

(19) 日本国特許庁 (JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2011-237795

(P2011-237795A)

(43) 公開日 平成23年11月24日 (2011.11.24)

(51) Int.Cl.	F I	テーマコード (参考)
G 1 0 L 13/08 (2006.01)	G 1 0 L 13/08 1 5 0 Z	
G 1 0 L 13/06 (2006.01)	G 1 0 L 13/06 1 3 0	
G 1 0 L 13/04 (2006.01)	G 1 0 L 13/04 Z	
	G 1 0 L 13/08 1 5 0 A	

審査請求 有 請求項の数 21 O L 外国語出願 (全 33 頁)

(21) 出願番号	特願2011-100487 (P2011-100487)	(71) 出願人	000003078
(22) 出願日	平成23年4月28日 (2011.4.28)		株式会社東芝
(31) 優先権主張番号	1007705.5		東京都港区芝浦一丁目1番1号
(32) 優先日	平成22年5月7日 (2010.5.7)	(74) 代理人	100108855
(33) 優先権主張国	英国 (GB)		弁理士 蔵田 昌俊
		(74) 代理人	100159651
			弁理士 高倉 成男
		(74) 代理人	100091351
			弁理士 河野 哲
		(74) 代理人	100088683
			弁理士 中村 誠
		(74) 代理人	100109830
			弁理士 福原 淑弘
		(74) 代理人	100075672
			弁理士 峰 隆司

最終頁に続く

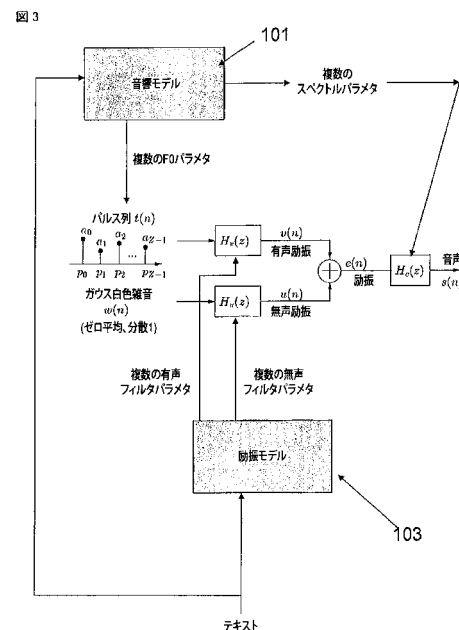
(54) 【発明の名称】 音声処理方法及び装置

(57) 【要約】 (修正有)

【課題】 音声波形の尤度を最大化するように、複数の音響モデルパラメータが複数の励振モデルパラメータと一緒に推定される音声合成方法を提供する。

【解決手段】 テキスト入力を受けており、前記テキスト入力に対応する音声を確率論的モデルを使用して出力し、前記確率論的モデルは音響モデル及び励振モデルを具備し、前記音響モデルは単語または単語の部分の特徴に関連づける複数の確率分布を記述する複数のモデルパラメータを有し、前記励振モデルは、前記特徴を使用して前記音声を出力するために、声帯及び肺をモデル化するために使用される励振モデルパラメータを具備し、ここで音響パラメータ及び励振パラメータは一緒に推定され、前記音声を出力することを具備する。

【選択図】 図3



【特許請求の範囲】**【請求項 1】**

テキスト入力を受け取り、確率論的モデルを使用して前記テキスト入力に対応する音声を出し、前記確率論的モデルは音響モデルと励振モデルとを具備し、前記音響モデルは単語または単語の部分のある特徴に関連づける複数の確率分布を記述する複数のモデルパラメータを有し、前記励振モデルは声帯及び肺をモデル化して特徴を使用して音声を出し、するために使用される励振モデルパラメータを具備し、

前記音響パラメータ及び前記励振パラメータは一緒に推定され、

前記音声を出力することを具備する音声処理方法。

【請求項 2】

前記テキスト入力は、前記音響モデルによって処理され、F 0 及びスペクトルの特徴を出し、

パルス列を形成するために前記 F 0 特徴を処理し、

励振信号を生成するために前記励振モデルから導かれる複数の励振パラメータを使用して前記パルス列をフィルタリングし、

前記スペクトル特徴から導かれる複数のフィルタパラメータを使用して前記励振信号をフィルタリングすることをさらに具備する請求項 1 に記載の音声合成方法。

【請求項 3】

音声合成に関する統計モデルをトレーニングする方法であって、

トレーニングデータを受け取り、前記トレーニングデータは音声と前記音声に対応するテキストとを具備し、

確率論的モデルをトレーニングし、前記確率論的モデルは音響モデルと励振モデルとを具備し、前記音響モデルは単語またはその単語の部分の特徴ベクトルに関連づける複数の確率分布を記述する複数のモデルパラメータを有して、前記励振モデルは声帯及び肺をモデル化して前記音声を出力する複数の励振モデルパラメータを具備し、

前記音響パラメータ及び前記励振パラメータはトレーニング処理の間に一緒に推定される方法。

【請求項 4】

前記音響モデルパラメータは、前記確率分布の平均と分散とを具備する請求項 3 に記載の方法。

【請求項 5】

前記音響モデルによって出力される前記特徴は F 0 特徴とスペクトル特徴を具備する請求項 3 に記載の方法。

【請求項 6】

前記励振モデルパラメータは、F 0 特徴から導かれるパルス信号をフィルタリングするフィルタ係数を具備する請求項 5 に記載の方法。

【請求項 7】

共同に推定する処理は、再帰的な処理を具備し、

1 つのステップでは励振パラメータが音響パラメータの最新の推定を使用して更新され、別のステップでは音響モデルパラメータが励振パラメータの前記最新の推定を使用して更新される請求項 3 に記載の方法。

【請求項 8】

共同に推定する処理は、最大尤度技術を使用する請求項 3 に記載の方法。

【請求項 9】

前記確率論的モデルは、マッピングモデルをさらに具備し、前記マッピングモデルは複数のマッピングモデルパラメータを具備し、前記マッピングモデルは複数のスペクトル特徴を、人間の声道を表現する複数のフィルタ係数にマッピングする請求項 5 に記載の方法。

【請求項 10】

前記パラメータは以下のように一緒に推定され、

10

20

30

40

50

【数 1】

$$\hat{\lambda} = \arg \max_{\lambda} p(s | \ell, \lambda),$$

ここで、 λ は最適化される前記励振モデルと音響モデルとのパラメータを表し、 s は自然な音声波形であり、

【数 2】

ℓ

は前記音声波形の表記である請求項 3 に記載の方法。

【請求項 1 1】

10

λ は、人間の声道を表現するために、複数のスペクトルパラメータをフィルタ関数にマッピングするマッピングモデルの複数のパラメータをさらに具備する請求項 1 0 に記載の方法。

【請求項 1 2】

前記スペクトル特徴とフィルタ係数との間の関係は、ガウスプロセスとしてモデル化される請求項 1 1 に記載の方法。

【請求項 1 3】

$p(s | \ell, \lambda)$ は以下のように表され、

【数 3】

$$p(s | \ell, \lambda) = \sum_{\forall q} \int \int p(s | H_c, q, \lambda_c) p(H_c | c, q, \lambda_h) p(c | q, \lambda_c) p(q | \ell, \lambda_c) dH_c dc. \quad (62)$$

20

ここで、 H_c は前記人間の声道をモデル化するために使用されるフィルタ関数であり、 q は状態であり、 λ_c は励振パラメータであり、 λ_h は音響モデルパラメータであり、 λ_h はマッピングモデルパラメータであり、 c はスペクトル特徴である請求項 1 1 に記載の方法。

【請求項 1 4】

q に渡る和は固定された状態シーケンスによって近似されて次を与え、

【数 4】

$$p(s | \ell, \lambda) \approx \int \int p(s | H_c, \hat{q}, \lambda_c) p(H_c | c, \hat{q}, \lambda_h) p(c | \hat{q}, \lambda_c) p(\hat{q} | \ell, \lambda_c) d\hat{q}, \quad (63)$$

30

ここで

【数 5】

$$\hat{q} = \{\hat{q}_0, \dots, \hat{q}_{T-1}\}$$

は状態シーケンスである請求項 1 3 に記載の方法。

【請求項 1 5】

前記全ての可能な H_c 及び c に渡る積分は、スペクトル応答及びインパルス応答のベクトルによって近似され、次を与え、

【数 6】

40

$$p(s | \ell, \lambda) \approx p(s | \hat{H}_c, \hat{q}, \lambda_c) p(\hat{H}_c | \hat{c}, \hat{q}, \lambda_h) p(\hat{c} | \hat{q}, \lambda_c) p(\hat{q} | \ell, \lambda_c), \quad (64)$$

ここで、

【数 7】

$$\hat{c} = [\hat{c}_1, \dots, \hat{c}_{T-1}]^T$$

は固定されたスペクトル応答である請求項 1 4 に記載の方法。

【請求項 1 6】

$p(s | \ell, \lambda)$ の対数尤度関数は以下

【数 8】

$$\mathcal{L} = \log p(s | \hat{H}_e, \hat{q}, \lambda_e) + \log p(\hat{H}_e | \hat{c}, \hat{q}, \lambda_h) + \log p(\hat{c} | \hat{q}, \lambda_c) + \log p(\hat{q} | \ell, \lambda_e). \quad (65)$$

によって与えられる請求項 15 に記載の方法。

【請求項 17】

コンピュータを制御して、請求項 1 から請求項 16 のいずれか 1 項の方法を実行させるコンピュータ読み取り可能な指示を運ぶキャリア媒体。

【請求項 18】

単語のシーケンスを具備するテキスト入力を受け取るレシーバーと、

プロセッサと、を具備し、前記プロセッサは前記入力テキストに対応する出力音声の尤度を確率論的モデルを使用して判定し、前記確率論的モデルは音響モデルと励振モデルを具備し、前記音響モデルは単語または単語の部分と特徴とを関連づける複数の確率分布を記述する複数のモデルパラメータを有して、前記励振モデルは、前記声帯及び肺をモデル化して前記特徴を使用して前記音声を出力するために使用される複数の励振モデルパラメータを具備し、

前記音響パラメータ及び励振パラメータは一緒に推定され、前記装置は前記音声のための出力を具備する音声処理装置。

【請求項 19】

入力音声認識部と、翻訳部と、請求項 18 に記載の音声合成装置と、を具備する音声翻訳システム。

【発明の詳細な説明】

【技術分野】

【0001】

ここに記述された本発明の実施形態は一般に、音声合成の分野に関係がある。

【背景技術】

【0002】

音響モデルは音声合成のバックボーンとして使用される。音響モデルは、複数の単語の順序または複数の単語の部分を、複数の特徴ベクトルの配列に関連づけるために使用される。統計のパラメータの音声合成では、励振モデルは音響モデルと組み合わせて使用される。励振モデルは、より自然な音声を出力するために、肺と声帯の作用をモデル化するために使用される。

【0003】

既知の統計音声合成では、ケプストラム係数のような特徴は、音声波形及びそれらのトラジェクトリから抽出され、隠れマルコフモデル(HMM)のような統計モデルによってモデル化される。統計モデルの複数のパラメータは、トレーニングデータへのその尤度を最大化するか、あるいはトレーニングデータと生成された複数の特徴との間のエラーを最小化するように推定される。合成段階では、センテンスレベルモデルは、入力テキストにしたがって、推定された統計モデルから構成され、次に、複数の特徴は、そのようなセンテンスモデルから生成され、その結果それらの出力確率を最大化するか、または目的関数を最小化する。

【発明の概要】

【0004】

【図面の簡単な説明】

【0005】

本発明は、下記の限定されない実施形態を参照して記述されるだろう。

【図 1】非常に基本的な音声合成システムの概略図である。

【図 2】テキスト音声合成のために形成されたプロセッサのアーキテクチャの概略図である。

【図 3】音声合成システムのブロック図であり、その複数のパラメータは本発明の実施形態に従って推定される。

【図４】特別の単語あるいはその部分を観測に関連づけるガウス分布のプロットである。

【図５】本発明の実施形態に従って音声合成モデルをトレーニングする方法での初期化ステップを示すフローダイアグラムである。

【図６】本発明の実施形態に従って音声合成モデルをトレーニングする方法での反復ステップを示すフローダイアグラムである。

【図７】本発明の実施形態に従う音声合成の方法を示すフローダイアグラムである。

【発明を実施するための形態】

【０００６】

現在の音声合成システムは、しばしばソースフィルタモデルを使用する。このモデルでは、励振信号は生成されフィルタリングされる。スペクトル特徴シーケンスは、音声から抽出され、別々に音響モデルと励振モデルのパラメータを評価するために利用される。したがって、複数のスペクトル特徴は、励振モデルを考慮に入れることによって、最適化されない。また逆の場合も同じである。

10

【０００７】

本発明の発明者は、音響及び励振モデルのパラメータを推定する問題への完全に異なるアプローチをとっており、ある実施形態では、音声波形の尤度を最大化するように、複数の音響モデルパラメータが複数の励振モデルパラメータと一緒に推定される方法を提供する。

【０００８】

実施形態によれば、複数のスペクトル包絡線特徴から導かれるゆっくり変化する声道インパルス応答フィルタと、励振源との畳み込みによって音声が表示されることが推定される。提案されたアプローチでは、複数のスペクトル特徴の抽出は、音響及び励振のモデルの組み合わせられたトレーニングに融合する。最大尤度（ML: maximum likelihood）基準に基づく問題となっている複数のモデルの複数のパラメータの推定は、自然及び合成された音声波形との間での距離の暗黙の最小化で、本格的な波形レベル閉ループトレーニングとして見られ得る。

20

【０００９】

実施形態では、統計的なパラメータ音声合成に関する、音響及び励振のモデルの共同の推定は、最大尤度に基づいている。

【００１０】

30

生じるシステムは、トラジェクトリHMMで解析された要因として解釈されるものになる。共同の音響及び励振のモデルの複数のパラメータの推定に関してなされた複数の近似は、１つの最良のスペクトル係数ベクトルのトレーニング及び導出に沿って固定される状態シーケンスを固定することを具備する。

【００１１】

実施形態では、音響モデルの複数のパラメータは、励振モデルを考慮に入れることにより更新される。また、後者の複数のパラメータは音響モデルから生成されたスペクトルを仮定して計算される。生じるシステムは、トラジェクトリHMMで分析された要因に類似する手法で、スペクトル包絡線パラメータ抽出と、励振信号モデルとを結びつける。提案されたアプローチは、自然及び合成された音声との間での距離を最小化するための波形レベル閉ループトレーニングとして解釈されうる。

40

【００１２】

実施形態では、音響及び励振のモデルは、統計の枠組みで直接、音声波形から一緒に最適化される。

【００１３】

したがって、複数のパラメータは、次のように一緒に推定される。

【数１】

$$\hat{\lambda} = \arg \max_{\lambda} p(s | \ell, \lambda),$$

【００１４】

50

ここで λ は最適化される励振モデル及び音響モデルのパラメータを表わし、 s は自然な音声波形であり、

【数 2】

ℓ

【0 0 1 5】

は音声波形の表記である。

【0 0 1 6】

実施形態では、上記のトレーニング方法は、統計パラメトリック原理にしたがって構築された、テキストから音声への (TTS : text-to-speech) 合成器に適用することができる。従って、それはまた、音声翻訳及び会話のシステムのような、上記 TTS システムが埋め込まれている任意のタスクに適用されうる。

10

【0 0 1 7】

1 つの実施形態では、ソースフィルタモデルが使用され、ここでは、前記テキスト入力が入記音響モデルによって処理され、 F_0 (基本周波数) 及びスペクトルの特徴を出力する。方法は、パルス列を形成するために複数の前記 F_0 特徴を処理し、励振信号を生成するために前記励振モデルから導かれる複数の励振パラメータを使用して前記パルス列をフィルタリングし、複数の前記スペクトル特徴から導かれる複数のフィルタパラメータを使用して前記励振信号をフィルタリングすることをさらに具備する。

【0 0 1 8】

複数の音響モデルパラメータは、複数の前記確率分布の複数の平均及び複数の分散を具備してもよい。前記音響モデルによって出力される複数の特徴の例は、複数の F_0 特徴及び複数のスペクトル特徴である。

20

【0 0 1 9】

複数の励振モデルパラメータは、複数の F_0 特徴及び白色雑音から導かれるパルス信号をフィルタリングする複数のフィルタ係数を具備してもよい。

【0 0 2 0】

実施形態では、前記共同推定処理は、再帰的な処理を具備する。この処理では、あるステップで複数の励振パラメータが複数の音響パラメータの最新の推定を使用して更新され、別のステップで複数の音響モデルパラメータが複数の励振パラメータの最新の推定を使用して更新される。好ましくは、前記共同推定処理は最大尤推定法技術を使用する。

30

【0 0 2 1】

さらなる実施形態では、前記確率論的モデルはマッピングモデルをさらに具備し、前記マッピングモデルは複数のマッピングモデルパラメータを具備し、前記マッピングモデルは、人間の声道を表す複数の係数をフィルタリングするために、複数のスペクトル特徴をマップするように構成される。好ましくは、複数のスペクトル特徴と複数のフィルタ係数との関係は、ガウスプロセスとしてモデル化される。

【0 0 2 2】

本発明の実施形態は、ハードウェアでも汎用コンピュータでのソフトウェアでも実行されうる。さらに、本発明は、ハードウェアとソフトウェアとの組み合わせで実行されうる。本発明はまた、単一処理装置あるいは複数の処理装置の分散型ネットワークによって実行されうる。

40

【0 0 2 3】

本発明はソフトウェアによって実行されうるので、本発明は、任意の適切なキャリア媒体で汎用コンピュータに提供されるコンピュータコードを包含する。キャリア媒体は、フロッピー (登録商標) ディスク、CD ROM、磁気デバイスあるいはプログラム可能メモリ装置のような任意の記憶媒体、あるいは任意の信号 (例えば電氣的、光学的、またはマイクロ波信号) のような任意の一時的な媒体 (transient medium) を具備することができる。

【0 0 2 4】

図 1 は非常に基本的な音声処理システムの概略図であり、図 1 のシステムは音声合成の

50

ために形成された。テキストはユニット 1 を介して受け取られる。ユニット 1 は、インターネットへの接続、プロセッサから出力されたテキストへの接続、音声から音声言語処理モジュール（携帯電話など）への入力、であってもよい。ユニット 1 は、以前に保存されたテキストデータを含んでいるメモリによって置き換えられる。

【0025】

その後、テキスト信号は、図 2 を参照してより詳細に記述される音声プロセッサ 3 に導かれる。

【0026】

音声プロセッサ 3 は、テキスト信号を取り込み、それをテキスト信号に対応する音声に変える。出力の様々な形式が利用可能である。例えば、出力は、スピーカーに出力する直接の音声出力 5 の形式でもよい。これは、自動車電話、衛星ナビゲーションシステムなどで実行されうる。あるいは、出力は、オーディオファイルとして保存されてもよいし、メモリに向けられてもよい。さらに、出力は、さらなるシステム 9 に提供される電子音声信号の形式になりうる。

【0027】

図 2 は、テキストを音声に変換するシステム 5 1 の基礎的なアーキテクチャを示す。テキストを音声に変換するシステム 5 1 は、プログラム 5 5 を実行するプロセッサ 5 3 を具備する。テキストを音声に変換するシステム 5 1 はさらに記憶装置 5 7 を含む。記憶装置 5 7 は、テキストを音声に変換するプログラム 5 5 によって使用されるデータを格納する。テキストを音声に変換するシステム 5 1 は、入力モジュール 6 1 及び出力モジュール 6 3 をさらに具備する。入力モジュール 6 1 はテキスト入力 6 5 に接続される。テキスト入力 6 5 はテキストを受け取る。テキスト入力 6 5 は例えばキーボードでもよい。

【0028】

あるいは、テキスト入力 6 5 は、外部記憶媒体あるいはネットワークからテキストデータを受け取る手段でもよい。

【0029】

オーディオ 6 7 のための出力が出力モジュール 6 3 に接続される。音声出力 6 7 は、テキスト入力からテキスト入力 6 3 に変換された音声信号を出力するために使用される。音声出力 6 7 は、例えば直接の音声出力（例えば、記憶媒体、ネットワーク等に送られる音声データファイルのための出力、またはスピーカー）でもよい。

【0030】

使用中では、テキストを音声に変換するシステム 5 1 は、テキスト入力 6 3 を介してテキストを受け取る。プロセッサ 5 3 で実行されるプログラム 5 5 は、記憶装置 5 7 に格納されるデータを使用して、テキストを音声データに変換する。音声は、音声出力 6 7 に出力モジュール 6 5 を介して出力される。

【0031】

図 3 は音声生成のモデルの概略図である。モデルには 2 つのサブモデルがある。音響モデル 101、及び励振モデル 103 である。

【0032】

単語あるいはその部分が特徴または特徴ベクトルに変換される場合、音響モデルは、音声合成の技術においてよく知られている。この実施形態では、隠れマルコフモデル（HMM）に基づく音響モデルが使用される。しかしながら、他のモデルを使用することができるともかもしれない。

【0033】

この実施形態で使用される実際のモデルは、標準モデルである。モデルの詳細はこの特許出願の範囲外である。しかしながら、モデルは、単語あるいはその部分に関係している特徴ベクトルによって表わされる観測の可能性に関する確率密度関数（pdfs）の提供を要求するだろう。一般に、この確率分布は n 次元空間でのガウス分布になるだろう。

【0034】

10

20

30

40

50

一般的なガウス分布の概要の例は図4に示される。ここで、水平軸は1次元での入力ベクトルのパラメータに対応する。また確率分布は、観測に関係のある特別の単語あるいはその部分に関する。例えば、図4では、特徴ベクトル x に対応する観測は、その確率分布が図4に示される単語に対応する確率 p_1 を有する。ガウス分布の形状及び位置はその平均と分散とによって定義される。これらのパラメータは、語いに関するトレーニングの間に決定される。それらは音響モデルに関する「モデルパラメータ」と呼ばれるだろう。

【0035】

音声へ出力されることになっているテキストは、最初に複数の電話ラベルに変換される。電話ラベルは、その音素に関する文脈情報を備えた音素を含む。文脈情報の例は、先行し後続する複数の音素、音素の単語内での位置、文での単語の位置などである。その後、音素ラベルは音響モデルに入力される。

10

【0036】

音響モデルHMMの出力、一旦複数のモデルパラメータが決定されると、複数の単語の順序または複数の単語の部分に対応する観測の順序の尤度を決定するために、モデルを使用することができる。

【0037】

特にこの実施形態では、音響モデル101の出力である特徴は F_0 特徴及びスペクトルの特徴である。この実施形態では、複数のスペクトル特徴はケプストラム係数である。しかしながら、他の実施形態では、他の複数のスペクトル特徴が、線形予測係数(LPC)、線スペクトルペア(LSP)及びそれらの周波数が歪んだバージョンのように、使用される。

20

【0038】

複数のスペクトル特徴は、 $h_c(n)$ として表現された声道フィルタ係数を形成するために変換される。

【0039】

生成された複数の F_0 特徴はパルス列シーケンス $t(n)$ に変換される。また、 F_0 値によれば、パルス列間の期間は決定される。

【0040】

例えば、パルス列は時間領域での信号のシーケンスである：例えば

0 1 0 0 0 1 0 0 0 0 1 0 0

30

ここで1はパルスである。人間の声帯は、有声発話のための複数の周期信号を生成するために振動する。パルス列シーケンスはこれらの周期信号に近づけるために使用される。

【0041】

白色雑音励起シーケンス $w(n)$ は白色ノイズ発生器(図示せず)から生成される。

【0042】

パルス列 $t(n)$ 及び白色雑音シーケンス $w(n)$ は、励振モデルパラメータ $H_v(z)$ 及び $H_u(z)$ によってそれぞれフィルタリングされる。励振モデルパラメータは励振モデル105から生成される。 $H_v(z)$ は、有声インパルス応答フィルタ係数を表わし、それが声門の動作を表わすので、時々「声門フィルタ」と呼ばれる。 $H_u(z)$ は無声フィルタ応答係数を表わす。 $H_v(z)$ 及び $H_u(z)$ はともに肺と声帯をモデル化する励振パラメータである。

40

【0043】

時間領域信号である有声励振信号 $v(n)$ は、フィルタリングされたパルス列から生成される。また、さらに時間領域信号である無声励振信号 $u(n)$ は、白色雑音 $w(n)$ から生成される。これらの信号 $v(n)$ 及び $u(n)$ は、混合されて(加算され)て、時間領域での複数の混合励振信号 $e(n)$ からなる。

【0044】

最後に、複数の励振信号 $e(n)$ は、上で説明したように導かれる複数のスペクトル特徴から導かれるインパルス応答 $H_c(z)$ によってフィルタリングされ、音声波形 $s(n)$ を得る。

50

【 0 0 4 5 】

音声合成ソフトウェア製品では、製品は、平均と分散のような音響モデルパラメータと共に、 $H_v(z)$ 及び $H_u(z)$ の係数を含んでいるメモリを具備する。製品はさらに、音響モデルから出力された複数のスペクトル特徴が $H_c(z)$ に変換されることを可能にするデータを含むだろう。複数のスペクトル特徴が複数のケプストラム係数である場合、 $H_c(z)$ へのスペクトル特徴の変換は決定論的で、確率論的モデルをトレーニングするために使用されるデータの性質に依存しない。

【 0 0 4 6 】

しかしながら、スペクトル特徴が、線形予測係数 (LPC)、線スペクトルペア (LSP) 及びそれらの周波数が歪んだバージョンのような他の特徴を具備する場合は、スペクトル特徴と $H_c(z)$ との間でのマッピングは、決定論的でなく、音響及び励振パラメータが推定される場合に推定される必要がある。しかしながら、スペクトル特徴と $H_c(z)$ との間でのマッピングが、決定論的であるか、マッピングモデルを使用して推定されるかに関わらず、好ましい実施形態では、ソフトウェア合成製品は、スペクトル特徴を $H_c(z)$ に変換するために必要とされる情報をちょうど具備するだろう。

【 0 0 4 7 】

音声合成システムのトレーニングは、複数のモデルの複数のパラメータを評価することを含んでいる。上記のシステムでは、音響、励振及び写像のモデルパラメータは、推定されることになっている。しかしながら、マッピングモデルパラメータを取り除くことができることは注意されるべきである。また、これは後で記述されるだろう。

【 0 0 4 8 】

本発明の実施形態に従うトレーニング方法では、音響モデルパラメータ及び励振モデルパラメータは、同じプロセスで同時に推定される。

【 0 0 4 9 】

違いを理解するために、これらのパラメータを推定するための従来のフレームワークが最初に記述される。

【 0 0 5 0 】

既知の統計パラメータの音声合成では、最初に、音声特徴の「スーパーベクトル」

【 数 3 】

$$c = [c_0^T \cdots c_{T-1}^T]^T$$

【 0 0 5 1 】

が音声波形から抽出される。ここで、

【 数 4 】

$$c_t = [c_t(0) \cdots c_t(C)]^T$$

【 0 0 5 2 】

フレーム t での C 次の音声パラメータベクトルであり、 T は全フレーム数である。音響モデルパラメータの推定は、ML 基準を介して通常行われる：

【 数 5 】

$$\hat{\lambda}_c = \arg \max_{\lambda_c} p(c | \ell, \lambda_c), \quad (1)$$

【 0 0 5 3 】

ここで、

【 数 6 】

$$\ell$$

【 0 0 5 4 】

は音声波形の表記であり、 λ_c は 1 セットの音響モデルパラメータを示す。

【 0 0 5 5 】

合成中に、音声特徴ベクトル c' は、合成された

【 数 7 】

$$\ell$$

【 0 0 5 6 】

になるある与えられたテキストに関して、その出力確率を最大化するように、合成されるために生成される。

【 数 8 】

$$\hat{c}' = \arg \max_{c'} p(c' | \ell', \hat{\lambda}_c). \quad (2)$$

10

【 0 0 5 7 】

F 0 及び恐らくは期間と共のこれらの特徴は、図 3 を参照して説明されるように、ソースフィルタ生成アプローチを使用することによって、音声波形を生成するために利用される。

【 0 0 5 8 】

本発明の実施形態に従うトレーニング方法は、異なるアプローチを使用する。任意の音声合成装置の意図が、できる限り音声波形を模倣することであるので、本発明の実施形態では波形レベルで定義された統計モデルが提案される。提案されたモデルのパラメータは、波形それ自身の尤度を最大化するように推定される。すなわち、

20

【 数 9 】

$$\hat{\lambda} = \arg \max_{\lambda} p(s | \ell, \lambda), \quad (3)$$

【 0 0 5 9 】

ここで、

【 数 1 0 】

$$s = [s(0) \cdots s(N-1)]^T$$

30

【 0 0 6 0 】

は音声波形全体を含んでいるベクトルである。 $s(n)$ はサンプル n の波形値である。 N はサンプル数である。また、 λ は、共同の音響励振モデルのパラメータのセットを示す。

【 0 0 6 1 】

2 つの隠れた変数：状態シーケンス $q = \{q_0, \dots, q_{T-1}\}$ (離散的)、スペクトルパラメータ

【 数 1 1 】

$$c = [c_0^T \cdots c_{T-1}^T]^T$$

【 0 0 6 2 】

(連続的)を導入することによって、式(3)は次のように書き直すことができる：

【 数 1 2 】

$$\hat{\lambda} = \arg \max_{\lambda} \sum_{\forall q} \int p(s, c, q | \ell, \lambda) dc \quad (4)$$

$$= \arg \max_{\lambda} \sum_{\forall q} \int p(s | c, q, \lambda) p(c | q, \lambda) p(q | \ell, \lambda) dc, \quad (5)$$

40

【 0 0 6 3 】

ここで q_t はフレーム t での状態である。

50

【 0 0 6 4 】

式 (5) の項、 $p (s | c , q , \lambda)$ 、 $p (c | q , \lambda)$ 、 $p (q | l , \lambda)$ を、以下のように別々に分析することができる。

【 0 0 6 5 】

・ $p (s | c , q , \lambda)$ ：この確率は、スペクトル特徴とある与えられた状態シーケンスからの音声波形生成に関する。 λ に関するこの確率の最大化は、スペクトルモデルパラメタの M L 推定と密接な関係がある。この確率は想定された音声信号生成モデルに関連する。

【 0 0 6 6 】

・ $p (c | q , \lambda)$ ：この確率は、HMMあるいは隠れた半マルコフモデル (H S M M) がその音響モデルとして使用される場合、複数の音声パラメータベクトルの複数の状態出力確率の積として与えられる。トラジェクトリ H M M が使用される場合、この確率は全体の音声パラメータベクトルの状態シーケンス出力確率として与えられる。

【 0 0 6 7 】

・ $p (q | l , \lambda)$ ：この確率は表記

【 数 1 3 】

ℓ

【 0 0 6 8 】

に関して状態シーケンス q の確率を与える。H S M M またはトラジェクトリ H S M M が使用される場合、それは状態遷移及び状態持続の両方の確率を含む。

【 0 0 6 9 】

H M M、H S M M あるいはトラジェクトリ H M M のような既存する音響モデルを使用して $p (c | q , \lambda)$ 、 $p (q | l , \lambda)$ をモデル化することが可能である。問題は $p (s | c , q , \lambda)$ をどのようにモデル化するかである。

【 0 0 7 0 】

音声信号が図 4 のダイアグラムによって生成されると仮定されている、つまり、

【 数 1 4 】

$$s(n) = h_c(n) * [h_v(n) * t(n) + h_u(n) * w(n)], \quad (6)$$

【 0 0 7 1 】

ここで * は線形の畳み込みを示し、

・ $h_c (n)$ ：声道フィルタインパルス応答である。

【 0 0 7 2 】

・ $t (n)$ ：パルス列である。

【 0 0 7 3 】

・ $w (n)$ ：平均 0 及び分散 1 を備えたガウスの白色雑音シーケンスである。

【 0 0 7 4 】

・ $h_v (n)$ ：有声フィルタインパルス応答である。

【 0 0 7 5 】

・ $h_u (n)$ ：無声フィルタ・インパルス応答である

ここで、声道、有声及び無声フィルタは、 z - 変換領域で次の形をそれぞれ持つと仮定される：

10

20

30

40

【数 1 5】

$$H_c(z) = \sum_{p=0}^P h_c(p) z^{-p} \quad (7)$$

$$H_v(z) = \sum_{m=-\frac{M}{2}}^{\frac{M}{2}} h_v(m) z^{-m}, \quad (8)$$

$$H_u(z) = \frac{K}{1 - \sum_{l=1}^L g(l) z^{-l}}, \quad (9)$$

10

【0 0 7 6】

ここで P、M 及び L はそれぞれ $H_c(z)$ 、 $H_v(z)$ 及び $H_u(z)$ のオーダーである。フィルタ $H_c(z)$ は、最小位相応答を有していると見なされる。なぜならそれが声道フィルタのインパルス応答を表わすからである。さらに、 $H_u(z)$ の係数が、R. Maia, T. Toda, H. Zen, Y. Nakaku, and K. Tokuda, "An excitation model for HMM-based s

20

【数 1 6】

$$\{p_0, \dots, p_{Z-1}\}, \{a_0, \dots, a_{Z-1}\}$$

【0 0 7 7】

(Z はパルス数) を具備する。 $H_v(z)$ 及び $H_u(z)$ を推定するいくつかの方法があるが、この報告書は、R. Maia, T. Toda, H. Zen, Y. Nakaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modelling," in Proc. of the 6th ISCA Workshop on Speech Synthesis, 2007 に述べられていた方法に基づくだろう。

30

【0 0 7 8】

それぞれ行列及びベクトルを意味する大文字及び小文字の頭文字と共に、行列表記法を使用して、式 (6) は次のように書くことができる。

【数 1 7】

$$s = H_c H_v t + s_u, \quad (10)$$

40

【0 0 7 9】

ここで、

【数 1 8】

$$\mathbf{s} = \left[s\left(-\frac{M}{2}\right) \quad \cdots \quad s\left(N + \frac{M}{2} + P - 1\right) \right]^\top, \quad (11)$$

$$\mathbf{H}_c = \left[\tilde{\mathbf{h}}_c^{(0)} \quad \cdots \quad \tilde{\mathbf{h}}_c^{(N+M-1)} \right], \quad (12)$$

$$\mathbf{h}_c^{(i)} = \left[\underbrace{0 \quad \cdots \quad 0}_i \quad h_c(0) \quad \cdots \quad h_c(P) \quad \underbrace{0 \quad \cdots \quad 0}_{N+M-i-1} \right]^\top, \quad (13)$$

$$\mathbf{H}_v = \left[\tilde{\mathbf{h}}_v^{(0)} \quad \cdots \quad \tilde{\mathbf{h}}_v^{(N-1)} \right], \quad (14)$$

$$\tilde{\mathbf{h}}_v^{(i)} = \left[\underbrace{0 \quad \cdots \quad 0}_i \quad h_v\left(-\frac{M}{2}\right) \quad \cdots \quad h_v\left(\frac{M}{2}\right) \quad \underbrace{0 \quad \cdots \quad 0}_{N-i-1} \right]^\top, \quad (15)$$

10

【0 0 8 0】

ベクトル $\mathbf{s}_u$ は

【数 1 9】

$$s_u(n) = h_c(n) * h_u(n) * w(n), \quad (17)$$

【0 0 8 1】

のサンプルを含んでいる。そして、ベクトル $\mathbf{s}_u$ は、次の共分散行列と共に、音声信号の
 有声領域に関するモデルのエラーとして解釈されうる。

20

【数 2 0】

$$\Phi = \mathbf{H}_c (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{H}_c^\top, \quad (19)$$

【0 0 8 2】

ここで、

【数 2 1】

$$\mathbf{G} = \left[\tilde{\mathbf{g}}^{(0)} \quad \cdots \quad \tilde{\mathbf{g}}^{(N+M-1)} \right], \quad (20)$$

$$\tilde{\mathbf{g}}^{(i)} = \left[\underbrace{0 \quad \cdots \quad 0}_i \quad \frac{1}{K} \quad \frac{g(1)}{K} \quad \cdots \quad \frac{g(L)}{K} \quad \underbrace{0 \quad \cdots \quad 0}_{N+M-i-1} \right]^\top. \quad (21)$$

30

【0 0 8 3】

$w(n)$ がガウスの白色雑音であるので、 $u(n) = h_u(n) * w(n)$ は正規分布した
 確率過程になる。ベクトル表記法の使用によって、確率 u は以下になる。

【数 2 2】

$$p(\mathbf{u} | \mathbf{G}) = \mathcal{N}(\mathbf{u}; \mathbf{0}, (\mathbf{G}^\top \mathbf{G})^{-1}), \quad (22)$$

【0 0 8 4】

ここで $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は、平均ベクトル $\boldsymbol{\mu}$ 及び共分散行列 $\boldsymbol{\Sigma}$ を備えた $\mathbf{x}$ のガウス分布で
 ある。したがって、

40

【数 2 3】

$$u(n) = H_c^{-1}(z) [s(n) - h(n) * h_v(n) * t(n)], \quad (23)$$

【0 0 8 5】

なので、音声ベクトル $\mathbf{s}$ の確率は以下になる。

【数 2 4】

$$p(\mathbf{s} | \mathbf{H}_c, \mathbf{H}_v, \mathbf{G}, t) = \mathcal{N}(\mathbf{s}; \mathbf{H}_c \mathbf{H}_v t, \mathbf{H}_c (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{H}_c^\top). \quad (24)$$

【0 0 8 6】

50

H c の最後の P 行が無視される（それは、サンプル

【数 2 5】

$$\left\{ s\left(N+\frac{M}{2}\right), \dots, s\left(N+\frac{M}{2}+P-1\right) \right\}$$

【0087】

を生成する H c (z) のゼロインパルス応答を無視することを意味する) 場合、H c は、式 (2 4) が次式 (2 5) に書き直される次元 (N + M) x (N + M) を有する正方になる。

【数 2 6】

10

$$p(s | H_c, \lambda_e) = |H_c|^{-1} \mathcal{N}\left(H_c^{-1}s; H_v t, (G^T G)^{-1}\right), \quad (25)$$

【0088】

ここで、

【数 2 7】

$$\lambda_e = \{H_v, G, t\}$$

【0089】

は音声生成モデルの励起モデル部分のパラメータである。H c (z) による逆フィルタリングを介して音声信号 s (n) から抽出される項

20

【数 2 8】

$$H_c^{-1}s$$

【0090】

が残りのシーケンスに対応することに注意することは興味深い。

【0091】

R. Maia, T. Toda, H. Zen, Y. Nakaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modelling," in Proc. of the 6th ISCA

30

Workshop on Speech Synthesis, 2007で提案されたように、H v と H u が状態依存パラメータ結束構造を有すると仮定することによって、式 (2 5) は次式に書き直される。

【数 2 9】

$$p(s | H_c, q, \lambda_e) = |H_c|^{-1} \mathcal{N}\left(H_c^{-1}s; H_{v,q} t, (G_q^T G_q)^{-1}\right), \quad (26)$$

【0092】

ここで、H v , q 及び G q は、状態シーケンス q に関して、それぞれ有声フィルタ及び逆無声フィルタインパルス応答行列である。

【0093】

声道インパルス応答 H c (あるいは H c (z) の複数の係数) とスペクトル特徴 c との間に 1 対 1 の関係が通常ある。しかしながら、いくつかのスペクトル特徴の表現に関する閉じた形式での c から、H c を計算することは難しい。この問題に取り組むために、確率近似が c と H c との関係をモデル化するために導入される。c と H c との間のマッピングが、確率 p (H c | c , q , λ h) を有するガウスプロセス (ここで λ h はスペクトル特徴を声道フィルタインパルス応答にマッピングするモデルのパラメータセットである) によって表現され则认为られるならば、p (s | c , q , λ e) は以下になる。

40

【数 3 0】

$$p(s | c, q, \lambda_c) = \int p(s | H_c, q, \lambda_c) p(H_c | c, q, \lambda_h) dH_c \quad (27)$$

$$= \int \mathcal{N}(H_c^{-1}s; H_{c,q}t, (G_q^T G_q)^{-1}) \mathcal{N}(H_c; f_q(c), \Omega_q), \quad (28)$$

【0094】

ここで、 $f_q(c)$ は c を H_c に変換するための近似された関数であり、 Ω_q は問題となっているガウス分布の共分散行列である。この表現は、 H_c が c から閉じた形式でその特別な場合として計算されうる場合（すなわち、 $f_q(c)$ は閉じた形式でマッピング関数になり、 Ω_q はゼロ行列になる）を含んでいる。結果として生じるモデルは、 $f_q(c)$ に関する線形関数が利用され、それは q に依存するパラメータ共有構造を有している場合に、共有の因子分析モデルに非常に類似していることに注目することは興味深い。

【0095】

トラジェクトリ HMM が音響モデル $p(c | l, \lambda_c)$ として使用される場合に、 $p(c | q, \lambda_c)$ 及び $p(q | l, \lambda_c)$ は次のように定義され得る。

【数 3 1】

$$p(c | q, \lambda_c) = \mathcal{N}(c; \bar{c}_q, P_q), \quad (29)$$

$$p(q | l, \lambda_c) = \pi_{q_0} \prod_{t=1}^{T-1} \alpha_{q_t q_{t+1}}, \quad (30)$$

【0096】

ここで、 π_i は状態 i の初期状態確率であり、 α_{ij} は状態 i から状態 j への状態遷移確率であり、

【数 3 2】

$$\bar{c}_q$$

【0097】

及び P_q は q に関するトラジェクトリ HMM の平均ベクトル及び共分散行列に対応する。式 (29) では、

【数 3 3】

$$\bar{c}_q$$

【0098】

及び P_q は以下のように与えられる。

【数 3 4】

$$R_q \bar{c}_q = r_q, \quad (31)$$

$$R_q = W^T \Sigma_q^{-1} W = P_q^{-1}, \quad (32)$$

$$r_q = W^T \Sigma_q^{-1} \mu_q, \quad (33)$$

【0099】

ここで W は典型的には、力学的な特徴（速度及び加速度の特徴）と c に付加する $3T(C+1) \times T(C+1)$ ウィンドウ行列である。例えば、 c_t 、 $\Delta^{(0)} c_t$ 、 $\Delta^{(1)} c_t$ 及び $\Delta^{(2)} c_t$ の静的な、速度の、加速度の特徴は次のように計算される。

【数 3 5】

$$\Delta^{(0)} z_t = z_t, \quad (34)$$

$$\Delta^{(1)} z_t = (z_{t+1} - z_{t-1}) / 2, \quad (35)$$

$$\Delta^{(2)} z_t = z_{t-1} - 2z_t + z_{t+1}, \quad (36)$$

【0 1 0 0】

このとき、Wは以下のとおりである。

【数 3 6】

$$\begin{bmatrix} \vdots \\ \Delta^{(0)} c_{t-1} \\ \Delta^{(1)} c_{t-1} \\ \Delta^{(2)} c_{t-1} \\ \Delta^{(0)} c_t \\ \Delta^{(1)} c_t \\ \Delta^{(2)} c_t \\ \Delta^{(0)} c_{t+1} \\ \Delta^{(1)} c_{t+1} \\ \Delta^{(2)} c_{t+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & 0 & I & 0 & 0 & \cdots \\ \cdots & -I/2 & 0 & I/2 & 0 & \cdots \\ \cdots & I & -2I & I & 0 & \cdots \\ \cdots & 0 & 0 & I & 0 & \cdots \\ \cdots & 0 & -I/2 & 0 & I/2 & \cdots \\ \cdots & 0 & I & -2I & I & \cdots \\ \cdots & 0 & 0 & 0 & I & \cdots \\ \cdots & 0 & 0 & -I/2 & 0 & \cdots \\ \cdots & 0 & 0 & I & -2I & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ c_{t-2} \\ c_{t-1} \\ c_t \\ c_{t+1} \\ \vdots \end{bmatrix} \quad (37)$$

10

20

【0 1 0 1】

ここでIと0は、 $(C+1) \times (C+1)$ の単位行列及びゼロ行列に対応する。式(32)及び式(33)での μ_q 及び

【数 3 7】

$$\sum_q^{-1}$$

【0 1 0 2】

は、状態シーケンスqに関して、 $3T(C+1) \times 1$ の平均パラメータベクトル及び $3T(C+1) \times 3T(C+1)$ の精度パラメータ行列に対応していて、次式で与えられる。

【数 3 8】

$$\mu_q = \left[\mu_{q_0}^\top \cdots \mu_{q_{T-1}}^\top \right]^\top, \quad (38)$$

$$\Sigma_q^{-1} = \text{diag} \left\{ \Sigma_{q_0}^{-1}, \dots, \Sigma_{q_{T-1}}^{-1} \right\}, \quad (39)$$

【0 1 0 3】

ここで、 μ_i と Σ_i は、状態iに関して、 $3(C+1)$ の平均パラメータベクトル及び $3(C+1) \times 3(C+1)$ の精度パラメータ行列が対応し、 $Y = \text{diag} \{ X_1, \dots, X_D \}$ は行列 $\{ X_1, \dots, X_D \}$ がYの対角なサブ行列であることを意味する。平均パラメータベクトル及び精度パラメータ行列は以下で定義される。

【数 3 9】

$$\mu_i = \left[\Delta^{(0)} \mu_i^\top \Delta^{(1)} \mu_i^\top \Delta^{(2)} \mu_i^\top \right]^\top, \quad (40)$$

$$\Sigma_i^{-1} = \text{diag} \left\{ \Delta^{(0)} \Sigma_i^{-1}, \Delta^{(1)} \Sigma_i^{-1}, \Delta^{(2)} \Sigma_i^{-1} \right\}, \quad (41)$$

【0 1 0 4】

ここで $\Delta^{(j)} \mu_i$ 及び

30

40

50

【数 4 0】

$$\Delta^{(i)} \Sigma_i^{-1}$$

【0 1 0 5】

は、状態 i と関連して、 $(C + 1) \times 1$ の平均パラメータベクトル及び $(C + 1) \times (C + 1)$ の精度パラメータ行列に対応する。

【0 1 0 6】

最終のパラメータモデルは、マッピングモデルを介して音響及び励振のモデルを組み合わせることによって以下のように得られる：

【数 4 1】

10

$$p(s | \ell, \lambda) = \sum_q \iint p(s | H_c, q, \lambda_e) p(H_c | c, q, \lambda_h) p(c | q, \lambda_c) p(q | \ell, \lambda_c) dH_c dc, \quad (42)$$

【0 1 0 7】

ここで、

【数 4 2】

$$p(s | H_c, q, \lambda_e) = |H_c|^{-1} \mathcal{N}\left(H_c^{-1} s; H_{v,q} t, (G_q^T G_q)^{-1}\right), \quad (43)$$

$$p(H_c | c, q, \lambda_h) = \mathcal{N}(H_c; f_q(c), \Omega_q), \quad (44)$$

$$p(c | q, \lambda_c) = \mathcal{N}(c; \bar{c}_q, P_q), \quad (45)$$

20

$$p(q | \ell, \lambda_c) = \pi_{q_0} \prod_{t=1}^{T-1} \alpha_{q_t q_{t+1}}, \quad (46)$$

【0 1 0 8】

ここで、 $\lambda = \{ \lambda_e, \lambda_h, \lambda_c \}$

ケプストラム係数、線形予測係数 (LPC)、線スペクトルペア (LSP) 及びそれらの周波数が歪んだバージョンのような様々な可能なスペクトル特徴がある。この実施形態では、ケプストラム係数は特殊な場合と見なされる。ケプストラム係数ベクトル

【数 4 3】

30

$$c_t = [c_t(0), \dots, c_t(C)]^T$$

【0 1 0 9】

からその対応する声道フィルタインパルス応答ベクトル

【数 4 4】

$$h_{c,t} = [h_{c,t}(0), \dots, h_{c,t}(P)]^T$$

【0 1 1 0】

へのマッピングは以下のような閉じた形式で書かれうる。

【数 4 5】

$$h_{c,t} = D_s^* \cdot \text{EXP}[D_s c_t], \quad (47)$$

40

【0 1 1 1】

ここで $\text{EXP}[\cdot]$ は、 $[\cdot]$ の要素の指数関数をとることによって導かれる行列を意味し、 D は $(P + 1) \times (C + 1)$ の DFT (離散フーリエ変換) 行列であり、

【数 4 6】

$$D_s = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_{P+1} & \cdots & W_{P+1}^C \\ \vdots & \vdots & & \vdots \\ 1 & W_{P+1}^P & \cdots & W_{P+1}^{PC} \end{bmatrix}, \quad (48)$$

【 0 1 1 2】

ここで

10

【数 4 7】

$$W_{P+1} = e^{-\frac{2\pi}{P+1}j}, \quad (49)$$

【 0 1 1 3】

また、 D^* は、次の形式を備えた $(P+1) \times (P+1)$ の I D F T (逆離散フーリエ変換) 行列である。

【数 4 8】

$$D_s^* = \frac{1}{P+1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_{P+1}^{-1} & \cdots & W_{P+1}^{-P} \\ \vdots & \vdots & & \vdots \\ 1 & W_{P+1}^{-P} & \cdots & W_{P+1}^{-P^2} \end{bmatrix}. \quad (50)$$

20

【 0 1 1 4】

ケプストラム係数と声道フィルタ応答との間のマッピングとして、閉じた形式でのコンピュータがあり得る。 c と Hc との間の確率近似を使用する必要はない。式 (10) の生成モデルに現れる声道フィルタインパルス応答関連の項は Hc であり hc でない。式 (12) 及び (13) として与えられた Hc と hc との関係は以下で与えられる。

【数 4 9】

30

$$h_c = \begin{bmatrix} h_{c,0}^\top & \cdots & h_{c,T-1}^\top \end{bmatrix}^\top \quad (51)$$

$$h_{c,t} = \begin{bmatrix} h_{c,t}(0) & \cdots & h_{c,t}(P) \end{bmatrix}^\top \quad (52)$$

【 0 1 1 5】

hc 、 t は t 番目のフレームの合成フィルタインパルス応答であり、 T は全フレーム数であり、以下のように書かれる。

【数 5 0】

40

$$H_c = \sum_{n=0}^{N-1} J_n B h_c j_n^\top. \quad (53)$$

【 0 1 1 6】

式 (53) では、 N は (データベースの) サンプル数であり、

【数 5 1】

$$j_n = \left[\underbrace{0 \cdots 0}_n \quad 1 \quad \underbrace{0 \cdots 0}_{N-1-n} \right]^T, \quad (54)$$

$$B = \begin{bmatrix} I_{P+1} & 0_{P+1,P+1} & \cdots & 0_{P+1,P+1} \\ \vdots & \vdots & & \vdots \\ I_{P+1} & 0_{P+1,P+1} & \cdots & 0_{P+1,P+1} \\ \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & & \vdots \\ \cdots & \cdots & \cdots & \cdots \\ 0_{P+1,P+1} & 0_{P+1,P+1} & \cdots & I_{P+1} \\ \vdots & \vdots & & \vdots \\ 0_{P+1,P+1} & 0_{P+1,P+1} & \cdots & I_{P+1} \end{bmatrix}, \quad (55)$$

10

【 0 1 1 7】

ここで B は、フレーム基底 h c ベクトルをサンプル基底にマッピングするための $N(P+1) \times T(P+1)$ 行列である。

【 0 1 1 8】

20

最後の P 行を無視することにより、H c の正方バージョンが考慮されることは注目すべきである。 $N \times N(P+1)$ 行列 J_n は以下のように構築される。

【数 5 2】

$$J_0 = \begin{bmatrix} I_{P+1} & 0_{P+1,N(P+1)-P-1} \\ \cdots & \cdots \\ 0_{N-P-1,P+1} & 0_{N-P-1,N(P+1)-P-1} \end{bmatrix}, \quad (56)$$

$$J_1 = \begin{bmatrix} 0_{1,P+1} & 0_{1,P+1} & 0_{1,N(P+1)-2P-2} \\ \cdots & \cdots & \cdots \\ 0_{P+1,P+1} & I_{P+1} & 0_{P+1,N(P+1)-2P-2} \\ \cdots & \cdots & \cdots \\ 0_{N-P-2,P+1} & 0_{N-P-2,P+1} & 0_{N-P-2,N(P+1)-2P-2} \end{bmatrix}, \quad (57)$$

30

$$\vdots$$

$$J_{N-1} = \begin{bmatrix} 0_{N-1,N(P+1)-P-1} & 0_{N-1,1} & 0_{N-1,P} \\ \cdots & \cdots & \cdots \\ 0_{1,N(P+1)-P-1} & 1 & 0_{1,P} \end{bmatrix}, \quad (58)$$

40

【 0 1 1 9】

ここで、 $0_{x,y}$ は X 行 Y 列でのゼロ要素の行列を意味し、 I_x は X サイズの単位行列である。

【 0 1 2 0】

モデルのトレーニングは今、図 5 及び 6 を参照して説明される。

【 0 1 2 1】

トレーニングは、共同モデル λ のパラメータが次のように推定されることを可能にする

:

【数 5 3】

$$\hat{\lambda} = \arg \max_{\lambda} p(s | \ell, \lambda), \quad (59)$$

【0 1 2 2】

ここで、励振モデルのパラメータに対応する $\lambda_e = \{H_v, G, t\}$ で $\lambda = \{\lambda_e, \lambda_h, \lambda_c\}$ であり、音響モデルのパラメータを構成する $\lambda = \{m, \sigma\}$ 、

【数 5 4】

$$m = [\mu_0^T \cdots \mu_{S-1}^T]^T, \quad (60)$$

$$\sigma = \text{vdiag} \{ \text{diag} \{ \Sigma_0^{-1}, \dots, \Sigma_{S-1}^{-1} \} \}, \quad (61)$$

10

【0 1 2 3】

ここで S は状態数である。 m と σ はそれぞれ、全ての状態の逆共分散行列の全ての平均及び対角を連結させることによって形成されるベクトルである。ここで $\text{vdiag} \{ [\cdot] \}$ は $[\cdot]$ の対角要素によって形成されるベクトルを意味している。スペクトル特徴としてのケプストラム係数を仮定する尤度関数 $p(s | \ell, \lambda)$ は、以下ようになる。

【数 5 5】

$$p(s | \ell, \lambda) = \sum_{\forall q} \iint p(s | H_c, q, \lambda_e) p(H_c | c, q, \lambda_h) p(c | q, \lambda_c) p(q | \ell, \lambda_c) dH_c dc. \quad (62)$$

20

【0 1 2 4】

残念ながら、期待値最大化 (EM: expectation-maximization) を介するこのモデルの推定は、解決困難である。したがって、近似の再帰的なアプローチが採用される。

【0 1 2 5】

式 (62) での全ての可能な q に渡る和が固定された状態シーケンスによって近似される場合には、上記の尤度関数は以下になる。

【数 5 6】

$$p(s | \ell, \lambda) \approx \iint p(s | H_c, \hat{q}, \lambda_e) p(H_c | c, \hat{q}, \lambda_h) p(c | \hat{q}, \lambda_c) p(\hat{q} | \ell, \lambda_c) dc, \quad (63)$$

【0 1 2 6】

30

ここで、

【数 5 7】

$$\hat{q} = \{\hat{q}_0, \dots, \hat{q}_{T-1}\}$$

【0 1 2 7】

は状態シーケンスである。さらに、全ての可能な c に渡る積分がスペクトルベクトル及びインパルス応答ベクトルによって近似される場合には、式 (63) は以下になる。

【数 5 8】

$$p(s | \ell, \lambda) \approx p(s | \hat{H}_c, \hat{q}, \lambda_e) p(\hat{H}_c | \hat{c}, \hat{q}, \lambda_h) p(\hat{c} | \hat{q}, \lambda_c) p(\hat{q} | \ell, \lambda_c), \quad (64)$$

40

【0 1 2 8】

ここで、

【数 5 9】

$$\hat{c} = [\hat{c}_1, \dots, \hat{c}_{T-1}]^T$$

【0 1 2 9】

は固定スペクトル応答ベクトルである。

【0 1 3 0】

式 (64) の対数を取ること、または最大化されるコスト関数によって、音響、励振、及びマッピングモデルパラメータの更新を介して、以下が得られる。

50

【数 6 0】

$$\mathcal{L} = \log p(s | \hat{H}_c, \hat{q}, \lambda_e) + \log p(\hat{H}_c | \hat{c}, \hat{q}, \lambda_h) + \log p(\hat{c} | \hat{q}, \lambda_c) + \log p(\hat{q} | \ell, \lambda_c). \quad (65)$$

【0 1 3 1】

最適化問題は2つの部分へ分割されうる：初期化と反復である。以下は、各部分で行なわれる計算について説明する。初期化は、図5を参照して説明され、反復は図6を参照して説明される。

【0 1 3 2】

モデルは、対応するテキストデータを備えた音声データであり、ステップS210で入力されるトレーニングデータを使用してトレーニングされる。

10

【0 1 3 3】

パート1 - 初期化

1. ステップS203では、音声データは最初の1つのケプストラム係数ベクトルから抽出される。

【数 6 1】

$$\mathbf{c} = [\mathbf{c}_0^T \quad \dots \quad \mathbf{c}_{T-1}^T], \quad (66)$$

$$\mathbf{c}_t = [c_t(0) \quad \dots \quad c_t(C)]^T. \quad (67)$$

【0 1 3 4】

2. ステップS205では、トラジェクトリHMMパラメータ λ_c は $\mathbf{c}$ を使用してトレーニングされる。

20

【数 6 2】

$$\hat{\lambda}_c = \arg \max_{\lambda_c} p(\mathbf{c} | \lambda_c). \quad (68)$$

【0 1 3 5】

3. ステップS207では、最良の状態シーケンス

【数 6 3】

$$\hat{q}$$

30

【0 1 3 6】

は、H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequence," Computer Speech and Language, vol. 21, pp. 153-173, Jan. 2007のアルゴリズムを使用することによって、トレーニングされたモデルからビタビパスとして決定される。

【数 6 4】

$$\hat{q} = \arg \max_q p(\mathbf{c}, q | \lambda_c). \quad (69)$$

【0 1 3 7】

4. ステップS209では、

40

【数 6 5】

$$\hat{q}$$

【0 1 3 8】

及び $\mathbf{c}$ を仮定して、マッピングモデルパラメータ λ_h は推定される。

【数 6 6】

$$\hat{\lambda}_h = \arg \max_{\lambda_h} p(\mathbf{H}_c | \mathbf{c}, \hat{q}, \lambda_h). \quad (70)$$

【0 1 3 9】

50

5 . ステップ S 2 1 1 では、R. Maia, T. Toda, H. Zen, Y. Nakaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modelling," in Proc. of the 6th ISCA Workshop on Speech Synthesis, 2007で記述されているアルゴリズムの1回の反復を使用することによって、励振パラメータ λ_e は

【数 6 7】

$$\hat{q}$$

【 0 1 4 0 】

及び c を仮定して、評価される。

【数 6 8】

$$\hat{\lambda}_e = \arg \max_{\lambda_e} p(s | H_c, \hat{q}, \lambda_e). \quad (71)$$

10

【 0 1 4 1 】

部分 2 : 反復

1 . 図 6 のステップ S 2 1 3 では、最良のケプストラム係数ベクトル

【数 6 9】

$$\hat{c}$$

【 0 1 4 2 】

は式 (6 5) の対数尤度関数を使用して推定される。

20

【数 7 0】

$$\hat{c} = \arg \max_c \mathcal{L}. \quad (72)$$

【 0 1 4 3 】

2 . ステップ S 2 1 5 では、声道フィルタインパルス応答 H_c は

【数 7 1】

$$\hat{q}$$

【 0 1 4 4 】

及び

30

【数 7 2】

$$\hat{c}$$

【 0 1 4 5 】

を仮定して推定される。

【数 7 3】

$$\hat{H}_c = \arg \max_{H_c} \mathcal{L}. \quad (73)$$

【 0 1 4 6 】

3 . ステップ S 2 1 7 では、励振モデルパラメータ λ_e は、

40

【数 7 4】

$$\hat{q}$$

【 0 1 4 7 】

及び

【数 7 5】

$$\hat{H}_c$$

【 0 1 4 8 】

を仮定して更新される。

50

【数 7 6】

$$\hat{\lambda}_c = \arg \max_{\lambda_c} \mathcal{L}. \quad (74)$$

【0 1 4 9】

4 . ステップ S 2 1 9 では、音響モデルパラメータが更新される。

【数 7 7】

$$\hat{\lambda}_c = \arg \max_{\lambda_c} \mathcal{L}. \quad (75)$$

【0 1 5 0】

5 . ステップ S 2 2 1 では、マッピングモデルパラメータが更新される。

10

【数 7 8】

$$\hat{\lambda}_h = \arg \max_{\lambda_h} \mathcal{L}. \quad (76)$$

【0 1 5 1】

再帰的なステップは数回繰り返されてもよい。以下では、それらの内のそれぞれ 1 つが詳細に説明される。

【0 1 5 2】

反復は収束までに終了する。収束は多くの異なった方法で決定されてもよい。一実施形態では、尤度での変化が所定の最小値よりも小さい場合に収束は生じたと見なされるだろう。例えば、尤度 L での変化が 5 % 未満である。

20

【0 1 5 3】

反復のステップ 1 S 2 1 3 では、ケプストラム係数がスペクトル特徴として使用される場合には、式 (6 5) の尤度関数は次のように書かれる。

【数 7 9】

$$\begin{aligned} \mathcal{L} = & -\frac{1}{2} \left\{ (N+M) \log(2\pi) + \log |G_q^\top G_q| - 2 \log |H_c| + s^\top H_c^{-\top} G_q^\top G_q H_c^{-1} s - \right. \\ & - 2 s^\top H_c^{-\top} G_q^\top G_q H_{v,q} t + t^\top H_{v,q}^\top G_q^\top G_q H_{v,q} t + T(C+1) \log(2\pi) - \\ & \left. - \log |R_q| + c^\top R_q c - 2 r_q^\top c + r_q^\top R_q^{-1} r_q \right\}, \quad (77) \end{aligned}$$

30

【0 1 5 4】

c に依存する項が選択され、以下によって与えられるコスト関数 h c を構成する。

【数 8 0】

$$\mathcal{L}_c = -\frac{1}{2} s^\top H_c^{-\top} G_q^\top G_q H_c^{-1} s + \log |H_c| + s^\top H_c^{-\top} G_q^\top G_q H_{v,q} t - \frac{1}{2} c^\top R_q c + r_q^\top c. \quad (78)$$

【0 1 5 5】

最良のケプストラム係数ベクトル

【数 8 1】

$$\hat{c}$$

40

【0 1 5 6】

は、コスト関数 h c を最大にするものとして定義されうる。最急勾配傾斜アルゴリズム (steepest gradient ascent algorithm) (例えば J. Nocedal and S. J. Wright, Numerical Optimization. Springer, 1999 を参照)、または Broyden Fletcher Goldfarb Shanno (BFGS) アルゴリズムのような別の最適化方法を利用することによって、

【数 8 2】

$$\hat{c}$$

【0 1 5 7】

50

に関するそれぞれの更新は以下によって計算することができる。

【数 8 3】

$$\hat{c}^{(i+1)} = \hat{c}^{(i)} + \gamma \frac{\partial \mathcal{L}_c}{\partial c}, \quad (79)$$

【0 1 5 8】

ここで γ は収束ファクタ（定数）であり、 i は反復インデックスであり、そして、

【数 8 4】

$$\frac{\partial \mathcal{L}}{\partial c} = D^T \text{DIAG}(\text{EXP}[Dc]) D^{*T} B^T \left\{ \sum_{n=0}^{N-1} J_n^T H_c^{-T} [G_q^T G_q (e - v) e^T - I] j_n \right\} - R_q c + r_q. \quad (80)$$

10

【0 1 5 9】

ここで、 $\text{Diag}([\cdot])$ はベクトル $[\cdot]$ の要素で形成される対角行列を意味し、

【数 8 5】

$$e = H_c^{-1} s, \quad (81)$$

$$v = H_{v,q} t, \quad (82)$$

$$D = \text{diag} \left\{ \underbrace{D_s, \dots, D_s}_T \right\}, \quad (83)$$

$$D^* = \text{diag} \left\{ \underbrace{D_s^*, \dots, D_s^*}_T \right\}. \quad (84)$$

20

【0 1 6 0】

上記では、反復過程は収束まで継続する。好ましい実施形態では、連続する反復間の差が 5 % 未満である場合に、収束が生じている。

【0 1 6 1】

再帰的手続きのステップ 3 S 2 1 7 では、励振パラメータ $\lambda_e = \{ H_v, G, G \}$ は、R. Maia, T. Toda, H. Zen, Y. Nakaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modelling," in Proc. Of the 6th ISCA Workshop on Speech Synthesis, 2007で説明されるアルゴリズムの 1 つの反復の使用に
より計算される。

30

【0 1 6 2】

この場合、推定されたケプストラム係数ベクトル

【数 8 6】

$$\hat{c}$$

【0 1 6 3】

は、残差ベクトル

【数 8 7】

$$e = H_c^{-1} s$$

40

【0 1 6 4】

を逆フィルタリングを介して抽出するために使用される。

【0 1 6 5】

ステップ 4 S 2 1 9 では、音響モデルパラメータの推定 $\lambda = \{ m, \sigma \}$ は、観測として最良に推定されたケプストラムベクトル

【数 8 8】

$$\hat{c}$$

【0 1 6 6】

50

を利用することによって、H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequence," Computer Speech and Language, vol. 21, pp. 153-173, Jan. 2007に記述されたようになされる。

【0167】

上記のトレーニング方法は、マッピングモデルの1セットのモデルパラメータ λ_h を使用して、 $f_q(c)$ によって予測された H_c の不確定性について記述する。

【0168】

しかしながら、他の実施形態では、決定論的な場合が仮定され、 $f_q(c)$ が完全に H_c を予測する。この実施形態では、 H_c と $f_q(c)$ との間に不確定性はなく、したがって、 λ_h はもはや必要ではない。

10

【0169】

そのようなシナリオでは、マッピングモデルパラメータは図5のステップS209でゼロに設定され、図6のステップS221で再度推定されない。

【0170】

図7は、本発明の実施形態で用いる音声合成方法のフローチャートである。

【0171】

テキストはステップS251で入力される。音響モデルはこのテキスト上で実行され、スペクトル特徴及びF0特徴を含む特徴はステップS253で抽出される。

【0172】

インパルス応答フィルタ関数は、ステップS253で抽出されたスペクトル特徴からステップS255で生成される。

20

【0173】

入力テキストもまた励振モデルに入力される。また、励振モデルパラメータはステップS257で入力テキストから生成される。

【0174】

ステップS253で抽出された特徴に戻って、この段階で抽出されたF0特徴はステップS259でのパルス列に変換される。パルス列はステップS257で生成された有声フィルタ関数を使用してフィルタリングされる。

【0175】

白色雑音は白色ノイズ発生器によって生成される。その後、白色雑音はステップS257で生成された無声フィルタ関数を使用して、ステップS263でフィルタリングされる。

30

【0176】

その後、ステップS261で生成された有声励振信号、及びステップS263で生成された無声励振信号は、ステップS265での混合励振信号を生成するために混合される。

【0177】

その後、混合励振信号は、ステップS255で生成されたインパルス応答を使用して、ステップS267でフィルタリングされる。また、音声信号が出力される。

【0178】

共同の最適化を介して音響及び励振モデルをトレーニングすることによって、位相情報のような音声パラメータ抽出の間に失われる情報は、自然な音声により近づいて聞こえる合成音声になって、実行時間で回復されるかもしれない。したがって、統計パラメータのテキストを音声に変換するシステムは、自然な音声に非常に似ているように聞こえる合成音声を生産する能力で生産されうる。

40

【0179】

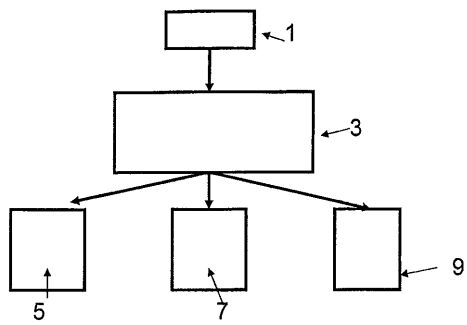
ある実施形態は記述されているが、これらの実施形態は例のみによって提示されていて、発明の範囲を制限するようには意図されない。実際、ここに記述された新しい方法及びシステムは様々な他の形式で具体化されてもよい。更に、ここに記述された方法及びシステムの形式での様々な省略、置換及び変更は、添付されたクレームによって定義されるよ

50

うな発明の範囲から外れずになされてもよい。

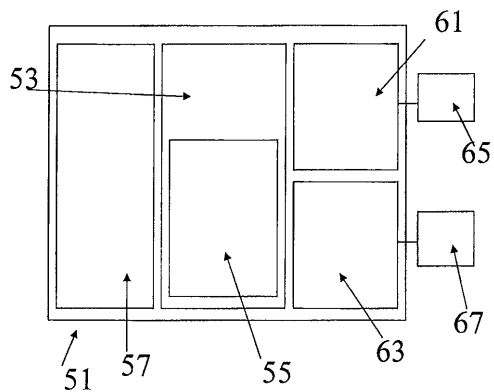
【図 1】

図 1



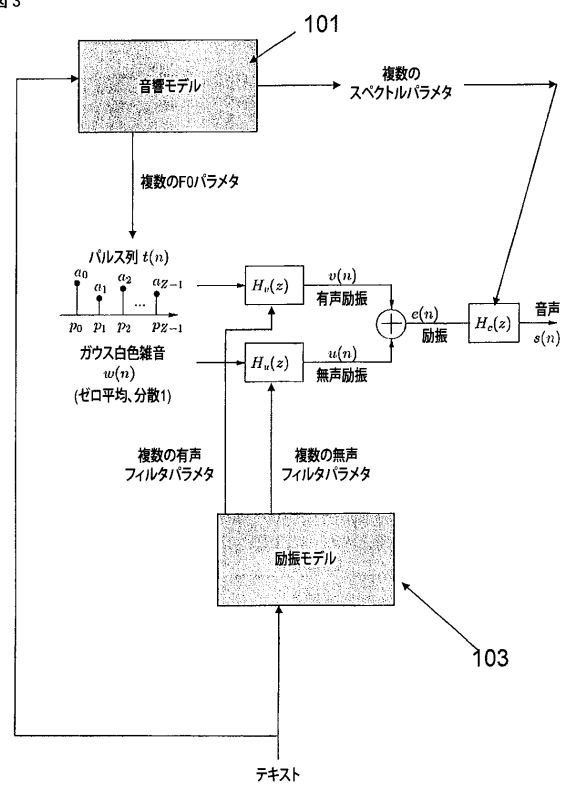
【図 2】

図 2



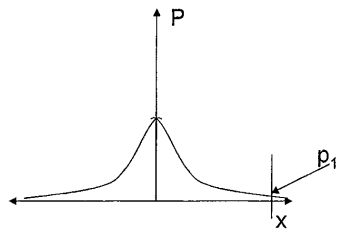
【図 3】

図 3



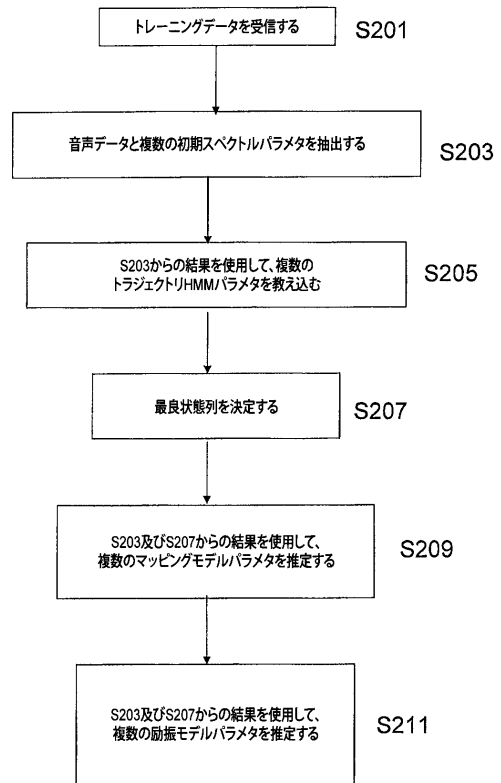
【図 4】

図 4



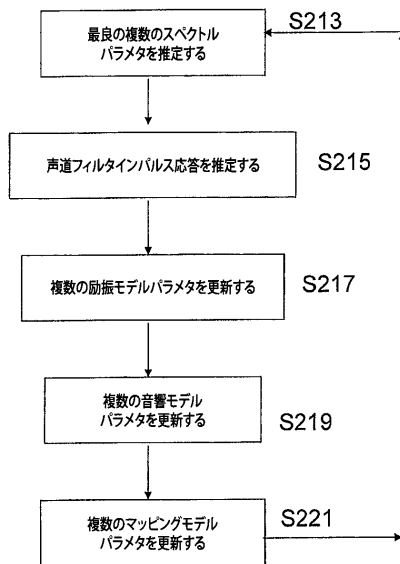
【図 5】

図 5



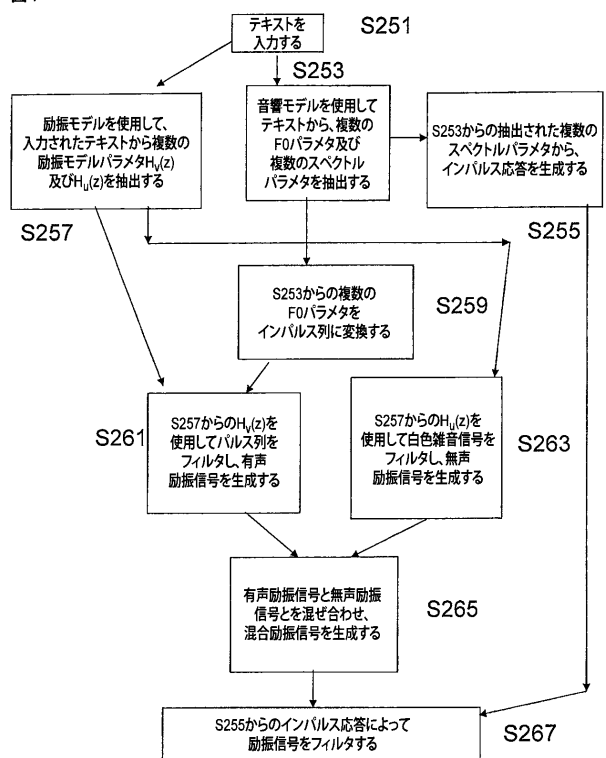
【図 6】

図 6



【図 7】

図 7



【手続補正書】

【提出日】平成23年6月24日(2011.6.24)

【手続補正 1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【特許請求の範囲】

【請求項 1】

テキスト入力を受け取り、確率論的モデルを使用して前記テキスト入力に対応する音声
を出力し、前記確率論的モデルは音響モデルと励振モデルとを具備し、前記音響モデルは
単語または単語の部分のある特徴に関連づける複数の確率分布を記述する複数のモデルパ
ラメータを有し、前記励振モデルは声帯及び肺をモデル化して特徴を使用して音声を出力
するために使用される励振モデルパラメータを具備し、

前記音響パラメータ及び前記励振パラメータは一緒に推定され、

前記音声を出力することを具備する音声処理方法。

【請求項 2】

前記テキスト入力は、前記音響モデルによって処理され、 F_0 及びスペクトルの特徴を
出力し、

パルス列を形成するために前記 F_0 特徴を処理し、

励振信号を生成するために前記励振モデルから導かれる複数の励振パラメータを使用して
前記パルス列をフィルタリングし、

前記スペクトル特徴から導かれる複数のフィルタパラメータを使用して前記励振信号を
フィルタリングすることをさらに具備する請求項 1 に記載の音声処理方法。

【請求項 3】

音声合成に関する統計モデルをトレーニングする方法であって、

トレーニングデータを受け取り、前記トレーニングデータは音声と前記音声に対応する
テキストとを具備し、

確率論的モデルをトレーニングし、前記確率論的モデルは音響モデルと励振モデルとを
具備し、前記音響モデルは単語またはその単語の部分の特徴ベクトルに関連づける複数の
確率分布を記述する複数のモデルパラメータを有していて、前記励振モデルは声帯及び肺
をモデル化して前記音声を出力する複数の励振モデルパラメータを具備し、

前記音響パラメータ及び前記励振パラメータはトレーニング処理の間に一緒に推定され
る方法。

【請求項 4】

前記音響モデルパラメータは、前記確率分布の平均と分散とを具備する請求項 3 に記載
の方法。

【請求項 5】

前記音響モデルによって出力される前記特徴は F_0 特徴とスペクトル特徴を具備する請
求項 3 に記載の方法。

【請求項 6】

前記励振モデルパラメータは、 F_0 特徴から導かれるパルス信号をフィルタリングする
フィルタ係数を具備する請求項 5 に記載の方法。

【請求項 7】

共同に推定する処理は、再帰的な処理を具備し、

1 つのステップでは励振パラメータが音響パラメータの最新の推定を使用して更新され
、別のステップでは音響モデルパラメータが励振パラメータの前記最新の推定を使用して
更新される請求項 3 に記載の方法。

【請求項 8】

共同に推定する処理は、最大尤度技術を使用する請求項 3 に記載の方法。

【請求項 9】

前記確率論的モデルは、マッピングモデルをさらに具備し、前記マッピングモデルは複数のマッピングモデルパラメータを具備し、前記マッピングモデルは複数のスペクトル特徴を、人間の声道を表現する複数のフィルタ係数にマッピングする請求項 5 に記載の方法。

【請求項 10】

前記パラメータは以下のように一緒に推定され、

【数 1】

$$\hat{\lambda} = \arg \max_{\lambda} p(s | \ell, \lambda),$$

ここで、 λ は最適化される前記励振モデルと音響モデルとのパラメータを表し、 s は自然な音声波形であり、

【数 2】

ℓ

は前記音声波形の表記である請求項 3 に記載の方法。

【請求項 11】

λ は、人間の声道を表現するために、複数のスペクトルパラメータをフィルタ関数にマッピングするマッピングモデルの複数のパラメータをさらに具備する請求項 10 に記載の方法。

【請求項 12】

前記スペクトル特徴とフィルタ係数との間の関係は、ガウスプロセスとしてモデル化される請求項 11 に記載の方法。

【請求項 13】

$p(s | \ell, \lambda)$ は以下のように表され、

【数 3】

$$p(s | \ell, \lambda) = \sum_{\forall q} \iint p(s | H_c, q, \lambda_c) p(H_c | c, q, \lambda_h) p(c | q, \lambda_c) p(q | \ell, \lambda_c) dH_c dc. \quad (62)$$

ここで、 H_c は前記人間の声道をモデル化するために使用されるフィルタ関数であり、 q は状態であり、 λ_c は励振パラメータであり、 λ_h は音響モデルパラメータであり、 λ はマッピングモデルパラメータであり、 c はスペクトル特徴である請求項 11 に記載の方法。

【請求項 14】

q に渡る和は固定された状態シーケンスによって近似されて次を与え、

【数 4】

$$p(s | \ell, \lambda) \approx \iint p(s | H_c, \hat{q}, \lambda_c) p(H_c | c, \hat{q}, \lambda_h) p(c | \hat{q}, \lambda_c) p(\hat{q} | \ell, \lambda_c) d\hat{q} dc, \quad (63)$$

ここで

【数 5】

$$\hat{q} = \{\hat{q}_0, \dots, \hat{q}_{T-1}\}$$

は状態シーケンスである請求項 13 に記載の方法。

【請求項 15】

前記全ての可能な H_c 及び c に渡る積分は、スペクトル応答及びインパルス応答のベクトルによって近似され、次を与え、

【数 6】

$$p(s | \ell, \lambda) \approx p(s | \hat{H}_c, \hat{q}, \lambda_e) p(\hat{H}_c | \hat{c}, \hat{q}, \lambda_h) p(\hat{c} | \hat{q}, \lambda_c) p(\hat{q} | \ell, \lambda_c), \quad (64)$$

ここで、

【数 7】

$$\hat{c} = [\hat{c}_1, \dots, \hat{c}_{T-1}]^T$$

は固定されたスペクトル応答である請求項 14 に記載の方法。

【請求項 16】

$p(s | \ell, \lambda)$ の対数尤度関数は以下

【数 8】

$$\mathcal{L} = \log p(s | \hat{H}_c, \hat{q}, \lambda_e) + \log p(\hat{H}_c | \hat{c}, \hat{q}, \lambda_h) + \log p(\hat{c} | \hat{q}, \lambda_c) + \log p(\hat{q} | \ell, \lambda_c). \quad (65)$$

によって与えられる請求項 15 に記載の方法。

【請求項 17】

単語のシーケンスを具備するテキスト入力を受け取るレシーバーと、

プロセッサと、を具備し、前記プロセッサは前記入力テキストに対応する出力音声の尤度を確率論的モデルを使用して判定し、前記確率論的モデルは音響モデルと励振モデルを具備し、前記音響モデルは単語または単語の部分と特徴とを関連づける複数の確率分布を記述する複数のモデルパラメータを有して、前記励振モデルは、前記声帯及び肺をモデル化して前記特徴を使用して前記音声を出力するために使用される複数の励振モデルパラメータを具備し、

前記音響パラメータ及び励振パラメータは一緒に推定され、前記装置は前記音声のための出力を具備する音声処理装置。

【請求項 18】

入力音声認識部と、翻訳部と、請求項 18 に記載の音声処理装置と、を具備する音声翻訳システム。

【請求項 19】

コンピュータに、

テキスト入力を受け取る手順、確率論的モデルを使用して前記テキスト入力に対応する音声を出力する手順、前記確率論的モデルは音響モデルと励振モデルとを具備し、前記音響モデルは単語または単語の部分とある特徴とを関連づける複数の確率分布を記述する複数のモデルパラメータを有し、前記励振モデルは声帯及び肺をモデル化して特徴を使用して音声を出力するために使用される励振モデルパラメータを具備し、

前記音響パラメータ及び前記励振パラメータは一緒に推定され、

前記音声を出力する手順を実行させるための音声処理プログラム。

【請求項 20】

音声合成に関する統計モデルをトレーニングするためのプログラムであって、

コンピュータに、

トレーニングデータを受け取る手順、前記トレーニングデータは音声と前記音声に対応するテキストとを具備し、

確率論的モデルをトレーニングする手順を実行させるためのものであり、前記確率論的モデルは音響モデルと励振モデルとを具備し、前記音響モデルは単語またはその単語の部分と特徴ベクトルとを関連づける複数の確率分布を記述する複数のモデルパラメータを有して、前記励振モデルは声帯及び肺をモデル化して前記音声を出力する複数の励振モデルパラメータを具備し、

前記音響パラメータ及び前記励振パラメータはトレーニング処理の間に一緒に推定されるプログラム。

【請求項 21】

コンピュータを、

単語のシーケンスを具備するテキスト入力を受け取るレシーバー手段と、

プロセッサ手段と、として機能させ、前記プロセッサ手段は前記入力テキストに対応する出力音声の尤度を確率論的モデルを使用して判定し、前記確率論的モデルは音響モデルと励振モデルを具備し、前記音響モデルは単語または単語の部分と特徴とを関連づける複数の確率分布を記述する複数のモデルパラメータを有して、前記励振モデルは、前記声帯及び肺をモデル化して前記特徴を使用して前記音声を出力するために使用される複数の励振モデルパラメータを具備し、

前記音響パラメータ及び励振パラメータは一緒に推定され、前記装置は前記音声のための出力手段として機能させるための音声処理プログラム。

フロントページの続き

- (74)代理人 100095441
弁理士 白根 俊郎
- (74)代理人 100084618
弁理士 村松 貞男
- (74)代理人 100103034
弁理士 野河 信久
- (74)代理人 100119976
弁理士 幸長 保次郎
- (74)代理人 100153051
弁理士 河野 直樹
- (74)代理人 100140176
弁理士 砂川 克
- (74)代理人 100158805
弁理士 井関 守三
- (74)代理人 100124394
弁理士 佐藤 立志
- (74)代理人 100112807
弁理士 岡田 貴志
- (74)代理人 100111073
弁理士 堀内 美保子
- (74)代理人 100134290
弁理士 竹内 将訓
- (72)発明者 マイア・ラニーリー
イギリス国、 ビーエス１・４エヌデー、 ブリストル、 クウィーンスクエア ３ ２
- (72)発明者 チュン・ビュン・ハ
イギリス国、 ビーエス１・４エヌデー、 ブリストル、 クウィーンスクエア ３ ２

【外国語明細書】
2011237795000001.pdf