(19)日本国特許庁(JP)

# (12)特許公報(B2)

(11)特許番号
特許第7362976号
(P7362976)

(45)発行日　令和5年10月18日(2023.10.18)

(24)登録日　令和5年10月10日(2023.10.10)

(51)国際特許分類

| | | | | |
|---|---|---|---|---|
| *Ｇ１０Ｌ* | *13/06* | *(2013.01)* | Ｇ１０Ｌ | 13/06　１２０Ｚ |
| *Ｇ１０Ｌ* | *13/10* | *(2013.01)* | Ｇ１０Ｌ | 13/10　１１１Ｆ |
| *Ｇ１０Ｌ* | *25/30* | *(2013.01)* | Ｇ１０Ｌ | 13/10　１１１Ｅ |
| | | | Ｇ１０Ｌ | 13/10　１１１Ａ |
| | | | Ｇ１０Ｌ | 13/10　１１１Ｃ |

ＦＩ

請求項の数　6　（全27頁）　最終頁に続く

(21)出願番号　特願2022-49374(P2022-49374)
(22)出願日　令和4年3月25日(2022.3.25)
(62)分割の表示　特願2018-113433(P2018-113433)の分割
原出願日　平成30年6月14日(2018.6.14)
(65)公開番号　特開2022-81691(P2022-81691A)
(43)公開日　令和4年5月31日(2022.5.31)
審査請求日　令和4年3月25日(2022.3.25)

(73)特許権者　000004352
日本放送協会
東京都渋谷区神南２丁目２番１号
(73)特許権者　399060908
一般財団法人ＮＨＫ財団
東京都世田谷区用賀四丁目１０番１号
(74)代理人　100121119
弁理士　花村 泰伸
(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　栗原 清
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　熊野 正
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)【特許請求の範囲】

【請求項１】

音声合成対象のテキストを言語分析し、言語特徴量を求める言語分析部と、

前記言語分析部により求めた前記言語特徴量に、音響の特徴を調整するための調整パラメータの調整量情報を追加する調整量追加部と、

前記調整量追加部により前記調整量情報が追加された前記言語特徴量に基づき、予め学習された統計モデルを用いて、音響特徴量を推定する音響特徴量推定部と、

前記音響特徴量推定部により推定された前記音響特徴量に基づいて、音声信号を合成し、前記テキストに対して前記調整パラメータによる調整が加えられた音声信号を出力する音声生成部と、を備えた音声合成装置であって、

前記音響特徴量推定部が用いる統計モデルは、

予め設定されたテキストを言語分析し、学習言語特徴量を求める学習言語分析部と、

前記テキストに対応する音声信号を音響分析し、学習音響特徴量を求める音声分析部と、

前記学習言語分析部により求めた前記学習言語特徴量及び前記音声分析部により求めた前記学習音響特徴量を時間的に対応付ける対応付け部と、

前記対応付け部により対応付けられた前記学習言語特徴量に、音響の特徴を調整するための調整パラメータの調整量情報を追加する学習調整量追加部と、

前記対応付け部により対応付けられた前記学習音響特徴量を、前記調整パラメータの前記調整量情報に従って調整する学習音響特徴量調整部と、

前記学習調整量追加部により前記調整量情報が追加された前記学習言語特徴量及び前記

学習音響特徴量調整部により調整された前記学習音響特徴量を用いて、統計モデルを学習する学習部と、
を備えた学習装置によって、予め学習された統計モデルである
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、
前記統計モデルは、ニューラルネットワークで構成された時間長モデル及び音響モデルからなり、
前記音響特徴量推定部は、
前記時間長モデルを用いて、音素毎の前記言語特徴量を前記時間長モデルの入力データとして、前記時間長モデルの出力データである音素毎の時間長を推定し、
音素毎の前記時間長からフレーム毎の時間長を生成し、
前記音響モデルを用いて、フレーム毎の前記言語特徴量及びフレーム毎の前記時間長を入力データとし、前記音響モデルの出力データであるフレーム毎の前記音響特徴量を推定する、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、
前記調整パラメータを、話速または時間長、パワー、ピッチ、及び抑揚の４つのパラメータのうちのいずれか１つまたは２つ以上の組み合わせとする、ことを特徴とする音声合成装置。
【請求項４】
請求項１または２に記載の音声合成装置において、
前記調整パラメータを、話速または時間長、パワー、ピッチ、及び抑揚の４つのパラメータとし、
当該４つのパラメータのうちのいずれか１つのパラメータの調整量は、所定範囲内の任意の値が指定され、他の３つのパラメータの調整量は、固定値が用いられる、ことを特徴とする音声合成装置。
【請求項５】
請求項１または２に記載の音声合成装置において、
前記調整パラメータを、話速または時間長、パワー、ピッチ、及び抑揚の４つのパラメータとし、
当該４つのパラメータにおけるそれぞれの調整量は、それぞれの所定範囲内の任意の値が指定される、ことを特徴とする音声合成装置。
【請求項６】
コンピュータを、請求項１から５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、テキストから音声信号を合成するための統計モデルを用いて音声信号を合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、テキストとこれに対応する音声信号を用いて統計モデルを学習し、任意のテキストに対する合成音声を得る方法として、ディープニューラルネットワーク（ＤＮＮ：Deep Neural Network）を用いた深層学習（ＤＬ：Deep Learing）に基づく技術が知られている（例えば、非特許文献１を参照）。
【０００３】
一方、音声信号の読み上げ方を調整する方法として、音声分析生成処理に基づく技術が知られている（例えば、非特許文献２を参照）。

【０００４】
図１５は、非特許文献１に記載された従来の学習方法及び合成方法を示す説明図である。この学習方法を実現する学習装置は、事前に用意された音声コーパスのテキストとこれに対応する音声信号を用いて、テキストについては言語分析処理により言語特徴量を抽出する（ステップＳ１５０１）。また、学習装置は、音声信号について音声分析処理により音響特徴量を抽出する（ステップＳ１５０２）。
【０００５】
学習装置は、言語特徴量と音響特徴量の時間対応付けを行い（ステップＳ１５０３）、言語特徴量と音響特徴量を用いて統計モデルを学習する（ステップＳ１５０４）。
【０００６】
また、この合成方法を実現する音声合成装置は、任意のテキストを入力し、テキストの言語分析処理により言語特徴量を抽出する（ステップＳ１５０５）。そして、音声合成装置は、学習装置により学習された統計モデルを用いて、言語特徴量から音響特徴量を推定し（ステップＳ１５０６）、音声生成処理により、音響特徴量から音声信号波形を求める（ステップＳ１５０７）。これにより、任意のテキストに対応する合成音声信号を得ることができる。
【０００７】
図１６は、非特許文献２に記載された従来の音声信号調整方法を示す説明図である。この音声信号調整方法を実現する音声調整装置は、音声分析処理により、音声信号からフレーム毎の音響特徴量を抽出し（ステップＳ１６０１）、調整パラメータに基づいて、音響特徴量の所望の部分に所望の調整を加える（ステップＳ１６０２）。
【０００８】
音声調整装置は、音声生成処理により、調整が加えられたフレーム毎の音響特徴量から音声信号を生成する（ステップＳ１６０３）。これにより、調整を加えた音声信号を得ることができる。
【先行技術文献】
【非特許文献】
【０００９】
【文献】Zhizheng Wu, Oliver Watts, Simon King,“ Merlin：An Open Source Neural Network Speech Synthesis System”, in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA.
M. Morise, F. Yokomori, and K. Ozawa,“ WORLD：a vocoder-based high-quality speech synthesis system for real-time applications”, IEICE transactions on information and systems, vol. E99-D, no, 7, pp. 1877-1884, 2016
【発明の概要】
【発明が解決しようとする課題】
【００１０】
例えば、放送番組等のコンテンツ制作に合成音声信号を利用する際に、演出効果として、テキストの特定部分の読み上げ方を調整した合成音声信号が求められることがある。
【００１１】
前述の非特許文献１の方法は、任意のテキストに対して合成音声信号を得るものであり、同一のテキストに対して常に同一の合成音声信号が得られる。また、前述の非特許文献２の方法は、音声信号の読み上げ方を調整するものである。
【００１２】
そこで、テキストの特定部分の読み上げ方を調整した合成音声信号を求める方法として、前述の非特許文献１，２を組み合わせることが想定される。
【００１３】
図１７は、非特許文献１，２の従来技術を組み合わせた想定例を示す説明図である。この想定例の学習方法は、図１５に示したステップＳ１５０１～Ｓ１５０４と同様である（ステップＳ１７０１～Ｓ１７０４）。

【００１４】
この想定例の合成方法は、図１５に示したステップＳ１５０５～Ｓ１５０７の処理に、図１６に示したステップＳ１６０２の処理を挿入したものである。具体的には、音声合成装置は、任意のテキストから言語特徴量を抽出し（ステップＳ１７０５）、統計モデルを用いて言語特徴量から音響特徴量を推定する（ステップＳ１７０６）。
【００１５】
音声合成装置は、調整パラメータに基づいて、音響特徴量の所望の部分に所望の調整を加える（ステップＳ１７０７）。音声合成装置は、音声生成処理により、調整が加えられたフレーム毎の音響特徴量から音声信号を生成する（ステップＳ１７０８）。これにより、任意のテキストに対応する合成音声信号を得ることができる。
【００１６】
しかしながら、この想定例では、ステップＳ１７０６にて統計モデルを用いて言語特徴量から推定した音響特徴量は、実際の音声信号から音声分析処理により抽出した音響特徴量とは異なり、時間的に平滑化された特性を持っている。このため、ステップＳ１７０７にて統計モデルを用いて推定した音響特徴量に調整を加え、ステップＳ１７０８にて調整後のフレーム毎の音響特徴量から合成音声信号を得ると、合成音声信号に音質劣化を生じてしまう。
【００１７】
このように、図１７に示した想定例では、高品質の合成音声信号を得ることができないという問題があった。このため、テキストの特定部分の読み上げ方を調整した、高品質の合成音声信号を得るために、新たな手法が所望されていた。
【００１８】
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、テキストの特定部分の読み上げ方を調整した合成音声信号を生成する際に、高品質の合成音声信号を得ることが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】
【００１９】
前記課題を解決するために、請求項１の音声合成装置は、音声合成対象のテキストを言語分析し、言語特徴量を求める言語分析部と、前記言語分析部により求めた前記言語特徴量に、音響の特徴を調整するための調整パラメータの調整量情報を追加する調整量追加部と、前記調整量追加部により前記調整量情報が追加された前記言語特徴量に基づき、予め学習された統計モデルを用いて、音響特徴量を推定する音響特徴量推定部と、前記音響特徴量推定部により推定された前記音響特徴量に基づいて、音声信号を合成し、前記テキストに対して前記調整パラメータによる調整が加えられた音声信号を出力する音声生成部と、を備えた音声合成装置であって、前記音響特徴量推定部が用いる統計モデルは、予め設定されたテキストを言語分析し、学習言語特徴量を求める学習言語分析部と、前記テキストに対応する音声信号を音響分析し、学習音響特徴量を求める音声分析部と、前記学習言語分析部により求めた前記学習言語特徴量及び前記音声分析部により求めた前記学習音響特徴量を時間的に対応付ける対応付け部と、前記対応付け部により対応付けられた前記学習言語特徴量に、音響の特徴を調整するための調整パラメータの調整量情報を追加する学習調整量追加部と、前記対応付け部により対応付けられた前記学習音響特徴量を、前記調整パラメータの前記調整量情報に従って調整する学習音響特徴量調整部と、前記学習調整量追加部により前記調整量情報が追加された前記学習言語特徴量及び前記学習音響特徴量調整部により調整された前記学習音響特徴量を用いて、統計モデルを学習する学習部と、を備えた学習装置によって、予め学習された統計モデルであることを特徴とする。
【００２０】
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、前記統計モデルが、ニューラルネットワークで構成された時間長モデル及び音響モデルからなり、前記音響特徴量推定部が、前記時間長モデルを用いて、音素毎の前記言語特徴量を前記時間長モデルの入力データとして、前記時間長モデルの出力データである音素毎の時間長を

推定し、音素毎の前記時間長からフレーム毎の時間長を生成し、前記音響モデルを用いて、フレーム毎の前記言語特徴量及びフレーム毎の前記時間長を入力データとし、前記音響モデルの出力データであるフレーム毎の前記音響特徴量を推定する、ことを特徴とする。

【００２１】

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、前記調整パラメータを、話速または時間長、パワー、ピッチ、及び抑揚の４つのパラメータのうちのいずれか１つまたは２つ以上の組み合わせとする、ことを特徴とする。

【００２２】

また、請求項４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、前記調整パラメータを、話速または時間長、パワー、ピッチ、及び抑揚の４つのパラメータとし、当該４つのパラメータのうちのいずれか１つのパラメータの調整量は、所定範囲内の任意の値が指定され、他の３つのパラメータの調整量は、固定値が用いられる、ことを特徴とする。

【００２３】

また、請求項５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、前記調整パラメータを、話速または時間長、パワー、ピッチ、及び抑揚の４つのパラメータとし、当該４つのパラメータにおけるそれぞれの調整量は、それぞれの所定範囲内の任意の値が指定される、ことを特徴とする。

【００２４】

また、請求項６のプログラムは、コンピュータを、請求項１から５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】

【００２５】

以上のように、本発明によれば、テキストの特定部分の読み上げ方を調整した合成音声信号を生成する際に、高品質の合成音声信号を得ることが可能となる。

【図面の簡単な説明】

【００２６】

【図１】本発明の実施形態による学習装置の構成例を示すブロック図である。
【図２】学習装置による事前学習処理例を示すフローチャートである。
【図３】言語特徴量のデータ構成例を説明する図である。
【図４】音声分析部による音声分析処理例を示すフローチャートである。
【図５】音響特徴量のデータ構成例を説明する図である。
【図６】時間情報が追加された言語特徴量のデータ構成例を説明する図である。
【図７】調整量情報が追加された言語特徴量のデータ構成例を説明する図である。
【図８】時間長モデルの学習処理例を説明する図である。
【図９】音響モデルの学習処理例を説明する図である。
【図１０】本発明の実施形態による音声合成装置の構成例を示すブロック図である。
【図１１】音声合成装置による音声合成処理例を示すフローチャートである。
【図１２】時間長モデルを用いた時間長推定処理例を説明する図である。
【図１３】音響モデルを用いた音響特徴量推定処理例を説明する図である。
【図１４】音声生成部による音声合成処理例を説明する図である。
【図１５】非特許文献１に記載された従来の学習方法及び合成方法を示す説明図である。
【図１６】非特許文献２に記載された従来の音声信号調整方法を示す説明図である。
【図１７】非特許文献１，２の従来技術を組み合わせた想定例を示す説明図である。

【発明を実施するための形態】

【００２７】

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。

〔学習装置〕

まず、本発明の実施形態による学習装置について説明する。図１は、学習装置の構成例を示すブロック図であり、図２は、学習装置による事前学習処理例を示すフローチャート

である。
【００２８】
この学習装置１は、記憶部１０，１７、言語分析部１１、音声分析部１２、対応付け部１３、調整量追加部１４、音響特徴量調整部１５及び学習部１６を備えている。音声信号はモノラルであり、標本化周波数４８ｋＨｚ及びビット数１６で標本化されているものとする。
【００２９】
記憶部１０には、予め設定された音声コーパスが格納されている。音声コーパスは、テキストと、これに対応する音声信号から構成される。例えば、ＡＴＲ（株式会社国際電気通信基礎技術研究所）により作成された音素バランス５０３文を利用する場合、テキストと、これを読み上げた音声信号は、５０３対からなる。音声コーパスについては、以下の文献を参照されたい。
磯健一、渡辺隆夫、桑原尚夫、「音声データベース用文セットの設計」、音講論（春）、pp.89-90（1988.3）
【００３０】
言語分析部１１は、記憶部１０から音声コーパスの各テキストを読み出し、テキストについて既知の言語分析処理を行い、音素毎の所定情報からなる言語特徴量を求める（ステップＳ２０１）。そして、言語分析部１１は、音素毎の言語特徴量を対応付け部１３に出力する。
【００３１】
具体的には、言語分析部１１は、言語分析処理により、文を構成する音素毎に、音素情報、アクセント情報、品詞情報、アクセント句情報、呼気段落情報及び総数情報を求め、これらの情報からなる言語特徴量を求める。
【００３２】
言語分析処理としては、例えば以下に記載された形態素解析処理が用いられる。
“MeCab：Yet Another Part-of-Speech and Morphological Analyzer”，インターネット＜ＵＲＬ：http://taku910.github.io/mecab/＞
また、言語分析処理としては、例えば以下に記載された係り受け解析処理が用いられる。
“CaboCha/南瓜：Yet Another Japanese Dependency Structure Analyzer”，インターネット＜ＵＲＬ：https://taku910.github.io/cabocha/＞
【００３３】
図３は、言語特徴量のデータ構成例を説明する図である。図３に示すように、言語特徴量は、音素毎に、音素情報、アクセント情報、品詞情報、アクセント句情報、呼気段落情報及び総数情報から構成される。
【００３４】
図１及び図２に戻って、音声分析部１２は、記憶部１０から音声コーパスの各テキストに対応する各音声信号を読み出し、フレーム毎に音声信号を切り出し、フレーム毎の音声信号について既知の音響分析処理を行う。そして、音声分析部１２は、フレーム毎の所定情報からなる音響特徴量を求め（ステップＳ２０２）、フレーム毎の音響特徴量を対応付け部１３に出力する。音響特徴量は、後述するように、１９９次元のデータから構成される。
【００３５】
音響分析処理としては、例えば以下に記載された音響分析処理が用いられる。
“A high-quality speech analysis, manipulation and synthesis system”，インターネット＜ＵＲＬ：https://github.com/mmorise/World＞
また、音響分析処理としては、例えば以下に記載された音声信号処理が用いられる。
“Speech Signal Processing Toolkit(SPTK) Version 3.11 December 25, 2017”，インターネット＜ＵＲＬ：http://sp-tk.sourceforge.net/＞
“REFERENCE MANUAL for Speech Signal Processing Toolkit Ver. 3.9”
【００３６】

図４は、音声分析部１２による音声分析処理例を示すフローチャートである。音声分析部１２は、記憶部１０から音声コーパスの各音声信号を読み出し、フレーム長２５ｍｓの音声信号をフレームシフト５ｍｓ毎に切り出す（ステップＳ４０１）。そして、音声分析部１２は、フレーム毎の音声信号について音響分析処理を行い、スペクトル、ピッチ周波数及び非周期成分を求める（ステップＳ４０２）。

【００３７】

音声分析部１２は、スペクトルをメルケプストラム分析してメルケプストラム係数ＭＧＣを求める（ステップＳ４０３）。また、音声分析部１２は、ピッチ周波数から有声／無声判定情報ＶＵＶを求め、ピッチ周波数の有声区間を対数化し、無声及び無音区間については前後の有声区間の情報を用いて補間することにより、対数ピッチ周波数ＬＦ０を求める（ステップＳ４０４）。また、音声分析部１２は、非周期成分をメルケプストラム分析して帯域非周期成分ＢＡＰを求める（ステップＳ４０５）。

【００３８】

これにより、静特性の音響特徴量として、フレーム毎に、メルケプストラム係数ＭＧＣ、有声／無声判定情報ＶＵＶ、対数ピッチ周波数ＬＦ０及び帯域非周期成分ＢＡＰが得られる。

【００３９】

音声分析部１２は、メルケプストラム係数ＭＧＣの１次差分Δを算出して１次差分メルケプストラム係数ΔＭＧＣを求め（ステップＳ４０６）、２次差分$\Delta^2$を算出して２次差分メルケプストラム係数$\Delta^2$ＭＧＣを求める（ステップＳ４０７）。

【００４０】

音声分析部１２は、対数ピッチ周波数ＬＦ０の１次差分Δを算出して１次差分対数ピッチ周波数ΔＬＦ０を求め（ステップＳ４０８）、２次差分$\Delta^2$を算出して２次差分対数ピッチ周波数$\Delta^2$ＬＦ０を求める（ステップＳ４０９）。

【００４１】

音声分析部１２は、帯域非周期成分ＢＡＰの１次差分Δを算出して１次差分帯域非周期成分ΔＢＡＰを求め（ステップＳ４１０）、２次差分$\Delta^2$を算出して２次差分帯域非周期成分$\Delta^2$ＢＡＰを求める（ステップＳ４１１）。

【００４２】

これにより、動特性の音響特徴量として、フレーム毎に、１次差分メルケプストラム係数ΔＭＧＣ、２次差分メルケプストラム係数$\Delta^2$ＭＧＣ、１次差分対数ピッチ周波数ΔＬＦ０、２次差分対数ピッチ周波数$\Delta^2$ＬＦ０、１次差分帯域非周期成分ΔＢＡＰ及び２次差分帯域非周期成分$\Delta^2$ＢＡＰが得られる。

【００４３】

音声分析部１２は、フレーム毎の静特性及び動特性の所定情報からなる音響特徴量を対応付け部１３に出力する。

【００４４】

図５は、音響特徴量のデータ構成例を説明する図である。図５に示すように、音響特徴量は、フレーム毎に、静特性のメルケプストラム係数ＭＧＣ、対数ピッチ周波数ＬＦ０及び帯域非周期成分ＢＡＰ、動特性の１次差分メルケプストラム係数ΔＭＧＣ、１次差分対数ピッチ周波数ΔＬＦ０、１次差分帯域非周期成分ΔＢＡＰ、２次差分メルケプストラム係数$\Delta^2$ＭＧＣ、２次差分対数ピッチ周波数$\Delta^2$ＬＦ０及び２次差分帯域非周期成分$\Delta^2$ＢＡＰ、並びに静特性の有声／無声判定情報ＶＵＶから構成される。この音響特徴量は、後述するように、１９９次元のデータから構成される。

【００４５】

図１及び図２に戻って、対応付け部１３は、言語分析部１１から音素毎の言語特徴量を入力すると共に、音声分析部１２からフレーム毎の音響特徴量を入力する。そして、対応付け部１３は、既知の音素アラインメントの技術を用いて、音素毎の言語特徴量とフレーム毎の音響特徴量とを時間的に対応付けることで、テキストの文を構成する各音素が音声信号のどの時刻に位置（対応）するのかを算出する（ステップＳ２０３）。

【００４６】
対応付け部１３は、音素毎に、対応する開始フレームの番号及び終了フレームの番号からなる時間情報を生成し、言語特徴量を構成する音素毎の所定情報に時間情報を追加すると共に、音素の時間長（フレーム数）を求める。そして、対応付け部１３は、対応付けた音素毎の時間情報を追加した言語特徴量を調整量追加部１４に出力する。また、対応付け部１３は、音素毎の時間長を音響特徴量に含め、対応付けたフレーム毎の音響特徴量（時間長については音素毎のデータ）を音響特徴量調整部１５に出力する。
【００４７】
ここで、言語特徴量に追加される時間情報は、ミリ秒単位の情報である。また、音素毎の時間長は、後述する統計モデルにおける時間長モデルの出力データに用いられ、音素におけるミリ秒単位の時間の長さをフレームシフト５ｍｓで除算した５ｍｓフレーム単位の数値、すなわち音素のフレーム数が用いられる。
【００４８】
音素アラインメントの技術としては、例えば以下に記載された音声認識処理が用いられる。
“The Hidden Markov Model Toolkit（HTK）”，インターネット＜ＵＲＬ：http://htk.eng.cam.ac.uk＞
“Speech Signal Processing Toolkit(SPTK) Version 3.11 December 25, 2017”
【００４９】
尚、対応付け部１３は、言語特徴量及び音響特徴量の時間的な対応付け処理の後に、各文の文頭及び文末の無音区間を削除する。
【００５０】
図６は、時間情報が追加された言語特徴量のデータ構成例を説明する図である。図６に示すように、時間情報が追加された言語特徴量は、図３に示した言語特徴量に時間情報を追加して構成される。具体的には、この言語特徴量は、音素毎に、時間情報、音素情報、アクセント情報、品詞情報、アクセント句情報、呼気段落情報及び総数情報から構成される。
【００５１】
図１及び図２に戻って、調整量追加部１４は、対応付け部１３から音素毎の言語特徴量を入力すると共に、所定の調整パラメータを入力する。そして、調整量追加部１４は、言語特徴量を構成する音素毎の所定情報に、調整パラメータの調整量情報を追加する（ステップＳ２０４）。調整量追加部１４は、音素毎の調整量情報を追加した言語特徴量を学習部１６に出力する。
【００５２】
所定の調整パラメータは、音声信号を調整する（音響の特徴を調整する）ためのパラメータであり、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$のうちのいずれか、またはこれらの組み合わせとし、ユーザにより選択されるものとする。また、調整パラメータは、学習部１６において学習データの一部として用いられる。
【００５３】
話速$R_{ST}$は話速の調整量を示し、パワー$R_{PW}$はパワー（声の大きさ）の調整量を示し、$R_{PT}$はピッチ（声の高さ）の調整量を示し、抑揚$R_{PD}$は抑揚（声の高さの変化幅）の調整量を示す。尚、話速の代わりに、時間長を用いるようにしてもよい。
【００５４】
話速$R_{ST}$の範囲（話速の調整量範囲）は、例えば以下のとおりとする。
（遅い）0.5〈=$R_{ST}$〈=4.0（速い）
これは、話速$R_{ST}$は0.5から4.0までの範囲において、0.5に近いほど遅く、4.0に近いほど速いことを意味する。
【００５５】
パワー$R_{PW}$の範囲（パワーの調整量範囲）は、例えば以下のとおりとする。
（小さい）1.0E-5〈=$R_{PW}$〈=2.0（大きい）

これは、パワー$R_{PW}$は1.0E-5から2.0までの範囲において、1.0E-5に近いほど小さく、2.0に近いほど大きいことを意味する。

【００５６】

ピッチ$R_{PT}$の範囲（ピッチの調整量範囲）は、例えば以下のとおりとする。

（低い）0.5〈=$R_{PT}$〈=2.0（高い）

これは、ピッチ$R_{PT}$は0.5から2.0までの範囲において、0.5に近いほど低く、2.0に近いほど高いことを意味する。

【００５７】

抑揚$R_{PD}$の範囲（抑揚の調整量範囲）は、例えば以下のとおりとする。

（小さい）1.0E-5〈=$R_{PD}$〈=2.0（大きい）

これは、抑揚$R_{PD}$は1.0E-5から2.0までの範囲において、1.0E-5に近いほど小さく、2.0に近いほど大きいことを意味する。話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$の標準値は、いずれも1.0とする。

【００５８】

また、これらの調整パラメータのそれぞれは、例えば以下に示す１１個のデータから選択されるものとする。すなわち、学習装置１における話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$の調整パラメータは、それぞれ１１個のデータのいずれかが使用される。

［数１］

$$R_{ST} \in \{0.5,0.6,0.7,0.8,0.9,1.0,1.6,2.2,2.8,3.4,4.0\}$$
$$R_{PW} \in \{0.00001,0.2,0.4,0.6,0.8,1.0,1.2,1.4,1.6,1.8,2.0\}$$
$$R_{PT} \in \{0.5,0.6,0.7,0.8,0.9,1.0,1.2,1.4,1.6,1.8,2.0\} \quad \cdots (1)$$
$$R_{PD} \in \{0.00001,0.2,0.4,0.6,0.8,1.0,1.2,1.4,1.6,1.8,2.0\}$$

【００５９】

ここで、４つの調整パラメータを以下の調整ベクトルで表現する。

$$R = (R_{ST},\ R_{PW},\ R_{PT},\ R_{PD})$$

話速、パワー等の調整量を変化させないで元の話速、パワー等を維持する場合、調整ベクトルは以下のとおりである。

$$R = (1.0,\ 1.0,\ 1.0,\ 1.0)$$

【００６０】

４つの調整パラメータにおいて、それぞれ１１個のデータから１個のデータが選択されるものとすると、全ての組み合わせ数は、$11^4$＝14,641となる。このため、統計モデルを学習するためには、膨大なデータ量が必要となることから、学習の負荷が高くなり、時間もかかってしまう。

【００６１】

そこで、本発明の実施形態では、ユーザは、４つの調整パラメータのうちの１つの調整パラメータについて、所定範囲の１１個のデータから１個のデータを選択し、他の３つの調整パラメータについては、標準値1.0を固定値として用いるようにしてもよい。音響特徴量調整部１５、及び後述する図１０の音声合成装置２についても同様である。

【００６２】

例えば、ユーザは、話速$R_{ST}$について１１個のデータから１個のデータを選択し、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$について標準値1.0を固定値として用いるものとすると、調整ベクトルは以下のとおりである。

$$R = (R_{ST},\ 1.0,\ 1.0,\ 1.0)$$

この場合、調整量追加部１４は、調整パラメータとして、ユーザにより１１個のデータのうち１個のデータが選択された話速$R_{ST}$、並びに、標準値1.0を固定値としたパワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$を入力する。

【００６３】

このように、４つの調整パラメータのうちの１つの調整パラメータについては１１個のデータから１個のデータが選択され、他の３つの調整パラメータについては標準値である1.0を固定値として用いることは、調整ベクトルＲのいずれか１つの要素の軸方向のみに調整量をプロットしたことと等価である。この場合の組み合わせ数は、10×4＋1＝41となる。これにより、統計モデルを学習する際に、学習データの数を減らすことができるから、学習処理の負荷を低減し、学習処理の時間を短縮することができる。

【００６４】

また、本発明の実施形態における他の例として、ユーザは、４つの調整パラメータを１１段階で連動させて選択するようにしてもよい。音響特徴量調整部１５、及び後述する図１０の音声合成装置２についても同様である。

【００６５】

この場合、調整量追加部１４は、調整パラメータとして、予め設定された１１種類のパターンのうち、ユーザにより選択されたいずれかのパターンの話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$を入力する。１１種類のパターンの調整ベクトルは以下のとおりである。

$$R=(\mathrm{a1},\ \mathrm{b1},\ \mathrm{c1},\ \mathrm{d1}),(\mathrm{a2},\ \mathrm{b2},\ \mathrm{c2},\ \mathrm{d2}),\cdots,(\mathrm{a11},\ \mathrm{b11},\ \mathrm{c11},\ \mathrm{d11})$$

a1，b1，・・・，c11，d11は、対応する調整パラメータの調整量範囲に含まれる値とする。

【００６６】

この場合の組み合わせ数は、１１となる。これにより、統計モデルを学習する際に、学習データの数を一層減らすことができるから、その負荷を一層低減し、その時間を一層短縮することができる。

【００６７】

尚、調整量追加部１４は、文章単位、呼気段落単位またはアクセント句単位で、異なる調整パラメータを入力するようにしてもよい。音響特徴量調整部１５、及び後述する音声合成装置２についても同様である。

【００６８】

図７は、調整量情報が追加された言語特徴量のデータ構成例を説明する図である。図７に示すように、調整量情報が追加された言語特徴量は、図６に示した言語特徴量に、調整パラメータの調整量情報を追加して構成される。具体的には、この言語特徴量は、音素毎に、時間情報、音素情報、アクセント情報、品詞情報、アクセント句情報、呼気段落情報、総数情報及び調整量情報から構成される。

【００６９】

調整量情報は、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$の調整パラメータにおける調整量が反映された情報である。

【００７０】

前述のとおり、調整量追加部１４は、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$のうちのどれか、またはこれらの組み合わせの調整パラメータを入力する。調整量追加部１４は、例えば話速$R_{ST}$のみの調整パラメータを入力した場合、言語特徴量に、入力した話速$R_{ST}$、並びに固定値である標準値1.0のパワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$の調整量情報を追加する。また、調整量追加部１４は、例えば話速$R_{ST}$及びパワー$R_{PW}$の調整パラメータを入力した場合、言語特徴量に、入力した話速$R_{ST}$及びパワー$R_{PW}$、並びに固定値である標準値1.0のピッチ$R_{PT}$及び抑揚$R_{PD}$の調整量情報を追加する。

【００７１】

図１及び図２に戻って、音響特徴量調整部１５は、対応付け部１３から、調整量追加部１４が入力する音素毎の言語特徴量に対応するフレーム毎の音響特徴量（時間長については音素毎のデータ）を入力する。また、音響特徴量調整部１５は、調整量追加部１４と同様の所定の調整パラメータを入力する。

【００７２】

音響特徴量調整部１５は、調整パラメータに従ってフレーム毎の音響特徴量を調整し、調整後のフレーム毎の音響特徴量（時間長については音素毎のデータ）を学習部１６に出力する。

【００７３】

話速$R_{ST}$の調整パラメータに従い話速が調整される場合、音響特徴量調整部１５は、以下の式のとおり、対応付け部１３から入力した時間長DURに話速$R_{ST}$の逆数を乗算し、乗算結果を整数化し、新たな時間長DUR’を求めることで、時間長を調整する。

［数２］

$$DUR' = \mathrm{int}(DUR \times 1 / R_{ST}) \quad \cdots (2)$$

対応付け部１３から入力した時間長をDUR、調整後の時間長をDUR’とする。

【００７４】

尚、話速$R_{ST}$の代わりに時間長の調整パラメータ$R_{DR}$（＝１／$R_{ST}$）に従い時間長が調整される場合、音響特徴量調整部１５は、対応付け部１３から入力した時間長DURに対し、話速$R_{ST}$の逆数の代わりに、時間長の調整パラメータ$R_{DR}$を乗算し、乗算結果を整数化し、新たな時間長DUR’を求めることで、時間長を調整する。

【００７５】

音響特徴量調整部１５は、調整後の時間長に応じて、対応付け部１３から入力したフレームの音響特徴量を繰り返しまたは間引きして、音響特徴量のフレーム数を揃えることで、音響特徴量を調整する。このように、音素毎の時間長の調整に応じて、音響特徴量のフレーム数が揃えられる。

【００７６】

尚、音響特徴量調整部１５は、調整後の時間長に応じて、対応するフレームの音響特徴量を繰り返しまたは間引くことで音響特徴量を調整する際に、前後のフレームの音響特徴量を用いて補間を行うようにしてもよい。これにより、高品質の音響特徴量を得ることができる。また、話速$R_{ST}$の調整パラメータ及び他の調整パラメータに従い話速等が調整される場合、音響特徴量調整部１５は、話速を調整する前に、他の調整パラメータによる調整を行う。

【００７７】

また、パワー$R_{PW}$の調整パラメータに従い音声のパワーが調整される場合、音響特徴量調整部１５は、対応付け部１３から入力した音響特徴量に含まれる静特性のメルケプストラム係数ＭＧＣにおける０次元目の値MGC[0]に、パワー$R_{PW}$を対数化した値を加算する。

【００７８】

音響特徴量調整部１５は、以下の式のとおり、加算した値と０とを比較して大きい方を、新たな静特性のメルケプストラム係数ＭＧＣにおける０次元目の値MGC[0]’として求めることで、音響特徴量を調整する。

［数３］

$$MGC[0]' = \max(0, MGC[0] + \log R_{PW}) \quad \cdots (3)$$

対応付け部１３から入力した音響特徴量に含まれる静特性のメルケプストラム係数ＭＧＣにおける０次元目の値をMGC[0]、調整後の値をMGC[0]’とする。

【００７９】

また、ピッチ$R_{PT}$の調整パラメータに従い音声のピッチ周波数が調整される場合、音響特徴量調整部１５は、対応付け部１３から入力した音響特徴量に含まれる静特性の対数ピッチ周波数ＬＦ０における０次元目の値LF0[0]に、ピッチ$R_{PT}$を対数化した値を加算する。

【００８０】
音響特徴量調整部１５は、以下の式のとおり、加算した値と０とを比較して大きい方を、新たな静特性の対数ピッチ周波数ＬＦ０における０次元目の値LF0[0]’として求めることで、音響特徴量を調整する。
［数４］
$$LF0[0]' = \max(0, LF0[0] + \log R_T) \quad \cdots (4)$$
対応付け部１３から入力した音響特徴量に含まれる静特性の対数ピッチ周波数ＬＦ０における０次元目の値をLF0[0]、調整後の値をLF0[0]’とする。
【００８１】
また、抑揚$R_{PD}$の調整パラメータに従い音声の抑揚が調整される場合、音響特徴量調整部１５は、対応付け部１３から入力した音響特徴量に含まれる静特性の対数ピッチ周波数ＬＦ０から、予め算出しておいた平均値$\mu_{LF0}$を減算する。そして、音響特徴量調整部１５は、減算結果を、予め算出しておいた標準偏差$\Sigma_{LF0}$で除算し、除算結果を求める。平均値$\mu_{LF0}$は、対応付け部１３から入力した音響特徴量に含まれる静特性の対数ピッチ周波数ＬＦ０の平均値であり、標準偏差$\Sigma_{LF0}$はその標準偏差である。
【００８２】
音響特徴量調整部１５は、以下の式のとおり、対応付け部１３から入力した音響特徴量に含まれる静特性の対数ピッチ周波数ＬＦ０について、その平均値$\mu_{LF0}$及び標準偏差$\Sigma_{LF0}$を文毎に算出しておくものとする。Ｎは、文に対応するフレーム数である。
［数５］
$$\mu_{LF0} = \frac{1}{N}\sum_{i=1}^{N} LF0_i \quad \cdots (5)$$

［数６］
$$\Sigma_{LF0} = \left(\frac{1}{N}\sum_{i=1}^{N}(LF0_i - \mu_{LF0})^2\right)^{\frac{1}{2}} \quad \cdots (6)$$

【００８３】
音響特徴量調整部１５は、標準偏差$\Sigma_{LF0}$に、抑揚$R_{PD}$を対数化した値を加算し、加算結果と０とを比較して大きい方を求める。そして、音響特徴量調整部１５は、前記除算結果に、大きい方の値を乗算し、乗算結果に平均値$\mu_{LF0}$を加算する。
【００８４】
音響特徴量調整部１５は、加算した値と０とを比較して大きい方を、新たな静特性の対数ピッチ周波数ＬＦ０’として求める。音響特徴量調整部１５による演算処理の式は以下のとおりである。
［数７］
$$LF0' = \max(0, ((LF0 - \mu_{LF0}) / \Sigma_{LF0}) \times \max(0, \Sigma_{LF0} + \log R_{PD}) + \mu_{LF0}) \quad \cdots (7)$$
対応付け部１３から入力した音響特徴量に含まれる静特性の対数ピッチ周波数をLF0、その平均値を$\mu_{LF0}$、その標準偏差を$\Sigma_{LF0}$、調整後の静特性の対数ピッチ周波数をLF0’とする。
【００８５】
音響特徴量調整部１５は、前記のように各調整パラメータに従い算出された新たな静特性の１次差分Δを算出して新たな動特性の１次差分を求める。また、音響特徴量調整部１５は、２次差分$\Delta^2$を算出して新たな動特性の２次差分を求める。このようにして、音響特徴量調整部１５は、音響特徴量を調整する。
【００８６】

尚、音響特徴量調整部１５による音響特徴量の調整処理は、調整量追加部１４による調整量情報の言語特徴量への追加処理と連動するものとする。

【００８７】

学習部１６は、調整量追加部１４から音素毎の言語特徴量を入力すると共に、音響特徴量調整部１５からフレーム毎の音響特徴量（時間長については音素毎のデータ）を入力する。そして、学習部１６は、これらのデータを標準化し、統計モデルである時間長モデル及び音響モデルを学習する。

【００８８】

（時間長モデルの学習）

次に、学習部１６による時間長モデルの学習処理について説明する。図８は、時間長モデルの学習処理例を説明する図である。学習部１６は、調整量追加部１４から入力した音素毎の言語特徴量に基づいて、言語特徴を表す３１２次元のバイナリ値及び１３次元の数値データ、並びに１次元の調整データを生成する。１次元の調整データは話速データであり、言語特徴量の次元数は３２６である。

【００８９】

ここで、言語特徴量における３１２次元のバイナリ値及び１３次元の数値データは、言語特徴量に含まれる音素情報、アクセント情報、品詞情報、アクセント句情報、呼気段落情報及び総数情報に基づいて生成される。言語特徴量における１次元の調整データは、言語特徴量に含まれる調整量情報（話速の調整量、パワーの調整量、ピッチの調整量及び抑揚の調整量）のうち、話速の調整量に基づいて生成される。

【００９０】

学習部１６は、言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び１次元の調整データ（話速データ）からなる３２６次元のデータを、時間長モデルの入力データとして扱う（ステップＳ８０１）。

【００９１】

学習部１６は、言語特徴量の３２６次元の全てのデータを用いて、次元毎に、最大値及び最小値を求めて記憶部１７に格納すると共に、全てのデータのそれぞれについて、次元毎の最大値及び最小値を用いて標準化する（ステップＳ８０２）。

【００９２】

また、学習部１６は、音響特徴量調整部１５から入力したフレーム毎の音響特徴量（時間長については音素毎のデータ）のうちの音素毎の時間長について、当該時間長の１次元のデータを、時間モデルの出力データとして扱う（ステップＳ８０３）。この時間長は、５ｍｓ単位のフレーム数であり、テキストを表現する音素毎に１次元の整数値からなる。

【００９３】

学習部１６は、時間長の１次元の全てのデータを用いて、平均値及び標準偏差を求めて記憶部１７に格納すると共に、全てのデータのそれぞれについて、平均値及び標準偏差を用いて標準化する（ステップＳ８０４）。

【００９４】

学習部１６は、ステップＳ８０２，Ｓ８０４から移行して、音素毎に、言語特徴量の３２６次元の標準化されたデータを入力データとし、時間長の１次元の標準化されたデータを出力データとして時間長モデルを学習する（ステップＳ８０５）。そして、学習部１６は、学習済みの時間長モデルを記憶部１７に格納する。

【００９５】

ステップＳ８０５における時間長モデルの学習の際には、以下のサイトに記載された技術が用いられる。

“CSTR-Edinburgh/merlin”，インターネット＜ＵＲＬ：https://github.com/CSTR-Edinburgh/merlin＞

後述する図９のステップＳ９０５における音響モデルの学習の場合も同様である。

【００９６】

時間長モデルは、例えば入力層を３２６次元、隠れ層を１０２４次元の６層、出力層を

１次元とした順伝播型のニューラルネットワークで構成される。隠れ層における活性化関数は双曲線正接関数が用いられ、損失誤差関数は平均二乗誤差関数が用いられる。また、ミニバッチ数を６４、エポック数を１００、dropout（ドロップアウト）率を０．５、学習係数の最適化方法として確率的勾配降下法、開始学習率を０．０１、１０エポックを過ぎてからエポック毎に学習率を指数減衰させ、誤差逆伝播法にて学習するものとする。尚、１５エポックを過ぎてから、５エポック連続して評価誤差が減少しない場合は学習を早期終了するものとする。

【００９７】

これにより、記憶部１７には、統計モデルとして時間長モデルが格納される。また、記憶部１７には、統計モデルとして、時間長モデルの入力データである言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び１次元の調整データ（話速データ）からなる３２６次元のデータに関する次元毎の最大値及び最小値が格納される。また、記憶部１７には、統計モデルとして、時間長モデルの出力データである時間長の１次元のデータに関する平均値及び標準偏差が格納される。

【００９８】

（音響モデルの学習）

次に、学習部１６による音響モデルの学習処理について説明する。図９は、音響モデルの学習処理例を説明する図である。学習部１６は、調整量追加部１４から入力した音素毎の言語特徴量に基づいて、言語特徴を表す３１２次元のバイナリ値、１３次元の数値データ、４次元の時間データ及び３次元の調整データを生成する。

【００９９】

４次元の時間データは、当該フレームに対応する音素のフレーム数（１次元のデータ）、及び当該フレームの音素内における位置（３次元のデータ）からなる。３次元の調整データは、パワーデータ、ピッチデータ及び抑揚データである。これらの調整データは、言語特徴量に含まれる調整量情報（話速の調整量、パワーの調整量、ピッチの調整量及び抑揚の調整量）のうち、パワーの調整量、ピッチの調整量及び抑揚の調整量に基づいて生成される。また、言語特徴量の次元数は３３２である。

【０１００】

学習部１６は、音素毎の言語特徴量における３１２次元のバイナリ値、１３次元の数値データ、４次元の時間データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３３２次元のデータから、フレーム毎の言語特徴量における３３２次元のデータを生成する。

【０１０１】

学習部１６は、フレーム毎の言語特徴量について、言語特徴量の３１２次元のバイナリ値、１３次元の数値データ、４次元の時間データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３３２次元のデータを、音響モデルの入力データとして扱う（ステップＳ９０１）。

【０１０２】

学習部１６は、言語特徴量の３３２次元の全てのデータを用いて、次元毎に、最大値及び最小値を求めて記憶部１７に格納すると共に、全てのデータのそれぞれについて、次元毎の最大値及び最小値を用いて標準化する（ステップＳ９０２）。

【０１０３】

また、学習部１６は、音響特徴量調整部１５から入力したフレーム毎の音響特徴量（時間長については音素毎のデータ）のうちの時間長を除く音響特徴量について、１９９次元のデータを、音響モデルの出力データとして扱う（ステップＳ９０３）。

【０１０４】

ここで、前述のとおり、時間長を除く音響特徴量は、静特性のメルケプストラム係数ＭＧＣ、対数ピッチ周波数ＬＦ０及び帯域非周期成分ＢＡＰ、動特性の１次差分メルケプストラム係数ΔＭＧＣ、１次差分対数ピッチ周波数ΔＬＦ０、１次差分帯域非周期成分ΔＢＡＰ、２次差分メルケプストラム係数$\Delta^2$ＭＧＣ、２次差分対数ピッチ周波数$\Delta^2$ＬＦ０及

び２次差分帯域非周期成分$\Delta^2$ＢＡＰ、並びに静特性の有声／無声判定情報ＶＵＶからなる。

【０１０５】

具体的には、時間長を除く音響特徴量は、静特性の６０次元のメルケプストラム係数、１次元の対数ピッチ周波数及び５次元の帯域非周期成分を併せた静特性の６６次元のデータと、これらの静特性のデータを１次差分及び２次差分して得られた動特性の１３２次元のデータと、１次元の有声／無声判定データとからなる。つまり、時間長を除く音響特徴量の次元数は１９９である。

【０１０６】

学習部１６は、音響特徴量の１９９次元の全てのデータを用いて、次元毎に、平均値及び標準偏差を求めて記憶部１７に格納すると共に、全てのデータのそれぞれについて、次元毎の平均値及び標準偏差を用いて標準化する（ステップＳ９０４）。

【０１０７】

学習部１６は、ステップＳ９０２，Ｓ９０４から移行して、フレーム毎に、言語特徴量の３３２次元の標準化されたデータを入力データとし、音響特徴量の１９９次元の標準化されたデータを出力データとして音響モデルを学習する（ステップＳ９０５）。そして、学習部１６は、学習済みの音響モデルを記憶部１７に格納する。

【０１０８】

音響モデルは、例えば入力層を３３２次元、隠れ層を１０２４次元の６層、出力層を１９９次元とした順伝播型のニューラルネットワークで構成される。隠れ層における活性化関数は双曲線正接関数が用いられ、損失誤差関数は平均二乗誤差関数が用いられる。また、ミニバッチ数を２５６、エポック数を１００、dropout（ドロップアウト）率を０．５学習係数の最適化方法として確率的勾配降下法、開始学習率を０．００１、１０エポックを過ぎてからエポック毎に学習率を指数減衰させ、誤差逆伝播法にて学習するものとする。尚、１５エポックを過ぎてから、５エポック連続して評価誤差が減少しない場合は学習を早期終了するものとする。

【０１０９】

これにより、記憶部１７には、統計モデルとして音響モデルが格納される。また、記憶部１７には、統計モデルとして、音響モデルの入力データである言語特徴量の３１２次元のバイナリ値、１３次元の数値データ、４次元の時間データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３３２次元のデータに関する次元毎の最大値及び最小値が格納される。また、記憶部１７には、統計モデルとして、音響モデルの出力データである音響特徴量の１９９次元のデータに関する次元毎の平均値及び標準偏差が格納される。

【０１１０】

以上のように、本発明の実施形態の学習装置１によれば、言語分析部１１は、音声コーパスのテキストについて既知の言語分析処理を行い、音素毎の言語特徴量を求める。音声分析部１２は、音声コーパスのテキストに対応する音声信号をフレーム毎に切り出し、フレーム毎の音声信号について既知の音響分析処理を行い、フレーム毎の音響特徴量を求める。

【０１１１】

対応付け部１３は、既知の音素アラインメントの技術を用いて、音素毎の言語特徴量とフレーム毎の音響特徴量とを時間的に対応付け、音素毎の時間長を求める。そして、対応付け部１３は、時間情報を追加した音素毎の言語特徴量を生成し、対応付けたフレーム毎の音響特徴量（時間長については音素毎のデータ）を生成する。

【０１１２】

調整量追加部１４は、時間情報を追加した音素毎の言語特徴量に、調整パラメータの調整量情報を追加する。音響特徴量調整部１５は、調整パラメータに従って、フレーム毎の音響特徴量（時間長については音素毎のデータ）を調整する。

【０１１３】

学習部１６は、言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び１次元の調整データ（話速データ）からなる３２６次元のデータに基づいて、次元毎に、最大値及び最小値を求め、全てのデータのそれぞれを標準化する。また、学習部１６は、時間長の１次元のデータに基づいて平均値及び標準偏差を求め、時間長の１次元のデータを標準化する。

【０１１４】

学習部１６は、音素毎に、言語特徴量の３２６次元の標準化されたデータを入力データとし、時間長の１次元の標準化されたデータを出力データとして時間長モデルを学習する。

【０１１５】

学習部１６は、言語特徴量の３１２次元のバイナリ値、１３次元の数値データ、４次元の時間データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３３２次元のデータに基づいて、次元毎に、最大値及び最小値を求め、全てのデータのそれぞれを標準化する。また、学習部１６は、音響特徴量の１９９次元のデータに基づいて、次元毎に、平均値及び標準偏差を求め、全てのデータのそれぞれを標準化する。

【０１１６】

学習部１６は、フレーム毎に、言語特徴量の３３２次元の標準化されたデータを入力データとし、音響特徴量の１９９次元の標準化されたデータを出力データとして音響モデルを学習する。

【０１１７】

これにより、記憶部１７には、学習済みの統計モデルとして、調整パラメータの調整量情報が反映された時間長モデル、音響モデル及び最大値等が格納される。

【０１１８】

そして、後述の音声合成装置２により、調整パラメータの調整量情報が反映された学習モデルを用いて、調整パラメータの調整量情報が追加された言語特徴量に基づき音響特徴量が推定され、フレーム毎の音響特徴量から合成音声信号が生成される。

【０１１９】

図１７に示した非特許文献１，２の従来技術を組み合わせた想定例では、学習モデルを用いた推定により時間的に平滑化された特性を有する音響特徴量に調整を加え、調整後のフレーム毎の音響特徴量から合成音声信号を生成することから、合成音声信号に音質劣化が生じてしまう。さらに、入力文章の特定部分に対応する音響特徴量に調整を加え、調整後のフレーム毎の音響特徴量から合成音声信号を生成することから、調整を加えた部分と、これに隣接する調整を加えていない部分との間の接続部分において、合成音声信号に不連続を生じてしまう。

【０１２０】

これに対し、本発明の実施形態による音声合成装置２は、調整パラメータの調整量情報が反映された学習モデルを用いて音響特徴量を推定し、合成音声信号を生成するから、学習モデルを用いた推定により時間的に平滑化された特性を有する音響特徴量に調整を加える必要がない。また、入力文章の特定部分に対応する言語特徴量を調整したものを学習モデルに入力して音響特徴量を求め、合成音声信号を生成することから、調整を加えた部分と、これに隣接する調整を加えていない部分との間の接続部分において、合成音声信号に不連続を生じることがない

【０１２１】

したがって、テキストの特定部分の読み上げ方を調整した合成音声信号を生成する際に、高品質の合成音声信号を得ることができる。

【０１２２】

また、本発明の実施形態では、調整パラメータは、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$のうちのどれか、またはこれらの組み合わせであり、ユーザにより選択される。この場合、ユーザは、例えば４つの調整パラメータのうちの１つの調整パラメータについて、１１個のデータから１個のデータを選択し、他の３つの調整パラメータについては、標準値1.0を固定値として用いる。または、ユーザは、例えば４つの調整パラメータ

を１１段階で連動させて選択する。

【０１２３】

このように、調整パラメータの選択範囲を限定することにより、統計モデルを学習する際の学習データを少なくすることができ、低負荷かつ短時間で、統計モデルを学習することができる。

【０１２４】

〔音声合成装置〕

次に、本発明の実施形態による音声合成装置について説明する。図１０は、音声合成装置の構成例を示すブロック図であり、図１１は、音声合成装置による音声合成処理例を示すフローチャートである。

【０１２５】

この音声合成装置２は、言語分析部２０、調整量追加部２１、音響特徴量推定部２２、記憶部１７及び音声生成部２３を備えている。記憶部１７は、図１に示した記憶部１７に相当し、学習装置１により学習された統計モデルとして、時間長モデル、音響モデル及び最大値等が格納されている。

【０１２６】

尚、学習装置１により学習された統計モデルは、学習装置１に備えた記憶部１７から読み出され、音声合成装置２に備えた記憶部１７に格納されるようにしてもよい。また、音声合成装置２は、インターネットを介して、学習装置１に備えた記憶部１７へ直接アクセスするようにしてもよい。

【０１２７】

言語分析部２０は、音声合成対象のテキストを入力し、図１に示した言語分析部１１と同様に、テキストについて既知の言語分析処理を行い、音素毎の所定情報からなる言語特徴量を求める（ステップＳ１１０１）。そして、言語分析部２０は、音素毎の言語特徴量を調整量追加部２１に出力する。

【０１２８】

調整量追加部２１は、言語分析部２０から音素毎の言語特徴量を入力すると共に、所定の調整パラメータを入力する。そして、調整量追加部２１は、図１に示した調整量追加部１４と同様に、言語特徴量を構成する音素毎の所定情報に、調整パラメータの調整量情報を追加する（ステップＳ１１０２）。調整量追加部２１は、音素毎の調整量情報を追加した言語特徴量を音響特徴量推定部２２に出力する。

【０１２９】

所定の調整パラメータは、前述と同様に、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$のうちのどれか、またはこれらの組み合わせとし、ユーザにより指定されるものとする。調整パラメータの値は、前述した調整の範囲において任意の実数とする。つまり、所定の調整パラメータは、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$のうちのいずれか１つまたは２つ以上の組み合わせとする。

【０１３０】

尚、所定の調整パラメータは、話速$R_{ST}$、パワー$R_{PW}$、ピッチ$R_{PT}$及び抑揚$R_{PD}$とし、これらの４つのパラメータのうちのいずれか１つのパラメータの調整量は、所定範囲内の任意の値が指定され、他の３つのパラメータの調整量は、固定値が用いられるようにしてもよい。また、所定の調整パラメータは、前述の４つのパラメータとし、それぞれの調整量は、それぞれの所定範囲内の任意の値が指定されるようにしてもよい。

【０１３１】

尚、調整量追加部２１は、図１に示した調整量追加部１４と同様に、文章単位、呼気段落単位またはアクセント句単位で、異なる調整パラメータを入力するようにしてもよい。

【０１３２】

音響特徴量推定部２２は、調整量追加部２１から音素毎の言語特徴量を入力し、記憶部１７に格納された最大値等を用いて標準化及び逆標準化の処理を行い、時間長モデルを用いて音素毎の時間長を推定する。

【０１３３】
音響特徴量推定部２２は、記憶部１７に格納された最大値等を用いて標準化及び逆標準化の処理を行い、音響モデルを用いてフレーム毎の音響特徴量を推定する（ステップＳ１１０３）。音響特徴量推定部２２は、フレーム毎の音響特徴量を音声生成部２３に出力する。
【０１３４】
（時間長モデルを用いた時間長の推定）
次に、音響特徴量推定部２２による時間長モデルを用いた時間長の推定処理について説明する。図１２は、時間長モデルを用いた時間長推定処理例を説明する図である。音響特徴量推定部２２は、調整量追加部２１から入力した音素毎の言語特徴量に基づいて、言語特徴を表す３１２次元のバイナリ値及び１３次元の数値データ、並びに１次元の調整データ（話速データ）を生成する。言語特徴量の次元数は３２６である。
【０１３５】
音響特徴量推定部２２は、言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び１次元の調整データ（話速データ）からなる３２６次元のデータを、時間長モデルの入力データとして扱う（ステップＳ１２０１）。
【０１３６】
音響特徴量推定部２２は、記憶部１７から、時間長モデルの入力データである言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び１次元の調整データ（話速データ）からなる３２６次元のデータに関する次元毎の最大値及び最小値を読み出す。そして、音響特徴量推定部２２は、言語特徴量の３２６次元のデータのそれぞれについて、次元毎に、最大値及び最小値を用いて標準化を行う（ステップＳ１２０２）。
【０１３７】
音響特徴量推定部２２は、記憶部１７に格納された時間長モデルを用いて、言語特徴量の３２６次元の標準化されたデータを時間長モデルの入力データとして、時間長モデルの出力データである時間長の１次元の標準化されたデータを推定する（ステップＳ１２０３）。
【０１３８】
音響特徴量推定部２２は、記憶部１７から、時間長モデルの出力データである時間長の１次元のデータに関する平均値及び標準偏差を読み出す。そして、音響特徴量推定部２２は、ステップＳ１２０３にて推定した時間長の１次元の標準化されたデータについて、平均値及び標準偏差を用いて逆標準化を行い（ステップＳ１２０４）、時間長の１次元のデータを求める（ステップＳ１２０５）。
【０１３９】
これにより、記憶部１７に格納された時間長モデル、時間長モデルの入力データである言語特徴量の３２６次元のデータに関する次元毎の最大値及び最小値、並びに、時間長モデルの出力データである時間長の１次元のデータに関する平均値及び標準偏差を用いて、音素毎の言語特徴量の３２６次元のデータから、音素毎の時間長の１次元のデータを得ることができる。
【０１４０】
（音響モデルを用いた音響特徴量の推定）
次に、音響特徴量推定部２２による音響モデルを用いた音響特徴量の推定処理について説明する。図１３は、音響モデルを用いた音響特徴量推定処理例を説明する図である。音響特徴量推定部２２は、ステップＳ１２０５にて求めた音素毎の時間長の１次元のデータに基づいて、図９のステップＳ９０１と同様に、音素に対応する複数フレームのそれぞれについて、時間データの４次元のデータを生成する（ステップＳ１３０１）。
【０１４１】
音響特徴量推定部２２は、調整量追加部２１から入力した音素毎の言語特徴量に基づいて、言語特徴を表す３１２次元のバイナリ値、１３次元の数値データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）を生成する。そして、音響特徴量推

定部２２は、音素毎の言語特徴量における３１２次元のバイナリ値、１３次元の数値データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３２８次元のデータから、フレーム毎の言語特徴量における３２８次元のデータを生成する。

【０１４２】

音響特徴量推定部２２は、フレーム毎の言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３２８次元のデータ、並びにステップＳ１３０１にて生成した時間データの４次元のデータを、音響モデルの入力データとして扱う（ステップＳ１３０２）。

【０１４３】

音響特徴量推定部２２は、記憶部１７から、音響モデルの入力データである言語特徴量の３１２次元のバイナリ値、１３次元の数値データ、４次元の時間データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３３２次元のデータに関する次元毎の最大値及び最小値を読み出す。そして、音響特徴量推定部２２は、言語特徴量の３２８次元のデータ及び時間データの４次元のデータからなる３３２次元のデータのそれぞれについて、次元毎に、最大値及び最小値を用いて標準化を行う（ステップＳ１３０３）。

【０１４４】

音響特徴量推定部２２は、記憶部１７に格納された音響モデルを用いて、言語特徴量の３２８次元の標準化されたデータ及び時間データの４次元の標準化されたデータからなる３３２次元の標準化されたデータを音響モデルの入力データとして、音響モデルの出力データである音響特徴量の１９９次元の標準化されたデータを推定する（ステップＳ１３０４）。

【０１４５】

音響特徴量推定部２２は、記憶部１７から、音響モデルの出力データである音響特徴量の１９９次元のデータに関する平均値及び標準偏差を読み出す。そして、音響特徴量推定部２２は、ステップＳ１３０４にて推定した音響特徴量の１９９次元の標準化されたデータについて、次元毎に、平均値及び標準偏差を用いて逆標準化を行う（ステップＳ１３０５）。音響特徴量推定部２２は、フレーム毎の音響特徴量の１９９次元のデータを生成する（ステップＳ１３０６）。

【０１４６】

このようにして推定され逆標準化された音響特徴量は、フレーム毎に離散的な値をとる。そこで、音響特徴量推定部２２は、連続するフレーム毎の音響特徴量の１９９次元のデータに対して、最尤推定または移動平均をとり、新たなフレーム毎の音響特徴量の１９９次元のデータを求める。これにより、フレーム毎の音響特徴量は滑らかな値となる。

【０１４７】

これにより、記憶部１７に格納された音響モデル、音響モデルの入力データである言語特徴量の３３２次元のデータに関する次元毎の最大値及び最小値、並びに、音響モデルの出力データである音響特徴量の１９９次元のデータに関する平均値及び標準偏差を用いて、フレーム毎の言語特徴量の３２８次元のデータ及び時間データの４次元のデータから、フレーム毎の音響特徴量の１９９次元のデータを得ることができる。

【０１４８】

図１０及び図１１に戻って、音声生成部２３は、音響特徴量推定部２２からフレーム毎の音響特徴量を入力し、フレーム毎の音響特徴量に基づいて音声信号を合成する（ステップＳ１１０４）。そして、音声生成部２３は、音声合成対象のテキストに対して調整パラメータによる調整が加えられた音声信号を出力する。

【０１４９】

図１４は、音声生成部２３による音声合成処理例を説明する図である。音声生成部２３は、音響特徴量推定部２２から入力したフレーム毎の音響特徴量のうち、フレーム毎のメルケプストラム係数ＭＧＣ、対数ピッチ周波数ＬＦ０及び帯域非周期成分ＢＡＰである静特性の音響特徴量を選択する（ステップＳ１４０１）。

【０１５０】
音声生成部２３は、メルケプストラム係数ＭＧＣをメルケプストラムスペクトル変換し、スペクトルを求める（ステップＳ１４０２）。また、音声生成部２３は、対数ピッチ周波数ＬＦ０から有声／無声判定情報ＶＵＶを求め、対数ピッチ周波数ＬＦ０の有声区間を指数化し、無声及び無音区間についてはゼロとし、ピッチ周波数を求める（ステップＳ１４０３）。また、音声生成部２３は、帯域非周期成分ＢＡＰをメルケプストラムスペクトル変換し、非周期成分を求める（ステップＳ１４０４）。
【０１５１】
音声生成部２３は、ステップＳ１４０２にて求めたフレーム毎のスペクトル、ステップＳ１４０３にて求めたフレーム毎のピッチ周波数、及びステップＳ１４０４にて求めたフレーム毎の非周期成分を用いて連続的に音声波形を生成し（ステップＳ１４０５）、音声信号を出力する（ステップＳ１４０６）。
【０１５２】
これにより、音声合成対象のテキストに対して所定の調整パラメータによる調整が加えられた音声信号を得ることができる。
【０１５３】
以上のように、本発明の実施形態の音声合成装置２によれば、言語分析部２０は、音声合成対象のテキストについて既知の言語分析処理を行い、音素毎の言語特徴量を求め、調整量追加部２１は、音素毎の言語特徴量に、調整パラメータの調整量情報を追加する。
【０１５４】
音響特徴量推定部２２は、言語特徴量の３１２次元のバイナリ値、１３次元の数値データ及び１次元の調整データ（話速データ）からなる３２６次元のデータを、記憶部１７に格納された最大値等を用いて標準化する。そして、音響特徴量推定部２２は、記憶部１７に格納された時間長モデルを用いて、これらの標準化されたデータを入力データとして、出力データである時間長の１次元の標準化されたデータを推定する。
【０１５５】
音響特徴量推定部２２は、時間長の１次元の標準化されたデータを、記憶部１７に格納された平均値等を用いて逆標準化し、フレーム毎の時間データを求める。音響特徴量推定部２２は、言語特徴量の３２９次元のデータのうち３１２次元のバイナリ値、１３次元の数値データ及び３次元の調整データ（パワーデータ、ピッチデータ及び抑揚データ）からなる３２８次元のデータ、並びに時間データの４次元のデータを、記憶部１７に格納された最大値等を用いて標準化する。そして、音響特徴量推定部２２は、記憶部１７に格納された音響モデルを用いて、これらの標準化されたデータを入力データとして、出力データである音響特徴量の１９９次元の標準化されたデータを推定する。
【０１５６】
音響特徴量推定部２２は、音響特徴量の１９９次元の標準化されたデータを、記憶部１７に格納された平均値等を用いて逆標準化し、フレーム毎の音響特徴量を求める。そして、音声生成部２３は、フレーム毎の音響特徴量に基づいて音声信号を合成し、合成音声信号を生成する。
【０１５７】
図１７に示した非特許文献１，２の従来技術を組み合わせた想定例では、学習モデルを用いた推定により時間的に平滑化された特性を有する音響特徴量に調整を加え、調整後のフレーム毎の音響特徴量から合成音声信号を生成することから、合成音声信号に音質劣化を生じてしまう。さらに、入力文章の特定部分に対応する音響特徴量に調整を加え、調整後のフレーム毎の音響特徴量から合成音声信号を生成することから、調整を加えた部分と、これに隣接する調整を加えていない部分との間の接続部分において、合成音声信号に不連続を生じてしまう。
【０１５８】
これに対し、本発明の実施形態による音声合成装置２は、調整パラメータの調整量情報が反映された学習モデルを用いて音響特徴量を推定し、合成音声信号を生成するから、学

習モデルを用いた推定により時間的に平滑化された特性を有する音響特徴量に調整を加える必要がない。また、入力文章の特定部分に対応する言語特徴量を調整したものを学習モデルに入力して音響特徴量を求め、合成音声信号を生成することから、調整を加えた部分と、これに隣接する調整を加えていない部分との間の接続部分において、合成音声信号に不連続を生じることがない。

【０１５９】

したがって、テキストの特定部分の読み上げ方を調整した合成音声信号を生成する際に、高品質の合成音声信号を得ることができる。

【０１６０】

以上、実施形態を挙げて本発明を説明したが、本発明は前記実施形態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。

【０１６１】

尚、本発明の実施形態による学習装置１及び音声合成装置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。学習装置１及び音声合成装置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される。

【０１６２】

学習装置１に備えた記憶部１０，１７、言語分析部１１、音声分析部１２、対応付け部１３、調整量追加部１４、音響特徴量調整部１５及び学習部１６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される。また、音声合成装置２に備えた言語分析部２０、調整量追加部２１、音響特徴量推定部２２、記憶部１７及び音声生成部２３の各機能も、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される。

【０１６３】

これらのプログラムは、前記記憶媒体に格納されており、ＣＰＵに読み出されて実行される。また、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。

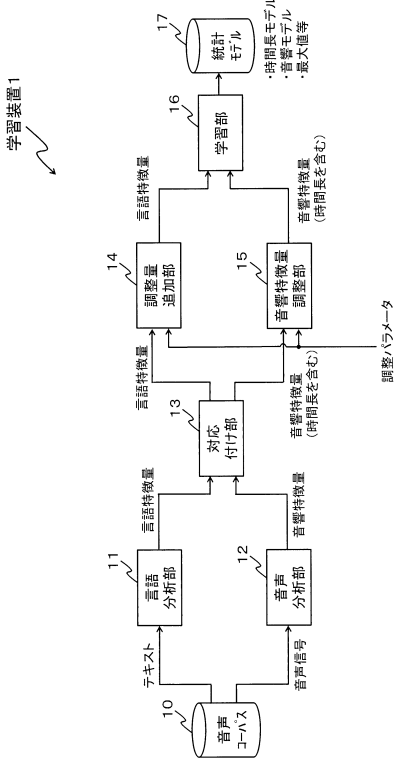
【符号の説明】

【０１６４】

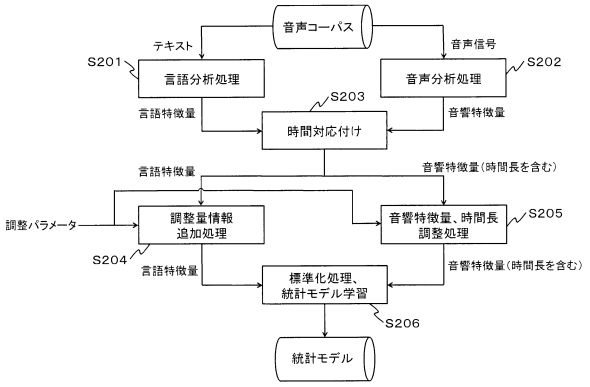
１　学習装置
２　音声合成装置
１０，１７　記憶部
１１，２０　言語分析部
１２　音声分析部
１３　対応付け部
１４，２１　調整量追加部
１５　音響特徴量調整部
１６　学習部
２２　音響特徴量推定部
２３　音声生成部

【図面】

【図１】

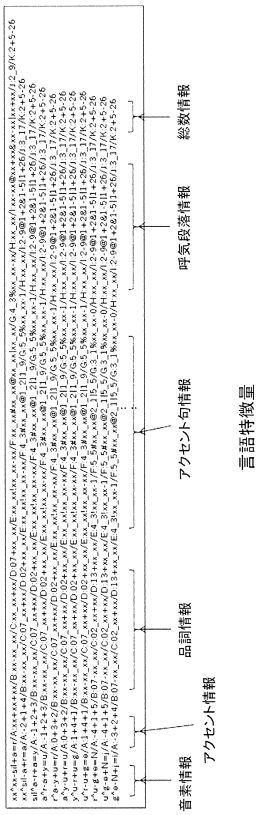



【図２】

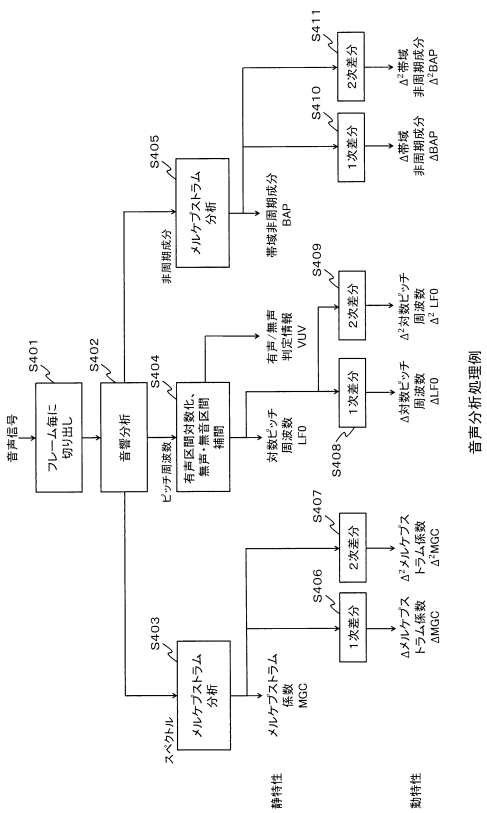



学習装置1による事前学習処理例

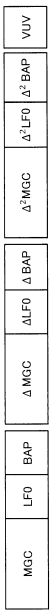
【図３】



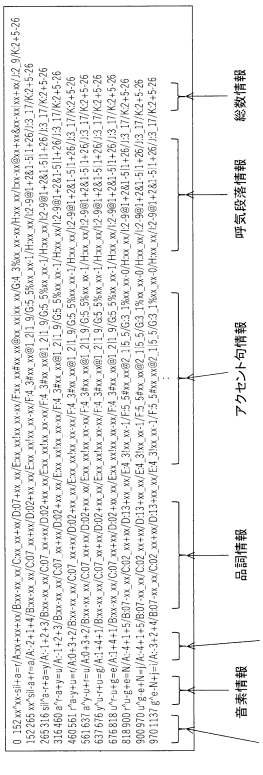

【図４】



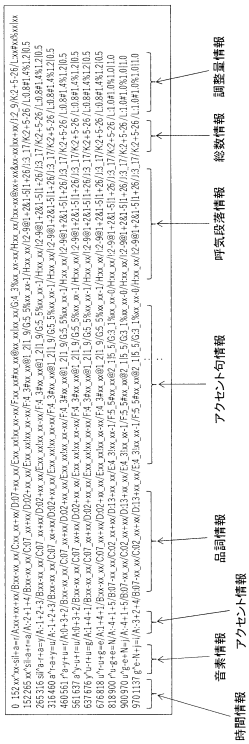
【図５】



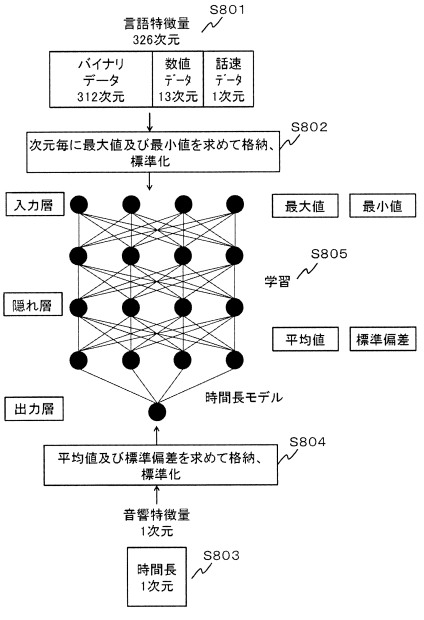

【図６】




【図７】

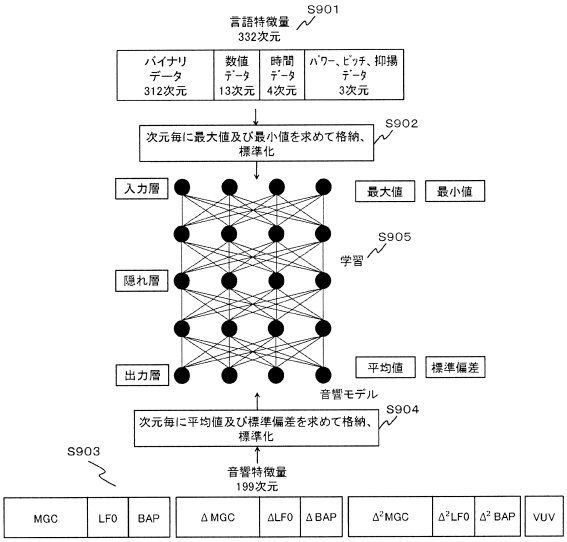



【図８】



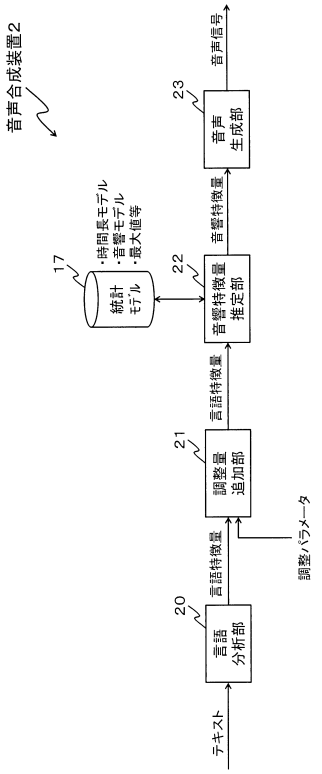

時間長モデルの学習処理例

【図９】

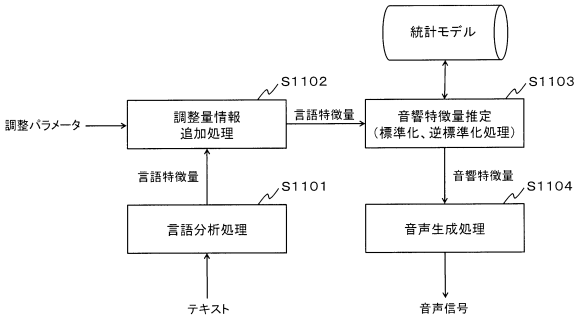



音響モデルの学習処理例

【図１０】

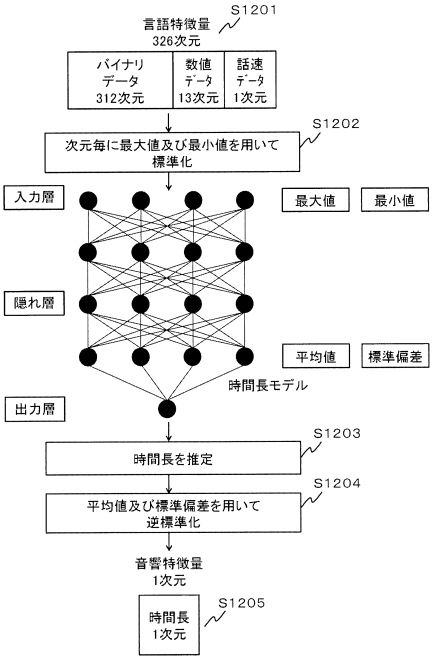



【図１１】



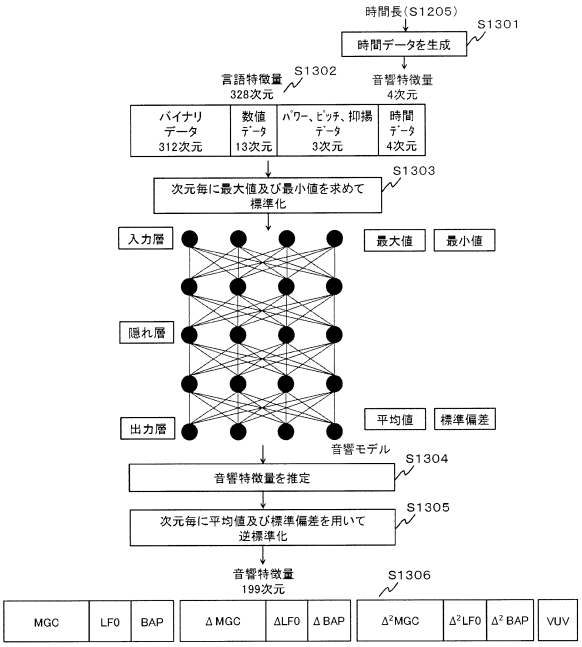

音声合成装置2による音声合成処理例

【図１２】



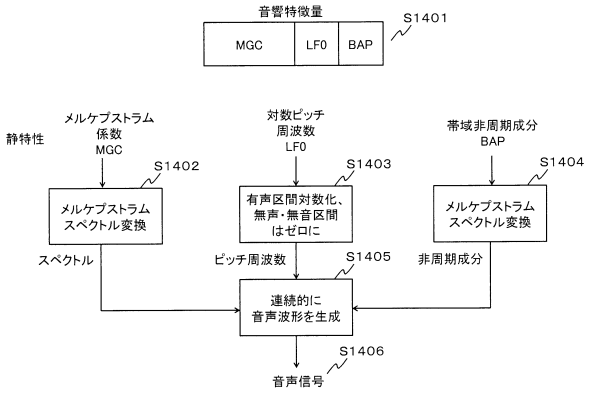

時間長モデルを用いた時間長推定処理例

【図１３】



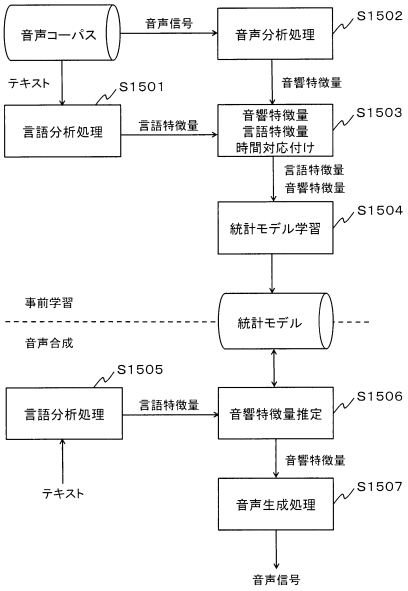

音響モデルを用いた音響特徴量推定処理例

【図１４】



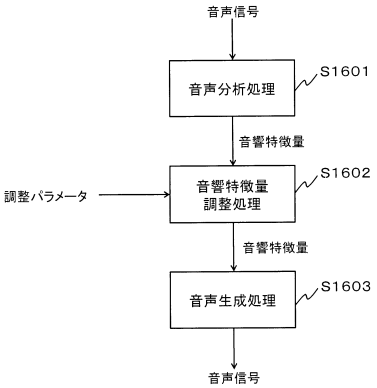

音声合成処理例

【図１５】



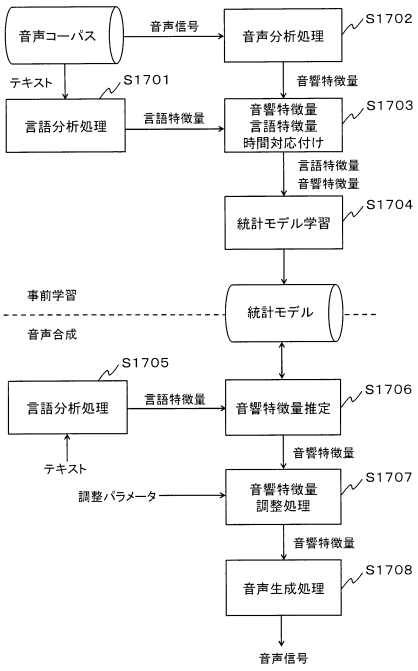

従来技術（非特許文献1）

【図１６】




従来技術（非特許文献2）

【図１７】




従来技術（非特許文献1，2）を組み合わせた想定例

フロントページの続き

(51)国際特許分類　　ＦＩ
　　Ｇ１０Ｌ　　25/30

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
審査官　大野 弘

(56)参考文献　再公表特許第２００９／１０７４４１（ＪＰ，Ａ１）
特開２０１７－０３２８３９（ＪＰ，Ａ）
山田 修平 Shuhei YAMADA，テーラーメイド音声合成のための差分特徴量を用いたＤＮＮに基づくＦ０制御，日本音響学会 ２０１７年 春季研究発表会講演論文集ＣＤ－ＲＯＭ［ＣＤ－ＲＯＭ］，日本，2023年02月17日，PP271-274

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１３／０６
Ｇ１０Ｌ　１３／１０
Ｇ１０Ｌ　２５／３０