(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号

第2975586号

(45)発行日　平成11年(1999)11月10日　(24)登録日　平成11年(1999)９月３日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ１０Ｌ　5/04		Ｇ１０Ｌ　5/04	Ｆ
3/00		3/00	Ｈ

請求項の数15(全 23 頁)

(21)出願番号　特願平10－51925

(22)出願日　平成10年(1998)３月４日

(65)公開番号　特開平11－249695

(43)公開日　平成11年(1999)９月17日

審査請求日　平成10年(1998)３月４日

(73)特許権者　593118597
株式会社エイ・ティ・アール音声翻訳通信研究所
京都府相楽郡精華町大字乾谷小字三平谷５番地

(72)発明者　藤澤　謙
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール音声翻訳通信研究所内

(72)発明者　ニック・キャンベル
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール音声翻訳通信研究所内

(74)代理人　弁理士　青山　葆　(外２名)

審査官　涌井　智則

最終頁に続く

(54)【発明の名称】　音声合成システム

1

(57)【特許請求の範囲】

【請求項１】　第１の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、

上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する第１の音声分析手段と、

上記第１の音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、

上記第２の記憶手段によって記憶された第１の音響的特徴パラメータと韻律的特徴パラメータとに基づいて、同

2

一の音素種類の１つの目標音素とそれ以外の音素候補との間の第２の音響的特徴パラメータにおける音響的距離を計算し、上記計算した音響的距離に基づいて各音素候補に対して上記第２の音響的特徴パラメータ毎に所定の統計的解析を実行することにより、各音素候補に対する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定する重み係数学習手段と、

上記重み係数学習手段によって決定された上記第２の音響的特徴パラメータにおける各目標音素毎の重み係数ベクトルを記憶する第３の記憶手段と、

上記第１の言語とは異なる第２の言語の自然発話の音声波形信号の音声セグメントを記憶する第４の記憶手段と、

上記第４の記憶手段によって記憶された音声波形信号の

音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する第２の音声分析手段と、

上記第２の音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第５の記憶手段と、

上記第３の記憶手段によって記憶された各目標音素毎の重み係数ベクトルと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する第１の音声単位選択手段と、

上記第１の音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応する第１の言語の音声信号波形データを合成して出力する第１の音声合成手段と、

上記第１の音声合成手段から出力される音声信号波形データからケプストラム係数データを抽出して出力する抽出手段と、

上記抽出手段から出力されるケプストラム係数データと、上記第５の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、上記入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する第２の音声単位選択手段と、

上記第２の音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第４の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応しかつ第２の言語の音声セグメントとによる音声信号波形を合成して出力する第２の音声合成手段とを備えたことを特徴とする音声合成システム。

【請求項２】　それぞれ互いに異なる話者の第１の言語の自然発話の音声波形信号の音声セグメントを記憶する複数の第１の記憶手段と、

上記複数の第１の記憶手段に記憶された異なる話者の第１の言語の自然発話の音声波形信号の音声セグメントと、上記第４の記憶手段に記憶された第２の言語の自然発話の音声波形信号の音声セグメントとに基づいて、所定の特徴パラメータの選択基準を用いて、第２の言語の自然発話の音声波形信号により声質が近い第１の言語の自然発話の音声波形信号の話者を選択して、選択した話者の第１の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段を上記第１の音声合成手段に接続する話者選択手段とをさらに備えたことを特徴とする請求項１記載の音声合成システム。

【請求項３】　上記特徴パラメータの選択基準に用いる特徴パラメータは、話者の性別及び基本周波数の平均値であることを特徴とする請求項２記載の音声合成システム。

【請求項４】　上記第１の音声分析手段は、入力される音声波形信号に基づいて上記音声波形信号に対応する音素列を予測する音素予測手段を備えたことを特徴とする請求項１乃至３のうちの１つに記載の音声合成システム。

【請求項５】　上記重み係数学習手段は、上記計算した音響的距離に基づいて、最良の上位複数Ｎ１個の音素候補を抽出した後、上記第２の音響的特徴パラメータの各々に対して線形回帰分析することにより、各音素候補に関する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定することを特徴とする請求項１乃至４のうちの１つに記載の音声合成システム。

【請求項６】　上記重み係数学習手段は、上記計算した音響的距離に基づいて、最良の上位複数Ｎ１個の音素候補を抽出した後、上記第２の音響的特徴パラメータの各々に対して所定のニューラルネットワークを用いた統計的解析を実行することにより、各音素候補に関する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定することを特徴とする請求項１乃至４のうちの１つに記載の音声合成システム。

【請求項７】　上記第１と第２の音声単位選択手段はそれぞれ、上記目標コストと上記連結コストとを含むコストが最良の上位複数Ｎ２個の音素候補を抽出した後、コストが最小となる音素候補の組み合わせを検索することを特徴とする請求項１乃至６のうちの１つに記載の音声合成システム。

【請求項８】　上記第１の音響的特徴パラメータは、ケプストラム係数と、デルタケプストラム係数と、音素ラベルとを含むことを特徴とする請求項１乃至７のうちの１つに記載の音声合成システム。

【請求項９】　上記第１の音響的特徴パラメータは、フォルマントパラメータと、声道音源パラメータとを含むことを特徴とする請求項１乃至８のうちの１つに記載の音声合成システム。

【請求項１０】　上記韻律的特徴パラメータは、音素時間長と、音声基本周波数F_0と、パワーとを含むことを特徴とする請求項１乃至９のうちの１つに記載の音声合

成システム。

【請求項１１】　上記第２の音響的特徴パラメータは、ケプストラム距離を含むことを特徴とする請求項１乃至１０のうちの１つに記載の音声合成システム。

【請求項１２】　入力される第１の言語の音声信号と、それに対応する第１の言語の音素列に基づいて、上記第１の言語の音素列に対応しかつ上記第１の言語とは異なる第２の言語の音声セグメントによる音声信号波形を合成して出力する音声合成システムであって、

上記第２の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、

上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する音声分析手段と、

上記音声分析手段から出力される索引情報と、上記音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、

上記入力される第１の言語の音声信号を音声信号波形データに変換して、変換された音声信号波形データからケプストラム係数データを抽出して出力する抽出手段と、

上記抽出手段から出力されるケプストラム係数データと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、上記入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する音声単位選択手段と、上記音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応しかつ第２の言語の音声セグメントによる音声信号波形を合成して出力する音声合成手段とを備えたことを特徴とする音声合成システム。

【請求項１３】　上記音声単位選択手段はそれぞれ、上記目標コストと上記連結コストとを含むコストが最良の上位複数Ｎ２個の音素候補を抽出した後、コストが最小となる音素候補の組み合わせを検索することを特徴とする請求項１２に記載の音声合成システム。

【請求項１４】　上記音響的特徴パラメータは、ケプストラム係数と、デルタケプストラム係数と、音素ラベルとを含むことを特徴とする請求項１２又は１３に記載の音声合成システム。

【請求項１５】　上記韻律的特徴パラメータは、音素時

間長と、音声基本周波数F_0と、パワーとを含むことを特徴とする請求項１２乃至１４のうちの１つに記載の音声合成システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、自然発話の音声波形信号の音声セグメントを連結することにより任意の音素列を音声合成する自然発話音声波形信号接続型音声合成装置を用いて、第１の言語のネイティブの人の音声波形データベースによる第１の言語の自然発話文の音声合成信号データに基づいて、第１の言語とは異なる第２の言語のネイティブの人の音声波形データベースによる第１の言語の自然発話文の音声合成信号を発生する音声合成システム、並びに、然発話の音声波形信号の音声セグメントを連結することにより任意の音素列を音声合成する自然発話音声波形信号接続型音声合成装置を用いて、入力される第１の言語のネイティブの人の自然発話文の音声信号に基づいて、第１の言語とは異なる第２の言語のネイティブの人の音声波形データベースによる第１の言語の自然発話文の音声合成信号を発生する音声合成システムに関する。ここで、ある言語のネイティブの人とは、その言語を話す国又は地域で生まれて育った人、もしくは、その言語を話す国又は地域で比較的長期間にわたって滞在して、生まれて育った人と同等にネイティブに話す人などをいう。

【０００２】

【従来の技術】図２は、第１の従来例の音声合成システムのブロック図である。図２に示すように、学習用話者の信号波形データに対して例えばＬＰＣ分析を実行し、１６次ケプストラム係数を含む特徴パラメータを抽出する。抽出された特徴パラメータは、バッファメモリである特徴パラメータメモリ６２に記憶された後、当該メモリ６２からパラメータ時系列生成部５２に入力される。次いで、パラメータ時系列生成部５２は、抽出された特徴パラメータに基づいて、時間正規化、及びメモリ６３内の韻律制御規則を用いたパラメータ時系列の生成処理などの信号処理を実行することにより、音声合成に必要な、例えば１６次のケプストラム係数などのパラメータ時系列を生成して音声合成部５３に出力する。

【０００３】音声合成部５３は公知の音声合成装置であって、有声音を発生するためのパルス発生器５３ａと、無声音を発生するための雑音発生器５３ｂと、フィルタ係数を変更可能なフィルタ５３ｃとを備え、入力されるパラメータ時系列に基づいて、パルス発生器５３ａによって発生される有声音と、雑音発生器５３ｂによって発生される無声音とを切り換え、かつその振幅を制御し、さらには、フィルタ５３ｃの伝達関数に対応するフィルタ係数を変化することにより、音声合成された音声信号を発生して、スピーカ５４からその音声を出力させる。

【０００４】しかしながら、第１の従来例の音声合成装

置では、韻律制御規則を用いた信号処理を必要とするために、また、処理された特徴パラメータに基づいて音声合成しているために、声質がきわめて悪いという問題点があった。

【０００５】以上の問題点を解決するために、本特許出願人は、特願平９－１２３８２２号の特許出願において、韻律制御規則を使わず、信号処理を実行することなく、任意の音素列を発声音声に変換することができ、しかも従来例に比較して自然に近い声質を得ることができる音声合成装置（以下、第２の従来例という。）を提案している。この第２の従来例の音声合成装置では、自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する音声分析手段と、上記音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、上記第２の記憶手段によって記憶された第１の音響的特徴パラメータと韻律的特徴パラメータとに基づいて、同一の音素種類の１つの目標音素とそれ以外の音素候補との間の第２の音響的特徴パラメータにおける音響的距離を計算し、上記計算した音響的距離に基づいて各音素候補に対して上記第２の音響的特徴パラメータ毎に所定の統計的解析を実行することにより、各音素候補に対する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定する重み係数学習手段と、上記重み係数学習手段によって決定された上記第２の音響的特徴パラメータにおける各目標音素毎の重み係数ベクトルを記憶する第３の記憶手段と、上記第３の記憶手段によって記憶された各目標音素毎の重み係数ベクトルと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、入力される自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する音声単位選択手段と、上記音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された音素列に対応する音声を合成して出力する音声合成手段とを備えて構成している。

【０００６】

【発明が解決しようとする課題】ところで、多言語翻訳電話装置で、日本語から英語への変換を考えた場合、現状では、日本人が日本語で話した言葉を英語に翻訳後、（ａ）英語話者すなわち元話者とは全く異なった声の合成音声を出力するか、もしくは、（ｂ）元話者の日本語ＤＢから英語のローマ字読みの合成音声を出力するしか方法がない。上記（ａ）の方法では、元話者とは全く異なった音声になり、上記（ｂ）の方法では、元話者の声ではあるものの、いわゆるカタカナ英語の発音となる。すなわち、例えば、日本人の声による英語の音声合成信号を発生する装置はなかった。

【０００７】本発明の目的は以上の問題点を解決し、第２の言語のネイティブの人の声による第１の言語の自然発話文の音声合成信号を、自然に近い声質で発生することができる音声合成音声合成システムを提供することにある。

【０００８】

【課題を解決するための手段】本発明に係る請求項１記載の音声合成システムは、第１の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する第１の音声分析手段と、上記第１の音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、上記第２の記憶手段によって記憶された第１の音響的特徴パラメータと韻律的特徴パラメータとに基づいて、同一の音素種類の１つの目標音素とそれ以外の音素候補との間の第２の音響的特徴パラメータにおける音響的距離を計算し、上記計算した音響的距離に基づいて各音素候補に対して上記第２の音響的特徴パラメータ毎に所定の統計的解析を実行することにより、各音素候補に対する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定する重み係数学習手段と、上記重み係数学習手段によって決定された上記第２の音響的特徴パラメータにおける各目標音素毎の重み係数ベクトルを記憶する第３の記憶手段と、上記第１の言語とは異なる第２の言語の自然発話の音声波形信号の音声セグメントを記憶する第４の記憶手段と、上記第４の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する第２の音声分析手段と、上記第２の音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する

第５の記憶手段と、上記第３の記憶手段によって記憶された各目標音素毎の重み係数ベクトルと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する第１の音声単位選択手段と、上記第１の音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応する第１の言語の音声信号波形データを合成して出力する第１の音声合成手段と、上記第１の音声合成手段から出力される音声信号波形データからケプストラム係数データを抽出して出力する抽出手段と、上記抽出手段から出力されるケプストラム係数データと、上記第５の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、上記入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する第２の音声単位選択手段と、上記第２の音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第４の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応しかつ第２の言語の音声セグメントによる音声信号波形を合成して出力する第２の音声合成手段とを備えたことを特徴とする。

【０００９】また、請求項２記載の音声合成システムは、請求項１記載の音声合成システムにおいて、それぞれ互いに異なる話者の第１の言語の自然発話の音声波形信号の音声セグメントを記憶する複数の第１の記憶手段と、上記複数の第１の記憶手段に記憶された異なる話者の第１の言語の自然発話の音声波形信号の音声セグメントと、上記第４の記憶手段に記憶された第２の言語の自然発話の音声波形信号の音声セグメントとに基づいて、所定の特徴パラメータの選択基準を用いて、第２の言語の自然発話の音声波形信号により声質が近い第１の言語の自然発話の音声波形信号の話者を選択して、選択した話者の第１の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段を上記第１の音声合成手段に接続する話者選択手段とをさらに備えたことを特徴とする。さらに、請求項３記載の音声合成システムは、請求項２記載の音声合成システムにおいて、上記特徴パラメータの選択基準に用いる特徴パラメータは、話者の性別及び基本周波数の平均値であることを特徴とする。

【００１０】また、請求項４記載の音声合成システムは、請求項１乃至３のうちの１つに記載の音声合成システムにおいて、上記第１の音声分析手段は、入力される音声波形信号に基づいて上記音声波形信号に対応する音素列を予測する音素予測手段を備えたことを特徴とする。また、請求項５記載の音声合成システムは、請求項１乃至４のうちの１つに記載の音声合成システムにおいて、上記重み係数学習手段は、上記計算した音響的距離に基づいて、最良の上位複数Ｎ１個の音素候補を抽出した後、上記第２の音響的特徴パラメータの各々に対して線形回帰分析することにより、各音素候補に関する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定することを特徴とする。さらに、請求項６記載の音声合成システムは、請求項１乃至４のうちの１つに記載の音声合成システムにおいて、上記重み係数学習手段は、上記計算した音響的距離に基づいて、最良の上位複数Ｎ１個の音素候補を抽出した後、上記第２の音響的特徴パラメータの各々に対して所定のニューラルネットワークを用いた統計的解析を実行することにより、各音素候補に関する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定することを特徴とする。また、請求項７記載の音声合成システムは、請求項１乃至６のうちの１つに記載の音声合成システムにおいて、上記第１と第２の音声単位選択手段はそれぞれ、上記目標コストと上記連結コストとを含むコストが最良の上位複数Ｎ２個の音素候補を抽出した後、コストが最小となる音素候補の組み合わせを検索することを特徴とする。

【００１１】また、請求項８記載の音声合成システムは、請求項１乃至７のうちの１つに記載の音声合成システムにおいて、上記第１の音響的特徴パラメータは、ケプストラム係数と、デルタケプストラム係数と、音素ラベルとを含むことを特徴とする。さらに、請求項９記載の音声合成システムは、請求項１乃至８のうちの１つに記載の音声合成システムにおいて、上記第１の音響的特徴パラメータは、フォルマントパラメータと、声道音源パラメータとを含むことを特徴とする。またさらに、請求項１０記載の音声合成システムは、請求項１乃至９のうちの１つに記載の音声合成システムにおいて、上記韻律的特徴パラメータは、音素時間長と、音声基本周波数F_0と、パワーとを含むことを特徴とする。さらに、請求項１１記載の音声合成システムは、請求項１乃至１０のうちの１つに記載の音声合成システムにおいて、上記第２の音響的特徴パラメータは、ケプストラム距離を含むことを特徴とする。

【００１２】本発明に係る請求項１２記載の音声合成システムは、入力される第１の言語の音声信号と、それに対応する第１の言語の音素列に基づいて、上記第１の言

語の音素列に対応しかつ上記第１の言語とは異なる第２の言語の音声セグメントによる音声信号波形を合成して出力する音声合成システムであって、上記第２の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する音声分析手段と、上記音声分析手段から出力される索引情報と、上記音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、上記入力される第１の言語の音声信号を音声信号波形データに変換して、変換された音声信号波形データからケプストラム係数データを抽出して出力する抽出手段と、上記抽出手段から出力されるケプストラム係数データと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、上記入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する音声単位選択手段と、上記音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応しかつ第２の言語の音声セグメントによる音声信号波形を合成して出力する音声合成手段とを備えたことを特徴とする。

【００１３】また、請求項１３記載の音声合成システムは、請求項１２記載の音声合成システムにおいて、上記音声単位選択手段はそれぞれ、上記目標コストと上記連結コストとを含むコストが最良の上位複数Ｎ２個の音素候補を抽出した後、コストが最小となる音素候補の組み合わせを検索することを特徴とする。さらに、請求項１４記載の音声合成システムは、請求項１２又は１３記載の音声合成システムにおいて、上記音響的特徴パラメータは、ケプストラム係数と、デルタケプストラム係数と、音素ラベルとを含むことを特徴とする。またさらに、請求項１５記載の音声合成システムは、請求項１２乃至１４のうちの１つに記載の音声合成システムにおいて、上記韻律的特徴パラメータは、音素時間長と、音声基本周波数F_0と、パワーとを含むことを特徴とする。

【００１４】

【発明の実施の形態】以下、図面を参照して本発明に係る実施形態について説明する。

【００１５】図１は、本発明に係る一実施形態である音声合成システムの構成を示すブロック図である。この実施形態の音声合成システムは、一般的に言えば、ある言語の音声波形データベースから別の言語の音声を合成する際に、一度上記ある言語のネイティブの音声波形データベースを用いて音声合成信号波形データを発生し、そのケプストラム情報に基づいてそれをターゲットとして、ネイティブ以外の別の言語の音声波形データベースを用いて音声を合成することを特徴としている。具体的には、当該実施形態の音声合成システムは、大きく分けて、英語音声による音声合成装置１と、日本語音声による音声合成装置２とを備えて構成され、英語の音声波形データベースから日本語の言語の音声を合成する際に、英語音声による音声合成装置１において、一度英語のネイティブの音声波形データベース（メモリ２１内）を用いて音声合成信号波形データを発生し、そのケプストラム情報に基づいて、日本語音声による音声合成装置２において、上記ケプストラム情報をターゲットとして、ネイティブ以外の別の言語の音声波形データベース（メモリ１２１内）を用いて音声を合成することを特徴としている。すなわち、本実施形態では、日本人による英語の音声合成を実現する。

【００１６】例えば図２に示した第１の従来例の音声合成装置では入力された発声音声に対応するテキスト抽出から音声波形信号の生成までが一連の処理として行なわれるのに対して、本実施形態の英語音声による音声合成装置１では、大きく分類すれば、次の４つの処理部に分類される。

（１）英語の音声波形信号データベースメモリ２１内の音声波形信号データベースの音声波形信号データの音声分析、具体的には、英語の音素記号系列の生成、音素のアラインメント、特徴パラメータの抽出を含む処理を実行する音声分析部１０。

（２）最適重み係数を学習しながら決定する重み係数学習部１１。

（３）入力される英語の音素列に基づいて音声単位の選択を実行して入力音素列に対応する音声波形信号データの索引情報を出力する音声単位選択部１２。

（４）音声単位選択部１２から出力される索引情報に基づいて英語の音声波形信号データベースメモリ２１内の音声波形信号データベースをランダムにアクセスして最適とされた各音素候補の音声波形信号を再生してバッファンメモリ１４を介して日本語音声による音声合成装置２のケプストラム抽出部１１１に出力する音声合成部１３。

【００１７】また、日本語音声による音声合成装置２では、大きく分類すれば、次の４つの処理部に分類される。

（１）日本語の音声波形信号データベースメモリ１２１内の音声波形信号データベースの音声波形信号データの音声分析、具体的には、日本語の音素記号系列の生成、音素のアラインメント、特徴パラメータの抽出を含む処

理を実行する音声分析部１１０。
（２）音声合成部１３からバッファメモリ１４を介して入力される英語の音声合成信号波形データに基づいて、ケプストラム係数データを抽出するケプストラム抽出部１１１。
（３）上記と同じ入力される英語の音素列に基づいて、抽出されたケプストラム係数データを目標として音声単位の選択を実行して入力音素列に対応する音声波形信号データの索引情報を出力する音声単位選択部１１２。
（４）音声単位選択部１１２から出力される索引情報に基づいて日本語の音声波形信号データベースメモリ１２１内の音声波形信号データベースをランダムにアクセスして最適とされた各音素候補の音声波形信号を再生してスピーカ１１４に出力する音声合成部１１３。
【００１８】具体的には、英語音声による音声合成装置１において、音声分析部１０は、入力される英語の自然発話の音声波形信号の音声セグメントと、上記音声波形信号に対応する英語の音素列とに基づいて、英語の音素隠れマルコフモデルメモリ（以下、隠れマルコフモデルをHMMという。）２３内のHMMを参照して、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の第１の韻律的特徴パラメータとを抽出して出力する。英語の特徴パラメータメモリ３０は、上記音声分析部１０から出力される索引情報と、上記第１の音響的特徴パラメータと、上記第１の韻律的特徴パラメータとを記憶する。次いで、重み係数学習部１１は、英語の特徴パラメータメモリ３０に記憶された第１の音響的特徴パラメータと上記第１の韻律的特徴パラメータとに基づいて、同一の音素種類の１つの目標音素とそれ以外の音素候補との間の第２の音響的特徴パラメータにおける音響的距離を計算し、上記計算した音響的距離に基づいて各音素候補に対して上記第２の音響的特徴パラメータ毎に所定の統計的解析を実行することにより、各音素候補に対する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定する。英語の重み係数ベクトルメモリ３１は、重み係数学習部１１によって決定された上記第２の音響的特徴パラメータにおける各目標音素毎の重み係数ベクトルと、予め与えられた、各音素候補に関する第２の韻律的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルとを記憶する。さらに、音声単位選択部１２は、英語の重み係数ベクトルメモリ３１に記憶された各目標音素毎の重み係数ベクトルと、英語の特徴パラメータメモリ３０に記憶された第１の韻律的特徴パラメータとに基づいて、入力される英語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する。そして、音声合成部１３は、音声単位選択部１２から出力される索引情報に基づいて、当該索引情報に対応する英語の音声波形信号の音声セグメントを英語の音声波形信号データベースメモリ２１から逐次読み出して連結してバッファメモリ１４を介してケプストラム抽出部１１１に出力する。
【００１９】次いで、日本語音声による音声合成装置２において、音声分析部１１０は、入力される日本語の自然発話の音声波形信号の音声セグメントと、上記音声波形信号に対応する日本語の音素列とに基づいて、日本語のHMMメモリ１２３内のHMMを参照して、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の第１の韻律的特徴パラメータとを抽出して出力する。日本語の特徴パラメータメモリ１３０は、上記音声分析部１１０から出力される索引情報と、上記第１の音響的特徴パラメータと、上記第１の韻律的特徴パラメータとを記憶する。一方、ケプストラム抽出部１１１は、入力される英語の音声合成信号波形データから音素毎に例えば１２次のメルケプストラム係数などのケプストラム係数データを抽出して音声単位選択部１１２に出力する。さらに、音声単位選択部１１２は、入力されるケプストラム係数データを目標音素データとして用いて、日本語の特徴パラメータメモリ１３０に記憶されたケプストラム係数データのみならず、（１）音素接続点におけるケプストラム距離、（２）対数パワーの差の絶対値、及び、（３）音声基本周波数F_0の差の絶対値を含む音響的特徴パラメータに基づいて、入力される英語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する。そして、音声合成部１１３は、音声単位選択部１２から出力される索引情報に基づいて、当該索引情報に対応する日本語の音声波形信号の音声セグメントを日本語の音声波形信号データベースメモリ１２１から逐次読み出して連結してスピーカ１４に出力することにより、音声合成装置は、上記入力された英語の音素列に対応する日本人による（すなわち、日本語の音声波形データベースメモリ１２１内の音声波形データベースの日本人の音声セグメントによる）音声を合成して出力する。
【００２０】ここで、音声分析部１０及び１１０の処理は新しい各音声波形信号データベース（メモリ２１及び１２１内）に対しては必ず一度行なう必要があり、重み係数学習部１１の処理は、一般に一度の処理でよく、重み係数学習部１１によって求めた最適重み係数は異なる音声合成条件に対しても再利用が可能である。さらに、

音声単位選択部１２及び１１２と、ケプストラム抽出部１１１と、音声合成部１３の処理は、音声合成すべき入力音素列が変われば、その都度実行される。

【００２１】本実施形態の英語音声による音声合成装置１は、与えられたレベルの入力に基づいて必要とする、すべての特徴パラメータを予測し、所望の音声の特徴に最も近いサンプル（すなわち、音素候補の音声波形信号）をメモリ２１内の音声波形信号データベースの中から選び出す。最低限、音素ラベルの系列が与えられれば処理は可能であるが、音声基本周波数F_0や音素時間長が予め与えられていれば、さらに高品質の合成音声が得られる。なお、入力として単語の情報だけが与えられた場合には、例えば音素HMMなどの辞書や規則に基づいて音素系列を予測する必要がある。また、韻律特徴が与えられなかった場合には音声波形信号データベース中のいろいろな環境における音素の既知の特徴を基に標準的な韻律を生成する。

【００２２】また、日本語音声による音声合成装置２は、音声合成装置１から入力されるケプストラム係数データを目標音素データとして、所望の音声の特徴に最も近いサンプル（すなわち、音素候補の音声波形信号）をメモリ１２１内の音声波形信号データベースの中から選び出す。最低限、音素ラベルの系列が与えられれば処理は可能であるが、音声基本周波数F_0や音素時間長が予め与えられていれば、さらに高品質の合成音声が得られる。なお、入力として単語の情報だけが与えられた場合には、例えば音素HMMなどの辞書や規則に基づいて音素系列を予測する必要がある。また、韻律特徴が与えられなかった場合には音声波形信号データベース中のいろいろな環境における音素の既知の特徴を基に標準的な韻律を生成する。

【００２３】本実施形態では、音声波形信号データベースメモリ２１及び１２１内の録音内容を少なくとも正書法で記述されたテキストデータが例えば、テキストデータベースメモリ２２及び１２２内のテキストデータベースのように存在するならば、あらゆる音声波形信号データベースが合成用の音声波形信号データとして利用可能であるが、出力音声の品質は録音状態、音声波形信号データベース中の音素のバランス等に大きく影響を受け、＊メモリ２１及び１２１内の音声波形信号データベースが豊富な内容であれば、より多様な音声が合成でき、反対に音声波形信号データベースが貧弱であれば、合成音声は不連続感が強く、ブツブツしたものになる。

【００２４】次いで、自然な発話音声に対する音素ラベル付けについて説明する。音声単位の選択の善し悪しは音声波形信号データベース中の音素のラベル付けと検索の方法に依存する。ここで、好ましい実施形態においては、音声単位は、音素である。まず、録音された音声に付与された正書法の発話内容を音素系列に変換し、さらに音声波形信号に割り当てる。韻律的特徴パラメータの抽出はこれに基づいて行なわれる。音声分析部１０及び１１０の入力はそれぞれメモリ２２及び１２２内の音素表記を伴ったメモリ２１及び１２１内の音声波形信号データであり、出力は特徴ベクトル又は特徴パラメータである。この特徴ベクトルは音声波形信号データベース中で音声サンプルを表す基本単位となり、最適な音声単位の選択に用いられる。

【００２５】音声分析部１０及び１１０の処理における第１段階においては、正書法で書かれた発話内容が実際の音声波形信号データでどのように発音されているかを記述するための正書法テキストから音素記号への変換である。次いで、第２段階においては、韻律的及び音響的特徴を計測するために各音素の開始及び終了時点を決めるために、各音素記号を音声波形信号に対応付ける処理である（以下、当該処理を、音素のアラインメント処理という。）。さらに、第３段階においては、各音素の特徴ベクトル又は特徴パラメータを生成することである。この特徴ベクトルには、必須項目として音素ラベル、メモリ２１及び１２１内の音声波形信号データベース中の各ファイルにおける当該音素の開始時刻（開始位置）、音声基本周波数F_0、音素時間長、パワーの情報が記憶され、さらに、特徴パラメータのオプションとしてストレス、アクセント型、韻律境界に対する位置、スペクトル傾斜等の情報が記憶される。以上の特徴パラメータを整理すると、例えば、次の表１のようになる。

【００２６】

【表１】

索引情報：
　索引番号（１つのファイルに対して付与）
　メモリ２１及び１２１内の音声波形信号データベース中の各ファイルにおける当該音素の開始時刻（開始位置）

第１の音響的特徴パラメータ：
　１２次メルケプストラム係数
　１２次Δメルケプストラム係数
　音素ラベル
　弁別素性：

母音性（vocalic）（＋）／非母音性（non-vocalic）（－）
子音性（consonantal）（＋）／非子音性（non-consonantal）（－）
中断性（interrupted）（＋）／連続性（continuant）（－）
抑止性（checked）（＋）／非抑止性（unchecked）（－）
粗擦性（strident）（＋）／円熟性（mellow）（－）
有声（voiced）（＋）／無声（unvoiced）（－）
集約性（compact）（＋）／拡散性（diffuse）（－）
低音調性（grave）（＋）／高音調性（acute）（－）
変音調性（flat）（＋）／常音調性（plain）（－）
嬰音調性（sharp）（＋）／常音調性（plain）（－）
緊張性（tense）（＋）／弛緩性（lax）（－）
鼻音性（nasal）（＋）／口音性（oral）（－）

第１の韻律的特徴パラメータ：
音素時間長
音声基本周波数F_0
パワー

【００２７】とって代わって、第１の音響的特徴パラメータは、好ましくは、フォルマントパラメータと、声道音源パラメータであってもよい。

【００２８】上記索引情報内の開始時刻（開始位置）、第１の音響的特徴パラメータ及び第１の韻律的特徴パラメータは、各音素毎に特徴パラメータメモリ３０及び１３０に記憶される。ここで、音素ラベルに付与される、例えば１２個の弁別素性の特徴パラメータは各項目別に（＋）又は（－）のパラメータ値が与えられる。さらに、例えば、音声分析部１０の出力結果である特徴パラメータの一例を表２に示す。ここで、索引番号は、音声波形信号データベースメモリ２１において、例えば複数の文からなる１つのパラグラフ又は１つの文のファイル毎に、索引番号が付与され、そして、１つの索引番号が付与されたファイル中の任意の音素の位置を示すために当該ファイル内の開始時刻から計時された当該音素の開始時刻及びその当該音素の音素時間長とを付与することにより、当該音素の音声波形信号の音声セグメントを特定することができる。

【００２９】

【表２】音声分析部１０の出力結果である特徴パラメータの一例

索引番号Ｘ０００５

音素	時間長	基本周波数	パワー	……
#	１２０	９０	４．０	……
s	１７５	９８	４．７	……
ei	９５	１０２	６．５	……
dh	３０	１１４	４．９	……
ih	７５	１４３	６．９	……
s	１５０	１４０	５．７	……
p	８７	１３７	５．１	……
l	３４	１０７	４．９	……
ii	１５０	９８	６．３	……
z	１４０	８７	５．８	……
#	２５３	８７	４．０	……

【００３０】表２において、＃はポーズを示す。音声単位を選択する場合に、音響的及び韻律的な各特徴パラメータがそれぞれの音素でどれだけの寄与をするかを予め調べておくことが必要であり、第４段階では、このために音声波形信号データベース中のすべての音声サンプルを用いて各特徴パラメータの重み係数を決定する。

【００３１】音声分析部１０及び１１０における音素記号系列の生成処理においては、上述した通り、本実施形態では、少なくとも録音内容が正書法で記述されたものがあれば、あらゆる音声波形信号データベースが合成用の音声波形信号データとして利用可能である。入力として単語の情報だけが与えられた場合には辞書や規則に基づいて音素系列を予測する必要がある。また、音声分析部１０及び１１０における音素のアラインメント処理においては、読み上げ音声の場合、各単語がそれぞれの標準の発音に近く発音されることが多く、躊躇したり、言い淀んだりすることもまれである。このような音声波形信号データの場合には簡単な辞書検索によって音素ラベリングが正しく行なわれ、音素アラインメント用の音素HMMの音素モデルの学習が可能となる。

【００３２】音素アラインメント用の音素モデルの学習では完全な音声認識の場合と異なり、学習用の音声波形信号データとテスト用の音声波形信号データとを完全に分離する必要はなく、すべての音声波形信号データを用いて学習を行なうことができる。まず、別の話者用のモデルを初期モデルとし、すべての単語について標準発音

か限られた発音変化のみを許し、適切なセグメンテーションが行なわれるように、全音声波形信号データを用いてビタビの学習アルゴリズムを用いて音素のアライメントを行ない、特徴パラメータの再推定を行なう。単語間のポーズは単語間ポーズ生成規則によって処理するが、単語内にポーズがあってアライメントが失敗した場合には人手により修正する必要がある。

【００３３】どういう音素ラベルを音素表記として用いるかは選択が必要である。もし良く学習されたHMMモデルが利用できるような音素セットが存在するなら、それを用いることが有利である。反対に、音声合成装置が完全な辞書を持っているなら、音声波形信号データベースのラベルを完全に辞書と照合する方法も有効である。我々は、重み係数の学習に対して選択の余地があるから、後で音声合成装置が予測したものと等価なものを音声波形信号データベースの中から照合できるかどうかを最も重要な基準とすれば良い。発音の微妙な違いはその発音の韻律的環境によって自動的に把握されるため、特に手作業で音素のラベル付けを行なう必要はない。

【００３４】前処理の次の段階として、個々の音素の調音的な特徴を記述するための韻律特徴パラメータの抽出を行なう。従来の音声学では、調音位置や調音様式といった素性で言語音を分類した。これに対して、ファース（Ｆｉｒｔｈ）学派のような韻律を考慮した音声学では、韻律的文脈の違いから生ずる細かな音質の違いをとらえるために、明瞭に調音されている箇所や強調が置かれている箇所を区別する。これらの違いを記述する方法はいろいろなものがあるが、ここでは以下の２つの方法を用いる。まず低次のレベルでは、１次元の特徴を求めるために、パワー、音素時間長の伸び及び音声基本周波数F_0を、ある音素について平均した値を用いる。一方、高次のレベルでは、韻律特徴における上記の違いを考慮した韻律境界や強調箇所をマークする方法を用いる。これらの２種類の特徴は相互に密接に関係しているため一方から他方を予測することができるが、両者は共に各音素の特徴に強い影響を与えている。

【００３５】音声波形信号データベースを記述するための音素セットの規定法に自由度があるのと同様に、韻律的特徴パラメータの記述方法についても自由度があるが、これらの選び方は音声合成装置の予測能力に依存する。もし音声波形信号データベースが予めラベリングされているなら、音声合成装置の仕事は内部表現から音声波形信号データベース中の実音声をいかに行なうかを適切に学習することである。これに対して、もし音声波形信号データベースが音素のラベル付けがなされていないなら、どのような特徴パラメータを使えば音声合成装置が最も適切な音声単位を予測できるかから検討することが必要となる。この検討及び最適な特徴パラメータの重みの決定学習は、各特徴パラメータに対する重み係数を学習しながら決定する重み係数学習部１１において実行される。

【００３６】次いで、重み係数学習部１１によって実行される重み係数学習処理について述べる。与えられた目標音声の音響的及び韻律的な環境に最適なサンプルを音声波形信号データベースから選択するために、まずどの特徴がどれだけ寄与しているかを音素的及び韻律的な環境の違いによって決める必要がある。これは音素の性質によって重要な特徴パラメータの種類が変化するためで、例えば、音声基本周波数F_0は有声音の選択には極めて有効であるが、無声音の選択にはほとんど影響がない。また、摩擦音の音響的特徴は前後の音素の種類によって影響が変わる。最適な音素を選択するためにそれぞれの特徴にどれだけの重みを置くかを最適重み決定処理、すなわち重み係数学習処理で自動的に決定する。

【００３７】重み係数学習部１１によって実行される最適重み係数の決定処理で、最初に行なわれることは音声波形信号データベース中で該当するすべての発話サンプルの中から最適なサンプルを選ぶときに使われる特徴をリストアップすることである。ここでは、調音位置や調音様式等の音素的特徴と先行音素、当該音素、及び後続音素の音声基本周波数F_0、音素時間長、パワー等の韻律的特徴パラメータ等を用いる。具体的には、詳細後述する第２の韻律的特徴パラメータを用いる。次いで、第２段階では各音素毎に、最適な候補を選ぶ際にどの特徴パラメータがどれだけ重要かを決定するために、１つの音声サンプル（又は音素の音声波形信号）に着目し、他のすべての音素サンプルとの音素時間長の差をも含む音響的距離を求め、上位Ｎ２個の最良の類似音声サンプル、すなわちＮ２ベストの音素候補の音声波形信号の音声セグメントを選び出す。

【００３８】さらに、第３段階では線形回帰分析を行ない、それらの類似音声サンプルを用いて種々の音響的及び韻律的環境におけるそれぞれの特徴パラメータの重要度を示す重み係数を求める。当該線形回帰分析処理における韻律的特徴パラメータとして、例えば、次の特徴パラメータ（以下、第２の韻律的特徴パラメータという。）を用いる。

（１）処理すべき当該音素から１つだけ先行する先行音素（以下、先行音素という。）の第１の韻律的特徴パラメータ；

（２）処理すべき当該音素から１つだけ後続する後続音素（以下、後続音素という。）の音素ラベルの第１の韻律的特徴パラメータ；

（３）当該音素の音素時間長；

（４）当該音素の音声基本周波数F_0；

（５）先行音素の音声基本周波数F_0；及び、

（６）後続音素の音声基本周波数F_0。

【００３９】ここで、先行音素は、当該音素から１つだけ先行する音素としているが、これに限らず、複数の音素だけ先行する音素を含んでもよい。また、後続音素

は、当該音素から１つだけ後続する音素としているが、これに限らず、複数の音素だけ後続する音素を含んでもよい。さらに、後続音素の音声基本周波数F_0を除外してもよい。

【００４０】以上の実施形態においては、線形回帰分析を行って、重み係数を求めているが、本発明はこれに限らず、例えば、所定のニューラルネットワークを用いた統計的解析などの種々の統計的解析を用いて、重み係数を求めてもよい。

【００４１】次いで、自然な音声サンプルの選択を行う音声単位選択部１２の処理について説明する。従来例の音声合成装置では目的の発話に対して音素系列を決定し、さらに韻律制御のためのF_0と音素時間長の目標値が計算された。これに対して、本実施形態では最適の音声サンプルを適切に選択するために韻律が計算されるだけで、直接韻律を制御することは行なわれない。

【００４２】図３は、図１の音声単位選択部１２の処理の入力は、目的発話の音素系列と、それぞれの音素毎に求めた各特徴パラメータに対する重みベクトル及び音声波形信号データベース中の全サンプルを表す特徴ベクトルである。一方、出力は音声波形信号データベース中での音素サンプルの位置を表す索引情報であって、音声波形信号の音声セグメントを接続するためのそれぞれの音声単位（具体的には音素、場合により複数の音素の系列が連続して選択され、一つの音声単位となることがある）の開始位置と音声単位時間長を示したものである。

【００４３】最適な音声単位は目的発話との差の近似コストを表す目標コストと、隣接音声単位間での不連続性の近似コストを表す連結コストの和を最小化するパスとして求められる。経路探索には公知のビタビの学習アルゴリズムが利用される。目的とする目標音声$t_1^n=(t_1, \cdots, t_n)$に対しては、目標コストと連結コストの和を最小化することで、各特徴が目的音声に近く、しかも音声単位間の不連続性が少ない音声波形信号データベース中の音声単位の組合せ$u_1^n=(u_1, \cdots, u_n)$を選ぶことができ、これらの音声単位の音声波形信号データベース内での位置を示すことにより、任意の発話内容の音声合成が可能になる。

【００４４】音声単位の選択コストは、図３に示すように、目標コスト$C^t(u_i, t_i)$と連結コスト$C^c(u_{i-1}, u_i)$からなり、目標コスト$C^t(u_i, t_i)$は、音声波形信号データベース中の音声単位（音素候補）u_iと、合成音声として実現したい音声単位（目標音素）t_iの間の差の予測値であり、連結コスト$C^c(u_{i-1}, u_i)$は接続単位（接続する２つの音素）u_{i-1}とu_iとの間の接続で起こる不連続の予測値である。例えば、本出願人によって研究実用化された従来のＡＴＲν－Ｔａｌｋ音声合成システムも目標コストと連結コストを最小化するという点では類似の考え方を取っていたが、韻律的な特徴パラメータを直接に単位選択に用いるということは本実施形態の音声合成装置の新しい特徴となっている。

【００４５】次いで、コストの計算について述べる。目標コストは実現したい音声単位の特徴ベクトルt_iと音声波形信号データベース中から選ばれた候補の音声単位の特徴ベクトルu_iの各要素の差の重み付き合計であり、各目標サブコスト$C^t_j(t_i, u_i)$の重み係数w^t_jが与えられた場合、目標コスト$C^t(t_i, u_i)$は次式で計算することができる。

【００４６】

【数１】

$$C^t(t_1, u_1) = \sum_{j=1}^{p} w^t_j C^t_j(t_1, u_1)$$

【００４７】ここで、特徴ベクトルの各要素の差はｐ個の目標サブコスト$C^t_j(t_i, u_i)$（ただし、ｊは１からｐまでの自然数である。）で表され、特徴ベクトルの次元数ｐは、好ましい実施例においては、２０から３０の範囲で可変としている。より好ましい実施例においては、次元数ｐ＝３０であり、目標サブコスト$C^t(t_i, u_i)$及び重み係数w^t_jにおける変数ｊの特徴ベクトル又は特徴パラメータは、上述の第２の韻律的特徴パラメータである。

【００４８】一方、連結コスト$C^c(u_{i-1}, u_i)$も同様にｑ個の連結サブコスト$C^c_j(u_{i-1}, u_i)$（ただし、ｊは１からｑまでの自然数である。）の重み付き合計で表される。連結サブコストは接続する音声単位u_{i-1}とu_iの音響的特徴から決定することができる。好ましい実施形態においては、連結サブコストとしては、（１）音素接続点におけるケプストラム距離、（２）対数パワーの差の絶対値、（３）音声基本周波数F_0の差の絶対値の３種類を用いており、すなわち、ｑ＝３である。これら３種類の音響的特徴パラメータと、先行音素の音素ラベルと、後続音素の音素ラベルとを、第３の音響的特徴パラメータという。各連結サブコスト$C^c_j(u_{i-1}, u_i)$の重みw^c_jは予め経験的に（又は実験的に）与えられ、この場合、連結コスト$C^c(u_{i-1}, u_i)$は次式で計算することができる。

【００４９】

【数２】

$$C^c(u_{1-1}, u_1) = \sum_{j=1}^{q} w^c_j C^c_j(u_{1-1}, u_1)$$

【００５０】もし、音素候補u_{i-1}とu_iが音声波形信号データベース中の連続する音声単位であった場合には、接続は自然であり、連結コストは０になる。ここで、好ましい実施例においては、連結コストは、特徴パラメータメモリ３０内の第１の音響的特徴パラメータと第１の

韻律的特徴パラメータに基づいて決定され、連続量である上記３つの第３の音響的特徴パラメータを取り扱うから例えば０から１までの任意のアナログ量をとる一方、目標コストは、それぞれの先行あるいは後続音素の弁別素性が一致するか否かなどを示す上記３０個の第２の韻律的特徴パラメータを取り扱うから、例えば０（特徴が一致しているとき）又は１（特徴が一致していないとき）のデジタル量で表される要素を含む。そして、Ｎ個の音声単位の連結コストはそれぞれの音声単位の目標コストと連結コストの和となり、次式で表される。

【００５１】

【数３】

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S)$$

＊

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{p} C^t(t_i, u_i) + \sum_{i=2}^{n}\sum_{j=1}^{q} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S)$$

＊【００５２】このとき、Ｓはポーズを表しており、C^c（Ｓ，u_1）及びC^c（u_n，Ｓ）はポーズから最初の音声単位へ及び最後の音声単位からポーズへの接続における連結コストを表している。この表現からも明らかなように、本実施形態ではポーズも音声波形信号データベース中の他の音素とまったく同じ扱い方をしている。さらに上の式をサブコストで直接表現すると次式のようになる。

【００５３】

【数４】

【００５４】音声単位選択処理は上式で決まる全体のコストを最小にするような音声単位の組合せ／u_1^nを決定するためのものである。ここで、日本出願の明細書では、オーバーラインを記述することができないために、オーバーラインの代わりに／を用いる。

【００５５】

【数５】$$/u_1^n = \min_{u_1, u_2, \cdots, u_n} C(t_1^n, u_1^n)$$

【００５６】上記数５において、関数ｍｉｎは、当該関数の引数である$C(t_1^n, u_1^n)$を最小にする音素候補の組み合わせ（すなわち、音素列候補）$u_1, u_2, \cdots, u_n$＝／u_1^nを表わす関数である。

【００５７】図１の重み係数学習部１１における重み係数の学習処理について以下説明する。目標サブコストの重みは音響的距離に基づく線形回帰分析を用いて決定する。重み係数の学習処理ではすべての音素毎に異なる重み係数を決めることもできるし、音素カテゴリ（例えば、すべての鼻音）毎に重み係数を決めることもできる。また、すべての音素について共通の重み係数を決めることもできるが、ここでは各音素で別々の重み係数を用いることとする。特徴パラメータメモリ３０内のデータベースにおける各トークン（又は各音声サンプル）は、各トークンの音響的特徴に関係する第１の音響的特徴パラメータと第１の韻律的特徴パラメータの組で記述されている。重み係数は、第１の音響的特徴パラメータと第１の韻律的特徴パラメータの各パラメータと、トークン又はコンテキストにおける音素の第２の音響的特徴パラメータにおける差又は音響的距離との間の関係の強さ（寄与度）を決定するために学習される。

【００５８】以下に線形回帰分析における処理の流れを示す。

【００５９】＜１＞現在学習を行なっている音素種類（又は音素カテゴリ）に属する音声波形信号データベース中のすべてのサンプルについて繰り返し以下の４つの処理（ａ）乃至（ｄ）を実行する。

（ａ）取り上げた音声サンプルを目的の発話内容と見なす。

（ｂ）音声波形信号データベース中の同一の音素種類（カテゴリ）に属する他のすべてのサンプルと当該音声サンプルとの音響的距離を計算する。

（ｃ）目標音素に近いもの上位Ｎ１個（例えば、Ｎ１＝２０個である。）の最良の音素候補を選び出す。

（ｄ）目標音素自身t_iと上記（ｃ）で選んだ上位Ｎ１個のサンプルについて目標サブコストC^t_j（t_i，u_i）を求める。

＜２＞すべての目標音素t_iと上位Ｎ１個の最適サンプルについて音響的距離と目標サブコストC^t_j（t_i，u_i）を求める。

＜３＞ｐ個の目標サブコストに対して線形回帰分析を実行することにより、上記目標音素を表わす第１の音響的特徴パラメータと第１の韻律的特徴パラメータの各特徴パラメータにおける寄与度を予測して、当該音素種類

（カテゴリ）に対する、ｐ個の目標サブコストの線形重み係数を求める。
この重み係数を用いて上記コストを計算する。そして、＜１＞から＜３＞までの処理をすべての音素種類（カテゴリ）について繰り返す。
【００６０】もし仮に目的音声単位の音響的距離が直接求められた場合に最も近い音声サンプルを選び出すためにはそれぞれの目標サブコストにどのような重み係数をかければ良いのかを決定するのが、この重み係数学習部１１の目的である。本実施形態の利点は音声波形信号データベース中の音声波形信号の音声セグメントを直接的に利用できることである。
【００６１】さらに、音声単位選択部１２は、英語の重み係数ベクトルメモリ３１に記憶された各目標音素毎の重み係数ベクトルと、英語の特徴パラメータメモリ３０に記憶された第１の韻律的特徴パラメータとに基づいて、入力される英語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する。そして、音声合成部１３は、音声単位選択部１２から出力される索引情＊報に基づいて、当該索引情報に対応する英語の音声波形信号の音声セグメントを英語の音声波形信号データベースメモリ２１から逐次読み出して連結してバッファメモリ１４を介してケプストラム抽出部１１１に出力する。
【００６２】以上のように構成された英語音声による音声合成装置１では、目標コストとして実現したい音声単位の基本周波数、音韻継続長、対数パワーなどを要素とした特徴ベクトルと、メモリ２１内の音声波形データーベース中から選ばれた候補の音声単位の特徴ベクトルの差を用いる。
【００６３】次いで、日本語音声による音声合成装置２の構成及び動作について詳述する。例えば、英語の音声波形データベースで他言語である日本語の音声を合成する場合、合成する日本語の音素が音声波形データベースに無い場合がある。そこで，各言語毎に、公知の通り国際的に取り決められたＩＰＡ（the International Phonetic Alphabet）で定義される調音位置や調音様式による分類を用いて音素をクラスタリングし、各クラスタに所属する音素を合成のための音素候補とする。次の表は、日本語の音素を調音様式で分類した例を示す。
【００６４】
【表３】

日本語音素分類

調音様式	日本語の音素例
母音	／ａ／，／ｉ／，／ｕ／，／ｅ／，／ｏ／
半母音	／ｊ／，／ｗ／
破裂音	／ｐ／，／ｂ／，／ｔ／，／ｄ／，／ｋ／，／ｇ／
摩擦音	／ｆ／，／ｈ／，／ｓ／，／ｓｈ／，／ｚ／，／ｒ／
破擦音	／ｔｓ／
鼻音	／ｎ／，／ｍ／

【００６５】音声単位選択部１１２の目標コストの計算では、目標として音声合成装置１で使われていた韻律情報ではなく、音声合成装置１で音声合成した英語音声信号をケプストラム抽出部１１１によって抽出されたケプストラム情報を用いて音素候補を選択することを特徴としている。
【００６６】日本語の音声波形データベースメモリ１２１は、英語の音声波形データベースメモリ２１と同様に日本語の音声波形データベースを予め記憶し、日本語のテキストデータベースメモリ１２２は、英語のテキストデータベースメモリ２２と同様に、日本語の音声波形データベースメモリ１２１の内容と対応して日本語のテキストデータベースを予め記憶し、日本語の音素HMMメモリ１２３は、英語の音素HMMメモリ２３と同様に、日本語音素HMMを予め記憶する。音声分析部１１０は、音声分析部１０の処理と同様に動作して、上記第１の音響的特徴パラメータと上記第１の韻律的特徴パラメータを抽出して日本語の特徴パラメータメモリ１３０に記憶する。
【００６７】音声単位選択部１１２の処理の入力は、目的発話の英語の音素系列と、目標特徴パラメータとなるケプストラム抽出部１１１から入力されたケプストラム係数データと、メモリ１３０内の音声波形信号データベース中の全サンプルを表す特徴ベクトルである。一方、出力は音声波形信号データベース中での音素サンプルの位置を表す索引情報であって、音声波形信号の音声セグメントを接続するためのそれぞれの音声単位（具体的には音素、場合により複数の音素の系列が連続して選択され、一つの音声単位となることがある）の開始位置と音声単位時間長を示したものである。
【００６８】最適な音声単位は、目的発話との差の近似コストを表す目標コストと、隣接音声単位間での不連続

性の近似コストを表す連結コストの和を最小化するパスとして求められる。経路探索には公知のビタビの学習アルゴリズムが利用される。目的とする目標音声$t_1^n=(t_1, \cdots, t_n)$に対しては、目標コストと連結コストの和を最小化することで、各特徴が目的音声に近く、しかも音声単位間の不連続性が少ない音声波形信号データベース中の音声単位の組合せ$u_1^n=(u_1, \cdots, u_n)$を選ぶことができ、これらの音声単位の音声波形信号データベース内での位置を示すことにより、任意の発話内容の音声合成が可能になる。

【００６９】音声単位の選択コストは、図３に示すように、目標コスト$C^t(u_i, t_i)$と連結コスト$C^c(u_{i-1}, u_i)$からなり、目標コスト$C^t(u_i, t_i)$は、音声波形信号データベース中の音声単位（音素候補）u_iと、合成音声として実現したい音声単位（目標音素であり、音声合成装置２では、ケプストラム係数データを用いる。））t_iの間の差の予測値であり、連結コスト$C^c(u_{i-1}, u_i)$は接続単位（接続する２つの音素）u_{i-1}とu_iとの間の接続で起こる不連続の予測値である。例えば、本特許出願人によって研究実用化された従来のＡＴＲν－Ｔａｌｋ音声合成システムも目標コストと連結コストを最小化するという点では類似の考え方を取っていたが、韻律的な特徴パラメータを直接に単位選択に用いるということは本実施形態の音声合成装置の新しい特徴となっている。

【００７０】次いで、コストの計算について述べる。目標コストは実現したい音声単位の特徴ベクトル（ケプストラム係数データ）t_iと音声波形信号データベース中から選ばれた候補の音声単位の特徴ベクトル（ケプストラム係数データ）u_iの各要素の差の重み付き合計であり、各目標サブコスト$C^t_j(t_i, u_i)$の重み係数w^t_jが与えられた場合、目標コスト$C^t(t_i, u_i)$は次式で計算することができる。

【００７１】

【数６】

$$C^t(t_i, u_i)=\sum_{j=1}^{p} C^t_j(t_i, u_i)$$

【００７２】ここで、特徴ベクトルの各要素の差はｐ個の目標サブコスト$C^t_j(t_i, u_i)$（ただし、ｊは１からｐまでの自然数である。）で表され、特徴ベクトルの次元数ｐは、好ましい実施例においては、２０から３０の範囲で可変としている。より好ましい実施形態においては、次元数ｐ＝３０であり、目標サブコスト$C^t(t_i, u_i)$の特徴パラメータは、ケプストラム係数データである。

【００７３】一方、連結コスト$C^c(u_{i-1}, u_i)$も同様にｑ個の連結サブコスト$C^c_j(u_{i-1}, u_i)$（ただし、ｊは１からｑまでの自然数である。）の重み付き合計で表される。連結サブコストは接続する音声単位u_{i-1}とu_iの音響的特徴から決定することができる。好ましい実施形態においては、連結サブコストとしては、（１）音素接続点におけるケプストラム距離、（２）対数パワーの差の絶対値、（３）音声基本周波数F_0の差の絶対値の３種類を用いており、すなわち、ｑ＝３である。各連結サブコスト$C^c_j(u_{i-1}, u_i)$の重みw^c_jは予め経験的に（又は実験的に）与えられ、この場合、連結コスト$C^c(u_{i-1}, u_i)$は次式で計算することができる。

【００７４】

【数７】

$$C^c(u_{i-1}, u_i)=\sum_{j=1}^{q} w^c_j C^c_j(u_{i-1}, u_i)$$

【００７５】もし、音素候補u_{i-1}とu_iが音声波形信号データベース中の連続する音声単位であった場合には、接続は自然であり、連結コストは０になる。ここで、好ましい実施例においては、連結コストは、特徴パラメータメモリ３０内の第１の音響的特徴パラメータと第１の韻律的特徴パラメータに基づいて決定され、連続量である上記３つの第３の音響的特徴パラメータを取り扱うから例えば０から１までの任意のアナログ量をとる一方、目標コストは、それぞれの先行あるいは後続音素の弁別素性が一致するか否かなどを示す上記３０個の第２の韻律的特徴パラメータを取り扱うから、例えば０（特徴が一致しているとき）又は１（特徴が一致していないとき）のデジタル量で表される要素を含む。そして、Ｎ個の音声単位の連結コストはそれぞれの音声単位の目標コストと連結コストの和となり、次式で表される。

【００７６】

【数８】

$$C(t_1^n, u_1^n)=\sum_{i=1}^{n} C^t(t_i, u_i)+\sum_{i=2}^{n} C^c(u_{i-1}, u_i)+C^c(S, u_1)+C^c(u_n, S)$$

【００７７】このとき、Ｓはポーズを表しており、$C^c(S, u_1)$及び$C^c(u_n, S)$はポーズから最初の音声単位へ及び最後の音声単位からポーズへの接続における連結コストを表している。この表現からも明らかなように、本実施形態ではポーズも音声波形信号データベース中の他の音素とまったく同じ扱い方をしている。さらに上の式をサブコストで直接表現すると次式のようになる。

【００７８】

【数９】

29 30

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{p} C^t(t_i, u_i) + \sum_{i=2}^{n}\sum_{j=1}^{q} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S)$$

【００７９】音声単位選択処理は上式で決まる全体のコストを最小にするような音声単位の組合せ／u_1^nを決定するためのものである。ここで、日本出願の明細書では、オーバーラインを記述することができないために、オーバーラインの代わりに／を用いる。

【００８０】

【数１０】$$／u_1^n = \min_{u_1, u_2, \cdots, u_n} C(t_1^n, u_1^n)$$

【００８１】上記数１０において、関数ｍｉｎは、当該関数の引数であるＣ（t_1^n, u_1^n）を最小にする音素候補の組み合わせ（すなわち、音素列候補）$u_1, u_2, \cdots, u_n$＝／u_1^nを表わす関数である。従って、音声単位選択部１１２は、ケプストラム抽出部１１１から入力されるケプストラム係数データと、日本語の特徴パラメータメモリ３０に記憶された第１の韻律的特徴パラメータとに基づいて、入力される英語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する。そして、音声合成部１１３は、音声単位選択部１１２から出力される索引情報に基づいて、当該索引情報に対応する日本語の音声波形信号の音声セグメントを日本語の音声波形信号データベースメモリ１２１から逐次読み出して連結してスピーカ１１４を介して出力することにより、英語の入力音素列に基づいて、日本語の音声波形データベースによる日本人の声による英語の音声合成信号の音声を出力することができる。

【００８２】以上のように構成された図１の音声合成システムにおいて、音声分析部１０及び１１０と、重み係数学習部１１と、ケプストラム抽出部１１１と、音声単位選択部１２及び１１２と、音声合成部１１３とは、例えば、マイクロプロセッシングユニット（ＭＰＵ）などのデジタル計算機又は演算制御装置によって構成される一方、テキストデータベースメモリ２２及び１２２と、音素HMMメモリ２３及び１２３と、特徴パラメータメモリ３０及び１３０と、重み係数ベクトルメモリ３１とは例えばハードディスクなどの記憶装置で構成される。ここで、好ましい実施例においては、音声波形信号データベースメモリ２１及び１２１は、ＣＤ－ＲＯＭの形式の記憶装置である。

【００８３】以下、以上のように構成された図１の音声合成装置の各処理部１０乃至１３における処理について説明する。

【００８４】図４は、図１の音声分析部１０によって実行される音声分析処理のフローチャートである。図４において、まず、ステップＳ１１で、音声波形信号データベースメモリ２１から自然発話の音声波形信号の信号を入力してＡ／Ｄ変換してデジタル音声波形信号データに変換するとともに、当該音声波形信号の音声文を書き下したテキストデータをテキストデータベースメモリ２２内のテキストデータベースから入力する。ここで、テキストデータはなくてもよく、ない場合は、音声波形信号から公知の音声認識装置を用いて音声認識してテキストデータを得てもよい。なお、Ａ／Ｄ変換した後のデジタル音声波形信号データは、例えば１０ミリ秒毎の音声セグメントに分割されている。そして、ステップＳ１２で、音素列が予測されているか否かが判断され、音素列が予測されていないときは、ステップＳ１３で例えば音素HMMを用いて音素列を予測して記憶した後、ステップＳ１４に進む。ステップＳ１２で音素列が予測されている又は予め与えられている、もしくは手作業で音素ラベルが付与されているときは、直接にステップＳ１４に進む。

【００８５】ステップＳ１４では、各音素セグメントに対する、音声波形信号の複数の文又は１つの文からなるファイルにおける開始位置と終了位置を記録し、当該ファイルに索引番号を付与する。次いで、ステップＳ１５では、各音素セグメントに対する上記第１の音響的特徴パラメータを例えば公知のピッチ抽出法を用いて抽出する。そして、ステップＳ１６では、各音素セグメントに対して音素ラベル付けを実行して、音素ラベルとそれに対する第１の音響的特徴パラメータを記録する。さらに、ステップＳ１７では、各音素セグメントに対する第１の音響的特徴パラメータと、音素ラベルと、音素ラベルに対する上記第１の韻律的特徴パラメータを、ファイルの索引番号と、ファイル内の開始位置と時間長とともに、特徴パラメータメモリ３０に記憶する。最後に、ステップＳ１８で、各音素セグメントに対して、ファイルの索引番号とファイル内の開始位置と時間長とを含む索引情報を付与して、当該索引情報を特徴パラメータメモリ３０に記憶して、当該音声分析処理を終了する。

【００８６】また、音声分析部１１０は、図４の音声分析処理と同様の処理を日本語について実行する。

【００８７】図５及び図６は、図１の重み係数学習部１１によって実行される重み係数学習処理のフローチャートである。図５において、まず、ステップＳ２１で、特

徴パラメータメモリ３０から１個の音素種類を選択する。次いで、ステップＳ２２で、選択された音素種類と同一の音素種類を有する音素の第１の音響的特徴パラメータから第２の音響的特徴パラメータを取り出して目標音素の第２の音響的特徴パラメータとする。そして、ステップＳ２３で、同一の音素種類を有する目標音素以外の残りの音素と、第２の音響的特徴パラメータにおける目標音素との間の、音響的距離であるユークリッドケプストラム距離と、底を２とする対数音素時間長とを計算する。ステップＳ２４では、すべての残りの音素についてステップＳ２２及びＳ２３の処理をしたか否かが判断され、処理が完了していないときは、ステップＳ２５で別の残りの音素を選択してステップＳ２３からの処理を繰り返す。

【００８８】一方、ステップＳ２４で処理が完了しているときは、ステップＳ２６で、ステップＳ２３で得られた距離及び時間長に基づいて、上位Ｎ１個の最良の音素候補を選択する。次いで、ステップＳ２７で選択された上位Ｎ１個の最良の音素候補について１番目からＮ１番目までランク付けする。そして、ステップＳ２８で、ランク付けされたＮ１個の最良の音素候補に対して各距離から中間値を引いてスケール変換値を計算する。そして、ステップＳ２９において、すべての音素種類及び音素についてステップＳ２２からＳ２８までの処理を完了したか否かが判断され、完了していないときは、ステップＳ３０で別の音素種類又は音素を選択した後、ステップＳ２２からの処理を繰り返す。一方、ステップＳ２９で処理が完了しているときは、図６のステップＳ３１に進む。

【００８９】図６において、ステップＳ３１では、１個の音素種類を選択する。次いで、ステップＳ３２では、選択された音素種類に対して各音素の第２の音響的特徴パラメータを抽出する。そして、ステップＳ３３で、選択された音素種類に対するスケール変換値に基づいて線形回帰分析を行うことにより、各第２の音響的特徴パラメータにおけるスケール変換値に対する寄与度を計算し、計算された寄与度を目標音素毎の重み係数として重み係数ベクトルメモリ３１に記憶する。ステップＳ３４では、すべての音素種類について上記ステップＳ３２及びＳ３３の処理を完了したか否かが判断され、完了していないときは、ステップＳ３５で別の音素種類を選択した後、ステップＳ３２からの処理を繰り返す。一方、ステップＳ３４で処理が完了しているときは、当該重み係数学習処理を終了する。なお、各第２の韻律的特徴パラメータにおける寄与度は経験的に（又は実験的に）予め与えられて、当該寄与度を目標音素毎の重み係数ベクトルとして重み係数ベクトルメモリ３１に記憶する。

【００９０】図７は、図１の音声単位選択部１２によって実行される音声単位選択処理のフローチャートである。図７において、まず、ステップＳ４１で、入力された音素列のうち最初から１個目の音素を選択する。次いで、ステップＳ４２で、選択された音素と同一の音素種類を有する音素の重み係数ベクトルを重み係数ベクトルメモリ３１から読み出し、目標サブコスト及び必要な特徴パラメータを特徴パラメータメモリ３０から読み出してリストアップする。そして、ステップＳ４３ですべての音素について処理したか否かが判断され、完了していないときはステップＳ４４で次の音素を選択した後、ステップＳ４２の処理を繰り返す。一方、ステップＳ４３で完了していないときは、ステップＳ４５に進む。

【００９１】ステップＳ４５では、入力された音素列に対して数４を用いて各音素候補における全体のコストを計算する。次いで、ステップＳ４６では、計算されたコストに基づいて、上位Ｎ２個の最良の音素候補をそれぞれの目標音素に対して選択する。そして、ステップＳ４７では、数５を用いてビタビサーチにより、全体のコストを最小にする音素候補の組み合わせの索引情報と、その各音素の開始時刻と時間長とともに検索した後、音声合成部１３に出力して、当該音声単位選択処理を終了する。

【００９２】図８は、図１の音声単位選択部１１２によって実行される音声単位選択処理のフローチャートである。図８において、まず、ステップＳ５１で、入力された音素列のうち最初から１個目の音素を選択する。次いで、ステップＳ５２で、選択された音素と同一の音素種類を有する音素のケプストラム係数データをケプストラム抽出部１１１から入力し、目標サブコスト及び必要な特徴パラメータを特徴パラメータメモリ１３０から読み出してリストアップする。そして、ステップＳ５３ですべての音素について処理したか否かが判断され、完了していないときはステップＳ５４で次の音素を選択した後、ステップＳ５２の処理を繰り返す。一方、ステップＳ５３で完了していないときは、ステップＳ５５に進む。

【００９３】ステップＳ５５では、入力された音素列に対して数８を用いて各音素候補における全体のコストを計算する。次いで、ステップＳ５６では、計算されたコストに基づいて、上位Ｎ２個の最良の音素候補をそれぞれの目標音素に対して選択する。そして、ステップＳ５７では、数１０を用いてビタビサーチにより、全体のコストを最小にする音素候補の組み合わせの索引情報と、その各音素の開始時刻と時間長とともに検索した後、音声合成部１１３に出力して、当該音声単位選択処理を終了する。

【００９４】さらに、音声合成部１１３は、音声単位選択部１１２から出力される索引情報と、その各音素の開始時刻と時間長とに基づいて、音声波形信号データベースメモリ１２１に対してアクセスして単位選択された音素候補のデジタル音声波形信号データを読み出して、逐次Ｄ／Ａ変換して変換後のアナログ音声信号をスピーカ

１１４を介して出力する。これにより、入力された英語の音素列に対応する日本語の音声波形データベースによる日本人の声により音声合成された音声がスピーカ１１４から出力される。

【００９５】本実施形態においては、音声波形信号の圧縮や音声基本周波数F_0や音素時間長の修正は不要になったが、代わって音声サンプルを注意深くラベル付けし、大規模な音声波形信号データベースの中から最適なものを選択することが必要となる。本実施形態の音声合成方法の基本単位は音素であり、これは辞書やテキスト－音素変換プログラムで生成されるが、同一の音素であっても音声波形信号データベース中に音素の十分なバリエーションを含んでいることが要求される。音声波形信号データベースからの音声単位選択処理では目的の韻律的環境に適合し、しかも接続したときに隣接音声単位間での不連続性が最も低い音素サンプルの組合せが選ばれる。このために、音素毎に各特徴パラメータの最適重み係数が決定される。

【００９６】本実施形態の音声合成装置の特徴は、次の通りである。

＜単位選択基準としての韻律的情報の利用＞

スペクトル的特徴は韻律的特徴と不可分であるとの立場から、音声単位の選択基準に韻律的な特徴を導入した。

＜音響的及び韻律的特徴パラメータの重み係数の自動学習＞

音素環境や音響的特徴、韻律的特徴等の各種の特徴量が音声単位の選択にどれだけの寄与があるかを音声波形信号データベース中の全音声サンプルを利用することで自動的に決定し、コーパスを基本とする音声合成装置を構築した。

＜音声波形信号の直接接続＞

上記の自動学習により、大規模音声波形信号データベースから最適な音声サンプルを選び出すことにより、何らの信号処理も利用しない任意音声合成装置を構築した。

＜音声波形信号データベースの外部情報化＞

音声波形信号データベースを完全に外部情報として取り扱うことにより、単にＣＤ－ＲＯＭ等に記憶した音声波形信号データを取り替えることで任意の言語、任意の話者に利用できる音声合成装置を構築した。

【００９７】以上説明したように、本実施形態によれば、英語音声による音声合成装置１と、日本語音声による音声合成装置２とを備えて音声合成システムを構成したので、英語の音声波形データベースを用いて英語の音声合成の音声波形データを発生した後、それに基づいて日本語の音声波形データベースを用いた英語の音声合成の音声を得ることができるので、例えば、日本語のネイティブの日本人の声による英語の音声合成の音声を得ることができる。

【００９８】＜変形例＞

図１の音声合成システムにおいて用いる英語の音声波形データベースは、以下のように、図７の話者選択装置２００で予め話者選択されたものであることが好ましい。図７において、話者選択部２００には、互いに異なる性別及び異なる人の複数Ｎ個の英語の音声波形データベースを記憶した音声波形データベースメモリ２１－１乃至２１－Ｎがスイッチ ＳＷを介して接続されるとともに、日本語の音声波形データベースメモリ１２１が接続される。話者選択部２０１は、音声合成したい日本語の音声波形データベースの声質に近い英語の音声波形データベースを選択して音声合成部１３に接続する。選択基準として、性別、基本周波数のレンジ、音韻継続長などがあげられる。ここで、選択基準として用いることが好ましいのは、性別及び基本周波数の平均値である。

【００９９】話者選択部２０１は、複数Ｎの英語の音声波形データベースの各登録話者に対して、スイッチ ＳＷを順次切り換えて、動的計画法マッチング（ＤＴＷ）で時間整合したメモリ１２１内の目的話者の学習音声スペクトル時系列と、メモリ２１－１乃至２１－Ｎ内の各登録話者の学習音声スペクトル時系列との距離（すなわち、音響的特徴パラメータの距離）を求め、２乗誤差最小基準により最も距離の小さい登録話者を選択する。そして、話者選択部２０１は、スイッチ ＳＷを制御して、選択した登録話者の英語の音声波形データベースメモリ２１を音声合成部１３に接続する。当該変形例では、日本語話者の自然発話の声質に近い英語話者の自然発話の音声合成を得ることができ、これにより、より声質が近くより自然な発話による日本人の声による英語の音声合成の音声を得ることができる。

【０１００】以上の実施形態においては、英語音声による音声合成装置１と、日本語音声による音声合成装置２とを備えて音声合成システムを構成し、英語の音声波形データベースを用いて英語の音声合成の音声波形データを発生した後、それに基づいて日本語の音声波形データベースを用いた英語の音声合成の音声を得ることにより、日本語のネイティブの日本人の声による英語の音声合成の音声を得ている。本発明はこれに限らず、２つの言語は英語と日本語に限らず、異なる２つの言語であってもよい。従って、音声合成部１３の前に、公知の音声認識装置を接続することにより、英語の自然発話文の音声を音声認識した後、本実施形態の音声合成システムにより、元の音声に対応する、日本人の声による英語の音声合成の音声を発生することができる。

【０１０１】以上の実施形態においては、英語音声による音声合成装置１と、日本語音声による音声合成装置２とを備えて音声合成システムを構成しているが、音声合成装置２と、マイクロホン（図示せず。）とＡ／Ｄ変換器（図示せず。）とを備えて変形例の音声合成システムを構成してもよい。すなわち、上記の実施形態では、英語の音声合成信号波形データは、音声合成装置１により発生しているが、これに代えて、英語の音声信号の音声

を、例えば英語のネイティブの人により発声して、それをマイクロホンに入力する。マイクロホンはその音声を音声信号に変換し、次いで、Ａ／Ｄ変換器により音声信号データに変換した後、図１のケプストラム抽出部１１１に入力して、以下、音声合成装置２の処理を実行する。これにより、入力される英語の音声に基づいて、音声合成装置２により日本語の音声波形データベースを用いた英語の音声合成の音声を得ることにより、日本語のネイティブの日本人の声による英語の音声合成の音声を得ることができる。ここで、もちろん、２つの言語は英語と日本語に限らず、異なる２つの言語であってもよい。

【０１０２】

【発明の効果】以上詳述したように本発明に係る請求項１記載の音声合成システムによれば、第１の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する第１の音声分析手段と、上記第１の音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、上記第２の記憶手段によって記憶された第１の音響的特徴パラメータと韻律的特徴パラメータとに基づいて、同一の音素種類の１つの目標音素とそれ以外の音素候補との間の第２の音響的特徴パラメータにおける音響的距離を計算し、上記計算した音響的距離に基づいて各音素候補に対して上記第２の音響的特徴パラメータ毎に所定の統計的解析を実行することにより、各音素候補に対する上記第２の音響的特徴パラメータにおける寄与度を表わす各目標音素毎の重み係数ベクトルを決定する重み係数学習手段と、上記重み係数学習手段によって決定された上記第２の音響的特徴パラメータにおける各目標音素毎の重み係数ベクトルを記憶する第３の記憶手段と、上記第１の言語とは異なる第２の言語の自然発話の音声波形信号の音声セグメントを記憶する第４の記憶手段と、上記第４の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の第１の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する第２の音声分析手段と、上記第２の音声分析手段から出力される索引情報と、上記第１の音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第５の記憶手段と、上記第３の記憶手段によって記憶された各目標音素毎の重み係数ベクトルと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する第１の音声単位選択手段と、上記第１の音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応する第１の言語の音声信号波形データを合成して出力する第１の音声合成手段と、上記第１の音声合成手段から出力される音声信号波形データからケプストラム係数データを抽出して出力する抽出手段と、上記抽出手段から出力されるケプストラム係数データと、上記第５の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、上記入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する第２の音声単位選択手段と、上記第２の音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第４の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応しかつ第２の言語の音声セグメントによる音声信号波形を合成して出力する第２の音声合成手段とを備える。従って、第２の言語のネイティブの人の声による第１の言語の自然発話文の音声合成信号の音声を、自然に近い声質で発生することができる。

【０１０３】また、請求項２記載の音声合成システムによれば、請求項１記載の音声合成システムにおいて、それぞれ互いに異なる話者の第１の言語の自然発話の音声波形信号の音声セグメントを記憶する複数の第１の記憶手段と、上記複数の第１の記憶手段に記憶された異なる話者の第１の言語の自然発話の音声波形信号の音声セグメントと、上記第４の記憶手段に記憶された第２の言語の自然発話の音声波形信号の音声セグメントとに基づいて、所定の特徴パラメータの選択基準を用いて、第２の言語の自然発話の音声波形信号により声質が近い第１の言語の自然発話の音声波形信号の話者を選択して、選択した話者の第１の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段を上記第１の音声合成手段に接続する話者選択手段とをさらに備える。ここで、上記特徴パラメータの選択基準に用いる特徴パラメータは、好ましくは、話者の性別及び基本周波数の平

均値である。従って、第２の言語の話者の自然発話の声質に近い第１の言語の話者の自然発話の音声合成を得ることができ、これにより、より声質が近くより自然な発話による第２の言語のネイティブの人の声による第１の言語の音声合成の音声を得ることができる。

【０１０４】さらに、本発明に係る請求項１２記載の音声合成システムによれば、入力される第１の言語の音声信号と、それに対応する第１の言語の音素列に基づいて、上記第１の言語の音素列に対応しかつ上記第１の言語とは異なる第２の言語の音声セグメントによる音声信号波形を合成して出力する音声合成システムであって、上記第２の言語の自然発話の音声波形信号の音声セグメントを記憶する第１の記憶手段と、上記第１の記憶手段によって記憶された音声波形信号の音声セグメントと、上記音声波形信号に対応する音素列とに基づいて、上記音声波形信号における音素毎の索引情報と、上記索引情報によって示された音素毎の音響的特徴パラメータと、上記索引情報によって示された音素毎の韻律的特徴パラメータとを抽出して出力する音声分析手段と、上記音声分析手段から出力される索引情報と、上記音響的特徴パラメータと、上記韻律的特徴パラメータとを記憶する第２の記憶手段と、上記入力される第１の言語の音声信号を音声信号波形データに変換して、変換された音声信号波形データからケプストラム係数データを抽出して出力する抽出手段と、上記抽出手段から出力されるケプストラム係数データと、上記第２の記憶手段によって記憶された韻律的特徴パラメータとに基づいて、上記入力される第１の言語の自然発話文の音素列に対して、目標音素と音素候補との間の近似コストを表わす目標コストと、隣接して連結されるべき２つの音素候補間の近似コストを表わす連結コストとを含むコストが最小となる、音素候補の組み合わせを検索して、検索した音素候補の組み合わせの索引情報を出力する音声単位選択手段と、上記音声単位選択手段から出力される索引情報に基づいて、当該索引情報に対応する音声波形信号の音声セグメントを上記第１の記憶手段から逐次読み出して連結して出力することにより、上記入力された第１の言語の音素列に対応しかつ第２の言語の音声セグメントによる音声信号波形を合成して出力する音声合成手段とを備える。従って、第１の言語のネイティブの人の音声に基づいて、第２の言語のネイティブの人の声による第１の言語の自然発話文の音声合成信号の音声を、自然に近い声質で発生することができる。

【図面の簡単な説明】

【図１】　本発明に係る一実施形態である音声合成システムのブロック図である。

【図２】　第１の従来例の音声合成装置の構成を示すブロック図である。

【図３】　図１の音声単位選択部１２，１１２によって計算される音声単位選択コストの定義を示すモデル図である。

【図４】　図１の音声分析部１０，１１０によって実行される音声分析処理のフローチャートである。

【図５】　図１の重み係数学習部１１によって実行される重み係数学習処理の第１の部分のフローチャートである。

【図６】　図１の重み係数学習部１１によって実行される重み係数学習処理の第２の部分のフローチャートである。

【図７】　図１の音声単位選択部１２によって実行される音声単位選択処理のフローチャートである。

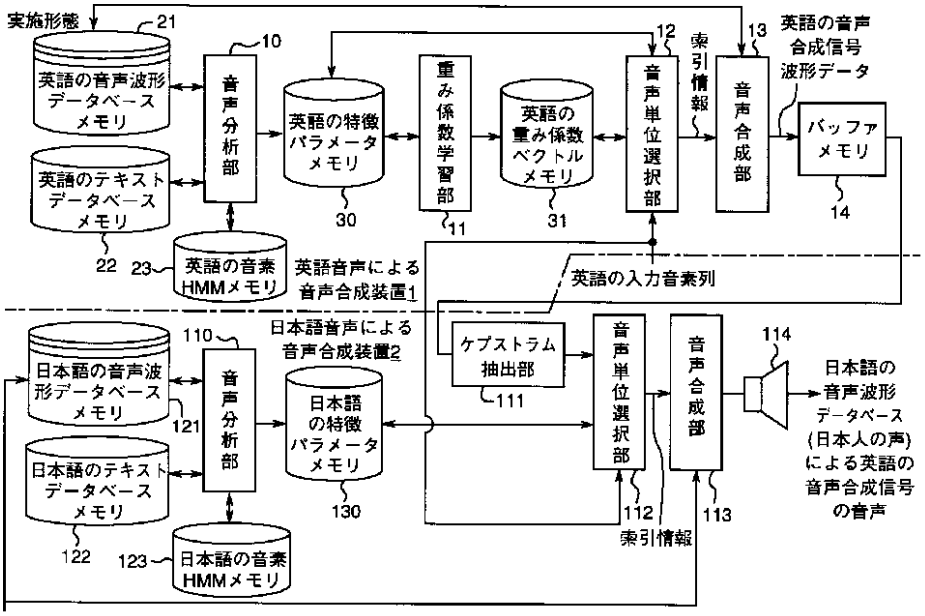
【図８】　図１の音声単位選択部１１２によって実行される音声単位選択処理のフローチャートである。

【図９】　本発明に係る変形例の付加装置である話者選択装置２００の構成を示すブロック図である。

【符号の説明】

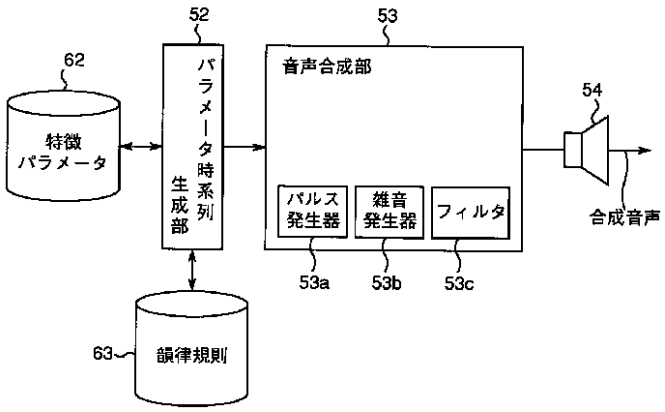
１…英語音声による音声合成装置、
２…日本語音声による音声合成装置、
１０…音声分析部、
１１…重み係数学習部、
１２…音声単位選択部、
１３…音声合成部、
１４…バッファメモリ、
２１，２１－１乃至２１－N…英語の音声波形信号データベースメモリ、
２２…英語のテキストデータベースメモリ、
２３…英語の音素HMMメモリ、
３０…英語の特徴パラメータメモリ、
３１…英語の重み係数ベクトルメモリ、
１１０…音声分析部、
１１１…ケプストラム抽出部、
１１２…音声単位選択部、
１１３…音声合成部、
１１４…スピーカ、
１２１…日本語の音声波形信号データベースメモリ、
１２２…日本語のテキストデータベースメモリ、
１２３…日本語の音素HMMメモリ、
１３０…日本語の特徴パラメータメモリ、
２００…話者選択装置、
２０１…話者選択部、
SW…スイッチ。

【図１】



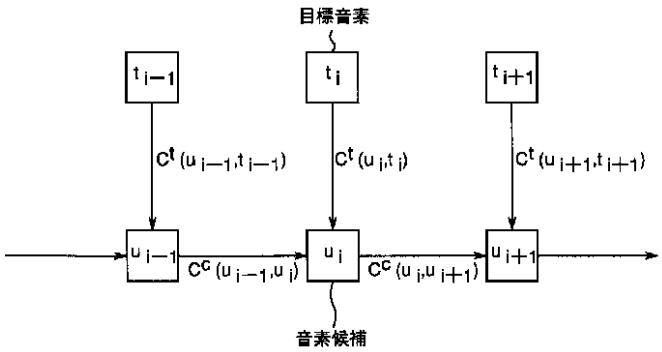
【図２】

第1の従来例

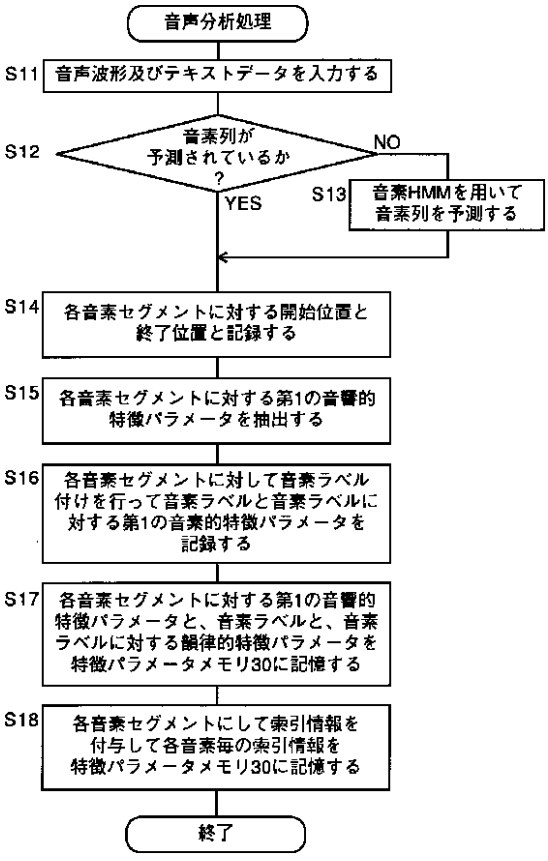


【図３】

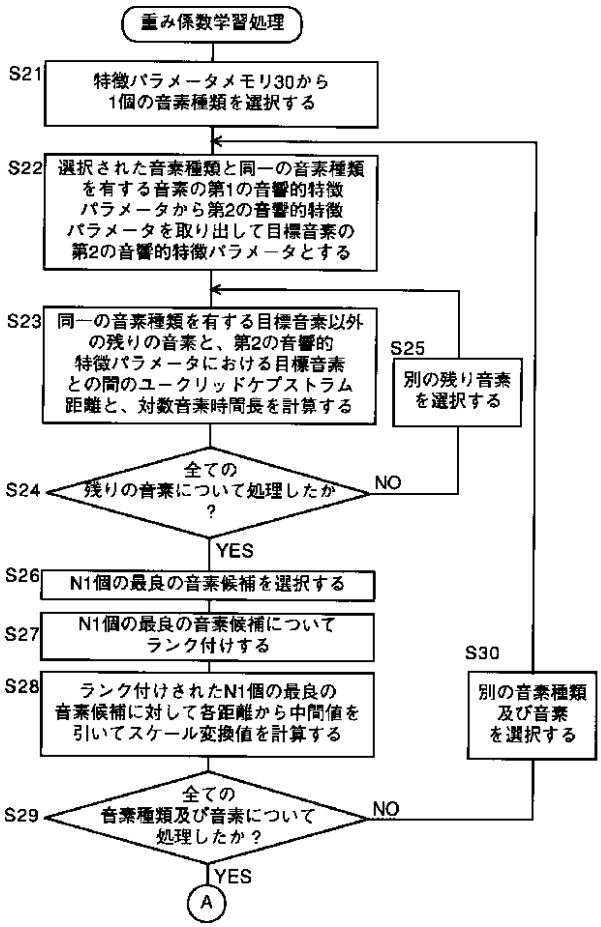
音声単位選択コストの定義



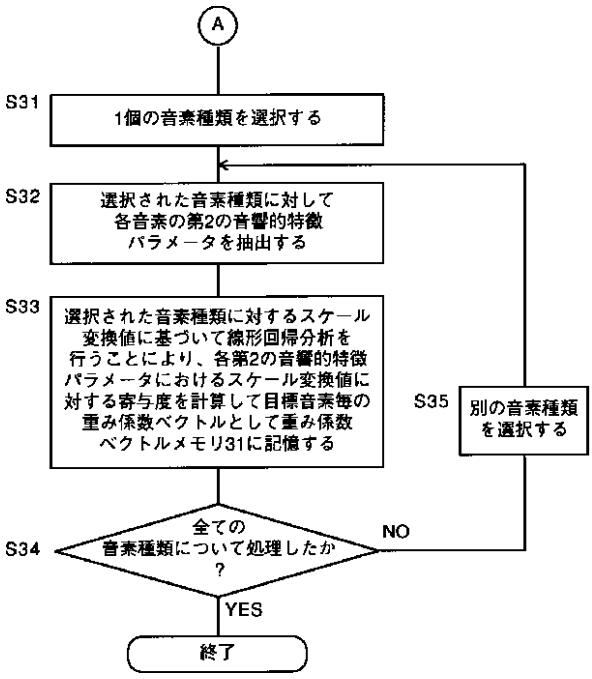
【図４】



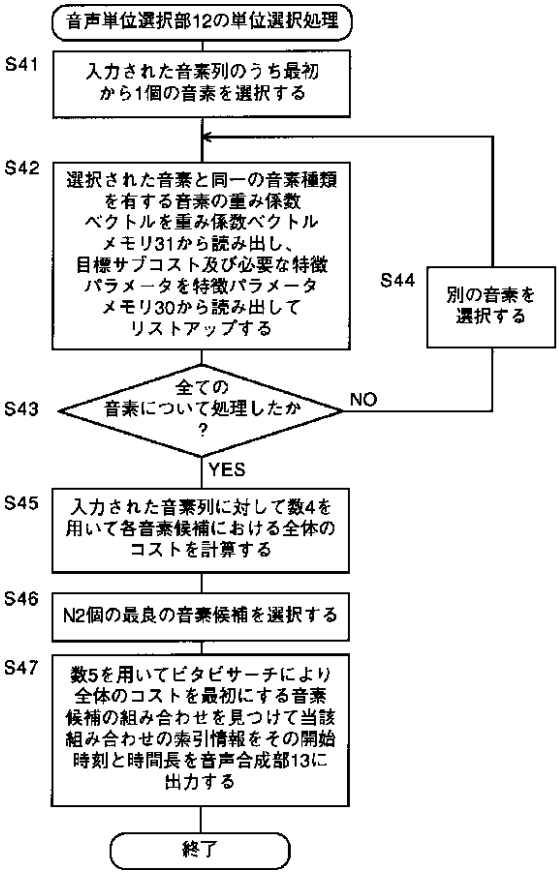
【図５】


重み係数学習処理
S21
特徴パラメータメモリ30から
1個の音素種類を選択する
S22
選択された音素種類と同一の音素種類
を有する音素の第1の音響的特徴
パラメータから第2の音響的特徴
パラメータを取り出して目標音素の
第2の音響的特徴パラメータとする
S23
同一の音素種類を有する目標音素以外
の残りの音素と、第2の音響的
特徴パラメータにおける目標音素
との間のユークリッドケプストラム
距離と、対数音素時間長を計算する
S25
別の残り音素
を選択する
S24
全ての
残りの音素について処理したか
？
NO
YES
S26
N1個の最良の音素候補を選択する
S27
N1個の最良の音素候補について
ランク付けする
S28
ランク付けされたN1個の最良の
音素候補に対して各距離から中間値を
引いてスケール変換値を計算する
S30
別の音素種類
及び音素
を選択する
S29
全ての
音素種類及び音素について
処理したか？
NO
YES
A

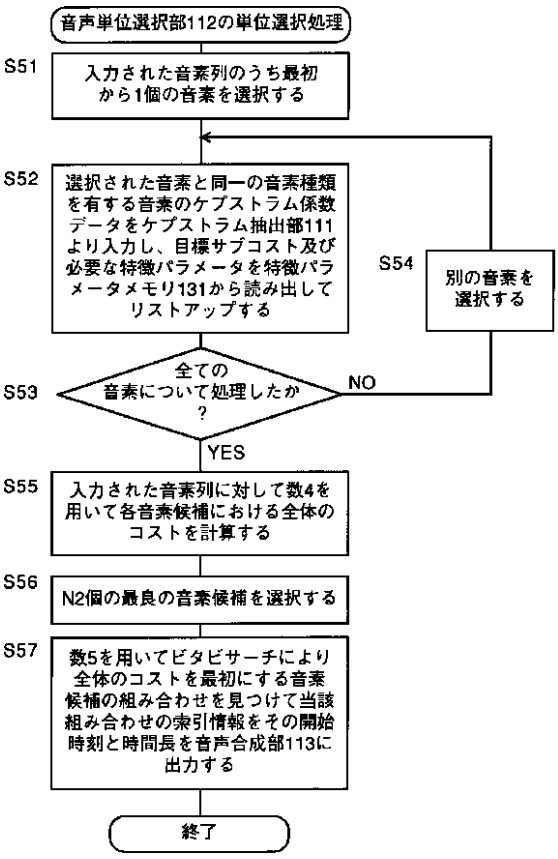
【図６】


A
S31
1個の音素種類を選択する
S32
選択された音素種類に対して
各音素の第2の音響的特徴
パラメータを抽出する
S33
選択された音素種類に対するスケール
変換値に基づいて線形回帰分析を
行うことにより、各第2の音響的特徴
パラメータにおけるスケール変換値に
対する寄与度を計算して目標音素毎の
重み係数ベクトルとして重み係数
ベクトルメモリ31に記憶する
S35
別の音素種類
を選択する
S34
全ての
音素種類について処理したか
？
NO
YES
終了

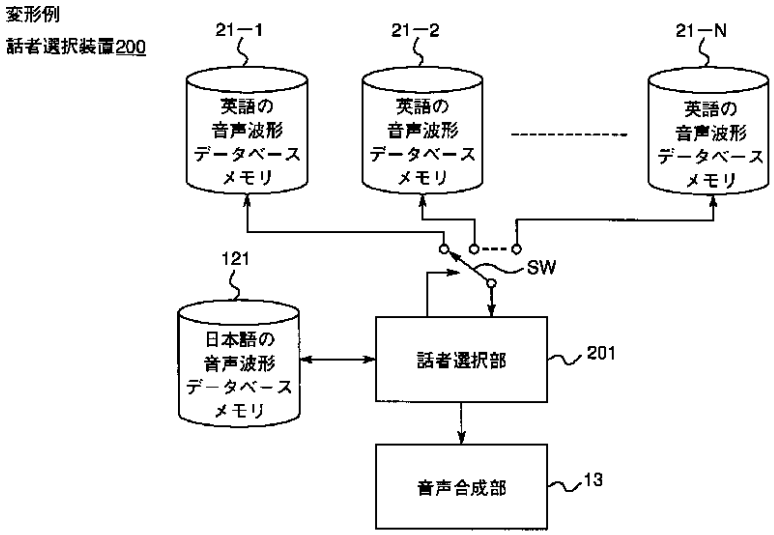
【図７】


音声単位選択部12の単位選択処理
S41
入力された音素列のうち最初
から1個の音素を選択する
S42
選択された音素と同一の音素種類
を有する音素の重み係数
ベクトルを重み係数ベクトル
メモリ31から読み出し、
目標サブコスト及び必要な特徴
パラメータを特徴パラメータ
メモリ30から読み出して
リストアップする
S44
別の音素を
選択する
S43
全ての
音素について処理したか
？
NO
YES
S45
入力された音素列に対して数4を
用いて各音素候補における全体の
コストを計算する
S46
N2個の最良の音素候補を選択する
S47
数5を用いてビタビサーチにより
全体のコストを最初にする音素
候補の組み合わせを見つけて当該
組み合わせの索引情報をその開始
時刻と時間長を音声合成部13に
出力する
終了

【図８】


音声単位選択部112の単位選択処理
S51
入力された音素列のうち最初から1個の音素を選択する
S52
選択された音素と同一の音素種類を有する音素のケプストラム係数データをケプストラム抽出部111より入力し、目標サブコスト及び必要な特徴パラメータを特徴パラメータメモリ131から読み出してリストアップする
S54
別の音素を選択する
S53
全ての音素について処理したか？
NO
YES
S55
入力された音素列に対して数4を用いて各音素候補における全体のコストを計算する
S56
N2個の最良の音素候補を選択する
S57
数5を用いてビタビサーチにより全体のコストを最小にする音素候補の組み合わせを見つけて当該組み合わせの索引情報をその開始時刻と時間長を音声合成部113に出力する
終了

【図９】


変形例
話者選択装置200
21−1
21−2
21−N
英語の音声波形データベースメモリ
英語の音声波形データベースメモリ
英語の音声波形データベースメモリ
121
SW
日本語の音声波形データベースメモリ
話者選択部
201
音声合成部
13

フロントページの続き

(56)参考文献　特開　平６－332494（ＪＰ，Ａ）
特開　平９－146585（ＪＰ，Ａ）
特開　昭62－18600（ＪＰ，Ａ）
特開　昭62－174800（ＪＰ，Ａ）
藤沢ら「入力音声の韻律を用いた音声合成」、日本音響学会平成10年度春季研究発表会講演論文集、ｐｐ191－192（1998）

(58)調査した分野(Int.Cl.[6]，ＤＢ名)
G10L　3/00 - 9/20
ＪＩＣＳＴファイル（ＪＯＩＳ）