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

(11) EP 1 184 839 A2

(12) EUROPÄISCHE PATENTANMELDUNG

(43) Veröffentlichungstag:
06.03.2002 Patentblatt 2002/10

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

(21) Anmeldenummer: **01117869.6**

(22) Anmeldetag: **23.07.2001**

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

(30) Priorität: **31.08.2000 DE 10042944**

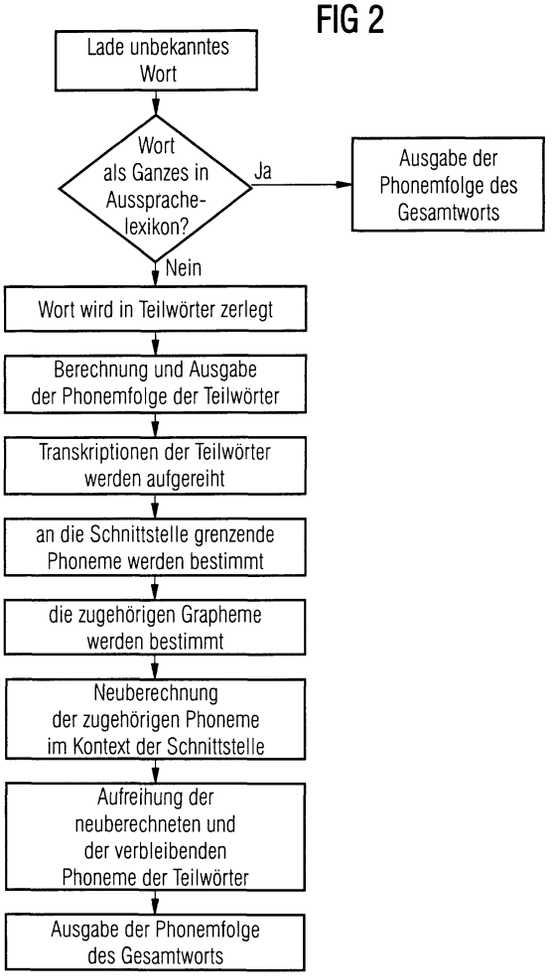
(71) Anmelder: **SIEMENS AKTIENGESELLSCHAFT**
80333 München (DE)

(72) Erfinder: **Hain, Horst-Udo**
81825 München (DE)

(54) Graphem-Phonem-Konvertierung

(57) Bei dem Verfahren zur Graphem-Phonem-Konvertierung eines Worts, das als Ganzes nicht in einem Aussprachelexikon enthalten ist, wird das Wort zunächst in Teilwörter zerlegt. Die Teilwörter werden transkribiert und verkettet. Dadurch bilden sich Schnittstellen zwischen den Transkriptionen der Teilwörter. Die Phoneme an den Schnittstellen müssen häufig geändert werden. Daher werden sie einer erneuten Berechnung unterzogen.

FIG 2



EP 1 184 839 A2

Beschreibung

[0001] Die Erfindung betrifft ein Verfahren, ein Computerprogrammprodukt, einen Datenträger und ein Computersystem zur Graphem-Phonem-Konvertierung eines Worts, das als Ganzes nicht in einem Aussprachelexikon enthalten ist.

[0002] Sprachverarbeitungsverfahren im Allgemeinen sind beispielsweise aus US 6 029 135, US 5 732 388, DE 19636739 C1 und DE 19719381 C1 bekannt. Bei einem Sprachsynthese-System ist die Schrift-zu-Sprache- bzw. Graphem-Phonem-Konvertierung der zu sprechenden Wörter von entscheidender Bedeutung. Fehler bei Lauten, Silbengrenzen und der Wortbetonung sind direkt hörbar, können zur Unverständlichkeit führen und im schlimmsten Fall sogar den Sinn einer Aussage verdrehen.

[0003] Die beste Qualität erhält man, wenn das zu sprechende Wort in einem Aussprachelexikon enthalten ist. Die Verwendung solcher Lexika bereitet jedoch Probleme. Auf der einen Seite erhöht die Anzahl der Einträge den Suchaufwand. Auf der anderen Seite ist es gerade bei Sprachen wie dem Deutschen nicht möglich, alle Wörter in einem Lexikon zu erfassen, da die Möglichkeiten der Kompositabildung nahezu unbeschränkt sind.

[0004] Abhilfe kann in diesem Fall eine morphologische Zerlegung schaffen. Ein Wort, das nicht im Lexikon gefunden wird, wird in seine morphologischen Bestandteile wie Präfixe, Stämme und Suffixe zerlegt, und diese Bestandteile werden im Lexikon gesucht. Eine morphologische Zerlegung ist jedoch gerade bei langen Wörtern problematisch, weil die Anzahl der möglichen Zerlegungen mit der Wortlänge steigt. Sie erfordert außerdem ein ausgezeichnetes Wissen über die Wortbildungsgrammatik einer Sprache. Daher werden Wörtern, die nicht in einem Aussprachelexikon gefunden werden, mit Out-Of-Vocabulary-Verfahren (OOV-Verfahren), z.B. mit Neuronalen Netzen, transkribiert. Solche OOV-Behandlungen sind allerdings relativ rechenintensiv und führen in aller Regel zu schlechteren Ergebnissen als die phonetische Konvertierung ganzer Wörter mit Hilfe eines Aussprachelexikons. Zur Bestimmung der Aussprache eines Worts, das nicht in einem Aussprachelexikon enthalten ist, kann das Wort auch in Teilwörter zerlegt werden. Die Teilwörter können mit Hilfe eines Aussprachelexikons oder eines OOV-Verfahrens transkribiert werden. Die gefundenen Teiltranskriptionen können aneinander gehängt werden. Dies führt jedoch zu Fehlern an den Trennstellen zwischen den Teiltranskriptionen.

[0005] Aufgabe der Erfindung ist es, das Aneinanderfügen von Teiltranskriptionen zu verbessern. Diese Aufgabe wird durch ein Verfahren, ein Computerprogrammprodukt, einen Datenträger und ein Computersystem gemäß den unabhängigen Ansprüchen gelöst.

[0006] Dabei wird unter einem Computerprogrammprodukt das Computerprogramm als handelbares Produkt verstanden, in welcher Form auch immer, z.B. auf Papier, auf einem computerlesbaren Datenträger, über ein Netz verteilt, etc.

[0007] Erfindungsgemäß wird bei der Graphem-Phonem-Konvertierung eines Worts, das als Ganzes nicht in einem Aussprachelexikon enthalten ist, zunächst das Wort in Teilwörter zerlegt. Anschließend wird eine Graphem-Phonem-Konvertierung der Teilwörter durchgeführt.

[0008] Die Transkriptionen der Teilwörter werden hintereinander aufgereiht, wobei sich mindestens eine Schnittstelle zwischen den Transkriptionen der Teilwörter ergibt. Die an die mindestens eine Schnittstelle grenzenden Phoneme der Teilwörter werden bestimmt.

[0009] Dabei besteht die Möglichkeit, nur das letzte Phonem des in der zeitlichen Reihenfolge der Aussprache vor der Schnittstelle liegenden Teilworts zu berücksichtigen. Besser ist es jedoch, wenn sowohl das genannte als auch das erste Phonem der folgenden Silbe für die erfindungsgemäße Sonderbehandlung ausgewählt werden. Noch bessere Ergebnisse werden erzielt, wenn weitere angrenzende Phoneme einbezogen werden, z.B. ein oder zwei Phoneme vor der Schnittstelle und zwei nach der Schnittstelle.

[0010] Anschließend werden diejenigen Grapheme der Teilwörter bestimmt, die die an die mindestens eine Schnittstelle grenzenden Phoneme erzeugen. Dies kann mittels eines Lexikons erfolgen, das angibt, durch welche Grapheme diese Phoneme erzeugt wurden. Wie das Lexikon zu erstellen ist, ist in Horst-Udo Hain: "Automation of the Training Procedures for Neural Networks Performing Multi-Lingual Grapheme to Phoneme Conversion", Eurospeech 1999, S. 2087-2090, ausgeführt.

[0011] Danach wird die Graphem-Phonem-Konvertierung der bestimmten Grapheme im Kontext, das heißt in Abhängigkeit des Kontexts, der jeweiligen Schnittstelle neu berechnet. Dies ist nur möglich, weil klar ist, welches Phonem durch welches Graphem bzw. welche Grapheme erzeugt wurde.

[0012] Die Schnittstellen zwischen den Teiltranskriptionen werden somit gesondert behandelt. Gegebenenfalls werden Änderungen an den vorher ermittelten Teiltranskriptionen vorgenommen. Ein für ein Sprachsynthese-System nicht unerheblicher Vorteil der Erfindung ist die Beschleunigung der Berechnung. Während Neuronale Netze für die Konvertierung der 310000 Wörter eines typischen Lexikons für die deutsche Sprache ca. 80 Minuten benötigen, geschieht dies mit dem erfindungsgemäßen Ansatz in nur 25 Minuten.

[0013] In einer vorteilhaften Weiterbildung der Erfindung kann die Graphem-Phonem-Konvertierung der Grapheme im Kontext der jeweiligen Schnittstelle mittels eines Neuronalen Netzes neu berechnet werden. Ein Aussprachelexikon

hat den Vorteil, die "richtige" Transkription zu liefern. Es versagt jedoch, wenn unbekannte Wörter auftreten. Neuronale Netze können hingegen für jede beliebige Zeichenkette eine Transkription liefern, machen dabei aber unter Umständen erhebliche Fehler. Die Weiterbildung der Erfindung kombiniert die Sicherheit des Lexikons mit der Flexibilität der Neuronalen Netze.

[0014] Die Transkription der Teilwörter kann auf verschiedene Weise erfolgen, z.B. mittels einer Out-of-Vocabulary-Behandlung (OOV-Behandlung). Ein recht zuverlässiger Weg besteht darin, für das Wort in einer Datenbank, die phonetische Transkriptionen von Wörtern enthält, nach Teilwörtern zu suchen. Als Transkription wird dann für ein in der Datenbank gefundenes Teilwort die in der Datenbank verzeichnete phonetische Transkription gewählt. Dies führt für die meisten Wörter bzw. Teilwörter zu brauchbaren Ergebnissen.

[0015] Falls das Wort neben dem gefundenen Teilwort mindestens einen weiteren Bestandteil aufweist, der nicht in der Datenbank verzeichnet ist, kann dieser mittels einer OOV-Behandlung phonetisch transkribiert werden. Die OOV-Behandlung kann mittels eines statistischen Verfahrens, z.B. mittels eines Neuronalen Netzes, oder regelbasiert erfolgen. Vorteilhafterweise wird das Wort in Teilwörter einer gewissen Mindestlänge zerlegt, damit möglichst große Teilwörter gefunden werden und entsprechend wenig Nachbesserungen anfallen.

[0016] Weitere vorteilhafte Weiterbildungen der Erfindung sind in den Unteransprüchen gekennzeichnet.

[0017] Im folgenden wird die Erfindung anhand von Ausführungsbeispielen näher erläutert, die in den Figuren schematisch dargestellt sind. Im einzelnen zeigt:

Fig. 1 ein zur Graphem-Phonem-Konvertierung geeignetes Computersystem; und
Fig. 2 eine schematische Darstellung des erfindungsgemäßen Verfahrens.

[0018] Fig. 1 zeigt ein zur Graphem-Phonem-Konvertierung eines Worts geeignetes Computersystem. Dies weist einen Prozessor (processor, CPU) 20, einen Arbeitsspeicher (RAM) 21, einen Programmspeicher (programm memory, ROM) 22, einen Festplatten-Controller (hard disc controller, HDC) 23, der eine Festplatte (hard disk) 30 steuert, und einen Schnittstellen-Controller (I/O controller) 24 auf. Prozessor 20, Arbeitsspeicher 21, Programmspeicher 22, Festplatten-Controller 23 und Schnittstellen-Controller 24 sind über einen Bus, den CPU-Bus 25, zum Austausch von Daten und Befehlen miteinander gekoppelt. Ferner weist der Computer einen Ein-/Ausgabe-Bus (I/O Bus) 26 auf, der verschiedene Ein- und Ausgabeeinrichtungen mit dem Schnittstellen-Controller 24 koppelt. Zu den Ein- und Ausgabeeinrichtungen zählen z.B. eine allgemeine Ein- und Ausgabe-Schnittstelle (I/O interface) 27, eine Anzeigeeinrichtung (display) 28, eine Tastatur (keyboard) 29 und eine Maus 31.)

[0019] Betrachten wir als Beispiel für die Graphem-Phonem-Konvertierung das deutsche Wort "überflüssigerweise".

[0020] Zunächst wird versucht, das Wort in Teilwörter zu zerlegen, die Bestandteile eines Aussprache-Lexikons sind. Um die Anzahl der möglichen Zerlegungen auf ein sinnvolles Maß zu beschränken, wird für die gesuchten Bestandteile eine Mindestlänge vorgegeben. Für die deutsche Sprache haben sich 6 Buchstaben als Mindestlänge in der Praxis bewährt.

[0021] Alle gefundenen Bestandteile werden in einer verketteten Liste abgespeichert. Bei mehreren Möglichkeiten wird immer der längste Bestandteil bzw. der Pfad mit den längsten Bestandteilen verwendet.

[0022] Werden nicht alle Teile des Worts als Teilwörter im Aussprachelexikon gefunden, so werden die verbleibenden Lücken im bevorzugten Ausführungsbeispiel durch ein Neuronales Netz geschlossen. Im Gegensatz zur Standardanwendung des Neuronalen Netzes, bei der die Transkription für das ganze Wort erstellt werden muss, ist die Aufgabe beim Auffüllen der Lücken einfacher, weil zumindest der linke Phonemkontext als sicher angenommen werden kann, da er ja aus dem Aussprachelexikon stammt. Die Eingabe der vorhergehenden Phoneme stabilisiert somit die Ausgabe des Neuronalen Netzes für die zu füllende Lücke, da das zu generierende Phonem nicht nur von den Buchstaben, sondern auch vom vorhergehenden Phonem abhängt.

[0023] Ein Problem beim Aneinanderhängen der Transkriptionen aus dem Lexikon sowie bei der Bestimmung der Transkription für die Lücken mittels eines Neuronalen Netzes besteht darin, daß in einigen Fällen der letzte Laut der vorhergehenden, linken Transkription verändert werden muss. Dies ist bei dem betrachteten Wort "überflüssigerweise" der Fall. Es wird im Lexikon als Ganzes nicht gefunden, dafür aber das Teilwort "überflüssig" und das Teilwort "erweise".

[0024] Im Folgenden werden Grapheme zur besseren Unterscheidung in spitzen Klammern <> eingeschlossen und Phoneme in eckigen Klammern [].

[0025] Die Endung <-ig> am Silbenende wird gesprochen wie [IC], dargestellt in der Lautschrift SAMPA, also wie [I] (ungespannter kurzer ungerundeter vorderer Vokal) gefolgt vom Ich-Laut [C] (stimmloser palataler Frikativ). Die Vorsilbe <er-> wird gesprochen wie [Er], mit einem [E] (ungespannter kurzer ungerundeter halboffener vorderer Vokal, offenes "e") und einem [r] (zentraler Sonorant).

[0026] Beim einfachen Verketten der Transkriptionen wird sinnvollerweise automatisch eine Silbengrenze zwischen den beiden Wörtern eingefügt, dargestellt durch einen Bindestrich "-". Es ergibt sich somit als Gesamttranskription des Worts <überflüssigerweise>

[y: - b6 - flY - sIC - Er - val - z@]

statt richtigerweise

[y: - b6 - flY - sI - g6 - val - z@]

mit einem [g] (stimmhafter velarer Plosiv) und einem [6] (nichtbetonter zentraler halboffener Vokal mit velarer Färbung) sowie einer verschobenen Silbengrenze. Somit wären an der Wortgrenze Laut und Silbengrenze falsch.

[0027] Abhilfe kann hier geschaffen werden, indem ein Neuronales Netz den letzten Laut der linken Transkription berechnet. Dabei stellt sich aber die Frage, welche Buchstaben am Ende der linken Transkription zur Bestimmung des letzten Lautes herangezogen werden sollen.

[0028] Für diese Entscheidung wird ein spezielles Aussprachelexikon benutzt. Die Besonderheit an diesem Lexikon besteht darin, daß es die Information enthält, welche Graphemgruppe zu welchem Laut gehört. Wie das Lexikon zu erstellen ist, ist in Horst-Udo Hain: "Automation of the Training Procedures for Neural Networks Performing Multi-Lingual Grapheme to Phoneme Conversion". Eurospeech 1999, S. 2087-2090, ausgeführt.

[0029] Der Eintrag für "überflüssig" hat in diesem Lexikon die Form

ü	-	b	er	-	f	l	ü	-	ss	i	g
y:	-	b	6	-	f	l	y	-	s	I	C

[0030] Damit kann eindeutig bestimmt werden, aus welcher Graphemgruppe der letzte Laut entstanden ist, nämlich aus dem <g>.

[0031] Das Neuronale Netz kann nun mit Hilfe des jetzt vorhandenen rechten Kontextes <erweise> neu über Phonem und Silbengrenze am Wortende entscheiden. Das Ergebnis ist in diesem Falle das Phonem [g], vor dem eine Silbengrenze gesetzt wird.

[0032] Jetzt ist die Silbengrenze an der richtigen Stelle und das <g> wird auch als [g] transkribiert und nicht als [C]. Der erste Laut der rechten Transkription wird nach dem gleichen Schema neu bestimmt. Die richtige Transkription für <er-> von <erweise> ist an dieser Stelle [6] und nicht [Er]. Hier sind gleich zwei Laute zu revidieren, weshalb im bevorzugten Ausführungsbeispiel stets zwei Laute revidiert werden.

[0033] Im Ergebnis erhält man die korrekte phonetische Transkription an dieser Schnittstelle.

[0034] Weitere Verbesserungen sind zu erzielen, wenn für das Ausfüllen der Transkriptionslücken nicht das Standardnetz verwendet wird, das zur Konvertierung ganzer Wörter trainiert wurde, sondern ein speziell zum Ausfüllen der Lücken trainiertes Netz. Zumindest in den Fällen, bei denen der rechte Phonemkontext auch vorhanden ist, bietet sich ein Spezialnetz an, das unter Verwendung des rechten Phonemkontextes über den zu generierenden Laut entscheidet.

Patentansprüche

1. Verfahren zur Graphem-Phonem-Konvertierung eines Wortes, das als Ganzes nicht in einem Aussprachelexikon enthalten ist, mit folgenden Schritten:

   a) das Wort wird in Teilwörter zerlegt,
   b) eine Graphem-Phonem-Konvertierung der Teilwörter wird durchgeführt,
   c) die durch die Konvertierung erhaltenen Transkriptionen der Teilwörter werden hintereinander aufgereiht, wobei sich mindestens eine Schnittstelle zwischen den Transkriptionen der Teilwörter ergibt,
   d) die an die mindestens eine Schnittstelle grenzenden Phoneme der Teilwörter werden bestimmt,
   e) es werden diejenigen Grapheme der Teilwörter bestimmt, die die an die mindestens eine Schnittstelle grenzenden Phoneme erzeugen,
   f) die Graphem-Phonem-Konvertierung der bestimmten Grapheme wird im Kontext der jeweiligen Schnittstelle neu berechnet.

2. Verfahren nach Anspruch 1,
   dadurch gekennzeichnet,
   dass die Graphem-Phonem-Konvertierung der bestimmten Grapheme im Kontext der jeweiligen Schnittstelle mittels eines Neuronalen Netzes neu berechnet werden.

3. Verfahren nach Anspruch 1,
dadurch gekennzeichnet,
dass die Graphem-Phonem-Konvertierung der bestimmten Grapheme im Kontext der jeweiligen Schnittstelle mittels eines Lexikons neu berechnet werden.

4. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
dass für das Wort in einer Datenbank, die phonetische Transkriptionen von Wörtern enthält, nach Teilwörtern des Worts gesucht wird; und- dass für ein in der Datenbank gefundenes Teilwort die in der Datenbank verzeichnete phonetische Transkription gewählt wird.

5. Verfahren nach Anspruch 4,
dadurch gekennzeichnet,
dass das Wort neben dem gefundenen Teilwort mindestens einen weiteren Bestandteil aufweist, der nicht in der Datenbank verzeichnet ist; und- dass dieser weitere Bestandteil mittels einer OOV-Behandlung phonetisch transkribiert wird.

6. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet, dass das Wort in Teilwörter einer gewissen Mindestlänge zerlegt wird.

7. Computerprogrammprodukt, das durch einen Computer ausführbar ist und dabei die Schritte nach einem der Ansprüche 1 bis 6 ausführt.

8. Computerprogrammprodukt, das auf einem computergeeigneten Medium gespeichert ist und computerlesbare Programmmittel umfaßt, die es einem Computer ermöglichen, das Verfahren nach einem der Ansprüche 1 bis 6 auszuführen.

9. Datenträger, auf dem ein Computerprogramm gespeichert ist, das es einem Computer ermöglicht, durch einen Ladeprozess das Verfahren nach einem der Ansprüche 1 bis 6 auszuführen.

10. Computersystem mit Mitteln zum Ausführen des Verfahrens nach einem der Ansprüche 1 bis 6.

11. Computersystem zur Graphem-Phonem-Konvertierung eines Worts, das als Ganzes nicht in einem Aussprachelexikon enthalten ist,

- einer Speichereinrichtung (22, 30) zum Speichern eines Computerprogramms auf einem Speichermedium;
- einer Verarbeitungseinheit (20) zum Laden des Computerprogramms aus der Speichereinrichtung und zum Ausführen des Computerprogramms;
- mit Mitteln zum Zerlegen des Worts in Teilwörter;
- mit Mitteln zum hintereinander Aufreihen der Transkriptionen der Teilwörter, wobei sich mindestens eine Schnittstelle zwischen den Transkriptionen der Teilwörter ergibt;
- mit Mitteln zum Bestimmen der an die mindestens eine Schnittstelle grenzenden Phoneme der Teilwörter;
- mit Mitteln zum Bestimmen derjenigen Grapheme der Teilwörter, die die an die mindestens eine Schnittstelle grenzenden Phoneme erzeugen;
- mit Mitteln zum erneuten Berechnen der Graphem-Phonem-Konvertierung der bestimmten Grapheme im Kontext der jeweiligen Schnittstelle; und
- mit Mitteln zum anschließenden Schreiben der an der Schnittstelle neu berechneten Phoneme in eine zweite Speichereinrichtung.

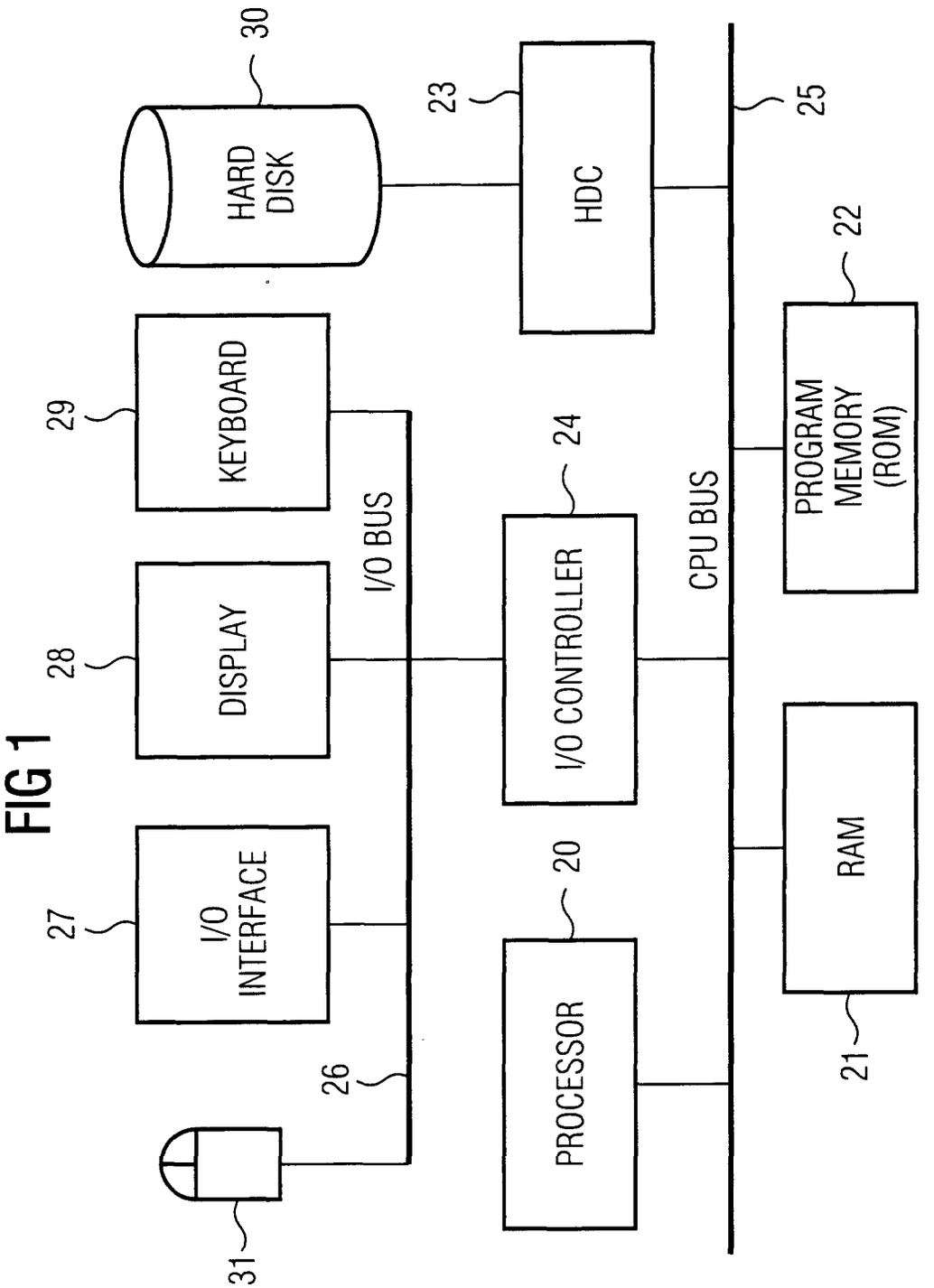
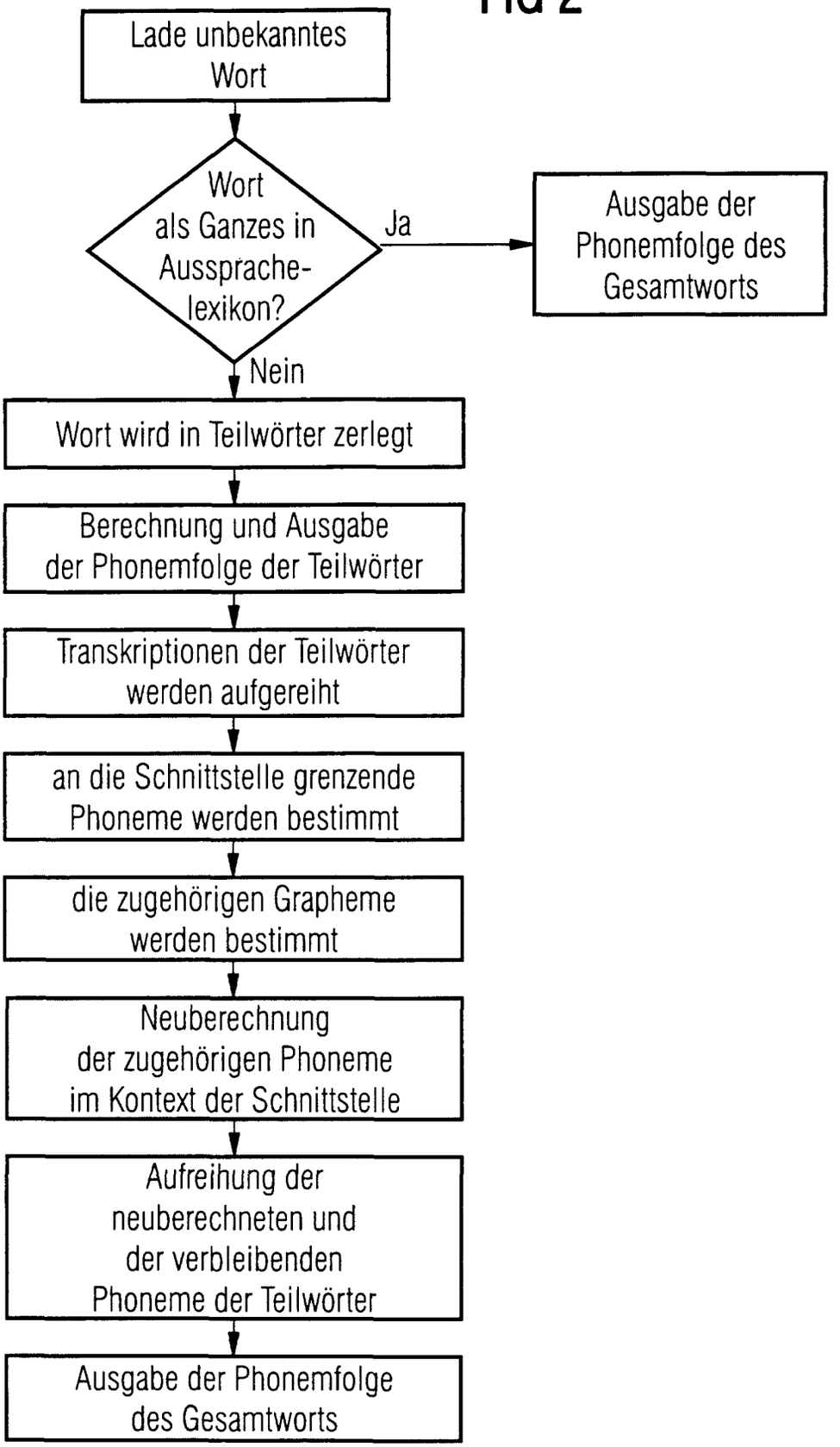

HARD DISK
30
HDC
23
25
KEYBOARD
29
PROGRAM MEMORY (ROM)
22
I/O BUS
DISPLAY
28
I/O CONTROLLER
24
CPU BUS
I/O INTERFACE
27
PROCESSOR
20
RAM
26
21
31

FIG 1

FIG 2


Lade unbekanntes Wort
Wort als Ganzes in Aussprache-lexikon?
Ja
Ausgabe der Phonemfolge des Gesamtworts
Nein
Wort wird in Teilwörter zerlegt
Berechnung und Ausgabe der Phonemfolge der Teilwörter
Transkriptionen der Teilwörter werden aufgereiht
an die Schnittstelle grenzende Phoneme werden bestimmt
die zugehörigen Grapheme werden bestimmt
Neuberechnung der zugehörigen Phoneme im Kontext der Schnittstelle
Aufreihung der neuberechneten und der verbleibenden Phoneme der Teilwörter
Ausgabe der Phonemfolge des Gesamtworts