(19) **Bundesrepublik Deutschland**
Deutsches Patent- und Markenamt

(10) **DE 600 05 326 T2** 2004.07.22

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 171 871 B1**
(21) Deutsches Aktenzeichen: **600 05 326.1**
(86) PCT-Aktenzeichen: **PCT/EP00/01965**
(96) Europäisches Aktenzeichen: **00 909 331.1**
(87) PCT-Veröffentlichungs-Nr.: **WO 00/58945**
(86) PCT-Anmeldetag: **07.03.2000**
(87) Veröffentlichungstag der PCT-Anmeldung: **05.10.2000**
(97) Erstveröffentlichung durch das EPA: **16.01.2002**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **17.09.2003**
(47) Veröffentlichungstag im Patentblatt: **22.07.2004**

(51) Int Cl.[7]: **G10L 15/26**

(30) Unionspriorität:
99200949 26.03.1999 EP

(84) Benannte Vertragsstaaten:
AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LI, LU, MC, NL, PT, SE

(73) Patentinhaber:
Philips Intellectual Property & Standards GmbH, 20099 Hamburg, DE; Koninklijke Philips Electronics N.V., Eindhoven, NL

(74) Vertreter:
Meyer, M., Dipl.-Ing., Pat.-Ass., 52076 Aachen

(72) Erfinder:
THELEN, Eric, NL-5656 AA Eindhoven, NL;
BESLING, Stefan, NL-5656 AA Eindhoven, NL;
ULLRICH, Meinhard, NL-5656 AA Eindhoven, NL

(54) Bezeichnung: **ERKENNUNGSEINHEITEN MIT KOMPLEMENTÄREN SPRACHMODELLEN**

Beschreibung

[0001] Die vorliegende Erfindung bezieht sich auf ein Spracherkennungssystem mit großem Vokabular zur Erkennung einer Sequenz von gesprochenen Worten, wobei das System Eingabemittel umfasst, um ein zeitsequentielles, die Sequenz der gesprochenen Worte darstellendes Eingabemuster zu empfangen, und einen Spracherkenner mit großem Vokabular, der unter Verwendung eines dem Spracherkenner zugeordneten Erkennungsmodells mit großem Vokabular das eingegebene Muster als eine Sequenz von Worten erkennt.

[0002] Aus der US-amerikanischen Patentschrift 5.819.220 ist ein System zur Erkennung von Sprache in einer Internetumgebung bekannt. Das System ist insbesondere auf den Zugriff auf Informationsressourcen im World Wide Web (WWW) mittels Sprache ausgerichtet. Die mit dem Aufbau eines Spracherkennungssystems als einer Schnittstelle zum Web verbundenen Probleme unterscheiden sich sehr von denen, die in den Bereichen herkömmlicher Spracherkennung auftreten. Hauptproblem ist das große Vokabular, das das System unterstützen muss, da ein Benutzer auf praktisch jedes beliebige Dokument zu jedem beliebigen Thema zugreifen kann. Ein geeignetes Erkennungsmodell, wie beispielsweise ein Sprachmodell für derartig großen Vokabulare aufzubauen, ist äußerst schwierig, wenn nicht unmöglich. In dem bekannten System wird ein vorgegebenes Erkennungsmodell, einschließlich eines statistischen N-gramm-Sprachmodells und eines akustischen Modells, verwendet. Das Erkennungsmodell wird dynamisch unter Verwendung eines vom Web ausgelösten Wortschatzes verändert. Ein HTML-Dokument (HyperText Mark-up Language) enthält Verknüpfungen, wie beispielsweise Hypertext-Links, die dazu verwendet werden, einen Wortschatz zu identifizieren, der in dem endgültigen Wortschatz enthalten sein soll, um die Wahrscheinlichkeit bei der Worterkennungssuche zu erhöhen. Auf diese Weise wird der für die Berechnung der Spracherkennungsresultate verwendete Wortschatz durch die Einbeziehung des durch das Web ausgelösten Wortschatzes im vorhinein beeinflusst (Biasing).

[0003] Das bekannte System erfordert ein geeignetes, großes Vokabularmodell als ein Ausgangsmodell, um nach der Adaption ein vorbeeinflusstes Modell erhalten zu können. Tatsächlich kann das vorbeeinflusste Modell als ein konventionelles Modell mit großem Vokabular gesehen werden, das für den aktuellen Erkennungskontext optimiert wurde. Wie bereits angemerkt, ist es sehr schwierig, ein geeignetes Modell mit großem Vokabular aufzubauen, auch wenn es nur als ein Ausgangsmodell benutzt wird. Ein weiteres Problem tritt bei gewissen Erkennungsaufgaben auf, zum Beispiel bei der Erkennung der Eingabe für bestimmte Websites oder HTML-Dokumente, wie man sie beispielsweise bei Suchmaschinen oder großen elektronischen Einkaufsläden wie Buchhandlungen vorfindet. In solchen Fällen ist die Anzahl der Wörter, die man äußern kann, riesig. Ein konventionelles Modell mit großem Vokabular wird im Allgemeinen nicht in der Lage sein, die gesamte Bandbreite möglicher Wörter effektiv abzudecken. Bei einem Ausgangsmodell mit relativ wenigen Wörtern wird die Vorbeeinflussung nicht zu einem guten Erkennungsmodell führen. Wenn das Ausgangsmodell bereits annehmbar gut wäre, würde eine geeignete Vorbeeinflussung einen großen zusätzlichen Wortschatz und einen signifikanten Verarbeitungsaufwand erfordern.

[0004] Die vorliegende Erfindung hat zur Aufgabe, ein Erkennungssystem zu schaffen, das große Vokabulare besser handhaben kann.

[0005] Zur Realisierung dieser Aufgabe ist das System dadurch gekennzeichnet, dass es eine Vielzahl von N Spracherkennern mit großem Vokabular umfasst, die jeweils zu einem entsprechenden, unterschiedlichen Erkennungsmodell mit großem Vokabular gehören, wobei jedes der Erkennungsmodelle auf einen bestimmten Teil des großen Vokabulars ausgerichtet ist, und wobei das System einen Controller umfasst, der das Eingabemuster einer Vielzahl der Spracherkenner zuführt und aus den von der Vielzahl von Spracherkennern erkannten Wortsequenzen eine erkannte Wortsequenz auswählt.

[0006] Durch die Verwendung mehrerer Spracherkenner, jeweils mit einem spezifischen, auf einen Teil des großen Vokabulars ausgerichteten Erkennungsmodell, wird die Aufgabe, ein Erkennungsmodell für ein großes Vokabular aufzubauen, in die zu bewältigende Aufgabe zerlegt, Modelle mit großem Vokabular für spezifische Kontexte zu erstellen. Zu solchen Kontexten können Gesundheit, Unterhaltung, Computer, Kunst, Geschäft, Bildung, Regierung, Wissenschaft, Nachrichten, Reisen usw. gehören. Hervorzuheben ist, dass sich normalerweise alle derartigen Kontexte im Vokabular überlappen, zum Beispiel in den allgemeinen Wörtern der Sprache. Die Kontexte werden sich sowohl hinsichtlich der Statistiken über solche allgemeinen Wörter als auch hinsichtlich der spezifischen Fachsprache solcher Kontexte unterscheiden. Indem man mehrere dieser Modelle zur Erkennung der Eingabe benutzt, kann ein größerer Bereich von Ausdrücken mittels sorgfältig trainierter Modelle erkannt werden. Ein weiterer Vorteil bei der Verwendung mehrerer Modelle liegt darin, dass dies eine bessere Unterscheidung während der Erkennung erlaubt. Wenn man ein großes Vokabular benutzte, wurden bestimmte Ausdrücke nur in einer bestimmten Bedeutung (und Schreibweise) erkannt. Wenn ein Benutzer zum Beispiel ein Wort ausspricht, das wie "color" klingt, werden die meisten erkannten Wortsequenzen das sehr allgemeine Wort "color" beinhalten. Weniger wahrscheinlich wird sein, dass das Wort "collar" (aus einem Modekontext) erkannt wird, oder "collar" von "collared hering" (Nahrungsmittelkontext), oder "collar-bone" (Gesundheitskontext). Solche spezifischen Wörter haben keine große

Chance in einem großen Vokabular erkannt zu werden, das unvermeidlich von häufig auftauchenden Wortsequenzen aus allgemeinen Wörtern dominiert wird. Bei Verwendung mehrerer Modelle wird jedes Modell eine oder mehrere in Frage kommende Wortsequenzen identifizieren, unter denen eine Auswahl getroffen werden kann. Selbst wenn in dieser abschließenden Auswahl eine Wortsequenz wie "color" ausgewählt wird, können dem Benutzer die alternativen Wortsequenzen mit "collar" präsentiert werden.

[0007] Vorzugsweise arbeiten die Spracherkenner parallel in dem Sinne, dass der Benutzer keine signifikante Verzögerung bei der Erkennung bemerkt. Dies kann durch Verwendung separater Spracherkennungsmaschinen (Recognition Engines) mit jeweils eigenen Verarbeitungsressourcen erreicht werden. Alternativ lässt sich dies durch Verwendung eines ausreichend leistungsfähigen Serienprozessors erzielen, der die Erkennungsaufgaben "parallel" mit Hilfe konventioneller Zeitscheibenverfahren verarbeitet.

[0008] Es sollte darauf hingewiesen werden, dass die Verwendung paralleler Spracherkennungsmaschinen bekannt ist. In der US-amerikanischen Patentschrift 5.754.978 wird die parallele Verwendung von Spracherkennungsmaschinen beschrieben. Alle Maschinen haben eine relativ hohe Genauigkeit von beispielsweise 95%. Die Genauigkeit lässt sich verbessern, wenn sich die 5%-Ungenauigkeit der Maschinen nicht überlappt. Um sicherzustellen, dass sich die Ungenauigkeiten nicht vollständig überlappen, können die Maschinen unterschiedlich sein. Alternativ können die Maschinen identisch sein, in welchem Fall das Eingangssignal für eine der Maschinen etwas gestört wird oder eine der Maschinen etwas gestört wird. Ein Komparator vergleicht den erkannten Text und akzeptiert ihn basierend auf dem Grad der Übereinstimmung zwischen der Ausgabe der Maschinen oder weist ihn zurück. Da das System präzise Spracherkennungsmaschinen erfordert, die es für große Vokabulare nicht gibt, bietet dieses System keine Lösung für die Spracherkennung mit großen Vokabularen. Noch verwendet das System verschiedene Modelle, die auf spezifische Bereiche eines großen Vokabulars ausgerichtet sind.

[0009] In der Patentschrift WO 98/10413 wird ein Dialogsystem mit einer optionalen Anzahl von Spracherkennungsmodulen beschrieben, die parallel arbeiten können. Die Module sind auf eine spezifische An der Spracherkennung ausgerichtet, wie beispielsweise die Erkennung einzelner Ziffern, die Erkennung fortlaufender Nummern, die Worterkennung mit kleinem Vokabular, die isolierte Erkennung mit großem Vokabular, die fortlaufende Worterkennung, die Schlüsselworterkennung, die Wortsequenzerkennung, die Alphabeterkennung usw. Das Dialogsystem weiß im Voraus, welche An von Eingabe der Benutzer liefern wird, und aktiviert dementsprechend ein oder mehrere der spezifischen Module. Wenn der Benutzer beispielsweise eine Nummer sprechen muss, aktiviert die Dialogmaschine die Erkennung einzelner Ziffern und die Erkennung fortlaufender Nummern, so dass der Benutzer die Nummer als Ziffern oder als fortlaufende Nummer sprechen kann. Das System bietet keine Lösung für die Handhabung großer Vokabulare.

[0010] Die Spracherkennungsmodelle des erfindungsgemäßen Systems können vorgegeben sein. Vorzugsweise wird, wie im abhängigen Anspruch 2 definiert, ein Modellselektor verwendet, um mindestens eines der aktiv für die Spracherkennung benutzten Modelle dynamisch auszuwählen. Die Auswahl ist vom Kontext der Benutzereingabe abhängig, wie beispielsweise dem Abfrage- oder Diktatthema. Vorzugsweise wählt der Modellselektor viele der Spracherkennungsmodule aus. In der Praxis wird mindestens eines der Modelle das normale Alltagsvokabular über allgemeine Themen repräsentieren. Ein solches Modell wird in der Regel immer benutzt.

[0011] In einer Ausführungsform gemäß dem abhängigen Anspruch 3 definiert das Dokument den Spracherkennungskontext. Wie im abhängigen Anspruch 5 definiert, kann dies dadurch erfolgen, dass man die im Dokument vorhandenen Wörter scannt und anschließend bestimmt, welches oder welche der Spracherkennungsmodelle am besten zur Erkennung dieser Wörter geeignet ist/sind (z. B. diejenigen Modelle, die die meisten Wörter oder Wortsequenzen mit dem Dokument gemeinsam haben).

[0012] In einer Ausführungsform gemäß dem abhängigen Anspruch 4 wird der Kontext (bzw. die Kontexte) in einer Web-Seite angegeben, zum Beispiel unter Verwendung einer eingebetteten Markierung (englisch „tag"), die den Kontext identifiziert. Die Seite kann auch den Kontext (oder den Kontextidentifizierer) angeben, beispielsweise über einen Link.

[0013] In einer Ausführungsform, wie sie im abhängigen Anspruch 6 definiert ist, versucht das System aktiv diejenigen Spracherkennungsmodelle zu identifizieren, die für die aktuelle Erkennungsaufgabe geeignet sind. Zusätzlich zu den Spracherkennungsmodellen, die momentan aktiv für die Spracherkennung benutzt werden, werden die anderen Modelle auf ihre Eignung überprüft. Diese Überprüfung kann als Hintergrundaufgabe durch Verwendung einer oder mehrerer zusätzlicher Spracherkenner erfolgen, die prüfen, ob mit den nicht benutzten Modellen ein besseres Ergebnis als mit den aktiv benutzten Modellen erzielt worden wäre. Alternativ können die aktuellen Spracherkenner verwendet werden, um die Testmodelle in Momenten zu überprüfen, in denen der Spracherkenner über ausreichende Leistungsreserven verfügt, beispielsweise, wenn der Benutzer nicht spricht. Die Überprüfung kann die gesamte Eingabe des Benutzers umfassen. Insbesondere, wenn bereits eine umfangreiche Spracheingabe durch den Benutzer vorliegt, wird die Überprüfung vorzugsweise auf die jüngste Eingabe begrenzt. Auf diese Weise können, wann immer der Benutzer das Thema schnell wechselt, besser geeignete Modelle gewählt

werden. Ein Kriterium um festzulegen, welche Modelle am besten geeignet sind, das heißt die höchste Genauigkeit bei der Spracherkennung bieten, basiert vorzugsweise auf Leistungshinweisen bei der Erkennung, wie zum Beispiel Resultaten oder Vertrauenswerten.

[0014] In einer Ausführungsform, wie sie im abhängigen Anspruch 7 definiert ist, sind die Erkennungsmodelle hierarchisch angeordnet. Dies vereinfacht die Auswahl geeigneter Modelle. Vorzugsweise beginnt die Erkennung mit einer Anzahl relativ generischer (allgemeiner) Modelle. Wenn sich herausstellt, dass ein bestimmtes generisches Modell ein gutes Erkennungsresultat liefert, können speziellere Modelle getestet werden, um die Erkennung noch weiter zu verbessern. Manche der spezielleren Modelle können mit mehreren allgemeineren Modellen gemeinsam genutzt werden. Falls die Erkennungsresultate eines spezielleren Modells zu einem bestimmten Zeitpunkt schlechter werden, können mehrere der allgemeineren Modelle ausprobiert werden, die in der Hierarchie über dem speziellen Modell angeordnet sind. Dies gestattet einen lückenlosen Übergang von einem Kontext zum anderen. Ein Benutzer beginnt beispielsweise damit, einen generischen Kontext über Gesundheit einzugeben. In einem bestimmten Moment wird möglicherweise festgestellt, dass sich der Benutzer hauptsächlich auf den spezielleren Kontext von medizinischen Zentren oder Instituten konzentriert, und sogar bis hin zu dem sehr speziellen Kontext von Gesundheitsfarmen. Insbesondere, wenn sich die Gesundheitsfarm in einer attraktiven Gegend befindet, kann dies den Benutzer dazu veranlassen, zu dem allgemeineren Kontext von Ferien oder Reisen, oder genauer das Reisen in der Gegend der Gesundheitsfarm, überzugehen.

[0015] Wie im abhängigen Anspruch 8 definiert, kann die Spracherkennung durch einen separaten Erkennungsserver erfolgen. Im Kontext des Internets könnte ein solcher Server eine separate Station im Internet sein, oder in vorhandene Stationen, wie beispielsweise eine Suchmaschine, integriert sein, oder ein Diensteanbieter sein, wie beispielsweise eine elektronische Buchhandlung. Insbesondere Erkennungsserver, die für viele Benutzer arbeiten, müssen ein Vokabular unterstützen, das für die meisten Benutzer geeignet ist. Die Verwendung mehrerer Modelle mit großem Vokabular sorgt dafür, dass ein solches System besser geeignet ist, diese Aufgabe mit einer hohen Erkennungsgenauigkeit durchzuführen.

[0016] Diese und andere Aspekte der Erfindung ergeben sich aus den nachfolgend beschriebenen Ausführungsformen und werden durch diese sowie die begleitenden Zeichnungen näher erläutert.

[0017] **Fig.** 1 zeigt die Struktur einer Spracherkennung mit großem Vokabular;

[0018] **Fig.** 2 veranschaulicht ein vollständiges Wortmodell;

[0019] **Fig.** 3 zeigt das Blockdiagramm eines erfindungsgemäßen Systems;

[0020] **Fig.** 4 zeigt eine Hierarchie von Erkennungsmodellen; und

[0021] **Fig.** 5 zeigt das Blockdiagramm eines erfindungsgemäßen verteilten Systems.

[0022] Spracherkennungssysteme, wie beispielsweise kontinuierliche Spracherkennungssysteme mit großem Vokabular, verwenden in der Regel eine Reihe von Erkennungsmodellen, um ein Eingabemuster zu erkennen. Beispielsweise können ein Akustikmodell und ein Vokabular verwendet werden, um Wörter zu erkennen, und ein Sprachmodell kann benutzt werden, um die grundlegenden Erkennungsresultate zu verbessern. **Fig.** 1 veranschaulicht die typische Struktur eines kontinuierlichen Spracherkennungssystems mit großem Vokabular **100** [siehe L. Rabiner, B-H. Juang, "Fundamentals of speech recognition", Prentice Hall 1993, Seiten 434–454]. Das System **100** umfasst ein Spektralanalyse-Teilsystem **110** und ein Einheitenabgleich-Teilsystem **120**. Im Spektralanalyse-Teilsystem **110** wird das Spracheingabesignal (speech input signal, SIS) spektral und/oder temporär analysiert, um einen repräsentativen Merkmalsvektor (Beobachtungsvektor, OV) zu berechnen. In der Regel wird das Sprachsignal digitalisiert (z. B. mit einer Rate von 6,67 kHz abgetastet) und vorverarbeitet, beispielsweise indem eine Preemphasis angewandt wird. Nachfolgende Abtastungen werden in Rahmen gruppenweise (blockweise) zusammengefasst, die beispielsweise 32 ms des Sprachsignals entsprechen. Aufeinander folgende Rahmen überlappen sich teilweise, zum Beispiel um 16 ms. Als Spektralanalyseverfahren wird häufig Linear Predictive Coding (LPC) eingesetzt, um für jeden Rahmen einen repräsentativen Merkmalsvektor (Beobachtungsvektor) zu berechnen. Der Merkmalsvektor kann beispielsweise 24, 32 oder 63 Komponenten haben. Die standardmäßige Vorgehensweise bei kontinuierlichen Spracherkennungssystemen mit großem Vokabular besteht in der Annahme eines Wahrscheinlichkeitsmodells der Sprachproduktion, wobei eine spezifische Wortsequenz $W = w_1w_2w_3 ... w_q$ eine Sequenz von akustischen Beobachtungsvektoren $Y = y_1y_2y_3 ... y_T$ erzeugt. Der Erkennungsfehler lässt sich statistisch minimieren, indem man die Sequenz der Wörter $w_1w_2w_3 ... w_q$ bestimmt, die aller Wahrscheinlichkeit nach die beobachtete Sequenz von Beobachtungsvektoren $y_1y_2y_3 ... y_T$ (über die Zeit t = 1, ..., T) verursacht haben, wobei die Beobachtungsvektoren das Resultat des Spektralanalyse-Teilsystems **110** sind. Dies führt zur Bestimmung der maximalen a-posteriori-Wahrscheinlichkeit:

max. P(W|Y), für alle möglichen Wortsequenzen W

[0023] Indem man Bayes' Theorem über bedingte Wahrscheinlichkeiten anwendet, ist P(W|Y) gegeben durch:

$$P(W|Y) = P(Y|W)\cdot P(W)/P(Y)$$

[0024] Da P(Y) unabhängig von W ist, ist die wahrscheinlichste Wortsequenz gegeben durch:

arg max. P(W|Y)·P(W) für alle möglichen Wortsequenzen W (1).

[0025] Im Einheitenabgleich-Teilsystem **120** liefert ein Akustikmodell den ersten Term der Gleichung (1). Das Akustikmodell wird benutzt, um die Wahrscheinlichkeit P(W|Y) einer Sequenz von Beobachtungsvektoren Y für eine gegebene Wortkette W zu schätzen. Bei einem System mit großem Vokabular geschieht dies in der Regel dadurch, dass die Beobachtungsvektoren mit einem Verzeichnis von Spracherkennungseinheiten abgeglichen werden. Eine Spracherkennungseinheit wird durch eine Sequenz aus akustischen Referenzen repräsentiert. Es können verschiedene Formen von Spracherkennungseinheiten verwendet werden. Beispielsweise kann ein ganzes Wort oder auch eine Gruppe von Wörtern durch eine einzige Spracherkennungseinheit repräsentiert werden. Ein Wortmodell (WM) bietet für jedes Wort eines gegebenen Vokabulars eine Transkription in Form einer Sequenz akustischer Referenzen. In Systemen, bei denen ein ganzes Wort durch eine Spracherkennungseinheit repräsentiert wird, besteht eine direkte Beziehung zwischen dem Wortmodell und der Spracherkennungseinheit. Andere Systeme, insbesondere Systeme mit großem Vokabular, können für die Spracherkennungseinheit sowohl linguistisch basierte Subworteinheiten, wie beispielsweise Phone, Diphone oder Silben, als auch Ableitungseinheiten, wie beispielsweise Fenene und Fenone, verwenden. Bei solchen Systemen wird ein Wortmodell durch ein Lexikon **134**, das die zu einem Wort des Vokabulars gehörende Sequenz der Subworteinheiten beschreibt, und die Subwortmodelle **132**, die die Sequenzen der akustischen Referenzen der betroffenen Spracherkennungseinheit beschreiben, vorgegeben. Ein Wortmodell-Zusammensetzer **136** setzt das Wortmodell basierend auf dem Subwortmodell **132** und dem Lexikon **134** zusammen. **Fig.** 2 veranschaulicht ein Wortmodell **220** für ein auf Subworteinheiten basierendes System, wobei das gezeigte Wort durch eine Sequenz aus drei Subwortmodellen (**250**, **260** und **270**) jeweils mit einer Sequenz aus vier akustischen Referenzen (**251**, **252**, **253**, **254**; **261** bis **264**; **271** bis **274**) gebildet wird. Das in **Fig.** 2 gezeigte Wortmodell basiert auf den Hidden-Markov-Modellen (HMMs), die häufig bei der stochastischen Bildung von Sprachsignalen eingesetzt werden. Bei diesem Modell ist jede Erkennungseinheit (Wortmodell oder Subwortmodell) typischerweise durch ein HMM gekennzeichnet, dessen Parameter anhand eines Trainingsdatensatzes geschätzt werden. Bei Spracherkennungssystemen mit großem Vokabular wird normalerweise ein begrenzter Satz von beispielsweise 40 Subworteinheiten verwendet, da sehr viele Trainingsdaten erforderlich wären, um ein HMM für größere Einheiten zu trainieren. Ein HMM-Status entspricht einer akustischen Referenz. Für die Bildung einer Referenz sind verschiedene Verfahren bekannt, einschließlich diskreter oder kontinuierlicher Wahrscheinlichkeitsdichten. Jede Sequenz der akustischen Referenzen, die zu einer bestimmten Äußerung gehört, wird auch als eine akustische Transkription der Äußerung bezeichnet. Hervorzuheben ist, dass bei Verwendung anderer Erkennungsverfahren als HMMs Details der akustischen Transkription unterschiedlich sein werden.

[0026] Ein Abgleichsystem auf Wortebene **130** in **Fig.** 1 gleicht die Beobachtungsvektoren mit allen Sequenzen der Spracherkennungseinheiten ab und liefert die Wahrscheinlichkeit einer Übereinstimmung zwischen dem Vektor und einer Sequenz. Bei Verwendung von Subworteinheiten können den Übereinstimmungen unter Verwendung eines Lexikons **134** Beschränkungen auferlegt werden, um die möglichen Sequenzen von Subworteinheiten auf Sequenzen im Lexikon **134** zu begrenzen. Dies reduziert das Resultat auf mögliche Sequenzen von Wörtern.

[0027] Für eine vollständige Erkennung wird vorzugsweise auch ein Abgleichsystem auf Satzebene **140** verwendet, das, basierend auf einem Sprachmodell (LM), weitere Beschränkungen für die Übereinstimmung auferlegt, so dass die untersuchten Pfade definitiv zu Wortsequenzen gehören, die vom Sprachmodell als geeignete Sequenzen spezifiziert wurden. Das Sprachmodell an sich liefert den zweiten Term P(W) der Gleichung (1). Indem man die Resultate des akustischen Modells mit dem Sprachmodell kombiniert, erhält man ein Ergebnis des Einheitenabgleich-Teilsystems **120**, das ein erkannter Satz (RS) **152** ist. Das bei der Mustererkennung verwendete Sprachmodell kann syntaktische und/oder semantische Beschränkungen **142** der Sprache und der Erkennungsaufgabe enthalten. Ein auf syntaktischen Beschränkungen basierendes Sprachmodell wird üblicherweise als eine Grammatik **144** bezeichnet. Die vom Sprachmodell benutzte Grammatik **144** liefert die Wahrscheinlichkeit einer Wortsequenz W = $w_1w_2w_3 ... w_q$, die im Prinzip gegeben ist durch:

$$P(W) = P(w_1)P(w_2|w_1)\cdot P(w_1)P(w_3|w_1w_2) ...$$
$$P(w_q|w_1w_2w_3 ... w_q).$$

[0028] Da es in der Praxis unmöglich ist, die bedingten Wortwahrscheinlichkeiten für alle Wörter und alle Sequenzlängen in einer gegebenen Sprache zuverlässig zu schätzen, werden häufig n-gramm-Wortmodelle verwendet. Bei einem n-gramm-Wortmodell nähert man sich dem Term $P(w_j|w_1w_2w_3 ... w_{j-1})$ durch $P(w_j|w_{j-N+1} ... w_{j-1})$ an. In der Praxis werden Bigramme oder Trigramme verwendet. In einem Trigramm nähert man sich dem Term $P(w_j|w_1w_2w_3 ... w_{j-1})$ durch $P(w_j|w_{j-2}w_{j-1})$ an.

[0029] **Fig.** 3 zeigt das Blockdiagramm eines Spracherkennungssystems **300** gemäß der Erfindung. Beispiele für die Arbeitsweise des Systems werden insbesondere für eine Anwendung beschrieben, bei der erkannte Sprache in eine textuale oder ähnliche Darstellung konvertiert wird. Eine solche textuale Darstellung kann für Diktatzwecke verwendet werden, bei der die Textdarstellung in ein Dokument, z.

B. in einem Textverarbeitungsprogramm, oder in ein Textfeld, z. B. zur Spezifizierung eines Datenbankfelds, eingegeben wird. Für das Diktieren unterstützen Spracherkenner mit großem Vokabular ein aktives Vokabular und Lexikon mit bis zu 60.000 Wörtern. Es ist schwierig, genügend relevante Daten zu erhalten, um Modelle aufzubauen, die eine ausreichend genaue Erkennung für eine viel größere Anzahl von Wörtern bieten. In der Regel kann der Benutzer eine begrenzte Anzahl von Wörtern zum aktiven Vokabular/Lexikon hinzufügen. Diese Wörter können von einem Hintergrundvokabular mit 300.000 bis 500.00 Wörtern (das auch eine akustische Transkription der Wörter enthält) abgerufen werden. Für das Diktieren oder ähnliche Zwecke umfasst ein großes Vokabular beispielsweise mindestens 100.000 aktive Wörter oder sogar mehr als 300.000 aktive Wörter. Hervorzuheben ist, dass insbesondere für eine Internetumgebung, in der durch einen Klick auf einen Link ein vollständig unterschiedlicher Kontext geschaffen werden kann, vorzugsweise zahlreiche der Wörter des Hintergrundvokabulars aktiv erkannt werden können. Für andere Erkennungsaufgaben, wie beispielsweise das Erkennen von Namen, die normalerweise aus einer einfachen, mit einer gewissen Form von früherer Namenswahrscheinlichkeit verbundenen Liste bestehen, für die jedoch kein hochwertiges Sprachmodell existiert, kann ein Vokabular mit mehr als 50.000 Wörtern bereits als groß eingestuft werden.

[0030] Es wird verständlich sein, dass das Erkennungsresultat nicht für Diktatzwecke verwendet werden muss. Es kann genauso gut als Eingabe für andere Systeme, wie zum Beispiel Dialogsysteme, verwendet werden, bei denen, je nach erkannter Sprache, Informationen aus einer Datenbank abgerufen werden oder eine Operation durchgeführt wird, wie beispielsweise eine Buchbestellung oder eine Reisereservierung.

[0031] In **Fig.** 3 wird ein Einzelsystem **300** gezeigt, das vorzugsweise auf einem Computer, wie beispielsweise einem PC, realisiert wird. Das Element **310** stellt eine Anschlussverbindung für den Empfang eines sprachrepräsentativen Signals von einem Benutzer dar. Zum Beispiel kann ein Mikrofon an der Verbindung **310** angeschlossen werden. Es ist zu beachten, dass das Sprachsignal auch bereits aufgezeichnet sein oder von einer entfernten Stelle, z. B. über ein Telefon oder ein Netzwerk, abgerufen werden kann. Das System **300** umfasst eine Schnittstelle **320**, um die Eingabe des Benutzers zu empfangen. Dies kann beispielsweise mit Hilfe einer konventionellen Soundkarte realisiert werden. Falls die Schnittstelle einen Eingang für den Empfang von Sprache in analoger Form hat, umfasst sie vorzugsweise einen A/D-Umsetzer, um die analoge Sprache in digitale Abtastwerte mit einem für die Weiterverarbeitung durch ein Spracherkennungssystem **330** geeigneten Format zu konvertieren. Wenn die Schnittstelle einen Eingang für den Empfang der Sprache in einem digitalen Format hat, ist der Umsetzer vorzugsweise in der Lage, das digitale Signal in ein für die Weiterverarbeitung geeignetes digitales Format zu konvertieren. Das Spracherkennungssystem **330** analysiert in der Regel das Eingangssignal, wie beispielsweise für das Spektralanalyse-Teilsystem **110** aus **Fig.** 1 beschrieben. Gemäß der Erfindung umfasst das Spracherkennungssystem **330** eine Vielzahl von Spracherkennern mit großem Vokabular, die jeweils einem entsprechenden, unterschiedlichen Erkennungsmodell mit großem Vokabular zugeordnet sind. Bei einer typischen Erkennung, wie sie in **Fig.** 1 dargestellt ist, können die einzelnen Spracherkenner das modellunabhängige Spektralanalyse-Teilsystem **110** aus **Fig.** 1 mit anderen teilen, wie in **Fig.** 3 unter dem Bezugszeichen **335** gezeigt. **Fig.** 3 veranschaulicht die Verwendung von drei separaten Spracherkennenn **331**, **332** und **333**. Die Spracherkenner können denselben Algorithmus verwenden, wobei die Unterschiede in den jeweils benutzten Modellen liegen, wie beispielsweise dem Vokabular und dem Sprachmodell. Die Spracherkennung ist vorzugsweise sprecherunabhängig und gestattet eine kontinuierliche Spracheingabe. Die Spracherkennung an sich ist bekannt und wurde in verschiedenen Dokumenten beschrieben, zum Beispiel in der Patentschrift EP 92202782.6, entsprechend US-Seriennummer 08/425.304 (PHD 91.136), in der Patentschrift EP 92202783.4, entsprechend US-Seriennummer 08/751.377 (PHD 91.138), und in der Patentschrift EP 94200475.5, entsprechend US-Patent 5.634.083 (PHD 93.034), alle von dem Anmelder der vorliegenden Erfindung. Die Spracherkenner arbeiten "parallel" in dem Sinne, dass sie unabhängig voneinander dieselbe Spracheingabe im beinahe selben Moment erkennen. Dies lässt sich realisieren, indem für jeden der Spracherkenner separate Ressourcen verwendet werden, wie beispielsweise ein separater Prozessor oder eine separate Verarbeitungseinheit in einem "parallel" arbeitenden Prozessor, z. B. einem VLIW-Prozessor. Eine ähnlich "parallele" Durchführung kann auch bei einem herkömmlichen sequentiellen Prozessor mit ausreichend großer Leistungsfähigkeit erreicht werden, wobei jede Spracherkennung als eine separate Aufgabe ausgeführt wird. Vorzugsweise erfolgt die Spracherkennung in "Echtzeit", in dem Sinne, dass bei der Erkennung eines Wortes keine signifikante Verzögerung auftritt, nachdem das Wort vom System empfangen wurde.

[0032] Erfindungsgemäß ist jeder der Spracherkenner mit großem Vokabular einem entsprechenden, unterschiedlichen Erkennungsmodell mit großem Vokabular zugeordnet, wobei jedes der Erkennungsmodelle auf einen spezifischen Teil des großen Vokabulars ausgerichtet ist. Die Modelle werden vorzugsweise aus einem Speicher **340** geladen. Für die Beschreibung hier wird unter einem Spracherkennungsmodell ein kohärenter Satz von Modellen verstanden, der für eine einzelne Erkennungsaufgabe verwendet wird. Bezug nehmend auf **Fig.** 1 besteht das Erken-

nungsmodell beispielsweise aus einem Wortmodell (Lexikon **134** und Subwortmodell **132**) und einem Sprachmodell (Grammatik **144** und semantische Beschränkungen **142**) für einen bestimmten Teil des großen Vokabulars. Natürlich kann und wird zwischen den verschiedenen Erkennungsmodellen im Normalfall eine Überlappung bestehen. Üblicherweise wird eine derartige Überlappung im Teil des Vokabulars auftreten. Das Sprachmodell kann außerdem teilweise oder sogar vollständig dasselbe sein. In einem einfachen System entspricht die Anzahl der Erkennungsmodelle der Anzahl der Spracherkenner; jeder Spracherkenner ist dabei in einer festen Eins-zu-Eins-Beziehung einem exklusiven Erkennungsmodell zugeordnet. Vorzugsweise umfasst das System mehr Modelle als aktive Spracherkenner, wie unten noch ausführlich beschrieben werden wird. Die Figur zeigt acht Modelle **341** bis **348**.

[0033] Die Ausgabe der Spracherkenner wird an einen Controller **350** geleitet, um die abschließende Auswahl einer erkannten Wortsequenz zu treffen. Die einzelnen Spracherkenner **331** bis **333** produzieren eventuell nur eine erkannte Wortsequenz. Alternativ können auch mehrere Sequenzen (z. B. dargestellt durch einen Wortgraph) produziert werden. Vorzugsweise beinhaltet das Resultat der einzelnen Spracherkenner Informationen, wie beispielsweise Mutmaßlichkeit oder Vertrauenswerte, die es dem Controller **350** ermöglichen, die wahrscheinlichste Wortsequenz auszuwählen. Der Controller **350** ist zudem dafür verantwortlich, die Spracheingabe den Spracherkennern zuzuführen. Diese Zuführung kann festgelegt sein, wenn die Anzahl der aktiven Spracherkenner konstant ist, in welchem Fall der Controller **350** keine spezifische Aufgabe für die Zuführung hat.

[0034] In einer bevorzugten Ausführungsform umfasst das System mehr Erkennungsmodelle (M) als aktive Spracherkenner (N). Ein Modellselektor **360** dient dazu, in Abhängigkeit von einem Erkennungskontext für mindestens einen der Spracherkenner das zugehörige Erkennungsmodell aus den M Modellen auszuwählen. Der Modellselektor **360** kann für jeden der aktiven Spracherkenner ein Modell auswählen. Zu bevorzugen ist allerdings, dass ein das allgemein benutzte Vokabular abdeckende Basiserkennungsmodell ständig aktiv ist. In einem solchen Falle braucht mindestens ein Modell nicht vom Modellselektor **360** ausgewählt zu werden und kann einem Spracherkenner fest zugeordnet sein.

[0035] In einer weiteren Ausführungsform wird mindestens ein Erkennungsmodell auf der Basis eines Kontextes ausgewählt, der durch ein Dokument bestimmt wird, auf das sich die Spracheingabe bezieht. Wenn ein Benutzer beispielsweise ein Dokument zum Thema Gesundheit diktiert, kann ein Spracherkenner mit einem spezifischen Erkennungsmodell geladen werden, das für die Erkennung von Sprache mit Bezug auf das Thema Gesundheit optimiert ist. Der Benutzer kann den Kontext für das Dokument explizit angeben, beispielsweise durch Auswählen aus einer Liste möglicher Kontexte, die den Modellen des Systems entsprechen. In diesem Fall kann das System **300** dem Benutzer eine solche Liste auf herkömmliche Weise anbieten, beispielsweise über ein Auswahlfeld in einem Fenster. Das System kann auch den Kontext automatisch bestimmen, in dem es zum Beispiel den im Dokument bereits vorhandenen oder bis dahin gesprochenen Text scannt und überprüft, welches der Modelle am besten für die Erkennung eines solchen Textes geeignet ist (z. B. bei welchem Modell die meisten Wörter oder Wortsequenzen mit dem bisherigen Text übereinstimmen). Weiterhin kann dem Dokument ein Kontextidentifizierer zugeordnet sein und vom System **300** erhalten werden, um das am besten geeignete Modell zu bestimmen. Bei Sprache mit Bezug auf Web-Seiten, wie beispielsweise eine HTML-Seite, sollte(n) der (oder die) Kontexte) des Dokuments vorzugsweise im Dokument oder in Verbindung mit dem Dokument spezifiziert sein. Dies kann in Form von einer Markierung (englisch „tag") geschehen, die von dem Autor der ursprünglichen Web-Seite eingefügt wird, auf die sich die Sprache bezieht. Die Markierung kann den Kontext explizit angeben, zum Beispiel in Form eines textualen Themas wie Sport, Gesundheit; Unterhaltung usw. Die Spezifizierung kann auch indirekt erfolgen, beispielsweise in Form eines Identifizierers wie einer Kontextnummer oder auch einer Verknüpfung (z. B. Hypertextlink) zu einer den Kontext spezifizierenden Stelle. In letzterem Fall kann das System **300** den eigentlichen Kontext von der impliziten Kontextspezifizierung ableiten (z. B. indem man eine Kontextnummer auf eines der Erkennungsmodelle abbildet oder auf den Hypertextlink zugreift und die Kontextinformation erhält).

[0036] In einer bevorzugten Ausführungsform versucht der Modellselektor **360** aktiv die Spracherkennung zu verbessern, indem er prüft, welches der verfügbaren Erkennungsmodelle am besten für die jeweilige Spracherkennung geeignet ist. Zu diesem Zweck steuert der Modellselektor **360** mindestens einen Testerkenner; dargestellt ist der Erkenner **334**. Der Testerkenner **334** ist mit einem der Erkennungsmodelle gekoppelt, das noch nicht vom aktiven Erkenner **331** bis **333** benutzt wird. Die empfangene Sprache wird teilweise (oder sogar vollständig) auch in den Testerkenner eingespeist. Das Resultat der Testerkennung wird mit dem Resultat der vom Controller **350** getroffenen Auswahl oder dem Resultat der einzelnen aktiven Erkenner **331** bis **333** verglichen. Falls das Erkennungsresultat des Testerkenners **334** besser als das Erkennungsresultat eines der aktiven Erkenner **331** bis **333** ist, wird das Testerkennungsmodell (d. h. das momentan vom Testerkenner **334** benutzte Modell) für die Benutzung durch einen der aktiven Erkenner geladen. Vorzugsweise wird das Modell, das die schlechtesten Erkennungsresultate ergab, ersetzt (möglichst mit Ausnahme des Basiserkennungsmodells, das immer benutzt werden könnte).

[0037] Vorzugsweise sind die Erkennungsmodelle hierarchisch angeordnet, von Modellen mit einem eher generischen (allgemeineren) Kontext hin zu Modellen mit einem eher spezifischen Kontext. **Fig.** 4 zeigt eine derartige Hierarchie mit vier sehr generischen Modellen **410**, **420**, **430** und **440**, die beispielsweise die jeweils allgemeinen Themen Unterhaltung, Gesundheit, Reisen und Computer abdecken. Ein generisches Modell wird aufgebaut, indem man repräsentative Texte für alle Fragestellungen zu einem Thema analysiert. Wie man Modelle aus repräsentativen Texten aufbauen kann, ist an sich allgemein bekannt. Das generische Gesundheitsmodell kann mit hierarchisch untergeordneten (d. h. spezifischeren) Modellen verbunden sein, wie beispielsweise mit Bezug auf Medizin, Chirurgie, Nahrungsmittel/Ernährung, Krankenhäuser/medizinische Zentren. Jedes dieser Modelle wird unter Verwendung von Texten erstellt, die sich auf solche spezielleren Themen beziehen. In der Figur kann sich das Modell **422** auf Krankenhäuser/medizinische Zentren beziehen. Innerhalb dieses Kontexts kann eine weitere Unterteilung getroffen werden, bei der beispielsweise das Modell **424** Gesundheitsfarmen abdeckt. Indem Texte, die sich auf Gesundheitsfarmen beziehen, analysiert werden, wird automatisch ein Erkennungsmodell erstellt, das sich auch für die Erkennung von Sprache mit Bezug auf bestimmte Reisethemen eignet, weil in Dokumenten über Gesundheitsfarmen üblicherweise auch die Umgebung beschrieben wird. Dadurch ist dasselbe Modell auch geeignet, um als ein dem Modell **432** in der Kategorie Reisemodelle hierarchisch untergeordnetes Modell benutzt zu werden. Der Modellselektor **360** kann die Erkennung mit einem spezielleren Modell aktivieren, falls die Erkennung mit einem bestimmten Modell gute Erkennungsresultate ergibt. Ein solches spezielleres (d. h. hierarchisch untergeordnetes) Modell kann als Ersatz für das allgemeinere Modell benutzt werden. Es kann auch zusätzlich zu dem allgemeineren Modell benutzt werden. Vorzugsweise findet die zusätzliche Erkennung mit spezielleren Modellen nur dann statt, wenn das allgemeinere Modell gute Resultate im Vergleich zu anderen Modellen ergibt, die sich ohne hierarchische Beziehung auf derselben Hierarchieebene wie das allgemeinere Modell befinden. Wenn beispielsweise ein Sport- und ein Gesundheitsmodell keine hierarchische Beziehung haben (z. B. beide auf der höchsten Ebene) und die Verwendung des Sportmodells bessere Resultate ergibt, dann kann das speziellere Modell benutzt werden. Es besteht kein Anlass, noch speziellere Gesundheitsmodelle zu benutzen. Falls das Erkennungsresultat des Gesundheitsmodells tatsächlich sehr gering ausfällt, dann kann die Erkennung mit diesem Modell zugunsten einer zusätzlichen Erkennung mit einem spezielleren Sportmodell beendet werden. Wenn mehrere speziellere Sportmodelle existieren, z. B. für Fußball, Baseball, Leichtathletik, Autorennen usw., dann kann jedes dieser Modelle getestet werden. Die Auswahl kann auch einfach auf der Übereinstimmung der Vokabulare der verschiedenen Modelle mit der bereits erkannten Sprache beruhen. Falls die Erkennung mit einem speziellen Modell zu einem bestimmten Zeitpunkt geringe Resultate ergibt, wird die Erkennung vorzugsweise mit mindestens einem dem speziellen Modell hierarchisch übergeordneten Modell fortgesetzt.

[0038] In einer bevorzugten Ausführungsform, wie sie in **Fig.** 5 dargestellt ist, ist das Erkennungssystem verteilt. Das verteilte System umfasst eine Serverstation **540** und mindestens eine Benutzerstation. Dargestellt sind drei Benutzerstationen **510**, **520** und **530**, wobei weitere Details nur für die Benutzerstation **520** gezeigt werden. Die Stationen können unter Verwendung konventioneller Computertechnologie realisiert werden. Beispielsweise kann die Benutzerstation **520** durch einen Desktop-PC oder eine Workstation gebildet werden, während die Serverstation **540** durch einen PC-Server oder einen Workstation-Server gebildet werden kann. Die Computer arbeiten unter der Kontrolle eines geeigneten, in den Prozessor des Computers geladenen Programms. Die Serverstation **540** und die Benutzerstationen **510**, **520** und **530** sind über ein Netzwerk **550** miteinander verbunden. Das Netzwerk **550** kann jedes geeignete Netzwerk sein, wie ein lokales Netz (LAN), beispielsweise in einer Büroumgebung, oder ein weiträumiges Netz (WAN), vorzugsweise das Internet. Die Stationen umfassen Kommunikationsmittel **522** bzw. **542**, um über das Netzwerk **550** miteinander zu kommunizieren. Dabei können alle für die Benutzung in Verbindung mit dem Netzwerk **550** geeigneten Kommunikationsmittel verwendet werden. Üblicherweise bestehen die Kommunikationsmittel aus einer Kombination aus Hardware, wie beispielsweise eine Kommunikationsschnittstelle oder ein Modem, und Software in Form eines Softwaretreibers, der ein spezifisches Kommunikationsprotokoll unterstützt, wie beispielsweise die TCP/IP-Protokolle des Internets. Die Benutzerstation **520** umfasst Mittel, um Sprache von einem Benutzer, beispielsweise über eine Schnittstelle **528**, zu empfangen. Weiterhin umfasst die Benutzerstation **520** Mittel zur Vorverarbeitung der Sprache, damit sie für die Übertragung zur Serverstation **540** geeignet ist. Die Benutzerstation kann beispielsweise ein Spektralanalyse-Teilsystem **526** ähnlich dem Spektralanalyse-Teilsystem **110** aus **Fig.** 1 umfassen. Die Serverstation **540** führt alle anderen Aufgaben aus, wie für das System **300** aus **Fig.** 3 beschrieben. Die Serverstation **540** kann beispielsweise ein Erkennungssystem **543** mit einer Vielzahl von Erkennern (ähnlich dem Erkennungssystem **335** aus **Fig.** 3), einen Controller (ähnlich dem Controller **350** aus **Fig.** 3), einen Modellselektor **545** (ähnlich dem Modellselektor **360** aus **Fig.** 3) und einen Speicher **546** zur Speicherung der Modelle (ähnlich dem Speicher **340** aus **Fig.** 3) umfassen.

Patentansprüche

1. Spracherkennungssystem mit großem Vokabular zur Erkennung einer Sequenz von gesprochenen Worten, wobei das System Folgendes umfasst:
Eingabemittel, um ein zeitsequentielles, die Sequenz der gesprochenen Worte darstellendes Eingabemuster zu empfangen; und
Einen Spracherkenner mit großem Vokabular, der unter Verwendung eines dem Spracherkenner zugeordneten Erkennungsmodells mit großem Vokabular das eingegebene Muster als eine Sequenz von Worten erkennt,
dadurch gekennzeichnet, dass
das System eine Vielzahl von N Spracherkennern mit großem Vokabular umfasst, die jeweils zu einem entsprechenden, unterschiedlichen Erkennungsmodell mit großem Vokabular gehören, wobei jedes der Erkennungsmodelle auf einen bestimmten Teil des großen Vokabulars ausgerichtet ist; und
das System einen Controller umfasst, der das Eingabemuster einer Vielzahl der Spracherkenner zuführt und aus den von der Vielzahl von Spracherkennern erkannten Wortsequenzen eine erkannte Wortsequenz auswählt.

2. System nach Anspruch 1, wobei das System M Erkennungsmodelle mit großem Vokabular umfasst, wobei M > N, und das System einen Modellselektor umfasst, der dazu dient, in Abhängigkeit von einem Erkennungskontext für mindestens einen der Spracherkenner das zugehörige Erkennungsmodell aus den M Modellen auszuwählen.

3. System nach Anspruch 2, wobei ein Dokument, auf das sich die Spracheingabe bezieht, mindestens einen Erkennungskontext bestimmt.

4. System nach Anspruch 3, wobei das Dokument eine Web-Seite ist, wie beispielsweise eine HTML-Seite, und der (oder die) Kontexte) des Dokuments vorzugsweise im Dokument oder in Verbindung mit dem Dokument spezifiziert werden.

5. System nach Anspruch 3, wobei der Modellselektor dazu dient, das Erkennungsmodell in Abhängigkeit von Wörtern auszuwählen, die in dem Dokument oder mit diesem in Zusammenhang stehen.

6. System nach Anspruch 3, wobei der Modellselektor zu Folgendem dient:
Auswahl eines Testerkennungsmodells aus den M-N Erkennungsmodellen, die noch nicht von einem der Erkenner benutzt werden;
Steuerung eines Testerkenners, um das Eingabemuster zumindest teilweise mit dem Testerkennungsmodell zu erkennen; und
Aktivierung der Erkennung mit dem Testerkennungsmodell, wenn das Erkennungsresultat des Testerkenners besser als das Erkennungsresultat eines der Erkenner ausfällt.

7. System nach Anspruch 1, wobei die Erkennungsmodelle hierarchisch angeordnet sind, von Modellen mit einem eher generischen Kontext hin zu Modellen mit einem eher spezifischen Kontext, und wobei der Modellselektor dazu dient, die Erkennung mit einem spezielleren Modell zu aktivieren, wenn die Erkennung mit dem allgemeineren Modell ohne hierarchische Beziehung auf einer höheren Hierarchieebene gute Erkennungsresultate im Vergleich zu Resultaten von mindestens einem Erkenner ergibt, der zu anderen Erkennungsmodellen gehört.

8. System nach Anspruch 1, wobei das System eine Benutzerstation und eine Serverstation umfasst, die über ein Netzwerk wie dem Internet miteinander verbunden sind;
wobei die Benutzerstation dazu dient, das Eingabemuster von einem Benutzer zu empfangen und ein das Eingabemuster repräsentierendes Signal an die Serverstation zu übertragen; wobei die Serverstation die Erkenner und den Controller umfasst.

Es folgen 3 Blatt Zeichnungen

Anhängende Zeichnungen

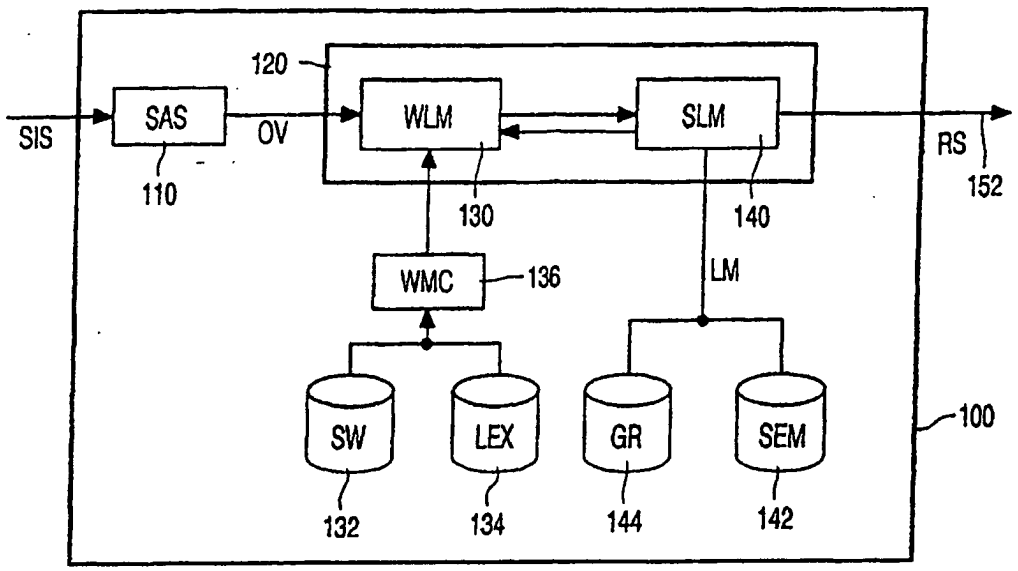

FIG. 1

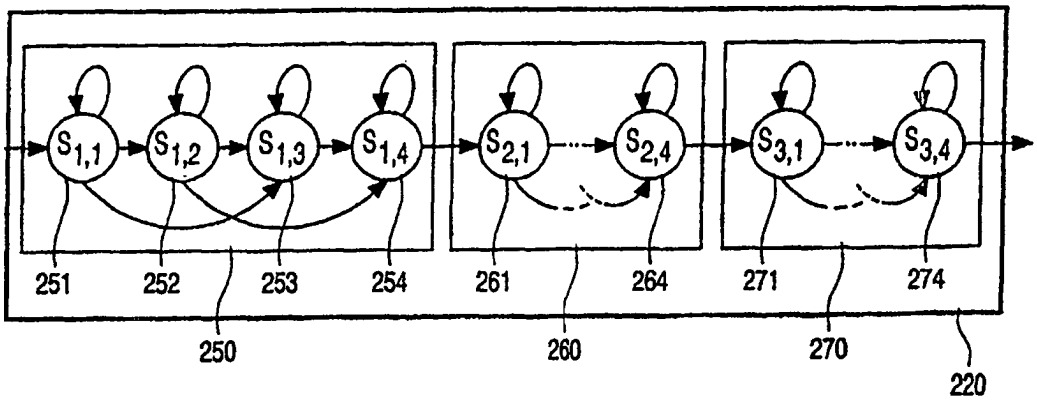

FIG. 2

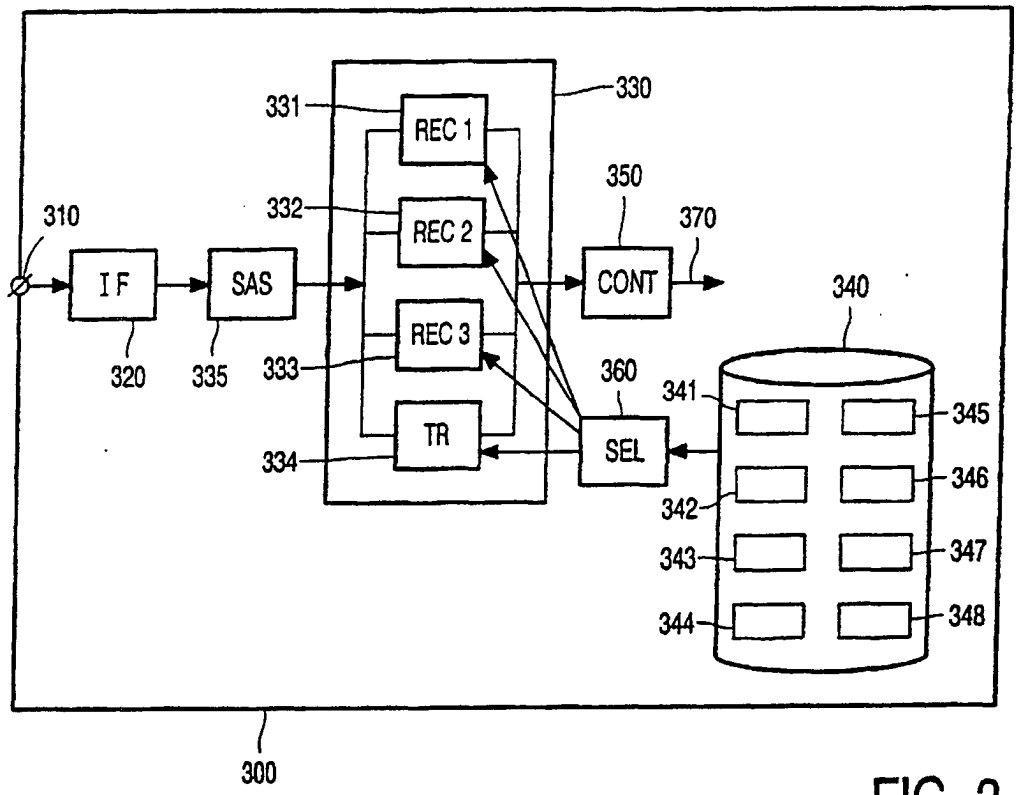
330
331
REC 1
332
REC 2
333
REC 3
334
TR
310
IF
320
SAS
335
350
CONT
370
340
360
SEL
341
342
343
344
345
346
347
348
300

FIG. 3

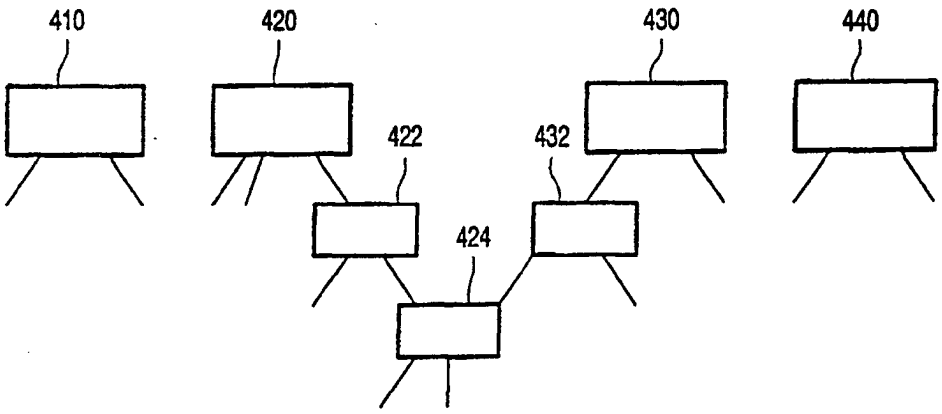
410
420
422
424
430
432
440

FIG. 4

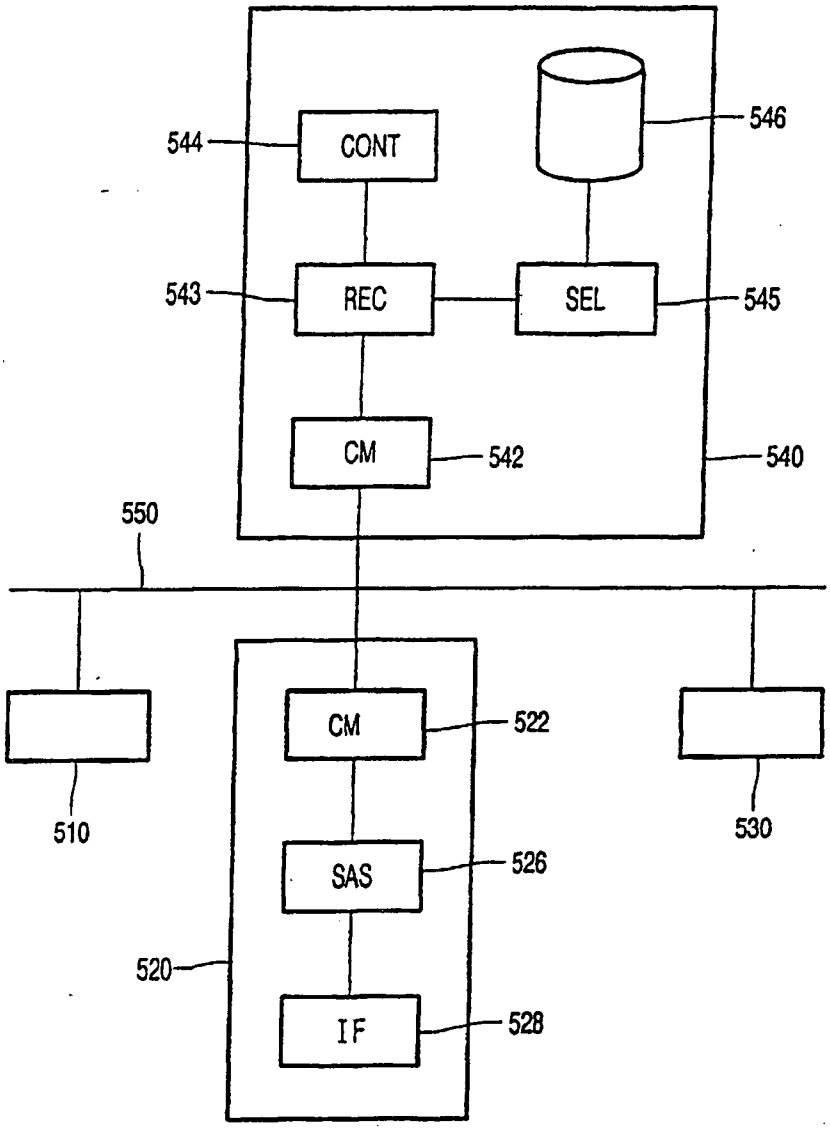
544
CONT
546
543
REC
SEL
545
CM
542
540
550
CM
522
SAS
526
520
IF
528
510
530

FIG. 5