(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2010 年 9 月 16 日(16.09.2010)

PCT

(10) 国際公開番号
WO 2010/104040 A1

(51) 国際特許分類:
G10L 13/06 (2006.01) *G10L 13/00* (2006.01)

(21) 国際出願番号: PCT/JP2010/053802

(22) 国際出願日: 2010 年 3 月 8 日(08.03.2010)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2009-055784 2009 年 3 月 9 日(09.03.2009) JP

(71) 出願人 (米国を除く全ての指定国について): 国立大学法人豊橋技術科学大学(National University Corporation TOYOHASHI UNIVERSITY OF TECHNOLOGY) [JP/JP]; 〒4418580 愛知県豊橋市天伯町雲雀ヶ丘１－１ Aichi (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 新田 恒雄 (NITTA Tsuneo) [JP/JP]; 〒4418580 愛知県豊橋市天伯町雲雀ヶ丘１－１国立大学法人豊橋技術科学大学内 Aichi (JP).

(74) 代理人: 井川浩文, 外(IKAWA Hirofumi et al.); 〒4400814 愛知県豊橋市前田町１－２－１１柴田法律特許事務所 Aichi (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

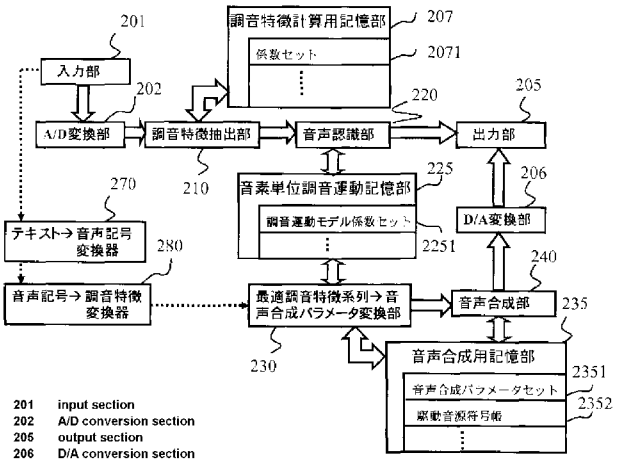
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[続葉有]

(54) **Title**: VOICE SYNTHESIS APPARATUS BASED ON SINGLE-MODEL VOICE RECOGNITION SYNTHESIS, VOICE SYNTHESIS METHOD AND VOICE SYNTHESIS PROGRAM

(54) 発明の名称 : １モデル音声認識合成に基づく音声合成装置、音声合成方法および音声合成プログラム

[図5]



201 input section
202 A/D conversion section
205 output section
206 D/A conversion section
207 storage section for articulatory feature calculation
2071 coefficient set
210 articulatory feature extraction section
220 voice recognition section
225 voice unit articulatory movement storage section
2251 articulatory movement model coefficient set
230 optimum articulatory characteristic series → voice synthesis parameter conversion section
235 storage section for voice synthesis
2351 voice synthesis parameter set
2352 drive sound source codebook
240 voice synthesis section
270 text → voice symbol converter
280 voice symbol → articulatory feature converter

(57) **Abstract**: Disclosed are a voice synthesis apparatus, voice synthesis method and voice synthesis program capable of implementing voice synthesis of a specified individual with high quality using few items of learned voice data. The voice synthesis apparatus learns a transition model (225) of articulatory movement stored for each of fixed voice units such as phonemes, from an unspecified large number of speakers. The voice synthesis apparatus is provided with means (230) for converting to voice synthesis parameters that carry vocal tract shape information whereby a series of articulatory features is adapted to individuals and an optimum voice unit series is obtained at the same time by comparing this model with the input voice. In addition, the voice synthesis apparatus obtains high-quality synthesised voice for a specified individual by registering sound source code in a state transition model of articulatory movement using closed loop learning employing a drive sound source codebook.

(57) 要約:

[続葉有]

WO 2010/104040 A1

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

【課題】 少ない学習音声データで、高品質な特定個人の合成音声を実現できる音声合成装置、音声合成方法および音声合成プログラムを提供する。 【解決手段】 音声合成装置では、音素など一定の音声単位毎に記憶された調音運動の遷移モデル２２５を不特定多数の話者から学習しておき、このモデルと入力音声を比較して、最適音声単位系列を得ると同時に、調音特徴系列を個人に特化した声道形状情報を担う音声合成パラメータに変換する手段２３０を設け、さらに駆動音源符号帳による閉ループ学習を使用して、音源符号を調音運動の状態遷移モデルに登録することで、特定個人の高品質合成音声を得る。

明　細　書

発明の名称 ：

１モデル音声認識合成に基づく音声合成装置、音声合成方法および音声合成プログラム

技術分野

[0001] 本発明は、１モデル音声認識合成に基づく音声合成装置、１モデル音声認識合成に基づく音声合成方法および１モデル音声認識合成に基づく音声合成プログラムに関する。より詳細には、音声発話から調音特徴を抽出し、音声認識に供することのできる調音運動に係る状態遷移モデルを構築するとともに、同じ調音運動の状態遷移モデルを用いて音声を合成する１モデル音声認識合成に基づく音声合成装置、音声合成方法および音声合成プログラムに関する。なお、１モデルとは、音声認識と音声合成の双方に共通の（すなわち１つの）状態遷移モデルを使用することを意味する。

背景技術

[0002] 音声入出力を用いたユーザインタフェースとして音声認識技術と音声合成技術の二つが知られている。音声認識技術では、周波数スペクトルなどの特徴分析処理結果をもとに、音素・音節・単語などを認識単位とするパターン認識処理を行うことが一般に行われてきた。これは、人間の聴覚神経系がスペクトル分析能力を持ち、スペクトル時系列に対して大脳で高次言語処理が行われるという推測に基づいている。これまでに開発された音声認識装置は、スペクトル時系列からなる音響特徴を基に単語もしくは単語列の分類を行うものであった。

[0003] 次に音声合成技術では、主に波形接続方式とボコーダ方式が利用されている。波形接続方式は、音素等を単位とする波形素片を基にこれらを接続して音声を生成する。またボコーダ方式は、人間の音声生成における調音運動を模擬した方式であり、発声器官の動作情報と声帯振動などの音源情報を分離して利用する。具体的には、音声から発声器官の動きすなわち調音運動を反

映するパラメータをＰＡＲＣＯＲ分析等により抽出し、これらのスペクトル包絡情報からなる素片を接続するとともに、励振源にピッチパルスもしくは雑音系列を加えて音声を生成する。

[0004] このように、現在の音声認識および音声合成は異なる二つのシステムとして実現されている。これに対して近年の脳研究から、人間は音響信号としての音声ではなく、調音運動としての音声を知覚しているとする仮説が有力視されつつある（非特許文献１参照）。

[0005] 人間の脳における音声言語の処理に関しては、まず発話の際に調音器官の筋肉の動きを支配するブローカ野が深く関わることが１８６１年にフランスのＰ．Ｐ．Ｂｒｏｃａによって発見された。この部分が損傷すると、発話の流暢性が失われるブローカ失語（運動失語）が観測されるため、主に音声生成システムを担うと考えられた。続いて、発話内容の理解に関わるウェルニッケ野が、１８８４年にドイツのＣ．Ｗｅｒｎｉｃｋｅによって発見された。この部分の疾患では、流暢ではあるが誤りだらけの文を発話するウェルニッケ失語（感覚失語）が観測されるため、主に音声理解システムに関わる部位と考えられた。このように人間の場合には、発話器官と聴覚器官の二つが存在し、さらに上記したように二つの脳部位の異なる働きが観測されたこともあり、２－ｓｙｓｔｅｍ説が優勢とされた。先に説明した音声合成におけるボコーダも、１９２８年にＨ．Ｄｕｄｌｅｙが最初に装置化した際には、脳からの調音指令を図に示し、発声器官の動きを帯域フィルター群で抽出し、同時に音源を抽出して伝送する装置を真空管回路で実現している。このボコーダの考えは、その後、１９６９年にＦ．ＩｔａｋｕｒａとＢ．Ａｔａｌによって線形予測符号化（Ｌｉｎｅａｒ Ｐｒｅｄｉｃｔｉｖｅ Ｃｏｄｉｎｇ：ＬＰＣ）として完成され、現在の音声通信の基礎となっている。

[0006] その後、１９７６年にＨ．ＭｃＧｕｒｋによりマクガーク効果が発見された。これは、例えば画面上に／ｇａ／と発話している映像を表示し、同時にスピーカから／ｂａ／という音声を呈示すると、／ｄａ／もしくは／ｇａ／と判断したという実験で、人間の音声発話と理解が脳では調音運動を担う１

－ｓｙｓｔｅｍによって処理されているという説を支持するものであった。人間の音声生成と理解は１－ｓｙｓｔｅｍか２－ｓｙｓｔｅｍかという論争は、その後も長く続いたが、近年になってｆＭＲＩ等により脳研究が大きく進展し、現在までの知見によると、音声の発話と理解にはブローカ野とウェルニッケ野の連携を含む大域的な処理機構が関係しているとされ、１－ｓｙｓｔｅｍ説が優勢になっている。近年は、調音運動に関する指令を正確に抽出する研究が音声認識の分野で盛んな一方、調音指令からの音声合成に関してｆＭＲＩ等による観測が行われている段階である。

[0007] このように、１－ｓｙｓｔｅｍ説が有力になりつつあるが、こうしたシステムを実用化する上で障害が多々ある。実現に最も近いシステムとして、隠れマルコフモデル（Ｈｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ；以下、ＨＭＭと記述する場合がある）合成がある（非特許文献２参照）。

[0008] この方式は、音声認識で現在標準的に用いられているＨＭＭを応用するもので、システムの動作を図１に示す。図に記載のないＨＭＭの学習部は、スペクトルパラメータ列（ここではメルケプストラム（Ｍｅｌ　Ｆｒｅｑｕｅｎｃｙ　Ｃｅｐｓｔｒｕｍ　Ｃｏｅｆｆｉｃｉｅｎｔ；以下、ＭＦＣＣと記述する場合がある）を使用）およびピッチパラメータを多空間上の確立分布に基づいたＨＭＭによってＢａｕｍ－Ｗｅｌｃｈアルゴリズムを用いて学習する。その際、特定話者のスペクトラム列を表現したＨＭＭ１０１に対して、これを連続学習する際に得られるトレリスなどから状態継続長分布を構成する。合成部では、テキストが入力され、テキスト解析によって韻律情報を付与した後、状態継続長分布を元にＨＭＭの各状態を連続し、得られるスペクトルおよびピッチから生成される励振波形をＭＬＳＡ（Ｍｅｌ　Ｌｏｇ：メル対数）合成フィルタ１０２に通して合成音声波形を得る。

[0009] 一方、人間は幼児の時から、親の音声波形という極少ない人間の声のみを聴取することで、その他、不特定多数の人間の音声を聞き取ることができる。この事実は、人間の脳が音声を調音運動という不変的な特徴パターンに変換して聴いていることを示唆する。

先行技術文献

非特許文献

[0010] 非特許文献1：柏野牧夫、音声知覚の運動理論をめぐって、日本音響学会誌、Ｖｏｌ．６２，Ｎｏ．５，ｐｐ．３９１－３９６（２００６年（平成１８年））

非特許文献2：徳田恵一、隠れマルコフモデルの音声合成への応用、電子情報通信学会技術研究報告、ＳＰ９９－６１，Ｎｏ．２５５，ｐｐ．４７－５４（２００８年（平成２０年））

非特許文献3：福田隆、新田恒雄、“OrthogonalizedDistinctive Phonetic Feature Extraction for Noise-robust Automatic Speech Recognition”、電子情報通信学会英文論文誌、Ｖｏｌ．Ｅ８７－Ｄ，Ｎｏ．５，ｐｐ．１１１０－１１１８（２００４年（平成１６年）

非特許文献4：M.R.Schroeder、B.S.Atal、Code-Excited Linear Prediction(CELP)：High-quality speech at very low bit rates、Proc.ICASSP’85,２５－１－１，ｐｐ．９３７－９４０（１９８５）

非特許文献5：F.J.Charpentier、M.G.Stella、“Diphone synthesis using an overlap-add Technique for speech waveforms concatenation”、Proc.IEEE-ICASSP’83,ｐｐ．１３２８－１３１１（１９８６）

非特許文献6：板橋秀一編、音声工学、森北出版（１９７３年（平成４８年））ｐｐ．６－１０（２．１．１．音声・音素・音節（表２．２　日本語の弁別素性）

非特許文献7：坂和正敏、田中雅博、ニューロコンピューティング入門、森北出版（１９９７年（平成９年））

発明の概要

発明が解決しようとする課題

[0011] 上記非特許文献２に開示される方式は、特定話者の音声スペクトル情報から作成した特定話者ＨＭＭで合成部を構成するため、高品質音声を実現する

には、特定話者の多大な音声データを必要とするという欠点がある。また、このＨＭＭを音声認識で利用する場合、特定話者の音声で設計したＨＭＭのため、その話者以外の多数話者に対して低い音声認識結果しか得られないものであった。

[0012] 本発明は、上記の問題点を解消するためになされたものであり、不特定話者に対する高い音声認識性能と特定個人に対する明瞭な音声合成という、これまでの方式では相反する機能を実現する１モデル音声認識合成に基づく音声合成装置、音声合成方法および音声合成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

[0013] 上述の問題点を解決するために、請求項１に係る発明の音声合成装置では、一定の音声単位毎に記憶された調音運動の状態遷移モデルを予め記憶する音素単位調音運動記憶部と、前記状態遷移モデルを参照しつつ音声認識を行う音声認識部と、前記状態遷移モデルから最適調音系列を取得しつつ音声合成を行う音声合成部とを備えた１モデル音声認識合成に基づく音声合成装置であって、音声認識部は、音声を取得する音声取得手段と、前記音声取得手段にて取得された音声の調音特徴を抽出する調音特徴抽出手段と、前記調音特徴抽出手段にて抽出された調音特徴を記憶手段に記憶する第１の記憶制御手段と、前記調音特徴の記憶手段から読み出された調音特徴時系列データと前記状態遷移モデルとを比較し最適音声単位系列を識別する最適音声単位系列識別手段を含み、音声合成部は、前記最適音声単位系列から調音運動に関する最適状態系列を推定し調音特徴系列を生成する最適調音特徴系列生成手段と、前記最適調音特徴系列生成手段にて生成された最適調音特徴系列データを記憶手段に記憶する第２の記憶制御手段と、前記最適調音特徴系列データの記憶手段から読み出された調音特徴系列データを音声合成パラメータ系列に変換する音声合成パラメータ系列変換手段と、前記音声合成パラメータ系列変換手段にて変換された音声合成パラメータ系列を記憶手段に記憶する第３の記憶制御手段と、前記音声合成パラメータ系列の記憶手段から読み出

された音声合成パラメータと駆動音源信号から音声を合成する手段とを含むことを特徴としている。

[0014] また、請求項２に係る発明の音声合成装置では、前記音素単位調音運動記憶部は、調音運動を表現した隠れマルコフモデル（ＨＭＭ）の係数セットが記憶され、前記音声認識部の最適音声単位系列識別手段および前記音声合成部の最適調音特徴系列生成手段から参照可能であることを特徴としている。

[0015] また、請求項３に係る発明の音声合成装置では、前記調音特徴抽出手段は、音声のデジタル信号をフーリエ分析する分析フィルタと、時間軸微分特徴抽出部および周波数軸微分特徴抽出部を有する局所特徴抽出部と、多層ニューラルネットワークを一段または複数段に構成された弁別的音素特徴抽出部とを備えたことを特徴としている。

[0016] また、請求項４に係る音声合成装置では、前記状態遷移モデルが、多数話者音声を用いて作成されるとともに、前記調音特徴系列データを音声合成パラメータ系列に変換する手段を、特定話者の音声のみ、もしくは不特定話者で作成した前記調音特徴系列データを音声合成パラメータ系列に変換する手段を、特定話者の音声で適応学習して作成されること
を特徴としている。

[0017] また、請求項５に係る発明の音声合成装置では、前記音声合成パラメータと駆動音源信号から音声を合成する手段において、駆動音源符号帳を設けるとともに、音声合成パラメータと駆動音源符号から合成された音声を元の学習音声と比較して最適な駆動音源を選択する手段と、前記選択された駆動音源符号を対応する調音運動の状態遷移モデルに登録する手段を備えたことを特徴としている。

[0018] 請求項６に係る発明の音声合成方法では、一定の音声単位毎に記憶された調音運動の状態遷移モデルを予め記憶する音素単位調音運動記憶部と、前記状態遷移モデルを参照しつつ音声認識を行う音声認識部と、前記状態遷移モデルから最適調音系列を取得しつつ音声合成を行う音声合成部とを備えた１モデル音声認識合成に基づく音声合成方法であって、

音声認識部は、音声を取得する音声取得ステップと、前記音声取得ステップにて取得された音声の調音特徴を抽出する調音特徴抽出ステップと、前記調音特徴抽出ステップにて抽出された調音特徴を記憶手段に記憶する第１の記憶制御ステップと、前記調音特徴の記憶手段から読み出された調音特徴時系列データと前記状態遷移モデルとを比較し最適音声単位系列を識別する最適音声単位系列識別ステップを含み、音声合成部は、前記最適音声単位系列から調音運動に関する最適状態系列を推定し調音特徴系列を生成する最適調音特徴系列生成ステップと、前記最適調音特徴系列生成ステップにて生成された最適調音特徴系列データを記憶手段に記憶する第２の記憶制御ステップと、前記最適調音特徴系列データの記憶手段から読み出された調音特徴系列データを音声合成パラメータ系列に変換する音声合成パラメータ系列変換ステップと、前記音声合成パラメータ系列変換ステップにて変換された音声合成パラメータ系列を記憶手段に記憶する第３の記憶制御ステップと、前記音声合成パラメータ系列の記憶手段から読み出された音声合成パラメータと駆動音源信号から音声を合成するステップとを含むことを特徴としている。

[0019] また、請求項７に係る発明の音声合成方法では、前記音素単位調音運動記憶部は、調音運動を表現した隠れマルコフモデル（ＨＭＭ）の係数セットが記憶され、前記音声認識部の最適音声単位系列識別ステップおよび前記音声合成部の最適調音特徴系列生成ステップにおいて参照可能であることを特徴としている。

[0020] また、請求項８に係る発明の音声合成方法では、前記調音特徴抽出ステップは、音声のデジタル信号をフーリエ分析する分析フィルタと、時間軸微分特徴抽出ステップおよび周波数軸微分特徴抽出ステップを有する局所特徴抽出ステップと、多層ニューラルネットワークにより処理される弁別的音素特徴抽出ステップとを備えたことを特徴としている。

[0021] また、請求項９に係る発明の音声合成方法では、前記状態遷移モデルが、多数話者音声を用いて作成されるとともに、前記調音特徴系列データを音声合成パラメータ系列に変換するステップを、特定話者の音声のみ、もしくは

不特定話者で作成した前記調音特徴系列データを音声合成パラメータ系列に変換する手段を、特定話者の音声で適応学習して作成されることを特徴としている。

[0022] また、請求項１０に係る発明の音声合成方法では、前記音声合成パラメータと駆動音源信号から音声を合成するステップにおいて、駆動音源符号帳を設けるとともに、音声合成パラメータと駆動音源符号から合成された音声を元の学習音声と比較して最適な駆動音源を選択するステップと、前記選択された駆動音源符号を対応する調音運動の状態遷移モデルに登録するステップを備えたことを特徴としている。

[0023] 請求項１１に係る発明の音声合成プログラムでは、請求項１ないし５のいずれかに記載の音声合成装置の各処理手段としてコンピュータを駆動させている。

[0024] また、請求項１２に係る発明の音声合成プログラムでは、請求項６ないし１０のいずれかに記載の音声合成方法の各処理ステップとしてコンピュータを駆動させている。

発明の効果

[0025] 請求項１に係る発明の音声合成装置は、従来のＨＭＭ合成装置が使用していた特定話者の「スペクトルに基づく情報」と異なり、「調音運動に基づく情報」を抽出してＨＭＭ合成装置を構成する。このため、ＨＭＭ合成の部分を調音運動という話者に対して基本的に不変なパラメータから構成するため、ＨＭＭ部分に関して個々の話者の学習音声データが不要もしくは極少量で済むという利点がある。また、音声を生成するには、調音運動を特定話者の発話器官の運動に変換する必要があるが、この部分は少量の音声データで実現できる。話者の音声は調音運動の状態遷移モデルとして不変量と見做し、特定話者の発話動作は音声合成パラメータ系列に変換されることから、両者を分離して把握することができる。このように、音声合成を、不変量と見做すことのできる発話器官への調音動作指令部分（調音運動の状態遷移モデルおよび音素単位調音運動記憶部）と、個人毎に異なる発話器官とその動作に

係わる部分（最適音声単位系列識別手段および最適調音特徴系列生成手段）に分離したことにより、個人の発話器官の特性に合わせた高品質な音声合成装置を実現することができる。

[0026] 特に、従来の音声スペクトル由来の特徴を使用する音声認識では、話者や発話時の文脈または周囲の騒音等によって、スペクトルが大きく変動してしまうため、音響的な尤度を求める際に使用するＨＭＭの設計に多くの音声データを必要としていた。※これに対し、調音特徴をＨＭＭへの入力特徴とする場合、少ない学習話者でも十分な音素認識性能を得ることができ、かつＨＭＭの混合分布数も少なくて済むという利点を有する。※

[0027] 請求項２に係る発明の音声合成装置は、音素単位調音運動記憶部に調音運動を表現したＨＭＭの係数セットが記憶されていることから、これを参照する最適音声単位系列識別手段および最適調音特徴系列生成手段では、話者に対して基本的に不変なパラメータにより音声認識処理および音声合成処理が実現される。

[0028] 請求項３に係る発明の音声合成装置は、局所特徴抽出部と弁別的音素特徴抽出部とによって調音特徴抽出部が構成されていることから、調音運動に基づく弁別特徴をＨＭＭへの入力特徴とすることができ、少ない学習話者により十分な音素認識性能を得ることができる。

[0029] 請求項４に係る発明の音声合成装置は、従来のＨＭＭ合成装置が使用していた「特定話者のスペクトルに基づく情報」ではなく、「不特定多数話者の調音運動の基づく情報」を抽出してＨＭＭ合成装置を構成するものである。これにより、上記発明の効果に加えて、ＨＭＭ合成の部分を話者に対し共通化することができ、個々の話者はＨＭＭ部分に関して学習音声データが原則不要にできるという利点がある。また、音声合成を、発話器官への調音動作指令部分と、個人毎に異なる発話器官とその動作に係わる部分に分離し、かつ前者を多数話者の調音特徴データを使用して、話者に対しより不変な調音動作指令として構成したことにより、個人の発話器官の特性に合わせた高品質音声合成と、高い音声認識性能の双方を達成することができる。

[0030] また、個人の音声に適応した合成音を少ないデータで得られることを可能にするため、高い音素認識性能の実現と相俟って、音声対話で問題となっている未知語に、人間同士が行っていると同様の対応を可能にする。すなわち、未知語が出現した際、未知語部分に対応する調音特徴系列を利用し、問い返しの確認発話を容易に合成することができる。

[0031] 請求項５に係る発明の音声合成装置は、合成音の音質に大きな影響を与える駆動音源信号に、音声通信で広く利用されているＣＥＬＰ（Ｃｏｄｅ Ｅｘｃｉｔｅｄ Ｌｉｎｅａｒ Ｐｒｅｄｉｃｔｉｏｎ）の閉ループ学習の考え方（非特許文献４参照）と、同じく波形合成に広く利用されているＰＳＯＬＡ（Ｐｉｔｃｈ Ｓｙｎｃｈｒｏｎｏｕｓ Ｏｖｅｒｌａｐ ａｎｄ Ａｄｄ）の技術（非特許文献５参照）を導入することにより、上記発明の効果に加えて、最適な駆動音源符号を選択して対応する調音運動の状態遷移モデルに登録し、これを参照しつつ音声合成することによって高品質音声を得ることができる。

[0032] 請求項６に係る発明の音声合成方法は、従来のＨＭＭ合成方法が使用していた特定話者の「スペクトルに基づく情報」と異なり、「調音運動に基づく情報」を抽出してＨＭＭ合成方法を構成する。このため、ＨＭＭ合成の部分を調音運動という話者に対して基本的に不変なパラメータから構成するため、個々の話者はＨＭＭ部分に関して学習音声データが不要もしくは極少量で済むという利点がある。また、音声を生成するには、調音運動を特定話者の発話器官の運動に変換する必要があるが、この部分は少量の音声データで実現できる。話者の音声は調音運動の状態遷移モデルとして不変量と見做し、特定話者の発話動作は音声合成パラメータ系列に変換されることから、両者を分離して把握することができる。このように、音声合成を、不変量と見做すことのできる発話器官への調音動作指令部分（調音運動の状態遷移モデルおよび音素単位調音運動記憶部）と、個人毎に異なる発話器官とその動作に係わる部分（最適音声単位系列識別ステップおよび最適調音特徴系列生成ステップ）に分離したことにより、個人の発話器官の特性に合わせた高品質な

音声合成方法を実現することができる。

[0033] 特に、従来の音声スペクトル由来の特徴を使用する音声認識では、話者や発話時の文脈または周囲の騒音等によって、スペクトルが大きく変動してしまうため、音響的な尤度を求める際に使用するＨＭＭの設計に多くの音声データを必要としていた。これに対し、調音特徴をＨＭＭへの入力特徴とする場合、少ない学習話者でも十分な音素認識性能を得ることができ、かつＨＭＭの混合分布数も少なくて済むという利点を有する。

[0034] 請求項７に係る発明の音声合成方法は、音素単位調音運動記憶部に調音運動を表現したＨＭＭの係数セットが記憶されていることから、これを参照する最適音声単位系列識別ステップおよび最適調音特徴系列生成ステップでは、話者に対して基本的に不変なパラメータにより音声認識処理および音声合成処理が実現される。

[0035] 請求項８に係る発明の音声合成方法は、局所特徴抽出ステップと弁別的音素特徴抽出ステップとによって調音特徴抽出ステップが構成されていることから、調音運動に基づく弁別特徴をＨＭＭへの入力特徴とすることができ、少ない学習話者により十分な音素認識性能を得ることができる。

[0036] 請求項９に係る発明の音声合成方法は、従来のＨＭＭ合成方法が使用していた「特定話者のスペクトルに基づく情報」ではなく、「不特定多数話者の調音運動の基づく情報」を抽出してＨＭＭ合成方法を構成するものである。これにより、上記発明の効果に加えて、ＨＭＭ合成の部分を話者に対し共通化することができ、個々の話者はＨＭＭ部分に関して学習音声データが原則不要にできるという利点がある。また、音声合成を、発話器官への調音動作指令部分と、個人毎に異なる発話器官とその動作に係わる部分に分離し、かつ前者を多数話者の調音特徴データを使用して、話者に対しより不変な調音動作指令として構成したことにより、個人の発話器官の特性に合わせた高品質音声合成と、高い音声認識性能の双方を達成することができる。

[0037] また、個人の音声に適応した合成音を少ないデータで得られることを可能にするため、高い音素認識性能の実現と相俟って、音声対話で問題となって

いる未知語に、人間同士が

行っていると同様の対応を可能にする。すなわち、未知語が出現した際、未知語部分に対応する調音特徴系列を利用し、問い返しの確認発話を容易に合成することができる。

[0038] 請求項１０に係る発明の音声合成方法は、合成音の音質に大きな影響を与える駆動音源信号に、音声通信で広く利用されているＣＥＬＰの閉ループ学習の考え方（非特許文献４参照）と、同じく波形合成に広く利用されているＰＳＯＬＡの技術（非特許文献５参照）を導入することにより、最適な駆動音源符号を選択して対応する調音運動の状態遷移モデルに登録し、これを参照しつつ音声合成することによって高品質音声を得ることができる。

[0039] 請求項１１に係る発明の音声合成プログラムは、請求項１ないし５のいずれかに記載の音声合成処理手段としてコンピュータを駆動させることが可能となるから、請求項１ないし５に係る発明の効果を奏することができる。

[0040] 請求項１２に係る発明の音声合成プログラムは、請求項６ないし１０のいずれかに記載の音声合成方法の各処理ステップとしてコンピュータを駆動させることが可能となるから、請求項６ないし１０に係る発明の効果を奏することができる。

図面の簡単な説明

[0041] [図1]特定話者のスペクトル情報に基づくＨＭＭ音声合成処理を示す模式図である。

[図2]音声合成装置の電気的構成を示す模式図である。

[図3]調音特徴を表す弁別的音素特徴の一例を示す図である。

[図4]ＭＦＣＣ特徴と調音特徴を用いた際の音素認識性能を比較した図である。

[図5]音声合成装置にて実行される音声合成処理を示す機能ブロック図である。

[図6]調音特徴抽出部の機能詳細を示すブロック図である。

[図7]弁別的音素特徴抽出部にて得られる調音特徴の一例を示す図である。

[図8]調音特徴に基づくＨＭＭ音声合成の動作を説明する図である。

[図9]音声合成で利用する駆動音源符号帳からの符号選択を説明する図である。

[図10]音声合成部で用いた音源波形を原音声の残差としての音源波形と比較した図である。

[図11]音声合成部で生成された合成音声のスペクトル包絡と原音声のスペクトル包絡を比較した図である。

[図12]音声合成部で生成された合成音声波形と原音声を比較した図である。

[図13]１モデル音声認識合成システムの構成例を示した図である。

発明を実施するための最良の形態

[0042] 以下、本明の音声合成装置および音声合成方法の実施の形態について、図面を参照して説明する。なお、これらの図面は、本発明が採用しうる技術的特徴を説明するために用いられるものであり、記載されている装置の構成、各種処理のフローなどは、特に特定的な記載がない限り、それのみに限定する趣旨ではなく、単なる説明例である。

[0043] はじめに、図２を参照し、音声合成装置１の電気的構成について説明する。図２は、音声合成装置１の電気的構成を示している。この図に示すように、音声合成装置１は、中央演算処理装置１１、入力装置１２、出力装置１３、記憶装置１４および外部記憶装置１５から構成されている。

[0044] 中央演算処理装置１１は、数値演算・制御などの処理を行うために設けられており、本実施の形態において説明する処理手順に従って演算・処理を行う。例えばＣＰＵ等が使用可能である。入力装置１２は、マイクロホンやキーボード等で構成され、利用者が発声した音声やキー入力された文字列が入力される。出力装置１３は、ディスプレイやスピーカ等で構成され、音声合成結果、あるいは音声合成結果を処理することによって得られた情報が出力される。記憶装置１４は、中央演算処理装置１１によって実行される処理手順（音声合成プログラム）や、その処理に必要な一時データが格納される。

例えば、ＲＯＭ（リード・オンリー・メモリ）やＲＡＭ（ランダム・アクセス・メモリ）が使用可能である。

[0045] また、外部記憶装置１５は、音声合成処理に使用される調音特徴系列セット、調音特徴抽出処理に使用されるニューラルネットの重み係数セット、調音特徴系列データから音声合成パラメータ系列への変換処理に使用されるニューラルネットの重み係数セット、調音運動のＨＭＭ状態遷移モデルセット、最適調音特徴系列データ、音声認識処理に必要なモデル、入力された音声のデータ、音声合成パラメータ系列データ、駆動音源用符号帳セット、解析結果データ等を記憶するために設けられている。例えば、ハードディスクドライブ（ＨＤＤ）が使用可能である。そして、これらは、互いにデータの送受信が可能なように、バス２２を介して電気的に接続されている。

[0046] なお、本発明の音声合成装置１のハードウエア構成は、図２に示す構成に限定されるものではない。従って、インターネット等の通信ネットワークと接続する通信Ｉ／Ｆを備えていても構わない。

[0047] また、本実施の形態では、音声合成装置１および音声合成プログラムは他のシステムから独立した構成を有しているが、本発明はこの構成に限定されるものではない。従って、他の装置の一部として組込まれた構成や、他のプログラムの一部として組込まれた構成とすることも可能である。また、その場合における入力は、上述の他の装置やプログラムを介して間接的に行われることになる。

[0048] 次に、外部記憶装置１５に記憶されている記憶データについて説明する。記憶データは各領域に区分されて外部記憶装置１５に記憶されており、図２に示すように、調音特徴が記憶されている調音特徴記憶領域１６、隠れマルコフモデルが記憶されている隠れマルコフモデル記憶領域１７、最適調音特徴系列が記憶されている最適調音特徴系列記憶領域１８、入力された音声が記憶される入力音声記憶領域１９、音声合成パラメータが記憶される音声合成パラメータ記憶領域２０、合成された音声が記憶される合成音声記憶領域２１、処理後のデータが記憶される処理結果記憶領域２２、各処理時に使用

される係数が記憶されている係数記憶領域２３、およびその他の領域が設けられている。

[0049] 調音特徴記憶領域１６には、音声の弁別的特徴系列が記憶されている。弁別特徴は、調音に関わる構造的な特徴を基に音素（音韻）を分類するために提案されたもので、有声性／非有声性／連続性／半母音性／破裂性／摩擦性／破擦性／舌端性／鼻音性／高舌性／低舌性／（舌の盛上る位置が）前方性／後方性／・・・；（Ｄｉｓｔｉｎｃｔｉｖｅ Ｆｅａｔｕｒｅ：ＤＦ）などがある。また、音声から弁別的特徴などの調音特徴を直接抽出する方法も、ニューラルネットワークを利用する手法など多く提案されている（非特許文献６参照）。

[0050] 隠れマルコフモデル記憶領域１７には、中央演算処理装置１１において音声認識や音声合成が行われる場合に参照される隠れマルコフモデルが記憶されている。最適調音特徴系列記憶領域１８には、中央演算処理装置１１において隠れマルコフモデルを参照して探索した結果の最適な調音特徴系列が記憶されている。入力音声記憶領域１９には、入力装置１２を介して入力された音声データが記憶される。音声合成パラメータ記憶領域２０には、中央演算処理装置１１においてニューラルネットの重み係数（係数記憶領域２３）を参照して計算された結果の音声合成パラメータが記憶されている。合成音声記憶領域２１には、中央演算処理１１において音声合成パラメータ２０と係数記憶領域２３上の駆動音源用符号帳セットを参照して計算された結果の合成音声データが記憶される。処理結果記憶領域２２には、中央演算処理装置１１において実行される各種処理の結果得られたデータが記憶される。係数記憶領域２３には、調音特徴抽出のためのニューラルネットの重み係数セット、調音特徴系列データから音声合成パラメータへの変換処理に使用されるニューラルネットの重み係数セット、および音声合成に使用される駆動音源用符号帳セットが記憶される。なお、これらのデータの詳細は後述する。

[0051] ここで、調音特徴記憶領域１６に記憶されている弁別的特徴系列に使用される弁別的音素特徴について詳述する。日本語の音素を例として、その弁別

的音素特徴（Ｄｉｓｔｉｎｃｔｉｖｅ　Ｐｈｏｎｅｍｉｃ　Ｆｅａｔｕｒｅ；以下、ＤＰＦと記述する場合がある）を図３に示す。ここで、弁別的音素特徴とは、調音特徴の表現方法の一つである。図は、縦欄が弁別的特徴を示しており、横欄が個々の音素を示している。図中（＋）は各音素についての弁別的特徴を有していることを意味し、（－）はその特徴を有しないことを意味する。なお、日本語以外の言語について弁別的音素特徴を把握する場合には、これらの弁別的特徴および音素に加えて、当該言語に特有の弁別的特徴または音素についても考慮されることとなる。

[0052] そして、この表から一つの音素を生成する際に必要な発声器官の動作を知ることができる。図３のうちｎｉｌ（高／低）は、高舌性／低舌性のどちらにも属さない音素に対して弁別特徴を割り当て、ｎｉｌ（前／後）は、（舌の盛上る位置が）前方性／後方性のどちらにも属さない音素に対して弁別特徴を割り当てるためのものであり、新たに追加した特徴であることを示す。このように、音素間のバランスをとることで、音声認識性能が向上することが知られている。

[0053] なお、調音特徴の表現としては、国際音声記号（Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｐｈｏｎｅｔｉｃ　Ａｌｐｈａｂｅｔ；以下、ＩＰＡと称する）として広く使用されている表に記載されたものを用いてもよい。このＩＰＡの表は、子音と母音の表に分かれ、子音では、調音位置および調音方法で分類されている。調音位置とは、唇、歯茎、硬口蓋、軟口蓋、声門などであり、調音方法とは破裂、摩擦、破擦、弾音、鼻音、半母音などである。また、それぞれについて有声と無声がある。例えば、／ｐ／は、子音で、無声音、唇音、破裂音に分類される。一方、母音では、舌が最も盛上る場所および舌と口蓋との空間の広さで分類されている。舌が最も盛上る場所は、前（前舌）、後（後舌）または中（中舌）に区別され、舌と口蓋との空間の広さは、狭、半狭、半広または広に区分される。例えば、／ｉ／は、前舌母音で狭母音（せまぼいん）である。ＩＰＡを使用する場合は、図３に示した弁別特徴の表と同様に、調音特徴のある個所（／ｐ／を例にとると、子音、無声音、唇音

、破裂音の個所）が＋となり、それ以外では－となる。

[0054] 従来の音声スペクトル由来の特徴を使用する音声認識では、話者や発話時の文脈、周囲

騒音等によってスペクトルが大きく変動してしまうため、音響的な尤度を求める際に使用するＨＭＭの設計に多くの音声データを必要としていた。近年のＨＭＭに基づく音声認識装置では、音声スペクトルを入力特徴として使用し、個々のベクトル要素の変動を複数の正規分布から表現する。なお、実際に多用される音声スペクトルは、音声スペクトルを聴覚特性に合わせて周波数をメル尺度化するとともに、スペクトルの対数値を離散コサイン変換（ＤＣＴ）したメルケプストラム（ＭＦＣＣ）が使用される。また、複数の正規分布は混合分布と呼ばれ、この数は前述した様々な変形に対処するため、近年では６０～７０の分布を使用するものが現れている。このように、厖大なメモリと演算が必要になった原因は、音声中に隠された変数を特定せずに、音素や単語を分類しようとした結果といえる。これに対し、調音特徴を用いると、ＨＭＭの混合数を数個程度で済ませることができる（非特許文献３参照）。

[0055] そこで、図４にＭＦＣＣを用いて音素単位のＨＭＭを学習した際の音素認識性能と、調音特徴（具体的には弁別特徴（ＤＰＦ、後述）を使用）をＨＭＭへの入力特徴とした場合の音素認識性能とを比較したグラフを示す。この図において、横軸はＨＭＭを表現する際に必要とした分布の混合数（左から１、２、４、８、１６）を示しており、混合数が増加するほど認識に必要な演算量も増加している。混合数毎に示した棒グラフは、ＨＭＭ学習に用いた男性話者の数を示し、それぞれの混合数毎に左から１名、２名、４名、８名、３３名で×印は１００名である。この時の変化を折れ線グラフで示す（破線がＭＦＣＣで、実線がＤＰＦを示す）。この図から明らかなとおり、従来法では、学習人数を増やすほど、音素認識性能も向上するが、ＨＭＭの分布混合数を増やさないと性能は飽和していくことがわかる。このように、従来のＭＦＣＣを特徴パラメータとする音声認識は、高い音素認識を達成するた

めに、多くの話者データを必要とするとともに、認識に必要とされる演算量も膨大であった。これに対し、ＤＰＦを使用した場合では、図からも明らかなとおり、少ない学習話者（１名）でも十分な音素認識性能を示しており、また、ＨＭＭの混合分布数も少なくて済むことが明らかである。音声認識では、話者の違いのほかに、騒音の重畳等があるため、これらに対してＨＭＭの混合数を上げる必要はあるものの、図示のように、少なくとも話者に対しては調音特徴が不変量であることを理解することができる。そこで、このような不変量の調音特徴を調音運動の状態遷移モデル（ＨＭＭ）として記憶させ、音声認識および音声合成において共通に参照可能にしているのである。

[0056] 次に、音声合成装置１にて実行される音声認識処理および音声合成処理について、図５～図１２を参照して説明する。図５は、音声合成装置１にて実行される音声認識および音声合成の処理を示す機能ブロック図である。この図に示すように、音声合成装置１において実行される音声認識処理および音声合成処理に必要な機能ブロックとして、入力部２０１、Ａ／Ｄ変換部２０２、調音特徴抽出部２１０、音声認識部２２０、最適調音特徴・音声合成パラメータ変換部（図では、最適調音特徴系列（右矢印）音声合成パラメータ変換部と記載している）２３０、音声合成部２４０、Ｄ／Ａ変換部２０６、出力部２０５、調音特徴計算用記憶部２０７、音素単位調音運動記憶部２２５および音声合成用記憶部２３５が設けられている。

[0057] 調音特徴計算用記憶部２０７には、音声分析のための各種係数セット２０７１、調音特徴計算のためのニューラルネット重み係数セット等が記憶されている。音素単位調音運動記憶部２２５には、調音運動を表現したＨＭＭモデルの係数セット２２５１が記憶され、ここに記憶されている係数セット２２５１は、音声認識部２２０、および、最適調音特徴系列・音声合成パラメータ変換部２３０より参照可能な状態となっている。音声合成用記憶部２３５には、最適調音特徴系列・音声合成パラメータ変換部２３０の計算結果である音声合成パラメータセット２３５１と、駆動音源符号帳２３５２が記憶されている。そして、音声合成部２４０は、音声合成パラメータ（声道形状

の変化に相当）を係数とするデジタルフィルタを構成し、駆動音源符号帳２３５２から読み出された駆動音源入力により
音声を合成する。合成音声はＤ／Ａ変換部２０６を経て、出力部２０５に送られ、スピーカから音声を送出する。

[0058] 入力部２０１は、外部から入力される音声を受け付け、アナログ電気信号に変換するために設けられている。Ａ／Ｄ変換部２０２は、入力部２０１にて受け付けられたアナログ信号をデジタル信号に変換するために設けられている。調音特徴抽出部２１０は、音声認識のために必要となる所定の特徴量を抽出するために設けられ、また、分析フィルタにより抽出された特徴量の時系列データから、調音特徴の時系列データ（以下、「調音特徴系列」という）を抽出するために設けられている。音声認識部２２０は、調音特徴抽出部２１０より得られる調音特徴系列から、音声に含まれる音素・音節・単語などを探索するために設けられている。この探索の際には、音素単位調音運動記憶部２２５の調音運動モデル係数セット２２５１が参照される。出力部２０５は、音声認識部２２０において探索された結果の音素・音節・単語（列）を出力すると同時に、後述する合成音声を出力するために設けられている。

[0059] 音声認識処理では、入力部２０１から入力された未知の音声がＡ／Ｄ変換部２０２を通して離散化され、デジタル信号に変換される。そして、変換されたデジタル信号は、調音特徴抽出部２１０に出力される。デジタル信号から調音特徴を抽出する調音特徴抽出部２１０は、図６に示すように、分析フィルタ２１１、局所特徴抽出部２１２および弁別的（音素）特徴抽出部２１３から構成されている。

[0060] 分析フィルタ２１１では、はじめに、Ａ／Ｄ変換部２０２にて変換されたデジタル信号がフーリエ分析（窓幅２４～３２ｍｓｅｃのハミング窓使用）される。次いで、２４チャンネル程度の帯域通過フィルタに通されて周波数成分が抽出される。これにより、５～１０ｍｓｅｃ間隔の音声スペクトル系列および音声パワー系列が抽出される。そして、得られた音声スペクトル系

列および音声パワー系列は、局所特徴抽出部２１２に対して出力される。

[0061] 局所特徴抽出部２１２では、時間軸微分特徴抽出部２１２１および周波数軸微分特徴抽出部２１２２により、時間軸方向および周波数方向の微分特徴が抽出される。また、図示していないが、別途音声パワー系列の時間軸微分特徴が計算される。これらの微分特徴（以下、「局所特徴」という）の抽出にあたっては、ノイズ変動などの影響を抑えるため線形回帰演算が用いられる。抽出された局所特徴は、弁別的音素特徴抽出部２１３に出力される。なお、弁別的音素特徴抽出部２１３に出力されるデータとしては、上述の局所特徴以外にも、性能は若干劣るが、音声スペクトル、あるいは音声スペクトルを直交化したケプストラム（実際には周波数軸をメル尺度化して求めるメルケプストラムが用いられる）を使用してもよい。

[0062] 弁別的音素特徴抽出部２１３では、局所特徴抽出部２１２にて抽出された局所特徴に基づき、調音特徴系列が抽出される。弁別的音素特徴抽出部２１３は、二段のニューラルネットワーク２１３１，２１３２で構成されている。

[0063] この弁別的音素特徴抽出部２１３を構成するニューラルネットワークは、図６に示されているように、初段の第一多層ニューラルネット２１３１と、次段の第二多層ニューラルネット２１３２との二段から構成される。第一多層ニューラルネット２１３１では、音声スペクトル系列および音声パワー系列より求めた局所特徴間の相関から、調音特徴系列を抽出する。また、第二多層ニューラルネット２１３２では、調音特徴系列が持つ文脈情報、すなわちフレーム間の相互依存関係から意味のある部分空間を抽出し、精度の高い調音特徴系列を求める。

[0064] 弁別的音素特徴抽出部２１３にて算出された調音特徴抽出結果の一例を図７に示す。この図は、「人工衛星」の日本語読みである「ｊｉｎｋｏｅｓｅ」という発話に対して求められた調音特徴抽出結果を示している。このように、二段のニューラルネットワーク２１３１，２１３２により抽出された調音特徴は、高い精度であることが理解される。

[0065] なお、調音特徴系列を求めるニューラルネットワークの構成は、図６にて示した二段構成のほかに、性能を犠牲にすることとなるが一段構成とすることも可能である（非特許文献３参照）。個々のニューラルネットワークは階層構造を持っており、入力層と出力層を除く隠れ層を１から２層持っている（これを多層ニューラルネットワークという）。また、出力層や隠れ層から入力層にフィードバックする構造を持ついわゆるリカレントニューラルネットワークが利用されることもある。調音特徴抽出に対する性能という点で比較すると、其々のニューラルネットワークにおいて算出された結果にそれほど大きな差はない。これらのニューラルネットワークは、非特許文献７に示される重み係数の学習を通して調音特徴抽出器として機能する（非特許文献７参照）。

[0066] また、弁別的音素特徴抽出部２１３のニューラルネットワークでの学習は、入力層に音声の局所特徴データを加え、出力層には、音声の調音特徴を教師信号として与えることで行われる。

[0067] このように、調音特徴抽出部２１０によって抽出された調音特徴系列は、音声認識部２２０に出力され、音素単位調音運動記憶部２２５の調音運動モデル係数セット２２５１を参照しつつ最適音声単位系列が得られると同時に、後述の音声合成パラメータによる音声合成に使用され、調音特徴系列を個人に特化した音声に合成される（図５参照）。

[0068] 以上が音声認識部に関する説明である。上記説明において、入力部２０１が音声合成装置にかかる発明の音声取得手段に相当し、調音特徴抽出部２１０が調音特徴抽出手段に相当する。また、音声認識部２２０が最適音声単位系列識別手段に相当し、中央演算処理装置１１が各記憶制御手段に、外部記憶装置１５が各記憶手段に相当する。そして、音素単位調音運動記憶部２２５が音素単位調音運動記憶部に相当し、これに記憶されている不特定話者の調音特徴に基づくＨＭＭが、調音運動の状態遷移モデルに相当する。さらに、これらの機能に基づいて処理されるステップは、音声合成方法にかかる発明の音声認識部における各ステップに相当する。

[0069] 次に、調音特徴に基づくＨＭＭ音声合成の動作について説明する。図５において示したように、音声合成処理では、最適調音特徴系列・音声合成パラメータ変換部２３０が、音素単位調音運動記憶部２２５に記憶されている調音運動を表現したＨＭＭモデルの係数セット２２５１を参照しつつ、音声合成パラメータを生成し、音声合成部２４０に出力する。なお、合成の対象となるデータは、入力部２０１で入力されたテキストデータ（または音声データ）が使用される。

[0070] 図８は、ＨＭＭ音声合成における最適調音特徴系列・音声合成パラメータ変換部２３０の動作説明図である。この図に示すように、不特定話者の調音特徴に基づくＨＭＭから、Ｖｉｔｅｒｂｉパス上の最適調音特徴系列が与えられると、次に時刻ｔを挟んで前後の計３フレームの調音特徴を３層ニューラルネットワークに入力し、対応するＰＡＲＣＯＲ係数を教師データとして、調音特徴系列・音声合成パラメータ（ここではＰＡＲＣＯＲ係数）変換部２３０が構成されている。

[0071] ＨＭＭは、複数の定常信号源間を状態遷移することで、非定常な時系列信号を表現する確率モデルで、音声のように様々な要因で変動する時系列の表現に適している。出力確率分布としては、多次元正規分布の重み付き和で表わされる多次元正規混合分布が用いられ
ることが多く、本実施形態も同様である。これによって、話者や前後環境に起因する複雑な変動を細かくモデル化することが可能である。

[0072] すなわち、ＨＭＭのモデルパラメータλの学習は、与えられた学習のベクトル系列Ｏに対して、観測尤度Ｐ（Ｏ｜λ）を最大にするλを求める形で数１に示すように定式化されている。

[0073] [数1]

$$\lambda_{\max} = \arg\max P(O \mid \lambda)$$

[0074] なお、このλは、ＥＭ（Ｅｘｐｅｃｔａｔｉｏｎ Ｍａｘｉｍｉｚａｔｉｏｎ）アルゴリズムに基づいて導出できる。

[0075] 音素の初期モデルは、学習用音声データに音素ラベルが付与されていれば、セグメンタルｋ－ｍｅａｎｓ法によって得ることができる。また、音素境界が与えられていない場合には、ラベルが付与された少量のデータから初期モデルを作成し、その後、音素境界の付与されていない大量の音素データを使用して連結学習を行うことができる。音声認識では、未知のベクトル系列Ｏが観測されたとき、それがどのモデルλから生成されたかを推定する（Ｐ（Ｏ｜λ））。これはベイズの判定式から求めることができる。

[0076] 次に、音声合成について説明する。音声合成の場合は、あるモデルλが最も高い確率で生成するパラメータ時系列を与える問題になる。連続出力分布型ＨＭＭλが与えられたとき、λから長さＴの出力ベクトル系列（数２参照）を生成するため、尤度最大の意味で最適な音声パラメータ列を求めると、数３に示す式を得る。

[0077] [数2]

$$O = \left[o_1', o_2', \ldots, o_T'\right]$$

ただし、’は転置行列である。

[0078]

[数3]

$$P[O|\lambda,T]=\sum_{allQ}P[Q,O|\lambda,T]$$

ただし、

$$Q=\{(q_1,i_1),(q_2,i_2),\cdots,(q_T,i_T)\}$$

は、サブステート列を表し、さらに、

$$(q_T,i_T)$$

は、ｔ番目のフレームが状態ｑ$_t$の第ｉ$_t$混合に属することを表す。

[0079] さらに、ここでは、問題を簡単化するため、混合分布サブステートに分解した上でＶｉｔｅｒｂｉパス上の確率を示すと、数４の式となり、この式において、Ｏに関して最大化する。

[0080] [数4]

$$\overline{P}[O|\lambda,T]=\max P[Q,O|\lambda,T]$$

[0081] なお、ｏ$_T$は、数５に示す静的特徴ｃ$_t$のみを考慮する場合、個々のフレームでの出力は、前後のフレームでの出力とは独立に、そのフレームに対応する分布の平均となるため、ある状態から次の状態に遷移する部分でスペクトルに不連続が生じる。

[0082] [数5]

$$c_t=[c_t(1),c_t(2),\ldots,c_t(M)]'$$

ただし、Ｍは次数である。

[0083] このような不連続を回避するために、出力パラメータに動的特徴を導入することが行われる。

[0084] 図８において図示される駆動音源は、学習音声データにより、ＨＭＭ学習を行う際、調音特徴系列と駆動音源符号のマルチストリームで作成する。この際、図９に示すように、ＣＥＬＰの符号帳選択で使用される閉ループ学習アルゴリズムを適用することで、誤差最小の（残差）素片を選択し、同時に対応する調音運動の状態に駆動音源符号を登録することにより、高音質の合成音声を得ることができる。すなわち、全ての駆動音源を合成フィルタ（ＰＡＲＣＯＲ合成フィルタ）に通して得られる音声波形を元の波形と比較し、誤差の少ない駆動音源符号を選択する。駆動音源符号帳は、学習音声データからクラスタリングにより代表素片を登録するとともに、登録符号帳を木構造化することにより、コンパクトで効率のよい符号帳を構成できる。

[0085] 以上が音声合成部に関する説明である。上記説明において、最適調音特徴系列・音声合成パラメータ変換部２３０のうち、ＨＭＭの係数セット２２５１を参照して最適調音特徴系列を取得する部分（図８参照）が、音声合成装置にかかる本発明の最適調音特徴系列生成手段に相当し、ＰＡＲＣＯＲ係数変換部が音声合成パラメータ系列変換手段に相当する。また、音声合成部（ＰＡＲＣＯＲ合成フィルタ）２４０が、音声合成パラメータと駆動音源信号から音声を合成する手段に相当する。なお、中央演算処理装置１１が各記憶制御手段に、外部記憶装置１５が各記憶手段にそれぞれ相当し、音素単位調音運動記憶部２２５が音素単位調音運動記憶部に相当し、これに記憶されている不特定話者の調音特徴に基づくＨＭＭが、調音運動の状態遷移モデルに相当する点は、音声認識装置の場合と同様である。さらに、これらの機能に基づいて処理されるステップは、音声合成方法にかかる発明の音声合成部における各ステップに相当する。

[0086] 本実施形態のように駆動音源符号帳から作成された音源波形と元の波形と

を比較した。図１０のうち（ａ）は原音声から抽出した残差の音源波形、（ｂ）は従来用いられていたパルス列と雑音から近似した音声波形、（ｃ）は本実施形態の駆動音源符号帳から作成した音源波形を示している。音源符号帳から作成した音源波形は、原音声をＰＡＲＣＯＲ分析した際の残差波形に近いことが分かる。

[0087] また、本実施形態による合成音声と原音声のＰＡＲＣＯＲ分析した際のスペクトラムを比較した。図１１のうち（ａ）は原音声のスペクトラムを示し、（ｂ）は音声から求めた調音特徴により調音特徴系列を音声合成パラメータ（ＰＡＲＣＯＲ係数列）に変換した合成音声のスペクトラムを示し、（ｃ）は、本実施形態の合成音声（ＨＭＭ／ＤＰＦ・ＰＡＲＣＯＲ分析）のスペクトラムを示す。図１１の（ａ）と（ｃ）を比較して明らかなとおり、本実施形態の合成音声は、ＨＭＭのスムージングにより、高域のスペクトルが平滑されているが、比較的少ない学習音声データによって十分に元の音声スペクトル形状を保っていることが分かる。また、（ｂ）のスペクトラムも（ｃ）に近似しており、音声認識結果を確認する際のトークバックなどにおいて、入力音声の調音特徴抽出結果を知る際に利用することができる。

[0088] さらに、合成音声波形を比較した。図１２のうち（ａ）は原音声波形、（ｂ）はパルス列と雑音から近似した音源波形を用いて合成した音声波形、（ｃ）および（ｄ）は駆動音源符号帳を用いて合成した際の音声波形である。なお、（ｃ）は特定話者の駆動音源符号帳によるものであり、（ｄ）は不特定話者の駆動音源符号帳によるものである。この図から明らかなとおり、（ｃ）と（ｄ）は元の音声に近い波形を得ている。ただし、（ｄ）は不特定多数の話者の音声から駆動音源符号帳を作成しており、特定話者の音声（調音特徴を抽出し、音声合成パラメータ変換の多層ニューラルネット学習に用いた話者）のみから作成した符号帳の場合（ｃ）と比較すれば、（ｄ）に若干の劣化が見られる。従って、特定話者にチューニングさせる処理が必要となる。そこで、多量の不特定多数の話者音声から作成した符号帳に、少量の特定話者音声を符号帳に含めて学習することで、音質を改善することができる

。また、同時に調音特徴を音声合成パラメータに変換する多層ニューラルネットについても、多量の不特定話者音声に対して、利用者となる特定話者音声を少量学習することで、変換精度を向上させることができる。

[0089] 以上の説明では、音声を取得し、調音特徴系列を抽出し、ＨＭＭの調音運動モデルから、最適調音系列を取得し、さらに音声合成パラメータに変換して、合成音声を出力した。

しかし、本発明は、こうした利用に限られるものではなく、キーボードから入力された漢字かな混じり文に対しても、通常の音声合成器が行っているように、かな系列に変換した後、音声記号を取得すれば、調音特徴としての弁別的音素特徴は、容易に分かるようにかな文字と一対一に対応しており、かな文字・調音特徴系列の変換を通して、音声を容易に合成することができる。

[0090] 図１３は、第１に、キーボードからのテキスト入力によって音声を合成する利用形態、第２に、音声から音声認識を経て認識結果のテキストをディスプレイに表示するとともに、認識結果を再合成して音声で認識する利用形態、第３に、調音特徴抽出部４０からの出力（抽出された調音特徴）を調音特徴・声道パラメータ変換部４３で変換して音声確認を行う利用形態（図のパス４７）が可能である。

[0091] 第１の利用形態では、図１３のテキスト音素変換部４６において、図示されない単語辞書を利用し、テキストを音素系列に変換する。単語辞書中には、単語表記項目毎に「読み、品詞、アクセント」が格納されており、テキストは最初に単語辞書を参照して形態素（単語）に分割され、続いて単語の読みから音素系列とアクセント位置、および文全体のイントネーションなどが決定される。音素と韻律の系列は、調音特徴・声道パラメータ変換部４３に送られ、音素単位の格納された話者共通の調音モデル４２、すなわちＨＭＭの各状態から調音特徴と音源の素片が読み出される（図８および図９参照）。続いて、調音特徴はＰＡＲＣＯＲ係数などの音道パラメータに変換され、これと駆動音源（残差信号）が音声合成部４５に送られ、合成音声に変換

される。

[0092] 第２の利用形態では、音声認識された結果のテキストを出力するとともに、キー操作されたテキストと同様に処理されることとなるから、第１の利用形態と同じく認識結果のテキスト（単語もしくは文（単語列））から、上記第１の利用形態と同じ処理過程を経て合成音声を利用者に返すことになる。

[0093] 第３の利用形態では、前記したように、調音特徴がパス４７（図１３）で示すように与えられているため、調音特徴・声道パラメータ変換部４３を経由して、声道パラメータが得られる。音声合成器に必要なもう一方の音源信号については、図示されていない残差信号計算部（音声をＰＡＲＣＯＲ分析した際の残差を計算する）で、入力音声から残差信号が抽出され、上記声道パラメータと共に音声合成部４５に送られて合成音声が得られる。この第３の利用形態では、コンピュータが利用者の音声が、正しい調音動作として抽出されたか否かを知ることができるため、利用者が音声認識処理の誤判定に関する情報を得ることができるほか、積極的な利用として発音訓練（特に外国語の発音訓練）などへ応用できるというメリットがある。

符号の説明

[0094] １　音声合成装置

１１　中央演算処理装置

１２　入力装置

１３　出力装置

１４　記憶装置

１５　外部記憶装置

２０１　入力部

２０２　Ａ／Ｄ変換部

２０５　出力部

２０６　Ｄ／Ａ変換部

２０７　調音特徴計算用記憶部

２１０　調音特徴抽出部

２１１　分析フィルタ

２１２　局所特徴抽出部

２１３　弁別的音素特徴抽出部

２２０　音声認識部

２３０　最適調音特徴系列・音声合成パラメータ変換部

２３５　音声合成用記憶部

２４０　音声合成部

請求の範囲

[請求項1] 一定の音声単位毎に記憶された調音運動の状態遷移モデルを予め記憶する音素単位調音運動記憶部と、前記状態遷移モデルを参照しつつ音声認識を行う音声認識部と、前記状態遷移モデルから最適調音系列を取得しつつ音声合成を行う音声合成部とを備えた１モデル音声認識合成に基づく音声合成装置であって、

音声認識部は、音声を取得する音声取得手段と、前記音声取得手段にて取得された音声の調音特徴を抽出する調音特徴抽出手段と、前記調音特徴抽出手段にて抽出された調音特徴を記憶手段に記憶する第１の記憶制御手段と、前記調音特徴の記憶手段から読み出された調音特徴時系列データと前記状態遷移モデルとを比較し最適音声単位系列を識別する最適音声単位系列識別手段を含み、

音声合成部は、前記最適音声単位系列から調音運動に関する最適状態系列を推定し調音特徴系列を生成する最適調音特徴系列生成手段と、前記最適調音特徴系列生成手段にて生成された最適調音特徴系列データを記憶手段に記憶する第２の記憶制御手段と、前記最適調音特徴系列データの記憶手段から読み出された調音特徴系列データを音声合成パラメータ系列に変換する音声合成パラメータ系列変換手段と、前記音声合成パラメータ系列変換手段にて変換された音声合成パラメータ系列を記憶手段に記憶する第３の記憶制御手段と、前記音声合成パラメータ系列の記憶手段から読み出された音声合成パラメータと駆動音源信号から音声を合成する手段とを含むことを特徴とする音声合成装置。

[請求項2] 前記音素単位調音運動記憶部は、調音運動を表現した隠れマルコフモデル（ＨＭＭ）の係数セットが記憶され、前記音声認識部の最適音声単位系列識別手段および前記音声合成部の最適調音特徴系列生成手段から参照可能であることを特徴とする請求項１記載の音声合成装置。

[請求項3] 前記調音特徴抽出手段は、音声のデジタル信号をフーリエ分析する分析フィルタと、時間軸微分特徴抽出部および周波数軸微分特徴抽出部を有する局所特徴抽出部と、多層ニューラルネットワークを一段または複数段に構成された弁別的音素特徴抽出部とを備えたことを特徴とする請求項１又は２に記載の音声合成装置。

[請求項4] 前記状態遷移モデルが、多数話者音声を用いて作成されるとともに、前記調音特徴系列データを音声合成パラメータ系列に変換する手段を、特定話者の音声のみ、もしくは不特定話者で作成した前記調音特徴系列データを音声合成パラメータ系列に変換する手段を、特定話者の音声で適応学習して作成されることを特徴とする請求項１ないし３のいずれかに記載の音声合成装置。

[請求項5] 前記音声合成パラメータと駆動音源信号から音声を合成する手段において、駆動音源符号帳を設けるとともに、音声合成パラメータと駆動音源符号から合成された音声を元の学習音声と比較して最適な駆動音源を選択する手段と、前記選択された駆動音源符号を対応する調音運動の状態遷移モデルに登録する手段を備えたことを特徴とする請求項１ないし４のいずれかに記載の音声合成装置。

[請求項6] 一定の音声単位毎に記憶された調音運動の状態遷移モデルを予め記憶する音素単位調音
運動記憶部と、前記状態遷移モデルを参照しつつ音声認識を行う音声認識部と、前記状態遷移モデルから最適調音系列を取得しつつ音声合成を行う音声合成部とを備えた１モデル音声認識合成に基づく音声合成方法であって、

音声認識部は、音声を取得する音声取得ステップと、前記音声取得ステップにて取得された音声の調音特徴を抽出する調音特徴抽出ステップと、前記調音特徴抽出ステップにて抽出された調音特徴を記憶手段に記憶する第１の記憶制御ステップと、前記調音特徴の記憶手段から読み出された調音特徴時系列データと前記状態遷移モデルとを比較

し最適音声単位系列を識別する最適音声単位系列識別ステップを含み、

音声合成部は、前記最適音声単位系列から調音運動に関する最適状態系列を推定し調音特徴系列を生成する最適調音特徴系列生成ステップと、前記最適調音特徴系列生成ステップにて生成された最適調音特徴系列データを記憶手段に記憶する第２の記憶制御ステップと、前記最適調音特徴系列データの記憶手段から読み出された調音特徴系列データを音声合成パラメータ系列に変換する音声合成パラメータ系列変換ステップと、前記音声合成パラメータ系列変換ステップにて変換された音声合成パラメータ系列を記憶手段に記憶する第３の記憶制御ステップと、前記音声合成パラメータ系列の記憶手段から読み出された音声合成パラメータと駆動音源信号から音声を合成するステップとを含むことを特徴とする音声合成方法。

[請求項7] 前記音素単位調音運動記憶部は、調音運動を表現した隠れマルコフモデル（ＨＭＭ）の係数セットが記憶され、前記音声認識部の最適音声単位系列識別ステップおよび前記音声合成部の最適調音特徴系列生成ステップにおいて参照可能であることを特徴とする請求項６記載の音声合成方法。

[請求項8] 前記調音特徴抽出ステップは、音声のデジタル信号をフーリエ分析する分析フィルタと、時間軸微分特徴抽出ステップおよび周波数軸微分特徴抽出ステップを有する局所特徴抽出ステップと、多層ニューラルネットワークにより処理される弁別的音素特徴抽出ステップとを備えたことを特徴とする請求項６又は７に記載の音声合成方法。

[請求項9] 前記状態遷移モデルが、多数話者音声を用いて作成されるとともに、前記調音特徴系列データを音声合成パラメータ系列に変換するステップを、特定話者の音声のみ、もしくは不特定話者で作成した前記調音特徴系列データを音声合成パラメータ系列に変換するステップを、特定話者の音声で適応学習して作成されることを特徴とする請求項６

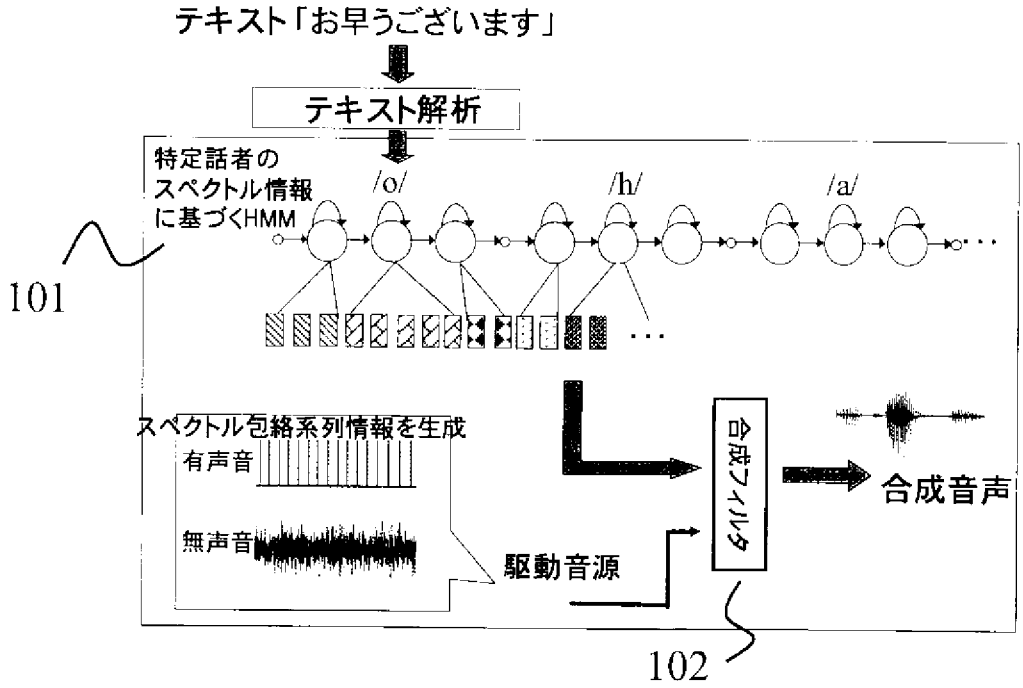
ないし８のいずれかに記載の音声合成方法。

[請求項10] 前記音声合成パラメータと駆動音源信号から音声を合成するステップにおいて、駆動音源符号帳を設けるとともに、音声合成パラメータと駆動音源符号から合成された音声を元の学習音声と比較して最適な駆動音源を選択するステップと、前記選択された駆動音源符号を対応する調音運動の状態遷移モデルに登録するステップを備えたことを特徴とする請求項６ないし９のいずれかに記載の音声合成方法。

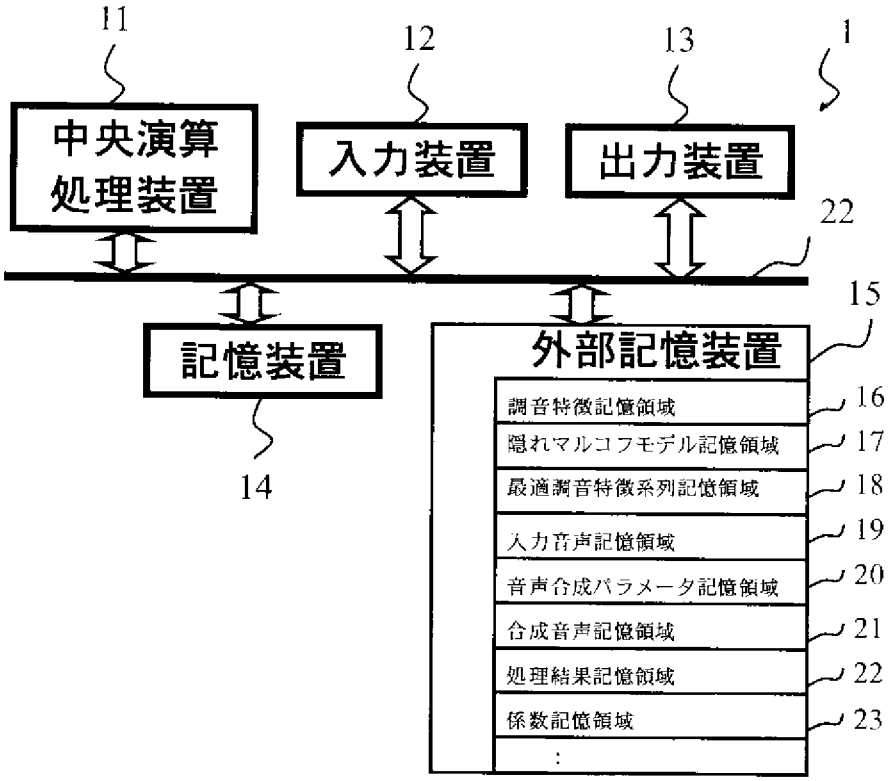
[請求項11] 請求項１ないし５のいずれかに記載の音声合成装置の各処理手段としてコンピュータを駆動させるための音声合成プログラム。

[請求項12] 請求項６ないし１０のいずれかに記載の音声合成方法の各処理ステップとしてコンピュータを駆動させるための音声合成プログラム。

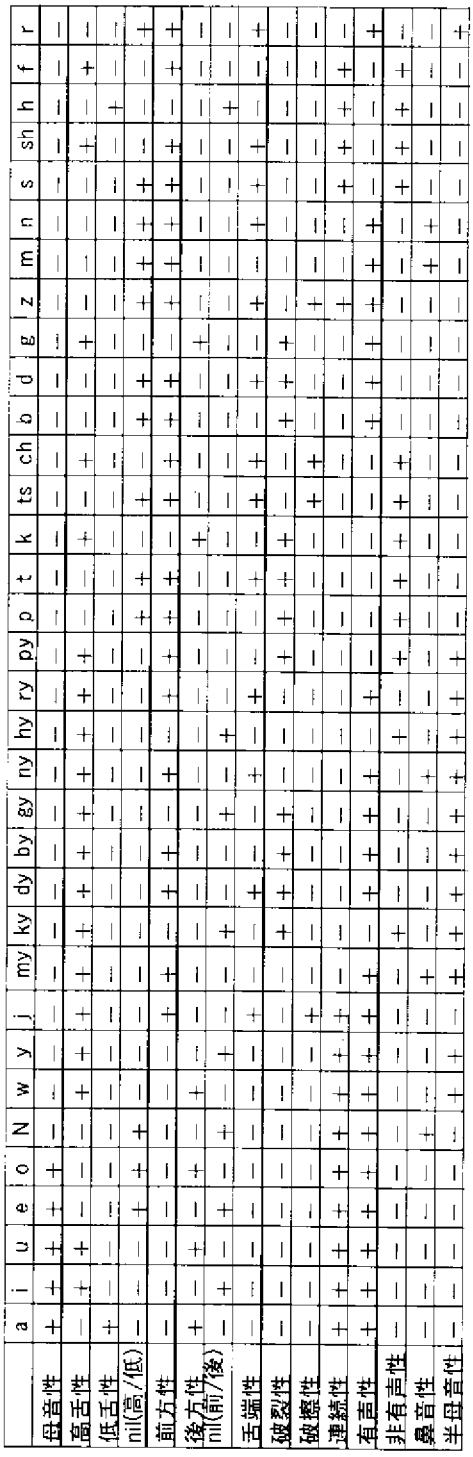
[図1]


テキスト「お早うございます」
テキスト解析
特定話者の
スペクトル情報
に基づくHMM
/o/
/h/
/a/
101
スペクトル包絡系列情報を生成
有声音
無声音
駆動音源
合成フィルタ
合成音声
102

[図2]

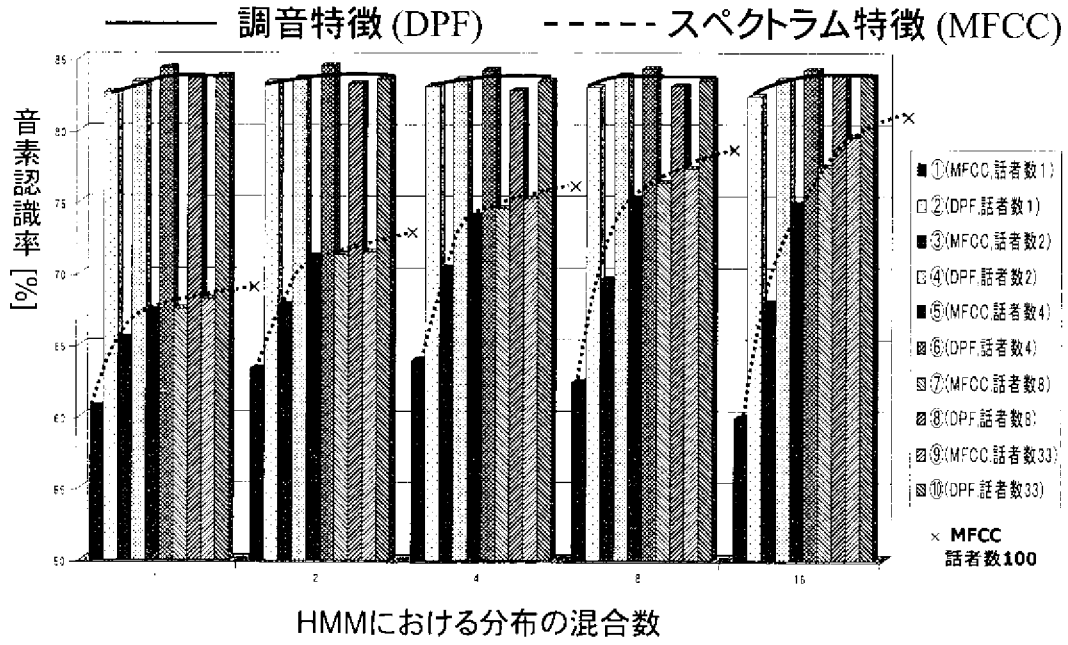

11
12
13
1
中央演算
処理装置
入力装置
出力装置
22
15
記憶装置
外部記憶装置
調音特徴記憶領域
16
隠れマルコフモデル記憶領域
17
最適調音特徴系列記憶領域
18
入力音声記憶領域
19
音声合成パラメータ記憶領域
20
合成音声記憶領域
21
処理結果記憶領域
22
係数記憶領域
23
:
14

[図3]

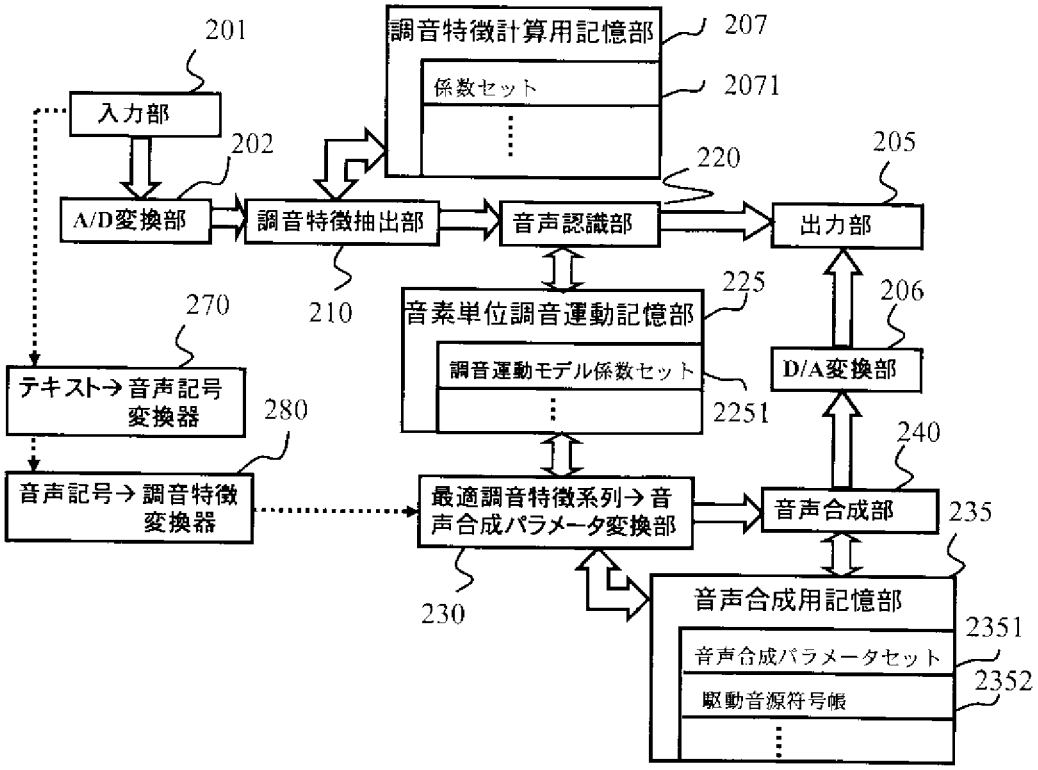


	a	i	u	e	o	N	w	y	j	my	ky	dy	by	gy	ny	hy	ry	py	p	t	k	ts	ch	b	d	g	z	m	n	s	sh	h	f	r
母音性	+	+	±	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
高舌性	−	+	+	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	−	+	+	−	−	+	−	−	−	−	−	+	−
低舌性	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
nil(高/低)	−	−	−	+	±	+	−	−	−	−	−	−	−	−	−	−	−	−	+	±	−	−	±	−	+	−	−	−	+	+	−	−	−	+
前方性	−	−	−	−	−	−	−	±	+	−	−	+	+	−	+	−	+	+	+	+	+	−	+	+	+	+	−	+	+	+	+	+	+	+
後方性 nil(前/後)	+	−	+	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−	−
	−	+	−	+	−	+	−	+	−	−	+	−	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−
舌端性	−	−	−	−	−	−	−	−	+	−	−	+	−	−	+	−	+	−	−	+	+	−	+	+	−	+	−	+	−	+	+	−	−	+
破裂性	−	−	−	−	−	−	−	−	−	+	−	+	+	+	−	−	−	−	+	+	+	+	−	−	+	+	+	−	−	−	−	−	−	−
破擦性	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	+	−	−	−	−	−	−
連続性	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−
有声性	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	−	+	−	−	−	−	−	−	−	−	+	+	+	+	+	−	+	−	+
非有声性	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	+	+	+	+	+	+	+	−	−	−	−	−	+	+	+	+	−
鼻音性	−	−	−	−	−	+	...	−	−	+	−	−	...	...	+	...	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−
半母音性	−	−	−	−	−	−	+	+	−	±	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+

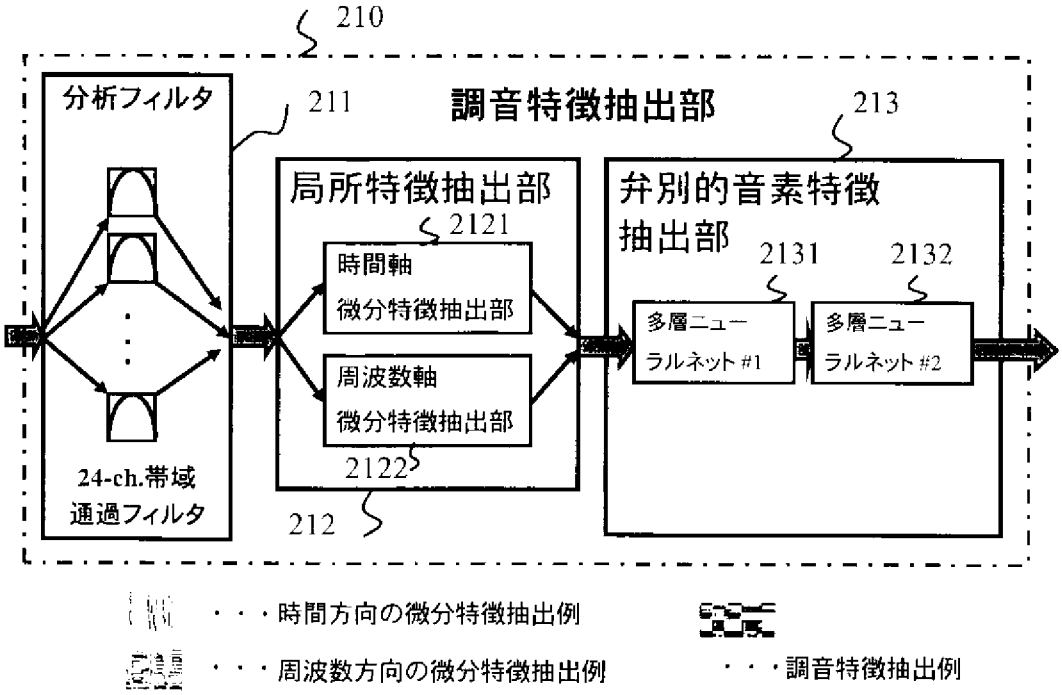
[図4]


調音特徴 (DPF)
スペクトラム特徴 (MFCC)
音素認識率 [%]
①(MFCC,話者数1)
②(DPF,話者数1)
③(MFCC,話者数2)
④(DPF,話者数2)
⑤(MFCC,話者数4)
⑥(DPF,話者数4)
⑦(MFCC,話者数8)
⑧(DPF,話者数8)
⑨(MFCC,話者数33)
⑩(DPF,話者数33)
× MFCC
話者数100
HMMにおける分布の混合数

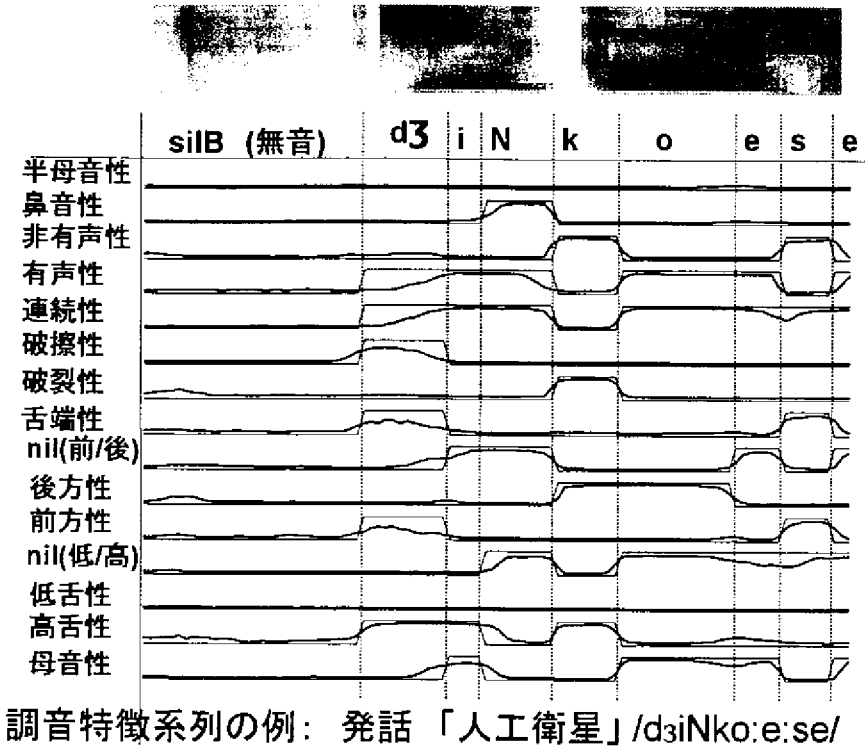
[図5]


201
入力部
202
A/D変換部
調音特徴計算用記憶部
207
係数セット
2071
調音特徴抽出部
210
220
音声認識部
205
出力部
音素単位調音運動記憶部
225
調音運動モデル係数セット
2251
206
D/A変換部
270
テキスト→音声記号変換器
280
音声記号→調音特徴変換器
最適調音特徴系列→音声合成パラメータ変換部
230
240
音声合成部
235
音声合成用記憶部
2351
音声合成パラメータセット
2352
駆動音源符号帳

[図6]

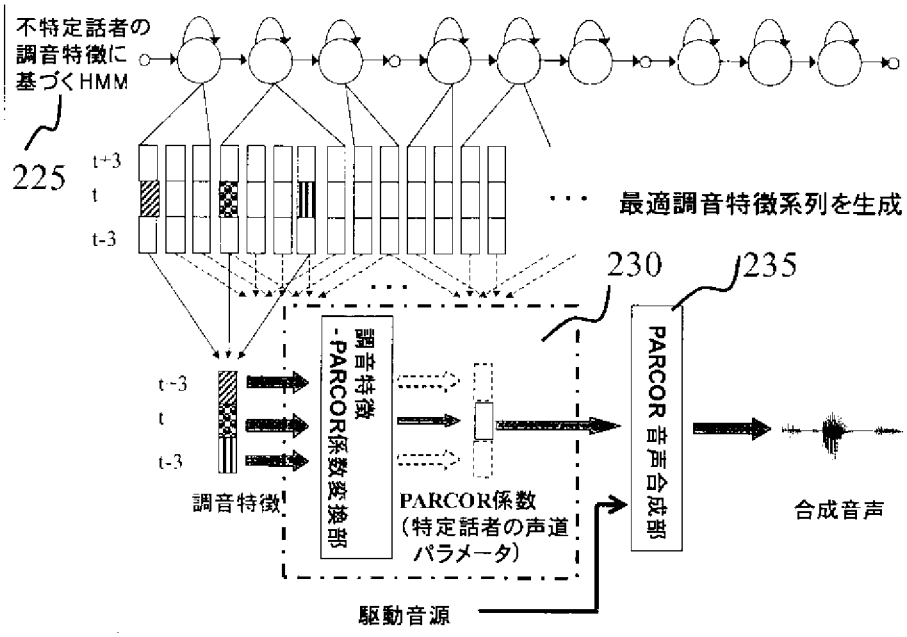


[図7]

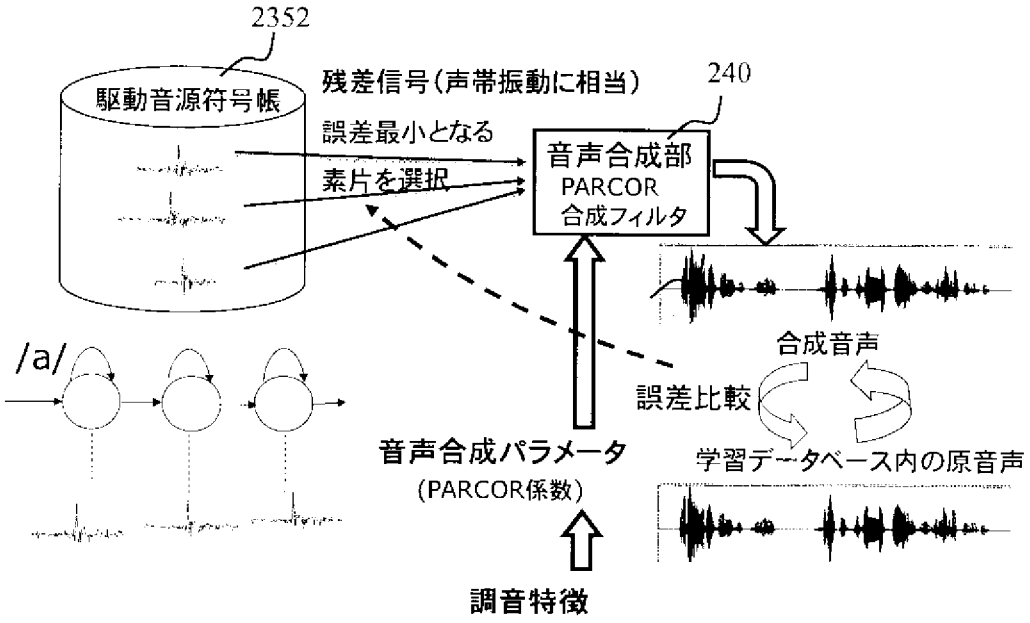


調音特徴系列の例： 発話「人工衛星」/dʒiNko:e:se/

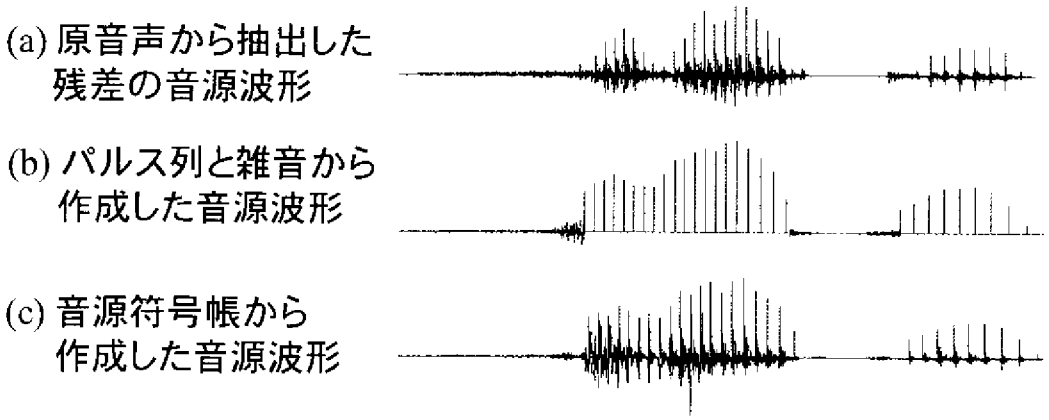
[図8]


不特定話者の
調音特徴に
基づくHMM
225
t+3
t
t-3
・・・ 最適調音特徴系列を生成
230
235
調音特徴
-PARCOR係数変換部
PARCOR 音声合成部
t-3
t
t-3
調音特徴
PARCOR係数
(特定話者の声道
パラメータ)
合成音声
駆動音源

[図9]

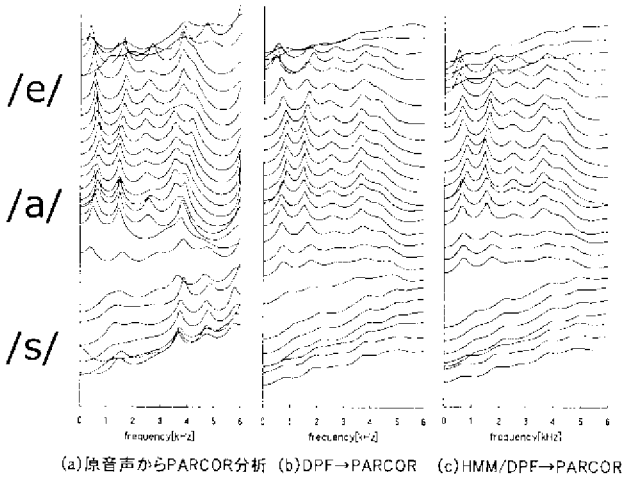

2352
駆動音源符号帳
残差信号(声帯振動に相当)
誤差最小となる
素片を選択
240
音声合成部
PARCOR
合成フィルタ
合成音声
/a/
誤差比較
音声合成パラメータ
(PARCOR係数)
学習データベース内の原音声
調音特徴

[図10]



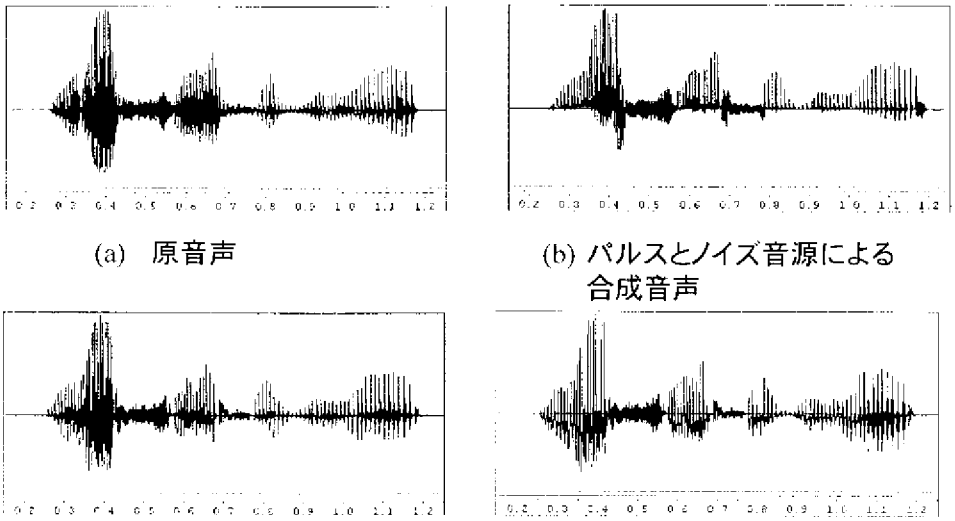
音源波形の比較

[図11]



生成されたスペクトル包絡(周波数特性)の比較:発話 /sae/

[図12]



(a) 原音声

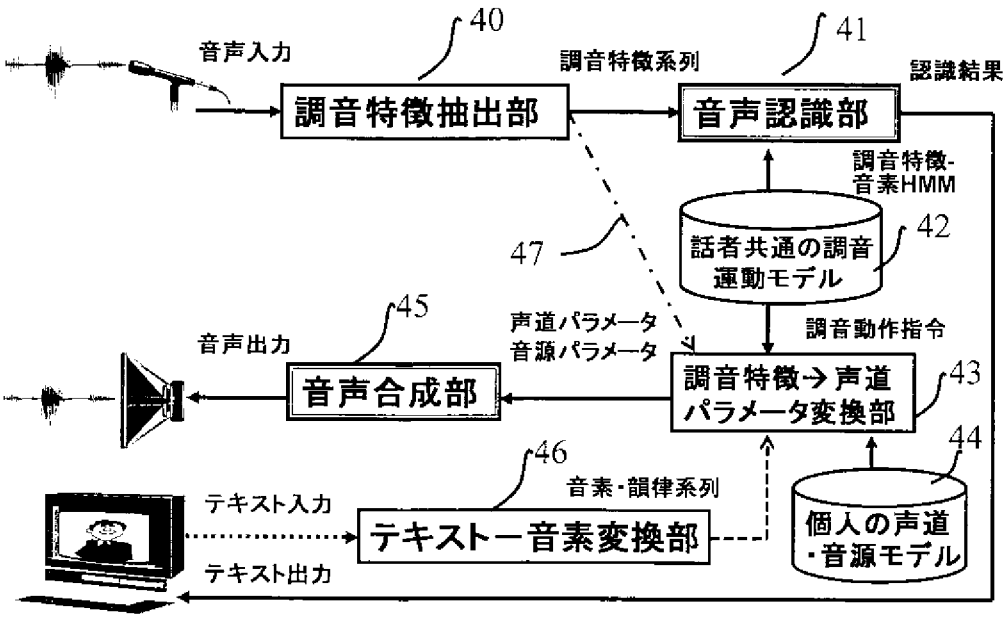
(b) パルスとノイズ音源による合成音声

(c) 特定話者の駆動音源符号帳による合成音声

(d) 不特定話者の駆動音源符号帳による合成音声

合成音声波形の比較：発話「嬉しいはずが」/ureshiihazuga/

[図13]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2010/053802

A. CLASSIFICATION OF SUBJECT MATTER
G10L13/06(2006.01)i, *G10L13/00*(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L13/00-15/28

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2010
Kokai Jitsuyo Shinan Koho 1971-2010 Toroku Jitsuyo Shinan Koho 1994-2010

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WPI, CiNii

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2004-12584 A (Nippon Telegraph And Telephone Corp.), 15 January 2004 (15.01.2004), entire text; all drawings (Family: none)	1-12
A	JP 2003-271182 A (Toshiba Corp.), 25 September 2003 (25.09.2003), entire text; all drawings & US 2003/0177005 A1	1-12
A	JP 2002-351791 A (Mitsubishi Electric Corp.), 06 December 2002 (06.12.2002), entire text; all drawings (Family: none)	1-12

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
07 June, 2010 (07.06.10)	22 June, 2010 (22.06.10)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2010/053802

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2000-66694 A (Sanyo Electric Co., Ltd.), 03 March 2000 (03.03.2000), entire text; all drawings (Family: none)	1-12
A	Keiichi TOKUDA, "Speech Syntehsis Based on Hidden Markov Models", IEICE Technical Report, 05 August 1999 (05.08.1999), vol.99, no.255, SP99-61, pages 47 to 54	1-12
A	Jun HIROI et al., "Very Low Bit Rate Speech Coding Based on HMMs", IEICE Technical Report, 11 September 1998 (11.09.1998), vol.98, no.264, SP98-63, pages 39 to 44	1-12

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１０／０５３８０２

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G10L13/06(2006.01)i, G10L13/00(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G10L13/00-15/28

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１０年
日本国実用新案登録公報	１９９６－２０１０年
日本国登録実用新案公報	１９９４－２０１０年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

WPI, CiNii

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X	JP 2004-12584 A (日本電信電話株式会社) 2004.01.15, 全文, 全図 (ファミリーなし)	１－12
A	JP 2003-271182 A (株式会社東芝) 2003.09.25, 全文, 全図 & US 2003/0177005 A1	１－12
A	JP 2002-351791 A (三菱電機株式会社) 2002.12.06, 全文, 全図 (ファミリーなし)	１－12

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０７．０６．２０１０	２２．０６．２０１０

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｚ　８９４６
山下　剛史
電話番号　０３－３５８１－１１０１　内線　３５８９

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2000-66694 A（三洋電機株式会社） 2000.03.03，全文，全図（ファミリーなし）	１－12
A	徳田恵一，隠れマルコフモデルの音声合成への応用， 電子情報通信学会技術研究報告， 1999.08.05，VoL.99，No.255，SP99-61，p.47-54	１－12
A	広井順他，ＨＭＭに基づいた極低ビットレート音声符号化， 電子情報通信学会技術研究報告， 1998.09.11，Vol.98，No.264，SP98-63，p.39-44	１－12