(19) Deutsches Patent- und Markenamt

(10) **DE 10 2017 125 475 B4** 2023.05.25

(12) # Patentschrift

(21) Aktenzeichen: **10 2017 125 475.7**
(22) Anmeldetag: **30.10.2017**
(43) Offenlegungstag: **20.09.2018**
(45) Veröffentlichungstag der Patenterteilung: **25.05.2023**

(51) Int Cl.: ***G10L 13/06*** (2013.01)
G10L 13/07 (2013.01)

(30) Unionspriorität:
PCT/GR2017/000012 14.03.2017 GR

(73) Patentinhaber:
GOOGLE LLC, Mountain View, Calif., US

(74) Vertreter:
Betten & Resch Patent- und Rechtsanwälte PartGmbB, 80333 München, DE

(72) Erfinder:
Agiomyrgiannakis, Ioannis, Mountain View, Calif., US

(56) Ermittelter Stand der Technik:

US	**9 240 178**	**B1**
US	**2014 / 0 257 818**	**A1**

(54) Bezeichnung: **Verfahren und System zur Sprachsyntheseeinheitenauswahl**

(57) Hauptanspruch: Nichttransitorisches Computerspeichermedium, das mit Anweisungen codiert ist, die dann, wenn sie durch einen oder mehrere Computer eines Text-zu-Sprache-Systems (116) ausgeführt werden, bewirken, dass der eine oder die mehreren Computer Operationen ausführen, die umfassen:

Empfangen (302) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems von Daten, die Text zur Sprachsynthese angeben;

Bestimmen (304) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems einer Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, wobei die Folge von Texteinheiten wenigstens eine erste Texteinheit gefolgt von einer zweiten Texteinheit enthält;

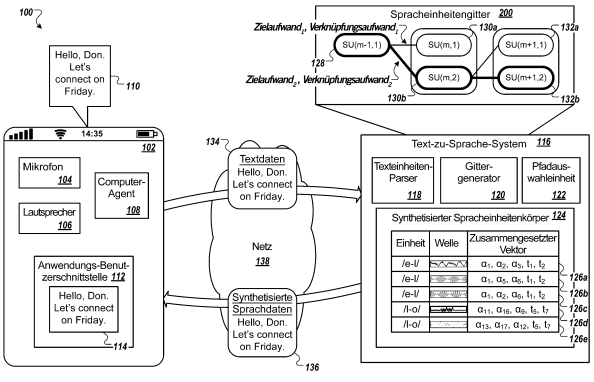
Bestimmen (306) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems mehrerer Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, wobei das Bestimmen der mehreren Pfade von Spracheinheiten umfasst:

Auswählen (308) aus einem Spracheinheitenkörper (124) einer vorbestimmten Menge L erster Spracheinheiten (202a-202f), die Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren; und

Definieren von Pfaden für eine vorbestimmte Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) durch:

Auswählen (310), für jede erste Spracheinheit der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f), einer vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, aus dem Spracheinheitenkörper (124), wobei jede zweite Spracheinheit der vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f) basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit der jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird; und

Definieren (314) von Pfaden von jeder der ersten Spracheinheiten der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) zu jeder zweiten Spracheinheit der jeweiligen vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die in die mehreren ...



Beschreibung

Hintergrund

[0001] Ein Text-zu-Sprache-System kann Textdaten zur hörbaren Präsentation für einen Anwender synthetisieren. Beispielsweise kann das Text-zu-Sprache-System eine Anweisung empfangen, die angibt, dass das Text-zu-Sprache-System Synthesedaten für eine Textnachricht oder eine E-Mail erzeugen sollte. Das Text-zu-Sprache-System kann die Synthesedaten für einen Lautsprecher bereitstellen, um eine hörbare Präsentation des Inhalts aus der Textnachricht oder E-Mail für einen Anwender zu bewirken.

[0002] US 9,240,178 B1 offenbart ein Text-to-Speech (TTS)-System, das mit mehreren Sprachkörpern konfiguriert ist, die zum Synthetisieren von Sprache verwendet werden.

[0003] US 2014/0257818 A1 offenbart Systeme, Verfahren und nicht flüchtige computerlesbare Speichermedien für die Sprachsynthese.

Zusammenfassung

[0004] Die der vorliegenden Erfindung zugrundeliegende technische Aufgabe wird mittels der Gegenstände der unabhängigen Patentansprüche gelöst. Die abhängigen Patentansprüche beschreiben einige beispielhafte Ausführungsformen.

[0005] In einigen Implementierungen synthetisiert ein Text-zu-Sprache-System Audiodaten unter Verwendung eines Einheitenauswahlprozesses. Das Text-zu-Sprache-System kann eine Folge von Spracheinheiten bestimmen und die Spracheinheiten verketten, um synthetisierte Audiodaten zu bilden. Als Teil des Einheitsauswahlprozesses erzeugt das Text-zu-Sprache-System ein Gitter, das mehrere Kandidatenspracheinheiten für jedes phonetische Element, das synthetisiert werden soll, enthält. Das Erzeugen des Gitters beinhaltet Verarbeitung, um die Kandidatenspracheinheiten für das Gitter aus einem großen Körper von Spracheinheiten auszuwählen. Um zu bestimmen, welche Kandidatenspracheinheiten in das Gitter aufgenommen werden sollen, kann das Text-zu-Sprache-System sowohl einen Zielaufwand als auch einen Verknüpfungsaufwand verwenden. Allgemein gibt der Zielaufwand an, wie genau eine spezielle Spracheinheit die phonetische Einheit, die synthetisiert werden soll, repräsentiert. Der Verknüpfungsaufwand kann angeben, wie gut die akustischen Eigenschaften der speziellen Spracheinheit zu einer oder mehreren anderen Spracheinheiten passen, die in dem Gitter repräsentiert sind. Unter Verwendung eines Verknüpfungsaufwands, um die Kandidatenspracheinheiten für das Gitter auszuwählen, kann das Text-zu-Sprache-System ein Gitter erzeugen, das Pfade enthält, die natürlicher klingende synthetisierte Sprache repräsentieren.

[0006] Das Text-zu-Sprache-System kann Spracheinheiten auswählen, die in ein Gitter aufgenommen werden sollen, unter Verwendung eines Abstands zwischen Spracheinheiten, Akustikparametern für andere Spracheinheiten in einem aktuell ausgewählten Pfad, einem Zielaufwand oder einer Kombination aus zwei oder mehr daraus. Beispielsweise kann das Text-zu-Sprache-System Akustikparameter einer oder mehrerer Spracheinheiten in einem aktuell ausgewählten Pfad bestimmen. Das Text-zu-Sprache-System kann die bestimmten Akustikparameter und Akustikparameter für eine Kandidatenspracheinheit verwenden, um einen Verknüpfungsaufwand zu bestimmen, z. B. unter Verwendung einer Abstandsfunktion, um die Kandidatenspracheinheit zu dem aktuell ausgewählten Pfad einer einen oder mehreren Spracheinheiten hinzuzufügen. In einigen Beispielen kann das Text-zu-Sprache-System einen Zielaufwand zum Hinzufügen der Kandidatenspracheinheit zu dem aktuell ausgewählten Pfad unter Verwendung von Linguistikparametern bestimmen. Das Text-zu-Sprache-System kann Linguistikparameter einer Texteinheit, für die die Kandidatenspracheinheit Sprachsynthesedaten enthält, bestimmen und kann Linguistikparameter der Kandidatenspracheinheit bestimmen. Das Text-zu-Sprache-System kann einen Abstand zwischen der Texteinheit und der Kandidatenspracheinheit als einen Zielaufwand unter Verwendung der Linguistikparameter bestimmen. Das Text-zu-Sprache-System kann irgendeine geeignete Abstandsfunktion zwischen Akustikparametervektoren oder Linguistikparametervektoren, die Spracheinheiten repräsentieren, verwenden. Einige Beispiele für Abstandsfunktionen enthalten wahrscheinlichkeitstheoretische Funktionen, Funktionen mit mittlerem quadriertem Fehler und Lp-Norm-Funktionen.

[0007] Das Text-zu-Sprache-System kann einen Gesamtaufwand eines Pfads, z. B. des aktuell ausgewählten Pfads und anderer Pfad mit unterschiedlichen Spracheinheiten, als eine Kombination der Aufwände für die Spracheinheiten in dem jeweiligen Pfad bestimmen. Das Text-zu-Sprache-System kann die Gesamtaufwände mehrerer unterschiedlicher Pfade vergleichen, um einen Pfad mit einem optimalen Aufwand zu bestimmen, z. B. einen Gesamtpfad mit einem niedrigsten Aufwand oder einem höchsten Aufwand. In einigen Beispielen können die Gesamtaufwände die Verknüpfungsaufwände oder eine Kombination aus dem Verknüpfungsaufwände und dem Zielaufwand sein. Das Text-zu-Sprache-System kann den Pfad mit dem optimalen Aufwand auswählen und die Einheiten aus dem Pfad mit optimalem Aufwand verwenden, um synthetisierte Sprache zu erzeugen. Das Text-zu-Sprache-System kann die synthetisierte Sprache zur Ausgabe bereitstellen, z.

B. durch Bereitstellen von Daten für die synthetisierte Sprache für eine Anwendervorrichtung oder Präsentieren der synthetisierten Sprache auf einem Lautsprecher.

[0008] Das Text-zu-Sprache-System kann einen sehr großen Körper von Spracheinheiten aufweisen, die zur Sprachsynthese verwendet werden können. Ein sehr großer Körper von Spracheinheiten kann Daten für mehr als dreißig Stunden von Spracheinheiten oder in einigen Implementierungen Daten für mehr als Hunderte von Stunden von Spracheinheiten enthalten. Einige Beispiele von Spracheinheiten enthalten Diphone, Phone und irgendeinen Typ linguistischer Atome, wie z. B. Worte, Audioblöcke oder eine Kombination aus zwei oder mehr davon. Die linguistischen Atome, die Audioblöcke oder beides können von fester oder variabler Größe sein. Ein Beispiel für einen Audioblock fester Größe ist ein Audiorahmen von fünf Millisekunden.

[0009] Im Allgemeinen kann ein innovativer Aspekt des in dieser Spezifikation beschriebenen Gegenstands in Verfahren verwirklicht sein, die die Aktionen zum Empfangen durch einen oder mehrere Computer eines Text-zu-Sprache-Systems von Daten, die Text zur Sprachsynthese angeben; Bestimmen durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems einer Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, wobei die Folge von Texteinheiten wenigstens eine erste Texteinheit gefolgt von einer zweiten Texteinheit enthält; Bestimmen durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems mehrerer Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, wobei das Bestimmen der mehreren Pfade von Spracheinheiten enthält: Auswählen aus einem Spracheinheitenkörper einer ersten Spracheinheit, die Sprachsynthesedaten enthält, die die erste Texteinheit repräsentieren; Auswählen aus dem Spracheinheitenkörper mehrerer zweiter Spracheinheiten, die Sprachsynthesedaten enthalten, die die zweite Texteinheit repräsentieren, wobei jede aus den mehreren zweiten Spracheinheiten basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit einer ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird; und Definieren von Pfaden von der ausgewählten ersten Spracheinheit zu jeder aus den mehreren zweiten Spracheinheiten, die in die mehreren Pfade von Spracheinheiten aufgenommen werden sollen; und Bereitstellen durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems synthetisierter Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist, enthalten. Andere Ausführungsformen dieses Aspekts enthalten entsprechende Computersysteme, Einrichtungen und Computerprogramme, die auf einer oder mehreren Computerspeichervorrichtungen gespeichert sind, von denen jedes konfiguriert ist, die Aktionen der Verfahren auszuführen. Ein System aus einem oder mehreren Computern kann konfiguriert sein, spezielle Operationen oder Aktionen aufgrund dessen, dass es Software, Firmware oder Hardware oder irgendeiner Kombination daraus auf dem System installiert aufweist, auszuführen, die im Betrieb bewirkt oder bewirken, dass das System die Aktionen ausführt. Ein oder mehrere Computerprogramme können konfiguriert sein, spezielle Operationen oder Aktionen aufgrund dessen auszuführen, dass sie Anweisungen enthalten, die dann, wenn sie durch eine Datenverarbeitungseinrichtung ausführt werden, bewirken, dass die Einrichtung die Aktionen ausführt.

[0010] Die vorstehende und andere Ausführungsformen können jeweils optional eines oder mehrere aus den folgenden Merkmalen allein oder in Kombination enthalten. Bestimmen der Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, kann Bestimmen der Folge von Texteinheiten enthalten, die jeweils einen unterscheidbaren Abschnitt des Texts enthalten, getrennt von den Abschnitten des Texts, die durch die anderen Texteinheiten repräsentiert sind. Bereitstellen der synthetisierten Sprachdaten gemäß dem Pfad, der aus den mehreren Pfaden ausgewählt ist, kann Bereitstellen der synthetisierten Sprachdaten, um zu bewirken, dass eine Vorrichtung hörbare Daten für den Text erzeugt, enthalten.

[0011] In einigen Implementierungen kann das Verfahren Auswählen aus dem Spracheinheitenkörper von zwei oder mehreren Anfangsspracheinheiten enthalten, die jeweils Sprachsynthesedaten enthalten, die eine Anfangstexteinheit in der Folge von Texteinheiten mit einem Ort an einem Anfang der Textfolge repräsentieren. Auswählen der zwei oder mehr Anfangsspracheinheiten kann Auswählen einer vorbestimmten Anzahl von Anfangsspracheinheiten enthalten. Bestimmen der mehreren Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, kann Bestimmen der vorbestimmten Anzahl von Pfaden enthalten. Das Verfahren kann Auswählen aus der vorbestimmten Anzahl von Pfaden des Pfads, für den die synthetisierten Sprachdaten bereitgestellt werden sollen, enthalten. Die mehreren zweiten Spracheinheiten können zwei oder mehr zweite Spracheinheiten enthalten. Das Definieren von Pfaden von der ausgewählten ersten Spracheinheit zu jeder aus den mehreren zweiten Spracheinheiten kann Bestimmen für eine weitere erste Spracheinheit, die Sprachsynthesedaten enthält, die die erste Texteinheit repräsentieren, nicht irgendwelche zusätzlichen Spracheinheiten zu einem Pfad hinzuzufügen, der die andere erste Spracheinheit enthält, enthalten. Das Verfah-

ren kann Auswählen für die erste Texteinheit der vorbestimmten Anzahl erster Spracheinheiten, die jeweils Sprachsynthesedaten enthalten, die die erste Texteinheit repräsentieren; und Auswählen für die zweite Texteinheit der vorbestimmten Anzahl zweiter Spracheinheiten, die jeweils Sprachsynthesedaten enthalten, die die zweite Texteinheit repräsentieren, enthalten, wobei jede aus der vorbestimmten Anzahl zweiter Spracheinheiten basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit einer jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird.

[0012] In einigen Implementierungen kann das Verfahren Bestimmen für eine zweite vorbestimmte Anzahl von zweiten Spracheinheiten, die jeweils Sprachsynthesedaten enthalten, die die zweite Einheit repräsentieren, (i) eines Verknüpfungsaufwands, um die zweite Spracheinheit mit einer jeweiligen ersten Spracheinheit zu verketten, und (ii) eines Zielaufwands, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, enthalten. Die zweite vorbestimmte Anzahl kann größer sein als die vorbestimmte Anzahl. Das Auswählen der vorbestimmten Anzahl zweiter Spracheinheiten kann Auswählen der vorbestimmten Anzahl zweiter Spracheinheiten aus der zweiten vorbestimmten Anzahl zweiter Spracheinheiten unter Verwendung der bestimmten Verknüpfungsaufwände und der bestimmten Zielaufwände enthalten. Die erste Texteinheit kann einen ersten Ort in der Folge von Texteinheiten aufweisen. Die zweite Texteinheit kann einen zweiten Ort in der Folge von Texteinheiten aufweisen, der dem ersten Ort ohne irgendwelche dazwischenliegenden Orte nachfolgt. Das Auswählen von mehreren zweiten Spracheinheiten aus dem Spracheinheitenkörper kann Auswählen der mehreren zweiten Spracheinheiten aus dem Spracheinheitenkörper unter Verwendung (i) eines Verknüpfungsaufwands, um die zweite Spracheinheit mit Daten für die ersten Spracheinheit und einer entsprechenden Anfangsspracheinheit aus den zwei oder mehr Anfangsspracheinheiten zu verketten, und (ii) eines Zielaufwands, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, enthalten. Das Verfahren kann Bestimmen eines Pfads, der eine ausgewählte Spracheinheit enthält, für jede aus den Texteinheiten in der Folge von Texteinheiten bis zu dem ersten Ort, wobei die ausgewählten Spracheinheiten die erste Spracheinheit und die entsprechende Anfangsspracheinheit enthalten; Bestimmen erster Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad; und Bestimmen für jede aus den mehreren zweiten Spracheinheiten des Verknüpfungsaufwands unter Verwendung der ersten Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad und der zweiten Akustikparameter für die zweite Spracheinheit enthalten. Das Bestimmen für jede aus den mehreren zweiten Spracheinheiten des Verknüpfungsaufwands kann gleichzeitiges Bestimmen für jede aus zwei oder mehr zweiten Spracheinheiten des Verknüpfungsaufwands unter Verwendung der ersten Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad und der zweiten Akustikparameter für die zweite Spracheinheit enthalten.

[0013] Der in dieser Spezifikation beschriebene Gegenstand kann in verschiedenen Ausführungsformen implementiert sein und kann zu einem oder mehreren der folgenden Vorteile führen. In einigen Implementierungen kann ein Text-zu-Sprache-System lokale Minima oder lokale Maxima beim Bestimmen eines Pfads, der Spracheinheiten zur Sprachsynthese von Text identifiziert, überwinden. In einigen Implementierungen verbessert das Bestimmen eines Pfads unter gemeinsamer Verwendung sowohl eines Zielaufwands als auch eines Verknüpfungsaufwands die Ergebnisse eines Text-zu–Sprache-Prozesses, z. B. um ein leichter verständliches oder natürlicher klingendes Text-zu–Sprache-Ergebnis zu bestimmen verglichen mit Systemen, die Vorauswahl oder Gitteraufbau nur unter Verwendung des Zielaufwands ausführen. Beispielsweise kann in einigen Fällen eine spezielle Spracheinheit mit einem gewünschten phonetischen Element gut übereinstimmen, z. B. einen niedrigen Zielaufwand aufweisen, jedoch mit anderen Einheiten in dem Gitter schlecht zusammenpassen, z. B. einen hohen Verknüpfungsaufwand aufweisen. Systeme, die Verknüpfungsaufwände nicht berücksichtigen, wenn sie ein Gitter aufbauen, können durch den Zielaufwand übermäßig beeinflusst werden und die spezielle Einheit zum Nachteil der Gesamtqualität der Äußerung aufnehmen. Mit den hier offenbarten Techniken kann das Verwenden von Verknüpfungsaufwänden, um das Gitter aufzubauen, Besetzen des Gitters mit Spracheinheiten, die den Zielaufwand auf Kosten der Gesamtqualität minimieren, verhindern. Mit anderen Worten kann das System den Beitrag der Verknüpfungsaufwände und der Zielaufwände ausgleichen, wenn es jede Einheit auswählt, die in das Gitter aufgenommen werden soll, um Einheiten hinzuzufügen, die nicht die besten Übereinstimmungen für einzelne Einheiten sein können, jedoch zusammen arbeiten, um eine bessere Gesamtqualität der Synthese bereitzustellen, z. B. einen niedrigeren Gesamtaufwand.

[0014] In einigen Implementierungen kann die Qualität einer Text-zu-Sprache-Ausgabe durch Aufbauen eines Gitters unter Verwendung eines Verknüpfungsaufwands, der Akustikparameter für alle Spracheinheiten in einem Pfad durch das Gitter verwendet, verbessert werden. Einige Implementierungen der vorliegenden Techniken bestimmen einen Verknüpfungsaufwand für das Hinzufügen einer aktuellen

Einheit nach der unmittelbar vorhergehenden Einheit. Zusätzlich oder als eine Alternative bauen einige Implementierungen ein Gitter auf unter Verwendung von Verknüpfungsaufwänden, die repräsentieren, wie gut eine hinzugefügte Einheit zu mehreren Einheiten in dem Pfad durch das Gitter passt. Beispielsweise kann ein Verknüpfungsaufwand, der verwendet wird, um Einheiten für das Gitter auszuwählen, die Eigenschaften eines gesamten Pfads von einer Spracheinheit in dem Gitter, die den Anfang der Äußerung repräsentiert, bis zu dem Punkt in dem Gitter, wo die neue Einheit hinzugefügt wird, berücksichtigen. Das System kann bestimmen, ob eine Einheit zu der gesamten Folge von Einheiten passt, und kann das Ergebnis des Viterbi-Algorithmus für den Pfad verwenden, um eine Einheit auszuwählen, die in das Gitter aufgenommen werden soll. Auf diese Weise kann die Auswahl von Einheiten, die in das Gitter aufgenommen werden sollen, von der Viterbi-Suchanalyse abhängen. Zusätzlich kann das System Einheiten zu dem Gitter hinzufügen, um mehrere unterschiedliche Pfade fortzusetzen, die mit derselben oder unterschiedlichen Einheiten in dem Gitter beginnen können. Das erhält die Verschiedenartigkeit von Pfaden durch das Gitter und kann dazu beitragen, lokale Minima oder lokale Maxima zu vermeiden, die andernfalls die Qualität der Synthese für die Äußerung als Ganzes beeinträchtigen könnten.

[0015] In einigen Implementierungen können die Systeme und Verfahren, die nachstehend beschrieben sind und die ein Gitter mit einem Zielaufwand und einem Verknüpfungsaufwand gemeinsam erzeugen, bessere Sprachsyntheseergebnisse erzeugen als andere Systeme mit einem großen Körper synthetisierter Sprachdaten, z. B. mehr als dreißig oder Hunderte von Stunden von Sprachdaten. In vielen Systemen wird die Qualität von Text-zu-Sprache-Ausgabe gesättigt, wenn die Größe des Körpers von Spracheinheiten zunimmt. Viele Systeme sind nicht fähig, die Beziehungen unter der Akustik von Spracheinheiten während der Vorauswahl oder der Gitteraufbauphase zu berücksichtigen, und sind deshalb nicht fähig, die große Menge verfügbarer Spracheinheiten vollständig auszunutzen. Mit den vorliegenden Techniken kann das Text-zu-Sprache-System die Verknüpfungsaufwände und die Akustikeigenschaften von Spracheinheiten berücksichtigen, wenn das Gitter konstruiert wird, was eine feiner granulare Auswahl ermöglicht, die Folgen von Einheiten aufbaut, die eine natürlicher klingende Sprache repräsentieren.

[0016] In einigen Implementierungen können die nachstehend beschriebenen Systeme und Verfahren die Qualität von Text-zu-Sprache-Synthese erhöhen, während sie die Berechnungskomplexität und andere Hardware-Anforderungen begrenzen. Beispielsweise kann das Text-zu-Sprache-System eine vorbestimmte Anzahl von Pfaden auswählen, die Folgen von Spracheinheiten identifizieren, und eine Grenze für eine Gesamtzahl von zu irgendeiner Zeit analysierten Pfaden und eine Speichermenge, die erforderlich ist, um die Daten für diese Pfade zu speichern, setzen. In einigen Implementierungen rufen die nachstehend beschriebenen Systeme und Verfahren im Voraus aufgezeichnete Äußerungen oder Teile von Äußerungen aus einem Körper von Spracheinheiten wieder auf, um die Qualität der Erzeugung synthetisierter Sprache in einer eingeschränkten Textdomäne zu verbessern. Beispielsweise kann ein Text-zu-Sprache-System die im Voraus aufgezeichneten Äußerungen oder Teile von Äußerungen wieder aufrufen, um eine maximale Qualität zu erreichen, wann immer die Textdomäne eingeschränkt ist, z. B. in GPS-Navigationsanwendungen.

[0017] Die Einzelheiten einer oder mehrerer Implementierungen des in dieser Spezifikation beschriebenen Gegenstands sind in den begleitenden Zeichnungen und der nachstehenden Beschreibung dargelegt. Andere Merkmale, Aspekte und Vorteile des Gegenstands werden aus der Beschreibung, den Zeichnungen und den Ansprüchen offensichtlich.

Figurenliste

Fig. 1 ist ein Beispiel einer Umgebung, in der eine Anwendervorrichtung Sprachsynthesedaten von einem Text-zu-Sprache-System anfordert.

Fig. 2 ist ein Beispiel eines Spracheinheitengitters.

Fig. 3 ist ein Ablaufdiagramm eines Prozesses zum Bereitstellen synthetisierter Sprachdaten.

Fig. 4 ist ein Blockdiagramm eines Berechnungssystems, das zusammen mit in diesem Dokument beschriebenen computerimplementierten Verfahren verwendet werden kann.

[0018] Gleiche Bezugszeichen und Bezeichnungen in den verschiedenen Zeichnungen geben gleiche Elemente an.

Ausführliche Beschreibung

[0019] **Fig. 1** ist ein Beispiel einer Umgebung 100, in der eine Anwendervorrichtung 102 Sprachsynthesedaten von einem Text-zu-Sprache-System 116 anfordert. Die Anwendervorrichtung 102 kann die Sprachsynthesedaten anfordern, so dass die Anwendervorrichtung 102 eine hörbare Präsentation des Textinhalts, wie z. B. einer E-Mail, einer Textnachricht, einer Nachricht, die durch einen digitalen Assistenten bereitgestellt werden soll, einer Kommunikation von einer Anwendung oder eines anderen Inhalts, erzeugen kann. In **Fig. 1** ist das Text-zu-

Sprache-System 116 von der Anwendervorrichtung 10ß2 getrennt. In einigen Implementierungen ist das Text-zu-Sprache-System 116 in der Anwendervorrichtung 102 enthalten, z. B. auf der Anwendervorrichtung 102 implementiert.

[0020] Die Anwendervorrichtung 102 kann bestimmen, Textinhalt hörbar zu präsentieren, z. B. für einen Anwender. Beispielsweise kann die Anwendervorrichtung 102 einen computerimplementierten Agenten 108 enthalten, der bestimmt, Textinhalt hörbar zu präsentieren. Der computerimplementierte Agent 108 kann einen Anwender darauf hinweisen, dass „eine ungelesene Textnachricht für dich vorhanden ist“. Der computerimplementierte Agent 108 kann Daten für einen Lautsprecher 106 bereitstellen, um die Präsentation des Hinweises zu bewirken. In Reaktion darauf kann der computerimplementierte Agent 108 ein Audiosignal von einem Mikrofon 104 empfangen. Der computerimplementierte Agent 108 analysiert das Audiosignal, um eine oder mehrere Äußerungen zu bestimmen, die in dem Audiosignal enthalten sind, und ob irgendeine dieser Äußerungen ein Befehl ist. Beispielsweise kann der computerimplementierte Agent 108 bestimmen, dass das Audiosignal eine Äußerung „Lies mit die Textnachricht vor“ enthält.

[0021] Der computerimplementierte Agent 108 ruft Textdaten, z. B. für die Textnachricht, aus einem Speicher ab. Beispielsweise kann der computerimplementierte Agent 108 eine Nachricht zu einer Textnachrichtenanwendung senden, die die Daten für die Textnachricht anfordert. Die Textnachrichtenanwendung kann die Daten für die Textnachricht aus einem Speicher abrufen und die Daten für den computerimplementierten Agenten 108 bereitstellen. In einigen Beispielen kann die Textnachrichtenanwendung den computerimplementierte Agenten 108 mit einem Bezeichner versorgen, der einen Speicherort angibt, an dem die Daten für die Textnachricht gespeichert sind.

[0022] Der computerimplementierte Agent 108 stellt die Daten für den Text, z. B. die Textnachricht, in einer Kommunikation 134 für das Text-zu-Sprache-System 116 bereit. Beispielsweise ruft der computerimplementierte Agent 108 die Daten für den Text „Hello, Don. Let's connect on Friday“ („Hallo Don. Lass uns am Freitag zusammenkommen“) aus einem Speicher ab und erzeugt die Kommunikation 134 unter Verwendung der abgerufenen Daten. Der computerimplementierte Agent 108 stellt die Kommunikation 134 für das Text-zu-Sprache-System 116 bereit, z. B. unter Verwendung eines Netzes 138.

[0023] Das Text-zu-Sprache-System 116 stellt wenigstens einige der Daten aus der Kommunikation 134 für einen Texteinheiten-Parser 118 bereit. Beispielsweise stellt das Text-zu-Sprache-System 116 Daten für den gesamten Text für „Hello, Don. Let's connect on Friday“ für den Texteinheiten-Parser 118 bereit. In einigen Beispielen kann das Text-zu-Sprache-System 116 Daten für einigen, jedoch nicht den gesamten, Text für den Texteinheiten-Parser 118 bereitstellen, z. B. abhängig von einer Größe des Texts, den der Texteinheiten-Parser 118 analysieren wird.

[0024] Der Texteinheiten-Parser 118 erzeugt eine Folge von Texteinheiten für Textdaten. Die Texteinheiten können irgendein geeigneter Typ von Texteinheiten wie z. B. Diphone, Phone, irgendein Typ eines linguistischen Atoms, z. Worte oder Audio-Blöcke oder eine Kombination aus zwei oder mehr davon sein. Beispielsweise erzeugt der Texteinheiten-Parser eine Folge von Texteinheiten für die Textnachricht. Ein Beispiel für eine Folge von Texteinheiten für das Wort „Hello“ enthält drei Texteinheiten: „h-e“, „e-l“ und „l-o“.

[0025] Die Folge von Texteinheiten kann einen Abschnitt eines Worts, ein Wort, eine Phrase, z. B. zwei oder mehr Worte, einen Abschnitt eines Satzes, einen Satz, mehrere Sätze, einen Absatz oder eine andere geeignete Textgröße repräsentieren. Der Texteinheiten-Parser 118 oder eine andere Komponente des Text-zu-Sprache-Systems 116 kann den Text für die Folge von Texteinheiten unter Verwendung eines oder mehrerer aus einer Verzögerung für die Präsentation von hörbarem Inhalt, einer gewünschten Wahrscheinlichkeit dafür, wie gut synthetisierte Sprache natürlich artikulierte Sprache repräsentiert, oder beidem auswählen. Beispielsweise kann das Text-zu-Sprache-System 116 eine Größe des Texts bestimmen, der für den Texteinheiten-Parser 118 bereitgestellt werden soll, unter Verwendung einer Verzögerung für die Präsentation von hörbarem Inhalt, z. B. so dass kleinere Textgrößen eine Verzögerung von der Zeit, zu der der computerimplementierte Agent 108 bestimmt, hörbaren Inhalt zu präsentieren, bis zu der Zeit, wenn der hörbare Inhalt auf dem Lautsprecher 106 präsentiert wird, reduziert, und stellt den Text für den Texteinheiten-Parser 118 bereit, um zu bewirken, dass der Texteinheiten-Parser 118 eine entsprechende Folge von Texteinheiten erzeugt.

[0026] Der Texteinheiten-Parser 118 stellt die Folge von Texteinheiten für einen Gittergenerator 120 bereit, der Spracheinheiten, die Sprachsynthesedaten enthalten, die entsprechende Texteinheiten aus einer Folge von Texteinheiten entsprechen, aus einem synthetisierten Spracheinheitenkörper 124 auswählt. Beispielsweise kann der synthetisierte Spracheinheitenkörper 124 eine Datenbank sein, die mehrere Einträge 126a-e enthält, die jeweils Daten für eine Spracheinheit enthalten. Der synthetisierte Spracheinheitenkörper 124 kann Daten für mehr als dreißig Stunden von Spracheinheiten ent-

halten. In einigen Beispielen kann der synthetisierte Spracheinheitenkörper 124 Daten für mehr als Hunderte Stunden von Spracheinheiten enthalten.

[0027] Jeder aus den Einträgen 126-e für eine Spracheinheit identifiziert eine Texteinheit, der der Eintrag entspricht. Beispielsweise kann ein erster, zweiter und dritter Eintrag 126a-c jeweils eine Texteinheit „/e-l/“ identifizieren, und ein vierter und fünfter Eintrag 126d-e kann jeweils eine Texteinheit „/l-o/“ identifizieren.

[0028] Jeder aus den Einträgen 126a-e für eine Spracheinheit identifiziert Daten für eine Wellenform zur hörbaren Präsentation der jeweiligen Texteinheit. Ein System, z. B. die Anwendervorrichtung 102, kann die Wellenform in Kombination mit anderen Wellenformen für andere Texteinheiten verwenden, um eine hörbare Präsentation des Texts, z. B. der Textnachricht, zu erzeugen. Ein Eintrag kann Daten für die Wellenform, z. B. Audiodaten, enthalten. Ein Eintrag kann einen Bezeichner enthalten, der einen Ort angibt, an dem die Wellenform gespeichert ist, z. B. in dem Text-zu-Sprache-System 116 oder in einem anderen System.

[0029] Die Einträge 126a-e für Spracheinheiten enthalten Daten, die mehrere Parameter der Wellenform, die durch den jeweiligen Eintrag identifiziert ist, angeben. Beispielsweise kann jeder der Einträge 126a-e Akustikparameter, Linguistikparameter oder beides für die entsprechende Wellenform enthalten. Der Gittergenerator 120 verwendet die Parameter für einen Eintrag, um zu bestimmen, ob der Eintrag als Kandidatenspracheinheit für eine entsprechende Texteinheit ausgewählt werden soll, wie nachstehend genauer beschrieben ist.

[0030] Akustikparameter können den Klang der entsprechenden Wellenform für die Spracheinheit repräsentieren. In einigen Beispielen können sich die Akustikparameter auf eine tatsächliche Realisierung der Wellenform beziehen und können von der Wellenform für die Spracheinheit abgeleitet werden. Beispielsweise können die Akustikparameter Informationen über die tatsächliche Nachricht, die in dem Text geführt ist, z. B. Informationen über die Identität der gesprochenen Phoneme, transportieren. Akustikparameter können den Abstand, die Grundfrequenz, Spektrumsinformationen und/oder Informationen über die Spektrumseinhüllende enthalten, die in Repräsentationen parametrisiert sein können, wie z. B. Mel-Frequenz-Koeffizienten, Intonation, Dauer, Spracheinheitkontext oder eine Kombination aus zwei oder mehr daraus. Ein Spracheinheitkontext kann andere Spracheinheiten angeben, die zu der Wellenform benachbart waren, z. B. davor oder danach oder beides, als die Wellenform erzeugt wurde. Die Akustikparameter können eine Emotion repräsentieren, die in der Wellenform ausgedrückt ist z. B. glücklich, nicht glücklich, traurig, nicht traurig, unglücklich oder eine Kombination aus zwei oder mehr davon. Die Akustikparameter können einen Stress, der in der Wellenform enthalten ist, repräsentieren, z. B. gestresst, nicht gestresst oder beides. Die Akustikparameter können eine Geschwindigkeit angeben, in der die Sprache, die in einer Wellenform enthalten ist, gesprochen wurde. Der Gittergenerator 120 kann mehrere Spracheinheiten mit der gleichen oder einer ähnlichen Geschwindigkeit auswählen, so dass sie den Texteinheiten in einer Folge von Texteinheiten entsprechen, z. B. so dass die synthetisierte Sprache natürlicher ist. Die Akustikparameter können angeben, ob die Wellenform eine Betonung enthält. In einigen Beispielen können die Akustikparameter angeben, ob die Wellenform geeignet ist, einen Text zu synthetisieren, der eine Frage ist. Beispielsweise kann der Gittergenerator 120 bestimmen, dass eine Folge von Texteinheiten eine Frage repräsentiert, z. B. für einen Anwender der Anwendervorrichtung 102, und eine Spracheinheit aus dem synthetisierten Spracheinheitenkörper 124 mit Akustikparametern auswählen, die angeben, dass die Spracheinheit eine geeignete Intonation zum Synthetisieren einer hörbaren Frage aufweist, z. B. ein ansteigender Tonfall. Die Akustikparameter können angeben, ob die Wellenform geeignet ist, einen Text zu synthetisieren, der ein Ausruf ist.

[0031] Linguistikparameter können Daten repräsentieren, die von Text abgeleitet sind, dem eine Einheit, z. B. eine Texteinheit oder eine Spracheinheit, entspricht. Der entsprechende Text kann ein Wort, eine Phrase, ein Satz, ein Absatz oder Teil eines Worts sein. In einigen Beispielen kann ein System Linguistikparameter aus dem Text ableiten, der gesprochen wurde, um die Wellenform für die Spracheinheit zu erzeugen. In einigen Implementierungen kann ein System Linguistikparameter für Text durch Inferenz bestimmen. Beispielsweise kann ein System Linguistikparameter für eine Spracheinheit aus einem Phonem oder einer Hidden Markov-Modell-Repräsentation des Texts, der die Spracheinheit enthält, ableiten. In einigen Beispielen kann ein System Linguistikparameter für eine Spracheinheit unter Verwendung eines neuronalen Netzes ableiten, z. B. unter Verwendung eines überwachten, halbüberwachten oder nicht überwachten Prozesses. Linguistikparameter können Stress, Prosodie, ob eine Texteinheit Teil einer Frage ist, ob eine Texteinheit Teil eines Ausrufs ist oder eine Kombination aus zwei oder mehr daraus enthalten. In einigen Beispielen können einige Parameter sowohl Akustikparameter als auch Linguistikparameter sein, wie z. B. Stress, ob eine Texteinheit Teil einer Frage ist, ob eine Texteinheit Teil eines Ausrufs ist oder zwei oder mehr daraus.

[0032] In einigen Implementierungen kann ein System einen oder mehrere Akustikparameter, einen oder mehrere Linguistikparameter oder eine Kombination aus beiden für eine Wellenform und entsprechende Spracheinheit unter Verwendung von Daten aus einem Wellenformanalysesystem, z. B. einem Wellenformanalysesystem mit künstlicher Intelligenz, unter Verwendung von Anwendereingabe oder beidem bestimmen. Beispielsweise kann ein Audiosignal ein Flag aufweisen, das angibt, dass der Inhalt, der in dem Audiosignal codiert ist, „glücklich“ ist. Das System kann mehrere Wellenformen für unterschiedliche Texteinheiten in dem Audiosignal erzeugen, z. B. durch Segmentieren des Audiosignals in mehrere Wellenformen, und jede der Spracheinheiten für die Wellenformen einem Parameter zuordnen, der angibt, dass die Spracheinheit synthetisierte Sprache mit einem glücklichen Klang enthält.

[0033] Der Gittergenerator 120 erzeugt ein Spracheinheitengitter 200, das nachstehend genauer beschrieben ist, durch Auswählen mehrerer Spracheinheiten für jede Texteinheit in der Folge von Texteinheiten unter Verwendung eines Verknüpfungsaufwands, eines Zielaufwand oder beidem für jede aus den mehreren Spracheinheiten. Beispielsweise kann der Gittergenerator 120 eine erste Spracheinheit, die die erste Texteinheit in der Folge von Texteinheiten, z. B. „h-e“, repräsentiert, unter Verwendung eines Zielaufwands auswählen. Der Gittergenerator 120 kann zusätzliche Spracheinheiten, wie z. B. eine zweite Spracheinheit, die eine zweite Texteinheit repräsentiert, z. B. „e-l“, und eine dritte Spracheinheit, die eine dritte Texteinheit repräsentiert, z. B. „l-o“, unter Verwendung von sowohl eines Zielaufwands als auch eines Verknüpfungsaufwands für jede aus den zusätzlichen Spracheinheiten auswählen.

[0034] Das Spracheinheitengitter 200 enthält mehrere Pfade durch das Spracheinheitengitter 200, die jeweils nur eine Spracheinheit für jede entsprechende Texteinheit in einer Folge von Texteinheiten enthalten. Ein Pfad identifiziert eine Folge von Spracheinheiten, die die Folge von Texteinheiten repräsentiert. Ein Beispielpfad enthält die Spracheinheiten 128, 130b und 132a, und ein weiterer Beispielpfad enthält die Spracheinheiten 128, 130b und 132b.

[0035] Jede aus den Spracheinheiten, die in dem Pfad identifiziert sind, kann einer einzigen Texteinheit an einem einzigen Ort in der Folge von Texteinheiten entsprechen. Beispielsweise kann mit der Folge von Texteinheiten „Hello, Don. Let's connect on Friday“ die Folge von Texteinheiten „D-o“, „o-n“, „l-e“, „t-s“, „c-o“, „n-e“, „c-t“ und „o-n“ unter anderen Texteinheiten enthalten. Der Gittergenerator 120 wählt eine Spracheinheit für jede dieser Texteinheiten aus. Obwohl der Pfad zwei Instanzen von „o-n“ - eine erste für das Dort „Don“ und eine zweite für das Wort „on“ - enthält, wird der Pfad zwei Texteinheiten identifizieren, eine für jede Instanz der Texteinheit „o-n“. Der Pfad kann die gleiche Spracheinheit für jede aus den zwei Texteinheiten „o-n“ identifizieren oder kann unterschiedliche Spracheinheiten identifizieren, z. B. abhängig von dem Zielaufwand, dem Verknüpfungsaufwand oder beiden für Spracheinheiten, die diesen Texteinheiten entsprechen.

[0036] Eine Anzahl von Spracheinheiten in einem Pfad ist kleiner als eine oder gleich einer Anzahl von Spracheinheiten in der Folge von Texteinheiten. Beispielsweise wenn der Gittergenerator 120 einen Pfad nicht fertiggestellt hat, enthält der Pfad weniger Spracheinheiten als die Anzahl von Texteinheiten in der Folge von Texteinheiten. Wenn der Gittergenerator 120 einen Pfad fertiggestellt hat, enthält dieser Pfad eine Spracheinheit für jede Texteinheit in der Folge von Texteinheiten.

[0037] Ein Zielaufwand für eine Spracheinheit gibt einen Grad dafür an, dass die Spracheinheit einer Texteinheit in einer Folge von Texteinheiten entspricht, z. B. beschreibt, wie gut die Wellenform für die Texteinheit die beabsichtigte Nachricht des Texts transportiert. Der Gittergenerator 120 kann einen Zielaufwand für eine Spracheinheit unter Verwendung der Linguistikparameter der Kandidatenspracheinheit und der Linguistikparameter der Zieltexteinheit bestimmen. Beispielsweise gibt ein Zielaufwand für die dritte Spracheinheit einen Grad dafür an, dass die dritte Spracheinheit der dritten Texteinheit, z. B. „l-o“, entspricht. Der Gittergenerator 120 kann einen Zielaufwand als einen Abstand zwischen den Linguistikparametern einer Kandidatenspracheinheit und den Linguistikparametern der Zieltexteinheit bestimmen. Der Gittergenerator 120 kann eine Abstandsfunktionen wie z. B. eine wahrscheinlichkeitstheoretische Funktion, einen mittleren quadratischen Fehler oder eine Lp-Norm verwenden.

[0038] Ein Verknüpfungsaufwand gibt einen Aufwand dafür an, eine Spracheinheit mit einer oder mehreren anderen Spracheinheiten in einem Pfad zu verketten. Beispielsweise beschreibt ein Verknüpfungsaufwand, wie gut sich eine Wellenform, z. B. eine synthetisierte Äußerung, als natürlich artikulierte Sprache verhält bei einer gegebenen Verkettung der Wellenform für eine Spracheinheit mit anderen Wellenformen für die anderen Spracheinheiten, die in einem Pfad sind. Der Gittergenerator 120 kann einen Verknüpfungsaufwand für eine Kandidatenspracheinheit unter Verwendung der Akustikparameter für die Spracheinheit und der Akustikparameter für eine oder mehrere Spracheinheiten in dem Pfad, zu dem erwogen wird, die Kandidatenspracheinheit hinzuzufügen, bestimmen. Beispielsweise kann der Verknüpfungsaufwand zum Hinzufügen der dritten Spracheinheit 132b zu einem Pfad, der

eine erste Spracheinheit 128 und eine zweite Spracheinheit 130b enthält, den Aufwand zum Kombinieren der dritten Spracheinheit 132b mit der zweiten Spracheinheit 130b, z. B. wie gut diese Kombination wahrscheinlich eine natürlich artikulierte Sprache repräsentiert, repräsentieren, oder kann den Aufwand zum Kombinieren der dritten Spracheinheit 132b mit der Kombination aus der ersten Spracheinheit 128 und der zweiten Spracheinheit 130b angeben. Der Gittergenerator 120 kann einen Verknüpfungsaufwand als einen Abstand zwischen den Akustikparametern der Kandidatenspracheinheit und der Spracheinheit oder Spracheinheiten in dem Pfad, zu dem erwogen wird, die Kandidatenspracheinheit hinzuzufügen, bestimmen. Der Gittergenerator 120 kann eine wahrscheinlichkeitstheoretische Abstandfunktion, eine Abstandsfunktion mit mittlerem quadratischem Fehler oder eine Lp-Norm-Abstandsfunktion verwenden.

[0039] Der Gittergenerator 120 kann bestimmen, ob ein Zielaufwand, ein Verknüpfungsaufwand oder beide verwendet werden sollen, wenn eine Spracheinheit ausgewählt wird, unter Verwendung eines Typs von Zieldaten, die für den Gittergenerator 120 verfügbar sind. Beispielsweise kann, wenn der Gittergenerator 120 nur Linguistikparameter für eine Zieltexteinheit besitzt, z. B. für eine Anfangstexteinheit in einer Folge von Texteinheiten, der Gittergenerator 120 einen Zielaufwand bestimmen, um eine Spracheinheit zu einem Pfad für die Folge von Texteinheiten hinzuzufügen. Wenn der Gittergenerator 120 sowohl Akustikparameter für eine vorhergehende Spracheinheit und Linguistikparameter für eine Zielspracheinheit besitzt, kann der Gittergenerator 120 sowohl einen Zielaufwand als auch einen Verknüpfungsaufwand zum Hinzufügen einer Kandidatenspracheinheit zu einem Pfad bestimmen.

[0040] Wenn der Gittergenerator 120 sowohl einen Zielaufwand als auch einen Verknüpfungsaufwand während der Analyse, ob eine Kandidatenspracheinheit 130a zu einem Pfad hinzugefügt werden soll, verwendet, kann der Gittergenerator 120 einen zusammengesetzten Vektor aus Parametern für die Kandidatenspracheinheit 130a verwenden, um den Gesamtaufwand zu bestimmen, der eine Kombination aus dem Zielaufwand und dem Verknüpfungsaufwand ist. Beispielsweise kann der Gittergenerator 120 einen zusammengesetzten Zielvektor durch Kombinieren eines Vektors aus Linguistikparametern für eine Zieltexteinheit z. B. Ziel(m), mit einem Vektor aus Akustikparametern für eine Spracheinheit 128 in einem Pfad, zu dem erwogen wird, die Kandidatenspracheinheit hinzuzufügen, bestimmen, z. B. SU (m-1,1). Der Gittergenerator 120 kann die Linguistikparameter für die Zieltexteinheit aus einem Speicher empfangen, z. B. einer Datenbank, die Linguistikparameter für Zieltexteinheiten enthält. Der Gittergenerator 120 kann die Akustikparameter für die Spracheinheit 128 aus dem synthetisierten Spracheinheitenkörper 124 empfangen.

[0041] Der Gittergenerator 120 kann einen zusammengesetzten Vektor für die Kandidatenspracheinheit 130a, z. B. SU(m,1), aus dem synthetisierten Spracheinheitenkörper 124 empfangen. Beispielsweise wenn der Gittergenerator 120 einen zusammengesetzten Vektor für einen ersten Eintrag 126a in dem synthetisierten Spracheinheitenkörper 124 empfängt, enthält der zusammengesetzte Vektor die Akustikparameter α_1, α_2, α_3 und die Linguistikparameter t_1, t_2 unter anderen Parametern für die Kandidatenspracheinheit 130a.

[0042] Der Gittergenerator 120 kann einen Abstand zwischen dem zusammengesetzten Zielvektor und dem zusammengesetzten Vektor für die Kandidatenspracheinheit 130a als einen Gesamtaufwand für die Kandidatenspracheinheit bestimmen. Wenn die Kandidatenspracheinheit 130a SU(m,1) ist, ist der Gesamtaufwand auf die Kandidatenspracheinheit SU(m,1) eine Kombination aus Zielaufwand, und Verknüpfungsaufwand$_1$. Der Zielaufwand kann als ein einziger numerischer, z. B. dezimaler, Wert repräsentiert sein. Der Gittergenerator 120 kann Zielaufwand, und Verknüpfungsaufwand$_1$ getrennt, z. B. parallel, bestimmen und dann die Werte kombinieren, um den Gesamtaufwand zu bestimmen. In einigen Beispiele kann der Gittergenerator 120 den Gesamtaufwand bestimmen, z. B. ohne den Zielaufwand, oder den Verknüpfungsaufwand$_1$ zu bestimmen.

[0043] Der Gittergenerator 120 kann eine weitere Kandidatenspracheinheit 130b, z. B. SU(m,2), bestimmen, um sie für ein potentielles Hinzufügen zu dem Pfad, der die ausgewählte Spracheinheit 128 enthält, z. B. SU(m-1,1), zu analysieren. Der Gittergenerator 120 kann den gleichen zusammengesetzten Zielvektor für die andere Kandidatenspracheinheit 130b verwenden, weil die Zieltexteinheit und die Spracheinheit 128 in dem Pfad, zu dem erwogen wird, die andere Kandidatenspracheinheit 130b hinzuzufügen, gleich sind. Der Gittergenerator 120 kann einen Abstand zwischen dem zusammengesetzten Zielvektor und einem weiteren zusammengesetzten Vektor für die andere Kandidatenspracheinheit 130b bestimmen, um einen Gesamtaufwand für das Hinzufügen der anderen Kandidatenspracheinheit zu dem Pfad zu bestimmen. Wenn die andere Kandidatenspracheinheit 130b SU(m,2) ist, ist der Gesamtaufwand auf die Kandidatenspracheinheit SU(m,2) eine Kombination aus Zielaufwand$_2$ und Verknüpfungsaufwand$_2$.

[0044] In einigen Implementierungen kann ein zusammengesetzter Zielvektor Daten für mehrere Spracheinheiten in einem Pfad, zu dem erwogen wird, die Kandidatenspracheinheit hinzuzufügen,

enthalten. Beispielsweise kann, wenn der Gittergenerator 120 Kandidatenspracheinheiten bestimmt, die zu dem Pfad, der die ausgewählte Spracheinheit 128 und die ausgewählte andere Kandidatenspracheinheit 130b enthält, hinzugefügt werden sollen, ein neuer zusammengesetzter Zielvektor Akustikparameter für sowohl die ausgewählte Spracheinheit 128 als auch die ausgewählte andere Spracheinheit 130b enthalten. Der Gittergenerator 120 kann einen zusammengesetzten Vektor für eine neue Kandidatenspracheinheit 132b abrufen und den neuen zusammengesetzten Zielvektor mit dem neuen zusammengesetzten Vektor vergleichen, um einen Gesamtaufwand für das Hinzufügen der neuen Kandidatenspracheinheit 132b zu dem Pfad zu bestimmen.

[0045] In einigen Implementierungen, wenn ein Parameter ein Akustikparameter oder ein Linguistikparameter sein kann, kann ein Eintrag 126a-e für eine Spracheinheit einen zusammengesetzten Vektor mit Daten für die Parameter enthalten, die den Parameter einmal codieren. Der Gittergenerator 120 kann bestimmen, ob der Parameter in einer Aufwandsberechnung für eine Spracheinheit basierend auf den Parametern für eine Zieltexteinheit, den Akustikparametern für ausgewählte Spracheinheiten in dem Pfad oder beidem verwenden werden soll. In einigen Beispiele, wenn ein Parameter ein Akustikparameter und ein Linguistikparameter sein kann, kann ein Eintrag 126a-e für eine Spracheinheit einen zusammengesetzten Vektor mit Daten für die Parameter enthalten, die den Parameter zweimal codieren, einmal als einen Linguistikparameter und einmal als einen Akustikparameter..

[0046] In einigen Implementierungen sind spezielle Typen von Parametern nur Linguistikparameter oder Akustikparameter oder sind beides. Beispielsweise wenn ein spezieller Parameter ein Linguistikparameter ist, kann dieser spezielle Parameter möglicherweise kein Akustikparameter sein. Wenn ein spezieller Parameter ein Akustikparameter ist, kann dieser spezielle Parameter möglicherweise kein Linguistikparameter sein.

[0047] **Fig. 2** ist ein Beispiel eines Spracheinheitengitters 200. Der Gittergenerator 120 kann der Reihe nach das Gitter 200 mit einer vorbestimmten Menge von L Spracheinheiten für jede Texteinheit in der Folge von Texteinheiten besetzen. Jede in **Fig. 2** dargestellte Spalte repräsentiert eine Texteinheit und entsprechende Spracheinheiten. Für jede Texteinheit setzt der Gittergenerator eine vorbestimmte Anzahl von Pfaden K fort, die durch das Spracheinheitengitter 200 repräsentiert sind. An jeder Texteinheit oder wenn jede dargestellte Spalte besetzt wird, bewertet der Gittergenerator 120 erneut, welche K Pfade fortgesetzt werden sollten. Nachdem das Gitter 200 konstruiert ist, kann das Text-zu-Sprache-System 116 das Spracheinheitengitter 200 verwenden, um synthetisierte Sprache für die Folge von Texteinheiten zu bestimmen. In einigen Beispielen kann der Gittergenerator 120 in das Gitter 200 und für jede Texteinheit eine vorbestimmte Menge L von Spracheinheiten aufnehmen, die größer ist als die vorbestimmte Menge K von Pfaden, die ausgewählt sind, um fortgesetzt zu werden, bei jedem Übergang von einer Texteinheit zur nächsten. Zusätzlich kann ein Pfad, der als einer der besten K Pfade für eine spezielle Texteinheit identifiziert ist, erweitert oder in zwei oder mehr Pfade für die nächste Texteinheit verzweigt werden.

[0048] Im Allgemeinen kann das Gitter 200 so konstruiert sein, um eine Folge von M Texteinheiten zu repräsentieren, wobei m eine einzelne Texteinheit in der Folge {1, ..., M} repräsentiert. Der Gittergenerator 120 füllt einen anfänglichen Gitterabschnitt oder eine Spalte, die die Anfangstexteinheit (m=1) in der Folge repräsentiert. Das kann durch Auswählen aus einem Spracheinheitenkörper der Menge L von Spracheinheiten, die den niedrigsten Zielaufwand in Bezug auf die m=1 Texteinheit aufweisen, ausgeführt werden. Für jede zusätzliche Texteinheit in der Folge (m = {2, ..., M}) füllt der Gittergenerator 120 ebenfalls die entsprechende Spalte mit L Spracheinheiten. Für diese Spalten kann die Gruppe von L Spracheinheiten aus unterscheidbaren Gruppen nächster Nachbarn bestehen, die für unterschiedliche Pfade durch das Gitter 200 identifiziert sind. Insbesondere kann der Gittergenerator 120 die besten K Pfade durch das Gitter 200 identifizieren und eine Gruppe nächster Nachbarn für jeden aus den besten K Pfaden bestimmen. Die besten K Pfade können beschränkt sein, so dass jeder an einer unterschiedlichen Spracheinheit in dem Gitter 200 endet, z. B. enden die besten K Pfade an K unterschiedlichen Spracheinheiten. Die nächsten Nachbarn für einen Pfad können unter Verwendung (i) des Zielaufwands für die aktuelle Texteinheit und (ii) des Verknüpfungsaufwands in Bezug auf die letzte Spracheinheit in dem Pfad und/oder andere Spracheinheiten in dem Pfad bestimmt werden. Nachdem die Gruppe von L Spracheinheiten für eine gegebene Texteinheit ausgewählt worden ist, kann der Gittergenerator 200 eine Iteration des Viterbi-Algorithmus oder eines anderen geeigneten Algorithmus ablaufen lassen, um die K besten Pfade zu identifizieren, die verwendet werden sollen, wenn Spracheinheiten ausgewählt werden, die in das Gitter 200 für die nächste Texteinheit aufgenommen werden sollen.

[0049] Im Allgemeinen wählt der Gittergenerator 120 mehrere Kandidatenspracheinheiten aus, um sie in das Gitter 1200 für jede Texteinheit, z. B. Phon oder Diphon, des Texts, der synthetisiert werden soll, aufzunehmen, z. B. für jede Texteinheit in der Folge von Texteinheiten. Die Anzahl von Spracheinheiten, die für jede Texteinheit ausgewählt wird,

kann auf eine vorbestimmte Anzahl beschränkt sein, z. B. die vorbestimmte Menge L.

[0050] Beispielsweise kann der Gittergenerator 120 vor der Zeitspanne T_1 die vorbestimmten Menge L erster Spracheinheiten 202a-f für eine erste Texteinheit „h-e“ in einer Folge von Texteinheiten auswählen. Der Gittergenerator 120 kann die L besten Spracheinheiten für die ersten Spracheinheiten 202a-f auswählen. Beispielsweise kann der Gittergenerator 120 einen Zielaufwand für jede aus den ersten Spracheinheiten 202a-f verwenden, um zu bestimmen, welche aus den ersten Spracheinheiten 202a-f ausgewählt werden sollen. Falls die erste Einheit „h-e“ die Anfangstexteinheit am Anfang einer Äußerung, die synthetisiert wird, repräsentiert, kann nur der Zielaufwand in Bezug auf die Texteinheit verwendet werden. Falls die erste Einheit „h-e“ die Mitte einer Äußerung repräsentiert, wie z. B. das zweite oder nachfolgende Wort in der Äußerung, kann der Zielaufwand zusammen mit einem Verknüpfungsaufwand verwendet werden, um zu bestimmen, welche Spracheinheiten ausgewählt und in das Gitter 200 aufgenommen werden sollen. Der Gittergenerator 120 wählt eine vorbestimmte Anzahl K der vorbestimmten Menge L der ersten Spracheinheiten 202a-f aus. Die ausgewählte vorbestimmte Anzahl K der ersten Spracheinheiten 202a-f, z. B. die ausgewählten ersten Spracheinheiten 202a-c, sind in **Fig. 2** mit Schraffur gezeigt. In einigen Beispielen kann der Gittergenerator 120 die vorbestimmte Anzahl K erster Spracheinheiten 202a-f bestimmen, die als die Startspracheinheiten für Pfade ausgewählt werden sollen, die die Folge von Texteinheiten repräsentieren, z. B. mit oder ohne Auswählen der L ersten Spracheinheiten 202a-f.

[0051] Wenn die erste Texteinheit die Anfangstexteinheit der Folge repräsentiert, kann der Gittergenerator 120 die ersten Spracheinheiten 202a-c als die vorbestimmte Anzahl K von Spracheinheiten auswählen, die den besten Zielaufwand für die erste Texteinheit aufweisen. Der beste Zielaufwand kann der niedrigste Zielaufwand sein, z. B. wenn niedrigere Werte eine bessere Übereinstimmung zwischen den jeweiligen ersten Spracheinheit 202a-f und der Texteinheit „h-e“ repräsentieren, z. B. Ziel(m-1). In einigen Beispielen kann der beste Zielaufwand ein kürzester Abstand zwischen Linguistikparametern für die erste Kandidatenspracheinheit und Linguistikparametern für die Zieltexteinheit sein. Der beste Zielaufwand kann ein höchster Zielaufwand sein, z. B. wenn höhere Werte eine bessere Übereinstimmung zwischen den jeweiligen ersten Spracheinheit 202a-f und der Texteinheit „h-e“ repräsentieren. Wenn der Gittergenerator 120 einen niedrigsten Zielaufwand verwendet, repräsentieren niedrigere Verknüpfungsaufwände natürlicher artikulierte Sprache für die Zieleinheit. Wenn der Gittergenerator 120 einen höchsten Zielaufwand verwendet, repräsentieren höhere Verknüpfungsaufwände natürlicher artikulierte Sprache für die Zieleinheit.

[0052] Während der Zeit T_1 bestimmt der Gittergenerator 120 für jeden aus den aktuellen Pfaden, z. B. für jede aus den ausgewählten ersten Einheiten 202a-c, eine oder mehrere Kandidatenspracheinheiten unter Verwendung eines Verknüpfungsaufwands, eines Zielaufwand oder von beidem für die Kandidatenspracheinheiten. Der Gittergenerator 120 kann die zweiten Kandidatenspracheinheiten 204a-f aus dem synthetisierten Spracheinheitenkörper 124 bestimmen. Der Gittergenerator 120 kann die gesamte vorbestimmte Menge L der Kandidatenspracheinheiten 204a-f bestimmen. Der Gittergenerator 120 kann für jeden aus den K aktuellen Pfaden eine Anzahl von Kandidatenspracheinheiten unter Verwendung beider Werte L und K bestimmen. Die K aktuellen Pfade sind in **Fig. 2** durch die ausgewählten ersten Spracheinheiten 202a-c angegeben, gezeigt mit Schraffur, und die Verbindungen zwischen den ausgewählten ersten Spracheinheiten 202a-c sind mit Pfeilen zwischen den ausgewählten ersten Spracheinheiten 202a-c und den zweiten Kandidatenspracheinheiten 204a-f gezeigt, z. B. jede aus den zweiten Kandidatenspracheinheiten 204a-f ist für eine aus den ausgewählten ersten Spracheinheiten 202a-c spezifisch. Beispielsweise kann der Gittergenerator 120 L/K Kandidatenspracheinheiten für jeden aus den K Pfaden bestimmen. Wie in **Fig. 2** gezeigt ist, kann mit K = 3 und L = 6 der Gittergenerator 120 insgesamt zwei zweite Kandidatenspracheinheiten 204 für jeden aus den aktuellen Pfaden, die durch die ausgewählten ersten Spracheinheiten 202a-c identifiziert sind, bestimmen. Der Gittergenerator 120 kann zwei zweite Kandidatenspracheinheiten 204a-b für den Pfad, der die erste Spracheinheit 202a enthält, zwei zweite Kandidatenspracheinheiten 204cd für den Pfad, der die erste Spracheinheiten 202b enthält, und zwei zweite Kandidatenspracheinheit 204e-f für den Pfad, der die erste Spracheinheit 202c enthält, bestimmen.

[0053] Der Gittergenerator 120 wählt mehrere Kandidatenspracheinheiten aus den zweiten Kandidatenspracheinheiten 204a-f zum Hinzufügen zu den Definitionen der K Pfade aus, die der zweiten Texteinheit „e-l“ entsprechen, z. B. Ziel(m). Der Gittergenerator 120 kann die mehreren Kandidatenspracheinheiten aus den zweiten Kandidatenspracheinheiten 204a-f unter Verwendung des Verknüpfungsaufwands, des Zielaufwand oder von beiden für die Kandidatenspracheinheiten auswählen. Beispielsweise kann der Gittergenerator 120 die besten K zweiten Kandidatenspracheinheiten 204a-f auswählen, die z. B. niedrigere oder höhere Aufwände als die anderen Spracheinheiten in den zweiten Kandidatenspracheinheiten 204a-f aufweisen. Wenn niedrigere Aufwände eine bessere Übereinstimmung mit der entsprechenden ausge-

wählten ersten Spracheinheit repräsentieren, kann der Gittergenerator 120 die K zweiten Kandidatenspracheinheiten 204a-f mit den niedrigsten Aufwänden auswählen. Wenn höhere Aufwände eine bessere Übereinstimmung mit der entsprechenden ausgewählten ersten Spracheinheit repräsentieren, kann der Gittergenerator 120 die K zweiten Kandidatenspracheinheiten 204a-f mit den höchsten Aufwänden auswählen.

[0054] Der Gittergenerator 120 wählt die zweiten Kandidatenspracheinheiten 204b-d während der Zeitspanne T_1 aus, um die besten K Pfade zu der zweiten Texteinheit „e-l“ zu repräsentieren. Die ausgewählten zweiten Spracheinheiten 204b-d sind in **Fig. 2** mit Schraffur gezeigt. Der Gittergenerator 120 fügt die zweite Kandidatenspracheinheit 204b als eine ausgewählte zweite Spracheinheit zu dem Pfad hinzu, der die erste Spracheinheit 202a enthält. Der Gittergenerator 120 fügt die zweiten Kandidatenspracheinheiten 204c-d als ausgewählte zweite Spracheinheiten zu dem Pfad hinzu, der die erste Spracheinheit 202b enthält, um zwei Pfade zu definieren. Beispielsweise enthält der erste Pfad, der die erste Spracheinheit 202b enthält, außerdem die ausgewählte zweite Spracheinheit 204c für die zweite Texteinheit „e-l“. Der zweite Pfad, der die erste Spracheinheit 202b enthält, enthält die ausgewählte zweite Spracheinheit 204d für die zweite Texteinheit „e-l“.

[0055] In diesem Beispiel enthält der Pfad, der vorher die erste Spracheinheit 202c enthielt, keine aktuelle Spracheinheit, ist z. B. nach der Zeit T_1 kein aktueller Pfad. Weil die Aufwände für beide Kandidatenspracheinheiten 204e-f schlechter waren als die Aufwände für die ausgewählten zweiten Spracheinheiten 204b-d, hat der Gittergenerator 120 keine aus den Kandidatenspracheinheiten 204e-f ausgewählt und bestimmt, das Hinzufügen von Spracheinheiten zu dem Pfad, der die erste Spracheinheit 202c enthält, anzuhalten.

[0056] Während der Zeitspanne T_2 bestimmt der Gittergenerator 120 für jede aus den ausgewählten zweiten Spracheinheiten 204b-d, die die besten K Pfade bis zu der „e-l“-Texteinheit repräsentieren, mehrere dritte Kandidatenspracheinheiten 206a-f für die Texteinheit „l-o“, z. B. Ziel(m+1). Der Gittergenerator 120 kann die dritten Kandidatenspracheinheiten 206a-f aus dem synthetisierten Spracheinheitenkörper 124 bestimmen. Der Gittergenerator 120 kann einen Prozess ähnlich dem Prozess, der verwendet wird, um die zweiten Kandidatenspracheinheiten 204a-f zu bestimmen, wiederholen, um die dritten Kandidatenspracheinheiten 206a-f zu bestimmen. Beispielsweise kann der Gittergenerator 120 die dritten Kandidatenspracheinheiten 206a-b für die ausgewählte zweite Spracheinheit 204b, die dritten Kandidatenspracheinheiten 206c-d für die ausgewählte zweite Spracheinheit 204c und die dritten Kandidatenspracheinheiten 206e-f für die ausgewählte zweite Spracheinheit 204d bestimmen. Der Gittergenerator 120 kann einen Zielaufwand, einen Verknüpfungsaufwand oder beide, z. B. einen Gesamtaufwand, verwenden, um die dritten Kandidatenspracheinheiten 206a-f zu bestimmen.

[0057] Der Gittergenerator 120 kann dann mehrere Spracheinheiten aus den dritten Kandidatenspracheinheiten 206a-f unter Verwendung eines Zielaufwands, eines Verknüpfungsaufwands oder von beidem auswählen, um sie zu den Spracheinheitenpfaden hinzuzufügen. Beispielsweise kann der Gittergenerator 120 die dritten Kandidatenspracheinheiten 206a-c auswählen, um Pfade für die Folge von Texteinheiten zu definieren, die Spracheinheiten für die Texteinheit „l-o“ enthalten. Der Gittergenerator 120 kann die dritten Kandidatenspracheinheiten 206a-c zum Hinzufügen zu den Pfaden auswählen, weil die Gesamtaufwände für diese Spracheinheiten besser sind als die Gesamtaufwände für die anderen dritten Kandidatenspracheinheiten 206d-f.

[0058] Der Gittergenerator 120 kann den Prozess zum Auswählen mehrerer Spracheinheiten für jede Texteinheit unter Verwendung von Verknüpfungsaufwänden, Zielaufwände oder beidem für alle Texteinheiten in der Folge von Texteinheiten fortsetzen. Beispielsweise kann die Folge von Texteinheiten „h-e“, „e-l“ und „l-o“ am Anfang der Folge, wie mit Bezug auf **Fig. 1** beschrieben ist, in der Mitte der Folge, z. B. „Don - hello...“ oder am Ende der Folge enthalten.

[0059] In einigen Implementierungen kann der Gittergenerator 120 einen Zielaufwand, einen Verknüpfungsaufwand oder beide für eine oder mehrere Kandidatenspracheinheiten in Bezug auf eine nicht ausgewählte Spracheinheit bestimmen. Beispielsweise kann der Gittergenerator 120 Aufwände für die zweiten Kandidatenspracheinheiten 204a-f in Bezug auf die nicht ausgewählten ersten Spracheinheiten 202d-f bestimmen. Falls der Gittergenerator 120 bestimmt, dass ein Gesamtpfadaufwand für eine Kombination aus einer aus den zweiten Kandidatenspracheinheiten 204a-f mit einer aus den nicht ausgewählten ersten Spracheinheiten 202d-f angibt, dass dieser Pfad einer aus den besten K Pfaden ist, kann der Gittergenerator 120 die entsprechende zweite Spracheinheit zu der nicht ausgewählten ersten Spracheinheit hinzufügen. Beispielsweise kann der Gittergenerator bestimmen, dass ein Gesamtpfadaufwand für einen Pfad, der die nicht ausgewählte erste Spracheinheit 202f und die zweite Kandidatenspracheinheit 204 enthält, einer aus den besten K Pfaden ist, und diesen Pfad verwenden, um eine dritte Spracheinheit 206 auszuwählen.

[0060] **Fig. 2** stellt verschiedene signifikante Aspekte des Prozesses zum Aufbauen des Gitters 200 dar. Der Gittergenerator 120 kann das Gitter 200 auf eine sequenzielle Weise aufbauen, indem er eine erste Gruppe von Spracheinheiten auswählt, um die erste Texteinheit in dem Gitter 200 zu repräsentieren, und dann eine zweite Gruppe von Spracheinheiten auswählt, um die zweite Texteinheit in dem Gitter 200 zu repräsentieren, und so weiter. Die Auswahl der Spracheinheiten für jede Texteinheit kann von den Spracheinheiten abhängen, die in dem Gitter 200 für vorhergehende Texteinheiten enthalten sind. Der Gittergenerator 120 wählt mehrere Spracheinheiten aus, um sie in das Gitter 200 für jede Texteinheit aufzunehmen, z. B. L = 6 Spracheinheiten pro Texteinheit in dem Beispiel von **Fig. 2**.

[0061] Der Gittergenerator 120 kann die Spracheinheiten für das Gitter 200 auf eine Weise auswählen, die die existierenden besten Pfade für das Gitter 200 fortsetzt oder darauf aufbaut. Anstatt einen einzigen besten Pfad oder nur Pfade, die eine einzelne Spracheinheit durchlaufen, fortzusetzen, setzt der Gittergenerator 120 Pfade durch mehrere Spracheinheiten in dem Gitter für jede Texteinheit fort. Der Gittergenerator 120 kann eine Viterbi-Analyse jedes Mal neu ablaufen lassen, wenn eine Gruppe von Spracheinheiten zu dem Gitter 200 hinzugefügt wird. Als ein Ergebnis kann sich die spezifische Beschaffenheit der Pfade von einem Auswahlschritt zum nächsten ändern.

[0062] In **Fig. 2** enthält jede Spalte sechs Spracheinheiten, und nur drei aus den Spracheinheiten in einer Spalte werden verwendet, um zu bestimmen, welche Spracheinheiten in die nächste Spalte aufgenommen werden sollen. Der Gittergenerator 120 wählt eine vorbestimmte Anzahl von Spracheinheiten, z. B. die Einheiten 202a-202c, für die Texteinheit „h-e“ aus, die die besten Pfade durch das Gitter 200 zu diesem Punkt repräsentieren. Diese können die Spracheinheiten sein, die einem niedrigsten Gesamtaufwand zugeordnet sind. Für eine spezielle Spracheinheit in dem Gitter 200 kann der Gesamtaufwand die kombinierten Verknüpfungsaufwände und Zielaufwände in einem besten Pfad durch das Gitter 200 repräsentieren, der (i) an irgendeiner Spracheinheit in dem Gitter 200, die die Anfangstexteinheit der Texteinheitenfolge repräsentiert, beginnt und (ii) an der speziellen Spracheinheit endet.

[0063] Um die Spracheinheiten für eine aktuelle Texteinheit auszuwählen, kann der Viterbi-Algorithmus ablaufen, um den besten Pfad und zugeordneten Gesamtaufwand für jede Spracheinheit in dem Gitter 200, die eine frühere Texteinheit repräsentiert, zu bestimmen. Eine vorbestimmte Anzahl von Spracheinheiten mit dem niedrigsten Gesamtpfadaufwand, z. B. K = 3 in dem Beispiel von **Fig. 2**, kann als die besten K Spracheinheiten für die frühere Texteinheit ausgewählt werden. Diese besten K Spracheinheiten für die frühere Texteinheit können während der Analyse verwendet werden, die ausgeführt wird, um die Spracheinheiten auszuwählen, die die aktuelle Texteinheit repräsentieren sollen. Jede aus den besten Spracheinheiten kann einem Abschnitt des begrenzten Raums in dem Gitter für die aktuelle Texteinheit zugewiesen werden, z. B. dem Raum für L = 6 Spracheinheiten.

[0064] Für jede aus den besten K Spracheinheiten für die frühere Texteinheit kann eine vorbestimmte Anzahl von Texteinheiten zu dem Gitter hinzugefügt werden, um die aktuelle Texteinheit zu repräsentieren. Beispielsweise können L / K Spracheinheiten, z. B. 6 / 3 = 2 Spracheinheiten, für jede aus den besten K Spracheinheiten für die frühere Spracheinheit hinzugefügt werden. Für die Spracheinheit 202a, die als eine der besten K Spracheinheiten für die Texteinheit „h-e“ bestimmt ist, werden die Spracheinheiten 204a und 204b basierend auf ihren Zielaufwänden in Bezug auf die Texteinheit „e-l“ und basierend auf ihren Verknüpfungsaufwänden in Bezug auf die Spracheinheit 202a ausgewählt und hinzugefügt. Ähnlich werden für die Spracheinheit 202b, die ebenfalls als eine der besten K Spracheinheiten für die Texteinheit „h-e“ bestimmt ist, die Spracheinheiten 204c und 204d basierend auf ihren Zielaufwänden in Bezug auf die Texteinheit „e-l“ und basierend auf ihren Verknüpfungsaufwänden in Bezug auf die Spracheinheit 202b ausgewählt und hinzugefügt. Die erste Gruppe von Spracheinheiten 204a und 204b kann gemäß etwas anderen Kriterien als die zweite Gruppe von Spracheinheiten 204c und 204d ausgewählt werden, da die zwei Gruppen unter Verwendung von Verknüpfungsaufwänden in Bezug auf unterschiedliche frühere Spracheinheiten bestimmt werden.

[0065] Das Beispiel von **Fig. 2** zeigt, dass für eine aktuelle Spalte des Gitters 200, die besetzt wird, Pfade durch einige der Spracheinheiten in der vorhergehenden Spalte effektiv abgeschnitten oder ignoriert werden und nicht verwendet werden, um Verknüpfungsaufwände zum Hinzufügen von Spracheinheiten zu der aktuellen Spalte zu bestimmen. Zusätzlich wird ein Pfad durch eine aus den besten K Spracheinheiten in der vorhergehenden Spalte verzweigt oder aufgespalten, so dass zwei oder mehr Spracheinheiten in der aktuellen Spalte den Pfad getrennt fortsetzen. Als ein Ergebnis zweigt der Auswahlprozess für jede Texteinheit effektiv die besten Pfade mit dem niedrigsten Aufwand ab, während er die Berechnungskomplexität durch Einschränken der Anzahl von Kandidatenspracheinheiten für jede Texteinheit begrenzt.

[0066] Zurück zu **Fig. 1**, wenn der Gittergenerator 120 Spracheinheiten für alle Texteinheiten in der Folge von Texteinheiten bestimmt hat, z. B. K Pfade

von Texteinheiten bestimmt hat, stellt der Gittergenerator 120 Daten für jeden aus den Pfaden für eine Pfadauswahleinheit 122 bereit. Die Pfadauswahleinheit 122 analysiert jeden aus den Pfaden, um einen besten Pfad zu bestimmen. Der beste Pfad kann einen niedrigsten Aufwand aufweisen, wenn niedrigere Werte für den Aufwand eine bessere Übereinstimmung zwischen Spracheinheiten und Texteinheiten repräsentieren. Der beste Pfad kann einen höchsten Aufwand aufweisen, wenn höhere Werte eine bessere Übereinstimmung zwischen Spracheinheiten und Texteinheiten repräsentieren.

[0067] Beispielsweise kann die Pfadauswahleinheit 122 jeden aus den K Pfaden, die durch den Gittergenerator 120 erzeugt werden, analysieren und einen Pfad unter Verwendung eines Zielaufwands, eines Verknüpfungsaufwands oder eines Gesamtaufwands für die Spracheinheiten in dem Pfad auswählen. Die Pfadauswahleinheit 122 kann einen Pfadaufwand durch Kombinieren der Aufwände für jede aus den ausgewählten Spracheinheiten in dem Pfad bestimmen. Beispielsweise kann, wenn ein Pfad drei Spracheinheiten enthält, die Pfadauswahleinheit 122 eine Summe der Aufwände bestimmen, die verwendet werden, um jede aus den drei Spracheinheiten auszuwählen. Die Aufwände können Zielaufwände, Verknüpfungsaufwände oder eine Kombination aus beiden sein. In einigen Beispielen können die Aufwände eine Kombination von zwei oder mehr aus den Zielaufwänden, den Verknüpfungsaufwänden oder den Gesamtaufwänden sein.

[0068] In dem Spracheinheitengitter 200, das in **Fig. 2** gezeigt ist, wählt die Pfadauswahleinheit 122 einen Pfad, der die Spracheinheit(m-1,1) 202a, die Spracheinheit(m,2) 204b und die Spracheinheit(m +1,2) 206b enthält, zur Synthese des Worts „hello“ aus, wie durch die fetten Linien angegeben ist, die diese Spracheinheiten umgeben und verbinden. Die ausgewählten Spracheinheiten können einen niedrigsten Pfadaufwand oder einen höchsten Pfadaufwand aufweisen, abhängig davon, ob niedrigere oder höhere Werte eine bessere Übereinstimmung zwischen Spracheinheiten und Texteinheiten und zwischen mehreren Spracheinheiten in demselben Pfad angeben.

[0069] Zurück zu **Fig. 1** erzeugt das Text-zu-Sprache-System 116 eine zweite Kommunikation 136, die synthetisierte Sprachdaten für den ausgewählten Pfad identifiziert. In einigen Implementierungen können die synthetisierten Sprachdaten Anweisungen enthalten, um zu bewirken, dass eine Vorrichtung, z. B. ein Lautsprecher, synthetisierte Sprache für die Textnachricht erzeugt.

[0070] Das Text-zu-Sprache-System 116 stellt die zweite Kommunikation 136 für die Anwendervorrichtung 102 bereit, z. B. unter Verwendung des Netzes 138. Die Anwendervorrichtung 102, z. B. der computerimplementierte Agent 108, stellt eine hörbare Präsentation 110 der Textnachricht auf einem Lautsprecher 106 unter Verwendung von Daten aus der zweiten Kommunikation 136 bereit. Die Anwendervorrichtung 102 kann die hörbare Präsentation 110 bereitstellen, während sie sichtbaren Inhalt 114 der Textnachricht in einer Anwendungs-Benutzerschnittstelle 112, z. B. einer Textnachrichtenanwendung-Benutzerschnittstelle, auf einer Anzeigevorrichtung präsentiert.

[0071] In einigen Implementierungen kann die Folge von Texteinheiten für ein Wort, einen Satz oder einen Absatz sein. Beispielsweise kann der Texteinheiten-Parser 118 Daten empfangen, die einen Absatz identifizieren, und den Absatz in Sätze unterteilen. Der erste Satz kann „Hello, Don“ sein, und der zweite Satz kann „Let's connect on Friday“ sein. Der Texteinheiten-Parser 118 kann getrennte Folgen von Texteinheiten für jeden der Sätze für den Gittergenerator 120 bereitstellen, um zu bewirken, dass die Auswahleinheit für synthetisierte Daten Pfade für jede aus den Folgen von Texteinheiten getrennt erzeugt.

[0072] Der Texteinheiten-Parser 118 und das Text-zu-Sprache-System 116 können eine Länge der Folge von Texteinheiten unter Verwendung einer Zeit, zu der die synthetisierten Sprachdaten präsentiert werden sollen, eines Maßes, das angibt, wie wahrscheinlich sich synthetisierte Sprachdaten als natürlich artikulierte Sprache verhalten, oder beidem bestimmen. Beispielsweise um zu bewirken, dass der Lautsprecher 106 hörbaren Inhalt schneller präsentiert, kann der Texteinheiten-Parser 118 kürzere Folgen von Texteinheiten auswählen, so dass das Text-zu-Sprache-System 116 die Anwendervorrichtung 102 schneller mit der zweiten Kommunikation 136 versorgen kann. In diesen Beispielen kann das Text-zu-Sprache-System 116 die Anwendervorrichtung 102 mit mehreren zweiten Kommunikationen versorgen, bis das Text-zu-Sprache-System 116 Daten für die gesamte Textnachricht oder andere Textdaten bereitgestellt hat. In einigen Beispielen kann der Texteinheiten-Parser 118 längere Folgen von Texteinheiten auswählen, um die Wahrscheinlichkeit zu erhöhen, dass sich die synthetisierten Sprachdaten wie natürlich artikulierte Sprache verhalten.

[0073] In einigen Implementierungen weist der computerimplementierte Agent 108 vorbestimmte Sprachsynthesedaten für eine oder mehrere vordefinierte Nachrichten auf. Beispielsweise kann der computerimplementierte Agent 108 vorbestimmte Sprachsynthesedaten für den Hinweis „Es ist eine ungelesene Textnachricht für dich vorhanden“ enthalten. In diesen Beispielen sendet der computerimplementierte Agent 108 Daten für die ungelesene

Textnachricht zu dem Text-zu-Sprache-System 116, weil der computerimplementierte Agent 108 keine vorbestimmten Sprachsynthesedaten für die ungelesene Textnachricht aufweist. Beispielsweise ist die Folge von Worten und Sätzen in der ungelesenen Textnachricht nicht gleich einer der vordefinierten Nachrichten für den computerimplementierten Agenten 108.

[0074] In einigen Implementierungen kann die Anwendervorrichtung 102 hörbare Präsentation von Inhalt ohne die Verwendung des computerimplementierten Agenten 108 bereitstellen. Beispielsweise kann die Anwendervorrichtung 102 eine Textnachrichtenanwendung oder eine andere Anwendung enthalten, die die hörbare Präsentation der Textnachricht bereitstellt.

[0075] Das Text-zu-Sprache-System 116 ist ein Beispiel eines Systems, das als Computerprogramme auf einem oder mehreren Computern an einem oder mehreren Orten implementiert ist, in dem Systeme, Komponenten und Techniken, die in diesem Dokument beschrieben sind, implementiert sind. Die Anwendervorrichtung 102 kann Personalcomputer, mobile Kommunikationsvorrichtungen und andere Vorrichtungen enthalten, die Daten über das Netz 138 senden und empfangen können. Das Netz 138 wie z. B. ein lokales Netz (LAN), ein Weitbereichsnetz (WAN), das Internet oder eine Kombination daraus verbindet die Anwendervorrichtung 102 und das Text-zu-Sprache-System 116. Das Text-zu-Sprache-System 116 kann einen einzelnen Server-Computer oder mehrere Server-Computer, die zusammen arbeiten, verwenden, die beispielsweise eine Gruppe entfernter Computer enthalten, die als Cloud-Berechnungsdienst verteilt sind.

[0076] **Fig. 3** ist ein Ablaufdiagramm eines Prozesses 300 zum Bereitstellen synthetisierter Sprachdaten. Beispielsweise kann der Prozess 300 durch das Text-zu-Sprache-System 116 aus der Umgebung 100 verwendet werden.

[0077] Ein Text-zu-Sprache-System empfängt Daten, die Text zur Sprachsynthese angeben (302). Beispielsweise empfängt das Text-zu-Sprache-System Daten von einer Anwendervorrichtung, die Text aus einer Textnachricht oder E-Mail angeben. Die Daten können den Typ des Texts wie z. B. E-Mail oder Textnachricht zum Gebrauch zum Bestimmen von Synthesedaten identifizieren.

[0078] Das Text-zu-Sprache-System bestimmt eine Folge von Texteinheiten, die jeweils einen entsprechenden Abschnitt des Texts repräsentieren (304). Jede aus den Texteinheiten kann einen unterscheidbaren Abschnitt des Texts repräsentieren, getrennt von den Abschnitten von Text, die durch die anderen Texteinheiten repräsentiert sind. Das Text-zu-Sprache-System kann eine Folge von Texteinheiten für den gesamten empfangenen Text bestimmen. In einigen Beispielen kann das Text-zu-Sprache-System eine Folge von Texteinheiten für einen Abschnitt des empfangenen Texts bestimmen.

[0079] Das Text-zu-Sprache-System bestimmt mehrere Pfade von Spracheinheiten, die jeweils eine Folge von Texteinheiten repräsentieren (306). Beispielsweise kann das Text-zu-Sprache-System einen oder mehrere der Schritte 308 bis 314 ausführen, um die Pfade der Spracheinheiten zu bestimmen.

[0080] Das Text-zu-Sprache-System wählt aus einem Spracheinheitenkörper eine erste Spracheinheit aus, die Sprachsynthesedaten umfasst, die die erste Texteinheit repräsentieren (308). Die erste Texteinheit kann einen Ort am Anfang der Folge von Texteinheiten aufweisen. In einigen Beispielen kann die erste Texteinheit einen unterschiedlichen Ort in der Folge von Texteinheiten aufweisen, der nicht der letzte Ort in der Folge von Texteinheiten ist. In einigen Beispielen kann das Text-zu-Sprache-System zwei oder mehr erste Spracheinheiten auswählen, die jeweils unterschiedliche Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren.

[0081] Das Text-zu-Sprache-System bestimmt für jede aus mehreren zweiten Spracheinheiten in dem Spracheinheitenkörper (i) einen Verknüpfungsaufwand, um die zweite Spracheinheit mit der ersten Spracheinheit zu verketten, und (ii) einen Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit einer zweiten Texteinheit entspricht (310). Die zweite Texteinheit kann einen zweiten Ort in der Folge von Texteinheiten aufweisen, der dem Ort für die erste Texteinheit ohne irgendwelche dazwischenliegenden Orte in der Folge von Texteinheiten nachfolgt. In einigen Implementierungen kann das Text-zu-Sprache-System einen Verknüpfungsaufwand bestimmen, um die zweite Spracheinheit mit der ersten Spracheinheit und einer oder mehreren zusätzlichen Spracheinheiten in dem Pfad zu verknüpfen, die z. B. eine Anfangsspracheinheit in dem Pfad enthält, die eine andere Spracheinheit als die erste Spracheinheit ist.

[0082] Das Text-zu-Sprache-System kann erste Akustikparameter für jede ausgewählte Spracheinheit in dem Pfad bestimmen. Das Text-zu-Sprache-System kann erste Linguistikparameter für die zweite Texteinheit bestimmen Das Text-zu-Sprache-System kann einen zusammengesetzten Zielvektor bestimmen, der Daten für die ersten Akustikparameter und die ersten Linguistikparameter enthält. Das Text-zu-Sprache-System muss die ersten Akustikparameter, die ersten Linguistikparameter und den zusammengesetzten Zielvektor nur einmal für die Gruppe aus

mehreren zweiten Spracheinheiten bestimmen. In einigen Beispielen kann das Text-zu-Sprache-System die ersten Akustikparameter, die ersten Linguistikparameter und den Zielvektor getrennt für jede zweite Spracheinheit bestimmen.

[0083] Das Text-zu-Sprache-System kann einen jeweiligen Verknüpfungsaufwand für eine spezielle zweite Spracheinheit unter Verwendung der ersten Akustikparameter und zweiten Akustikparameter für die spezielle zweite Spracheinheit bestimmen. Das Text-zu-Sprache-System kann einen jeweiligen Zielaufwand für eine spezielle zweite Spracheinheit unter Verwendung der ersten Linguistikparameter und zweiten Linguistikparameter für die spezielle zweite Spracheinheit bestimmen. Wenn das Text-zu-Sprache-System sowohl einen Verknüpfungsaufwand als auch einen Zielaufwand für eine spezielle zweite Spracheinheit bestimmt, kann das Text-zu-Sprache-System nur einen Gesamtaufwand für die spezielle zweite Spracheinheit bestimmen, der sowohl den Verknüpfungsaufwand als auch den Zielaufwand für das Hinzufügen der speziellen zweiten Spracheinheit zu einem Pfad repräsentiert.

[0084] In einigen Implementierungen kann das Text-zu-Sprache-System einen oder mehrere Aufwände für mehrere zweite Spracheinheiten gleichzeitig bestimmen. Beispielsweise kann das Text-zu-Sprache-System gleichzeitig für jede aus zwei oder mehr zweiten Spracheinheiten den Verknüpfungsaufwand und die Zielaufwände bestimmen, z. B. als getrennte Aufwände oder als einen einzigen Zielaufwand für die jeweilige zweite Spracheinheit.

[0085] Das Text-zu-Sprache-System wählt aus den mehreren zweiten Spracheinheiten mehrere dritte Spracheinheiten, die Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, unter Verwendung des jeweiligen Verknüpfungsaufwands und Zielaufwands aus (312). Beispielsweise kann das Text-zu-Sprache-System die besten K zweiten Spracheinheiten bestimmen. Das Text-zu-Sprache-System kann den Aufwand für jede aus den zweiten Spracheinheiten mit den Aufwänden für die anderen zweiten Spracheinheiten vergleichen, um die besten K zweiten Spracheinheiten zu bestimmen.

[0086] Das Text-zu-Sprache-System definiert Pfade von der ausgewählten ersten Spracheinheit zu jeder aus den mehreren zweiten Spracheinheiten, die in die mehreren Pfade von Spracheinheiten aufgenommen werden sollen (314). Das Text-zu-Sprache-System kann K Pfade unter Verwendung der bestimmten besten K zweiten Spracheinheiten erzeugen, wobei jede aus den besten K zweiten Spracheinheiten eine letzte Spracheinheit für den jeweiligen Pfad ist.

[0087] Das Text-zu-Sprache-System stellt synthetisierte Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist, bereit (316). Das Bereitstellen der synthetisierten Sprachdaten für eine Vorrichtung kann bewirken, dass die Vorrichtung eine hörbare Präsentation der synthetisierten Sprachdaten erzeugt, die dem gesamten oder einem Teil des empfangenen Texts entspricht.

[0088] In einigen Implementierungen kann der Prozess 300 zusätzliche Schritte oder weniger Schritte enthalten, oder einige Schritte können in mehrere Schritte unterteilt sein. Beispielsweise kann das Text-zu-Sprache-System die Schritte 302 bis 304 und 310 bis 314 ausführen, ohne die Schritte 306, 308 oder 316 auszuführen.

[0089] Ausführungsformen der Gegenstands und der funktionalen Operationen, die in dieser Spezifikation beschrieben sind, können in einer digitalen elektronischen Schaltungsanordnung oder in materiell ausgeführter Computer-Software oder Firmware, in Computer-Hardware, die die Strukturen, die in dieser Spezifikation offenbart sind, und ihre strukturellen Äquivalente enthalten, oder in Kombinationen aus einem oder mehreren daraus implementiert sein. Ausführungsformen des Gegenstands, der in dieser Spezifikation beschrieben ist, können als ein oder mehrere Computerprogramme implementiert sein, d. h. ein oder mehrere Module aus Computerprogrammanweisungen, die auf einem materiellen Nichttransitorischen Programmträger codiert sind, zur Ausführung durch oder zur Steuerung des Betriebs einer Datenverarbeitungseinrichtung. Alternativ oder zusätzlich können die Programmanweisungen auf einem künstlich erzeugten verbreiteten Signal codiert sein, z. B. einem maschinenerzeugten elektrischen, optischen oder elektromagnetischen Signal, das erzeugt wird, um Informationen zur Übertragung zu geeigneten Empfängereinrichtungen zur Ausführung durch eine Datenverarbeitungseinrichtung zu codieren. Das Computerspeichermedium kann eine maschinenlesbare Speichervorrichtung, ein maschinenlesbares Speichersubstrat, eine Speichervorrichtung für Direktzugriff oder seriellen Zugriff oder eine Kombination aus einem oder mehreren daraus sein.

[0090] Der Begriff „Datenverarbeitungseinrichtung“ bezieht sich auf Datenverarbeitungs-Hardware und umfasst alle Arten von Einrichtungen, Geräten und Maschinen zur Verarbeitung von Daten, die als Beispiel einen programmierbaren Prozessor, einen Computer oder mehrere Prozessoren oder Computer enthalten. Die Einrichtung kann außerdem Spezial-Logikschaltungsanordnung, z. B. ein FGPA (feldprogrammierbares Gatterfeld) oder eine ASIC (anwendungsspezifische integrierte Schaltung), sein oder ferner enthalten. Die Einrichtung kann optional zusätzlich zu Hardware Code enthalten, der eine Ausführungsumgebung für Computerprogramme erzeugt, z. B. Code, der Prozessor-Firm-

ware, einen Protokollstack, ein Datenbankmanagementsystem, ein Betriebssystem oder eine Kombination aus einem oder mehreren davon bildet.

[0091] Ein Computerprogramm, das auch als Programm, Software, eine Software-Anwendung, ein Modul, ein Software-Modul, ein Skript oder Code bezeichnet oder beschrieben sein kann, kann in irgendeiner Form einer Programmiersprache geschrieben sein, die kompilierte oder interpretierte Sprachen oder deklarative oder prozedurale Sprachen enthält, und es kann in irgendeiner Form verteilt werden, die als ein eigenständiges Programm oder als ein Modul, eine Komponente, eine Subroutine oder eine andere Einheit, die zum Gebrauch in einer Berechnungsumgebung geeignet ist, enthält. Ein Computerprogramm kann, muss jedoch nicht, einer Datei in einem Dateisystem entsprechen. Ein Programm kann in einem Abschnitt einer Datei, die andere Programme oder Daten enthält, z. B. ein oder mehrere Skripte, die in einem Auszeichnungssprachen-Dokument gespeichert sind, in einer einzelnen Datei, die für das fragliche Programm dediziert ist, oder in mehreren koordinierten Dateien, z. B. Dateien, die ein oder mehrere Module, Unterprogramme oder Code-Abschnitte speichern, gespeichert sein. Ein Computerprogramm kann verteilt werden, um auf einem Computer oder auf mehreren Computern, die sich an einem Standort oder verteilt über mehrere Standorte befinden und durch ein Kommunikationsnetz miteinander verbunden sind, ausgeführt zu werden.

[0092] Die Prozesse und Logikabläufe, die in dieser Spezifikation beschrieben sind, können durch einen oder mehrere programmierbare Computer ausgeführt werden, die ein oder mehrere Computerprogramme ablaufen lassen, um Funktionen durch Arbeiten auf Eingabedaten und Erzeugen einer Ausgabe auszuführen. Die Prozesse und Logikabläufe können auch durch eine Spezial-Logikschaltungsanordnung, z. B. ein FGPA (feldprogrammierbares Gatterfeld) oder eine ASIC (anwendungsspezifische integrierte Schaltung) ausgeführt werden, und die Einrichtung kann dadurch implementiert sein.

[0093] Computer, die für die Ausführung eines Computerprogramms geeignet sind, enthalten als Beispiel Allzweck- oder Spezial-Mikroprozessoren oder beides oder irgendeine andere Art von zentraler Verarbeitungseinheit. Allgemein wird eine zentrale Verarbeitungseinheit Anweisungen und Daten aus einem Festwertspeicher oder einem Direktzugriffsspeicher oder beiden empfangen. Die wesentlichen Elemente eines Computers sind eine zentrale Verarbeitungseinheit zum Ausführen oder Durchführen von Anweisungen und eine oder mehrere Speichervorrichtungen zum Speichern von Anweisungen und Daten. Allgemein wird ein Computer auch eine oder mehrere Massenspeichervorrichtungen zum Speichern von Daten, z. B. magnetische, magneto-optische Platten oder optische Platten, enthalten oder betriebstechnisch damit gekoppelt sein, um Daten von ihnen zu empfangen, zu ihnen zu übertragen oder beides. Ein Computer muss jedoch solche Vorrichtungen nicht aufweisen. Außerdem kann ein Computer in eine weitere Vorrichtung eingebettet sein, z. B. in ein Mobiltelefon, ein Smartphone, einen persönlichen digitalen Assistenten (PDA), ein mobiles Audio- oder Videoabspielgerät, eine Spielkonsole, einen Empfänger des globalen Positionierungssystems (GPS-Empfänger) oder eine tragbare Speichervorrichtung, z. B. ein Flashlaufwerk über den universellen seriellen Bus (USB-Flashlaufwerk), um nur einige wenige zu nennen.

[0094] Computerlesbare Medien, die zum Speichern von Computerprogrammanweisungen und Daten geeignet sind, enthalten alle Formen von nichtflüchtigem Speicher, Medien und Speichervorrichtungen, die als Beispiel Halbleiterspeichervorrichtungen, z. B. EPROM, EEPROM und Flash-Speichervorrichtungen; Magnetplatten, z. B. interne Festplatten oder herausnehmbare Platten; magneto-optische Platten; und CD-ROM und DVD-ROM-Platten enthalten. Der Prozessor und der Speicher können durch eine Spezial-Logikschaltungsanordnung ergänzt oder darin integriert sein.

[0095] Um die Interaktion mit einem Anwender bereitzustellen, können Ausführungsformen des in dieser Spezifikation beschriebenen Gegenstands auf einem Computer implementiert sein, der eine Anzeigevorrichtung, z. B. einen LCD-Monitor (Flüssigkristallanzeige-Monitor), einen OLED-Monitor (Monitor mit organischer Leuchtdiode) oder einen anderen Monitor zum Anzeigen von Informationen für den Anwender und eine Tastatur und eine Zeigevorrichtung, z. B. eine Maus oder einen Trackball, durch die der Anwender Eingaben für den Computer bereitstellen kann, aufweist. Andere Arten von Vorrichtungen können verwendet werden, um ebenfalls Interaktion mit Anwender einem bereitzustellen; beispielsweise kann eine für den Anwender bereitgestellte Rückmeldung irgendeine Form sensorischer Rückmeldung sein, z. B. visuelle Rückmeldung, hörbare Rückmeldung oder tastbare Rückmeldung; und eine Eingabe von dem Anwender kann in irgendeiner Form empfangen werden, die akustische, Sprach- oder tastbare Eingabe enthält. Zusätzlich kann ein Computer mit einem Anwender interagieren durch Senden von Dokumenten zu einer Vorrichtung und Empfangen von Dokumenten von einer Vorrichtung, die durch den Anwender verwendet wird; beispielsweise durch Senden von Web-Seiten zu einem Web-Browser auf einer Vorrichtung eines Anwenders in Reaktion auf Anforderungen, die von dem Web-Browser empfangen werden.

[0096] Ausführungsformen des in dieser Spezifikation beschriebenen Gegenstands können in einem Berechnungssystem implementiert sein, das eine Backend-Komponente, z. B. als ein Daten-Server, enthält oder das eine Middleware-Komponente, z. B. einen Anwendungsserver, enthält, oder der eine Frontend-Komponente, z. B. einen Client-Computer, enthält, der eine grafische Anwenderschnittstelle oder einen Web-Browser aufweist, durch den ein Anwender mit einer Implementierung des in dieser Spezifikation beschriebenen Gegenstands interagieren kann, oder irgendeine Kombination eines oder mehrerer solcher Backend-, Middleware- oder Frontend-Komponenten. Die Komponenten des Systems können durch irgendeine Form oder irgendein Medium zur digitalen Datenkommunikation, z. B. ein Kommunikationsnetz, miteinander verbunden sein. Beispiele für Kommunikationsnetze enthalten ein lokales Netz (LAN) und ein Weitbereichsnetz (WAN), z. B. das Internet.

[0097] Das Berechnungssystem kann Clients und Server enthalten. Ein Client und ein Server sind im Allgemeinen voneinander entfernt und interagieren typischerweise über ein Kommunikationsnetz. Die Beziehung von Client und Server entsteht aufgrund der Computerprogramme, die auf den jeweiligen Computern laufen und eine Client-Server-Beziehung miteinander aufweisen. In einigen Ausführungsformen sendet ein Server Daten, z. B. eine Seite mit HyperText-Auszeichnungssprache (HTML-Seite) zu einer Anwendervorrichtung, z. B. zum Zweck der Anzeige der Daten für einen Anwender und Empfangen von Anwendereingabe von einem Anwender, der mit der Anwendervorrichtung, die als ein Client arbeitet, interagiert. Daten, die in der Anwendervorrichtung erzeugt werden, z. B. ein Ergebnis einer Anwenderinteraktion, können von der Anwendervorrichtung in dem Server empfangen werden.

[0098] Fig. 4 ist ein Blockdiagramm von Berechnungsvorrichtungen 400, 450, die verwendet werden können, um die Systeme und Verfahren, die in diesem Dokument beschrieben sind, entweder als ein Client oder als ein Server oder mehrere Server zu implementieren. Die Berechnungsvorrichtung 400 ist vorgesehen, um verschiedene Formen von digitalen Computern, wie z. B. Laptops, Desktops, Workstations, persönliche digitale Assistenten, Server, Blade-Server, Mainframes und andere geeignete Computer zu repräsentieren. Die Berechnungsvorrichtung 450 ist vorgesehen, um verschiedene Formen mobiler Vorrichtungen zu repräsentieren, wie z. B. persönliche digitale Assistenten, Mobiltelefone, Smartphones, Smartwatches, am Kopf getragene Vorrichtungen und andere ähnliche Berechnungsvorrichtungen. Die hier gezeigten Komponenten, ihre Verbindungen und Beziehungen und ihre Funktionen sollen nur beispielhaft sein und sind nicht so gemeint, in diesem Dokument beschriebene und beanspruchte Implementierungen einzuschränken.

[0099] Die Berechnungsvorrichtung 400 enthält einen Prozessor 402, einen Speicher 404, eine Speichervorrichtung 406, eine Hochgeschwindigkeitsschnittstelle 408, die mit dem Speicher 404 und Hochgeschwindigkeitserweiterungsanschlüssen 410 verbindet, und eine Niedergeschwindigkeitsschnittstelle 412, die mit dem Niedergeschwindigkeitsbus 414 und der Speichervorrichtung 406 verbindet. Jede aus den Komponenten 402, 404, 406, 408, 410 und 412 ist unter Verwendung verschiedener Busse miteinander verbunden und kann auf einer gemeinsamen Hauptplatine oder auf andere Weise wie jeweils anwendbar montiert sein. Der Prozessor 402 kann Anweisungen zur Ausführung innerhalb der Berechnungsvorrichtung 400 verarbeiten, die Anweisungen enthalten, die in dem Speicher 404 oder auf der Speichervorrichtung 406 gespeichert sind, um grafische Informationen für eine GUI auf einer externen Eingabe/Ausgabevorrichtung, wie z. B. einer Anzeigevorrichtung 416, die mit der Hochgeschwindigkeitsschnittstelle 408 gekoppelt ist, anzuzeigen. In anderen Implementierungen können mehrere Prozessoren und/oder mehrere Busse wie jeweils anwendbar zusammen mit mehreren Speichern und Speichertypen verwendet werden. Außerdem können mehrere Berechnungsvorrichtungen 400 verbunden sein, wobei jede Vorrichtung Abschnitte der notwendigen Operationen bereitstellt (z. B. als eine Server-Bank, eine Gruppe von Blade-Servern oder ein Mehrprozessorsystem).

[0100] Der Speicher 404 speichert Informationen innerhalb der Berechnungsvorrichtung 400. In einer Implementierung ist der Speicher 404 ein computerlesbares Medium. In einer Implementierung ist der Speicher 404 eine flüchtige Speichereinheit oder -einheiten. In einer weiteren Implementierung ist der Speicher 404 eine nichtflüchtige Speichereinheit oder -einheiten.

[0101] Die Speichervorrichtung 406 ist zum Bereitstellen von Massenspeicher für die Computervorrichtung 400 fähig. In einer Implementierung ist die Speichervorrichtung 406 ein computerlesbares Medium. In verschiedenen unterschiedlichen Implementierungen kann die Speichervorrichtung 406 eine Diskettenvorrichtung, eine Festplattenvorrichtung, eine optische Plattenvorrichtung oder ein Bandvorrichtung, ein Flash-Speicher oder eine andere ähnlicher Festkörperspeichervorrichtung oder eine Gruppe von Vorrichtungen sein, die Vorrichtungen in einem Speicherbereichsnetz oder anderen Konfigurationen enthält. In einer Implementierung ist ein Computerprogrammprodukt in einem Informationsträger materiell verwirklicht. Das Computerprogrammprodukt enthält Anweisungen, die dann, wenn sie zum Ablauf gebracht werden, ein oder mehrere Verfahren wie z.

B. diejenigen, die vorstehend beschrieben sind, ausführen. Der Informationsträger ist ein computer- oder maschinenlesbares Medium wie z. B. der Speicher 404, die Speichervorrichtung 406 oder der Speicher auf dem Prozessor 402.

[0102] Die Hochgeschwindigkeitssteuereinheit 408 managt bandbreitenintensive Operationen für die Berechnungsvorrichtung 400, während die Niedergeschwindigkeitssteuereinheit 412 Operationen mit geringerer Bandbreitenintensität managt. Eine solche Zuweisung von Aufgaben ist nur beispielhaft. In einer Implementierung ist die Hochgeschwindigkeitssteuereinheit 408 mit dem Speicher 404, der Anzeigevorrichtung 416 (z. B. über einen Grafikprozessor oder -beschleuniger) und Hochgeschwindigkeitserweiterungsanschlüssen 410, die verschiedene Erweiterungskarten aufnehmen können (nicht gezeigt), gekoppelt. In der Implementierung ist die Niedergeschwindigkeitssteuereinheit mit der Speichervorrichtung 406 und dem Niedergeschwindigkeitserweiterungsanschluss 414 gekoppelt. Der Niedergeschwindigkeitserweiterungsanschluss, der verschiedene Kommunikationsanschlüsse (z. B. USB, Bluetooth, Ethernet, drahtloses Ethernet) enthalten kann, kann mit einer oder mehreren Eingabe/Ausgabevorrichtungen, wie z. B. einer Tastatur, einer Zeigevorrichtung, einem Scanner oder einer Vernetzungsvorrichtung wie z. B. einem Verteiler oder einem Router, z. B. über einen Netzadapter, gekoppelt sein.

[0103] Die Berechnungsvorrichtung 400 kann in einer Anzahl unterschiedlicher Formen implementiert sein, wie in der Figur gezeigt ist. Beispielsweise kann sie als ein Standard-Server 420 oder mehrfach in einer Gruppe aus solchen Servern implementiert sein. Sie kann auch als Teil eines Rack-Server-Systems 424 implementiert sein. Zusätzlich kann sie in einem Personalcomputer wie z. B. einem Laptop-Computer 422 implementiert sein. Alternativ können Komponenten aus der Berechnungsvorrichtung 400 mit anderen Komponenten in einer mobilen Vorrichtung (nicht gezeigt) wie z. B. der Vorrichtung 450 kombiniert sein. Jede solcher Vorrichtungen kann eine oder mehrere Berechnungsvorrichtungen 400, 450 beinhalten, und ein Gesamtsystem kann aus mehreren Berechnungsvorrichtungen 400, 450, die miteinander kommunizieren, bestehen.

[0104] Die Berechnungsvorrichtung 450 enthält einen Prozessor 452, einen Speicher 464, eine Eingabe/Ausgabevorrichtung wie z. B. eine Anzeigevorrichtung 454, eine Kommunikationsschnittstelle 466 und einen Sender/Empfänger 468 unter anderen Komponenten. Die Vorrichtung 450 kann außerdem mit einer Speichervorrichtung wie z. B. einem Mikrolaufwerk oder einer anderen Vorrichtung versehen sein, um zusätzliches Speichern bereitzustellen. Jede aus den Komponenten 450, 452, 464, 454, 466 und 468 ist unter Verwendung verschiedener Busse miteinander verbunden, und mehrere der Komponenten können auf einer gemeinsamen Hauptplatine oder auf andere Weise wie jeweils anwendbar montiert sein.

[0105] Der Prozessor 452 kann Anweisungen zur Ausführung innerhalb der Berechnungsvorrichtung 450 verarbeiten, die Anweisungen enthalten, die in dem Speicher 464 gespeichert sind. Der Prozessor kann außerdem getrennte analoge und digitale Prozessoren enthalten. Der Prozessor kann beispielsweise Koordination der anderen Komponenten der Vorrichtung 450 bereitstellen, wie z. B. Steuerung von Anwenderschnittstellen, Anwendungen, die auf der Vorrichtung 450 ablaufen, und drahtloser Kommunikation durch die Vorrichtung 450.

[0106] Der Prozessor 452 kann mit einem Anwender über die Steuerschnittstelle 458 und die Anzeigeschnittstelle 456, die mit einer Anzeigevorrichtung 454 gekoppelt ist, kommunizieren. Die Anzeigevorrichtung 454 kann beispielsweise eine TFT-LCD-Anzeigevorrichtung oder eine OLED-Anzeigevorrichtung oder eine andere geeignete Anzeigetechnologie sein. Die Anzeigeschnittstelle 456 kann eine geeignete Schaltungsanordnung zum Ansteuern der Anzeigevorrichtung 454 sein, um grafische oder andere Informationen für einen Anwender zu präsentieren. Die Steuerschnittstelle 458 kann Befehle von einem Anwender empfangen und sie zur Übertragung zu dem Prozessor 452 umsetzen. Zusätzlich kann eine externe Schnittstelle 462 in Kommunikation mit dem Prozessor 452 bereitgestellt sein, um Nahbereichskommunikation der Vorrichtung 450 mit anderen Vorrichtungen zu ermöglichen. Die externe Schnittstelle 462 kann beispielsweise drahtgebundene Kommunikation (z. B. über eine Docking-Prozedur) oder drahtlose Kommunikation (z. B. über Bluetooth oder andere solche Technologien) bereitstellen.

[0107] Der Speicher 464 speichert Informationen innerhalb der Berechnungsvorrichtung 450. In einer Implementierung ist der Speicher 464 ein computerlesbares Medium. In einer Implementierung ist der Speicher 464 eine flüchtige Speichereinheit oder -einheiten. In einer weiteren Implementierung ist der Speicher 464 eine nichtflüchtige Speichereinheit oder -einheiten. Ein Erweiterungsspeicher 474 kann ebenfalls bereitgestellt und mit der Vorrichtung 450 über die Erweiterungsschnittstelle 472, die beispielsweise eine SIMM-Karten-Schnittstelle enthalten kann, verbunden sein. Ein solcher Erweiterungsspeicher 474 kann zusätzlichen Speicherplatz für die Vorrichtung 450 bereitstellen oder kann außerdem Anwendungen oder andere Informationen für die Vorrichtung 450 speichern. Insbesondere kann der Erweiterungsspeicher 474 Anweisungen enthalten, die vorstehend beschriebenen Prozesse auszufüh-

ren oder zu ergänzen, und kann außerdem sichere Informationen enthalten. Somit kann beispielsweise der Erweiterungsspeicher 474 als ein Sicherheitsmodul für die Vorrichtung 450 bereitgestellt sein und kann mit Anweisungen programmiert sein, die sichere Verwendung der Vorrichtung 450 erlauben. Zusätzlich können sichere Anwendungen über die SIMM-Karten zusammen mit zusätzlichen Informationen bereitgestellt sein, wie z. B. Platzieren von Identifizierungsinformationen auf der SIMM-Karte auf eine nicht hackbare Weise.

[0108] Der Speicher kann beispielsweise Flash-Speicher und/oder MRAM-Speicher enthalten, wie nachstehend diskutiert ist. In einer Implementierung ist ein Computerprogrammprodukt in einem Informationsträger materiell verwirklicht. Das Computerprogrammprodukt enthält Anweisungen, die dann, wenn sie zum Ablauf gebracht werden, ein oder mehrere Verfahren wie z. B. diejenigen, die vorstehend beschrieben sind, ausführen. Der Informationsträger ist ein computer- oder maschinenlesbares Medium wie z. B. der Speicher 464, der Erweiterungsspeicher 474 oder der Speicher auf dem Prozessor 452.

[0109] Die Vorrichtung 450 kann drahtlos über die Kommunikationsschnittstelle 466 kommunizieren, die wenn notwendig eine Schaltungsanordnung zur digitalen Signalverarbeitung enthalten kann. Die Kommunikationsschnittstelle 466 kann Kommunikation unter verschiedenen Arten oder Protokollen bereitstellen, wie z. B. GSM-Sprachanrufe, SMS, EMS- oder MMS-Nachrichtenübermittlung, CDMA, TDMA, PDC, WCDMA, CDMA2020 oder GPRS unter anderen. Eine solche Kommunikation kann beispielsweise über den Hochfrequenz-Sender/Empfänger 468 stattfinden. Zusätzlich kann Nahbereichskommunikation stattfinden, wie z. B. unter Verwendung eines Bluetooth-, WiFi- oder eines anderen solchen Sender/Empfängers (nicht gezeigt). Zusätzlich kann ein GPS-Empfängermodul 470 zusätzliche drahtlose Daten für die Vorrichtung 450 bereitstellen, die wie jeweils anwendbar durch Anwendungen, die auf der Vorrichtung 450 ablaufen, verwendet werden können.

[0110] Die Vorrichtung 450 kann außerdem unter Verwendung eines Audio-Codec 460, der gesprochene Informationen von einem Anwender empfangen und sie in verwendbare digitale Informationen umsetzen kann, hörbar kommunizieren. Der Audio-Codec 460 kann auf ähnliche Weise hörbaren Schall für einen Anwender wie z. B. über einen Lautsprecher, z. B. in einem Kopfhörer der Vorrichtung 450, erzeugen. Ein solcher Schall kann Schall aus Sprachtelefonanrufen enthalten, kann aufgezeichneten Schall (z. B. Sprachnachrichten, Musikdateien usw.) enthalten und kann außerdem Schall enthalten, der durch Anwendungen, die auf der Vorrichtung 450 arbeiten, erzeugt wird.

[0111] Die Berechnungsvorrichtung 450 kann in einer Anzahl unterschiedlicher Formen implementiert sein, wie in der Figur gezeigt ist. Beispielsweise kann sie als ein Mobiltelefon 480 implementiert sein. Sie kann außerdem als ein Teil eines Smartphone 482, eines persönlichen digitalen Assistenten oder einer anderen ähnlichen mobilen Vorrichtung implementiert sein.

[0112] Gemäß beispielhaften Ausführungsformen, Verfahren, Systemen und Einrichtungen, die Computerprogramme enthalten, die auf Computerspeichermedien codiert sind, zum Auswählen von Einheiten zur Sprachsynthese. Eines der Verfahren enthält Empfangen durch einen oder mehreren Computer eines Text-zu-Sprache-Systems von Daten, die Text zur Sprachsynthese angeben; Bestimmen durch den einen oder die mehreren Computer einer Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, wobei die Folge von Texteinheiten wenigstens eine erste Texteinheit gefolgt von einer zweiten Texteinheit enthält; Bestimmen durch den einen oder die mehreren Computer mehrerer Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, wobei das Bestimmen der mehreren Pfade von Spracheinheiten enthält: Auswählen aus einem Spracheinheitenkörper einer ersten Spracheinheit, die Sprachsynthesedaten enthält, die die erste Texteinheit repräsentieren; Auswählen aus dem Spracheinheitenkörper mehrerer zweiter Spracheinheiten, die Sprachsynthesedaten enthalten, die die zweite Texteinheit repräsentieren, wobei jede aus den mehreren zweiten Spracheinheiten basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit einer ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird; und Definieren von Pfaden von der ausgewählten ersten Spracheinheit zu jeder aus den mehreren zweiten Spracheinheiten, die in die mehreren Pfade von Spracheinheiten aufgenommen werden sollen; und Bereitstellen durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems synthetisierter Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist.

[0113] Verschiedene Implementierungen der Systeme und Techniken, die hier beschrieben sind, können in digitaler elektronischer Schaltungsanordnung, integrierter Schaltungsanordnung, speziell konstruierten ASICs (anwendungsspezifischen integrierten Schaltungen), Computer-Hardware, Firmware, Software und/oder Kombinationen daraus realisiert sein. Diese verschiedenen Implementierungen können eine Implementierung in einem oder mehreren Computerprogrammen enthalten, die auf einem programmierbaren System ausführbar oder interpretierbar sind, das wenigstens einen programmierbaren

Prozessor, der ein Spezial- oder Allzweckprozessor sein kann, der gekoppelt ist, um Daten und Anweisungen von einem Speichersystem zu empfangen und Daten und Anweisungen zu ihm zu übertragen, wenigstens eine Eingabevorrichtung und wenigstens eine Ausgabevorrichtung enthält.

[0114] Diese Computerprogramme (auch als Programme, Software, Software-Anwendungen oder Code bezeichnet) enthalten Maschinenanweisungen für einen programmierbaren Prozessor und können in einer prozeduralen Hochsprache und/oder objektorientierten Programmiersprache oder in Assembler/Maschinensprache implementiert sein. Wie hier verwendet beziehen sich die Begriffe „maschinenlesbares Medium“, „computerlesbares Medium“ auf irgendein Computerprogrammprodukt, eine Einrichtung und/oder Vorrichtung (z. B. Magnetplatten, optische Platten, Speicher, programmierbare Logikvorrichtungen (PLDs)), das/die verwendet wird, um Maschinenanweisungen und/oder Daten für einen programmierbaren Prozessor bereitzustellen, einschließlich eines maschinenlesbaren Mediums, das Maschinenanweisungen als ein maschinenlesbares Signal empfängt. Der Begriff „maschinenlesbares Signal“ bezieht sich auf irgendein Signal, das verwendet wird, um Maschinenanweisungen und/oder für einen programmierbaren Prozessor bereitzustellen.

[0115] Um die Interaktion mit einem Anwender bereitzustellen, können die Systeme und Techniken, die hier beschrieben sind, auf einem Computer implementiert sein, der eine Anzeigevorrichtung (z. B. einen CRT- (Kathodenstrahlröhren-) oder LCD-Monitor (Flüssigkristallanzeige-Monitor)) zum Anzeigen von Informationen für den Anwender und eine Tastatur und eine Zeigevorrichtung (z. B. eine Maus oder einen Trackball) durch die der Anwender Eingaben für den Computer bereitstellen kann, aufweist. Andere Arten von Vorrichtungen können verwendet werden, um ebenfalls Interaktion mit Anwender einem bereitzustellen; beispielsweise kann eine für den Anwender bereitgestellte Rückmeldung irgendeine Form sensorischer Rückmeldung sein (z. B. visuelle Rückmeldung, hörbare Rückmeldung oder tastbare Rückmeldung); und eine Eingabe von dem Anwender kann in irgendeiner Form empfangen werden, die akustische, Sprach- oder tastbare Eingabe enthält.

[0116] Die hier beschriebenen Systeme und Techniken können in einem Berechnungssystem implementiert sein, das eine Backend-Komponente (z. B. als ein Daten-Server) enthält oder das eine Middleware-Komponente (z. B. einen Anwendungsserver) enthält, oder der eine Frontend-Komponente (z. B. einen Client-Computer, der eine grafische Anwenderschnittstelle oder einen Web-Browser aufweist, durch den ein Anwender mit einer Implementierung der hier beschriebenen Systeme und Techniken interagieren kann) enthält oder irgendeine Kombination solcher Backend-, Middleware- oder Frontend-Komponenten. Die Komponenten des Systems können durch irgendeine Form oder irgendein Medium zur digitalen Datenkommunikation (z. B. ein Kommunikationsnetz) miteinander verbunden sein. Beispiele für Kommunikationsnetze enthalten ein lokales Netz („LAN“), ein Weitbereichsnetz („WAN“) und das Internet.

[0117] Das Berechnungssystem kann Clients und Server enthalten. Ein Client und ein Server sind im Allgemeinen voneinander entfernt und interagieren typischerweise über ein Kommunikationsnetz. Die Beziehung von Client und Server entsteht aufgrund der Computerprogramme, die auf den jeweiligen Computern laufen und eine Client-Server-Beziehung miteinander aufweisen.

[0118] Obwohl diese Spezifikation viele spezifische Implementierungseinzelheiten beinhaltet, sollten diese nicht als Einschränkungen für den Schutzbereich dessen, was beansprucht sein kann, gedeutet werden, sondern vielmehr als Beschreibungen von Merkmalen, die für spezielle Ausführungsformen spezifisch sein können. Spezielle Merkmale, die in dieser Spezifikation im Kontext getrennter Ausführungsformen beschrieben sind, können auch in Kombination in einer einzigen Ausführungsform implementiert sein. Umgekehrt können verschiedene Merkmale, die im Kontext einer einzigen Ausführungsform beschrieben sind, auch in mehreren Ausführungsformen getrennt oder in irgendeiner geeigneten Unterkombination implementiert sein. Außerdem können, obwohl Merkmale vorstehend als in speziellen Kombinationen arbeitend beschrieben und anfangs sogar als solche beansprucht sind, ein oder mehrere Merkmale aus einer beanspruchten Kombination in einigen Fällen aus der Kombination herausgenommen sein, und die beanspruchte Kombination kann sich auf eine Unterkombination oder eine Variation einer Unterkombination richten.

[0119] Ähnlich sollte, obwohl Operationen in den Zeichnungen in einer speziellen Reihenfolge abgebildet sind, das nicht so verstanden werden, dass es erforderlich ist, dass solche Operationen in der speziellen gezeigten Reihenfolge oder in sequentieller Reihenfolge ausgeführt werden oder dass alle dargestellten Operationen ausgeführt werden, um wünschenswerte Ergebnisse zu erreichen. Unter speziellen Umständen können Multitasking und Parallelverarbeitung vorteilhaft sein. Außerdem sollte die Trennung verschiedener Systemmodule und Komponenten in den vorstehend beschriebenen Ausführungsformen nicht so verstanden werden, dass eine solche Trennung in allen Ausführungsformen erforderlich ist, und es sollte verstanden werden, dass die beschriebenen Programmkomponen-

ten und Systeme im Allgemeinen gemeinsam in einem einzigen Software-Produkt oder in mehrere Software-Produkte paketiert integriert sein können.

[0120] Spezielle Ausführungsformen des Gegenstands sind beschrieben worden. Andere Ausführungsformen sind innerhalb des Schutzbereichs der folgenden Ansprüche. Beispielsweise können die Aktionen, die in den Ansprüchen vorgetragen sind, in einer anderen Reihenfolge ausgeführt werden und immer noch wünschenswerte Ergebnisse erreichen. Als ein Beispiel erfordern die in den begleitenden Figuren abgebildeten Prozesse nicht notwendigerweise die spezielle gezeigte Reihenfolge oder sequentielle Reihenfolge, um wünschenswerte Ergebnisse zu erreichen. In einigen Fällen können Multitasking und Parallelverarbeitung vorteilhaft sein.

Patentansprüche

1. Nichttransitorisches Computerspeichermedium, das mit Anweisungen codiert ist, die dann, wenn sie durch einen oder mehrere Computer eines Text-zu-Sprache-Systems (116) ausgeführt werden, bewirken, dass der eine oder die mehreren Computer Operationen ausführen, die umfassen:
Empfangen (302) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems von Daten, die Text zur Sprachsynthese angeben;
Bestimmen (304) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems einer Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, wobei die Folge von Texteinheiten wenigstens eine erste Texteinheit gefolgt von einer zweiten Texteinheit enthält;
Bestimmen (306) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems mehrerer Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, wobei das Bestimmen der mehreren Pfade von Spracheinheiten umfasst:
Auswählen (308) aus einem Spracheinheitenkörper (124) einer vorbestimmten Menge L erster Spracheinheiten (202a-202f), die Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren; und
Definieren von Pfaden für eine vorbestimmte Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) durch:
Auswählen (310), für jede erste Spracheinheit der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f), einer vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, aus dem Spracheinheitenkörper (124), wobei jede zweite Spracheinheit der vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f) basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit der jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird; und
Definieren (314) von Pfaden von jeder der ersten Spracheinheiten der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) zu jeder zweiten Spracheinheit der jeweiligen vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die in die mehreren Pfade von Spracheinheiten aufgenommen werden sollen, wobei zu Pfaden, die eine in der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) enthaltene erste Spracheinheit aufweisen, die nicht von der vorbestimmten Anzahl K (202a-202c) umfasst ist, keine zusätzlichen Spracheinheiten hinzugefügt werden; und
Bereitstellen (316) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems synthetisierter Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist.

2. Computerspeichermedium nach Anspruch 1, wobei das Bestimmen der Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, Bestimmen der Folge von Texteinheiten umfasst, die jeweils einen unterscheidbaren Abschnitt des Texts repräsentieren, getrennt von den Abschnitten des Texts, die durch die anderen Texteinheiten repräsentiert sind.

3. Computerspeichermedium nach Anspruch 1 oder 2, wobei das Bereitstellen der synthetisierten Sprachdaten gemäß dem Pfad, der aus den mehreren Pfaden ausgewählt ist, Bereitstellen der synthetisierten Sprachdaten, um zu bewirken, dass eine Vorrichtung hörbare Daten für den Text erzeugt, umfasst.

4. Computerspeichermedium nach einem der Ansprüche 1 bis 3, wobei die Operationen umfassen:
Auswählen aus dem Spracheinheitenkörper (124) von zwei oder mehreren Anfangsspracheinheiten (202a-202c), die jeweils Sprachsynthesedaten umfassen, die eine Anfangstexteinheit in der Folge von Texteinheiten mit einem Ort an einem Anfang der Textfolge repräsentieren.

5. Computerspeichermedium nach Anspruch 4, wobei:
das Auswählen der zwei oder mehr Anfangsspracheinheiten (202a-202c) Auswählen einer vorbestimmten Menge von Anfangsspracheinheiten umfasst; und
Bestimmen der mehreren Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, Bestimmen einer vorbestimmten Menge

von Pfaden umfasst, wobei die Operationen umfassen:
Auswählen aus der vorbestimmten Menge von Pfaden des Pfads, für den die synthetisierten Sprachdaten bereitgestellt werden sollen.

6. Computerspeichermedium nach Anspruch 5, wobei:
die mehreren zweiten Spracheinheiten (204a-204b) zwei oder mehr zweite Spracheinheiten umfassen.

7. Computerspeichermedium nach Anspruch 6, wobei die Operationen umfassen:
Auswählen für die erste Texteinheit einer vorbestimmten Menge erster Spracheinheiten (202a-202f), die jeweils Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren; und
Auswählen für die zweite Texteinheit einer vorbestimmten Menge zweiter Spracheinheiten (204a-204f), die jeweils Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, wobei jede aus der vorbestimmten Menge zweiter Spracheinheiten basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit einer jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird.

8. Computerspeichermedium nach Anspruch 7, wobei die Operationen umfassen:
Bestimmen für eine zweite vorbestimmte Menge zweiter Spracheinheiten, die jeweils Sprachsynthesedaten umfassen, die die zweite Einheit repräsentieren, (i) eines Verknüpfungsaufwands, um die zweite Spracheinheit mit einer jeweiligen ersten Spracheinheit zu verketten, und (ii) eines Zielaufwands, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, wobei:
die zweite vorbestimmte Menge größer ist als die vorbestimmte Anzahl; und
das Auswählen der vorbestimmten Menge zweiter Spracheinheiten (204a-204f) Auswählen der vorbestimmten Menge zweiter Spracheinheiten aus der zweiten vorbestimmten Menge zweiter Spracheinheiten unter Verwendung der bestimmten Verknüpfungsaufwände und der bestimmten Zielaufwände umfasst.

9. Computerspeichermedium nach einem der Ansprüche 4 bis 8, wobei:
die erste Texteinheit einen ersten Ort in der Folge von Texteinheiten aufweist;
die zweite Texteinheit einen zweiten Ort in der Folge von Texteinheiten aufweist, der dem ersten Ort ohne irgendwelche dazwischenliegenden Orte nachfolgt; und
das Auswählen aus dem Spracheinheitenkörper (124) von mehreren zweiten Spracheinheiten Auswählen aus dem Spracheinheitenkörper der mehreren zweiten Spracheinheiten unter Verwendung (i) eines Verknüpfungsaufwands, um die zweite Spracheinheit mit Daten für die ersten Spracheinheit und einer entsprechenden Anfangsspracheinheit aus den zwei oder mehr Anfangsspracheinheiten zu verketten, und (ii) eines Zielaufwands, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, umfasst.

10. Computerspeichermedium nach Anspruch 9, wobei die Operationen umfassen:
Bestimmen eines Pfads, der eine ausgewählte Spracheinheit enthält, für jede aus den Texteinheiten in der Folge von Texteinheiten bis zu dem ersten Ort, wobei die ausgewählten Spracheinheiten die erste Spracheinheit und die entsprechende Anfangsspracheinheit enthalten;
Bestimmen erster Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad; und
Bestimmen für jede aus den mehreren zweiten Spracheinheiten des Verknüpfungsaufwands unter Verwendung der ersten Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad und zweiter Akustikparameter für die zweite Spracheinheit.

11. Computerspeichermedium nach Anspruch 10, wobei das Bestimmen für jede aus den mehreren zweiten Spracheinheiten des Verknüpfungsaufwands gleichzeitiges Bestimmen für jede aus zwei oder mehr zweiten Spracheinheiten des Verknüpfungsaufwands unter Verwendung der ersten Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad und zweiter Akustikparameter für die zweite Spracheinheit umfasst.

12. Text-zu-Sprache-System (116), das einen oder mehrere Computer und eine oder mehrere Speichervorrichtungen umfasst, auf denen Anweisungen gespeichert sind, die arbeiten, wenn sie durch den einen oder die mehreren Computer ausgeführt werden, um zu bewirken, dass der eine oder die mehreren Computer Operationen ausführen, die umfassen:
Empfangen (302) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems von Daten, die Text zur Sprachsynthese angeben;
Bestimmen (304) durch denen einen oder die mehreren Computer des Text-zu-Sprache-Systems einer Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, wobei die Folge von Texteinheiten wenigstens eine erste Texteinheit gefolgt von einer zweiten Texteinheit enthält;
Bestimmen (306) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems mehrerer Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, wobei das

Bestimmen der mehreren Pfade von Spracheinheiten umfasst:
Auswählen (308) aus einem Spracheinheitenkörper (124) einer vorbestimmten Menge L erster Spracheinheiten (202a-202f), die Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren; und
Definieren von Pfaden für eine vorbestimmte Anzahl K(202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) durch:
Auswählen (310), für jede erste Spracheinheit der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f), einer vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, aus dem Spracheinheitenkörper (124), wobei jede zweite Spracheinheit der vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f) basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit der jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird; und
Definieren (314) von Pfaden von jeder der ersten Spracheinheiten der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) zu jeder zweiten Spracheinheit der jeweiligen vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die in die mehreren Pfade von Spracheinheiten aufgenommen werden sollen, wobei zu Pfaden, die eine in der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) enthaltene erste Spracheinheit aufweisen, die nicht von der vorbestimmten Anzahl K (202a-202c) umfasst ist, keine zusätzlichen Spracheinheiten hinzugefügt werden; und
Bereitstellen (316) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems synthetisierter Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist.

13. Text-zu-Sprache-System (116) nach Anspruch 12, wobei das Bestimmen der Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, Bestimmen der Folge von Texteinheiten, die jeweils einen unterscheidbaren Abschnitt des Texts enthalten, getrennt von den Abschnitten des Texts, die durch die anderen Texteinheiten repräsentiert sind, umfasst.

14. Text-zu-Sprache-System (116) nach Anspruch 12 oder 13, wobei das Bereitstellen der synthetisierten Sprachdaten gemäß dem Pfad, der aus den mehreren Pfaden ausgewählt ist, Bereitstellen der synthetisierten Sprachdaten, um zu bewirken, dass eine Vorrichtung hörbare Daten für den Text erzeugt, umfasst.

15. Text-zu-Sprache-System (116) nach einem der Ansprüche 12 bis 14, wobei die Operationen umfassen:
Auswählen aus dem Spracheinheitenkörper (124) von zwei oder mehreren Anfangsspracheinheiten (202a-202c), die jeweils Sprachsynthesedaten umfassen, die eine Anfangstexteinheit in der Folge von Texteinheiten mit einem Ort an einem Anfang der Textfolge repräsentieren.

16. Text-zu-Sprache-System (116) nach Anspruch 15, wobei:
das Auswählen der zwei oder mehr Anfangsspracheinheiten (202a-202c) Auswählen einer vorbestimmten Menge von Anfangsspracheinheiten umfasst; und
Bestimmen der mehreren Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, Bestimmen einer vorbestimmten Menge von Pfaden umfasst, wobei die Operationen umfassen:
Auswählen aus der vorbestimmten Menge von Pfaden des Pfads, für den die synthetisierten Sprachdaten bereitgestellt werden sollen.

17. Text-zu-Sprache-System (116) nach Anspruch 16, wobei:
die mehreren zweiten Spracheinheiten zwei oder mehr zweite Spracheinheiten umfassen.

18. Text-zu-Sprache-System (116) nach Anspruch 17, wobei die Operationen umfassen:
Auswählen für die erste Texteinheit einer vorbestimmten Menge erster Spracheinheiten (202a-202f), die jeweils Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren; und
Auswählen für die zweite Texteinheit einer vorbestimmten Menge zweiter Spracheinheiten (204a-204f), die jeweils Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, wobei jede aus der vorbestimmten Menge zweiter Spracheinheiten basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit einer jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird.

19. Text-zu-Sprache-System (116) nach Anspruch 18, wobei die Operationen umfassen:
Bestimmen für eine zweite vorbestimmte Menge zweiter Spracheinheiten, die jeweils Sprachsynthesedaten umfassen, die die zweite Einheit repräsentieren, (i) eines Verknüpfungsaufwands, um die zweite Spracheinheit mit einer jeweiligen ersten Spracheinheit zu verketten, und (ii) eines Zielauf-

wands, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, wobei:
die zweite vorbestimmte Menge größer ist als die vorbestimmte Menge; und
das Auswählen der vorbestimmten Menge zweiter Spracheinheiten Auswählen der vorbestimmten Menge zweiter Spracheinheiten aus der zweiten vorbestimmten Menge zweiter Spracheinheiten unter Verwendung der bestimmten Verknüpfungsaufwände und der bestimmten Zielaufwände umfasst.

20. Text-zu-Sprache-System (116) nach einem der Ansprüche 15 bis 19, wobei:
die erste Texteinheit einen ersten Ort in der Folge von Texteinheiten aufweist;
die zweite Texteinheit einen zweiten Ort in der Folge von Texteinheiten aufweist, der dem ersten Ort ohne irgendwelche dazwischenliegenden Orte nachfolgt; und
das Auswählen aus dem Spracheinheitenkörper (124) von mehreren zweiten Spracheinheiten Auswählen aus dem Spracheinheitenkörper (124) der mehreren zweiten Spracheinheiten unter Verwendung (i) eines Verknüpfungsaufwands, um die zweite Spracheinheit mit Daten für die ersten Spracheinheit und einer entsprechenden Anfangsspracheinheit aus den zwei oder mehr Anfangsspracheinheiten zu verketten, und (ii) eines Zielaufwands, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, umfasst.

21. Text-zu-Sprache-System (116) nach Anspruch 20, wobei die Operationen umfassen:
Bestimmen eines Pfads, der eine ausgewählte Spracheinheit enthält, für jede aus den Texteinheiten in der Folge von Texteinheiten bis zu dem ersten Ort, wobei die ausgewählten Spracheinheiten die erste Spracheinheit und die entsprechende Anfangsspracheinheit enthalten;
Bestimmen erster Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad; und
Bestimmen für jede aus den mehreren zweiten Spracheinheiten des Verknüpfungsaufwands unter Verwendung der ersten Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad und zweiter Akustikparameter für die zweite Spracheinheit.

22. Text-zu-Sprache-System (116) nach Anspruch 21, wobei das Bestimmen für jede aus den mehreren zweiten Spracheinheiten des Verknüpfungsaufwands gleichzeitiges Bestimmen für jede aus zwei oder mehr zweiten Spracheinheiten des Verknüpfungsaufwands unter Verwendung der ersten Akustikparameter für jede aus den ausgewählten Spracheinheiten in dem Pfad und zweiter Akustikparameter für die zweite Spracheinheit umfasst.

23. Computerimplementiertes Verfahren, das umfasst:
Empfangen (302) durch einen oder mehrere Computer des Text-zu-Sprache-Systems von Daten, die Text zur Sprachsynthese angeben;
Bestimmen (304) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems einer Folge von Texteinheiten, die jeweils einen jeweiligen Abschnitt des Texts repräsentieren, wobei die Folge von Texteinheiten wenigstens eine erste Texteinheit gefolgt von einer zweiten Texteinheit enthält;
Bestimmen (306) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems mehrerer Pfade von Spracheinheiten, die jeweils die Folge von Texteinheiten repräsentieren, wobei das Bestimmen der mehreren Pfade von Spracheinheiten umfasst:
Auswählen (308) aus einem Spracheinheitenkörper (124) einer vorbestimmten Menge L erster Spracheinheiten (202a-202f), die Sprachsynthesedaten umfassen, die die erste Texteinheit repräsentieren; und
Definieren von Pfaden für eine vorbestimmte Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) durch:
Auswählen (310), für jede erste Spracheinheit der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f), einer vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, aus dem Spracheinheitenkörper (124), wobei jede zweite Spracheinheit der vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f) basierend auf (i) einem Verknüpfungsaufwand, um die zweite Spracheinheit mit der jeweiligen ersten Spracheinheit zu verketten, und (ii) einem Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit der zweiten Texteinheit entspricht, bestimmt wird; und
Definieren (314) von Pfaden von jeder der ersten Spracheinheiten der vorbestimmten Anzahl K (202a-202c) der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) zu jeder zweiten Spracheinheit der jeweiligen vorbestimmten Anzahl größer eins von zweiten Spracheinheiten (204a-204b, 204c-204d, 204e-204f), die in die mehreren Pfade von Spracheinheiten aufgenommen werden sollen, wobei zu Pfaden, die eine in der vorbestimmten Menge L der ersten Spracheinheiten (202a-202f) enthaltene erste Spracheinheit aufweisen, die nicht von der vorbestimmten Anzahl K (202a-202c) umfasst ist, keine zusätzlichen Spracheinheiten hinzugefügt werden; und
Bereitstellen (316) durch den einen oder die mehreren Computer des Text-zu-Sprache-Systems syn-

thetisierter Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist.

Es folgen 4 Seiten Zeichnungen

Anhängende Zeichnungen

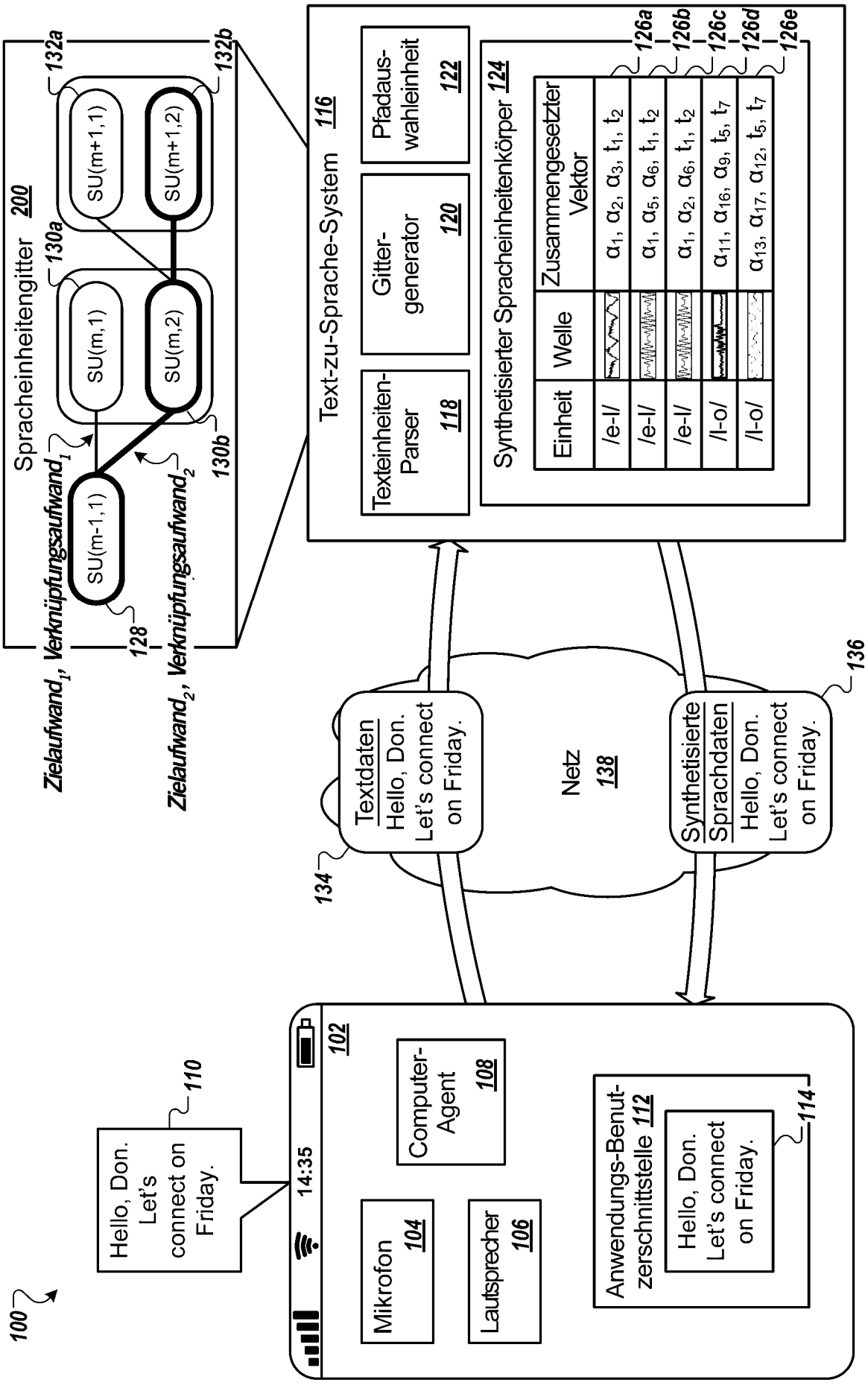


FIG. 1

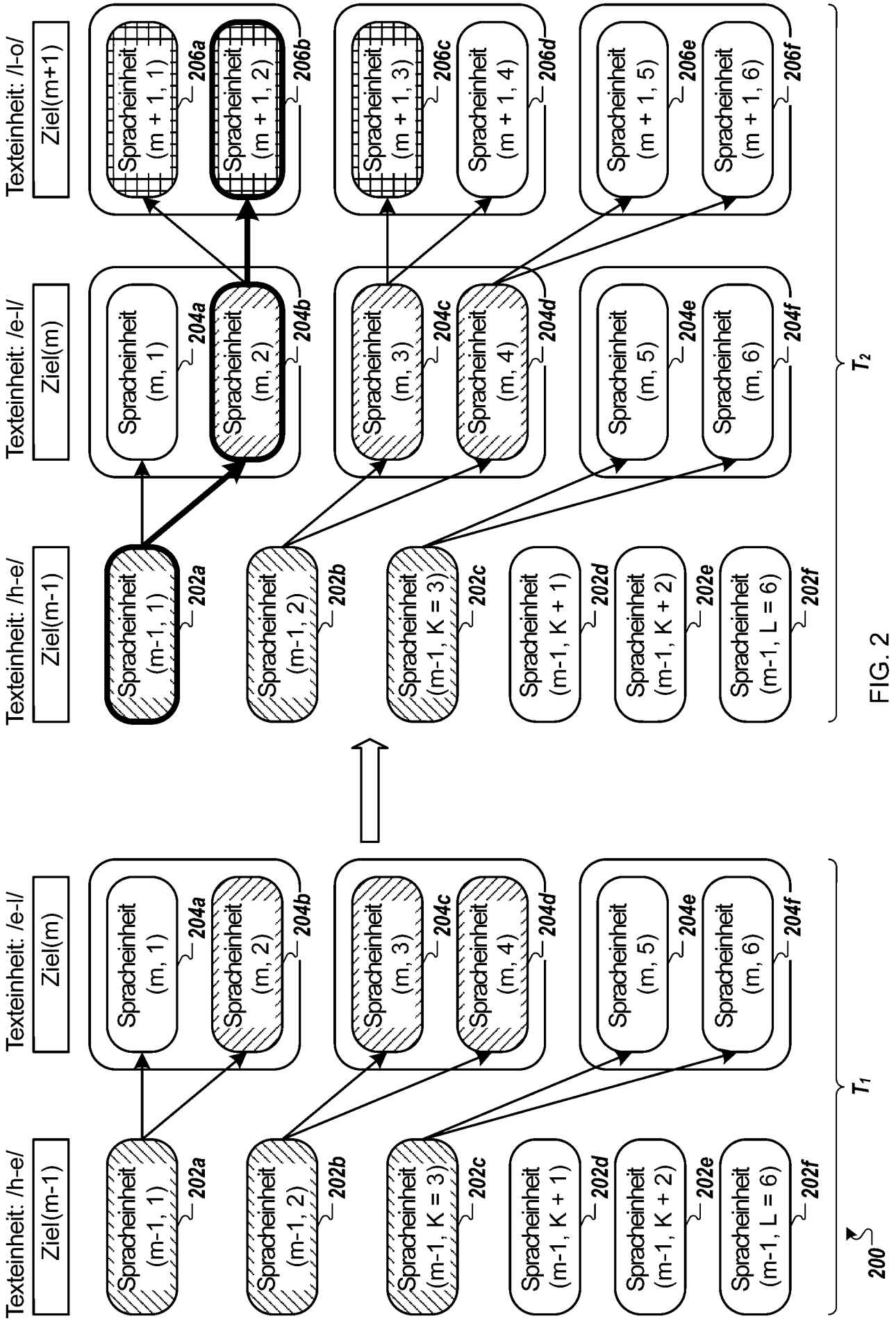

Texteinheit: /h-e/
Ziel(m-1)
Texteinheit: /e-l/
Ziel(m)
Texteinheit: /l-o/
Ziel(m+1)
Spracheinheit (m-1, 1)
202a
Spracheinheit (m-1, 2)
202b
Spracheinheit (m-1, K = 3)
202c
Spracheinheit (m-1, K + 1)
202d
Spracheinheit (m-1, K + 2)
202e
Spracheinheit (m-1, L = 6)
202f
Spracheinheit (m, 1)
204a
Spracheinheit (m, 2)
204b
Spracheinheit (m, 3)
204c
Spracheinheit (m, 4)
204d
Spracheinheit (m, 5)
204e
Spracheinheit (m, 6)
204f
Spracheinheit (m + 1, 1)
206a
Spracheinheit (m + 1, 2)
206b
Spracheinheit (m + 1, 3)
206c
Spracheinheit (m + 1, 4)
206d
Spracheinheit (m + 1, 5)
206e
Spracheinheit (m + 1, 6)
206f
T1
T2
200

FIG. 2

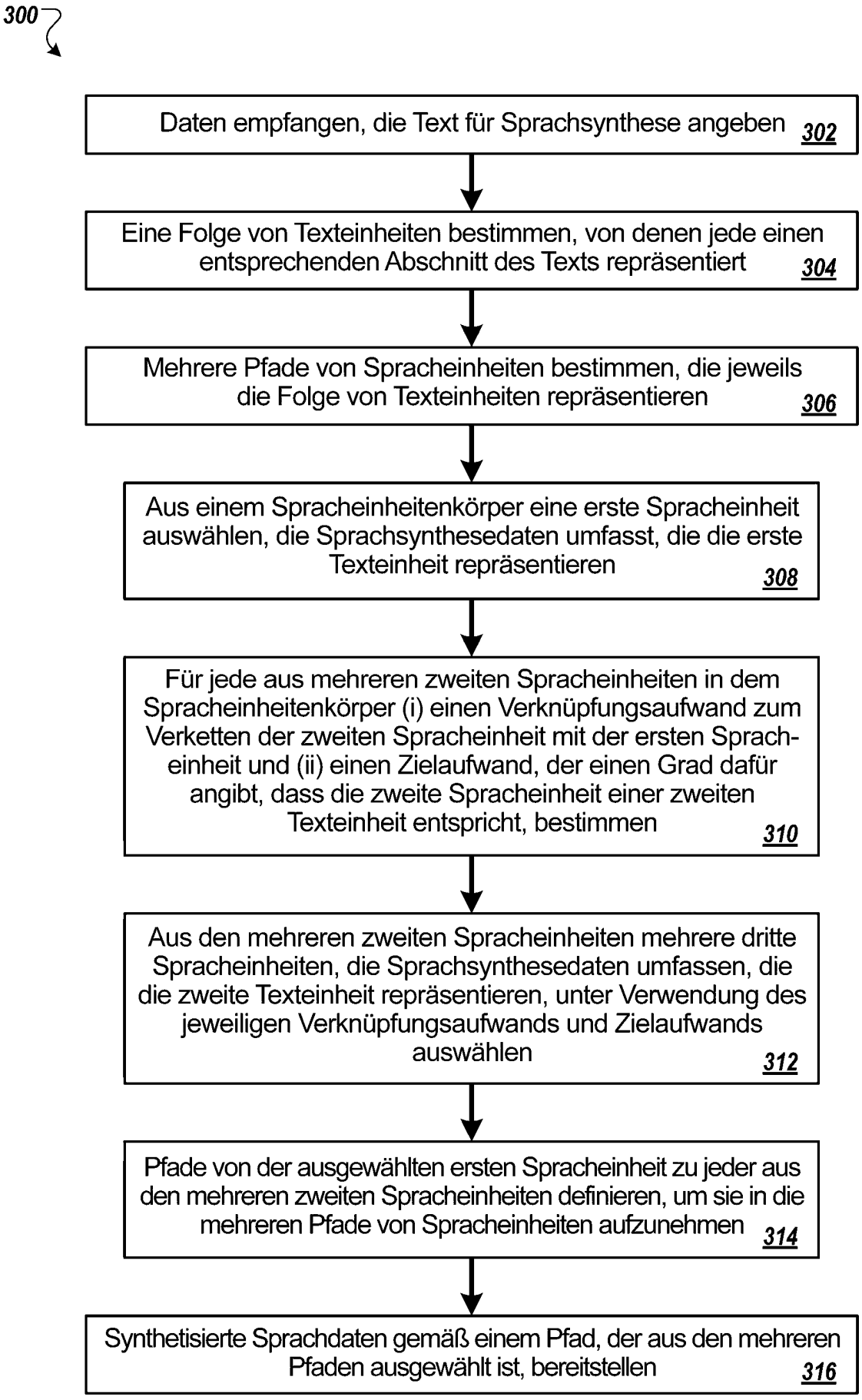

300
Daten empfangen, die Text für Sprachsynthese angeben 302
Eine Folge von Texteinheiten bestimmen, von denen jede einen entsprechenden Abschnitt des Texts repräsentiert 304
Mehrere Pfade von Spracheinheiten bestimmen, die jeweils die Folge von Texteinheiten repräsentieren 306
Aus einem Spracheinheitenkörper eine erste Spracheinheit auswählen, die Sprachsynthesedaten umfasst, die die erste Texteinheit repräsentieren 308
Für jede aus mehreren zweiten Spracheinheiten in dem Spracheinheitenkörper (i) einen Verknüpfungsaufwand zum Verketten der zweiten Spracheinheit mit der ersten Sprach-einheit und (ii) einen Zielaufwand, der einen Grad dafür angibt, dass die zweite Spracheinheit einer zweiten Texteinheit entspricht, bestimmen 310
Aus den mehreren zweiten Spracheinheiten mehrere dritte Spracheinheiten, die Sprachsynthesedaten umfassen, die die zweite Texteinheit repräsentieren, unter Verwendung des jeweiligen Verknüpfungsaufwands und Zielaufwands auswählen 312
Pfade von der ausgewählten ersten Spracheinheit zu jeder aus den mehreren zweiten Spracheinheiten definieren, um sie in die mehreren Pfade von Spracheinheiten aufzunehmen 314
Synthetisierte Sprachdaten gemäß einem Pfad, der aus den mehreren Pfaden ausgewählt ist, bereitstellen 316

FIG. 3

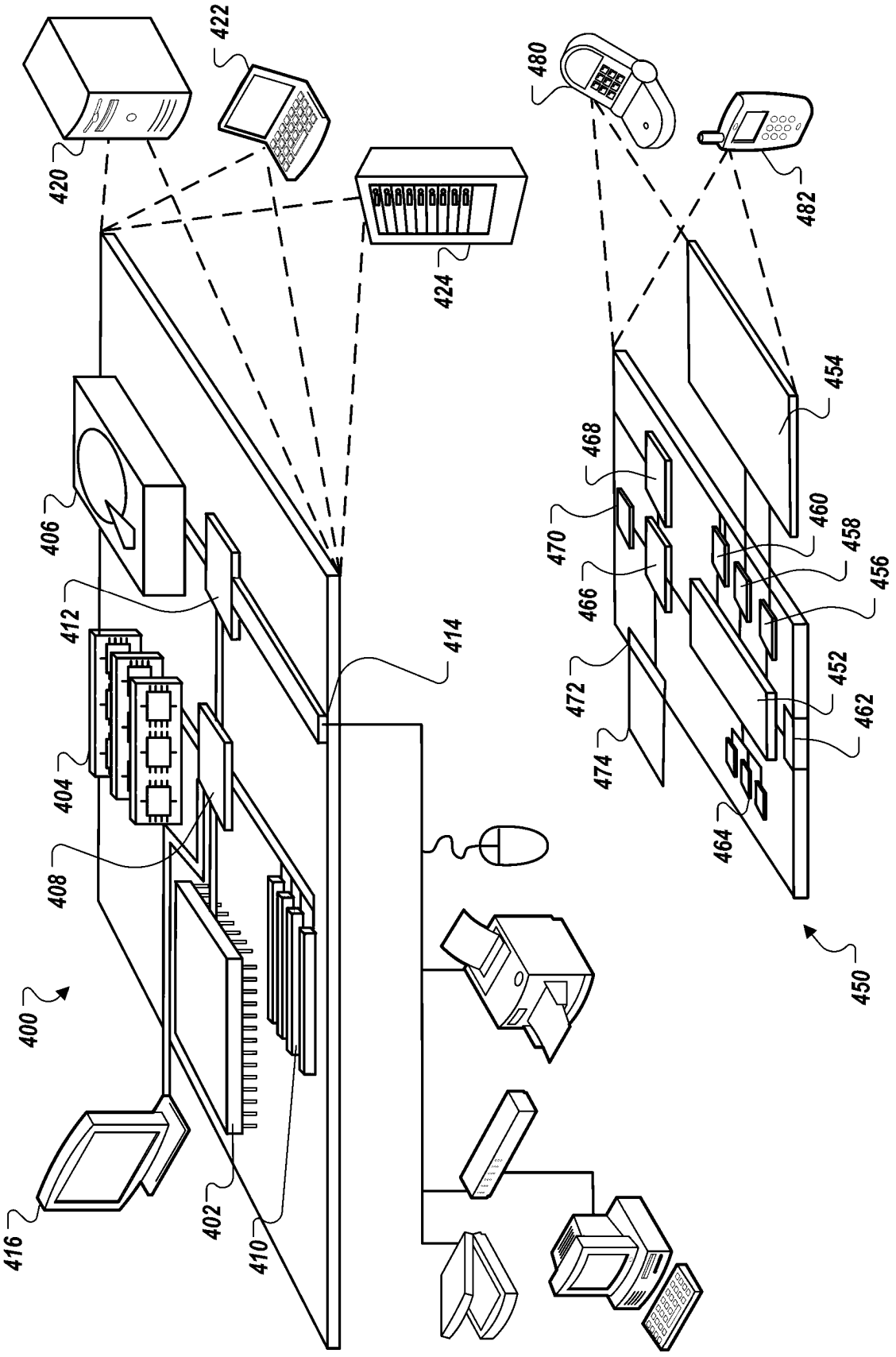


FIG. 4

