

(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2010-243914

(P2010-243914A)

(43) 公開日 平成22年10月28日 (2010. 10. 28)

(51) Int. Cl.	F I	テーマコード (参考)
G 1 0 L 15/06 (2006.01)	G 1 0 L 15/06 3 0 0 C	5 D 0 1 5
G 1 0 L 15/18 (2006.01)	G 1 0 L 15/18 2 0 0 D	
	G 1 0 L 15/06 3 0 0 Y	

審査請求 未請求 請求項の数 6 O L (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)
 特許法第30条第1項適用申請有り 平成21年3月10日 社団法人日本音響学会発行の「日本音響学会2009年春季研究発表会講演論文集」に発表

(71) 出願人 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
 Fターム(参考) 5D015 GG05

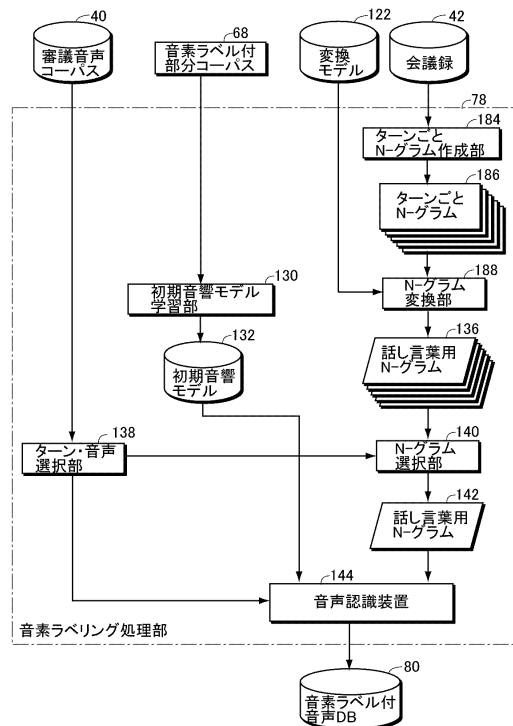
(54) 【発明の名称】 音響モデル学習装置、音声認識装置、及び音響モデル学習のためのコンピュータプログラム

(57) 【要約】

【課題】過去に整形済の文書スタイルテキストDBが存在している類の話し言葉の音声認識・書き起こしのための音響モデルを、効果的に作成する音響モデル学習装置を提供する。

【解決手段】音響モデル学習装置78は、音声DB（例えば審議音声コーパス）40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（例えば会議録）42により学習した言語モデル186から、実際の発言内容に忠実な書き起こしの言語モデル136を推定する言語モデル推定部188と、初期音響モデル130及び言語モデル推定部188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136を用いた音声認識により、音声DB40に書き起こしとその音素ラベルを付して音素ラベル付音声DB80を出力する音素ラベリング部144と、音素ラベル付音声DB80を学習データとして、音響モデルの学習を行なう音響モデル学習部とを含む。

【選択図】図3



【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。

10

【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、

20

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、

30

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、

40

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

50

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。

10

【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。

20

【0005】

この問題に対処するため、Lame1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。

30

【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。

40

【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。

50

【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」、INTER-SPEECH, 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、必要なハードウェア及びソフトウェ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。

10

【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、発話中に頻りに言直し、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。

20

【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るれば、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。

30

【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。

40

【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。

50

【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る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。

10

【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）があつたとしても、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。

20

【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。

30

【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。

40

【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

50

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。

10

20

30

40

50

【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ェ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以下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

10

20

30

40

50

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。

【 0 0 4 2 】

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。

【 0 0 4 3 】

【 数 1 】

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

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

10

【 0 0 4 4 】

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。

【 0 0 4 5 】

20

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。

【 0 0 4 6 】

30

【 数 2 】

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

【 0 0 4 7 】

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。

40

【 0 0 4 8 】

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。

【 0 0 4 9 】

【 数 3 】

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

【 0 0 5 0 】

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 $w \rightarrow v$ 、

50

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 $p(v|w)$ 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、パターンの隣接单語も考慮する。例えば、フィラー「あー」は、 $\{w = (w_{-1}, w_{+1}) \quad v = (w_{-1}, \text{あー}, w_{+1})\}$ 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

10

20

30

40

50

、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。

10

【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。

20

【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 $w = \text{「} \langle sp \rangle \text{この 法案} \text{」}$ ($\langle sp \rangle$ 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 $v = \text{「} \langle sp \rangle \text{えー この 法案} \text{」}$ となっていた(フィル「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。

30

【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。

40

【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

50

N - グラム 1 8 6 が得られる。ターンごと N - グラム 1 8 6 内の各エントリについて、変換モデル 1 2 2 を適用することによって話し言葉用 N - グラム 1 3 6 がターンごとに得られる。

【 0 0 6 3 】

音素ラベリング処理部 7 8 はさらに、審議音声コーパス 4 0 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 1 3 8 と、ターン・音声選択部 1 3 8 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 1 3 6 の中から、そのターンに対応する N - グラム 1 4 2 を選択するための N - グラム選択部 1 4 0 と、初期音響モデル 1 3 2 及び N - グラム 1 4 2 を用い、特に N - グラム 1 4 2 を言語モデルとして用いて、ターン・音声選択部 1 3 8 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 8 0 に出力するための音声認識装置 1 4 4 とを含む。

10

【 0 0 6 4 】

音声認識装置 1 4 4 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 1 4 4 は、発話中のポーズにより、最長で 3 0 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。

【 0 0 6 5 】

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 8 0 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 4 0 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。

20

【 0 0 6 6 】

したがって、この音素ラベル付音声データベース 8 0 から、図 1 に示す音響モデル学習部 8 2 によって通常の方法で音響モデル 4 8 を作成すると、音声認識装置 5 2 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。

【 0 0 6 7 】

一方、音声認識装置 5 2 が使用する言語モデル 5 0 も、会議録 4 2 中に出現する N - グラムについて、変換モデル 1 2 2 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。

30

【 0 0 6 8 】

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 4 8 及び言語モデル 5 0 を使用するため、音声認識装置 5 2 は、話し言葉においてよく発生する事象、すなわちフィルターの挿入、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 4 0 の高精度な書き起こしを出力することができる。

【 0 0 6 9 】

図 4 は、対応付けコーパス 7 6 中の 2 つの文例を示す。図 4 において、審議音声コーパス 4 0 では発話されているが会議録 4 2 では削除されている音声を図 4 (A) の発話 1 6 0 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 4 0 では発話されていないが会議録 4 2 では挿入されている音声は、図 4 (B) の発話 1 6 2 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 4 0 の発話での表現が会議録 4 2 では他の表現に変えられている部分は、発話 1 6 0 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 4 0 での表現を「/」の前に、会議録 4 2 での表現を「/」の後に、それぞれ示してある。

40

【 0 0 7 0 】

この対応付けコーパスは、書き起こし 7 0 と部分会議録 7 2 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

50

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。

【 0 0 7 1 】

[話し言葉 / 書き言葉変換モデル学習部 1 2 0 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 1 2 0 による変換モデル 1 2 2 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 1 9 0 と、対応付けコーパス 7 6 のファイルをオープンするステップ 1 9 2 と、繰返し変数 i に 0 を代入するステップ 1 9 4 とを含む。

【 0 0 7 2 】

繰返し変数 i は、対応付けコーパス 7 6 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。

【 0 0 7 3 】

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 7 6 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 1 9 6 と、ステップ 1 9 6 の判定結果が N O のときに実行され、対応付けコーパス 7 6 の中で、部分会議録 7 2 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 1 9 8 と、変数 i に 1 を加算して制御をステップ 1 9 6 に戻すステップ 2 0 0 とを含む。ステップ 1 9 6 からステップ 2 0 0 の処理を、対応付けコーパス 7 6 中の全単語に対して実行することにより、部分会議録 7 2 の N - グラムモデルが作成される。

【 0 0 7 4 】

このプログラムは更に、ステップ 1 9 6 での判定結果が Y E S のときに実行され、対応付けコーパス 7 6 の読出位置を先頭に再設定するステップ 2 0 2 と、ステップ 2 0 2 に続き、部分会議録 7 2 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 7 0 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 1 2 2 を計算するステップ 2 0 4 と、ステップ 2 0 4 で計算された変換モデル 1 2 2 をファイルとして出力し、処理を終了するステップ 2 0 6 とを含む。

【 0 0 7 5 】

[ターンごと N - グラム作成部 1 8 4 及び N - グラム変換部 1 8 8 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 1 8 4 及び N - グラム変換部 1 8 8 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 2 1 0 と、繰返し変数 i に 0 を代入するステップ 2 1 2 と、繰返し変数 i を処理対象の部分会議録 7 2 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 2 1 4 とを含む。

【 0 0 7 6 】

このプログラムはさらに、ステップ 2 1 4 の判定結果が N O の場合に実行され、ターン (i) の会議録を部分会議録 7 2 から読出すステップ 2 1 6 と、ステップ 2 1 6 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 2 1 8 と、ステップ 2 1 8 に続き、繰返し変数 i の値に 1 を加算し、制御をステップ 2 1 4 に戻すステップ 2 2 0 とを含む。

【 0 0 7 7 】

このプログラムはさらに、ステップ 2 1 4 の判定結果が Y E S の場合に実行され、変換モデル 1 2 2 を外部記憶媒体から主記憶装置に読出すステップ 2 2 2 と、繰返し変数 i に 0 を代入するステップ 2 2 4 と、繰返し変数 i の値と部分会議録 7 2 に含まれるターン数との比較により、部分会議録 7 2 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 2 2 6 と、ステップ 2 2 6 において、部分会議録 7 2 の内

10

20

30

40

50

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。

10

【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。

20

【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及び一对のスピーカ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。

30

【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。

40

【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、

50

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 RAM 280 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及び一对のスピーカ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。

10

【0086】

会議録作成用コンピュータシステム300のハードウェア構成は、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。

20

【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。

30

【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)。

40

【0092】

この処理は例えば以下に行なう。部分会議録72内に、N-グラムのトライグラム $w = \langle s p \rangle$ の「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

50

$v = \langle \text{sp} \rangle \text{えー この 法案}$ となっているとする。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 $w = \langle \text{sp} \rangle \text{この 法案}$ 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 $w = \langle \text{sp} \rangle \text{この 法案}$ 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、「 $\langle \text{sp} \rangle \text{えー この 法案}$ 」が50、「 $\langle \text{sp} \rangle \text{この 法案}$ 」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)。

10

【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。

20

【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。

30

【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。

40

【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】

50

- 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、通常の手話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。

10

【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。

20

【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、後は自動的に処理で音響モデル 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。

30

【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。

40

【0105】

これらプログラムは、例えば DVD 282 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

50

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ェ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。

10

【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)、MFCC、MFCC、パワー、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)。

20

30

【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ルターのエントリのみを追加したモデル(「会議録、フィルター」)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。

40

【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	—	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

10

【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。

20

【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。

30

【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。

10

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。

20

【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。

30

【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。

40

【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

10

【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。

20

【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。

30

【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。

40

【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の全ての変更を含む。

【符号の説明】

【0128】

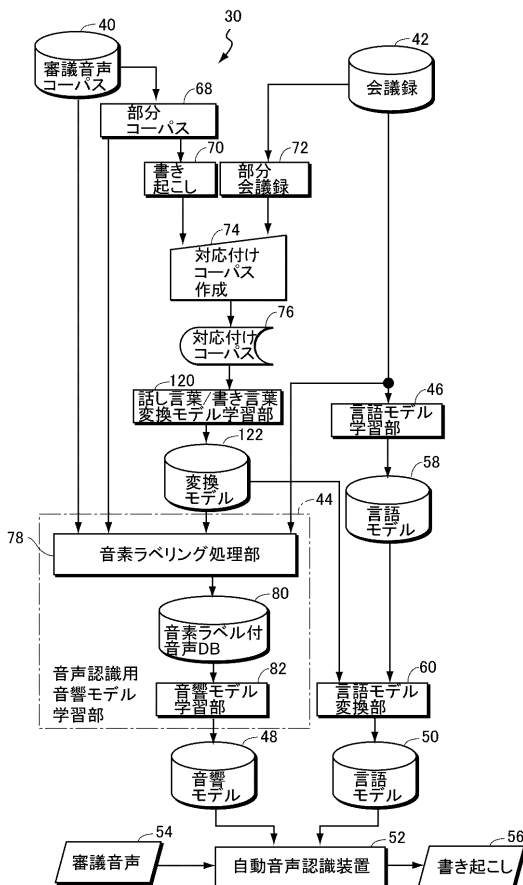
50

- 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

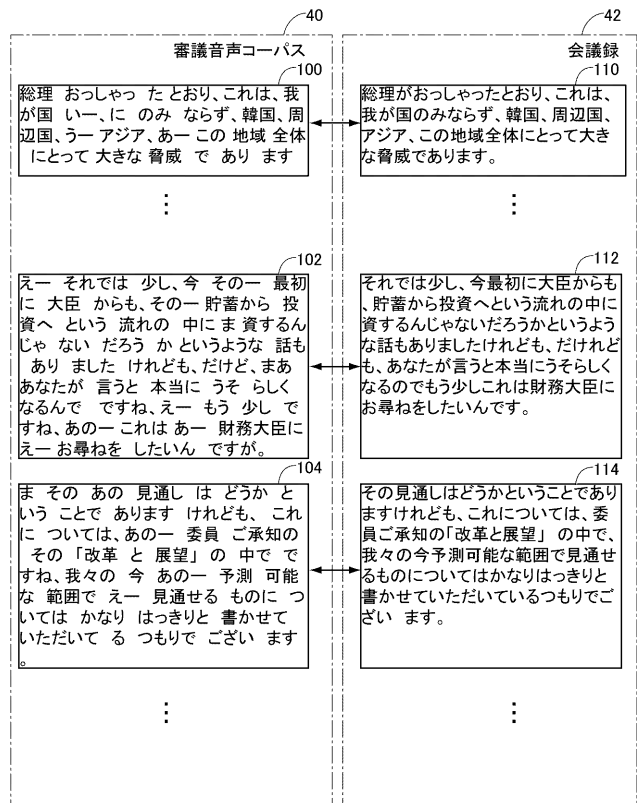
10

20

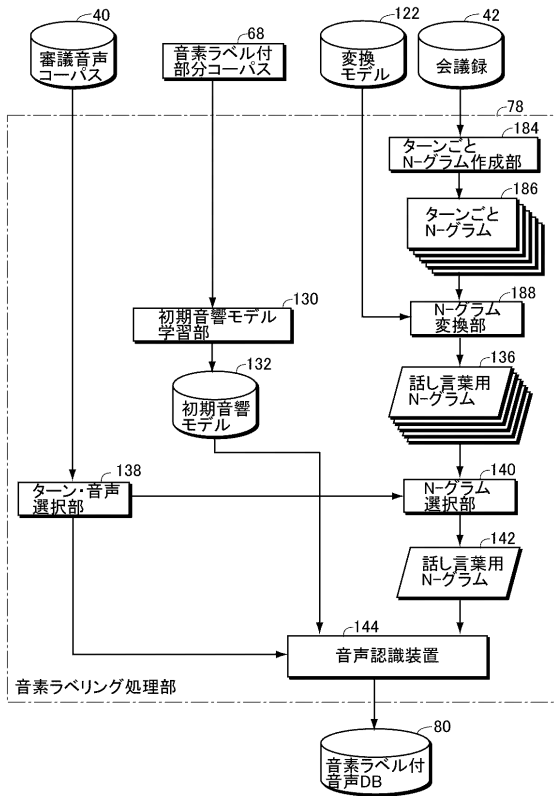
【図1】



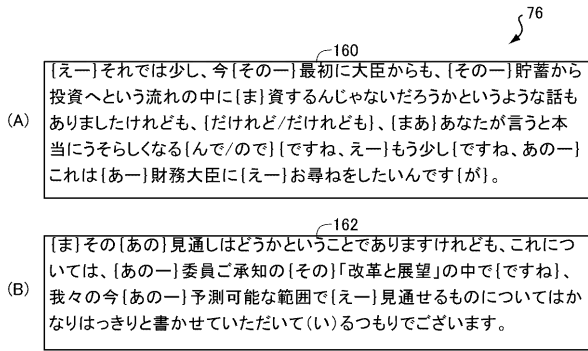
【図2】



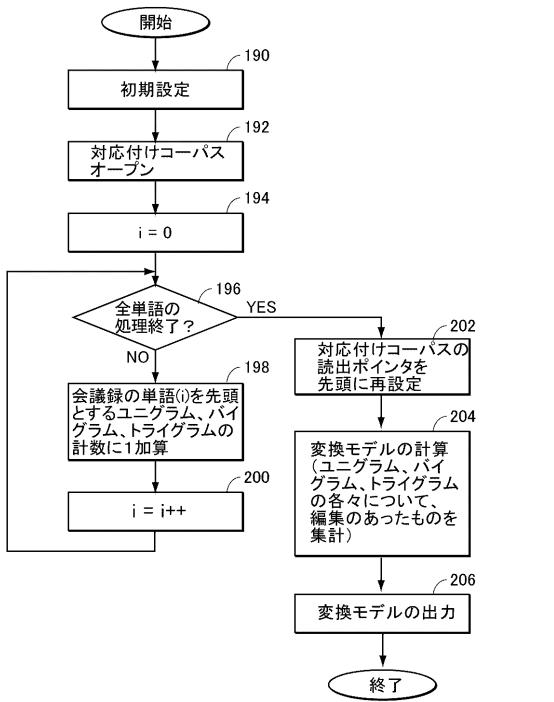
【 図 3 】



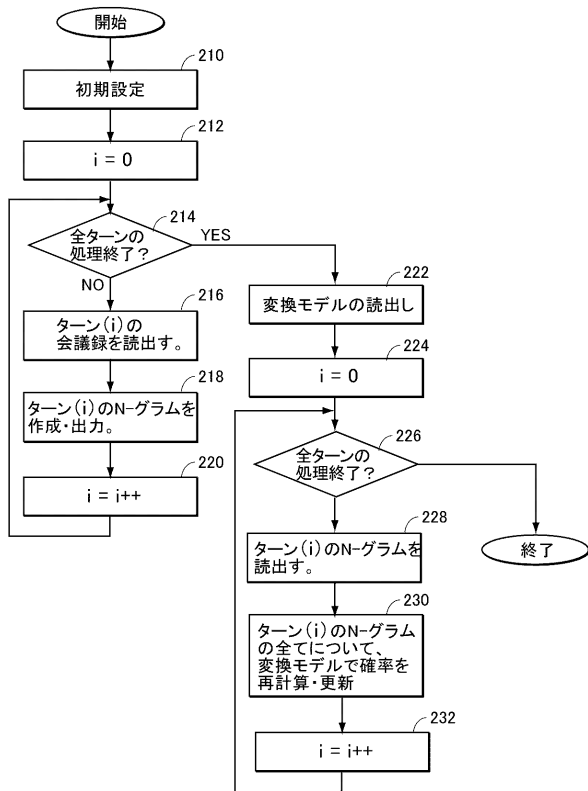
【 図 4 】



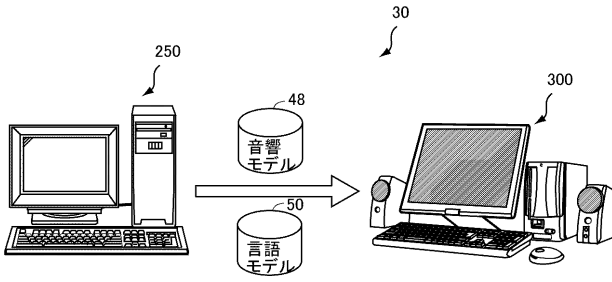
【 図 5 】



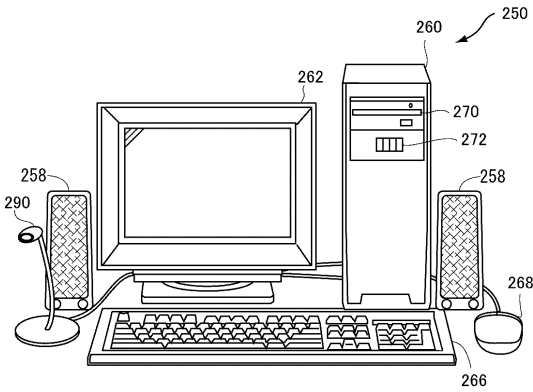
【 図 6 】



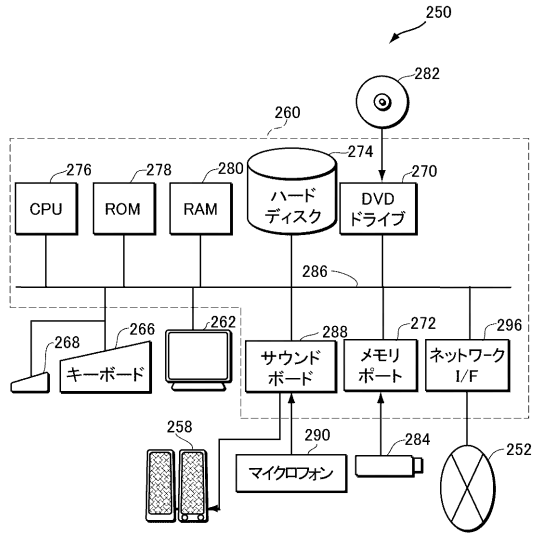
【図7】



【図8】



【図9】



【図10】

