(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2018-4997
(P2018-4997A)
(43)公開日　平成30年1月11日(2018.1.11)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 13/06	(2013.01)	G10L	13/06	130
G10L 13/06	(2013.01)	G10L	13/08	150A

審査請求 未請求　請求項の数 4　OL　（全 13 頁）

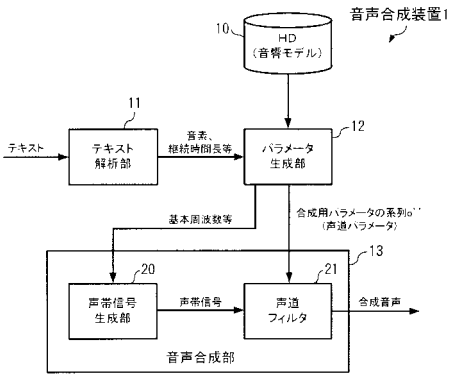
(21)出願番号　特願2016-132586 (P2016-132586)
(22)出願日　平成28年7月4日 (2016.7.4)

(71)出願人　000004352
日本放送協会
東京都渋谷区神南２丁目２番１号
(71)出願人　591053926
一般財団法人ＮＨＫエンジニアリングシステム
東京都世田谷区砧一丁目１０番１１号
(74)代理人　100121119
弁理士　花村　泰伸
(72)発明者　尾上　和穂
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内
(72)発明者　齋藤　礼子
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内

最終頁に続く

(54)【発明の名称】音声合成装置及びプログラム

(57)【要約】
【課題】合成音声が自然の音に近くなるように、合成用パラメータを生成する。
【解決手段】音声合成装置１のパラメータ生成部１２は、テキストの音素列について、音素の各状態に対応するパラメータの系列ｏを音響モデルから生成する。パラメータ生成部１２は、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布の尤度が最大となるパラメータの系列ｏ’を生成し、さらに、合成用パラメータの系列ｏ’に対し、ＭＦＣＣ等の分散に基づいた正規乱数（ガウスノイズ）を付加し、合成用パラメータの系列ｏ”を生成する。音声合成部１３は、ピッチの基本周波数等に基づいて、テキストに対応する声帯信号を生成し、声帯信号に対し、合成用パラメータの系列ｏ”を用いて声道フィルタによるフィルタ処理を施すことで、合成音声を生成する。
【選択図】図１



実施例1

【特許請求の範囲】
【請求項１】
予め学習した音響モデルを用いて、テキストに対応した合成音声を生成する音声合成装置において、
前記音響モデルを用いて、テキストに対応した合成用パラメータを生成するパラメータ生成部と、
前記パラメータ生成部により生成された前記合成用パラメータを用いて、前記テキストに対応した声帯信号に対し、声道フィルタのフィルタ処理を施して合成音声を生成する音声合成部と、を備え、
前記パラメータ生成部は、
前記テキストを構成する音素の各状態に対応する特徴量を、前記音響モデルから読み出し、前記特徴量の確率分布の尤度が最大となるパラメータを生成するパラメータ生成手段と、
前記テキストを構成する音素の各状態に対応する特徴量の分散を、前記音響モデルから読み出し、前記パラメータ生成手段により生成された前記パラメータに、前記分散に基づいた正規乱数を付加し、前記合成用パラメータを生成する分散付加手段と、を備えたことを特徴とする音声合成装置。
【請求項２】
請求項１に記載の音声合成装置において、
前記テキストを構成する音素の各状態に対応する特徴量を、ＭＦＣＣ、ＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔとする、ことを特徴とする音声合成装置。
【請求項３】
請求項１または２に記載の音声合成装置において、
前記パラメータ生成部に備えた前記分散付加手段に代わる新たな分散付加手段は、
前記テキストを構成する音素の各状態に対応する特徴量の分散を、前記音響モデルから読み出し、前記パラメータ生成手段により生成された前記パラメータに、前記分散に基づいた正規乱数を付加し、複数の合成用パラメータを生成し、
前記音声合成部に代わる新たな音声合成部は、
前記新たな分散付加手段により生成された前記複数の合成用パラメータを用いて、前記テキストに対応した声帯信号に対して声道フィルタのフィルタ処理を施し、前記複数の合成用パラメータのそれぞれに対応する複数の合成音声を生成し、前記複数の合成音声を平均化する、ことを特徴とする音声合成装置。
【請求項４】
コンピュータを、請求項１から３までのいずれか一項に記載の音声合成装置として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、統計的音響モデルを用いて、音声合成のための合成用パラメータを生成する音声合成装置及びプログラムに関する。
【背景技術】
【０００２】
従来、テキストを分析して言語情報を生成し、言語情報からテキストの文に対応する音声波形を生成する音声合成手法が知られている。この音声合成手法は、大きく２種類に分類することができ、一方が波形接続音声合成手法であり、他方が統計的音声合成手法である。
【０００３】
波形接続音声合成手法は、得られた音韻の系列に従って音声素片を選択し、韻律情報に従って音声素片のピッチ及び継続時間長を変形して接続することで、合成音声を生成する。これに対し、統計的音声合成手法は、予め学習しておいた統計量に基づいて、最適な合

成用パラメータを生成することで、合成音声を生成する。

【０００４】

具体的には、統計的音声合成手法を用いる音声合成装置は、入力したテキストを音素に変換し、予め学習しておいた各音素の状態毎の統計量を有する音響モデルを用いて、確率分布の尤度が最も高い合成用パラメータを生成する（例えば、非特許文献１を参照）。

【０００５】

ここで、音響モデルをλ、音素の状態の系列をｑ、音響モデルλから観測されるパラメータの系列をｏ、音素の状態の系列ｑと音響モデルλにおいて、パラメータの系列ｏが観測される確率分布をＰとすると、合成用パラメータは、以下の式にて推定される。

【数１】

$$o' = \arg\max_{o} P(o \mid q, \lambda) \qquad \cdots (1)$$

前記数式（１）は、音素の状態の系列ｑと音響モデルλにおいて、パラメータの系列ｏが観測される確率分布Ｐの尤度が最大となるパラメータの系列ｏを、合成用パラメータの系列ｏ’として求めることを表している。

【０００６】

音響モデルを生成する際には、大量の学習用音声から、音素の状態毎にＭＦＣＣ（Mel-Frequency Cepstrum Coefficients：メル周波数ケプストラム係数）、その一次回帰係数Δ及び二次回帰係数ΔΔを含む特徴量が学習される。そして、音響モデルには、これらのパラメータの平均及び分散、及び、音素の状態間の遷移確率、並びに、出力確率が格納される。このように、合成用パラメータの系列は、ＭＦＣＣの一次回帰係数Δ及び二次回帰係数ΔΔが考慮されて生成されるため、発話文章全体で平均をトレースするような滑らか値となる。

【０００７】

図６は、従来技術において、音響モデルを用いて生成された合成用パラメータの系列を示す図である。横軸は時間ｔを示し、縦軸は、ＭＦＣＣ　ｃ及びＭＦＣＣの一次回帰係数Δｃの値を示す。

【０００８】

図６に示すように、従来の統計的音声合成手法により生成される合成用パラメータの系列ｏ’（ｃ，Δｃ）は、音素を構成する複数の状態について、実線で表す平均を基準にして、滑らかに変化する値となる。四角の点線は分散の範囲を示す。

【０００９】

しかしながら、このような統計的音声合成手法により生成される合成用パラメータの系列ｏ’は、図６に示したとおり、平均を基準にして滑らかに変化する値となっており、自然な音声から得られる揺らぎを持つパラメータの振る舞いとは一致しない。つまり、この合成用パラメータの系列ｏ’により生成される合成音声は、不自然な音となってしまう。

【００１０】

一方で、音素間で合成用パラメータの値が不連続とならないように、滑らかに変化する自然な合成用パラメータの系列を生成する手法が開示されている（例えば、特許文献１を参照）。この手法は、言語レベル（音素、音節、単語等）を単位とする言語区間のスペクトルパラメータを算出し、複数の言語区間のそれぞれに対するスペクトルパラメータを、言語情報に基づいて複数のクラスタに分類する。そして、同一クラスタに属する複数のスペクトルパラメータの特徴を示す音響モデルをクラスタ毎に生成する。合成用パラメータの系列を生成する際には、テキストの言語情報に基づいて、クラスタ毎の音響モデルからテキストに応じた音響モデルを選択する。

【先行技術文献】

【特許文献】

【００１１】

【特許文献１】特開２０１０－２３７３２３号公報

【非特許文献】
【００１２】
【非特許文献１】Keiichi Tokuda, Heiga Zen, “Fundamentals and recent advances in HMM-based speech synthesis”［online］, ［平成２８年５月２０日検索］, インターネット＜ＵＲＬ：http://hts.sp.nitech.ac.jp/?Tutorial＞
【発明の概要】
【発明が解決しようとする課題】
【００１３】
前述の特許文献１の手法は、言語区間に対応したクラスタ毎の音響モデルから、テキストに応じた音響モデルを選択することで、音素間で不連続点のない滑らかに変化する合成用パラメータの系列を生成することができる。
【００１４】
しかしながら、人間の発声には毎回揺らぎが存在し、実際の音声信号から得られる特徴量は、滑らかに変化する特性を有さない。このため、特許文献１の手法により生成される合成用パラメータの系列は、発話文章全体で平均をトレースするような滑らかな値となることに変わりがなく、前述の非特許文献１の手法と同様に、合成音声の音は、実際の人間の発声に特有な揺らぎが表現できず、不自然であるという問題があった。
【００１５】
そこで、本発明は前記課題を解決するためになされたものであり、その目的は、自然の音に近い合成音声を生成可能な音声合成装置及びプログラムを提供することにある。
【課題を解決するための手段】
【００１６】
前記課題を解決するために、請求項１の音声合成装置は、予め学習した音響モデルを用いて、テキストに対応した合成音声を生成する音声合成装置において、前記音響モデルを用いて、テキストに対応した合成用パラメータを生成するパラメータ生成部と、前記パラメータ生成部により生成された前記合成用パラメータを用いて、前記テキストに対応した声帯信号に対し、声道フィルタのフィルタ処理を施して合成音声を生成する音声合成部と、を備え、前記パラメータ生成部が、前記テキストを構成する音素の各状態に対応する特徴量を、前記音響モデルから読み出し、前記特徴量の確率分布の尤度が最大となるパラメータを生成するパラメータ生成手段と、前記テキストを構成する音素の各状態に対応する特徴量の分散を、前記音響モデルから読み出し、前記パラメータ生成手段により生成された前記パラメータに、前記分散に基づいた正規乱数を付加し、前記合成用パラメータを生成する分散付加手段と、を備えたことを特徴とする。
【００１７】
また、請求項２の音声合成装置は、請求項１に記載の音声合成装置において、前記テキストを構成する音素の各状態に対応する特徴量を、ＭＦＣＣ、ＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔとする、ことを特徴とする。
【００１８】
また、請求項３の音声合成装置は、請求項１または２に記載の音声合成装置において、前記パラメータ生成部に備えた前記分散付加手段に代わる新たな分散付加手段が、前記テキストを構成する音素の各状態に対応する特徴量の分散を、前記音響モデルから読み出し、前記パラメータ生成手段により生成された前記パラメータに、前記分散に基づいた正規乱数を付加し、複数の合成用パラメータを生成し、前記音声合成部に代わる新たな音声合成部が、前記新たな分散付加手段により生成された前記複数の合成用パラメータを用いて、前記テキストに対応した声帯信号に対して声道フィルタのフィルタ処理を施し、前記複数の合成用パラメータのそれぞれに対応する複数の合成音声を生成し、前記複数の合成音声を平均化する、ことを特徴とする。
【００１９】
さらに、請求項４のプログラムは、コンピュータを、請求項１から３までのいずれか一項に記載の音声合成装置として機能させることを特徴とする。

【発明の効果】
【００２０】
以上のように、本発明によれば、自然の音に近い合成音声を生成することが可能となる。
【図面の簡単な説明】
【００２１】
【図１】実施例１の音声合成装置の構成例を示すブロック図である。
【図２】パラメータ生成部において、合成用パラメータの系列を生成するための構成例を示すブロック図である。
【図３】音響モデルを用いて生成された合成用パラメータの系列ｏ”の例を示す図である。
【図４】音響モデルのパラメータの系列ｏ、合成用パラメータの系列ｏ’及び合成用パラメータの系列ｏ”の例を示す図である。
【図５】実施例２の音声合成装置の構成例を示すブロック図である。
【図６】従来技術において、音響モデルを用いて生成された合成用パラメータの系列を示す図である。
【発明を実施するための形態】
【００２２】
以下、本発明を実施するための形態について図面を用いて詳細に説明する。本発明は、合成音声の合成用パラメータを生成する際に、予め学習しておいた音響モデルの統計量である分散を用いて、揺らぎを付加した合成用パラメータを生成することを特徴とする。
【００２３】
つまり、本発明では、音響モデルの統計量である分散に基づいた正規乱数（ガウスノイズ）を、人間の発声の揺らぎとみなし、この揺らぎを合成用パラメータへ反映する。これにより、合成用パラメータを用いて生成される合成音声は、人間の発声の音に近くなる。つまり、自然の音に近い合成音声を生成することが可能となる。
【００２４】
（音声合成装置／実施例１）
まず、実施例１の音声合成装置について説明する。図１は、実施例１の音声合成装置の構成例を示すブロック図である。この音声合成装置１は、ＨＤ１０、テキスト解析部１１、パラメータ生成部１２及び音声合成部１３を備えている。
【００２５】
ＨＤ１０には、予め学習しておいた音響モデル（統計的音響モデル、ＨＭＭ（隠れマルコフモデル））が格納されている。この音響モデルは、音素を構成する複数の状態（音素の始まり、中間及び終わりの各状態）に対するガウス分布のパラメータ、及び、各状態間の遷移確率、並びに、各状態における出力確率により構成される。ここで、ガウス分布のパラメータとは、例えば、人間の音声知覚の特徴を考慮した声道特性を表す特徴量であるメル周波数ケプストラム係数（ＭＦＣＣ）、このＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔからなるスペクトルパラメータの平均及び分散から構成される。
【００２６】
図示しない音響モデル学習部は、音響モデルを以下の手順により生成する。すなわち、音響モデル学習部は、学習対象の音声信号に対し、当該音声信号を構成する各フレームのＭＦＣＣを算出し、音素区間等の複数のＭＦＣＣをベクトル化したスペクトルパラメータを算出する。そして、音響モデル学習部は、複数のスペクトルパラメータを近似するガウス分布のパラメータ、及び、各状態間の遷移確率、並びに、各状態における出力確率を算出する。これにより、音素を構成する複数の状態のそれぞれについてのパラメータからなる音響モデルを生成する。
【００２７】
テキスト解析部１１は、合成音声の生成対象であるテキストを入力し、テキストに対して形態素解析等の処理を行う。これにより、テキストの音素列、音素の開始時間及び終了

時間、アクセントの有無、有声音及び無声音の区別情報等を生成する。そして、テキスト解析部１１は、音素毎に、その開始時間及び終了時間から継続時間長を算出する。テキスト解析部１１は、テキストの音素列、音素毎の継続時間長、アクセントの有無、有声音及び無声音の区別情報等をパラメータ生成部１２に出力する。

【００２８】

パラメータ生成部１２は、テキスト解析部１１からテキストの音素列、音素毎の継続時間長、アクセントの有無、有声音及び無声音の区別情報等を入力する。そして、パラメータ生成部１２は、まず、音素列の音素の各状態に対応するパラメータの系列ｏを、ＨＤ１０の音響モデルから生成する。つまり、パラメータ生成部１２は、音素列の音素の各状態に対応するそれぞれのパラメータをＨＤ１０の音響モデルから読み出し、パラメータの系列ｏを生成する。次に、パラメータ生成部１２は、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される確率分布Ｐの尤度が最大となり、かつＭＦＣＣ等の分散が反映された合成用パラメータの系列ｏ”を生成する。パラメータ生成部１２は、合成用パラメータの系列ｏ”を音声合成部１３に出力する。

【００２９】

また、パラメータ生成部１２は、音素列の各音素について、その継続時間長及びアクセントの有無等の情報に基づいて、ピッチの基本周波数を生成する。そして、パラメータ生成部１２は、音素列の各音素についてのピッチの基本周波数、有声音及び無声音の区別情報等を音声合成部１３に出力する。

【００３０】

図２は、パラメータ生成部１２において、合成用パラメータの系列を生成するための構成例を示すブロック図である。このパラメータ生成部１２は、パラメータ生成手段３０及び分散付加手段３１を備えている。

【００３１】

パラメータ生成手段３０は、テキスト解析部１１から、テキストの音素列及び音素の継続時間長を入力し、ＨＤ１０のＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔからなるスペクトルパラメータから生成した音響モデルから、音素の各状態に対応するパラメータの系列ｏを生成する。そして、パラメータ生成手段３０は、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布Ｐの尤度が最大となる合成用パラメータの系列ｏ’を生成する。そして、パラメータ生成手段３０は、合成用パラメータの系列ｏ’を分散付加手段３１に出力する。

【００３２】

具体的には、パラメータ生成手段３０は、音響モデルをλ、音素の状態の系列をｑ、音響モデルλから観測されるパラメータの系列をｏ、音素の状態の系列ｑと音響モデルλにおいて、パラメータの系列ｏが観測される音響モデルの確率分布をＰとして、音素の各状態について、前記数式（１）により、音素の状態の系列ｑと音響モデルλにおいて、パラメータの系列ｏが観測される音響モデルλの確率分布Ｐの尤度が最大となるパラメータの系列ｏを合成用パラメータの系列ｏ’として算出する。

【００３３】

これにより、合成用パラメータの系列ｏ’は、ＭＦＣＣの一次回帰係数Δ及び二次回帰係数ΔΔが考慮されて生成されるため、発話文章全体で平均をトレースするような滑らか値となる。

【００３４】

分散付加手段３１は、パラメータ生成手段３０から合成用パラメータの系列ｏ’を入力すると共に、ＨＤ１０の音響モデルから、音素列の音素の各状態について合成用パラメータの系列ｏ’に対応するＭＦＣＣ等の分散を読み出す。

【００３５】

分散付加手段３１は、音素の各状態の合成用パラメータの系列ｏ’に対し、合成用パラメータの系列ｏ’に対応する分散に基づいた正規乱数（ガウスノイズ）を付加し、合成用

パラメータの系列ｏ”を生成する。そして、分散付加手段３１は、合成用パラメータの系列ｏ”を音声合成部１３に出力する。

【００３６】

この分散に基づいた正規乱数（ガウスノイズ）は、人間の発声の揺らぎを表現するものであり、分散付加手段３１により生成される合成用パラメータの系列ｏ”は、人間の発声の揺らぎの成分が付加されたパラメータの系列となる。

【００３７】

具体的には、分散付加手段３１は、ＭＦＣＣ等の分散をσ_{λ}、分散σ_{λ}に基づいたガウスノイズをＮ（σ_{λ}）として、合成用パラメータの系列ｏ’に、分散σ_{λ}に基づいたガウスノイズＮ（σ_{λ}）を加算することで、合成用パラメータの系列ｏ”を求める。つまり、分散付加手段３１は、合成用パラメータの系列ｏ”を、以下の数式（２）にて算出する。

［数２］

$$o'' = o' + N(\sigma_{\lambda}) \quad \cdots (2)$$

【００３８】

図３は、音響モデルを用いて生成された合成用パラメータの系列ｏ”の例を示す図である。横軸は時間ｔを示し、縦軸はＭＦＣＣ　ｃを示す。図３には、音響モデルのパラメータの系列ｏ（ＭＦＣＣ　ｃの平均）、合成用パラメータの系列ｏ’及び合成用パラメータの系列ｏ”の特性を示す。四角の点線は分散を示す。

【００３９】

図４は、音響モデルのパラメータの系列ｏ（ＭＦＣＣ　ｃの平均）、合成用パラメータの系列ｏ’及び合成用パラメータの系列ｏ”の例を示す図であり、図３に示す分散の範囲の領域α，βを拡大した図である。図４（１）は音響モデルのパラメータの系列ｏ（ＭＦＣＣ　ｃの平均）の例であり、図４（２）は合成用パラメータの系列ｏ’の例であり、図４（３）は合成用パラメータの系列ｏ”の例である。

【００４０】

図４（１）は、音響モデルのパラメータの系列ｏ（ＭＦＣＣ　ｃの平均）が、音素の状態毎に一定値であることを示す。図４（２）は、合成用パラメータの系列ｏ’が、図６に示した合成用パラメータの系列ｏ’の特性と同じであることを示す。この合成用パラメータの系列ｏ’は、ＭＦＣＣの一次回帰係数Δ及び二次回帰係数ΔΔが考慮されて生成されるため、音響モデルのパラメータの系列ｏであるＭＦＣＣ　ｃの平均を基準にして、滑らかに変化する値となる。この合成用パラメータの系列ｏ’は、図２に示したパラメータ生成手段３０により生成される。

【００４１】

図４（３）は、合成用パラメータの系列ｏ”が、図４（２）に示した合成用パラメータの系列ｏ’に対し、人間の発声の揺らぎ成分を表現する、ＭＦＣＣ等の分散に基づいた正規乱数（ガウスノイズ）が付加された特性となることを示す。この合成用パラメータの系列ｏ”は、図２に示した分散付加手段３１により生成される。

【００４２】

このように、パラメータ生成部１２は、音素の各状態について、ＭＦＣＣ等の分散を反映した合成用パラメータの系列ｏ”、すなわち、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布Ｐの尤度を最大とし、かつＭＦＣＣ等の分散を反映した合成用パラメータの系列ｏ”を生成する。この合成用パラメータの系列ｏ”は、ＭＦＣＣ等の分散を考慮した揺らぎが与えられ、後述する声道フィルタ２１は、これを声道パラメータとして用いる。

【００４３】

図１に戻って、音声合成部１３は、パラメータ生成部１２から、音素の各状態におけるピッチの基本周波数、有声音及び無声音の区別情報等、並びに合成用パラメータの系列ｏ”を入力する。そして、音声合成部１３は、基本周波数、有声音及び無声音の区別情報等に基づいて、テキストに対応する声帯信号を生成し、声帯信号に対し、合成用パラメータの系列ｏ”を用いて声道フィルタによるフィルタ処理を施すことで、合成音声を生成する

。音声合成部１３は、生成した合成音声を出力する。
【００４４】
図１に示すように、音声合成部１３は、声帯信号生成部２０及び声道フィルタ２１を備えている。声帯信号生成部２０は、パラメータ生成部１２から、音素列の各音素についてのピッチの基本周波数、有声音及び無声音の区別情報等を入力する。そして、声帯信号生成部２０は、対象区間が有声音である場合、基本周波数の逆数であるピッチ周期のパルス列を声帯信号として生成し、対象区間が無声音である場合、白色雑音を声帯信号として生成する。声帯信号生成部２０は、生成した声帯信号を声道フィルタ２１に出力する。
【００４５】
声道フィルタ２１は、声帯信号生成部２０から声帯信号を入力すると共に、パラメータ生成部１２から合成用パラメータの系列ｏ”を入力し、声帯信号に対し、合成用パラメータの系列ｏ”を用いて声道フィルタによるフィルタ処理を施すことで、合成音声を生成する。声道フィルタ２１は、生成した合成音声を出力する。尚、声道フィルタ２１による合成音声の生成手法は既知であり、非特許文献１等を参照されたい。
【００４６】
以上のように、実施例１の音声合成装置１によれば、パラメータ生成部１２は、テキストの音素列について、音素の各状態に対応するパラメータの系列ｏを音響モデルから生成する。そして、パラメータ生成部１２は、前記数式（１）にて、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布の尤度が最大となる合成用パラメータの系列ｏ’を生成する。さらに、パラメータ生成部１２は、前記数式（２）にて、合成用パラメータの系列ｏ’に対し、ＭＦＣＣ等の分散に基づいた正規乱数（ガウスノイズ）を付加し、合成用パラメータの系列ｏ”を生成する。
【００４７】
音声合成部１３は、ピッチの基本周波数等に基づいて、テキストに対応する声帯信号を生成し、声帯信号に対し、合成用パラメータの系列ｏ”を用いて声道フィルタによるフィルタ処理を施すことで、合成音声を生成する。
【００４８】
このように、音響モデルの統計量である分散を用いて、人間の発声の揺らぎを表現する成分を合成用パラメータの系列ｏ”に反映するようにした。これにより、音響モデルの学習時に用いる音声信号の分布に適合した合成用パラメータの系列ｏ”を生成することができる。つまり、人間の音声のパラメータに近い合成用パラメータの系列ｏ”である声道パラメータを生成することができるため、この声道パラメータを用いて生成される合成音声は、実際に人間が発声する音に近くなる。すなわち、自然の音に近い合成音声を生成することができ、人間の発話に近い統計的合成音声の提供が可能となる。
【００４９】
（音声合成装置／実施例２）
次に、実施例２の音声合成装置について説明する。図５は、実施例２の音声合成装置の構成例を示すブロック図である。この音声合成装置２は、ＨＤ１０、テキスト解析部１１、パラメータ生成部１４及び音声合成部１５を備えている。
【００５０】
図１に示した実施例１の音声合成装置１と、図５に示す実施例２の音声合成装置２とを比較すると、両音声合成装置１，２は、同じテキスト解析部１１を備えている点で共通する。一方、実施例２の音声合成装置２は、実施例１の音声合成装置１に備えたパラメータ生成部１２及び音声合成部１３とは異なるパラメータ生成部１４及び音声合成部１５を備えている点で、実施例１の音声合成装置１と相違する。
【００５１】
テキスト解析部１１は、図１に示したテキスト解析部１１と同一であるから、ここでは説明を省略する。
【００５２】

パラメータ生成部１４は、テキスト解析部１１からテキストの音素列、音素毎の継続時間長、アクセントの有無、有声音及び無声音の区別情報等を入力する。そして、パラメータ生成部１２は、音素列の音素の各状態に対応するパラメータの系列ｏを、ＨＤ１０の音響モデルから生成する。

【００５３】

パラメータ生成部１４は、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布の尤度が最大となり、かつＭＦＣＣ等の分散が反映された複数の合成用パラメータの系列ｏ”を生成し、複数の合成用パラメータの系列ｏ”を音声合成部１５に出力する。具体的には、パラメータ生成部１４は、前記数式（１）にて、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布の尤度が最大となる合成用パラメータの系列ｏ’を生成する。さらに、パラメータ生成部１４は、前記数式（２）にて、合成用パラメータの系列ｏ’に対し、ＭＦＣＣ等の分散に基づいた正規乱数（ガウスノイズ）を付加し、複数の合成用パラメータの系列ｏ”を生成する。

【００５４】

また、パラメータ生成部１４は、図１に示したパラメータ生成部１２と同様に、ピッチの基本周波数を生成し、音素列の各音素についてのピッチの基本周波数等を音声合成部１５に出力する。

【００５５】

例えば、パラメータ生成部１４は、合成用パラメータの系列ｏ’に対し、ＭＦＣＣ等の分散に基づいた正規乱数（ガウスノイズ）を付加することで、異なる合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”を生成する。ここで、合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”は、同一の合成用パラメータの系列ｏ’に、ＭＦＣＣ等の分散に基づく異なる正規乱数（ガウスノイズ）を付加して生成する。尚、ＭＦＣＣ等の分散を用いるから、合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”は、生成処理毎に異なる値となる。

【００５６】

音声合成部１５は、パラメータ生成部１４から、音素列の各音素についてのピッチの基本周波数等、及び複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”をそれぞれ入力する。そして、音声合成部１５は、ピッチの基本周波数等に基づいて、テキストに対応する声帯信号を生成し、声帯信号に対し、複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”のそれぞれを用いて声道フィルタによるフィルタ処理を施すことで、複数の合成音声ｗ１，ｗ２，ｗ３を生成する。尚、複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３” のそれぞれを用いて声道フィルタによるフィルタ処理を施すことで生成された複数の合成音声ｗ１，ｗ２，ｗ３は、時間が揃っている。音声合成部１５は、生成した複数の合成音声ｗ１，ｗ２，ｗ３の時間波形を平均化し、平均化した時間波形を合成音声として出力する。

【００５７】

図５に示すように、音声合成部１５は、声帯信号生成部２０、声道フィルタ２２及び平均化部２３を備えている。声帯信号生成部２０は、図１に示した声帯信号生成部２０と同一であるため、ここでは説明を省略する。

【００５８】

声道フィルタ２２は、声帯信号生成部２０から声帯信号を入力すると共に、パラメータ生成部１４から複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”をそれぞれ入力する。そして、声道フィルタ２２は、声帯信号に対し、複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”のそれぞれを用いて声道フィルタによるフィルタ処理を施すことで、複数の合成音声ｗ１，ｗ２，ｗ３を生成する。声道フィルタ２２は、生成した複数の合成音声ｗ１，ｗ２，ｗ３を平均化部２３に出力する。尚、声道フィルタ２２による合成音声の生成手法は既知であるため、ここでは詳細な説明を省略する。

【００５９】

声道フィルタ２２は、第１の合成音声ｗ１、第２の合成音声ｗ２及び第３の合成音声ｗ

３におけるそれぞれの時間波形を平均化部２３に出力する。ここで、第１の合成音声ｗ１、第２の合成音声ｗ２及び第３の合成音声ｗ３は、同じテキストに対応した音声である。一方、合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”が異なるため、異なる時間波形となる。

【００６０】

平均化部２３は、声道フィルタ２２から複数の合成音声ｗ１，ｗ２，ｗ３の時間波形を入力し、複数の合成音声ｗ１，ｗ２，ｗ３の時間波形を平均化し、平均化した時間波形を合成音声として出力する。

【００６１】

以上のように、実施例２の音声合成装置２によれば、パラメータ生成部１４は、テキストの音素列について、音素の各状態に対応するパラメータの系列ｏを音響モデルから生成する。そして、パラメータ生成部１４は、前記数式（１）及び前記数式（２）にて、音素の各状態について、音素の状態の系列と音響モデルにおいて、パラメータの系列ｏが観測される音響モデルの確率分布の尤度が最大となり、かつＭＦＣＣ等の分散が反映された複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”を生成する。

【００６２】

音声合成部１５は、ピッチ基本周波数等に基づいて、テキストに対応する声帯信号を生成し、声帯信号に対し、複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”を用いて声道フィルタによるフィルタ処理を施すことで、複数の合成音声ｗ１，ｗ２，ｗ３を生成する。そして、音声合成部１５は、複数の合成音声ｗ１，ｗ２，ｗ３の時間波形を平均化する。

【００６３】

このように、音響モデルの統計量である分散を用いて、人間の発声の揺らぎを表現する成分を複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”に反映するようにした。これにより、これらの複数の合成用パラメータの系列ｏ１”，ｏ２”，ｏ３”を用いて生成される合成音声は、人間の発声の音に一層近くなる。すなわち、自然の音に一層近い合成音声を生成することができる。

【００６４】

尚、本発明の実施例１，２による音声合成装置１，２のハードウェア構成としては、通常のコンピュータを使用することができる。音声合成装置１，２は、ＣＰＵ、ＲＡＭ等の揮発性の記憶媒体、ＲＯＭ等の不揮発性の記憶媒体、及びインターフェース等を備えたコンピュータによって構成できる。音声合成装置１に備えたＨＤ１０、テキスト解析部１１、パラメータ生成部１２及び音声合成部１３の各機能は、これらの機能を記述したプログラムをＣＰＵに実行させることによりそれぞれ実現できる。また、音声合成装置２に備えたＨＤ１０、テキスト解析部１１、パラメータ生成部１４及び音声合成部１５の各機能は、これらの機能を記述したプログラムをＣＰＵに実行させることによりそれぞれ実現できる。

【００６５】

これらのプログラムは、磁気ディスク（フロッピー（登録商標）ディスク、ハードディスク等）、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤ等）、半導体メモリ等の記憶媒体に格納して頒布することもでき、ネットワークを介して送受信することもできる。

【００６６】

以上、実施例１，２を挙げて本発明を説明したが、本発明は前記実施例１，２に限定されるものではなく、その技術思想を逸脱しない範囲で種々変形可能である。例えば、前記実施例１において、音声合成装置１のパラメータ生成部１２は、音響モデルとして学習されたＭＦＣＣ等の統計量を用いて、ＭＦＣＣ等の分散が反映された合成用パラメータの系列ｏ”を生成するようにした。

【００６７】

この場合、パラメータ生成部１２は、ＭＦＣＣの統計量を用いて、ＭＦＣＣの分散が反映された合成用パラメータの系列ｏ”を生成するようにしてもよい。具体的には、パラメ

ータ生成部１２のパラメータ生成手段３０は、前記数式（１）にて、音素の状態の系列と、ＭＦＣＣを用いて学習した音響モデルにおいて、パラメータの系列ｏが観測される確率分布の尤度が最大となる合成用パラメータの系列ｏ’を生成する。そして、分散付加手段３１は、前記数式（２）にて、合成用パラメータの系列ｏ’に対し、ＭＦＣＣを用いて学習した音響モデルにおけるＭＦＣＣの分散に基づいた正規乱数（ガウスノイズ）を付加し、合成用パラメータの系列ｏ”を生成する。

【００６８】

また、パラメータ生成部１２は、音響モデルとして学習されたＭＦＣＣ及びＭＦＣＣの一次回帰係数Δの統計量を用いて、ＭＦＣＣの分散及びＭＦＣＣの一次回帰係数Δの分散が反映された合成用パラメータの系列ｏ”を生成するようにしてもよい。具体的には、パラメータ生成部１２のパラメータ生成手段３０は、前記数式（１）にて、音素の状態の系列と、ＭＦＣＣ及びＭＦＣＣの一次回帰係数Δを用いて学習した音響モデルにおいて、パラメータの系列ｏが観測される確率分布の尤度が最大となる合成用パラメータの系列ｏ’を生成する。そして、分散付加手段３１は、前記数式（２）にて、合成用パラメータの系列ｏ’に対し、ＭＦＣＣ及びＭＦＣＣの一次回帰係数Δを用いて学習した音響モデルにおけるＭＦＣＣの分散及びＭＦＣＣの一次回帰係数Δの分散に基づいた正規乱数（ガウスノイズ）を付加し、合成用パラメータの系列ｏ”を生成する。

【００６９】

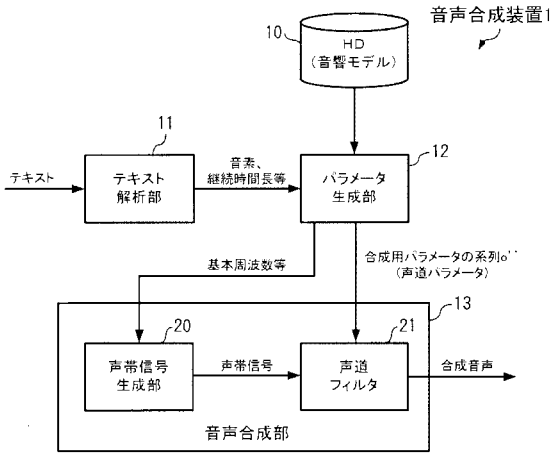
また、パラメータ生成部１２は、音響モデルとして学習されたＭＦＣＣ、ＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔの統計量を用いて、ＭＦＣＣの分散、ＭＦＣＣの一次回帰係数Δの分散及びＭＦＣＣの二次回帰係数ΔΔの分散が反映された合成用パラメータの系列ｏ”を生成するようにしてもよい。具体的には、パラメータ生成部１２のパラメータ生成手段３０は、前記数式（１）にて、音素の状態の系列と、ＭＦＣＣ、ＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔを用いて学習した音響モデルにおいて、パラメータの系列ｏが観測される確率分布の尤度が最大となる合成用パラメータの系列ｏ’を生成する。そして、分散付加手段３１は、前記数式（２）にて、合成用パラメータの系列ｏ’に対し、ＭＦＣＣ、ＭＦＣＣの一次回帰係数Δ及びＭＦＣＣの二次回帰係数ΔΔを用いて学習した音響モデルにおけるＭＦＣＣの分散、ＭＦＣＣの一次回帰係数Δの分散及びＭＦＣＣの二次回帰係数ΔΔの分散に基づいた正規乱数（ガウスノイズ）を付加し、合成用パラメータの系列ｏ”を生成する。実施例２についても同様である。

【符号の説明】

【００７０】

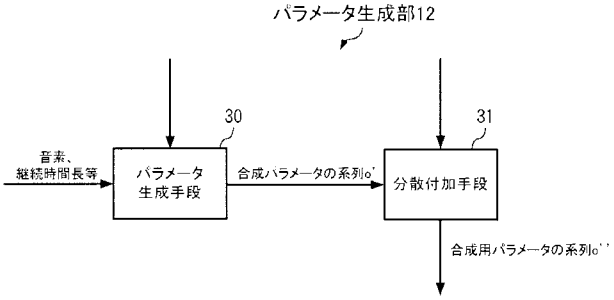
１，２　音声合成装置
１０　ＨＤ（ハードディスク）
１１　テキスト解析部
１２，１４　パラメータ生成部
１３，１５　音声合成部
２０　声帯信号生成部
２１，２２　声道フィルタ
２３　平均化部
３０　パラメータ生成手段
３１　分散付加手段

【図１】

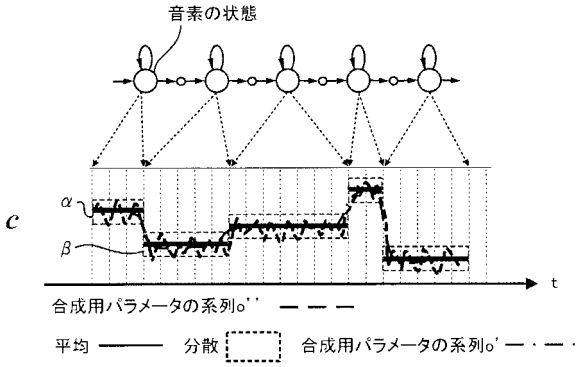


実施例1

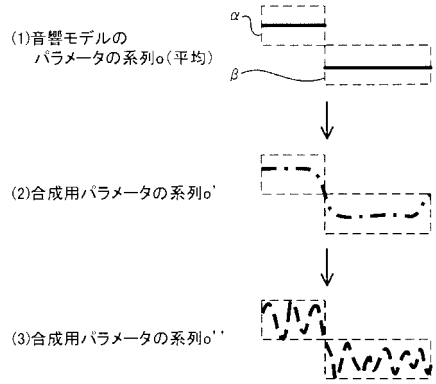
【図２】



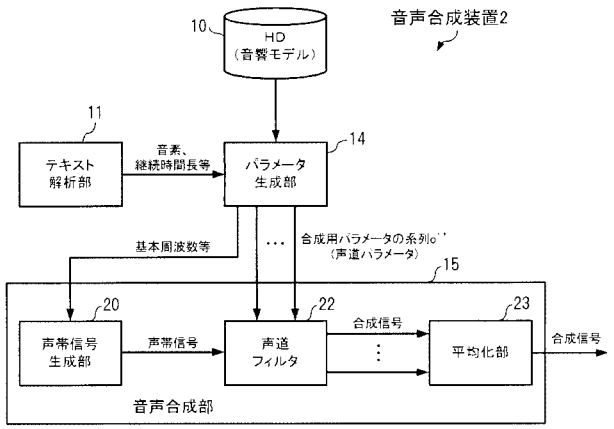
【図３】



【図４】

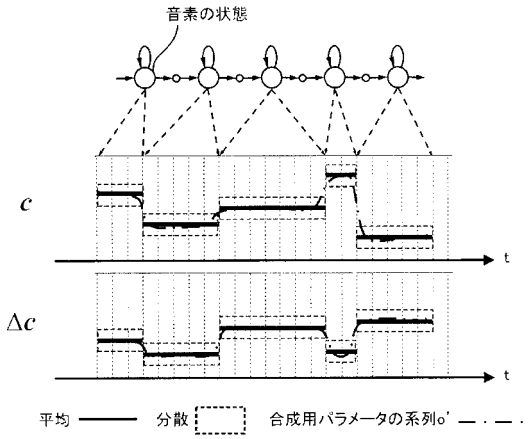


【図５】



実施例2

【図６】



従来技術

フロントページの続き

(72)発明者　清山　信正
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内
(72)発明者　今井　篤
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内
(72)発明者　都木　徹
東京都世田谷区砧一丁目１０番１１号　一般財団法人ＮＨＫエンジニアリングシステム内