

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号  
特許第7557085号  
(P7557085)

(45)発行日 令和6年9月26日(2024.9.26)

(24)登録日 令和6年9月17日(2024.9.17)

|                   |                               |          |                         |         |
|-------------------|-------------------------------|----------|-------------------------|---------|
| (51)国際特許分類        |                               | F I      |                         |         |
| G 1 0 L           | 13/08 (2013.01)               | G 1 0 L  | 13/08                   | 1 1 0 A |
| G 1 0 L           | 13/00 (2006.01)               | G 1 0 L  | 13/00                   | 1 0 0 M |
|                   |                               | G 1 0 L  | 13/08                   | 1 2 4   |
| 請求項の数 24 (全27頁)   |                               |          |                         |         |
| (21)出願番号          | 特願2023-553359(P2023-553359)   | (73)特許権者 | 502208397               |         |
| (86)(22)出願日       | 令和4年2月28日(2022.2.28)          |          | グーグル エルエルシー             |         |
| (65)公表番号          | 特表2024-508033(P2024-508033 A) |          | G o o g l e L L C       |         |
| (43)公表日           | 令和6年2月21日(2024.2.21)          |          | アメリカ合衆国 カリフォルニア州 9 4    |         |
| (86)国際出願番号        | PCT/US2022/018219             |          | 0 4 3 マウンテン ビュー アンフィシ   |         |
| (87)国際公開番号        | WO2022/187168                 |          | アター パークウェイ 1 6 0 0      |         |
| (87)国際公開日         | 令和4年9月9日(2022.9.9)            |          | 1 6 0 0 Amphitheatre P  |         |
| 審査請求日             | 令和5年10月30日(2023.10.30)        |          | arkway 9 4 0 4 3 Mounta |         |
| (31)優先権主張番号       | 17/190,456                    | (74)代理人  | 100108453               |         |
| (32)優先日           | 令和3年3月3日(2021.3.3)            |          | 弁理士 村山 靖彦               |         |
| (33)優先権主張国・地域又は機関 | 米国(US)                        | (74)代理人  | 100110364               |         |
| 早期審査対象出願          |                               |          | 弁理士 実広 信哉               |         |
|                   |                               | (74)代理人  | 100133400               |         |
|                   |                               |          | 弁理士 阿部 達彦               |         |
|                   |                               |          | 最終頁に続く                  |         |

(54)【発明の名称】 対話中のテキスト-音声の瞬時学習

(57)【特許請求の範囲】  
【請求項1】

データ処理ハードウェア(610)上で実行されると前記データ処理ハードウェア(610)に動作を行わせるコンピュータ実装方法(500)であって、前記動作が、

ユーザによって話される質問(12)に存在する特定の単語のユーザ発音(202)を受信するステップと、

テキスト-音声(TTS)入力(152)に存在する同じ特定の単語のTTS発音(204)を受信するステップであり、前記TTS入力(152)が前記質問(12)への応答のテキスト表現を備え、前記特定の単語の前記TTS発音(204)が前記特定の単語の前記ユーザ発音(202)とは異なる、ステップと、

前記特定の単語の前記ユーザ発音(202)と関連付けられるユーザ発音関連特徴(210)を得るステップと、

前記特定の単語の前記TTS発音(204)と関連付けられるTTS発音関連特徴(230)を得るステップと、

入力として前記ユーザ発音関連特徴(210)および前記TTS発音関連特徴(230)を受信するように構成される発音決定モデル(200)からの出力として、TTSオーディオ(154)に使用するために最高信頼度と関連付けられる前記特定の単語の前記ユーザ発音(202)または前記特定の単語の前記TTS発音(204)の一方を選択する発音決定(250)を生成するステップと、

前記ユーザと関連付けられるユーザデバイス(110)からの可聴出力のために、前記発音決定モデル(200)から出力される前記発音決定(250)によって選択された前記特定の単語

に対する前記ユーザ発音(202)または前記特定の単語に対する前記TTS発音(204)の前記一方を使用して前記質問(12)への前記応答の合成音声表現を備える前記TTSオーディオ(154)を提供するステップとを含む、コンピュータ実装方法(500)。

【請求項2】

前記動作が、

前記ユーザによって話される前記質問(12)に対応するオーディオデータ(14)を受信するステップと、

自動音声認識器(ASR(140))を使用して、前記オーディオデータ(14)を処理して前記質問(12)の表記(142)を生成するステップとを更に含む、請求項1に記載の方法(500)。

【請求項3】

前記特定の単語の前記ユーザ発音(202)を受信するステップが、  
前記オーディオデータ(14)から前記特定の単語のユーザ音響表現を抽出するステップであり、前記ユーザ音響表現が前記特定の単語の前記ユーザ発音(202)を伝える、ステップか、  
前記オーディオデータ(14)を処理して、前記特定の単語の前記ユーザ発音(202)を伝えるユーザ音素表現を生成するステップかのうちの少なくとも1つを含む、請求項2に記載の方法(500)。

【請求項4】

前記特定の単語の前記ユーザ発音(202)と関連付けられる前記ユーザ発音関連特徴(210)が、前記ASR(140)が前記オーディオデータ(14)に前記特定の単語を認識することと関連付けられる1つまたは複数の信頼度特徴を含む、請求項2または3に記載の方法(500)。

【請求項5】

前記特定の単語の前記ユーザ発音(202)と関連付けられる前記ユーザ発音関連特徴(210)が、

前記質問(12)が前記ユーザによって話されたときの前記ユーザの地理的地域か、

前記ユーザと関連付けられる言語人口学的情報か、

前記ユーザおよび/または他のユーザによって話された以前の質問(12)内の前記特定の単語を発音するときに前記ユーザ発音(202)を使用する頻度かの少なくとも1つを含む、請求項1から4のいずれか一項に記載の方法(500)。

【請求項6】

前記特定の単語の前記TTS発音(204)を受信するステップが、

TTSシステム(150)への入力として、前記質問(12)への前記応答のテキスト表現を備える前記TTS入力(152)を受信するステップと、

前記TTSシステム(150)からの出力として、前記質問(12)への前記応答の初期合成音声表現を備えるTTSオーディオ(154)の初期サンプルを生成するステップと、

前記TTSオーディオ(154)の前記初期サンプルから前記特定の単語のTTS音響表現を抽出するステップであり、前記TTS音響表現が前記特定の単語の前記TTS発音(204)を伝える、ステップとを含む、請求項1から5のいずれか一項に記載の方法(500)。

【請求項7】

前記特定の単語の前記TTS発音(204)を受信するステップが、前記質問(12)への前記応答の前記テキスト表現を処理して、前記特定の単語の前記TTS発音(204)を伝えるTTS音素表現を生成するステップを含む、請求項1から6のいずれか一項に記載の方法(500)。

【請求項8】

前記特定の単語の前記TTS発音(204)と関連付けられる前記TTS発音関連特徴(230)が、前記特定の単語に対する確認済みの好ましい発音か、

1つまたは複数の補助情報源からの発音マイニングを使用して推定される前記特定の単語に対する未確認の発音か、

前記特定の単語を発音するためのいくらかでも他の異形が存在するかどうかを示す発音異形特徴か、

前記特定の単語のユーザ誤発音の可能性を示す発音複雑性特徴かの少なくとも1つを含む、請求項1から7のいずれか一項に記載の方法(500)。

10

20

30

40

50

## 【請求項 9】

前記動作が、前記特定の単語の前記ユーザ発音(202)または前記特定の単語の前記TTS発音(204)の前記一方を選択する前記発音決定(250)を生成した後に、

以降のTTS出力(204)内の前記特定の単語を発音するために前記ユーザが前記特定の単語の前記ユーザ発音(202)または前記特定の単語の前記TTS発音(204)のどちらの方を好むかを示す前記ユーザからの明示的なフィードバックを受信するステップと、

前記ユーザからの前記明示的なフィードバックに基づいて前記発音決定モデル(200)を更新するステップとを更に含む、請求項1から8のいずれか一項に記載の方法(500)。

## 【請求項 10】

前記動作が、前記ユーザからの前記明示的なフィードバックが前記特定の単語の前記ユーザ発音(202)を前記ユーザが好むことを示すとき、前記特定の単語を含むTTSオーディオ(154)を生成するときに前記特定の単語の前記ユーザ発音(202)を使用するようにTTSシステム(150)を更新するステップを更に含む、請求項9に記載の方法(500)。

## 【請求項 11】

前記動作が、前記ユーザデバイス(110)からの可聴出力のために前記TTSオーディオ(154)を提供した後に、

前記特定の単語を含む前記ユーザまたは別のユーザによって話される次の質問(12)に対応するオーディオデータ(14)を受信するステップと、

前記発音決定(250)によって選択された前記特定の単語に対する前記ユーザ発音(202)または前記特定の単語に対する前記TTS発音(204)の前記一方と同じように前記ユーザまたは前記別のユーザが前記次の質問(12)内の前記特定の単語を発音したか否かを示す黙示的なユーザフィードバックを判定するステップと、

前記黙示的なユーザフィードバックに基づいて前記発音決定モデル(200)を更新するステップとを更に含む、請求項1から10のいずれか一項に記載の方法(500)。

## 【請求項 12】

前記動作が、前記黙示的なフィードバックが前記発音決定(250)によって選択された前記特定の単語に対する前記ユーザ発音(202)または前記特定の単語に対する前記TTS発音(204)の前記一方と同じように前記ユーザが前記次の質問(12)内の前記特定の単語を発音したことを示すとき、前記黙示的なユーザフィードバックに基づいてTTSシステム(150)を更新するステップを更に含む、請求項11に記載の方法(500)。

## 【請求項 13】

データ処理ハードウェア(610)と、

前記データ処理ハードウェア(610)と通信しているメモリハードウェア(620)とを備えるシステム(600)であって、前記メモリハードウェア(620)が、前記データ処理ハードウェア(610)上で実行されると前記データ処理ハードウェア(610)に動作を行わせる命令を記憶しており、前記動作が、

ユーザによって話される質問(12)に存在する特定の単語のユーザ発音(202)を受信することと、

テキスト-音声(TTS)入力(152)に存在する同じ特定の単語のTTS発音を受信することであり、前記TTS入力(152)が前記質問(12)への応答のテキスト表現を備え、前記特定の単語の前記TTS発音(204)が前記特定の単語の前記ユーザ発音(202)とは異なる、ことと、

前記特定の単語の前記ユーザ発音(202)と関連付けられるユーザ発音関連特徴(210)を得ることと、

前記特定の単語の前記TTS発音(204)と関連付けられるTTS発音関連特徴(230)を得ることと、

入力として前記ユーザ発音関連特徴(210)および前記TTS発音関連特徴(230)を受信するように構成される発音決定モデル(200)からの出力として、TTSオーディオ(154)に使用するために最高信頼度と関連付けられる前記特定の単語の前記ユーザ発音(202)または前記特定の単語の前記TTS発音(204)の一方を選択する発音決定(250)を生成することと、

前記ユーザと関連付けられるユーザデバイス(110)からの可聴出力のために、前記発音

10

20

30

40

50

決定モデル(200)から出力される前記発音決定(250)によって選択された前記特定の単語に対する前記ユーザ発音(202)または前記特定の単語に対する前記TTS発音(204)の前記一方を使用して前記質問(12)への前記応答の合成音声表現を備える前記TTSオーディオ(154)を提供することを含む、システム(600)。

【請求項14】

前記動作が、

前記ユーザによって話される前記質問(12)に対応するオーディオデータ(14)を受信することと、

自動音声認識器(ASR)(140)を使用して、前記オーディオデータ(14)を処理して前記質問(12)の表記(142)を生成することとを更に含む、請求項13に記載のシステム(600)。

10

【請求項15】

前記特定の単語の前記ユーザ発音(202)を受信することが、  
前記オーディオデータ(14)から前記特定の単語のユーザ音響表現を抽出することであり、  
前記ユーザ音響表現が前記特定の単語の前記ユーザ発音(202)を伝える、ことか、

前記オーディオデータ(14)を処理して、前記特定の単語の前記ユーザ発音(202)を伝えるユーザ音素表現を生成することかのうちの少なくとも1つを含む、請求項14に記載のシステム(600)。

【請求項16】

前記特定の単語の前記ユーザ発音(202)と関連付けられる前記ユーザ発音関連特徴(210)が、前記ASR(140)が前記オーディオデータ(14)に前記特定の単語を認識することと関連付けられる1つまたは複数の信頼度特徴を含む、請求項14または15に記載のシステム(600)。

20

【請求項17】

前記特定の単語の前記ユーザ発音(202)と関連付けられる前記ユーザ発音関連特徴(210)が、

前記質問(12)が前記ユーザによって話されたときの前記ユーザの地理的地域か、

前記ユーザと関連付けられる言語人口学的情報か、

前記ユーザおよび/または他のユーザによって話された以前の質問(12)内の前記特定の単語を発音するときに前記ユーザ発音(202)を使用する頻度かの少なくとも1つを含む、請求項13から16のいずれか一項に記載のシステム(600)。

30

【請求項18】

前記特定の単語の前記TTS発音(204)を受信することが、

TTSシステム(150)への入力として、前記質問(12)への前記応答のテキスト表現を備える前記TTS入力(152)を受信することと、

前記TTSシステム(150)からの出力として、前記質問(12)への前記応答の初期合成音声表現を備えるTTSオーディオ(154)の初期サンプルを生成することと、

前記TTSオーディオ(154)の前記初期サンプルから前記特定の単語のTTS音響表現を抽出することであり、前記TTS音響表現が前記特定の単語の前記TTS発音(204)を伝える、こととを含む、請求項13から17のいずれか一項に記載のシステム(600)。

【請求項19】

40

前記特定の単語の前記TTS発音(204)を受信することが、前記質問(12)への前記応答の前記テキスト表現を処理して、前記特定の単語の前記TTS発音(204)を伝えるTTS音素表現を生成することを含む、請求項13から18のいずれか一項に記載のシステム(600)。

【請求項20】

前記特定の単語の前記TTS発音(204)と関連付けられる前記TTS発音関連特徴(230)が、  
前記特定の単語に対する確認済みの好ましい発音か、

1つまたは複数の補助情報源からの発音マイニングを使用して推定される前記特定の単語に対する未確認の発音か、

前記特定の単語を発音するためのいくらかでも他の異形が存在するかどうかを示す発音異形特徴か、

50

前記特定の単語のユーザ誤発音の可能性を示す発音複雑性特徴かの少なくとも1つを含む、請求項13から19のいずれか一項に記載のシステム(600)。

【請求項21】

前記動作が、前記特定の単語の前記ユーザ発音(202)または前記特定の単語の前記TTS発音(204)の前記一方を選択する前記発音決定(250)を生成した後に、

以降のTTS出力(204)内の前記特定の単語を発音するために前記ユーザが前記特定の単語の前記ユーザ発音(202)または前記特定の単語の前記TTS発音(204)のどちらの方を好むかを示す前記ユーザからの明示的なフィードバックを受信することと、

前記ユーザからの前記明示的なフィードバックに基づいて前記発音決定モデル(200)を更新することとを更に含む、請求項13から20のいずれか一項に記載のシステム(600)。

10

【請求項22】

前記動作が、前記ユーザからの前記明示的なフィードバックが前記特定の単語の前記ユーザ発音(202)を前記ユーザが好むことを示すとき、前記特定の単語を含むTTSオーディオ(154)を生成するときに前記特定の単語の前記ユーザ発音(202)を使用するようにTTSシステム(150)を更新することを更に含む、請求項21に記載のシステム(600)。

【請求項23】

前記動作が、前記ユーザデバイス(110)からの可聴出力のために前記TTSオーディオ(154)を提供した後に、

前記特定の単語を含む前記ユーザまたは別のユーザによって話される次の質問(12)に対応するオーディオデータ(14)を受信することと、

20

前記発音決定(250)によって選択された前記特定の単語に対する前記ユーザ発音(202)または前記特定の単語に対する前記TTS発音(204)の前記一方と同じように前記ユーザまたは前記別のユーザが前記次の質問(12)内の前記特定の単語を発音したか否かを示す黙示的なユーザフィードバックを判定することと、

前記黙示的なユーザフィードバックに基づいて前記発音決定モデル(200)を更新することとを更に含む、請求項13から22のいずれか一項に記載のシステム(600)。

【請求項24】

前記動作が、前記黙示的なフィードバックが前記発音決定(250)によって選択された前記特定の単語に対する前記ユーザ発音(202)または前記特定の単語に対する前記TTS発音(204)の前記一方と同じように前記ユーザが前記次の質問(12)内の前記特定の単語を発音したことを示すとき、前記黙示的なユーザフィードバックに基づいてTTSシステム(150)を更新することを更に含む、請求項23に記載のシステム(600)。

30

【発明の詳細な説明】

【技術分野】

【0001】

本開示は、対話中のテキスト-音声の瞬時学習に関する。

【背景技術】

【0002】

ユーザは、デジタルアシスタントを通して、スマートフォン、スマートウォッチおよびスマートスピーカなどの、音声対応デバイスと度々相互作用する。これらのデジタルアシスタントは、ユーザとの対話を提供して、ユーザが全て自然な会話的相互作用を通じてタスクを完了し、自分が抱く疑問への回答を得ることを可能にする。理想的には、ユーザとデジタルアシスタントとの間の対話の間、ユーザは、デジタルアシスタントを走らせている自分の音声対応デバイスに向けられる口頭の質問を介して、あたかもユーザが別の人に話しているかのように、やりとりをすることができべきである。デジタルアシスタントは、これらの口頭の質問を自動音声認識器(ASR)システムに提供して、行為を行うことができるように口頭の要求を処理および認識するであろう。追加的に、デジタルアシスタントは、テキスト-音声(TTS)システムも利用して、質問への応答のテキスト表現をユーザの音声対応デバイスからの可聴出力のために合成音声へ変換するであろう。しばしば、デジタルアシスタント対話中に口頭の質問と対応するTTS応答との間に語彙の重なりがあり、

40

50

それによって口頭の質問内の単語のユーザ発音が、合成音声として可聴に出力されるときに質問へのデジタルアシスタント応答に存在する同じ単語のTTS発音とは異なる。

【先行技術文献】

【非特許文献】

【0003】

【文献】「Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions」、J. Shenら著、例えば<https://arxiv.org/abs/1712.05884>において入手可能

【発明の概要】

【課題を解決するための手段】

【0004】

本開示の一態様は、データ処理ハードウェア上で実行されるとデータ処理ハードウェアに、特定の単語のユーザ発音または特定の単語のテキスト-音声発音のどちらの方がテキスト-音声オーディオに使用するためにより信頼できるかを選択するための動作を行わせるコンピュータ実装方法を提供する。動作は、ユーザによって話される質問に存在する特定の単語のユーザ発音を受信するステップを含む。動作は、テキスト-音声(TTS)入力に存在する同じ特定の単語のTTS発音を受信するステップも含む。TTS入力は、質問への応答のテキスト表現を含み、そして特定の単語のTTS発音は、特定の単語のユーザ発音とは異なる。動作は、特定の単語のユーザ発音と関連付けられるユーザ発音関連特徴を得るステップと、特定の単語のTTS発音と関連付けられるTTS発音関連特徴を得るステップとをも含む。動作は、入力としてユーザ発音関連特徴およびTTS発音関連特徴を受信するように構成される発音決定モデルからの出力として、TTSオーディオに使用するために最高信頼度と関連付けられる特定の単語のユーザ発音または特定の単語のTTS発音の一方を選択する発音決定を生成するステップも含む。動作は、ユーザと関連付けられるユーザデバイスからの可聴出力のために、発音決定モデルから出力される発音決定によって選択された特定の単語に対するユーザ発音または特定の単語に対するTTS発音の一方を使用して質問への応答の合成音声表現を含むTTSオーディオを提供するステップも含む。

【0005】

本開示の実装例は、以下の任意選択の特徴の1つまたは複数を含んでよい。一部の实装例において、動作は、ユーザによって話される質問に対応するオーディオデータを受信するステップと、自動音声認識器(ASR)を使用して、オーディオデータを処理して質問の表記を生成するステップとを更に含む。これらの実装例において、特定の単語のユーザ発音を受信するステップは、ASRを使用してオーディオデータを処理する間にASRの間欠状態から特定の単語のユーザ発音を抽出するステップか、特定の単語のユーザ発音を伝えるオーディオデータからの特定の単語のユーザ音響表現を抽出するステップか、オーディオデータを処理して、特定の単語のユーザ発音を伝えるユーザ音素表現を生成するステップかの少なくとも1つを含む。特定の単語のユーザ発音と関連付けられるユーザ発音関連特徴は、ASRがオーディオデータに特定の単語を認識することと関連付けられる1つまたは複数の信頼度特徴も含んでよい。

【0006】

一部の例では、特定の単語のユーザ発音と関連付けられるユーザ発音関連特徴は、質問がユーザによって話されたときのユーザの地理的地域か、ユーザと関連付けられる言語人口学的情報か、ユーザおよび/または他のユーザによって話された以前の質問内の特定の単語を発音するときにユーザ発音を使用する頻度かの少なくとも1つを含む。特定の単語のTTS発音を受信するステップは、TTSシステムへの入力として、質問への応答のテキスト表現を含むTTS入力を受信するステップと、TTSシステムからの出力として、質問への応答の初期合成音声表現を含むTTSオーディオの初期サンプルを生成するステップと、TTSオーディオの初期サンプルから特定の単語のTTS音響表現を抽出するステップであり、TTS音響表現が特定の単語のTTS発音を伝える、ステップとを含んでよい。

【0007】

10

20

30

40

50

任意選択で、特定の単語のTTS発音を受信するステップは、質問への応答のテキスト表現を処理して、特定の単語のTTS発音を伝えるTTS音素表現を生成するステップを含んでよい。一部の例では、特定の単語のTTS発音と関連付けられるTTS発音関連特徴は、特定の単語に対する確認済みの好ましい発音か、1つまたは複数の補助情報源からの発音マイニングを使用して推定される特定の単語に対する未確認の発音か、特定の単語を発音するためのいくらかでも他の異形が存在するかどうかを示す発音異形特徴か、特定の単語のユーザ誤発音の可能性を示す発音複雑性特徴かの少なくとも1つを含む。

#### 【0008】

一部の実装例において、特定の単語のユーザ発音または特定の単語のTTS発音の一方を選択する発音決定を生成した後に、動作は、以降のTTS出力内の特定の単語を発音するためにユーザが特定の単語のユーザ発音または特定の単語のTTS発音のどちらの方を好むかを示すユーザからの明示的なフィードバックを受信するステップと、ユーザからの明示的なフィードバックに基づいて発音決定モデルを更新するステップとを更に含む。ここで、ユーザからの明示的なフィードバックが特定の単語のユーザ発音をユーザが好むことを示すとき、動作は、特定の単語を含むTTSオーディオを生成するときに特定の単語のユーザ発音を使用するようにTTSシステムを更新するステップを更に含む。一部の例では、ユーザデバイスからの可聴出力のためにTTSオーディオを提供した後に、動作は、特定の単語を含むユーザまたは別のユーザによって話される次の質問に対応するオーディオデータを受信するステップと、発音決定によって選択された特定の単語に対するユーザ発音または特定の単語に対するTTS発音の一方と同じようにユーザまたは他のユーザが次の質問内の特定の単語を発音したか否かを示す黙示的なユーザフィードバックを判定するステップと、黙示的なユーザフィードバックに基づいて発音決定モデルを更新するステップとを更に含む。

#### 【0009】

本開示の別の態様は、データ処理ハードウェアと、データ処理ハードウェアと通信しており、かつデータ処理ハードウェア上で実行されるとデータ処理ハードウェアに、特定の単語のユーザ発音または特定の単語のテキスト-音声発音のどちらの方がテキスト-音声オーディオに使用するためにより信頼できるかを選択するための動作を行わせる命令を記憶しているメモリハードウェアとを含むシステムを提供する。動作は、ユーザによって話される質問に存在する特定の単語のユーザ発音を受信することを含む。動作は、テキスト-音声(TTS)入力に存在する同じ特定の単語のTTS発音を受信することを含む。TTS入力は、質問への応答のテキスト表現を含み、そして特定の単語のTTS発音は、特定の単語のユーザ発音とは異なる。動作は、特定の単語のユーザ発音と関連付けられるユーザ発音関連特徴を得ることと、特定の単語のTTS発音と関連付けられるTTS発音関連特徴を得ることとをも含む。動作は、入力としてユーザ発音関連特徴およびTTS発音関連特徴を受信するように構成される発音決定モデルからの出力として、TTSオーディオに使用するために最高信頼度と関連付けられる特定の単語のユーザ発音または特定の単語のTTS発音の一方を選択する発音決定を生成することを含む。動作は、ユーザと関連付けられるユーザデバイスからの可聴出力のために、発音決定モデルから出力される発音決定によって選択された特定の単語に対するユーザ発音または特定の単語に対するTTS発音の一方を使用して質問への応答の合成音声表現を含むTTSオーディオを提供することを含む。

#### 【0010】

この態様は、以下の任意選択の特徴の1つまたは複数を含んでよい。一部の実装例において、動作は、ユーザによって話される質問に対応するオーディオデータを受信することと、自動音声認識器(ASR)を使用して、オーディオデータを処理して質問の表記を生成することとを更に含む。これらの実装例において、特定の単語のユーザ発音を受信することは、ASRを使用してオーディオデータを処理する間にASRの間欠状態から特定の単語のユーザ発音を抽出することか、特定の単語のユーザ発音を伝えるオーディオデータからの特定の単語のユーザ音響表現を抽出することか、オーディオデータを処理して、特定の単語のユーザ発音を伝えるユーザ音素表現を生成することかの少なくとも1つを含む。特定の

10

20

30

40

50

単語のユーザ発音と関連付けられるユーザ発音関連特徴は、ASRがオーディオデータに特定の単語を認識することと関連付けられる1つまたは複数の信頼度特徴を含んでよい。

【0011】

一部の例では、特定の単語のユーザ発音と関連付けられるユーザ発音関連特徴は、質問がユーザによって話されたときのユーザの地理的地域か、ユーザと関連付けられる言語人口学的情報か、ユーザおよび/または他のユーザによって話された以前の質問内の特定の単語を発音するときにユーザ発音を使用する頻度かの少なくとも1つを含む。特定の単語のTTS発音を受信することは、TTSシステムへの入力として、質問への応答のテキスト表現を含むTTS入力を受信することと、TTSシステムからの出力として、質問への応答の初期合成音声表現を含むTTSオーディオの初期サンプルを生成することと、TTSオーディオの初期サンプルから特定の単語のTTS音響表現を抽出することであり、TTS音響表現が特定の単語のTTS発音を伝える、こととを含んでよい。

10

【0012】

任意選択で、特定の単語のTTS発音を受信することは、質問への応答のテキスト表現を処理して、特定の単語のTTS発音を伝えるTTS音素表現を生成することを含んでよい。一部の例では、特定の単語のTTS発音と関連付けられるTTS発音関連特徴は、特定の単語に対する確認済みの好ましい発音か、1つまたは複数の補助情報源からの発音マイニングを使用して推定される特定の単語に対する未確認の発音か、特定の単語を発音するためのいくらかでも他の異形が存在するかどうかを示す発音異形特徴か、特定の単語のユーザ誤発音の可能性を示す発音複雑性特徴かの少なくとも1つを含む。

20

【0013】

一部の実装例において、特定の単語のユーザ発音または特定の単語のTTS発音の一方を選択する発音決定を生成した後に、動作は、以降のTTS出力内の特定の単語を発音するためにユーザが特定の単語のユーザ発音または特定の単語のTTS発音のどちらの方を好むかを示すユーザからの明示的なフィードバックを受信することと、ユーザからの明示的なフィードバックに基づいて発音決定モデルを更新することとを更に含む。ここで、ユーザからの明示的なフィードバックが特定の単語のユーザ発音をユーザが好むことを示すとき、動作は、特定の単語を含むTTSオーディオを生成するときに特定の単語のユーザ発音を使用するようにTTSシステムを更新することとを更に含む。一部の例では、ユーザデバイスからの可聴出力のためにTTSオーディオを提供した後に、動作は、特定の単語を含むユーザまたは別のユーザによって話される次の質問に対応するオーディオデータを受信することと、発音決定によって選択された特定の単語に対するユーザ発音または特定の単語に対するTTS発音の一方と同じようにユーザまたは他のユーザが次の質問内の特定の単語を発音したか否かを示す黙示的なユーザフィードバックを判定することと、黙示的なユーザフィードバックに基づいて発音決定モデルを更新することとを更に含む。

30

【0014】

本開示の1つまたは複数の実装例の詳細が添付図面および以下の説明に明らかにされる。他の態様、特徴および利点は、同説明および図面から、ならびに請求項から明らかであろう。

【図面の簡単な説明】

40

【0015】

【図1】音声環境例の概略図である。

【図2】特定の単語のユーザ発音または特定の単語のテキスト-音声発音のどちらの方がテキスト-音声オーディオに使用するためにより信頼できるかを決定するための発音決定モデル例の概略図である。

【図3】TTSオーディオを生成するためのテキスト-音声モデル例の概略図である。

【図4】特定の単語に対するユーザ発音が同じ単語に対するTTS発音とは異なりかつより信頼できる対話例の概略図である。

【図5】ユーザとデジタルアシスタントとの間の対話中のテキスト-音声の瞬時学習の方法のための動作の配置例のフローチャートである。

50

【図6】本明細書に記載されるシステムおよび方法を実装するために使用され得るコンピューティングデバイス例の概略図である。

【発明を実施するための形態】

【0016】

様々な図面中の同様の参照記号は、同様の要素を示す。

【0017】

ユーザが行為を行うためにデジタルアシスタントに対して口頭の質問を發し、そしてデジタルアシスタントが質問に関連するテキスト-音声(TTS)応答を可聴に出力するときなど、ユーザとデジタルアシスタントとの間の対話中に、しばしば口頭の質問およびTTS応答の語彙の重なりがある。例えば、ユーザは音声検索質問「ベア(Bexar)はどこにある?」を話すことができ、対応するオーディオデータが自然言語理解(NLU)モジュールによる解釈のための対応する表記への変換のために自動音声認識器(ASR)に渡され、そして検索質問として検索エンジンに提供されて結果を取り出す。取り出された結果を使用して、デジタルアシスタントは、質問への応答のテキスト表現を含むTTS入力を生成してよい。ここで、応答のテキスト表現は「ベアはテキサス州に位置する」となり、これは、ユーザと関連付けられるデバイスからの可聴出力のために合成音声を含む対応するTTSオーディオを生成するためのTTSシステムへのTTS入力に与えられる。

【0018】

一部の事例では、ユーザは、口頭の質問内の単語を、TTS入力として提供される質問への応答のテキスト表現に存在する同じ単語をTTSシステムが発音することになる仕方とは異なって発音する。競合する発音の違いは、幾つか挙げれば、一般に別々に訓練されて、しばしば更新の点で同期しなくなるASRシステムとTTSシステムとの間の正規化の欠如、TTSシステムが発音するように訓練されていない珍しい単語/用語(例えば、連絡先名、新語等)、ユーザが知らずに誤発音しており、ASRシステムは口バストであり正確に認識しかつTTSシステムは正しく発音する単語、およびユーザが故意に誤発音してASRシステムを誘導する単語など、各種の理由に起因し得る。

【0019】

本明細書における実装例は、質問中にユーザによって話される用語のユーザ発音が質問への応答に存在する同じ用語のTTS発音とは異なる時を識別し、そして発音決定モデルを使用して、合成音声としてユーザに応答を伝えるTTSオーディオ内の単語を発音するためにユーザ発音またはTTS発音のどちらの方を使用すべきかを判定することに向けられる。以下に更に詳細に記載されるように、発音決定モデルは、TTS発音関連特徴および/またはユーザ発音関連特徴に基づいて同じ単語の競合する発音間で選択するように訓練および連続的に更新されてよい。すなわち、特定の単語のユーザ発音と関連付けられるユーザ発音関連特徴および特定の単語のTTS発音と関連付けられるTTS発音関連特徴に基づいて、発音決定モデルは、特定の単語のユーザ発音の信頼度および同じ特定の単語のTTS発音の信頼度を推定して、ユーザ発音またはTTS発音のどちらの方がより信頼できるかを判定してよい。

【0020】

図1を参照すると、一部の实装例において、音声環境100が、音声対応デバイス110(デバイス110またはユーザデバイス110とも称される)に質問12を話すユーザ10を含む。具体的には、ユーザ10は、ユーザデバイス110上で実行しているデジタルアシスタント115(『デジタルアシスタントインタフェース』とも称される)と対話をしており、ユーザ10によって話される質問12がデジタルアシスタント115に動作を行うことを要求する。ユーザ10(すなわち、質問12の話者)は、質問12を話して、デジタルアシスタント115に応答を求めてまたはデジタルアシスタント115に質問12によって指定されるタスクを実行させてよい。デバイス110は、音声環境100内の1人または複数のユーザ10からの音を捕捉するように構成される。デバイス110のまたはデバイス110と関連付けられる音声対応システム(例えば、デジタルアシスタント115)は、質問12に対応し、質問12によって指定される1つまたは複数の動作を行い、そしてデバイス110からの可聴出力のために合成音声154

として質問への応答を提供してよい。

【0021】

ここで、デバイス110は、ユーザ10による口頭の質問12に対応するオーディオデータ14を捕捉する。デバイス110は、ユーザ10と関連付けられかつオーディオデータ14を受信することが可能な任意のコンピューティングデバイスに相当してよい。ユーザデバイス110の一部の例には、モバイルデバイス(例えば、移動電話、タブレット、ラップトップ、電子書籍リーダ等)、コンピュータ、ウェアラブルデバイス(例えば、スマートウォッチ)、音楽プレーヤ、キャストリングデバイス、スマート家電(例えば、スマートテレビ)およびモノのインターネット(IoT)デバイス、リモートコントロール、スマートスピーカ等を含むが、これらに限定されない。デバイス110は、データ処理ハードウェア112、およびデータ処理ハードウェア112と通信しており、かつデータ処理ハードウェア112によって実行されると、データ処理ハードウェア112に音声および/またはテキスト処理に関連する1つまたは複数の動作を行わせる命令を記憶しているメモリハードウェア114を含む。一部の例では、デバイス110は、1つまたは複数のアプリケーション(すなわち、ソフトウェアアプリケーション)を含んでおり、各アプリケーションがデバイス110と関連付けられる1つまたは複数の音声処理システム/モデル140、150、200を活用してアプリケーション内の様々な機能を行ってよい。例えば、デバイス110は、ユーザ10に合成再生オーディオ154(合成音声154とも称される)を伝達するように構成されるデジタルアシスタント115を実行して、ユーザ10と会話して様々なタスクの履行を支援する。

【0022】

デバイス110は、音声環境100内のオーディオデータ14を捕捉して電気信号へ変換するためのオーディオ捕捉デバイス(例えば、マイクロホン)116および可聴オーディオ信号(例えば、デバイス110からの合成再生信号154)を伝達するための音声出力デバイス(例えば、スピーカ)118を伴うオーディオサブシステムを更に含む。図示した例ではデバイス110が単一のオーディオ捕捉デバイス116を実装するのに対して、デバイス110は、本開示の範囲から逸脱することなくオーディオ捕捉デバイス116のアレイを実装してよく、それによってアレイ内の1つまたは複数のオーディオ捕捉デバイス116が物理的にデバイス110に常駐するのではなく、オーディオサブシステムと通信してよい(例えば、デバイス110の周辺機器)。例えば、デバイス110は、車両の全体にわたって配設されるマイクロホンのアレイを活かす車両インフォテインメントシステムに相当してよい。同様に、音声出力デバイス118は、1つまたは複数のスピーカを含んでよく、デバイス110に常駐するか、それと通信しているか、組合せとして1つまたは複数のスピーカがデバイス110に常駐し、そして1つまたは複数の他のスピーカが物理的にデバイス110から除かれるがデバイス110と通信しているかである。

【0023】

更には、デバイス110は、リモートシステム130とネットワーク120を介して通信するように構成されてよい。リモートシステム130は、リモートデータ処理ハードウェア134(例えば、リモートサーバもしくはCPU)および/またはリモートメモリハードウェア136(例えば、リモートデータベースもしくは他のストレージハードウェア)などのリモートリソース132を含んでよい。デバイス110は、リモートリソース132を活用して、音声処理および/または合成再生伝達に関連する様々な機能性を行ってよい。例えば、デバイス110は、自動音声認識(ASR)システム140を使用して音声認識を、および/またはテキスト-音声(TTS)システム150を使用してテキストから音声への変換を行うように構成される。追加的に、発音決定モデル200が、ユーザ10によって話される質問12にも質問12への応答のテキスト表現に対応するTTS入力152にも存在する特定の単語に対するユーザ発音202またはTTS発音204間で選択する発音決定250を生成してよい。TTSシステム150は、質問12への応答の合成音声表現を含むTTSオーディオ154を作成し、それによってTTSオーディオ154は、発音決定モデル200から出力される発音決定250によって選択されたユーザ発音202またはTTS発音204の一方を使用して特定の単語を発音する。

【0024】

これらのシステム/モデル140、150、200は、デバイス110に常駐し(オンデバイスシステムと称される)、またはリモートで常駐する(例えば、リモートシステム130に常駐する)がデバイス110と通信してよい。一部の例では、これらのシステム140、150、200の一部がローカルまたはオンデバイスに常駐する一方で、他はリモートで常駐する。言い換えれば、これらのシステム140、150、200のいずれも、ローカル、リモート、または任意の組合せで両方でよい。例えば、システム140、150、200のサイズまたは処理要件がかなり大きい場合、システム140、150、200は、リモートシステム130に常駐してよい。更にデバイス110が1つまたは複数のシステム140、150、200のサイズまたは処理要件をサポートし得る場合、1つまたは複数のシステム140、150、200は、データ処理ハードウェア112および/またはメモリハードウェア114を使用してデバイス110に常駐してよい。任意選択で、システム140、150、200の1つまたは複数は、ローカル/オンデバイスにもリモートでも常駐してよい。例えば、デバイス110とリモートシステム130との間のネットワーク120への接続が利用可能であるときには、システム140、150、200の1つまたは複数がデフォルトでリモートシステム130上で実行してよいが、接続が失われるまたはネットワーク120が利用不可能であると、システム140、150、200は、代わりにデバイス110上でローカルに実行する。

#### 【0025】

図示した例では、ASRシステム140は、入力として、質問12に対応するオーディオデータ14を受信し、そしてオーディオデータ14を処理して、出力として、質問12の表記142を生成する。ASRシステム140は、自然言語理解(NLU)機能性を含んで表記142に質問解釈(例えば、意味解析)を行ってよい。表記142は、デジタルアシスタント115が質問12への応答を生成する、より詳細には、質問12への応答のテキスト表現に対応するTTS入力152を生成するために使用してよいテキストの系列によって表現される書記素表現を含む。例えば、前からの例を続けると、ユーザ10は、質問「Where is Bexar?(Bexarはどこにある?)」を話すことができ、これが捕捉されて、対応する表記142を生成するASRシステム140による処理のために対応するオーディオデータ14へ変換される。単語Bexarに対する標準的な発音がテキサス州の文脈では「

#### 【0026】

##### 【数1】

“BAY-ər”

#### 【0027】

」であるのに対して、ユーザ10は、知らずに「X」をはっきり発音することによって単語Bexarをその英語スペリングに基づいて「

#### 【0028】

##### 【数2】

“Bək-sar”

#### 【0029】

」として誤発音した。特に、ASRシステム140は口バストであり、テキサス州の都市としての単語Bexarのユーザ誤発音を正確に認識する。

#### 【0030】

表記142を生成した(そして任意のNLU機能性を行った)後に、デジタルアシスタント115は、次いで表記142を使用して質問12への応答を判定してよい。例えば、都市名(例えば、Bexar)に対して場所を判定するために、デジタルアシスタント115は、検索エンジンに表記142(「Where is Bexar?」)または表記の識別部分(例えば、「where」および「Bexar」)を含む検索文字列を渡してよい。検索エンジンは、次いでデジタルアシスタント115が質問12への応答のテキスト表現を含むTTS入力152を生成するために解釈する1つまたは複数の検索結果を返してよい。特に、単語「Bexar」が質問12の表記142にもTTS入力152にも存在しており、それによって表記142もTTS入力152のテキスト表現も単語「Be

xar」の同じ書記素表現を共有する。テキスト表現は、特定の自然言語の書記素/文字の系列を含んでよい。書記素/文字の系列は、字母、数字、句読点および/または他の特殊文字を含むことができる。

【0031】

TTSシステム150は、TTS入力152を、デバイス110がユーザ10に質問12への応答を伝達するために可聴に出力することになる対応するTTSオーディオ154(例えば、合成音声)へ変換してよい。TTSオーディオ154の可聴出力に先立ち、TTSシステム150および/または発音決定モデル200は、質問12に存在する特定の単語のユーザ発音202がTTS入力152に存在する同じ特定の単語のTTS発音204とは異なることを識別してよい。図示した例では、単語Bexarのユーザ発音202が「

10

【0032】

【数3】

“Bək-sar”

【0033】

」である一方で、単語BexarのTTS発音204は「

【0034】

【数4】

“BAY-ər”

20

【0035】

」である。一部の実装例において、発音決定モデル200は、特定の単語のユーザ発音202と関連付けられるユーザ発音関連特徴(UPF)210および特定の単語のTTS発音204と関連付けられるTTS発音関連特徴(TTSPF)230を得る。図2を参照しつつ以下に更に詳細に記載されるように多数の異なるUPF210およびTTSPF230が得られて発音決定250を生成するためのベースを形成してよいが、図1の本例に描かれる発音決定モデル200は、ユーザがBexarを度々誤発音することを示すユーザ統計を含むUPF210の他に、Bexarがテキサス州の文脈では広く「

【0036】

【数5】

“BAY-ər”

30

【0037】

」として発音されることを示すTTSPF230を得る。したがって、発音決定モデル200は、ユーザが通常Bexarを誤発音しており、かつテキサス州の文脈で話される場合TTS発音204が広く採用されるという事実により、BexarのTTS発音204に対して推定される信頼度より低いBexarのユーザ発音202に対する信頼度を推定してよい。特に、TTS応答のテキスト表現に存在する単語Bexarに対する「

【0038】

【数6】

“BAY-ər”

40

【0039】

」のTTS発音204は、確認されるのではなく、代わりに1つまたは複数の補助情報源からの発音マイニングを通じてTTSシステム150によって学習される。例えば、補助情報源は、TTSシステム150によって使用するための様々な文脈内の様々な用語の発音のためにマイニングされ得るオーディオおよび/またはビデオウェブソースを含んでよい。

【0040】

TTSシステム150がTTS入力152からTTSオーディオ154を生成すると、TTS出力154は、質問12への応答のテキスト表現を含むTTS入力152を定める書記素/文字の系列によ

50

って形成される単語を人間が発音するであろう仕方に近似する合成音声を含む。例を続けると、発音決定250が特定の単語BexarのTTS発音204を選択する結果として、TTSオーディオ154でのTTS発音204の使用を維持し、そしてユーザデバイス110からの可聴出力のためにTTSオーディオ154を提供することになる。そのため、TTSオーディオ154は、質問12への応答の合成音声表現として可聴に出力され、かつTTS入力152に存在する単語Bexarのテキスト表現に対して「

【0041】

【数7】

“BAY-or”

【0042】

」のTTS発音204を使用する。

【0043】

一部の例では、特定の単語に対してユーザ発音202またはTTS発音204の選択された一方を使用して可聴出力のためにTTSオーディオ154を提供した後に、発音決定モデル200および/またはTTSシステム150は、ユーザ10が発音決定250に同意するか、代わりに発音決定250に選択されなかったユーザ発音202またはTTS発音204の他方を好むかを示す明示的なフィードバックを提供するようにユーザ10に促す。より詳細には、特定の単語に対する好ましい発音を示す明示的なフィードバックは、特定の単語の競合するユーザおよびTTS発音に対して次の発音決定250をするときに発音決定モデル200によって得られるTTSPF230として使用されてよい。ユーザ10は、モデル200に近い(例えば、閾値信頼度範囲内である)ユーザおよびTTS発音202、204に対して信頼度を推定するときに、ならびに/または信頼度が信頼度閾値を満足させることができないときに、明示的なフィードバックを提供するように促されてよい。明示的なフィードバックは、発音決定モデル200および/またはTTSシステム150を更新するために使用されてよい。発音決定モデル200および/またはTTSシステム150を更新するのに先立ち、特定の単語に対する好ましい発音が妥当な異形であることを確認する、および/または明示的なフィードバックが敵対的なユーザ訂正でないことを確認するために、追加の確認ステップが実施されてよい。

【0044】

上の例の代替の実装例において、TTSPF230が単語Bexarに対する「

【0045】

【数8】

“Bək-sar”

【0046】

」のユーザ発音202が以前に好ましい発音としてユーザ10によって指定されたことを示せば、発音決定モデル200は、おそらくTTS発音204よりユーザ発音202に対して高い信頼度を推定し、その結果としてTTSシステム150がTTSオーディオ154を変更してユーザ発音202を採用するであろう。この実装例において、TTSシステム150がユーザに特有のローカル/カスタムTTSシステム150であれば、TTSPF230が単語Bexarに対して好ましい発音として「

【0047】

【数9】

“Bək-sar”

【0048】

」のユーザ発音202を示す結果として、TTSシステム150が最初に、TTS入力152に存在するときに単語Bexarを「

【0049】

【数10】

10

20

30

40

50

“Bək-sar”

【0050】

」として発音することによってTTSオーディオ154を生成するようにTTSシステム150を再訓練することになってよい。

【0051】

TTSシステム150は、入力テキスト(例えば、TTS入力)をTTSオーディオ154として対応する合成音声表現へ変換することが可能な任意の種類のTTSシステムを含んでよい。例えば、TTSシステム150は、パラメトリックTTSモデル、またはディープニューラルネットワーク(例えば、アテンションベースのTacotronネットワーク)を活用して合成音声としての可聴出力のためにTTSオーディオ154を生成するTTSモデル300(図3のTTSモデル)を利用してよい。一部の実装例において、TTSモデルは、音声特徴(例えば、TTS入力152の特徴)のエンコードされた表現である埋込みを処理して、TTSオーディオ154を表現するオーディオ波形(例えば、時間とともにオーディオ信号の振幅を定める時間領域オーディオ波形)を生成する。一旦生成されると、TTSシステム150は、TTSオーディオ154をデバイス110に通信して、デバイス110が質問12への応答の合成音声表現としてTTSオーディオ154を出力できるようにする。例えば、デバイス110は、1つまたは複数のスピーカ118から、モデル200から出力される発音決定250によって選択される

10

【0052】

【数11】

20

Bək-sar

【0053】

のTTS発音204を使用して「Bexar is in Texas(Bexarはテキサス州にある)」のTTSオーディオ154を可聴に出力する。ここで、TTSシステム150のTTSモデル300は、合成音声154の音声関連属性を制御するように構成される。言い換えれば、TTSモデル300は、自然度の点で人間の話者の声をシミュレートする一方で、詳細潜在特徴をモデル化することによって多様な合成音声を生成することもできるように構成される。図1がデジタルアシスタント115に対する応用の文脈でのTTSシステム150の一例を描くが、TTSシステム150(例えば、TTSモデル300を使用)は、例えば音声検索、ナビゲーションまたは文書読上げなどの、他のテキスト-音声シナリオに応用可能である。

30

【0054】

一部の实装例において、TTSシステム150と関連付けられるTTSモデル300がユーザデバイス110に常駐するが、地理的地域内のユーザ全体に共有されるグローバルモデルとして訓練される場合、モデル300を更新/再訓練するために連合学習技術が利用される。例えば、ユーザデバイス110は、ローカルに実行しているTTSモデル300のオンデバイスバージョンのパラメータを再訓練/更新し、次いで更新されたパラメータまたは訓練損失を、グローバルTTSモデル300を更新するためのサーバと共有してよい。そうする際に、グローバルTTSモデル300は、ユーザ人口全体のユーザに自分のオーディオデータ、質問内容、またはユーザが共有したくない他の情報を共有することを要求することなく複数ユーザから受信されるパラメータ更新/訓練損失を使用して再訓練/更新されてよい。

40

【0055】

図3を参照すると、一部の例では、TTSモデル300は、エンコーダ302およびデコーダ304を有するエンコーダデコーダネットワークアーキテクチャを含む。一部の实装例において、エンコーダデコーダ302、304の構造は、Tacotron 2の系列間リカレントニューラルネットワーク(RNN)に相当する(例えば、Shen、Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions、<https://arxiv.org/pdf/1712.05884.pdf>において入手可能、に記載されており、参照により本明細書に組み込まれる)。一部の構成では、エンコーダ302は、入力として、TTS入力152またはTTS入力152に対応する埋込み(例えば、文字埋込み)を受信し、そして出力として、デコーダ304が後に

50

生成することになる各メル周波数スペクトログラムに対して文脈ベクトル $V_c$ を生成するように構成される。文脈ベクトル $V_c$ は、固定長であり、一般にTTS入力152のテキスト表現を形成する文字の系列に対応する特定の位置に現れる特徴を定めてよい。一部の構成では、TTS入力152は、エンコーダ302へ入力されるのに先立ちまず対応する音素系列へ(例えば、書記素音素モデルなどの正規化エンジンを介して)変換される書記素系列を含む。

【0056】

エンコーダ302は、双方向長短期記憶(LSTM)層が続く1つまたは複数の畳込み層を含んでよい。各畳込み層におけるニューロンは、前の層におけるニューロンの小さなサブセットから入力を受信してよい。この点で、ニューロン結合性により、畳込み層は、TTS入力152のテキスト表現に対応する文字の系列の位置に特定の隠れ特徴が現れると活性化するフィルタを学習できるようになる。一部の実装例において、各畳込み層におけるフィルタは、一連の文字(例えば、4、5または6文字)にわたってよい。各畳込み層に、バッチ正規化および整流化線形ユニット(RELU)が続いてよい。エンコーダ302が1つまたは複数の畳込み層を含む場合、双方向LSTM層がこれらの畳込み層に続いてよい。ここで、双方向LSTMは、TTS入力152に対応する文字の系列の順次特徴表現を生成するために最終畳込み層によって生成される隠れ特徴を処理するように構成される。順次特徴表現は、特徴ベクトルの系列を含んでよい。

【0057】

一部の实装例において、エンコーダ302は、エンコーダ302から順次特徴表現を受信するように、そして順次特徴表現を処理して各デコーダ出力ステップに対して文脈ベクトル $V_c$ を生成するように構成されるアテンションネットワークも含む。すなわち、アテンションネットワークは、デコーダ304が後に生成することになるメル周波数スペクトログラムの各フレームに対して固定長文脈ベクトル $V_c$ を生成できる。フレームは、入力信号の小部分(例えば、10ミリ秒サンプル)に基づくメル周波数スペクトログラムの単位を指す。アテンションネットワークのアーキテクチャは、特定のTTSシステム150に応じて変化してよい。アテンションネットワークの一部の例には、加法アテンションネットワーク、位置有感アテンションネットワーク、ガウス混合モデル(GMM)アテンションネットワーク(例えば、長い発話に対する一般化を改善する)、前方アテンションネットワーク、段階的単調アテンションネットワークまたは動的畳込みアテンションネットワークを含む。アテンションネットワークにより、TTSモデル300は、文脈ベクトル $V_c$ を生成するために特定のアテンション重みを受信する追加の入力(例えば、音声埋込み $e$ を伴う)に基づいて出力系列(例えば、出力ログメルスペクトログラムフレームの系列)を生成することができよう。

【0058】

デコーダ304は、意図された音声関連属性(例えば、TTS入力152に存在する各単語に対するTTS発音、意図された韻律および/または音声特性)を含む表出的な音声の出力オーディオ信号AS(例えば、出力系列メル周波数スペクトログラム)を発生させるようにニューラルネットワーク(例えば、自己回帰リカレントニューラルネットワーク)として構成される。例えば、文脈ベクトル $V_c$ に基づいて、デコーダ304は、エンコーダ302によって生成されるエンコードされた表現から音声信号の表現(例えば、メルフレームまたはスペクトログラムフレーム)を予測する。すなわち、デコーダ304は、入力として、1つまたは複数の文脈ベクトル $V_c$ を受信するように構成され、そして各文脈ベクトル $V_c$ に対して、メル周波数スペクトログラムが音の周波数領域表現であるメル周波数スペクトログラムの対応するフレームを生成してよい。一部の例では、デコーダ304は、Tacotron 2と類似のアーキテクチャを含む。言い換えれば、デコーダ304は、プリネット、長短期記憶(LSTM)サブネットワーク、線形投影および畳込みポストネットを有するアーキテクチャを含んでよい。

【0059】

一部の構成では、TTSモデル300は、音声合成器306(シンセサイザ306とも称される)も含む。シンセサイザ306は、メル周波数スペクトログラムを受信するように、そして合成音声としてメル周波数スペクトログラムに基づいてTTSオーディオ154の出力サンプルを生成するように構成される任意のネットワークであることができる。一部の他の実装例

10

20

30

40

50

では、シンセサイザ306は、ボコーダを含む。例えば、音声合成器306は、WaveRNNボコーダを含んでよい(例えば、「Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions」、J. Shenら著、例えば<https://arxiv.org/abs/1712.05884>において入手可能、によって記載されているように)。ここで、WaveRNNボコーダは、TTSモデル300によって予測されるスペクトログラムを条件にする24kHzでサンプリングされる16ビット信号を発生させてよい。一部の他の実装例では、シンセサイザ306は、訓練可能なスペクトログラム波形インバータである。シンセサイザ306が波形を生成した後に、オーディオサブシステムが、波形を使用してTTSオーディオ154を生成し、そして合成音声としての可聴再生(例えば、デバイス110上)のためにTTSオーディオ154を提供する、または生成された波形を別のシステムに提供して、他のシステムがTTSオーディオ154を生成および再生できるようにすることができる。一般的に言えば、シンセサイザ306は、合成音声154の結果的な発音、韻律および/または文体への影響をほとんどまたは全く有さず、実際には、シンセサイザ306が音声信号の表現(例えば、デコーダ304によって出力されるメルフレームまたはスペクトログラムフレーム)を波形へ変換するのでTTSオーディオ154のオーディオ忠実度に影響を与えるだけである。

#### 【0060】

図2は、特定の単語のユーザ発音202または同じ特定の単語の異なるTTS発音204のどちらの方がTTSオーディオ154に使用するためにより信頼できるかを選択する発音決定250を生成する発音決定モデル例200の概略図を示す。発音決定モデル200は、ユーザデバイス110、リモートシステム130またはその組合せ上で実行してよい。一部の实装例において、発音決定モデル200は、ユーザ10とデジタルアシスタント115との間の対話/相互作用からの学習に基づいてユーザ10に固有の発音決定250をするために発音決定モデル200がカスタマイズされるという点でユーザ固有である。他の実装例において、発音決定モデル200は、複数ユーザによって共有され、そして発音決定250を全てのユーザ10が自分のデジタルアシスタント115とする対話に基づかせることを瞬時に学習する。

#### 【0061】

発音決定モデル200は、様々な技術を使用して特定の単語に対するユーザ発音202を受信してよい。例えば、特定の単語のユーザ発音202は、ASRシステム140がユーザによって話される質問12に対応するオーディオデータ14を処理している間にASRシステム140の間欠状態から抽出されてよい。別の実装例において、モデル200は、オーディオデータから特定の単語のユーザ音響表現を抽出することによって特定の単語のユーザ発音202を受信する。ここで、ユーザ音響表現は、特定の単語に対するユーザ発音202を伝える。更に別の実装例において、質問12に対応するオーディオデータ14が処理されて、特定の単語のユーザ発音を伝えるユーザ音素表現を生成する。提示される開示は、単語のユーザ発音を得るためのいずれの具体的な技術にも限定されない。

#### 【0062】

発音決定モデル200は、様々な技術を使用して特定の単語に対するTTS発音204を受信してよい。一例では、発音決定モデル200は、特定の単語のTTS発音204を伝える音素表現を受信する。例えば、TTSシステム150は、TTS入力152の書記素表現に対応する音素表現へ変換する書記素音素モデル(例えば、図3の正規化エンジン)を含んでよい。別の例では、TTSシステム150によって生成される(が可聴に出力されない)TTSオーディオの初期サンプルから特定の単語のTTS音響表現(例えば、波形)が抽出される。ここで、音響表現は、特定の単語のTTS発音204を伝える。TTS音響表現は、図3のTTSモデル300のデコーダ304からの出力として発生されるオーディオ信号As(例えば、出力系列メル周波数スペクトログラム)からも抽出され得る。更に別の例では、TTS発音204は、TTS入力152のテキスト表現を音声への変換のために処理する間にTTSシステム150の間欠状態から抽出される。例えば、TTSシステム150の抽出される間欠状態は、図3のTTSモデル300のエンコーダ302から出力される系列文脈ベクトルVcを含んでよい。

#### 【0063】

ユーザ10によって話される質問12にも質問12への応答のテキスト表現を含むTTS入力1

52にも存在する特定の単語に対して、TTSシステム150または発音決定モデル200は、特定の単語のユーザ発音202およびTTS発音204が互いと異なり、そのため合成音声としての可聴出力のために提供されるTTSオーディオ154内の特定の単語を発音するために最終的にどちらの方が使用されることになるかに関して競合している時を識別してよい。発音決定モデル200は、特定の単語と関連付けられる1つもしくは複数のユーザ発音関連特徴(UPF)210および/または1つもしくは複数のTTS発音関連特徴(TTSPF)230に基づいて発音決定250を生成するように訓練される。ここで、発音決定250は、質問12への応答を伝えるTTSオーディオ154内の特定の単語を発音するために使用するべき最高信頼度と関連付けられるユーザ発音202またはTTS発音204の一方を選択する。一部の例では、発音決定モデル200は、ユーザ発音信頼度推定子220およびTTS発音推定子240を含む。これらの例では、ユーザ発音信頼度推定子220は、ユーザ発音202が好ましいという可能性を示す特定の単語のユーザ発音202に対するユーザ発音信頼度222を推定する。他方で、TTS発音推定子240は、TTS発音204が好ましいという可能性を示す特定の単語のTTS発音204に対するTTS発音信頼度242を推定する。一部の実装例において、発音決定モデル200は、ユーザ発音信頼度222とTTS発音信頼度242との間の対数尤度比として発音決定250を生成する。

#### 【0064】

特定の単語のユーザ発音202と関連付けられるUPF210には、質問12中にユーザ10によって話されるときに特定の単語の発音と関連付けられる信頼性(または非信頼性)を確かめる際に発音決定モデル200にとって有益な情報を伝える多数の異なる特徴を含んでよい。例えば、質問12の表記142は、特定の単語の単語型および表記内に認識される他の単語に鑑みて特定の単語の文脈を伝え得る。追加的に、ユーザ10と関連付けられる言語人口学的情報は、ユーザの特定の言語の熟練度を伝え得る。例えば、母国語としてアメリカ英語を話すユーザ10がイタリアンレストランの名前を誤発音しがちであり得る一方で、英語およびイタリア語の2カ国語話者はイタリア語の名前/単語を誤発音しそうにない。UPFは、ユーザ10と関連付けられる、年齢および性別などの、任意の種類の人口的情報を提供し得る。質問12が話されたときにユーザ10が位置する地理的地域は、特定の単語のユーザ発音202と関連付けられるUPF210として役立ち得る。例えば、珍しい名前を伴う街路アドレスへのナビゲーション命令のための質問12では、TTSシステム150を訓練するために使用される訓練例に珍しい名前の存在が乏しいまたは実在しないことがあるので、TTSシステム150は、珍しい名前を誤発音する傾向があり得る。したがって、ユーザの地理的地域が珍しい名前を伴って要求された街路アドレスに近いことを特定するUPF210は、珍しい名前のユーザ発音202が珍しい名前のTTS発音204より信頼でき得るという強力な指標の役目をし得る。ユーザは、ユーザの地理的地域へのアクセスを明示的に許可してよく、そして地理的地域へのアクセスをいつでも拒否してよい。このシナリオでは、UPF210によって提供されるユーザ10の地理的地域と結合され、TTSPF230により、TTSシステム150が訓練されている場所/方言が要求された街路アドレスと関連付けられる地理的地域の外側であり得ることを示し、それによって更にユーザ発音信頼度222を高めかつ/またはTTS発音信頼度242を下げる。

#### 【0065】

その上、ユーザおよび/または他のユーザによって話された以前の質問内の特定の単語を発音するときにユーザ発音202を使用する頻度を提供する履歴情報が、発音決定モデル200への入力として提供されるUPF210として役立ち得る。例えば、特定の単語に対して地理的地域全体の複数ユーザが度々同じユーザ発音202を使用することで、TTSシステム150および/またはASRシステム140が訓練されていない新生の発音および/または未知の方言変化を示し得る。モデル200によって生成される信頼度発音決定250を改善することに加えて、特定の単語のユーザ発音202を反映するTTSオーディオ154を作成することを学習するようにTTSシステム150を更新するために、特定の単語を発音する際の新生の発音および/または未知の方言変化の知識が使用されてよい。

#### 【0066】

一部の例では、以前の質問12内の特定の単語のユーザ発音202を使用してユーザ10の閾値頻度を確立することで、特定の単語の好ましい発音としてユーザ発音202を示す默示的なユーザフィードバックの役目をする。例えば、決定モデル200がTTSオーディオ154内の特定の単語を発音するために異なるユーザ発音202よりもTTS発音204を選択した1つまたは複数の以前の対話セッションにおいて、ユーザ10および/または他のユーザが特定の単語を発音するためにユーザ発音202を使用し続けた以降の質問12が、ユーザ10および/または他のユーザがTTS発音204よりもユーザ発音202を好むことを示し得る。この默示的なフィードバックは、TTSシステム150を訓練/更新するために更に使用されてよい。

【0067】

一部の追加の実装例において、特定の単語のユーザ発音202と関連付けられるUPF210には、オーディオデータ14におけるASRシステム140による特定の単語の認識が正しいという可能性を示す1つまたは複数のASR信頼度特徴を含む。発音決定モデル200は、質問12の表記142を生成するために使用されるASRシステム140から直接にASR信頼度特徴を得てよい。ASR信頼度特徴には、限定することなく、特定の単語を認識することと関連付けられる可能性/事後スコア、特定の単語が正しく認識されたという可能性を示す信頼度スコア(ASRシステム140または補助信頼度モデルによって得られる)、および単語を認識することと関連付けられる任意の認識混同性を含んでよい。ここで、認識混同性は、表記142に対する候補認識結果の事後単語ラティスおよび/またはASRシステム140のニューラルネットワークモデルの可能な認識結果の事後分布のエントロピーから確かめられてよい。

【0068】

一例では、TTS発音204を使用して発音されたときのASRシステム140による特定の単語の以前の誤認識に続き、代わりにユーザ発音202を使用して単語を発音した特定の単語の次の認識成功は、ユーザ10が単にASRシステム140を欺き誤認識エラーを回避するためにユーザ発音202を選んだおそれがあるので、ユーザ発音202が信頼できないことを示し得る。そのような知識は、特定の単語のTTS発音204を認識する際に口バストになるようにASRシステム140を更新するために使用されてよい。追加的または代替的に、ユーザ発音202を使用して発音されたときのASR140による特定の単語の以前の誤認識に、代わりにTTS発音204を使用して発音された特定の単語の次の認識が続いたという知識は、ユーザ発音202を認識する際に口バストになるようにASRシステム140を更新するために使用されてよい。

【0069】

UPF210と同様に、特定の単語のTTS発音204と関連付けられるTTSPF230には、TTSシステム150によって受信されるTTS入力152に存在する特定の単語を発音するためにTTS発音204を使用するための信頼性(または非信頼性)を確かめる際に発音決定モデル200にとって有益な情報を伝える多数の異なる特徴を含んでよい。例えば、質問12への応答のテキスト表現を含むTTS入力152は、特定の単語の単語型およびTTS入力152に存在する他の言葉に鑑みて特定の単語の文脈を伝え得る。追加的に、TTSシステム150が訓練されている地理的地域/方言は、発音決定モデル200にTTSPF230として提供されてよい。

【0070】

一部の例では、TTSPF230は、TTS発音204が特定の単語に対する確認済みの好ましい発音を含むかどうかを示す。好ましい発音として確認されるTTS発音204は、TTS発音204に対してTTS発音信頼度推定子240によって推定されるTTS発音信頼度242を上昇させかつユーザ発音202に対してユーザ発音信頼度推定子220によって推定されるユーザ発音信頼度222を低下させるための強力な指標の役目をし得る。特定の単語に対する好ましい発音は、特定の単語の好ましい発音のオーディオを対応する書記素表現でマッピングする訓練サンプルペアに関してTTSシステム150を訓練することによって、TTSシステム150の訓練中に手動で確認されてよい。例えば、母国語としてスウェーデン語を話さないユーザ10がスウェーデンを海外旅行しており、スウェーデンの都市への行き方のための質問12を話すか都市名を誤発音すると、TTSシステム150がスウェーデン語に対して訓練されておりかつ都市名のTTS発音204が確認されていることをTTSPF230が示すことで、TTS発

音204がユーザ発音202より信頼できるという信頼度を上昇させることになる。

【0071】

好ましい発音には、ユーザ10がユーザによって指定される単語に対する好ましい発音を提供し、そしてTTSオーディオ154内の単語を発音するために好ましい発音を使用するようにTTSシステム150を構成してなる、ユーザ確認の好ましい発音も含んでよい。このシナリオは、さもなければTTSシステム150が単に名前の書記素表現だけに基づいて異なって発音するであろう固有の発音を有し得るカスタム名または連絡先名を発音するために共通である。同様に、ASRシステム140は、これらの連絡先名を正確に認識するのは、それらの固有の発音により困難であり得る。特定の単語(例えば、名前または他の固有名詞)に対する好ましい発音を提供するために、ユーザは、好ましい発音を使用して単語を話してよく、そしてTTSシステム150は、好ましい発音を対応する書記素表現にマッピングしてよく、その結果、対応する書記素表現がTTS入力152に存在するときに、TTSシステム150が好ましい発音を使用して単語を発音する。

10

【0072】

他の例では、ユーザ10は、デジタルアシスタント115との以前の対話セッション中に明示的なフィードバックを介して特定の単語に対する好ましい発音を提供する。これらの例では、ユーザ10によって話された質問中に単語が発音された仕方とは異なって特定の単語が発音される、質問への応答を伝えるTTSオーディオ154の可聴出力の後に、デジタルアシスタント115は、次のTTSオーディオ154内の単語を発音するためにユーザがどちらの発音を好むかを示すために明示的なフィードバックを提供するようにユーザ10に促してよい。別のシナリオでは、促されることなく、ユーザ10は、TTSオーディオ154内の特定の単語の発音に満足していなければ、「正しく言わなかった」または「[好ましい発音を使用して]ということ?」などの応答を発することによって明示的なフィードバックを提供してよい。特定の単語のTTS発音204がユーザ確認の好ましい発音を含むことをTTSPF230が示すことは、ユーザが質問12を話すときに特定の単語の異なるユーザ発音202を使用することを意識的に選ぶ理由の見解に矛盾するように見えるが、ユーザ10は、ASRシステム140を欺き誤認識エラーを回避しようとして、特定の単語を意図的に誤発音し、または少なくともユーザ10によって好まれる発音と矛盾して発音し得る。

20

【0073】

一部の実装例において、TTSPF230は、TTS発音204が特定の単語に対する未確認の発音を含むかどうかを示す。未確認の発音は、TTS発音の信頼性の信頼度を上昇させてよいが、その程度は確認された発音よりは小さい。特定の単語に対するTTS発音204は、発音が1つまたは複数の補助情報源からの発音マイニングを介してTTSシステム150によって学習/推定される場合は未確認でよい。例えば、上の例では、TTSシステム150は、単語/用語Bexarに対するTTS発音204を、テキサス州の文脈で標準的な発音「

30

【0074】

【数12】

“BAY-ər”

【0075】

」を正しく使用するオーディオおよび/またはビデオ情報源からの発音マイニングを使用して推定される未確認の発音として学習してよい。

40

【0076】

追加的または代替的に、発音決定250をするための発音決定モデル200に入力される他のTTSPF230には、限定することなく、発音異形および発音複雑性特徴を含む。発音異形特徴は、特定の単語を発音するためのいくらかでも他の異形が存在するかどうかを示してよい。特定の単語の発音のこれらの異形は、時間とともに発音決定モデル200によって学習されてよい。異なる発音は、異なる文脈にマッピングされてよい。例えば、大多数の話者は特定の単語の特定の発音を使用してよい一方で、特定の地理的領域の話者が同じ単語を排他的に異なって発音してよい。他方で、発音複雑性特徴は、特定の単語のユーザ誤発

50

音の可能性を示してよい。TTSシステム150は、ユーザが発音するのが困難として認められ得る特定の音素系列/表現を識別することによってユーザ誤発音の強い可能性を示す発音複雑性特徴を提供してよい。TTSPF230は、ユーザ人口が質問内の特定の単語を誤発音する頻度を示す誤発音統計を更に示してよい。

【0077】

図4は、質問12に存在する特定の単語に対するユーザ発音202が、質問12への応答を伝えるTTS入力152に存在する同じ単語のTTS発音204とは異なる、ユーザ10とユーザデバイス110上で実行しているデジタルアシスタント115との間の対話例の概略図400を提供する。図示した例では、ユーザ10は、デジタルアシスタント115に向けられる質問12「Play songs by A\$AP Rocky(A\$APロッキーによる曲を再生して)」を話す。この例では、単語「A\$AP」のユーザ発音202はA-S-A-Pであり、そしてASRシステム140は口バストであり、ユーザ発音202を「A\$AP」と正しく認識する。ここで、ASRシステム140は、質問12の表記142を生成し、そしてNLU機能性を適用して、デジタルアシスタント115が音楽プレーヤーアプリケーションにA\$AP Rockyという名のアーティストによる音楽曲を可聴に出力させる要求として質問12を解釈する。デジタルアシスタント115は、デジタルアシスタント115が質問12を理解しており、質問12を満たしている最中であることを確認する質問12への応答も生成する。ここで、デジタルアシスタント115は、質問12への応答「Playing latest songs by A\$AP Rocky(A\$APロッキーによる最新曲を再生中)」のテキスト表現を生成し、TTSオーディオ154への変換のためにTTSシステム150にTTS入力152として提供される。

【0078】

しかしながら、このシナリオでは、TTSシステム150は、TTS入力152に存在する単語「A\$AP」の発音の仕方を知らず、したがって単語「A\$AP」に対してTTS発音204は存在しない。ユーザ発音202を使用して発音される単語A\$APを認識するための強い信頼度があることを示す1つまたは複数のASR信頼度特徴を含むUPF210に基づいて、発音決定モデル200は、ユーザ発音信頼度222が高いと結論し、したがってTTSオーディオ154内の単語A\$APを発音するためにA-S-A-Pのユーザ発音202を選択する発音決定250を生成してよい。図示した例では、TTSオーディオ154は、ユーザデバイス110から可聴に出力され、単語「A\$AP」を発音するためにユーザ発音202を使用する質問12への応答の合成音声表現を含む。

【0079】

発音決定250が或る閾値を満たさない信頼度と関連付けられれば、モデル200および/またはTTSシステム150は、ユーザ発音202を使用することに対して多少の疑いがあり得ると判定し、したがってユーザが発音決定250に同意するか異なる発音を好むかを示すようにユーザ10に促してよい。プロンプトへのユーザの応答は、上記したように明示的なフィードバックを含んでよい。プロンプトは、発音決定に同意するか異なる発音を好むかをユーザ10に尋ねる合成音声をTTSシステム150が出力する可聴プロンプトでよい。プロンプトは、ユーザデバイス110と通信している画面上に通知を表示する視覚プロンプトを追加的または代替的に含んでよい。ここで、ユーザは、好ましい発音を示す、または最低でも、発音決定250に関する満足もしくは不満を示すグラフィック要素を選択するユーザ入力指示を提供し得る。

【0080】

発音決定250がユーザ発音202をTTS発音204より信頼できるとして選択する(またはTTS発音204が利用可能でない)と、TTSシステム150がTTSオーディオ154を作成することができない、またはTTSオーディオ154が、特定の単語を発音するためにユーザ発音202を使用して、さもなければ品質基準を満たさない合成品質を含むであろう、シナリオが存在し得る。このシナリオは、特定の単語のユーザ発音202がTTSシステム150(および関連付けられたTTSモデル300)が訓練されたアクセント/方言とは異なるアクセント/方言で話されるときに発生し得る。TTSシステム150は、ユーザ発音202をオンザフライで使用してTTSオーディオ154を作成するための各種の異なる技術を利用してよい。一部の例では

、TTSシステム150は、ユーザによって話される質問12に対応するオーディオデータ14から抽出される特定の単語の音響表現を使用する。これらの例では、オーディオデータ14から抽出される特定の単語の音響表現は、TTSオーディオ154へ挿入されてよい。他方で、TTSシステム150は、特定の単語の音響表現から音素表現を更に抽出し、そして音素表現を使用して特定の単語のユーザ発音202でTTSオーディオ154を作成してよい。

【0081】

追加の例では、TTSシステム150は、特定の単語のユーザ発音202を含む口頭の質問12に対応するオーディオデータ14の一部分から導出される潜在表現を得る。これらの例では、潜在表現は、ユーザ発音202を使用して特定の単語を発音するTTSオーディオ154を作成するようにTTSシステム150(および関連付けられたTTSモデル)を誘導する。更に別の例では、ユーザ発音202を含む口頭の質問12に対応するオーディオデータ14の部分に音声変換技術が適用されて、特定の単語を合成音声で発音するためにユーザ発音202を使用するTTSオーディオ154を作成する。

【0082】

ソフトウェアアプリケーション(すなわち、ソフトウェアリソース)は、コンピューティングデバイスにタスクを行わせるコンピュータソフトウェアを指してよい。一部の例では、ソフトウェアアプリケーションは、「アプリケーション」、「アプリ」または「プログラム」と称されてよい。アプリケーション例には、システム診断アプリケーション、システム管理アプリケーション、システム保守アプリケーション、ワードプロセッシングアプリケーション、スプレッドシートアプリケーション、メッセージングアプリケーション、メディアストリーミングアプリケーション、ソーシャルネットワーキングアプリケーションおよびゲーミングアプリケーションを含むが、これらに限定されない。

【0083】

非一時的メモリは、コンピューティングデバイスによる使用のために一時的または永続的にプログラム(例えば、命令の系列)またはデータ(例えば、プログラム状態情報)を記憶するために使用される物理デバイスでよい。非一時的メモリは、揮発性および/または不揮発性アドレス可能半導体メモリでよい。不揮発性メモリの例には、フラッシュメモリおよびリードオンリメモリ(ROM)/プログラマブルリードオンリメモリ(PROM)/消去可能プログラマブルリードオンリメモリ(EPROM)/電子的消去可能プログラマブルリードオンリメモリ(EEPROM)(例えば、ブートプログラムなどのファームウェアのために典型的に使用される)を含むが、これらに限定されない。揮発性メモリの例には、ランダムアクセスメモリ(RAM)、ダイナミックランダムアクセスメモリ(DRAM)、スタティックランダムアクセスメモリ(SRAM)、相変化メモリ(PCM)の他にディスクまたはテープを含むが、これらに限定されない。

【0084】

図5は、口頭の質問12にも質問12への応答のテキスト表現を含むTTS入力152にも存在する単語のユーザ発音202または異なるTTS発音204の一方を選択する方法500のための動作の配置例のフローチャートである。動作502で、方法500は、ユーザ10によって話される質問12に存在する特定の単語のユーザ発音202を受信することを含む。動作504で、方法500は、TTS入力152に存在する同じ特定の単語のTTS発音204を受信することを含む。TTS入力152は、質問12への応答のテキスト表現を含み、そして特定の単語のTTS発音204は、特定の単語のユーザ発音202とは異なる。動作506で、方法500は、特定の単語のユーザ発音202と関連付けられるユーザ発音関連特徴210を得ることを含む。動作508で、方法500は、特定の単語のTTS発音204と関連付けられるTTS発音関連特徴230を得ることを含む。動作510で、方法500は、入力としてユーザ発音関連特徴210およびTTS発音関連特徴230を受信するように構成される発音決定モデル200からの出力として、TTSオーディオ154に使用するために最高信頼度と関連付けられる特定の単語のユーザ発音202または特定の単語のTTS発音204の一方を選択する発音決定250を生成することを含む。動作512で、方法500は、ユーザ10と関連付けられるユーザデバイス110からの可聴出力のために、発音決定モデル200から出力される発音決定250によって選択された特定の

単語に対するユーザ発音202または特定の単語に対するTTS発音204の一方を使用して質問12への応答の合成音声表現を含むTTSオーディオ154を提供することを含む。

【0085】

図6は、本文書に記載されるシステムおよび方法を実装するために使用され得るコンピューティングデバイス例600(例えば、システム600)の概略図である。コンピューティングデバイス600は、ラップトップ、デスクトップ、ワークステーション、携帯情報端末、サーバ、ブレードサーバ、メインフレームおよび他の適切なコンピュータなど、様々な形態のデジタルコンピュータを表現すると意図される。ここで図示される部品、それらの接続および関係、ならびにそれらの機能は、単に例証的であると意味され、本文書に記載および/または特許請求される本発明の実装例を限定するとは意味されない。

10

【0086】

コンピューティングデバイス600は、プロセッサ610(例えば、データ処理ハードウェア610)、メモリ620(例えば、メモリハードウェア620)、ストレージデバイス630、メモリ620および高速拡張ポート650に接続する高速インタフェース/コントローラ640、ならびに低速バス670およびストレージデバイス630に接続する低速インタフェース/コントローラ660を含む。部品610、620、630、640、650および660の各々は、様々なバスを使用して相互接続され、かつ共通のマザーボードにまたは適宜他の方式で搭載されてよい。プロセッサ610(すなわち、ユーザデバイス110のデータ処理ハードウェア112またはリモートシステム130のデータ処理ハードウェア134)は、メモリ620にまたはストレージデバイス630に記憶される命令を含め、コンピューティングデバイス600内での実行のための命令を処理して、高速インタフェース640に結合されるディスプレイ680などの外部出力デバイスにグラフィカルユーザインタフェース(GUI)のためのグラフィック情報を表示できる。メモリ620およびストレージデバイス630は、ユーザデバイス110のメモリハードウェア114またはリモートシステム130のメモリハードウェア136を含んでよい。他の実装例では、複数プロセッサおよび/または複数バスが、適宜、複数メモリおよび複数種類のメモリと共に使用されてよい。同じく、複数コンピューティングデバイス600が接続されて、各デバイスが必要な動作の一部を提供してよい(例えば、サーババンク、一群のブレードサーバまたはマルチプロセッサシステムとして)。

20

【0087】

メモリ620は、コンピューティングデバイス600内に非一時的に情報を記憶する。メモリ620は、コンピュータ可読媒体、揮発性メモリユニットまたは不揮発性メモリユニットでよい。非一時的メモリ620は、コンピューティングデバイス600による使用のために一時的または永続的にプログラム(例えば、命令の系列)またはデータ(例えば、プログラム状態情報)を記憶するために使用される物理デバイスでよい。不揮発性メモリの例には、フラッシュメモリおよびリードオンリメモリ(ROM)/プログラマブルリードオンリメモリ(PROM)/消去可能プログラマブルリードオンリメモリ(EPROM)/電子的消去可能プログラマブルリードオンリメモリ(EEPROM)(例えば、ブートプログラムなどのファームウェアのために典型的に使用される)を含むが、これらに限定されない。揮発性メモリの例には、ランダムアクセスメモリ(RAM)、ダイナミックランダムアクセスメモリ(DRAM)、スタティックランダムアクセスメモリ(SRAM)、相変化メモリ(PCM)の他にディスクまたはテープを含むが、これらに限定されない。

30

40

【0088】

ストレージデバイス630は、コンピューティングデバイス600のために大容量記憶を提供することが可能である。一部の实装例において、ストレージデバイス630は、コンピュータ可読媒体である。様々な異なる実装例において、ストレージデバイス630は、フロッピーディスクデバイス、ハードディスクデバイス、光ディスクデバイスもしくはテープデバイス、フラッシュメモリもしくは他の同様のソリッドステートメモリデバイス、またはストレージエリアネットワークもしくは他の構成のデバイスを含め、デバイスのアレイでよい。追加の実装例において、コンピュータプログラム製品が情報担体に有形に具現化される。コンピュータプログラム製品は、実行されると、上記したものなど、1つまたは複

50

数の方法を行う命令を含む。情報担体は、メモリ620、ストレージデバイス630またはプロセッサ610上のメモリなど、コンピュータまたは機械可読媒体である。

【0089】

高速コントローラ640がコンピューティングデバイス600のための帯域幅集約動作を管理する一方で、低速コントローラ660は低帯域幅集約動作を管理する。そのような責務の割当ては単に例証的である。一部の実装例において、高速コントローラ640は、メモリ620、ディスプレイ680(例えば、グラフィックプロセッサまたはアクセラレータを通じて)に、および様々な拡張カード(図示せず)を受けることができる高速拡張ポート650に結合される。一部の实装例において、低速コントローラ660は、ストレージデバイス630および低速拡張ポート690に結合される。低速拡張ポート690は、様々な通信ポート(例えば、USB、Bluetooth(登録商標)、イーサネット、無線イーサネット)を含んでよく、キーボード、ポインティングデバイス、スキャナ、または例えばネットワークアダプタを通じて、スイッチもしくはルータといったネットワークデバイスなど、1つまたは複数の入力デバイスに結合されてよい。

【0090】

コンピューティングデバイス600は、図に図示されるように、幾つかの異なる形態で実装されてよい。例えば、それは、標準サーバ600aもしくは重ねて一群のそのようなサーバ600aとして、ラップトップコンピュータ600bとして、またはラックサーバシステム600cの一部として実装されてよい。

【0091】

本明細書に記載されるシステムおよび技術の様々な実装例は、デジタル電子および/もしくは光回路網、集積回路網、特別設計の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、ならびに/またはその組合せで実現できる。これらの様々な実装例は、ストレージシステム、少なくとも1つの入力デバイスおよび少なくとも1つの出力デバイスからデータおよび命令を受信するように、ならびにそれらにデータおよび命令を送信するように結合される、専用または汎用でよい、少なくとも1つのプログラマブルプロセッサを含むプログラマブルシステム上で実行および/または解釈可能である1つまたは複数のコンピュータプログラムによる実装例を含むことができる。

【0092】

これらのコンピュータプログラム(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またはコードとしても知られる)は、プログラマブルプロセッサ用の機械命令を含み、そして高水準手続き型および/もしくはオブジェクト指向プログラミング言語で、ならびに/またはアセンブリ/機械語で実装できる。本明細書で使用される場合、用語「機械可読媒体」および「コンピュータ可読媒体」は、機械可読信号として機械命令を受信する機械可読媒体を含め、プログラマブルプロセッサに機械命令および/またはデータを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、非一時的コンピュータ可読媒体、装置および/またはデバイス(例えば、磁気ディスク、光ディスク、メモリ、プログラマブル論理デバイス(PLD))を指す。用語「機械可読信号」は、プログラマブルプロセッサに機械命令および/またはデータを提供するために使用される任意の信号を指す。

【0093】

本明細書に記載されるプロセスおよび論理フローは、1つまたは複数のコンピュータプログラムを実行して、入力データを演算して出力を生成することによって機能を行う、1つまたは複数の、データ処理ハードウェアとも称される、プログラマブルプロセッサによって行うことができる。プロセスおよび論理フローは、専用論理回路網、例えばFPGA(フィールドプログラマブルゲートアレイ)またはASIC(特定用途向け集積回路)によって行うこともできる。コンピュータプログラムの実行のために適切なプロセッサは、例として、汎用および専用の両マイクロプロセッサ、ならびに任意の種類のデジタルコンピュータの任意の1つまたは複数のプロセッサを含む。一般に、プロセッサは、リードオンリメモリまたはランダムアクセスメモリまたは両方から命令およびデータを受信することになる。コンピュータの必須要素は、命令を行うためのプロセッサ、ならびに命令およびデータを

10

20

30

40

50

記憶するための1つまたは複数のメモリデバイスである。一般に、コンピュータは更に、データを記憶するための1つまたは複数の大容量ストレージデバイス、例えば、磁気、光磁気ディスクまたは光ディスクを含む、または作動的に結合されて、それからデータを受信するもしくはそれにデータを転送するもしくは両方行うことになる。しかしながら、コンピュータは、そのようなデバイスを有する必要はない。コンピュータプログラム命令およびデータを記憶するために適切なコンピュータ可読媒体は、例として、半導体メモリデバイス、例えばEPROM、EEPROMおよびフラッシュメモリデバイス、磁気ディスク、例えば内部ハードディスクまたはリムーバブルディスク、光磁気ディスク、ならびにCD-ROMおよびDVD-ROMディスクを含め、全ての形態の不揮発性メモリ、媒体およびメモリデバイスを含む。プロセッサおよびメモリは、専用論理回路網によって補足、またはそれに組み込みできる。

10

#### 【0094】

ユーザとの相互作用に備えるために、本開示の1つまたは複数の態様は、ユーザに情報を表示するためのディスプレイデバイス、例えばCRT(陰極線管)、LCD(液晶表示)モニタまたはタッチスクリーンならびに任意選択でユーザがコンピュータに入力を提供できるキーボードおよびポインティングデバイス、例えばマウスまたはトラックボールを有するコンピュータ上に実装できる。ユーザとの相互作用を提供するために他の種類のデバイスも使用でき、例えば、ユーザに提供されるフィードバックは任意の形態の感覚フィードバック、例えば視覚フィードバック、聴覚フィードバックまたは触覚フィードバックであることができ、そしてユーザからの入力は、音響、音声または触覚入力を含め、任意の形態で受信できる。加えて、コンピュータは、ユーザによって使用されるデバイスに文書を送信し、それから文書を受信することによって、例えばユーザのクライアントデバイス上のウェブブラウザから受信される要求に応答してウェブブラウザにウェブページを送信することによって、ユーザと相互作用できる。

20

#### 【0095】

幾つかの実装例が記載された。それにもかかわらず、本開示の趣旨および範囲から逸脱することなく様々な変更がなされ得ることが理解されるであろう。したがって、他の実装例は以下の特許請求の範囲内である。

#### 【符号の説明】

#### 【0096】

- 10 ユーザ
- 12 質問
- 14 オーディオデータ
- 100 音声環境
- 110 音声対応デバイス
- 112 データ処理ハードウェア
- 114 メモリハードウェア
- 115 デジタルアシスタント
- 116 オーディオ捕捉デバイス
- 118 音声出力デバイス
- 120 ネットワーク
- 130 リモートシステム
- 132 リモートリソース
- 134 リモートデータ処理ハードウェア
- 136 リモートメモリハードウェア
- 140 自動音声認識(ASR)システム
- 142 表記
- 150 テキスト-音声(TTS)システム
- 152 TTS入力
- 154 TTSオーディオ

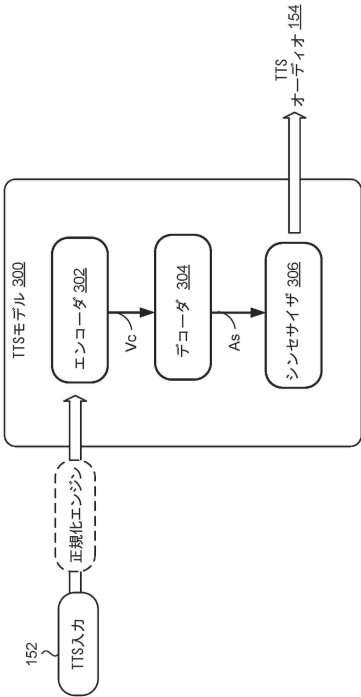
30

40

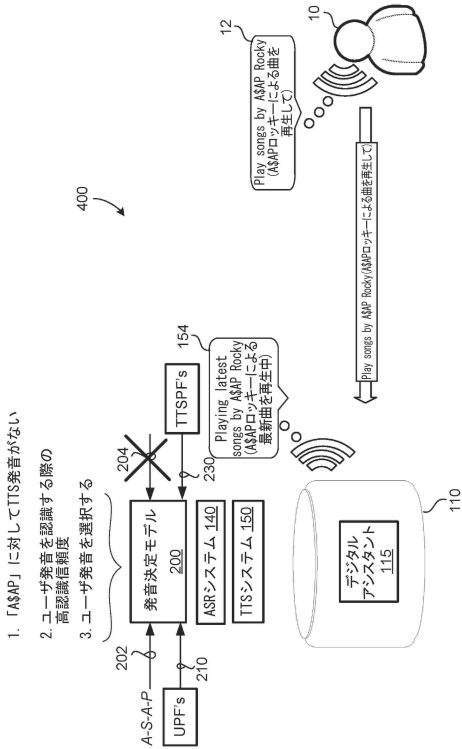
50



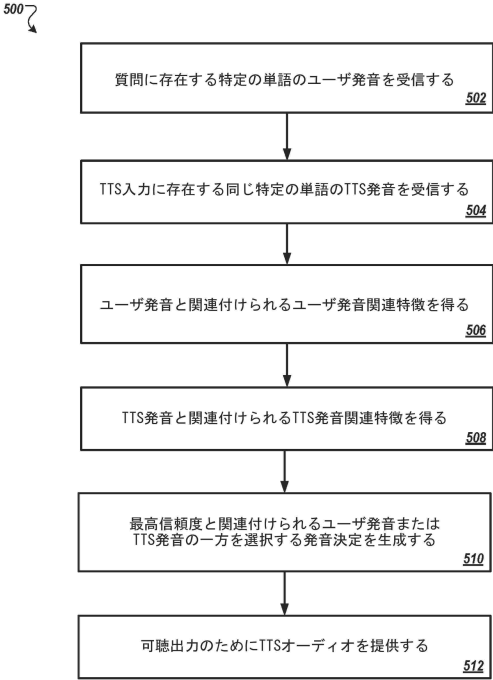
【図 3】



【図 4】



【図 5】



【図 6】

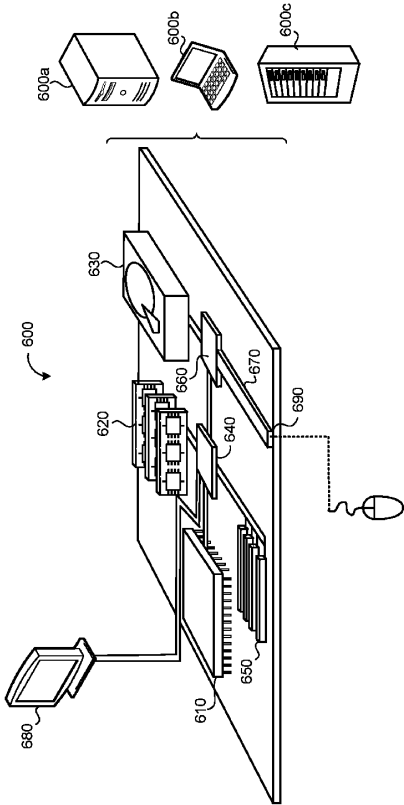


FIG. 6

## フロントページの続き

- (72)発明者 ビジャヤディティヤ・ペディンチ  
アメリカ合衆国・カリフォルニア・94043・マウンテン・ビュー・アンフィシアター・パーク  
ウェイ・1600
- (72)発明者 ブヴァナ・ラマバドラン  
アメリカ合衆国・ニューヨーク・10549・マウント・キスコ・ビクトリア・ドライブ・400  
2
- (72)発明者 アンドリュー・ローゼンバーグ  
アメリカ合衆国・カリフォルニア・94043・マウンテン・ビュー・アンフィシアター・パーク  
ウェイ・1600
- (72)発明者 マテウシュ・ゴレビエフスキ  
アメリカ合衆国・カリフォルニア・94043・マウンテン・ビュー・アンフィシアター・パーク  
ウェイ・1600
- 審査官 菊池 智紀
- (56)参考文献 特開2007-086404 (JP, A)  
特開2007-248886 (JP, A)  
米国特許出願公開第2016/0147740 (US, A1)  
米国特許出願公開第2016/0307569 (US, A1)  
米国特許出願公開第2005/0273337 (US, A1)  
米国特許出願公開第2014/0365216 (US, A1)  
国際公開第2021/006920 (WO, A1)
- (58)調査した分野 (Int.Cl., DB名)  
G10L 13/00-15/34