(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6900536号
(P6900536)
(45)発行日　令和3年7月7日(2021.7.7)　(24)登録日　令和3年6月18日(2021.6.18)

(51)Int.Cl.　F I
G10L 13/047 (2013.01)　G10L 13/047 Z

請求項の数 14　(全 21 頁)

(21)出願番号　特願2020-56285(P2020-56285)
(22)出願日　令和2年3月26日(2020.3.26)
(65)公開番号　特開2021-56489(P2021-56489A)
(43)公開日　令和3年4月8日(2021.4.8)
審査請求日　令和2年3月26日(2020.3.26)
(31)優先権主張番号　201910927040.3
(32)優先日　令和1年9月27日(2019.9.27)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　513224353
バイドゥ　オンライン　ネットワーク　テクノロジー　(ペイジン)　カンパニー　リミテッド
中華人民共和国、１０００８５　ペイジン　ハイディエン　ディストリクト、シャンディ　１０ティーエイチ　ストリート、バイドゥ　キャンパス、ナンバー　１０、３／フロア
(74)代理人　100118913
弁理士　上田　邦生
(74)代理人　100142789
弁理士　柳　順一郎
(74)代理人　100163050
弁理士　小栗　眞由美

最終頁に続く

(54)【発明の名称】音声合成モデルのトレーニング方法、装置、電子機器及び記憶媒体

(57)【特許請求の範囲】
【請求項１】
音声合成モデルのトレーニング方法であって、
前記方法は、
現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップと、
符号化標記された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを融合して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの一つの加重組み合わせを取得するステップと、
前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得するステップと、
前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を前記トレーニング対象モデルのデコーダの入力とし、前記デコーダの出力端で前記現在のサンプルの音声メルスペクトル出力を取得するステップと、を含むことを特徴とする、音声合成モデルのトレーニング方法。
【請求項２】

前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップは、

前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを一つの共有のエンコーダに入力するステップと、

前記共有のエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップと、を含むことを特徴とする、請求項１に記載の音声合成モデルのトレーニング方法。

【請求項３】

前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップは、

前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをそれぞれ取得するステップと、

畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とし、前記シーケンス変換ニューラルネットワークモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップと、を含むことを特徴とする、請求項１に記載の音声合成モデルのトレーニング方法。

【請求項４】

前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップは、

前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力するステップと、

各独立したエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップと、を含むことを特徴とする、請求項１に記載の音声合成モデルのトレーニング方法。

【請求項５】

前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とする前に、前記方法は、

前記現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換するステップと、

前記音節のベクトル標記と前記漢字のベクトル標記とを前記音素のベクトル標記と同じ長さに変換して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを取得し、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを前記トレーニング対象モデルのエンコーダの入力とする操作を実行するステップと、をさらに含むことを特徴とする、請求項１に記載の音声合成モデルのトレーニング方法。

【請求項６】

前記音素入力シーケンスは、声調入力シーケンスと、児化音入力シーケンスと、句読点入力シーケンスと、３５個の独立した韻母の入力シーケンスとを含み、

前記音素入力シーケンスは、１０６個の音素単位を含み、各音素単位は、１０６ビットを含み、前記１０６ビットにおいて、有効ビットの値は１であり、非有効ビットの値は０であり、

前記漢字入力シーケンスは、３０００個の漢字の入力シーケンスを含み、
前記音節入力シーケンスは、５０８個の音節の入力シーケンスを含むことを特徴とする、請求項１に記載の音声合成モデルのトレーニング方法。
【請求項７】
音声合成モデルのトレーニング装置であって、
前記装置は、入力モジュールと、融合モジュールと、出力モジュールと、を含み、
前記入力モジュールは、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得し、
前記融合モジュールは、符号化標記された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを融合して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの一つの加重組み合わせを取得し、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得し、
前記出力モジュールは、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を前記トレーニング対象モデルのデコーダの入力とし、前記デコーダの出力端で前記現在のサンプルの音声メルスペクトル出力を取得することを特徴とする、音声合成モデルのトレーニング装置。
【請求項８】
前記入力モジュールは、具体的には、
前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを一つの共有のエンコーダに入力し、
前記共有のエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得することを特徴とする、請求項７に記載の音声合成モデルのトレーニング装置。
【請求項９】
前記入力モジュールは、具体的には、
前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをそれぞれ取得し、
畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とし、前記シーケンス変換ニューラルネットワークモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得することを特徴とする、請求項７に記載の音声合成モデルのトレーニング装置。
【請求項１０】
前記入力モジュールは、具体的には、
前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力し、
各独立したエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得することを特徴とする、請求項７に記載の音声合成モデルのトレーニング装置。
【請求項１１】
前記装置は、変換モジュールをさらに含み、
前記変換モジュールは、前記現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換し、前記音節のベクトル標記と前記漢字のベクト

ル標記とを前記音素のベクトル標記と同じ長さに変換して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを取得し、
前記入力モジュールは、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを前記トレーニング対象モデルのエンコーダの入力とする操作を実行することを特徴とする、請求項７に記載の音声合成モデルのトレーニング装置。
【請求項１２】
前記音素入力シーケンスは、声調入力シーケンスと、児化音入力シーケンスと、句読点入力シーケンスと、３５個の独立した韻母の入力シーケンスとを含み、
前記音素入力シーケンスは、１０６個の音素単位を含み、
各音素単位は、１０６ビットを含み、前記１０６ビットにおいて、有効ビットの値は１であり、非有効ビットの値は０であり、
前記漢字入力シーケンスは、３０００個の漢字の入力シーケンスを含み、
前記音節入力シーケンスは、５０８個の音節の入力シーケンスを含むことを特徴とする、請求項７に記載の音声合成モデルのトレーニング装置。
【請求項１３】
電子機器であって、
少なくとも一つのプロセッサと、
前記少なくとも一つのプロセッサと通信可能に接続されるメモリと、を含み、
前記メモリには前記少なくとも一つのプロセッサによって実行可能な命令が記憶され、
前記命令が前記少なくとも一つのプロセッサによって実行される場合に、前記少なくとも一つのプロセッサが請求項１から６のいずれかに記載の方法を実行することを特徴とする、電子機器。
【請求項１４】
コンピュータ命令が記憶されている非一時的なコンピュータ読み取り可能な記憶媒体であって、
前記コンピュータ命令は、前記コンピュータに請求項１から６のいずれかに記載の方法を実行させることを特徴とする、非一時的なコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、人工知能技術の分野に関し、さらに、コンピュータ知能音声の分野に関し、特に、音声合成モデルのトレーニング方法、装置、電子機器及び記憶媒体に関する。
【背景技術】
【０００２】
音声合成の分野では、ＷａｖｅＮｅｔやＷａｖｅＲＮＮなどのニューラルネットワークに基づく方法は、合成音声の音質及び自然度を大きく改善する。このような方法は、通常、フロントエンドシステムテキストに基づいて言語特徴を抽出し、基本周波数及び時間などの情報を予測する必要がある。グーグル（Ｇｏｏｇｌｅ）（登録商標）が提供するエンドツーエンドモデリングのＴａｃｏｔｒｏｎモデルは、大量の専門知識を必要とする複雑なフロントエンドシステムを脱却し、シーケンス変換モデルを介してサウンドライブラリにおける音声のリズム及び感情などの情報を自動的に学習するため、合成された音声は、表現力の面で特に優れている。しかし、Ｔａｃｏｔｒｏｎモデルの中国語への応用は、多くの挑戦がある。主に中国語では漢字の数が多く、一般的に使用される漢字は数千個あり、同じ音の漢字の現象が非常に一般的であり、同音異字の発音方式には違いがあり、同じ漢字であっても異なる単語又は文脈中では発音方式も異なるからである。
【０００３】
従来技術において、Ｔａｃｏｔｒｏｎモデルを中国語での応用に実現した技術案では、多くが三つに分けられる。（１）Ｔａｃｏｔｒｏｎの英語での応用と類似し、直接漢字を

入力要素とし、（２）漢字を音節に縮めて入力要素とし、（３）音節を音素に分割して入力要素とする。上記の技術案（１）を採用すると、漢字の数が多いため、通常、音声合成のトレーニングに使用されるサウンドライブラリは数時間から数十時間程度の規模であるので、直接漢字をモデルの入力要素とすると、データが少ないため、多くの低頻度の漢字の発音が十分に学習することができない。上記の技術案（２）及び（３）を採用して、音素又は音節を入力要素とすると、漢字の数が少ないという問題を解決することができ、漢字中の同じ音の漢字は、共有ユニットによってより十分なトレーニングを取得することができる。しかし、異なる漢字は、同じ発音であっても、発音方式が明確に違い、よく見られる虚語では、通常、発音が弱く、実語は発音が明確である。このため、Ｔａｃｏｔｒｏｎモデルでは、虚語の発音方式で学習する傾向があるため、合成効果が良くないという問題がある。また、音素を入力要素とすると、ある韻母は、単独に一つの完全な音節とすることができる場合があり、この二つのケースでは、韻母の発音が実際に一定の違いがあるので、独立した音節の韻母としてはより完全な発音プロセスが必要となるが、音素に基づくモデルは、この二つの場合を区別できないため、韻母が独立した場合の発音が不十分になるという別の問題も存在する。

【発明の概要】

【発明が解決しようとする課題】

【０００４】

これを考慮して、本出願の実施例は、発音効果を効果的に改善し、音声製品に高い表現力と高い自然度の中国語合成音声を提供することができる音声合成モデルのトレーニング方法、装置、電子機器及び記憶媒体を提供する。

【課題を解決するための手段】

【０００５】

第１の態様では、本出願の実施例は、音声合成モデルのトレーニング方法を提供し、前記方法は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの符号化標記を取得するステップと、符号化標記された前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスを融合して、前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの一つの加重組み合わせを取得するステップと、前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの加重組み合わせの各時点における加重平均を取得するステップと、前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの加重組み合わせの各時点における加重平均を前記トレーニング対象モデルのデコーダの入力とし、前記デコーダの出力端で前記現在のサンプルの音声メル（Ｍｅｌ）スペクトル出力を取得するステップと、を含む。

【０００６】

上記の実施例は、トレーニング対象モデルの入力端の入力テキスト及びトレーニング対象モデルの出力端の出力音声によって、トレーニング対象モデルのエンコーダ及びデコーダの共同トレーニングを実現する。本出願では、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合する技術的手段を採用するため、従来技術では、音節入力シーケンス又は音素入力シーケンス又は漢字入力シーケンスのみしか採用しないために音声合成効果が良くないという技術問題を克服し、発音効果を効果的に改善し、音声製品に高い表現力と高い自然度の中国語合成音声の技術的効果を提供するという利点又は有益な効果を奏する。

【０００７】

上記の実施例では、前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エン

コーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップは、前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスを一つの共有のエンコーダに入力するステップと、前記共有のエンコーダの出力端で前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの符号化標記を取得するステップと、を含む。

【０００８】

上記の実施例は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを一つの共有のエンコーダに入力することによって、共有のエンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができるという利点又は有益な効果を奏する。

【０００９】

上記の実施例では、前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの符号化標記を取得するステップは、前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスを入力到三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスをそれぞれ取得するステップと、畳み込み変換された前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスをシーケンス変換ニューラルネットワークモジュールの入力とし、前記シーケンス変換ニューラルネットワークモジュールの出力端で前記音節入力シーケンス、前記音素入力シーケンス、及び前記漢字入力シーケンスの符号化標記を取得するステップと、を含む。

【００１０】

上記の実施例は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを、三つの独立した畳み込み層変換モジュールにそれぞれ入力し、畳み込み変換された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とすることにより、シーケンス変換ニューラルネットワークモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができるという利点又は有益な効果を奏する。

【００１１】

上記の実施例では、前記現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップは、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力するステップと、各独立したエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップと、を含む。

【００１２】

上記の実施例は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力し、各独立したエンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得するという利点又は有益な効果を有する。実験により、音素の独立した韻母及び音素、音節、漢字を融合する三つの技術案は、エンドツーエンドの中国語音声合成の問題をある程度で解決することができ、そのうち、独立エンコーダの効果が最適である。測定の結果は、発音問題の割合が２％から０．４％に減少することを示す。

【００１３】

上記の実施例では、前記現在のサンプルの音節入力シーケンスと、音素入力シーケンス

と、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とする前に、前記方法は、前記現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換するステップと、前記音節のベクトル標記と前記漢字のベクトル標記とを前記音素のベクトル標記と同じ長さに変換して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを取得し、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスと、を前記トレーニング対象モデルのエンコーダの入力とする操作を実行するステップと、をさらに含む。

【００１４】

上記の実施例は、音節のベクトル標記と漢字のベクトル標記を音素のベクトル標記と同じ長さに変換して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを取得できるため、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とする操作を実行することができるという利点又は有益な効果を有する。

【００１５】

上記の実施例では、前記音素入力シーケンスは、声調入力シーケンスと、児化音入力シーケンスと、句読点入力シーケンスと、３５個の独立した韻母の入力シーケンスとを含み、前記音素入力シーケンスは、１０６個の音素単位を含み、各音素単位は、１０６ビットを含む。前記１０６ビットにおいて、有効ビットの値は１であり、非有効ビットの値は０であり、前記漢字入力シーケンスは、３０００個の漢字の入力シーケンスを含み、前記音節入力シーケンスは、５０８個の音節の入力シーケンスを含む。

【００１６】

上記の実施例は、それぞれ音節及び漢字から有効な情報を抽出して発音効果を改善し、特に、同音異字の場合は、発音問題を著しく減少するという利点又は有益な効果を有する。本出願は、製品に高い表現力高い自然度の中国語合成音声を提供することができ、ユーザのヒューマン-コンピュータ・インタラクション体験を効果的に向上させることができ、ユーザの粘着性を向上させ、バイドゥＡＰＰ、スマートスピーカ、及びマップナビゲーションシステムのプロモーションに有利である。

【００１７】

第２の態様では、本出願は、音声合成モデルのトレーニング装置をさらに提供する。前記装置は、入力モジュールと、融合モジュールと、出力モジュールと、を含む。

前記入力モジュールは、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

前記融合モジュールは、符号化標記された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを融合して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの一つの加重組み合わせを取得し、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得する。

前記出力モジュールは、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を前記トレーニング対象モデルのデコーダの入力とし、前記デコーダの出力端で前記現在のサンプルの音声メルスペクトル出力を取得する。

【００１８】

上記の実施例では、前記入力モジュールは、具体的には、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを一つの共有のエンコーダに入力するステップと、前記共有のエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得するステップと、を含

む。

【００１９】

上記の実施例では、前記入力モジュールは、具体的には、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを、三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをそれぞれ取得し、畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とし、前記シーケンス変換ニューラルネットワークモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

【００２０】

上記の実施例では、前記入力モジュールは、具体的には、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力し、各独立したエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

【００２１】

上記の実施例では、前記装置は、前記現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換し、前記音節のベクトル標記と前記漢字のベクトル標記とを前記音素のベクトル標記と同じ長さに変換して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを取得するための変換モジュールをさらに含み、前記入力モジュールは、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを前記トレーニング対象モデルのエンコーダの入力とする操作を実行する。

【００２２】

上記の実施例では、前記音素入力シーケンスは、声調入力シーケンスと、児化音入力シーケンスと、句読点入力シーケンスと、３５個の独立した韻母の入力シーケンスとを含み、前記音素入力シーケンスは、１０６個の音素単位を含む。各音素単位は、１０６ビットを含み、前記１０６ビットにおいて、有効ビットの値は１であり、非有効ビットの値は０であり、前記漢字入力シーケンスは、３０００個の漢字の入力シーケンスを含み、前記音節入力シーケンスは、５０８個の音節の入力シーケンスを含む。

【００２３】

第３の態様では、本出願の実施例は、電子機器を提供する。電子機器は、一つ又は複数のプロセッサと、一つ又は複数のプログラムを記憶するためのメモリと、を含み、前記一つ又は複数のプログラムが前記一つ又は複数のプロセッサによって実行される場合、前記一つ又は複数のプロセッサが、本出願の任意の実施例に記載の音声合成モデルのトレーニング方法を実現する。

【００２４】

第４の態様では、本出願の実施例は、コンピュータ命令が記憶されている記憶媒体を提供する。記憶媒体は、当該プログラムがプロセッサによって実行される場合に、本出願の任意の実施例に記載の音声合成モデルのトレーニング方法が実現される。

【００２５】

上記の出願中の一つの実施例は、以下のような利点又は有益な効果を有する。本出願で提供される音声合成モデルのトレーニング方法、装置、電子機器及び記憶媒体は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得し、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得し、再び音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケン

スとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得し、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得する。つまり、本出願は、トレーニング対象モデルの入力端の入力テキスト及びトレーニング対象モデルの出力端の出力音声により、トレーニング対象モデルのエンコーダ及びデコーダの共同トレーニングを実現する。本出願では、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合する技術的手段を採用するため、従来技術では音節入力シーケンス又は音素入力シーケンス又は漢字入力シーケンスのみを採用するために音声合成効果が良くないという技術問題を克服し、発音効果を効果的に改善し、音声製品に高い表現力と高い自然度の中国語合成音声の技術的効果を提供することができる。また、本出願の実施例の技術案は、簡単で便利に実現され、普及しやすいので、適用範囲がより広くなる。

【００２６】

上記の選択可能な方式が有する他の効果は、以下で具体的な実施例を組み合わせて説明される。

【図面の簡単な説明】

【００２７】

図面は、本技術案をよりよく理解するために使用され、本出願の構成を限定するものではない。

【図１】本出願の実施例１により提供される音声合成モデルのトレーニング方法の概略フローチャートである。

【図２】本出願の実施例２により提供される音声合成モデルのトレーニング方法の概略フローチャートである。

【図３】本出願の実施例２により提供されるＴａｃｏｔｒｏｎモデルの概略構成図である。

【図４】本出願の実施例３により提供される音声合成モデルのトレーニング装置の概略構成図である。

【図５】本出願の実施例の音声合成モデルのトレーニング方法を実現するための電子機器のブロック図である。

【発明を実施するための形態】

【００２８】

以下、図面を組み合わせて本出願の例示的な実施例を説明し、理解を容易にするために本出願の実施例の様々な詳細を含んでいるが、それらは単なる例示であると見なすべきである。したがって、当業者は、本出願の範囲及び精神から逸脱することなく、ここで説明される実施例に対して様々な変更と修正を行うことができることを認識されたい。同様に、明確及び簡潔にするために、以下の説明では、周知の機能及び構造の説明を省略する。

【００２９】

実施例１

図１は、本出願の実施例１により提供される音声合成モデルのトレーニング方法の概略フローチャートである。当該方法は、音声合成モデルのトレーニング装置又は電子機器により実行することができ、当該装置又は電子機器は、ソフトウェア及び／又はハードウェアの方式によって実現することができ、当該装置又は電子機器は、任意のネットワーク通信機能を有するスマートデバイスに集積することができる。図１に示すように、音声合成モデルのトレーニング方法は、以下のステップを含むことができる。

Ｓ１０１：現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得する。

【００３０】
本出願の具体的な実施例では、電子機器は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができる。具体的には、電子機器は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを一つの共有のエンコーダに入力し、共有のエンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができる。好ましくは、電子機器は、さらに、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを入力到三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをそれぞれ取得し、畳み込み変換された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とし、シーケンス変換ニューラルネットワークモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができる。ここのシーケンス変換ニューラルネットワークは、ＲＮＮ、ＬＳＴＭ、ＧＲＵ、Ｔｒａｎｓｆｏｒｍｅｒを含むが、これらに限定されない。好ましくは、電子機器は、さらに、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力するステップと、各独立したエンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得するステップを含むことができる。

【００３１】
Ｓ１０２：符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得する。

【００３２】
本出願の具体的な実施例では、電子機器は、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得することができる。例えば、電子機器は、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを線形的に重ね合わせて、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとのある一つの加重組み合わせを取得することができる。

【００３３】
Ｓ１０３：音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得する。

【００３４】
本出願の具体的な実施例では、電子機器は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得することができる。加重平均値は、各値を対応する重みに乗算し、次に合計を加算して全体の値を取得し、次に全体の単位数で除算することであり、加重平均値の大きさは、全体の各単位の値（変数値）の大きさだけでなく、各値が出現する回数（頻度）にも依存する。

【００３５】
Ｓ１０４：音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得する。

【００３６】
本出願の具体的な実施例では、電子機器は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得することができる。
【００３７】
本出願の具体的な実施例では、Ｔａｃｏｔｒｏｎモデルは、古典的なエンコーダ・デコーダ構造に基づく。エンコーダは、入力された要素シーケンス（英語では、通常アルファベット及び句読点など）に基づき、畳み込み及びシーケンス変換ニューラルネットワークの変換を経て、各入力要素の符号化標記を取得する。デコーダは、前のフレームの出力メルスペクトルを入力とし、アテンションメカニズムを利用してエンコーダ出力の一つの加重組み合わせ表現を取得し、次に、ＬＳＴＭなどの変換を経て、二つの出力を生成する。一つは、現在のフレーム出力のメルスペクトルであり、もう一つは、終了するか否かを判断する停止確率である。停止確率が５０％より大きい場合には合成が終了し、そうでなければ現在の出力を次のフレームの入力とし、この自己回帰プロセスを継続する。このモデルでは、エンコーダは、各入力要素の符号化を担当し、デコーダは、符号化に基づいて現在合成された音を決定するとともに、ＬＳＴＭの記憶機能を利用して秩序的に生成する。当該モデルは、典型的な一対多のマッピングモデルであり、同じ内容は、異なるリズム、異なる感情の音声に対応する。トレーニングセットにおける異なる音声（出力）が同じ文字（入力）に対応する場合、モデルが最終的に学習した発音は、統計的な平均効果を反映する。本出願は、このような一対多のマッピング関係を減少するために、モデルが異なる文脈で適切な発音方式で合成することを学習できるようにする。音素シーケンスが最高のカバー能力を有することを考慮すると、セット以外の発音要素が発生する問題はなく、１０６個の音素単位をモデル入力の基礎要素として選択することができ、各要素は、十分なデータを取得して十分にトレーニングすることができる。入力はｏｎｅ－ｈｏｔの形式であり、ｅｍｂｅｄｄｉｎｇ層を経て固定次元の稠密なベクトル標記に変換され、声調、児化音、及び句読点などの特徴は、同様にｅｍｂｅｄｄｉｎｇを経て同じ次元のベクトルに変換されて、音素ベクトルと加算されてニューラルネットワークに送られる。韻母が独立した場合の発音特性をよりよく学習するために、本出願は、音素における３５個の独立した韻母を単独でモデリングし、声母後に出現された韻母とは、二つの異なる要素と見なされる。実験により、このようなモデリング戦略は、独立した韻母発音が不明確であるという問題をうまく解決する。さらに、同音異字の発音特性を区分するために、本出願は、音節及び漢字を補助情報としてネットワークに入力し、補助モデルが異なる字の発音特性を区分できるようにサポートする。そのうち、無調音節の数は５０８であり、漢字要素は、トレーニングセットにおける高頻度３０００漢字及び５０８音節の計３５０８個の要素を選択し、ある漢字が高頻度３０００字に属していない場合には、対応する音節要素に縮め、より高いカバー率を確保する。
【００３８】
本出願の実施例により提供される音声合成モデルのトレーニング方法は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得し、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得し、再び音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得し、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出

力を取得する。つまり、本出願は、トレーニング対象モデルの入力端の入力テキスト及びトレーニング対象モデルの出力端の出力音声によって、トレーニング対象モデルのエンコーダ及びデコーダの共同トレーニングを実現する。本出願では、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合する技術的手段を採用するため、従来技術では、音節入力シーケンス又は音素入力シーケンス又は漢字入力シーケンスのみが採用されるために音声合成効果が良くないという技術問題を克服し、発音効果を効果的に改善し、音声製品に高い表現力と高い自然度の中国語合成音声の技術的効果を提供する。また、本出願の実施例の技術案は、簡単で便利に実現されるので、普及しやすく、適用範囲がより広くなる。

【００３９】

実施例２

図２は、本出願の実施例２により提供される音声合成モデルのトレーニング方法の概略フローチャートである。図２に示すように、音声合成モデルのトレーニング方法は、以下のようなステップを含む。

Ｓ２０１：現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換する。

【００４０】

本出願の具体的な実施例では、電子機器は、現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換することができる。具体的には、電子機器は、現在のサンプルにおける音素を第１の長さのベクトル標記に変換することができ、現在のサンプルにおける音節及び漢字を第２の長さのベクトル標記に変換することができる。第１の長さが第２の長さより大きい。

【００４１】

Ｓ２０２：音節のベクトル標記と漢字のベクトル標記を音素のベクトル標記と同じ長さに変換して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを取得する。

【００４２】

本出願の具体的な実施例では、電子機器は、音節のベクトル標記と漢字のベクトル標記を音素のベクトル標記と同じ長さに変換して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを取得することができる。具体的には、電子機器は、第１の長さの音素のベクトル標記を音素入力シーケンスとし、音節のベクトル標記と漢字のベクトル標記を第２の長さから第１の長さに変換し、変換された音節のベクトル標記及び漢字のベクトル標記を音節入力シーケンス及び漢字入力シーケンスとすることができる。

【００４３】

Ｓ２０３：現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得する。

【００４４】

本出願の具体的な実施例では、電子機器は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができる。具体的には、電子機器は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを一つの共有のエンコーダに入力し、共有のエンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができる。好ましくは、電子機器は、さらに、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを、三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをそれぞれ取得し、畳み込み変換された音節入力シーケンスと、音

素入力シーケンスと、漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とし、シーケンス変換ニューラルネットワークモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得することができる。ここのシーケンス変換ニューラルネットワークは、ＲＮＮ、ＬＳＴＭ、ＧＲＵ、Ｔｒａｎｓｆｏｒｍｅｒを含むが、これらに限定されない。好ましくは、電子機器は、さらに、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力するステップと、各独立したエンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得するステップとを含むことができる。

【００４５】

Ｓ２０４：符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得する。

【００４６】

本出願の具体的な実施例では、電子機器は、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得することができる。例えば、電子機器は、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを線形的に重ね合わせて、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得することができる。

【００４７】

Ｓ２０５：音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得する。

【００４８】

本出願の具体的な実施例では、電子機器は、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得することができる。

【００４９】

Ｓ２０６：音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得する。

【００５０】

図３は、本出願の実施例２により提供されるＴａｃｏｔｒｏｎモデルの概略構成図である。図３に示すように、Ｔａｃｏｔｒｏｎモデルは、古典的なエンコーダ・デコーダ構造に基づいている。現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得し、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得し、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得し、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得する。

【００５１】

本出願の具体的な実施例では、音素入力シーケンスは、声調入力シーケンス、児化音入力シーケンスと、句読点入力シーケンスと、３５個の独立した韻母の入力シーケンスとを含むことができる。音素入力シーケンスは、１０６個の音素単位を含み、各音素単位は、１０６ビットをむ。１０６ビットにおいて、有効ビットの値は１であり、非有効ビットの値は０である。漢字入力シーケンスは、３０００個の漢字の入力シーケンスを含み、音節入力シーケンスは、５０８個の音節の入力シーケンスを含む。

【００５２】

実験により、音素の独立した韻母と音素、音節、漢字を融合する三つの技術案は、エンドツーエンドの中国語音声合成の問題をある程度で解決することができる。そのうち、独立エンコーダの効果が最も優れている。測定の結果は、発音問題の割合が２％から０．４％に減少することを示す。詳細の分析結果は、音素が発音のタイプを基本的に決定するが、場合によって、音節を変更すると、発音に一定の影響があり、漢字を変更すると発音方式にのみ影響する。これらの結果から、モデルは、それぞれ音節及び漢字から有効な情報を抽出して発音効果を改善し、特に、同音異字の場合は、発音問題を著しく減少することを証明する。本出願は、製品に高い表現力高い自然度の中国語合成音声を提供することができ、ユーザのヒューマン-コンピュータ・インタラクション体験を効果的に向上させることができるので、ユーザの粘着性を向上させ、バイドゥＡＰＰ、スマートスピーカ、及びマップナビゲーションシステムのプロモーションに有利である。

【００５３】

本出願の実施例により提供される音声合成モデルのトレーニング方法は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得し、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得し、再び音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得し、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得する。つまり、本出願は、トレーニング対象モデルの入力端の入力テキスト及びトレーニング対象モデルの出力端の出力音声によって、トレーニング対象モデルのエンコーダ及びデコーダの共同トレーニングを実現する。本出願では、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合する技術的手段を採用したため、従来技術では、音節入力シーケンス又は音素入力シーケンス又は漢字入力シーケンスのみを採用したことで音声合成効果が良くないという技術問題を克服し、発音効果を効果的に改善し、音声製品に高い表現力と高い自然度の中国語合成音声の技術的効果を提供することができる。また、本出願の実施例の技術案は、簡単で便利に実現され、普及しやすく、適用範囲がより広くなる。

【００５４】

実施例３

図４は、本出願の実施例３により提供される音声合成モデルのトレーニング装置の概略構成図である。図４に示すように、前記装置４００は、入力モジュール４０１と、融合モジュール４０２と、出力モジュール４０３と、を含む。

前記入力モジュール４０１は、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、前記エンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

前記融合モジュール４０２とは、符号化標記された前記音節入力シーケンスと、前記音

素入力シーケンスと、前記漢字入力シーケンスとを融合して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの一つの加重組み合わせを取得し、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得する。

前記出力モジュール４０３は、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの加重組み合わせの各時点における加重平均を前記トレーニング対象モデルのデコーダの入力とし、前記デコーダの出力端で前記現在のサンプルの音声メルスペクトル出力を取得する。

【００５５】

さらに、前記入力モジュール４０１は、具体的には、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを一つの共有のエンコーダに入力し、前記共有のエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

【００５６】

さらに、前記入力モジュール４０１は、具体的には、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを、三つの独立した畳み込み層変換モジュールにそれぞれ入力し、各独立した畳み込み層変換モジュールの出力端で畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをそれぞれ取得し、畳み込み変換された前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとをシーケンス変換ニューラルネットワークモジュールの入力とし、前記シーケンス変換ニューラルネットワークモジュールの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

【００５７】

さらに、前記入力モジュール４０１は、具体的には、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを三つの独立したエンコーダにそれぞれ入力し、各独立したエンコーダの出力端で前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとの符号化標記を取得する。

【００５８】

さらに、前記装置は、前記現在のサンプルにおける音素と、音節と、漢字とを個々の固定次元のベクトル標記にそれぞれ変換し、前記音節のベクトル標記と前記漢字のベクトル標記とを前記音素のベクトル標記と同じ長さに変換して、前記音節入力シーケンスと、前記音素入力シーケンスと、前記漢字入力シーケンスとを取得するための変換モジュール４０４（図示せず）をさらに含む。

前記入力モジュール４０１は、前記音節入力シーケンスと、前記音素入力シーケンスと、記漢字入力シーケンスとを前記トレーニング対象モデルのエンコーダの入力とする操作を実行する。

【００５９】

さらに、前記音素入力シーケンスは、声調入力シーケンスと、児化音入力シーケンスと、句読点入力シーケンスと、３５個の独立した韻母の入力シーケンスとを含む。前記音素入力シーケンスは、１０６個の音素単位を含み、各音素単位は、１０６ビットを含む。前記１０６ビットにおいて、有効ビットの値は１であり、非有効ビットの値は０である。前記漢字入力シーケンスは、３０００個の漢字の入力シーケンスを含み、前記音節入力シーケンスは、５０８個の音節の入力シーケンスを含む。

【００６０】

上記の音声合成モデルのトレーニング装置は、本発明の任意の実施例により提供される方法を実行することができ、実行方法に対応する機能モジュール及び有益な効果を有する。本実施例で詳細に説明されていない技術的詳細は、本発明の任意の実施例により提供さ

れる音声合成モデルのトレーニング方法を参照することができる。

【００６１】

実施例４

本出願の実施例によれば、本出願は、電子機器及び読み取り可能な記憶媒体をさらに提供する。

【００６２】

図５には、本出願の実施例に係る音声合成モデルのトレーニング方法の電子機器のブロック図が示されている。電子機器は、ラップトップコンピュータ、デスクトップコンピュータ、ワークステーション、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、大型コンピュータ、及び他の適切なコンピュータなどの様々な形式のデジタルコンピュータを表すことを目的とする。電子機器は、パーソナルデジタル処理、携帯電話、スマートフォン、ウェアラブルデバイス、他の同様のコンピューティングデバイスなどの様々な形式のモバイルデバイスを表すこともできる。本明細書で示されるコンポーネント、それらの接続と関係、及びそれらの機能は単なる例であり、本明細書の説明及び／又は要求される本出願の実現を制限することを意図しない。

【００６３】

図５に示すように、当該電子機器は、一つ又は複数のプロセッサ５０１と、メモリ５０２と、高速インターフェースと低速インターフェースを含む各コンポーネントを接続するためのインターフェースと、を含む。各コンポーネントは、異なるバスで相互に接続され、共通のマザーボードに取り付けられるか、又は必要に応じて他の方式で取り付けることができる。プロセッサは、外部入力／出力装置（インターフェースに結合されたディスプレイデバイスなど）にＧＵＩの図形情報をディスプレイするためにメモリに記憶されている命令を含む、電子機器内に実行される命令を処理することができる。他の実施方式では、必要であれば、複数のプロセッサ及び／又は複数のバスを、複数のメモリと複数のメモリとともに使用することができる。同様に、複数の電子機器を接続することができ、各機器は、部分的な必要な操作（例えば、サーバアレイ、ブレードサーバ、又はマルチプロセッサシステムとする）を提供することができる。図５では、一つのプロセッサ５０１を例とする。

【００６４】

メモリ５０２は、本出願により提供される非一時的なコンピュータ読み取り可能な記憶媒体である。その中、前記メモリには、少なくとも一つのプロセッサによって実行される命令を記憶して、前記少なくとも一つのプロセッサが本出願により提供される音声合成モデルのトレーニング方法を実行することができるようにする。本出願の非一時的なコンピュータ読み取り可能な記憶媒体は、コンピュータが本出願により提供される音声合成モデルのトレーニング方法を実行するためのコンピュータ命令を記憶する。

【００６５】

メモリ５０２は、非一時的なコンピュータ読み取り可能な記憶媒体として、本出願の実施例における音声合成モデルのトレーニング方法に対応するプログラム命令／モジュール（例えば、図４に示す入力モジュール４０１、融合モジュール４０２、出力モジュール４０３）ように、非一時的なソフトウェアプログラム、非一時的なコンピュータ実行可能なプログラム及びモジュールを記憶するために用いられる。プロセッサ５０１は、メモリ５０２に記憶されている非一時的なソフトウェアプログラム、命令及びモジュールを実行することによって、サーバの様々な機能アプリケーション及びデータ処理を実行する。すなわち上記の方法の実施例における音声合成モデルのトレーニング方法を実現する。

【００６６】

メモリ５０２は、ストレージプログラム領域とストレージデータ領域とを含むことができる。ストレージプログラム領域は、オペレーティングシステムや、少なくとも一つの機能に必要なアプリケーションプログラムを記憶することができ、ストレージデータ領域は、音声合成モデルのトレーニング方法に基づく電子機器の使用によって作成されたデータなどを記憶することができる。また、メモリ４０２は、高速ランダム存取メモリを含むこ

とができ、非一時的なメモリをさらに含むことができる。例えば、少なくとも一つのディスクストレージデバイス、フラッシュメモリデバイス、又は他の非一時的なソリッドステートストレージデバイスである。いくつかの実施例では、メモリ５０２は、プロセッサ５０１に対して遠隔に設置されたメモリを含むことができ、これらの遠隔メモリは、ネットワークを介して音声合成モデルのトレーニング方法の電子機器に接続されることができる。上記のネットワークの例は、インターネット、イントラネット、ローカルエリアネットワーク、モバイル通信ネットワーク、及びその組み合わせを含むが、これらに限定しない。

【００６７】

音声合成モデルのトレーニング方法の電子機器は、入力装置５０３と出力装置５０４とをさらに含むことができる。プロセッサ５０１と、メモリ５０２と、入力装置５０３と、出力装置５０４とは、バス又は他の方式を介して接続することができ、図５では、バスを介して接続することを例示している。

【００６８】

入力装置５０３は、入力された数字又は文字情報を受信することができ、及び音声合成モデルのトレーニング方法の電子機器のユーザ設置及び機能制御に関するキー信号入力を生成することができる。例えば、タッチスクリーン、キーパッド、マウス、トラックパッド、タッチパッド、指示杆、一つ又は複数のマウスボタン、トラックボール、ジョイスティックなどの入力装置である。出力装置５０４は、ディスプレイデバイス、補助照明デバイス（例えば、ＬＥＤ）、及び触覚フィードバックデバイス（例えば、振動モータ）などを含むことができる。当該ディスプレイデバイスは、液晶ディスプレイ（ＬＣＤ）、発光ダイオード（ＬＥＤ）ディスプレイ、及びプラズマディスプレイを含むことができるが、これらに限定しない。いくつかの実施方式では、ディスプレイデバイスは、タッチスクリーンであってもよい。

【００６９】

本明細書で説明されるシステムと技術の様々な実施方式は、デジタル電子回路システム、集積回路システム、特定用途向けＡＳＩＣ（特定用途向け集積回路）、コンピュータハードウェア、ファームウェア、ソフトウェア、及び／又はそれらの組み合わせで実現することができる。これらの様々な実施方式は、一つ又は複数のコンピュータプログラムで実施されることを含むことができ、当該一つ又は複数のコンピュータプログラムは、少なくとも一つのプログラマブルプロセッサを含むプログラム可能なシステムで実行及び／又は解釈されることができ、当該プログラマブルプロセッサは、特定用途向け又は汎用プログラマブルプロセッサであってもよく、ストレージシステム、少なくとも一つの入力装置、及び少なくとも一つの出力装置からデータ及び命令を受信し、データ及び命令を当該ストレージシステム、当該少なくとも一つの入力装置、及び当該少なくとも一つの出力装置に伝送することができる。

【００７０】

これらのコンピューティングプログラム（プログラム、ソフトウェア、ソフトウェアアプリケーション、又はコードとも呼ばれる）は、プログラマブルプロセッサの機械命令、高レベルのプロセス及び／又はオブジェクト指向プログラミング言語、及び／又はアセンブリ／機械言語でこれらのコンピューティングプログラムを実施することを含む。本明細書に使用されるように、用語「機械読み取り可能な媒体」及び「コンピュータ読み取り可能な媒体」は、機械命令及び／又はデータをプログラマブルプロセッサに提供するために使用される任意のコンピュータプログラム製品、機器、及び／又は装置（例えば、磁気ディスク、光ディスク、メモリ、プログラマブルロジックデバイス（ＰＬＤ））を指し、機械読み取り可能な信号である機械命令を受信する機械読み取り可能な媒体を含む。用語「機械読み取り可能な信号」は、機械命令及び／又はデータをプログラマブルプロセッサに提供するための任意の信号を指す。

【００７１】

ユーザとのインタラクションを提供するために、コンピュータ上で、ここで説明されて

いるシステム及び技術を実施することができる。当該コンピュータは、ユーザに情報を表示するためのディスプレイ装置（例えば、ＣＲＴ（陰極線管）又はＬＣＤ（液晶ディスプレイ）モニタ）と、キーボード及びポインティングデバイス（例えば、マウス又はトラックボール）とを有し、ユーザは、当該キーボード及び当該ポインティングデバイスによって入力をコンピュータに提供することができる。他の種類の装置は、ユーザとのインタラクションを提供するために用いられることもでき、例えば、ユーザに提供されるフィードバックは、任意の形式のセンシングフィードバック（例えば、視覚フィードバック、聴覚フィードバック、又は触覚フィードバック）であってもよく、任意の形式（音響入力と、音声入力と、触覚入力とを含む）でユーザからの入力を受信することができる。

【００７２】

ここで説明されるシステム及び技術は、バックエンドコンポーネントを含むコンピューティングシステム（例えば、データサーバとする）、又はミドルウェアコンポーネントを含むコンピューティングシステム（例えば、アプリケーションサーバ）、又はフロントエンドコンポーネントを含むコンピューティングシステム（例えば、グラフィカルユーザインタフェース又はウェブブラウザを有するユーザコンピュータ、ユーザは、当該グラフィカルユーザインタフェース又は当該ウェブブラウザによってここで説明されるシステム及び技術の実施方式とインタラクションする）、又はこのようなバックエンドコンポーネントと、ミドルウェアコンポーネントと、フロントエンドコンポーネントの任意の組み合わせを含むコンピューティングシステムで実施することができる。任意の形式又は媒体のデジタルデータ通信（例えば、通信ネットワーク）によってシステムのコンポーネントを相互に接続されることができる。通信ネットワークの例は、ローカルエリアネットワーク（ＬＡＮ）と、ワイドエリアネットワーク（ＷＡＮ）と、インターネットとを含む。

【００７３】

コンピュータシステムは、クライアントとサーバとを含むことができる。クライアントとサーバは、一般に、互いに離れており、通常に通信ネットワークを介してインタラクションする。対応するコンピュータ上で実行され、互いにクライアント-サーバ関係を有するコンピュータプログラムによってクライアントとサーバとの関係が生成される。

【００７４】

本出願の実施例の技術案によれば、現在のサンプルの音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとをトレーニング対象モデルのエンコーダの入力とし、エンコーダの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの符号化標記を取得し、符号化標記された音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合して、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの一つの加重組み合わせを取得し、再び音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせをアテンションモジュールの入力とし、アテンションモジュールの出力端で音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均を取得し、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとの加重組み合わせの各時点における加重平均をトレーニング対象モデルのデコーダの入力とし、デコーダの出力端で現在のサンプルの音声メルスペクトル出力を取得する。つまり、本出願は、トレーニング対象モデルの入力端の入力テキスト及びトレーニング対象モデルの出力端の出力音声によって、トレーニング対象モデルのエンコーダ及びデコーダの共同トレーニングを実現する。本出願では、音節入力シーケンスと、音素入力シーケンスと、漢字入力シーケンスとを融合する技術的手段を採用するため、従来技術では音節入力シーケンス又は音素入力シーケンス又は漢字入力シーケンスのみを採用したために音声合成効果が良くないという技術問題を克服し、発音効果を効果的に改善し、音声製品に高い表現力と高い自然度の中国語合成音声の技術的効果を提供する。また、本出願の実施例の技術案は、簡単で便利に実現され、普及しやすく、適用範囲がより広くなる。

【００７５】

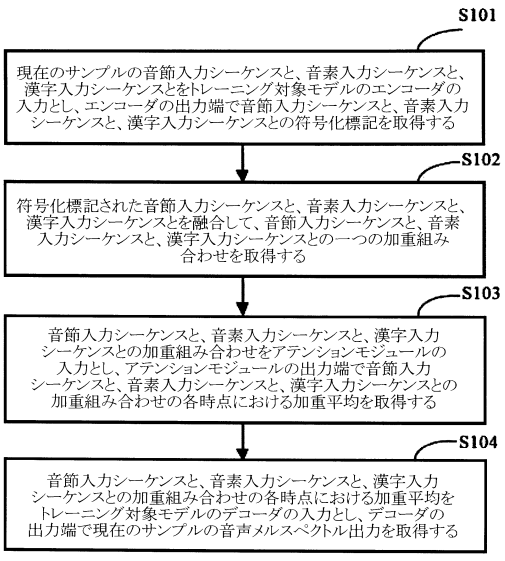
上記に示される様々な形式のフローを使用して、ステップを並べ替え、追加、又は削除

することができることを理解されたい。例えば、本出願に記載されている各ステップは、並列に実行されてもよいし、順次的に実行されてもよいし、異なる順序で実行されてもよく、本出願で開示されている技術案が所望の結果を実現することができれば、本明細書では限定されない。

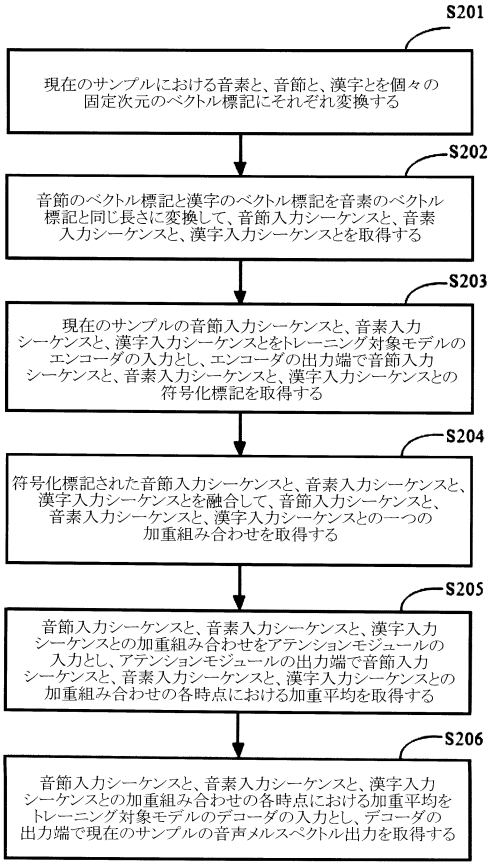
【００７６】

上記の具体的な実施方式は、本出願に対する保護範囲の制限を構成するものではない。当業者は、設計要求と他の要因に応じて、様々な修正、組み合わせ、サブコンビネーション、及び代替を行うことができる。任意の本出願の精神と原則内で行われる修正、同等の置換、及び改善などは、いずれも本出願の保護範囲内に含まれなければならない。

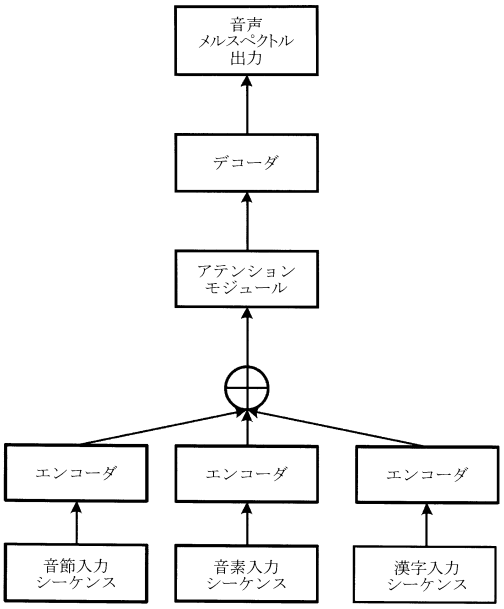
【図１】



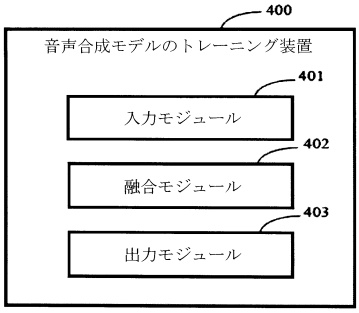
【図２】



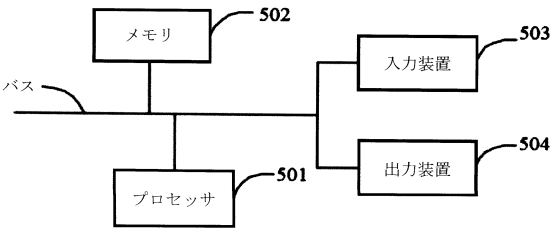
【図３】


音声
メルスペクトル
出力
デコーダ
アテンション
モジュール
エンコーダ
エンコーダ
エンコーダ
音節入力
シーケンス
音素入力
シーケンス
漢字入力
シーケンス

【図４】


400
音声合成モデルのトレーニング装置
401
入力モジュール
402
融合モジュール
403
出力モジュール

【図５】


メモリ
502
バス
入力装置
503
出力装置
504
プロセッサ
501

フロントページの続き

(74)代理人 100201466
弁理士 竹内 邦彦
(72)発明者 チェン, ジーペン
中華人民共和国 １０００８５ ベイジン ハイディエン ディストリクト シャンディ １０ティーエイチ ストリート バイドゥ キャンパス ナンバー １０ ３／フロア
(72)発明者 バイ, ジンフェン
中華人民共和国 １０００８５ ベイジン ハイディエン ディストリクト シャンディ １０ティーエイチ ストリート バイドゥ キャンパス ナンバー １０ ３／フロア
(72)発明者 ジア, レイ
中華人民共和国 １０００８５ ベイジン ハイディエン ディストリクト シャンディ １０ティーエイチ ストリート バイドゥ キャンパス ナンバー １０ ３／フロア

審査官 渡部 幸和

(56)参考文献 LU Chunhui et al., Self-attention Based Prosodic Boundary Prediction for Chinese Speech Synthesis, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019年04月17日
DING Chuang et al., Automatic prosody prediction for Chinese speech synthesis using BLASTM-RNN and embedding features, 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2016年02月11日
YASUDA Yusuke et al., Investigation of Enhanced Tacotron Text-to-Speech Synthesis Systems with Self-attention for Pitch Accent Language, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019年04月17日

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