(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6580911号
(P6580911)
(45)発行日　令和1年9月25日(2019.9.25)　(24)登録日　令和1年9月6日(2019.9.6)

(51)Int.Cl.　F I
G10L 13/06 (2013.01)　G10L 13/06 130
G10L 13/06 120Z

請求項の数 10　(全 14 頁)

(21)出願番号　特願2015-174715(P2015-174715)
(22)出願日　平成27年9月4日(2015.9.4)
(65)公開番号　特開2017-49535(P2017-49535A)
(43)公開日　平成29年3月9日(2017.3.9)
審査請求日　平成30年2月13日(2018.2.13)

(73)特許権者　000208891
ＫＤＤＩ株式会社
東京都新宿区西新宿二丁目３番２号
(74)代理人　100092772
弁理士　阪本　清孝
(74)代理人　100119688
弁理士　田邊　壽二
(72)発明者　西澤　信行
埼玉県ふじみ野市大原二丁目１番１５号
株式会社ＫＤＤＩ研究所内

審査官　山下　剛史

最終頁に続く

(54)【発明の名称】音声合成システムならびにその予測モデル学習方法および装置

(57)【特許請求の範囲】
【請求項１】
音声データに基づいて音声合成用の予測モデルを学習する装置において、
前記音声データから複数種の音声合成パラメータを抽出する手段と、
一の音声合成パラメータから生成した標準ベクトルおよび他の一の音声合成パラメータから生成した追加ベクトルに基づいて拡張ベクトルを生成する手段と、
前記拡張ベクトルを音素ごとにモデル化する手段と、
音素モデルの集合に対して、その拡張ベクトルを評価規準としてモデル尤度が最大となる分割条件をノード毎に決定することを繰り返し、各リーフノードに各音声合成パラメータの分布情報が登録された決定木を構築する手段と、
前記決定木の各リーフノードから前記追加ベクトルに対応した分布情報を削除する手段とを具備し、
前記追加ベクトルが、音声合成の際に分布情報を用いられない音声合成パラメータのベクトルであることを特徴とする予測モデル学習装置。
【請求項２】
前記標準ベクトルが、メルケプストラム係数の特徴ベクトルであり、前記追加ベクトルがLSP係数の特徴ベクトルであることを特徴とする請求項１に記載の予測モデル学習装置。
【請求項３】
前記標準ベクトルが、所定の音声合成パラメータに関する所定の時間長の特徴ベクトル

であり、前記追加ベクトルが、前記所定の音声合成パラメータに関して前記所定の時間長の前後少なくとも一方に連続する時間長部分の特徴ベクトルであることを特徴とする請求項１に記載の予測モデル学習装置。

【請求項４】

音声データに基づいて音声合成用の予測モデルを学習する予測モデル学習装置および入力テキストの音素ラベル列を前記予測モデルに適用して音声を合成する音声合成装置を備えた音声合成システムにおいて、

前記予測モデル学習装置が、

前記音声データから複数種の音声合成パラメータを抽出する手段と、

一の音声合成パラメータから生成した標準ベクトルおよび他の一の音声合成パラメータから生成した追加ベクトルを連結して拡張ベクトルを生成する手段と、

前記拡張ベクトルを音素ごとにモデル化する手段と、

音素モデルの集合に対して、その拡張ベクトルを評価規準としてモデル尤度が最大となる分割条件をノード毎に決定することを繰り返し、各リーフノードに各音声合成パラメータの分布情報が登録された決定木を構築する手段と、

前記決定木の各リーフノードから前記追加ベクトルに対応した分布情報を削除する手段とを具備し、

前記追加ベクトルが、音声合成の際に分布情報を用いられない音声合成パラメータのベクトルであり、

前記音声合成装置は、リーフノードに前記標準ベクトルに対応した分布情報のみが残った決定木を用いて音声合成を行うことを特徴とする音声合成システム。

【請求項５】

前記音声合成装置が、

入力テキストからコンテキスト依存の音素ラベル列を生成する手段と、

前記音素ラベル列を決定木に適用し、尤度が最大となる分布情報の時系列を生成する手段と、

前記分布情報の時系列に基づいて音声を合成する手段とを具備したことを特徴とする請求項４に記載の音声合成システム。

【請求項６】

前記標準ベクトルが、メルケプストラム係数の特徴ベクトルであり、前記追加ベクトルがLSP係数の特徴ベクトルであることを特徴とする請求項４または５に記載の音声合成システム。

【請求項７】

前記標準ベクトルが、所定の音声合成パラメータに関する所定の時間長の特徴ベクトルであり、前記追加ベクトルが、前記所定の音声合成パラメータに関して前記所定の時間長の前後少なくとも一方に連続する時間長部分の特徴ベクトルであることを特徴とする請求項４または５に記載の音声合成システム。

【請求項８】

音声データに基づいて音声合成用の予測モデルを学習する方法おいて、

前記音声データから複数種の音声合成パラメータを抽出する手順と、

一の音声合成パラメータに基づいて標準ベクトルを生成する手順と、

他の一の音声合成パラメータに基づいて追加ベクトルを生成する手順と、

前記標準ベクトルおよび追加ベクトルに基づいて拡張ベクトルを生成する手順と、

前記拡張ベクトルを音素ごとにモデル化する手順と、

音素モデルの集合に対して、その拡張ベクトルを評価規準としてモデル尤度が最大となる分割条件をノード毎に決定することを繰り返し、各リーフノードに各音声合成パラメータの分布情報が登録された決定木を構築する手順と、

前記決定木の各リーフノードから前記追加ベクトルに対応した分布情報を削除する手順とを含み、

前記追加ベクトルが、音声合成の際に分布情報を用いられない音声合成パラメータのベ

クトルであることを特徴とする音声合成装置の予測モデル学習方法。
【請求項９】
前記標準ベクトルが、メルケプストラム係数の特徴ベクトルであり、前記追加ベクトルがLSP係数の特徴ベクトルであることを特徴とする請求項８に記載の予測モデル学習方法。
【請求項１０】
前記標準ベクトルが、所定の音声合成パラメータに関する所定の時間長の特徴ベクトルであり、前記追加ベクトルが、前記所定の音声合成パラメータに関して前記所定の時間長の前後少なくとも一方に連続する時間長部分の特徴ベクトルであることを特徴とする請求項８に記載の予測モデル学習方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声合成システムならびにその予測モデル学習方法および装置に係り、特に、多数の音声データを学習させた決定木で予測モデルを構築し、入力テキストに対応する音声合成パラメータ時系列を予測して音声を合成する音声合成システムならびにその予測モデル学習方法および装置に関する。
【背景技術】
【０００２】
音声合成技術の代表的な利用例として、任意のテキストを自動的に音声に変換するTTS(Text-To-Speech) システムが知られている。TTSシステムでは、入力されたテキストから自然言語解析処理により音素系列データを生成し、この音素系列データから音声波形生成のためのパラメータ（音声合成パラメータ）の時系列データを生成する処理（以下、音声合成パラメータ時系列データ生成処理と表現する）が必要となる。音声合成パラメータの時系列データからは、信号処理や事前音声素片蓄積に対する素片選択および接続処理により音声波形が生成される。
【０００３】
ここで、音素とは便宜的に用いる用語で、音声学的な定義による必要はなく、時間軸方向に音声を区分する何らかの統一された単位の総称である。加えて、その出現環境、例えば先行・後続の音素の種類、韻律的特徴、TTSシステムにおける入力テキスト中で対応する個所の言語情報等を区別した非常に細かい音素分類が行われる。このような出現環境は、一般にコンテキストと呼ばれる。
【０００４】
音声合成パラメータは、音声波形生成で必要な複数の特徴の組み合わせで表現されるが、一般的には、特定時刻におけるスペクトル情報、基本周波数情報、有声・無声切り替え情報等を連結して構成されるベクトルが用いられる。
【０００５】
スペクトル情報としては、例えばそれ自身も多次元のベクトルであるメルケプストラム係数やLSP（Linear Spectrum Pairs）係数が用いられる。そして、各時刻の音声合成パラメータのベクトルを１フレームとし、それを５msといった一定時間間隔で並べたものを波形生成のための音声合成パラメータ時系列データとしている。
【０００６】
音声合成パラメータ時系列データ生成処理の実現では、単純には、予め各音素に対応した音声合成パラメータ時系列データを準備しておき、入力された音素情報系列の各音素に対応する音声合成パラメータ時系列データを連結し、それを出力とすれば良い。
【０００７】
しかしながら、実際の音声合成処理ではコンテキストを考慮した非常に細かい音素分類を行うことから、全ての音素の種類に対応した音声合成パラメータ時系列データを事前に準備しておくことは不可能である。そこで、実際には各音素の情報から音声合成パラメータ時系列データを予測する処理が必要となる。

【０００８】
例えば、隠れマルコフモデル（HMM：Hidden Markov Model）に基づくHMM音声合成では、各音素がHMMでモデル化される。より具体的には、１音素を時間方向に５状態程度に分割し、各状態内では定常な音声合成パラメータが出力されるというモデルを置き、HMMのパラメータである、音素内の状態遷移確率、および各状態における音声合成パラメータ（代表的には、平均ベクトルおよび分散共分散行列）の出力分布を、実際の音声データから予め求めている。ここで、出力分布のモデルとしては、正規分布が広く用いられている。このようなHMMのパラメータの推定には「Baum-Welchアルゴリズム」を利用できる。
【０００９】
しかしながら、必要なすべての種類の音素に対してこの処理を事前に行っておくことは不可能なため、音素情報から対応するHMMの各パラメータを予測する処理を行うことで、全ての種類の音素に対して適当なHMMを得ている。
【００１０】
この予測では、予めHMMの音声データに含まれている限られた種類の音素から学習したHMMを用いて、そのHMMの音素のコンテキストと、HMMの各パラメータの関係をモデル化するような決定木を予測モデルとして構築しておく。この際、決定木のリーフノードには、予測値となるHMMのパラメータの値を結びつけておく。そして、構築された予測モデルの決定木に対して、入力されたコンテキストが、決定木の各ノードでそれぞれ分割されたコンテキスト空間のいずれに属するかを選択する処理を、ルートノードからリーフノード方向に繰り返し行い、最終的にリーフノードに結び付けられた値を得ることで、任意のコンテキストに対して、音声合成パラメータをモデル化したHMMのパラメータの予測値を得ることができる。
【００１１】
ただし、実際には少ない音声データで予測モデルを構築するために、スペクトル情報や基本周波数といった音声合成パラメータの種類ごとに異なるベクトルとして扱い、それぞれの種類ごとに異なる予測モデルを作成し、それぞれ用いる方法が用いられる。
【００１２】
また、HMMの状態間のパラメータの不連続を抑えるために、各時刻のパラメータ時系列データ単独（以下、静的特徴という）だけでなく、その一階差分（傾きに相当）や二階差分（傾きの変化に相当）の系列（以下、動的特徴という）を音声合成パラメータとして追加する方法も用いられる。
【００１３】
一般に出力分布の平均値を出力するのがモデル上で確率最大となるため、確率最大の基準で静的特徴しか考慮しないと、HMMの一状態内では同じ値が出力される。この場合、最終的な音声合成パラメータ時系列データでは、HMMの状態が切り替わった際に、出力される値が大きく変化する。すなわち、時系列データが段状となるので、品質劣化の原因となる。
【００１４】
これに対して、非特許文献２には、HMMで静的特徴に加えて動的特徴もモデル化しておき、静的特徴と動的特徴の双方を考慮した確率最大の基準で音声合成パラメータ時系列データを求めることにより、HMMの状態が切り替わった場合でも、動的特徴のモデルが制約となって急激な値の変化が抑えられ、滑らかな時系列データを得られる技術が開示されている。
【００１５】
また、予測モデルの構築では、その木構造を大きくし過ぎると、各リーフノードに対応付けられる出力分布が少ないデータのサンプル数から推定されることになって分布の信頼性が下がり、予測精度が逆に低下してしまう。このため、実際には状態の分割をある程度の段階で止める必要がある。非特許文献１には、分割を停止させる基準としてMDL（最小記述長）を用いる技術が開示されている。
【先行技術文献】

【非特許文献】
【００１６】
【非特許文献１】吉村貴克、徳田恵一、益子貴史、小林隆夫、北村正、「HMMに基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化」、電子情報通信学会論文誌(D-II), J83-D-II, 11, pp.2099-2107, Nov.2000.
【非特許文献２】益子貴史、徳田恵一、小林隆夫、今井 聖、「動的特徴を用いたHMMに基づく音声合成」、電子情報通信学会論文誌(D-II), J79-D-II, 12, pp.2184-2190, Dec. 1996.
【非特許文献３】戸田智基、徳田恵一、「HMM音声合成のための系列内変動を考慮した音声パラメータ生成アルゴリズム」、電子情報通信学会技術報告, SP2005-52, pp. 1-6, Aug. 2005.
【発明の概要】
【発明が解決しようとする課題】
【００１７】
入力音素情報から対応する音声合成パラメータを予測するための予測モデルの構築では、予測モデルが出力する音声合成パラメータの値の分布を評価規準に用いていた。したがって、出力する音声合成パラメータへの影響が相対的に小さいコンテキストの差異は合成音声に反映されにくくなる。
【００１８】
しかしながら、コンテキスト空間の分割の観点で考えると、音声合成に用いる音声合成パラメータの値で最適化することが、最も適切な分割にはならない可能性がある。一般に音声合成パラメータには、主観的な品質に影響する様々なコンテキストの影響が含まれているが、例えば、主観的には差異が大きいが、音声合成パラメータの値として見ると、他のコンテキストの影響で生じる差異よりも相対的に値の変化が小さい、といったコンテキストの影響は、適切に取り扱うことができない。
【００１９】
また、例えばHMM音声合成のパラメータの予測を行う場合、基本的にHMMの状態に対応した短時間の特徴分布のみを考慮したクラスタリングが行われる。動的特徴を考慮することで、例えば前後１フレームといったような、静的特徴よりも長時間の特徴変化を考慮できるが、非特許文献２に記載されたアルゴリズムからも明らかなように、考慮するフレーム数を長く取るほど、音声合成パラメータ時系列データの計算コストが増加してしまう。
【００２０】
一方、音素の分類では、前後音素やさらにその前後の音素を考慮する等、音素HMMの出力フレーム周期や動的特徴の導入により考慮される区間よりも、時間的に長い区間の特徴として表れるコンテキストが考慮されている。
【００２１】
出力の短時間分布の分散には、データの揺らぎによる影響だけでなく、コンテキストに依存した、より長時間の時間変化の影響が含まれている可能性がある。しかしながら、短時間の特徴のみを考慮したクラスタリングでは両者を区別することが難しく、コンテキスト空間が適切に分割されない可能性があった。
【００２２】
この場合、例えば本来は独立した２つのクラスタとすべきものが１つのクラスタになるといった、不適切なクラスタリングが行われる可能性が高くなる。これは結果的に予測誤差を増加させ、最終的な合成音声の品質低下の原因となってしまう。
【００２３】
非特許文献３には、予測モデルとは別にパラメータ時系列の長時間変動に関するパラメータを求めておいて、最終的なパラメータ時系列計算の際に、そのパラメータも考慮した計算を行う方法が開示されている。しかしながら、そのような考慮を行うと、パラメータ時系列を演繹的に求めることができなくなり、逐次近似が必要となってしまう。
【００２４】

本発明の目的は、上記の技術課題を解決し、主観的な品質に影響を与えるコンテキストを決定木に反映させることで予測モデルの精度を実用において高めることにより、音声合成パラメータ時系列データの計算コストを増加させることなく、最終的な合成音声の品質を向上させることができる音声合成システムならびにその予測モデル学習方法および装置を提供することにある。

【課題を解決するための手段】

【００２５】

上記の目的を達成するために、本発明は、音声合成システムならびにその予測モデル学習方法および装置において、以下の構成を具備した点に特徴がある。

【００２６】

(1) 本発明の予測モデル学習装置は、音声データから複数種の音声合成パラメータを抽出する手段と、一の音声合成パラメータから生成した標準ベクトルおよび他の一の音声合成パラメータから生成した追加ベクトルに基づいて拡張ベクトルを生成する手段と、拡張ベクトルを音素ごとにモデル化する手段と、音素モデルの集合に対して、その拡張ベクトルを評価規準としてモデル尤度が最大となる分割条件をノード毎に決定することを繰り返し、各リーフノードに各音声合成パラメータの分布情報が登録された決定木を構築する手段と、決定木の各リーフノードから前記追加ベクトルに対応した分布情報を削除する手段とを具備した。

【００２７】

(2) 本発明の音声合成システムは、前記予測モデル学習装置に加えて、決定木のリーフノードに前記標準ベクトルに対応した分布情報のみが残った決定木を用いて音声合成を行う音声合成装置を具備した。

【００２８】

(3) 本発明の予測モデル学習方法は、音声データから複数種の音声合成パラメータを抽出する手順と、一の音声合成パラメータに基づいて標準ベクトルを生成する手順と、他の一の音声合成パラメータに基づいて追加ベクトルを生成する手順と、標準ベクトルおよび追加ベクトルに基づいて拡張ベクトルを生成する手順と、拡張ベクトルを音素ごとにモデル化する手順と、音素モデルの集合に対して、その拡張ベクトルを評価規準としてモデル尤度が最大となる分割条件をノード毎に決定することを繰り返し、各リーフノードに各音声合成パラメータの分布情報が登録された決定木を構築する手順と、決定木の各リーフノードから前記追加ベクトルに対応した分布情報を削除する手順とを備えた。

【００２９】

(4) 標準ベクトルとして、メルケプストラム係数の特徴ベクトルを採用し、追加ベクトルとして、LSP係数の特徴ベクトルを採用した。

【００３０】

(5) 標準ベクトルとして、所定の音声合成パラメータに関する所定の時間長の特徴ベクトルを採用し、追加ベクトルとして、所定の音声合成パラメータに関して所定の時間長の前後少なくとも一方に連続する時間長部分の特徴ベクトルを採用した。

【発明の効果】

【００３１】

本発明によれば、以下のような効果が達成される。

【００３２】

(1)請求項１，８の発明によれば、予測部の構築におけるコンテキスト空間の分割において、予測部が出力する音声合成パラメータでは小さい変化しか生じさせないが、別の種類の音声合成パラメータでは大きい変化として表れるコンテキストの違いを捉えることができる。予測部が出力する音声合成パラメータのみに注目する場合と比較し、予測部が出力する種類の音声合成パラメータの小さい変化がコンテキストの差異に由来するものなのか、あるいは単なるデータの揺らぎによるものなのかを分離できるので、より適切なコンテキスト空間の分割が得られ、予測モデルの精度を実用において高めることができる。

【００３３】

また、予測モデルの決定木が、当該予測モデルが直接出力しない音声合成パラメータも考慮して学習されるので、音声合成には使わないが主観的な品質との相関の高い音声合成パラメータを決定木学習時に考慮することで、主観的な品質に影響を与えるコンテキストの影響を決定木に反映させ、予測モデルの精度を実用において高めることができる。

【００３４】

(2) 請求項４，５の発明によれば、音声合成処理時のパラメータ時系列データの計算処理は従来技術と同じなので、音声合成パラメータ時系列データの計算コストを増加させることなく、最終的な合成音声の品質を向上させることができる。

【００３５】

(3) 請求項２，６，９の発明によれば、メルケプストラム係数に基づく標準ベクトルとLSP係数に基づく追加ベクトルとを連結した拡張ベクトルを規準にしてコンテキストクラスタリングを実施し、最終的な音声合成パラメータはメルケプストラム係数のみに限定するので、LSP係数を考慮したクラスタリングとLSP係数を考慮しない低演算量な音声合成とを両立できるようになる。したがって、LSP係数により捉えられる音声の特徴の差を反映させつつ、音声合成時には雑音発生抑制の処理が不要となり、また、スペクトル強調も容易な音声合成が可能となる。

【００３６】

(4) 請求項３，７，１０の発明によれば、少ない計算コストで、長時間特徴の影響を反映させた音声合成パラメータ時系列データを得ることができる。

【図面の簡単な説明】

【００３７】

【図１】本発明の一実施形態に係る予測モデル学習装置の主要部の構成を示した機能ブロック図である。

【図２】本発明の一実施形態に係る音声合成システムの主要部の構成を示した機能ブロック図である。

【図３】拡張ベクトルを評価規準としてコンテキストクラスタリング用の決定木を構築する手順を示したフローチャートである。

【図４】拡張ベクトルを評価規準としてコンテキストクラスタリング用の決定木を構築する様子を模式的に示した図である。

【図５】標準ベクトルに、追加した時間長部分のフレームに相当する追加ベクトルを連結して拡張ベクトルを構成する方法を模式的に示した図である。

【発明を実施するための形態】

【００３８】

以下、図面を参照して本発明の実施の形態について詳細に説明する。図１は、本発明の一実施形態に係る予測モデル学習装置１０の主要部の構成を示した機能ブロック図であり、図２は、この予測モデル学習装置１０に音声合成装置３０を追加した音声合成システム１の主要部の構成を示した機能ブロック図である。

【００３９】

初めに図１の予測モデル学習装置１０を参照し、音声データベース１０１には、音素別にコンテキスト依存でラベル付けされた多数の音声データが記憶されている。音声合成パラメータ抽出部１０２は、基本周波数(F0)抽出部１０２ａおよびスペクトル情報算出部１０２ｂを含み、音声データから複数種の音声合成パラメータを抽出、計算する。

【００４０】

前記基本周波数(F0)抽出処理部１０２ａは、音声データベース１０１に記憶された音声のフレームごとに基本周波数(F0)を抽出する。前記スペクトル情報算出部１０２ｂは、音声データベース１０１に記憶された音声のフレームごとに、例えばMFCC (Mel Frequency Cepstrum Coefficient)やLSP (line spectral pairs)などのスペクトル情報を算出する。

【００４１】

特徴ベクトル生成部１０３は、スペクトル情報として、例えばメルケプストラム係数の

ベクトル(静的特徴)並びにその動的特徴である1階差分ベクトルおよび2階差分ベクトルを連結してスペクトル特徴ベクトル（以下、標準ベクトルx）を生成する。同様に、対数基本周波数の値(静的特徴)並びにその動的特徴である1階差分ベクトルおよび2階差分ベクトルを連結してF0特徴ベクトル（以下、標準ベクトルy）を生成する。

【００４２】

前記特徴ベクトル生成部１０３のベクトル拡張部１０３ａは、前記音声合成パラメータ抽出部１０２が音声データから抽出した複数種の音声合成パラメータのうち、前記スペクトル情報および対数基本周波数とは異なる他の音声合成パラメータに関して、それぞれ追加用のベクトル（以下、追加ベクトル）x', y'を同様に生成する。

【００４３】

前記ベクトル拡張部１０３ａはさらに、前記標準ベクトルxに追加ベクトルx'を連結して拡張ベクトルXを生成し、前記標準ベクトルyに追加ベクトルy'を連結して拡張ベクトルYを生成する。これらの拡張ベクトルX, Yは、コンテキストクラスタリング用の決定木を構築する際の評価規準として用いられる。

【００４４】

以下、拡張ベクトルの生成方法について説明する。音声合成パラメータを構成する各特徴パラメータについて、その時間変化に関する要素を含むフレームiの標準ベクトルxiは、適宜の行列Wを用いて次式(1)で表せる。

【００４５】

【数１】

$$\mathbf{x}_i = \mathbf{W}\mathbf{c}_i \quad \cdots(1)$$

【００４６】

ただし、ciは行列Wの列数と等しい次元の、フレームiを中心とする特徴パラメータの時系列データで構成されるベクトルであり、行列Wが次式(2)で与えられるとき、標準ベクトルxiは次式(3)で与えられる。標準ベクトルxiの各要素は、フレームｉの特徴パラメータの元の値（静的特徴）、その一階差分および二階差分（動的特徴）となる。

【００４７】

【数２】

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 \\ -0.5 & 0 & 0.5 \\ 1 & -2 & 1 \end{pmatrix} \quad \cdots(2)$$

【００４８】

【数３】

$$\mathbf{x}_i = \mathbf{W}\begin{pmatrix} c_{i-1} & c_i & c_{i+1} \end{pmatrix}^T \quad \cdots(3)$$

【００４９】

ここで、異なる種類の特徴パラメータの値diから同様に生成される追加ベクトルx'iを次式(4)で表すものとする。

【００５０】

【数４】

$$\mathbf{x}_i = \mathbf{W}'\mathbf{d}_i \quad \cdots(4)$$

【００５１】

そして、本実施形態では次式(5)で計算される拡張ベクトルXiに対するモデル尤度をクラスタリングの評価規準に用いる。

【００５２】

【数５】

$$\mathbf{X}_i = \left(\mathbf{x}_i \quad \mathbf{x}'_i\right)^{\mathrm{T}} \qquad \cdots(5)$$

【００５３】

この際、モデルの尤度関数において追加ベクトルのx'iの影響を標準ベクトルxiの影響よりも小さくする、あるいは大きくする方法を用いることができる。モデルに多次元正規分布を仮定したとき、Xiに対するモデルの対数尤度は一般に次式(6)で与えられる。

【００５４】

【数６】

$$-\frac{(n+n')\ln 2\pi + \ln|\Sigma|}{2} - \frac{(\mathbf{X}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}_i-\boldsymbol{\mu})}{2} \qquad \cdots(6)$$

【００５５】

ただし、nおよびn'はそれぞれ標準ベクトルの次元数、追加ベクトルの次元数で、μとΣはモデルの平均ベクトルおよび分散共分散行列である。これに対し、決定木学習では対数尤度関数として上式の代わりに次式(7)を用いても良い。

【００５６】

【数７】

$$-\frac{(n+n')\ln 2\pi + \ln|\Sigma|}{2} - \frac{(\mathbf{X}_i-\boldsymbol{\mu})^T\mathbf{S}^T\Sigma^{-1}\mathbf{S}(\mathbf{X}_i-\boldsymbol{\mu})}{2} \qquad \cdots(7)$$

$$\mathbf{S} = diag\left(\underbrace{1 \quad \cdots \quad 1}_{n} \quad \underbrace{\alpha \quad \cdots \quad \alpha}_{n'}\right)$$

【００５７】

ここで、αは追加ベクトルの成分に対する重み係数であり、αを１以下にすると、決定木学習における追加ベクトルの影響がより小さくなる。

【００５８】

なお、クラスタリング結果に結び付ける音声合成パラメータの分布情報（ここでは、平均ベクトルおよび分散共分散行列）は標準ベクトルxiの分布のみとし、最終的な音声合成パラメータ時系列データの計算では行列Wを考慮する。

【００５９】

コンテキスト依存HMM学習部１０４は、音声データベース１０１に記憶された音声データを音素ごとにHMMでモデル化し、フレームごとにHMM状態の集合を入力として、前記拡張ベクトルXiを評価規準としてモデル尤度が最大となる分割条件をノード毎に決定することを繰り返し、各リーフノードに各音声合成パラメータの分布情報が登録された決定木を構築する。

【００６０】

分布情報編集部１０５は、前記決定木の各リーフノードから前記追加ベクトルx'i（y'i）に対応した分布情報を削除する。編集後の学習結果は決定木としてHMM記憶部２０に蓄積される。

【００６１】

図３は、HMM学習部１０４が前記拡張ベクトルを評価規準としてコンテキストクラスタリング用の決定木を構築する手順を示したフローチャートであり、図４は、その様子を模式的に示した図である。

【００６２】

ステップS１では、弁別素性などに基づいて予め用意された音韻に関する分割条件の質問集合Qが取得される。ステップS２では、音声データの音素系列状態の全てを包含する拡張ベクトル集合Sがルートノードに割り当てられる。

【００６３】
ステップＳ３では、リーフノードの集合から、その1つが選択される。なお、一番最初の状態では、ルートノードが唯一のリーフノードとなる。このリーフノードの選択では、例えばそのノードに結び付けられたモデルの平均尤度が最も小さい、すなわちモデルと実際のデータが最も合っていないリーフノードを選べばよい。
【００６４】
ステップＳ４では、質問集合Qから今回の質問qiが選択され、当該質問qiで前記集合SがSq+，Sq-に２分割される。ステップＳ５では、今回の分割前後におけるモデル尤度が前記拡張ベクトルX（Y）を評価規準として計算される。
【００６５】
ステップＳ６では、全ての質問による分割結果に対して前記評価計算が終了したか否かが判定される。終了していなければステップＳ４へ戻り、質問を残りの他の質問に切り替えながら分割及び評価計算が繰り返される。
【００６６】
ステップＳ７では、モデル尤度の上昇が最も大きい最尤の質問が選択されて分割対象のノードに割り当てられ、当該質問によりノードが２つのリーフノードに分割される。このとき、元の分割対象のノードは中間ノードになる。
【００６７】
ステップＳ８では、例えばMDL（最小記述長）基準に基づいて分割を終了するか否かが判定され、分割停止条件が満足されるまでは、ステップＳ３へ戻って上記の各処理が繰り返される。分割停止要件が満足されるとステップＳ９へ進み、決定木の各リーフノードから、前記追加ベクトルx'i（y'i）に対応した音声合成パラメータの分布情報が、前記分布情報編集部１０５により削除される。
【００６８】
以上の処理により、ルートノードおよび各中間ノードには、拡張ベクトルを反映した分割条件が紐付けられる。このとき、末端の各リーフノードには、標準ベクトルxに関する分布情報と追加ベクトルx'に関する分布情報とが登録されることになるが、本実施形態では、追加ベクトルx'に関する分布情報が削除され、標準ベクトルxに関する分布情報のみが対応付けられる。これらの学習結果は、拡張ベクトルXで最適化された木構造としてHMM記憶部２０に記憶される。
【００６９】
図２を参照し、音声合成装置３０において、テキスト解析部３０１は、入力テキストに対して自然言語解析を行ない、合成音声が持つべき韻律情報等が付されたコンテキスト依存の音素ラベル列を出力する。パラメータ生成部３０２は、前記音素ラベル列の音素ごとに、そのコンテキストに対応した決定木をHMM記憶部２０から選択し、当該各決定木に各コンテキストを適用することにより、最も適合したHMMを抽出、連結することにより、音声合成用のスペクトル情報系列および対数基本周波数系列を生成する。
【００７０】
音源生成部３０３は、対数基本周波数系列に基づいて音源信号を生成する。合成フィルタ３０４は、パラメータ生成部３０２により生成されたスペクトル情報系列に基づいて、音源生成部３０３により生成された音源信号をフィルタリングすることにより合成音声信号を生成する。
【００７１】
本実施形態によれば、音声合成に用いる音声合成パラメータのみならず、音声合成に用いない音声合成パラメータをも考慮して決定木が構築される。これにより、音声合成に用いる音声合成パラメータ上では値の変化が小さいが、主観品質との相関が高い別の種類の音声合成パラメータでは大きな値の変化として容易に捉えることができるコンテキストの影響を、決定木の構造に反映できる。
【００７２】
このように、本実施形態によれば、主観的な品質に影響する様々なコンテキストの影響

を合成音声に反映させることができ、その結果、入力テキストのコンテキストにより適した合成音声を出力できるようになる。

【００７３】

なお、上記の実施形態では、決定木が出力しない、換言すれば音声合成に用いられない音声合成パラメータのベクトルを追加して拡張ベクトルXを生成するものとして説明したが、このような追加ベクトルとしては、例えば標準ベクトルがメルケプストラム係数から計算されるスペクトル情報であれば、LSP係数から計算されるベクトルを採用できる。

【００７４】

一般に、メルケプストラム係数よりもLSP係数の方がスペクトルの急峻なピークを捉えられるが、LSP係数に基づく音声合成では、隣接するLSP係数の値が交差すると合成フィルタが不安定になって雑音が発生する。このため、LSP係数を特徴ベクトルとして採用する際には、このような現象が生じないようにするための追加の処理が必要となる。

【００７５】

また、合成音声の明瞭性を高めるためのスペクトル強調処理に関しても、メルケプストラム係数に対しては、０次以外の係数を定数倍するだけで比較的簡単に行えるのに対して、LSP係数に対しては、より複雑な処理が必要となる。したがって、LSP係数をスペクトル情報として採用すると音声合成処理が複雑化してしまう。

【００７６】

これに対して、本実施形態によれば、クラスタリングの際は、メルケプストラム係数に基づく標準ベクトルxiとLSP係数に基づく追加ベクトルx'iとを連結した拡張ベクトルXiを規準にしてコンテキストクラスタリングを実施し、最終的な音声合成パラメータはメルケプストラム係数のみに限定するので、LSP係数を考慮したクラスタリングとLSP係数を考慮しない低演算量な音声合成とを両立できるようになる。

【００７７】

すなわち、LSP係数により捉えられる音声の特徴の差を反映させつつ、音声合成時には雑音発生抑制の処理が不要となり、また、スペクトル強調も容易に分離された音声合成が可能となる。この際、W'はWと同じ行列であっても良いし、異なる行列であっても良い。

【００７８】

さらに、前記拡張ベクトルXを生成するための追加ベクトルx'としては、決定木が直接出力する音声合成パラメータの時間長を超える時間長部分の特徴ベクトルを採用することができる。

【００７９】

この場合には、前記行列Wよりも長時間の影響を考慮した、すなわち列幅の大きい行列W''を置き、次式(8)で計算される拡張ベクトルxi''をクラスタリングの評価規準に用いる。

【００８０】

【数８】

$$\mathbf{x}_i'' = \left(\mathbf{W}\mathbf{c}_i \quad \mathbf{W}''\mathbf{c}_i''\right)^{\mathrm{T}} \left(= \left(\mathbf{x}_i \quad \alpha''\mathbf{y}_i''\right)^{\mathrm{T}}\right) \quad \cdots(8)$$

【００８１】

ただし、ci"は行列W''の列数と等しい次元の、フレームｉを中心とする、ciと同じ特徴のパラメータの時系列で構成されるベクトルである。例えば、連続する５フレームの時間変化に関する特徴を生成する行列として、次式(9)の行列W"が挙げられる。

【００８２】

【数９】

$$\mathbf{W}'' = \begin{pmatrix} -1 & -\sqrt{3}/2 & 0 & \sqrt{3}/2 & 1 \\ -1 & 0 & 1 & 0 & -1 \end{pmatrix} \quad \cdots(9)$$

【００８３】

図５は、標準ベクトルxiに２フレーム分の音声合成パラメータを追加ベクトルx''iとして連結して拡張ベクトルXiの構成する方法を模式的に示した図であり、標準ベクトルxiは連続する３つのフレームDt2, Dt3, Dt4の特徴量から構成されるのに対して、拡張ベクトルXiでは、その前後にフレームDt1, Dt5が更に連結されている。すなわち、拡張ベクトルXiはその要素として標準ベクトルxiの要素を全て含んでいる。

【００８４】

このような時間長の長い拡張ベクトルXiを用いれば、従来の標準ベクトルxiのみを用いたクラスタリングが、連続する３フレームの変化の特徴しか反映できないのに対し、連続する５フレームの変化の特徴を反映させることができる。

【００８５】

一方、クラスタリング結果に結び付ける特徴パラメータの分布情報は、標準ベクトルxiに対応した分布情報のみとし、最終的な音声合成パラメータ時系列データの計算では、行列Wのみを考慮する。

【００８６】

長時間変化の影響は、予測モデルが予測する合成パラメータの分布自体は例の場合はHMMの状態単位で切り替わるが、標準ベクトルxiではなく追加ベクトルxi''を考慮した拡張ベクトルXiでクラスタリングを行うことにより、長時間変化が大きく異なる場合は異なるクラスタにクラスタリングされるので、予測モデルでは長時間変化の影響も含めて予測できる。

【００８７】

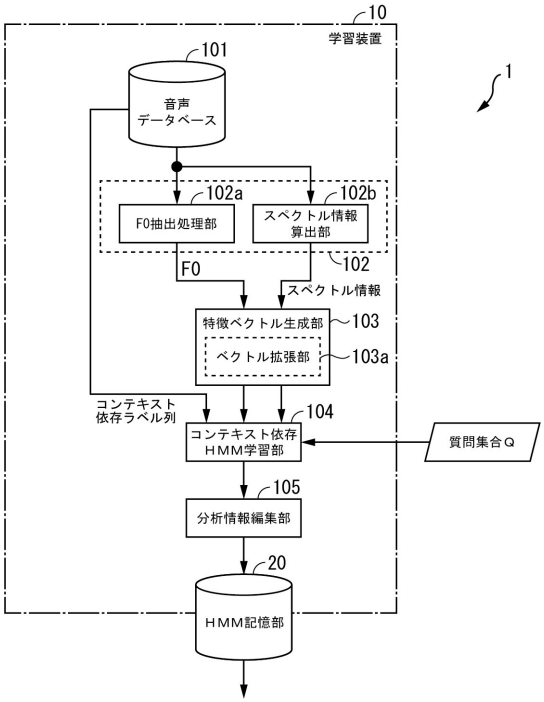
これにより、従来の行列Wより大きな行列W''を、最終的な音声合成パラメータ時系列データの生成処理では考慮する必要が無いので、少ない計算コストで、長時間特徴の影響を反映させた音声合成パラメータ時系列データを得ることができる。

【符号の説明】

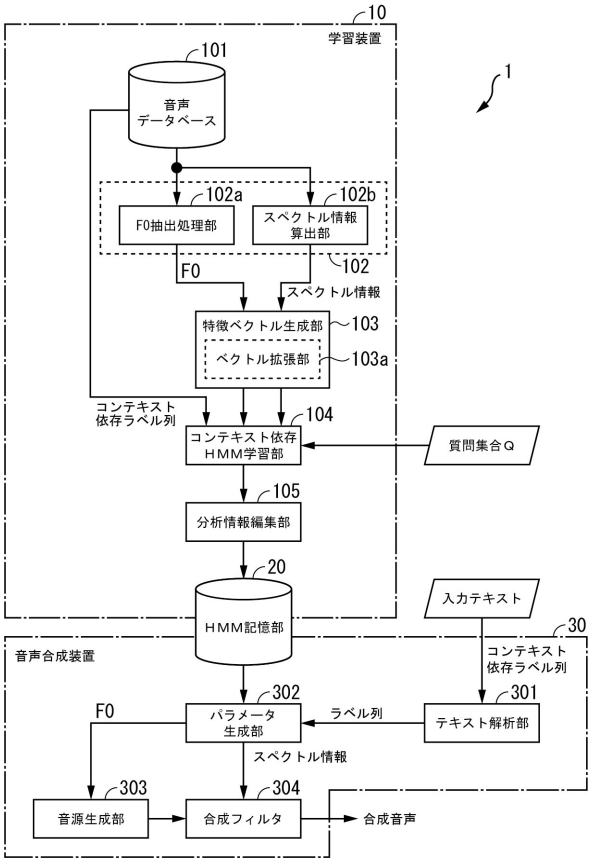
【００８８】

１…音声合成システム，１０１…学習装置，２０…HMM記憶部，３０…音声合成装置，１０１…音声データベース，１０２…音声合成パラメータ抽出部，１０２ａ…基本周波数(F0)抽出処理部，１０２ｂ…スペクト情報算出部，１０３…特徴ベクトル生成部，１０３ａ…特徴ベクトル拡張部，１０４…コンテキスト依存HMM学習部，１０５…分布情報編集部，３０１…テキスト解析部，３０２…パラメータ生成部，３０３…音源生成部，３０４…合成フィルタ

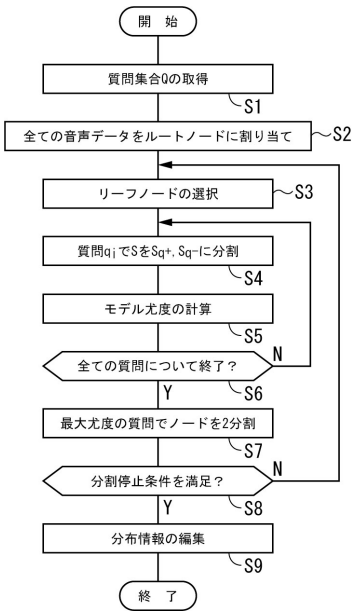
【図１】


10
学習装置
1
101
音声
データベース
102a
F0抽出処理部
102b
スペクトル情報
算出部
102
F0
スペクトル情報
特徴ベクトル生成部
103
ベクトル拡張部
103a
コンテキスト
依存ラベル列
104
コンテキスト依存
HMM学習部
質問集合Q
105
分析情報編集部
20
HMM記憶部

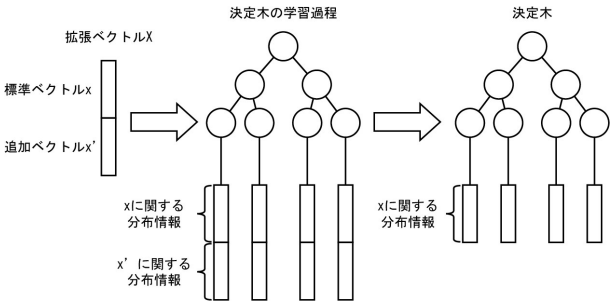
【図２】


10
学習装置
1
101
音声
データベース
102a
F0抽出処理部
102b
スペクトル情報
算出部
102
F0
スペクトル情報
特徴ベクトル生成部
103
ベクトル拡張部
103a
コンテキスト
依存ラベル列
104
コンテキスト依存
HMM学習部
質問集合Q
105
分析情報編集部
20
HMM記憶部
入力テキスト
30
コンテキスト
依存ラベル列
音声合成装置
302
パラメータ
生成部
F0
ラベル列
301
テキスト解析部
スペクトル情報
303
音源生成部
304
合成フィルタ
合成音声

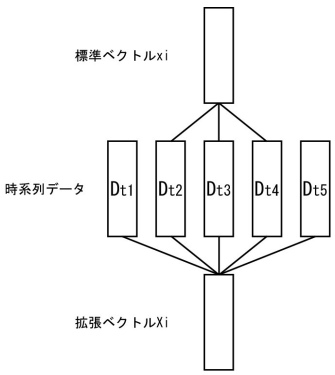
【図３】


開　始
質問集合Qの取得
S1
全ての音声データをルートノードに割り当て
S2
リーフノードの選択
S3
質問qjでSをSq+, Sq-に分割
S4
モデル尤度の計算
S5
全ての質問について終了？
N
Y
S6
最大尤度の質問でノードを2分割
S7
分割停止条件を満足？
N
Y
S8
分布情報の編集
S9
終　了

【図４】


決定木の学習過程
決定木
拡張ベクトルX
標準ベクトルx
追加ベクトルx'
xに関する
分布情報
x'に関する
分布情報
xに関する
分布情報

【図５】


標準ベクトルxi
時系列データ
Dt1
Dt2
Dt3
Dt4
Dt5
拡張ベクトルXi

フロントページの続き

(56)参考文献 特開２０１０－２３７３２３（ＪＰ，Ａ）
特開２０１２－５８３４３（ＪＰ，Ａ）
特開２０１４－５６２３５（ＪＰ，Ａ）
米国特許出願公開第２０１４／０３４３９３４（ＵＳ，Ａ１）
能勢隆他，HMM音声合成のための動的特徴量を用いた音素継続長モデリングの検討，電子情報通信学会技術研究報告，２０１１年１２月，Vol.111，No.364，p.197-202

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １３／００－９９／００