライン	ー	82.3	79.5
会議録	会議	83.6	81.3
biased LM	会議	86.5	83.9
会議録、話し言葉変換	会議	86.3	83.7
会議録	ターン	86.1	83.5
会議録、フィラー	ターン	88.7	86.2
会議録、話し言葉変換	ターン	94.0	92.1

【０１１３】

テーブル１を参照して、会議単位の条件では、ｂｉａｓｅｄ ＬＭ及び上記実施の形態の手法で話し言葉スタイルに対処した場合、会議録単独のモデルよりも高い単語認識精度が得られた。ただし、２６の会議に対し、上記実施の形態の手法ではコンパクトなサイズでモデルが構築できた（１００ＭＢ）のに対し、ｂｉａｓｅｄ ＬＭでは極めて大きなサイズを要した（１．６ＧＢ）。したがって、ｂｉａｓｅｄ ＬＭをターン単位の処理に適用するのは非現実的と考えられる。

【０１１４】

ターン単位の条件では、会議単位の場合よりも全体に高い精度が得られた。本実施の形態に係る手法では、会議録のみを用いた場合よりも認識精度で８．６ポイント高くなった。会議録から得られた単語モデルにフィラーを追加したモデル（会議録、フィラー）は、簡易な話し言葉向け言語モデルとなっており、話し言葉の現象のうちフィラーの挿入のみに対応し、かつ文脈を考慮しない場合に相当する。本実施の形態に係る手法では、「会議

録、フィラー」モデルを認識精度で５．９ポイント上回った。統計的変換モデルにより、会議録から適切に話し言葉向け言語モデルが推定できていることが分かる。本実施の形態の手法では、精度で９２．１％、単語正解率で９４．０％を実現した。

【０１１５】

本実施の形態により作成された音素ラベルの例を以下に示す。

【０１１６】

【表２】

テーブル2

発話
総理 おっしゃっ た とおり 、これ は 、我が 国 いー 、に のみ ならず 、韓国 、周辺 国 、うー アジア 、あー この 地域 全体 に とって 大きな 脅威 で あり ま す。
会議録
総理がおっしゃったとおり、これは、我が国のみならず、韓国、周辺国 、アジア、この地域全体にとって大きな脅威であります。
作成された音素ラベル
総理 おっしゃっ た とおり 、これ は 、我が 国 いー 、のみ ならず 、韓国 、周辺 国 、うー アジア 、あー この 地域 全体 に とって 大き な 脅威 で あり ま す。

【０１１７】

この例では、助詞「が」の脱落、「いー」などのフィラーの挿入について、本実施の形態に係る手法により正しい音素ラベルが得られた。助詞「に」の挿入については不正解だったが、このパターンはそもそも変換規則に存在しなかったため、言語モデルで予測できるものではなかったと考えられる。

【０１１８】

－音声認識実験－

上記実施の形態に係る手法により作成した音素ラベルを用いて学習データを追加し、この学習データを使用して音響モデルの学習を行なった。学習済の音響モデルを用いて以下のような音声認識実験を行なった。

【０１１９】

ベースラインモデルは、２００３年、２００４年のデータ（１３４時間）を用いて人手の書き起こし音素ラベルにより学習を行なった音響モデルによる。追加データは、上記「音素ラベル作成実験」で音素ラベルを付与した２００６年及び２００７年の９１時間分である。比較のため、同じデータに対して人手の音素ラベルにより学習を行なった場合も評価する。学習はＭＬ（最尤基準）及びＭＰＥ（Ｍｉｎｉｍｕｍ Ｐｈｏｎｅ Ｅｒｒｏｒ）基準の２つの基準により行なう。HMMの状態数は５０００、混合数は３２である。特徴量にはＣＭＮ，ＣＶＮ及びＶＴＬＮ（Ｖｏｃａｌ Ｔｒａｃｔ Ｌｅｎｇｔｈ Ｎｏｒｍａｌｉｚａｔｉｏｎ）を適用した。テストセットは２００８年2月２６日及び２９日の衆議院予算委員会（２．４時間、１２１ターン）及び２００８年１０月７日の衆議院予算委員会（３．９時間、２１１ターン）である。

【０１２０】

この実験で得られた単語認識精度をテーブル３に示す。

【０１２１】

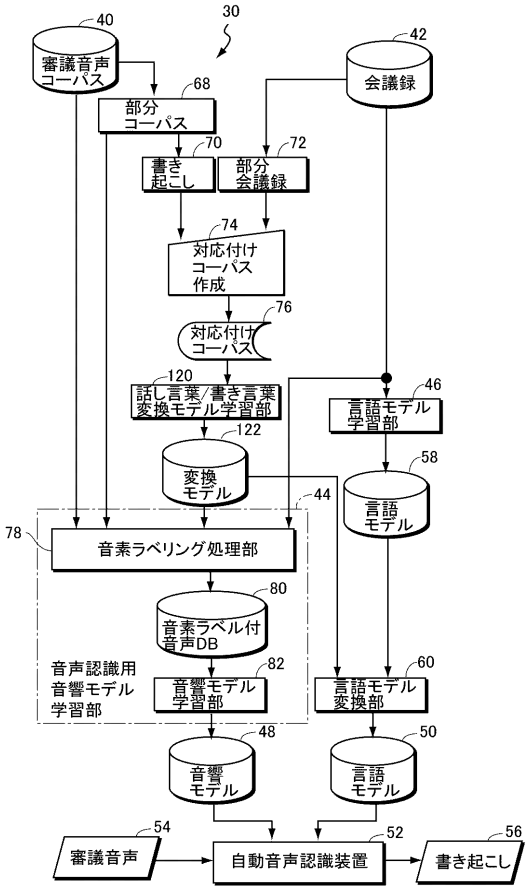
【表３】
テーブル3

	学習基準	2008/2/26, 29予算委員会	2008/10/7予算委員会
ベースライン	ML	85.4	77.6
本実施の形態	ML	85.8	78.4
人手	ML	85.9	78.5
ベースライン	MPE	86.8	79.2
本実施の形態	MPE	87.4	80.3
人手	MPE	87.6	80.4

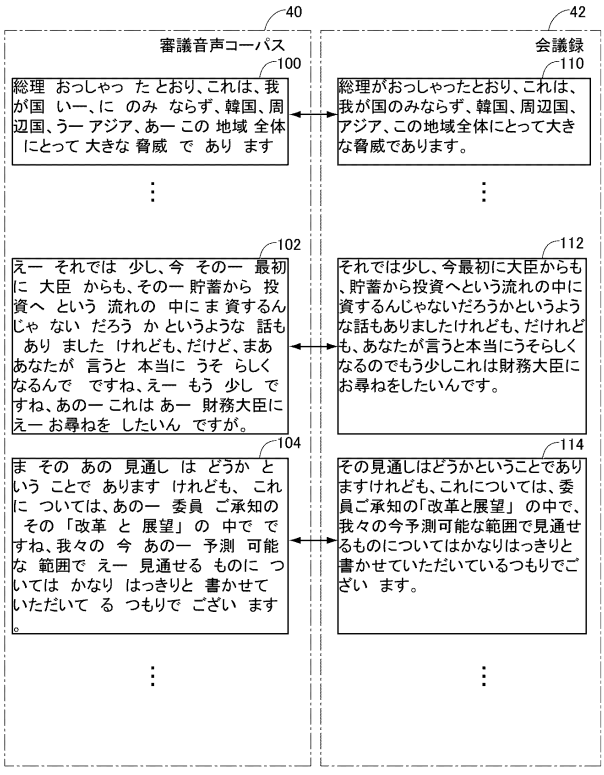
【０１２２】
テーブル３を参照して、ＭＬ学習の場合には、いずれのテストセットに対しても本実施の形態に係る手法を用いることでベースラインより高い精度が得られ、人手による音素ラベル付けの場合とほとんど変わらない水準となったことが分かる。ＭＰＥ学習の場合にも、ベースラインより精度が向上し、この場合にも人手による音素ラベル付けとほとんど変わらない水準となっている。
【０１２３】
以上のように本発明によれば、統計的話し言葉変換を用いた準教師付学習により、低コストで音響モデルを構築し、更新することが可能となった。したがって、音響モデルの学習のための音声コーパスにデータを追加したり入替えたりしても、音響モデルを容易に、かつ低コストで再構築することができる。その結果、内閣改造や総選挙などによる話者の変更、各話者の話し方の変化にも容易に対応することができる。
【０１２４】
上記実施の形態は、国会の委員会審議録を自動的に作成するシステムに関するものである。しかし本発明はそのような実施の形態には限定されない。例えば、放送番組の字幕や大学の講義録の作成などにこのシステムを適用することもできる。
【０１２５】
また、上記実施の形態では、音響モデル４８及び言語モデル５０の学習をコンピュータシステム２５０で行ない、会議録作成用コンピュータシステム３００では音響モデル４８及び言語モデル５０を受取って会議録作成のみを行なっている。しかし本発明はそのような実施の形態には限定されない。例えば、１つのコンピュータシステム内に上記した全ての機能を組込んでもよい。また、コンピュータシステム２５０内で実行されるプログラムのうち、音素ラベリング処理部７８の機能のみを別のコンピュータで実行し、音素ラベル付音声データベース８０をコンピュータシステム２５０で受けて音響モデル４８の学習を行なうようにしてもよい。同様に、話し言葉／書き言葉変換モデル学習部１２０の機能を別システムで実現してもよい。
【０１２６】
上記実施の形態の会議録作成システム３０は、一般には音声認識システムと呼ばれるべきものであり、音声認識によって、審議の発話内容に忠実な書き起こしを生成することができる。審議音声コーパスは、より一般的には、審議内における発話を収録した音声データベースであり、その名称はどのようなものでもよい。また、会議録は文書スタイルテキストの一例であって、発話内容を人間が書き起こし、整形したものであればどのようなものでもよい。
【０１２７】
今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内での全ての変更を含む。
【符号の説明】
【０１２８】
３０　会議録作成システム

４０　審議音声コーパス
４２　会議録
４４　音声認識用音響モデル学習部
４６　言語モデル学習部
４８　音響モデル
５０　言語モデル
５２，１４４　音声認識装置
５４　審議音声
５６　書き起こし
５８　言語モデル
６０　言語モデル変換部
６８　部分コーパス
７０　書き起こし
７２　部分会議録
７６　対応付けコーパス
７８　音素ラベリング処理部
８０　音素ラベル付音声データベース
１３０　初期音響モデル学習部
１３２　初期音響モデル
１３６　話し言葉用Ｎ－グラム
１３８　ターン・音声選択部
１８６　ターンごとＮ－グラム
１８８　Ｎ－グラム変換部

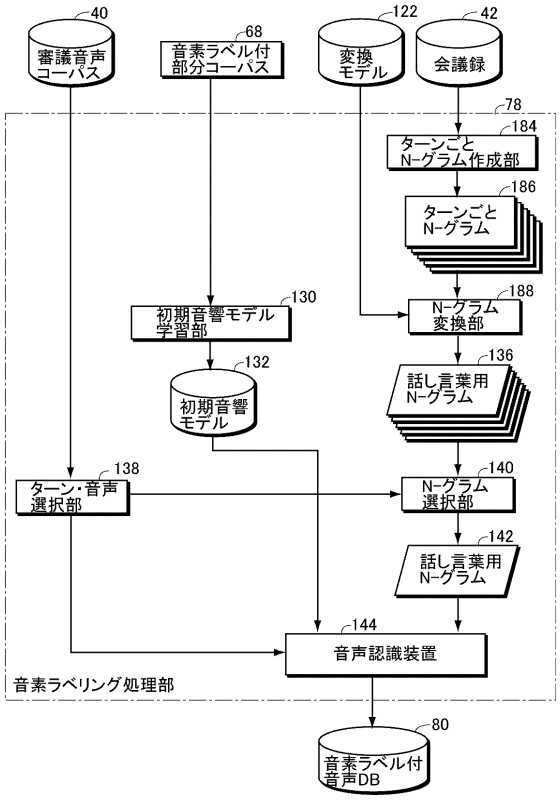
【図１】



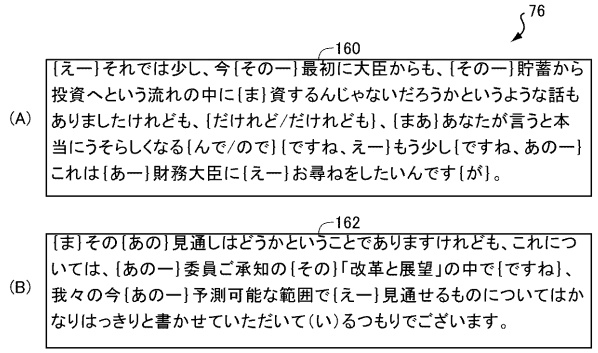
【図２】



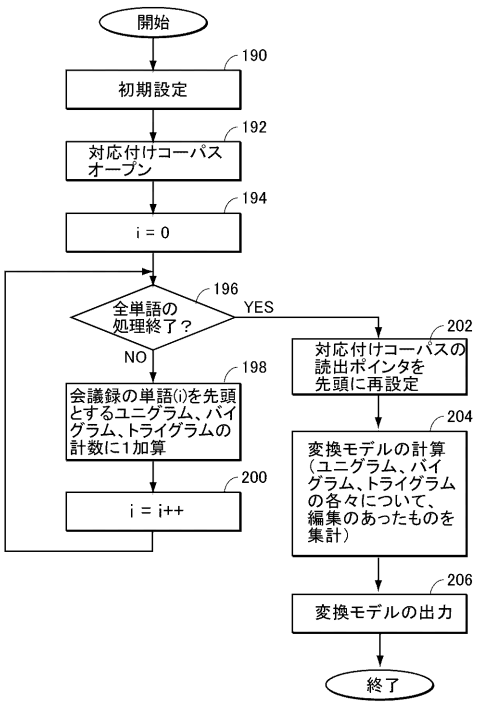
【図３】


40
審議音声コーパス
68
音素ラベル付部分コーパス
122
変換モデル
42
会議録
78
184
ターンごとN-グラム作成部
186
ターンごとN-グラム
130
初期音響モデル学習部
188
N-グラム変換部
132
初期音響モデル
136
話し言葉用N-グラム
138
ターン・音声選択部
140
N-グラム選択部
142
話し言葉用N-グラム
144
音声認識装置
音素ラベリング処理部
80
音素ラベル付音声DB

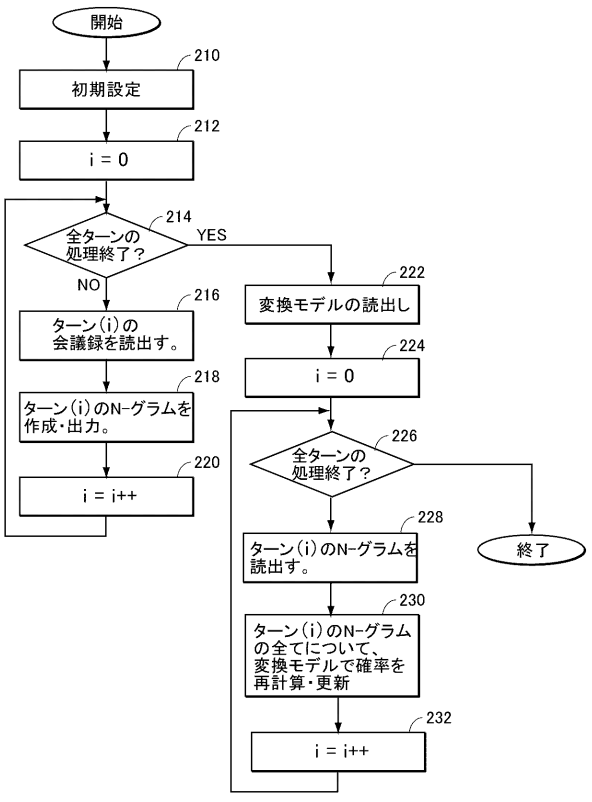
【図４】


76
160
(A)
{えー}それでは少し、今{その一}最初に大臣からも、{その一}貯蓄から投資へという流れの中に{ま}資するんじゃないだろうかというような話もありましたけれども、{だけれど/だけれども}、{まあ}あなたが言うと本当にうそらしくなる{んで/ので}{ですね、えー}もう少し{ですね、あの一}これは{あー}財務大臣に{えー}お尋ねをしたいんです{が}。
162
(B)
{ま}その{あの}見通しはどうかということでありますけれども、これについては、{あの一}委員ご承知の{その}「改革と展望」の中で{ですね}、我々の今{あの一}予測可能な範囲で{えー}見通せるものについてはかなりはっきりと書かせていただいて(い)るつもりでございます。

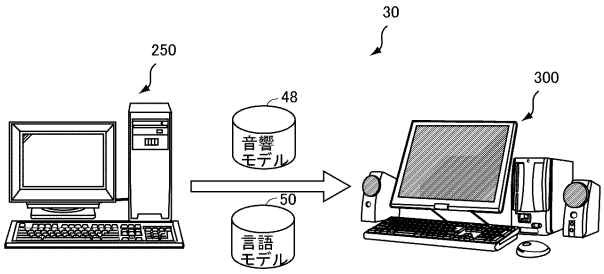
【図５】


開始
190
初期設定
192
対応付けコーパスオープン
194
i = 0
196
全単語の処理終了？
YES
NO
198
会議録の単語(i)を先頭とするユニグラム、バイグラム、トライグラムの計数に1加算
200
i = i++
202
対応付けコーパスの読出ポインタを先頭に再設定
204
変換モデルの計算（ユニグラム、バイグラム、トライグラムの各々について、編集のあったものを集計）
206
変換モデルの出力
終了

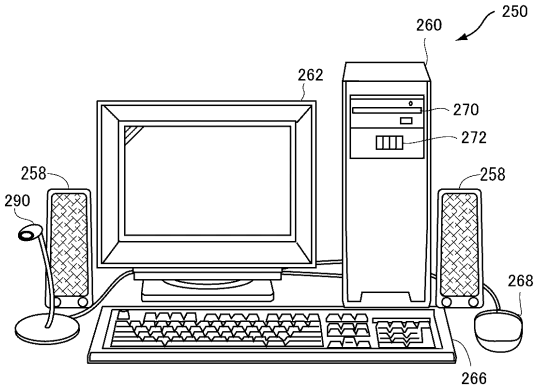
【図６】


開始
210
初期設定
212
i = 0
214
全ターンの処理終了？
YES
NO
216
ターン(i)の会議録を読出す。
218
ターン(i)のN-グラムを作成・出力。
220
i = i++
222
変換モデルの読出し
224
i = 0
226
全ターンの処理終了？
228
ターン(i)のN-グラムを読出す。
230
ターン(i)のN-グラムの全てについて、変換モデルで確率を再計算・更新
232
i = i++
終了

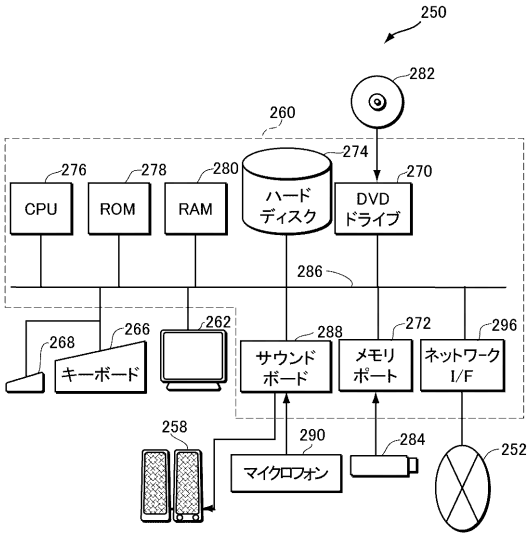
【図７】


30
250
48
音響
モデル
50
言語
モデル
300

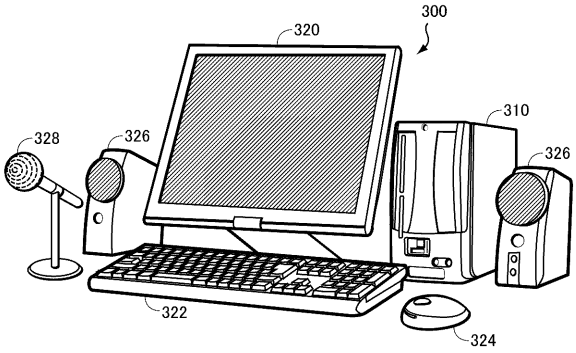
【図８】


250
260
262
270
272
258
258
290
268
266

【図９】


250
282
260
274
276
278
280
270
CPU
ROM
RAM
ハード
ディスク
DVD
ドライブ
286
262
266
268
キーボード
288
272
296
サウンド
ボード
メモリ
ポート
ネットワーク
I/F
258
290
284
252
マイクロフォン

【図１０】


300
320
310
328
326
326
322
324

フロントページの続き

(56)参考文献 特開２００７－２０６６０３（ＪＰ，Ａ）
特開２００４－２７１６１５（ＪＰ，Ａ）
特開２００３－１３２０４７（ＪＰ，Ａ）
特開２００２－９１９６７（ＪＰ，Ａ）
秋田祐哉他，"会議録作成支援のための国会審議の音声認識システム"，電子情報通信学会技術研究報告，２００８年１２月，Vol.108，No.338，pp.121-126
秋田祐哉他，"統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換"，電子情報通信学会技術研究報告，２００５年１２月，Vol.105，No.496，pp.19-24
秋田祐哉他，"言語モデルと発音辞書の統計的話し言葉変換に基づく国会音声認識"，電子情報通信学会技術研究報告，２００７年１２月，Vol.107，No.406，pp.61-66

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４
ＮＩＩ論文情報ナビゲータ（ＣｉＮｉｉ）