(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 058 235 A2

(12) **EUROPÄISCHE PATENTANMELDUNG**

(43) Veröffentlichungstag:
06.12.2000 Patentblatt 2000/49

(51) Int. Cl.[7]: **G10L 13/06**, G10L 13/08

(21) Anmeldenummer: **00108486.2**

(22) Anmeldetag: **19.04.2000**

(84) Benannte Vertragsstaaten:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Benannte Erstreckungsstaaten:
AL LT LV MK RO SI

(30) Priorität: **05.05.1999 DE 19920501**

(71) Anmelder:
NOKIA MOBILE PHONES LTD.
02150 Espoo (FI)

(72) Erfinder:
- **Buth, Peter**
  44795 Bochum (DE)
- **Dufhues, Frank**
  44866 Bochum (DE)

(74) Vertreter: **Stendel, Klaus**
Nokia GmbH,
Patentabteilung,
Postfach 101823
44718 Bochum (DE)

(54) **Wiedergabeverfahren für sprachgesteuerte Systeme mit text-basierter Sprachsynthese**

(57) Erfindungsgemäß wird ein einfaches und aussprachever besserte s Wiedergabeverfahren für sprachgesteuerte Systeme mit basierter Sprachsynthese angegeben, auch wenn die hinterlegte und zu synthetisierende Zeichenkette nicht den allgemeinen Regeln der Sprachwiedergabe folgt. Auch wird nach der Erfindung ein im Stand der Technik teilweises angewendetes "Hineinkopieren" des originalen Spracheingabetextes in den sonst synthetisierten Wiedergabetext vermeiden, wodurch durch das erfindungsgemäße Verfahren die Akzeptanz des Anwenders des sprachgesteuerten System wesentlich verbessert wird. Im einzelnen wird zunächst bei Vorliegen einer tatsächlich gesprochenen und mit einer gespeicherten Zeichenkette korrespondierenden Spracheingabe vor einer Wiedergabe der nach allgemeinen Regeln phonetisch beschriebenen und in eine rein synthetische Form gewandelten Zeichenkette die gewandelte Zeichenkette mit der Spracheingabe verglichen. Bei Feststellung einer oberhalb einer Schwelle liegenden Abweichung der gewandelten Zeichenkette von der Spracheingabe wird dann wenigstens eine Variante von der gewandelten Zeichenkette gebildet. Diese Variante wird dann, sofern diese bei einem Vergleich mit der Spracheingabe eine unterhalb der Schwelle liegende Abweichung aufweist, anstelle der gewandelten Zeichenkette ausgegeben.

EP 1 058 235 A2

Beschreibung

Technisches Gebiet

[0001] Die Erfindung befasst sich mit der Verbesserung von sprachgesteuerten Systemen mit text-basierter Sprachsynthese, insbesondere mit der Verbesserung der synthetischen Wiedergabe von gespeichert vorliegenden, aber bei der Aussprache bestimmten Eigentümlichkeiten unterliegenden Zeichenketten.

Stand der Technik

[0002] Bei der Bedienung von technischen Geräten gewinnt die Sprache zunehmend an Bedeutung. Dies betrifft sowohl die Eingabe von Daten und Kommandos wie auch die Ausgabe von Meldungen. Systeme, bei denen die Kommunikation zwischen Benutzer und Maschine in beiden Richtungen mit Hilfe akustischer Signale in Form von Sprache erfolgt, werden als Sprachdialogsysteme bezeichnet. Die vom jeweiligen System ausgegebenen Äußerungen können entweder zuvor aufgezeichnete natürliche Sprache sein oder entsprechend dem Gegenstand der vorliegenden Erfindung synthetisch erzeugt werden. Auch sind Anordnungen bekannt, bei denen die jeweiligen Äußerungen Kombinationen aus synthetischer und zuvor aufgezeichneter natürlicher Sprache sind.

[0003] Um die Erfindung besser zu verstehen, seien einige allgemeine Erläuterungen und Definitionen zur Sprachsynthese vorausgeschickt.

[0004] Gegenstand der Sprachsynthese ist die maschinelle Transformation der symbolischen Repräsentation einer Äußerung in ein akustisches Signal, welches von einem menschlichen Sprecher als der menschlichen Sprache hinreichend ähnlich anerkannt wird.

[0005] Im Bereich der Sprachsynthese gilt es, zwei unterschiedliche Systeme zu unterscheiden:

1) Ein Sprachsynthesesystem produziert, ausgehend von einem Text, gesprochene Sprache.
2) Ein Sprachsynthetisator produziert, ausgehend von gewissen Kontrollparametern gesprochene Sprache. Der Sprachsynthetisator stellt damit die letzte Stufe eines Sprachsynthesesystems dar.

[0006] Eine Sprachsynthesetechnik ist eine Technik, die den Bau eines Sprachsynthetisators erlaubt. Beispiele für Sprachsynthesetechniken sind die direkte Synthese, die Synthese mittels eines Modells und die Simulation des Vokaltraktes.

[0007] Bei der direkten Synthese werden entweder Teilstücke des Sprachsignals ausgehend von abgespeicherten Signalstücken (z. B. eines je Phonem) zu den entsprechenden Wörtern zusammengesetzt oder die Transferfunktion des Vokaltraktes, welcher beim Menschen für die Spracherzeugung benutzt wird, durch Energie eines Signals in bestimmten Frequenzbereichen nachgebildet. So werden hier beispielsweise stimmhafte Laute durch eine quasiperiodische Anregung einer bestimmten Frequenz repräsentiert.

[0008] Das oben erwähnte Phonem ist die kleinste bedeutungsunterscheidende, aber selbst nicht bedeutungstragende Einheit der Sprache. Zwei Wörter verschiedener Bedeutung, die sich nur durch ein Phonem unterscheiden (z. B. Fisch - Tisch; Wald - Wild), bilden ein Minimalpaar. Die Anzahl der Phoneme einer Sprache ist verhältnismäßig klein (zwischen 20 und 60). So existieren etwa im Deutschen 45 Phoneme.

[0009] Um die charakteristischen Übergänge zwischen den Phonemen zu berücksichtigen, werden bei der direkten Synthese meist Diphone verwendet. Vereinfacht kann ein Diphon als der Bereich vom invariablen Teil des ersten Phonems bis zum invariablen Teil des folgenden Phonems definiert werden.

[0010] Phoneme bzw. Sequenzen von Phonemen werden mit Hilfe des Internationalen Phonetischen Alphabets (IPA) notiert. Die Umsetzung eines Textes in eine Abfolge von Zeichen des Phonetischen Alphabetes wird als Phonetische Transkription bezeichnet.

[0011] Bei der Synthese mittels eines Modells wird ein Produktionsmodell gebildet, welches meist auf der Minimierung der Differenz zwischen einem digitalisierten menschlichen Sprachsignal (Originalsignal) und einem prädizierten Signal basiert.

[0012] Eine weitere Methode besteht in der Simulation des Vokaltraktes, bei der dessen Form sowie die Lage der einzelnen Artikulationsorgane (Zunge, Kiefer, Lippen) nachgebildet wird. Dazu wird ein mathematisches Modell der Strömungsverhältnisse in einem derartig definierten Vokaltrakt erzeugt und das Sprachsignal mit Hilfe dieses Modells berechnet.

[0013] Im folgenden sollen weitere Begriffe und Verfahren, die im Zusammenhang mit der Sprachsynthese Verwendung finden, kurz erläutert werden.

[0014] Die bei der direkten Synthese verwendeten Phoneme bzw. Diphone müssen zunächst durch Segmentierung aus natürlicher Sprache gewonnen werden. Hierbei können zwei Ansätze unterschieden werden:

[0015] Bei der impliziten Segmentierung werden nur die im Sprachsignal selbst enthaltenen Informationen zur Segmentierung verwendet.
Die explizite Segmentierung hingegen nutzt zusätzliche Informationen, wie z.B. die Anzahl der in der Äußerung enthaltenen Phoneme.

[0016] Zur Segmentierung müssen zunächst Merkmale aus dem Sprachsignal extrahiert werden, anhand derer eine Unterscheidung der Segmente möglich wird. Anschließend werden diese Merkmale in Klassen eingeordnet.
Möglichkeiten zur Merkmalsextraktion bieten unter anderem Spektralanalysen, Filterbankanalysen oder das Verfahren der Linearen Prädiktion.

[0017] Für die Klassifikation können beispielsweise

Hidden Markov Modelle, künstliche neuronale Netze oder Dynamic Time Warping (ein Verfahren zur Zeilnormalisierung) benutzt werden.

[0018] Das Hidden-Markov-Modell (HMM) ist ein zweistufiger stochastischer Prozess. Er besteht aus einer Markov-Kette mit einer meist geringen Zahl von Zuständen, denen Wahrscheinlichkeiten bzw. Wahrscheinlichkeitsdichten zugeordnet sind. Beobachtbar sind die Sprachsignale bzw. deren durch Wahrscheinlichkeitsdichten beschriebene Parameter. Die durchlaufende Zustandsfolge selbst bleibt verborgen. HMMs haben sich wegen ihrer Leistungsfähigkeit, Robustheit und guten Trainierbarkeit in der Spracherkennung weithin durchgesetzt.

[0019] Mit Hilfe des sogenannten Viterbi-Algorithmus kann die Übereinstimmung mehrerer HMMs bestimmt werden.

In neueren Ansätzen werden zur Klassifikation vielfach selbstorganisierende Merkmalskarten (Kohonen-Maps) verwendet. Diese spezielle Art eines Künstlichen Neuronalen Netzes ist in der Lage, die im menschlichen Gehirn ablaufenden Vorgänge nachzubilden.

[0020] Ein verbreiteter Ansatz ist die Klassifizierung in Stimmhaft / Stimmlos / Stille - gemäß der verschiedenen Anregungsformen bei der Erzeugung von Sprache im Vokaltrakt.

[0021] Gleichgültig, welche der eben genannten Synthesetechniken auch angewendet wird, bleibt bei text-basierten Syntheseanordnungen das Problem, dass, auch wenn zwischen der Aussprache einer als Text vorliegenden bzw. gespeicherten Zeichenfolge eine relativ große Korrelation gegeben ist, in jeder Sprache Worte vorhanden sind, bei denen aus der Schreibweise nicht ohne weiteres auf deren Aussprache geschlossen werden kann. Insbesondere für Eigennamen ist es vielfach nicht möglich, allgemeine phonetische Regeln zur Aussprache anzugeben. So haben zum Beispiel die beiden Städtenamen Itzehoe und Laboe die gleiche Endung, wenngleich Itzehoe mit "oe" und Laboe mit "ö" ausgesprochen wird. Liegen die jeweiligen Worte, die zur synthetischen Wiedergabe bestimmt sind, als Zeichenfolge vor, führt die Anwendung einer allgemeinen Regel dazu, dass in dem obigen Beispiel beiden Städtenamen entweder durchgängig mit "ö" oder "oe" ausgesprochen werden, was im Falle der "ö-Version" für Itzehoe und im Fall der "oe-Version" für Laboe aussprachetechnisch falsch wäre. Will man diese Besonderheiten berücksichtigen, ist es notwendig, dass die entsprechenden Worte dieser Sprache zur Wiedergabe einer besonderen Behandlung unterzogen werden müssen. Dies bedeutet aber gleichzeitig, dass keine rein text-basierte Eingabe der zur späteren Wiedergabe vorgesehenen Wörter mehr möglich ist.

[0022] Da die besondere Behandlung von bestimmten Wörtern einer Sprache außerordentlich aufwendig ist, ist man bei sprachgesteuerten Anordnungen dazu übergegangen, die Ansage, welche eine Anordnung angeben soll, aus einem Mix von gesprochener und synthetisierter Sprache zu bilden. Dazu wird beispielsweise bei einem Routefinder der gewünschte Zielort, welcher gegenüber den übrigen Worten der entsprechenden Sprache oftmals aussprachetechnische Besonderheiten ausweist und welcher bei sprachgesteuerten Anordnungen von einem Benutzer vorgegeben wird, aufgenommen und in die entsprechende Zielansage hineinkopiert. Dies führt dann dazu, dass bei der Zielansage "*In drei Kilometern erreichen sie* Itzehoe" nur der kursiv geschriebene Teil synthetisiert wurde und der restliche Teil "Itzehoe" aus der Zieleingabe des Benutzers entnommen wurde. Die gleichen Gegebenheiten treten auch bei der Einrichtung vom Mailboxen auf, bei denen bei der Einrichtung der Nutzer seinen Namen eingeben muss. Dort wird auch zur Vermeidung des Aufwands der entsprechende Ansagetext, der bei Verbindung eines Anrufers mit der Mailbox wiedergeben wird, aus dem synthetisierten Teil "Sie sind verbunden mit der Mailbox von" und dem originalen - bei der Einrichtung der Mailbox aufgenommenen - Teil "Otto Berger" gebildet.

[0023] Abgesehen davon, dass zusammengesetzte Ansagen der vorbeschriebenen Art einen eher wenig professionellen Eindruck hinterlassen, können sie auch durch die Einbindung der Originalsprache zu Abhörproblemen führen. In diesem Zusammenhang sei nur auf die Spracheingabe in lärmbelasteter Umgebung hingewiesen. Daher liegt der Erfindung die Aufgabe zugrunde, ein Wiedergabeverfahren für sprachgesteuerte Systeme mit text-basierter Sprachsynthese anzugeben, bei welchem die im Stand der Technik gegebenen Nachteile beseitigt werden.

Darstellung der Erfindung

[0024] Diese Aufgabe wird mit den in Anspruch 1 angegebenen Merkmalen gelöst. Vorteilhafte Aus- und Weiterbildungen der Erfindung sind den Ansprüchen 2 bis 9 entnehmbar.

[0025] Wird gemäß Anspruch 1 beim Vorliegen einer tatsächlich gesprochenen und mit einer gespeicherten Zeichenkette korrespondierenden Spracheingabe vor einer tatsächlichen Wiedergabe der nach allgemeinen Regeln phonetisch beschriebenen und in eine rein synthetische Form gewandelten Zeichenkette die gewandelte Zeichenkette mit der gesprochenen Spracheingabe verglichen und erfolgt die tatsächliche Wiedergabe der gewandelten Zeichenkette erst dann, wenn der Vergleich dieser Zeichenkette mit der tatsächlich gesprochenen Spracheingabe eine unterhalb einer Schwelle liegenden Abweichung zeigt, wird die Verwendung der Originalsprache bei der Wiedergabe entsprechend dem Stand der Technik überflüssig. Dies ist selbst dann der Fall, wenn das gesprochene Wort von der diesem Wort entsprechenden, gewandelten Zeichenfolge erheblich abweicht. Hierbei muss lediglich sichergestellt werden, dass von der gewandelten Zei-

chenkette wenigstens eine Variante gebildet wird und dass die gebildete Variante, sofern diese bei einem Vergleich mit der originalen Spracheingabe eine unterhalb der Schwelle liegende Abweichung aufweist, anstelle der -ursprünglich- gewandelten Zeichenkette ausgegeben wird.

[0026] Wird das Verfahren gemäß Anspruch 2 durchgeführt, ist der Rechen- und Speicheraufwand relativ gering. Dies ist darauf zurückzuführen, dass immer nur eine Variante gebildet und untersucht werden muss.

[0027] Werden gemäß Anspruch 3 wenigstens zwei Varianten gebildet und wird aus den hergestellten Varianten diejenige herausgesucht, welche die geringsten Abweichungen zur originalen Spracheingabe hat, ist im Gegensatz zur Verfahrensführung gemäß Anspruch 2 immer eine der originalen Spracheingabe entsprechende synthetische Wiedergabe möglich.

[0028] Die Verfahrensführung wird vereinfacht, wenn gemäß Anspruch 4 eine Segmentierung der Spracheingabe und der gewandelten Zeichenkette bzw. der daraus gebildeten Varianten erfolgt. Diese Segmentierung erlaubt es, Segmente, in denen keine bzw. unter der Schwelle liegende Unterschiede festgestellt werden, von der weiteren Behandlung auszuschließen.

[0029] Wird gemäß Anspruch 5 ein gleicher Segmentierungsansatz verwendet, ist der Vergleich besonders einfach, da eine direkte Zuordnung der jeweiligen Segmente gegeben ist.

[0030] Wie Anspruch 6 zeigt, können auch verschiedene Segmentierungsansätze verwendet werden. Dies hat insbesondere bei der Betrachtung der originalen Spracheingabe Vorteile, weil dort zur Segmentierung zwingend die im Sprachsignal enthaltenen und nur in einem sehr aufwendigen Schritt ermittelbaren Informationen genutzt werden müssen, während bei der Segmentierung von Zeichenketten sehr einfach die bekannte Anzahl der in der Äußerung enthaltenen Phoneme genutzt werden kann.

[0031] Sehr rationell wird die Verfahrensführung dann, wenn gemäß Anspruch 8 die Segmente ausgeschieden werden, in denen ein hohes Maß an Übereinstimmung besteht, und nur noch das Segment der Zeichenkette, welches zu dem korrespondierenden Segment der originalen Spracheingabe eine oberhalb der Schwelle liegenden Abweichung zeigt, dadurch variiert wird, indem das in dem Segment der Zeichenkette vorliegende Phonem durch ein Ersatzphonem ersetzt wird.

[0032] Eine besonders einfache Verfahrensführung wird erreicht, wenn gemäß Anspruch 9 zu jedem Phonem wenigstens ein diesem Phonem ähnliches Ersatzphonem verknüpft bzw. in einer Liste abgelegt ist.

[0033] Die Rechenarbeit wird weiter verringert, wenn gemäß Anspruch 10 bei einer als wiedergabewürdig ermittelten Variante einer Zeichenkette die Besonderheiten, die mit der Wiedergabe der Zeichenkette verbunden sind, zusammen mit der Zeichenkette abgespeichert werden. In diesem Fall ist dann die besondere Aussprache der jeweiligen Zeichenkette bei späterer Nutzung ohne großen Aufwand sofort aus dem Speicher abrufbar.

Kurze Darstellung der Figuren

[0034] Es zeigen:

Fig. 1 einen schematischen Ablauf gemäß der Erfindung

Fig. 2 einen Vergleich von segmentierten Äußerungen

Wege zum Ausführen der Erfindung

[0035] Die Erfindung soll nun anhand der beiden Figuren näher erläutert werden.

[0036] Um die Wirkungen der Erfindung besser darlegen zu können, wird von einem sprachgesteuerten System mit text-basierter Sprachsynthese ausgegangen. Derartige Systeme sind beispielsweise in Routefindern oder Mailboxanordnungen realisiert, so dass sich wegen der hohen Verbreitung derartiger Systeme deren Darstellung auf die Dinge beschränken kann, die für die Ausführung der Erfindung zwingend notwendig sind.

[0037] Allen diesen Systemen ist ein Speicher gemein, in welchem eine Mehrzahl von Zeichenketten abgelegt sind. Bei diesen Zeichenketten kann es sich bei einem Routefinder beispielsweise um Straßen- oder Ortsnamen handeln. In einer Mailboxanwendung können dies wie in einem Telefonbuch die Namen von Anschlussinhabern sein. Damit die Speicher leicht mit den entsprechenden Informationen beladen bzw. die gespeicherten Informationen leicht upgedatet werden können, liegen die jeweiligen Zeichenketten als Text vor.

[0038] In Fig. 1, die den schematischen Ablauf entsprechend dem erfinderischen Verfahren zeigt, ist ein solcher Speicher mit 10 bezeichnet. Dieser Speicher 10, welcher für die Darstellung der Erfindung die deutschen Städtenamen enthalten soll, gehört zu einem Routefinder 11. Außerdem umfasst dieser Routefinder 11 eine Anordnung 12, mit welcher natürliche Spracheingaben aufgenommen und temporär gespeichert werden können. Vorliegend ist dies so realisiert, dass die jeweilige Spracheingabe von einem Mikrophon 13 erfasst und in einem Sprachspeicher 14 abgelegt wird. Wird nun ein Benutzer vom Routefinder 11 aufgefordert, seine Zieleingabe zu machen, wird der jeweils vom Benutzer ausgesprochene Zielort z. B. "Bochum" oder "Itzehoe" vom Mikrophon 13 erfasst und an den Sprachspeicher 14 weitergeben. Da der Routefinder 11 entweder seinen derzeitigen Standort mitgeteilt bekommen hat oder aber ihn noch kennt, wird er zunächst anhand der gewünschten Zieleingabe und dem derzeitigen Standort die entsprechende Fahrtroute zum Zielort ermitteln. Soll der Routefinder 11 die entsprechende

Fahrtroute nicht nur graphisch zeigen, sondern gesprochene Ansage liefern, werden die textlich hinterlegten Zeichenketten der jeweiligen Ansage nach allgemeinen Regeln phonetisch beschrieben und anschließend für die Sprachausgabe in eine rein synthetische Form gewandelt. In dem in Fig. 1 gezeigten Ausführungsbeispiel erfolgt die phonetische Beschreibung der hinterlegten Zeichenketten im Umsetzer 15 und die Synthetisierung in der nachfolgend angeordneten Sprachsynthetisieranordnung 16.

[0039] Solange die über die Spracheingabe aufgerufenen und zur Wiedergabe bestimmten Zeichenketten in bezug auf ihre jeweilige Aussprache den Regeln der phonetischen Transkription der Sprache, in welcher der Dialog zwischen dem Benutzer und dem Routefinder 11 geführt werden soll, folgen, kann die jeweilige Zeichenkette, wenn sie dem Umsetzer 15 und die Sprachsynthetisieranordnung 16 durchlaufen hat, als ein den phonetischen Gegebenheiten der jeweiligen Sprache entsprechendes Wort mittels eines Lautsprechers 17 an die Umwelt abgegeben und von dieser als solches auch verstanden werden. Dies bedeutet für einen Routefinder 11 der vorbeschriebenen Art, dass beispielsweise der aus einer Mehrzahl von Zeichenketten bestehende, zur Wiedergabe bestimmte und über die Spracheingabe initiierte Wiedergabetext "An der nächsten Kreuzung rechts abbiegen!" problemlos, d.h. entsprechend den phonetischen Gegebenheiten der Sprache über den Lautsprecher 17 abgegeben und auch verstanden werden kann, da diese Information keinen Eigentümlichkeiten bei der Wiedergabe unterliegt.

[0040] Soll aber beispielsweise dem Benutzer nach Eingabe des Zielorts die Möglichkeit eingeräumt werden, die Richtigkeit seiner Zieleingabe zu überprüfen, wird der Routefinder 11 nach der Zieleingabe etwa folgenden Satz wiedergeben: "Sie haben als Ziel Berlin gewählt. Sofern dies nicht Ihren Vorstellungen entspricht, geben sie jetzt ein neues Ziel ein." Auch wenn diese Information nach allgemeinen Regeln phonetisch richtig wiedergegeben werden kann, treten dann Probleme auf, wenn das Ziel nicht Berlin, sondern Laboe sein soll. Wird die Zeichenkette, welche die textliche Darstellung des Zielortes Laboe im Umsetzer 15 nach allgemeinen Regeln phonetisch geschrieben und anschließend in der Sprachsynthetisieranordnung 16 zur Ausgabe über den Lautsprecher 17 wie der übrige Teil der obigen Information in eine synthetische Form gebracht, wäre das über den Lautsprecher 17 abgegebene Ergebnis nur dann richtig, wenn nach allgemeinen Regeln die Endung "oe" grundsätzlich als "ö" wiedergeben wird. Die Richtigkeit der Wiedergabe des Zielorts Laboe im letzten Fall führt aber dann zwangsläufig zu einer fehlerhaften Wiedergabe, wenn der Benutzer als Zielort Itzehoe wählt, denn wegen der grundsätzlichen Aussprache der "oe" als "ö" würde der Zielort dann phonetisch falsch als "Itzehö" wiedergegeben.

[0041] Um dies zu vermeiden, ist zwischen der Sprachsynthetisieranordnung 16 und dem Lautsprecher 17 eine Vergleichsanordnung 18 angeordnet. Dieser Vergleichsanordnung 18 werden der tatsächlich vom Benutzer gesprochene Zielort und die dem Zielort entsprechende Zeichenkette, nachdem sie den Umsetzer 15 und die Sprachsynthetisieranordnung 16 durchlaufen hat, zugeführt und anschließend verglichen. Zeigt die synthetisierte Zeichenkette eine hohe - oberhalb einer Schwelle liegenden - Übereinstimmung mit dem original gesprochenen Zielort, wird für die Wiedergabe die synthetisierte Zeichenkette verwendet. Kann diese Übereinstimmung nicht festgestellt werden, wird in der Sprachsynthetisieranordnung 16 eine Variante der ursprünglichen Zeichenkette gebildet und im Vergleicher 18 erneut ein Vergleich zwischen dem original gesprochenen Zielort und der gebildeten Variante durchgeführt.

[0042] Ist der Routefinder 11 so ausgebildet, dass sobald eine Zeichenkette bzw. eine Variante die geforderte Übereinstimmung mit dem Original aufweist, deren Wiedergabe über den Lautsprecher 17 erfolgt, werden weitere Variantenbildungen sofort gestoppt. Auch kann der Routefinder 11 so modifiziert sein, dass eine Mehrzahl von Varianten gebildet werden und dann aus den Varianten diejenige Variante ausgewählt wird, die die größte Übereinstimmung mit dem Original zeigt.

[0043] Wie der Vergleich im Vergleicher 18 ausgeführt wird, wird im Zusammenhang mit Fig. 2a und b näher gezeigt. Dort ist in Fig. 2a ein Sprachsignal im Zeitbereich des tatsächlich von einem Benutzer gesprochenen Wortes Itzehoe dargestellt. Fig. 2b zeigt ebenfalls ein Sprachsignal im Zeitbereich des Wortes Itzehoe, wobei jedoch im in Fig. 2b gezeigten Fall das Wort Itzehoe aus einer entsprechend vorliegenden Zeichenkette zunächst im Umsetzer 15 nach allgemeinen Regeln phonetisch beschrieben und dann anschließend in der Sprachsynthetisieranordnung 16 in eine synthetische Form gebracht wurde. Deutlich ist der Darstellung gemäß Fig. 2b entnehmbar, dass bei Anwendung der allgemeinen Regeln die Endung "oe" des Wortes Itzehoe als "ö" wiedergeben wird. Um jedoch diese fehlerhafte Wiedergabe auszuschließen, werden die gesprochene und die synthetisierte Form in einem Vergleicher 18 miteinander verglichen.

[0044] Um diesen Vergleich zu vereinfachen, werden sowohl die gesprochene als auch die synthetisierte Form in Segmente 19, 20 unterteilt und dann der Vergleich zwischen korrespondierenden Segmenten 19/20 durchgeführt. In dem in Fig. 2a und b gezeigten Ausführungsbeispiel zeigt sich, dass lediglich in den beiden letzten Segmenten 19.6, 20.6 eine starke Abweichung gegeben ist, während der Vergleich der übrigen Segmentpaare 19.1/20.1, 19.2/20.2 ... 19.5/20.5 eine relativ große Übereinstimmung zeigen. Wegen der starken Abweichung in dem Segmentpaar 19.6/20.6 wird die phonetische Beschreibung im Segment 20.6 anhand einer in einem Speicher 21 (Fig. 1) hinterlegten Liste, welche besser passende bzw. ähnliche Phoneme enthält, verändert. Da vorliegend das fragliche Phonem "ö"

ist und die Liste mit ähnlichen Phonemen die Ersatzphoneme "o" und "oh" vorsieht, wird das Phonem "ö" gegen das Ersatzphonem "o" ausgetauscht. Dazu wird die hinterlegte Zeichenkette in einem Umsetzer 15' (Fig. 1) erneut phonetisch beschrieben, in der Sprachsynthetisieranordnung 16 in eine synthetische Form gebracht und erneut mit der tatsächlich gesprochenen Zieleingabe im Vergleicher 18 verglichen.

[0045] Nur der Vollständigkeit halber sei darauf hingewiesen, dass der Umsetzer 15' in einem anderen - nicht dargestellten - Ausführungsbeispiel auch vom Umsetzer 15 gebildet sein kann.

[0046] Zeigt sich, dass die entsprechend modifizierte Zeichenkette, welche im Zusammenhang mit dieser Anmeldung auch als Variante bezeichnet wird, keine oberhalb einer Schwelle liegende Übereinstimmung mit dem gesprochenen Wort hat wird die Prozedur mit einem weiteren Ersatzphonem nochmals ausgeführt. Liegt der Grad der Übereinstimmung dann oberhalb der Schwelle, wird das entsprechend synthetisierte Wort über den Lautsprecher 17 ausgegeben.

[0047] Auch kann der Verfahrensablauf modifiziert sein. Wird festgestellt, dass eine Abweichung zwischen der gesprochenen und der ursprünglichen synthetischen Form gegeben ist, und liegen eine Mehrzahl von Ersatzphonemen in der im Speicher 21 ablegten Liste vor, können auch gleichzeitig eine Mehrzahl von Varianten gebildet und mit dem tatsächlich gesprochenen Wort verglichen werden. Wiedergegeben wird dann diejenige Variante, die die größte Übereinstimmung mit dem gesprochenen Wort zeigt.

[0048] Soll vermieden werden, dass bei der mehrfachen Benutzung von Worten, die die obige Prozedur auslösen können, immer die richtige -synthetische- Aussprache aufwendig ermittelt werden muss, kann, wenn beispielweise die richtige synthetische Aussprache zum Beispiel des Wortes Itzehoe ermittelt worden ist, die entsprechende Modifikation mit Hinweis auf die Zeichenkette Itzehoe gespeichert werden. Dies bedeutet, dass bei eine erneuten Anforderung der Zeichenkette Itzehoe gleichzeitig zur richtigen Aussprache dieses Wortes die von der phonetischen Beschreibung nach allgemeinen Regeln abweichenden Besonderheiten berücksichtigt werden, so dass der Vergleichsschritt im Vergleicher 18 entfallen kann. Um diese Modifikation sichtbar zu machen, wurde in Fig. 1 ein Zusatzspeicher 22 gestrichelt angedeutet, in welchem die auf Modifikationen von hinterlegten Zeichenketten hinweisenden Informationen abgelegt werden.

[0049] Nur der Vollständigkeit halber sei auch darauf hingewiesen, dass der Zusatzspeicher 22 nicht nur auf die Aufnahme von Informationen zur richtigen Aussprache von hinterlegten Zeichenketten beschränkt ist. Ergibt beispielsweise ein Vergleich im Vergleicher 18, dass zwischen der gesprochenen und der synthetisierten Form eines Wortes keine bzw. unterhalb einer Schwelle liegende Abweichung gegeben sind, kann im Zusatzspeicher 22 für dieses Wort ein Hinweis hinterlegt werden, welcher bei der künftigen Verwendung dieses Wortes einen aufwendigen Vergleich im Vergleicher 18 ausschließt.

[0050] Auch ist den Fig. 2a und b entnehmbar, dass die Segmente 19 gemäß Fig. 2a und die Segmente 20 gemäß Fig. 2b kein gleiches Format besitzen. So hat beispielsweise das Segment 20.1 im Vergleich zum Segment 19.1 eine größere Breite, während das Segment 20.2 gegenüber dem korrespondierenden Segment 19.2. wesentlich schmaler ausgebildet ist. Dies ist darauf zurückzuführen, dass die "Sprechlänge" der verschiedenen zum Vergleich anstehenden Phoneme unterschiedlich lang sein kann. Da aber derart unterschiedliche lange Sprechzeiten nicht ausgeschlossen werden können, ist die Vergleichsanordnung 18 so ausgelegt, dass verschieden lange Aussprechzeiten eines Phonemes noch keine gegenseitige Abweichung indizieren.

[0051] Nur der Vollständigkeit halber sei darauf hingewiesen, dass bei der Verwendung von verschiedenen Segmentierungsverfahren für das gesprochene und das synthetisierte Format auch eine unterschiedliche Anzahl von Segmenten 19, 20 berechnet werden können. Tritt dies ein, sollte dann ein bestimmtes Segment 19, 20 nicht nur mit einem korrespondieren Segment 19, 20 verglichen werden, sondern ebenfalls mit dem Vorgänger und Nachfolger des korrespondierenden Segments 19, 20. Somit ist es auch möglich, ein Phonem durch zwei andere Phoneme zu ersetzen. Dieses Vorgehen ist in umgekehrter Richtung ebenfalls möglich. Gibt es keine Übereinstimmung für ein Segment 19, 20, so kann dieses ausgeschlossen, oder durch zwei besser passende ersetzt werden.

Patentansprüche

1. Wiedergabeverfahren für sprachgesteuerte Systeme mit text-basierter Sprachsynthese, **dadurch gekennzeichnet,**

   dass beim Vorliegen einer tatsächlich gesprochenen und mit einer gespeicherten Zeichenkette korrespondierenden Spracheingabe vor einer Wiedergabe der nach allgemeinen Regeln phonetisch beschriebenen und in eine rein synthetische Form gewandelten Zeichenkette die gewandelte Zeichenkette mit der Spracheingabe verglichen wird,

   dass bei Feststellung einer oberhalb einer Schwelle liegenden Abweichung der gewandelten Zeichenkette von der Spracheingabe wenigstens eine Variante der gewandelten Zeichenkette gebildet wird und

   dass eine der gebildeten Varianten, sofern diese bei einem Vergleich mit der Spracheingabe eine unterhalb der Schwelle liegende

Abweichung aufweist, anstelle der gewandelten Zeichenkette ausgegeben wird.

2. Wiedergabeverfahren nach Anspruch 1, **dadurch gekennzeichnet,**

dass in Schritt zwei jeweils immer nur eine Variante gebildet wird und

dass, sofern in Schritt drei ein Vergleich der Variante mit der Spracheingabe immer eine oberhalb der Schwelle liegende Abweichung zeigt, Schritt zwei mindestens noch einmal zur Bildung einer neuen Variante durchgeführt wird.

3. Wiedergabeverfahren nach Anspruch 1, **dadurch gekennzeichnet,**

dass in Schritt zwei wenigstens zwei Varianten gebildet werden und

dass beim Vorliegen von Varianten, die jeweils im Vergleich zur Spracheingabe eine unterhalb der Schwelle liegende Abweichung haben, immer diejenige Variante wiedergegeben wird, die die geringste Abweichung zur Spracheingabe besitzt.

4. Verfahren nach einem der Ansprüche 1 bis 3, **dadurch gekennzeichnet,**

dass vor einem Vergleich der Spracheingabe mit der gewandelten Zeichenkette bzw. der daraus gebildeten Variante(n) eine Segmentierung der Spracheingabe und der gewandelten Zeichenkette bzw. der gebildeten Variante(n) erfolgt.

5. Wiedergabeverfahren nach Anspruch 4, **dadurch gekennzeichnet,**

dass das sowohl zur Segmentierung der Spracheingabe und der gewandelten Zeichenkette bzw. der daraus abgeleiteten Variante(n) ein gleicher Segmentierungsansatz verwendet wird.

6. Wiedergabeverfahren nach Anspruch 4, **dadurch gekennzeichnet,**

dass das sowohl zur Segmentierung der Spracheingabe und der gewandelten Zeichenkette bzw. der daraus abgeleiteten Variante(n) jeweils ein verschiedener Segmentierungsansatz verwendet wird.

7. Wiedergabeverfahren nach Anspruch 4, **dadurch gekennzeichnet,**

dass zur Segmentierung der gewandelten Zeichenkette bzw. der daraus abgeleiteten Variante(n) ein explizierter und zur Segmentierung der Spracheingabe ein implizierter Segmentierungsansatz verwendet wird.

8. Wiedergabeverfahren nach einem der Ansprüche 4 bis 7, **dadurch kennzeichnet,**

dass die in segmentierter Form vorliegende gewandelte Zeichenkette und die segmentierte Spracheingabe in den entsprechenden Segmenten auf Gemeinsamkeiten untersucht wird und
dass, wenn in zwei korrespondierenden Segmenten eine oberhalb eines Schwellwerts liegende Abweichung vorliegt, das in dem Segment der gewandelten Zeichenkette vorliegende Phonem durch ein Ersatzphonem ersetzt wird.

9. Wiedergabeverfahren nach Anspruch 8, **dadurch gekennzeichnet,**

dass mit jedem Phonem wenigstens ein diesem Phonem ähnliches Ersatzphonem verknüpft ist.

10. Wiedergabeverfahren nach einem der Ansprüche 1 bis 9, **dadurch gekennzeichnet,**

dass, sobald eine Variante einer Zeichenkette als wiedergabewürdig ermittelt wird, die Besonderheiten, die mit der Wiedergabe der Zeichenkette verbunden sind, im Zusammenhang mit der Zeichenkette abgespeichert werden.

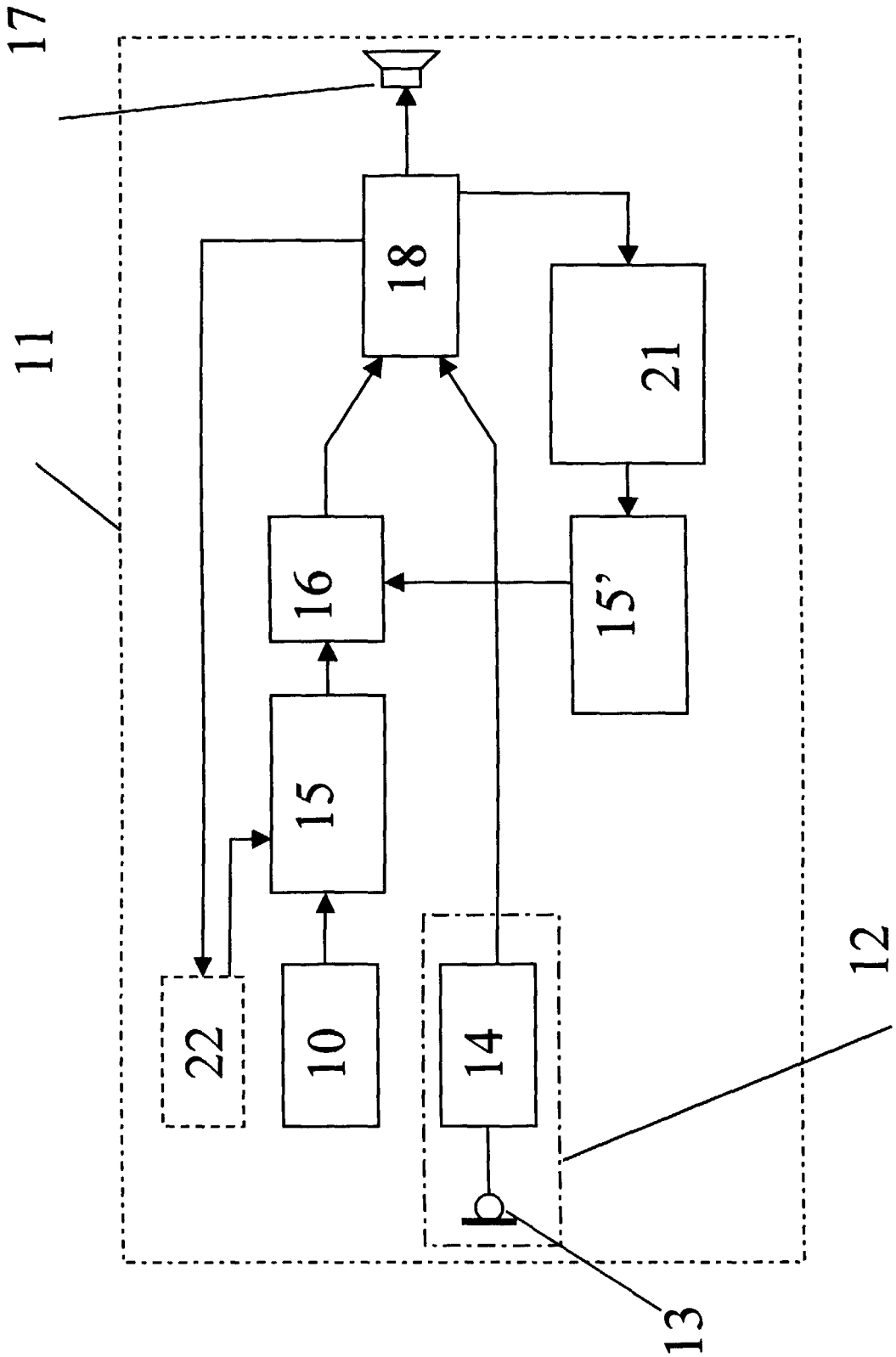

Fig. 1

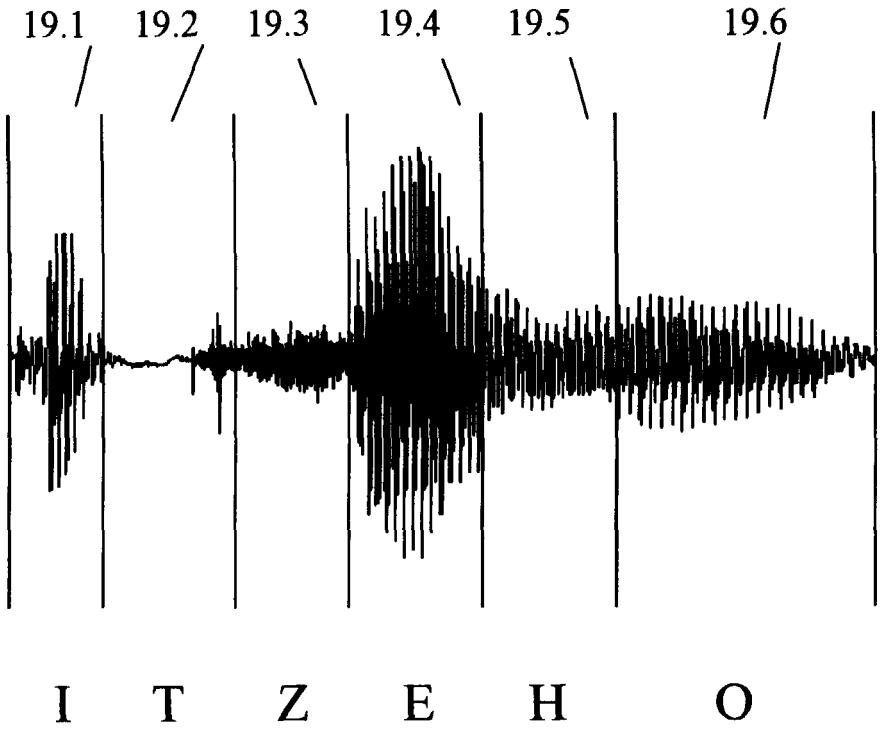

Fig. 2a

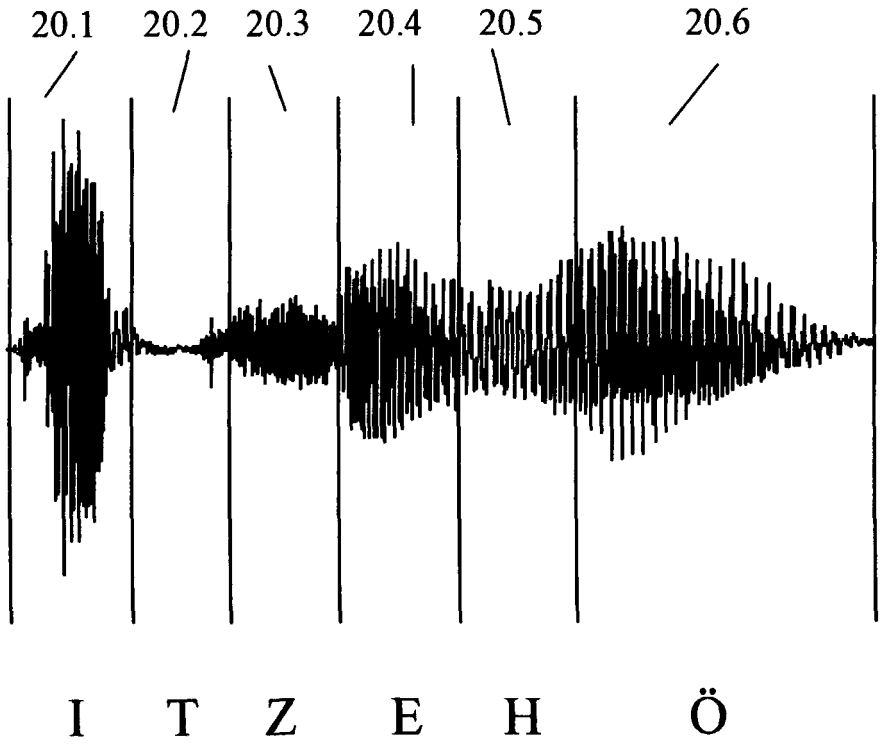

Fig. 2b