(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7691027号
(P7691027)

(45)発行日　令和7年6月11日(2025.6.11)

(24)登録日　令和7年6月3日(2025.6.3)

(51)国際特許分類

		FI		
Ｇ１０Ｌ	*21/007(2013.01)*	Ｇ１０Ｌ	21/007	
Ｇ１０Ｌ	*13/02 (2013.01)*	Ｇ１０Ｌ	13/02	１１０Ｚ
Ｇ１０Ｌ	*15/20 (2006.01)*	Ｇ１０Ｌ	15/20	３００

請求項の数　10　（全28頁）

(21)出願番号　特願2024-504041(P2024-504041)
(86)(22)出願日　令和4年3月1日(2022.3.1)
(86)国際出願番号　PCT/JP2022/008597
(87)国際公開番号　WO2023/166557
(87)国際公開日　令和5年9月7日(2023.9.7)
審査請求日　令和6年8月6日(2024.8.6)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100104765
弁理士　江上 達夫
(74)代理人　100107331
弁理士　中村 聡延
(74)代理人　100131015
弁理士　三輪 浩誉
(72)発明者　カク レイ
東京都港区芝五丁目７番１号 日本電気株式会社内
(72)発明者　山本 仁
東京都港区芝五丁目７番１号 日本電気株式会社内

最終頁に続く

(54)【発明の名称】　音声認識システム、音声認識方法、及び記録媒体

(57)【特許請求の範囲】

【請求項１】

話者が発話したリアル発話データを取得する発話データ取得手段と、

前記リアル発話データをテキストデータに変換するテキスト変換手段と、

前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成する音声合成手段と、

前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成する変換モデル生成手段と、

前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、

を備える音声認識システム。

【請求項２】

前記変換モデル生成手段は、前記入力音声と、前記音声認識手段の認識結果と、を用いて前記変換モデルのパラメータを調整する、

請求項１に記載の音声認識システム。

【請求項３】

前記対応合成音声を含むデータを用いて音声認識モデルを生成する音声認識モデル生成手段を更に備え、

前記音声認識手段は、前記音声認識モデルを用いて音声認識する、

請求項１又は２に記載の音声認識システム。

【請求項４】

前記音声認識モデル生成手段は、前記変換モデルを用いて変換された前記合成音声と、前記音声認識手段の認識結果と、を用いて前記音声認識モデルのパラメータを調整する、
請求項３に記載の音声認識システム。
【請求項５】
前記話者の属性を示す属性情報を取得する属性取得手段を更に備え、
前記音声合成手段は、前記属性情報を用いて音声合成を行うことで前記対応合成音声を生成する、
請求項１から４のいずれか一項に記載の音声認識システム。
【請求項６】
所定の条件ごとに前記リアル発話データを記憶する複数のリアル発話音声コーパスを更に備え、
前記発話データ取得手段は、前記複数のリアル発話音声コーパスから１つを選択して前記リアル発話データを取得する、
請求項１から５のいずれか一項に記載の音声認識システム。
【請求項７】
前記テキストデータ及び前記対応合成音声の少なくとも一方にノイズを付与するノイズ付与手段を更に備える、
請求項１から６のいずれか一項に記載の音声認識システム。
【請求項８】
手話データを取得する手話データ取得手段と、
前記手話データをテキストデータに変換するテキスト変換手段と、
前記テキストデータを用いた音声合成により、前記手話データに対応する対応合成音声を生成する音声合成手段と、
前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成する変換モデル生成手段と、
前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、
を備える音声認識システム。
【請求項９】
少なくとも１つのコンピュータによって、
話者が発話したリアル発話データを取得し、
前記リアル発話データをテキストデータに変換し、
前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、
前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、
前記変換モデルを用いて変換された前記合成音声を音声認識する、
音声認識方法。
【請求項１０】
少なくとも１つのコンピュータに、
話者が発話したリアル発話データを取得し、
前記リアル発話データをテキストデータに変換し、
前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、
前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、
前記変換モデルを用いて変換された前記合成音声を音声認識する、
音声認識方法を実行させるコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】

この開示は、音声認識システム、音声認識方法、及び記録媒体の技術分野に関する。
【背景技術】
【０００２】
この種のシステムとして、合成音声を生成するものが知られている。例えば特許文献１では、音声の声色を表す特徴量を学習済みの変換モデルによって変換するなどして、合成音声を生成することが開示されている。特許文献２では、音声認識結果として取得されたテキストデータからターゲット言語の文を生成し、そのターゲット言語の文から合成音声を生成することが開示されている。
【０００３】
その他の関連する技術として、例えば特許文献３では、学習用コーパスを用いて音声変換モデルの学習を行うことが開示されている。
【先行技術文献】
【特許文献】
【０００４】
【文献】国際公開第２０２１／０３３６８５号
【文献】国際公開第２０１４／０１０４５０号
【文献】特開２０２０－１６６２２４号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
この開示は、先行技術文献に開示された技術を改善することを目的とする。
【課題を解決するための手段】
【０００６】
この開示の音声認識システムの一の態様は、話者が発話したリアル発話データを取得する発話データ取得手段と、前記リアル発話データをテキストデータに変換するテキスト変換手段と、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成する音声合成手段と、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成する変換モデル生成手段と、前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、を備える。
【０００７】
この開示の音声認識システムの一の態様は、手話データを取得する手話データ取得手段と、前記手話データをテキストデータに変換するテキスト変換手段と、前記テキストデータを用いた音声合成により、前記手話データに対応する対応合成音声を生成する音声合成手段と、前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成する変換モデル生成手段と、前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、を備える
【０００８】
この開示の音声認識方法の一の態様は、少なくとも１つのコンピュータによって、話者が発話したリアル発話データを取得し、前記リアル発話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する。
【０００９】
この開示の記録媒体の一の態様は、少なくとも１つのコンピュータに、話者が発話したリアル発話データを取得し、前記リアル発話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識す

る、音声認識方法を実行させるコンピュータプログラムが記録されている。
【図面の簡単な説明】
【００１０】
【図１】第１実施形態に係る音声認識システムのハードウェア構成を示すブロック図である。
【図２】第１実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図３】第１実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。
【図４】第１実施形態に係る音声認識システムによる音声認識動作の流れを示すフローチャートである。
【図５】第２実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図６】第２実施形態に係る音声認識システムによる変換モデル学習動作の流れを示すフローチャートである。
【図７】第３実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図８】第３実施形態に係る音声認識システムによる音声認識モデル生成動作の流れを示すフローチャートである。
【図９】第４実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図１０】第４実施形態に係る音声認識システムによる音声認識モデル学習動作の流れを示すフローチャートである。
【図１１】第５実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図１２】第５実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。
【図１３】第６実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図１４】第６実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。
【図１５】第７実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【図１６】第７実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。
【図１７】第７実施形態の変形例に係る音声認識システムの機能的構成を示すブロック図である。
【図１８】第７実施形態の変形例に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。
【図１９】第８実施形態の変形例に係る音声認識システムの機能的構成を示すブロック図である。
【図２０】第８実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。
【図２１】第８実施形態に係る音声認識システムによる音声認識動作の流れ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】
【００１１】
以下、図面を参照しながら、音声認識システム、音声認識方法、及び記録媒体の実施形態について説明する。
【００１２】
＜第１実施形態＞
第１実施形態に係る音声認識システムについて、図１から図４を参照して説明する。
【００１３】
（ハードウェア構成）
まず、図１を参照しながら、第１実施形態に係る音声認識システムのハードウェア構成について説明する。図１は、第１実施形態に係る音声認識システムのハードウェア構成を示すブロック図である。

【００１４】
図１に示すように、第１実施形態に係る音声認識システム１０は、プロセッサ１１と、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）１２と、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）１３と、記憶装置１４とを備えている。音声認識システム１０は更に、入力装置１５と、出力装置１６と、を備えていてもよい。上述したプロセッサ１１と、ＲＡＭ１２と、ＲＯＭ１３と、記憶装置１４と、入力装置１５と、出力装置１６とは、データバス１７を介して接続されている。
【００１５】
プロセッサ１１は、コンピュータプログラムを読み込む。例えば、プロセッサ１１は、ＲＡＭ１２、ＲＯＭ１３及び記憶装置１４のうちの少なくとも一つが記憶しているコンピュータプログラムを読み込むように構成されている。或いは、プロセッサ１１は、コンピュータで読み取り可能な記録媒体が記憶しているコンピュータプログラムを、図示しない記録媒体読み取り装置を用いて読み込んでもよい。プロセッサ１１は、ネットワークインタフェースを介して、音声認識システム１０の外部に配置される不図示の装置からコンピュータプログラムを取得してもよい（つまり、読み込んでもよい）。プロセッサ１１は、読み込んだコンピュータプログラムを実行することで、ＲＡＭ１２、記憶装置１４、入力装置１５及び出力装置１６を制御する。本実施形態では特に、プロセッサ１１が読み込んだコンピュータプログラムを実行すると、プロセッサ１１内には、音声認識を行うための機能ブロックが実現される。即ち、プロセッサ１１は、音声認識システム１０における各制御を実行するコントローラとして機能してよい。
【００１６】
プロセッサ１１は、例えばＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）、ＧＰＵ（Ｇｒａｐｈｉｃｓ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）、ＦＰＧＡ（ｆｉｅｌｄ－ｐｒｏｇｒａｍｍａｂｌｅ ｇａｔｅ ａｒｒａｙ）、ＤＳＰ（Ｄｅｍａｎｄ－Ｓｉｄｅ Ｐｌａｔｆｏｒｍ）、ＡＳＩＣ（Ａｐｐｌｉｃａｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔ）として構成されてよい。プロセッサ１１は、これらのうち一つで構成されてもよいし、複数を並列で用いるように構成されてもよい。
【００１７】
ＲＡＭ１２は、プロセッサ１１が実行するコンピュータプログラムを一時的に記憶する。ＲＡＭ１２は、プロセッサ１１がコンピュータプログラムを実行している際にプロセッサ１１が一時的に使用するデータを一時的に記憶する。ＲＡＭ１２は、例えば、Ｄ－ＲＡＭ（Ｄｙｎａｍｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）や、ＳＲＡＭ(Ｓｔａｔｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ)であってよい。また、ＲＡＭ１２に代えて、他の種類の揮発性メモリが用いられてもよい。
【００１８】
ＲＯＭ１３は、プロセッサ１１が実行するコンピュータプログラムを記憶する。ＲＯＭ１３は、その他に固定的なデータを記憶していてもよい。ＲＯＭ１３は、例えば、Ｐ－ＲＯＭ（Ｐｒｏｇｒａｍｍａｂｌｅ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）や、ＥＰＲＯＭ(Ｅｒａｓａｂｌｅ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ)であってよい。また、ＲＯＭ１３に代えて、他の種類の不揮発性 メモリが用いられてもよい。
【００１９】
記憶装置１４は、音声認識システム１０が長期的に保存するデータを記憶する。記憶装置１４は、プロセッサ１１の一時記憶装置として動作してもよい。記憶装置１４は、例えば、ハードディスク装置、光磁気ディスク装置、ＳＳＤ（Ｓｏｌｉｄ Ｓｔａｔｅ Ｄｒｉｖｅ）及びディスクアレイ装置のうちの少なくとも一つを含んでいてもよい。
【００２０】
入力装置１５は、音声認識システム１０のユーザからの入力指示を受け取る装置である。入力装置１５は、例えば、キーボード、マウス及びタッチパネルのうちの少なくとも一つを含んでいてもよい。入力装置１５は、スマートフォンやタブレット等の携帯端末として構成されていてもよい。入力装置１５は、例えばマイクを含む音声入力が可能な装置で

あってもよい。
【００２１】
出力装置１６は、音声認識システム１０に関する情報を外部に対して出力する装置である。例えば、出力装置１６は、音声認識システム１０に関する情報を表示可能な表示装置（例えば、ディスプレイ）であってもよい。また、出力装置１６は、音声認識システム１０に関する情報を音声出力可能なスピーカ等であってもよい。出力装置１６は、スマートフォンやタブレット等の携帯端末として構成されていてもよい。また、出力装置１６は、画像以外の形式で情報を出力する装置であってもよい。例えば、出力装置１６は、音声認識システム１０に関する情報を音声で出力するスピーカであってもよい。
【００２２】
なお、図１では、複数の装置を含んで構成される音声認識システム１０の例を挙げたが、これらの全部又は一部の機能を、１つの装置（音声認識装置）として実現してもよい。その場合、音声認識装置は、例えば上述したプロセッサ１１、ＲＡＭ１２、ＲＯＭ１３のみを備えて構成され、その他の構成要素（即ち、記憶装置１４、入力装置１５、出力装置１６）については、音声認識装置に接続される外部の装置が備えるようにしてもよい。また、音声認識装置は、一部の演算機能を外部の装置（例えば、外部サーバやクラウド等）によって実現するものであってもよい。
【００２３】
（機能的構成）
次に、図２を参照しながら、第１実施形態に係る音声認識システム１０の機能的構成について説明する。図２は、第１実施形態に係る音声認識システムの機能的構成を示すブロック図である。
【００２４】
図２に示すように、第１実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、音声変換部２１０と、音声認識部２２０と、を備えて構成されている。発話データ取得部１１０、テキスト変換部１２０、音声合成部１３０、変換モデル生成部１４０、音声変換部２１０、音声認識部２２０の各々は、例えば上述したプロセッサ１１（図１参照）によって実現される処理ブロックであってよい。
【００２５】
発話データ取得部１１０は、話者が発話したリアル発話データを取得可能に構成されている。リアル発話データは、音声データ（例えば、波形データ）であってよい。リアル発話データは、例えば複数のリアル発話データを蓄積するデータベース（リアル発話音声コーパス）から取得されてよい。発話データ取得部１１０で取得されたリアル発話データは、テキスト変換部１２０及び変換モデル生成部１４０に出力される構成となっている。
【００２６】
テキスト変換部１２０は、発話データ取得部１１０で取得されたリアル発話データをテキストデータに変換可能に構成されている。即ち、テキスト変換部１２０は、音声データをテキスト変換する処理を実行可能に構成されている。なお、テキスト変換の具体的な手法については、既存の技術が適宜採用されてよい。テキスト変換部１２０で変換されたテキストデータ（即ち、リアル発話データに対応するテキストデータ）は、音声合成部１３０に出力される構成となっている。
【００２７】
音声合成部１３０は、テキスト変換部１２０で変化されたテキストデータを音声合成することで、リアル発話データに対応する対応合成音声を生成可能に構成されている。なお、音声合成の具体的な手法については、既存の技術を適宜採用することができる。音声合成部１３０で生成された対応合成音声は、変換モデル生成部１４０に出力される構成となっている。なお、対応合成音声は、複数の対応合成を蓄積可能なデータベース（合成音声コーパス）に蓄積されてから、変換モデル生成部１４０に出力されてもよい。
【００２８】

変換モデル生成部１４０は、発話データ取得部１１０で取得されたリアル発話データと、音声合成部１３０で合成された対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成可能に構成されている。変換モデルは、例えば、話者が発話した入力音声（即ち、人間の音声）を、合成音声（即ち、機械的な音声）に近づくように変換する。変換モデル生成部１４０は、例えばＧＡＮ（Ｇｅｎｅｒａｔｉｖｅ Ａｄｖｅｒｓａｒｉａｌ Ｎｅｔｗｏｒｋ：敵対的生成ネットワーク）を用いて、変換モデルを生成するように構成されてよい。変換モデル生成部１４０で生成された変換モデルは、音声変換部２１０に出力される構成となっている。

【００２９】

音声変換部２１０は、変換モデル生成部１４０で生成された変換モデルを用いて、入力音声を合成音声に変換可能に構成されている。音声変換部２１０に入力される入力音声は、例えばマイク等を用いて入力される音声であってよい。音声変換部２１０で変換された合成音声は、音声認識部２２０に出力される構成となっている。

【００３０】

音声認識部２２０は、音声変換部２１０で変換された合成音声を音声認識することが可能に構成されている。即ち、音声認識部２２０は、合成音声をテキスト化する処理を実行可能に構成されている。音声認識部２２０は、合成音声の音声認識結果を出力可能に構成されてよい。なお、音声認識結果の利用方法については特に限定されない。

【００３１】

（変換モデル生成動作）

次に、図３を参照しながら、第１実施形態に係る音声認識システム１０による変換モデルを生成する際の動作（以下、適宜「変換モデル生成動作」と称する）の流れについて説明する。図３は、第１実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。

【００３２】

図３に示すように、第１実施形態に係る音声認識システム１０による変換モデル生成動作が開始されると、まず発話データ取得部１１０が、リアル発話データを取得する（ステップＳ１０１）。そして、テキスト変換部１２０が、発話データ取得部１１０で取得されたリアル発話データをテキストデータに変換する（ステップＳ１０２）。

【００３３】

続いて、音声合成部１３０が、テキスト変換部１２０で変換されたテキストデータを音声合成し、リアル発話データに対応する対応合成音声を生成する（ステップＳ１０３）。そして、変換モデル生成部１４０が、発話データ取得部１１０で取得されたリアル発話データ及び音声合成部１３０で生成された対応合成音声に基づいて、変換モデルを生成する（ステップＳ１０４）。その後、変換モデル生成部１４０は、生成した変換モデルを音声変換部２１０に出力する（ステップＳ１０５）。

【００３４】

（変換認識動作）

次に、図４を参照しながら、第１実施形態に係る音声認識システム１０による音声認識を行う際の動作（以下、適宜「音声認識動作」と称する）の流れについて説明する。図３は、第１実施形態に係る音声認識システムによる音声認識動作の流れを示すフローチャートである。

【００３５】

図４に示すように、第１実施形態に係る音声認識システム１０による音声認識動作が開始されると、まず音声変換部２１０が入力音声を取得する（ステップＳ１５１）。そして、音声変換部２１０は、変換モデル生成部１４０で生成された変換モデルを読み込む（ステップＳ１５２）。その後、音声変換部２１０は、読み込んだ変換モデルを用いて音声変換を行い、入力音声を合成音声に変換する（ステップＳ１５３）。

【００３６】

続いて、音声認識部２２０は、音声認識モデル（即ち、音声認識をするためのモデル）

を読み込む（ステップＳ１５４）。そして、音声認識部２２０は、読み込んだ音声認識モデルを用いて、音声変換部２１０で合成された合成音声を音声認識する（ステップＳ１５５）。その後、音声認識部２２０は、音声認識結果を出力する（ステップＳ１５６）。

【００３７】

（技術的効果）

次に、第１実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。

【００３８】

図１から図４で説明したように、第１実施形態に係る音声認識システム１０では、変換モデルを生成する際に、リアル発話データ及びリアル発話データに対応する対応合成音声が用いられる。そして特に、リアル発話データに対応する対応合成音声は、リアル発話データをテキスト変換し、テキストデータを音声合成することで生成される。このようにすれば、リアル発話データと、それに対応する合成音声と、の両方を用意する必要がなくなる（即ち、リアル発話データのみ用意すれば、対応合成音声を生成できる）ため、変換モデルを生成するのに要するコストを抑制することができる。その結果、低コストで認識精度の高い音声認識を実現することが可能となる。

【００３９】

＜第２実施形態＞

第２実施形態に係る音声認識システム１０について、図５及び図６を参照して説明する。なお、第２実施形態は、上述した第１実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１実施形態と同一であってよい。このため、以下では、すでに説明した第１実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。

【００４０】

（機能的構成）

まず、図５を参照しながら、第２実施形態に係る音声認識システム１０の機能的構成について説明する。図５は、第２実施形態に係る音声認識システムの機能的構成を示すブロック図である。なお、図５では、図２で示した構成要素と同様の要素に同一の符号を付している。

【００４１】

図５に示すように、第２実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、音声変換部２１０と、音声認識部２２０と、を備えて構成されている。そして第２実施形態では特に、変換モデル生成部１４０に、音声変換部２１０に入力される入力音声及び音声認識部２２０による認識結果が入力される構成となっている。第２実施形態に係る変換モデル生成部１４０は、音声変換部２１０に入力される入力音声及び音声認識部２２０による認識結果に基づいて、変換モデルの学習を実行可能に構成されている。

【００４２】

（変換モデル学習動作）

次に、図６を参照しながら、第２実施形態に係る音声認識システム１０による変換モデルを学習する際の動作（以下、適宜「変換モデル学習動作」と称する）の流れについて説明する。図６は、第２実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。

【００４３】

図６に示すように、第２実施形態に係る音声認識システム１０による変換モデル学習動作が開始されると、まず変換モデル生成部１４０が、音声変換部２１０に入力される入力音声を取得する（ステップＳ２０１）。そして、変換モデル生成部１４０は更に、その入力音声が入力された際の音声認識結果（即ち、図４に示すステップＳ１５６で出力される音声認識結果）を取得する（ステップＳ２０２）。

【００４４】
続いて、変換モデル生成部１４０は、取得した入力音声及び音声認識結果に基づいて、変換モデルを学習する（ステップＳ２０３）。この際、変換モデル生成部１４０は、すでに生成していた変換モデルのパラメータ調整を行ってよい。その後、変換モデル生成部１４０は、学習した変換モデルを音声変換部２１０に出力する（ステップＳ２０４）。
【００４５】
（技術的効果）
次に、第２実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。
【００４６】
図５及び図６で説明したように、第２実施形態に係る音声認識システム１０では、入力音声及び音声認識結果に基づいて変換モデルが学習される。このようにすれば、入力音声が実際にどのように音声認識されるかを考慮して学習が行われるため、より適切な音声変換が行えるように変換モデルを学習できる。具体的には、音声変換した合成音声を用いて行う音声認識の精度が向上するように、変換モデルを学習できる。
【００４７】
＜第３実施形態＞
第３実施形態に係る音声認識システム１０について、図７及び図８を参照して説明する。なお、第３実施形態は、上述した第１及び第２実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１及び第２実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。
【００４８】
（機能的構成）
まず、図７を参照しながら、第３実施形態に係る音声認識システム１０の機能的構成について説明する。図７は、第３実施形態に係る音声認識システムの機能的構成を示すブロック図である。なお、図７では、図２で示した構成要素と同様の要素に同一の符号を付している。
【００４９】
図７に示すように、第３実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、音声変換部２１０と、音声認識部２２０と、音声認識モデル生成部３１０と、を備えて構成されている。即ち、第３実施形態に係る音声認識システム１０は、第１実施形態の構成（図２参照）に加えて、音声認識モデル生成部３１０を更に備えている。なお、音声認識モデル生成部３１０は、例えば上述したプロセッサ１１（図１参照）によって実現される処理ブロックであってよい。
【００５０】
音声認識モデル生成部３１０は、入力音声を合成音声に変換する音声認識モデルを生成可能に構成されている。具体的には、音声認識モデル生成部３１０は、音声合成手段で生成された対応合成音声を用いて、音声認識モデルを生成可能に構成されている。なお、音声認識モデルは、対応合成音声と、それ以外の合成音声とを用いて、音声認識モデルを生成してもよい。音声認識モデル生成部３１０は、音声合成部１３０から直接対応合成音声を取得するよう構成されてもよいし、音声合成手段で生成された対応合成音声を複数記憶する合成音声コーパスから対応合成音声を取得するように構成されてもよい。音声認識モデル生成部３１０で生成された音声認識モデルは、音声認識部２２０に出力される構成となっている。
【００５１】
（音声認識モデル生成動作）
次に、図８を参照しながら、第３実施形態に係る音声認識システム１０による音声認識モデルを生成する際の動作（以下、適宜「音声認識モデル生成動作」と称する）の流れに

ついて説明する。図８は、第３実施形態に係る音声認識システムによる音声認識モデル生成動作の流れを示すフローチャートである。

【００５２】

図８に示すように、第３実施形態に係る音声認識システム１０による音声認識モデル生成動作が開始されると、まず音声認識モデル生成部３１０が、音声合成部１３０で生成された対応合成音声を取得する（ステップＳ３０１）。

【００５３】

続いて、音声認識モデル生成部３１０は、取得した対応合成音声を用いて音声認識モデルを生成する（ステップＳ３０２）。その後、音声認識モデル生成部３１０は、生成した音声認識モデルを音声認識部２２０に出力する（ステップＳ３０３）。

【００５４】

（技術的効果）

次に、第３実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。

【００５５】

図７及び図８で説明したように、第３実施形態に係る音声認識システム１０では、対応合成音声を用いて音声認識モデルが生成される。このようにすれば、音声認識モデルを生成するための合成音声を別途用意する必要がない（即ち、音声変換モデルを生成するために用いた対応合成音声を利用できる）ため、効率的に音声認識モデルを生成することが可能である。

【００５６】

＜第４実施形態＞

第４実施形態に係る音声認識システム１０について、図９及び図１０を参照して説明する。なお、第４実施形態は、上述した第第３実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１から第３実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。

【００５７】

（機能的構成）

まず、図９を参照しながら、第４実施形態に係る音声認識システム１０の機能的構成について説明する。図９は、第４実施形態に係る音声認識システムの機能的構成を示すブロック図である。なお、図９では、図７で示した構成要素と同様の要素に同一の符号を付している。

【００５８】

図９に示すように、第４実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、音声変換部２１０と、音声認識部２２０と、音声認識モデル生成部３１０と、を備えて構成されている。そして第４実施形態では特に、音声認識モデル生成部３１０に、音声変換部２１０で変換された合成音声及び音声認識部２２０による認識結果が入力される構成となっている。第４実施形態に係る音声認識モデル生成部３１０は、音声変換部２１０で変換された合成音声及び音声認識部２２０による認識結果に基づいて、音声認識モデルの学習を実行可能に構成されている。

【００５９】

（音声認識モデル学習動作）

次に、図１０を参照しながら、第４実施形態に係る音声認識システム１０による音声認識モデルを学習する際の動作（以下、適宜「音声認識モデル学習動作」と称する）の流れについて説明する。図１０は、第３実施形態に係る音声認識システムによる音声認識モデル学習動作の流れを示すフローチャートである。

【００６０】

図１０に示すように、第４実施形態に係る音声認識システム１０による音声認識モデル

学習動作が開始されると、まず音声認識モデル生成部３１０が、音声変換部２１０で変換された合成音声（即ち、音声認識部２２０に入力される合成音声）を取得する（ステップＳ４０１）。そして、音声認識モデル生成部３１０は更に、その合成音声の音声認識結果（即ち、図４に示すステップＳ１５６で出力される音声認識結果）を取得する（ステップＳ４０２）。

【００６１】

続いて、音声認識モデル生成部３１０は、取得した合成音声及び音声認識結果に基づいて、音声認識モデルを学習する（ステップＳ４０３）。この際、音声認識モデル生成部３１０は、すでに生成していた変換モデルのパラメータ調整を行ってよい。その後、音声認識モデル生成部３１０は、学習した音声認識モデルを音声変換部２１０に出力する（ステップＳ４０４）。

【００６２】

（技術的効果）

次に、第４実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。

【００６３】

図９及び図１０で説明したように、第４実施形態に係る音声認識システム１０では、合成音声及び音声認識結果に基づいて変換モデルが学習される。このようにすれば、合成音声が実際にどのように音声認識されるかを考慮して学習が行われるため、より適切な音声認識が行えるように音声認識モデルを学習できる。具体的には、音声認識の精度が向上するように、音声認識モデルを学習できる。

【００６４】

＜第５実施形態＞

第５実施形態に係る音声認識システム１０について、図１１及び図１２を参照して説明する。なお、第５実施形態は、上述した第１から第４実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１から第４実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。

【００６５】

（機能的構成）

まず、図１１を参照しながら、第５実施形態に係る音声認識システム１０の機能的構成について説明する。図１１は、第５実施形態に係る音声認識システムの機能的構成を示すブロック図である。なお、図１１では、図２で示した構成要素と同様の要素に同一の符号を付している。

【００６６】

図１１に示すように、第５実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、属性情報取得部１５０と、音声変換部２１０と、音声認識部２２０と、を備えて構成されている。即ち、第５実施形態に係る音声認識システム１０は、第１実施形態の構成（図２参照）に加えて、属性情報取得部１５０を更に備えている。なお、属性情報取得部１５０は、例えば上述したプロセッサ１１（図１参照）によって実現される処理ブロックであってよい。

【００６７】

属性情報取得部１５０は、リアル発話データの話者に関する属性情報を取得可能に構成されている。属性情報は、例えば話者の性別、年齢、職業等に関する情報を含んでいてよい。属性情報取得部１５０は、例えば話者が保有する端末やＩＤカード等から属性情報を取得可能に構成されてよい。或いは、属性情報取得部１５０は、話者が入力した属性情報を取得するように構成されてよい。属性情報取得部１５０で取得された属性情報は、音声合成部１３０に出力される構成になっている。属性情報は、リアル発話データに紐付けた状態でリアル発話音声コーパスに記憶されてもよい。この場合、属性情報は、リアル発話

音声コーパスから音声合成部１３０に出力されるように構成されればよい。
【００６８】
（変換モデル生成動作）
次に、図１２を参照しながら、第５実施形態に係る音声認識システム１０による変換モデル生成動作の流れについて説明する。図１２は、第５実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。なお、図１２では、図３に示した処理と同様の処理に同一の符号を付している。
【００６９】
図１２に示すように、第５実施形態に係る音声認識システム１０による変換モデル生成動作が開始されると、まず発話データ取得部１１０が、リアル発話データを取得する（ステップＳ１０１）。そして、属性情報取得部１５０が、リアル発話データの話者に関する属性情報を取得する（ステップＳ５０１）。なお、ステップＳ１０１とＳ１０２の処理は相前後して実行されてもよいし、同時に並行して実行されてもよい。
【００７０】
続いて、テキスト変換部１２０が、発話データ取得部１１０で取得されたリアル発話データをテキストデータに変換する（ステップＳ１０２）。その後、音声合成部１３０が、テキスト変換部１２０で変換されたテキストデータを音声合成し、リアル発話データに対応する対応合成音声を生成するが、本実施形態では特に、属性情報も用いて音声合成を行う（ステップＳ５０２）。例えば、音声合成部１３０は、リアル発話データの話者の性別や年齢、職業等を考慮した音声合成を行ってよい。
【００７１】
続いて、変換モデル生成部１４０が、発話データ取得部１１０で取得されたリアル発話データ及び音声合成部１３０で生成された対応合成音声（ここでは、属性情報に基づいて音声合成された合成音声）に基づいて、変換モデルを生成する（ステップＳ１０４）。なお、変換モデル生成部１４０に入力されるリアル発話データ及び対応合成音声の組には、属性情報が付与されていてよい。その場合、変換モデル生成部１４０は、属性情報も考慮して、変換モデルを生成してよい。その後、変換モデル生成部１４０は、生成した変換モデルを音声変換部２１０に出力する（ステップＳ１０５）。
【００７２】
（技術的効果）
次に、第５実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。
【００７３】
図１１及び図１２で説明したように、第５実施形態に係る音声認識システム１０では、話者の属性情報を用いて対応合成音声が生成される。このようにすれば、話者の属性が考慮された状態で対応合成音声が生成されるため、より適切な音声変換モデルを生成することが可能となる。また、上述した第３実施形態のように、対応合成音声を用いて音声認識モデルを生成する場合（図７及び図８参照）も、属性が考慮された対応合成音声が用いられることで、より適切な音声認識モデルを生成することが可能となる。
【００７４】
＜第６実施形態＞
第６実施形態に係る音声認識システム１０について、図１３及び図１４を参照して説明する。なお、第６実施形態は、上述した第１から第５実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１から第５実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。
【００７５】
（機能的構成）
まず、図１３を参照しながら、第６実施形態に係る音声認識システム１０の機能的構成について説明する。図１３は、第６実施形態に係る音声認識システムの機能的構成を示す

ブロック図である。なお、図１３では、図１１で示した構成要素と同様の要素に同一の符号を付している。

【００７６】

図１３に示すように、第６実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、複数のリアル発話音声コーパス１０５ａ、１０５ｂ、及び１０５ｃ（以下、適宜まとめて「リアル発話音声コーパス１０５」と称する）と、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、音声変換部２１０と、音声認識部２２０と、を備えて構成されている。即ち、第６実施形態に係る音声認識システム１０は、第１実施形態の構成（図２参照）に加えて、複数のリアル発話音声コーパス１０５を更に備えている。なお、複数のリアル発話音声コーパス１０５は、例えば上述した記憶装置１４（図１参照）によって構成されてよい。

【００７７】

複数のリアル発話音声コーパス１０５は、リアル発話データを所定の条件ごとに記憶している。ここでの「所定の条件」は、例えばリアル発話データを分類するために設定される条件である。例えば、複数のリアル発話音声コーパス１０５の各々は、分野別にリアル発話データを記憶するものであってよい。この場合、リアル発話音声コーパス１０５ａが法律の分野に関するリアル発話データを記憶し、リアル発話音声コーパス１０５ｂが科学の分野に関するリアル発話データを記憶し、リアル発話音声コーパス１０５ｃが医療の分野に関するリアル発話データを記憶するように構成されてよい。なお、ここでは説明の便宜上３つのリアル発話音声コーパス１０５を図示しているが、リアル発話音声コーパス１０５の数は特に限定されるものではない。

【００７８】

第６実施形態に係る発話データ取得部１１０は、上述した複数のリアル発話音声コーパス１０５から１つを選択してリアル発話データを取得可能に構成されている。なお、ここで選択されたリアル発話音声コーパス１０５に関する情報（具体的には、所定の条件に関する情報）は、リアル発話データと共に変換モデル生成部１４０に出力されてよい。そして、変換モデル生成部１４０は、変換モデルを生成する際に選択されたリアル発話音声コーパス１０５に関する情報を用いてもよい。また、上述した第３実施形態のように、音声認識モデルを生成する構成では、選択されたリアル発話音声コーパス１０５に関する情報が、音声認識モデル生成部３１０に出力されてもよい。そして、音声認識モデル生成部３１０は、音声認識モデルを生成する際に選択されたリアル発話音声コーパス１０５に関する情報を用いてもよい。

【００７９】

（変換モデル生成動作）

次に、図１４を参照しながら、第６実施形態に係る音声認識システム１０による変換モデル生成動作の流れについて説明する。図１４は、第６実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。なお、図１４では、図１２に示した処理と同様の処理に同一の符号を付している。

【００８０】

図１４に示すように、第６実施形態に係る音声認識システム１０による変換モデル生成動作が開始されると、まず発話データ取得部１１０が、複数のリアル発話音声コーパス１０５の中から、発話データを取得するコーパスを選択する（ステップＳ６０１）。そして、発話データ取得部１１０は、選択したリアル発話音声コーパスから、リアル発話データを取得する（ステップＳ６０２）。

【００８１】

続いて、テキスト変換部１２０が、発話データ取得部１１０で取得されたリアル発話データをテキストデータに変換する（ステップＳ１０２）。そして、音声合成部１３０が、テキスト変換部１２０で変換されたテキストデータを音声合成し、リアル発話データに対応する対応合成音声を生成する（ステップＳ１０３）。

【００８２】

続いて、変換モデル生成部１４０が、発話データ取得部１１０で取得されたリアル発話データ及び音声合成部１３０で生成された対応合成音声に基づいて、変換モデルを生成するが、本実施形態では特に、選択されたリアル発話音声コーパスに関する情報も用いられる（ステップＳ６０６）。その後、変換モデル生成部１４０は、生成した変換モデルを音声変換部２１０に出力する（ステップＳ１０５）。

【００８３】

（技術的効果）

次に、第６実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。

【００８４】

図１３及び図１４で説明したように、第６実施形態に係る音声認識システム１０では、変換モデルを生成する際に、リアル発話データを取得する際に選択したリアル発話音声コーパス１０５に関する情報が用いられる。このようにすれば、リアル発話データの分類に用いられた所定の条件（例えば、分野）が考慮されることになるため、より適切な変換モデルを生成することが可能となる。

【００８５】

＜第７実施形態＞

第７実施形態に係る音声認識システム１０について、図１５及び図１６を参照して説明する。なお、第７実施形態は、上述した第１から第６実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１から第６実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。

【００８６】

（機能的構成）

まず、図１５を参照しながら、第７実施形態に係る音声認識システム１０の機能的構成について説明する。図１５は、第７実施形態に係る音声認識システムの機能的構成を示すブロック図である。なお、図１５では、図２で示した構成要素と同様の要素に同一の符号を付している。

【００８７】

図１５に示すように、第７実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、ノイズ付与部１６０と、音声変換部２１０と、音声認識部２２０と、を備えて構成されている。即ち、第７実施形態に係る音声認識システム１０は、第１実施形態の構成（図２参照）に加えて、ノイズ付与部１６０を更に備えている。なお、ノイズ付与部１６０は、例えば上述したプロセッサ１１（図１参照）によって実現される処理ブロックであってよい。

【００８８】

ノイズ付与部１６０は、テキスト変換部１２０で生成されるテキストデータにノイズを付与可能に構成されている。ノイズ付与部１６０は、例えば、テキスト変換前のリアル発話データにノイズを付与することで、テキストデータにノイズが付与されるようにしてもよいし、テキスト変換後のテキストデータにノイズを付与するようにしてもよい。或いは、ノイズ付与部１６０は、テキスト変換部１２０がリアル発話データをテキスト変換する際にノイズを付与するようにしてよい。ノイズ付与部１６０は、予め設定されたノイズを付与するようにしてもよいし、ランダムに設定したノイズを付与するようにしてもよい。

【００８９】

（変換モデル生成動作）

次に、図１６を参照しながら、第７実施形態に係る音声認識システム１０による変換モデル生成動作の流れについて説明する。図１６は、第７実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。なお、図１６では、図３に示した処理と同様の処理に同一の符号を付している。

【００９０】
図１６に示すように、第７実施形態に係る音声認識システム１０による変換モデル生成動作が開始されると、まず発話データ取得部１１０が、リアル発話データを取得する（ステップＳ１０１）。ここで本実施形態では特に、ノイズ付与部１６０がテキスト変換部１２０にノイズ情報を出力する（ステップＳ７０１）。そして、テキスト変換部１２０は、発話データ取得部１１０で取得されたリアル発話データを、ノイズが付与されたテキストデータに変換する（ステップＳ７０２）。
【００９１】
続いて、音声合成部１３０が、テキスト変換部１２０で変換されたテキストデータ（ここでは、ノイズが付与されたテキストデータ）を音声合成し、リアル発話データに対応する対応合成音声を生成する（ステップＳ１０３）。そして、変換モデル生成部１４０が、発話データ取得部１１０で取得されたリアル発話データ及び音声合成部１３０で生成された対応合成音声に基づいて、変換モデルを生成する（ステップＳ１０４）。その後、変換モデル生成部１４０は、生成した変換モデルを音声変換部２１０に出力する（ステップＳ１０５）。
【００９２】
（技術的効果）
次に、第７実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。
【００９３】
図１５及び図１６で説明したように、第７実施形態に係る音声認識システム１０では、リアル発話データが、ノイズが付与されたテキストデータに変換される。このようにすれば、ノイズを含むデータを用いて変換モデルが生成されることになるため、ノイズに強い変換モデル（例えば、入力音声にノイズが含まれていても適切に音声変換できる変換モデル）を生成することが可能である。
【００９４】
＜第７実施形態の変形例＞
第７実施形態の変形例に係る音声認識システム１０について、図１７及び図１８を参照して説明する。なお、第７実施形態の変形例は、上述した第７実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１から第７実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。
【００９５】
（機能的構成）
まず、図１７を参照しながら、第７実施形態の変形例に係る音声認識システム１０の機能的構成について説明する。図１７は、第７実施形態の変形例に係る音声認識システムの機能的構成を示すブロック図である。なお、図１７では、図１５で示した構成要素と同様の要素に同一の符号を付している。
【００９６】
図１７に示すように、第７実施形態の変形例に係る音声認識システム１０は、その機能を実現するための構成要素として、発話データ取得部１１０と、テキスト変換部１２０と、音声合成部１３０と、変換モデル生成部１４０と、ノイズ付与部１６０と、音声変換部２１０と、音声認識部２２０と、を備えて構成されている。ただし、第７実施形態の変形例に係る音声認識システム１０では、ノイズ付与部１６０が、音声合成部１３０にノイズ情報を出力可能に構成されている。即ち、第７実施形態の変形例では、音声合成部１３０による音声合成の際にノイズが付与される構成となっている。
【００９７】
（変換モデル生成動作）
次に、図１８を参照しながら、第７実施形態の変形例に係る音声認識システム１０による変換モデル生成動作の流れについて説明する。図１８は、第７実施形態の変形例に係る

音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。なお、図１８では、図１６に示した処理と同様の処理に同一の符号を付している。

【００９８】

図１８に示すように、第７実施形態の変形例に係る音声認識システム１０による変換モデル生成動作が開始されると、まず発話データ取得部１１０が、リアル発話データを取得する（ステップＳ１０１）。そして、テキスト変換部１２０が、発話データ取得部１１０で取得されたリアル発話データをテキストデータに変換する（ステップＳ１０２）。

【００９９】

続いて、本実施形態では特に、ノイズ付与部１６０が音声合成部１３０にノイズ情報を出力する（ステップＳ７５１）。そして、音声合成部１３０は、テキスト変換部１２０で変換されたテキストデータを音声合成し、ノイズが付与された対応合成音声を生成する（ステップＳ７５２）。

【０１００】

続いて、変換モデル生成部１４０が、発話データ取得部１１０で取得されたリアル発話データ及び音声合成部１３０で生成された対応合成音声（ここでは、ノイズが付与された対応合成音声）に基づいて、変換モデルを生成する（ステップＳ１０４）。その後、変換モデル生成部１４０は、生成した変換モデルを音声変換部２１０に出力する（ステップＳ１０５）。

【０１０１】

（技術的効果）

次に、第７実施形態の変形例に係る音声認識システム１０によって得られる技術的効果について説明する。

【０１０２】

図１７及び図１８で説明したように、第７実施形態の変形例に係る音声認識システム１０では、ノイズが付与された対応合成音声が生成される。このようにすれば、ノイズを含むデータを用いて変換モデルが生成されることになるため、ノイズに強い変換モデル（例えば、入力音声にノイズが含まれていても適切に音声変換できる変換モデル）を生成することが可能である。

【０１０３】

＜第８実施形態＞

第８実施形態に係る音声認識システム１０について、図１９から図２１を参照して説明する。なお、第８実施形態は、上述した第１から第７実施形態と一部の構成及び動作が異なるのみであり、その他の部分については第１から第７実施形態と同一であってよい。このため、以下では、すでに説明した各実施形態と異なる部分について詳細に説明し、その他の重複する部分については適宜説明を省略するものとする。

【０１０４】

（機能的構成）

まず、図１９を参照しながら、第８実施形態に係る音声認識システム１０の機能的構成について説明する。図１９は、第８実施形態に係る音声認識システムの機能的構成を示すブロック図である

【０１０５】

図１９に示すように、第８実施形態に係る音声認識システム１０は、その機能を実現するための構成要素として、手話データ取得部４１０と、テキスト変換部４２０と、音声合成部４３０と、変換モデル生成部４４０と、音声変換部５１０と、音声認識部５２０と、を備えて構成されている。手話データ取得部４１０、テキスト変換部４２０、音声合成部４３０、変換モデル生成部４４０、音声変換部５１０、音声認識部５２０の各々は、例えば上述したプロセッサ１１（図１参照）によって実現される処理ブロックであってよい。

【０１０６】

手話データ取得部４１０は、手話発話データを取得可能に構成されている。手話データは、例えば手話の動画データであってよい。手話データは、例えば複数の手話データを蓄

積するデータベース（手話コーパス）から取得されてよい。手話データ取得部４１０で取得された手話データは、テキスト変換部１２０及び変換モデル生成部１４０に出力される構成となっている。

【０１０７】

テキスト変換部４２０は、手話データ取得部４１０で取得された手話データをテキストデータに変換可能に構成されている。即ち、テキスト変換部４２０は、手話データに含まれる手話の内容をテキスト変換する処理を実行可能に構成されている。なお、テキスト変換の具体的な手法については、既存の技術が適宜採用されてよい。テキスト変換部４２０で変換されたテキストデータ（即ち、手話データに対応するテキストデータ）は、音声合成部４３０に出力される構成となっている。

【０１０８】

音声合成部４３０は、テキスト変換部４２０で変化されたテキストデータを音声合成することで、手話データに対応する対応合成音声を生成可能に構成されている。なお、音声合成の具体的な手法については、既存の技術を適宜採用することができる。音声合成部４３０で生成された対応合成音声は、変換モデル生成部４４０に出力される構成となっている。なお、対応合成音声は、複数の対応合成を蓄積可能なデータベース（合成音声コーパス）に蓄積されてから、変換モデル生成部４４０に出力されてもよい。

【０１０９】

変換モデル生成部４４０は、手話データ取得部４１０で取得された手話データと、音声合成部４３０で合成された対応合成音声を用いて、入力手話を合成音声に変換する変換モデルを生成可能に構成されている。変換モデルは、例えば、入力される入力手話（例えば、手話の動画）を、合成音声（即ち、機械的な音声）に変換する。変換モデル生成部４４０は、例えばＧＡＮを用いて、変換モデルを生成するように構成されてよい。変換モデル生成部４４０で生成された変換モデルは、音声変換部５１０に出力される構成となっている。

【０１１０】

音声変換部５１０は、変換モデル生成部４４０で生成された変換モデルを用いて、入力手話を合成音声に変換可能に構成されている。音声変換部５１０に入力される入力手話は、例えばカメラ等を用いて入力される動画であってよい。音声変換部５１０で変換された合成音声は、音声認識部５２０に出力される構成となっている。

【０１１１】

音声認識部５２０は、音声変換部５１０で変換された合成音声を音声認識することが可能に構成されている。即ち、音声認識部５２０は、合成音声をテキスト化する処理を実行可能に構成されている。音声認識部５２０は、合成音声の音声認識結果を出力可能に構成されてよい。なお、音声認識結果の利用方法については特に限定されない。

【０１１２】

（変換モデル生成動作）

次に、図２０を参照しながら、第８実施形態に係る音声認識システム１０による変換モデル生成動作の流れについて説明する。図２０は、第８実施形態に係る音声認識システムによる変換モデル生成動作の流れを示すフローチャートである。

【０１１３】

図２０に示すように、第８実施形態に係る音声認識システム１０による変換モデル生成動作が開始されると、まず手話データ取得部４１０が、手話データを取得する（ステップＳ８０１）。そして、テキスト変換部４２０が、手話データ取得部４１０で取得された手話データをテキストデータに変換する（ステップＳ８０２）。

【０１１４】

続いて、音声合成部４３０が、テキスト変換部４２０で変換されたテキストデータを音声合成し、手話データに対応する対応合成音声を生成する（ステップＳ４０３）。そして、変換モデル生成部１４０が、手話データ取得部４１０で取得された手話データ及び音声合成部４３０で生成された対応合成音声に基づいて、変換モデルを生成する（ステップＳ

８０４）。その後、変換モデル生成部４４０は、生成した変換モデルを音声変換部５１０に出力する（ステップＳ８０５）。

【０１１５】

（変換認識動作）

次に、図２１を参照しながら、第８実施形態に係る音声認識システム１０による音声認識動作の流れについて説明する。図２１は、第８実施形態に係る音声認識システムによる音声認識動作の流れを示すフローチャートである。

【０１１６】

図２１に示すように、第１実施形態に係る音声認識システム１０による音声認識動作が開始されると、まず音声変換部５１０が入力手話を取得する（ステップＳ８５１）。そして、音声変換部５１０は、変換モデル生成部４４０で生成された変換モデルを読み込む（ステップＳ８５２）。その後、音声変換部２１０は、読み込んだ変換モデルを用いて音声変換を行い、入力手話を合成音声に変換する（ステップＳ８５３）。

【０１１７】

続いて、音声認識部５２０は、音声認識モデルを読み込む（ステップＳ８５４）。そして、音声認識部５２０は、読み込んだ音声認識モデルを用いて、音声変換部５１０で合成された合成音声を音声認識する（ステップＳ８５５）。その後、音声認識部５２０は、音声認識結果を出力する（ステップＳ８５６）。

【０１１８】

（技術的効果）

次に、第８実施形態に係る音声認識システム１０によって得られる技術的効果について説明する。

【０１１９】

図１９から図２１で説明したように、第８実施形態に係る音声認識システム１０では、変換モデルを生成する際に、手話データ及び手話データに対応する対応合成音声が用いられる。そして特に、手話データに対応する対応合成音声は、手話データをテキスト変換し、テキストデータを音声合成することで生成される。このようにすれば、手話データと、それに対応する合成音声と、の両方を用意する必要がなくなる（即ち、手話データのみ用意すれば、対応合成音声を生成できる）ため、変換モデルを生成するのに要するコストを抑制することができる。その結果、低コストで認識精度の高い音声認識を実現することが可能となる。

【０１２０】

上述した各実施形態の機能を実現するように該実施形態の構成を動作させるプログラムを記録媒体に記録させ、該記録媒体に記録されたプログラムをコードとして読み出し、コンピュータにおいて実行する処理方法も各実施形態の範疇に含まれる。すなわち、コンピュータ読取可能な記録媒体も各実施形態の範囲に含まれる。また、上述のプログラムが記録された記録媒体はもちろん、そのプログラム自体も各実施形態に含まれる。

【０１２１】

記録媒体としては例えばフロッピー（登録商標）ディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、磁気テープ、不揮発性メモリカード、ＲＯＭを用いることができる。また該記録媒体に記録されたプログラム単体で処理を実行しているものに限らず、他のソフトウェア、拡張ボードの機能と共同して、ＯＳ上で動作して処理を実行するものも各実施形態の範疇に含まれる。更に、プログラム自体がサーバに記憶され、ユーザ端末にサーバからプログラムの一部または全てをダウンロード可能なようにしてもよい。

【０１２２】

＜付記＞

以上説明した実施形態に関して、更に以下の付記のようにも記載されうるが、以下には限られない。

【０１２３】

（付記１）
付記１に記載の音声認識システムは、話者が発話したリアル発話データを取得する発話データ取得手段と、前記リアル発話データをテキストデータに変換するテキスト変換手段と、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成する音声合成手段と、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成する変換モデル生成手段と、前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、を備える音声認識システムである。
【０１２４】
（付記２）
付記２に記載の音声認識システムは、前記変換モデル生成手段は、前記入力音声と、前記音声認識手段の認識結果と、を用いて前記変換モデルのパラメータを調整する、付記１に記載の音声認識システムである。
【０１２５】
（付記３）
付記３に記載の音声認識システムは、前記対応合成音声を含むデータを用いて音声認識モデルを生成する音声認識モデル生成手段を更に備え、前記音声認識手段は、前記音声認識モデルを用いて音声認識する、付記１又は２に記載の音声認識システムである。
【０１２６】
（付記４）
付記４に記載の音声認識システムは、前記音声認識モデル生成手段は、前記変換モデルを用いて変換された前記合成音声と、前記音声認識手段の認識結果と、を用いて前記音声認識モデルのパラメータを調整する、付記３に記載の音声認識システムである。
【０１２７】
（付記５）
付記５に記載の音声認識システムは、前記話者の属性を示す属性情報を取得する属性取得手段を更に備え、前記音声合成手段は、前記属性情報を用いて音声合成を行うことで前記対応合成音声を生成する、付記１から４のいずれか一項に記載の音声認識システムムである。
【０１２８】
（付記６）
付記６に記載の音声認識システムは、所定の条件ごとに前記リアル発話データを記憶する複数のリアル発話音声コーパスを更に備え、前記発話データ取得手段は、前記複数のリアル発話音声コーパスから１つを選択して前記リアル発話データを取得する、付記１から５のいずれか一項に記載の音声認識システムである。
【０１２９】
（付記７）
付記７に記載の音声認識システムは、前記テキストデータ及び前記対応合成音声の少なくとも一方にノイズを付与するノイズ付与手段を更に備える、付記１から６のいずれか一項に記載の音声認識システムである。
【０１３０】
（付記８）
付記８に記載の音声認識システムは、手話データを取得する手話データ取得手段と、前記手話データをテキストデータに変換するテキスト変換手段と、前記テキストデータを用いた音声合成により、前記手話データに対応する対応合成音声を生成する音声合成手段と、前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成する変換モデル生成手段と、前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、を備える音声認識システムである。
【０１３１】
（付記９）

付記９に記載の音声認識方法は、少なくとも１つのコンピュータによって、話者が発話したリアル発話データを取得し、前記リアル発話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する、音声認識方法である。

【０１３２】

（付記１０）

付記１０に記載の記録媒体は、少なくとも１つのコンピュータに、話者が発話したリアル発話データを取得し、前記リアル発話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する、音声認識方法を実行させるコンピュータプログラムが記録された記録媒体である。

【０１３３】

（付記１１）

付記１１に記載のコンピュータプログラムは、少なくとも１つのコンピュータに、話者が発話したリアル発話データを取得し、前記リアル発話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成し、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する、音声認識方法を実行させるコンピュータプログラムである。

【０１３４】

（付記１２）

付記１２に記載の音声認識装置は、話者が発話したリアル発話データを取得する発話データ取得手段と、前記リアル発話データをテキストデータに変換するテキスト変換手段と、前記テキストデータを用いた音声合成により、前記リアル発話データに対応する対応合成音声を生成する音声合成手段と、前記リアル発話データ及び前記対応合成音声を用いて、入力音声を合成音声に変換する変換モデルを生成する変換モデル生成手段と、前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、を備える音声認識装置である。

【０１３５】

（付記１３）

付記１３に記載の音声認識方法は、少なくとも１つのコンピュータによって、手話データを取得し、前記手話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記手話データに対応する対応合成音声を生成し、前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する、音声認識方法である。

【０１３６】

（付記１４）

付記１４に記載の記録媒体は、少なくとも１つのコンピュータに、手話データを取得し、前記手話データをテキストデータに変換し、前記テキストデータを用いた音声合成により、前記手話データに対応する対応合成音声を生成し、前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する、音声認識方法を実行させるコンピュータプログラムが記録された記録媒体である。

【０１３７】

（付記１５）

付記１５に記載のコンピュータプログラムは、少なくとも１つのコンピュータに、手話データを取得し、前記手話データをテキストデータに変換し、前記テキストデータを用い

た音声合成により、前記手話データに対応する対応合成音声を生成し、前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成し、前記変換モデルを用いて変換された前記合成音声を音声認識する、音声認識方法を実行させるコンピュータプログラムである。

【０１３８】

（付記１６）

付記１６に記載の音声認識装置は、手話データを取得する手話データ取得手段と、前記手話データをテキストデータに変換するテキスト変換手段と、前記テキストデータを用いた音声合成により、前記手話データに対応する対応合成音声を生成する音声合成手段と、前記手話データ及び前記対応合成音声を用いて、入力される手話を合成音声に変換する変換モデルを生成する変換モデル生成手段と、前記変換モデルを用いて変換された前記合成音声を音声認識する音声認識手段と、を備える音声認識装置である。

【０１３９】

この開示は、請求の範囲及び明細書全体から読み取ることのできる発明の要旨又は思想に反しない範囲で適宜変更可能であり、そのような変更を伴う音声認識システム、音声認識方法、及び記録媒体もまたこの開示の技術思想に含まれる。

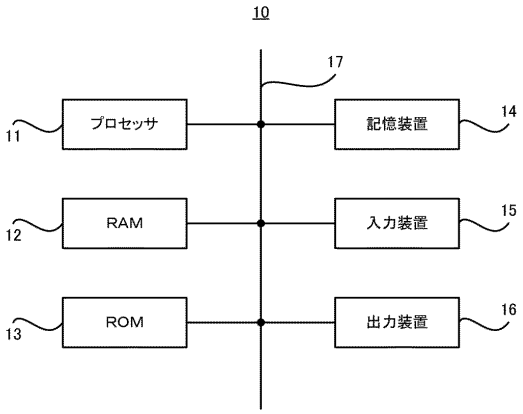
【符号の説明】

【０１４０】

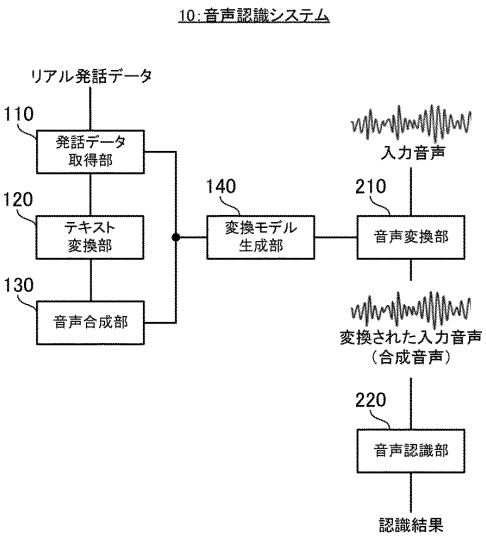
１０　音声認識システム
１１　プロセッサ
１４　記憶装置
１０５　リアル発話音声コーパス
１１０　発話データ取得部
１２０　テキスト変換部
１３０　音声合成部
１４０　変換モデル生成部
１５０　属性情報取得部
１６０　ノイズ付与部
２１０　音声変換部
２２０　音声認識部
３１０　音声認識モデル生成部
４１０　手話データ取得部
４２０　テキスト変換部
４３０　音声合成部
４４０　変換モデル生成部
５１０　音声変換部
５２０　音声認識部

【図面】

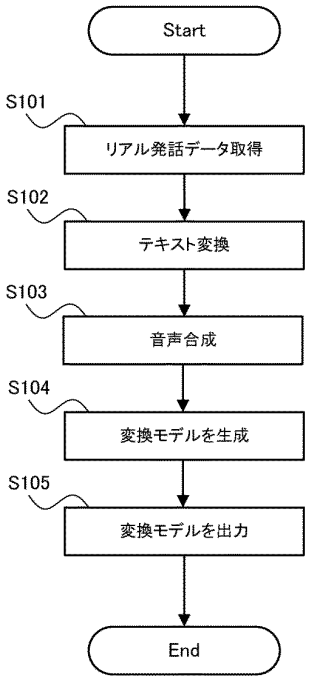
【図１】


10
17
プロセッサ
11
記憶装置
14
RAM
12
入力装置
15
ROM
13
出力装置
16

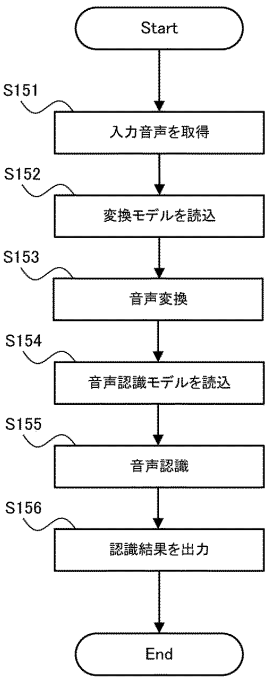
【図２】


10:音声認識システム
リアル発話データ
110
発話データ取得部
120
テキスト変換部
130
音声合成部
140
変換モデル生成部
入力音声
210
音声変換部
変換された入力音声（合成音声）
220
音声認識部
認識結果

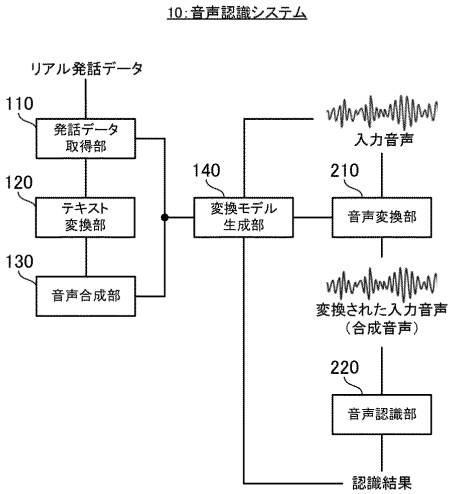
【図３】


Start
S101
リアル発話データ取得
S102
テキスト変換
S103
音声合成
S104
変換モデルを生成
S105
変換モデルを出力
End

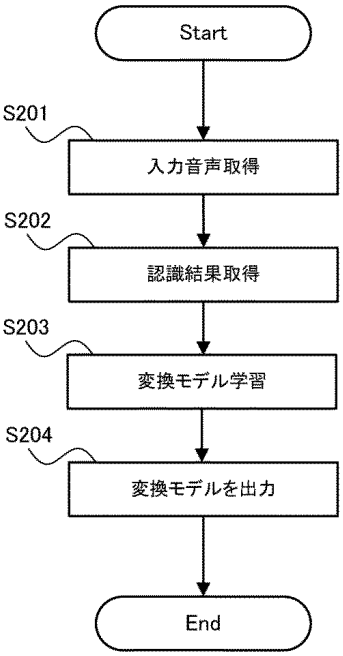
【図４】


Start
S151
入力音声を取得
S152
変換モデルを読込
S153
音声変換
S154
音声認識モデルを読込
S155
音声認識
S156
認識結果を出力
End

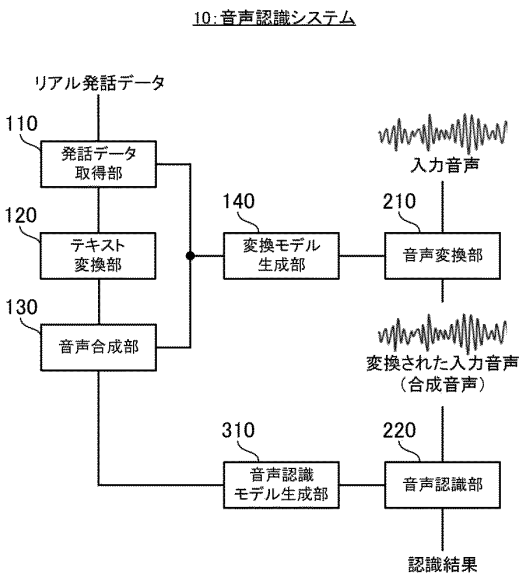
【図５】


10:音声認識システム
リアル発話データ
110
発話データ取得部
120
テキスト変換部
130
音声合成部
140
変換モデル生成部
入力音声
210
音声変換部
変換された入力音声（合成音声）
220
音声認識部
認識結果

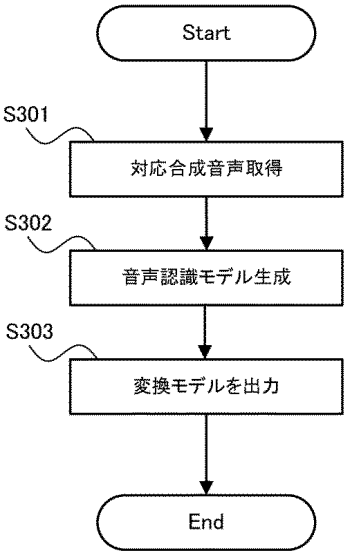
【図６】


Start
S201
入力音声取得
S202
認識結果取得
S203
変換モデル学習
S204
変換モデルを出力
End

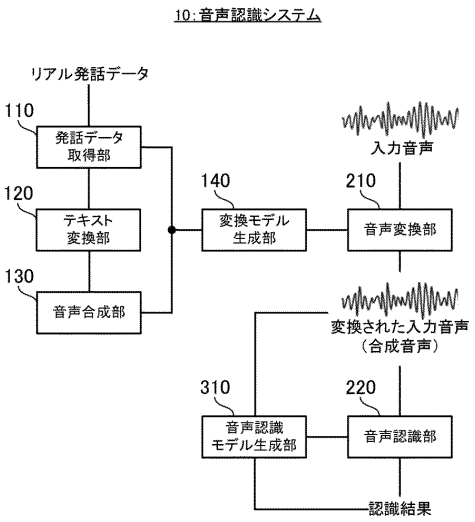
【図７】


10:音声認識システム
リアル発話データ
110
発話データ取得部
120
テキスト変換部
130
音声合成部
140
変換モデル生成部
入力音声
210
音声変換部
変換された入力音声（合成音声）
310
音声認識モデル生成部
220
音声認識部
認識結果

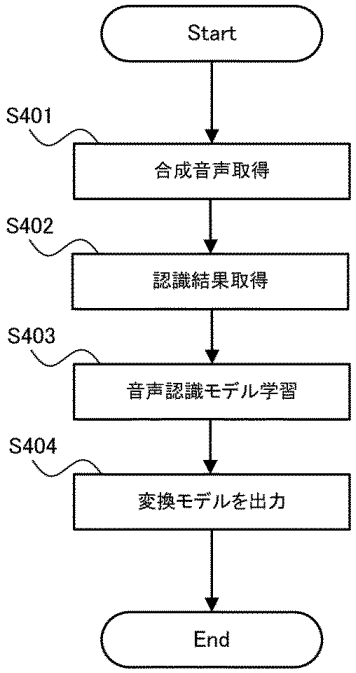
【図８】


Start
S301
対応合成音声取得
S302
音声認識モデル生成
S303
変換モデルを出力
End

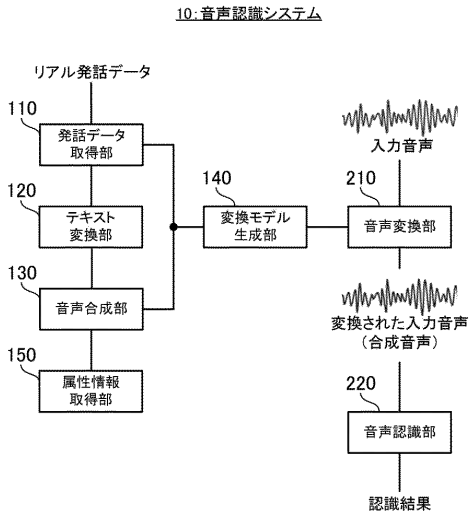
【図９】



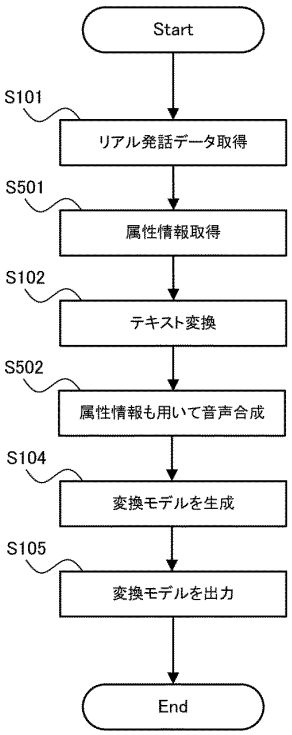
【図１０】



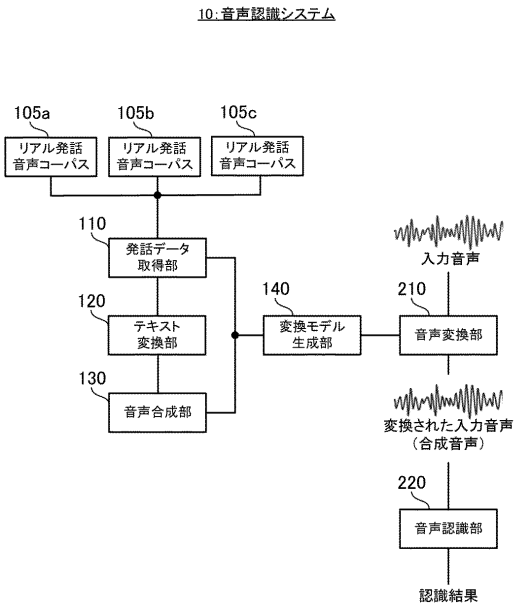
【図１１】



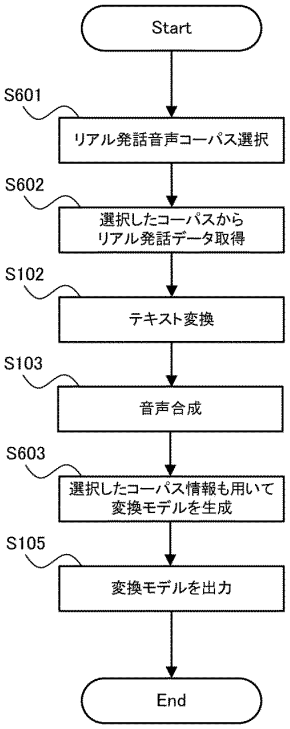
【図１２】



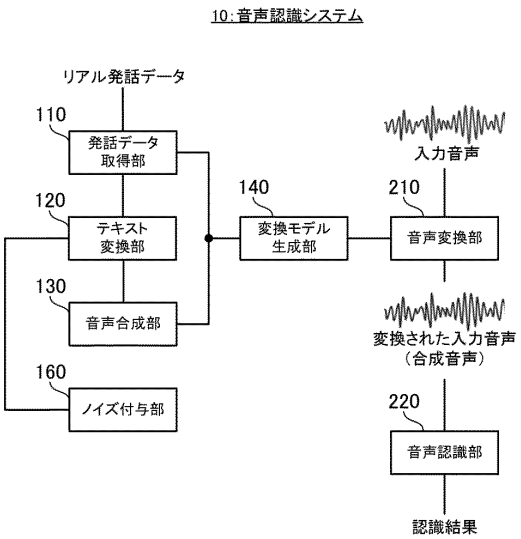
【図１３】



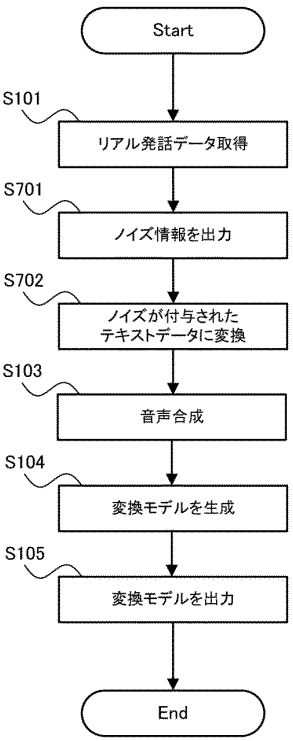
【図１４】



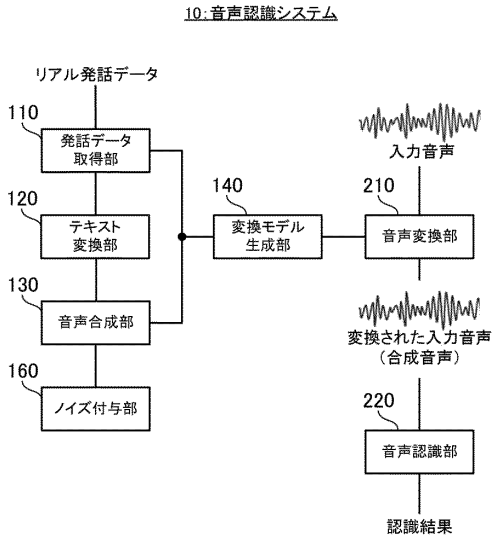
【図１５】



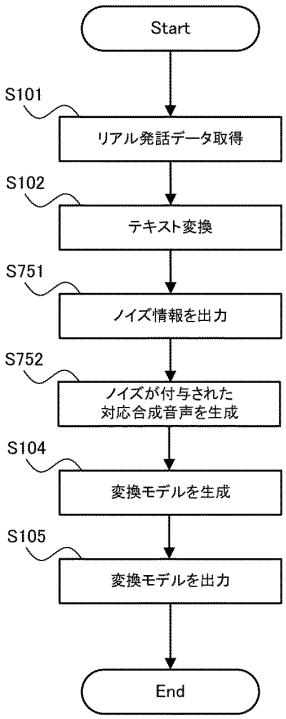
【図１６】



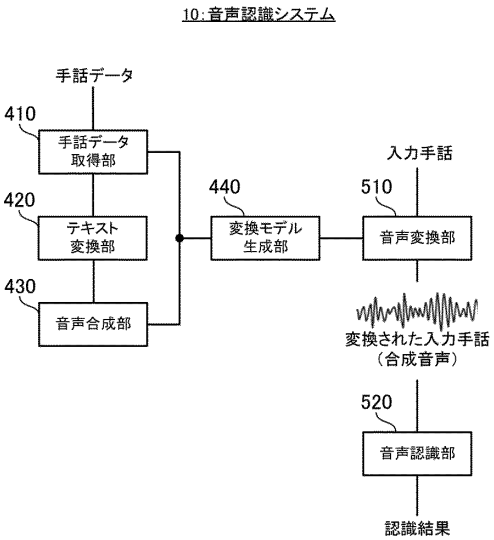
【図１７】



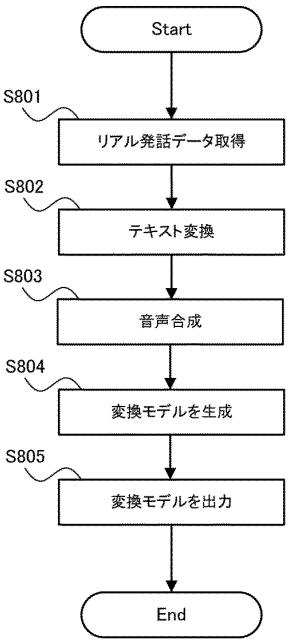
【図１８】



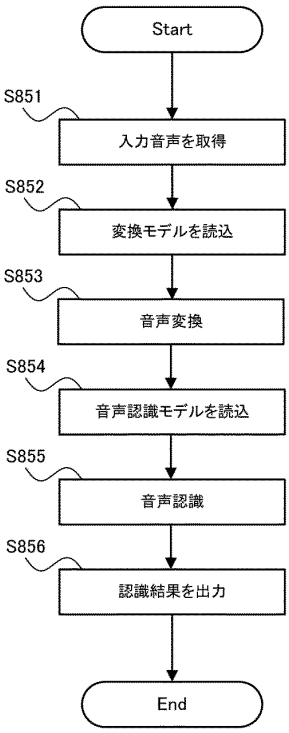
【図１９】



【図２０】



【図２１】


Start
S851
入力音声を取得
S852
変換モデルを読込
S853
音声変換
S854
音声認識モデルを読込
S855
音声認識
S856
認識結果を出力
End

フロントページの続き

審査官　大野 弘
(56)参考文献　特開２０１９－００８１２０（ＪＰ，Ａ）
特表２００３－５２２９７８（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　２１／００７
Ｇ１０Ｌ　１３／０２
Ｇ１０Ｌ　１５／２０