(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6903613号
(P6903613)
(45)発行日　令和3年7月14日(2021.7.14)　(24)登録日　令和3年6月25日(2021.6.25)

(51)Int.Cl.　F I
G10L 15/06 (2013.01)　G10L 15/06 300Y

請求項の数 13　(全 14 頁)

(21)出願番号　特願2018-168708(P2018-168708)
(22)出願日　平成30年9月10日(2018.9.10)
(65)公開番号　特開2020-42130(P2020-42130A)
(43)公開日　令和2年3月19日(2020.3.19)
審査請求日　令和2年7月1日(2020.7.1)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(72)発明者　丁　寧
東京都港区芝浦一丁目１番１号　株式会社東芝内

審査官　中村　天真

最終頁に続く

(54)【発明の名称】音声認識装置、音声認識方法及びプログラム

(57)【特許請求の範囲】
【請求項１】
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する生成部と、
前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付与する重みを前記個数に応じて決定する決定部と、
前記重みに基づいて前記音声データを選択する選択部と、
前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する適応部と、
を備える音声認識装置。
【請求項２】
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する生成部と、
前記ラベルを用いて、前記音声データに含まれる音声フレームと、前記音声データに含まれる非音声フレームとの比率を計算する計算部と、
前記音声フレームの比率が所定の選択範囲以内である音声データを選択する選択部と、
前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する適応部と、
を備える音声認識装置。
【請求項３】

言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する生成部と、
前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付与する重みを前記個数に応じて決定する決定部と、
前記ラベルを用いて、前記音声データに含まれる音声フレームと、前記音声データに含まれる非音声フレームとの比率を計算する計算部と、
前記重みと、前記音声フレームの比率とに基づいて、前記音声データを選択する選択部と、
前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する適応部と、
を備える音声認識装置。
【請求項４】
前記決定部は、前記個数が多いほど、前記重みを小さく決定する、
請求項１に記載の音声認識装置。
【請求項５】
前記選択部は、前記重みが閾値より大きいか否かを判定し、前記閾値よりも大きい重みが付与された音声データを選択する、
請求項１に記載の音声認識装置。
【請求項６】
前記言語モデルと前記第２音響モデルとを用いて、前記音声データの音声認識を行う認識部、
を更に備える請求項１に記載の音声認識装置。
【請求項７】
前記言語モデル及び前記第１音響モデルは、音声の言語的な特徴、及び、音声の音響的な特徴の両方を区別せずにモデル化された１つの音声認識ベースモデルにより表され、
前記適応部は、前記選択部により選択された音声データを用いて、前記音声認識ベースモデルを適応させる、
請求項１に記載の音声認識装置。
【請求項８】
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成するステップと、
前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付与する重みを前記個数に応じて決定するステップと、
前記重みに基づいて前記音声データを選択するステップと、
前記選択するステップにより選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成するステップと、
を含む音声認識方法。
【請求項９】
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成するステップと、
前記ラベルを用いて、前記音声データに含まれる音声フレームと、前記音声データに含まれる非音声フレームとの比率を計算するステップと、
前記音声フレームの比率が所定の選択範囲以内である音声データを選択する選択部と、
前記選択するステップにより選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成するステップと、
を含む音声認識方法。
【請求項１０】
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成するステップと、
前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付

与する重みを前記個数に応じて決定するステップと、
前記ラベルを用いて、前記音声データに含まれる音声フレームと、前記音声データに含まれる非音声フレームとの比率を計算するステップと、
前記重みと、前記音声フレームの比率とに基づいて、前記音声データを選択するステップと、
前記選択するステップにより選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成するステップと、
を含む音声認識方法。
【請求項１１】
コンピュータを、
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する生成部と、
前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付与する重みを前記個数に応じて決定する決定部と、
前記重みに基づいて前記音声データを選択する選択部と、
前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する適応部、
として機能させるためのプログラム。
【請求項１２】
コンピュータを、
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する生成部と、
前記ラベルを用いて、前記音声データに含まれる音声フレームと、前記音声データに含まれる非音声フレームとの比率を計算する計算部と、
前記音声フレームの比率が所定の選択範囲以内である音声データを選択する選択部と、
前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する適応部、
として機能させるためのプログラム。
【請求項１３】
コンピュータを、
言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する生成部と、
前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付与する重みを前記個数に応じて決定する決定部と、
前記ラベルを用いて、前記音声データに含まれる音声フレームと、前記音声データに含まれる非音声フレームとの比率を計算する計算部と、
前記重みと、前記音声フレームの比率とに基づいて、前記音声データを選択する選択部と、
前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する適応部、
として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は音声認識装置、音声認識方法及びプログラムに関する。
【背景技術】
【０００２】
音響モデルと言語モデルとを用いて音声データを認識し、音声データに含まれる発話のテキストを出力する音声認識技術が従来から知られている。音響モデルは予め大量（例えば数百時間以上）のデータを用いて学習される。しかし、どのような条件で用いても高い

認識率（例えば８５％以上）が得られるような音響モデルを学習することは困難である。例えば、クリーンな環境で収録された音声データを用いて学習された音響モデルが用いられた場合、残響が大きい会議室での認識率が劣化してしまう。認識率の劣化を防ぐ有効な方法の一つとして、音響モデルの適応がある。

【先行技術文献】

【特許文献】

【０００３】

【特許文献１】特許第５８５２５５０号公報

【発明の概要】

【発明が解決しようとする課題】

【０００４】

しかしながら、従来の技術では、音響モデルの適応を行った場合、悪影響も生じていた。例えば、同じ内容の発話が繰り返された場合、音響モデルの適応によって、この発話を認識しやすくなるが、他の発話を認識しにくくなる。また例えば、音声データには音声及び非音声の両方が含まれているが、非音声の部分が多い場合、音響モデルの適応によって、非音声の認識結果が出やすくなり、音声の認識結果が出にくくなる。本発明が解決しようとする課題は、音響モデルの適応による悪影響を抑制できる音声認識装置、音声認識方法及びプログラムを提供することである。

【課題を解決するための手段】

【０００５】

実施形態の音声認識装置は、生成部と決定部と選択部と適応部とを備える。生成部は、言語モデルと第１音響モデルとを用いて音声データを認識し、前記音声データに含まれる発話を識別するラベルを生成する。決定部は、前記ラベルを用いて、同じ発話を含む音声データの個数を特定し、前記音声データに付与する重みを前記個数に応じて決定する。選択部は、前記重みに基づいて前記音声データを選択する。適応部は、前記選択部により選択された音声データを用いて、前記第１音響モデルを適応させることにより、第２音響モデルを生成する。

【図面の簡単な説明】

【０００６】

【図１】第１実施形態の音声認識装置の機能構成の例を示すブロック図。

【図２】第１実施形態のラベル情報の例を示す図。

【図３】第１実施形態の音声認識装置の動作方法の例を示すフローチャート。

【図４】第２実施形態の音声認識装置の機能構成の例を示すブロック図。

【図５】第２実施形態の音声データの例を示す図。

【図６】第３実施形態の音声認識装置の機能構成の例を示すブロック図。

【図７】第４実施形態の音声認識装置の機能構成の例を示すブロック図。

【図８】第５実施形態の音声認識装置の機能構成の例を示すブロック図。

【図９】第１乃至第５実施形態の音声認識装置のハードウェア構成の例を示す図。

【発明を実施するための形態】

【０００７】

以下に添付図面を参照して、音声認識装置、音声認識方法及びプログラムの実施形態を詳細に説明する。

【０００８】

はじめに、音響モデルの適応について説明する。音響モデルの適応は、学習された音響モデルに基づき、適応データを用いて当該音響モデルを再学習することにより行われる。以下、はじめに学習された音響モデルをベース音響モデル（第１音響モデル）と呼び、適応させた音響モデルを適応音響モデル（第２音響モデル）と呼ぶ。

【０００９】

音響モデルを適応させる方法は、おおむね二種類ある（教師あり適応及び教師なし適応）。教師あり適応では、音声データ、及び、音声データの正解ラベルの両方を含む適応デ

ータが使用される。教師なし適応では、音声データのみを含む適応データが使用される（正解ラベルがない。）。

【００１０】

教師あり適応は正解ラベルがあるため適応に対して良いが、書き起こしなどによって正解ラベルを作成する必要があるので、コストが高い。

【００１１】

一方、教師なし適応は正解ラベルの作成が要らないためコストが低い。教師なし適応では、音声データを認識し、音声認識結果をラベルとして用いる。音声認識結果の誤りは適応に悪影響を及ぼす可能性があるため、基本的には、音声認識精度は高いほどよい。従来の教師なし適応方法では、言語モデルとベース音響モデルとを用いて音声を認識し、ラベル、信頼度及び音響尤度を出力する。従来の教師なし適応方法では、信頼度がより高く、かつ、音響尤度がより小さい音声データを選択して、音響モデルの適応を行う。

【００１２】

（第１実施形態）

はじめに、第１実施形態の音声認識装置１０の機能構成の例について説明する。

【００１３】

［機能構成の例］

図１は第１実施形態の音声認識装置１０の機能構成の例を示す図である。第１実施形態の音声認識装置１０は、生成部１、決定部２、選択部３及び適応部４を備える。音声認識装置１０の一部又は全ての機能は、ソフトウェア（プログラム）で実現されても良いし、ハードウェアで実現されても良い。

【００１４】

また、第１実施形態の音声認識装置１０は、言語モデル１０１、ベース音響モデル１０２及び適応音響モデル１０３を記憶する。言語モデル１０１は、音声の言語的な特徴をモデル化したデータである。ベース音響モデル１０２及び適応音響モデル１０３は、音声の音響的な特徴をモデル化したデータである。ベース音響モデル１０２は、はじめに学習されたデータである。適応音響モデル１０３は、適応データを用いてベース音響モデル１０２を再学習することにより得られたデータである。なお、言語モデル１０１、ベース音響モデル１０２及び適応音響モデル１０３を記憶する記憶部は、外部の装置に備えられていてもよい。

【００１５】

生成部１は、言語モデル１０１とベース音響モデル１０２とを用いて音声データを認識し、ラベルを生成する。音声データは、例えば発話毎に区切られたデータである。ラベルは、音声データの音声認識結果から変換されたデータである。ラベルは、音声データに含まれる発話を識別する情報である。

【００１６】

決定部２は、ラベルを用いて、同じ発話を含む音声データの個数を特定し、音声データに付与する重みを、当該個数に応じて決定する。

【００１７】

ラベル及び当該ラベルの個数は、例えば図２に示すラベル情報として、音声認識装置１０に記憶される。

【００１８】

図２は第１実施形態のラベル情報の例を示す図である。第１実施形態のラベル情報は、音声データ、音声認識結果、ラベル、カウント数及び重みを含む。

【００１９】

音声認識結果は、音声データの認識結果である。図２の例では、ラベルは、音声認識結果をひらがなに変換したデータである。なお、ラベルは、ひらがなに限らずローマ字等でもよい。

【００２０】

カウント数は、ラベルの個数を示す。例えば、発話－１、発話－３及び発話－５のラベ

ルは同じである。発話－１のラベル生成時には、当該ラベルのカウント数は１となる。発話－３のラベル生成時には、当該ラベルのカウント数は２となる。発話－５のラベル生成時には、当該ラベルのカウント数は３となる。

【００２１】

重みは、ラベルの重みを示す。図２の例では、ラベルのカウント数が大きいほど、当該ラベルの重みは小さくなる。

【００２２】

生成部１は、例えば下記式（１）により、ラベルの重みを決定する。

【００２３】

$$\mu = e^{1-x} \cdots (1)$$

【００２４】

ここで、μは重みであり、ｘはカウント数である。図２の例では、式（１）により重みが決定されている。例えば発話－１、発話－３及び発話－５のラベルの重みは、それぞれ１．００、０．３７、０．１４である。発話－２、発話－４及び発話－６のラベルの重みは、１．００である。

【００２５】

なお、重みを決定する式は、上述の式（１）に限られず、他の減少関数でも良い。

【００２６】

図１に戻り、選択部３は、生成部１により生成されたラベル情報に含まれる重みに基づいて、適応データとして使用する音声データ（発話）を選択する。適応データの中で同じ内容の発話が複数存在する場合、適応によって、同じ内容の発話の事後確率が高くなって、当該発話の認識がしやすくなる。一方、この場合、他の発話の事後確率が低くなるため、他の発話を認識しにくくなる。

【００２７】

したがって、選択部３は、各発話の重みと重み閾値とを比較し、重み閾値より大きい発話を適応データとして選択する。これにより、適応データを使用して生成された適応音響モデル１０３を使用して、音声認識をする場合の悪影響を抑制することができる。

【００２８】

重み閾値は、例えば下記式（２）により決定される。

【００２９】

$$\theta = e^{1-\alpha n} \cdots (2)$$

【００３０】

ここで、θは重み閾値であり、αは発話係数であり、ｎは全発話数である。つまり、同じ内容の発話については、カウント数ｘが全発話数ｎのα倍より小さい場合（$x < \alpha n$）、適応データとして選択される。

【００３１】

発話係数αは、例えば０．２である。図２の例では、全発話数ｎは６であるため、重み閾値θは０．８２になる。発話－１、発話－２、発話－３及び発話－５の重みは、重み閾値θより大きいため、選択部３により適応データとして選択される。一方、発話－４及び発話－６の重みは、重み閾値θより小さいため、選択部３により適応データとして選択されない。

【００３２】

なお、第１実施形態の説明では、発話係数が０．２の場合について説明したが、必要に応じて、発話係数を１以下の他の数値を設定しても良い。また、全発話数ｎの比率αｎではなく、絶対発話数（全発話数ｎ）に基づいて、重み閾値θを決定してもよい。この場合、上述の式（２）のαｎをｎに変更すればよい。

【００３３】

適応部４は、選択部３により選択された適応データを用いて、ベース音響モデル１０２を適応させることにより、適応音響モデル１０３を生成する。具体的には、ベース音響モデル１０２の適応は、ベース音響モデル１０２のパラメータを、適応データを用いて最適

化することにより行われる。ベース音響モデル１０２を適応させる方法は、例えばＤＮＮ（Ｄｅｅｐ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）、ＣＮＮ（Ｃｏｎｖｏｌｕｔｉｏｎａｌ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）、及び、ＲＮＮ（Ｒｅｃｕｒｒｅｎｔ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）などを使用する方法がある。適応音響モデル１０３は、音声認識装置１０の外部の記憶部に記憶されるようにしてもよい。

【００３４】

［動作方法の例］

図３は第１実施形態の音声認識装置１０の動作方法の例を示すフローチャートである。はじめに、生成部１が、言語モデル１０１とベース音響モデル１０２とを用いて音声データを認識する（ステップＳ１）。次に、生成部１が、ステップＳ１の処理により認識された音声データに含まれる発話を識別するラベルを生成する（ステップＳ２）。

【００３５】

次に、決定部２が、ラベルを用いて、同じ発話を含む音声データの個数を特定し、当該音声データに付与する重みを当該個数に応じて決定する（ステップＳ３）。次に、選択部３が、適応データとして使用する音声データを、重みに基づいて選択する（ステップＳ４）。次に、適応部４が、選択部３により選択された音声データ（適応データ）を用いて、ベース音響モデル１０２を適応させることにより、適応音響モデル１０３を生成する（ステップＳ５）。

【００３６】

以上説明したように、第１実施形態の音声認識装置１０では、生成部１が、言語モデル１０１とベース音響モデル１０２（第１音響モデル）とを用いて音声データを認識し、当該音声データに含まれる発話を識別するラベルを生成する。決定部２が、ラベルを用いて、同じ発話を含む音声データの個数を特定し、当該音声データに付与する重みを当該個数に応じて決定する。選択部３が、重みに基づいて音声データを選択する。そして、適応部４が、選択部３により選択された音声データ（適応データ）を用いて、ベース音響モデル１０２（第１音響モデル）を適応させることにより、適応音響モデル１０３（第２音響モデル）を生成する。

【００３７】

これにより第１実施形態の音声認識装置１０によれば、音響モデルを適応させた場合に生じる音声認識に与える悪影響を抑制することができる。

【００３８】

（第２実施形態）

次に第２実施形態について説明する。第２実施形態の説明では、第１実施形態と同様の説明については省略する。

【００３９】

適応データに含まれる非音声の部分が多いほど、当該適応データを用いた適応によって、非音声の確率が高くなる（音声の確率が低くなる）ため、音声の認識結果が非音声になることが多くなる。一方、適応データに含まれる非音声の部分が少ないほど、当該適応データを用いた適応によって、非音声の確率が低くなる（音声の確率が高くなる）ため、非音声の認識結果が音声になることが多くなる。

【００４０】

教師あり学習の場合、音声データから手動で発話ごとに切り出すため、非音声の部分のデータ量を制御できる。一方、教師なしの学習の場合、基本的にＶＡＤ（ｖｏｉｃｅ　ａｃｔｉｖｉｔｙ　ｄｅｔｅｃｔｉｏｎ）等の音声区間検出処理により、自動的に発話を切り出すため、非音声の部分のデータ量の制御が困難である。

【００４１】

第２実施形態では、音声データに含まれる音声（または非音声）の部分が多い場合でも、適応の悪影響を抑制できる構成について説明する。

【００４２】

［機能構成の例］

図４は第２実施形態の音声認識装置１０－２の機能構成の例を示すブロック図である。第２実施形態の音声認識装置１０－２は、生成部１、選択部３－２、適応部４及び計算部５を備える。生成部１及び適応部４の説明は、第１実施形態と同様なので省略する。

【００４３】

計算部５は、生成部１により生成されたラベルを用いて、音声データに含まれる音声フレームと、当該音声データに含まれる非音声フレームとの比率を計算する。

【００４４】

図５は第２実施形態の音声データの例を示す図である。図５の例では、音声データに含まれるフレームの数が２０である場合を示す。１、２、１８、１９及び２０番目のフレームは、非音声フレームの一例である。なお、ｓｉｌは、ｓｉｌｅｎｃｅの略である。３～１７番目のフレームは、音声フレームである。図５の音声データに含まれる発話は、「おはようございます」であり、当該発話のラベルも「おはようございます」である。

【００４５】

計算部５は、フレームごとの音素を表すために、生成したラベルを用いてアライメントを行う。発音の長さによって二つ以上のフレームに、一つの音素が対応することもある。図５の例では、例えば、フレーム４及び５に対応する音素は同じになる。

【００４６】

計算部５は、音声フレームと非音声フレームとの比率を計算する。図５の例では、音声フレームの比率は１５／２０＝０．７５である。非音声フレームの比率は５／２０＝０．２５である。

【００４７】

図４に戻り、選択部３－２は、音声フレームの比率が所定の選択範囲以内である音声データを、適応データとして選択する。所定の選択範囲は、例えば０．３～０．９である。図５の例では、音声フレームの比率は０．７５であるので、当該音声フレームは選択部３－２により適応データとして選択される。

【００４８】

所定の選択範囲は、適応の目的に応じて設定すれば良い。音声認識装置１０－２から、できるだけ音声の認識結果を出力したい場合、所定の選択範囲として、値がより高い区間の範囲を使用する（例えば、０．４～１．０）。一方、音声データに背景雑音が入っているため、音声認識装置１０－２から、できるだけ背景雑音の認識結果を出力したくない場合、所定の選択範囲として、値がより低い区間の範囲を使用する（例えば、０．０～０．５）。

【００４９】

以上説明したように、第２実施形態の音声認識装置１０－２によれば、例えば非音声フレームの比率が高い（例えば０．７以上）音声データが含まれている場合でも、選択部３－２により、当該音声データが選択されない。これにより、適応音声モデル１０３を使用した音声認識結果への悪影響を抑制できる。

【００５０】

（第３実施形態）

次に第３実施形態について説明する。第３実施形態の説明では、第１及び第２実施形態と同様の説明については省略する。第３実施形態では、第１及び第２実施形態を組み合わせる場合の動作について説明する。

【００５１】

［機能構成の例］

図６は第３実施形態の音声認識装置１０－３の機能構成の例を示すブロック図である。第３実施形態の音声認識装置１０－３は、生成部１、決定部２、選択部３－３、適応部４及び計算部５を備える。生成部１、決定部２及び適応部４の説明は、第１実施形態と同様なので省略する。計算部５の説明は、第２実施形態と同様なので省略する。

【００５２】

第１実施形態の選択部３による適応データの選択方法を選択方法Ａとし、第２実施形態

の選択部３－２による適応データの選択方法を選択方法Ｂとする。選択方法Ａ及びＢは独立である。そのため、選択方法Ａ及びＢの組み合わせによって、適応データとして使用する音声データを選択することが可能である。

【００５３】

選択部３－３は、決定部２により決定された重みと、計算部５により計算された音声フレームの比率とに基づいて、適応データとして使用する音声データを選択する。具体的には、選択部３－３は、例えば選択方法Ａにより適応データ候補を選択し、次に、適応データ候補から選択方法Ｂにより適応データを選択する。また例えば、選択部３－３は、選択方法Ｂにより適応データ候補を選択し、次に、適応データ候補から選択方法Ａにより適応データを選択する。

【００５４】

これにより第３実施形態の音声認識装置１０－３によれば、第１及び第２実施形態の効果を得ることができる。

【００５５】

（第４実施形態）

次に第４実施形態について説明する。第４実施形態の説明では、第１実施形態と同様の説明については省略する。第４実施形態では、適応音響モデル１０３を使用して、音声認識をする構成について説明する。

【００５６】

［機能構成の例］

図７は第４実施形態の音声認識装置１０－４の機能構成の例を示す図である。第４実施形態の音声認識装置１０－４は、生成部１、決定部２、選択部３、適応部４及び認識部６を備える。生成部１、決定部２、選択部３及び適応部４の説明は、第１実施形態と同様なので省略する。

【００５７】

認識部６は、言語モデル１０１及び適応音響モデル１０３を用いて、音声データの音声認識を行う。例えば、適応データが取得された環境と類似する環境で取得された音声データの音声認識をする場合、適応音響モデル１０３のパラメータは、ベース音響モデル１０２のパラメータより好ましい。また例えば、適応データに含まれる発話の話者と類似する話者（または同じ話者）の音声データの音声認識をする場合、適応音響モデル１０３のパラメータは、ベース音響モデル１０２のパラメータより好ましい。そのため、適応音響モデル１０３を用いて音声認識を行う場合、より高い音声認識精度が得られる。

【００５８】

（第５実施形態）

次に第５実施形態について説明する。第５実施形態の説明では、第１実施形態と同様の説明については省略する。第１実施形態では、言語モデル１０１及びベース音響モデル１０２の２種類のモデルを用いて適応を行っていた。第５実施形態では、言語モデル１０１及びベース音響モデル１０２を区別せずに、Ｅｎｄ－ｔｏ－Ｅｎｄの音声認識方法により、適応を行う場合の構成について説明する。

【００５９】

［機能構成の例］

図８は第５実施形態の音声認識装置１０－５の機能構成の例を示す図である。第５実施形態の音声認識装置１０－５は、生成部１－２、決定部２、選択部３及び適応部４－２を備える。決定部２及び選択部３の説明は、第１実施形態と同様なので省略する。

【００６０】

第５実施形態の音声認識装置１０－５は、音声認識ベースモデル１０４及び音声認識適応モデル１０５を記憶する。音声認識ベースモデル１０４は、音声の言語的な特徴、及び、音声の音響的な特徴の両方を区別せずにモデル化したデータである。

【００６１】

生成部１－２は、音声認識ベースモデル１０４を用いて音声データを認識し、ラベルを

生成する。第５実施形態では、音声認識ベースモデル１０４が、言語モデル１０１及びベース音響モデル１０２の役割を果たす。ラベルの生成方法の説明は、第１実施形態と同じなので省略する。

【００６２】

適応部４－２は、選択部３により選択された適応データを用いて、音声認識ベースモデル１０４を適応させることにより、音声認識適応モデル１０５を生成する。具体的には、音声認識ベースモデル１０４の適応は、音声認識ベースモデル１０４のパラメータを、適応データを用いて最適化することにより行われる。音声認識ベースモデル１０４を適応させる方法は、例えばＤＮＮ、ＣＮＮ及びＲＮＮ（Ｒｅｃｕｒｒｅｎｔ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ）などを使用する方法がある。音声認識適応モデル１０５は、音声認識装置１０の外部の記憶部に記憶されるようにしてもよい。

【００６３】

最後に、第１乃至第５実施形態の音声認識装置１０（１０－２，１０－３，１０－４，１０－５）のハードウェア構成の例について説明する。

【００６４】

［ハードウェア構成の例］

図９は第１乃至第５実施形態の音声認識装置１０（１０－２，１０－３，１０－４，１０－５）のハードウェア構成の例を示す図である。以下では、第１実施形態の音声認識装置１０の場合を例にして説明する。なお、第２乃至第５実施形態の音声認識装置１０－２（１０－３，１０－４，１０－５））のハードウェア構成も、第１実施形態の音声認識装置１０のハードウェア構成と同様である。

【００６５】

第１実施形態の音声認識装置１０は、制御装置３０１、主記憶装置３０２、補助記憶装置３０３、表示装置３０４、入力装置３０５及び通信装置３０６を備える。制御装置３０１、主記憶装置３０２、補助記憶装置３０３、表示装置３０４、入力装置３０５及び通信装置３０６は、バス３１０を介して接続されている。

【００６６】

制御装置３０１は、補助記憶装置３０３から主記憶装置３０２に読み出されたプログラムを実行する。主記憶装置３０２は、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）、及び、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）等のメモリである。補助記憶装置３０３は、ＨＤＤ（Ｈａｒｄ Ｄｉｓｋ Ｄｒｉｖｅ）、及び、メモリカード等である。

【００６７】

表示装置３０４は表示情報を表示する。表示装置３０４は、例えば液晶ディスプレイ等である。入力装置３０５は、音声認識装置１０を操作するためのインタフェースである。入力装置３０５は、例えばキーボードやマウス等である。音声認識装置１０がスマートフォン及びタブレット型端末等のスマートデバイスの場合、表示装置３０４及び入力装置３０５は、例えばタッチパネルである。通信装置３０６は、他の装置と通信するためのインタフェースである。

【００６８】

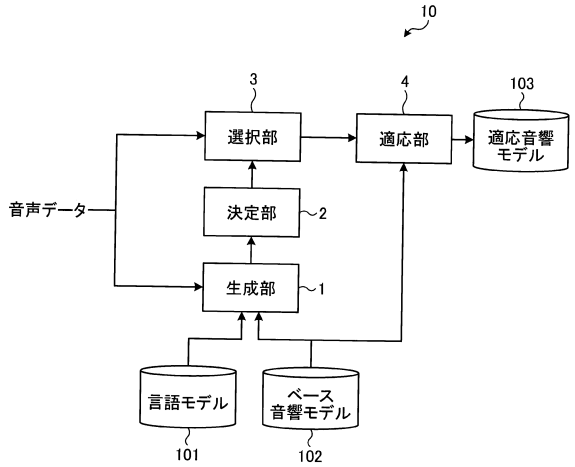
第１実施形態の音声認識装置１０で実行されるプログラムは、インストール可能な形式又は実行可能な形式のファイルでＣＤ－ＲＯＭ、メモリカード、ＣＤ－Ｒ及びＤＶＤ（Ｄｉｇｉｔａｌ Ｖｅｒｓａｔｉｌｅ Ｄｉｓｃ）等のコンピュータで読み取り可能な記憶媒体に記録されてコンピュータ・プログラム・プロダクトとして提供される。

【００６９】

また第１実施形態の音声認識装置１０で実行されるプログラムを、インターネット等のネットワークに接続されたコンピュータ上に格納し、ネットワーク経由でダウンロードさせることにより提供するように構成してもよい。また第１実施形態の音声認識装置１０で実行されるプログラムをダウンロードさせずにインターネット等のネットワーク経由で提供するように構成してもよい。

【００７０】
また第１実施形態の音声認識装置１０のプログラムを、ＲＯＭ等に予め組み込んで提供するように構成してもよい。
【００７１】
第１実施形態の音声認識装置１０で実行されるプログラムは、上述の機能ブロックのうち、プログラムによっても実現可能な機能ブロックを含むモジュール構成となっている。当該各機能ブロックは、実際のハードウェアとしては、制御装置３０１が記憶媒体からプログラムを読み出して実行することにより、上記各機能ブロックが主記憶装置３０２上にロードされる。すなわち上記各機能ブロックは主記憶装置３０２上に生成される。
【００７２】
なお上述した各機能ブロックの一部又は全部をソフトウェアにより実現せずに、ＩＣ（Ｉｎｔｅｇｒａｔｅｄ　Ｃｉｒｃｕｉｔ）等のハードウェアにより実現してもよい。
【００７３】
また複数のプロセッサを用いて各機能を実現する場合、各プロセッサは、各機能のうち１つを実現してもよいし、各機能のうち２以上を実現してもよい。
【００７４】
また第１実施形態の音声認識装置１０の動作形態は任意でよい。第１実施形態の音声認識装置１０を、例えばネットワーク上のクラウドシステムとして動作させてもよい。
【００７５】
本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。
【符号の説明】
【００７６】
１　生成部
２　決定部
３　選択部
４　適応部
５　計算部
６　認識部
１０１　言語モデル
１０２　ベース音響モデル
１０３　適応音響モデル
１０４　音声認識ベースモデル
１０５　音声認識適応モデル
３０１　制御装置
３０２　主記憶装置
３０３　補助記憶装置
３０４　表示装置
３０５　入力装置
３０６　通信装置
３１０　バス

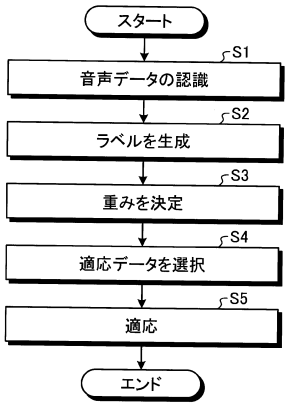
【図１】



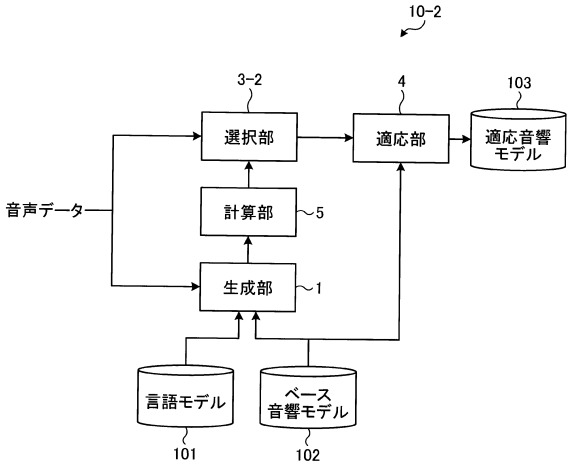
【図２】

音声データ	音声認識結果	ラベル	カウント数	重み
発話-1	あのー	あのー	1	1.00
発話-2	先週の進捗ですが	せんしゅうのしんちょくですが	1	1.00
発話-3	あのー	あのー	2	0.37
発話-4	いろんなトラブルがあったので	いろんなとらぶるがあったので	1	1.00
発話-5	あのー	あのー	3	0.14
発話-6	うまく進んでない	うまくすすんでない	1	1.00

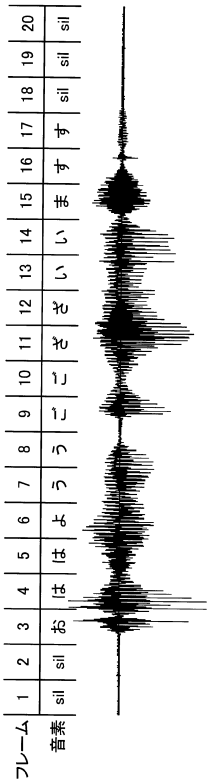
【図３】



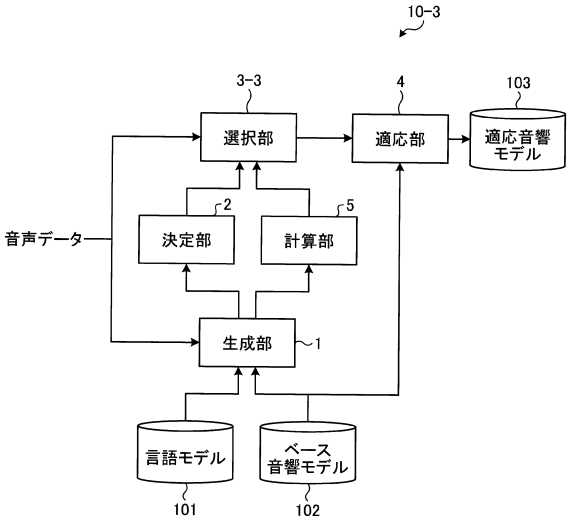
【図４】



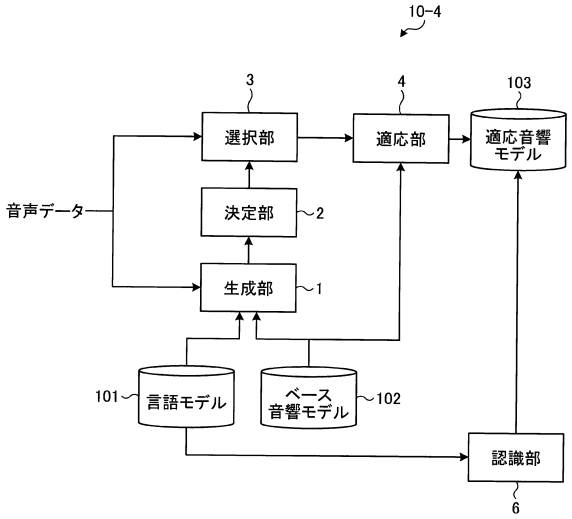
【図５】


フレーム
音素
1 sil
2 sil
3 お
4 は
5 は
6 よ
7 う
8 う
9 い
10 い
11 そ
12 そ
13 い
14 い
15 ま
16 す
17 す
18 sil
19 sil
20 sil

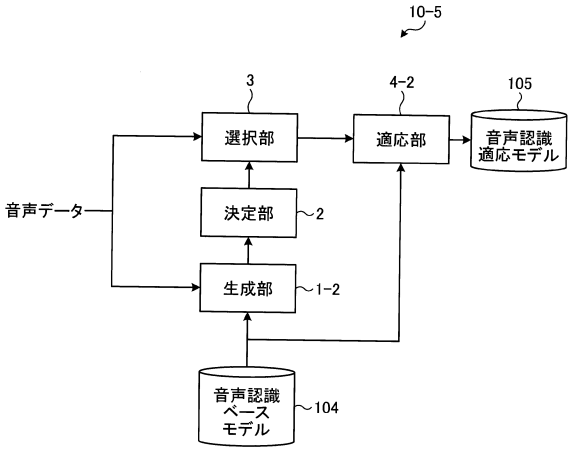
【図６】


10-3
3-3
4
103
選択部
適応部
適応音響モデル
音声データ
2
5
決定部
計算部
生成部
1
言語モデル
ベース音響モデル
101
102

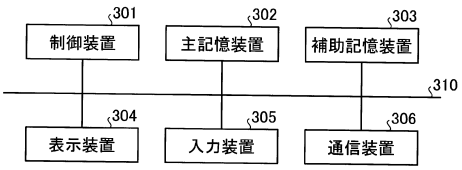
【図７】


10-4
3
4
103
選択部
適応部
適応音響モデル
音声データ
決定部
2
生成部
1
101
言語モデル
ベース音響モデル
102
認識部
6

【図８】


10-5
3
4-2
105
選択部
適応部
音声認識適応モデル
音声データ
決定部
2
生成部
1-2
音声認識ベースモデル
104

【図９】


301
制御装置
302
主記憶装置
303
補助記憶装置
310
304
表示装置
305
入力装置
306
通信装置

フロントページの続き

(56)参考文献 特開２０１４－１０２３４５（ＪＰ，Ａ）
特開２００２－１５６９９４（ＪＰ，Ａ）
特開２０００－３０５５８９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４
ＩＥＥＥ Ｘｐｌｏｒｅ