(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号 特許第3459600号 (P3459600)

(45)発行日　平成15年10月20日(2003.10.20)　(24)登録日　平成15年８月８日(2003.8.8)

(51)Int.Cl.⁷	識別記号	ＦＩ	
Ｇ１０Ｌ　13/06		Ｇ１０Ｌ　5/04	Ｅ
13/08		3/00	Ｈ

請求項の数６(全 24 頁)

(21)出願番号　特願平11－274233

(22)出願日　平成11年９月28日(1999.9.28)

(65)公開番号　特開2001－100775(P2001－100775A)

(43)公開日　平成13年４月13日(2001.4.13)

審査請求日　平成13年１月17日(2001.1.17)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
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　北川　敏
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　渡邊　聡

最終頁に続く

(54)【発明の名称】　音声合成装置のための音声データ量削減装置及び音声合成装置

(57)【特許請求の範囲】

【請求項１】　音素ラベルに対応した音声波形信号の音声セグメントのデータからなる音声波形データベースを記憶する記憶装置を備え、上記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音声合成装置のための音声データ量削減装置であって、

上記音声波形データベースに含まれる１対の音素のリストを生成する生成手段と、

上記生成された１対の音素のリストに基づいて各１対の音素に対する韻律的特徴パラメータと音響的特徴パラメータとに関する所定の類似度を計算し、上記計算された類似度が所定の第１のしきい値以上であるとき、当該各１対の音素のうちの一方の１対の音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベー

スから削除することにより音声データ量を削減するとともに、上記計算された類似度が所定の第１のしきい値以上であり、かつ上記１対の音素のリスト中の同一の１対の音素の数が所定の第２のしきい値以上であるときに、当該各１対の音素のうちの一方の１対の音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベースから削除する削減手段とを備えたことを特徴とする音声合成装置のための音声データ量削減装置。

【請求項２】　上記類似度は、それぞれ所定の重み係数で重み付けされた、上記韻律的特徴パラメータに関する類似度のスコアと、上記音響特徴パラメータに関する類似度のスコアとの線形結合の式を用いて計算されることを特徴とする請求項１記載の音声合成装置のための音声データ量削減装置。

【請求項３】　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、経過時間に対する音声基本周波数F_0の傾きと、パワーとを含むことを特徴とする請求項１又は２記載の音声合成装置のための音声データ量削減装置。

【請求項４】　上記音響的特徴パラメータは、スペクトラム情報を含む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装置のための音声データ量削減装置。

【請求項５】　請求項１乃至４のうちの１つに記載の音声合成装置のための音声データ量削減装置によって音声データ量が削減された音声波形データベースに基づいて、入力された自然発話文の音素列に対して、音素候補を上記音声波形データベースから検索して連結することにより音声合成を行なう音声合成手段を備えたことを特徴とする音声合成装置。

【請求項６】　請求項１乃至４のうちの１つに記載の音声合成装置のための音声データ量削減装置によって音声データ量が削減された音声波形データベースに基づいて、入力された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と、時間的に隣接して連結されるべき音素候補間の近似コストとを含むコストが最小となるように、音素候補を上記音声波形データベースから検索して連結することにより音声合成を行なう音声合成手段を備えたことを特徴とする音声合成装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、音素ラベルに対応した音声波形信号の音声セグメントのデータからなる音声波形データベースを用いて、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音声合成装置のための音声データ量削減装置及び音声合成装置に関する。

【０００２】

【従来の技術】例えば、特開平１０－０４９１９３号公報において、音素ラベルに対応した音声波形信号の音声セグメントのデータからなる音声波形データベースを用いて、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音声合成装置が開示されている。

【０００３】

【発明が解決しようとする課題】この従来例の音声合成装置においては、最小限の信号処理を用いて単純に音声波形の連結を行うために、適切な音響的パラメータと韻律的パラメータとを有する音声セグメントを選択するため、大規模な音声波形データベースを必要とする。この大規模な音声波形データベースでは、当該データベースを記憶するメモリの容量が大きくなり、また、それに伴って探索空間が大きくなるために、適切な音声セグメントを探索するときの探索速度を高めることができないという問題点があった。

【０００４】本発明の目的は以上の問題点を解決し、音声波形データベースを格納するメモリ容量を削減することができ、音声合成時の探索速度を高めることができる音声合成装置のための音声データ量削減装置及び音声合成装置を提供することにある。

【０００５】

【課題を解決するための手段】本発明に係る音声合成装置のための音声データ量削減装置は、音素ラベルに対応した音声波形信号の音声セグメントのデータからなる音声波形データベースを記憶する記憶装置を備え、上記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音声合成装置のための音声データ量削減装置であって、上記音声波形データベースに含まれる１対の音素のリストを生成する生成手段と、上記生成された１対の音素のリストに基づいて各１対の音素に対する韻律的特徴パラメータと音響的特徴パラメータとに関する所定の類似度を計算し、上記計算された類似度が所定の第１のしきい値以上であるとき、当該各１対の音素のうちの一方の１対の音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベースから削除することにより音声データ量を削減するとともに、上記計算された類似度が所定の第１のしきい値以上であり、かつ上記１対の音素のリスト中の同一の１対の音素の数が所定の第２のしきい値以上であるときに、当該各１対の音素のうちの一方の１対の音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベースから削除する削減手段とを備えたことを特徴とする。

【０００６】

【０００７】また、上記音声データ量削減装置において、上記類似度は、好ましくは、それぞれ所定の重み係数で重み付けされた、上記韻律的特徴パラメータに関する類似度のスコアと、上記音響特徴パラメータに関する類似度のスコアとの線形結合の式を用いて計算されることを特徴とする。

【０００８】さらに、上記音声データ量削減装置において、上記韻律的特徴パラメータは、好ましくは、音素時間長と、音声基本周波数F_0と、経過時間に対する音声基本周波数F_0の傾きと、パワーとを含み、上記音響的特徴パラメータは、スペクトラム情報を含む。

【０００９】本発明に係る音声合成装置は、上記音声データ量削減装置によって音声データ量が削減された音声波形データベースに基づいて、入力された自然発話文の音素列に対して、音素候補を上記音声波形データベースから検索して連結することにより音声合成を行なう音声合成手段を備えたことを特徴とする。

【００１０】また、本発明に係る音声合成装置は、上記音声データ量削減装置によって音声データ量が削減された音声波形データベースに基づいて、入力された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

似コストと、時間的に隣接して連結されるべき音素候補間の近似コストとを含むコストが最小となるように、音素候補を上記音声波形データベースから検索して連結することにより音声合成を行なう音声合成手段を備えたことを特徴とする。

【００１１】

【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。

【００１２】図１は、本発明に係る一実施形態である音声データ量削減処理装置のブロック図である。この実施形態の音声データ量削減装置は、図２の音声合成装置に提供する音素ラベルデータにてなるテキストデータベースと音声波形データベース内の音声データ量を削減するために、図３の音声データ量削減処理を用いて、各バイフォン（biphone；時間的に隣接する１対の音素をいう。）に対する評価韻律データとバイスペクトラム（詳細後述）とのデータを含む評価データ行列に基づいて、テキストデータベースと音声波形データベース内で所定の類似度以上の音声データを削除することにより音声データ量の削減を行う音声データ量削減処理部４５を備えたことを特徴とする。

【００１３】すなわち、本実施形態においては、韻律的特徴パラメータの領域及びスペクトル領域の両方における近似性の物理的尺度を使用して、出力される音声合成された音声の品質を維持しながら音声波形データベースにおける音声セグメントの数を減らして音声波形データベースにおける冗長さを小さくする方法について開示する。

【００１４】本実施形態では、上記物理的尺度の音響的特徴パラメータとして、バイスペクトラムを用いる。パワースペクトルは２次のデータによってのみ決定される＊＊ため、より高次の情報は無視される。もし音声がガウス分布を示すならば、２次のデータのみで完全に音声を復元合成することができる。しかしながら、実際は音声がガウス分布ではない。以上の２つの理由により、本発明者は音声セグメントの類似度を測定するための音響特徴パラメータの尺度として、バイスペクトルを使用して音声合成を評価する。

【００１５】ここで、本実施形態において用いるバイスペクトラム（bispectrum）について、従来技術文献１「C. L. Nikias et al., "Bispectrum Estimation: A DigitalSignal Processing Framework", Proceedings of the IEEE 76, pp.869-891,1987」を参照して説明する。この実施形態において、実離散処理を仮定しており、｛Ｘ（ｋ）｝を実離散ゼロ平均定常処理とすると、パワースペクトラムＰ（ω）は次式で定義される。

【００１６】

【数１】

$$P(\omega)=\sum_{\tau=-\infty}^{+\infty} r(\tau)\ \exp\{-j\ (\omega\tau)\},\ |\omega|<\pi$$

ここで、

【数２】$r(\tau)=E\{X(k)X(k+\tau)\}$

はその自己相関シーケンスである。もしＲ（ｍ，ｎ）が｛Ｘ（ｋ）｝の３次のモーメントのシーケンスを示すならば、すなわち、

【数３】$R(m,n)=E\{X(k)X(k+m)X(k+n)\}$

であるとき、そのバイスペクトラムは次式で定義される。

【００１７】

【数４】

$$B(\omega_1,\ \omega_2)=\sum_{m=-\infty}^{+\infty}\ \sum_{n=-\infty}^{+\infty} R(m,\ n)\ \exp\{-j\ (\omega_1 m+\omega_2 n)\}$$

【００１８】ここで、Ｒ（ｍ，ｎ）は上述の３次の自己相関関数であり、３次のモーメントと累積係数（累積率ともいう。）が同一であるので、バイスペクトラムは３次の累積係数のスペクトラムとなる。パワースペクトラムとバイスペクトラムの物理的な重要性は、次式のＸ（ｋ）のフーリエ・スティルチェスの表現式（クラメルのスペクトル表現式ともいう。）の成分ｄＺ（ω）の項を用いて表わすときに明らかになる。

【００１９】すべてのｋに対して、

【数５】

$$X(k)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{j\omega k}\ dZ(\omega)$$

であり、ここで、

【数６】$E\{dZ(\omega)\}=0$

【数７】$E\{dZ(\omega_1)\ dZ^{*}(\omega_2)\}$

$=0,\ \omega_1\neq\omega_2$のとき

$=2\pi P(\omega)\ d\omega,\ \omega_1=\omega_2=\omega$のとき

並びに、

【数８】$E\{dZ(\omega_1)\ dZ(\omega_2)\ dZ^{*}(\omega_3)\}$

$=B(\omega_1,\ \omega_2)\ d\omega_1 d\omega_2,\ \omega_1+\omega_2=\omega_3$

$=0,\ \omega_1+\omega_2\neq\omega_3$

である。

【００２０】従って、パワースペクトラムＰ（ω）は周波数が同一である２つのフーリエ成分の平均値の積に寄与することを表わす一方、バイスペクトラムＢ（ω_1，ω_2）は、１つの周波数が他の２つの周波数の和に等しいときの３つのフリーエ成分の平均値の積に寄与することを表わす。すなわち、バイスペクトラムは、２つのスペクトラム列に対する寄与度又は類似度を表わす。

【００２１】本実施形態では、２次元のフーリエ変換を使用してバイスペクトルを計算する。バイスペクトルの次元は一般に高いため、２次元のＤＣＴを使用してバイスペクトルをより低次の係数に圧縮して計算する。なお、本実施形態においては、音響特徴パラメータとして

バイスペクトラムを用いているが、本発明はこれに限らず、スペクトラム情報を用いてもよい。

【００２２】本実施形態では、上記物理的尺度の韻律的特徴パラメータとして、各音声セグメントの持続時間（又は音素時間長）、最大信号振幅（又はパワー）、平均基本周波数（又はピッチ周波数）F_0及び基本周波数F_0の傾き（又は傾斜）を用いる。１対の音声セグメント間の音声距離を決定するにはスペクトル領域における物理的尺度が必要であるが、音声波形データベースの多様性を維持しようとする場合には、１対の音声セグメント間の韻律的特徴パラメータの距離を決定することも必要である。スペクトル特性及び韻律的特性の両方で所定のしきい値内で類似する（又は重複する）音声セグメントは音声波形データベースから除去することが可能である。言い換えれば、音声波形データベースのある部分を削除しても、残った部分から削除前と実質的に同一の音声波形を合成できるならば、その部分を削除してもかまわないという技術的思想に基づいている。これにより音声波形データベースの量は減少するが、音声波形データ＊＊としてのカバー範囲は実質的に変化しないといえる。本実施形態においては、韻律的特徴パラメータの類似度を測定するため、音素サイズの各音声セグメントの上述の韻律的特徴パラメータを用いる。

【００２３】そして、本実施形態においては、各１対のバイフォンに対して、以下に示す類似度のスコアMを計算して所定のしきい値Ｍｔｈ以下のときに、１対のバイフォンは互いに類似していると判断し、かつこのバイフォンの数Nを予め決められたしきい値Ｎｔｈ未満とならないように、すなわち当該しきい値Ｎｔｈ以上であれば削除しても複製可能であり、それは冗長であると判断して、当該バイフォンの音素に関するデータをテキストデータベースメモリ２２及び音声波形データベースメモリ２１から削除することにより、音声データ量の削減を行う。

【００２４】１対のバイフォンに対する類似度のスコアMは次式で表わされる。

【００２５】

【数９】

$$M = w_1 \times (w_{11} \times S_{F0} + w_{12} \times S_P + w_{13} \times S_D + w_{14} \times S_{FI}) + w_2 \times S_{BS}$$

【００２６】ここで、

（ａ）w_1：韻律的特徴パラメータのスコアに対する重み係数であり、例えば０．５である。

（ｂ）w_{11} ：基本周波数F_0のスコアに対する重み係数であり、例えば０．３である。

（ｃ）S_{F0} ：１対のバイフォンに対する基本周波数F_0の差の絶対値である。

（ｄ）w_{12} ：パワーのスコアに対する重み係数であり、例えば０．２である。

（ｅ）S_P：１対のバイフォンに対するパワーの差の絶対値である。

（ｆ）w_{13} ：音素時間長のスコアに対する重み係数であり、例えば０．１である。

（ｇ）S_D：１対のバイフォンに対する音素時間長の比の値（又は比率）である。

（ｈ）w_{14} ：基本周波数F_0の傾きのスコアに対する重み係数であり、例えば０．３である。

（ｉ）S_{FI} ：１対のバイフォンに対する基本周波数F_0の傾きに係るスコアであり、２つのバイフォンの音声波形の変化パターンの組み合せから求める。

＜１＞まず、基本周波数F_0の変化を上昇、水平（実質的に変化せずを意味する。）、下降の３つのパターンに分類する。

＜２－１＞音声波形Ａと音声波形Ｂのパターンが同じならば、スコア＝０とする。

＜２－２＞音声波形Ａと音声波形Ｂのパターンが次の表に該当するときは、スコア＝１とする。

【表１】

波形Ａ	波形Ｂ
上昇	水平
水平	上昇
水平	下降
下降	水平

＜２－２＞音声波形Ａと音声波形Ｂのパターンが次の表に該当するときは、スコア＝１とする。

【表２】

波形Ａ	波形Ｂ
上昇	下降
下降	上昇

（ｊ）w_2：バイスペクトラムに対する重み係数であり、例えば，０．８である。

（ｋ）S_{BS} ：１対のバイフォンのスペクトラム列のベクトルの距離、すなわち、各スペクトラム要素の２乗和の平方根である。

【００２７】上記数９から明らかなように、類似度のスコアMが０に近いほど、類似度の度合いが高いといえる。

【００２８】次いで、図１を参照して音声データ量削減処理装置の構成及び動作について説明する。図１において、テキストデータベースメモリ２２は、自然発話の書

き下し文の音素ラベルデータにてなるテキストデータベースを記憶し、音声波形データベースメモリ２１は、上記テキストデータベースにおける自然発話の書き下し文の音素ラベルデータに対応する音声波形信号の音声セグメントからなる音声波形データベースを記憶する。基本韻律データ生成部４０は、これら２つのメモリ２１，２２内のデータに基づいて、各音素ラベルに対応して、例えばＬＰＣ分析法を用いて基本周波数F_0を検出するとともに、所定の音声波形分析を行うことによりパワーを検出することにより、基本韻律データとして生成して基本韻律データメモリ５０に格納する。

【００２９】次いで、評価韻律データ生成部４１は、テキストデータベースメモリ２２内の音素ラベルデータと、基本韻律データメモリ５０内の基本韻律データとに基づいて、基本周波数F_0とパワーに加えて、基本周波数F_0の時間方向の傾きと、音素時間長を計算してこれら４つのデータを評価韻律データとして評価韻律データメモリ５１に格納する。また、バイフォンリスト生成部４２は、テキストデータベースメモリ２２内の音素ラベルデータに基づいて、時間方向に隣接する２つの音素並びであるバイフォンを計数して、そのバイフォンと計数値をバイフォンリストメモリ５２に格納する。さらに、バイスペクトラムデータ生成部４３は、テキストデータベースメモリ２２内の音素ラベルデータと音声波形データベースメモリ２１内の音声波形データベースとに基づいて、上述のバイスペクトラムデータ、すなわち、１対のバイフォンのスペクトラム列のベクトルの距離、すなわち、各スペクトラム要素の２乗和の平方根を計算してバイスペクトラムデータメモリ５３に格納する。

【００３０】さらに、評価データ行列生成部４４は、評価韻律データメモリ５１内の評価韻律データと、バイフォンリストメモリ５２内のバイフォンと、バイスペクトラムデータメモリ５３内のバイスペクトラムデータとに基づいて、テキストデータベース内の各バイフォンの列に対して、上述の４つの評価韻律データと１つのバイスペクトラムデータを行方向並置することにより、評価データ行列を生成して評価データ行列メモリ５４に格納する。そして、音声データ量削減処理部４５は、評価データ行列メモリ５４内の評価データ行列と、バイフォンリストメモリ５２内のバイフォンに基づいて、図３の音声データ量削減処理を実行することにより、テキストデータベースメモリ２２内のテキストデータと音声波形データベースメモリ２１内の音声波形データベースの音声データ量を削減して、削減後のデータをテキストデータベースメモリ２２ａと音声波形データベースメモリ２１ａにコピーする。そして、図２の音声合成装置は、音声データ量が削減された後のテキストデータベースメモリ２２ａ及び音声波形データベースメモリ２１ａ内のデータを用いて音声合成処理を行う。

【００３１】図３は、図１のデータ量削減処理部によって実行されるデータ量削減処理を示すフローチャートである。図３において、まず、ステップＳ１においてバイフォンリストメモリ５２内のバイフォンリストから同一の１対のバイフォンを検索し、ステップＳ２において存在するか否かが判断され、ＹＥＳのときはステップＳ３に進む一方、ＮＯのときはステップＳ９に進む。ステップＳ３では、検索した１対バイフォンに対する評価韻律データとバイスペクトラムデータに基づいて，上記数９を用いて類似度のスコアMを計算する。

【００３２】そして、ステップＳ４において１対のバイフォンが類似しているか否かが判断され、具体的にはM≦Ｍｔｈであるか否かが判断され、ＹＥＳのときは類似していると判断してステップＳ５に進む一方、ＮＯのときは類似していないと判断してステップＳ８に進む。ここで、Ｍｔｈは予め決められたしきい値であり、例えば０．１である。次いで、ステップＳ５においてバイフォンリストメモリ５２内のバイフォンリストから同一のバイフォンの件数Nを計算し、ステップＳ６においてN≧Ｎｔｈか否かが判断され、ＹＥＳのときはステップＳ７に進む一方、ＮＯのときはステップＳ８に進む。ここで、Ｎｔｈは予め決められたしきい値であり、例えば５０であり、これは元のテキストデータベースのバイフォンの数にも依存する。ステップＳ６の判断では、音声波形データベースにおいて所定数の同一のバイフォンに対する音声波形データを確保して、音声合成後の音声の品質を所定以上に確保するために設けられる。さらに、ステップＳ７において当該１対のバイフォンのうちの一方のバイフォンの音素をバイフォンリストから削除し、当該バイフォンの音素のラベルデータ及び音声波形データをそれぞれテキストデータベース及び音声波形データベースから削除して、ステップＳ８に進む。ステップＳ８では、バイフォンリストから別の組み合わせの１対のバイフォンを検索して、ステップＳ２に戻る。

【００３３】ステップＳ２においてNOであるときは、ステップＳ９においてバイフォンリストから別の種類のバイフォンを検索し、ステップＳ１０において存在するか否かが判断され、ＹＥＳのときはステップＳ１に戻る一方、ＮＯのときはステップＳ１０ａに進む。ステップＳ１０ａでは、音声データ量が削減された後のテキストデータベースメモリ２２及び音声波形データベースメモリ２１内のデータをそれぞれ、テキストデータベースメモリ２２ａ及び音声波形データベースメモリ２１ａにコピーして当該音声データ量削減処理を終了する。

【００３４】以上の図３の実施形態において、ステップＳ４における判断とステップＳ６における判断がともにＹＥＳであるときに、当該バイフォンの音素のラベルデータと音声波形データとを削除しているが、本発明はこれに限らず、ステップＳ４における判断のみがＹＥＳであるときに、当該バイフォンの音素のラベルデータと音声波形データとを削除してもよい。

【００３５】次いで、図１の音声データ量削減処理装置で音声データ量が削減された、テキストデータベースメモリ２２ａ及び音声波形データベースメモリ２１ａ内のデータを用いて音声合成を行う音声合成装置について以下に説明する。

【００３６】図２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。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２１ａ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２１ａ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。

【００３７】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、音素HMMメモリ２３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。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２１ａ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。

【００３８】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。

【００３９】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）を音声波形データベースメモリ２１ａ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、例えば音素隠れマルコフモデルメモリ２３に格納された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。

【００４０】本実施形態では、音声波形信号データベースメモリ２１ａ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２２ａ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、音声波形データベースメモリ２１ａ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。

【００４１】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の入力はテキストデータベースメモリ２２ａ内の音素表記を伴った音声波形データベースメモリ２１ａ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。

【００４２】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_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、次の表のようになる。

【００４３】

【表３】

――――――――――――――――――――――――――――――――――

索引情報：

　索引番号（１つのファイルに対して付与）

　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）

――――――――――――――――――――――――――――――――――

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

――――――――――――――――――――――――――――――――――

【００４４】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。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、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次の表に示す。ここで、索引番号は、音声波形信号データベースメモリ２１ａ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。

【００４５】

【表４】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	………

【００４６】表４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。

【００４７】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。

【００４８】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。

【００４９】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よい。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。

【００５０】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。

【００５１】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否か、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。

【００５２】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。

【００５３】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。

【００５４】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。

【００５５】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。

【００５６】図４にはこの音声合成装置における図２の音声単位選択部１２の処理を示す。この処理の入力は、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。

【００５７】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，…，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，…，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。

【００５８】音声単位の選択コストは、図４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。

【００５９】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$C^t_j(t_i, u_i)$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。

【００６０】

【数１０】

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

【００６１】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。

【００６２】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。

【００６３】

【数１１】

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

【００６４】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。

【００６５】

【数１２】

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

【００６６】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。

【００６７】

【数１３】

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

【００６８】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。

【００６９】

【数１４】
$$\hat{u}_1^{n} = \min_{u_1, u_2, \cdots, u_n} C(t_1^{n}, u_1^{n})$$

【００７０】上記数１４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）$u_1, u_2, \cdots, u_n = \hat{u}_1^{n}$を表わす関数である。

【００７１】図２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。以下に線形回帰分析における処理の流れを示す。

【００７２】<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。

【００７３】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。

【００７４】以上のように構成された図２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２２ａ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２１ａ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２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。

【００７５】図５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５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２１ａ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２２ａ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メモリ２３内の音素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。

【００７６】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。

【００７７】図６及び図７は、図２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６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。

【００７８】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７のステップＳ３１に進む。

【００７９】図７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。

【００８０】図８は、図２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。

【００８１】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。

【００８２】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２１ａ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がスピーカ１４から出力される。

【００８３】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。

＜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。

＜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。
＜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２１ａの索引情報を作成する処理部。
＜音声合成部１３＞作成された索引情報に従って、音声波形データベースメモリ２１ａ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。
【００８４】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。
【００８５】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、テキストデータベース及び音声波形＊
＊データベー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。
【００８６】
【実施例】本発明者は、本実施形態の図１の音声データ量削減装置を用いて、音声波形データベースの音声データ量を削減し、かつ音声データ量が削減された音声波形データベースに基づいて、図２の音声合成装置を用いて音声合成した実験結果について以下に説明する。
【００８７】まず、本発明者らによる音素バランス調査について説明する。本特許出願人が所有するテキストデータベース及び音声波形データベース（以下、ＣＨＡＴＲという。）で使用される日本語話者のラベルデータに基づいて音素バランス調査を行った。この調査では、音素ラベルデータから抽出した音素列から、２音素ずつ組み合わせたバイフォンデータを作成した。例えば、ラベルデータの発話内容が「あらゆる」であった場合、抽出した音素列は、
【数１５】＃　ａ　ｒ　ａ　ｙ　ｕ　ｒ　ｕ　＃
となる。ここで、＃は発話開始記号又は発話終了記号を示す。この音素列から２音素ずつ組み合わせたバイフォンリストは、
【数１６】＃＿ａ　ａ＿ｒ　ｒ＿ａ　ａ＿ｙ　ｙ＿ｕ　ｕ＿ｒ　ｒ＿ｕ　ｕ＿＃
の８種類となる。
【００８８】次いで、バイフォンリストの作成例について説明する。この作成例の結果を次の表に示す。
【００８９】
【表５】

バイフォンリストの作成例
――――――――――――――――――――――――――――――――――――
発話内容：あらゆる現実を全て自分の方へねじ曲げたのだ。
――――――――――――――――――――――――――――――――――――
音素列：
＃ａｒａｙｕｒｕｇｅＮｊｉtsｕｏ
＃ｓｕｂｅｔｅ
＃ｊｉｂｕＮｎｏｈｏｏｅ
＃ｎｅｊｉｍａｇｅｔａｎｏｄａ＃
――――――――――――――――――――――――――――――――――――
バイフォンリスト：
#_a a_r r_a a_y y_u u_r r_u u_g g_e e_N N_j j_i i_ts ts_u u_o o_#
#_s s_u u_b b_e e_t t_e e_#

#_j j_i i_b b_u u_N N_n n_o o_h h_o o_o o_e e_#
#_n n_e e_j j_i i_m m_a a_g g_e e_t t_a a_n n_o o_d d_a a_#

【００９０】上記の場合、バイフォンリストは、５０件＊示す。４４種類となり、６件が重複したバイフォンであることがわかる。ここで、件数別バイフォンリストを次の表に＊

【００９１】
【表６】

件数別バイフォンリスト

１件：#_a a_r r_a a_y y_u u_r r_u u_g e_N N_j i_ts ts_u u_o o_#
#_s s_u u_b b_e t_e
#_j i_b b_u u_N N_n o_h h_o o_o o_e
#_n n_e e_j i_m m_a a_g t_a a_n o_d d_a a_#

２件：g_e, e_t, e_#, n_o

３件：j_i

【００９２】次いで、ＣＨＡＴＲ内のある日本語話者ＭＹＡに対してのバイフォンリストを作成し、以下のように出現頻度を集計することにより音素バランスを調べた。ここで、話者ＭＹＡの音声波形データベースの情報を次の表に示す。

【００９３】
【表７】話者ＭＹＡの音声波形データベースの情報

発話内容	ファイル数
５０３文章	５０３ファイル
２２文章	２２ファイル
旅行会話	６７５ファイル
合計	１２００ファイル

【００９４】日本語話者ＭＹＡに対する音素バランス調査の結果を以下に示す。

【００９５】
【表８】

バイフォンリスト集計結果

バイフォンの件数：６０，４０５件（ラベリングミス除く）
バイフォンの種類：４１５通り（ラベリングミス除く）

【００９６】
【表９】出現頻度の多いバイフォンリスト１０種類

バイフォン	件数
ｏ＿ｏ	１４３６
ａ＿＃	１３２６
ｄ＿ｅ	９９２
ｍ＿ａ	９８８
ａ＿ｉ	９７０
ｎ＿ｏ	９５５
ｋ＿ａ	８８６
ｔ＿ｏ	７８６
ｇ＿ａ	７７２
ｉ＿ｍ	７４４

【００９７】
【表１０】

出現頻度１件のバイフォン：４２通り

I_g I_n U_b U_r d_y f_o t_y I_kk a_ff a_gg ch_e dd_a e_dd e_gg e_hh e_ss ff_u
gg_a gg_u hh_a hh_i hh_o i_hh i_jj jj_i o_hh o_zz pp_U ss_o ss_u tt_U u_dd
zz_u U_tts cch_I cch_e cch_u e_cch o_tts ssh_I ssh_U tts_U

【００９８】日本語話者ＭＹＡに対する音素バランス調査におけるバイフォンリストの詳細結果を以下に示す。

【００９９】 ＊ ＊【表１１】

出現頻度５００件以上のバイフォンリスト
（バイフォン／件数）

o_o	1436	a_#	1326	d_e	992	m_a	988	a_i	970	n_o	955
k_a	886	t_o	786	g_a	772	i_m	744	a_r	738	t_a	716
w_a	696	n_i	678	a_s	651	o_#	604	o_k	603	n_a	594
s_U	586	r_u	582	k_u	581	o_n	573	y_o	563	a_k	551
r_i	544	i_#	536	o_r	517	k_o	507	sh_i	506	r_a	506
u_u	504	e_#	503	a_N	503						

【０１００】 ※ ※【表１２】

出現頻度１００件以上５００件未満のバイフォンリスト
（バイフォン／件数）

#_k	490	t_e	482	m_o	465	u_n	452	r_e	444	i_n	440
sh_I	424	e_N	415	s_a	412	o_m	405	a_n	397	s_u	394
I_t	394	i_r	365	a_sh	364	#_h	363	ts_u	358	h_a	356
#_s	356	g_o	355	k_i	349	o_d	348	e_s	348	o_t	340
e_r	340	i_N	332	y_u	326	d_a	324	U_#	317	u_#	313
d_o	312	u_r	311	y_a	310	o_sh	309	ch_i	299	s_o	299
s_e	288	i_t	286	a_t	281	a_m	280	U_k	279	j_i	276
h_o	274	o_g	271	N_d	267	k_e	262	n_e	261	m_i	259
o_N	254	e_k	254	e_e	248	i_k	243	sh_o	242	r_o	239
o_y	238	N_n	238	#_n	237	i_g	234	#_o	233	u_g	232
u_k	230	i_d	227	u_d	222	e_t	218	a_d	218	i_i	217
o_h	216	o_s	214	a_g	210	o_i	205	e_n	205	m_e	200
#_#	199	I_k	198	#_m	192	tt_e	188	z_a	188	u_m	188
k_y	188	#_g	187	a_tt	181	e_d	180	e_g	178	#_i	173
b_a	170	u_N	169	e_sh	167	e_i	164	a_a	162	#_d	160
#_j	159	#_a	159	#_t	157	z_u	156	e_m	155	N_k	153
o_ch	152	j_u	152	#_y	152	i_sh	149	o_z	145	b_u	140
i_o	139	#_sh	137	i_h	137	u_i	132	N_g	131	o_w	129
i_s	129	tt_a	127	k_U	126	e_w	125	u_o	124	i_w	124
e_o	123	a_ts	118	j_o	115	a_w	114	g_e	112	o_b	111
o_j	110	a_o	110	u_t	109	U_t	109	k_I	108	a_e	106
N_t	106	u_s	102								

【０１０１】 ★ ★【表１３】

出現頻度５０件以上１００件未満のバイフォンリスト
（バイフォン／件数）

u_y	99	m_u	99	i_y	99	u_sh	98	h_i	97	h_I	97
a_ch	95	i_ch	93	ts_U	92	ch_I	92	o_a	92	N_j	92
f_u	91	u_b	90	g_u	90	g_i	90	U_s	90	#_ch	88
i_b	83	sh_a	82	f_U	81	b_e	81	a_h	81	o_tt	80
N_w	80	i_ts	78	u_j	77	N_sh	76	a_y	76	b_i	75
#_w	74	a_z	73	a_b	73	o_e	71	i_a	71	b_o	71
sh_u	70	i_tt	69	e_b	68	N_b	68	N_#	68	tt_o	67
i_z	67	#_b	67	a_j	66	i_j	65	e_a	65	#_r	65

#_f 65	u_w 63	u_h 63	o_ts 62	u_z 62	h_y 61
N_m 61	r_y 60	e_ts 57	ch_o 57	h_e 56	N_s 55
e_y 51					

【０１０２】

＊ ＊【表１４】

出現頻度１０件以上５０件未満のバイフォンリスト
（バイフォン／件数）

z_e 47	u_e 47	o_f 47	N_r 47	N_o 47	I_ts 46
ch_u 45	i_e 45	i_pp 42	z_o 42	N_h 42	#_z 41
n_y 40	#_u 40	a_u 39	#_ts 38	o_u 34	n_u 34
j_a 34	g_y 34	pp_a 33	u_ts 32	p_a 32	I_s 32
#_e 32	kk_a 30	e_h 29	N_y 29	u_ch 28	e_tt 26
U_sh 26	i_u 26	N_p 26	i_kk 25	b_y 25	e_z 23
e_j 23	a_f 23	a_kk 22	U_n 21	u_tt 20	i_f 20
#_p 20	kk_u 19	o_kk 18	e_kk 18	u_f 18	i_ssh 17
pp_o 17	N_ch 17	I_ch 17	o_p 17	m_y 17	u_a 16
p_u 16	N_z 16	N_i 16	kk_o 15	ch_a 15	p_e 15
j_e 15	N_e 15	I_# 15	o_pp 14	t_i 14	U_ts 13
p_o 13	ssh_i 12	pp_u 12	kk_i 12	w_o 12	p_i 12
kk_y 11	a_pp 11	U_d 11			

【０１０３】

※ ※【表１５】

出現頻度１件以上１０件未満のバイフォンリスト
（バイフォン／件数）

u_kk 9	ss_e 9	i_p 9	e_u 9	N_f 9	ssh_a 8
a_ssh 8	u_pp 8	pp_i 8	u_p 8	r_U 8	f_i 8
pp_y 7	e_ch 7	U_tt 7	p_y 7	e_f 7	e_ssh 6
cch_i 6	kk_e 6	i_ss 6	N_ts 6	e_p 6	d_i 6
a_p 6	U_z 6	U_p 6	U_g 6	tts_u 5	ssh_o 5
o_cch 5	I_tt 5	I_sh 5	f_e 5	u_ssh 4	i_tts 4
i_cch 4	cch_a 4	u_ss 4	tt_i 4	p_U 4	U_h 4
N_u 4	N_a 4	a_cch 3	ss_a 3	pp_e 3	kk_U 3
dd_o 3	a_ss 3	U_kk 3	I_pp 3	j_I 3	f_a 3
I_d 3	u_cch 2	o_ssh 2	cch_o 2	sh_e 2	i_dd 2
hh_e 2	ff_e 2	e_pp 2	e_ff 2	a_hh 2	w_e 2
p_I 2	d_u 2	U_j 2	I_p 2	I_m 2	I_h 2
I_b 2	tts_U 1	ssh_U 1	ssh_I 1	o_tts 1	e_cch 1
cch_u 1	cch_e 1	cch_I 1	U_tts 1	zz_u 1	u_dd 1
tt_U 1	ss_u 1	ss_o 1	pp_U 1	o_zz 1	o_hh 1
jj_i 1	i_jj 1	i_hh 1	hh_o 1	hh_i 1	hh_a 1
gg_u 1	gg_a 1	ff_u 1	e_ss 1	e_hh 1	e_gg 1
e_dd 1	dd_a 1	ch_e 1	a_gg 1	a_ff 1	I_kk 1
t_y 1	f_o 1	d_y 1	U_r 1	U_b 1	I_n 1
I_g 1					

【０１０４】以上示したように、ある日本語話者に対して、バイフォンの分布にかなりバラツキがあることがわかる。これに基づいて、本発明者らは、音声合成後の品質を低下させずに、図３の処理を用いて音声データ量の削減を行った。

【０１０５】音声波形データベースの縮小の実験におい

ては、音素的にバランスのとれた文章セット（文数ｎ＝５２５）及び旅行用対話テキストのセット（文数ｎ＝６７５）を読み上げる男性話者ＭＹＡ（職業アナウンサーではない）による音声波形データを使用した。そして、本発明者らは、上述の冗長性尺度、すなわち類似度の判断基準に従って音声波形データベースを元のサイズの３分の２及び半分にまで、すなわち１２００発話から各々８０６及び６２５発話にまで音声データ量を削減した。ここで、それぞれのデータベースを第１減量ＭＹＡと第２減量ＭＹＡと呼ぶ。音声データ量を削減した後に、得られた音声波形データベースのスコアを、韻律的特徴パラメータ及び音響特徴パラメータにかかる距離に関して測定した。表１６は、３つの音声波形データベースの各々における音声ファイルの計数値を示している。表１７は韻律的特徴パラメータのスコアを示し、表１８は音響特徴パラメータのスコアを示している。

【０１０６】
【表１６】

音声波形データベースの縮小

データベース	音声ｾｸﾞﾒﾝﾄ数	音声波形ＤＢ内の文の数	テストされた文の数
ＭＹＡ	１２００	５２５／６７５	１１９８
第１減量ＭＹＡ	８０６	４７４／３３２	８０４
第２減量ＭＹＡ	６２５	３３８／２３７	６２３

（注）ＤＢはデータベースを示す。以下、同様である。

【０１０７】
【表１７】

韻律的特徴パラメータに基づくスコア

データベース	全体のスコア	音声波形ＤＢ内の文の数	スコアの平均値
ＭＹＡ	９５６０	１１９８	＊７．９８０５１８
第１減量ＭＹＡ	６９８３	８０４	８．６８５６２７
第２減量ＭＹＡ	５６０５	６２３	８．９９７５５７

（注）＊は最低のスコアを示す。

【０１０８】
【表１８】

バイスペクトルに基づくスコア

データベース	全体のスコア	音声波形ＤＢ内の文の数	スコアの平均値
ＭＹＡ	１５０２０	１１９８	１２．５３７７７０
第１減量ＭＹＡ	９７５７	８０４	＊１２．１３６１１９
第２減量ＭＹＡ	７５７１	６２３	１２．１５２６８７

（注）＊は最低のスコアを示す。

【０１０９】表１７及び表１８から明らかなように、音声波形データベースの音声データ量を削減した結果、韻律的特徴パラメータの変動性ではスコアで約０．７ポイント低下しているものの、音響特徴パラメータにおける平滑性の面ではスコアで０．４ポイント向上していることが分かる。

【０１１０】以上説明したように、本実施形態によれば、バイフォンリストに基づいて各１対のバイフォンに対する韻律的特徴パラメータと音響的特徴パラメータとに関する所定の類似度を計算し、上記計算された類似度が所定の第１のしきい値以上であり、かつ上記バイフォンのリスト中の同一のバイフォンの数が所定の第２のしきい値以上であるときに、当該１対のバイフォンのうちの一方のバイフォンの音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベースから削除することにより音声データ量を削減する。従って、音声合成時の音声品質を実質的に低下させることなく、音声波形データベースを格納するメモリ容量を削減することができ、音声合成時の探索速度を高めることができる。

【０１１１】
【発明の効果】以上詳述したように本発明に係る音声合成装置のための音声データ量削減装置によれば、１対の音素のリストに基づいて各１対の音素に対する韻律的特徴パラメータと音響的特徴パラメータとに関する所定の

類似度を計算し、上記計算された類似度が所定の第１のしきい値以上であるときに、当該各１対の音素のうちの一方の１対の音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベースから削除することにより音声データ量を削減する。従って、音声合成時の音声品質を実質的に低下させることなく、音声波形データベースを格納するメモリ容量を削減することができ、音声合成時の探索速度を高めることができる。

【０１１２】また、上記計算された類似度が所定の第１のしきい値以上でありかつ上記バイフォンのリスト中の同一のバイフォンの数が所定の第２のしきい値以上であるときに、当該各１対の音素のうちの一方の１対の音素に係る音声波形信号の音声セグメントのデータを上記音声波形データベースから削除することにより音声データ量を削減する。従って、音声波形データベースを格納するメモリ容量を削減することができ、音声合成時の探索速度を高めることができるとともに、音声波形データベースにおいて所定数の同一のバイフォンに対する音声波形データを確保して、音声合成後の音声の品質を所定以上に確保することができる。

【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データ量削減処理装置のブロック図である。

【図２】　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３】　図１の音声データ量削減処理部によって実行される音声データ量削減処理を示すフローチャートである。

【図４】　図２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５】　図２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６】　図２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。

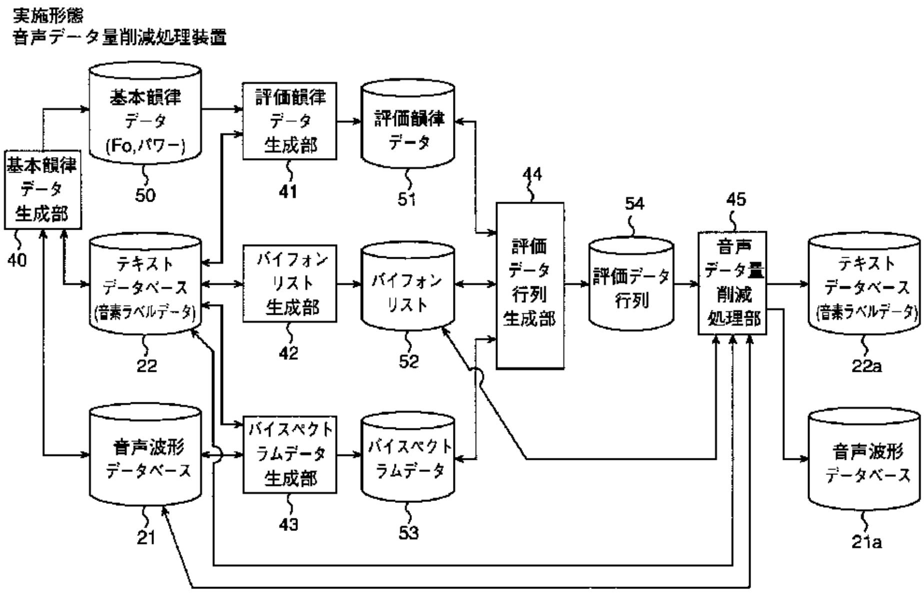
【図７】　図２の重み係数学習部によって実行される重み係数学習処理の第２の部分のフローチャートである。

【図８】　図２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。

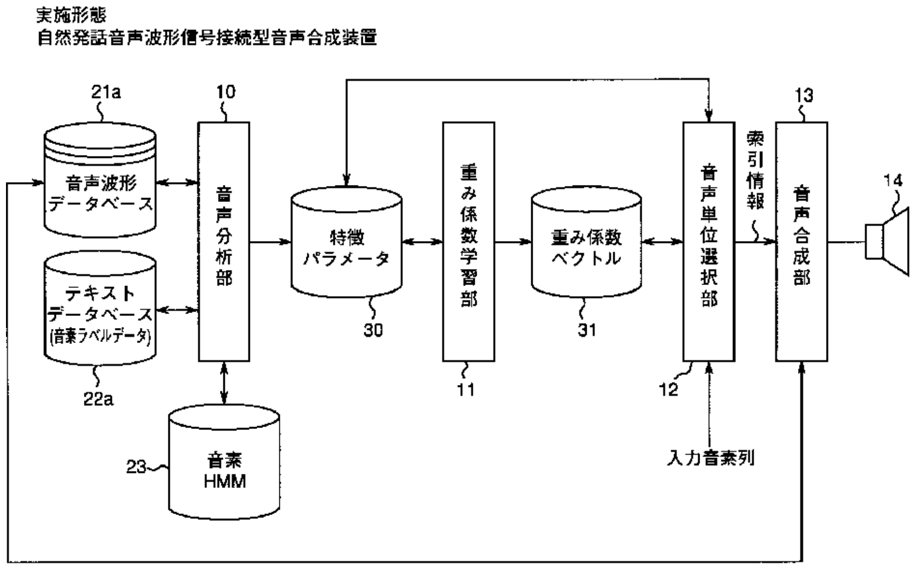
【符号の説明】

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、
２１，２１ａ…音声波形信号データベースメモリ、
２２，２２ａ…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、
４０…基本韻律データ生成部、
４１…評価韻律データ生成部、
４２…バイフォンリスト生成部、
４３…バイスペクトラムデータ生成部、
４４…評価データ行列生成部、
４５…音声データ削減処理部、
５０…基本韻律データメモリ、
５１…評価韻律データメモリ、
５２…バイフォンリストメモリ、
５３…バイスペクトラムデータメモリ、
５４…評価データ行列メモリ。

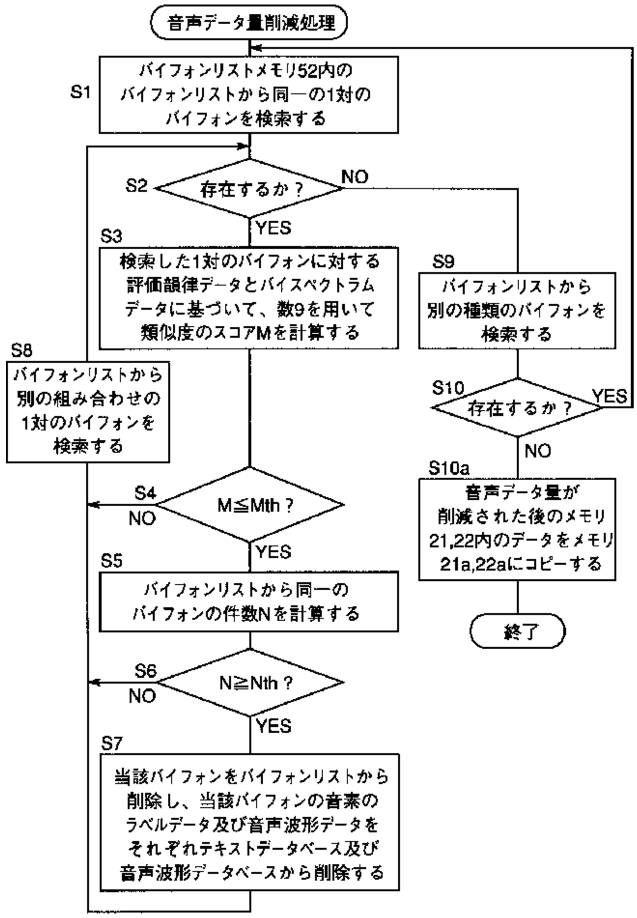
【図１】



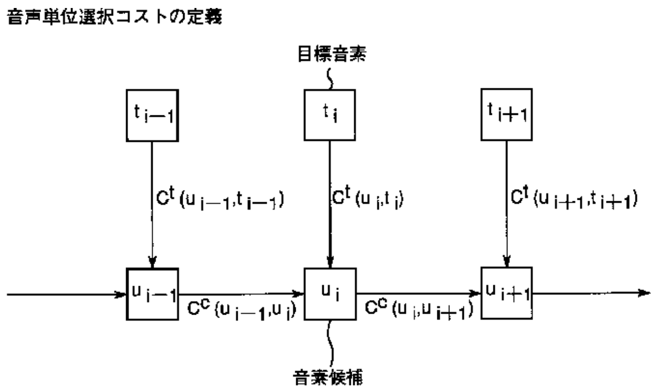
【図２】


実施形態
自然発話音声波形信号接続型音声合成装置
21a
音声波形データベース
テキストデータベース
(音素ラベルデータ)
22a
10
音声分析部
音素HMM
23
特徴パラメータ
30
重み係数学習部
11
重み係数ベクトル
31
音声単位選択部
12
入力音素列
索引情報
13
音声合成部
14

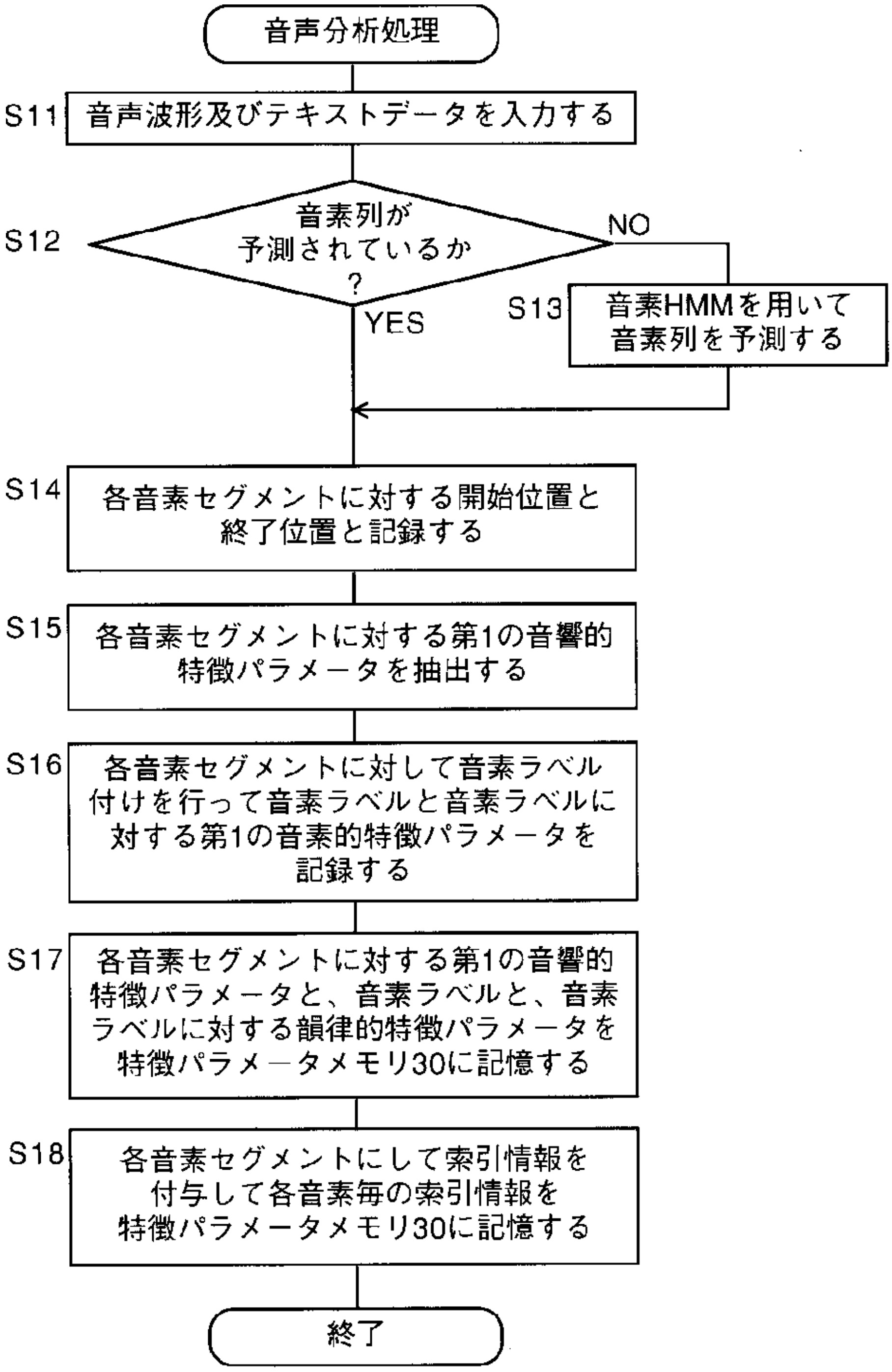
【図３】


音声データ量削減処理
S1
バイフォンリストメモリ52内のバイフォンリストから同一の1対のバイフォンを検索する
S2
存在するか？
NO
YES
S3
検索した1対のバイフォンに対する評価韻律データとバイスペクトラムデータに基づいて、数9を用いて類似度のスコアMを計算する
S8
バイフォンリストから別の組み合わせの1対のバイフォンを検索する
S4
M≦Mth？
NO
YES
S5
バイフォンリストから同一のバイフォンの件数Nを計算する
S6
N≧Nth？
NO
YES
S7
当該バイフォンをバイフォンリストから削除し、当該バイフォンの音素のラベルデータ及び音声波形データをそれぞれテキストデータベース及び音声波形データベースから削除する
S9
バイフォンリストから別の種類のバイフォンを検索する
S10
存在するか？
YES
NO
S10a
音声データ量が削減された後のメモリ21,22内のデータをメモリ21a,22aにコピーする
終了

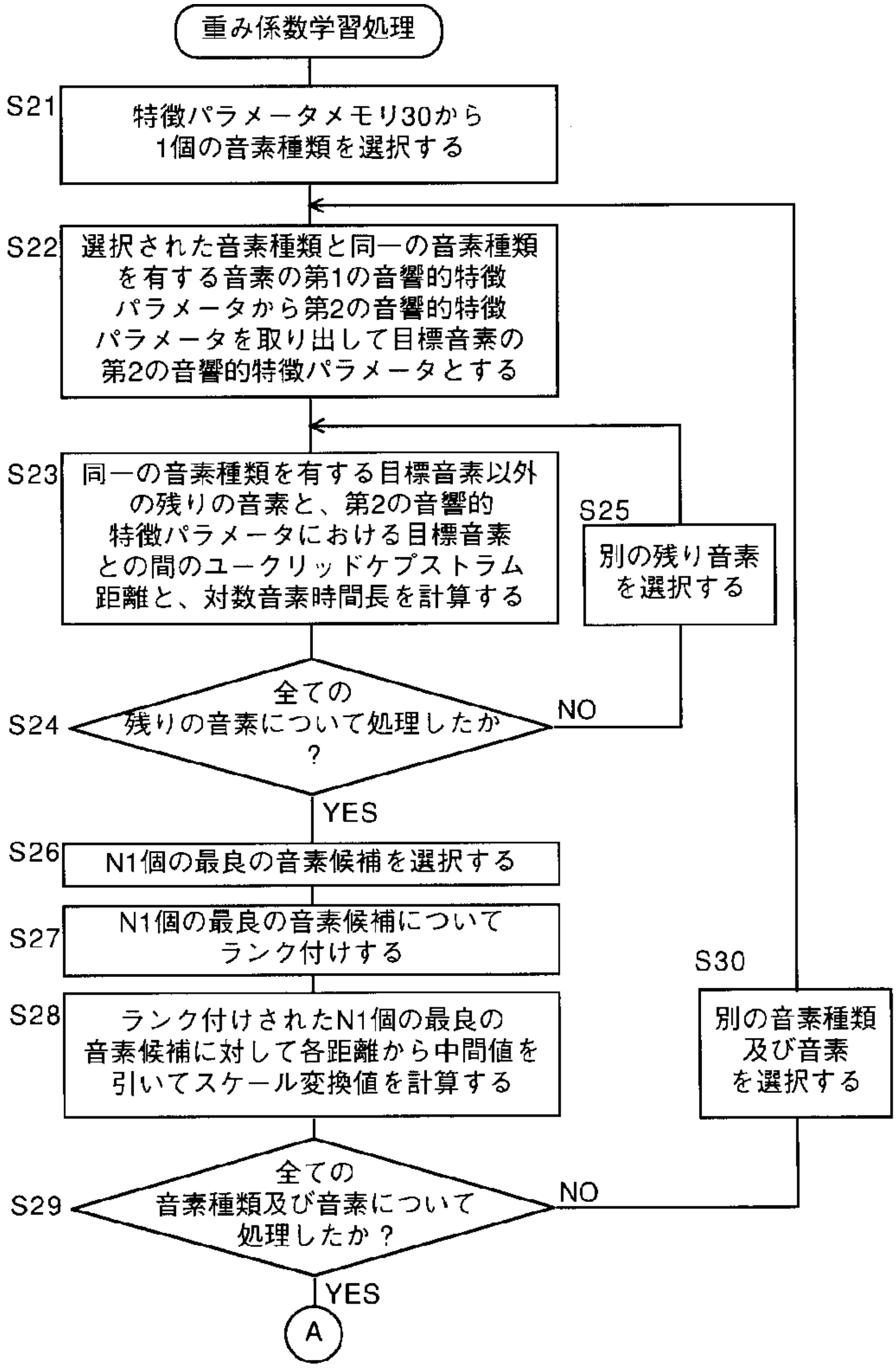
【図４】


音声単位選択コストの定義
目標音素
t i−1
t i
t i+1
Ct (u i−1,t i−1)
Ct (u i,t i)
Ct (u i+1,t i+1)
u i−1
u i
u i+1
Cc (u i−1,u i)
Cc (u i,u i+1)
音素候補

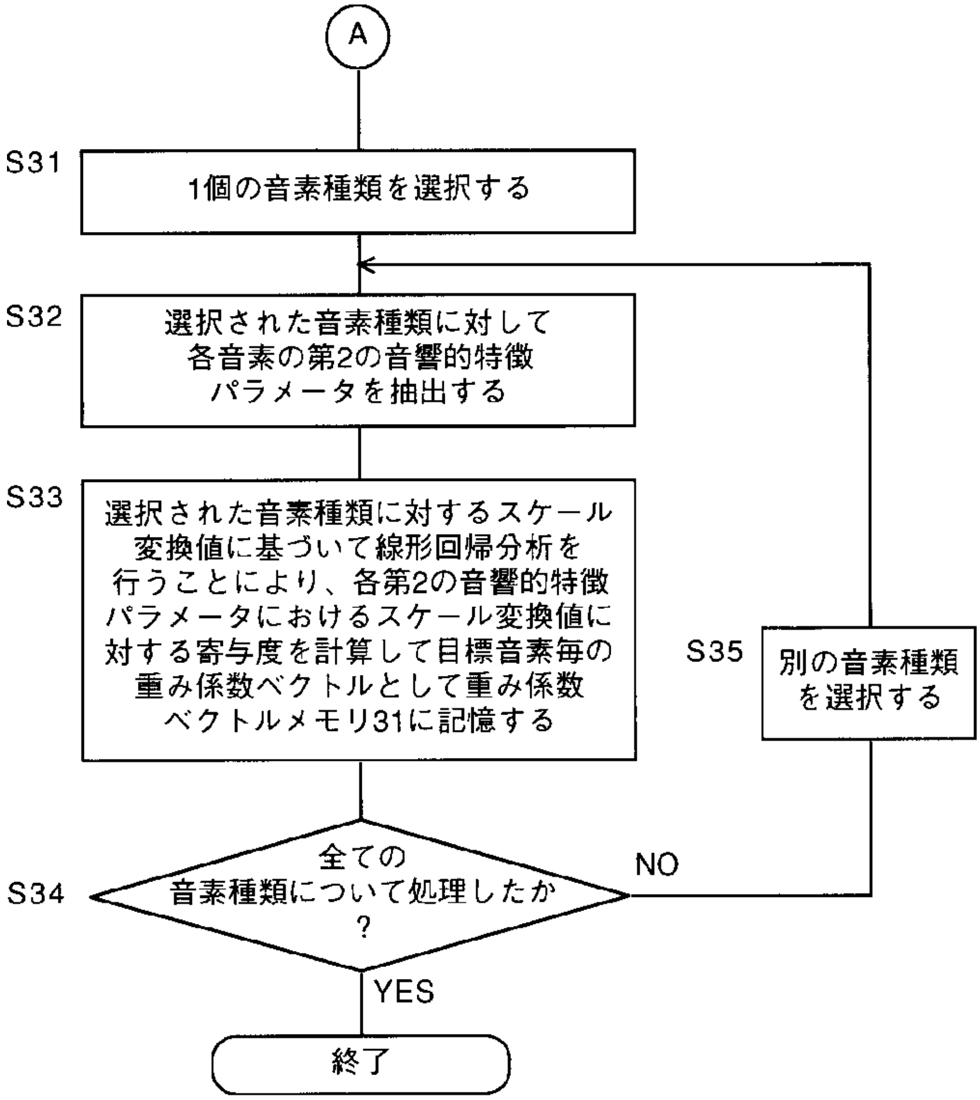
【図５】


音声分析処理
S11
音声波形及びテキストデータを入力する
S12
音素列が予測されているか？
NO
YES
S13
音素HMMを用いて音素列を予測する
S14
各音素セグメントに対する開始位置と終了位置と記録する
S15
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を抽出する
S16
各音素セグメントに対して音素ラベル付けを行って音素ラベルと音素ラベルに対する第1の音素的特徴パラメータを記録する
S17
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韻律的特徴パラメータを特徴パラメータメモリ30に記憶する
S18
各音素セグメントにして索引情報を付与して各音素毎の索引情報を特徴パラメータメモリ30に記憶する
終了

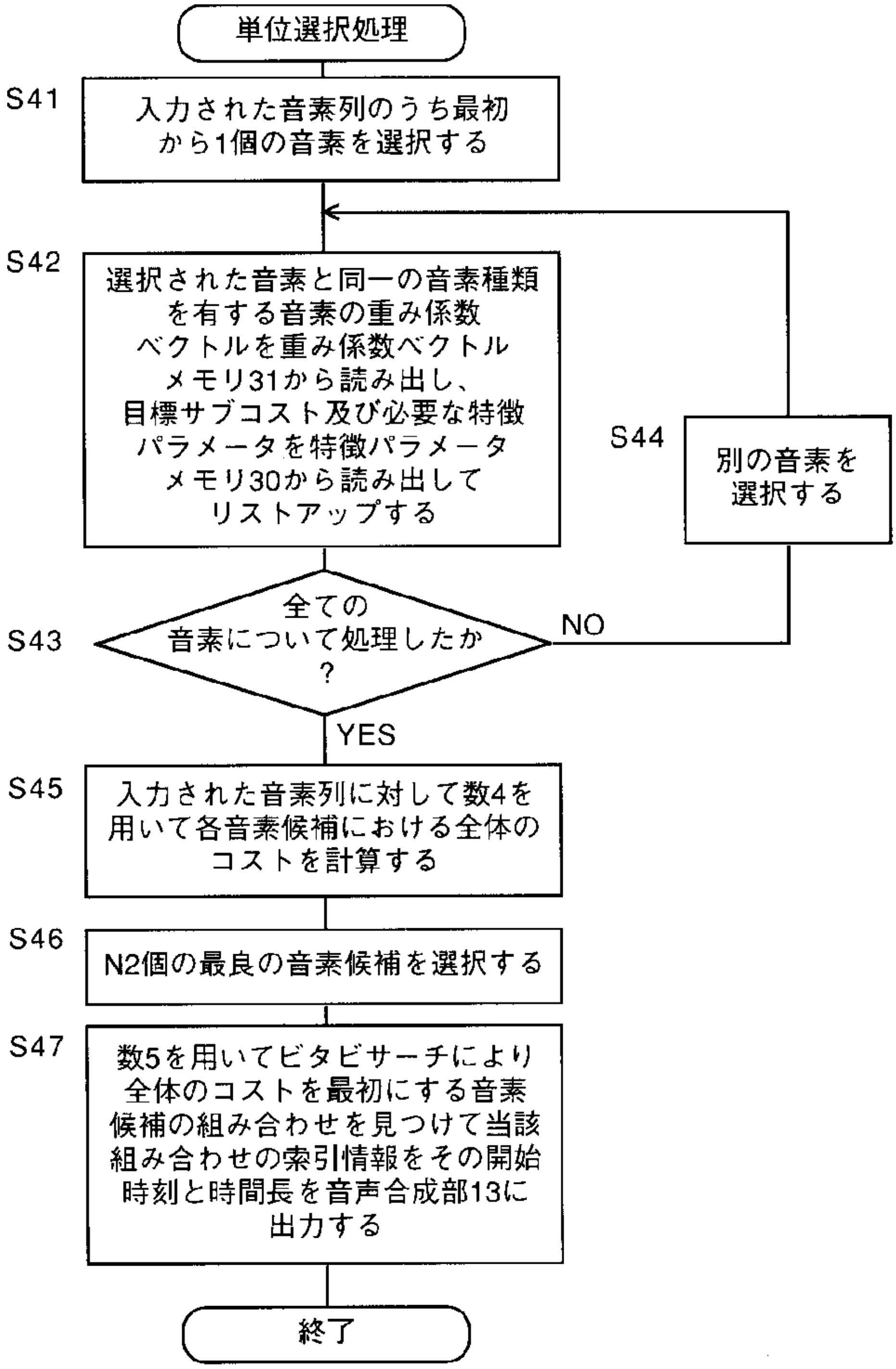
【図６】


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７】


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
S34
全ての
音素種類について処理したか
？
NO
S35
別の音素種類
を選択する
YES
終了

【図８】


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
S43
全ての音素について処理したか？
NO
S44
別の音素を選択する
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

フロントページの続き

(56)参考文献　特開　平11－95796（ＪＰ，Ａ）
特開　平10－39889（ＪＰ，Ａ）
特開　平８－248975（ＪＰ，Ａ）
特開　昭56－51800（ＪＰ，Ａ）
特開　平７－84590（ＪＰ，Ａ）
特開　平10－49193（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G10L　13/06