(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6846237号
(P6846237)
(45)発行日　令和3年3月24日(2021.3.24)　(24)登録日　令和3年3月3日(2021.3.3)

(51)Int.Cl.　F I
G10L 13/10 (2013.01)　G10L 13/10 114
G10L 13/033 (2013.01)　G10L 13/10 113B
G10L 13/10 112C
G10L 13/10 111F
G10L 13/10 113C
請求項の数 4　(全 17 頁)　最終頁に続く

(21)出願番号　特願2017-42169(P2017-42169)
(22)出願日　平成29年3月6日(2017.3.6)
(65)公開番号　特開2018-146803(P2018-146803A)
(43)公開日　平成30年9月20日(2018.9.20)
審査請求日　令和2年2月3日(2020.2.3)

(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　591053926
一般財団法人ＮＨＫエンジニアリングシステム
東京都世田谷区砧一丁目１０番１１号
(74)代理人　100121119
弁理士　花村　泰伸
(72)発明者　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　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
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】
事前に学習されたＤＮＮ（ディープニューラルネットワーク）を用いて、音声波形を合成する音声合成装置において、
音素の言語特徴量、話者を識別するための話者ラベル、及び感情を識別するための感情ラベルが入力層に与えられ、音素の時間長が出力層に与えられることで学習された時間長ＤＮＮと、
音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベルが入力層に与えられ、音素フレームの音響特徴量が出力層に与えられることで学習された音響特徴量ＤＮＮと、
テキスト、話者情報及び感情情報を入力し、前記時間長ＤＮＮ及び前記音響特徴量ＤＮＮを用いて、前記テキスト、前記話者情報及び前記感情情報に対応する音声波形を合成する合成処理部と、を備え、
前記合成処理部は、
前記テキストをテキスト解析して音素の言語特徴量を生成し、
前記時間長ＤＮＮを用いて、前記音素の言語特徴量、前記話者情報に付与した話者ラベル、及び前記感情情報に付与した感情ラベルに基づいて、音素の時間長を生成し、
前記音素の言語特徴量及び前記音素の時間長に基づいて、音素フレームの言語特徴量を生成し、
前記音響特徴量ＤＮＮを用いて、前記音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベルに基づいて、音素フレームの音響特徴量を生成し、

当該音素フレームの音響特徴量に基づいて、前記音声波形を合成する、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、
前記合成処理部は、
前記話者情報に前記話者ラベルを付与すると共に、前記感情情報に前記感情ラベルを付与する話者感情ラベル処理部と、
前記テキストをテキスト解析して前記音素の言語特徴量を生成し、前記音素の言語特徴量及び前記音素の時間長に基づいて、前記音素フレームの言語特徴量を生成するテキスト解析部と、
前記時間長ＤＮＮを用いて、前記テキスト解析部により生成された前記音素の言語特徴量、前記話者感情ラベル処理部により付与された前記話者ラベル及び前記感情ラベルに基づいて、前記音素の時間長を生成し、
前記音響特徴量ＤＮＮを用いて、前記テキスト解析部により生成された前記音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベルに基づいて、前記音素フレームの音響特徴量を生成する時間長及び音響特徴量生成部と、
前記時間長及び音響特徴量生成部により生成された前記音素フレームの音響特徴量に基づいて、前記音声波形を合成する音声波形合成部と、
を備えた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、
さらに、テキスト、話者情報、感情情報及び音声波形が格納された音声コーパスを用いて、前記時間長ＤＮＮ及び前記音響特徴量ＤＮＮを学習する学習部を備え、
前記学習部は、
前記音声コーパスから前記テキストを読み出し、当該テキストをテキスト解析して音素の言語特徴量を生成し、
前記音声コーパスから前記音声波形を読み出し、当該音声波形を音響分析して音素の区切り位置を求めると共に、音素の時間長を求め、
前記音声コーパスから前記話者情報及び前記感情情報を読み出し、話者ラベル及び感情ラベルをそれぞれ付与し、
前記音素の言語特徴量及び前記音素の時間長に基づいて、音素フレームの言語特徴量を生成し、
前記音素の言語特徴量、前記話者ラベル及び前記感情ラベル、並びに前記音素の時間長を用いて、前記時間長ＤＮＮを学習し、
前記音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベル、並びに前記音素フレームの音響特徴量を用いて、前記音響特徴量ＤＮＮを学習する、ことを特徴とする音声合成装置。
【請求項４】
コンピュータを、請求項１から３までのいずれか一項に記載の音声合成装置として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ＤＮＮ（Deep Neural Network：ディープニューラルネットワーク）を用いた音声合成装置及びプログラムに関する。
【背景技術】
【０００２】
従来、統計モデルを用いた音声合成技術が進展し、スマートフォンまたはパソコンを通して、身近なサービスとして使用できるようになっている。放送分野では、音声合成を用いて解説放送を補完する音声ガイドシステムの研究が進められている（例えば、非特許文

献１を参照）。音声ガイドシステムでは、ガイド音声を提示する際に、演出効果として、多様な話者性及び感情表現が求められている。

【０００３】

一方、統計モデルを用いた音声合成技術の主流は、ＨＭＭ（Hidden Markov Model：隠れマルコフモデル）方式とＤＮＮ方式とに分類される。ＨＭＭ方式では、話者性及び感情表現を制御可能な音声合成を実現している（例えば、非特許文献２を参照）。

【０００４】

ＤＮＮ方式は、一般にＨＭＭ方式よりも音質が良いと言われており、様々な手法で音声合成を実現しているが（例えば、非特許文献３～５を参照）、話者性及び感情表現の両方を制御可能な音声合成の実現に至っていない。

【先行技術文献】

【非特許文献】

【０００５】

【非特許文献１】今井他，電子情報通信学会総合大会講演論文集，H-4-11，Mar 2016
【非特許文献２】J.Yamagishi et al,vol.E88-D,no.3,pp.503-509,Mar 2005
【非特許文献３】Zhizheng Wu et al,ISCA SSW9,vol PS2-13,pp.218-223,Sep 2016
【非特許文献４】H.Zen et al,IEICE Trans.Inf. & Syst.,vol.E90-D, no.5,pp.825-834, May 2007
【非特許文献５】北条他，日本音響学会講演論文集，pp.215-218，Sep 2015

【発明の概要】

【発明が解決しようとする課題】

【０００６】

ＤＮＮ方式を用いた音声合成技術において、話者性及び感情表現を制御可能な音声合成を実現するためには、話者及び感情の組み合わせ毎に、ＤＮＮを用意する手法が想定される。例えば、怒りの感情を有する話者ａのＤＮＮ、喜びの感情を有する話者ａのＤＮＮ、・・・、怒りの感情を有する話者ｂのＤＮＮ、喜びの感情の有する話者ｂのＤＮＮ等を用意する必要がある。

【０００７】

しかしながら、この手法では、話者と感情との組み合わせが膨大であり、用意すべきＤＮＮの数が多くなり、実現が困難である。また、音声合成の際に、複数の異なるＤＮＮを用いる場合には、連続的な自然な読み上げ音声を生成することが困難となる。

【０００８】

このように、ＤＮＮ方式を用いた音声合成技術では、話者性及び感情表現を制御可能な音声合成を実現する際に、膨大な数のＤＮＮを用意する必要のない新たな手法が所望されていた。

【０００９】

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、簡易な構成にて、話者性及び感情表現を同時に制御可能な音声合成を実現する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】

【００１０】

前記課題を解決するために、請求項１の音声合成装置は、事前に学習されたＤＮＮ（ディープニューラルネットワーク）を用いて、音声波形を合成する音声合成装置において、音素の言語特徴量、話者を識別するための話者ラベル、及び感情を識別するための感情ラベルが入力層に与えられ、音素の時間長が出力層に与えられることで学習された時間長ＤＮＮと、音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベルが入力層に与えられ、音素フレームの音響特徴量が出力層に与えられることで学習された音響特徴量ＤＮＮと、テキスト、話者情報及び感情情報を入力し、前記時間長ＤＮＮ及び前記音響特徴量ＤＮＮを用いて、前記テキスト、前記話者情報及び前記感情情報に対応する音声波形を合成する合成処理部と、を備え、前記合成処理部が、前記テキストをテキスト解析して音素

の言語特徴量を生成し、前記時間長ＤＮＮを用いて、前記音素の言語特徴量、前記話者情報に付与した話者ラベル、及び前記感情情報に付与した感情ラベルに基づいて、音素の時間長を生成し、前記音素の言語特徴量及び前記音素の時間長に基づいて、音素フレームの言語特徴量を生成し、前記音響特徴量ＤＮＮを用いて、前記音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベルに基づいて、音素フレームの音響特徴量を生成し、当該音素フレームの音響特徴量に基づいて、前記音声波形を合成する、ことを特徴とする。

【００１１】

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、前記合成処理部が、前記話者情報に前記話者ラベルを付与すると共に、前記感情情報に前記感情ラベルを付与する話者感情ラベル処理部と、前記テキストをテキスト解析して前記音素の言語特徴量を生成し、前記音素の言語特徴量及び前記音素の時間長に基づいて、前記音素フレームの言語特徴量を生成するテキスト解析部と、前記時間長ＤＮＮを用いて、前記テキスト解析部により生成された前記音素の言語特徴量、前記話者感情ラベル処理部により付与された前記話者ラベル及び前記感情ラベルに基づいて、前記音素の時間長を生成し、前記音響特徴量ＤＮＮを用いて、前記テキスト解析部により生成された前記音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベルに基づいて、前記音素フレームの音響特徴量を生成する時間長及び音響特徴量生成部と、前記時間長及び音響特徴量生成部により生成された前記音素フレームの音響特徴量に基づいて、前記音声波形を合成する音声波形合成部と、を備えたことを特徴とする。

【００１２】

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、さらに、テキスト、話者情報、感情情報及び音声波形が格納された音声コーパスを用いて、前記時間長ＤＮＮ及び前記音響特徴量ＤＮＮを学習する学習部を備え、前記学習部が、前記音声コーパスから前記テキストを読み出し、当該テキストをテキスト解析して音素の言語特徴量を生成し、前記音声コーパスから前記音声波形を読み出し、当該音声波形を音響分析して音素の区切り位置を求めると共に、音素の時間長を求め、前記音声コーパスから前記話者情報及び前記感情情報を読み出し、話者ラベル及び感情ラベルをそれぞれ付与し、前記音素の言語特徴量及び前記音素の時間長に基づいて、音素フレームの言語特徴量を生成し、前記音素の言語特徴量、前記話者ラベル及び前記感情ラベル、並びに前記音素の時間長を用いて、前記時間長ＤＮＮを学習し、前記音素フレームの言語特徴量、前記話者ラベル及び前記感情ラベル、並びに前記音素フレームの音響特徴量を用いて、前記音響特徴量ＤＮＮを学習する、ことを特徴とする。

【００１３】

さらに、請求項４のプログラムは、コンピュータを、請求項１から３までのいずれか一項に記載の音声合成装置として機能させることを特徴とする。

【発明の効果】

【００１４】

以上のように、本発明によれば、話者及び感情の組み合わせ毎の膨大な数のＤＮＮを用意する必要がないから、簡易な構成にて、話者性及び感情表現を同時に制御可能な音声合成を実現することができる。

【図面の簡単な説明】

【００１５】

【図１】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２】事前学習部の構成例を示すブロック図である。
【図３】テキスト解析部の構成例を示すブロック図である。
【図４】音響分析部の構成例を示すブロック図である。
【図５】合成処理部の構成例を示すブロック図である。
【図６】時間長ＤＮＮの構成の概要を説明する図である。
【図７】音響特徴量ＤＮＮの構成の概要を説明する図である。
【図８】言語特徴量及び音響特徴量の関係について説明する図である。

【図９】音素の言語特徴量、話者ラベル及び感情ラベルの例を説明する図である。
【図１０】事前学習部の処理例を示すフローチャートである。
【図１１】合成処理部の処理例を示すフローチャートである。
【発明を実施するための形態】
【００１６】
以下、本発明を実施するための形態について図面を用いて詳細に説明する。本発明は、話者性及び感情表現と音声波形とを関連付けて、時間長ＤＮＮ及び音響特徴量ＤＮＮを事前に学習する。また、本発明は、事前に学習した時間長ＤＮＮ及び音響特徴量ＤＮＮを用いて、話者性及び感情表現を反映した音声合成を実現する。
【００１７】
時間長ＤＮＮは、音素の言語特徴量、話者ラベル及び感情ラベルを入力層の各ユニットに与え、音素の時間長を出力層のユニットに与えることで、音素毎に学習されたモデルである。音響特徴量ＤＮＮは、音素フレームの言語特徴量、話者ラベル及び感情ラベルを入力層の各ユニットに与え、音素フレームの音響特徴量を出力層の各ユニットに与えることで、音素フレーム毎に学習されたモデルである。
【００１８】
これにより、時間長ＤＮＮ及び音響特徴量ＤＮＮの２つのＤＮＮを用意すればよいから、話者及び感情の組み合わせ毎の膨大な数のＤＮＮを用意する必要がなく、簡易な構成にて、話者性及び感情表現を同時に制御可能な音声合成を実現することができる。
【００１９】
〔音声合成装置〕
まず、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１は、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。この音声合成装置１は、音声コーパスが格納された記憶部２、事前学習部３、時間長ＤＮＮ及び音響特徴量ＤＮＮが格納された記憶部４、及び合成処理部５を備えている。
【００２０】
記憶部２には、特定の文章が複数の話者と感情で読み上げられた音声に関する情報、すなわち、テキスト、話者情報、感情情報及び音声波形の各情報により構成された音声コーパスが格納されている。音声コーパスは、話者及び感情の組み合わせを単位としたデータベースである。
【００２１】
話者情報は、個々の発話者を識別するための情報であり、感情情報は、例えば喜び、怒り、悲哀、平静等の発話表現を識別するための情報であり、音声波形は、テキストに対する音声波形情報である。
【００２２】
事前学習部３は、記憶部２から、所定の音声コーパスのテキスト、話者情報、感情情報及び音声波形を読み出し、話者情報及び感情情報に話者ラベル及び感情ラベルをそれぞれ付与する。話者ラベルは、話者を識別するためのラベルであり、感情ラベルは、感情を識別するためのラベルである。
【００２３】
事前学習部３は、テキストに対し、所定のテキスト解析を行うと共に、音声波形に対し、所定の音響分析を行うことで、時間長ＤＮＮ及び音響特徴量ＤＮＮを学習するための言語特徴量及び音響特徴量等の情報を生成する。事前学習部３は、言語特徴量及び音響特徴量等の情報、並びに話者ラベル及び感情ラベルを用いて、記憶部４に格納された時間長ＤＮＮ及び音響特徴量ＤＮＮを事前に学習する。
【００２４】
テキスト解析の手法及び音響解析の手法は既知であるから、ここでは詳細な説明は省略する。時間長ＤＮＮ及び音響特徴量ＤＮＮの学習は、例えばＬＳＴＭ（Long Short Term Memory：長期短期記憶）方式にて行われる。
【００２５】

記憶部４には、事前学習部３により学習された時間長ＤＮＮ及び音響特徴量ＤＮＮが格納される。

【００２６】

図６は、時間長ＤＮＮの構成の概要を説明する図である。時間長ＤＮＮは、学習時に、音素の言語特徴量、話者（話者ラベル）及び感情（感情ラベル）が入力層の各ユニットに与えられ、音素の時間長が出力層のユニットに与えられることで、入力層、隠れ層及び出力層の各ユニットの重み等が計算され、音素単位の学習が行われる。

【００２７】

学習のための音素の言語特徴量は、例えば、音素ラベル、アクセントの位置、品詞情報、アクセント句の情報、呼気段落の情報等からなる。音素の時間長は、例えば音素を構成する音素フレームの数で表される。

【００２８】

また、後述する音声合成時には、時間長ＤＮＮの入力層の各ユニットに、音素の言語特徴量、話者ラベル及び感情ラベルが与えられることで、出力層のユニットから、当該音素の言語特徴量、話者ラベル及び感情ラベルに対応する音素の時間長が出力される。

【００２９】

図７は、音響特徴量ＤＮＮの構成の概要を説明する図である。音響特徴量ＤＮＮは、学習時に、音素フレームの言語特徴量、話者（話者ラベル）及び感情（感情ラベル）が入力層の各ユニットに与えられ、音素フレームの音響特徴量が出力層の各ユニットに与えられることで、入力層、隠れ層及び出力層の各ユニットの重み等が計算され、音素フレーム単位の学習が行われる。

【００３０】

学習のための音素フレームの言語特徴量は、例えば、音素の言語特徴量と同様の音素ラベル、アクセントの位置、品詞情報、アクセント句の情報、呼気段落の情報等に加え、音素を構成するフレームの番号（フレームの位置）、音素を構成するフレームの総数等の情報からなる。音素フレームの音響特徴量は、例えば、スペクトル係数、雑音性係数、ピッチ、有声／無声判定等の情報からなる。

【００３１】

また、後述する音声合成時には、音響特徴量ＤＮＮの入力層の各ユニットに、音素フレームの言語特徴量、話者ラベル及び感情ラベルが与えられることで、出力層の各ユニットから、当該音素フレームの言語特徴量、話者ラベル及び感情ラベルに対応する音素フレームの音響特徴量が出力される。

【００３２】

図１に戻って、合成処理部５は、合成対象の音声波形に対応するテキスト、話者情報及び感情情報を入力し、話者情報及び感情情報に基づいて、話者ラベル及び感情ラベルをそれぞれ付与する。そして、合成処理部５は、テキストに対し、所定のテキスト解析を行い、言語特徴量等の情報を生成し、記憶部４に格納された時間長ＤＮＮ及び音響特徴量ＤＮＮを用いて、時間長及び音響特徴量を生成し、音声波形を合成して出力する。

【００３３】

例えば、話者Ａによる喜びの感情を表現した音声波形を合成する場合、合成処理部５は、所定のテキスト、話者Ａを示す話者情報、及び喜びの感情を示す感情情報を入力し、時間長ＤＮＮ及び音響特徴量ＤＮＮを用いて、所定のテキストに対応する音声波形を合成する。

【００３４】

〔事前学習部３／構成〕

次に、図１に示した事前学習部３の構成について詳細に説明する。図２は、事前学習部３の構成例を示すブロック図である。この事前学習部３は、テキスト解析部１１、話者感情ラベル処理部１２及び音響分析部１３を備えている。

【００３５】

事前学習部３は、記憶部２から、事前学習対象の音声コーパスのテキスト、話者情報、

感情情報及び音声波形を読み出す。テキスト解析部１１は、記憶部２の音声コーパスから読み出されたテキストに対し、テキスト解析を行い、音素毎に音素の言語特徴量を生成し、音素の言語特徴量に含まれる音素ラベルを音響分析部１３に出力する。

【００３６】

図９は、音素の言語特徴量、話者ラベル及び感情ラベルの例を説明する図である。この音素の言語特徴量は、テキスト解析により生成された情報である。図９に示すように、テキスト解析により生成された音素の言語特徴量は、音素毎に、「音素ラベル」「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」の各種情報からなる。「音素ラベル」は、テキストを構成する音素を特定するための情報（音素情報）であり、当該音素に加え、前後の音素も含まれる。「話者ラベル」は、話者情報に付与された情報であり、「感情ラベル」は、感情情報に付与された情報である。

【００３７】

図２に戻って、テキスト解析部１１は、テキスト解析にて生成した音素の言語特徴量に基づいて、事前学習のための音素の言語特徴量を生成し、話者感情ラベル処理部１２から話者ラベル及び感情ラベルを入力する。そして、テキスト解析部１１は、事前学習のための音素の言語特徴量、話者ラベル及び感情ラベルを、記憶部４の時間長ＤＮＮにおける入力層の各ユニットに出力する。

【００３８】

事前学習のための音素の言語特徴量は、テキスト解析により生成した音素の言語特徴量の一部の情報、及び、テキスト解析により生成した音素の言語特徴量を加工した情報からなる。事前学習のための音素の言語特徴量は、例えば「音素ラベル」「音素情報」「有声音の有無」「アクセントの位置」等の各種情報からなる。

【００３９】

テキスト解析部１１は、音響分析部１３から音素の時間長を入力し、事前学習のための音素の言語特徴量及び音素の時間長に基づいて、音素の時間長が示す音素フレーム数分の音素フレームの言語特徴量を生成する。そして、テキスト解析部１１は、音素フレームの言語特徴量、話者ラベル及び感情ラベルを、記憶部４の音響特徴量ＤＮＮにおける入力層の各ユニットに出力する。

【００４０】

音素フレームの言語特徴量は、事前学習のための音素の言語特徴量の各種情報に加え、音素フレームを特定するための情報からなる。音素フレームの言語特徴量は、例えば「音素ラベル」「有声音の有無」「アクセントの位置」「フレームの番号」「フレームの総数」等の情報からなる。テキスト解析部１１の詳細については後述する。

【００４１】

話者感情ラベル処理部１２は、記憶部２の音声コーパスから読み出された話者情報及び感情情報に対し、話者ラベル及び感情ラベルをそれぞれ付与し、話者ラベル及び感情ラベルをテキスト解析部１１に出力する。

【００４２】

音響分析部１３は、テキスト解析部１１から音素ラベルを入力し、記憶部２の音声コーパスから読み出された音声波形に対し、所定の学習データを用いて音響分析を行うと共に、音素の区切り位置を求める。そして、音響分析部１３は、音素の区切り位置から音素の時間長を求めると共に、音素フレームの音響特徴量を生成する。音響分析部１３は、音素の時間長をテキスト解析部１１に出力すると共に、記憶部４の時間長ＤＮＮにおける出力層のユニットに出力する。音響分析部１３は、音素フレームの音響特徴量を、記憶部４の音響特徴量ＤＮＮにおける出力層の各ユニットに出力する。

【００４３】

音響分析により音素の区切り位置及び音素の時間長を求め、音素フレームの音響特徴量を生成する手法は既知であるから、ここでは詳細な説明は省略する。

【００４４】

音素フレームの音響特徴量は、例えば、「スペクトル係数」「雑音性係数」「ピッチ」

「有声／無声判定」等の情報からなる。音響分析部１３の詳細については後述する。
【００４５】
このような事前学習により、時間長ＤＮＮ及び音響特徴量ＤＮＮは、話者性及び感情表現と音声波形とを関連付けたモデルとなる。
【００４６】
〔事前学習部３／処理〕
次に、図２に示した事前学習部３の処理について説明する。図１０は、事前学習部３の処理例を示すフローチャートである。事前学習部３のテキスト解析部１１は、記憶部２の音声コーパスから読み出されたテキストに対し、テキスト解析を行い（ステップＳ１００１）、音素の言語特徴量を生成する。そして、テキスト解析部１１は、テキスト解析にて生成した音素の言語特徴量に基づいて、事前学習のための音素の言語特徴量を生成する（ステップＳ１００２）。
【００４７】
話者感情ラベル処理部１２は、記憶部２の音声コーパスから読み出された話者情報及び感情情報に対し、話者ラベル及び感情ラベルをそれぞれ付与する（ステップＳ１００３）。
【００４８】
音響分析部１３は、記憶部２の音声コーパスから読み出された音声波形に対し、音響分析を行い（ステップＳ１００４）、音素の区切り位置を求め、音素の時間長を求める（ステップＳ１００５）。
【００４９】
テキスト解析部１１は、事前学習のための音素の言語特徴量、話者ラベル及び感情ラベルを時間長ＤＮＮに出力すると共に、音響分析部１３は、音素の時間長を時間長ＤＮＮに出力する。これにより、時間長ＤＮＮの事前学習が行われる（ステップＳ１００６）。
【００５０】
音響分析部１３は、音声波形を音響分析することで、ステップＳ１００５にて求めた音素の区切り位置に基づいて、音素フレームの音響特徴量を生成する（ステップＳ１００７）。
【００５１】
テキスト解析部１１は、事前学習のための音素の言語特徴量、及び音響分析部１３により求めた音素の時間長に基づいて、音素フレームの言語特徴量を生成する（ステップＳ１００８）。
【００５２】
テキスト解析部１１は、音素フレームの言語特徴量、話者ラベル及び感情ラベルを音響特徴量ＤＮＮに出力すると共に、音響分析部１３は、音素フレームの音響特徴量を音響特徴量ＤＮＮに出力する。これにより、音響特徴量ＤＮＮの事前学習が行われる（ステップＳ１００９）。
【００５３】
〔言語特徴量と音響特徴量との間の関係〕
次に、時間長ＤＮＮ及び音響特徴量ＤＮＮの入出力データである音素の言語特徴量、音素フレームの言語特徴量、音素の時間長及び音素フレームの音響特徴量の関係について説明する。図８は、言語特徴量及び音響特徴量の関係について説明する図である。
【００５４】
テキストを「い」「ま」とし、「い」の音素ラベルを「ｉ」、「ま」の音素ラベルを「ｍ」「ａ」とする。また、音素ラベル「ｉ」「ｍ」「ａ」における音素の時間長をそれぞれ「１２」「８」「１５」とする。音素の時間長は、１音素あたりの音素フレームの数を示す。音素フレームの時間長は、例えば５msecである。
【００５５】
図８に示すように、音素ラベル「ｉ」の時間区間において、この１音素に対応して、１組の音素の言語特徴量（の各情報）が生成され、１２組の音素フレームの言語特徴量（の

各情報）が生成され、１２組の音素フレームの音響特徴量（の各情報）が生成される。
【００５６】
また、音素ラベル「ｍ」の時間区間において、この１音素に対応して、１組の音素の言語特徴量が生成され、８組の音素フレームの言語特徴量が生成され、８組の音素フレームの音響特徴量が生成される。
【００５７】
また、音素ラベル「ａ」の時間区間において、この１音素に対応して、１組の音素の言語特徴量が生成され、１５組の音素フレームの言語特徴量が生成され、１５組の音素フレームの音響特徴量が生成される。
【００５８】
このように、事前学習において、時間長ＤＮＮの入力層の各ユニットには、音素の言語特徴量、話者ラベル及び感情ラベルが与えられ、出力層のユニットには、音素の時間長が与えられ、この事前学習は音素を単位として行われる。つまり、時間長ＤＮＮには、音素毎に、音素の言語特徴量、話者ラベル、感情ラベル及び音素の時間長が与えられ、事前学習が行われる。音声合成においては、音素毎に、時間長ＤＮＮを用いて、音素の言語特徴量、話者ラベル及び感情ラベルに基づいて、音素の時間長が生成され出力される。
【００５９】
また、事前学習において、音響特徴量ＤＮＮの入力層の各ユニットには、音素フレームの言語特徴量、話者ラベル及び感情ラベルが与えられ、出力層の各ユニットには、音素フレームの音響特徴量が与えられ、この事前学習は音素フレームを単位として行われる。つまり、音響特徴量ＤＮＮには、音素フレーム毎に、音素フレームの言語特徴量、話者ラベル、感情ラベル及び音素フレームの音響特徴量が与えられ、事前学習が行われる。音声合成においては、音素フレーム毎に、音響特徴量ＤＮＮを用いて、音素フレームの言語特徴量、話者ラベル及び感情ラベルに基づいて、音素フレームの音響特徴量が生成され出力される。
【００６０】
（テキスト解析部１１）
次に、図２に示したテキスト解析部１１について詳細に説明する。図３は、テキスト解析部１１の構成例を示すブロック図である。このテキスト解析部１１は、テキスト解析手段３１、前処理手段３２及びフレーム処理手段３３を備えている。
【００６１】
テキスト解析手段３１は、記憶部２の音声コーパスから読み出されたテキストに対し、形態素解析等のテキスト解析を行い、音素毎に音素の言語特徴量を生成する。そして、テキスト解析手段３１は、音素の言語特徴量を前処理手段３２に出力する。
【００６２】
前処理手段３２は、テキスト解析手段３１から、テキスト解析により生成された音素の言語特徴量を入力すると共に、話者感情ラベル処理部１２から話者ラベル及び感情ラベルを入力する。そして、前処理手段３２は、図９に示したように、テキスト解析により生成された音素の言語特徴量、話者ラベル及び感情ラベルからなる情報群を生成する。
【００６３】
前処理手段３２は、テキスト解析により生成された音素の言語特徴量（図９を参照）に基づいて、事前学習のための音素の言語特徴量を生成する。そして、前処理手段３２は、事前学習のための音素の言語特徴量、話者ラベル及び感情ラベルをフレーム処理手段３３に出力すると共に、記憶部４の時間長ＤＮＮにおける入力層の各ユニットに出力する。
【００６４】
フレーム処理手段３３は、前処理手段３２から、事前学習のための音素の言語特徴量、話者ラベル及び感情ラベルを入力すると共に、音響分析部１３から音素の時間長を入力する。そして、フレーム処理手段３３は、事前学習のための音素の言語特徴量及び音素の時間長に基づいて、音素の時間長が示す音素フレーム数分の音素フレームの言語特徴量を生成する。

【００６５】
フレーム処理手段３３は、音素フレームの言語特徴量、話者ラベル及び感情ラベルを、記憶部４の音響特徴量ＤＮＮにおける入力層の各ユニットに出力する。
【００６６】
（音響分析部１３）
次に、図２に示した音響分析部１３について詳細に説明する。図４は、音響分析部１３の構成例を示すブロック図である。この音響分析部１３は、音素区切り処理手段３４及び音響分析手段３５を備えている。
【００６７】
音素区切り処理手段３４は、テキスト解析部１１から音素ラベルを入力し、記憶部２の音声コーパスから読み出された音声波形に対し、所定の学習データを用いて音響分析を行う。そして、音素区切り処理手段３４は、音素ラベルの示す音素が音声波形内でどの位置にあるかを特定し、音素の区切り位置を求める。また、音素区切り処理手段３４は、音素の区切り位置に基づいて、音素ラベルの示す音素の時間長を求める。前述のとおり、音素の時間長は、音素を構成する音素フレームの数で表される。
【００６８】
音素区切り処理手段３４は、音素の区切り位置を音響分析手段３５に出力し、音素の時間長をテキスト解析部１１に出力すると共に、記憶部４の時間長ＤＮＮにおける出力層のユニットに出力する。
【００６９】
音響分析手段３５は、音素区切り処理手段３４から音素の区切り位置を入力し、記憶部２の音声コーパスから読み出された音声波形に対し、音響分析を行い、音素を構成する複数の音素フレームのそれぞれについて、音素フレームの音響特徴量を生成する。
【００７０】
音響分析手段３５は、音素フレームの音響特徴量を、記憶部４の音響特徴量ＤＮＮにおける出力層の各ユニットに出力する。
【００７１】
〔合成処理部５〕
次に、図１に示した合成処理部５の構成について詳細に説明する。図５は、合成処理部５の構成例を示すブロック図である。この合成処理部５は、テキスト解析部２１、話者感情ラベル処理部２２、時間長及び音響特徴量生成部２３及び音声波形合成部２４を備えている。
【００７２】
テキスト解析部２１は、図２に示したテキスト解析部１１と同様の処理を行う。具体的には、テキスト解析部２１は、合成対象の音声波形に対応するテキストを入力し、テキストに対してテキスト解析を行い、音素毎に音素の言語特徴量を生成する。
【００７３】
テキスト解析部２１は、テキスト解析にて生成した音素の言語特徴量に基づいて、図２に示したテキスト解析部１１により生成された事前学習のための音素の言語特徴量と同様の音素の言語特徴量を生成し、話者感情ラベル処理部２２から話者ラベル及び感情ラベルを入力する。そして、テキスト解析部２１は、音素の言語特徴量、話者ラベル及び感情ラベルを時間長及び音響特徴量生成部２３に出力する。
【００７４】
テキスト解析部２１は、時間長及び音響特徴量生成部２３から、当該時間長及び音響特徴量生成部２３に出力した音素の言語特徴量、話者ラベル及び感情ラベルに対応する音素の時間長を入力し、音素の言語特徴量及び音素の時間長に基づいて、音素の時間長が示す音素フレーム数分の音素フレームの言語特徴量を生成する。そして、テキスト解析部２１は、音素フレームの言語特徴量、話者ラベル及び感情ラベルを、時間長及び音響特徴量生成部２３に出力する。
【００７５】

話者感情ラベル処理部２２は、図２に示した話者感情ラベル処理部１２と同様の処理を行う。具体的には、話者感情ラベル処理部２２は、話者情報及び感情情報を入力し、話者情報及び感情情報に対し、話者ラベル及び感情ラベルをそれぞれ付与し、話者ラベル及び感情ラベルをテキスト解析部２１に出力する。

【００７６】

時間長及び音響特徴量生成部２３は、テキスト解析部２１から音素の言語特徴量、話者ラベル及び感情ラベルを入力し、記憶部４の時間長ＤＮＮを用いて、音素の言語特徴量、話者ラベル及び感情ラベルに基づいて、音素の時間長を生成する。そして、時間長及び音響特徴量生成部２３は、音素の時間長をテキスト解析部２１に出力する。

【００７７】

時間長及び音響特徴量生成部２３は、テキスト解析部２１から音素フレームの言語特徴量、話者ラベル及び感情ラベルを入力し、記憶部４の音響特徴量ＤＮＮを用いて、音素フレームの言語特徴量、話者ラベル及び感情ラベルに基づいて、音素フレームの音響特徴量を生成する。

【００７８】

時間長及び音響特徴量生成部２３は、音素フレームの音響特徴量を音声波形合成部２４に出力する。

【００７９】

音声波形合成部２４は、時間長及び音響特徴量生成部２３から音素フレームの音響特徴量を入力し、音素フレームの音響特徴量に基づいて、音声波形を合成し、合成した音声波形を出力する。

【００８０】

具体的には、音声波形合成部２４は、音素フレームの音響特徴量に含まれるピッチ、雑音特性等の情報に基づいて、声帯音源波形を生成する。そして、音声波形合成部２４は、声帯音源波形に対し、音素フレームの音響特徴量に含まれるスペクトル係数等の情報に基づいて声道フィルタ処理を施し、音声波形を合成する。

【００８１】

音素フレームの音響特徴量に基づいて音声波形を合成する手法は既知であるから、ここでは詳細な説明を省略する。

【００８２】

このような音声合成により、話者情報及び感情情報と音声波形とを関連付けた時間長ＤＮＮ及び音響特徴量ＤＮＮを用いることで、テキスト、話者情報及び感情情報に対応する音声波形が合成される。

【００８３】

〔合成処理部５／処理〕

次に、図５に示した合成処理部５の処理について説明する。図１１は、合成処理部５の処理例を示すフローチャートである。合成処理部５のテキスト解析部２１は、合成対象の音声波形に対応するテキストに対し、テキスト解析を行い（ステップＳ１１０１）、音素の言語特徴量を生成する（ステップＳ１１０２）。

【００８４】

話者感情ラベル処理部２２は、話者情報及び感情情報に対し、話者ラベル及び感情ラベルをそれぞれ付与する（ステップＳ１１０３）。

【００８５】

時間長及び音響特徴量生成部２３は、時間長ＤＮＮを用いて、音素の言語特徴量、話者ラベル及び感情ラベルに基づき音素の時間長を生成する（ステップＳ１１０４）。そして、テキスト解析部２１は、音素の言語特徴量及び音素の時間長に基づいて、音素フレームの言語特徴量を生成する（ステップＳ１１０５）。

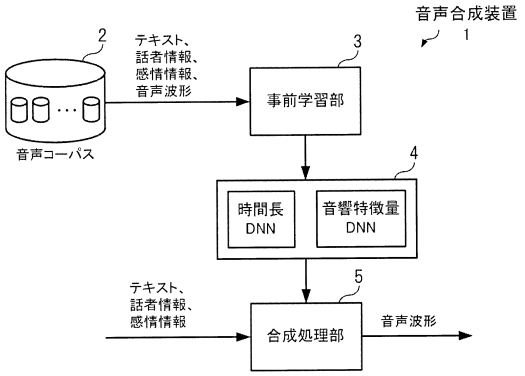
【００８６】

時間長及び音響特徴量生成部２３は、音響特徴量ＤＮＮを用いて、音素フレームの言語特徴量、話者ラベル及び感情ラベルに基づき音素フレームの音響特徴量を生成する（ステ

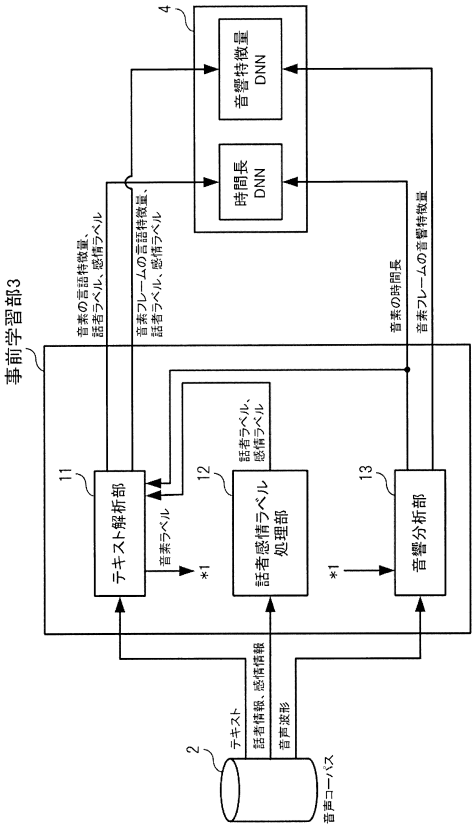
ップＳ１１０６）。
【００８７】
音声波形合成部２４は、音素フレームの音響特徴量に基づいて、音声波形を合成し、合成した音声波形を出力する（ステップＳ１１０７）。
【００８８】
以上のように、本発明の実施形態の音声合成装置１によれば、学習時に、事前学習部３のテキスト解析部１１は、音声コーパスから読み出されたテキストに対しテキスト解析を行い、音素の言語特徴量を生成する。音響分析部１３は、音声コーパスから読み出された音声波形に対して音響分析を行い、音素の区切り位置を求め、音素の時間長を求める。テキスト解析部１１は、音素の言語特徴量、話者ラベル及び感情ラベルを、時間長ＤＮＮにおける入力層の各ユニットに出力すると共に、音響分析部１３は、音素の時間長を、時間長ＤＮＮにおける出力層のユニットに出力する。これにより、時間長ＤＮＮの事前学習が行われる。
【００８９】
また、音響分析部１３は、音声波形を音響分析することで、音素の区切り位置に基づいて、音素フレームの音響特徴量を生成する。テキスト解析部１１は、音素の言語特徴量及び音素の時間長に基づいて、音素フレームの言語特徴量を生成し、音素フレームの言語特徴量、話者ラベル及び感情ラベルを、音響特徴量ＤＮＮにおける入力層の各ユニットに出力すると共に、音響分析部１３は、音素フレームの音響特徴量を、音響特徴量ＤＮＮにおける出力層の各ユニットに出力する。これにより、音響特徴量ＤＮＮの事前学習が行われる。
【００９０】
さらに、本発明の実施形態の音声合成装置１によれば、音声合成時に、合成処理部５のテキスト解析部２１は、対象のテキストに対しテキスト解析を行い、音素の言語特徴量を生成する。時間長及び音響特徴量生成部２３は、時間長ＤＮＮを用いて、音素の言語特徴量、話者ラベル及び感情ラベルに基づき音素の時間長を生成する。
【００９１】
テキスト解析部２１は、音素の言語特徴量及び音素の時間長に基づいて、音素フレームの言語特徴量を生成する。時間長及び音響特徴量生成部２３は、音響特徴量ＤＮＮを用いて、音素フレームの言語特徴量、話者ラベル及び感情ラベルに基づき音素フレームの音響特徴量を生成する。そして、音声波形合成部２４は、音素フレームの音響特徴量に基づいて、音声波形を合成する。
【００９２】
これにより、話者性及び感情表現と音声波形とを関連付けた時間長ＤＮＮ及び音響特徴量ＤＮＮを用いるようにしたから、ＨＭＭ方式よりも音質の良い音声合成を実現することができると共に、話者性及び感情表現を同時に制御することができる。このため、学習時には、話者性及び感情表現に対応した複雑な特徴抽出作業を行う必要がない。
【００９３】
また、時間長ＤＮＮ及び音響特徴量ＤＮＮからなる２つのＤＮＮを用いて音声合成を行うようにしたから、話者及び感情の組み合わせ毎の膨大なＤＮＮを用いる必要がなく、途切れることのない連続的なかつ自然な読み上げ音声を生成することができる。
【００９４】
したがって、簡易な構成にて、話者性及び感情表現を同時に制御可能な音声合成を実現することができる。
【００９５】
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。前記実施形態では、図１に示したように、音声合成装置１は、事前学習を行う事前学習部３と、音声合成を行う合成処理部５とを備えるようにした。これに対し、事前学習部３と合成処理部５とを、それぞれ異なる装置に備えるようにしてもよい。

【００９６】
具体的には、記憶部２を備えた記憶装置、事前学習部３を備えた学習装置、記憶部４を備えた記憶装置、及び合成処理部５を備えた合成装置により音声合成システムが構成される。この場合、学習装置と、記憶部２を備えた記憶装置及び記憶部４を備えた記憶装置とは、インターネットを介して接続されるようにしてもよい。また、合成装置と、記憶部４を備えた記憶装置とは、同様にインターネットを介して接続されるようにしてもよい。さらに、学習装置は、記憶部２、事前学習部３及び記憶部４を備え、合成装置は、記憶媒体を介して可搬された記憶部４、及び合成処理部５を備えるようにしてもよい。
【００９７】
また、前記実施形態では、時間長ＤＮＮ及び音響特徴量ＤＮＮにおけるそれぞれの入力層のユニットに、話者情報及び感情情報を与えるようにした。これに対し、これらの入力層のユニットに、複数の話者情報及び複数の感情情報を与えるようにしてもよい。例えば、話者が複数の観点から分類され、話者に対して複数の話者情報が紐付けられ、同様に、感情が複数の観点から分類され、感情に対して複数の感情情報が紐付けられ、これらを入力層のユニットに与えるようにしてもよい。
【００９８】
尚、本発明の実施形態による音声合成装置１のハードウェア構成としては、通常のコンピュータを使用することができる。音声合成装置１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。音声合成装置１に備えた事前学習部３及び合成処理部５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、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】
【００９９】
１　音声合成装置
２，４　記憶部
３　事前学習部
５　合成処理部
１１，２１　テキスト解析部
１２，２２　話者感情ラベル処理部
１３　音響分析部
２３　時間長及び音響特徴量生成部
２４　音声波形合成部
３１　テキスト解析手段
３２　前処理手段
３３　フレーム処理手段
３４　音素区切り処理手段
３５　音響分析手段

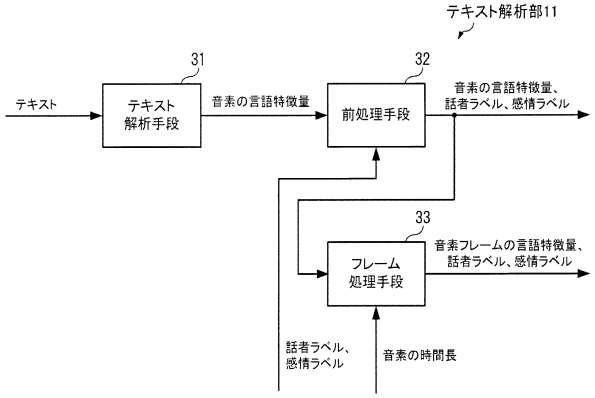
【図１】


音声合成装置
1
2
テキスト、
話者情報、
感情情報、
音声波形
3
事前学習部
音声コーパス
4
時間長
DNN
音響特徴量
DNN
5
テキスト、
話者情報、
感情情報
合成処理部
音声波形

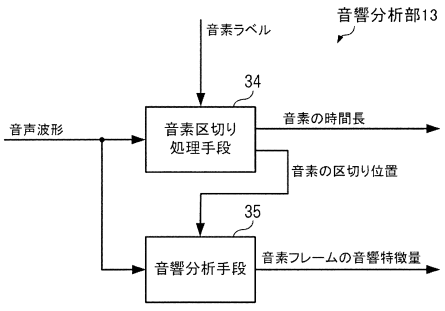
【図２】


4
音響特徴量
DNN
時間長
DNN
事前学習部3
音素の言語特徴量、
話者ラベル、感情ラベル
音素フレームの言語特徴量、
話者ラベル、感情ラベル
音素の時間長
音素フレームの音響特徴量
11
テキスト解析部
音素ラベル
*1
12
話者感情ラベル
処理部
話者ラベル、
感情ラベル
*1
13
音響分析部
テキスト
話者情報、感情情報
音声波形
2
音声コーパス

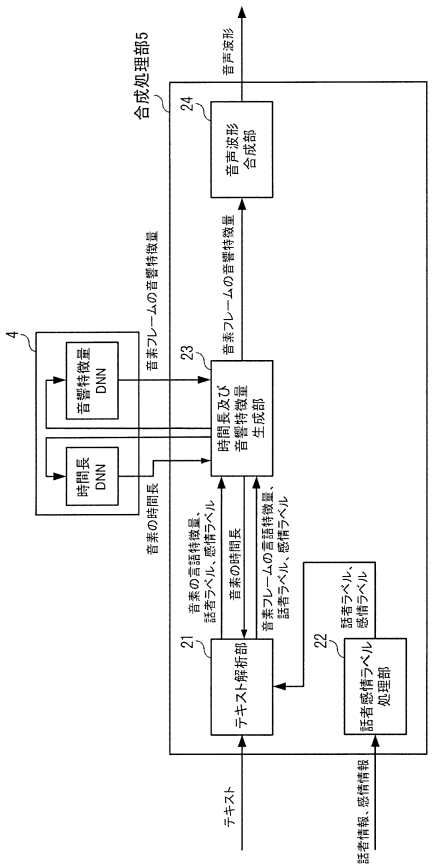
【図３】


テキスト解析部11
31
テキスト
テキスト
解析手段
音素の言語特徴量
32
前処理手段
音素の言語特徴量、
話者ラベル、感情ラベル
33
フレーム
処理手段
音素フレームの言語特徴量、
話者ラベル、感情ラベル
話者ラベル、
感情ラベル
音素の時間長

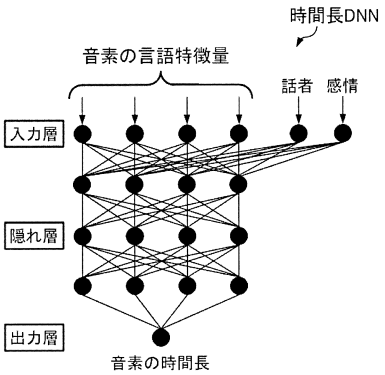
【図４】


音響分析部13
音素ラベル
34
音声波形
音素区切り
処理手段
音素の時間長
音素の区切り位置
35
音響分析手段
音素フレームの音響特徴量

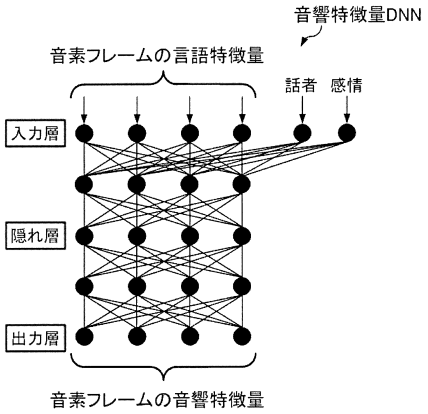
【図５】


合成処理部5
音声波形
24
音声波形
合成部
音素フレームの音響特徴量
4
音響特徴量
DNN
時間長
DNN
音素の時間長
23
時間長及び
音響特徴量
生成部
音素の言語特徴量、
話者ラベル、感情ラベル
音素の時間長
音素フレームの言語特徴量、
話者ラベル、感情ラベル
21
テキスト解析部
22
話者感情ラベル
処理部
話者ラベル、
感情ラベル
テキスト
話者情報、感情情報

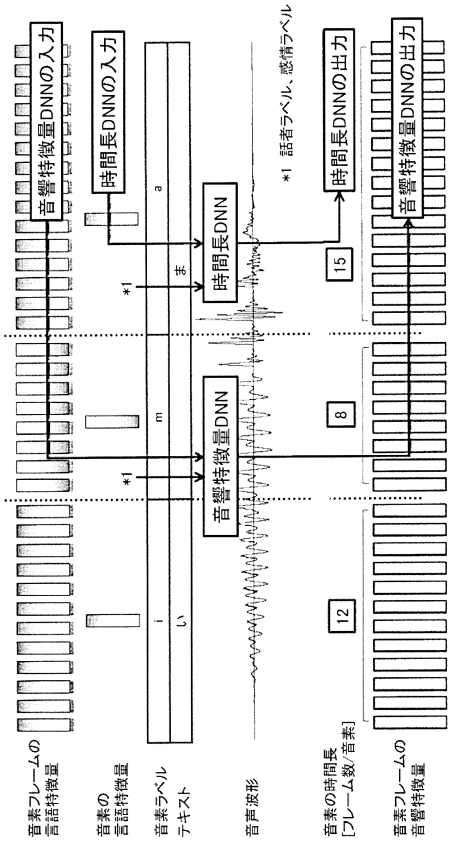
【図６】



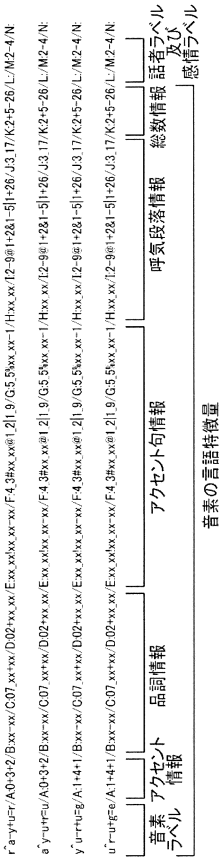
【図７】



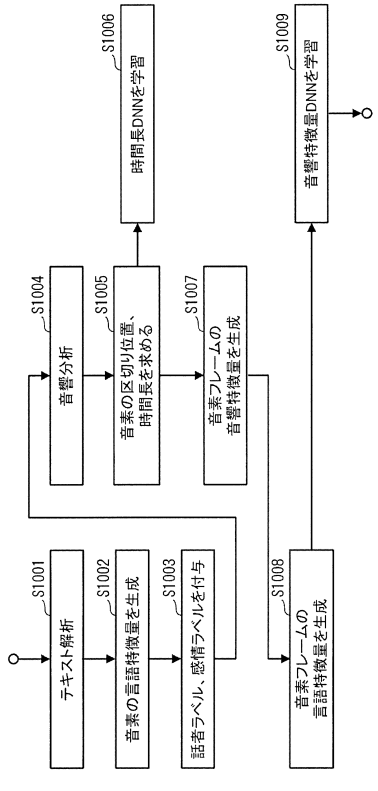
【図８】



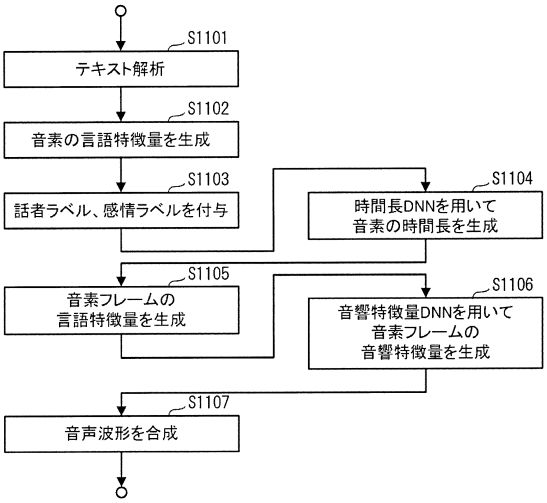
【図９】



【図１０】



【図１１】



合成処理部5の処理例

フロントページの続き

(51)Int.Cl.　　FI
　　G10L　13/033　　102B

(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　上田　雄

(56)参考文献　米国特許第08527276（US，B1）
特開平02－072399（JP，A）
中国特許出願公開第104538024（CN，A）
清山　信正，"解説02 音声合成技術の動向と放送・通信分野における応用展開"，NHK技研R&D，2017年　1月15日，No．161
高木　信二，"とてもDeepなテキスト音声合成"，電子情報通信学会技術研究報告，2017年　1月14日，Vol．116，No．414，pp．41－46
LUONG，Hieu Thi、外3名，"DNNに基づくテキスト音声合成における話者・ジェンダー・年齢コード利用の検討"，電子情報通信学会技術研究報告，2016年10月20日，Vol．116，No．279

(58)調査した分野(Int.Cl.，DB名)
　　G10L　13／00－13／10