(19) Deutsches Patent- und Markenamt

(10) **DE 10 2017 124 264 B4** 2021.08.05

(12) # Patentschrift

(21) Aktenzeichen: **10 2017 124 264.3**
(22) Anmeldetag: **18.10.2017**
(43) Offenlegungstag: **26.04.2018**
(45) Veröffentlichungstag der Patenterteilung: **05.08.2021**

(51) Int Cl.: ***G10L 15/06*** (2013.01)
G10L 13/00 (2006.01)
G10L 15/26 (2006.01)

Innerhalb von neun Monaten nach Veröffentlichung der Patenterteilung kann nach § 59 Patentgesetz gegen das Patent Einspruch erhoben werden. Der Einspruch ist schriftlich zu erklären und zu begründen. Innerhalb der Einspruchsfrist ist eine Einspruchsgebühr in Höhe von 200 Euro zu entrichten (§ 6 Patentkostengesetz in Verbindung mit der Anlage zu § 2 Abs. 1 Patentkostengesetz).

(30) Unionspriorität:
62/410,564 **20.10.2016** **US**

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

(74) Vertreter:
Venner Shipley LLP, 85521 Ottobrunn, DE

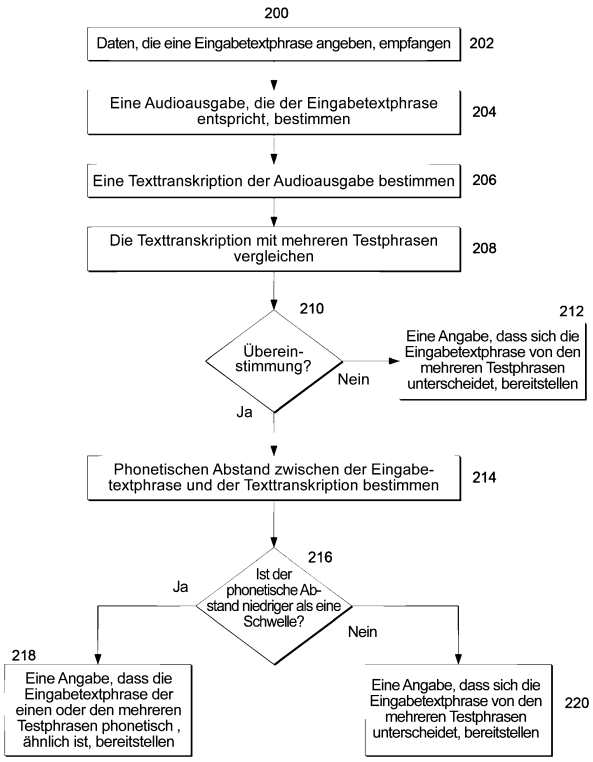
(72) Erfinder:
Rao, Nikhil Chandru, Mountain View, Calif., US;
Krishnakumaran, Saisuresh, Mountain View, Calif., US

(56) Ermittelter Stand der Technik:

US	**9 292 487**	**B1**
US	**2011 / 0 106 792**	**A1**

(54) Bezeichnung: **Computerimplementiertes Verfahren und Rechensystem zum Bestimmen phonetischer Beziehungen**

(57) Hauptanspruch: Computerimplementiertes Verfahren (200) zum Bestimmen einer phonetischen Beziehung zwischen zwei oder mehr Phrasen, wobei das Verfahren umfasst:
Empfangen (202) von Daten, die eine von einem Anwender eingegebene Eingabetextphrase angeben, durch eine oder mehrere Rechenvorrichtungen (104, 310);
Bestimmen (204) einer Audioausgabe, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht, durch die eine oder die mehreren Rechenvorrichtungen;
Bestimmen (206) einer Texttranskription der Audioausgabe der Eingabetextphrase durch die eine oder die mehreren Rechenvorrichtungen, wobei die Texttranskription eine textuelle Repräsentation der Audioausgabe umfasst; und
Vergleichen (208) der Texttranskription mit mehreren Testphrasen durch die eine oder die mehreren Rechenvorrichtungen, um eine Übereinstimmung zwischen der Texttranskription und mindestens einer Testphrase (112) zu identifizieren.



Beschreibung

Gebiet

[0001] Die vorliegende Erfindung bezieht sich im Allgemeinen auf ein Bestimmen phonetischer Beziehungen zwischen zwei oder mehr Phrasen.

Hintergrund

[0002] Die Spracherkennung ist eine weit verbreitete und häufig verwendete Art der Interaktion mit Rechenvorrichtungen geworden. Spracheingaben können bequemer und effizienter sein als herkömmliche Eingabearten wie etwa das Tippen auf einer Tastatur. Zum Beispiel können mobile Rechenvorrichtungen, alternativ zum Eingeben von Zeichen über eine virtuelle Tastatur auf einem Berührungsbildschirm, Spracherkennungsdienste als Eingabeart anbieten. Einige Rechenvorrichtungen sind dazu ausgelegt, Sprachbefehle von einem Anwender zu akzeptieren, beispielsweise als abgekürzter Weg zum Ausführen bestimmter Aktionen auf der Rechenvorrichtung. So können solche Rechenvorrichtungen dazu ausgelegt sein, die Sprachbefehle zu interpretieren und eine oder mehrere Aktionen basierend auf den Interpretationen der Sprachbefehle durchzuführen.

[0003] In einigen Fällen können Entwickler von Rechenanwendungen oder -diensten Sprachbefehle auswählen, die von einem oder mehreren Anwendern verwendet werden können, um die Rechenanwendung oder den Rechendienst auf einer Rechenvorrichtung aufzurufen. Es kann wichtig sein, sicherzustellen, dass solche ausgewählten Sprachbefehle phonetisch von anderen Sprachbefehlen verschieden sind, die bereits zum Aufrufen anderer Rechenanwendungen oder -dienste reserviert sind. So kann es vorteilhaft sein, einen Entwickler oder eine andere Partei darauf hinzuweisen, ob ein von dem Entwickler oder einer anderen Partei ausgewählter Sprachbefehl phonetisch einem oder mehreren reservierten Sprachbefehlen ähnelt. Die US 2011/0106792 A1 stellt ein Verfahren zum Abrufen ähnlich klingender Wörter aus einer elektronischen Datenbank bereit. Die US 9,292,487 B1 stellt ein Sprachmodell zur Spracherkennung bereit, das unterschiedlich beschnitten wird.

Zusammenfassung

[0004] Aspekte und Vorteile von Ausführungsformen der vorliegenden Offenbarung werden teilweise in der folgenden Beschreibung dargelegt oder können aus der Beschreibung oder durch die Umsetzung der Ausführungsformen erlernt werden.

[0005] Ein beispielhafter Aspekt der vorliegenden Offenbarung betrifft ein computerimplementiertes Verfahren zum Bestimmen einer phonetischen Beziehung zwischen zwei oder mehr Phrasen. Das Verfahren umfasst ein Empfangen von Daten, die eine von einem Anwender eingegebene Eingabetextphrase angeben, durch eine oder mehrere Rechenvorrichtungen. Das Verfahren umfasst ferner ein Bestimmen einer Audioausgabe, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht, durch die eine oder die mehreren Rechenvorrichtungen. Das Verfahren umfasst ferner ein Bestimmen einer Texttranskription der Audioausgabe der eingegebenen Textphrase durch die eine oder die mehreren Rechenvorrichtungen. Die Texttranskription enthält eine textuelle Darstellung der Audioausgabe. Das Verfahren umfasst ferner ein Vergleichen der Texttranskription mit mehreren Testphrasen durch das eine oder die mehreren Rechenvorrichtungen, um eine Übereinstimmung zwischen der Texttranskription und mindestens einer Testphrase zu identifizieren.

[0006] Andere beispielhafte Aspekte der vorliegenden Offenbarung betreffen Systeme, Vorrichtungen, konkrete, nichttransitorische computerlesbare Medien, Anwenderschnittstellen, Speichervorrichtungen und elektronische Vorrichtungen zum Bestimmen phonetischer Beziehungen.

[0007] Diese und andere Merkmale, Aspekte und Vorteile verschiedener Ausführungsformen sind unter Bezugnahme auf die folgende Beschreibung und die beigefügten Ansprüche besser zu verstehen. Die beigefügten Zeichnungen, die in diese Beschreibung aufgenommen sind und einen Teil Dieser Text wurde durch das DPMA aus Originalquellen übernommen. Er enthält keine Zeichnungen. Die Darstellung von Tabellen und Formeln kann unbefriedigend sein. davon bilden, veranschaulichen Ausführungsformen der vorliegenden Offenbarung und dienen zusammen mit der Beschreibung dazu, die zugehörigen Prinzipien zu erläutern.

Figurenliste

[0008] Eine genaue Erörterung der Ausführungsformen, die an Fachleute gerichtet ist, wird in der genauen Beschreibung, die auf die angehängten Figuren Bezug nimmt, gegeben, wobei:

Fig. 1 einen Überblick über ein beispielhaftes System zum Bestimmen phonetischer Beziehungen gemäß Ausführungsbeispielen der vorliegenden Offenbarung zeigt;

Fig. 2 eine beispielhafte Anwenderschnittstelle gemäß Ausführungsbeispielen der vorliegenden Offenbarung zeigt;

Fig. 3 ein Ablaufdiagramm eines beispielhaften Verfahrens zum Bestimmen phonetischer Bezie-

hungen gemäß Ausführungsbeispielen der vorliegenden Erfindung zeigt; und

Fig. 4 ein beispielhaftes System gemäß Ausführungsbeispielen der vorliegenden Offenbarung zeigt.

Genaue Beschreibung

[0009] Es wird nun genau auf die Ausführungsformen, von denen ein oder mehrere Beispiele in den Zeichnungen gezeigt sind, Bezug genommen. Jedes Beispiel wird beispielhaft für die Ausführungsformen gegeben und soll nicht für die vorliegende Erfindung einschränkend sein. Tatsächlich wird es für Fachleute offensichtlich sein, dass verschiedene Abwandlungen und Variationen an den Ausführungsformen vorgenommen werden können, ohne vom Geltungsbereich oder Gedanken der vorliegenden Erfindung abzuweichen. Zum Beispiel können als Teil einer Ausführungsform gezeigte oder beschriebene Merkmale zusammen mit einer weiteren Ausführungsform verwendet werden, um noch eine weitere Ausführungsform zu erhalten. Daher sollen Aspekte der vorliegenden Offenbarung solche Abwandlungen und Variationen abdecken.

[0010] Beispielaspekte der vorliegenden Offenbarung zielen darauf ab, eine phonetische Beziehung zwischen zwei oder mehr Phrasen zu bestimmen. Wie hierin verwendet kann der Begriff „Phrase“ als Folge von einem oder mehreren Buchstaben definiert sein. Zum Beispiel kann eine Phrase eine Folge von Buchstaben beinhalten, die ein oder mehrere Wörter ergeben. Eine Eingabetextphrase kann in eine Audioausgabe, die einer synthetisierten Sprachwiedergabe der Eingabetextphrase entspricht, umgewandelt werden. Eine Texttranskription der Audioausgabe kann dann beispielsweise unter Verwendung von Spracherkennungstechniken bestimmt werden. Die Texttranskription kann dann mit mehreren Testphrasen verglichen werden, um eine genaue Übereinstimmung zwischen der Texttranskription und einer oder mehreren der Testphrasen zu bestimmen. Wenn keine genaue Übereinstimmung gefunden wird, kann bestimmt werden, dass sich die Eingabetextphrase phonetisch von jeder der Testphrasen unterscheidet. So kann die Audioausgabe der Eingabetextphrase eine phonetische Aussprache der Eingabetextphrase darstellen. Durch Umwandeln der Eingabetextphrase in eine Audioausgabe und folgendes Umwandeln der Audioausgabe in ein Textformat, kann eine textuelle Darstellung der phonetischen Aussprache der Eingabetextphrase (z. B. die Texttranskription) bestimmt werden.

[0011] In einigen Implementierungen können Beispielaspekte der vorliegenden Offenbarung innerhalb eines Stimmbefehl-Verifizierungssystems eingebettet sein oder anderweitig diesem zugeordnet sein. Auf diese Weise kann eine Bestimmung der phonetischen Beziehung zwischen zwei oder mehreren Phrasen gemäß Beispielaspekten der vorliegenden Offenbarung dazu verwendet werden, um zu bestimmen, ob ein vorgeschlagener Stimmbefehl ein autorisierter Stimmbefehl ist, der phonetisch von mehreren reservierten Stimmbefehlen (z. B. Testphrasen), die schon durch eine Rechenplattform verwendet werden, verschieden ist. Auf diese Weise kann sichergestellt werden, dass ein einziger Stimmbefehl (oder mehrere phonetisch ähnliche Stimmbefehle) nicht dazu verwendet wird, um mehrere Rechenanwendungen oder Rechendienste, die der Rechenplattform zugeordnet sind, aufzurufen. Folglich kann die Anzahl von Operationen, die durch die Rechenplattform durchgeführt werden, verringert werden, beispielsweise eine Operation, die durch eine unbeabsichtigt aufgerufene Anwendung durchgeführt wird, oder eine Operation, um eine Klärung eines Stimmbefehls durch einen Anwender anzufordern, und ein Verarbeiten der Antwort. Der Betrieb des Prozessors der Rechenplattform kann reduziert werden und folglich kann ein Energieverbrauch der Rechenplattform verringert werden.

[0012] Als spezielles Beispiel kann durch einen Anwender eine Eingabetextphrase „Lehre“ eingegeben werden. Die Eingabetextphrase kann in eine Audioausgabe, die einer synthetisierten Sprachwiedergabe des Wortes „Lehre“ entspricht, umgewandelt werden. Eine Texttranskription der Audioausgabe kann bestimmt werden. Zum Beispiel kann die Texttranskription eine Transkription sein, die sich wie „Leere“ liest, was ein Homophon (z. B. phonetisch ähnlich) zu dem Wort „Lehre“ ist. Die Texttranskription kann mit einer Liste von Testphrasen verglichen werden, um eine Übereinstimmung zwischen der Texttranskription und einer oder mehreren Testphrasen zu identifizieren. Wenn die Liste von Testphrasen das Wort „Leere“ umfasst, kann eine Übereinstimmung gefunden werden und die Eingabetextphrase „Lehre“ kann als phonetisch ähnlich zu dem Wort „Leere“, wie es in der Liste der Testphrasen zu finden ist, identifiziert werden.

[0013] Genauer gesagt kann die Eingabetextphrase eine Folge von einem oder mehreren Wörtern, die durch einen Anwender in Textform eingegeben werden, sein. Die Eingabetextphrase kann durch einen Anwender zum Beispiel unter Verwendung einer Tastatur (z. B. einer Berührungstastatur oder einer physischen Tastatur) oder eines Tastenfelds, das einer Anwenderrechenvorrichtung wie beispielsweise einem Smartphone, einem Tablet, einer Laptoprechenvorrichtung, einer Desktoprechenvorrichtung, einer tragbaren Rechenvorrichtung oder anderen geeigneten Anwendervorrichtung zugeordnet ist, eingegeben werden. Die Eingabetextphrase kann in eine Audioausgabe aus künstlich erzeugter Sprache, die der Eingabetextphrase entspricht, umgewandelt werden. Die Audioausgabe kann eine Audiowellenform sein,

die dazu ausgelegt ist, durch eine Audiowiedergabevorrichtung abgespielt zu werden. Die Eingabetextphrase kann unter Verwendung verschiedener geeigneter Sprachsynthese- oder Text-zu-Sprache-Techniken in die Audioausgabe umgewandelt werden.

[0014] Zum Beispiel kann in einigen Implementierungen ein Umwandeln der Eingabetextphrase in eine Audioausgabe ein Bestimmen einer phonetischen Transkription der Eingabetextphrase, die einer symbolischen Repräsentation davon entspricht, wie eine gesprochene Wiedergabe des Texts klingen sollte, umfassen. Die phonetische Transkription kann eine Folge phonetischer Spracheinheiten wie beispielsweise Phonemen, Phonen oder anderen geeigneten phonetischen Spracheinheiten umfassen. Eine phonetische Spracheinheit kann einer akustischen Repräsentation eines Sprachsegments entsprechen. In einigen Implementierungen kann die Eingabetextphrase auf eine Folge von Wörtern heruntergebrochen werden und jedes Wort in der Eingabetextphrase kann in eine Folge von Phonemen umgewandelt werden, um die phonetische Transkription zu bestimmen. In einigen Implementierungen kann jedes Wort der Eingabetextphrase in eine Folge von Graphemen umgewandelt werden. Wie Fachleute verstehen werden, bezieht sich ein Graphem im Allgemeinen auf die kleinste Einheit eines Schriftsystems einer bestimmten Sprache. Die Grapheme können dann in eine Folge von Phonemen umgewandelt werden, um die phonetische Transkription zu bestimmen. In einigen Implementierungen können ferner ein oder mehrere prosodische Merkmale (z. B. mit linguistischen Funktionen wie beispielsweise Intonation, Ton, Betonung, Rhythmus etc. verbundene Merkmale) der Eingabetextphrase bestimmt werden.

[0015] Die phonetische Transkription kann dann einer Audioausgabe, die einer gesprochenen Wiedergabe einer phonetischen Transkription entspricht, zugeordnet werden. In einigen Implementierungen kann die phonetische Transkription der entsprechenden Audioausgabe zumindest teilweise basierend auf den prosodischen Merkmalen, die mit der phonetischen Transkription verknüpft sind, zugeordnet werden. Zum Beispiel kann die phonetische Transkription in einigen Implementierungen einem oder mehreren akustischen Merkmalen, die einer akustischen Wiedergabe der phonetischen Transkription entsprechen, zugeordnet werden. Die akustischen Merkmale können die Form von Merkmalsvektoren (z. B. Mel-Frequenz-Cepstrum-Koeffizienten oder anderen geeigneten Merkmalsvektoren) annehmen, die quantifizierbare Anteile von Sprachwellenformen wie beispielsweise Frequenzen und Spektralleistungen enthalten. Die akustischen Merkmale können dann in physikalische Eigenschaften, die eine Sprachwellenform der akustischen Merkmale darstellen, transformiert werden. Die Audioausgabe kann als eine Audiodatei erzeugt werden, die auf einem computerlesbaren Medium gespeichert oder aufgenommen werden kann. Zum Beispiel kann die Audiodatei zur anschließenden Wiedergabe der Audiodatei durch eine Audiowiedergabevorrichtung geeignet sein. Es ist zu beachten, dass verschiedene geeignete Sprachsynthesetechniken verwendet werden können, um die phonetische Transkription der Audioausgabe zuzuordnen, wie beispielsweise Verkettungssynthese, Einheitsauswahlsynthese, Diphonsynthese, domänenspezifische Synthese, Formatsynthese, artikulatorische Synthese, auf Hidden-Markov-Modellen basierende (HMM-basierte) Synthese, Sinusschwingungssynthese und/oder andere geeignete Sprachsynthesetechniken.

[0016] Die Audioausgabe kann anschließend unter Verwendung von einer oder mehreren geeigneten Spracherkennungstechniken in ein Textformat umgewandelt werden. Auf diese Weise kann eine Texttranskription der Audioausgabe bestimmte werden. Insbesondere kann ein Bestimmen einer Texttranskription der Audioausgabe ein Bestimmen einer oder mehrerer akustischer Merkmale, die der Audioausgabe zugeordnet sind, umfassen. Zum Beispiel kann die Audioausgabe in mehrere Segmente zerlegt werden und ein oder mehrere akustische Merkmale (z. B. Merkmalsvektoren) können für jedes Segment bestimmt werden. Die Merkmalsvektoren können einem oder mehreren Phonemen zugeordnet werden. Daten, die die zugeordneten Phoneme und/oder die Merkmalsvektoren angeben, können an ein oder mehrere Sprachmodelle (z. B. n-Gramm-Sprachmodelle oder andere geeignete Sprachmodelle) geliefert werden. Das eine oder die mehreren Sprachmodelle können dazu verwendet werden, die Transkription der Audioausgabe zu bestimmen. In einigen Implementierungen kann die Texttranskription an ein allgemeines Sprachmodell oder Basis-Sprachmodell geleifert werden. Ein solches allgemeines Sprachmodell kann mehrere üblicherweise verwendete Phrasen enthalten. Das allgemeine Sprachmodell kann ferner Wahrscheinlichkeitsschätzungen, die jeder Phrase zugeordnet sind, umfassen. Die Wahrscheinlichkeitsschätzungen können eine Schätzung der Wahrscheinlichkeit des Auftretens jeder Phrase in einer bestimmten Folge spezifizieren. Auf diese Weise kann das allgemeine Sprachmodell eine geschätzte Wahrscheinlichkeit eines Auftretens eines Wortes unter der Voraussetzung eines oder mehrerer vorher geäußerten Wörter definieren. In einigen Implementierungen können ein oder mehrere akustische Modelle (Hidden-Markov-Modelle, neuronale Netze etc.) ferner dazu verwendet werden, die Transkription der Audioausgabe zu bestimmen. Solche akustischen Modelle können statistische Beziehungen zwischen mehreren Audiosignalen und phonetischen Spracheinheiten definieren.

[0017] In einigen Implementierungen kann ein Bestimmen der Transkription der Audioausgabe ein Lie-

fern von Daten, die die bestimmten Merkmalsvektoren und/oder zugeordneten Phoneme angeben, an ein voreingenommenes oder spezialisiertes Sprachmodell umfassen. Zum Beispiel kann das voreingenommene Sprachmodell durch Beeinflussen des allgemeinen Sprachmodells hin zu mehreren Testphrasen erzeugt werden. Insbesondere kann das voreingenommene Sprachmodell durch ein Erhöhen der Wahrscheinlichkeitsschätzungen der Phrasen, die in den mehreren Testphrasen eingeschlossen sind, erzeugt werden. Auf diese Weise kann das voreingenommene Sprachmodell eine erhöhte geschätzte Auftretenswahrscheinlichkeit der Testphrasen spezifizieren.

[0018] So kann die Texttranskription der Audioausgabe zumindest teilweise basierend auf dem voreingenommenen Sprachmodell und/oder dem allgemeinen Sprachmodell bestimmt werden. In einigen Implementierungen kann eine erste Transkription unter Verwendung des allgemeinen Sprachmodells bestimmt werden und eine zweite Transkription unter Verwendung des voreingenommenen Sprachmodells bestimmt werden. Insbesondere kann eine erste Erkennungssicherheitspunktzahl für die erste Transkription und eine zweite Erkennungssicherheitspunktzahl für die zweite Transkription bestimmt werden. Die Erkennungssicherheitspunktzahlen können jeweils ein geschätztes Vertrauen in die Genauigkeit der Transkriptionen spezifizieren. Eine Transkription kann zumindest teilweise basierend auf den Erkennungssicherheitspunktzahlen ausgewählt werden. Zum Beispiel kann die ausgewählte Transkription die Transkription mit der höheren Erkennungssicherheitspunktzahl sein. In einigen Implementierungen können mehrere Texttranskriptionen bestimmt und ausgewählt werden, die alternative Schreibweisen eines oder mehrerer Wörter in den Transkriptionen darstellen. Zum Beispiel kann, um das Beispiel von oben weiterzuführen, unter Verwendung der Eingabetextphrase „Lehre“ eine erste Texttranskription des Wortes „Lehre“ und eine zweite Texttranskription des Wortes „Leere“ ausgewählt werden.

[0019] Nach Bestimmen der Texttranskription der Audioausgabe kann die Texttranskription mit mehreren Testphrasen verglichen werden, um zu bestimmen, ob die Texttranskription unter den mehreren Testphrasen enthalten ist. In Implementierungen, in denen mehrere Texttranskriptionen ausgewählt werden, kann jede Texttranskription mit mehreren Testphrasen verglichen werden. Auf diese Weise können die mehreren Testphrasen durchsucht werden, um eine direkte Übereinstimmung zwischen einer oder mehreren Testphrasen und der Texttranskription zu bestimmen. Wenn keine direkte Übereinstimmung gefunden wird, kann bestimmt werden, dass die Eingabetextphrase phonetisch von den mehreren Testphrasen verschieden ist. Wenn eine direkte Übereinstimmung gefunden wird, kann bestimmt werden, dass die Eingabetextphrase phonetisch der einen oder den mehreren Testphrasen gleich oder ähnlich ist.

[0020] In einigen Implementierungen kann dann, wenn eine direkte Übereinstimmung zwischen der Texttranskription und der einen oder den mehreren Testphrasen gefunden wird, ein phonetischer Abstand zwischen der Eingabetextphrase und der Texttranskription bestimmt werden. Der phonetische Abstand kann bestimmt werden, um zu bestimmen, ob die Eingabetextphrase der Texttranskription phonetisch ähnlich ist. Eine solche Bestimmung des phonetischen Abstands kann dazu verwendet werden, um die Bestimmung, dass die Eingabetextphrase phonetisch einen oder den mehreren Testphrasen gleich oder ähnlich ist, zu verifizieren. So kann dann, wenn die Eingabetextphrase der Texttranskription phonetisch ähnlich ist, bestimmt werden, dass, weil die Texttranskription der einen oder den mehreren Testphrasen phonetisch gleich bestimmt wurde und die Eingabetextphrase als der Texttranskription phonetisch ähnlich bestimmt wurde, die Eingabetextphrase der einen oder den mehreren Testphrasen phonetisch ähnlich ist.

[0021] Der phonetische Abstand kann durch Umwandeln der Eingabetextphrase und der Texttranskription in jeweilige Phonemfolgen bestimmt werden. Insbesondere kann eine erste Phonemfolge für die Eingabetextphrase bestimmt werden und eine zweite Phonemfolge für die Texttranskription bestimmt werden. Der phonetische Abstand kann dann zumindest teilweise basierend auf der ersten und zweiten Phonemfolge bestimmt werden. Zum Beispiel kann der phonetische Abstand durch Bestimmen einer Anzahl von Phonemen der zweiten Phonemfolge, die sich von der ersten Phonemfolge unterscheiden, (z. B. einer Anzahl von Phonemen aus der zweiten Phonemfolge, die geändert werden müssten, um mit der ersten Phonemfolge übereinzustimmen) bestimmt werden.

[0022] Wenn der phonetische Abstand geringer als eine vorbestimmte Schwelle ist, kann bestimmt werden, dass die Eingabetextphrase der Texttranskription ähnlich ist. Auf diese Weise kann bestimmt werden, dass die Eingabetextphrase der einen oder den mehreren Testphrasen ähnlich ist. Wenn der phonetische Abstand höher als die Schwelle ist, kann daraus abgeleitet werden, dass die Eingabetextphrase von den mehreren Testphrasen phonetisch verschieden ist. In einigen Implementierungen kann dann, wenn der phonetische Abstand höher als die Schwelle ist, der gesamte Prozess erneut durchgeführt werden.

[0023] Das Bestimmen phonetischer Beziehungen zwischen Phrasen gemäß beispielhaften Aspekten der vorliegenden Offenbarung kann es ermöglichen, dass solche phonetischen Beziehungen in Echtzeit

oder nahezu in Echtzeit bestimmt werden. So kann nach einer Eingabe einer Eingabetextphrase in eine Anwendervorrichtung eine Angabe der phonetischen Beziehung zwischen der Eingabetextphrase und den Testphrasen beispielsweise innerhalb einer Anwenderschnittstelle der Anwendervorrichtung in Echtzeit oder nahezu in Echtzeit an einen Anwender geliefert werden. Solche Bestimmungstechniken für phonetische Beziehungen können verglichen mit herkömmlichen Bestimmungstechniken für phonetische Beziehungen unter Verwendung von wenigen Verarbeitungsbetriebsmitteln, geringer Bandbreite und/oder geringer Datenübermittlung durchgeführt werden. Die Zeit und die Betriebsmittel, die zum Bestimmen der Beziehungen benötigt werden, sind gemäß Beispielaspekten der vorliegenden Offenbarung nicht von der Anzahl der Testphrasen unter den mehreren Testphrasen abhängig. Auf diese Weise können die mehreren Testphrasen eine beliebige geeignete Anzahl von Testphrasen enthalten, ohne die Qualität der Bestimmungstechniken zu opfern. Weiterhin stützen sich solche Bestimmungstechniken für phonetische Beziehungen nicht auf eine Stapelverarbeitung.

[0024] Unter Bezugnahme auf die Figuren werden nun beispielhafte Aspekte der vorliegenden Offenbarung genauer erörtert. Zum Beispiel zeigt **Fig. 1** eine Übersicht über ein beispielhaftes System **100** zum Bestimmen einer phonetischen Ähnlichkeit zwischen zwei oder mehr Phrasen. Das System **100** enthält eine Anwendervorrichtung **102** und einen Server **104**. Die Anwendervorrichtung **102** kann eine beliebige geeignete Anwendervorrichtung sein, wie z. B. ein Smartphone, ein Tablet, ein Laptop, ein Desktop-Computer, eine tragbares Rechenvorrichtung oder eine andere geeignete Anwendervorrichtung. Der Server **104** enthält einen Sprachsynthesizer **106**, einen Audiotranskriptor **108** und einen Bestimmer phonetischer Beziehungen **110**. Die Anwendervorrichtung **102** kann mit dem Server **104** beispielsweise über ein Netz kommunizieren. In einigen Implementierungen können eine oder mehrere Funktionen, die dem Sprachsynthesizer **106**, dem Audiotranskriptor **108** und/oder dem Bestimmer phonetischer Beziehungen **110** zugeordnet sind, lokal auf der Anwendervorrichtung **102** ausgeführt werden.

[0025] Die Anwendervorrichtung **102** kann dazu ausgelegt sein, eine Eingabe von dem Anwender zu empfangen, die eine Eingabetextphrase angibt. Insbesondere kann die Anwendervorrichtung **102** dazu ausgelegt sein, eine Anwenderschnittstelle beispielsweise auf einer Anzeigevorrichtung, die der Anwendervorrichtung zugeordnet ist, anzuzeigen. Die Anwenderschnittstelle kann den Anwender auffordern, die Eingabetextphrase einzugeben. Zum Beispiel zeigt **Fig. 2** eine beispielhafte Anwendervorrichtung **102**, die eine beispielhafte Anwenderschnittstelle **120** anzeigt, gemäß beispielhaften Ausführungsformen der vorliegenden Offenbarung. Die Anwenderschnittstelle **120** enthält ein Textfeld **122**, das zum Empfangen einer Texteingabe ausgelegt ist. Der Anwender kann die Eingabetextphrase z. B. unter Verwendung einer Tastatur **124** in das Textfeld **122** eingeben. Der Anwender kann die Anforderung durch Interaktion mit einem Übermittlungsschnittstellenelement **126** übermitteln. Die Tastatur **124** kann eine in der Anwenderschnittstelle **120** angezeigte Berührungstastatur sein. Es versteht sich, dass verschiedene andere geeignete Eingabevorrichtungen verwendet werden können, wie beispielsweise eine physische Tastatur, ein Tastenfeld oder eine andere geeignete Eingabevorrichtung.

[0026] Nach einer Bestimmung einer phonetischen Beziehung der Eingabetextphrase und einer oder mehrerer Testphrasen (z. B. ob die Eingabetextphrase von der einen oder den mehreren Testphrasen phonetisch verschieden ist oder diesen phonetisch ähnlich ist), kann die Anwenderschnittstelle **120** dazu ausgelegt sein, eine Angabe der phonetischen Beziehung an den Anwender zu liefern. Zum Beispiel kann die Anwenderschnittstelle in einigen Implementierungen eine geeignete Angabe anzeigen, die die phonetische Beziehung bezeichnet.

[0027] Nach Empfangen der Eingabetextphrase von dem Anwender kann die Anwendervorrichtung Daten an den Server **104** liefern, die die Eingabetextphrase angeben. Der Server **104** kann dann bestimmen, ob die Eingabetextphrase einer oder mehreren Testphrasen phonetisch ähnlich ist. Zum Beispiel kann unter Bezugnahme auf **Fig. 1** der Sprachsynthesizer **106** dazu ausgelegt sein, eine Audioausgabe, die einer synthetisierten gesprochenen Wiedergabe der Eingabetextphrase entspricht, zu bestimmen. Zum Beispiel kann eine Wiedergabe der Audioausgabe durch ein Audiovorrichtung wie eine menschliche Stimme klingen, die das Wort (die Wörter) der Eingabetextphrase spricht.

[0028] Insbesondere kann der Sprachsynthesizer **106** dazu ausgelegt sein, die Audioausgabe durch Bestimmen einer phonetischen Transkription der Eingabetextphrase zu bestimmen. Wie angegeben kann die phonetische Transkription eine Folge von phonetischen Spracheinheiten enthalten, die jeweils einer akustischen Repräsentation eines Sprachsegments entsprechen, das der Eingabetextphrase zugeordnet ist. In einigen Implementierungen kann die phonetische Transkription von Kontextinformationen begleitet sein, die eine korrekte und/oder beabsichtigte Sprachwiedergabe der phonetischen Spracheinheiten der phonetischen Transkription angeben. Zum Beispiel können die Kontextinformationen relative Positionen identifizierter Phoneme innerhalb einer Eingabefolge (z. B. linker Kontext, rechter Kontext usw.) umfassen. Die Kontextinformationen können ferner Zeitvorgabeinformationen enthalten, um beabsichtigte Dauern akustischer Wiedergaben iden-

tifizierter Phoneme und beabsichtigte relative Zeitvorgaben von Phonemen innerhalb längerer Wellenformen anzugeben. Die Kontextinformationen können ferner Zustandsinformationen enthalten, um akustische Phasen der Phoneme anzugeben.

[0029] Der Sprachsynthesizer **106** kann die phonetische Transkription einem oder mehreren vorhergesagten Merkmalsvektoren zuordnen, beispielsweise zumindest teilweise basierend auf den Kontextinformationen, die mit der phonetischen Transkription verknüpft sind. Der Sprachsynthesizer **106** kann eine Menge vorhergesagter Merkmalsvektoren, die der phonetischen Transkription entsprechen, zumindest teilweise basierend auf den Zuordnungen erzeugen. Die vorhergesagten Merkmalsvektoren können akustische Metriken umfassen, die akustische Eigenschaften einer entsprechenden Wellenform bestimmen. Auf diese Weise können die vorhergesagten Merkmalsvektoren in eine Wellenform, die der Audioausgabe entspricht, umgesetzt werden. Zum Beispiel können Merkmalsvektoren verschiedene geeignete akustische Metriken wie z. B. Mel-Cepstral-Koeffizienten, Linienspektralpaare, lineare Prädiktionskoeffizienten, Mel-verallgemeinerte Cepstral-Koeffizienten, Grundfrequenz (f0), aperiodische Maße, logarithmisches Leistungsspektrum oder Phase.

[0030] Wie angegeben kann der Sprachsynthesizer **106** die Merkmalsvektoren in eine Audioausgangswellenform übersetzen, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht. In einigen Implementierungen kann der Sprachsynthesizer die Audioausgabe durch Zuordnen der Merkmalsvektoren zu vordefinierten Sprachwellenformsegmente, die in einer Sprachdatenbank **114** gespeichert sind, bestimmen. Es versteht sich, dass der Sprachsynthesizer **106** verschiedene geeignete Sprachsynthesetechniken verwenden kann, um die phonetische Transkription der Audioausgabe zuzuordnen, beispielsweise Verkettungssynthese, Einheitsselektionsynthese, Diphonsynthese, domänenspezifische Synthese, Formatsynthese, artikulatorische Synthese, auf Hidden-Markov-Modellen basierend (HMM-basierte) Synthese, Sinusschwingungssynthese und/oder andere geeignete Sprachsynthesetechn iken.

[0031] In einigen Implementierungen kann die Audioausgabe unter Verwendung verschiedener Parameter wie etwa verschiedener geeigneter Stimmen, Sprachabtastraten usw. bestimmt werden. Auf diese Weise kann der Sprachsynthesizer **106** die Wiedergabetreue der Audioausgabe durch Anpassen solcher Parameter steuern.

[0032] Nach einer Bestimmung der Audioausgabe kann der Audiotranskriptor **108** dazu ausgelegt sein, eine Texttranskription der Audioausgabe unter Verwendung einer oder mehrerer geeigneter Spracherkennungstechniken zu bestimmen. Insbesondere kann der Audiotranskriptor **108** dazu ausgelegt sein, die Audioausgabewellenform in mehrere Segmente zu unterteilen und mehrere Merkmalsvektoren aus den mehreren Segmenten zu extrahieren. Der Audiotranskriptor **108** kann dann zumindest teilweise basierend auf einem oder mehreren Spracherkennungsmodellen **116** eine Wortfolge aus den Merkmalsvektoren erzeugen. Das eine oder die mehreren Spracherkennungsmodelle können ein oder mehrere akustische Modelle (z. B. HMMs, neuronale Netze, segmentale Modelle, Supersegmentmodelle, Modelle mit maximaler Entropie, bedingte Zufallsfelder usw.) und ein oder mehrere Sprachmodelle (z. B. Grammatik, n-Gramm-Sprachmodell, stochastisches Sprachmodell usw.) umfassen. Die akustischen Modelle können statistische Eigenschaften der Audioausgabe spezifizieren. Das Sprachmodell (die Sprachmodelle) können unter Vorgabe eines oder mehrerer zuvor bestimmter Wörter Wahrscheinlichkeitsschätzungen des Auftretens von Wörtern spezifizieren. Wie für den Fachmann ersichtlich ist, kann der Audiotranskriptor **108** eine Folge von einem oder mehreren Wörtern zumindest teilweise basierend auf dem/den Spracherkennungsmodell(en) **116** bestimmen, so dass die bestimmte Folge von Wörtern die Maximum A-posteriori-Wahrscheinlichkeit für die Eingabemerkmalsvektoren aufweist. Zum Beispiel kann der akustische Transkriptor **108** in einigen Implementierungen die Folge von Wörtern unter Verwendung eines Viterbi-Decodierers bestimmen.

[0033] In einigen Implementierungen können das/die Spracherkennungsmodell(e) **116** ein allgemeines Sprachmodell und ein voreingenommenes Sprachmodell umfassen. Auf diese Weise kann der Audiotranskriptor **108** die Texttranskription zumindest teilweise basierend auf dem allgemeinen Sprachmodell und/oder dem voreingenommenen Sprachmodell bestimmen. Wie angegeben kann das voreingenommene Sprachmodell zumindest teilweise basierend auf dem allgemeinen Sprachmodell bestimmt werden, beispielsweise durch Erhöhen der Wahrscheinlichkeitsschätzungen, die den mehreren Testphrasen **112** zugeordnet sind, relativ zu den Wahrscheinlichkeitsschätzungen für die mehreren Testphrasen **112**, wie sie im allgemeinen Sprachmodell angegeben sind. Auf diese Weise kann in einigen Implementierungen die Texttranskription zumindest teilweise basierend auf dem voreingenommenen Sprachmodell bestimmt werden.

[0034] In einigen Implementierungen kann der Audiotranskriptor **108** eine erste Transkription unter Verwendung des allgemeinen Sprachmodells und eine zweite Transkription unter Verwendung des voreingenommenen Sprachmodells bestimmen. Der Audiotranskriptor **108** kann ferner eine erste Erkennungssicherheitspunktzahl für die erste Transkription und eine zweite Erkennungssicherheitspunktzahl für die zweite Transkription bestimmen. Es kann zumindest

teilweise basierend auf den Sicherheitspunktzahlen entweder die erste Transkription oder die zweite Transkription ausgewählt werden. In einigen Implementierungen können eine oder mehrere zusätzliche Transkriptionen unter Verwendung eines oder mehrerer zusätzlicher Sprachmodelle bestimmt werden. Die zusätzlichen Transkriptionen können begleitende Erkennungssicherheitspunktzahlen aufweisen, so dass die ausgewählte Transkription zumindest teilweise basierend auf den Erkennungssicherheitspunktzahlen bestimmt wird. In einigen Implementierungen können mehrere Transkriptionen basierend auf alternativen Schreibweisen von Wörtern ausgewählt werden.

[0035] Nach einer Bestimmung der Texttranskription der Audioausgabe kann der Bestimmer phonetischer Ähnlichkeiten **110** das eine oder die mehreren Wörter der Texttranskription mit den mehreren Testphrasen **112** vergleichen, um eine Übereinstimmung zwischen der Texttranskription und einem oder mehreren Testphrasen zu bestimmen. Wenn eine Übereinstimmung bestimmt wird, kann der Bestimmer phonetischer Beziehungen **110** bestimmen, dass die Eingabetextphrase der einen oder den mehreren Testphrasen phonetisch ähnlich ist. Wenn keine Übereinstimmung bestimmt wird, kann der Bestimmer phonetischer Beziehungen **110** bestimmen, dass die Eingabetextphrase von jeder Testphrase phonetisch verschieden ist.

[0036] In einigen Implementierungen kann dann, wenn eine direkte Übereinstimmung zwischen der Texttranskription und einer oder mehreren Testphrasen bestimmt wird, der Bestimmer phonetischer Beziehungen **110** einen phonetischen Abstand zwischen der Eingabetextphrase und der Texttranskription der Audioausgabe bestimmen. Insbesondere kann der Bestimmer phonetischer Beziehungen **110** phonetische Transkriptionen für die Eingabetextphrase und die Texttranskription bestimmen. Die phonetischen Transkriptionen können eine Folge von phonetischen Spracheinheiten enthalten, die jeweils die Eingabetextphrase und die Texttranskription darstellen. Zum Beispiel kann der Bestimmer phonetischer Beziehungen **110** eine erste phonetische Transkription für die Eingabetextphrase und eine zweite phonetische Transkription für die Texttranskription bestimmen. Der Bestimmer phonetischer Beziehungen **110** kann dann einen phonetischen Abstand zumindest teilweise basierend auf der ersten und zweiten phonetischen Transkription bestimmen. Der phonetische Abstand kann eine Quantifizierung davon sein, wie unähnlich die Texttranskription der eingegebenen Textphrase ist. Der Bestimmer phonetischer Beziehungen1 10 kann den phonetischen Abstand bestimmen, indem er eine Anzahl von phonetischen Spracheinheiten in der zweiten phonetischen Transkription bestimmt, die sich von der ersten phonetischen Transkription unterscheiden.

[0037] Wenn der phonetische Abstand geringer als eine vorbestimmte Schwelle ist, kann der Bestimmer phonetischer Beziehungen **110** bestimmen, dass die Texttranskription der eingegebenen Textphrase phonetisch ähnlich ist und dass daher die Eingabetextphrase der oder den mehreren Testphrasen, mit denen die Texttranskription abgeglichen wurde, phonetisch ähnlich ist. Wenn der phonetische Abstand höher als die Schwelle ist, kann der Bestimmer phonetischer Beziehungen **110** bestimmen, dass die Texttranskription der Eingabetextphrase nicht phonetisch ähnlich ist und daher die Eingabetextphrase von den mehreren Testphrasen **112** phonetisch verschieden ist.

[0038] Nach einer Bestimmung der phonetischen Beziehung zwischen der Eingabetextphrase und der einen oder den mehreren Testphrasen kann der Server **104** eine Angabe der phonetischen Beziehung an die Anwendervorrichtung **102** liefern. Zum Beispiel kann der Server **104** ein oder mehrere Signale, die die phonetische Beziehung angeben, an die Anwendervorrichtung **102** liefern. Wenn beispielsweise bestimmt wird, dass die Eingabetextphrase von jeder Testphrase phonetisch verschieden ist, können das eine oder die mehreren Signale angeben, dass die Eingabetextphrase phonetisch verschieden ist. In Implementierungen, in denen die Bestimmungstechniken für phonetische Beziehungen einem Sprachbefehls-Verifikationssystem zugeordnet sind, können das eine oder die mehreren Signale eine Angabe enthalten, dass der vorgeschlagene Sprachbefehl (z. B. die Eingabetextphrase) von den reservierten Sprachbefehlen phonetisch verschieden ist und/oder dass der vorgeschlagene Sprachbefehl zur Verwendung autorisiert ist. Die Anwendervorrichtung **102** kann dann eine Angabe der phonetischen Beziehung an den Anwender liefern. Zum Beispiel kann die Anwendervorrichtung **102** die Angabe innerhalb der in **Fig. 2** dargestellten Anwenderschnittstelle **120** präsentieren.

[0039] **Fig. 3** zeigt ein Ablaufdiagramm eines beispielhaften Verfahrens (**200**) zum Bestimmen einer phonetischen Beziehung zwischen zwei oder mehr Phrasen. Das Verfahren (**200**) kann durch eine oder mehrere Rechenvorrichtungen wie etwa eine oder mehrere der Rechenvorrichtungen, die in **Fig. 1** dargestellt sind, implementiert werden. Außerdem zeigt **Fig. 3** zum Zweck der Veranschaulichung und Diskussion Schritte, die in einer bestimmten Reihenfolge ausgeführt werden. Fachleute werden unter Verwendung der hierin bereitgestellten Offenbarungen verstehen, dass die Schritte von jedem der hier diskutierten Verfahren auf verschiedene Arten angepasst, umgeordnet, erweitert, weggelassen oder abgewandelt werden können, ohne vom Umfang der vorliegenden Offenbarung abzuweichen .

[0040] Bei (**202**) kann das Verfahren (**200**) ein Empfangen von Daten, die eine Eingabetextphrase angeben, umfassen. Die Eingabetextphrase kann von einem Anwender beispielsweise unter Verwendung verschiedener geeigneter Texteingabetechniken auf einer Anwendervorrichtung eingegeben werden. Die Eingabetextphrase kann eine Folge von einem oder mehreren Wörtern sein, die der Anwender mit mehreren Testphrasen vergleichen möchte, um eine phonetische Beziehung zwischen der Eingabetextphrase und den Testphrasen zu bestimmen. Die phonetische Beziehung kann angeben, ob die Eingabetextphrase von den Testphrasen phonetisch verschieden oder diesen phonetisch ähnlich ist.

[0041] Bei (**204**) kann das Verfahren (**200**) ein Bestimmen einer Audioausgabe, die der Eingabetextphrase entspricht, umfassen. Insbesondere kann die Audioausgabe eine Wellenform sein, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht. Zum Beispiel kann eine Wiedergabe der Audioausgabe wie eine menschliche Stimme klingen, die das Wort (die Wörter) der Eingabetextphrase spricht. Die Audioausgabe kann eine phonetische Aussprache der Eingabetextphrase darstellen. So kann die phonetische Aussprache unabhängig von der Schreibweise des Wortes (der Wörter) sein, die in der Eingabetextphrase enthalten sind. Die Audioausgabe kann unter Verwendung einer beliebigen geeigneten Sprachsynthesetechnik bestimmt werden. Die Audioausgabe kann als eine beliebige geeignete Audiodatei gespeichert werden, die für eine Audiowiedergabe geeignet ist. Auf diese Weise kann die Audiowellenform als eine Audiodatei erzeugt werden, die auf einem Speichermedium gespeichert oder aufgezeichnet werden kann und die für eine anschließende Wiedergabe geeignet ist.

[0042] Bei (**206**) kann das Verfahren (**200**) ein Bestimmen einer Texttranskription der Audioausgabe umfassen. Die Texttranskription kann eine Darstellung der Audioausgabe in Textform sein. In einigen Fällen können ein oder mehrere Wörter der Texttranskription alternative Schreibweisen zu den entsprechenden Wörtern in der Eingabetextphrase aufweisen. Zum Beispiel kann die Texttranskription so bestimmt werden, dass sie alternative Schreibweisen des einen oder der mehreren Wörter, die auf der eingegebenen Textphrase basieren, umfasst. Die Texttranskription kann unter Verwendung einer beliebigen geeigneten Spracherkennungstechnik bestimmt werden. Zum Beispiel kann die Texttranskription unter Verwendung eines oder mehrerer akustischer Modelle und/oder eines oder mehrerer Sprachmodelle bestimmt werden. Wie angegeben können das eine oder die mehreren Sprachmodelle ein allgemeines Sprachmodell und/oder ein voreingenommenes Sprachmodell umfassen. Das voreingenommene Sprachmodell kann zumindest teilweise basierend auf dem allgemeinen Sprachmodell und den mehreren Testphrasen erzeugt werden.

[0043] Bei (**208**) kann das Verfahren (**200**) ein Vergleichen der Texttranskription mit mehreren Testphrasen umfassen. Zum Beispiel kann das Vergleichen der Texttranskription mit den Testphrasen ein Durchsuchen der Testphrasen, um zu bestimmen, ob die Texttranskription mit einer oder mehreren der Testphrasen übereinstimmt, umfassen. Bei (**210**) kann das Verfahren (**200**) ein Bestimmen, ob die Texttranskription mit einer oder mehreren der Testphrasen übereinstimmt, umfassen. Wenn die Texttranskription mit keiner der Testphrasen übereinstimmt, kann das Verfahren (**200**) bei (212) ein Liefern einer Angabe, dass die eingegebene Textphrase von den mehreren Testphrasen phonetisch verschieden ist, umfassen.

[0044] Wenn die Texttranskription mit einer oder mehreren der Testphrasen übereinstimmen, kann das Verfahren (**200**) bei (**214**) ein Bestimmen eines phonetischen Abstands zwischen der Eingabetextphrase und der Texttranskription umfassen. Wie angegeben kann das Bestimmen des phonetischen Abstands ein Bestimmen phonetischer Transkriptionen, die der Eingabetextphrase und der Texttranskription zugeordnet sind, und ein Vergleichen der phonetischen Transkriptionen zum Bestimmen einer oder mehrerer phonetischer Spracheinheiten, die verschieden sind, umfassen. Auf diese Weise kann der phonetische Abstand eine Anzahl von phonetischen Spracheinheiten angeben, die der Texttranskription zugeordnet sind und die von den entsprechenden phonetischen Spracheinheiten, die der Eingabetextphrase zugeordnet sind, verschieden sind.

[0045] Bei (**216**) kann das Verfahren (**200**) ein Bestimmen, ob der phonetische Abstand kleiner als eine vordefinierte phonetische Abstandsschwelle ist, umfassen. Wenn der phonetische Abstand bei (**218**) kleiner als die (oder gleich der) Schwelle ist, kann das Verfahren (**200**) ein Liefern einer Angabe, dass die Eingabetextphrase der Texttranskription und/oder der einen oder mehreren Testphrasen phonetisch ähnlich ist, umfassen. Wenn der phonetische Abstand größer als die Schwelle ist, kann das Verfahren (**200**) bei (**220**) ein Liefern einer Angabe, dass die Eingabetextphrase von den mehreren Testphrasen phonetisch verschieden ist, umfassen.

[0046] In einigen Implementierungen kann das Verfahren (**200**) dann, wenn bei (**210**) eine Übereinstimmung zwischen der Texttranskription und einer oder mehreren Testphrasen bestimmt wird, (**214**) und (**216**) umgehen und direkt zu (**218**) voranschreiten. Auf diese Weise kann bei einer Bestimmung einer Übereinstimmung zwischen der Texttranskription und der einen oder den mehreren Testphrasen bestimmt werden, dass die Eingabetextphrase der ei-

nen oder den mehreren Testphrasen phonetisch ähnlich ist, ohne dass der phonetische Abstand zwischen der Eingabetextphrase und der Texttranskription bestimmt werden muss.

[0047] Fig. 4 zeigt ein beispielhaftes Rechensystem **300**, das verwendet werden kann, um die Verfahren und Systeme gemäß beispielhaften Aspekten der vorliegenden Offenbarung zu implementieren. Das System **300** kann unter Verwendung einer Client-Server-Architektur implementiert werden, die einen Server **310** umfasst, der über ein Netz **340** mit einer oder mehreren Clientvorrichtungen **330** kommuniziert. Das System **300** kann unter Verwendung anderer geeigneter Architekturen implementiert werden, wie zum Beispiel einer einzelnen Rechenvorrichtung.

[0048] Das System **300** enthält einen Server **310**, beispielsweise einen Webserver. Der Server **310** kann unter Verwendung einer beliebigen geeigneten Rechenvorrichtung implementiert werden. Der Server **310** kann einen oder mehrere Prozessoren **312** und eine oder mehrere Speichervorrichtungen **314** aufweisen. Der Server **310** kann auch eine Netzschnittstelle enthalten, die zur Kommunikation mit einer oder mehreren Clientvorrichtungen **330** über das Netz **340** verwendet wird. Die Netzschnittstelle kann beliebige geeignete Komponenten zur Verbindung mit einem weiteren Netz enthalten, einschließlich beispielsweise Sendern, Empfängern, Anschlüssen, Controllern, Antennen oder anderen geeigneten Komponenten.

[0049] Der eine oder die mehreren Prozessoren **312** können eine beliebige geeignete Verarbeitungsvorrichtung wie etwa einen Mikroprozessor, einen Mikrocontroller, eine integrierte Schaltung, eine Logikvorrichtung oder eine andere geeignete Verarbeitungsvorrichtung umfassen. Die eine oder die mehreren Speichervorrichtungen **314** können ein oder mehrere computerlesbare Medien umfassen, einschließlich, jedoch nicht darauf beschränkt auf, nicht-transitorische computerlesbare Medien, RAM, ROM, Festplatten, Flash-Laufwerke oder andere Speichervorrichtungen. Die eine oder die mehreren Speichervorrichtungen **314** können Informationen speichern, auf die der eine oder die mehreren Prozessoren **312** zugreifen können und die computerlesbare Befehle **316** enthalten, die von dem einen oder den mehreren Prozessoren **312** ausgeführt werden können. Die Befehle **316** können ein beliebiger Satz von Befehlen sein, die, wenn sie von dem einen oder den mehreren Prozessoren **312** ausgeführt werden, veranlassen, dass der eine oder die mehreren Prozessoren **312** Operationen ausführen. Zum Beispiel können die Befehle **316** von dem einen oder den mehreren Prozessoren **312** ausgeführt werden, um den Sprachsynthesizer **106**, den Audiotranskriptor **108** und/oder den Bestimmer phonetischer Beziehungen **110**, die unter Bezugnahme auf **Fig. 1** beschrieben sind, zu implementieren.

[0050] Wie in **Fig. 4** gezeigt können die eine oder die mehreren Speichervorrichtungen **314** auch Daten **318** speichern, die von dem einen oder den mehreren Prozessoren **312** abgerufen, manipuliert, erzeugt oder gespeichert werden können. Die Daten **318** können beispielsweise ein oder mehrere Spracherkennungsmodelle, Audioausgabedaten, mehrere Testphrasen, Sprachdaten und andere Daten umfassen. Die Daten **318** können in einer oder mehreren Datenbanken gespeichert sein. Die eine oder die mehreren Datenbanken können durch ein LAN oder WAN mit hoher Bandbreite mit dem Server **310** verbunden sein oder können auch über das Netz **340** mit dem Server **310** verbunden sein. Die eine oder die mehreren Datenbanken können so aufgeteilt sein, dass sie sich an mehreren Regionen befinden.

[0051] Der Server **310** kann Daten mit einer oder mehreren Clientvorrichtungen **330** über das Netz **340** austauschen. In **Fig. 4** kann eine beliebige Anzahl von Clientvorrichtungen **330** mit dem Server **310** über das Netz **340** verbunden sein. Jede der Clientvorrichtungen **330** kann eine beliebige geeignete Art von Rechenvorrichtung sein, wie beispielsweise ein Allzweckcomputer, ein Spezialcomputer, ein Laptop, ein Desktop, eine Mobilvorrichtung, ein Navigationssystem, ein Smartphone, ein Tablet, eine tragbare Rechenvorrichtung, eine Anzeige mit einem oder mehreren Prozessoren oder eine andere geeignete Rechenvorrichtung.

[0052] Ähnlich wie der Server **310** kann eine Clientvorrichtung **330** einen oder mehrere Prozessoren **332** und einen Speicher **334** enthalten. Der eine oder die mehreren Prozessoren **332** können eine oder mehrere zentrale Verarbeitungseinheiten (CPUs), Grafikverarbeitungseinheiten (GPUs), die zum effizienten Rendern von Bildern oder zur Durchführung anderer spezieller Berechnungen vorgesehen sind, und/oder andere Verarbeitungsvorrichtungen umfassen. Der Speicher **334** kann ein oder mehrere computerlesbare Medien enthalten und kann Informationen speichern, auf die der eine oder die mehreren Prozessoren **332** zugreifen können und die Befehle **336**, die von dem einen oder den mehreren Prozessoren **332** ausgeführt werden können, und Daten **338** umfassen. Beispielsweise kann der Speicher **334** Befehle **336** zum Implementieren einer Anwenderschnittstelle wie z. B. der Anwenderschnittstelle **120** die in **Fig. 2** dargestellt ist, speichern.

[0053] Die Clientvorrichtung **330** von **Fig. 4** kann verschiedene Eingabe-/Ausgabe-Vorrichtungen **337** zum Liefern an und Empfangen von Informationen von einem Anwender, wie etwa einen Berührungsbildschirm, ein Berührungsfeld, Dateneingabetasten, Lautsprecher und/oder ein Mikrofon, das für die

Spracherkennung geeignet ist, umfassen. Zum Beispiel kann die Clientvorrichtung **330** eine Anzeigevorrichtung **335** zum Präsentieren einer Anwenderschnittstelle, wie z. B. der in **Fig. 2** dargestellten Anwenderschnittstelle **120**, aufweisen.

[0054] Die Clientvorrichtung **330** kann auch eine Netzschnittstelle enthalten, die verwendet wird, um mit einem oder mehreren entfernten Rechenvorrichtungen (z. B. den Server **310**) über das Netz **340** zu kommunizieren. Die Netzschnittstelle kann beliebige geeignete Komponenten zum Koppeln mit einem oder mehreren Netzen umfassen, einschließlich z. B. Sender, Empfänger, Anschlüsse, Controller, Antennen oder anderer geeigneter Komponenten.

[0055] Das Netz **340** kann eine beliebige Art von Kommunikationsnetz sein, wie etwa ein lokales Netz (z. B. Intranet), ein Weitbereichsnetz (z. B. das Internet), ein Mobilfunknetz oder eine Kombination davon. Das Netz **340** kann auch eine direkte Verbindung zwischen einer Clientvorrichtung **330** und dem Server **310** umfassen. Im Allgemeinen kann die Kommunikation zwischen dem Server **310** und einer Clientvorrichtung **330** über eine Netzschnittstelle unter Verwendung eines beliebigen Typs von drahtgebundener und/oder drahtloser Verbindung erfolgen, wobei eine Vielzahl von Kommunikationsprotokollen (z. B. TCP/IP, HTTP, SMTP, FTP), Codierungen oder Formaten (z. B. HTML, XML) und/oder Schutzschemata (z. B. VPN, sicheres HTTP, SSL) verwendet werden kann.

[0056] Systeme und Verfahren zum Bestimmen phonetischer Beziehungen werden bereitgestellt. Zum Beispiel können Daten, die eine von einem Anwender eingegebene Eingabetextphrase angeben, empfangen werden. Eine Audioausgabe, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht, kann bestimmt werden. Eine Texttranskription der Audioausgabe der Eingabetextphrase kann bestimmt werden. Die Texttranskription kann eine Textdarstellung der Audioausgabe sein. Die Texttranskription kann mit mehreren Testphrasen verglichen werden, um eine Übereinstimmung zwischen der Texttranskription und mindestens einer Testphrase zu identifizieren.

[0057] Die hierin diskutierte Technologie bezieht sich auf Server, Datenbanken, Softwareanwendungen und andere computerbasierte Systeme sowie auf Aktionen, die von solchen Systemen vorgenommen werden, und Informationen, die an solche und von solchen Systemen gesendet werden. Fachleute werden erkennen, dass die inhärente Flexibilität computergestützter Systeme eine große Vielfalt von möglichen Konfigurationen, Kombinationen und Aufteilungen von Aufgaben und Funktionen zwischen und unter Komponenten ermöglicht. Zum Beispiel können hierin erörterte Serverprozesse unter Verwendung eines einzelnen Servers oder mehrerer Server, die in Kombination arbeiten, implementiert werden. Datenbanken und Anwendungen können auf einem einzelnen System implementiert sein oder über mehrere Systeme verteilt sein. Verteilte Komponenten können sequentiell oder parallel arbeiten.

[0058] Während der vorliegende Gegenstand unter Bezugnahme auf spezifische beispielhafte Ausführungsformen davon genau beschrieben worden ist, gilt es zu verstehen, dass Fachleute nach Verstehen des Vorstehenden leicht Änderungen, Variationen und Äquivalente zu solchen Ausführungsformen erzeugen können. Dementsprechend ist der Umfang der vorliegenden Offenbarung lediglich beispielhaft und nicht einschränkend und die vorliegende Offenbarung schließt die Aufnahme solcher Abwandlungen, Variationen und/oder Ergänzungen zu dem vorliegenden Gegenstand, wie sie ohne Weiteres für Fachleute auf diesem Gebiet ersichtlich sind, nicht aus.

Patentansprüche

1. Computerimplementiertes Verfahren (200) zum Bestimmen einer phonetischen Beziehung zwischen zwei oder mehr Phrasen, wobei das Verfahren umfasst:
Empfangen (202) von Daten, die eine von einem Anwender eingegebene Eingabetextphrase angeben, durch eine oder mehrere Rechenvorrichtungen (104, 310);
Bestimmen (204) einer Audioausgabe, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht, durch die eine oder die mehreren Rechenvorrichtungen;
Bestimmen (206) einer Texttranskription der Audioausgabe der Eingabetextphrase durch die eine oder die mehreren Rechenvorrichtungen, wobei die Texttranskription eine textuelle Repräsentation der Audioausgabe umfasst; und
Vergleichen (208) der Texttranskription mit mehreren Testphrasen durch die eine oder die mehreren Rechenvorrichtungen, um eine Übereinstimmung zwischen der Texttranskription und mindestens einer Testphrase (112) zu identifizieren.

2. Computerimplementiertes Verfahren (200) nach Anspruch 1, das ferner ein Identifizieren (210-JA) einer Übereinstimmung zwischen der Texttranskription und einer ersten Testphrase der mehreren Testphrasen durch die eine oder die mehreren Rechenvorrichtungen zumindest teilweise basierend auf dem Vergleichen (210) umfasst.

3. Computerimplementiertes Verfahren (200) nach Anspruch 2, das ferner, als Antwort auf das Identifizieren der Übereinstimmung (210-JA), ein Liefern (218) einer Angabe, dass die Eingabetextphrase der ersten Testphrase phonetisch ähnlich ist, durch die

eine oder die mehreren Rechenvorrichtungen umfasst.

4. Computerimplementiertes Verfahren (200) nach Anspruch 2 oder Anspruch 3, das ferner, als Antwort auf das Identifizieren der Übereinstimmung (210-JA), ein Bestimmen (214) eines phonetischen Abstands zwischen der Texttranskription der Audioausgabe und der Eingabetextphrase durch die eine oder die mehreren Rechenvorrichtungen umfasst.

5. Computerimplementiertes Verfahren (200) nach Anspruch 4, das ferner dann, wenn der phonetische Abstand zwischen der Texttranskription und der Eingabetextphrase geringer als eine Schwelle ist (216-JA), ein Liefern (218) einer Angabe, dass die Eingabetextphrase der ersten Testphrase phonetisch ähnlich ist, durch die eine oder die mehreren Rechenvorrichtungen umfasst.

6. Computerimplementiertes Verfahren (200) nach Anspruch 4 oder Anspruch 5, wobei das Bestimmen (214) eines phonetischen Abstands zwischen der Texttranskription der Audioausgabe und der Eingabetextphrase durch die eine oder die mehreren Rechenvorrichtungen umfasst:
Bestimmen einer ersten phonetischen Transkription, die der Eingabetextphrase zugeordnet ist, und einer zweiten phonetischen Transkription, die der Texttranskription zugeordnet ist, durch die eine oder mehreren Rechenvorrichtungen, wobei die erste und die zweite phonetische Transkription jeweils mehrere phonetische Spracheinheiten enthalten; und
Bestimmen einer Anzahl von phonetischen Spracheinheiten in der zweiten phonetischen Transkription, die sich von der ersten phonetischen Transkription unterscheiden, durch die eine oder die mehreren Rechenvorrichtungen.

7. Computerimplementiertes Verfahren (200) nach einem der vorhergehenden Ansprüche, das ferner ein Bestimmen, dass die Texttranskription mit keiner Testphrase der mehreren Testphrasen übereinstimmt (210-NEIN), zumindest teilweise basierend auf dem Vergleichen durch die eine oder mehreren Rechenvorrichtungen umfasst.

8. Computerimplementiertes Verfahren (200) nach Anspruch 7, das ferner, als Antwort auf das Bestimmen, dass die Texttranskription nicht mit einer Testphrase der mehreren Testphrasen übereinstimmt (210-NEIN), ein Liefern (212) einer Angabe, dass die Eingabetextphrase von den mehreren Testphrasen phonetisch verschieden ist, durch die eine oder die mehreren Rechenvorrichtungen umfasst.

9. Computerimplementiertes Verfahren (200) nach einem der vorhergehenden Ansprüche, wobei das Bestimmen (206) einer Texttranskription der Audioausgabe der Eingabetextphrase durch die eine oder die mehreren Rechenvorrichtungen ein Bestimmen der Texttranskription zumindest teilweise basierend auf einem oder mehreren Sprachmodellen umfasst.

10. Computerimplementiertes Verfahren (200) nach Anspruch 9, wobei das eine oder die mehreren Sprachmodelle ein voreingenommenes Sprachmodell umfassen, das zumindest teilweise auf einem allgemeinen Sprachmodell und den mehreren Testphrasen basiert.

11. Computerimplementiertes Verfahren (200) nach einem der vorhergehenden Ansprüche, wobei die Audioausgabe der Eingabetextphrase eine Sprachwellenform enthält, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht.

12. Rechensystem (104, 310), das enthält:
einen oder mehrere Prozessoren (312); und
eine oder mehrere Speichervorrichtungen (314), wobei die eine oder die mehreren Speichervorrichtungen (314) computerlesbare Befehle speichern, die, wenn sie von dem einen oder den mehreren Prozessoren (312) ausgeführt werden, veranlassen, dass der eine oder die mehreren Prozessoren (312) Operationen ausführen, wobei die Operationen Folgendes umfassen:
Empfangen (202) von Daten, die eine von einem Anwender eingegebene Eingabetextphrase angeben;
Bestimmen (204) einer Audioausgabe, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht;
Bestimmen (206) einer Texttranskription der Audioausgabe der Eingabetextphrase, wobei die Texttranskription eine textuelle Repräsentation der Audioausgabe umfasst; und
Vergleichen (208) der Texttranskription mit mehreren Testphrasen, um eine Übereinstimmung zwischen der Texttranskription und mindestens einer Testphrase zu identifizieren.

13. Rechensystem (104, 310) nach Anspruch 12, wobei die Operationen ferner ein Identifizieren (210-JA) einer Übereinstimmung zwischen der Texttranskription und einer ersten Testphrase der mehreren Testphrasen zumindest teilweise basierend auf dem Vergleichen umfassen.

14. Rechensystem (104, 310) nach Anspruch 13, wobei die Operationen ferner, als Antwort auf das Identifizieren (210-JA) der Übereinstimmung, ein Liefern (218) einer Angabe, dass die Eingabetextphrase der ersten Testphrase phonetisch ähnlich ist, umfassen.

15. Rechensystem (104, 310) nach Anspruch 13 oder Anspruch 14, wobei die Operationen ferner, als Antwort auf das Identifizieren (210-JA) der Übereinstimmung, ein Bestimmen (214) eines phonetischen

Abstands zwischen der Texttranskription der Audioausgabe und der Eingabetextphrase umfassen.

16. Rechensystem (104, 310) nach Anspruch 15, wobei die Operationen ferner dann, wenn der phonetische Abstand zwischen der Texttranskription und der Eingabetextphrase geringer als eine Schwelle ist (216-JA), ein Liefern (218) einer Angabe, dass die Eingabetextphrase der ersten Testphrase phonetisch ähnlich ist, umfassen.

17. Rechensystem (104, 310) nach Anspruch 15 oder 16, wobei das Bestimmen (214) eines phonetischen Abstands zwischen der Texttranskription der Audioausgabe und der Eingabetextphrase umfasst:
Bestimmen einer ersten phonetischen Transkription, die der Eingabetextphrase zugeordnet ist, und einer zweiten phonetischen Transkription, die der Texttranskription zugeordnet ist, wobei die erste und die zweite phonetische Transkription jeweils mehrere phonetische Spracheinheiten enthalten; und
Bestimmen einer Anzahl von phonetischen Spracheinheiten in der zweiten phonetischen Transkription, die sich von der ersten phonetischen Transkription unterscheiden.

18. Ein oder mehrere konkrete, nichttransitorische computerlesbare Medien (314), die computerlesbare Befehle (316) speichern, die, wenn sie von einem oder mehreren Prozessoren (312) ausgeführt werden, veranlassen, dass der eine oder die mehreren Prozessoren (312) Operationen ausführen, wobei die Operationen umfassen:
Empfangen (202) von Daten, die eine von einem Anwender eingegebene Eingabetextphrase angeben;
Bestimmen (204) einer Audioausgabe, die einer gesprochenen Wiedergabe der Eingabetextphrase entspricht;
Bestimmen (206) einer Texttranskription der Audioausgabe der Eingabetextphrase, wobei die Texttranskription eine textuelle Repräsentation der Audioausgabe umfasst; und
Vergleichen (208) der Texttranskription mit mehreren Testphrasen, um eine Übereinstimmung zwischen der Texttranskription und mindestens einer Testphrase zu identifizieren.

19. Ein oder mehrere konkrete, nichttransitorische computerlesbare Medien (314) nach Anspruch 18, wobei die Operationen ferner ein Bestimmen (210-NEIN), dass die Texttranskription mit keiner Testphrase der mehreren Testphrasen übereinstimmt, zumindest teilweise basierend auf dem Vergleichen (208) umfassen.

20. Ein oder mehrere konkrete, nichttransitorische computerlesbare Medien (314) nach Anspruch 19, wobei die Operationen ferner, als Antwort auf das Bestimmen, dass die Texttranskription nicht mit einer Testphrase der mehreren Testphrasen übereinstimmt (210-NEIN), ein Liefern (212) einer Angabe, dass die Eingabetextphrase von den mehreren Testphrasen phonetisch verschieden ist, umfassen.

Es folgen 4 Seiten Zeichnungen

Anhängende Zeichnungen

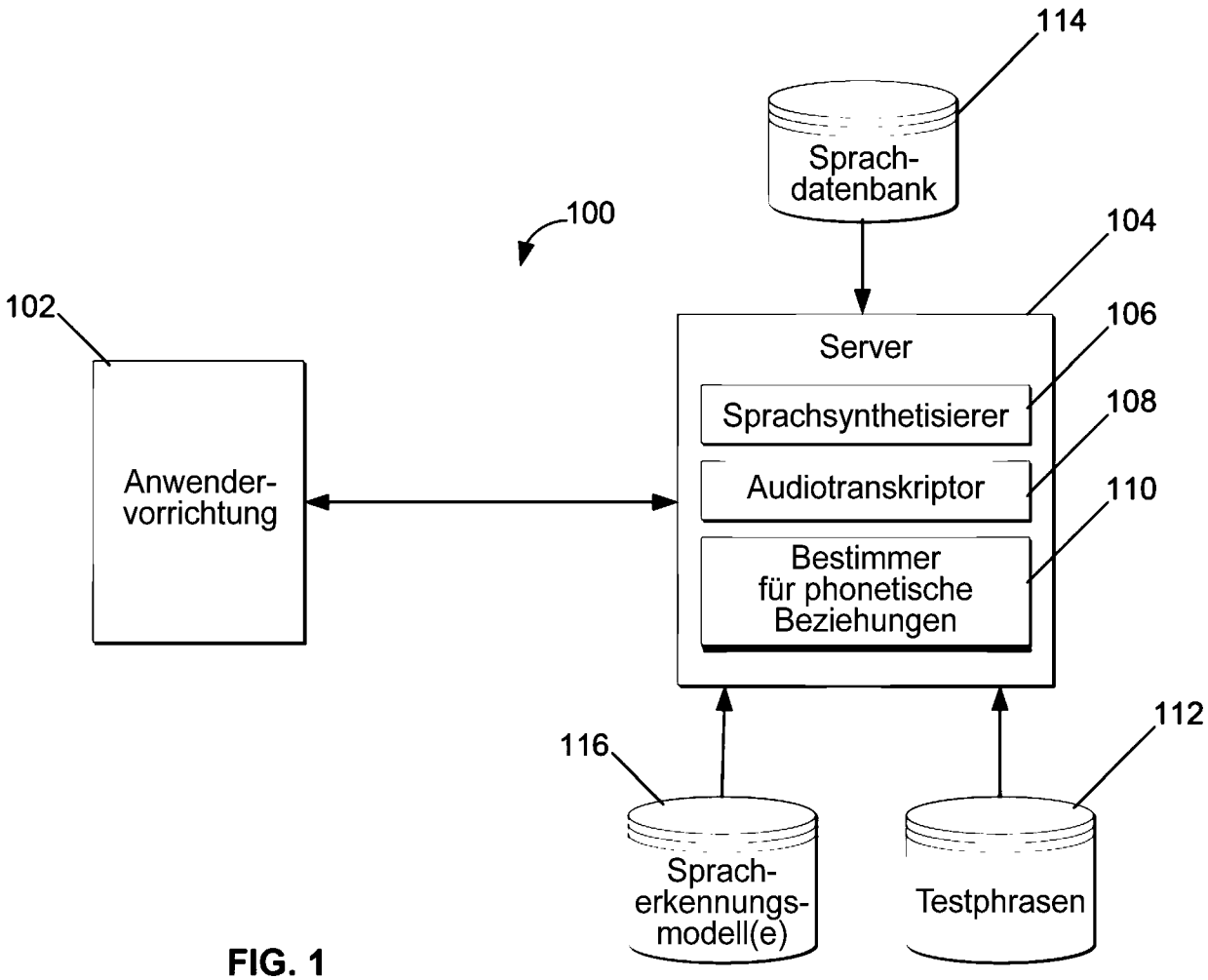


FIG. 1

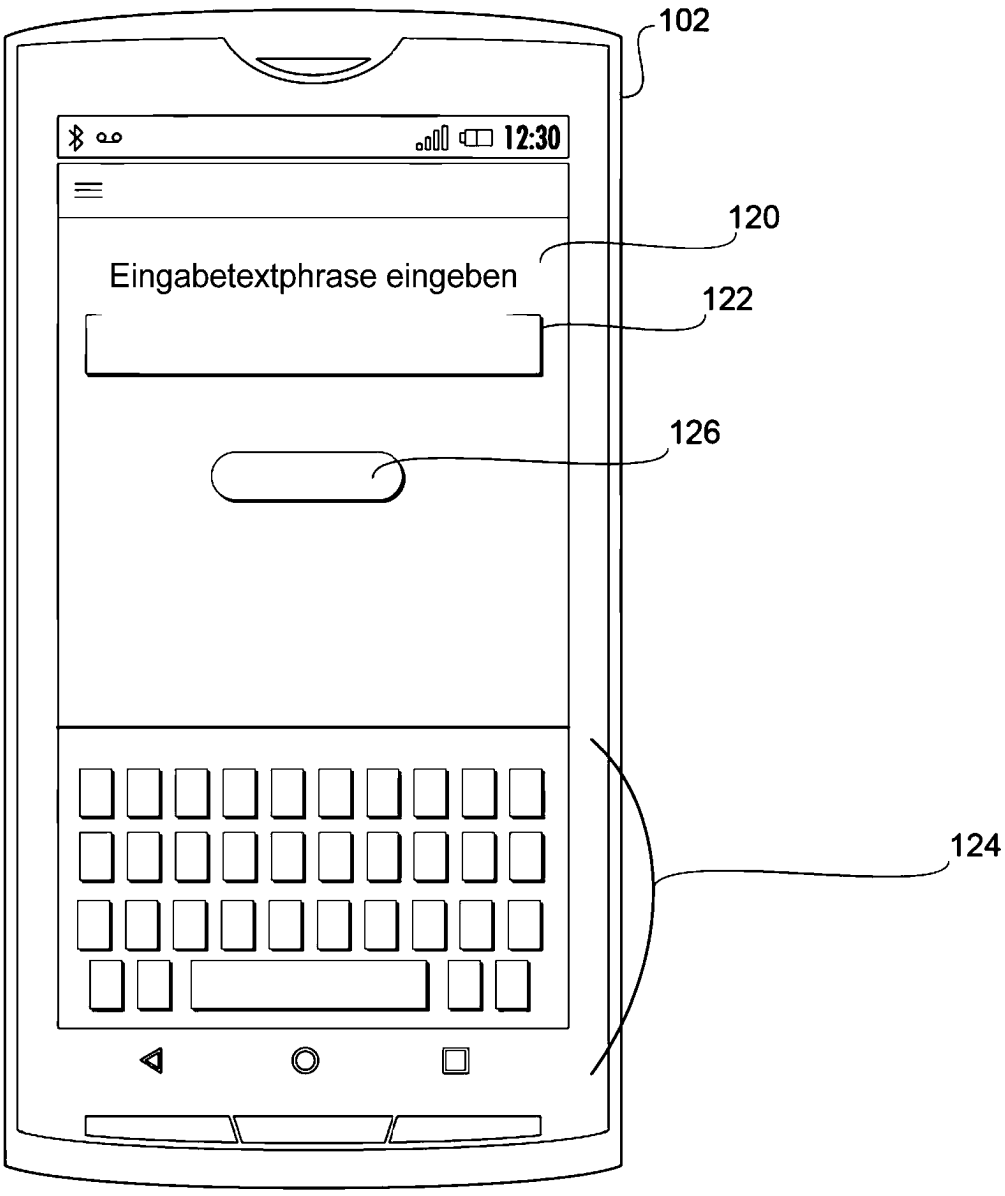


FIG. 2

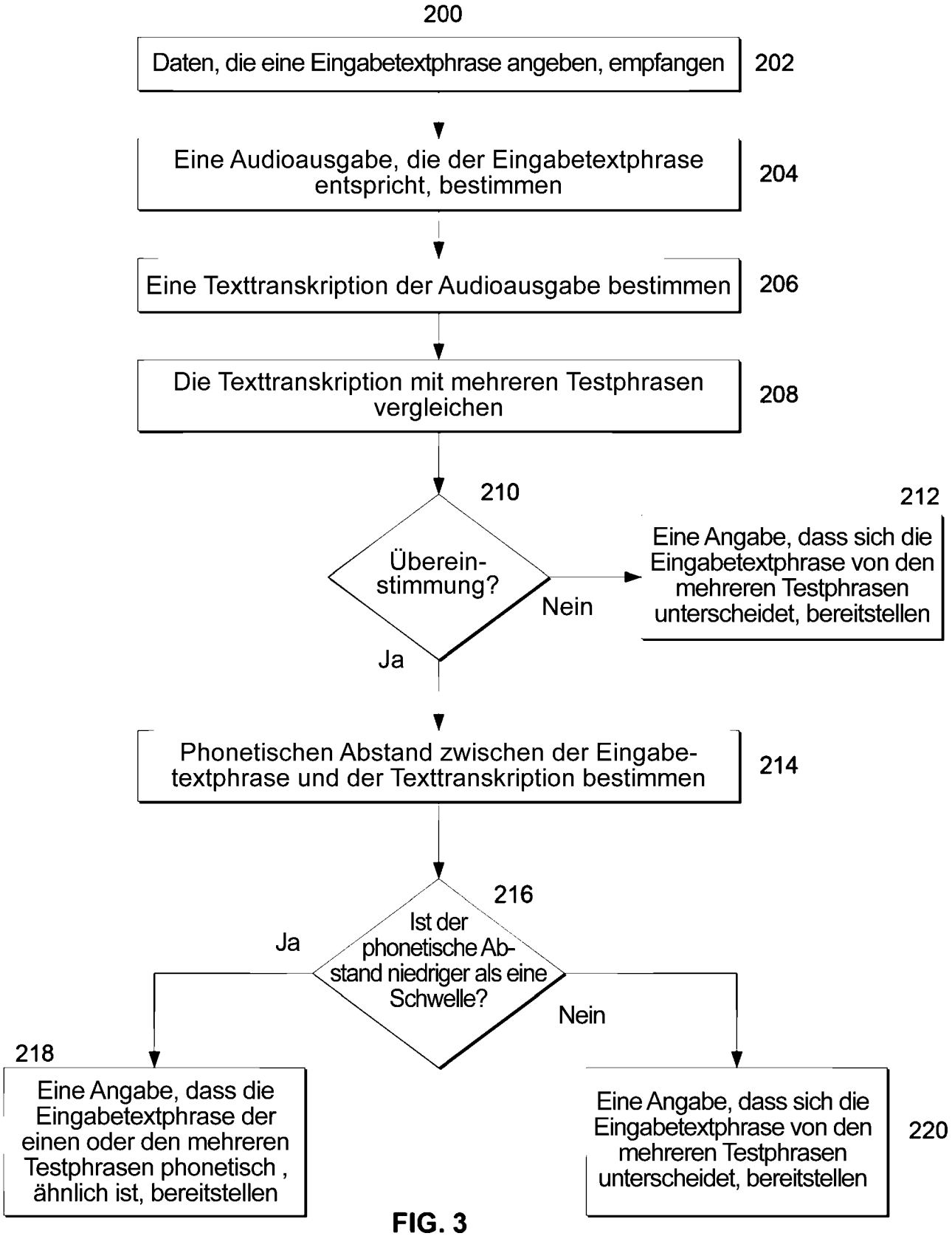

200
Daten, die eine Eingabetextphrase angeben, empfangen
202
Eine Audioausgabe, die der Eingabetextphrase entspricht, bestimmen
204
Eine Texttranskription der Audioausgabe bestimmen
206
Die Texttranskription mit mehreren Testphrasen vergleichen
208
210
Überein-stimmung?
Nein
212
Eine Angabe, dass sich die Eingabetextphrase von den mehreren Testphrasen unterscheidet, bereitstellen
Ja
Phonetischen Abstand zwischen der Eingabe-textphrase und der Texttranskription bestimmen
214
216
Ist der phonetische Ab-stand niedriger als eine Schwelle?
Ja
Nein
218
Eine Angabe, dass die Eingabetextphrase der einen oder den mehreren Testphrasen phonetisch , ähnlich ist, bereitstellen
Eine Angabe, dass sich die Eingabetextphrase von den mehreren Testphrasen unterscheidet, bereitstellen
220

FIG. 3

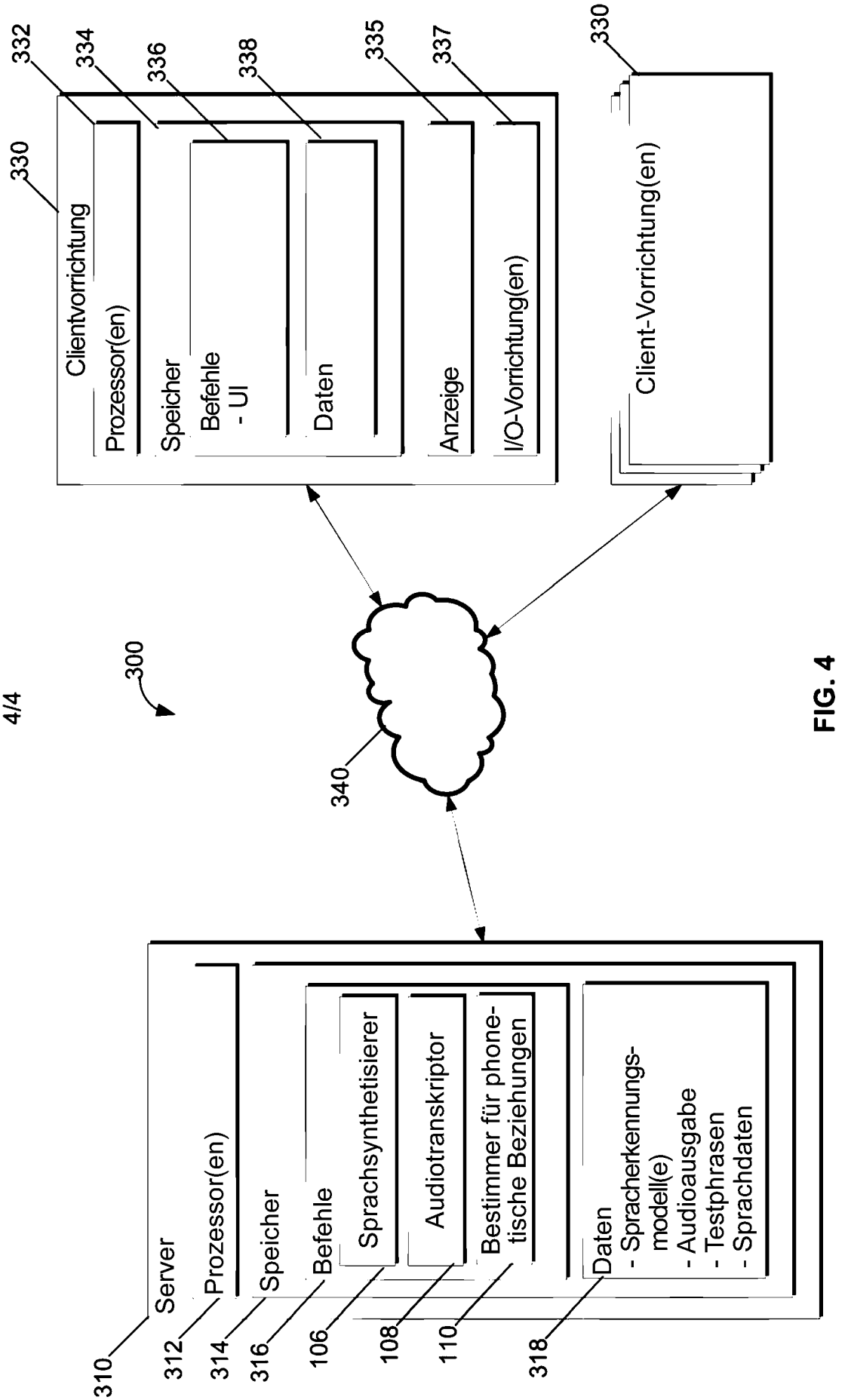


FIG. 4

