(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2010-55030
(P2010-55030A)
(43)公開日　平成22年3月11日(2010.3.11)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 15/06	(2006.01)	G10L	15/06	300Y	5D015
		G10L	15/06	300C	
		G10L	15/06	310Z	

審査請求 未請求　請求項の数 7　OL　（全 33 頁）

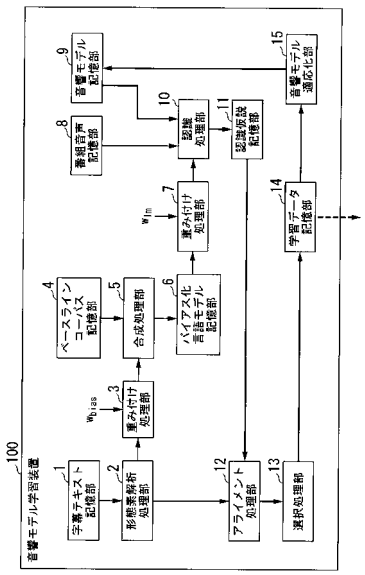
(21)出願番号　特願2008-222817 (P2008-222817)
(22)出願日　平成20年8月29日 (2008.8.29)

(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　100064908
弁理士　志賀　正武
(74)代理人　100108578
弁理士　高橋　詔男
(72)発明者　佐藤　圧衛
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　今井　亨
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

Fターム(参考)　5D015 AA01　GG01

(54)【発明の名称】音響処理装置およびプログラム

(57)【要約】
【課題】字幕と発話内容の一致率が低いオフライン字幕を利用した場合にも、高い認識率によって音響モデルの学習データを自動生成し、多様な発話スタイルに対応できる音響モデルを作成できる音響モデル学習装置を提供する。
【解決手段】一致区間のみを切り出して利用するのではなく、その他の区間から得られる情報も、音響モデル学習のために用いる。そのため、音声認識結果と書き起こし字幕テキストの一致区間を利用して学習データを得る際に、各形態素の信頼度を導入することにより、一致区間以外の音声も利用して学習データを自動生成させる。
【選択図】図１



【特許請求の範囲】
【請求項１】
音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部と、
前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理部と、
前記認識処理部が出力した認識結果データに基づき、教師系列データを取得し、前記教師系列データに含まれる各々の音素に信頼度データを付加し、前記信頼度データが付加された前記教師系列データと該教師系列データに対応する音響特徴量データとを、学習データとして出力する選択処理部と、
前記選択処理部が出力した学習データを読み込み、前記学習データに含まれる各々の前記音素の観測確率データを算出し、前記音素ごとに、前記学習データに含まれる前記音響特徴量データと算出された前記観測確率データとに基づき、且つ、前記信頼度データを重みとして乗じて得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化部と、
を具備することを特徴とする音響処理装置。
【請求項２】
請求項１に記載の音響処理装置であって、
前記選択処理部は、前記音声に対応するテキストデータを読み込み、前記認識結果データに含まれる最尤パスを前記教師系列データとして取得し、当該教師系列データと前記テキストデータが一致する一致区間に含まれる前記音素に対しては最高信頼度を表わす信頼度データを付加し、その他の区間に含まれる前記音素に対しては最低信頼度を表わす信頼度データを付加する、
ことを特徴とする音響処理装置。
【請求項３】
請求項１に記載の音響処理装置であって、
前記選択処理部は、前記音声に対応するテキストデータを読み込み、前記認識結果データに含まれる最尤パスと前記テキストデータが一致する一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝刈り処理の結果得られる系列を前記教師系列データとして取得し、当該教師系列データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる前記音素に対する信頼度データとして付加する、
ことを特徴とする音響処理装置。
【請求項４】
音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部と、
前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理部と、
前記認識処理部が出力した認識結果データと前記音声に対応するテキストデータとを読み込み、前記認識結果データに含まれる最尤パスと前記テキストデータが一致する一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝狩り後の認識結果データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる各々の音素に対する信頼度データとして付加した教師系列データを求め、前記教師系列データとこれに対応する音響特徴量データとを、学習データとして出力する選択処理部と、
前記選択処理部が出力した学習データを読み込み、フォワードバックワードアルゴリズムを用いて前記学習データに含まれる各々の音素の観測確率を求め、前記音素ごとに、前記観測確率で重み付けられた前記音響特徴量データとして得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化部と、
を具備することを特徴とする音響処理装置。
【請求項５】
請求項１に記載の音響処理装置であって、

前記選択処理部は、前記音声に対応するテキストデータを読み込み、前記認識結果データから、前記テキストデータとの一致が最大となるようなパスを選択して前記教師系列データとして取得するとともに、当該教師系列データと前記テキストデータが一致する一致区間に含まれる前記音素に対しては最高信頼度を表わす信頼度データを付加する、

ことを特徴とする音響処理装置。

【請求項６】

音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部と、

前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理部と、

前記認識処理部が出力した認識結果データと前記音声に対応するテキストデータとを読み込み、前記認識結果データから、前記テキストデータとの一致が最大となるようなパスを選択して前記教師系列データとして取得するとともに、当該教師系列データ内で前記認識結果データと前記テキストデータとが一致する一致区間を選択し、前記一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝狩り後の認識結果データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる各々の音素に対する信頼度データとして付加した教師系列データを求め、前記教師系列データとこれに対応する音響特徴量データとを、学習データとして出力する選択処理部と、

前記選択処理部が出力した学習データを読み込み、フォワードバックワードアルゴリズムを用いて前記学習データに含まれる各々の音素の観測確率を求め、前記音素ごとに、前記観測確率で重み付けられた前記音響特徴量データとして得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化部と、

を具備することを特徴とする音響処理装置。

【請求項７】

音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部を備えるコンピュータに、

前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理過程と、

前記認識処理過程で出力した認識結果データに基づき、教師系列データを取得し、前記教師系列データに含まれる各々の音素に信頼度データを付加し、前記信頼度データが付加された前記教師系列データと該教師系列データに対応する音響特徴量データとを、学習データとして出力する選択処理過程と、

前記選択処理過程で出力した学習データを読み込み、前記学習データに含まれる各々の前記音素の観測確率データを算出し、前記音素ごとに、前記学習データに含まれる前記音響特徴量データと算出された前記観測確率データとに基づき、且つ、前記信頼度データを重みとして乗じて得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化過程と、

の処理を実行させるプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、統計的処理に基づく音声処理に関する。特に、各音素の音響特徴量の統計量を音響モデルとして処理を行なう音響処理装置およびそのプログラムに関する。また特に、音響モデルの学習に関する。

【背景技術】

【０００２】

統計モデルを用いた音声認識では、各音素の統計量を得るために大量の学習データが必要になる。この学習データは、大量の音声データとその音声に対する書き起こしの対であり、従来人手により時間とコストをかけて作成されてきた。

【０００３】

現在、放送番組の一部には、高齢者や聴覚障害者の番組視聴など音声からの情報が十分に得られない環境でも番組内容が理解できるように、番組音声に対応する字幕テキストが付与されているものがある。この字幕テキストにはオンラインで付与された字幕テキストとオフラインで付与された字幕テキストの２種類があり、前者は番組中の発話内容とほぼ一致している。但し、後者は話し言葉の冗長性を除去し、簡潔で読みやすい字幕を付与できるため、発話内容と字幕との一致率が低い。

【０００４】

上記の一致率が高いオンライン字幕を利用できる場合には、音声区間の切り出しと、切り出された音声に対応する字幕区間の切りだしを行なうことで、音響モデルの学習データを容易に自動作成することができ、読み上げ音声が中心で比較的高い音声認識精度が得られるニュース番組を対象とした場合の認識精度の改善が報告されている（非特許文献１）。この方法では、字幕区間の切り出しに、字幕テキストで適応化した言語モデルを用いて得られた番組音声の認識結果を利用し、字幕テキストと認識結果の一致区間を音響モデルの学習データとしている。この番組音声の認識は、主に音声区間の切り出しと字幕と発話内容の不一致部分の検出を目的として行われ、言語モデルの適応化に用いる字幕テキストの重みは比較的小さいことが特徴である。

【０００５】

図１４は、従来技術による音響モデル学習装置の機能構成を示すブロック図である。この図に示す音響モデル学習装置２００は、字幕テキストを形態素単位に分割した後、字幕テキストにW_{bias}の重みを付けて、バイアスのある言語モデル（バイアス化言語モデル８６）を得る。この言語モデルを用いて番組音声を認識する（認識処理部９０）。本装置の主目的は、字幕と発話内容の不一致区間の検出と音声と字幕の切り出しであるため、バイアス重みW_{bias}（４程度）と言語重みW_{lm}（１０程度）は比較的小さな値を用いる。次に、番組全体で認識仮説と字幕テキストのＤＰマッチングを行い（アライメント処理部９２）、字幕テキストと認識結果が３単語以上連続して一致している区間を選択し、番組音声中の音声区間と対応する単語列である学習データ（学習データ記憶部９４）が得られる。

【非特許文献１】Long Nguyen, 外１名, “Light Supervision in Acoustic Model Training”, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP apos;04), ２００４年, Volume 1, ｐ. １７－２１

【発明の開示】

【発明が解決しようとする課題】

【０００６】

現在、放送番組の対談部分などの比較的自由なスタイルで発話された部分の認識精度は、読み上げ音声の認識精度に比べて低く、このような部分の認識精度を向上させるため、様々なスタイルで発話された音響モデルの学習データが必要とされている。

【０００７】

放送番組の対談部分などの比較的自由なスタイルでは、オフラインで作成される字幕テキストが用いられるが、前述の通りこの字幕テキストの作成の際は簡潔さが優先されるため、字幕テキストと実際の発話内容との一致率は低い。従来技術による音響モデル学習装置は字幕テキストと実際の発話内容との一致率が高いことを前提としており、従来技術による方法ではこの一致率の低いデータからは高い認識精度が得られず、その結果として音響モデルの学習データを効率よく多量に生成することができないという問題があった。

【０００８】

本発明のより具体的な課題は、次の通りである。対談などの自由発話の音声を学習する場合、議事録や放送番組の字幕テキストなど、一般的に入手容易な書き起こしテキストは不完全であることが多く、十分な学習効果が得られないことがある。例えば、対談の自由発話には、「あの」、「で」、「まあ」などの多くの不要語が含まれており、これらの不要語は書き起こしのテキストには盛り込まれないことが多い。よって、そのような不完全な書き起こしのテキストからは、これらの不要語が得られない。また、字幕テキストを作

成する目的に応じて、話し言葉に含まれる冗長な部分を読みやすいフレーズに置き換えて書き起こしのテキストが作成されることも多い。従って、その冗長な部分に対応するテキストが得られない。

【０００９】

また、従来のように一致部分の音声区間を切り出して学習する技術では、不一致部分の音声は学習には用いられない。そのため、音声と書き起こしテキストの一致率が低い場合には、少量の学習データしか得られない。また、音声認識の入力となる発話切り出し単位と学習音声の切り出し単位とが異なり、細かく切り出された音声を学習データとしなければならず、学習時と認識時の音声切り出し単位の不接合を避けられないという問題もあった。また、従来技術を用いた場合には、認識結果（最尤仮説）と書き起こしテキストとの間の不一致には、例えば、「こと」と「事」のような表記の揺らぎによる不一致や、「ですから」と「です／から」といったように形態素分割が異なるために起こる不一致が見られ、これらの区間の音声が学習データとしては用いられなかったという問題もあった。

【００１０】

本発明は、上記のような事情を考慮して為されたものであり、字幕と発話内容の一致率が低いオフライン字幕を利用した場合にも、高い認識率によって音響モデルの学習データを自動生成し、多様な発話スタイルに対応できる音響モデルを作成することを目的とするものである。また、本発明は、字幕と発話内容の一致率が低いオフライン字幕を利用した場合にも、より多くの量の音響モデルの学習データを生成することのできる音響処理装置およびプログラムを提供することを目的とする。

【課題を解決するための手段】

【００１１】

上記の課題を解決するために、本発明では、次のような手段を用いる。即ち、所定数の形態素が連続して一致している区間が一致区間であるが、そのような一致区間を切り出して利用するのではなく、その他の区間から得られる情報も、音響モデル学習のために用いる。より具体的には、音声認識結果と書き起こし字幕テキストの一致区間を利用して学習データを得る際に、各形態素（単語）の信頼度を導入することにより、一致区間以外の音声も利用して学習データを自動生成させる。このように一致区間に対応する音声を切り出さずに学習できるようになるため、認識時と学習時の発話単位の不整合が解消されることにより認識精度の向上が期待される。

【００１２】

より具体的には、本発明の特別な技術的特長（Special Technical Feature）は、次に述べる構成あるいはその部分集合による構成によるものである。その構成とは即ち、

１）音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部を備え、

２）認識処理部が、前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力し、

３）選択処理部は、前記認識処理部が出力した認識結果データに基づき、各々の音素に対する信頼度データを付加した教師系列データを求め、前記教師系列データとこれに対応する音響特徴量データとを、学習データとして出力し、そして、

４）音響モデル適応化部は、前記選択処理部が出力した学習データを読み込み、前記学習データに含まれる各々の音素の観測確率を求め、前記音素ごとに、音響モデルを用いて、前記音響モデル記憶部を更新する。

【００１３】

そして、本発明の第１の態様による音響処理装置は、音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部と、前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理部と、前記認識処理部が出力した認識結果データに基づき、教師系列データを取得し、前記教師系列データに含まれる各々の音素に信頼度データを付加し、前記信頼度データが付加された前記教師系列データと該教師系列データに対応する音響特徴

量データとを、学習データとして出力する選択処理部と、前記選択処理部が出力した学習データを読み込み、前記学習データに含まれる各々の前記音素の観測確率データを算出し、前記音素ごとに、前記学習データに含まれる前記音響特徴量データと算出された前記観測確率データとに基づき、且つ、前記信頼度データを重みとして乗じて得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化部とを具備することを特徴とするものである。

ここで、言語的単位とは、言語的にまとまりのある単位であり、例えば、形態素、単語などである。また、認識仮説とは、認識処理の結果得られる認識の仮説であり、ある時刻でみたときには、複数の仮説が並立する場合もある。例えば形態素を単位としてみたとき、認識仮説は、形態素がアークに対応し、形態素間の接続点がノードに対応する形のラティス構造（時間方向の半順序構造）を有する。ある時刻において複数のアーク（相互に対立するアーク）が存在するとき、それらに対する確率を与えることができるが、本態様ではその確率を信頼度データとしている。また、教師系列データとは、信頼度データが付与されたラティスである。

本態様では、音響処理装置は、認識仮説の最尤パスと（書き起こしの字幕などの）テキストデータとの一致に基づき音声の区間を切り出す代わりに、教師音素列の信頼度を導入し、信頼度に基づいて学習する部分と学習しない部分を特定する。選択処理部が出力する学習データには、音声と教師音素列に加えて、それらに対応する教師音素列の信頼度を含んでいる。そして音響モデル適応化部は、音響モデルの統計量を推定する際に、例えばフォワードバックワードアルゴリズム（Forward-Backward Algorithm）やビタビアルゴリズム（Viterbi Algorithm）により得られる教師音素列の各音素の観測確率に、上記の信頼度による重み付けを行なって推定する。

このような構成により、認識結果データの中の一致区間だけではなく、その他の部分の音声の音響特徴量データを用いて、音響モデルの学習（音響モデル記憶部の更新（適応化））を行なえる。また、信頼度データを用いて、それによって重み付けをしていることにより、信頼度に応じた適切な学習を行なうことが可能となる。

【００１４】

本発明の第２の態様による音響処理装置は、前記の音響処理装置において、前記選択処理部は、前記音声に対応するテキストデータを読み込み、前記認識結果データに含まれる最尤パスを前記教師系列データとして取得し、当該教師系列データと前記テキストデータが一致する一致区間に含まれる前記音素に対しては最高信頼度を表わす信頼度データを付加し、その他の区間に含まれる前記音素に対しては最低信頼度を表わす信頼度データを付加する、ことを特徴とする。

ここで、テキストデータとは、音声に対応するものであり、例えば書き起こしの字幕テキストのデータである。但し、音声（実際の発話内容）と字幕テキストの内容とは、必ずしも完全に一致しているとは限らない。また、信頼度データは、認識結果データに含まれるあるパス（パスは、言語的単位（形態素等）の系列に対応）が信頼できる度合いを表わす数値である。最高信頼度とは、その仮説の正しさを表わす確率が１である状態に対応する。最低信頼度とは、その仮説が正解に対して無情報である状態に対応する。信頼度データを０以上１以下の実数で表わしたとき、信頼度データが大きいほど信頼できる度合いが高く、最高信頼度は信頼度データ「１」に相当し、最低信頼度は信頼度データ「０」に相当する。

本態様では、音響処理装置は、最尤仮説とテキストデータの一致区間のみを学習するために、一致区間に含まれる形態素（単語）およびその形態素内の音素にのみ信頼度１を与え、それ以外の区間に含まれる形態素およびその形態素内の音素の信頼度を０とする。

これにより、一致区間に高い信頼度を与えて音響モデルの学習をすることができる。

【００１５】

本発明の第３の態様による音響処理装置は、前記の音響処理装置において、前記選択処理部は、前記音声に対応するテキストデータを読み込み、前記認識結果データに含まれる最尤パスと前記テキストデータが一致する一致区間と時間軸上で対立する区間を枝刈りす

る処理を行ない、この枝刈り処理の結果得られる系列を前記教師系列データとして取得し、当該教師系列データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる前記音素に対する信頼度データとして付加する、ことを特徴とする。

本態様においては、一致区間に対立する区間を枝刈りしているため、一致区間の事後確率は１となる。つまり、この一致区間に含まれる音素には最高信頼度を表わす信頼度データ（＝１）が付加される。そして、本態様では、音響処理装置は、認識結果データの形態素ラティスを利用し、最尤仮説とテキストデータとの一致区間以外の形態素にも非零の信頼度を付与する。これは、時間軸上で最尤仮説と対立する（重複する）仮説の枝刈り処理を行なった後の認識仮説のラティスを用いて、最尤仮説の各形態素の事後確率を算出し、この事後確率を信頼度として付与する方法である。

これにより、一致区間に最高信頼度を表わす信頼度データを付加し、その他の区間にも事後確率に応じた非零の信頼でデータを付加し、その重みを用いて音響モデルの学習を行なえる。

【００１６】

上記の各態様の音響処理装置は、最尤仮説に含まれる各形態素を教師とし、これに基づいて信頼度を付与し、音響モデルの学習を行なっている。

これに対して、本発明の第４の態様による音響処理装置は、音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部と、前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理部と、前記認識処理部が出力した認識結果データと前記音声に対応するテキストデータとを読み込み、前記認識結果データに含まれる最尤パスと前記テキストデータが一致する一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝狩り後の認識結果データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる各々の音素に対する信頼度データとして付加した教師系列データを求め、前記教師系列データとこれに対応する音響特徴量データとを、学習データとして出力する選択処理部と、前記選択処理部が出力した学習データを読み込み、フォワードバックワードアルゴリズムを用いて前記学習データに含まれる各々の音素の観測確率を求め、前記音素ごとに、前記観測確率で重み付けられた前記音響特徴量データとして得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化部とを具備することを特徴とするものである。

前記の第１の態様による音響処理装置と比較して、本態様の特徴は、音響処理装置が、観測確率で重み付けられた音響特徴量データとして得られる音響モデルを用いて音響モデル記憶部を更新する点である。また、選択処理部は、認識結果データに含まれる最尤パスとテキストデータが一致する一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝狩り後の認識結果データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる各々の音素に対する信頼度データとして付加する。

本態様では、音響処理装置は、枝刈り処理後の認識仮説のラティス（このラティスには、下で述べる観測確率に信頼度が含まれている）を教師系列として、このラティスに直接フォワードバックワードアルゴリズムを適用して、音響モデルを学習する。この方法では、単語の信頼度（事後確率）は直接用いず、フォワードバックワードアルゴリズムにより得られる観測確率に信頼度が含まれている。

【００１７】

本発明の第５の態様による音響処理装置は、前記の音響処理装置において、前記選択処理部は、前記音声に対応するテキストデータを読み込み、前記認識結果データから、前記テキストデータとの一致が最大となるようなパスを選択して前記教師系列データとして取得するとともに、当該教師系列データと前記テキストデータが一致する一致区間に含まれる前記音素に対しては最高信頼度を表わす信頼度データを付加することを特徴とするものである。

本態様では、音響処理装置は、認識仮説のラティスとテキストデータの一致が最大になる形態素系列を教師系列とする。そして、一致区間には最高信頼度を表わす信頼度データ

を付加し、その他の区間にも適宜信頼度データを付加する。つまり、最尤仮説ではないパス内に一致区間が存在する場合にもそのような一致区間に最高信頼度が与えられる。これにより、最尤仮説とテキストデータとの間の表記の揺らぎや形態素分割の異なりに起因して不一致区間と判定されてしまう区間を一致区間として利用し、音響モデルを学習することができる。

【００１８】

本発明の第６の態様による音響処理装置は、音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部と、前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理部と、前記認識処理部が出力した認識結果データと前記音声に対応するテキストデータとを読み込み、前記認識結果データから、前記テキストデータとの一致が最大となるようなパスを選択して前記教師系列データとして取得するとともに、当該教師系列データ内で前記認識結果データと前記テキストデータとが一致する一致区間を選択し、前記一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝狩り後の認識結果データに含まれる言語的単位ごとの事後確率を前記言語的単位に含まれる各々の音素に対する信頼度データとして付加した教師系列データを求め、前記教師系列データとこれに対応する音響特徴量データとを、学習データとして出力する選択処理部と、

前記選択処理部が出力した学習データを読み込み、フォワードバックワードアルゴリズムを用いて前記学習データに含まれる各々の音素の観測確率を求め、前記音素ごとに、前記観測確率で重み付けられた前記音響特徴量データとして得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化部とを具備することを特徴とする。

前記の第４の態様による音響処理装置が最尤パスとテキストデータとが一致する一致区間を基準として対立する区間を枝刈りするのに対して、この第６の態様による音響処理装置は、認識結果データの全体（最尤パス以外の仮説も含む）とテキストデータとが一致する一致区間を基準として対立する区間を枝刈りする点が特徴である。

【００１９】

また、本発明の一態様は、音素と該音素に対応する音響特徴量とを関連付けた音響モデルを記憶する音響モデル記憶部を備えるコンピュータに、前記音響モデル記憶部から読み出した音響モデルを用いて音声の認識処理を行い、認識仮説を含んだ認識結果データを出力する認識処理過程と、前記認識処理過程で出力した認識結果データに基づき、教師系列データを取得し、前記教師系列データに含まれる各々の音素に信頼度データを付加し、前記信頼度データが付加された前記教師系列データと該教師系列データに対応する音響特徴量データとを、学習データとして出力する選択処理過程と、前記選択処理過程で出力した学習データを読み込み、前記学習データに含まれる各々の前記音素の観測確率データを算出し、前記音素ごとに、前記学習データに含まれる前記音響特徴量データと算出された前記観測確率データとに基づき、且つ、前記信頼度データを重みとして乗じて得られる音響モデルを用いて、前記音響モデル記憶部を更新する音響モデル適応化過程との処理を実行させるプログラム。

【００２０】

また、さらに、次のＡ１からＡ５までの態様として上記課題を解決するようにしても良い。

【００２１】

［Ａ１］本発明の一態様は、上記の音響処理装置において、前記音響モデル記憶部は、前記音響モデルを番組ごとに記憶するものであり、前記認識処理部は、前記番組の音声を、当該番組に対応した前記音響モデルを用いて音声の認識処理を行ない、前記選択処理部は、前記番組ごとに前記学習データを出力し、前記音響モデル適応化部は、当該番組用の前記音響モデルを更新することを特徴とする。

この構成によれば、認識対象の音声は放送等の番組の音声であり、番組ごとに音響モデルを持ち、番組ごとの認識結果を用いて音響モデルの適応化を行なえる。よって、番組ごとに特有の音響的特徴がある場合に、その特有の特徴に応じた適応化を行なうことができ

、認識精度が上がり、より多量の学習モデルを獲得できる。

【００２２】

［Ａ２］本発明の一態様は、上記の音響処理装置において、言語モデルを記憶する言語モデル記憶部をさらに具備し、前記認識処理部は、前記言語モデル記憶部から読み出した前記言語モデルと前記音響モデル記憶部から読み出した前記音響モデルとを所定の重み比率値で重み付けして用いて、認識処理を行なうものであり、且つ、前記認識処理部は、音響モデル適応化部によって更新された前記音響モデルを用いて再度認識処理を行なうものであり、このとき使用する前記重み比率値は、前回の認識処理のときよりも前記言語モデルがより小さい重みで用いられる値とすることを特徴とする。

この構成によれば、認識処理と、認識処理結果を用いた音響モデルの適応化とを、繰り返し行なう場合に、徐々に、言語モデルを用いる重みが相対的に低下する。よって、当初の認識率を高めるために言語モデルの重みを大きめにとった場合にも、音響モデルの適応化の進展に合わせて、言語モデルの重みを減ずることができ、認識精度が上がり、より多量の学習モデルを獲得できる。

【００２３】

［Ａ３］本発明の一態様は、上記の音響処理装置において、前記音声の中の発話部分の特徴量を表わす発話モデルを記憶する発話モデル記憶部と、前記発話モデル記憶部から読み出した前記発話モデルを用いて前記音声の中の発話部分の音声区間を切り出す音声区間切り出し部と、前記選択処理部が出力した前記学習データを用いて、前記発話モデル記憶部に記憶される前記発話モデルを更新する発話モデル適応化部とをさらに具備し、前記認識処理部は、前記音声区間切り出し部によって切り出された前記音声区間の音声を認識処理することを特徴とする。

この構成によれば、発話モデルを用いて音声区間の中の発話部分の音声区間を切り出す。そして、認識処理の結果得られた学習データを用いて、発話モデルを適応化する。よって、発話部分の音声区間の切り出しと、切り出された音声区間の認識処理と、認識結果に基づく学習データを用いた発話モデルの適応化の処理を繰り返すことができる。よって、音声区間の切り出しの精度が向上し、認識精度が上がり、より多量の学習モデルを獲得できる。

【００２４】

［Ａ４］本発明の一態様は、上記の音響処理装置において、話者ごと又は話者属性ごとの統計量を表わす話者属性別発話モデルを記憶する話者属性別発話モデル記憶手段と、前記話者属性別発話モデル記憶手段から読み出した前記話者属性別発話モデルを用いて前記音声の中の所定音声区間における話者又は話者属性を推定する話者属性推定手段とをさらに具備し、前記音響モデル記憶部は、前記話者ごとまたは前記話者属性ごとに区別して前記音響モデルを記憶するものであり、前記認識処理部は、前記話者属性推定手段による推定結果に応じて、該当する前記話者用または前記話者属性用の前記音響モデルを前記音響モデル記憶部から読み出し、この読み出された前記音響モデルを用いて前記音声の認識処理を行なうものであり、さらに、前記話者属性推定手段による推定結果に対応する前記学習データを用いて、前記話者属性別発話モデル記憶手段に記憶された、当該推定結果に対応する前記話者ごとまたは前記話者属性ごとの話者属性別発話モデルを更新する話者属性別発話モデル適応化手段を備えることを特徴とする。

この構成によれば、話者属性別発話モデルを用いて話者又は話者属性を推定する。また、話者又は話者属性ごとの音響モデルを備えており、前記の推定結果に応じた音響モデルを用いた認識処理を行なう。この認識処理の結果得られる学習データを用いて話者属性別発話モデルの適応化を行なう。この処理を繰り返すことにより、話者属性別発話モデルの質が向上し、話者又は話者属性を推定する精度が向上し、即ち、話者又は話者属性ごとの音響モデルに適した音声区間を切り出す精度が向上する。よって、認識精度が上がり、より多量の学習モデルを獲得できる。

【００２５】

［Ａ５］本発明の一態様は、上記の音響処理装置において、前記音響モデル適応化部は

、前記話者属性推定手段による推定結果に対応する前記学習データを用いて、当該推定結果に対応する前記話者ごとまたは前記話者属性ごとの前記音響モデルを更新するものであることを特徴とする。

この構成によれば、前記の推定結果に対応した学習データを用いて、当該推定結果に対応した音響モデルを適応化する。よって、音響モデルの質がより向上し、認識精度が上がり、より多量の学習モデルを獲得できる。

【発明の効果】

【００２６】

本発明によると、字幕が付与されている様々な番組音声から、大量の音響モデルの学習データを効率的に得ることができる。また、字幕と実際の音声との一致率が低い番組音声からも、より効率的に、音響モデルの学習データを得ることが可能となる。特に、字幕と音声の一致しない区間の一部からも学習データを得ることができる。また、これまで認識できなかった番組や話者の認識が可能になったり、音声認識の認識精度の向上が可能になったりする。

【発明を実施するための最良の形態】

【００２７】

以下、図面を参照しながら本発明の実施形態について説明する。

［第１の実施の形態］

図１は、第１の実施形態による音響モデル学習装置の機能構成を示すブロック図である。図示するように、音響モデル学習装置１００は、字幕テキスト記憶部１と、形態素解析処理部２と、重み付け処理部３と、ベースラインコーパス記憶部４と、合成処理部５と、バイアス化言語モデル記憶部６（言語モデル記憶部）と、重み付け処理部７と、番組音声記憶部８と、音響モデル記憶部９と、認識処理部１０と、認識仮説記憶部１１と、アライメント処理部１２と、選択処理部１３と、学習データ記憶部１４と、音響モデル適応化部１５とを含んで構成される。

【００２８】

字幕テキスト記憶部１は、放送番組の字幕テキストのデータを記憶する。形態素解析処理部２は、字幕テキスト記憶部１から読み出した字幕テキストについて、既存技術を用いて形態素解析処理を行い、その結果を、形態素に区切られた列として出力する。

【００２９】

重み付け処理部３は、形態素解析された字幕テキストについて、ベースラインコーパスと合成するための重み付けを行なう。つまり、重み付け処理部３は、字幕テキストの出現頻度に重み値W_{bias}を乗ずるようなバイアスをかける。ベースラインコーパス記憶部４は、ベースラインコーパスを記憶する。ベースラインコーパスは、電子化された自然言語の大量の文章からなるデータベースである。例えば、蓄積された過去のニュース原稿のテキストをベースラインコーパスとして利用する。なお、ベースラインコーパスに対して予め統計処理を行い、後述する言語モデルに合う形式で記憶しておくようにしても良い。

【００３０】

合成処理部５は、重み付け処理部３によって重み付けされた字幕テキストのデータと、ベースラインコーパス記憶部４から読み出したベースラインコーパスとを用いて、これらを合成し、出現する語に関する統計的な数値とともに表わしたバイアス化言語モデルを作成する。バイアス化言語モデル記憶部６は、合成処理部５によって作成されたバイアス付きの言語モデルを記憶する。言語モデルの詳細については、後で図面を参照しながら説明する。

【００３１】

重み付け処理部７は、認識処理部１０による認識処理のために、バイアス化言語モデル記憶部６に記憶されている言語モデルに対して重み値W_{lm}による重み付けを行なう。これにより、認識処理部１０は、言語モデルと音響モデルとをW_{lm}：１の重みの比で用いることとなる。

【００３２】

番組音声記憶部８は、番組音声を記憶している。この番組音声は、字幕テキスト記憶部１に記憶される字幕テキストのデータと対応するものである。

音響モデル記憶部９は、音素とその音素に対応する音響特徴量とを関連付けた音響モデルのデータを記憶する。音響モデルのデータについては後でも説明する。

【００３３】

認識処理部１０は、音響モデル記憶部９から読み出した音響モデルと重み値W_{lm}による重み付けをつけた言語モデルとを用いて、番組音声記憶部８から読み出す音声の認識処理を行い、認識仮説（認識結果データ）を出力する。認識仮説記憶部１１は、この認識仮説を記憶する。

【００３４】

アライメント処理部１２は、形態素解析処理部２によって形態素解析処理済みの字幕テキストデータと、認識仮説記憶部１１から読み出す認識仮説との一致部分を見つける処理を行なう。具体的には、アライメント処理部１２は、これら両データをサーチし、所定数以上連続して語が一致しているか所定数以上連続して文字が一致している箇所を見つける処理を行なう。

【００３５】

選択処理部１３は、アライメント処理部１２によってアライメントされたデータに基づいて、認識仮説の各パスに信頼度データを付加する処理を行なう。

一例として、選択処理部１３は、最尤パスと字幕テキストデータとが一致する区間（およびその区間に含まれる音素）には信頼度１（最高信頼度を表わす）を与え、その他の区間（およびその区間に含まれる音素）には信頼度０（最低信頼度を表わす）を与える。そして、選択処理部１３は、信頼度データが付加された認識結果データ（これが教師系列データであり、この系列中に音素が含まれる）と、これに対応する音響特徴量データとを、対応付けて出力する。具体的には、選択処理部１３は、このデータを学習データ記憶部１４に書き込む。

【００３６】

学習データ記憶部１４は、選択処理部１３によって出力される学習データを記憶する。音響モデル適応化部１５は、学習データ記憶部１４から読み出した学習データを用いて、音響モデル記憶部９に記憶される音響モデルを適応化する（更新する）処理を行なう。

【００３７】

この音響モデル学習装置１００の特徴は、音響モデル適応化部１５が存在し、選択処理部１３によって出力される学習データを用いて、最大事後推定法（ＭＡＰ）や最尤線形回帰分析法（ＭＬＬＲ）等の適応化技術により、音響モデル記憶部９に記憶されている音響モデルを適応化（更新）し、さらに適応化された音響モデルを用いて認識処理を繰り返し行なうようにした点である。

【００３８】

また、従来技術と異なり、比較的大きな重み値W_{bias}（１０００程度）と重み値W_{lm}（１６程度）を用いて認識処理を行なう点も、特徴的である。字幕と発話の一致率が低く、自由発話の比率が高い番組では、上記のような重み値を用いることが、認識精度の向上を図りながらより多くの学習データを得られることにつながる。

これは、実験結果からも得られた適切な重み値である。具体的には、本願発明者らが、ＮＨＫ（日本放送協会）の実際に番組の音声および字幕テキストを用いて行なった実験では、W_{bias}＜５０００の領域において、W_{lm}の変化による単語誤認識率（ＷＥＲ）の差が小さい。また、W_{lm}＜１８の領域において、W_{bias}の変化による単語誤認識率の差が小さい。そして、W_{bias}＜５０００且つW_{lm}＜１８の場合に、単語誤認識率がそれほど上がらずに、且つ得られる学習データの量が多くなる。そして、W_{bias}が１０００程度でW_{lm}が１６程度のときに、特に、単語誤認識率がそれほど上がらずに、且つ得られる学習データの量が多くなる。

【００３９】

次に、言語モデルについて説明する。

図２は、バイアス化言語モデル記憶部６によって記憶され認識処理部１０によって認識処理の際に使用される言語モデルのデータの構成を示す概略図である。
図２（ａ）は、形態素単体での出現確率を表わすテーブルを示す。このテーブルは、表形式のデータであり、形態素と出現確率の項目（列）を有する。テーブルの行は、形態素の種類ごとに存在する。各形態素と、その形態素の出現確率とが関連付けられてこのテーブルに保持される。
図２（ｂ）は、形態素と形態素の接続確率を表わすテーブルである。このテーブルは、表形式のデータであり、形態素１と形態素２と接続確率の項目（列）を有する。テーブルの行は、形態素１と形態素２の組み合わせごとに存在する。各行は、形態素１が出現したときその形態素１に直ちに続いて形態素２が出現する（接続する）確率である接続確率の情報を表わしている。

【００４０】

次に、音響モデルについて説明する。
図３は、音響モデル記憶部９によって記憶され認識処理部１０によって認識処理の際に使用される音響モデルのデータの構成を示す概略図である。図示するように、音響モデルは、音素と、その音素が発話されるときの音響的特徴量とを関連付けて保持するデータである。音響的特徴量としては、例えば、１０ミリ秒ごとの間隔でそれぞれ開始する２５ミリ秒のフレーム内の周波数パワー分布を基にしたＭＦＣＣ（メル周波数ケプストラム係数, Mel-Frequency Cepstrum Coefficient）や、ＰＬＰ（Perceptual Linear Predictive）などを用いる。

【００４１】

次に、認識処理部１０による認識結果である認識仮説について説明する。
図４は、認識処理部１０によって出力され認識仮説記憶部１１によって記憶される認識仮説のデータ構成およびデータ例を示す概略図である。認識仮説は、論理的には、言語的単位（形態素）をアークとするラティス構造（半順序（partial order）構造）のグラフで表現される。

【００４２】

図４（ａ）は、表形式で表現される認識仮説のデータ構成を示す。図示するように、表形式で表わした認識仮説は、各行がラティス上の各アークに対応し、始点ノード、終点ノード、言語的単位、信頼度の各項目（列）を有する。始点ノードおよび終点ノードは、ラティス上のノードに便宜的に与えられたノード識別情報を値として持つ。言語的単位は、そのアークに対応する形態素である。信頼度は、認識結果におけるそのアークの信頼度（言語モデルおよび音響モデルを基にした統計的処理で得られる認識結果の信頼度）を表わす数値で表わされるものであり、その数値の範囲は０以上１以下である）。図示しているデータ例は、例えば１行目のデータに相当するアークの始点ノードは「１」であり、終点ノードは「２」であり、言語的単位は「Ｆ」である。また例えば２行目のデータに相当するアークの始点ノードは「１」であり、終点ノードは「２」であり、言語的単位は「Ｈ」である。３行目以下のデータも同様である。同図では信頼度数値の記載を省略している。

【００４３】

図４（ｂ）は、図４（ａ）の表形式のデータが表現するラティスを絵的に表現した概略図である。同図における丸印がそれぞれノードに相当し、アークはノード間を結ぶ矢印付きの線で表わされている。また、各アークに対応する言語的要素が付記されている。例えば一番左のノード（ノード「１」）を始点とするアークは３本あり、それらはそれぞれ、言語的要素が「Ｆ」で終点がノード「２」（左から二番目のノード）、言語的要素が「Ｈ」で終点がノード「２」、言語的要素が「Ｍ」で終点がノード「３」（左から三番目のノード）である。

【００４４】

次に、音響モデル学習装置１００の動作および処理手順について説明する。
図５は、音響モデル学習装置１００全体の処理手順を示すフローチャートである。
図示するように、ステップＳ０１において、まず形態素解析処理部２が、字幕テキスト

記憶部１から読み出した字幕テキストの形態素解析処理を行なう。形態素解析処理は、既存の技術により可能である。

【００４５】

次に、ステップＳ０２において重み値W_{bias}を用いて形態素解析処理結果に重み付けをするとともに、ステップＳ０３において重み付けされた形態素解析処理結果とベースラインコーパスとを合成する処理を行なう。具体的には、合成処理部５は、形態素解析処理部２による処理結果を統計処理し、各形態素の出現確率を算出するとともに、２つの連続する形態素間の接続確率を算出し、図２（ａ）に示した形式の出現確率データおよび図２（ｂ）に示した形式の接続確率データを得る。また、合成処理部５は、ベースラインコーパス記憶部４に記憶されているベースラインコーパスを基に、必要に応じて同様の統計処理を行い、図２（ａ）に示した形式の出現確率データおよび図２（ｂ）に示した形式の接続確率データを得る。これらのうち形態素解析処理部２による字幕テキスト１の処理結果には重み付け処理部３が重み値W_{bias}を用いた重み付けを行い、合成処理部５が両者を合成するための確率計算を行ない、その結果としてバイアス化言語モデルが算出され、これがバイアス化言語モデル記憶部６に書き込まれる。

【００４６】

次に、ステップＳ０４において重み値W_{lm}を用いてバイアス化言語モデル記憶部６から読み出される言語モデルに重み付けするとともに、ステップ０５において認識処理部１０は、重み付けされたバイアス化言語モデルと音響モデル記憶部９から読み出した音響モデルを用いて、番組音声記憶部８に記憶されている番組音声の認識処理を行い、その結果として生成される認識仮説を認識仮説記憶部に書き込む。なお、生成される認識仮説は文字列のデータであり、この文字列を形態素ごとに区切る情報を伴っている場合もある。ここで用いられる認識処理自体は、既存技術である。

【００４７】

次に、ステップＳ０６において、アライメント処理部１２は、形態素解析処理された字幕テキストと認識処理の結果得られた認識仮説とのアライメント処理を行なう。このアライメント処理は、両データが単語単位で一致する部分を探索することによって行ない、或いは両データ中の複数の単語が連続して一致する区間を探索することによって行なう。アライメント処理の結果、字幕テキストの中の区間と番組音声の中の区間がデータ的に対応付けられる。

【００４８】

次に、ステップＳ０７において、選択処理部１３は、認識仮説に信頼度を付加する処理を行なう。具体的な処理の一例は次の通りである。選択処理部１３は、まずアライメント処理部１２によるアライメント結果を用いて、元の字幕テキストと認識仮説に含まれる最尤パスとが３単語以上連続して一致している一致区間を選択し、その一致区間（およびその区間に含まれる各音素）に対しては、信頼度１を与える。そして、選択処理部１３は、その他の区間に対しては信頼度０を与える。そして、選択処理部１３は、そのようにして得られた学習データを、学習データ記憶部１４に書き込む。

【００４９】

次に、ステップＳ０８において、繰り返し処理を終了するか否かを判定する。この判定は、例えば、認識結果の精度が十分かどうかにより行なう。より具体的には、この判定は（ａ）ステップＳ０５～Ｓ０９のループを繰り返した回数（例えば、この回数を３～４回として良い場合もある）、（ｂ）番組音声全体のうちの選択処理部１３によって選択された一致区間の長さの比率、（ｃ）番組音声全体のうちの選択処理部１３によって選択された一致区間の前回からの増分比率、などのいずれかによって行なう。

判定の結果、繰り返し処理を終了しない場合（ステップＳ０８：ＮＯ）には、次のステップＳ０９に進む。

判定の結果、繰り返しを終了する場合（ステップＳ０８：ＹＥＳ）には、このフローチャート全体の処理を終了する。このとき、学習データ記憶部１４に累積的に書き込まれたデータが、本装置による学習処理の結果として得られた音響モデルである。

【００５０】
ステップＳ０９においては、音響モデル適応化部１５は、得られた学習データを用いて、音響モデル記憶部９に記憶されている音響モデルを適応化する（更新する）処理を行なう。なお、音響モデル適応化部１５が如何に学習データを用いて具体的に音響モデルの更新を行うかについては、後で詳述する。
【００５１】
上記のように、音響モデル学習装置１００は、選択処理部１３によって出力される学習データを用いて、音響モデル適応化部１５が、音響モデル記憶部９に記憶されている音響モデルを適応化し、さらに適応化された音響モデルを用いて認識処理を繰り返し行なうようにしている。
【００５２】
次に、上述した音響モデル適応化部１５による音響モデルの適応化処理について詳しく説明する。本実施形態では、音声認識による音声認識区間ｒをそのまま用いて音響モデルを学習するため、認識結果の各形態素や音素ｊの信頼度を導入する。音響モデル適応化部１５は、フォワードバックワードアルゴリズムを用いて、各時刻における音素ｊの観測確率に、信頼度による重みを与えて最尤推定を行ない、これにより得られる平均ベクトルを用いて音響モデル記憶部９に記憶されている音響モデルの更新を行なう。最尤推定により得られる音素（状態）ｊの平均ベクトルは、下の式（１）により計算される。
【００５３】
【数１】

$$\hat{\mu}_j = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} C_j^r L_j^r(t) o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} C_j^r L_j^r(t)} \quad \cdots \quad (1)$$

【００５４】
【数２】

$$o_t^r$$

【００５５】
は、音声認識による音声切り出し区間ｒの時刻ｔにおける特徴量の観測ベクトルである。
また、
【００５６】
【数３】

$$L_j^r(t)$$

【００５７】
は、音声切り出し区間ｒの時刻ｔにおける音素ｊの観測確率である。なお、音響モデル適応化部１５が、動的計画法の一種であるフォワードバックワードアルゴリズムによる手順を実行することにより、この観測確率を得ることができる。
また、

【００５８】
【数４】

$$C_j^r$$

【００５９】
は、音声切り出し区間ｒにおける音素ｊの信頼度である。
本実施形態では、選択処理部１３は、次の式（２）により信頼度を与えている。
【００６０】
【数５】

$$C_j^r = \begin{cases} 1 & (\text{最尤仮説と字幕の一致区間}) \\ 0 & (\text{その他の区間}) \end{cases} \quad \cdots (2)$$

【００６１】
つまり、認識結果の最尤仮説のパスと元の字幕テキストとが一致する区間においては信頼度は１（信頼度最大を表わす）であり、その他の区間においては信頼度０（信頼度最低を表わす）とする。
【００６２】
図６は、本実施形態において選択処理部１３が付加する信頼度の例を表わす概略図であり、最尤仮説のパスと書き起こしによる字幕テキストの一致を利用する方法による信頼度を表わしている。図６（ａ）は、書き起こしによる字幕テキストに対応するグラフを示す。図示するように、字幕テキストに対応するグラフは、形態素ａ、ｂ、ｃ、ｄ、ｅにそれぞれ対応するアークが直列につながって構成されている。図６（ｂ）は、最尤仮説の教師単語列に対応するグラフを示す。図示するように、最尤仮説に対応するグラフは、形態素ａ、ｆ、ｃ０、ｃ１、ｑ、ｅ´にそれぞれ対応するアークが直列につながって構成されている。また、図６（ｂ）に付記されている括弧内の数値は、信頼度である。本例において両者を比較すると、形態素ａおよびｄが一致しており、その他の形態素は不一致である。したがって、選択処理部１３は、形態素ａおよびｄの区間にはそれぞれ信頼度１．０を与え、その他の区間には信頼度０．０を与えている。
そして、選択処理部１３は、各々の区間の信頼度を、その区間に含まれる音素の信頼度として付加する。
【００６３】
本実施形態による音響処理装置の特徴は、音声認識による音声切り出し区間ｒをそのまま用いて音響モデルの学習を行なうため、認識結果に含まれる状態（単語や音素）ｊの信頼度を導入した点にある。そして、音響モデル適応化部１５が音響モデルの統計量の推定を行なう際に、フォワードバックワードアルゴリズムにより得られる音素ｊの観測確率に、信頼度による重みを与えている。そして、音響モデル適応化部１５は、上記の式（１）による計算で得られる音素ｊの平均ベクトルを用いて、音響モデル記憶部９に記憶されている音響モデルを更新する。
【００６４】
また、式（２）で表わされる信頼度を付与することにより、認識結果の最尤パスと字幕テキストの一致区間に含まれる音素に対応付けられた特徴量のみを、音響モデルの学習に利用することができる。
【００６５】
なお、式（１）による計算では最尤推定に基づく特徴量の平均ベクトルを得てこれを音

響モデルの適応化に用いたが、その代わりに、最尤推定に基づく分散や混合重みなどを用いて音響モデルを更新するようにしても良い。さらに、最尤推定に限らず、ＭＡＰ推定（maximum a posteriori estimation、最大事後確率推定）やＭＬＬＲ（Maximum Likelihood Linear Regression、最尤線形回帰）推定などを用いるようにしても良い。

【００６６】

ここで、仮に、

【００６７】

【数６】

$$C_j^r$$

【００６８】

をすべての区間において等しく１にすると、これは、認識結果と字幕テキストの一致に関する情報を全く利用しない教師なし学習となる。

【００６９】

また、上記の方法では、フォワードバックワードアルゴリズムにより得られる観測確率を用いたが、その代わりに、例えばビタビアルゴリズムを用いて得られる最尤パス（ビタビパス）を利用するようにしても良い。この場合は、最尤パス上の観測確率には１を与え、その他のパス上の観測確率には０を与えるようにする。

【００７０】

図７は、音響モデル学習装置１００による音響モデルの学習の概略を補足的に説明するための図である。

図７において、形態素解析結果は字幕テキストを形態素解析して得られたデータであり、このデータは形態素（単語）の一次元的な列である（図中のそれぞれの四角が形態素に相当）。また、認識結果は、認識処理部１０による音声認識処理の結果得られる認識仮説のラティスに含まれるパスの一部に相当するデータである（図中のそれぞれの四角が形態素に相当）。そして、この図においては形態素解析結果と最尤パスとが一致する一致区間をハッチングで表わしている。また、これらの各々の区間には、音響特徴量および信頼度が対応付いている。本実施形態では、上記の一致区間の信頼度は１．０であり、その他の区間の信頼度は０．０である。そして、各区間に含まれる音素と、対応する音響特徴量と、対応する信頼度とを関連付けたものが、選択処理部１３によって得られる学習データである。

【００７１】

本実施形態では、認識処理の結果である学習データを用いて、音響モデルの適応化を行なっている。そして、適応化された音響モデルを用いて再度認識を行い、学習データを生成する。この音響モデルの適応化と、適応化された音響モデルを用いた認識処理を繰り返すことにより、学習データの正確性が増し、より多くの量の学習データを効率的に得ることができる。

【００７２】

［第２の実施の形態］

次に第２の実施形態について説明する。本実施形態では、認識結果の単語ラティスを用いて、最尤パスと字幕テキストとの一致区間以外の形態素をも学習データとして利用する。

なお、以下では、本実施形態特有の部分のみを説明し、その他の部分については前記の実施形態１と同様であるため説明を省略する。

【００７３】

本実施形態においては、選択処理部１３は、認識仮説のラティスの枝刈り処理を行い、

枝刈りの後のラティス内の最尤パスの各形態素の事後確率を求め、この事後確率を各形態素および形態素内の各音素の信頼度として付与する。ここで、枝刈りの対称となるのは、ある時間において最尤パスと対立するすべてのパスである。このような枝刈り処理の結果として、最尤パスと字幕テキストが一致する区間の各形態素の信頼度は１（信頼度最大を表わす）となる。つまり、最尤パスと字幕テキストが一致する区間の形態素については、第１の実施形態と同様の学習を行なうこととなる。

【００７４】

次に、選択処理部１３による認識仮説の枝刈りの処理についてより詳細に説明する。

図８は、選択処理部１３によって枝刈りされた後の認識仮説のデータ例を示す概略図である。そして、図４に示したデータが、選択処理部１３によって枝刈りされる前の認識仮説のデータである。

まず、選択処理部１３は、図４に示した認識仮説のデータと字幕テキストのデータとの一致区間を選択する。ここでは選択処理部１３は１形態素以上連続して両者が一致する区間を一致区間として選択するものとする。その結果、形態素Ｂ（始点ノードが「２」で終点ノードが「３」）と形態素Ｃ（始点ノードが「３」で終点ノードが「４」）と形態素Ｅ（始点ノードが「５」で終点ノードが「６」）が一致区間として選択される。

【００７５】

その後、選択処理部１３は、上で選択された一致区間の時間を含み且つ字幕テキストのデータとは一致しない認識仮説を枝刈りする。具体的には、図４に示した形態素Ｂの区間の時間を含み且つ形態素Ｂとは一致しない認識仮説（言い換えれば、形態素Ｂと対立する認識仮説）である形態素ＩとＭとＮは、選択処理部１３による枝刈りの対象となる。同様に、一致区間として選択された形態素Ｃと対立する認識仮説である形態素ＪとＮとＯは、選択処理部１３による枝刈りの対象となる。同様に、一致区間として選択された形態素Ｅと対立する認識仮説である形態素ＬとＰは、選択処理部１３による枝刈りの対象となる。

つまり、図４に示すラティスのうち、形態素ＩとＪとＬとＭとＮとＯとＰのそれぞれのアークが選択処理部１３によって枝刈りされる。その枝刈りの結果として残るのは、形態素ＢとＣとＥとＦとＧとＨとＫのそれぞれのアークであり、これらのアークに対応するデータが、図８に示すデータである。

【００７６】

図８（ａ）は、そのようなラティス構造のグラフを表わす表形式のデータであり、この表は、前記の形態素ＢとＣとＥとＦとＧとＨとＫに対応する行のデータを有している。またこの表において、形態素ＢとＣとＥは、それぞれ一致区間に含まれていたものであるため、その結果として、選択処理部１３は、これらの形態素に対応する信頼度がいずれも最高の１となるように認識仮説のデータを更新している。また、その他の、形態素ＦとＨとＫとＧのアークについては、選択処理部１３は、枝刈り後の事後確率を算出し、その算出結果をそれぞれの信頼度データとして認識仮説のデータを更新する。図示する例では、形態素Ｆの信頼度は０．１、形態素Ｈの信頼度は０．９、形態素Ｋの信頼度は０．２、形態素Ｇの信頼度は０．８である。

図８（ｂ）は、図８（ａ）の表形式のデータが表現するラティスを絵的に示した概略図である。

【００７７】

なお、選択処理部１３が事後確率を算出する方法は次の通りである。選択処理部１３は、動的計画法の一種であるフォワードバックワードアルゴリズムを用いることで、デコーダー（認識処理部１０）の出力から得られる形態素ｍのスコアＬ（ｍ）を基に、事後確率を算出する。ここで、Ｌ（ｍ）としては、音響モデルの尤度、言語モデルの尤度、あるいはそれらの和を用いる。まず、形態素ｍの前向き確率α（ｍ）と後ろ向き確率β（ｍ）とを以下の式（３）～（７）の通り定義する。

【００７８】

【数７】

$$\alpha(b)=1.0 \qquad \cdots (3)$$

【００７９】
【数８】

$$\beta(\mathrm{e})=1.0 \qquad \cdots (4)$$

【００８０】
【数９】

$$\alpha(\mathrm{m})=\sum_{m'\in LEFT\,(m)}\alpha(m')L(m) \qquad \cdots (5)$$

【００８１】
【数１０】

$$\beta(\mathrm{m})=\sum_{m'\in RIGHT\,(m)}\beta(m')L(m) \qquad \cdots (6)$$

【００８２】
【数１１】

$$Z=\alpha(e)=\beta(b) \qquad \cdots (7)$$

【００８３】
ここでは、事後確率を計算するため、図４（ｂ）などに示したラティスの両端に、それぞれ、始端単語と終端単語に相当するアークを追加する。始端単語はその図のラティスの左端に追加され、「ｂ」と表わされる。終端単語はその図のラティスの右端に追加され、「ｅ」と表わされる。ここで前向きとは、始端から終端の方向であり、図のラティスでは左から右の方向である。また後ろ向きとは、終端から始端の方向であり、図のラティスでは右から左の方向である。また、ＬＥＦＴ（ｍ）は、形態素ｍに左から（つまり始端側から）接続する形態素の集合である。また、ＲＩＧＨＴ（ｍ）は、形態素ｍに右から（つまり終端側から）接続する形態素の集合である。
【００８４】
式（３）に表わすように、始端単語（ｂ）の前向き確率は１．０である。また、式（４）に表わすように、終端単語（ｅ）の後ろ向き確率は１．０である。
式（５）に表わすように、形態素ｍの前向き確率は、形態素ｍに左から接続する各形態素ｍ´の前向き確率α（ｍ´）に当該形態素ｍのスコアＬ（ｍ）を乗じた値の、左から接続する全ての形態素についての総和である。このように前向き確率は再帰的な定義となっているが、始端単語から前向きに順次計算していくことにより、ラティス中の全ての形態

素の前向き確率を算出できる。

式（６）に表わすように、形態素ｍの後ろ向き確率は、形態素ｍに右から接続する各形態素ｍ´の後ろ向き確率β（ｍ´）に当該形態素ｍのスコアＬ（ｍ）を乗じた値の、右から接続する全ての形態素についての総和である。このように後ろ向き確率は再帰的な定義となっているが、終端単語から後ろ向きに順次計算していくことにより、ラティス中の全ての形態素の後ろ向き確率を算出できる。

式（７）に表わすように、終端単語の前向き確率および始端単語の後ろ向き確率をＺとする。

【００８５】

そして、形態素ｍの事後確率γ（ｍ）は、α、β、Ｚを用いて、式（８）により算出することができる。

【００８６】

【数１２】

$$\gamma(m)=\frac{\alpha(m)\beta(m)L(m)}{Z} \quad \cdots \quad (8)$$

【００８７】

以上のように、選択処理部１３は、認識結果に含まれる最尤パスと字幕テキストとの一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝刈り処理の結果を教師系列データとして取得し、この教師系列データに含まれる形態素ごとの事後確率をその形態素に含まれる音素の信頼度データとして付加する。

そして、音響モデル適応化部１５は、選択処理部１３が得た信頼度データを用いて、第１の実施形態と同様に、例えば前記の式（１）を用いて音素に対応するベクトルの更新式を得て、これにより音響モデル記憶部９に記憶された音響モデルを更新する。

【００８８】

［第３の実施の形態］

次に第３の実施形態について説明する。本実施形態では、選択処理部１３が、第２の実施形態における枝刈り処理と同様の処理を行なう。そして、枝刈り後の認識結果のラティスに対して、直接フォワードバックワードアルゴリズムによる処理を行い、各音素の統計量を学習する。この場合、フォワードバックワードアルゴリズムの処理より得られる観測確率に、既に信頼度が含まれる。

【００８９】

第２の実施形態の場合と同様に本実施形態においても、選択処理部１３は、枝刈り処理を行なう結果、最尤パスと字幕テキストとの一致区間では信頼度１（信頼度最大を表わす）を付与する。一方で、最尤パスと字幕テキストが一致しない区間においては、選択処理部１３は、各々のパスの事後確率に応じた信頼度を付与する。

【００９０】

そして、音響モデル適応化部１５は、得られている観測ベクトルと、選択処理部１３により求められた観測確率とに基づき、下の式（９）を用いて音素ｊの平均ベクトルを計算する。

【００９１】

【数１３】

$$\hat{\mu}_j = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_j^r(t)\,o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_j^r(t)} \qquad \cdots \quad (9)$$

【００９２】
この式（９）においては、
【００９３】
【数１４】

$$L_j^r(t)$$

【００９４】
は、区間ｒの時刻ｔにおける音素ｊの観測確率であり、この観測確率は上述したフォワードバックワードアルゴリズムによる処理を用いた方法で選択処理部１３によって算出されたものであり、信頼度が含まれている。
【００９５】
本実施形態で上記の平均ベクトルを計算する場合、式（９）からも明らかなように、ある時刻において対立する複数のパス（候補）間で、共通して出現する度合いの高い音素については、対立候補それぞれにおける当該音素の観測確率がすべて加算されるため、当該音素に対応する観測ベクトルの影響が強く効いた平均ベクトルが算出される。一方で、同じく式（９）からわかるように、対立する複数のパスのいずれかにしか現れない音素については、そのパス中の音素に対応する観測ベクトルが平均ベクトルには強い影響を与えない。
【００９６】
図９は、対立する複数のパス間で共通して出現する音素の例を示す概略図である。この図に沿って本実施形態の特徴を補足説明する。この図における時間軸は左から右への方向である。また、この図に示す認識結果のラティスは対立する２系列のパスを含んでいる。図における上側のパスは、音素j_1、j_2、j_3からなる形態素と、音素j_4、j_5、j_6からなる形態素を含んでいる。また、下側のパスは、音素j_7、j_2、j_3からなる形態素と、音素j_8、j_9、j_{10}からなる形態素を含んでいる。ここで、各々の音素の観測確率が時刻ｔをパラメタとする関数になることは式（９）にも表わされているとおりであるが、上記の音素j_1～j_{10}のうちj_2とj_3だけは上側のパスと下側のパスの両方に共通して出現している。従って、式（９）によれば、このような音素j_2およびj_3は、認識結果のラティス全体の中の図示している部分から学習データを得るのに、他の音素（j_1と、j_4～j_{10}）に比べてより大きな重みで作用する。
【００９７】
以上述べたように、本実施形態では、選択処理部１３は、認識結果データに含まれる最尤パスと字幕テキストのデータが一致する一致区間と時間軸上で対立する区間を枝刈りする処理を行ない、この枝狩り後の認識結果データに含まれる形態素ごとの事後確率を当該形態素に含まれる各々の音素に対する信頼度データとして実質的に付加し、この信頼度データが付加された教師系列データ（ラティス）とそれに対応する音響特徴量データとを、学習データとして出力する処理を行なう。
そして、音響モデル適用化部１５は、学習データ記憶部１４からこの学習データを読み

込み、フォワードバックワードアルゴリズムを用いて、教師系列データ（ラティス）に含まれる各音素の観測確率を求め、この観測確率で重み付けられた観測ベクトルの平均ベクトルを音素ごとに算出し、得られた平均ベクトルを音響モデルとして用いて音響モデル記憶部に記憶されている音響モデルを適応化する。

【００９８】

このように、本実施形態による方法では、対立する複数の候補間で、含まれる音素のバリエーションに応じた結果が得られることとなる。

【００９９】

［第４の実施形態］

次に、第４の実施形態を説明する。第１～第３の実施形態が最尤パスと字幕テキストが一致する区間を基準として教師系列を求めていたのに対し、本実施形態の特徴は、認識結果のラティスと字幕テキストが一致する区間を基準として教師系列を作成する点である。

本実施形態では、選択処理部１３は、認識結果のラティスと字幕テキストとのビタビアライメントを取ることにより、認識結果のラティスと字幕テキストが一致する区間を基に教師系列を得る。

【０１００】

図６が最尤パスと字幕テキストとが一致する区間を用いた方法を説明しているのに対して、図１０は、本実施形態による選択処理部１３が、認識結果のラティスと字幕テキストの一致区間を抽出する処理を概念的に説明する概略図である。図示する例では、（ａ）の書き起こしの字幕テキストデータは、ａ－ｂ－ｃ－ｄ－ｅという５つの形態素に対応するアークが直列する形のグラフによって表わされる。一方で、（ｂ）の認識結果のラティスは、複数の対立候補のアーク（各々のアークは形態素に対応する）を含んだ構造を有しており、字幕テキストデータに対応する形態素ａ、ｂ、ｃ、ｅの他にも、形態素ｃ０やｃ１やｅ′や、その他（ｆ、ｇ、ｈ、ｉ、ｊ、ｋ、ｌ、ｍ、ｎ、・・・、以下記載省略）の形態素に対応するアークを含んでいる。そして、（ｃ）は、（ａ）の字幕テキストと（ｂ）の認識結果のラティスを基に、これらの一致区間を抽出して得られた教師形態素列を示している。（ｃ）の例は、ａ－ｂ－ｃ－ｇ－ｅという５つの形態素に対応するアークが直列する形のグラフを表わしている。ここで、形態素ａ、ｂ、ｃ、ｅは字幕テキストと認識結果のラティスが一致する区間であるため、選択処理部１３は、それらに信頼度１．０を付与している。また、字幕テキストにおける形態素ｄに一致するパスが認識結果のラティス内に存在しないため、選択処理部１３は、それに対応する時間の区間からは形態素ｇを選択するとともに、それに信頼度０．１を付与している。ここで、（ｂ）に示した認識結果のラティスにおける最尤パスは、形態素ｃ０－ｃ１の系列を含み、また形態素ｅ′の系列を含んでいるが、これらｃ０、ｃ１、ｅ′はいずれも字幕テキストと一致するものではない。そこで、選択処理部１３は、ｃ０－ｃ１の系列に対立する区間からｃを選び、また、ｅ′に対立する区間であるｅを選んでいる。これは、前述の通り、ｃおよびｅがそれぞれ字幕テキストと一致するためである。

【０１０１】

上記の処理に具体例を当てはめると次の通りである。

例えば、形態素ｃは「ですから」であり、形態素ｃ０－ｃ１の系列は「です／から」に対応する。本実施形態の方法では、認識結果における最尤パスが「です／から」を含むものであっても、字幕テキストと一致する「ですから」を選択して教師形態素列を作る。つまり、最尤パスにおける形態素分割の結果が字幕テキストと異なることに起因して不一致区間と判定されてしまう区間を、一致区間と同等に扱うことができるようになる。

また例えば、形態素ｅは「こと」に対応し、形態素ｅ′は「事」に対応する。本実施形態の方法では、認識結果における最尤パスが「事」を含むものであっても、字幕テキストと一致する「こと」を選択して教師形態素列を作る。つまり、表記の揺らぎに起因して不一致区間と判定されてしまう区間を、一致区間と同等に扱うことができるようになる。

【０１０２】

このように、本実施形態の方法では、選択処理部１３は、最尤仮説と字幕テキストとの

間では不一致となるものの、ラティス内の、その最尤仮説に対立する仮説の中に字幕テキストと一致するテキストがあると期待される場合に、認識結果のラティスが含む情報を最大限に活かすことが可能となる。

【０１０３】

［第５の実施形態］

次に、第５の実施形態を説明する。第５の実施形態は、第４の実施形態と同様に、認識結果のラティスと字幕テキストが一致する区間を基準として教師系列を作成する点である。

本実施形態では、選択処理部１３は、認識結果のラティスと字幕テキストとのビタビアライメントを取ることにより、認識結果のラティスと字幕テキストが一致する区間を得る。

そして、選択処理部１３は、得られた一致区間と対立する区間の枝刈り処理を行なう。ここで、枝刈り処理の手法そのものは第３の実施形態におけるそれと同様であるが、第３の実施形態が最尤パスと字幕テキストとの一致区間を基準としていたのに対し、本実施形態は、認識結果のラティスと字幕テキストが最大に一致する区間を基準とする。

そのように枝刈り後のラティスが得られた後の処理は、第３の実施形態における処理と同様である。つまり、音響モデル適用化部１５は、選択処理部１３が出力した学習データを読み込み、フォワードバックワードアルゴリズムを用いて学習データに含まれる各々の音素の観測確率を求め、音素ごとに、観測確率で重み付けられた音響特徴量データとして得られる音響モデルを用いて、前記音響モデル記憶部を更新する。

【０１０４】

次に、第１～第５のいずれかの実施形態を用いたさらなるバリエーションとして、複数の追加実施形態を説明する。要約すると、追加実施形態の１は、各々の番組用に、学習データを抽出する。また、追加実施形態の２は、言語モデルの重みを適宜変える。また、追加実施形態の３は、話者の交代を検出し、話者ごと、又は話者属性ごとの音響モデルを抽出する。

【０１０５】

［追加実施形態の１］

本発明の追加実施形態の１について説明する。

図１１は、同実施形態による音響モデル学習装置１０１の機能構成を示すブロック図である。図示するように、音響モデル学習装置１０１は、内部に、音響モデル学習装置（番組Ａ用）１０１Ａと音響モデル学習装置（番組Ｂ用）１０１Ｂと音響モデル学習装置（番組Ｃ用）１０１Ｃと音響モデル学習装置（番組Ｄ用）１０１Ｄとを含んで構成されている。これら音響モデル学習装置１０１Ａ～１０１Ｄの各々は、それぞれ単一の番組用のものである。

【０１０６】

ここで、番組とは、典型的にはテレビやラジオの放送番組であるが、公衆によって直接受信されることを目的とする無線通信の送信であるところの「放送」の番組に限らず、有線通信によるテレビ放送やラジオ放送の番組、インターネットを介してパケットデータとして配信される動画（音声を含む）による放送や音声放送の番組、通信衛星から多数の受信者に向けて送信される番組、視聴者或いは聴取者のリクエストに応じてオン・デマンド的に配信される番組など、音声を含むコンテンツを含んでいる。

なお、音響モデル学習装置１０１が内部に備える番組個別用の音響モデル学習装置の数は、４に限定されず、任意の自然数として構成しても良い。

【０１０７】

音響モデル学習装置１０１Ａ～１０１Ｄの各々は、図示するように、第１の実施形態で説明した音響モデル学習装置１００と同様の構成・機能を有し、同様の動作をする。

本実施形態においても、選択処理部１３は、アライメント処理部１２によってアライメントされたデータを用いて、認識結果データに信頼度を付加し、学習データとして出力する。

【０１０８】
本実施形態のポイントは、音響モデル学習装置１０１Ａ～１０１Ｄの各々が専用の番組音声記憶部８と音響モデル記憶部９を備え、個々の音響モデル記憶部９に各々専用の音響モデルを記憶し、この各々専用の音響モデルを用いて認識処理部１０が認識処理を行い、その結果得られる学習データが各々専用の学習データ記憶部１４に書き込まれ、この学習データが蓄積されて出力されるとともに、この得られた学習データを用いて音響モデル適応化部１５が当該番組用の音響モデル記憶部９を適応化する点であり、そのように番組ごとの音響モデルを用いて番組ごとの学習データを得る処理を繰り返す点である。
言い換えれば、音響モデル記憶部９は音響モデルを番組ごとに記憶するものであり、認識処理部１０は番組の音声を当該番組に対応した音響モデルを用いて音声の認識処理を行なうものであり、選択処理部１３は番組ごとに学習データを出力し、音響モデル適応化部１５は、当該番組用の音響モデルを更新する。
【０１０９】
なお、音響モデル学習装置１０１Ａ～１０１Ｄは、そのすべての構成要素をそれぞれが専用に備えるようにしても良いし、一部の構成要素を共用にしても良い。例えば、認識処理部１０とアライメント処理部１２と選択処理部１３と音響モデル適応化部１５の処理機能自体は共通のハードウェアを用いて、音響モデル記憶部９に記憶される音響データと認識仮説記憶部１１に記憶される認識仮説と学習データ記憶部１４に記憶される学習データとが、それぞれの番組専用の領域に分けて管理されるように装置を構成しても良い。
【０１１０】
本実施形態の利点は次の通りである。例えば、対談形式のテレビ放送番組の音声を認識させて学習データを得ることを考えた場合、そしてその番組の形式がレギュラー話者（毎回出演する話者）とゲスト話者（特定回のみ出演する話者）の２人による複数回にわたるシリーズ番組を対象とした場合（ここでは放送の各回が番組Ａ～Ｄに相当）、異なるゲスト話者の学習データが重要であるならば、番組回ごとの音響モデルの適応化処理を繰り返して行なったほうが、複数回に共通の音響モデルを適応化処理するよりも、話者適応の効果が得られると言える。この場合には、各回から得られた学習データを集めて最終的な音響モデルの学習データを得る。
【０１１１】
さらに、上記のような対談形式の番組に限らず、放送番組の中には、限られた数人の話者しかいない番組や、特定の話者が継続してレギュラー出演している番組などがある。本実施形態では、音響モデルの適応化を番組ごとに行なうため、他の番組の学習データが混ざらず、その結果として高い認識精度が得られる。
なお、この実施形態は、前述した第１から第５のいずれかの実施形態と組み合わせて実現してよい。言い換えれば、第１から第５のいずれかの実施形態で述べた、選択処理部１３および音響モデル適応化部１５の処理を用いて、番組ごとに音響モデルの学習を行なうようにする。
【０１１２】
［追加実施形態の２］
次に、本発明の追加実施形態の２について説明する。上記の実施形態では、認識処理部１０による認識処理の際の言語モデルと音響モデルそれぞれの重みに影響する重み値W_{lm}として固定的な値を用いていた。本実施形態では、音響モデルの適応度合いに応じて、この重み値W_{lm}を変化させる。
【０１１３】
なお、本実施形態においても、選択処理部１３は、アライメント処理部１２によってアライメントされたデータを用いて、認識結果データに信頼度を付加し、学習データとして出力する。
【０１１４】
図１２は、本実施形態による音響モデル学習装置の処理手順を示すフローチャートである。本フローチャートのステップＳ３１からＳ３９までは、図５に示したＳ０１からＳ０

９までに対応し、同様の処理手順であるのでここでは説明を省略する。そして、本フローチャートのステップＳ３９の処理に続いて、ステップＳ４０では、バイアス化言語モデルの重み付け係数を更新する。一例としては、ステップＳ３４からＳ４０までの繰り返し処理の第ｎ回目（ｎ＝１，２，３、・・・）における重み値（この重み値はｎの関数であるためＷ$_{lm}$（ｎ）と表わす）を、

Ｗ$_{lm}$（ｎ）＝１３－０．５ｎ　・・・　（１０）

とする。つまり、上の式（１０）に従えば、１回目の認識処理に用いる重み値Ｗ$_{lm}$（１）は１２．５であり、２回目の認識処理に用いる重み値Ｗ$_{lm}$（２）は１２．０であり、３回目の認識処理に用いる重み値Ｗ$_{lm}$（３）は１１．５であり、以下同様に、前回の認識処理のときよりも小さい重み値Ｗ$_{lm}$（ｎ）を用いる。これは、認識処理において、前回の認識処理のときよりも言語モデルによる制約の影響をより小さくすることを意味する。

ステップＳ４０の処理が終わると、ステップＳ３４の処理に戻って新たな重み付け係数Ｗ$_{lm}$による言語モデルの重み付けを行い、以下、ステップＳ３５以降の処理に続く。

【０１１５】

ここでは言語モデルに対する重み値Ｗ$_{lm}$を用いて認識処理を行なっているが、これは、言語モデルと音響モデルの重みの比率をＷ$_{lm}$：１の比としていることになる。そしてこれは、言語モデルの重み値を１に固定して音響モデルの重み値Ｗ$_{am}$を用いて（つまり言語モデルと音響モデルの重みの比率を１：Ｗ$_{am}$の比として）も相対的に同じことであり、この場合、本実施形態では認識処理の繰り返しごとに音響モデルの重み値Ｗ$_{am}$が徐々に大きくなるように変化させても、同様に、前回の認識処理のときよりも言語モデルによる制約の影響をより小さくすることを意味する。また、重み値Ｗ$_{lm}$とＷ$_{am}$とを用いて言語モデルと音響モデルの重みの比率をＷ$_{lm}$：Ｗ$_{am}$の比としても同様であり、本実施形態では、認識処理において前回の認識処理のときよりも言語モデルによる制約の影響がより小さくなるようにする。

【０１１６】

また、本実施例の変形例として、重み値Ｗ$_{lm}$を固定したままで、当該番組音声に対応する字幕テキストのための重み値Ｗ$_{bias}$を認識処理の繰り返しに伴って徐々に小さくしていっても、今回認識処理においては前回の認識処理のときよりも言語モデルが相対的に音響モデルよりもより小さい重みで用いられるという同様の作用が得られる。

さらにまた、本実施例の変形例として、重み値Ｗ$_{lm}$と重み値Ｗ$_{bias}$の両方を認識処理の繰り返しに伴って徐々に小さくしていっても、今回認識処理においては前回の認識処理のときよりも言語モデルが相対的に音響モデルよりもより小さい重みで用いられるという同様の作用が得られる。

【０１１７】

言い換えれば、本実施形態あるいはその変形例においては、認識処理部１０は、バイアス化言語モデル記憶部６から読み出した言語モデルと音響モデル記憶部９から読み出した前記音響モデルとを所定の重み比率値で重み付けして用いて、認識処理を行なうものである。また、認識処理部１０は、音響モデル適応化部１５によって更新された音響モデルを用いて再度認識処理を行なうものであり、このとき使用する前記重み比率値は、前回の認識処理のときよりも言語モデルがより小さい重みで用いられる値としている。

【０１１８】

これにより、音響モデルの適応度合いに応じて言語的な制約を減じることとなり、これによってより一層、音響モデルの学習の精度が向上する。

さらに詳細に述べると、前記の各実施形態では、認識精度を高めるため、従来技術による方法よりも強い言語的な制約を用いた認識処理を行なうようにしている。つまり、言語モデルの重み付けの度合いを比較的大きくしている。しかしながら、音響モデルを適応化する処理を繰り返すことにより、比較的小さな言語的な制約で認識精度を確保できるようになる。従って、本実施形態では、繰り返し処理による音響モデルの精度向上に応じて言語的制約を減ずるようにしている。この言語的制約の調整により、字幕と発話の不一致区

間の検出制度の向上が可能となる。

なお、この実施形態は、前述した第１から第５のいずれかの実施形態と組み合わせて実現してよい。言い換えれば、第１から第５のいずれかの実施形態で述べた、選択処理部１３および音響モデル適応化部１５の処理を用いながら、重み値W_{lm}を徐々に変化させる。

【０１１９】

［追加実施形態の３］

次に、本発明の追加実施形態の３について説明する。本実施形態では、話者の交代あるいは話者の属性の交代を検出し、話者に依存した音響モデル或いは話者の属性に依存した音響モデルを用いて認識処理を行なう点が特徴的である。

【０１２０】

図１３は、本実施形態による音響モデル学習装置の機能構成を示すブロック図である。図示する構成において、音響モデル学習装置１０２が有する字幕テキスト記憶部１と形態素解析処理部２と重み付け処理部３とベースラインコーパス処理部４と合成処理部５とバイアス化言語モデル記憶部６と重み付け処理部７と番組音声記憶部８は、前述した実施形態におけるそれらと同様のものである。さらに、本実施形態の音響モデル学習装置１０２は、その特有の構成として、音響モデル記憶部９Ｆおよび９Ｍと、認識処理部１０ＭＦと、認識仮説記憶部１１Ｆおよび１１Ｍと、アライメント処理部１２Ｆおよび１２Ｍと、選択処理部１３Ｆおよび１３Ｍと、学習データ記憶部１４Ｆおよび１４Ｍと、音響モデル適応化部１５Ｆおよび１５Ｍと、女声発話モデル記憶部１６Ｆ（発話モデル記憶部、話者属性別発話モデル記憶手段）と、男声発話モデル記憶部１６Ｍ（発話モデル記憶部、話者属性別発話モデル記憶手段）と、発話モデル適応化部１７Ｆおよび１７Ｍと、音声区間切り出し部１８（話者属性推定手段）とを含んで構成される。

【０１２１】

女声発話モデル記憶部１６Ｆは、女性の声の統計的な音響特徴量を含む女声発話モデルを記憶している。また、男声発話モデル記憶部１６Ｍは、男性の声の統計的な音響特徴量を含む男声発話モデルを記憶している。つまり、女声発話モデル記憶部１６Ｆと男声発話モデル記憶部１６Ｍは、話者ごと又は話者属性ごとの統計量を表わす話者属性別発話モデルを記憶するための話者属性別発話モデル記憶手段としても機能する。なお、女声および男声の統計的音響特徴量としてはＭＦＣＣを用いている。この統計的音響特徴量としては、他にも、ＰＬＰやＨＭＭ（隠れマルコフモデル）やホルマント周波数の分布などを用いることができる。

音声区間切り出し部１８は、女声発話モデル記憶部１６Ｆから読み出した女声発話モデルと男声発話モデル記憶部１６Ｍから読み出した男声発話モデルとを用いて、番組音声の中の、女声で発話されている部分と、男声で発話されている部分と、発話以外の部分（番組中の、例えば無音部分や、背景雑音のみの部分や、音楽の部分や、効果音の部分など）とを区別する。つまり音声区間切り出し部１８は、入力される音声がヒトによる発話であるか否かを検出するとともに、発話である場合にはその話者属性（ここでは、話者の性別）を検知する。つまり、音声区間切り出し部１８は、話者属性別発話モデルを用いて音声の中の所定音声区間における話者又は話者属性を推定する話者属性推定手段としても機能する。そして、音声区間切り出し部１８は、女声による音声区間と男声による音声区間とをそれぞれ切り出した形で認識処理部１０ＭＦに渡す。

【０１２２】

また、音響モデル学習装置１０２は、女声用と男声用の音響モデルを区別して管理し、記憶している。具体的には、音響モデル記憶部９Ｆには女声用の音響モデルが記憶され、音響モデル記憶部９Ｍには男声用の音響モデルが記憶されている。つまり、音響モデル記憶部９Ｆと９Ｍは、話者ごとまたは話者属性ごとに区別して音響モデルを記憶している。

【０１２３】

そして、認識処理部１０ＭＦは、音声区間切り出し部１８から渡された女声音声区間については音響モデル記憶部９Ｆから読み出した女声用音響モデルを用いて、また音声区間

切り出し部１８から渡された男声音声区間については音響モデル記憶部９Ｍから読み出した男声用音響モデルを用いて、それぞれ認識処理を行なう。つまり、認識処理部１０ＭＦは、音声区間切り出し部１８によって切り出された音声区間の音声を認識処理する。またつまり、認識処理部１０ＭＦは、話者属性推定手段による推定結果に応じて、該当する話者用または話者属性用の音響モデルを用いて前記音声の認識処理を行なう。そして、認識処理部１０ＭＦは、女声音声区間から得られた認識結果仮説を認識仮説記憶部１１Ｆに書き込み、男声音声区間から得られた認識結果仮説を認識仮説記憶部１１Ｍに書き込む。

【０１２４】

なお、男女別の音声の統計量（Male speech model, Female speech model）を用いて音声区間の検出と男女の話者交代を検出し、性別に依存する音響モデルを用いて認識を行なうには、Toru IMAI, Shoei SATO, Shinichi HOMMA, Kazuo ONOE, Akio KOBAYASHI「Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News」（IEICE Transactions on Information and Systems 2007 E90-D(8):1286-1291）に記載された方法を利用可能である。

【０１２５】

アライメント処理部１２Ｆは、認識仮説記憶部１１Ｆから読み出した女声音声区間の認識仮説のみを用いて、前述した実施形態と同様のアライメント処理を行なう。選択処理部１３Ｆは、アライメント処理部１２Ｆによる処理結果に基づき、女声音声区間の認識仮説に、前述の手法を用いて信頼度データを付加し、得られた学習データを学習データ記憶部１４Ｆに書き込む。

これと同様に、アライメント処理部１２Ｍは、認識仮説記憶部１１Ｍから読み出した男声音声区間の認識仮説のみを用いて、前述した実施形態と同様のアライメント処理を行なう。選択処理部１３Ｍは、アライメント処理部１２Ｍによる処理結果に基づき、男声音声区間の認識仮説に、前述の手法を用いて信頼度データを付加し、得られた学習データを学習データ記憶部１４Ｍに書き込む。

このように、音声区間切り出し部１８によって切り出された女声音声区間および男声音声区間それぞれに基づいて、女性用および男性用のそれぞれ専用の学習データが得られる。

【０１２６】

音響モデル適応化部１５Ｆは、学習データ記憶部１４Ｆから読み出した女声用学習データのみを用いて、音響モデル記憶部９Ｆに記憶されている女声用音響モデルを適応化（更新）する。また、音響モデル適応化部１５Ｍは、学習データ記憶部１４Ｍから読み出した男声用学習データのみを用いて、音響モデル記憶部９Ｍに記憶されている男声用音響モデルを適応化（更新）する。つまり、音響モデル適応化部１５Ｆと１５Ｍは、話者属性推定手段による推定結果に対応する学習データを用いて、当該推定結果に対応する話者ごとまたは話者属性ごとの音響モデルを更新するものである。

【０１２７】

発話モデル適応化部１７Ｆは、学習データ記憶部１４Ｆから読み出した女声用学習データのみを用いて、女声発話モデル記憶部１６Ｆに記憶されている女声発話モデルを適応化（更新）する。また、発話モデル適応化部１７Ｍは、学習データ記憶部１４Ｍから読み出した男声用学習データのみを用いて、男声発話モデル記憶部１６Ｆに記憶されている男声発話モデルを適応化（更新）する。つまり、発話モデル適応化部１７Ｆと１７Ｍは、話者属性推定手段による推定結果に対応する学習データを用いて、話者属性別発話モデル記憶手段に記憶された、当該推定結果に対応する話者ごとまたは話者属性ごとの話者属性別発話モデルを更新するものである。

【０１２８】

また、音響モデル学習装置１０２は、上記の一連の認識処理、アライメント処理、選択処理、そして音響モデル適応化処理と発話モデル適応化処理を、繰り返し行なう。

【０１２９】

上記のような構成および作用により、入力音声の統計的音響特徴量（女声発話モデルお

よび男声発話モデル）を利用して高精度な音声区間の切り出しを行なう方法を併用する場合において、得られた学習データを用いて切り出しに使用する上記の統計量も適応化することを繰り返すことができる。これにより、例えば雑音区間を発話区間の音声として認識してしまうような音声区間の切り出し誤りが減少し、学習データの質のさらなる向上が可能となる。

また、話者属性（性別）に依存した音響モデルを使用して認識処理を行なうことにより、話者属性非依存の音響モデルを用いた場合に比べ、高い認識精度を得ることができ、より多くの学習データを作成することができる。

また、入力音声の統計量を利用して話者或いは話者属性の推定手段（音声区間切り出し部１８）を備えた音声アルゴリズムを併用する場合に、得られた学習データを用いて話者推定用の統計量（女声発話モデルおよび男声発話モデル）を適応化するとともに、話者（或いは話者属性）ごとに音響モデルを設けて（音響モデル記憶部９Ｆおよび９Ｍ）、これらをそれぞれ適応化しているため、さらに、認識精度の向上が図れる。

【０１３０】

なお、この実施形態は、前述した第１から第５のいずれかの実施形態と組み合わせて実現してよい。言い換えれば、選択処理部１３Ｆおよび１３Ｍはそれぞれ、第１から第５のいずれかの実施形態で述べた選択処理部１３の処理を用いる。また、音響モデル適応化部１５Ｆおよび１５Ｍはそれぞれ、第１から第５のいずれかの実施形態で述べた音響モデル適応化部１５の処理を用いる。このようにして、話者ごと、あるいは、話者属性ごとの音響モデルの学習を行なう。

【０１３１】

なお、本実施形態の更なる変形例として、追加実施形態の１で説明した番組ごとの音響モデルを管理する構成（このとき、適宜、発話モデルも番組ごとに管理するようにしても良い）や、追加実施形態の２で説明した処理の繰り返しに伴って言語モデルの重み付けを徐々に変化させる構成を併用しても良い。

【０１３２】

また、性別ごとに音声区間切り出しのための発話モデルおよび認識処理のための音響モデルを設ける代わりに、或いは性別に加えて、他の話者属性ごと或いは話者個人ごとに、発話モデルや音響モデルを設けて、その話者属性ごと或いは話者個人ごとの音声区間切り出し処理や認識処理を行なうようにしても良い。「他の話者属性」とは、例えば、年齢層などである。このとき、話者の年齢層を例えば、少年期（５歳から１４歳）、青年期（１５歳から２４歳）、壮年期（２５歳から４４歳）、中年期（４５歳から６４歳）、高年期（６５歳以上）などに分類する。

【０１３３】

また、発話環境ごとに、発話モデルや音響モデルを設けるようにしても良い。ここで「発話環境ごと」とは、例えば、話者が原稿を読み上げている形式の場合、対談あるいは座談形式の場合、雑談の場合などである。

【０１３４】

また、本実施形態では、発話モデルを利用して音声区間を切り出す構成と、話者属性ごとに音響モデルを設けるとともに話者属性ごとに音声区間を切り出してそれぞれ専用の音響モデルを用いて認識を行い、音響モデルを適応化する構成との両方を用いているが、これらのいずれか一方のみの構成を含むようにしても良い。

【０１３５】

また、本実施形態では、認識仮説を、論理的には言語的単位（形態素）をアークとするラティス構造のグラフで表現して処理を行なったが、その代わりに、言語的単位として音素をアークとするラティス構造のグラフで表現して、同様の処理を行なうようにしても良い。

【０１３６】

＜コンピュータシステムを用いた実施形態＞

なお、上述した各実施形態における音響モデル学習装置の機能をコンピュータで実現す

るようにしても良い。その場合、この音響モデル学習の機能を実現するためのプログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することによって実現しても良い。なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものとする。また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時刻の間、動的にプログラムを保持するもの、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時刻プログラムを保持しているものも含んでも良い。また上記プログラムは、前述した機能の一部を実現するためのものであっても良く、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるものであっても良い。

【０１３７】

以上、この発明の実施形態について図面を参照して詳述してきたが、具体的な構成はこの実施形態に限られるものではなく、この発明の要旨を逸脱しない範囲の設計等も含まれる。

例えば、一致区間を選択する際に、選択処理部１３は字幕テキストと認識仮説が３単語以上連続して一致している区間を選択することを説明したが、「３」に限らず、適宜所定の単語数以上連続して位置している区間を選択するようにしても良い。また、語数を基準とする代わりに、所定の文字数（例えば、５文字）以上連続して一致している区間を選択するようにしても良い。

【０１３８】

＜認識実験の結果＞

ここで、本願発明者らが実際に行なった認識実験について説明する。

２００４年１０月１４日、および２００７年６月２５日、２７日、２８日（本放送と再放送を含む）にＮＨＫで放送された５回分の番組「きょうの健康」を評価音声として認識実験を行なった。各放送回には約２６００単語（形態素）の発話があり総計１２８０７単語を評価した。

【０１３９】

音響モデルの学習用には、２００４年から２００７年に放送されたＮＨＫ「きょうの健康」から、評価対象と同一のコンテンツを除く９３回分の音声と字幕を用いて、放送回ごとに下記の４手法（ＬＳ０～ＬＳ３）を適用して約２０時間分の学習データを生成し、性別依存不特定話者音響モデルをＭＬＬＲとＭＡＰで適応化して認識実験を行なった。

ＬＳ０：これは従来技術による手法であり、認識結果（最尤仮説）と書き起こしのテキストデータとの一致部分の音声を切り出して、学習音声とした。

ＬＳ１：信頼度を全て１として、一致区間情報を利用しない教師なし学習とした。

ＬＳ２：最尤仮説と書き起こしのテキストデータとの一致部分の信頼度を１とし、それ以外の部分の信頼度を０として、音声を切り出さずに一区間のみを用いた学習とした。

ＬＳ３：最尤仮説と書き起こしのテキストデータとの一致部分の信頼度を１とし、最尤仮説の単語の事後確率から不一致部分の信頼度を得て、学習した。

【０１４０】

このときの言語的制約の強さは、W_{bias}＝１０００、W_{lm}＝１６を用いた。こうして作成されたこの適応化言語モデルでのテストセットパープレキシティーは２２程度であった。また、上記の２０時間分の学習データのうち、認識結果と字幕の一致部分、つまり従来のＬＳ法で得られる学習音声の量は５～８時間程度であった。

言語モデルは、情報番組の書き起こしの字幕テキストデータを基に、番組ホームページから得られる各放送回の番組内容紹介テキストに２０倍の重みを付けて学習した番組依存言語モデルである。この言語モデルのテストセットパープレキシティーは５９であり、未

知語率は０．８％であった。

【０１４１】

認識実験および学習データの生成には、男女別のモノフォンＨＭＭを用いて音声区間を検出し、男女の自動判別を行ないながら性別依存トライフォンＨＭＭの両方を適応化する。したがって、適応化モデルにより発話の切り出し精度と男女の判別精度の向上も期待される。ベースラインの音響モデルは、３４０時間（男性）と２５０時間（女性）のニュース音声から学習され、モノフォンＨＭＭ（１２６状態３２混合分布）、状態共有トライフォンＨＭＭ（４０００状態、１６混合分布）が男女別に作成された。

【０１４２】

実験結果は次の通りである。最尤仮説を教師単語列とした場合のＬＳ０～ＬＳ３でのＷＥＲを考察する。まず、最尤仮説と書き起こしの字幕テキストの一致部分のみを学習するＬＳ０とＬＳ２を比較すると、ＬＳ２では、信頼度を用いることにより認識と同一の発話区間検出を利用できたため、より大きな改善が得られた。次に、ＬＳ０以外の３手法を比較すると、収束の速さに若干の差があるものの、最終的にはほぼ同等の改善が得られた。本実験条件では、教師単語列作成時の言語的制約（W_{bias}，W_{lm}）が事前に最適化されていたため、書きお越しの一致の有無にかかわらず、全ての音声を学習（ＬＳ１）しても改善が得られた。一方、ラティスと字幕テキストの一致が最大になる単語列を教師とした場合、提案法ＬＳ３以外のＬＳ１とＬＳ２のＷＥＲの増加が見られた。本実験条件では、仮説の信頼度に関わらず教師単語列を選択したため、学習データ生成時の認識誤りを教師とすることが多くなり、評価音声の認識精度が低下したと思われる。これに対して提案法ＬＳ３では、教師単語列の制度が低い場合でも頑健な学習結果が得られ、ＷＥＲの増加は小さかった。

【０１４３】

本発明は、不完全な書き起こしの字幕データから効率的に音響モデルを学習するため、教師単語系列に信頼度を与え、さらに認識ラティスの事後確率を信頼度とする方法を用いた。字幕テキストを用いた音響モデルの学習に本発明の方法を適用した結果、信頼度の導入により、従来法よりも低いＷＥＲが得られることが認識実験を通しても実証された。さらに事後確率を用いることで、頑健な学習が可能であることを示すことができた。

【図面の簡単な説明】

【０１４４】

【図１】本発明の第１の実施形態による音響モデル学習装置の機能構成を示したブロック図である。

【図２】同実施形態で用いる言語モデルのデータ構成を示した概略図である。

【図３】同実施形態で用いる音響モデルのデータ構成を示した概略図である。

【図４】同実施形態で用いる認識仮説記憶部のデータ構成およびデータ例、並びにそのデータが表わすラティス構造を示した概略図である。

【図５】同実施形態による音響モデル学習の処理手順を示したフローチャートである。

【図６】同実施形態により最尤仮説のパスと書き起こしによる字幕テキストの一致を利用して信頼度を付加する例を表わす概略図である。

【図７】同実施形態の音響モデル学習装置による音響モデルの学習の概略を説明するため概略図である。

【図８】第２の実施形態において枝刈りされた後の認識仮説のデータ例、並びにそのデータが表わすラティス構造を示した概略図である。

【図９】第３の実施形態において対立する複数のパス間で共通して出現する音素の例を説明するための概略図である。

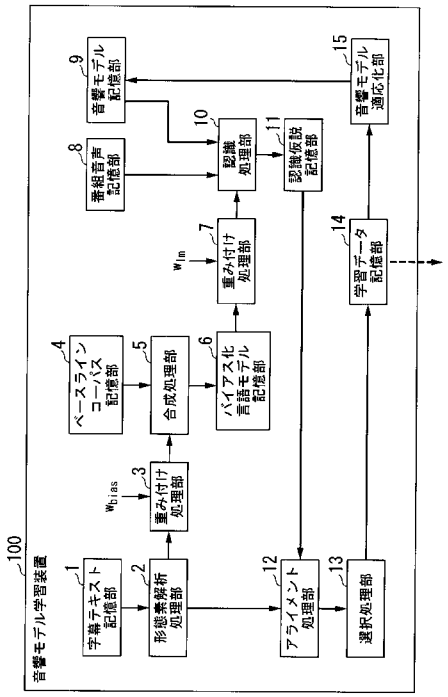
【図１０】第４の実施形態において、選択処理部１３が認識結果のラティスと字幕テキストの一致区間を抽出する処理を概念的に説明する概略図である。

【図１１】本発明の追加実施形態の１による音響モデル学習装置の機能構成を示したブロック図である。

【図１２】本発明の追加実施形態の２による音響モデル学習装置の処理手順を示したフロ

ーチャートである。
【図１３】本発明の追加実施形態の３による音響モデル学習装置の機能構成を示したブロック図である。
【図１４】従来技術による音響モデル学習装置の機能構成を示したブロック図である。
【符号の説明】
【０１４５】
１　字幕テキスト（Caption text）記憶部
２　形態素解析（Morphological analysis）処理部
３　重み付け処理部
４　ベースラインコーパス（Baseline corpus）記憶部
５　合成処理部
６　バイアス化言語モデル（Biased LM）記憶部（言語モデル記憶部）
７　重み付け処理部
８　番組音声（Program audio）記憶部
９，９Ｆ，９Ｍ　音響モデル（AM）記憶部
１０，１０ＭＦ　認識（Recognition）処理部
１１，１１Ｆ，１１Ｍ　認識仮説記憶部
１２，１２Ｆ，１２Ｍ　アライメント（Alignment）処理部
１３，１３Ｆ，１３Ｍ　選択（Selection）処理部
１４，１４Ｆ，１４Ｍ　学習データ（Transcripts）記憶部
１５，１５Ｆ，１５Ｍ　音響モデル適応化部（Adaptation）
１６Ｆ　女声発話モデル（Female speech model）記憶部（話者属性別発話モデル記憶手段）
１６Ｍ　男声発話モデル（Male speech model）記憶部（話者属性別発話モデル記憶手段）
１７Ｆ，１７Ｍ　発話モデル適応化部（Adaptation）
１８　音声区間切り出し部（話者属性推定手段）
１００，１０１，１０１Ａ～１０１Ｄ，１０２　音響モデル学習装置

【図１】



【図２】

(a)

形態素	出現確率
…	…

(b)

形態素1	形態素2	接続確率
…	…	…

【図３】

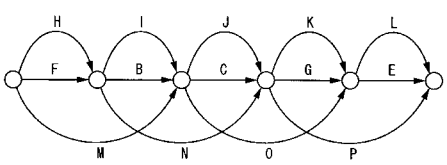
音素	特徴量
…	…

【図４】

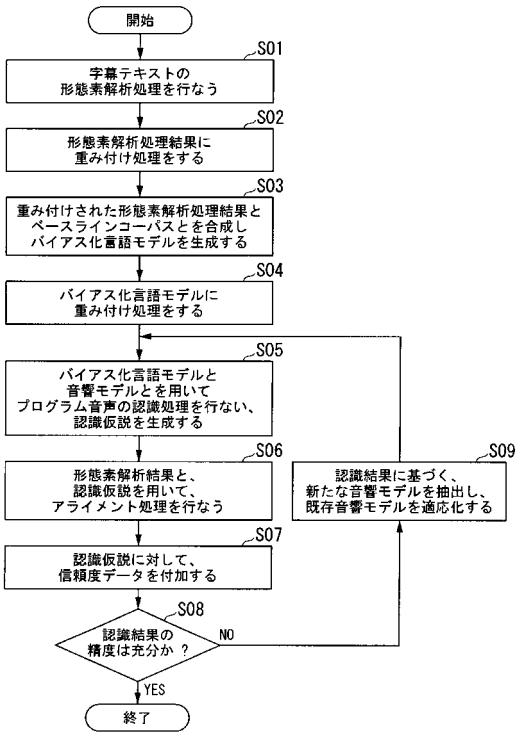
(a)

始点ノード	終点ノード	言語的単位（形態素）	信頼度
1	2	F	
1	2	H	
1	3	M	
2	3	B	
2	3	I	
2	4	N	
3	4	J	
3	4	C	
3	5	O	
4	5	K	
4	5	G	
4	6	P	
5	6	E	
5	6	L	
…	…	…	…

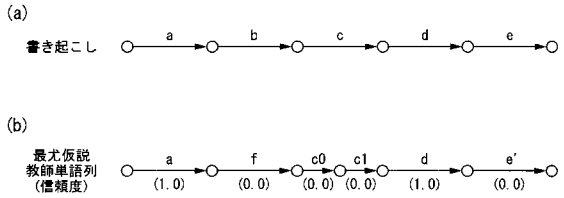
(b)



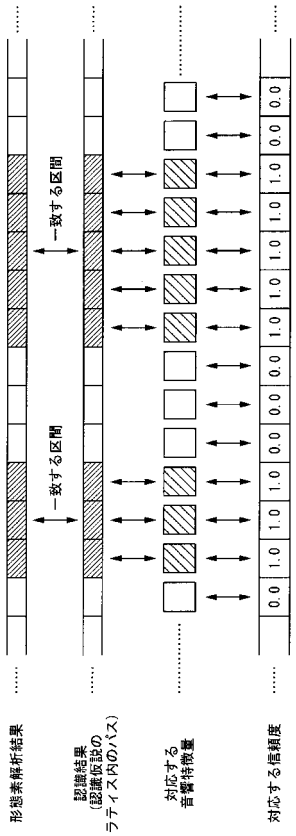
【図５】



【図６】



【図７】

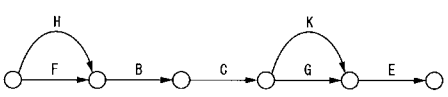


【図８】

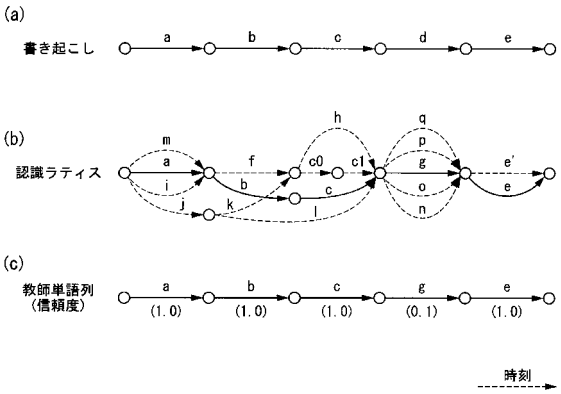
(a)

始点ノード	終点ノード	言語的単位(形態素)	信頼度
1	2	F	0.1
1	2	H	0.9
2	3	B	1
3	4	C	1
4	5	K	0.2
4	5	G	0.8
5	6	E	1
⋮	⋮	⋮	⋮

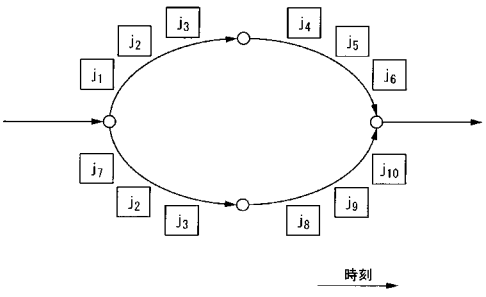
(b)



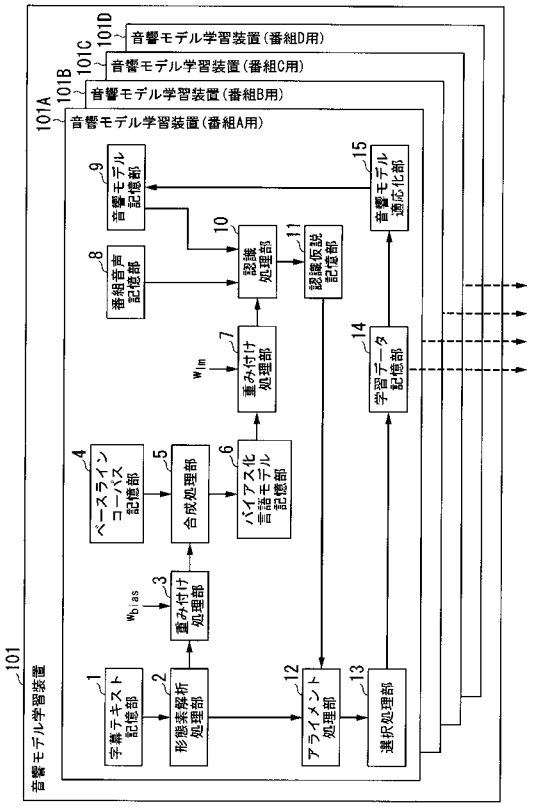
【図１０】



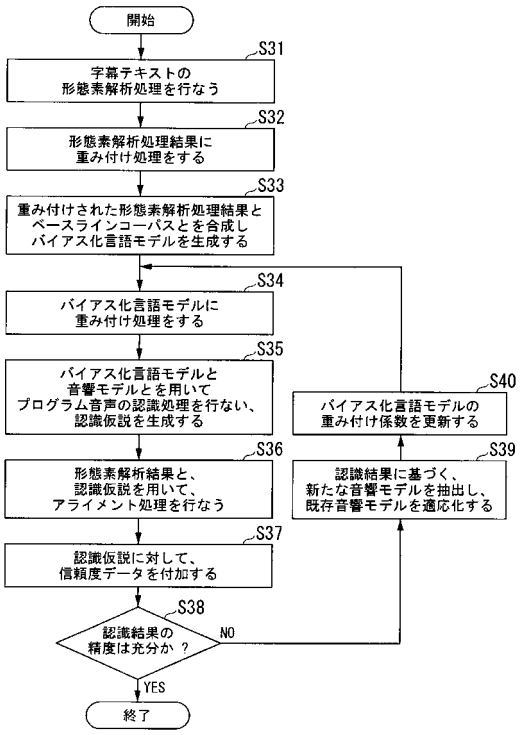
【図９】



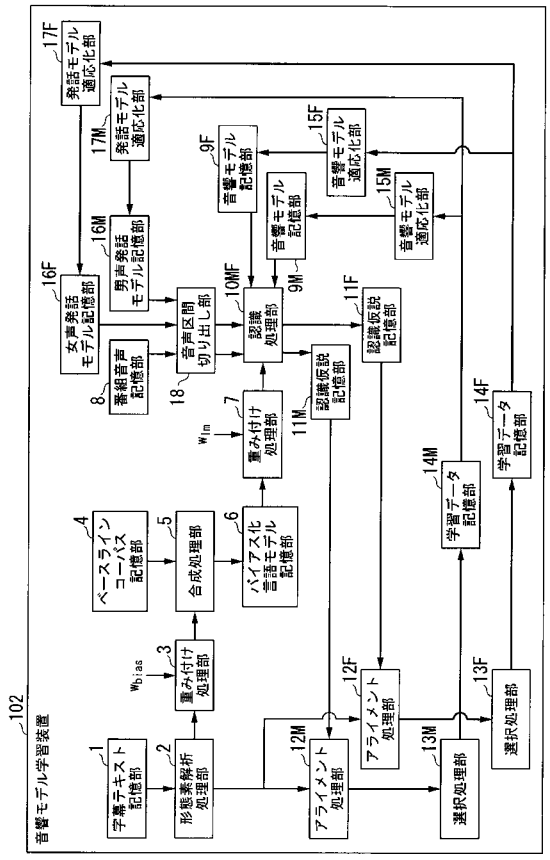
【図１１】



【図１２】



【図１３】



【図１４】