

明細書

音声認識装置

5 技術分野

本発明は、特定話者に固有の登録語と不特定話者に共通の固定語が混在した語彙に対し、音声認識が可能な音声認識装置および方法に関するものである。

背景技術

- 10 特定話者に固有の登録語を認識する技術を一般に特定話者音声認識と呼ぶ。特定話者音声認識においては、特定話者が予め音声認識させたい単語の音声を登録するという作業が発生する。この作業は具体的には、特定話者が予め発声した単語の音声サンプルをある特徴パラメータの系列（テンプレートと呼ぶ）に変換し、その系列を単語のラベルとともにメモリーやハードディスクなどの記憶装置に蓄積する。
- 15 音声サンプルをある特徴パラメータの系列に変換する方法としては、ケプストラム分析や線形予測分析などが知られており、「音声・音情報のデジタル信号処理」（鹿野清宏、中村哲、伊勢史郎共著、（株）昭晃堂）にも詳述されている。特定話者音声認識においては、記憶装置に蓄積された特徴パラメータの系列と入力された音声から変換された特徴パラメータの系列とを比較し、入力された音声から変換された特徴パラメータの系列に最も類似した特徴パラメータの系列を持つ単語のラベルを認識結果として出力する。
- 20

記憶装置に蓄積された特徴パラメータの系列と入力された音声から変換された特徴パラメータの系列とを比較する方法としては、動的計画法(dynamic programming) による時間軸伸縮マッチング(Dynamic Time Warping, D TW) が広く用いられており、上記「音声・音情報のデジタル信号処理」にも詳しく述べられている。

25

一方、不特定話者に共通の固定語を認識する技術を一般に不特定話者音声認識と呼ぶ。不特定話者音声認識においては、不特定話者に共通の固定語の特徴パラメータに関する情報が予め記憶装置に蓄積されているため、特定話者音声認識のようにユーザーが音声認識させたい単語を登録するという作業は発生しない。音声サンプルをある特徴パラメータの系列に変換する方法としては、特定話者音声認識と同様にケプストラム分析や線形予測分析などが知られている。また、不特定話者に共通の固定語の特徴パラメータに関する情報の作成およびその情報と入力された音声から変換された特徴パラメータの系列との比較には、隠れマルコフモデル (Hidden Markov Model, HMM) による方法が一般に用いられている。

- 10 HMMによる不特定話者音声認識についても、上記「音声・音情報のデジタル信号処理」に詳しく述べられている。例えば、日本語の場合、音声単位を上記「音声・音情報のデジタル信号処理」の第2章に記載されている音韻のセットとし、各音韻がHMMによりモデル化されているとする。表1に音韻のセットのラベルの一覧を示す。

15 【表1】

母音	a, i, u, e, o
摩擦音	f, z, s, zh, sh, h
破擦音	dz, ts, dh, ch
破裂音	b, p, d, t, g, k
半母音	w, r, y
鼻音	m, n, ng

この時、例えば「CD（シーディー）」は図2Aのような話者に共通の音韻ラベルのネットワーク（固定語ラベル系列と呼ぶ）でモデル化することができる。

また、例えば「MD（エムディー）」は図2Bのような固定語ラベル系列でモデル化することができる。HMMによる音韻モデルのデータと固定語ラベル系列を用意すれば、上記「音声・音情報のデジタル信号処理」の第4章に記載されているViterbiアルゴリズムにより、当業者は不特定話者音声認識装置を構成することができる。

音声認識装置において、特定話者に固有の登録語と不特定話者に共通の固定語が混在した語彙に対し、認識が可能であるという機能を提供したいというニーズがある。例えば、カーオーディオ装置においては、安全上の理由から「CD」、「MD」などの各機器の制御を音声で行いたいというニーズがある。これらの機器の名前は、不特定多数の話者で共通に設定することができるため、不特定話者音声認識技術で実現することができる。これにより、特定話者音声認識技術で必要な登録の作業を回避することができ、ユーザーインターフェースの面で好ましい。

一方で、例えば、CDチェンジャーに挿入された複数のCDを声で選択し再生するというニーズが存在する。この場合、CDチェンジャーに挿入されるCDのタイトル名、歌手名などはユーザーにより異なると考えられる。従って、従来の不特定話者音声認識技術ではなく、特定話者音声認識技術の適用が必要となる。すなわち、ユーザーがCDチェンジャーに挿入するCDのタイトル名、歌手名などを予め声により登録するという作業が必要となる。「CD」、「MD」などの各機器の名前とCDのタイトル名、歌手名などが混在した語彙に対し、音声認識を行うことが可能であれば、「CD」、「MD」などの不特定話者に共通の固定語を認識できるモードとCDのタイトル名、歌手名などの特定話者に固有の登録語を認識できるモードを切り替える必要がなく、ユーザーに対してより使いやすい音声認識機能を提供できると考えられる。

これまで、特定話者音声認識には、DTWによる方法が広く用いられてきた。また、不特定話者音声認識には、HMMによる方法が広く用いられてきた。上記ニーズを実現する1つの方法として、DTWによる特定話者音声認識とHMMによる不特定話者音声認識を併用する方法が考えられる。この2つの方法で入力音声のパラメータ系列と記憶装置に記憶された語彙単語のパラメータ系列の情報とを比較する際に用いる尺度は一般に異なっている。従って、DTWによる特定話者音声認識により最も入力音声に近いと判断された特定話者に固有の登録語とHMMによる不特定話者音声認識により最も入力音声に近いと判断された不特定話者に共通の固定語のいずれかが入力音声により近いのかを判断するのは容易ではない。

- 10 DTWによる特定話者音声認識において、固定語に対し多数の話者の音声を用いて複数のテンプレートを記憶装置に記憶させることにより、不特定話者音声認識を実現することも可能である。これにより、DTWを用いて上記ニーズを実現することができる。しかしながら、この方法では、各固定語のテンプレートを複数持つことによる記憶装置の容量の増加、複数のテンプレートとのDTWによる照合時間の増加、固定語を変更する際にあらためて多数話者の音声を収集する必要があるなどの不具合がある。

- 20 以上を要約すると、たとえば、カーオーディオ装置に音声認識装置を搭載する場合、不特定話者用の音声認識装置は、製造会社にとっては多数のユーザの声を個々に登録する必要がないので好都合である反面、個々のユーザにとっては、認識精度が特定話者の認識精度よりも若干落ちるというデメリットがある。

また、特定話者用の音声認識装置は認識精度において優れるものの、製造会社にとっては、個々のユーザの音声から特徴パラメータを取り出し、音声認識装置に予め記憶させることは至難である。ユーザ自身が自己の音声を登録するにしても、たとえば、多くの単語を登録するには多大の労力を必要とする。

- 25 さらに従来の特定制話者の音声認識に使用する方法と不特定話者の音声認識に使用する方法は種類内容が異なるので両方の音声認識方法を1つの装置に組み込み

すると装置が大型化してしまうという問題がある。

発明の開示

本発明の目的は、このような点を考慮して、不特定話者の音声を認識する音声認識方法を使用した場合でも、特定話者の発声する登録語を精度よく認識することができる音声認識装置および方法を提供することにある。

本発明によれば、入力音声を、単語より短い長さの音声単位のモデルの情報を用いて、単語認識を行う音声認識装置であって、不特定話者の入力音声の単語認識のために共通使用する固定語に対する、前記音声単位のラベル系列を蓄積しておく語彙ラベルネットワーク蓄積手段と、特定話者の入力音声から登録語に対する前記音声単位のラベル系列を作成する登録語ラベル系列抽出手段と、当該作成された登録語のラベル系列を前記語彙ラベルネットワーク蓄積手段に追加登録する登録手段とを具え、前記音声単位が音韻の隠れマルコフモデルの各状態を分割して生成した音響イベントである音声認識装置が提供される。

本発明によれば、音声認識装置は前記固定語を前記語彙ラベルネットワーク蓄積手段に登録する手段をさらに具えてもよい。

本発明によれば、音声認識装置では前記不特定話者の入力音声の単語認識のために共通使用する固定語に対する、音声単位のラベル系列と、作成された登録語のラベル系列とが並列のネットワークの形態で前記語彙ラベルネットワーク蓄積手段に記憶されてもよい。

本発明によれば、音声認識装置の製造会社では、多数の不特定のユーザを対象にした固定語のラベル系列を音声認識用データとして提供すればよく、ユーザは所望の単語について自己の音声を登録することができる。換言すると、不特定話者方式の音声認識方法に特定話者の音声認識機能を加味することが可能となる。これにより、ユーザおよび製造会社双方に好適な音声認識装置および方法を提供することが可能となる。

図面の簡単な説明

- 図1は本発明実施形態のシステム構成を示すブロック図である。
- 図2Aは「CD（シーディー）」の固定語ラベル系列を示す説明図である。
- 5 図2Bは「MD（エムディー）」の固定語ラベル系列を示す説明図である。
- 図3は音韻ネットワークの内容を示す説明図である。
- 図4Aは「ジャズ」の登録語ラベル系列を示す説明図である。
- 図4Bは「ポップス」の登録語ラベル系列を示す説明図である。
- 図5は固定語と登録語が混在した語彙ラベルネットワークを示す説明図である。
- 10 図6は固定語だけの語彙ラベルネットワークを示す説明図である。
- 図7は登録語だけの語彙ラベルネットワークを示す説明図である。
- 図8は音韻HMMの構造を示す説明図である。
- 図9は音響イベントHMM * 1の構造を示す説明図である。
- 図10は音響イベントHMM * 2の構造を示す説明図である。
- 15 図11は音響イベントHMM * 3の構造を示す説明図である。
- 図12は無制約音響イベントネットワークを示す説明図である。
- 図13は位置制約つき音響イベントネットワークを示す説明図である。

発明を実施するための最良の形態

- 20 本発明は、従来不特定話者音声認識に用いられてきたHMMを用いて、特定話者に固有の登録語と不特定話者に共通の固定語が混在した語彙に対し、認識が可能な音声認識装置を提供するものである。HMMを用いた不特定話者音声認識においては、音声単位（音節、半音節、音韻、音響イベントなど）のモデルデータ、不特定話者に共通の固定語に対応する音声単位のラベル系列および各固定語のラベル系
- 25 列相互の接続情報を用いて、先述のV i t e r b i アルゴリズムにより、入力音声に含まれる固定語の認識を行う。

特定話者に固有の登録語に対応する音声単位のラベル系列を何らかの手段で得られれば、そのラベル系列を上記ラベル系列相互の接続情報に付加することにより、先述のV i t e r b i アルゴリズムを用いて、特定話者に固有の登録語と不特定話者に共通の固定語が混在した語彙に対する認識を行うことが可能であると考えられる。

特定話者に固有の登録語に対応する音声単位のラベル系列を得る方法としては、各固定語のラベル系列相互の接続情報の代わりに音声単位が任意の順番で任意の長さで接続可能な接続情報を用いて、V i t e r b i アルゴリズムにより、入力音声に含まれる全ての音声単位の系列を取得することにより実現できる。

以下、図面を参照して、本発明の実施形態を詳細に説明する。図1は本発明実施形態の基本構成を示す。

図1において、マイクロホン（不図示）から入力されたアナログ音声信号11はADコンバータによる入力手段aによりデジタル信号22に変換される。デジタル信号22は変換手段bにより一定のフレーム周期で音響パラメータ33に変換される。音響（特徴）パラメータとしては、上記「音声・音情報のデジタル信号処理」にも記述されているケプストラムなどを用いればよい。

（特定話者の単語登録）

特定話者に固有の登録語を登録する場合には、音声認識装置の不図示のスイッチにより単語登録を指示し、変換手段bを登録語ラベル系列抽出手段dに接続する。

特定話者、すなわち、音声認識装置が接続されたカーオーディオのユーザ（特定話者）は登録したい音声をマイクロホンに入力する。

マイクロホンで発生したアナログ音声信号11は変換手段bにより音響パラメータ33に変換される。音響パラメータ33が登録語ラベル系列抽出手段dに送られる。登録語ラベル系列抽出手段dはハードディスクやメモリ等の認識用データ蓄積手段cに蓄積された認識用データ44を参照しながら、音声単位のラベル系列の認識を行い、音響パラメータ33を登録語ラベル系列55に変換する。認識用

データ蓄積手段cに蓄積された認識用データ44としては、隠れマルコフモデルでモデル化された音声単位のモデルのデータおよび音響パラメータ33を登録語ラベル系列55に変換するための音声単位のモデルのネットワークデータの2つが蓄積されている。

- 5 先行特許であるUSP 5, 732, 187号に述べられているように、音声単位としては、音節、半音節、音韻が考えられる。

以下では、音韻を例にとって、不特定話者に共通の固定語と特定話者に固有の登録語が混在した語彙を認識する音声認識装置および方法について説明する。音響パラメータ33を登録語ラベル系列55に変換するための音声単位（今の場合は省
10 略）のモデルのネットワークデータとは、許容される音韻の接続を表現したネットワークのことである。

例えば、図3の音韻ネットワークの場合、表1の27種類の音韻が任意の順序で任意の長さで接続しうることを示している。このネットワークデータとHMMによる音韻モデルのデータを用いれば、上記V i t e r b i アルゴリズムにより、特定
15 話者に固有の登録語の音声を音韻ラベルの系列に変換することができる。図4Aに、ある話者が登録のために発声した「ジャズ」の入力信号から得られる音韻ラベルのネットワークの例を示す。音韻の認識率が100%であれば、「ジャズ」と発声した音声からは、j + a + z + uという音韻ラベル系列が得られるはずであるが、この例では、最後のuが無声化して得られていない。

20 図4Bに、ある話者が登録のために発声した「ポップス」の入力信号から得られる音韻ラベルのネットワークの例を示す。音韻の認識率が100%であれば「ポップス」と発声した音声からは、p + o + p + u + s + uという音韻ラベル系列が得られるはずであるが、p、t、kの識別は一般に困難であるため、2つ目のpがkに置き換わっている。こうして、登録語ラベル系列抽出手段dにより得られた音韻
25 ラベルのネットワークを登録語ラベル系列55と呼ぶことにする。登録語ラベル系列55は登録手段iにより、ハードディスクやメモリ等の語彙ラベルネットワーク

蓄積手段 f に追加登録され、記憶される。

(不特定話者に共通の固定語の登録)

- 一方、不特定話者に共通の固定語のスペルから予めあるルールに従って音韻ラベルのネットワークを抽出することができる。あるいは、固定語に対する 1 つ以上の
- 5 不特定話者の音声サンプルから前記 (特定話者の単語登録) と同様にして 1 つ以上の音韻ラベルネットワークを作成することもできる。これらのネットワークを固定語ラベル系列 6 6 と呼び、そのデータを語彙ラベルネットワーク蓄積手段 f に転送するために一時記憶する手段、例えば、CPU および RAM を固定語ラベル系列記憶手段 e と呼ぶことにする。例えば、固定語「CD」、「MD」に対しては、図 2 A、
- 10 図 2 B に示される固定語ラベル系列を得ることができる。固定語ラベル系列についてのデータはフロッピーディスクや CD 等の記録媒体に製造会社側で記録しておき、この記録媒体から音声認識装置内の固定語レベル系列記憶手段 e に引き渡せばよい。固定語レベル系列記憶手段 e を介して語彙ラベルネットワーク蓄積手段 f に蓄積 (記憶) することにより不特定話者についての固定語ラベル系列を登録する。
- 15 登録語ラベル系列抽出手段 d から抽出され、登録手段 i により語彙ラベルネットワーク蓄積手段 f に追加登録される、登録語ラベル系列 5 5 と固定語ラベル系列記憶手段 e により記憶された固定語ラベル系列 6 6 から語彙ネットワーク蓄積手段 f に蓄積される語彙ラベルネットワーク 7 7 を構成する方法としては、以下の 3 つの方法が考えられる。
- 20 まず、1 つ目は、登録語ラベル系列 5 5 と固定語ラベル系列 6 6 の両方を含むネットワークを形成し、それを語彙ラベルネットワーク 7 7 とする方法である。図 5 に「CD」、「MD」の固定語ラベル系列と「ジャズ」、「ポップス」の登録語ラベル系列を接続した語彙ラベルネットワークを示す。
- このネットワークにより、「CD」、「MD」の固定語と「ジャズ」、「ポップス」
- 25 の登録語の 4 つの単語の内、いずれでも認識できる音声認識装置を実現することができる。

2つ目の方法は、登録語ラベル系列55のみを含むネットワークを形成し、それを語彙ラベルネットワーク77とする方法である。図6に「CD」、「MD」の固定語ラベル系列を接続した語彙ラベルネットワークを示す。このネットワークにより、「CD」、「MD」の固定語の2つの単語の内、いずれかを認識できる不特定話者音声認識装置を実現することができる。

3つ目の方法は、登録語ラベル系列55のみを含むネットワークを形成し、それを語彙ラベルネットワーク77とする方法である。図7に「ジャズ」、「ポップス」の登録語ラベル系列を接続した語彙ラベルネットワークを示す。このネットワークにより、「ジャズ」、「ポップス」の登録語の2つの単語の内、いずれかを認識できる特定話者音声認識装置を実現することができる。

なお、音声によりCDチェンジャーに動作指示を与える場合には、予め不特定話者の固定語のラベル系列と動作命令を予め関連付けておき、特定話者（ユーザ）が登録した登録語ラベル系列とを関連付けるとよい。たとえば、このための方法としては、予め、表示器に選択可能な動作指示内容を表示させた後、カーソルキー等で表示の動作指示を選択した後、この動作指示に関連付ける単語を音声でマイクロホンから入力して、登録語ラベル系列抽出手段dにより抽出された登録語ラベル系列55を登録手段iにより語彙ラベルネットワーク蓄積手段fに登録する。また、動作指示内容を示すコードと、対応する登録語ラベル系列とをテーブル（関連付けテーブル）形態でハードディスク等に保存しておく、以後、音声による動作指示が可能となる。この関連付けテーブルには製造会社が用意した固定語ラベル系列と、それらに対応する動作指示コードをも記憶してもよいことは勿論である。

（音声認識処理）

ユーザは不図示のモード指示スイッチにより音声認識モードを指示すると、変換手段bと認識手段gに接続が切り替わる。入力手段aから入力され、変換手段bで変換されて出力された音響パラメータ33と認識用データ蓄積手段cに蓄積された音韻モデルのデータからなる認識用データ45と、語彙ラベルネットワーク蓄積

手段 f に蓄積された語彙ラベルネットワーク 77 を用いて、認識手段 g において上述の V i t e r b i アルゴリズムにより、単語についての認識結果 88 を得る。認識結果 88 は、スピーカーやディスプレイなどの出力手段 h により、音声や画面による出力結果 99 として話者に出力される。また、音声により CD チェンジャーなど
5 に対して動作の指示を与える場合には、上述の関連付けテーブルにより音声認識結果に対応する動作指示の内容が判別されて動作指示に対応する制御プログラム等が実行される。

以上のシステム構成により、ユーザは頻繁に使用する動作指示については自己の音声で単語形態で登録しておくことで、より精度よく、動作指示を行うことができる。
10 このシステムのメリットは、ユーザ（特定話者）が登録した登録語ラベル系列がラベル認識（認識用データを使用した音響パラメータからラベルへの変換）において誤認識されても、影響がないという点である。

音声認識装置では、誤認識結果のラベル系列のままで、動作指示が登録され、音声認識モードで特定話者がその動作指示を音声で入力した場合にも、音声認識装置
15 は同じ内容で誤認識を行うので、かならず、予め関連付けられた動作指示が選択される。

一方、不特定話者用の固定語ラベル系列を音声認識に使用した場合には、ラベル認識に誤認識が生じると、動作指示に対応するラベル系列と上記誤認識の結果が一致しないので、音声の動作指示が正しく認識されない場合がある。このため、製造
20 会社が用意する固定語ラベル系列の中の特定のものと同じ内容の動作指示を特定話者が音声で登録することができる、また、特定話者の登録語ラベル系列により音声認識モードにおいて、正しく音声認識を行うことができる。

本実施形態では音声単位として音韻から分解されて生成された不特定話者用の音響イベントのモデルを用いることにより、特定話者の発声する登録語を言語に依
25 らず、精度良く認識する。

以下では、音韻よりも小さな音声単位として、音韻 HMM を構成する状態を用い

た場合について説明を加える。言語や音韻の種類を問わず、一般に音韻は図8に示される4つの状態からなるHMMでモデル化されることが多い。

状態1は初期状態であり、状態遷移は必ず状態1から開始される。状態4は最終状態であり、この状態に遷移するとそれ以上状態遷移は発生しない。 a_{ij} は状態
5 i から状態 j への遷移する確率を表し、遷移確率と呼ばれる。 $j = i + 1$ とすると、 $a_{ii} + a_{ij} = 1.0$ である。

$b_i(x)$ は状態 i から状態遷移する際に、観測ベクトル x が出力される確率を表し、出力確率と呼ばれる。観測ベクトルは、図1の音響パラメータ33に対応し、ケプストラムがよく用いられる。出力確率 $b_i(x)$ は1個以上の正規分布の重ね
10 合わせで表現される。また、出力確率は状態 i へ状態遷移する際に、観測ベクトル x が出力される確率 $b_i(x)$ として定式化される場合や、状態 i から状態 j へ状態遷移する際に、観測ベクトル x が出力される確率 $b_{ij}(x)$ として定式化される場合もある。

出力確率 $b_i(x)$ が1個以上の正規分布の重ね合わせで定式化される場合、H
15 MMを連続HMMと呼ぶ。その他に、半連続HMMや離散HMMで音韻をモデル化する場合があるが、本発明は同様に適用が可能である。

図8に示される4状態の音韻HMMを図9、図10、図11の3つの2状態からなるHMMに分割する。これらを音響イベントHMMと呼ぶこととし、状態番号に
合わせて、*、1、*、2、*、3と表すこととする。ここで、*は表1の音韻ラ
20 ベルを表す。

これらの新しく生成されたHMMは音韻の中の音響的なイベントを表現しているものと考えることができるため、音響イベントHMMと呼ぶことにする。これらの音響イベントHMMが任意の順番で任意の長さで接続しうるネットワークを図
12に示す。このネットワークを無制約音響イベントネットワークと呼ぶ。この無
25 制約音響イベントネットワークは、図3の音韻ネットワークに比べて、音響的なイベントの系列の自由度が高く、その分、音響的なイベント系列の表現能力が高いと

考えられる。

従って、図3の音韻ネットワークを用いてV i t e r b i アルゴリズムにより得られる音韻系列に比べて、図12の無制約音響イベントネットワークを用いてV i t e r b i アルゴリズムにより得られる音響イベント系列の方が、特定話者に固有の登録語の発声をより精度良く近似するものと期待できる。

そこで、無制約音響イベントネットワークを用いてV i t e r b i アルゴリズムにより得られる音響イベント系列を図1の登録語ラベル系列55として用いることにより、登録語に対してより高い認識性能が得られると思われる。

一方、図12の無制約音響イベントネットワークでは、特定話者に固有の登録語の発声を近似する音響イベント系列の候補が数多く存在するため、その選択に多くの処理時間を要することも考えられる。

そこで、図13に示すような状態の位置に関する制約を設けたネットワークを使用することも処理時間を削減する上で有効であると考えられる。図13のネットワークを位置制約つき音響イベントネットワークと呼ぶことにする。このネットワークでは、図12の無制約音響イベントネットワークに比べて、接続しうる音響イベントの数が3分の1になっているため、特定話者に固有の登録語を近似する音響イベント系列の候補数が少なくなっており、若干の認識性能の低下は起こるかも知れないが、特定話者に固有の登録語を近似する音響イベント系列を求める処理時間もそれに応じて短縮できると考えられる。

実際に、音韻系列と音響イベント系列で特定話者の登録語の認識性能の比較実験を行った。

日本語の音韻HMMと図3の音韻ネットワークを用いて、V i t e r b i アルゴリズムにより128単語の音韻系列を求め、それを認識語彙として、日本語、英語、仏語、独語、伊語についての特定話者単語認識を行うための男女各2名を選んだ。また、その時の音声認識の平均認識率および最低認識率を表2に示す。音声と雑音の比率はいずれも10dBである。

【表 2】

言語	日本語	英語	仏語	独語	伊語
平均認識率	93.6%	86.9%	88.7%	92.2%	91.4%
5 最低認識率	92.2%	82.8%	83.6%	85.2%	87.5%

音韻ネットワークを用いた場合の

特定話者単語認識性能

次に日本語の音韻HMMから上述の方法により分解されて生成された音響イベントHMMと図13の位置制約つき音響イベントネットワークを用いてViterbiアルゴリズムにより128単語の音響イベント系列を求め、それを認識対象とした場合の日本語、英語、仏語、独語、伊語についての特定話者単語認識の平均認識率および最低認識率を表3に示す。

15 【表 3】

言語	日本語	英語	仏語	独語	伊語
平均認識率	97.5%	92.9%	94.0%	94.0%	93.6%
最低認識率	96.1%	91.1%	92.2%	91.4%	91.4%

位置制約つき音響イベントネットワークを用いた場合の

20 特定話者単語認識性能

表2と表3を比較すると言語に依らず、音韻系列を使用するよりも音響イベント系列を使用する方が平均認識率が2-6%程度向上していることが分る。また、最低認識率も4-8%向上していることが分る。これらの結果から図3の音韻ネットワークを用いてViterbiアルゴリズムにより得られる音韻系列を使用した場合25 場合に比べ、図13の位置制約つき音響イベントネットワークを用いてViter

b i アルゴリズムにより得られる音響イベント系列を使用した方が、特定話者に固有の登録語に対してより高い性能が得ることが明らかになった。この音響イベント系列の優位性は、音響イベント系列の方が音韻系列よりも特定話者に固有の近似度が高いことによると考えられ、この点が本願発明の根幹をなす。

5 上述の実施形態については以下の形態が実施できる。

- 1) 図1に示したシステム構成は、デジタル回路で構成してもよい、マイクロコンピュータやパーソナルコンピュータのソフトウェア処理で実現することもできる。この場合、変換手段b、登録語ラベル系列抽出手段d、認識手段gについてCPUがソフトウェアプログラムを実行することによりそれらの機能が実現される。どのような回路構成とするかは、音声認識装置の用途に応じて適宜決定すればよい。たとえば、パーソナルコンピュータにおいて文字入力および制御キー入力を音声で行う場合には、パーソナルコンピュータに搭載されたCPU、ハードディスクを使用して図1のシステムを実現するとよい。
- 2) 上述の実施形態では、不特定話者に共通の固定語ラベル系列をCDROM等から音声認識装置に登録可能としているが、音声認識装置に内蔵のEEPROM等の不揮発性書き込み可能メモリに記憶してもよいことは勿論である。この場合には不足する単語、あるいはより認識精度を高めたい単語を特定話者により音声で登録する。

請求の範囲

1. 入力音声を、単語より短い長さの音声単位のモデルの情報を用いて、単語認識を行う音声認識装置において、
- 5 不特定話者の入力音声の単語認識のために共通使用する固定語に対する、前記音声単位のラベル系列を蓄積しておく語彙ラベルネットワーク蓄積手段と、
特定話者の入力音声から登録語に対する前記音声単位のラベル系列を作成する登録語ラベル系列抽出手段と、
当該作成された登録語のラベル系列を前記語彙ラベルネットワーク蓄積手段に
10 追加登録する登録手段と
を具え、前記音声単位が音韻の隠れマルコフモデルの各状態を分割して生成した音響イベントであることを特徴とする音声認識装置。
2. 請求項1に記載の音声認識装置において、前記固定語を前記語彙ラベルネットワーク蓄積手段に登録する手段をさらに具えたことを特徴とする音声認識装置。
- 15 3. 請求項1に記載の音声認識装置において、前記不特定話者の入力音声の単語認識のために共通使用する固定語に対する、音声単位のラベル系列と、作成された登録語のラベル系列とを並列のネットワークの形態で前記語彙ラベルネットワーク蓄積手段に記憶することを特徴とする音声認識装置。
4. 請求項3に記載の音声認識装置において、前記固定語を前記語彙ラベルネットワーク蓄積手段に登録する手段をさらに具えたことを特徴とする音声認識装置。
20
5. 入力音声を、単語より短い長さの音声単位のモデルの情報を用いて、単語認識を行う音声認識方法において、
不特定話者の入力音声の単語認識のために共通使用する固定語に対する前記音声単位のラベル系列を語彙ラベルネットワーク蓄積手段に蓄積しておき、
25 特定話者の入力音声から登録語に対する前記音声単位のラベル系列を作成し、
当該作成された登録語のラベル系列を前記語彙ラベルネットワーク蓄積手段に



追加登録し、

前記音声単位が音韻の隠れマルコフモデルの各状態を分解して生成した音響イベントであることを特徴とする音声認識方法。

6. 請求項5に記載の音声認識方法において、前記固定語を前記語彙ラベルネットワーク蓄積手段に登録可能とすることを特徴とする音声認識方法。
- 5 7. 請求項5に記載の音声認識方法において、前記不特定話者の入力音声の単語認識のために共通使用する固定語に対する、音声単位のラベル系列と、作成された登録語のラベル系列とを並列のネットワークの形態で前記語彙ラベルネットワーク蓄積手段に記憶することを特徴とする音声認識方法。
- 10 8. 請求項7に記載の音声認識方法において、前記固定語を前記語彙ラベルネットワーク蓄積手段に登録する手段をさらに具えたことを特徴とする音声認識方法。

