

(19)日本国特許庁 (J P)

(12) 特 許 公 報 (B 2)

(11)特許番号

特許第3050832号  
(P3050832)

(45)発行日 平成12年6月12日(2000.6.12)

(24)登録日 平成12年3月31日(2000.3.31)

(51)Int.Cl.<sup>7</sup>

識別記号

F I

G 1 0 L 13/06  
13/08  
15/14

G 1 0 L 5/04  
3/00

F  
H  
5 3 5 Z

請求項の数9(全 14 頁)

(21)出願番号 特願平9-123822

(22)出願日 平成9年5月14日(1997.5.14)

(65)公開番号 特開平10-49193

(43)公開日 平成10年2月20日(1998.2.20)

審査請求日 平成9年5月14日(1997.5.14)

(31)優先権主張番号 特願平8-120113

(32)優先日 平成8年5月15日(1996.5.15)

(33)優先権主張国 日本(J P)

(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光台二丁目2番地2

(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

(72)発明者  
アンドリュー・ハント  
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 100062144  
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】 自然発話音声波形信号接続型音声合成装置

1

(57)【特許請求の範囲】

【請求項1】 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

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たことを特徴とする自然発話音声波形信号接続型音声合成装置。

【請求項2】 上記音声分析手段は、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記載の音声合成装置。

【請求項3】 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又は2記載の音声合成装置。

【請求項4】 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又は2記載の音声合成装置。

【請求項5】 上記音声単位選択手段は、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4のうちの1つに記載の音声合成装置。

【請求項6】 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5のうちの1つに記載の音声合成装置。

【請求項7】 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5のうちの1つに記載の音声合成装置。

【請求項8】 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7のうちの1つに記載の音声合成装置。

【請求項9】 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8のうちの1つに記載の音声合成装置。

【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に関する。

【0002】

【従来の技術】図2は、従来例の音声合成装置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】

【発明が解決しようとする課題】しかしながら、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】本発明の目的は以上の問題点を解決し、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を提供することにある。

【0006】

【課題を解決するための手段】本発明に係る請求項1記載の自然発話音声波形信号接続型音声合成装置は、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たことを特徴とする。

【0007】また、請求項 2 記載の音声合成装置は、請求項 1 記載の音声合成装置において、上記音声分析手段は、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。

【0008】さらに、請求項 3 記載の音声合成装置は、請求項 1 又は 2 記載の音声合成装置において、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 N 1 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。

【0009】さらに、請求項 4 記載の音声合成装置は、請求項 1 又は 2 記載の音声合成装置において、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 N 1 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。

【0010】また、請求項 5 記載の音声合成装置は、請求項 1 乃至 4 のうちの 1 つに記載の音声合成装置において、上記音声単位選択手段は、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 N 2 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】さらに、請求項 6 記載の音声合成装置は、請求項 1 乃至 5 のうちの 1 つに記載の音声合成装置において、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 7 記載の音声合成装置は、請求項 1 乃至 5 のうちの 1 つに記載の音声合成装置において、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。

【0012】また、請求項 8 記載の音声合成装置は、請求項 1 乃至 7 のうちの 1 つに記載の音声合成装置において、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<sub>0</sub> と、パワーとを含むことを特徴とする。

【0013】さらに、請求項 9 記載の音声合成装置は、請求項 1 乃至 8 のうちの 1 つに記載の音声合成装置において、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。

【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。図 1 は、本発明に係る一実施形態である自然発話音声波形信号接続型音声合成装置のブロック図である。例えば図 2 に示した従来例の音声合成装置では入力された発聲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) 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、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スピーカ 14 に出力する音声合成部 13。

【0015】具体的には、音声分析部 10 は、入力され

る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第1の韻律的特徴パラメータとを抽出して出力する。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、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、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、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音声波形信号の音声セグメントを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、上記入力された音素列に対応する音声を合成して出力する。

【0016】ここで、音声分析部10の処理は新しい音声波形信号データベース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。

【0017】本実施形態の音声合成装置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、例えば音素隠れマルコフモデル(以下、隠れマルコフモデルをHMMという。)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。

【0018】本実施形態では、音声波形信号データベースメモリ21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。

【0019】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、好ましい実施例においては、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の入力はメモリ22内の音素表記を伴ったメモリ21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。

【0020】音声分析部10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30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<sub>0</sub>、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。  
【 0 0 2 1 】  
【 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 3 0 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：

- 1 2 次メルケプストラム係数
- 1 2 次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<sub>0</sub>
- パワー

【 0 0 2 2 】として代わって、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。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 3 0 に記憶される。ここで、音素ラベルに付与される、例えば 1 2 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、音声分析部 1 0 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 2 1 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。

【 0 0 2 3 】  
【 表 2 】音声分析部 1 0 の出力結果である特徴パラメータの一例  
索引番号 X 0 0 0 5

40	音素	時間長	基本周波数	パワー	.....
	#	1 2 0	9 0	4 . 0	.....
	s	1 7 5	9 8	4 . 7	.....
	e i	9 5	1 0 2	6 . 5	.....
	d h	3 0	1 1 4	4 . 9	.....
	i h	7 5	1 4 3	6 . 9	.....
	s	1 5 0	1 4 0	5 . 7	.....
	p	8 7	1 3 7	5 . 1	.....
	l	3 4	1 0 7	4 . 9	.....
50	i i	1 5 0	9 8	6 . 3	.....

			11	
z	140	87	5.8	.....
#	253	87	4.0	.....

【0024】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。

【0025】音声分析部10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。

【0026】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が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ピタピの学習アルゴリズムを用いて音素のアラインメントを行ない、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ン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。

【0027】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規準とすれば良い。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。

【0028】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

12

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。

【0029】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。

【0030】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、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。

【0031】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。

【0032】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 $F_0$ 。

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。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。

【0033】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な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。

【0034】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。

【0035】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 $t_1^n = (t_1, \dots, t_n)$ 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 $u_1^n = (u_1, \dots, u_n)$ 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。

【0036】音声単位の選択コストは、図3に示すように、目標コスト $C^1(u_i, t_i)$ と連結コスト $C^0(u_{i-1}, u_i)$ からなり、目標コスト $C^1(u_i, t_i)$ 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 $C^0(u_{i-1}, u_i)$ 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例のATR-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。

【0037】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 $C^1_j(t_i, u_i)$ の重み係数 $w^1_j$ が与えられた場合、目標コスト $C^1(t_i, u_i)$ は次式で計算することができる。

【0038】

【数1】

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

【0039】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 $C^1_j(t_i, u_i)$ (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 $C^1(t_i, u_i)$ 及び重み係数 $w^1_j$ 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。

【0040】一方、連結コスト $C^{\circ}(u_{i-1}, u_i)$ も同様に $q$ 個の連結サブコスト $C^{\circ}_j(u_{i-1}, u_i)$ （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 $C^{\circ}_j(u_{i-1}, u_i)$ の重み $w^{\circ}_j$ 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 $C^{\circ}(u_{i-1}, u_i)$ は次式で計算することができる。

【0041】

【数2】

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

【0042】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

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

【0046】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。

【0047】

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

【0048】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） $u_1, u_2, \dots, u_n = /u_1^n$ を表わす関数である。

【0049】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

\* 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の音響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。

【0043】

【数3】

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

【0044】このとき、 $S$ はポーズを表しており、 $C^{\circ}(S, u_1)$ 及び $C^{\circ}(u_n, S)$ 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のように、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。

【0045】

【数4】

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

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。以下に線形回帰分析における処理の流れを示す。

【0050】 $\langle 1 \rangle$  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

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 $C^i_j(t_i, u_i)$ 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 $C^i_j(t_i, u_i)$ 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。

【0051】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。

【0052】以上のように構成された図1の音声合成装置において、音声分析部10と、重み係数学習部11と、音声単位選択部12と、音声合成部13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と、音素HMMメモリ23と、特徴パラメータメモリ30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は、CD-ROMの形式の記憶装置である。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。

【0053】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。

【0054】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。

【0055】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。

【0056】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。

【0057】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。

【0058】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。

【0059】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S4

7では、数5を用いてピタピサーチにより、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。

【0060】さらに、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21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14を介して出力する。これにより、入力された音素列に対応する音声合成された音声はスピーカ14から出力される。

【0061】以上説明したように、本実施形態の音声合成装置においては、出力音声の自然性を最大にするために、大規模な自然音声のデータベースを用いて処理を最小に抑える方法について述べた。本実施形態は4つの処理部10乃至13から構成される。

20 <音声分析部10>正書法の書き起こしテキストを伴った任意の音声波形信号データを入力とし、この音声波形信号データベース中のすべての音素について、それらの性質を記述する特徴ベクトルを与える処理部。

20 <重み係数学習部11>音声波形信号データベースの特徴ベクトルと音声波形信号データベースの原波形を用いて、目的の音声を合成する場合に最も適するように音声単位を選ぶための、各特徴パラメータの最適重み係数を重みベクトルとして決定する処理部。

30 <音声単位選択部12>音声波形信号データベースの全音素の特徴ベクトルと重みベクトルと目的音声の発話内容の記述から音声波形信号データベースメモリ21の索引情報を作成する処理部。

30 <音声合成部13>作成された索引情報に従って、メモリ21内の音声波形信号データベース中の音声波形信号データの音声セグメントに飛び飛びにアクセスし、目的の音声波形信号の音声セグメントを連結しかつD/A変換してスピーカ14に出力して音声を合成する処理部。

40 【0062】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<sub>0</sub>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。

【0063】本実施形態の音声合成装置の特徴は、次の通りである。

<単位選択規準としての韻律的情報の利用>スペクトルの特徴は韻律の特徴と不可分であるとの立場から、音声単位の選択規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。

【0064】  
【実施例】本実施形態の音声合成装置を用いて、発明者はこれまで4カ国語を含む各種の音声波形信号データベースで評価を行なっている。周知の通り、従来は女性話者の音声を用いて高品質の音声を合成することは技術的に大変難しかったが、本実施形態の方法では性別や年齢などによる差がなくなった。現在のところ、日本語では若年の女性話者が短い物語を読み上げた音声波形信号データベースを用いた場合に最も高品質の合成音声を得られている。一方、ドイツ語については韻律ラベルと詳細な音素ラベルが付与された読み上げ文のCD-ROMデータを用いた合成音声を出力している。これは特別に音声合成用に録音した音声波形信号データではなく、本実施形態の音声合成装置が既存の各種の音声波形信号データを技術的には自由に利用できることを示している。また、英語ではボストン大学のニュース・コーパスのラジオ・アナウンサの45分間の音声波形信号データで最も良い音質が得られている。なお、韓国語については短い物語の読み上げ音声を用いている。

【0065】

【発明の効果】以上詳述したように本発明に係る請求項1記載の自然発話音声波形信号接続型音声合成装置によれば、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る。従って、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。

【0066】また、請求項2記載の音声合成装置においては、請求項1記載の音声合成装置において、上記音声分析手段は、好ましくは、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る。従って、予め音素列を与える必要がないので、手作業の部分を簡単化することができる。

【0067】さらに、請求項3記載の音声合成装置においては、請求項1又は2記載の音声合成装置において、上記重み係数学習手段は、好ましくは、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におい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。従って、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、しかもより自然に近い声質を得ることができる。

【0068】さらに、請求項4記載の音声合成装置にお

いては、請求項1又は2記載の音声合成装置において、上記重み係数学習手段は、好ましくは、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。従って、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、しかもより自然に近い声質を得ることができる。

【0069】またさらに、請求項5記載の音声合成装置においては、請求項1乃至4のうちの1つに記載の音声合成装置において、上記音声単位選択手段は、好ましくは、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。従って、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、しかもより自然に近い声質を得ることができる。

【図面の簡単な説明】

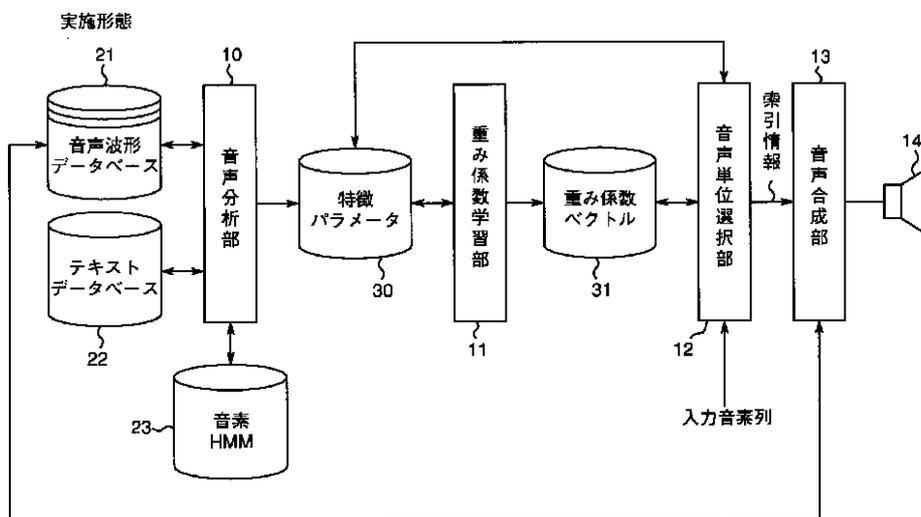
【図1】 本発明に係る一実施形態である自然発話音声波形信号接続型音声合成装置のブロック図である。 \*

- \* 【図2】 従来例の音声合成装置のブロック図である。
- 【図3】 図1の音声単位選択部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。
- 10 【図7】 図1の音声単位選択部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。

【符号の説明】

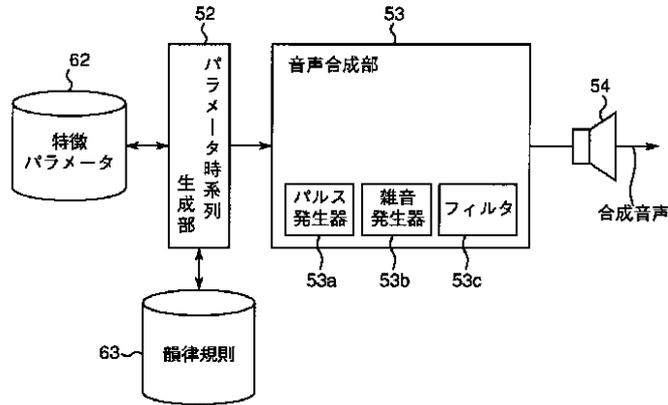
- 10...音声分析部、
- 11...重み係数学習部、
- 12...音声単位選択部、
- 13...音声合成部、
- 14...スピーカ、
- 21...音声波形信号データベースメモリ、
- 22...テキストデータベースメモリ、
- 23...音素HMMメモリ、
- 30...特徴パラメータメモリ、
- 31...重み係数ベクトルメモリ、
- 30...特徴パラメータメモリ、
- 31...重み係数ベクトル。

【図1】



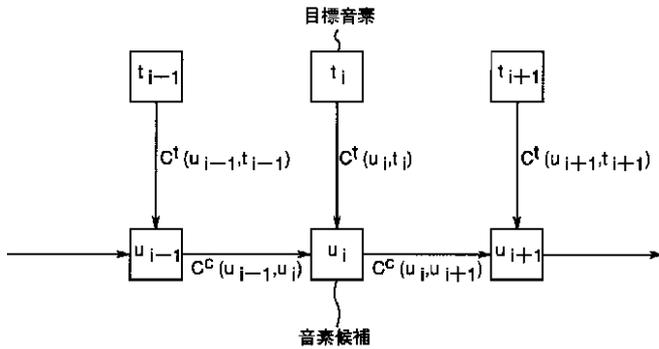
【図2】

従来例

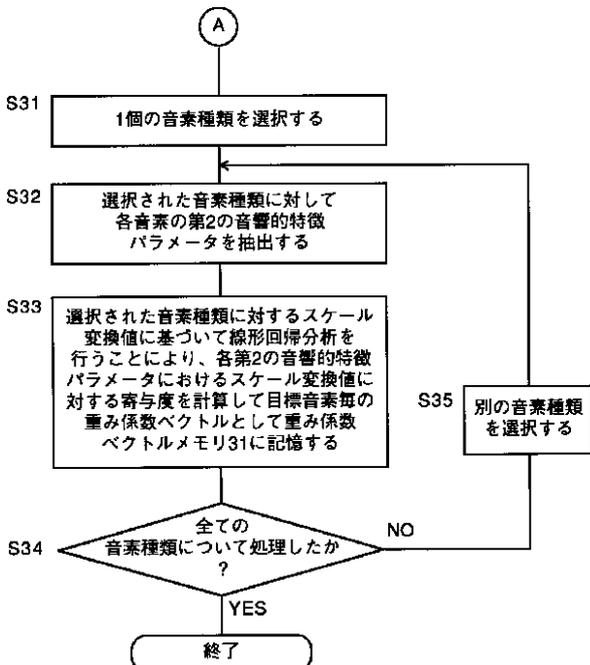


【図3】

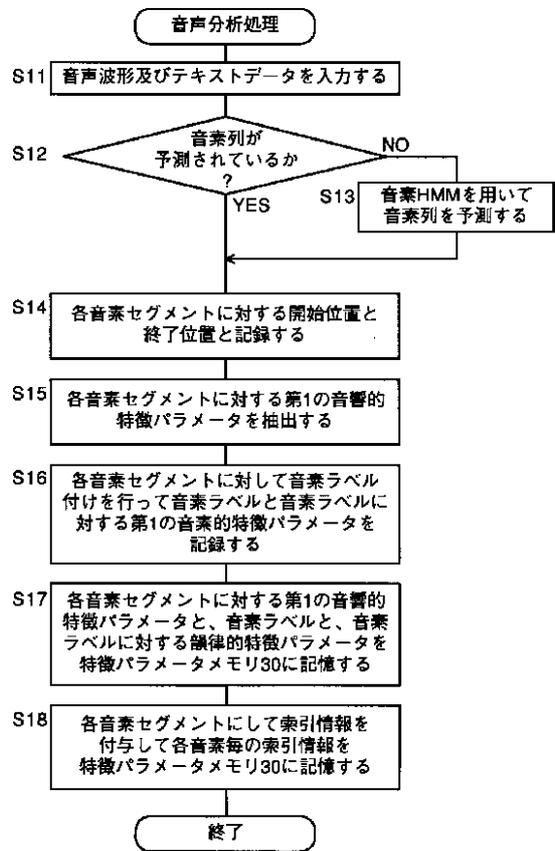
音声単位選択コストの定義



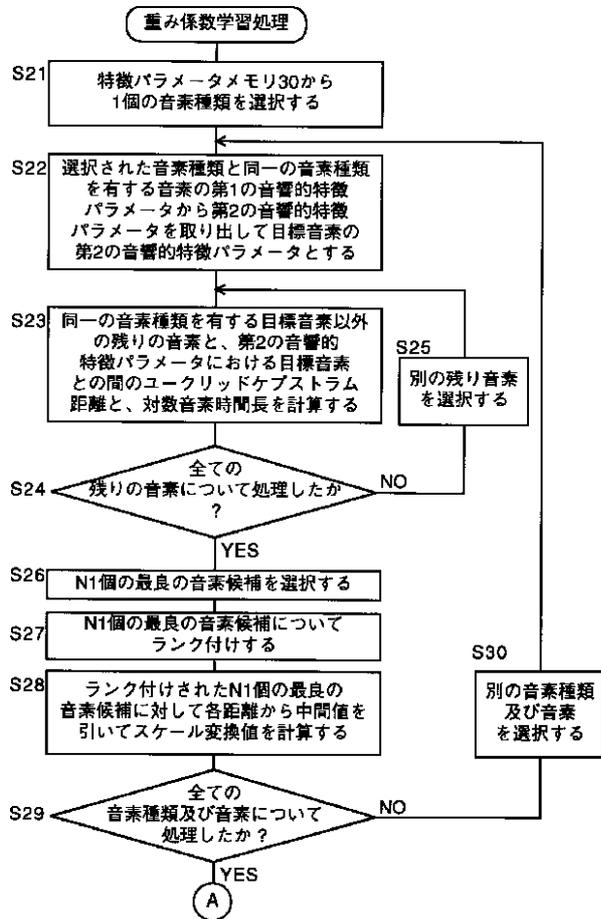
【図6】



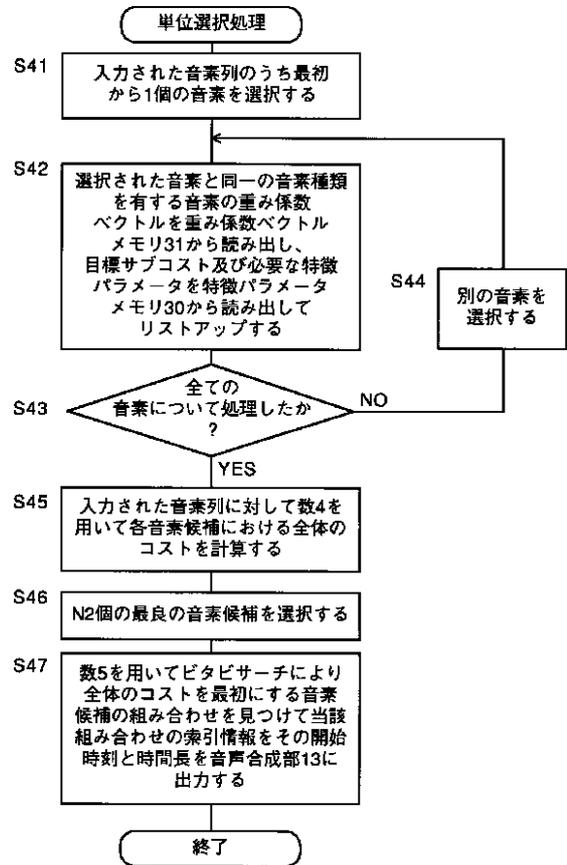
【図4】



【図5】



【図7】



フロントページの続き

(56) 参考文献 特開 昭64 - 78300 ( J P , A )  
 特開 平 7 - 319497 ( J P , A )  
 特開 平 7 - 92997 ( J P , A )  
 特開 平 8 - 87297 ( J P , A )  
 特開 平 6 - 266390 ( J P , A )  
 CAMPBELL N, et al.  
 「CHATR: 自然音声波形接続型任意音声合成システム」、信学技報 S P 96 - 7, Vol. 96, No. 39, pp. 45 - 52 ( 1996 )

(58) 調査した分野(Int.Cl.<sup>7</sup>, D B 名)  
 G10L 11/00 - 21/06  
 J I C S T ファイル ( J O I S )