

(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)公開番号
特開2023-38957
(P2023-38957A)

(43)公開日 令和5年3月20日(2023.3.20)

(51)国際特許分類		F I	
G 1 0 L	13/08 (2013.01)	G 1 0 L	13/08 1 2 4
G 1 0 L	13/10 (2013.01)	G 1 0 L	13/10 1 1 2 C
G 1 0 L	13/033 (2013.01)	G 1 0 L	13/033 1 0 2 A

審査請求		未請求	請求項の数	15	O L	(全31頁)
(21)出願番号	特願2021-145841(P2021-145841)	(71)出願人	000005108 株式会社日立製作所 東京都千代田区丸の内一丁目6番6号			
(22)出願日	令和3年9月8日(2021.9.8)	(74)代理人	110001678 藤央弁理士法人			
		(72)発明者	本間 健 東京都千代田区丸の内一丁目6番6号 株式会社日立製作所内			
		(72)発明者	孫 慶華 東京都千代田区丸の内一丁目6番6号 株式会社日立製作所内			
		(72)発明者	藤岡 拓也 東京都千代田区丸の内一丁目6番6号 株式会社日立製作所内			

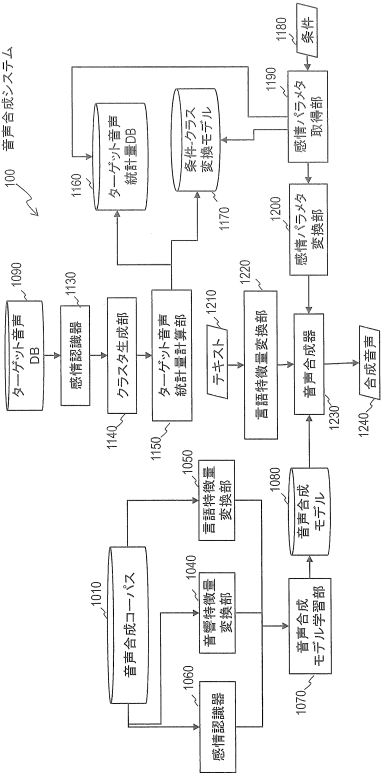
(54)【発明の名称】 音声合成システム及び音声合成する方法

(57)【要約】

【課題】感情音声合成において、人間の感情を模倣した
 音声を得る。

【解決手段】システムは、感情パラメタ及びテキストの
 言語特徴量に基づき、テキストの音響特徴量を生成する
 、音声合成モデルと発話条件をターゲット音声サンプル
 の感情パラメタの統計量と関連付ける変換モデルと、を
 格納する。ターゲット音声サンプルは、感情を模倣する
 対象であるターゲット話者の音声サンプルである。シス
 テムは、入力された現在発話条件を取得する。システム
 は、現在発話条件から、変換モデルを使用して、現在感
 情パラメタを生成する。システムは、現在感情パラメタ
 及び目的発話の言語特徴量に基づき、音声合成モデルを
 使用して、目的発話の音響特徴量を生成する。システム
 は、音響特徴量に基づき合成音声生成する。

【選択図】図1 A



【特許請求の範囲】

【請求項 1】

音声合成システムであって、
1 以上の演算装置と、
1 以上の記憶装置と、を含み、
前記 1 以上の記憶装置は、
感情パラメタ及びテキストの言語特徴量に基づき、前記テキストの音響特徴量を生成する、音声合成モデルと
発話条件をターゲット音声サンプルの感情パラメタの統計量と関連付ける変換モデルと、
を格納し、
前記ターゲット音声サンプルは、感情を模倣する対象であるターゲット話者の音声サンプルであり、
前記 1 以上の演算装置は、
入力された現在発話条件を取得し、
前記現在発話条件から、前記変換モデルを使用して、現在感情パラメタを生成し、
前記現在感情パラメタ及び目的発話の言語特徴量に基づき、前記音声合成モデルを使用して、前記目的発話の音響特徴量を生成し、
前記音響特徴量に基づき合成音声を生成する、音声合成システム。

10

【請求項 2】

請求項 1 に記載の音声合成システムであって、
前記音声合成モデルは、前記ターゲット話者と異なる話者の音声サンプルを使用して訓練されている、音声合成システム。

20

【請求項 3】

請求項 1 に記載の音声合成システムであって、
前記変換モデルは、
前記現在発話条件に対応するクラスを決定し、
前記クラスに対応する感情パラメタの統計量を決定し、
前記感情パラメタの統計量を前記現在感情パラメタとして出力する、音声合成システム。

【請求項 4】

請求項 3 に記載の音声合成システムであって、
前記 1 以上の演算装置は、
前記ターゲット音声サンプルの感情パラメタのクラスタリングを実行して、前記感情パラメタの統計量が属するクラスとの間の関係を決定し、
前記感情パラメタの統計量が属するクラスの情報を前記変換モデルに含める、音声合成システム。

30

【請求項 5】

請求項 4 に記載の音声合成システムであって、
前記クラスタリングが実行される感情パラメタは、同一話者のターゲット音声サンプルから得た感情パラメタの平均値を除いた感情パラメタである、音声合成システム。

40

【請求項 6】

請求項 1 に記載の音声合成システムであって、
前記変換モデルは、入力された前記現在発話条件から前記現在感情パラメタを出力する機械学習モデルである、音声合成システム。

【請求項 7】

請求項 6 に記載の音声合成システムであって、
前記機械学習モデルの学習データの感情パラメタは、同一話者のターゲット音声サンプルから得た感情パラメタの平均を除いた感情パラメタである、音声合成システム。

【請求項 8】

請求項 1 に記載の音声合成システムであって、

50

前記 1 以上の記憶装置は、前記音声合成モデルの学習データを格納し、

前記 1 以上の演算装置は、前記音響特徴量の生成前に、前記変換モデルの変換結果の分布が前記音声合成モデルの学習データの感情パラメタの分布と近づくように、前記現在感情パラメタを変換する、音声合成システム。

【請求項 9】

請求項 1 に記載の音声合成システムであって、

前記 1 以上の演算装置は、前記音響特徴量の生成前に、前記現在感情パラメタを変換パラメタに従って変換し、

合成音声の感情パラメタの分布が、予め指定された感情パラメタの分布に近づくように、前記変換パラメタを更新する、音声合成システム。

10

【請求項 10】

請求項 9 に記載の音声合成システムであって、

前記予め指定された感情パラメタの分布は、前記音声合成モデルの学習データの感情パラメタの分布、前記ターゲット音声サンプルの感情パラメタの分布、又は前記ターゲット音声サンプルの感情パラメタの前記変換パラメタによって変換された感情パラメタの分布、のいずれかである、音声合成システム。

【請求項 11】

請求項 9 に記載の音声合成システムであって、

前記 1 以上の演算装置は、前記合成音声の感情パラメタの分布と前記予め指定された感情パラメタの分布との差が閾値以下となるまで、前記変換パラメタの更新を繰り返す、音声合成システム。

20

【請求項 12】

請求項 9 に記載の音声合成システムであって、

前記 1 以上の演算装置は、前記合成音声の感情パラメタの分布の情報と前記予め指定された感情パラメタの分布情報を提示して、前記変換パラメタのユーザによる編集を受け付ける、音声合成システム。

【請求項 13】

請求項 1 に記載の音声合成システムであって、

前記 1 以上の演算装置は、前記音響特徴量の生成前に、前記変換モデルの変換結果の分布が前記ターゲット音声サンプルの話者の感情パラメタ分布に近づくように、前記現在感情パラメタを変換する、音声合成システム。

30

【請求項 14】

請求項 1 に記載の音声合成システムであって、

前記 1 以上の演算装置は、

前記音響特徴量の生成前に、前記現在感情パラメタを変換パラメタに従って変換し、

基準音声サンプル及び現在の前記変換パラメタによる合成音声の感情パラメタの分布の間の対応点のユーザからの指定を受け付け、前記対応点に基づき前記変換パラメタを更新する、音声合成システム。

【請求項 15】

システムが音声を合成する方法であって、

40

前記システムは、感情パラメタ及びテキストの言語特徴量に基づき、前記テキストの音響特徴量を生成する、音声合成モデルと

発話条件をターゲット音声サンプルの感情パラメタの統計量と関連付ける変換モデルと、を格納し、

前記ターゲット音声サンプルは、感情を模倣する対象であるターゲット話者の音声サンプルであり、

前記方法は、前記システムが、

入力された現在発話条件を取得し、

前記現在発話条件から、前記変換モデルを使用して、現在感情パラメタを生成し、

前記現在感情パラメタ及び目的発話の言語特徴量に基づき、前記音声合成モデルを使用

50

して、前記目的発話の音響特徴量を生成し、

前記音響特徴量に基づき合成音声を生成する、ことを含む方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、音声合成技術に関する。

【背景技術】

【0002】

人が機械とやりとりを行うヒューマンインタフェースにおいて、ユーザが機械と音声で会話できる対話システムが存在する。対話システムの用途は、人からの話しかけを受け付けるだけでなく、対話システムから人に適切に働きかけることによる人の生活のサポートも考えられる。

10

【0003】

対話システムが会話により人間に適切に働きかけるためには、対話システムが何を言うかだけでなく、こういった感情を音声に込めて言うかが重要である。人が人に何かをお願いするとき、声に乗せる感情を適切に制御しながら、相手の人に話しかける。対話システムにおいても、そういった音声に感情を載せることが望ましい。

【0004】

これまでも、感情を声に乗せて生成する感情音声合成技術は存在する。この技術では、指定された感情を表す多次元ベクトルにしたがって、感情を持った音声合成される。この感情指定に使われる入力を、感情パラメタと呼ぶこととする。

20

【0005】

対話システムが音声により適切に働きかけるためには、対話システムの用途と似たタスクを行っている人間がどのように発話しているかを調べ、その発話音声の感情を真似た音声を生成する方法が有効であると考えられる。

【0006】

この方法を実現する技術として、特許文献1では、マイクからの入力音声に対して、感情を1個のカテゴリとして認識し、そのカテゴリを音声合成の出力音声の感情として指定する技術が開示されている。これにより、人間が発話した音声サンプルを用意することで、その発話音声の感情を真似た音声を生成することができる。

30

【先行技術文献】

【特許文献】

【0007】

【特許文献1】特願2006-113546号公報

【発明の概要】

【発明が解決しようとする課題】

【0008】

従来の感情音声合成技術では、感情を決めるためには、開発者が感情音声を生成し、それを聞いて、所望する音声となるまで感情パラメタを調整する必要があった。開発者は、この調整を試行錯誤的に行うため、時間が長くかかる問題があった。

40

【0009】

他の方法として、人を真似た感情を持つ音声を生成したい場合、入力音声の一つの感情カテゴリを認識し、そのカテゴリを音声合成の出力音声の感情として指定することができる。しかし、人間が音声に乗せる感情は多様であり、感情の種類として少数のカテゴリから選ぶ構成では、人間の感情を十分に表現することは困難である。

【0010】

また、同一の状況において複数の人間が発話した場合、その音声に乗る変化は多様であり、かつ感情も多様である。人間が声によって行ってきた働きかけを機械で代替することを考えた場合、こういった発話者ごとのバラつきを考慮したうえで音声を合成すべきである。しかし、1個の音声に基づいて感情を決定する方法は、この要求を満たすことができ

50

ない。

【課題を解決するための手段】

【0011】

本発明の一態様の音声合成システムは、1以上の演算装置と、1以上の記憶装置と、を含む。前記1以上の記憶装置は、感情パラメタ及びテキストの言語特徴量に基づき、前記テキストの音響特徴量を生成する、音声合成モデルと発話条件をターゲット音声サンプルの感情パラメタの統計量と関連付ける変換モデルと、を格納する。前記ターゲット音声サンプルは、感情を模倣する対象であるターゲット話者の音声サンプルである。前記1以上の演算装置は、入力された現在発話条件を取得し、前記現在発話条件から、前記変換モデルを使用して、現在感情パラメタを生成し、前記現在感情パラメタ及び目的発話の言語特徴量に基づき、前記音声合成モデルを使用して、前記目的発話の音響特徴量を生成し、前記音響特徴量に基づき合成音声を生成する。

10

【発明の効果】

【0012】

本発明の一態様を用いれば、人間が話す音声の感情をより適切に再現する音声合成を行うことができる。

【図面の簡単な説明】

【0013】

【図1A】本明細書の一実施形態に係る音声合成システムの論理構成例を示す。

【図1B】本明細書の一実施形態に係る音声合成システムの他の論理構成例を示す。

20

【図2】音声合成コーパスの構成例を示す。

【図3】ターゲット音声データベースの構成例を示す。

【図4】ターゲット音声統計量データベースの構成例を示す。

【図5】条件ークラス変換モデルの構成例を示す。

【図6】音声合成システムのハードウェア構成例を模式的に示したブロック図である。

【図7】音声合成モデルの学習手順例のフローチャートである。

【図8】ターゲット音声データベースから、条件と感情パラメタを関連付けるモデルを生成する処理例のフローチャートを示す。

【図9A】クラスタリングベースモデル化の例を示すフローチャートである。

【図9B】平均値ベースモデル化の例を示すフローチャートである。

30

【図9C】機械学習ベースモデル化の例を示すフローチャートである。

【図10】音声合成の方法例を示すフローチャートである。

【図11】第2の実施形態に係る音声合成システムの論理構成例を模式的に示す。

【図12】機械学習モデルから得られた感情パラメタから統計量を求める手順を説明する図である。

【図13】第3の実施形態の音声合成システムの論理構成例を示す。

【図14】出力音声統計量データベースの構成例を示す。

【図15】感情パラメタの変換パラメタ更新処理例を示すフローチャートである。

【図16A】第4の実施形態において、音声合成用の感情パラメタに対して開発者が調整する機能を説明するための図である。

40

【図16B】第4の実施形態において、音声合成用の感情パラメタに対して開発者が調整する機能を説明するための図である。

【図16C】第4の実施形態において、音声合成用の感情パラメタに対して開発者が調整する機能を説明するための図である。

【図17】第5の実施形態において、感情パラメタの変換パラメタの調整方法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】

【0014】

以下、図面を用いて、本発明のいくつかの実施形態を説明する。なお、以下の実施形態においては便宜上その必要があるときは、複数のセクション又は実施形態に分割して説明

50

するが、特に明示した場合を除き、それらはお互い無関係なものではなく、一方は他方の一部又は全部の変形例、詳細、補足説明などの関係にある。各実施形態は、個別に実施しても良いが、組合せて実施しても良い。

【0015】

また、以下の実施形態において、要素の数など（個数、数値、量、範囲などを含む）に言及する場合、特に明示した場合及び原理的に明らかに特定の数に限定される場合などを除き、その特定の数に限定されるものではなく、特定の数以上でも以下でも良いものとする。

【0016】

さらに、以下の実施形態において、その構成要素（要素ステップなどを含む）は、特に明示した場合及び原理的に明らかに必須であると考えられる場合などを除き、必ずしも必須のものではないことは言うまでもない。

【0017】

同様に、以下の実施形態において、構成要素などの形状、位置関係等と言及するときは、特に明示した場合や原理的に明らかにそうでないと考えられる場合などを除き、実質的にその形状などに近似又は類似するものなどを含むものとする。このことは数値及び範囲についても同様である。

【0018】

また、以下の説明では、「 $x \times x$ テーブル」といった表現で説明することがあるが、当該情報は、どのような構造のデータでも良い。また、以下の説明において、各表の構成は一例であり、1つの表は、2以上の表に分割されても良いし、2以上のテーブルの全部又は一部が1つのテーブルであっても良い。また、以下の説明では、「 $x \times x$ 部」の表現にて機能を説明することがあるが、当該機能は、1以上の演算装置によって1以上のコンピュータプログラムが実行されることで実現されても良い。

【0019】

以下において説明される本明細書の一実施形態に係る音声合成システムは、感情を含む人間の声を合成し、出力する。これにより、人に対してより適切に働きかけ、人の活動をより適切にサポートすることができる。

【0020】

人間が音声に乗せる多様な感情は、例えば、多次元ベクトルによって表現することができる。多次元ベクトルのそれぞれの要素には、たとえば *Arousal*（活性度）、*Pleasantness*（感情価）、*Dominance*（支配度）といったものが対応する。多次元ベクトルは、多くの感情のバリエーションに対応できる。

【0021】

また、同一の状況において複数の人間が発話した場合、その音声に乗る変化は多様であり、かつ感情も多様である。人間が行ってきた声による働きかけを機械で代替することを考えた場合、こういった発話者ごとのバラつきを考慮したうえで合成すべきである。

【0022】

また、感情認識器が出力した感情パラメタを、直接音声合成器に指定したとしても、生成される音声の感情が開発者の所望のものであるとは限らない。なぜならば、音声合成の音源となった話者と、感情を真似るターゲットである話者が異なる場合があるためである。この話者の異なりが、音声合成による音声を不自然にする可能性がある。

【0023】

本明細書の一実施形態に係る音声合成システムは、音声合成モデル及び音声合成器を含む。音声合成モデルは、言語特徴量及び感情パラメタを入力として受け取り、音響特徴量を出力する。言語特徴量は音声の内容を示すテキストから得られる特徴量である。感情パラメタは、音声に乗せる感情を示すパラメタである。音響特徴量は音声（音）の特徴量である。音声合成器は、音声合成モデルからの音響特徴量から合成音声を生成する。

【0024】

本明細書の一実施形態の音声合成システムは、感情認識器を使用して、音声合成モデル

10

20

30

40

50

の学習コーパス（音声合成コーパス）における音声データから、言語特徴量及び感情パラメタを得る。上述のように、音声合成システムは、音声合成モデルの学習において、言語特徴量に加えて感情パラメタを入力として用いる。

【0025】

音声合成モデルは、入力された言語特徴量及び感情パラメタから、音響特徴量を出力する。学習コーパスにおける音声データから抽出された音響特徴量と、音声合成モデルから出力された音響特徴量の誤差に基づき、音声合成モデルのパラメタが更新される（音声合成モデルの学習又は訓練）。これにより、感情パラメタを制御できる音声合成モデルを得ることができる。

【0026】

本明細書の一実施形態の音声合成システムは、音声合成システムが感情を真似る対象であるターゲット音声サンプルを格納するターゲット音声データベースを保持し、ターゲット音声サンプルそれぞれに対して感情認識器を使い感情パラメタを得る。音声合成システムは、感情パラメタ統計量を算出して、データベースに格納する。

【0027】

音声合成システムは、感情サンプルの統計量と、発話対象者又は対話システムが置かれているシチュエーションを示す条件（発話条件）とを関係付けるモデルを生成する。感情パラメタの統計量を使用することで、より普遍的な感情を伴う合成音声を生成できる。

【0028】

音声合成システムは、発話対象者又は対話システムが置かれている現在のシチュエーションを示す条件を取得して、上記モデルにより感情パラメタを生成する。音声合成システムは、目的の発話内容の言語特徴量と生成した感情パラメタを学習済みの音声合成モデルに入力して、ターゲット音声を真似た感情を伴う音響特徴量を得る。音声合成器は、その音響特徴量から、発話される合成音声を生成する。

【0029】

本明細書の一実施形態の音声合成システムは、学習コーパスとターゲット音声データベースの感情パラメタの分布を求め、これらの差異に基づき、ターゲット音声から得られる感情パラメタの統計量を変更する。これにより、学習コーパスとターゲット音声との間の感情表現の異なりを吸収し、より自然な感情表現を伴う音声を合成できる。

【0030】

<第1の実施形態>

本実施形態では、学習コーパスである音声合成コーパスと、ターゲット音声と、に対して感情認識器を適用する。音声合成モデルの学習は、音声合成コーパスから得られた感情パラメタを、学習データとして用いる。音声合成モデルによる音声合成時には、ターゲット音声から得られた感情パラメタを、音声合成モデルに対する入力として用いる。

【0031】

音声合成コーパスは、例えば、発話することを専門とする声優の発話音声データ及び発話の内容を示すテキストを格納する。ターゲット音声は、音声合成システムが模倣を目的とする人の音声である。例えば、特定職業（例えば介護や接客）についている人々による発話の音声データ及びテキストが、データベースに格納される。これにより、発話対象者又は対話システムのシチュエーションに適した感情を表す、聞き取りやすい発話が実現できる。

【0032】

図1Aは、本明細書の一実施形態に係る音声合成システムの論理構成例を示す。音声合成システム100は、音声合成コーパス1010、音響特徴量変換部1040、言語特徴量変換部1050、感情認識器1060、音声合成モデル学習部1070、及び音声合成モデル1080を含む。

【0033】

音声合成システム100は、さらに、ターゲット音声データベース(DB)1090、感情認識器1130、クラスタ生成部1140、ターゲット音声統計量計算部1150、

10

20

30

40

50

ターゲット音声統計量データベース１１６０、条件ークラス変換モデル１１７０を含む。さらに、音声合成システム１００は、感情パラメタ取得部１１９０、感情パラメタ変換部１２００、言語特徴量変換部１２２０、及び音声合成器１２３０を含む。

【００３４】

音声合成システム１００は、対話システムが置かれているシチュエーションを示す条件（発話条件）１１８０を受け取り、その条件１１８０に対応する感情パラメタを算出する。算出する感情パラメタは、既存サンプルの感情パラメタの統計量に基づく値である。音声合成システム１００は、生成した感情パラメタと発話内容を示すテキスト１２１０から合成音声１２４０を生成する。これにより、シチュエーションに即した感情を伴う音声を発話することができる。

【００３５】

図１Ａに示す構成例において、感情パラメタ取得部１１９０は、条件ークラス変換モデル１１７０とターゲット音声統計量データベース１１６０を使用して、入力条件１１８０から、感情パラメタを生成する。条件ークラス変換モデル１１７０とターゲット音声統計量データベース１１６０は、入力条件とターゲット音声サンプルの感情パラメタ統計量とを関連付けるモデルである。後述するように、他の構成例において、音声合成システム１００は、機械学習モデルを使用して、入力条件１１８０から、感情パラメタを生成する。

【００３６】

図１Ｂは、本明細書の一実施形態に係る音声合成システム１００の他の論理構成例を示す。図１Ｂに示す音声合成システム１００は、図１Ａに示す構成例の条件ークラス変換モデル１１７０とターゲット音声統計量データベース１１６０に代えて、条件ー感情パラメタ変換モデル１１６５を含む。また、クラスタ生成部１１４０に代えて正規化部１１４５を含む。

【００３７】

条件ー感情パラメタ変換モデル１１６５は機械学習モデルであって、入力された条件１１８０に対して感情パラメタを出力する。条件ー感情パラメタ変換モデル１１６５は、ターゲット音声サンプルの感情パラメタを使用して訓練される。したがって、条件ー感情パラメタ変換モデル１１６５は、入力条件とターゲット音声サンプルの感情パラメタ統計量とを関連付けるモデルである。なお、図１Ａ及び１Ｂに示す双方の構成が一つの音声合成システムに含まれ、いずれか一方のモードが選択可能であってもよい。

【００３８】

図２は、音声合成コーパス１０１０の構成例を示す。音声合成コーパス１０１０は、音声合成モデル１０８０を学習するためのデータを格納する。音声合成コーパス１０１０は、音声サンプル欄１０２０及びテキスト欄１０３０を含む。各レコードは、特定の発話の音声データ及びその発話内容のテキストを示す。なお、音声合成コーパス１０１０は、音響特徴量変換部１０４０、言語特徴量変換部１０５０、感情認識器１０６０の少なくとも一部の変換結果を保持していてもよい。

【００３９】

図３は、ターゲット音声データベース１０９０の構成例を示す。ターゲット音声データベース１０９０は、ターゲット話者の発話の情報を格納している。具体的には、ターゲット音声データベース１０９０は、話者ＩＤ欄１１００、音声サンプル欄１１１０、テキスト欄１１２０、及び条件欄１１２５を含む。各レコードは、一つの発話の情報を示す。

【００４０】

話者ＩＤ欄１１００は、発話者のＩＤを示す。音声サンプル欄１１１０は、発話の音声データを示す。テキスト欄１１２０は、発話した内容のテキストを示す。条件欄１１２５は、音声が発話された条件を示す。

【００４１】

条件欄１１２５には、音声合成システム１００の想定される用途における、人の音声の感情を変える可能性がある、あらゆる条件を設定することができる。例えば、発話が実行された場所や時間帯が挙げられる。場所の例は、食堂、部屋、談話室等である。また、タ

10

20

30

40

50

ターゲット話者が相手にどういう目的で発話したかを示す条件も入れておいてもよい。

【0042】

例えば、相手を励ます目的、相手をいたわる目的、挨拶等を、異なる条件として扱うことができる。条件は、1変数で表現されてもよく、複数変数で表現されてもよい。複数変数での表現の場合、発話が行われた1つのシチュエーションは、複数の条件変数の組、すなわちタプルとして表される。シチュエーションは、発話の、場所、時間帯、発話目的等を含む。なお、ターゲット音声データベース1090は、感情認識器1060は、感情認識器1130の変換結果を格納していてもよく、テキストが省略されていてもよい。

【0043】

図4は、図1Aに示すターゲット音声統計量データベース1160の構成例を示す。ターゲット音声統計量データベース1160は、ターゲット音声データベース1090に格納されているターゲット音声（レコード）の統計情報を格納している。具体的には、ターゲット音声統計量データベース1160は、クラスID欄1161、及び感情パラメタ欄1162を含む。

【0044】

クラスID欄1161は、ターゲット音声データベース1090のターゲット音声それぞれが属するクラスのIDを示す。感情パラメタ欄1162は、各クラスにおけるターゲット音声の感情パラメタの統計量を示す。統計量は、例えば、平均値、最頻値、中央値等である。以下の説明において、感情パラメタの変数それぞれの平均値が、例として使用される。図4において、感情パラメタは多次元ベクトルで表され、変数それぞれが異なる感情を示す。感情パラメタ欄1162の感情パラメタは、対応するクラスにおける変数それぞれの平均値を示す。

【0045】

図5は、図1Aに示す条件ークラス変換モデル1170の構成例を示す。条件ークラス変換モデル1170は、合成音声の生成のために指定された発話の条件からクラスIDを決定する。出力されたターゲットIDに対応する感情パラメタが、ターゲット音声統計量データベース1160から決定される。図5に示す構成例において、条件ークラス変換モデル1170は決定木である。条件ークラス変換モデルの構成は任意である。

【0046】

図6は、音声合成システム100のハードウェア構成例を模式的に示したブロック図である。音声合成システム100は、演算性能を有する演算装置101と、演算装置101が実行するプログラム及び処理対象データを格納する記憶領域を与える主記憶装置102と、を含む。演算装置101は、例えば、1又は複数のコアを含むCPUであり、主記憶装置102は、例えば、揮発性記憶領域を含むRAMである。

【0047】

音声合成システム100は、さらに、他の計算機装置や外部記憶装置とデータ通信をおこなう通信インターフェース106と、HDD（Hard Disk Drive）やフラッシュメモリなどを利用した不揮発記憶領域を与える補助記憶装置103と、を含む。また、音声合成システム100は、ユーザからの操作を受け付ける入力装置104と、各プロセスでの出力結果をユーザに提示する出力装置105と、を含む。

【0048】

入力装置104は、例えば、キーボードやマウスを含み、出力装置105は、例えばモニタやプリンタを含む。音声合成システム100のこれら構成要素は、内部バス107を介して通信可能である。

【0049】

演算装置101が実行するプログラム及び処理対象のデータは、例えば、補助記憶装置103から主記憶装置102にロードされる。図1A又は1Bに示す機能部は、演算装置101が対応するプログラムを実行することによって実装することができる。図1A又は1Bに示すコーパスやデータベース等の各種データは、例えば、補助記憶装置103に格納され得る。

10

20

30

40

50

【0050】

音声合成システム100は、物理的な計算機システム（一つ以上の物理的な計算機）でもよいし、クラウド基盤のような計算リソース群（複数の計算リソース）上に構築されたシステムでもよい。音声合成システム100は、スマートフォンやタブレットなどの携帯機器でもよい。計算機システムあるいは計算リソース群は、1以上のインターフェース装置、1以上の記憶装置（例えば、主記憶装置及び補助記憶装置を含む）、及び、1以上の演算装置を含む。

【0051】

プログラムが演算装置によって実行されることで機能が実現される場合、定められた処理が、適宜に記憶装置及び／またはインターフェース装置等を用いながら行われるため、機能は演算装置の少なくとも一部とされてもよい。機能を主語として説明された処理は、演算装置あるいはそのプロセッサを有するシステムが行う処理としてもよい。

10

【0052】

プログラムは、プログラムソースからインストールされてもよい。プログラムソースは、例えば、プログラム配布計算機または計算機が読み取り可能な記憶媒体（例えば計算機読み取り可能な非一過性記憶媒体）であってもよい。各機能の説明は一例であり、複数の機能が一つの機能にまとめられたり、一つの機能が複数の機能に分割されたりしてもよい。

【0053】

図1A又は1Bに戻って、音声合成モデル学習部1070は、音声合成モデル1080を訓練する。この訓練のため、音声合成モデル学習部1070は、音声合成コーパス1010から取得したレコードのテキストを、言語特徴量変換部1050に入力して、言語特徴量を取得する。音声合成モデル学習部1070は、そのレコードの音声サンプルを音響特徴量変換部1040に入力して、音響特徴量を取得する。さらに、音声合成モデル学習部1070は、その音声サンプルを感情認識器1060に入力して、感情パラメタを取得する。

20

【0054】

音声合成モデル学習部1070は、音声合成モデル1080に、言語特徴量及び感情パラメタを入力する。音声合成モデル1080は、入力に対応する音響特徴量を出力する。音声合成モデル1080が出力した音響特徴量と音響特徴量変換部1040からの音響特徴量の誤差に基づき、音声合成モデル1080の構成パラメタが更新される。

30

【0055】

音声合成器1230は、音声合成モデル1080から出力された音響特徴量から、合成音声1240を生成する。従って、音声合成モデル1080は、最終的に合成音声1240の音声波形が生成されるように学習をすることになる。上述のように、本実施形態では、感情パラメタの形式として、多次元ベクトルを想定する。それぞれの次元は、感情の異なる側面の強さを表す。なお、ベクトルの次元数は1以上の任意の数であり、一つの次元が複数種類の感情を示すことができる。

【0056】

音声合成モデル1080の学習手順の詳細を説明する。図7は、音声合成モデル1080の学習手順例のフローチャートである。ステップS101において、音声合成モデル学習部1070は、音声合成コーパス1010内の各音声サンプルを感情認識器1060に入力し、感情パラメタを得る。

40

【0057】

ステップS102において、音声合成モデル学習部1070は、音声合成コーパス1010内の各音声サンプルを音響特徴量変換部1040に入力し、音響特徴量を得る。ステップS103において、音声合成モデル学習部1070は、音声合成コーパス1010内の各テキストを言語特徴量変換部1050に入力し、言語特徴量を得る。

【0058】

ステップS104において、音声合成モデル学習部1070は、音声合成モデル1080

50

0の学習（訓練）を行う。具体的には、音声合成モデル学習部1070は、音声合成モデル1080に、ステップS101及びS103で取得した感情パラメタと言語特徴量を入力し、出力された音響特徴量を取得する。音声合成モデル学習部1070は、出力された音響特徴量とステップS102で取得した音響特徴量の誤差が小さくなるように、音声合成モデル1080を訓練する。つまり、音声合成モデル1080に言語特徴量と感情パラメタを入力したときに、正しい音響特徴量が生成される確率が最大化されるように学習を行う。

【0059】

図1A又は1Bに戻って、ターゲット音声データベース1090にある音声サンプルは、感情認識器1130に入力される。感情認識器1130は、音声合成コーパス1010に対して使用する感情認識器1060と異なるプログラム又は同一のプログラムで実現される。一例において、感情認識器1130及び1060は、同一である。

10

【0060】

図1Aに示す構成例において、感情認識器1130が出力した感情パラメタは、クラスタ生成部1140に入力される。クラスタ生成部1140は、ターゲット音声データベース1090のレコードのクラスタリングを行い、レコードそれぞれが属するクラスを生成する。クラスの生成方法の詳細は後述する。

【0061】

図1Aの構成例において、クラスの生成結果は、ターゲット音声統計量計算部1150に入力される。ターゲット音声統計量計算部1150は、各クラスに属するレコードの統計量を計算して、ターゲット音声統計量データベース1160に保存する。図4を参照して説明したように、統計量は、各クラスのレコードの感情パラメタの平均値である。ターゲット音声統計量計算部1150は、さらに、条件－クラス変換モデル1170を生成する。生成された条件－クラス変換モデル1170は、補助記憶装置103に格納される。

20

【0062】

図1Bの構成例において、感情認識器1130が出力した感情パラメタは、正規化部1145に入力される。正規化部1145は、感情パラメタの正規化処理を行う。図1Aの構成例において、正規化処理は、クラスタ生成部1140により実行される。正規化処理の詳細は後述する。なお、正規化処理は、正規化モードが指定されている場合に実行される。

30

【0063】

図1Bの構成例において、ターゲット音声統計量計算部1150は、正規化部1145により正規化された感情パラメタ又は正規化が省略された感情パラメタと、ターゲット音声データベース1090の条件欄1125が示す条件との組み合わせによって、条件－感情パラメタ変換モデル1165の訓練（学習）を行う。

【0064】

ターゲット音声データベース1090に対する上記処理の詳細を説明する。図8は、ターゲット音声データベース1090から、条件と感情パラメタを関連付けるモデルを生成する処理例のフローチャートを示す。

【0065】

ステップS201において、感情認識器1130は、ターゲット音声データベース1090内の音声サンプルそれぞれから、感情パラメタを求める。ステップS202において、クラスタ生成部1140又は正規化部1145は、事前設定に従って、正規化モードを適用するか否かを判定する。正規化モードは、ターゲット話者の音声における話者性の影響を減らす処理を有効にするモードである。

40

【0066】

ステップS203及びS204は、正規化モードが有効であるときの処理である。ステップS203において、クラスタ生成部1140又は正規化部1145は、ターゲット音声データベース1090内の各話者の感情パラメタの平均値を求める。具体的には、クラスタ生成部1140は、話者IDが同一のレコードを抽出し、抽出した全てのレコードの

50

感情パラメタの各変数の平均値を求める。

【0067】

ステップS204において、クラスタ生成部1140又は正規化部1145は、感情認識器1130による各ターゲット音声サンプルの感情パラメタから、各ターゲット音声サンプルの話者の感情パラメタの平均値を減算する。これにより、各話者の基準レベルからの感情の変化を示す感情パラメタを得ることができる。

【0068】

ステップS205において、ターゲット音声統計量計算部1150は、ターゲット音声データベース1090の条件欄1125が示す条件と、ターゲット音声サンプルの感情パラメタの統計量との関係を、モデル化する。

【0069】

入力された条件から感情パラメタ統計量を計算する方法、つまり、入力条件と感情パラメタ統計量とを関連付けるモデルのいくつかの例が存在する。以下において、クラスタリングベースモデル化、平均値ベースモデル化、機械学習ベースモデル化を説明する。クラスタリングベースモデル化は図1Aの構成例に対応し、機械学習ベースモデル化は図1Bの構成例に対応する。平均値ベースモデル化は、図1Aの構成例からクラスタ生成部1140を省略した構成例に対応する。

【0070】

まず、クラスタリングベースモデル化を説明する。図9Aは、クラスタリングベースモデル化の例を示すフローチャートである。ステップS311において、クラスタ生成部1140は、感情認識器1130から取得したターゲット音声サンプルの感情パラメタに対して、非教師クラスタリングを実施する。これにより、いくつかのクラス（クラスタ）が生成される。

【0071】

ステップS312において、ターゲット音声統計量計算部1150は、各クラスに属するターゲット音声サンプルの感情パラメタの平均値（各変数の平均値）を計算する。平均値が計算される前の感情パラメタは、正規化モードにおいて、話者の感情パラメタ平均値が減算されたベクトルであり、非正規化モードにおいて感情認識器1130から出力されたベクトルである。

【0072】

ステップS313において、ターゲット音声統計量計算部1150は、ターゲット音声データベース1090の条件欄1125が示す条件を説明変数とし、ステップS311で決定されたターゲット音声サンプルそれぞれのクラスIDを出力とする決定木を、訓練する。ステップS314において、ターゲット音声統計量計算部1150は、訓練された決定木を条件－クラス変換モデル1170として保存する。

【0073】

ステップS315において、ターゲット音声統計量計算部1150は、ステップS312で計算した各クラスの感情パラメタの平均値をターゲット音声統計量データベース1160の感情パラメタ欄1162に格納する。対応するクラスIDは、クラスID欄1161に格納される。

【0074】

ステップS313で生成された決定木に対して、合成音声を生成するときのシチュエーションを表す条件1180を入力することにより、対応するクラスIDを同定することができる。同定されたクラスに属する感情パラメタと発話テキストの言語特徴量を音声合成モデル1080に入力することにより、そのシチュエーションにふさわしい感情の音声を生成することが可能となる。なお、条件を入力としてクラスIDを出力するモデルであれば、決定木と異なるモデリング手法を用いてよい。

【0075】

次に、平均値ベースモデル化を説明する。図9Bは、平均値ベースモデル化の例を示すフローチャートである。まず、ステップS321において、条件をクラスに変換するテ

10

20

30

40

50

ブルが用意される。このテーブルは開発者が作成する想定でよい。最も単純な場合、条件の組み合わせそれぞれに対して、1個のクラスが対応する関係を採用することができる。

【0076】

ステップS322において、ターゲット音声統計量計算部1150は、各クラスに属するターゲット音声サンプルの感情パラメタの平均値を計算する。ターゲット音声サンプルのクラスは、上記テーブルにより決定される。ステップS323において、ターゲット音声統計量計算部1150は、ステップS321で得られたテーブルを、条件－クラス変換モデル1170として保存する。

【0077】

ステップS324において、ターゲット音声統計量計算部1150は、ステップS322で求めた各クラスの感情パラメタの平均値を、ターゲット音声統計量データベース1160の感情パラメタ欄1162に格納する。対応するクラスIDは、クラスID欄1161に格納される。

【0078】

平均値ベースの手法は単純であるが、音声を生成したいシチュエーションの内訳が明確であり、かつそれぞれのシチュエーションにおけるターゲット話者の音声を採取可能である場合、本手法は特に有効である。

【0079】

次に、機械学習ベースモデル化を説明する。図9Cは、機械学習ベースモデル化の例を示すフローチャートである。ステップS331において、ターゲット音声統計量計算部1150は、条件を入力変数とし、感情パラメタを出力（すなわち予測対象変数）とするデータセットを作成する。ステップS332において、ターゲット音声統計量計算部1150は、データセットに基づき入力を変換する機械学習モデルを生成する。ステップS333において、ターゲット音声統計量計算部1150は、生成した機械学習モデルを、図1Bに示す、条件－感情パラメタ変換モデル1165として保存する。

【0080】

本手法では、シチュエーションを表す条件を指定すると、機械学習モデルが感情パラメタを直接出力する。よって、ターゲット音声統計量データベースを省略できる。訓練済みの機械学習モデルは、ターゲット音声データベース1090の音声サンプルの感情パラメタに基づく感情パラメタを出力する。機械学習モデルの出力は、音声サンプルの感情パラメタ統計量を表す。

【0081】

図1A又は1Bに戻り、学習済み音声合成モデル及びターゲット音声サンプルの感情パラメタの統計量に基づく音声合成を、説明する。条件1180は、対話システムによって音声が發せられるシチュエーションを表す変数である。条件1180は、ターゲット音声データベース1090の条件欄1125に規定される変数で表される。音声合成システム100の外部システム（不図示）から、条件1180は入力される。条件1180は、感情パラメタ取得部1190に入力される。

【0082】

図1Aの構成例において、感情パラメタ取得部1190は、条件－クラス変換モデル1170を使用し、指定された条件1180に対応するクラスを決定する。さらに、感情パラメタ取得部1190は、決定されたクラスに対応する感情パラメタを、ターゲット音声統計量データベース1160の感情パラメタ欄1162から読み込む。図1Bの構成例において、条件－感情パラメタ変換モデル1165によって、入力条件1180から感情パラメタが生成される。

【0083】

感情パラメタ変換部1200は、本実施形態ではとくに何も処理をしない。そのため、感情パラメタ取得部1190が得た感情パラメタを、無変換で、そのまま音声合成器1230に入力する。

【0084】

10

20

30

40

50

テキスト 1 2 1 0 は、音声合成で生成したい文を規定した文字列である。言語特徴量変換部 1 2 2 0 は、入力されたテキスト 1 2 1 0 を音声合成器 1 2 3 0 の入力となる言語特徴量に変換する。この処理は、言語特徴量変換部 1 0 5 0 の処理と同様である。

【 0 0 8 5 】

音声合成器 1 2 3 0 は、入力された言語特徴量及び感情パラメタを、音声合成モデル 1 0 8 0 に入力し、音響特徴量を取得する。音声合成器 1 2 3 0 は、音響特徴量から合成音声 1 2 4 0 を生成し、出力する。

【 0 0 8 6 】

音声合成の手順の詳細を説明する。図 1 0 は、音声合成の方法例を示すフローチャートである。ステップ S 4 0 1 において、ターゲット音声サンプルの感情パラメタに対する統計量計算の方法が、機械学習ベースか否かにより分岐する。

10

【 0 0 8 7 】

機械学習ベースと異なる手法が使用されている場合（S 4 0 1：NO）、フローはステップ S 4 0 2 に進む。感情パラメタ取得部 1 1 9 0 は、条件ークラス変換モデル 1 1 7 0 を使用して、入力された条件 1 1 8 0 からクラス ID を決定する。さらに、ステップ S 4 0 3 において、感情パラメタ取得部 1 1 9 0 は、クラス ID に対応する感情パラメタを、ターゲット音声統計量データベース 1 1 6 0 の感情パラメタ欄 1 1 6 2 から読み込む。

【 0 0 8 8 】

機械学習ベースの手法が使用されている場合（S 4 0 1：YES）、フローはステップ S 4 1 0 に進む。機械学習ベースを用いる場合、条件ークラス変換モデル 1 1 7 0 は、条件を入力すると感情パラメタを出力する機械学習モデルである。感情パラメタ取得部 1 1 9 0 は、条件ークラス変換モデル 1 1 7 0 に対して条件 1 1 8 0 を入力し、感情パラメタを得る。

20

【 0 0 8 9 】

ステップ S 4 0 4 において、感情パラメタ取得部 1 1 9 0 は、指定されたモードが正規化モードか否か判定する。正規化モードが有効である場合（S 4 0 4：YES）、フローはステップ S 4 0 5 に進む。正規化モードの場合には、ステップ S 4 0 3 またはステップ S 4 1 0 で得られた感情パラメタは、絶対値ではなく、1 話者の音声に含まれる平均的な感情パラメタからの変動（相対値）である。

【 0 0 9 0 】

よって、ステップ S 4 0 5 において、感情パラメタ取得部 1 1 9 0 は、この感情パラメタに対して、特定の話者の平均的な感情パラメタを加える。ここで加得る感情パラメタを、どの話者から得るかは任意である。例えば、ターゲット話者のなかでもっとも好ましい音声で話している話者を選ぶ方法や、音声合成コーパス 1 0 1 0 の音声を発話した話者を選び、その話者のさまざまな声から得た感情パラメタの平均値を加算することができる。

30

【 0 0 9 1 】

正規化モードが無効である（S 4 0 4：YES）又はステップ S 4 0 5 の後、ステップ S 4 0 6 において、感情パラメタ変換部 1 2 0 0 は、得られた感情パラメタにさらに変換を行う。ただし、本実施形態ではこの変換は実行されない。よって、無変換である。

【 0 0 9 2 】

ステップ S 4 0 7 において、言語特徴量変換部 1 2 2 0 は、入力されたテキスト 1 2 1 0 から、言語特徴量を生成、出力する。ステップ S 4 0 8 において、音声合成器 1 2 3 0 は、言語特徴量変換部 1 2 2 0 からの言語特徴量と、感情パラメタ変換部 1 2 0 0 からの感情パラメタとを、受け取る。

40

【 0 0 9 3 】

ステップ S 4 0 9 において、音声合成器 1 2 3 0 は、合成音声 1 2 4 0 を生成し、出力する。音声合成器 1 2 3 0 は、取得した言語特徴量及び感情パラメタを音声合成モデル 1 0 8 0 に入力し、出力された音響特徴量を取得する。音声合成器 1 2 3 0 は、音響特徴量から合成音声 1 2 4 0 を生成する。

【 0 0 9 4 】

50

以上、第1の実施形態では、音声合成コーパスに対して感情認識器を経て得た感情パラメタを使って音声合成モデルの学習を行い、ターゲット音声から感情認識器を使い得た感情パラメタの統計量を使用して音声合成を行う。これにより、ターゲット音声の感情を真似た音声合成を簡便に実現することができる。

【0095】

＜第2の実施形態＞

本実施形態では、ターゲット音声データベースの感情パラメタ分布が音声合成コーパスの感情パラメタ分布に近くなるように、ターゲット音声の感情パラメタが変換される。変換された感情パラメタが、音声合成器に入力される。

【0096】

これにより、学習に使用される音声合成コーパスの話者が持つ感情の話者内変動を模擬して、音声合成ができるようになる。結果、開発者が所望した感情により近い音声を合成することができる。本実施形態は、第1の実施形態に追加変更を行ったものであるため、第1の実施形態と異なる部分を主に説明する。

【0097】

図11は、第2の実施形態に係る音声合成システム100の論理構成例を模式的に示す。図1Aに示す構成例に加えて、音声合成システム100は、音声合成コーパス統計量計算部1250、音声合成コーパス統計量データベース1260を含む。さらに、音声合成システム100は、変換パラメタデータベース1280、及び変換パラメタ計算部1270を含む。なお、図1Bに示す構成例に対しても、本実施形態の構成が同様に適用される。

【0098】

図11を参照して、音声合成コーパス統計量計算部1250は、音声合成コーパス1010にある音声サンプルの感情パラメタにおける統計量、すなわち分布を規定するパラメタを計算する。音声合成コーパス統計量データベース1260は、音声合成コーパス統計量計算部1250で求めた統計量を保存する。

【0099】

いかなる分布形態も使うことができるが、ここでは正規分布を仮定し、平均と標準偏差を求めるものとする。なお、感情認識器が出力する感情パラメタは多次元ベクトルであることを仮定するため、平均および標準偏差もそれぞれベクトルとなる。

【0100】

変換パラメタ計算部1270は、音声合成コーパス1010及びターゲット音声データベース1090の両方を参照し、ターゲット音声の感情パラメタを音声合成コーパス1010の分布に近づけるための変換アルゴリズムで使用するパラメタを求める。

【0101】

ターゲット音声の感情パラメタを機械学習モデル（条件－感情パラメタ変換モデル）1165により求める場合、変換パラメタ計算部1270は、対話システムを使用する条件の一覧、つまり、条件1180として使用され得る条件の一覧を使用して、機械学習モデル1165に、それぞれの条件を与える。変換パラメタ計算部1270は、機械学習モデル1165から得られた感情パラメタから統計量を求め、それをターゲット音声統計量データベース1160に保存されている感情パラメタの代替として使用する。

【0102】

図12を参照して、この手順を説明する。ステップS501において、変換パラメタ計算部1270は、音声合成コーパス1010にあるそれぞれの音声サンプルを感情認識器1060に入力することにより、それぞれの音声サンプルの感情パラメタを求める。

【0103】

ステップS502において、変換パラメタ計算部1270は、感情パラメタから統計量を求める。ここでは統計量として、平均と標準偏差を求めることとする。各音声サンプルを w とし、音声サンプルの全体集合を W とすると、感情パラメタの平均値 a 及び標準偏差 σ は、以下の式で計算される。

10

20

30

40

50

【 0 1 0 4 】

【 数 1 】

$$a = \frac{1}{|W|} \sum_{w \in W} \text{SER}(w)$$

【 0 1 0 5 】

【 数 2 】

$$\sigma_a = \sqrt{\frac{1}{|W| - 1} \sum_{w \in W} (\text{SER}(w) - a)^2}$$

10

【 0 1 0 6 】

| W | は音声サンプルの個数である。S E R (w) は、音声サンプル w の感情パラメタを求める関数である。標準偏差を求める式における二乗と平方根は、ともにベクトルの各次元の値に適用される。

【 0 1 0 7 】

ステップ S 5 0 3 において、変換パラメタ計算部 1 2 7 0 は、多くの条件を含む条件セットを作成する。この条件セットには、対話システムを実際に使用する場面を表す条件を普遍的に含むことが望ましい。一つのシチュエーションを特徴づける条件変数は 1 個ないし複数あるため、この条件変数の組を条件タプルと呼ぶこととする。よって、条件セットには、複数の条件タプルが含まれている。さらに、実際の利用におけるそれぞれの条件の利用頻度に応じて、含める条件タプルの数の割合を調整してもよい。

20

【 0 1 0 8 】

条件セットを X と表し、格納されるそれぞれの条件タプルを x と表すと、以下の式が定義できる。| X | は条件タプルの数である。

【 0 1 0 9 】

【 数 3 】

$$x \in X = \{x_1, x_2, \dots, x_{|X|}\}$$

30

【 0 1 1 0 】

i 番目の条件タプル x i は、以下のように表現できる。q i 1 , q i 2 , … は、条件タプル x i を構成するそれぞれの条件であり、| Q | は条件の数である。

【 0 1 1 1 】

【 数 4 】

$$x_i = (q_{i1}, q_{i2}, \dots, q_{i|Q|})$$

【 0 1 1 2 】

ステップ S 5 0 4 は分岐であり、ターゲット音声の感情パラメタのモデル化に機械学習を使用しているか否かにより分岐する。

40

【 0 1 1 3 】

機械学習を使用していない場合 (S 5 0 4 : N O) 、フローはステップ S 5 0 5 に進む。ステップ S 5 0 5 において、変換パラメタ計算部 1 2 7 0 は、それぞれの条件タプルからクラス ID を求める。i 番目のクラス ID を c i とすると、c i は以下の式で表される。C L S は条件タプルをクラス ID に変換する関数である。

【 0 1 1 4 】

【 数 5 】

50

$$c_i = \text{CLS}(x_i) \quad i = 1, 2, \dots, |X|$$

【 0 1 1 5 】

ステップ S 5 0 6 において、変換パラメタ計算部 1 2 7 0 は、それぞれのクラスに対応した感情パラメタを得る。i 番目の感情パラメタを e_i とすると、 e_i は以下の式で表される。EP は、クラス ID を感情パラメタに変換する関数である。

【 0 1 1 6 】

【数 6】

$$e_i = \text{EP}(c_i) \quad i = 1, 2, \dots, |X|$$

10

【 0 1 1 7 】

機械学習が使用されている場合（S 5 0 4 : Y E S）、フローはステップ S 5 0 9 に進む。変換パラメタ計算部 1 2 7 0 は、以下の数式に示すように、条件タプルから直接感情パラメタを求める。EPM は、条件タプルを感情パラメタに変換する関数であり、機械学習モデルにより実現される。

【 0 1 1 8 】

【数 7】

$$e_i = \text{EPM}(x_i) \quad i = 1, 2, \dots, |X|$$

20

【 0 1 1 9 】

ステップ S 5 0 7 において、変換パラメタ計算部 1 2 7 0 は、感情パラメタ（ e_i ）の統計量を求める。ここでは、平均値と標準偏差を求める。 e および σ_e は、それぞれターゲット音声から求められた感情パラメタの平均値および標準偏差である。

【 0 1 2 0 】

【数 8】

$$e = \frac{1}{|X|} \sum_{i \in [1, |X|]} e_i$$

30

【 0 1 2 1 】

【数 9】

$$\sigma_e = \sqrt{\frac{1}{|X| - 1} \sum_{i \in [1, |X|]} (e_i - e)^2}$$

【 0 1 2 2 】

ステップ S 5 0 8 において、変換パラメタ計算部 1 2 7 0 は、以上求めた a 、 σ_a 、 e 、 σ_e が、感情パラメタ変換部 1 2 0 0 において使われる変換用のパラメタと決定する。これらを変換パラメタデータベース 1 2 8 0 に保存する。

40

【 0 1 2 3 】

次に、上記変換用パラメタを音声の合成時に使用方法を説明する。この処理は、感情パラメタ変換部 1 2 0 0 により、図 1 0 のフローチャートのステップ S 4 0 6 において実行される。音声合成時において、ある条件 x が指定され、感情パラメタ欄 1 1 6 2 から読み込まれた対応する感情パラメタが m であったと仮定する。

【 0 1 2 4 】

最初に、感情パラメタ変換部 1 2 0 0 は、感情パラメタ m から、ターゲット音声の感情分布における標準得点 z を計算する。

50

【 0 1 2 5 】

【 数 1 0 】

$$z = \frac{m - e}{\sigma_e}$$

【 0 1 2 6 】

次に、感情パラメタ変換部 1 2 0 0 は、標準得点 z に対応する音声合成コーパス 1 0 1 0 の感情分布における感情パラメタ b を計算する。

【 0 1 2 7 】

【 数 1 1 】

$$b = z\sigma_a + a$$

10

【 0 1 2 8 】

感情パラメタ変換部 1 2 0 0 は、この感情パラメタ b を音声合成器 1 2 3 0 に入力し、音声合成をする。つまり、ここでは、感情パラメタを線形変換している。

【 0 1 2 9 】

本実施形態では、ターゲット音声の感情パラメタを、音声合成コーパスの話者の感情分布の情報に基づいて変換するため、開発者が所望した感情により近い音声を合成することができる。

20

【 0 1 3 0 】

< 第 3 の実施形態 >

本実施形態は、音声合成器が生成した音声の感情パラメタ分布に基づいて、感情パラメタの変換パラメタを更新（補正）する。これにより、開発者が所望した感情により近い音声を合成することができる。感情パラメタの変換パラメタの初期値は、例えば、第 2 の実施形態において決定された変換パラメタであってもよく、予め設定された変換パラメタであってもよい。初期値は設定されていなくてもよい。

【 0 1 3 1 】

図 1 3 は、本実施形態の音声合成システム 1 0 0 の論理構成例を示す。図 1 1 に示す構成例に加えて、感情認識器 1 2 9 0、出力音声統計量計算部 1 3 0 0、及び出力音声統計量データベース 1 3 1 0 を含む。

30

【 0 1 3 2 】

本実施形態に係る構成要素を、図 1 3 を参照して説明する。音声合成器 1 2 3 0 が出力した合成音声 1 2 4 0 から、感情認識器 1 2 9 0 が、感情パラメタを求める。出力音声統計量計算部 1 3 0 0 は、感情認識器 1 2 9 0 からの感情パラメタを取得して、感情パラメタの統計量を算出する。出力音声統計量計算部 1 3 0 0 は、感情パラメタの統計量を、出力音声統計量データベース 1 3 1 0 に保存する。

【 0 1 3 3 】

図 1 4 は、出力音声統計量データベース 1 3 1 0 の構成例を示す。出力音声統計量データベース 1 3 1 0 は、音声合成器 1 2 3 0 が生成した合成音声の統計情報を格納する。具体的には、出力音声統計量データベース 1 3 1 0 は、クラス ID 欄 1 3 2 0、及び感情パラメタ欄 1 3 3 0 を含む。クラス ID 欄 1 3 2 0 は、音声合成器 1 2 3 0 が生成した合成音声それぞれが属するクラスの ID を示す。感情パラメタ欄 1 3 3 0 は、各クラスにおけるターゲット音声の感情パラメタの統計量を示す。ターゲット音声統計量データベース 1 1 6 0 と合わせて、感情パラメタの変数それぞれの平均値が格納される。

40

【 0 1 3 4 】

本実施形態における変換パラメタの補正処理を説明する。図 1 5 は、感情パラメタの変換パラメタ補正処理例を示すフローチャートである。ステップ S 6 0 1 において、変換パラメタ計算部 1 2 7 0 は、多くの条件を含む条件セットを作成する。この処理は、図 1 2 のフローチャートにおけるステップ S 5 0 3 と同様である。

50

【0135】

ステップS602は分岐であり、ターゲット音声の感情パラメタのモデル化に機械学習を使っているか否かにより分岐する。機械学習を使っていない場合（S602：NO）、フローはステップS604に進む。ステップS604において、変換パラメタ計算部1270は、それぞれの条件タプルからクラスIDを求める。この処理は、図12のフローチャートにおけるステップS505と同様である。ステップS605において、変換パラメタ計算部1270は、それぞれのクラスに対応した感情パラメタを得る。

【0136】

機械学習を使用する場合（S602：YES）、フローは、ステップS603に進む。ステップS603において、変換パラメタ計算部1270は、条件タプルから直接感情パラメタを求める。この処理は、図12のフローチャートにおけるステップS509と同様である。

10

【0137】

ステップS606において、変換パラメタ計算部1270は、感情パラメタを感情パラメタ変換部1200に通し、変換後の感情パラメタを得る。この処理は、第2の実施形態を実施したのちに本実施形態を実施する想定である。ただし、第1の実施形態と本実施形態を組み合わせることもできる。その際には、ステップS606では感情パラメタの変換は行わない。第2の実施形態の数式表現を踏襲するのであれば、この時点で得られる感情パラメタの個数は $|X|$ 個である。

【0138】

ステップS607において、変換パラメタ計算部1270は、ステップS606で得た感情パラメタを使用して、音声合成を実行する。音声合成の対象となるテキストは任意である。例えば、対話システムの利用場面においてよく発話される文を使用することができる。また、それぞれの感情パラメタを得た条件タプルそれぞれに対して、よく発話される文を個別に設計し、その文を入力してもよい。

20

【0139】

ステップS608において、変換パラメタ計算部1270は、ステップS607で得た合成音声を感情認識器1290に入力して、合成音声の感情パラメタを得る。ステップS609において、変換パラメタ計算部1270は、ステップS608で得た感情パラメタに対して、統計量を求める。すなわち、合成音声における感情パラメタの分布を求める。ここでは、平均値と標準偏差を求めることを想定する。この処理は、6050と同様の方法をとることができる。

30

【0140】

ステップS610は分岐であり、合成音声の感情パラメタ分布に対して比較対象とする感情パラメタ分布を決定する。いずれの感情パラメタ分布を選択するか、予めユーザにより指定されているものとする。なお、指定されている感情パラメタ分布は下記の例と異なるものであってもよい。また、指定され得る感情パラメタ分布は一つのみであってもよい。

【0141】

ステップS611において、変換パラメタ計算部1270は、比較対象分布を音声合成コーパス1010の感情パラメタ分布と決定する。この場合、変換パラメタ計算部1270は、音声合成コーパス統計量データベース1260に格納されている平均値及び標準偏差を読み込む。

40

【0142】

ステップS612において、変換パラメタ計算部1270は、比較対象分布をターゲット音声の変換前の感情パラメタ分布と決定する。すなわち、変換パラメタ計算部1270は、第2の実施形態で得た平均値 e 及び標準偏差 σ_e を、変換パラメタデータベース1280から読み込む。

【0143】

ステップS613において、変換パラメタ計算部1270は、比較対象分布をターゲッ

50

ト音声の変換後の感情パラメタ分布と決定する。すなわち、変換パラメタ計算部 1270 は、第 2 の実施形態で得られる変換後の感情パラメタ b をそれぞれの条件タプルで求め、条件タプル間での変換後感情パラメタの平均値及び標準偏差を計算し、読み込む。

【0144】

ステップ S614 において、変換パラメタ計算部 1270 は、合成音声の感情パラメタ分布と比較対象分布の差を求める。この差の計算には様々な方法をとることができるが、たとえば Kullback-Leibler divergence をとることもできるし、単に平均の差を計算する方法も取ることができる。

【0145】

ステップ S615 において、変換パラメタ計算部 1270 は、ステップ S614 で求めた分布間の差が所定の閾値であるか否かを判定する。差が閾値以下であれば（S615：YES）、処理が終了する。差が閾値を上回る場合（S615：NO）、フローはステップ S616 に進む。

10

【0146】

ステップ S616 において、変換パラメタ計算部 1270 は、分布の差が小さくなるように、変換パラメタを更新する。すなわち、第 2 の実施形態では、変換後の感情パラメタ b が一意に決定できたが、この変換式を変更し、変換に必要なとなるパラメタ σa と a を調整可能な値である $\sigma a'$ と a' として再定義する。

【0147】

【数 12】

20

$$b = z\sigma a' + a'$$

【0148】

変換パラメタ計算部 1270 は、この $\sigma a'$ と a' を調整し、分布の差が小さくなる方向に更新する。

【0149】

ステップ S617 において、変換パラメタ計算部 1270 は、変換後のパラメタを変換パラメタデータベース 1280 に保存する。さらに、ステップ S602 に戻り、変換パラメタ計算部 1270 は、分布の差が所定の閾値以下になるまで繰り返す。音声合成では、調整後の変換パラメタが使用され、音声合成に用いる感情パラメタが決定される。

30

【0150】

以上、本実施形態では、感情パラメタの変換部の挙動を決めるために、音声合成器が実際に生成した音声の感情認識結果に基づいて求めることが特徴となる。音声合成器が生成する音声に含まれる感情は、音声合成器に入力した感情パラメタと一致するとは限らない。よって、本実施形態で説明したとおり、実際の音声合成音に対する感情認識器による感情推定結果を使用することで、より所望の感情に近い音声を生成することができる。

【0151】

<第 4 の実施形態>

第 4 の実施形態では、第 1 から第 3 の実施形態において求めた音声合成用の感情パラメタに対して、開発者が調整する機能を提供する。本実施形態を図 16A から 16C を参照して説明する。

40

【0152】

図 16A から 16C のそれぞれのグラフは、音声群に対する感情の分布であり、横軸に感情の大きさ、縦軸に確率密度をプロットしている。横軸となる感情の大きさは、多次元で表現される感情パラメタの各次元の値であり、例えば活性度（Arousal）、感情価（Pleasantness）、支配度（Dominance）などである。

【0153】

図 16A は、ターゲット音声の感情パラメタから求めた条件間の感情分布を示す。条件のかわりにクラス ID 間の感情分布を用いてもよい。図 16B は、音声合成コーパスに含まれる音声サンプルの感情分布を示す。

50

【0154】

図16Aに示すターゲット音声の分布と、音声合成コーパスの分布8020は、異なるケースが多い。そこで、ターゲット音声の分布を音声合成コーパスの分布に近くなるような感情パラメタの変換を行った実施形態を説明した。よって、対話システムが生成する音声は、音声合成コーパスの分布8020に沿った形で感情が生成される。

【0155】

図16Cは、実際にアプリケーションで使用する音声の感情分布を示す。上記の場合には、音声合成コーパス1010の分布に沿った形となるため、分布は線8035のような形となる。たとえば、ターゲット音声では感情の大きさが x であったものが、合成時には x' の感情の大きさで合成される。この両者は、それぞれの分布上の標準得点が同一である。

10

【0156】

しかし、対話システムの用途によっては、この分布からさらに人手で感情を調整したいケースがある。たとえば、より感情の起伏をつけたい場合や、全体的に落ち着いて発話させたいといったケースがある。

【0157】

こういったケースに対応するため、本実施形態では、変換パラメタを開発者が変更する機能を提供する。たとえば、より感情の起伏をつけるために標準偏差を大きくし、全体的に落ち着いた声にするため平均値を下げるような操作を想定する。そのため、開発者が感情の変換アルゴリズムにおけるパラメタ($\sigma a'$ と a')を操作できる機能を提供する。

20

【0158】

図16Cにおいて、線8040は、調整後の分布の例を示す。この場合、ターゲット音声では感情の大きさが x であったものが、合成時には x'' の感情の大きさで合成される。この両者は、それぞれの感情の分布において同一の標準得点をとる。

【0159】

音声合成システム100は、パラメタ操作のためのGUI画面を出力装置105において表示する。音声合成システム100は、例えば、比較対象の感情パラメタ分布の情報を開発者に出力装置105において提示し、入力装置104からのユーザ入力に応じて、変換アルゴリズムにおけるパラメタ($\sigma a'$ と a')を変更する。

【0160】

例えば、GUIは、比較する二つの分布の間の対応点のユーザからの指定を受け付け、音声合成システム100は、それら対応点の位置から変換パラメタを計算する、例えば対応点が一致するように変換パラメタを決定してもよい。音声合成システム100は、例えば、変更前の分布図と変更後の分布図とを出力装置105において表示してもよい。変換パラメタは、感情パラメタの変数毎に調整可能であってもよい。

30

【0161】

比較する二つの分布の間の対応点をユーザが指定し、その対応点から変換パラメタを決める方法を説明する。ユーザは、入力装置104を使用して、基準音声サンプルとして、1個のターゲット音声サンプルを指定し、出力装置105によってそれを聴取する。次に、音声合成システム100は、音声合成器1230にテキストと感情パラメタを入力し、音声を合成する。ユーザは、それを出力装置105において聴取する。

40

【0162】

ユーザは、ターゲット音声及び合成音声を聞き比べ、合成音声の感情がターゲット音声の感情と近づくよう、入力装置104を介して音声合成器1230に入力する感情パラメタを変更する。音声合成システム100は、ユーザにより変更された感情パラメタにより再度合成を行い、ユーザは合成音声を聴取する。このプロセスを繰り返して、ターゲット音声の感情により類似する音声を生成するための音声合成器へ入力する感情パラメタが決定される。

【0163】

ここで比較対象としていたターゲット音声サンプルを例えば感情認識器1290に入力

50

して得た感情パラメタを x_i 、音声合成器 1230 に入力した感情パラメタを y_i とする。 i は、比較対象となったターゲット音声の順序を表す。ユーザは、複数のターゲット音声に対して上記の作業を行う。この作業を N 個のターゲット音声に対して行った場合、得られるターゲット音声の感情パラメタは x_1 、 x_2 、 $\dots$ 、 x_N となり、それぞれに対応する音声合成器 1230 に入力する感情パラメタは y_1 、 y_2 、 $\dots$ 、 y_N となる。

【0164】

感情パラメタ変換部 1200 は、ここで得た感情パラメタを使い、分布の変換を行う。ここでは x_i と y_i の変換を線形関数により行うと仮定する。ただし、他の関数形式による変換を用いてもよい。第 3 の実施形態の数式の変換式に則り x_i から y_i への変換を行う場合、以下の式を用いることができる。

【0165】

【数 13】

$$y_i = z \sigma_a' + a'$$

$$z = \frac{x_i - e}{\sigma_e}$$

【0166】

e および σ_e は、それぞれターゲット音声から求められた感情パラメタの平均値および標準偏差であり、既知の値である。よって、 x_i から y_i に変換するには、 a' および σ_a' が求まればよい。

【0167】

x_i と y_i との間に線形関係を仮定する場合、最小二乗法を用いて a' および σ_a' を求めることができる。すなわち、以下の式で定義される変換誤差 D を計算する。

【0168】

【数 14】

$$D = \sum_{i \in [1, N]} \left(y_i - \left(\frac{x_i - e}{\sigma_e} \sigma_a' + a' \right) \right)^2$$

【0169】

この変換誤差 D を最小化する a' および σ_a' を計算により求める。この計算は、既知の方法により行うことができる。

【0170】

本実施形態では、少数の変換パラメタを変える手段を提供することで、開発者が容易に音声の感情を変更することを実現する。

【0171】

< 第 5 の実施形態 >

本実施形態では、感情を開発者が調整する際に、音声の情報を視聴しつつ簡便に調整できる機能を提供する。図 17 は、調整方法例を示すフローチャートである。ステップ S704 において、開発者が 1 個のターゲット音声のクラスを入力装置 104 によって指定し、音声合成システム 100 がそれを受け取る。クラスの代替として、条件を指定してもよい。

【0172】

ステップ S705 において、音声合成システム 100 は、指定されたクラスが持つ感情パラメタを使い、音声合成を実施し、出力装置 105 で再生する。ステップ S707 において、開発者が音声を聴取した結果に基づき、音声合成時の感情パラメタを入力装置 104 によって変更する。音声合成システム 100 は、変更後の感情パラメタを受け取る。そして、ステップ S708 において、音声合成システム 100 は、変更した後の感情パラメタを使い合成音声を生成し、その音声を出力装置 105 によって再生し、開発者が聴取する。このプロセスを開発者が編集を終了するまで繰り返す（ステップ S706 の分岐が N

10

20

30

40

50

0のループ)。

【0173】

開発者が編集を終了した場合、音声合成システム100は、再度クラスの指定をするかを開発者に尋ねる。ここで開発者が編集の終了を指定した場合(S706: YES)、フローは、ステップS701を経て、ステップS702に進む。ステップS702において、音声合成システム100は、編集前の感情パラメタを編集後の感情パラメタに変換するよう変換パラメタを計算する。ステップS703では、音声合成システム100は、計算された変換パラメタを変換パラメタデータベース1280に保存する。

【0174】

本実施形態は、開発者が試行錯誤的に所望する音声合成音を設計する。本実施形態では、ユーザが指定した編集前・編集後の感情パラメタの関係から、感情パラメタの分布全体を変換する。すなわち、たとえば高々2個のクラスに基づいてユーザが感情を編集したとしても、その結果から、対話システムの音声全体の感情が、開発者の編集と矛盾しないように変更される。よって、開発者の試行錯誤の手間を大幅に減らすことができ、効率的な感情音声の設計を実現することができる。

【0175】

<第6の実施形態>

本実施形態は、第2の実施形態における感情パラメタ分布の変換を、別の観点で実現する。第2の実施形態は、感情パラメタ分布を変換する際に、条件間の感情パラメタのバラつきを音声合成コーパスの標準偏差に合わせる。これに代えて、本実施形態は、ターゲット話者個人の音声の感情に表れる条件間のバラつきに合わせるものである。

【0176】

本実施形態では、第2の実施形態における感情パラメタの統計量の計算(S507)において、さらに以下の処理を行う。ここで、ターゲット話者の全体集合をSと表す。s1, s2, ...は、それぞれのターゲット話者を表す。ターゲット話者の数は|S|である。

【0177】

【数15】

$$s \in S = \{s_1, s_2, \dots, s_{|S|}\}$$

【0178】

さらに、ターゲット話者sがクラスcに属する条件タプルにおいて発話した音声us, cと表すこととする。変換パラメタ計算部1270は、この音声us, cを使い、ターゲット話者1名の音声に表れるクラス間の感情パラメタの分散σus, cを計算する。以下で説明する乗算、加算、二乗計算、平方根計算、除算は、すべてベクトルの要素ごとに計算するものである。

【0179】

【数16】

$$u_s = \frac{1}{|X|} \sum_{c \in C} \text{SER}(u_{s,c})$$

【0180】

【数17】

$$\sigma_{u,s}^2 = \frac{1}{|X| - 1} \sum_{c \in C} (\text{SER}(u_{s,c}) - u_s)^2$$

【0181】

ただし、Cはクラスの一覧である。

【0182】

【数 1 8】

$$c \in C = \{c_1, c_2, \dots, c_{|X|}\}$$

【0 1 8 3】

さらに、変換パラメタ計算部 1 2 7 0 は、このターゲット話者それぞれの分散をターゲット話者全体で平均する。さらに平方根をとり、標準偏差 σ_u に変換する。

【0 1 8 4】

【数 1 9】

$$\sigma_u = \sqrt{\frac{1}{|S|} \sum_{s \in S} \sigma_{u,s}^2}$$

10

【0 1 8 5】

変換パラメタ計算部 1 2 7 0 は、ここで計算した σ_u を、変換パラメタデータベース 1 2 8 0 に保存する。

【0 1 8 6】

次に、変換パラメタを使って音声合成する方法を説明する。音声合成時において、ある条件 x が指定され、それに対する感情パラメタ欄 1 1 6 2 から読み込まれた感情パラメタが m であったと仮定する。この値を使い、音声合成器 1 2 3 0 に入力する感情パラメタ b を以下の式により求められる。

20

【0 1 8 7】

【数 2 0】

$$b = \lambda(m - e) + a$$

$$\lambda = \frac{\sigma_u}{\sigma_e}$$

【0 1 8 8】

この式における平均値 a としては、音声合成コーパスの感情パラメタから求めた平均値、またはターゲット音声の感情パラメタから求めた平均値を使うことができる。この式における係数である λ は、感情パラメタの次元ごとに標準偏差の値を除算することにより求められる。ただし、クラス数が少ない場合、標準偏差の各次元の値が 0 に近くなる可能性があり、除算の結果、極端に大きな値となることがありえる。その対策として、上記式で使われる λ を、以下の式を用いてスカラとして計算してもよい。

30

【0 1 8 9】

【数 2 1】

$$\lambda = \frac{1}{E} \sum_{i=1}^E \frac{\sigma_u[i]}{\sigma_e[i]}$$

【0 1 9 0】

または、以下の式も用いることができる。

40

【0 1 9 1】

【数 2 2】

$$\lambda = \sqrt{\frac{\sum_{i=1}^E (\sigma_u[i])^2}{\sum_{i=1}^E (\sigma_e[i])^2}}$$

【0 1 9 2】

ここで E は感情パラメタの次元数である。[i] により、ベクトルの i 番目の要素を参照することを表す。

50

【0193】

すなわち、第2の実施形態では、感情のバラつきを音声合成コーパスの音声に一致するようにする。一方、本実施形態は、ターゲット話者それぞれに表れるバラつき（分布）を模擬するように音声合成をする。これにより、第2の実施形態と比べて、よりターゲット話者の感情の操作を模した音声を合成することができる。実際には、第2の実施形態と本実施形態の音声を開発者が効き比べ、応用用途によって適する実施形態を選択してよい。

【0194】

<その他の実施>

以上の実施形態では、感情パラメタの変換において線形関数による変換を説明したが、他の非線形な関数を使用する変換が採用されてもよい。

【0195】

なお、本発明は上記した実施形態に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施形態は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明したすべての構成を備えるものに限定されるものではない。また、ある実施形態の構成の一部を他の実施形態の構成に置き換えることが可能であり、また、ある実施形態の構成に他の実施形態の構成を加えることも可能である。また、各実施形態の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

【0196】

また、上記の各構成・機能・処理部等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、SSD (Solid State Drive) 等の記録装置、または、ICカード、SDカード等の記録媒体に置くことができる。

【0197】

また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしもすべての制御線や情報線を示しているとは限らない。実際には殆どすべての構成が相互に接続されていると考えてもよい。

【符号の説明】

【0198】

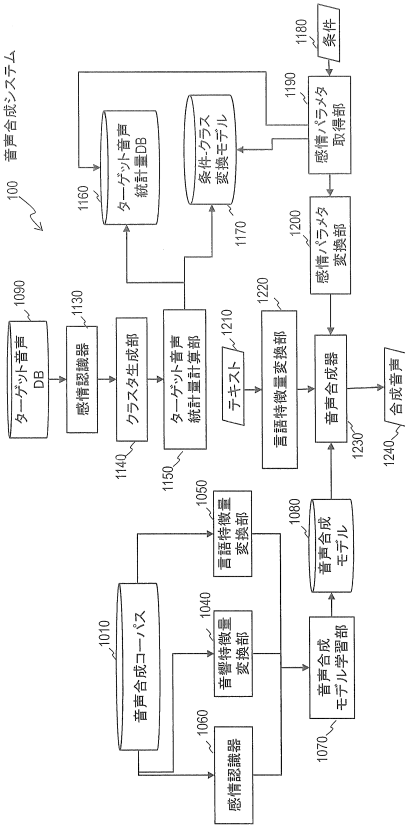
- 100 音声合成システム
- 101 演算装置
- 102 主記憶装置
- 103 補助記憶装置
- 104 入力装置
- 105 出力装置
- 106 通信インターフェース
- 1010 音声合成コーパス
- 1040 音響特徴量変換部
- 1050 言語特徴量変換部
- 1060、1130、1290 感情認識器
- 1070 音声合成モデル学習部
- 1080 音声合成モデル
- 1090 ターゲット音声データベース
- 1140 クラスタ生成部
- 1145 正規化部
- 1150 ターゲット音声統計量計算部
- 1160 ターゲット音声統計量データベース
- 1165 条件－感情パラメタ変換モデル
- 1170 条件－クラス変換モデル

- 1 1 9 0 感情パラメタ取得部
- 1 2 0 0 感情パラメタ変換部
- 1 2 2 0 言語特徴量変換部
- 1 2 3 0 音声合成器
- 1 2 5 0 音声合成コーパス統計量計算部
- 1 2 6 0 音声合成コーパス統計量データベース
- 1 2 7 0 変換パラメタ計算部
- 1 2 8 0 変換パラメタデータベース
- 1 3 0 0 出力音声統計量計算部
- 1 3 1 0 出力音声統計量データベース

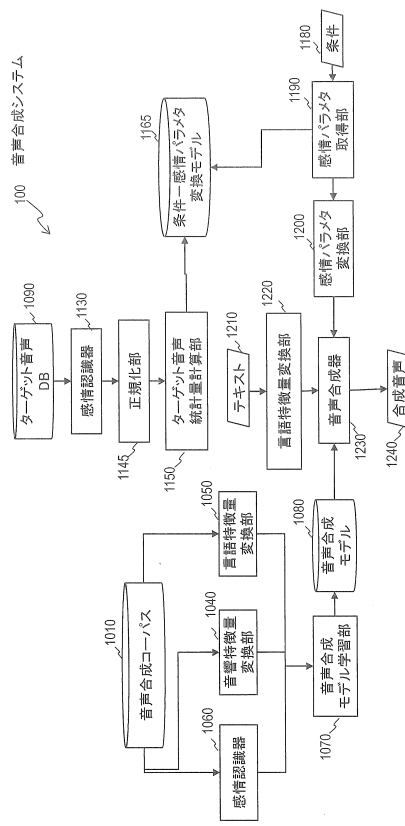
10

【図面】

【図 1 A】



【図 1 B】



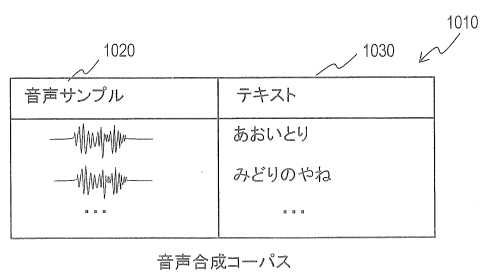
20

30

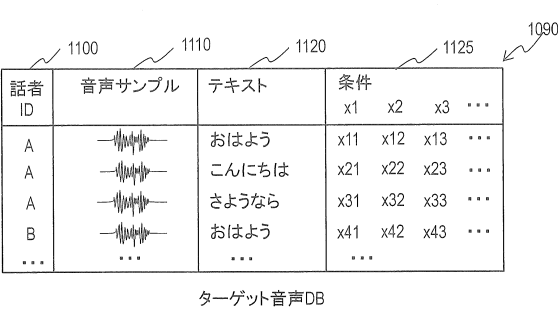
40

50

【図 2】

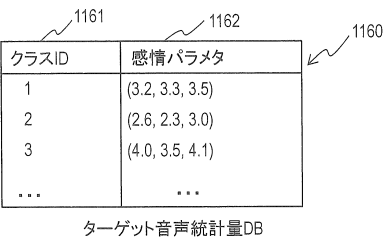


【図 3】

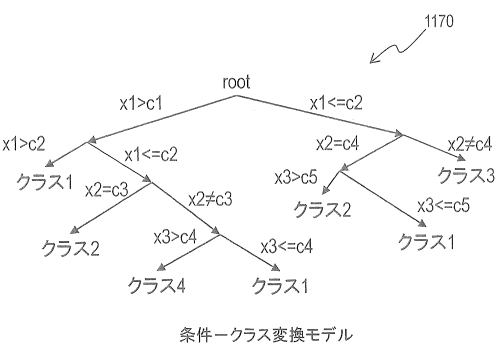


10

【図 4】



【図 5】



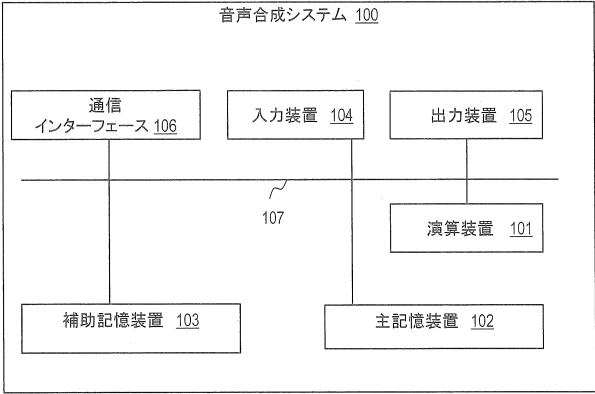
20

30

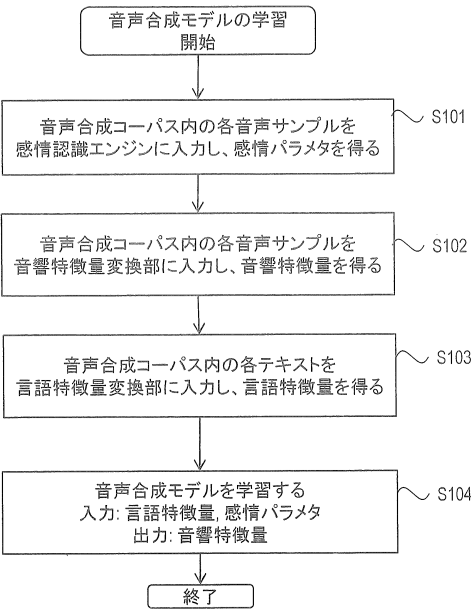
40

50

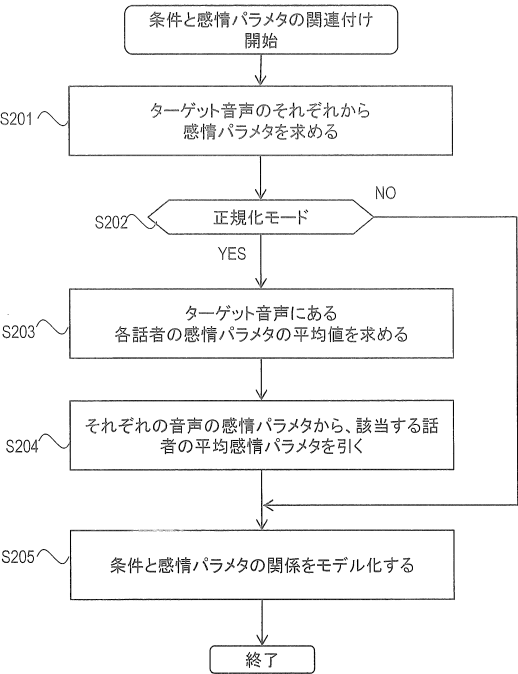
【図 6】



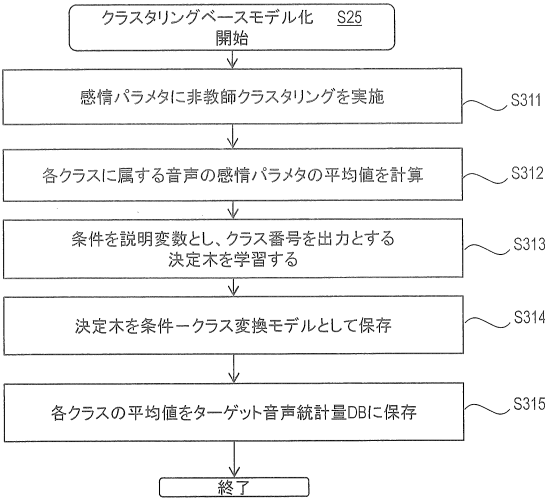
【図 7】



【図 8】



【図 9 A】



10

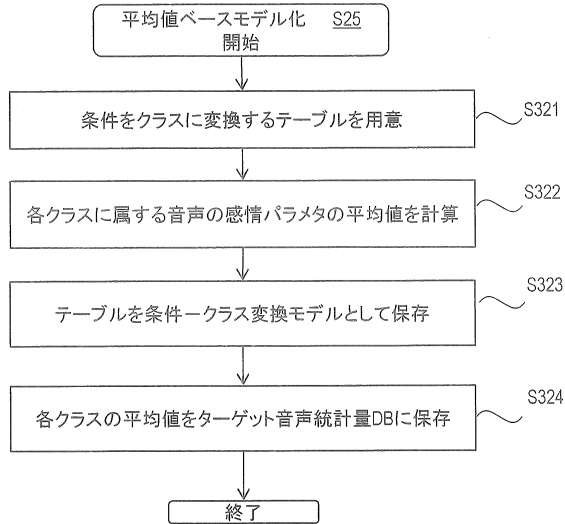
20

30

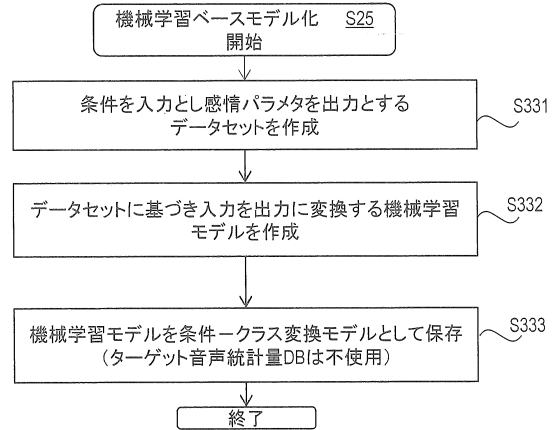
40

50

【図 9 B】

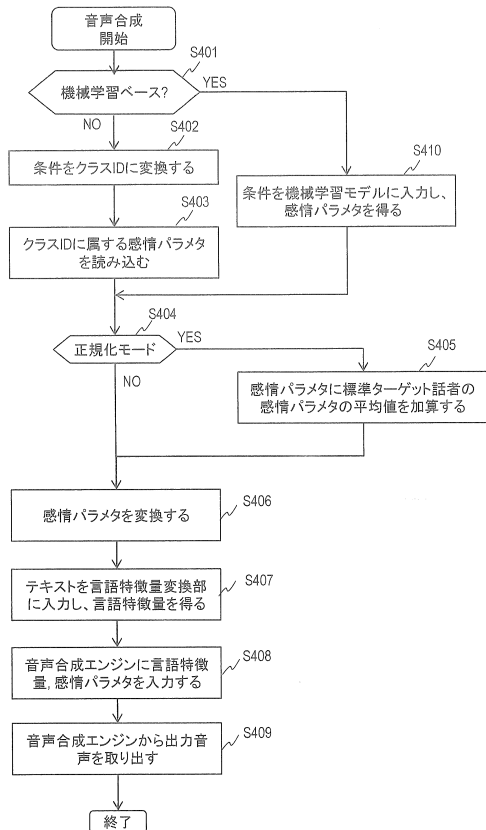


【図 9 C】

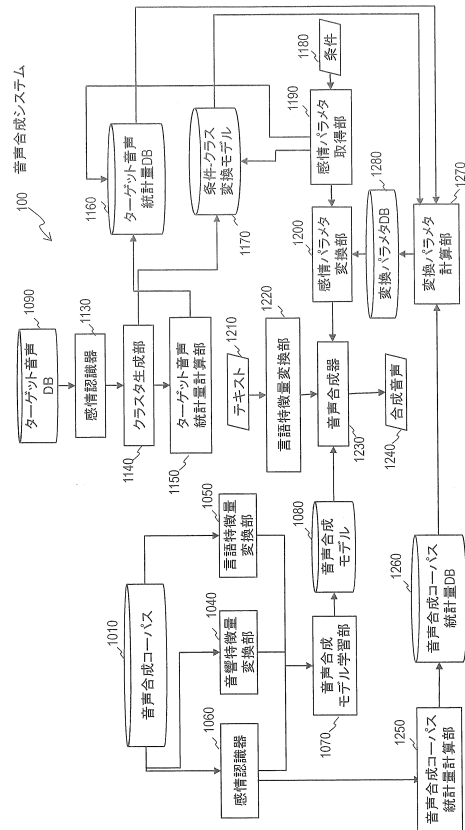


10

【図 10】



【図 11】



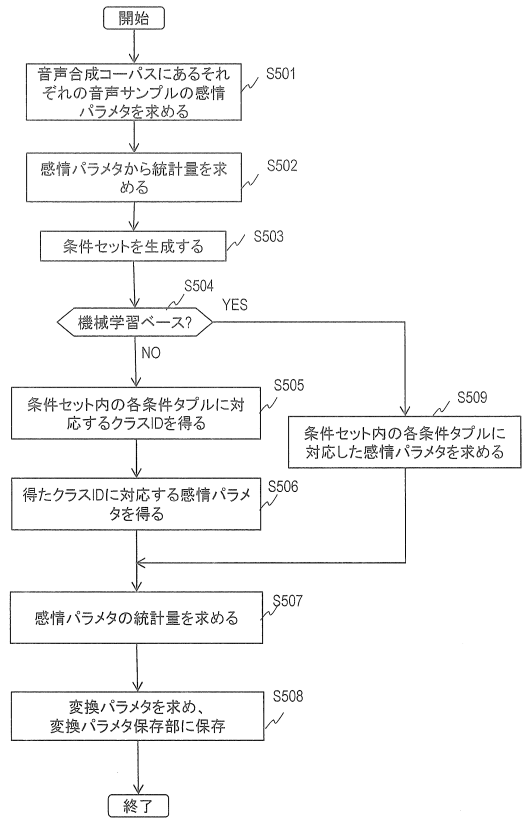
20

30

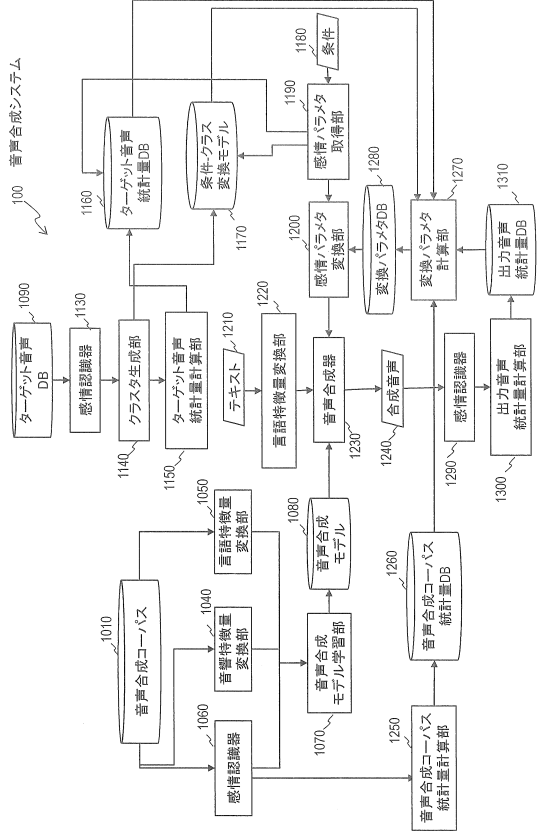
40

50

【図 1 2】



【図 1 3】

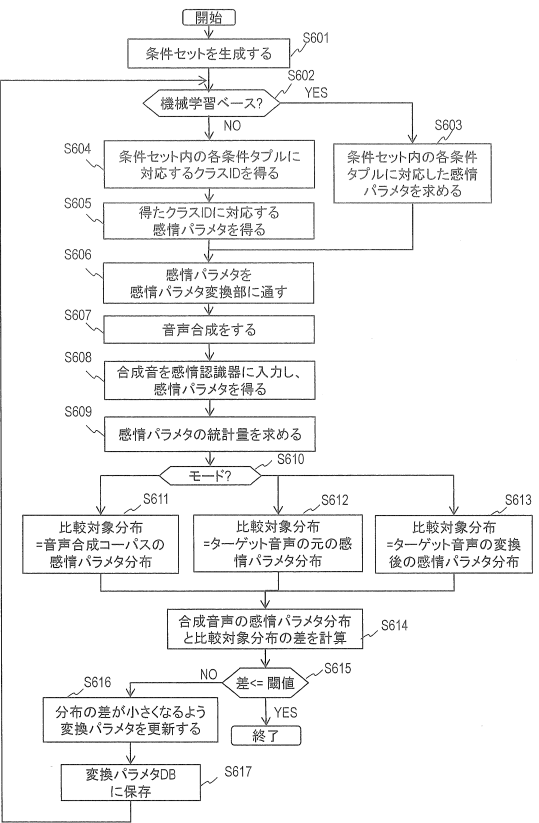


【図 1 4】

クラスID	感情パラメタ
1	(3.1, 3.1, 3.3)
2	(2.7, 2.5, 3.1)
3	(3.9, 3.3, 3.9)
...	...

出力音声統計量DB

【図 1 5】



10

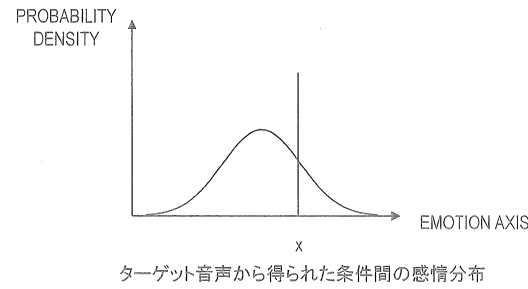
20

30

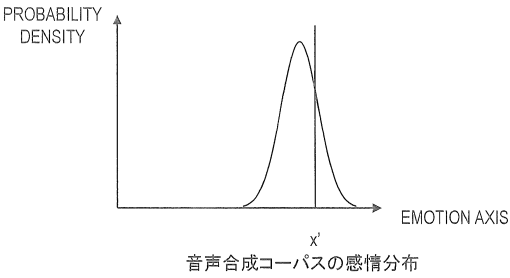
40

50

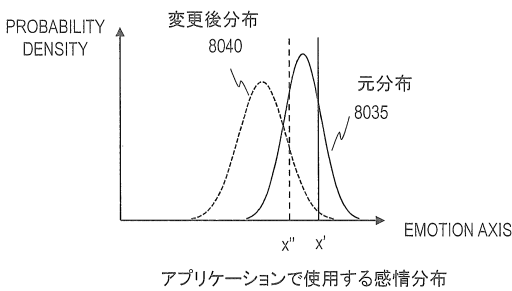
【図 1 6 A】



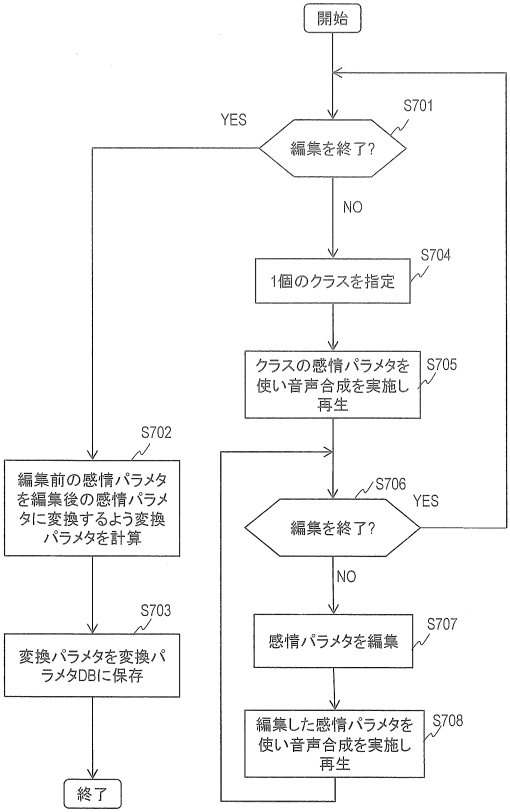
【図 1 6 B】



【図 1 6 C】



【図 1 7】



10

20

30

40

50