(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2011-186143
(P2011-186143A)
(43)公開日　平成23年9月22日(2011.9.22)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 13/08	(2006.01)	G10L 13/08	131Z	
G10L 11/00	(2006.01)	G10L 11/00	404	
G10L 13/00	(2006.01)	G10L 13/00	100K	

審査請求 未請求　請求項の数 20　OL　（全 24 頁）

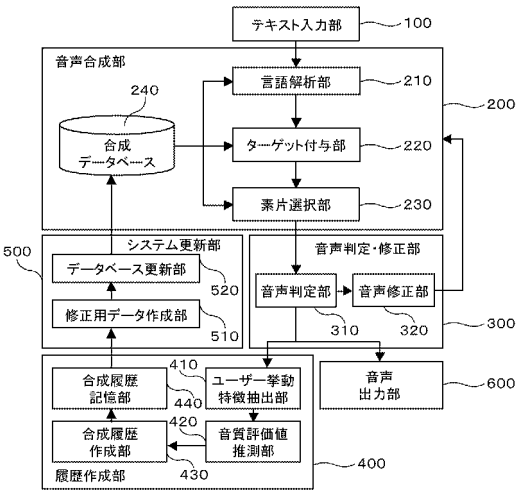
(21)出願番号　特願2010-50604(P2010-50604)
(22)出願日　平成22年3月8日(2010.3.8)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110000350
ポレール特許業務法人
(72)発明者　孫　慶華
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　永松　健司
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　藤田　雄介
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内

(54)【発明の名称】ユーザ挙動を学習する音声合成装置、音声合成方法およびそのためのプログラム

(57)【要約】
【課題】
ユーザの挙動に基づいて音声合成装置を自動修正することにより、ユーザに合わせて、ユーザに好まれる音声を合成できるようにする。
【解決手段】
音声合成装置において、音声を合成する音声合成部２００と、合成音声を提示し、提示された合成音声に対するユーザの挙動情報を出力する音声判定部３１０と、ユーザ挙動情報からユーザ挙動特徴を抽出するユーザ挙動特徴抽出部４１０と、抽出したユーザ挙動特徴から、合成音声に対するユーザの音質評価値を推測する音質評価値推測部４２０と、前記音質評価値に基づいて、前記音声合成部２００を修正する修正部５００とを備える。
【選択図】図１

図１



【特許請求の範囲】
【請求項１】
音声を合成する音声合成部と、
提示された合成音声に対するユーザの挙動からユーザ挙動特徴を抽出するユーザ挙動特徴抽出部と、
抽出したユーザ挙動特徴から、合成音声に対するユーザの音質評価値を推測する音質評価値推測部と、
前記音質評価値に基づいて、前記音声合成部を修正する修正部と
を備えることを特徴とする音声合成装置。
【請求項２】
請求項１記載の音声合成装置において、
前記ユーザ挙動特徴抽出部は、合成音声の分割されたセグメントごとに音質を評価することを特徴とする音声合成装置。
【請求項３】
請求項１記載の音声合成装置において、
前記音質評価値推測部は、専門家経験により作成した規則群、もしくは、音声主観評価実験から統計的手法で作成した規則群を用いて音質評価値を推測することを特徴とする音声合成装置。
【請求項４】
請求項１乃至３の何れか一つに記載の音声合成装置において、
前記ユーザ挙動特徴は、合成波形の拡大された場所、再生回数、行動時間、修正の有無、修正の種類、修正時間、ユーザ注目時間、集まるユーザの人数などからなることを特徴とする音声合成装置。
【請求項５】
請求項１記載の音声合成装置において、
前記音声合成部は、音声合成のための合成データベースを備え、
前記修正部は、前記合成データベースの音声合成用のデータを修正することを特徴とする音声合成装置。
【請求項６】
請求項５記載の音声合成装置において、
前記音声合成部は、言語モデルを用いてテキストを分析し、中間言語を生成する言語解析部と、ターゲット付与モデルを用いて中間言語から合成用ターゲットを作成するターゲット付与部と、合成用ターゲットを用いて、前記合成データベースから、コスト計算関数を使って、音声素片を選択し、接続する素片選択部を備えることを特徴とする音声合成装置。
【請求項７】
請求項６記載の音声合成装置において、
前記修正部は、ユーザの評価に応じて、前記音声合成部の、韻律モデルを修正する手段と、接続評価関数を修正する手段を有することを特徴とする音声合成装置。
【請求項８】
請求項５記載の音声合成装置において、
更に、合成音声波形を提示し、音声波形に対するユーザの挙動をユーザ挙動情報として出力するユーザインターフェースと、合成情報を提示し、合成音声に対するユーザの行った音声修正をユーザ挙動情報として出力するユーザインターフェースとを備えることを特徴とする音声合成装置。
【請求項９】
請求項５記載の音声合成装置において、
更に、前記音質評価値と合成情報に基づいて合成履歴を作成する合成履歴作成部を備え、
前記修正部は、合成履歴から生成した修正用データにより前記合成データベースの音声

合成用データを修正することを特徴とする音声合成装置。
【請求項１０】
請求項１記載の音声合成装置において、
前記ユーザ挙動特徴抽出部が、出力された合成音声に対するユーザの挙動を監視し、ユーザ挙動特徴を抽出するユーザ挙動監視装置であることを特徴とする音声合成装置。
【請求項１１】
請求項１記載の音声合成装置において、
前記音声合成部が、発話スタイルを指定する音声合成手法指定部と、指定されたスタイルで音声を合成する音声合成部とを備え、
前記修正部が、前記音声合成手法指定部で指定する発話スタイルを修正することを特徴とする音声合成装置。
【請求項１２】
請求項１記載の音声合成装置において、
更に、音声の提示手法を指定する音声提示手法指定部と、指定された手法で音声を提示する音声提示部とを備え、
前記修正部が、前記音声提示手法指定部で指定する音声の提示手法を修正することを特徴とする音声合成装置。
【請求項１３】
音声を合成しユーザに提示するステップと、
提示された合成音声に対するユーザの行動からユーザの挙動特徴を抽出するステップと、
抽出したユーザ挙動特徴から、合成音声に対するユーザの音質評価値を推測するステップと、
前記音質評価値に基づいて、前記音声合成を修正するステップと
を備えることを特徴とする音声合成方法。
【請求項１４】
請求項１３記載の音声合成方法において、
更に、合成音声波形を提示し、音声波形に対するユーザの挙動をユーザ挙動情報として出力するステップと、合成情報を提示し、合成音声に対するユーザの行った音声修正をユーザ挙動情報として出力するステップとを備えることを特徴とする音声合成方法。
【請求項１５】
請求項１３に記載された音声合成方法において、
前記音声を合成し提示するステップが、発話スタイルを指定するステップと、指定された発話スタイルで音声を合成するステップとを含み、
前記音声合成を修正するステップが、指定する発話スタイルを修正するものであることを特徴とする音声合成方法。
【請求項１６】
請求項１３に記載された音声合成方法において、
前記音声を合成し提示するステップが、音声の提示手法を指定するステップと、指定された音声の提示手法で音声を提示するステップとを含み、
前記音声合成を修正するステップが、指定する音声の提示手法を修正するものであることを特徴とする音声合成方法。
【請求項１７】
コンピュータに、音声を合成しユーザに提示するステップと、
提示された合成音声に対するユーザの行動からユーザの挙動特徴を抽出するステップと、
抽出したユーザ挙動特徴から、合成音声に対するユーザの音質評価値を推測するステップと、
前記音質評価値に基づいて、前記音声合成を修正するステップと
を実行させるためのプログラム。

【請求項１８】
請求項１７記載のプログラムにおいて、
更に、合成音声波形を提示し、音声波形に対するユーザの挙動をユーザ挙動情報として出力するステップと、合成情報を提示し、合成音声に対するユーザの行った音声修正をユーザ挙動情報として出力するステップとを備えることを特徴とするプログラム。
【請求項１９】
請求項１７に記載されたプログラムにおいて、
前記音声を合成し提示するステップが、発話スタイルを指定するステップと、指定された発話スタイルで音声を合成するステップとを含み、
前記音声合成を修正するステップが、指定する発話スタイルを修正するものであることを特徴とするプログラム。
【請求項２０】
請求項１７に記載されたプログラムにおいて、
前記音声を合成し提示するステップが、音声の提示手法を指定するステップと、指定された音声の提示手法で音声を提示するステップとを含み、
前記音声合成を修正するステップが、指定する音声の提示手法を修正するものであることを特徴とするプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声を合成する装置に係り、特に、ユーザに好まれる自然な音声を合成する技術に関する。
【背景技術】
【０００２】
近年、生活の多くの場面で音声合成処理された音声を耳にする機会が増えている。波形接続方式の導入などにより、音質的にもかなり改善が進んでおり、車載用ナビゲーション装置、公共施設における自動放送装置、メール読み上げ装置、自動通訳システムなど、音声を用いて自動的に情報を提供するサービスが広く普及している。
【０００３】
現在商品化されているTTS(Text-To-Speech：テキスト読み上げ)システムでは、いずれにしても、何らかの物理量を使って、合成音声の音質を客観評価し、最も音質の高い合成音声を出力する手法を用いている。しかし、従来の音質評価手法（評価関数）では、ユーザの主観評価と完全に一致させることは不可能である。従って、高品質でユーザに好まれる音声を合成するために、評価関数をユーザの主観評価に近づける必要がある。
【０００４】
すでに、いくつかのシステム学習手法が提案されている。たとえば、ユーザの修正履歴を記憶し、次回同じ（及び特徴の近い）テキストを合成するときに、自動的に修正履歴を呼び出し、適用する手法が提案された（特許文献１）。しかし、このような手法では、単なるユーザの行動パターンを記録し、それを自動的に再現することにしかならない。さらに、主観評価実験に基づいた評価関数の最適化手法も提案されている。その手法は、大量に合成した音声に対して、主観評価実験を行い、その結果を使って評価関数を最適化する手法である（非特許文献１）。しかし、このような手法は、大量な主観評価実験が必要となるため、実現するのは困難である。さらに、ユーザそれぞれの好みに合わせて、カスタマイズするのは、ほぼ不可能である。
【０００５】
人間の言語学習アルゴリズムについて、まだ完全に解明されていない部分が多いが、主に情報蓄積とルール帰納の繰り返しと考えられている。そして、自分の発話について、自分の耳で聞いて、他人の発話パターンと照合したり、誰かの指摘を受けたりして、何らかの方法でフィードバックを受け、自己学習することが可能である。さらに、人と話をするとき、相手の表情や動きなどの反応から、自分の音声に対する評価を読み取って（推測）

、それを用いて学習し、相手の好みの発話スタイルに変えることができる。よって本発明では、これらのような、人間の言語（発言、発話）学習プロセスを参考に、何らかの手法で合成音声に対するユーザの主観評価を取得し、それを用いてデータベースの拡張及び音質評価関数の再学習などを繰り返す、自動学習プロセスを提案する。

【０００６】

現行の多くのシステムは、ユーザが修正を行った場合のみに対して、音質を劣化させる原因を特定し、悪い音声を合成しないように、システムを修正する手法を用いている。しかし、場合によっては、音質劣化の原因を自動的に特定することは難しいため、システムの自動修正は困難である。本発明は、ユーザが良いと思われる音声を学習データに追加し、より良い音声を合成するものである。原因特定などを必要としないので、高精度な自動処理ができる。ここで言及した音声合成は、テキストから合成するものや中間言語から合成するもの、概念から合成するものなどいろいろ存在するほか、合成手法についても、HMM（Hidden Markov Model)を用いたパラメータ合成や単位選択手法を用いた素片接続型合成など、いろいろ存在する。本発明は、すべての音声合成システムに適用できる。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】特許第4326251号

【非特許文献】

【０００８】

【非特許文献１】T. Toda, H. Kawai, and M. Tsuzaki, “Optimizing Sub-cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis,” Proc. ICASSP, 2004.

【発明の概要】

【発明が解決しようとする課題】

【０００９】

ユーザの音声に対する好みはそれぞれであり、それに応えるために、ユーザに合わせて、合成システムをカスタマイズすることが必要となる。しかし、現在の技術では、簡単に実現することができない。

【００１０】

本発明は、ユーザに合わせてユーザに好まれる音声を合成できる音声合成装置および音声合成方法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、本発明の音声合成装置は、音声を合成する音声合成部と、提示された合成音声に対するユーザの挙動からユーザ挙動特徴を抽出するユーザ挙動特徴抽出部と、抽出したユーザ挙動特徴から、合成音声に対するユーザの音質評価値を推測する音質評価値推測部と、前記音質評価値に基づいて、前記音声合成部を修正する修正部とを備えることを特徴とする。

【００１２】

本発明の音声合成装置において、前記ユーザ挙動特徴抽出部は、合成音声の分割されたセグメントごとに音質を評価するものでよい。

また、本発明の音声合成装置において、前記音質評価値推測部は、専門家経験により作成した規則群、もしくは、音声主観評価実験から統計的手法で作成した規則群を用いて音質評価値を推測するものでよい。

また、本発明の音声合成装置において、前記ユーザ挙動特徴は、合成波形の拡大された場所、再生回数、行動時間、修正の有無、修正の種類、修正時間、ユーザ注目時間、集まるユーザの人数などからなるものでよい。

【００１３】

また、本発明の音声合成装置において、前記音声合成部は、音声合成のための合成デー

タベースを備え、前記修正部は、前記合成データベースの音声合成用のデータを修正するものでよい。

また、本発明の音声合成装置において、前記音声合成部は、言語モデルを用いてテキストを分析し、中間言語を生成する言語解析部と、ターゲット付与モデルを用いて中間言語から合成用ターゲットを作成するターゲット付与部と、合成用ターゲットを用いて、前記合成データベースから、コスト計算関数を使って、音声素片を選択し、接続する素片選択部を備えるものでよい。

また、本発明の音声合成装置において、前記修正部は、ユーザの評価に応じて、前記音声合成部の、韻律モデルを修正する手段と、接続評価関数を修正する手段を有するものでよい。

また、本発明の音声合成装置において、更に、合成音声波形を提示し、音声波形に対するユーザの挙動をユーザ挙動情報として出力するユーザインターフェースと、合成情報を提示し、合成音声に対するユーザの行った音声修正をユーザ挙動情報として出力するユーザインターフェースとを備えるものでよい。

また、本発明の音声合成装置において、更に、前記音質評価値と合成情報に基づいて合成履歴を作成する合成履歴作成部を備え、前記修正部は、合成履歴から生成した修正用データにより前記合成データベースの音声合成用データを修正するものでよい。

【００１４】

また、本発明の音声合成装置において、前記ユーザ挙動特徴抽出部が、出力された合成音声に対するユーザの挙動を監視し、ユーザ挙動特徴を抽出するユーザ挙動監視装置であるものでよい。

また、本発明の音声合成装置において、前記音声合成部が、発話スタイルを指定する音声合成手法指定部と、指定されたスタイルで音声を合成する音声合成部とを備え、前記修正部が、前記音声合成手法指定部で指定する発話スタイルを修正するものでよい。

また、本発明の音声合成装置において、更に、音声の提示手法を指定する音声提示手法指定部と、指定された手法で音声を提示する音声提示部とを備え、前記修正部が、前記音声提示手法指定部で指定する音声の提示手法を修正するものでよい。

【００１５】

本発明の音声合成方法は、音声を合成しユーザに提示するステップと、提示された合成音声に対するユーザの行動からユーザの挙動特徴を抽出するステップと、抽出したユーザ挙動特徴から、合成音声に対するユーザの音質評価値を推測するステップと、前記音質評価値に基づいて、前記音声合成を修正するステップとを備えることを特徴とするものである。

【００１６】

また、本発明の音声合成方法において、更に、合成音声波形を提示し、音声波形に対するユーザの挙動をユーザ挙動情報として出力するステップと、合成情報を提示し、合成音声に対するユーザの行った音声修正をユーザ挙動情報として出力するステップとを備えるものでよい。

また、本発明の音声合成方法において、前記音声を合成し提示するステップが、発話スタイルを指定するステップと、指定された発話スタイルで音声を合成するステップとを含み、前記音声合成を修正するステップが、指定する発話スタイルを修正するものでよい。

また、本発明の音声合成方法において、前記音声を合成し提示するステップが、音声の提示手法を指定するステップと、指定された音声の提示手法で音声を提示するステップとを含み、前記音声合成を修正するステップが、指定する音声の提示手法を修正するものでよい。

【００１７】

本発明のプログラムは、コンピュータに、上記の音声合成方法を実行させるものである。

【発明の効果】

【００１８】

本発明の音声合成装置および音声合成方法では、ユーザの挙動(反応)に基づいて、音声合成を自動修正することによって、よりユーザに好まれる音声を合成することができる。

【図面の簡単な説明】

【００１９】

【図１】本発明の第１の実施例における音声合成装置の全体構成を示すブロック図である。

【図２】本発明の第１の実施例における音声合成部の構成および動作を示す図である。

【図３】本発明の第１の実施例における音声判定・修正部の構成および動作を示す図である。

【図４】本発明の第１の実施例における履歴作成部の構成および動作を示す図である。

【図５】本発明の第１の実施例におけるシステム更新部の構成および動作を示す図である。

【図６】第１の実施例における合成した音声の波形を示す図である。

【図７】第１の実施例における音質評価値推定ルールの１例である。

【図８】第１の実施例における音質評価値推定ルール（決定木）の１例である。

【図９】本発明の第２の実施例における音声合成装置の全体構成を示すブロック図である。

【図１０】本発明の第２の実施例における音声合成装置の構成及び動作を示す図である。

【図１１】本発明の第１の実施例の音声合成装置をパソコンで構成した図である。

【図１２】本発明の第２の実施例の音声合成装置をパソコンで構成した図である。

【発明を実施するための形態】

【００２０】

以下、図面を参照して本発明の実施の形態を説明する。

【実施例１】

【００２１】

図１は、本発明の第１の実施例に係り、本発明の音声合成装置を示すブロック図である。

本実施例は、図示のとおり、テキスト入力部１００、音声合成部２００、ユーザインタフェースを持つ音声判定・修正部３００、履歴作成部４００、システム更新部５００、音声出力部６００を備えている。

【００２２】

テキスト入力部１００はテキストを得られる装置であり、キーボードでの入力やニュース配信サービスなどの手法が考えられる。

【００２３】

音声合成部２００は、言語モデルを用いて、テキスト入力部１００から入力されるテキストを分析し、形態素情報や韻律境界情報などの言語情報を含む中間言語（合成用記号列）を生成する言語解析部２１０と、ターゲット付与モデルを用いて、言語解析部から入力された中間言語から、韻律情報及びコンテキスト環境情報などを解析するターゲット付与部２２０と、ターゲット付与部から入力されたターゲット情報を用いて、素片データベースから素片波形を選択し接続する素片選択部２３０で構成されている。

【００２４】

音声判定・修正部３００では、ユーザに音声波形を提示し、ユーザの操作に応じて、波形を拡大したり、部分再生したりできるインタフェースを用いて、音声の良否を判定する音質判定部３１０と、ユーザが音質判定部で音声を修正する必要があると判定した場合に、合成音声の素片を変えたり、韻律を変えたり、読みを変えたりできるインタフェースを用いて、合成音声を修正できる音声修正部３２０で構成されている。

【００２５】

履歴作成部４００では、ユーザの挙動から特徴を抽出するユーザ挙動特徴抽出部４１０と、ユーザ挙動特徴からユーザが合成した音声の各セグメントに対する音質評価値を推測する音質評価値推測部４２０と、音質評価値および合成情報を用いて、合成履歴を生成す

る合成履歴生成部４３０と、合成履歴を保存する合成履歴記憶部４４０で構成されている。

【００２６】

システム更新部５００では、合成履歴からシステム更新情報を生成する修正用データ作成部５１０と、修正用データを用いて、システムのデータベースを修正するデータベース更新部５２０で構成されている。

【００２７】

図２を使って、音声合成部の動作について、説明する。

まず、テキスト入力部１００から得られた入力テキスト１０１が言語解析部２１０に入力される。言語解析部２１０では、事前に作成された言語モデル２４１を用いて、入力テキストを中間言語２１１に変換する。ここで、中間言語２１１に少なくとも、読みを示す発音記号が含まれる。ほかに、品詞情報、韻律境界情報、構文情報、アクセント型などを解析することが望ましい。さらに言語解析部２１０では、入力テキストに一部の中間言語情報が付けられている場合に、解析結果がそれと一致するようになる機能を備えていなければならない。つまり、解析結果の一部を指定することができる。そして、解析結果は修正可能となる言語モデルに依存する。

一例では、入力された「これは合成音声です。」は「(k%o)(r%e)/(w%a)#(g%oo)(s%ee)/(o%N)(s%ee)/(d%e)(s%u)」に変換される。ここでは、％は音素境界を示し、（）に囲まれるものはモーラであり、/は形態素境界を示し、＃はアクセント句境界を示している。

【００２８】

そして、言語解析部２１０で出力された中間言語２１１がターゲット付与部２２０に入力される。ターゲット付与部２２０では、事前に学習したターゲット付与モデル２４２を用いて、合成用ターゲット２２１を予測する。ここでの合成用ターゲット２２１とは、合成の目標となる特徴量のことである。一般的に、合成用ターゲット２２１には、基本周波数（F0）、パワー、継続長、音韻特徴（スペクトル）、コンテキスト特徴などが含まる。さらに、入力された中間言語の一部に合成用ターゲットの情報が付けられた場合、解析結果がそれと一致するようになる機能を備えていなければならない。つまり、解析結果の一部を指定することができる。そして、解析結果は修正可能となるターゲット付与モデルに依存する。

一例では、入力された中間言語「(k%o)(r%e)/(w%a)#(g%oo)(s%ee) /(o%N)(s%ee)/(d%e)(s%u)」を表１のような合成用ターゲットに変換する。

【００２９】

【表１】

第1表

合成用ターゲット

音素	韻律特徴			音韻特徴量	コンテキスト特徴
	F0開始・終了値	継続長	パワー		
k	100Hz,120Hz	20ms	50	2.5,0.7,1.8・・・	X-k-o-2-4-6-1・・・
o	120Hz,150Hz	100ms	80	0.4,0.3,0.1・・・	k-o-r-5-3-7-3・・・
r	150Hz,200Hz	50ms	75	0.6,0.5,0.9・・・	o-r-e-3-7-3-8・・・
e	200Hz,15Hz	80ms	85	1.8,0.3,0.1・・・	r-e-w-3-6-9-1・・・
・・・	・・・	・・・	・・・	・・・	・・・

【００３０】

最後に、予測された合成用ターゲットを素片選択部２３０に入力する。この素片選択部２３０では、音声素片データベース２４９から、合成音声を構成できる素片のすべての組み合わせについて、素片とターゲットとの距離（ターゲットコスト）および隣接素片間の距離（接続コスト）を計算し、総合的に距離（コスト）の小さい組み合わせを出力する。ターゲットコストは、F0ターゲットコスト、継続長ターゲットコスト、パワーターゲットコスト、音韻ターゲットコスト、コンテキストターゲットコストなどのサブコストの少なくとも一つ以上で構成される。接続コストは、F0接続コスト、パワー接続コスト、音韻接続コストなどのサブコストの少なくとも一つ以上で構成される。そして、入力された合成用ターゲットに使用素片リストおよび不使用素片リストが指定されている場合、それに応じて選択する機能を備えなければならない。さらに、解析結果が修正可能となる音声素片データベース２４９およびコスト計算関数群２４８に依存する。つまり、音声素片データベースで音素に付けられた情報を修正することや、コスト計算関数のパラメータを修正することで、素片の選択基準が変えられる。そして、得られた素片波形をつなぎ合わせ、合成情報２３１とともに合成音声２３２を音声判定・修正部３００に出力する。ここでの合成情報は、言語解析部で得られた中間言語、ターゲット付与部で得られた合成ターゲット、素片選択部で得られた素片系列および対応する各コストなどを含む。

一例では、表１のような合成用ターゲットを素片選択部に入力し、「k=435,o=32,r=436,・・・」のようなコスト素片番号系列に変換される。さらに、それらの番号に対応する音声波形をつなげ、図６のような合成音声が出力される。

【００３１】

図３を使って、音声判定・修正部３００の動作について、説明する。

まず、音声合成部から出力された合成情報３０１とともに、合成音声３０２を音声判定部３１０に入力する。音声判定部３１０では、ユーザインタフェース３１１を介して、ユーザが音声波形を選択したり、拡大したり、再生したりして、最終的に音質の良否を判定する。この過程でユーザの行ったすべて行動を記録し、ユーザ挙動情報３１５を履歴作成部４００に出力する。出力したユーザ行動には、拡大の有無、拡大された場所、各部分の再生回数、各作業の所要時間などを含む。さらに可能であれば、ユーザの心拍数変化、脳波変化、表情変化、目の動きなどの身体情報も出力すると良い。そして、判定結果が「可」となった場合、音声出力部６００から音声を出力する。一方、ユーザが「否」と判断した場合、音声修正部３２０に移り、合成音声の修正を行う。

【００３２】
ここで、入力された合成音声（発話内容は「これは音声合成システムです。」）の波形及び合成情報をユーザに提示する場合について、説明する。一例（Case1）では、最初に音声の波形と素片境界情報や韻律情報などの合成情報を表示するとともに、一回のみ音声の全体を再生する。次に、ユーザが「音声合成」の部分を選択し、この部分のみを2回再生した。最後に、「可」と判定し、音声を出力させた。この場合では、表２のようなユーザ行動を履歴作成装置４００に出力する。一例（Case2）では、最初に音声の波形と素片境界情報や韻律情報などの合成情報を表示するとともに、一回のみ音声の全体を再生する。次に、ユーザが音声合成の部分を選択し、この部分のみを１回再生した。さらに、「否」と判定し、修正を行った。この場合では、表３のような行動記録を履歴作成装置４００に出力する。

【００３３】
【表２】

第2表

合成音声	「これは合成音声です。」
行動1	「これは音声合成です。」を再生(一回)
行動2	「音声合成」を選択
行動3	「音声合成」を再生(2回)
行動4	「可」と判定

【００３４】
【表３】

第3表

合成音声	「これは合成音声です。」
行動1	「これは音声合成です。」を再生(一回)
行動2	「音声合成」を選択
行動3	「音声合成」を再生(1回)
行動4	「否」と判定

【００３５】
そして、音質判定部で「否」と判定された場合のみ、音声修正部３２０で音声修正を行う。この音声修正部３２０では、ユーザインタフェース３２１を介して、合成情報をユーザに提示し、読み修正、アクセント型修正、F0修正、継続長修正、パワー修正、接続点不連続解消修正、素片不使用指定、素片使用指定などの修正手法の、少なくとも一つが用意されている。ユーザが自分の好む傾向に修正手法を選択し、合成音声を修正する。ユーザインタフェースは、入力したユーザの声から特徴抽出して、修正手法や修正結果を自動検出する手法を用意すると、ユーザの負担が軽減される。修正された合成情報３２７に再合成用情報分析部３２８に入力するとともに、ユーザがこのステップで行ったすべての行動(ユーザ挙動情報３２６)を履歴作成部４００に出力する。ここで行動には、修正の種類、修正場所、修正前の値、修正後の値、作業所要時間、などを含む。さらに可能であれば、

ユーザの心拍数変化、脳波変化、表情変化、目の動きなどの身体情報も出力すると良い。

【００３６】

最後に、修正された合成情報３２７が再合成用情報分析部３２８に入力され、ユーザの修正に応じて、再合成情報を生成し、音声合成部２００の言語解析部２１０、ターゲット付与部２２０および素片選択部２３０のいずれかに出力する。音声修正部３２０では、読み、アクセント型、韻律境界を修正した場合、テキストに修正情報を加え、言語解析部２１０に入力する。F0、継続長、パワーなどの修正を行った場合、中間言語に修正情報を加え、ターゲット付与部２２０に出力する。「だめ素片」（この素片が不使用）、「だめ接続」（この接続が不使用）などを修正した場合、ターゲットにこれらの情報を加えて、素片選択部２３０に入力する。

【００３７】

合成音声（発話内容は「これは音声合成システムです。」）の場合(前記Case2)について、説明する。一例（Case2-1）では、ユーザが「音声」の発音を「オンセイ」から「オトコエ」に修正した場合、「これは合成音声(オトコエ)です。」のように一部の読みを指定したテキストを生成し、言語解析部に入力すると同時に表４で示すユーザ行動情報を出力する。一例（Case2-2）では、ユーザがモーラ「音」の音素「o」の開始F0を120Hzから200Hzに修正した場合、「(k%o)(r%e)/(w%a)#(g%oo)(s%ee{終了F0=200Hz}) /(o{開始F0=200Hz}%N)(s%ee)/(d%e)(s%u)」のように、中間言語に一部のF0ターゲット値を指定して、ターゲット付与部に出力するとともに、表５のようなユーザ行動情報を出力する。

【００３８】

【表４】

第4表

合成音声	「これは合成音声です。」
行動1	「音声」の発音を「オンセイ」から「オトコエ」に修正
行動2	修正所要時間3秒
・・・	・・・

【００３９】

【表５】

第5表

合成音声	「これは合成音声です。」
行動1	「こ」の「o」の開始F0を120Hzから200Hzに修正
行動2	修正所要時間10秒
・・・	・・・

【００４０】

一例（Case2-3）では、「れ」の音素「r」の素片（436番）を「不使用」と指定した場合、表７のような素片情報を指定した合成用ターゲットを生成し、単位選択部に入力するとともに、表６のようなユーザ行動情報を出力する。ここで、複数修正を行った場合、言語情報修正、合成ターゲット修正、素片情報修正の優先順位を付けて、合成システムに入力する。

【００４１】

【表６】

第6表

合成音声	「これは合成音声です。」
行動1	「れ」の「r」の素片(436番)を「不使用」と修正
行動2	修正所要時間2秒
・・・	・・・

【００４２】

【表７】

第7表

合成用ターゲット

音素	韻律特徴			音韻特徴量	コンテキスト特徴	不使用素片
	F0開始・終了値	継続長	パワー			
k	100Hz,120Hz	20ms	50	2.5,0.7,1.8・・・	X-k-o-2-4-6-1・・・	なし
o	120Hz,150Hz	100ms	80	0.4,0.3,0.1・・・	k-o-r-5-3-7-3・・・	なし
r	150Hz,200Hz	50ms	75	0.6,0.5,0.9・・・	o-r-e-3-7-3-8・・・	436
e	200Hz,15Hz	80ms	85	1.8,0.3,0.1・・・	r-e-w-3-6-9-1・・・	なし
・・・	・・・	・・・	・・・	・・・	・・・	・・・

【００４３】

この音声判定部３００と音声修正部３１０のプロセスは、合成音声が「可」と判定されるまで繰り返される。

【００４４】

図４を使って、履歴作成部４００の動作について、説明する。

まず、音声判定・修正部３００から入力されたユーザ挙動情報４０１と合成情報４０２がユーザ挙動特徴抽出部４１０に入力し、素片ごとに対応するユーザの挙動特徴４１１を抽出する。一回の合成で、複数回修正が繰り返された場合があるので、複数のユーザ行動情報とそれに対応する合成情報が存在する。中に、同じ素片が複数回に使われた場合は、すべての情報を統合する。ここで、ユーザ挙動特徴は拡大の有無、再生回数、確認所要時間、修正の有無、修正回数、修正の種類、修正場所、修正前の値、修正後の値、修正所要時間、素片使用回数などの特徴量を抽出する。可能であれば、心拍変化、脳波変化、表情変化などのユーザの身体情報も加える。

【００４５】

たとえば、前記の一例（Case1）では、表８で示したようなユーザ挙動特徴が抽出される。前記の一例（Case2-2）の後に再合成し、（Case1）になった場合(Case3)は、表９のようなユーザ挙動特徴情報が抽出される。

【００４６】

【表８】

第8表

k	素片番号:435 素片使用回数:１回、 再生回数:１回、 部分再生回数:0回、 修正:なし、 最終結果:使用 ・・・
○	・・・
・・・	・・・
○	素片番号:534 素片使用回数:1回 再生回数:3回、 部分再生回数:2回、 修正:なし、 最終結果:使用 ・・・
N	・・・
・・・	・・・

【００４７】

【表９】

第9表

k	素片番号:435 素片使用回数:2回、 1回目: 再生回数:1回、 部分再生回数:0回、 修正:なし、 2回目: 再生回数:1回、 部分再生回数:0回、 修正:なし、 最終結果:使用 ・・・
○	・・・
・・・	・・・
○	素片番号:534 素片使用回数:1回 再生回数:3回、 部分再生回数:2回、 修正:なし、 最終結果:使用 ・・・
○	素片番号:973 素片使用回数:1回 再生回数:2回、 部分再生回数:1回、 修正:あり、 修正種類:開始F0 修正前の値:120Hz 修正後の値:200Hz 修正所要時間:10秒 最終結果:不使用 ・・・
N	・・・
・・・	・・・

【００４８】

そして、ユーザ挙動特徴抽出部４１０で抽出されたユーザ挙動特徴４１１を、音質評価値推測部４２０に入力し、ユーザの合成音声に対する音質評価値４２２を推定する。この音質評価値推測部４２０では、事前に作成した音質評価値推定ルール４２１を用いて、すべての素片について、ユーザの音質評価値を推定する。この音質評価値推定ルール４２１は専門家の作成したものや学習で得られるものなどで入手できる。

【００４９】

図７は専門家で作成したルールの一例を示す。図８は主観評価実験から学習したルール（決定木）の一例を示す。

一例(Case1)では、表10のような結果となる。一例(Case3)では、表11のような結果となる。

【００５０】

【表１０】

第10表

k	素片番号:435 評価値:5.0
o	・・・
r	・・・
・・・	・・・
o	素片番号:534 評価値:4.5
N	・・・
・・・	・・・

【００５１】

【表１１】

第11表

k	素片番号:435 評価値:5.0
o	・・・
r	・・・
・・・	・・・
o	素片番号:534 評価値:4.5
o	素片番号:973 評価値:2.7
N	・・・
・・・	・・・

【００５２】

さらに、推定された音質評価値４２２を合成情報４０２とともに、合成履歴作成部４３０に入力し、合成履歴４３１を生成する。合成履歴４３１には、素片情報、素片評価、テキスト解析結果、ターゲット付与結果、コスト計算結果などが含まれている。

【００５３】

一例(Case1)では、表12のような合成履歴が生成される。一例(Case3)では、表13のような合成履歴が生成される。

【００５４】

【表１２】

第12表

k	素片番号:435 素片評価:5.0 素片韻律修正有無:なし 素片韻律コスト:・・・ 素片接続コスト:・・・ 素片コンテキスト:・・・ ・・・
o	・・・
r	・・・
・・・	・・・
o	素片番号:534 素片評価:4.5 素片韻律修正有無:なし 素片韻律コスト:・・・ 素片接続コスト:・・・ 素片コンテキスト:・・・ ・・・
N	・・・
・・・	・・・

【００５５】

【表１３】

第13表

k	素片番号:435 素片評価:5.0 素片韻律修正有無:なし 素片韻律コスト:・・・ 素片接続コスト:・・・ 素片コンテキスト:・・・ ・・・
o	・・・
r	・・・
・・・	・・・
o	素片番号:534 素片評価:4.5 素片韻律修正有無:あり 素片韻律コスト:・・・ 素片接続コスト:・・・ 素片コンテキスト:・・・ ・・・
o	素片番号:3334、 素片評価:2.7、 素片韻律修正有無:あり 修正種類:開始F0 素片韻律コスト:・・・ 素片接続コスト:・・・ 素片コンテキスト:・・・ ・・・
N	・・・
・・・	・・・

【００５６】

最後に、合成履歴記憶部４４０で合成履歴４４１を記憶装置に保存する。

【００５７】

図５を使って、システム更新部５００の動作について、説明する。

【００５８】

システム更新命令５０１を受けると、合成履歴４４１を修正用データ作成部５１０に入力し、システム更新部５００が作動する。このシステム更新命令はユーザが指定したものでもいいし、システムが自動的に出したものでも良い。

修正用データ作成部５１０では、音声合成履歴から一部および全部を用いて、更新用データを作成する。この修正用データ作成部５１０では、言語情報データ作成モジュール５１１、音韻情報データ作成モジュール５１２、パワー情報データ作成モジュール５１３、F0データ作成モジュール５１４、継続長データ作成モジュール５１５、単位選択データ作成モジュール５１６、素片情報データ作成モジュール５１７などから、少なくとも一つ以上で構成されている。これらのモジュールは、それぞれ対応している音声合成装置に用いる合成データベースにある各モデルやデータを修正するときに、必要なデータを作成する。

【００５９】
素片情報データ作成モジュール５１７では、音声判定部で「可」と判定された音声に対して、合成情報に含まれる中間言語が正解とし、言語情報データを作成し、学習データベース(合成データベース作成時に使われているデータベース)に追加する。そうすることで、言語モデル用データベースが拡大され、ユーザによく使われる単語、好む読み方などが増える。それによって作成した言語解析モデルによって、よりユーザに好まれる言語解析結果が得られる。
【００６０】
一例では、「これは合成音声です」について、ユーザが「音声」の読みを「オンセイ」から、「オトコエ」と修正し、最終出力が「コレハゴウセイオトコエデス」という読み方になったとする場合、ユーザの修正した「音声」に対して、単なる辞書から読み方を「オンセイ」から「オトコエ」にする手法が一般的である。しかし、このような手法は、今度「音声合成」の言語分析するときに、正しく分析されることが多い。提案した手法では、「これは合成音声です」の言語解析の正解は「コレハゴウセイオトコエデス」と与えて、言語モデルを学習することで、入力テキストに含まれた「音声」の読み方を文脈から推定することができる。
【００６１】
音韻情報データ作成モジュール５１２について、一例では、音質評価値がある閾値(たとえば4.7)以上の素片のみに対して、音韻情報データを作成し、学習データベース５０２に追加する。そうすると、データベースが拡張され、より高精度な音韻特徴推定ができる。もう一例では、すべての素片について、音韻情報データを作成し、音質評価とともに学習データベース５０２に追加する。ただし、学習時に音質評価値を用いることで、より評価値の高い予測結果が得られる。
【００６２】
パワー情報データ作成モジュール５１３、F0データ作成モジュール５１４、継続長データ作成モジュール５１５では、音韻情報データ作成モジュールと同じような手法でデータ作成を行う。
【００６３】
単位選択データ作成モジュール５１６では、単位選択時に使われる関数の修正用データを作成する。一例では、隣接する素片の音質評価値が両方ともある閾値（たとえば、4.7）以上であれば、その接続コストを学習データに追加し、接続コスト計算関数の修正を行う。一例では、すべての素片について、素片のコンテキスト、ターゲットコンテキストおよびその音質評価値を学習データに追加し、コンテキストコスト空間の分布を再計算する。そうすることで、その素片があるコンテキスト環境に置かれたときに、より正確に音質評価値を推測することができる。
【００６４】
素片情報データ作成モジュール５１７では、ユーザが「不使用」や「強制使用」などの情報を作成し、素片情報データベースを修正する。一例では、「r」の素片（436番）の音質が悪いため、「不使用」素片と判定された場合、そのような情報を作成し、データベース更新部５３０に渡す。
【００６５】
データベース更新部５３０では、修正用データ作成部５１０で作成したデータ（５２１～５２７）および学習データベース５０２を用いて、合成用データベース２４０の各モジュールを更新する。言語情報モデル修正モジュール５３１、音韻情報モデル修正モジュール５３２、パワーモデル修正モジュール５３３、F0モデル修正モジュール５３４、継続長モデル修正モジュール５３５、コスト計算関数モデル修正モジュール５３６、素片データベースモデル修正モジュール５３７などで構成されている。それぞれモジュールは拡張された学習データベース５０２から、各モデルおよび関数を学習し、新しい合成データベース２４０を生成する。
【００６６】

本発明の第１の実施例は、汎用のパソコンなどのコンピュータにプログラムを組み込むことにより実現することができる。図１１は、その構成を示すブロック図で、ＣＰＵ６１１およびメモリ６１２から成る制御装置６１０、プログラムを記憶した記憶装置６２０、キーボード６３１および通信装置６３２から成るテキスト入力装置６３０、スピーカ６４１および増幅回路６４２から成る音声出力装置６４０、およびこれらを接続するバスライン６５０から構成される。そして、記憶装置６２０に記憶されるプログラムは、言語解析部、ターゲット付与部、素片選択部、音質判定部、音声修正部、ユーザ挙動特徴抽出部、音質評価値推測部、合成履歴作成部、合成履歴記憶部、修正用データ作成部、データベース更新部、合成用データベース、学習用データベースなどとして動作する。

【実施例２】

【００６７】

図９は、本発明の第２の実施例に係り、本発明の音声合成装置を示すブロック図である。

本実施例は、図示のとおり、音声合成出力部７００、ユーザ挙動監視部７１０、音声評価推測部７２０、音声修正部７３０を備えている。

【００６８】

音声合成出力部７００は、テキスト及び中間言語（合成用記号列）を得られる入力部７０１と、合成音声の発話速度、声の高さ、メリハリの大きさ、イントネーション、感情、話者年齢、話者性別などの発話スタイルを指定できる音声合成手法指定部７０４と、指定したスタイルで入力したテキスト（もしくは記号列（中間言語））を音声に変換する音声合成部７０２と、音声を再生するときの大きさ、方向性、音質、タイミングなどを指定できる音声提示手法指定部７０５と、指定された提示方法で音声を再生する音声提示部７０３で構成されている。

ユーザ挙動監視部７１０では、ユーザの挙動を監視して、特徴を抽出する。カメラ・マイクなどからの入力や備え付けられたインタフェースから得られたユーザの操作などから、ユーザの挙動を得られる。

音声評価推測部７２０では、入力されたユーザ挙動特徴から、現在の音声合成装置の生成した音声（発話スタイルと提示手法を含む）に対する評価（好まれる度合いや注目される度合いなど）を推測する。

音声修正部７３０では、評価が高くなるように、音声合成部及び音声提示部を修正する装置である。

【００６９】

図１０を使って、本実施例の動作について、説明する。

まず、入力部７０１に入力されたテキスト及び中間言語７０６を音声合成部７０２に入力する。入力部では、キーボードでの入力、ニュース配信サービス及び事前に用意されているものなどから、入力が得られる。一例では、街角に設置されている音声広告案内装置について、入力の獲得はセンターから配信される及び事前に作成されるなどの手法が考えられる。

【００７０】

そして、音声合成手法指定部７０４では、音声の発話速度、声の高さ、メリハリの大きさ、イントネーション、感情、話者年齢、話者性別などの発話スタイルを指定する。ここで、初回に音声合成するとき、初期値として指定する必要がある。一例では、初期値として、「発話速度＝普通、声の高さ＝普通、メリハリ＝普通、イントネーション＝普通、感情＝平静、話者年齢＝３０、話者性別＝女性」のように、合成手法を指定する。

【００７１】

次に、音声合成部７０２では、指定された合成手法で入力されたテキスト及び中間言語７０６を合成音声７０７に変換し、音声提示部７０３に出力する。

【００７２】

さらに、音声提示手法指定部７０５では、音声を再生する時の大きさ、方向性、音質特性、タイミング、信号処理使用可否、信号処理手法などを指定する。ここで、初回に音声

合成するとき、初期値として指定する必要がある。一例では、初期値として、「大きさ＝普通、方向＝前方、音質＝低音強調、タイミング＝人が10メートル範囲内に近づいたとき、信号処理＝あり、信号処理手法＝白雑音混入」のように、音声提示手法を指定する。

【００７３】

次に、音声提示部７０３で、指定された音声提示手法で、音声をユーザに提示する（再生する）。

【００７４】

そして、音声合成出力部７００が動作している間に、ユーザ挙動監視部７１０では、ユーザ７０６の挙動（反応）を監視し、ユーザ挙動特徴７１１を抽出した後、音声評価推測部７２０に出力する。一例では、街角で設置されている音声広告案内装置について、カメラの入力した映像から、ある一定の時間内（たとえば、1時間）に、集まる人の性別、集まる人の注目した時間などを抽出することが考えられる。たとえば、「女性、10分注目；男性、3分注目；女性、15分注目；・・・」などのように出力する。

【００７５】

さらに、音声評価推測部７２０は、ユーザ挙動監視部から入力されたユーザ挙動特徴７１１から、評価推定ルール７２２を使って、ユーザが音声に対する評価値７２１を推定する。ここで、音声評価値推定ルールは、事前に専門家によって作成、統計によって学習などが考えられる。一例では、化粧品の音声広告を流れている場合、「評価値＝集まる人数×注目した時間×性別係数(男性１、女性３)」と定義した場合、「女性、10分注目；男性、3分注目；女性、15分注目」入力に対して、「評価値＝10*3+3*1+15*3＝78」のように計算される。

【００７６】

そして、推定された音声評価値７２１を音声修正部７３０に入力し、音声合成手法及び音声提示手法を修正し、修正結果７３１をそれぞれ音声合成出力部の音声合成手法指定部７０４と音声提示手法指定部７０５に入力する。音声修正部７３０は、入力された推定結果と今まで得られた修正履歴７３２と照合し、評価が上がるような方向で修正を行う。ただし、初めての修正では、ランダムに修正するか、予めに決めた修正を行う。さらに、修正結果を修正履歴７３２に追加する。

一例では、一回目の評価は78となった場合、話者を39歳女性から30歳女性に修正した。2回目の評価は82となった。年を若くしたことで、評価が上がったので、さらに20歳女性の声に変更し、評価を行う。このように、修正と評価を繰り返し、その時間帯、場所、発話内容、ユーザ好みにもっとも合った合成音声を合成できる。

【００７７】

図１２は、第２の実施例をコンピュータにプログラムを組み込んだ構成を示すブロック図で、ＣＰＵ６１１およびメモリ６１２から成る制御装置６１０、プログラムを記憶した記憶装置６２０、キーボード６３１および通信装置６３２から成るテキスト入力装置６３０、スピーカ６４１および増幅回路６４２から成る音声出力装置６４０、マイク６５１およびカメラ６６２から成るユーザ監視装置、およびこれらを接続するバスライン６５０から構成される。そして、記憶装置６２０に記憶されるプログラムは、入力部、音声合成部、音声提示部、ユーザ挙動監視部、音声合成手法指定部、音声提示手法指定部、音声評価推定部、音声修正部などとして動作する。

【符号の説明】

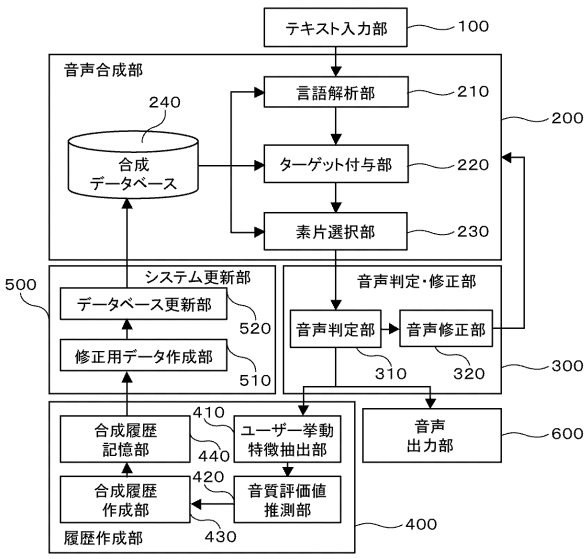
【００７８】

１００…テキスト入力部、１０１…入力テキスト、
２００…音声合成部、２１０…言語解析部、２１１…中間言語、２２０…ターゲット付与部、２３０…素片選択部、２３１…合成情報、２３２…合成音声、２４０…合成データベース、２４１…言語解析モデル、２４２…ターゲット付与モデル、２４３…単位選択データベース、２４８…コスト計算関数群、２４９…音声素片データベース、
３００…音声判定・修正部、３０１…合成情報、３０２…合成音声、３１０…音声判定部、３１１…音声判定ユーザインタフェース、３１５…ユーザ挙動情報、３２０…音声修正

部、３２１…音声修正ユーザインタフェース、３２６…ユーザ挙動情報、３２７…修正された合成情報、３２８…再合成用情報分析部、
４００…履歴作成部、４０１…ユーザ挙動情報、４０２…合成情報、４１０…ユーザ挙動特徴抽出部、４１１…ユーザ挙動特徴、４２０…音質評価値推測部、４２１…音質評価値推定ルール、４２２…推定された音質評価値、４３０…合成履歴作成部、４３１…合成履歴、４４０…合成履歴記憶部、４４１…合成履歴群、
５００…システム更新部、５０１…システム更新命令、５０２…学習データベース、５１０…修正用データ作成部、５２０…データベース更新部、
６００…音声出力部、６１０…制御装置、６１１…ＣＰＵ、６１２…メモリ、６２０…記憶装置、６３０…テキスト入力装置、６３１…キーボード、６３２…通信装置、６４０…音声出力装置、６４１…スピーカ、６４２…増幅回路、６５０…バスライン、６６０…ユーザ監視装置、６６１…マイク、６６２…カメラ、
７００…音声合成出力部、７０１…入力部、７０２…音声合成部、７０３…音声提示部、７０４…音声合成手法指定部、７０５…音声提示手法指定部、７０６…テキスト及び中間言語、７０７…合成音声、７１０…ユーザ挙動監視部、７１１…ユーザ挙動特徴、７２０…音声評価推測部、７２１…音声評価値、７２２…音声評価値推定ルール、７３０…音声修正部、７３１…修正結果。

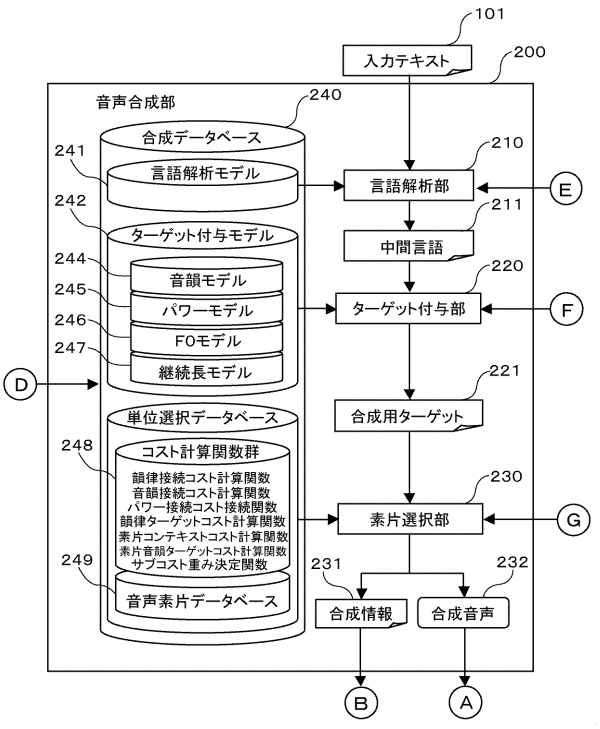
【図１】

図１



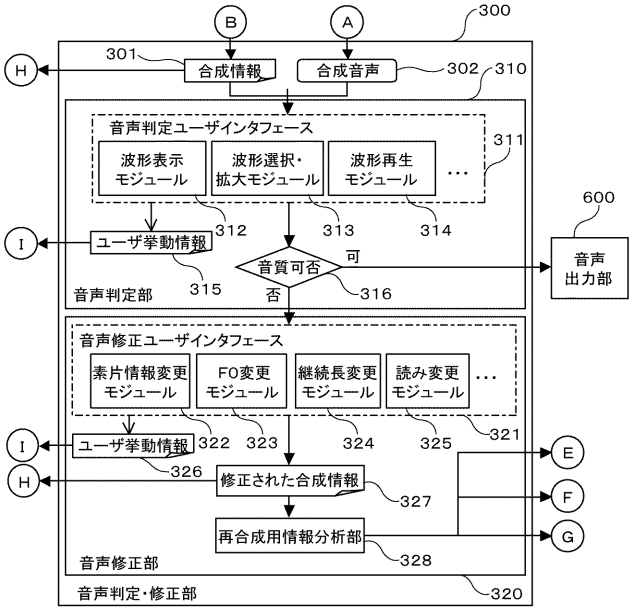
【図２】

図２



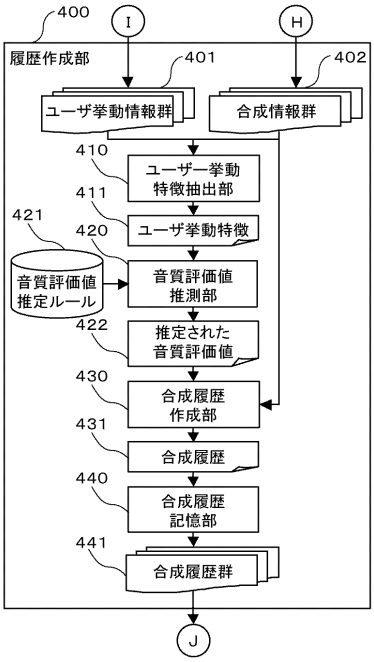
【図３】

図３

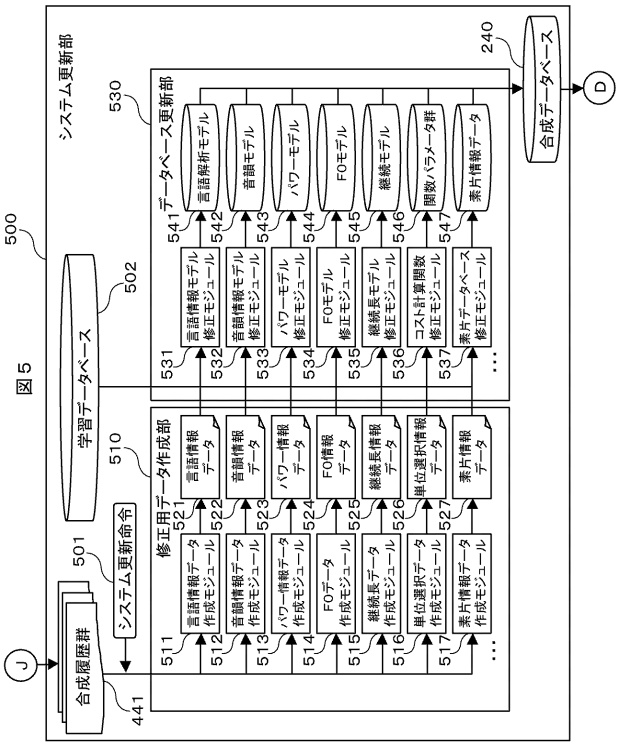


【図４】

図４

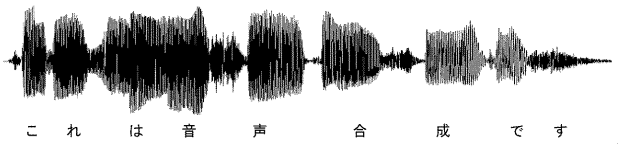


【図５】



【図６】

図６



【図７】

図７

音質評価値推定ルール

ルール1：文最初は5点、減点方式を行う。

ルール2：修正されたら、減点2.

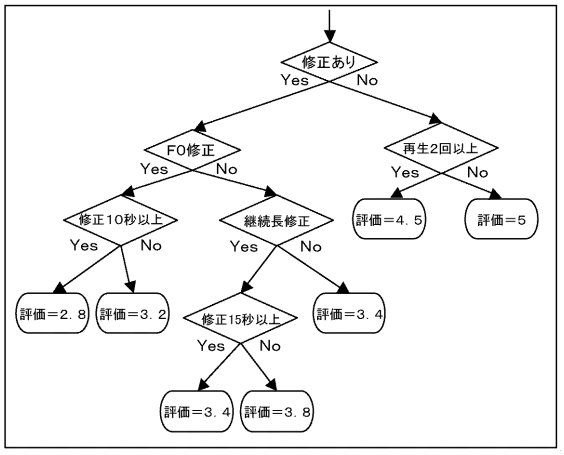
ルール3：全体再生、一回ごとに減点0．2点。ただし、0．6点までとする。

ルール4：部分再生、一回ごとに減点0．3点。ただし、0．9点までとする。

ルール5：・・・

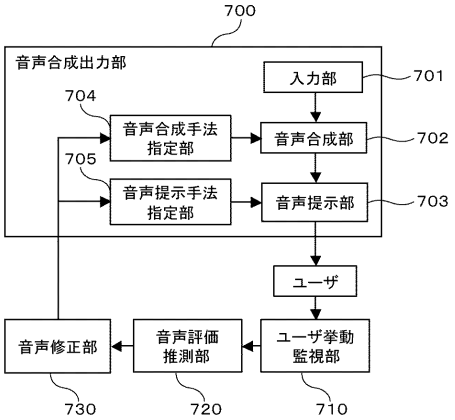
【図８】

図８



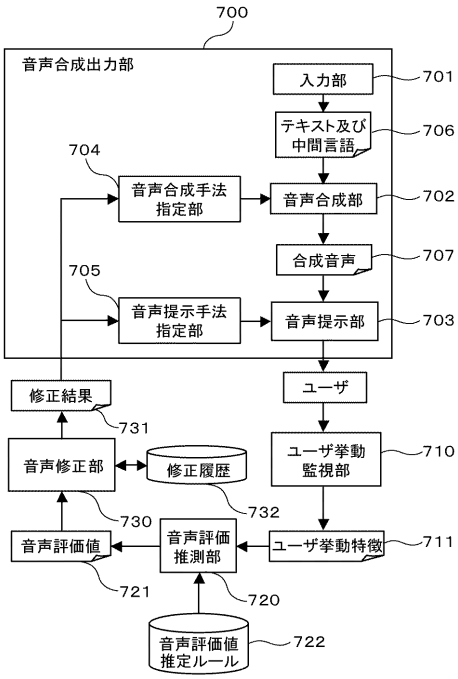
【図９】

図９



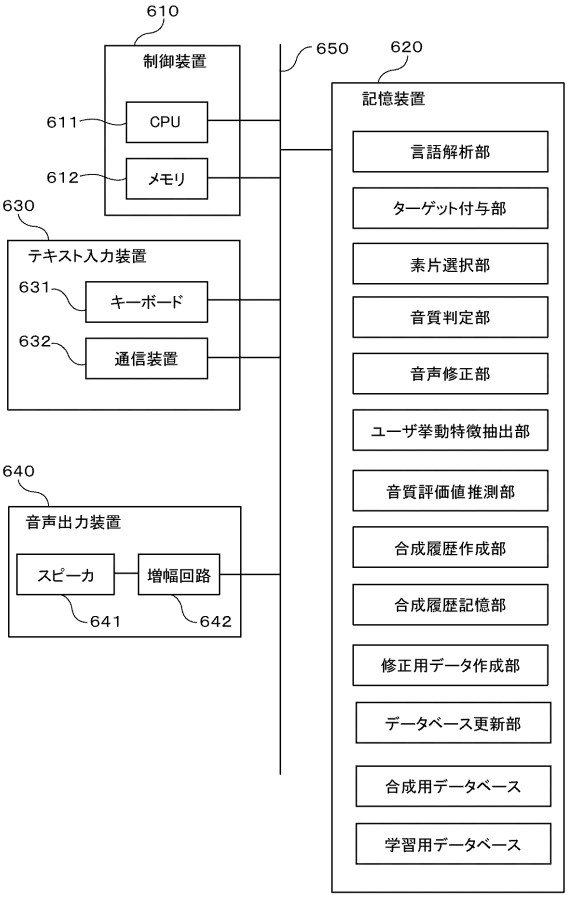
【図１０】

図１０



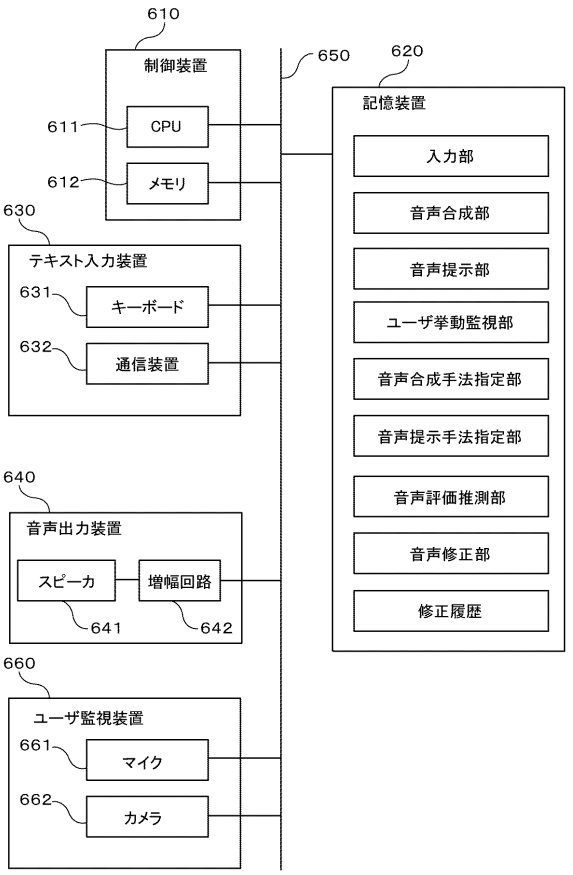
【図１１】

図１１



【図１２】

図１２


610
制御装置
611
CPU
612
メモリ
650
620
記憶装置
入力部
音声合成部
音声提示部
ユーザ挙動監視部
音声合成手法指定部
音声提示手法指定部
音声評価推測部
音声修正部
修正履歴
630
テキスト入力装置
631
キーボード
632
通信装置
640
音声出力装置
スピーカ
641
増幅回路
642
660
ユーザ監視装置
661
マイク
662
カメラ