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

(10) **DE 600 34 772 T2** 2008.01.31

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 159 735 B1**
(21) Deutsches Aktenzeichen: **600 34 772.9**
(86) PCT-Aktenzeichen: **PCT/US00/02903**
(96) Europäisches Aktenzeichen: **00 914 513.7**
(87) PCT-Veröffentlichungs-Nr.: **WO 2000/046791**
(86) PCT-Anmeldetag: **04.02.2000**
(87) Veröffentlichungstag der PCT-Anmeldung: **10.08.2000**
(97) Erstveröffentlichung durch das EPA: **05.12.2001**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **09.05.2007**
(47) Veröffentlichungstag im Patentblatt: **31.01.2008**

(51) Int Cl.[8]: ***G10L 15/10*** (2006.01)
G10L 15/22 (2006.01)

(30) Unionspriorität:
248513 **08.02.1999** **US**

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

(73) Patentinhaber:
Qualcomm, Inc., San Diego, Calif., US

(74) Vertreter:
WAGNER & GEYER Partnerschaft Patent- und Rechtsanwälte, 80538 München

(72) Erfinder:
BI, Ning, San Diego, CA 92128, US; CHANG, Chienchung, San Diego, CA 92131, US; GARUDADRI, Harinath, San Diego, CA 92129, US; DEJACO, Andrew P., San Diego, CA 95131, US

(54) Bezeichnung: **ZURÜCKWEISUNGSVERFAHREN IN DER SPRACHERKENNUNG**

Beschreibung

HINTERGRUND DER ERFINDUNG

I. Gebiet der Erfindung

[0001] Die vorliegende Erfindung betrifft allgemein das Gebiet der Kommunikationen, und insbesondere Spracherkennungssysteme.

II. Hintergrund

[0002] Spracherkennung bzw. Stimmerkennung (VR = voice recognition) stellt eine der wichtigsten Techniken dar, um eine Maschine bzw. Vorrichtung mit simulierter Intelligenz auszustatten, um Benutzer oder von Benutzern gesprochene Befehle zu erkennen und um das Bilden einer Mensch-Maschine-Schnittstelle zu erleichtern. VR stellt auch eine Schlüsseltechnik für das Verstehen menschlicher Sprache dar. Systeme, die Techniken einsetzen, um eine linguistische Nachricht aus einem akustischen Sprachsignal zu erhalten, werden als Spracherkennungsvorrichtungen bezeichnet. Eine Spracherkennungsvorrichtung weist typischerweise einen Akustikprozessor auf, welcher eine Sequenz von informationstragenden Merkmalen oder Vektoren extrahiert,, welche benötigt werden, um eine VR der eingehenden Rohsprache zu erreichen, und einen Wortdecodierer, welcher die Sequenz von Merkmalen oder Vektoren decodiert, um eine Ausgabeformat, das eine Bedeutung aufweist und wunschgemäß ist, wie beispielsweise eine Sequenz linguistischer Wörter entsprechend der Eingabeäußerung, zu erhalten. Um die Performance eines gegebenen Systems zu steigern, wird ein Training benötigt, um das System mit gültigen Parametern auszustatten. Mit anderen Worten muss das System lernen, bevor es optimal funktionieren kann.

[0003] Der Akustikprozessor stellt ein Front-End-Sprachanalyseuntersystem in einer Spracherkennungsvorrichtung dar. Ansprechend auf ein Eingabesprachsignal liefert der Akustikprozessor eine geeignete Darstellung, um das mit der Zeit variierende Sprachsignal zu charakterisieren. Der Akustikprozessor sollte irrelevante Information, wie beispielsweise Hintergrundrauschen, Kanalverzerrung, Sprechercharakteristika und die Art des Sprechens verwerfen. Eine effiziente akustische Verarbeitung stattet Spracherkennungsvorrichtungen mit erweiterter bzw. verbesserter akustischer Unterscheidungsleistung aus. Eine nützliche Charakteristik, die zu diesem Zweck analysiert werden sollte, ist die Kurzzeitspektralumhüllende. Zwei herkömmlicher Weise verwendete Spektralanalysetechniken zur Charakterisierung der Kurzzeitspektralumhüllenden sind eine Linearprädiktivcodierung (LPC = linear predictive coding) und eine filterbankbasierte Spektralmodellierung. Beispielhafte LPC-Techniken werden beschrieben in US-Patent Nr. 5,414,796, welches dem Anmelder der vorliegenden Erfindung zueigen ist, und in L.B. Rabiner & R.W. Schafer, Digital Processing of Speech Signals 396-453 (1978).

[0004] Die Verwendung von VR (im Allgemeinen auch als Spracherkennung bezeichnet) wird in zunehmendem Maß wichtig aus Sicherheitsgründen. Beispielsweise kann Spracherkennung verwendet werden, um die manuelle Aufgabe, Knöpfe auf einem Tastenfeld eines drahtlosen Telefons zu drücken, zu ersetzen. Dies ist insbesondere wichtig, wenn ein Benutzer einen Telefonanruf beginnt, während er ein Auto fährt. Wenn er ein Telefon ohne Spracherkennung verwendet, muss der Fahrer eine Hand vom Lenkrad entfernen und auf das Telefontastenfeld sehen, während er die Knöpfe drückt, um den Anruf zu wählen. Diese Handlungen erhöhen die Wahrscheinlichkeit eines Autounfalls. Ein sprachfähiges Telefon (d.h. ein Telefon, das für Spracherkennung konstruiert ist) würde es dem Fahrer gestatten, einen Anruf zu tätigen, während er kontinuierlich die Straße beobachtet. Und ein Autofreisprechsystem (hands free car kit) würde es zusätzlich dem Fahrer gestatten, während der Anrufinitialisierung beide Hände auf dem Lenkrad zu behalten.

[0005] Spracherkennungssysteme werden entweder als sprecherabhängige oder als sprecherunabhängige Einrichtungen bzw. Vorrichtungen eingeordnet. Sprecherunabhängige Einrichtungen sind in der Lage, Sprachbefehle von irgendeinem Benutzer zu akzeptieren. Sprecherabhängige Einrichtungen, welche üblicher sind, werden trainiert, um Befehle von bestimmten Benutzern zu erkennen. Eine sprecherabhängige VR-Einrichtung arbeitet typischerweise in zwei Phasen, einer Trainingsphase und einer Erkennungsphase. In der Trainingsphase veranlasst das VR-System den Benutzer, jedes der Wörter in dem Vokabular des Systems einmal oder zweimal zu sprechen, so dass das System die Charakteristika der Sprache des Benutzers für diese bestimmten Wörter oder Wendungen lernen kann. Alternativ wird, für eine phonetische VR-Einrichtung, das Training durchgeführt durch Vorlesen eines oder mehrerer kurzer Artikel, die speziell geschrieben wurden, um alle Phoneme in der Sprache abzudecken. Ein beispielhaftes Vokabular für ein Autofreisprechanlage kann die Ziffern auf der Tastatur beinhalten; die Stichwörter „anrufen bzw. rufe an", „senden bzw. sende", „wählen bzw. wähle", „abbrechen", „leeren bzw. leere", „hinzufügen bzw. füge hinzu", „Verlauf", „Programm", „ja" und „nein"; Namen von vordefinierten Nummern von üblicherweise angerufenen Kollegen, Freunden oder Familienmitgliedern. Sobald das Training abgeschlossen ist, kann der Benutzer in der Erkennungsphase Anrufe einleiten bzw. initialisieren durch Sprechen der trainierten Stichwörter. Wenn beispielsweise „John" einer der trainierten Namen wäre, dann könnte der Benutzer einen Anruf bei John

einleiten durch Aussprechen des Ausdrucks „Rufe John an". Das VR-System würde die Wörter „rufe an" und „John" erkennen, und würde die Nummer wählen, die der Benutzer zuvor als Johns Telefonnummer eingegeben hat.

[0006] Der Durchsatz eines VR-Systems kann als der Prozentsatz von Fällen definiert werden, in denen ein Benutzer die Erkennungsmaßnahme bzw. -aufgabe erfolgreich durchführt. Eine Erkennungsaufgabe weist typischerweise mehrere Schritte auf. Beispielsweise bezieht sich, bei Sprachwahl mit einem Drahtlostelefon, der Durchsatz auf den durchschnittlichen Prozentsatz der Male, die ein Benutzer erfolgreich einen Telefonanruf mit dem VR-System abschließt. Die Anzahl der Schritte, die notwendig sind, um einen erfolgreichen Telefonanruf mit VR zu erreichen, kann von einem Anruf zum anderen variieren. Im Allgemeinen hängt der Durchsatz eines VR-Systems hauptsächlich von zwei Faktoren ab, der Erkennungsgenauigkeit des VR-Systems, und der Mensch-Maschine-Schnittstelle. Die subjektive Wahrnehmung eines menschlichen Benutzers einer VR-Systemperformance basiert auf dem Durchsatz. Daher besteht ein Bedarf nach einem VR-System mit hoher Erkennungsgenauigkeit und einer intelligenten Mensch-Maschine-Schnittstelle, um den Durchsatz zu erhöhen.

[0007] US-Patent Nr. 4,827,520 und EP 0 867 861 A beschreiben Verfahren, die eine Äußerung mit gespeicherten Vorlagen vergleichen. Es werden Punktzahlen bzw. Werte als ein Ergebnis eines Vergleiches bestimmt und unter besonderen Umständen können die Differenzen zwischen den Punktzahlen bewertet werden.

ZUSAMMENFASSUNG DER ERFINDUNG

[0008] Die vorliegende Erfindung ist auf ein VR-System mit hoher Erkennungsgenauigkeit und eine intelligente Mensch-Maschine-Schnittstelle, zur Erhöhung des Durchsatzes, gerichtet. Demgemäß weist, gemäß einem Aspekt der Erfindung, ein Verfahren für das Erfassen einer Äußerung in einem Spracherkennungssystem auf vorteilhafte Weise die Schritte des Anspruchs 1 auf.

[0009] Gemäß einem weiteren Aspekt der Erfindung weist ein Spracherkennungssystem auf vorteilhafte Weise die Merkmale des Anspruchs 11 auf.

KURZE BESCHREIBUNG DER ZEICHNUNGEN

[0010] Fig. 1 ist ein Blockdiagramm eines Spracherkennungssystems.

[0011] **Fig.** 2 ist ein Graph der Punktzahl gegenüber der Veränderung der Punktzahl für ein Ablehnungsschema eines VR-Systems, das Ablehnungs-, N-Beste- und Akzeptanzbereiche darstellt.

DETAILLIERTE BESCHREIBUNG DER BEVORZUGTEN AUSFÜHRUNGSBEISPIELE

[0012] Gemäß einem wie in Fig. 1 dargestellten Ausführungsbeispiel weist ein Spracherkennungssystem **10** einen Analog-zu-Digital-Wandler bzw. A/D-Wandler (A/D) **12** auf, einen Akustikprozessor **14**, eine VR-Vorlagendatenbank **16**, eine Mustervergleichslogik **18** und eine Entscheidungslogik **20**. Das VR-System **10** kann sich z.B. in einem Drahtlostelefon oder einem Autofreisprechanlage befinden.

[0013] Wenn sich das VR-System **10** in der Spracherkennungsphase befindet, spricht eine Person (nicht gezeigt) ein Wort oder eine Wendung, wobei ein Sprachsignal erzeugt wird. Das Sprachsignal wird in ein elektrisches Sprachsignal s(t) mit einem herkömmlichen Transducer bzw. Wandler (auch nicht gezeigt) konvertiert. Das Sprachsignal s(t) wird dann an den A/D **12** geliefert, der das Sprachsignal s(t) in digitalisierte Sprachabtastungen bzw. -samples s(n) gemäß einem bekannten Abtastungs- bzw. Samplingverfahren, wie beispielsweise pulscodierter Modulation (PCM = pulse coded modulation), konvertiert.

[0014] Die Sprachabtastungen s(n) werden an den Akustikprozessor **14** für eine Parameterbestimmung geliefert. Der Akustikprozessor **14** erzeugt einen Satz von extrahierten Parametern, die die Charakteristika des Eingabesprachsignals s(t) modellieren. Die Parameter können gemäß irgendeiner einer Anzahl von bekannten Sprachparameterbestimmungstechniken bestimmt werden, einschließlich beispielsweise Codierung durch einen Sprachcodierer und unter Verwendung von Cepstrum-Koeffizienten, die auf einer Fast-Fourier-Transformation (FFT) basieren, wie beschrieben im zuvor erwähnten US-Patent Nr. 5,414,796. Der Akustikprozessor **14** kann als ein Digitalsignalprozessor (DSP) implementiert werden. Der DSP kann einen Sprachcodierer aufweisen. Alternativ kann der akustische Prozessor bzw. Akustikprozessor **14** als ein Sprachcodierer implementiert werden.

[0015] Die Parameterbestimmung wird auch während des Trainings des VR-Systems **10** durchgeführt, in dem ein Satz von Vorlagen für alle Vokabularwörter des VR-Systems **10** an die VR-Vorlagendatenbank **16** zur permanenten Speicherung darin weitergeleitet wird. Die VR-Vorlagendatenbank **16** wird vorteilhafter Weise als irgendeine herkömmliche Form eines nichtflüchtigen Speichermediums implementiert, wie beispielsweise als Flash-Speicher. Dies gestattet, dass die Vorlagen in der VR-Vorlagendatenbank **16** verbleiben, wenn die Leistung an das VR-System **10** abgeschaltet wird.

[0016] Der Satz von Parametern wird an die Mustervergleichslogik **18** geliefert. Die Mustervergleichslogik **18** detektiert auf vorteilhafte Weise die Start- und Endpunkte einer Äußerung, berechnet dynamische akustische bzw. Akustikmerkmale (wie beispielsweise Ableitungen nach der Zeit, zweite Ableitungen nach der Zeit usw.), komprimiert die Akustikmerkmale durch Auswählen relevanter Rahmen, und quantisiert die statischen und dynamischen Akustikmerkmale. Verschiedene Verfahren zur Endpunktdetektion, dynamischen Akustikmerkmalsableitung, Musterkomprimierung und Musterquantisierung werden beispielsweise in Lawrence Rabiner & Biing-Hwang Juang, Fundamentals of Speech Recognition (1993) beschrieben. Die Mustervergleichslogik **18** vergleicht den Satz der Parameter mit allen Vorlagen, die in der VR-Vorlagendatenbank **16** gespeichert sind. Die Vergleichsergebnisse, oder die Distanzen, zwischen dem Satz der Parameter und allen Vorlagen, die in der VR-Vorlagendatenbank **16** gespeichert sind, werden an die Entscheidungslogik **20** geliefert. Die Entscheidungslogik **20** kann (1) aus der VR-Vorlagendatenbank **16** die Vorlage auswählen, die am nächstliegendsten bzw. ehesten mit dem Satz der Parameter übereinstimmt, oder kann (2) einen „N-Beste"-Auswahlalgorithmus anwenden, der die N engsten Übereinstimmungen innerhalb einer vordefinierten Übereinstimmungsschwelle auswählt; oder kann (3) den Satz von Parametern zurückweisen. Wenn ein N-Beste-Algorithmus verwendet wird, dann wird die Person befragt, welche Auswahl beabsichtigt war. Die Ausgabe der Entscheidungslogik **20** ist die Entscheidung, welches Wort in dem Vokabular ausgesprochen wurde. In einer N-Beste-Situation beispielsweise könnte die Person „John Anders" sagen, und das VR-System **10** könnte erwidern „Sagten Sie John Andrews?" Die Person würde dann „John Anders" antworten. Das VR-System **10** könnte dann „Sagten Sie John Anders?" antworten. Die Person würde dann „Ja" antworten, an welchem Punkt das VR-System **10** das Wählen eines Telefonanrufs beginnen würde.

[0017] Die Mustervergleichslogik **18** und die Entscheidungslogik **20** können auf vorteilhafte Weise als ein Mikroprozessor implementiert werden. Alternativ können die Mustervergleichslogik **18** und die Entscheidungslogik **20** als eine herkömmliche Form eines Prozessors, eines Controllers oder einer Zustandsmaschine implementiert werden. Das VR-System **10** kann beispielsweise ein ASIC (ASIC = application specific integrated circuit) sein. Die Erkennungsgenauigkeit des VR-Systems **10** ist ein Maß, wie gut das VR-System **10** gesprochene Wörter oder Wendungen in dem Vokabular richtig erkennt. Beispielsweise zeigt eine Erkennungsgenauigkeit von 95% an, dass das VR-System **10** Wörter in dem Vokabular 95 von 100 Malen korrekt erkennt.

[0018] In einem Ausführungsbeispiel ist ein Graph der Punktzahl gegenüber der Veränderung der Punktzahl in Bereiche der Akzeptanz, der N-Besten und der Zurückweisung segmentiert, wie in **Fig.** 2 dargestellt. Die Bereiche werden durch Linien getrennt gemäß bekannten linearen Unterscheidungsanalysetechniken, welche beschrieben werden in Richard O. Duda & Peter E. Hart, Pattern Classification and Scene Analysis (1973). Jeder Äußerung, die in das VR-System **10** eingegeben wird, wird ein Vergleichsergebnis für, oder eine Distanz von, jeder Vorlage, die in der VR-Vorlagendatenbank **16** gespeichert ist, durch die Mustervergleichslogik **18** zugewiesen, wie oben beschrieben. Diese Distanzen oder „Punktzahlen" können auf vorteilhafte Weise Euklidische Distanzen zwischen Vektoren in einem N-dimensionalen Vektorraum sein, die über mehrere Rahmen summiert wurden. In einem Ausführungsbeispiel ist der Vektorraum ein vierundzwanzigdimensionaler Vektorraum, die Punktzahl wird kumuliert über zwanzig Rahmen, und die Punktzahl ist eine ganzzahlige Distanz. Der Fachmann wird verstehen, dass die Punktzahl auch gleich gut ausgedrückt werden kann als Bruch oder als anderer Wert. Der Fachmann wird auch verstehen, dass andere Metriken anstelle der Euklidschen Distanzen bzw. Abständen verwendet werden können, so dass die die Punktzahlen beispielsweise Wahrscheinlichkeitsmaße, Auftrittswahrscheinlichkeiten etc. sein können.

[0019] Für eine gegebene Äußerung und eine gegebene VR-Vorlage aus der VR-Vorlagendatenbank **16**, gilt, dass je niedriger die Punktzahl (d.h. je kleiner die Distanz zwischen der Äußerung und der VR-Vorlage), umso näher bzw. genauer ist die Übereinstimmung zwischen der Äußerung und der VR-Vorlage. Für jede Äußerung analysiert die Entscheidungslogik **20** die Punktzahl, die mit der nächstgelegenen Übereinstimmung in der VR-Vorlagendatenbank **16** assoziiert ist, und zwar in Bezug zur Differenz zwischen dieser Punktzahl und der Punktzahl, die mit der am zweitnächsten gelegenen bzw. zweitbesten Übereinstimmung in der VR-Vorlagendatenbank **16** assoziiert ist (d.h. der zweitniedrigsten Punktzahl). Wie im Graph der **Fig.** 2 dargestellt, wird die „Punktzahl" gegen die „Veränderung in der Punktzahl" aufgetragen, und drei Bereiche werden definiert. Der Zurückweisungsbereich stellt ein Gebiet dar, in dem eine Punktzahl relativ hoch ist und die Differenz zwischen dieser Punktzahl und der nächsten niedrigsten Punktzahl relativ gering ist. Wenn eine Äußerung in den Zurückweisungsbereich fällt, weist die Entscheidungslogik **20** die Äußerung zurück. Der Akzeptanzbereich stellt ein Gebiet dar, in dem eine Punktzahl relativ niedrig ist und die Differenz zwischen dieser Punktzahl und der nächsten niedrigsten Punktzahl relativ groß ist. Wenn eine Äußerung in den Akzeptanzbereich fällt, dann akzeptiert die Entscheidungslogik **20** die Äußerung. Der N-Beste-Bereich liegt zwischen dem Zurückweisungsbereich und dem Akzeptanzbereich. Die N-Beste-Region stellt ein Gebiet dar, in dem ent-

weder eine Punktzahl geringer ist als eine Punktzahl in dem Zurückweisungsbereich oder die Differenz zwischen dieser Punktzahl und der nächsten niedrigsten Punktzahl größer ist als die Differenz für eine Punktzahl in dem Zurückweisungsbereich. Die N-Beste-Region stellt auch ein Gebiet dar, in dem entweder eine Punktzahl größer ist als eine Punktzahl im Akzeptanzbereich oder die Differenz zwischen der Punktzahl und der nächsten niedrigsten Punktzahl geringer ist als die Differenz für eine Punktzahl in dem Akzeptanzbereich, vorausgesetzt die Differenz für die Punktzahl in dem N-Beste-Bereich ist größer als eine vordefinierter Schwellenwert für die Veränderung in der Punktzahl. Falls eine Äußerung in den N-Beste-Bereich fällt, dann wendet die Entscheidungslogik **20** einen N-Beste-Algorithmus auf die Äußerung an, wie oben beschrieben.

[0020] In dem mit Bezug zu **Fig.** 2 beschriebenen Ausführungsbeispiel trennt ein erstes Liniensegment den Zurückweisungsbereich von dem N-Beste-Bereich. Das erste Liniensegment schneidet die „Punktzahl"-Achse bei einem vorbestimmten Schwellenpunktzahlwert. Die Steigung des ersten Liniensegments ist auch vordefiniert. Ein zweites Liniensegment trennt den N-Beste-Bereich vom Akzeptanzbereich. Die Steigung des zweiten Liniensegments ist vorbestimmt mit der Steigung des ersten Liniensegments gleich zu sein, so dass die ersten und zweiten Liniensegmente parallel sind. Ein drittes Liniensegment erstreckt sich vertikal von einem vorbestimmten Schwellenveränderungswert auf der „Veränderung in der Punktzahl"-Achse, um einen Endpunkt des zweiten Liniensegments zu treffen. Es wird dem Fachmann klar sein, dass die ersten und zweiten Liniensegmente nicht parallel sein müssen, und jegliche zufällig zugewiesene Steigungen aufweisen könnten. Zudem muss das dritte Liniensegment nicht verwendet werden.

[0021] In einem Ausführungsbeispiel ist der Schwellenpunktzahlwert 375, der Schwellenveränderungswert ist 28, und wenn der Endpunkt des zweiten Liniensegments verlängert würde, würde das zweite Liniensegment die „Punktzahl"-Achse beim Wert 250 schneiden, so dass die Steigungen der ersten und zweiten Liniensegment jeweils 1 sind. Wenn der Punktzahlwert größer ist als der Veränderung-in-der-Punktzahl-Wert plus 375, dann wird die Äußerung zurückgewiesen. Anderenfalls, wenn entweder der Punktzahlwert größer ist als der Veränderung-in-der-Punktzahl-Wert plus 250 oder der Veränderung-in-der-Punktzahl-Wert geringer ist als 28, dann wird ein N-Beste-Algorithmus auf die Äußerung angewandt. Anderenfalls wird die Äußerung akzeptiert.

[0022] In dem mit Bezug zu **Fig.** 2 beschriebenen Ausführungsbeispiel werden zwei Dimensionen für die lineare Unterscheidungsanalyse verwendet. Die Dimension "Punktzahl" stellt die Distanz zwischen einer gegebenen Äußerung und einer gegebenen VR-Vorlage dar, wie aus den Ausgaben der Mehrfach-Bandpassfilter (nicht gezeigt) abgeleitet. Die Dimension "Veränderung in der Punktzahl" stellt die Differenz zwischen dem niedrigsten Wert, d.h. der Punktzahl der engsten Übereinstimmung, und dem nächsten niedrigsten Wert, d.h. der Punktzahl für die am nächstengsten übereinstimmende Äußerung, dar. In einem weiteren Ausführungsbeispiel stellt die Dimension "Punktzahl" die Distanz zwischen einer gegebenen Äußerung und einer gegebenen VR-Vorlage dar, wie aus den Cepstral-Koeffizienten der Äußerung abgeleitet. In einem weiteren Ausführungsbeispiel stellt die Dimension "Punktzahl" die Distanz zwischen einer gegebenen Äußerung und einer gegebenen VR-Vorlage dar, wie abgeleitet durch die Koeffizienten der Linearprädiktivcodierung bzw. LPC-Koeffizienten (LPC = linear predictive coding) der Äußerung. Techniken zum Herleiten der LPC-Koeffizienten und der Cepstral-Koeffizienten einer Äußerung werden beschrieben in dem zuvor erwähnten US-Patent Nr. 5,414,796.

[0023] In anderen Ausführungsbeispielen ist die lineare Unterscheidungsanalyse nicht auf zwei Dimensionen beschränkt. Demgemäß werden eine erste Punktzahl basierend auf Bandpassfilterausgaben, eine zweite Punktzahl basierend auf Cepstral-Koeffizienten, und eine Veränderung in der Punktzahl im Verhältnis zueinander analysiert. Alternativ werden eine erste Punktzahl basierend auf Bandpassfilterausgaben, eine zweite Punktzahl basierend auf Cepstral-Koeffizienten, eine dritte Punktzahl basierend auf LPC-Koeffizienten und eine Veränderung in der Punktzahl im Verhältnis zueinander analysiert. Wie der Fachmann leicht erkennen wird, muss die Anzahl der Dimensionen für die "Punktzahl" nicht auf irgendeine bestimmte Anzahl beschränkt sein. Der Fachmann wird erkennen, dass die Anzahl der Punktzahldimensionen nur durch die Anzahl der Wörter im Vokabular des VR-Systems beschränkt ist. Der Fachmann wird auch erkennen, dass die verwendeten Arten von Punktzahlen nicht auf irgendeine bestimmte Art von Punktzahl beschränkt sein muss, sondern dass sie jedes Punktzahl- bzw. Bewertungsverfahren enthalten kann, das auf dem Fachgebiet bekannt ist. Weiter wird dem Fachmann auch leicht ersichtlich sein, dass die Anzahl der Dimensionen für die "Veränderung in der Punktzahl" nicht auf Eins beschränkt sein muss, oder auf irgendeine bestimmte Zahl. Beispielsweise wird in einem Ausführungsbeispiel eine Punktzahl analysiert im Verhältnis zu einer Veränderung in der Punktzahl zwischen der nächstliegender Übereinstimmung und der nächsten nächstliegenden Übereinstimmung, und die Punktzahl wird auch analysiert im Verhältnis zu einer Veränderung zwischen der nächstliegenden Übereinstimmung und der drittnächsten Übereinstimmung. Für den Fachmann sollte es ersichtlich sein, dass die

Anzahl der Dimensionen der Veränderung in der Punktzahl nur durch die Anzahl der Wörter in dem Vokabular des VR-Systems beschränkt ist.

[0024] Somit ist ein neues und verbessertes Spracherkennungszurückweisungsschema, basierend auf linearer Unterscheidungsanalyse beschrieben worden. Der Fachmann wird verstehen, dass die verschiedenen illustrativen logischen Blöcke und Algorithmusschritte, die in Verbindung mit den hierin offenbarten Ausführungsbeispielen beschrieben wurden, mit einem Digitalsignalprozessor (DSP), einem ASIC (ASIC = application specific integrated circuit), diskretem Gatter oder Transistorlogik, diskreten Hardwarekomponenten, wie beispielsweise Registern und FIFO, einem Prozessor, der einen Satz von Firmware-Instruktionen ausführt, oder irgendeinem herkömmlichen programmierbaren Softwaremodul und einem Prozessor, implementiert oder ausgeführt werden können. Der Prozessor kann vorteilhafter Weise ein Mikroprozessor sein, aber alternativ kann der Prozessor jeder herkömmliche Prozessor, Controller, Mikrocontroller oder eine Zustandsmaschine sein. Das Softwaremodul kann sich in RAM-Speicher, Flash-Speicher, Registern oder jeder anderen Form von beschreibbarem Speichermedium, das auf dem Fachgebiet bekannt ist, befinden. Der Fachmann wird weiter erkennen, dass die Daten, Instruktionen, Befehle, Informationen, Signale, Bits, Symbole und Chips, auf die über die obige Beschreibung hinweg Bezug genommen wurde, auf vorteilhafte Weise durch Spannungen, Ströme, elektromagnetische Wellen, magnetische Felder oder Partikel, optische Felder oder Partikel, oder irgendeine Kombination davon, dargestellt werden können.

[0025] Somit sind bevorzugte Ausführungsbeispiele der vorliegenden Erfindung gezeigt und beschrieben worden. Es wird dem Fachmann jedoch klar sein, dass zahlreiche Änderungen an den Ausführungsbeispielen ausgeführt werden können, ohne vom Umfang der hierin offenbarten Erfindung abzuweichen.

[0026] Daher soll die vorliegende Erfindung nicht eingeschränkt werden, außer gemäß den folgenden Ansprüchen.

Patentansprüche

1. Ein Verfahren zum Erfassen einer Äußerung in einem Spracherkennungssystem (**10**), wobei das Verfahren folgende Schritte aufweist:
Vergleichen (**18**) der Äußerung mit einem ersten gespeicherten Wort, um eine erste Punktzahl bzw. Wert zu generieren;
Vergleichen (**18**) der Äußerung mit einem zweiten gespeicherten Wort, um eine zweite Punktzahl zu generieren; und
Bestimmen (**18**) einer Differenz zwischen der ersten Punktzahl und der zweiten Punktzahl;
Verarbeiten (**20**) der Äußerung, basierend auf der ersten Punktzahl und der bestimmten Differenz, und zwar durch:
Vergleichen der ersten Punktzahl mit einem ersten mit Steigung versehenen Schwellenwert bzw. Steigungsschwellenwert und Zurückweisen der Äußerung, wenn die erste Punktzahl größer ist als der erste mit Steigung versehene Schwellenwert;
andernfalls, Vergleichen der ersten Punktzahl mit einem zweiten mit Steigung versehenen Schwellenwert und Anwenden eines N-Beste-Algorithmus, um die Äußerung zu verifizieren, wenn die erste Punktzahl größer ist als der zweite mit Steigung versehene Schwellenwert; andernfalls, Akzeptieren der Äußerung;
wobei die ersten und zweiten mit Steigung versehenen Schwellenwerte bezüglich der bestimmten Differenz variieren.

2. Verfahren nach Anspruch 1, wobei der Schritt des Vergleichens der ersten Punktzahl mit einem zweiten mit Steigung versehenen Schwellenwert weiterhin die bestimmte Differenz mit einer Differenzschwelle vergleicht, und der N-Beste-Algorithmus angewendet wird, um die Äußerung zu verifizieren, wenn die erste Punktzahl größer ist als der zweite mit Steigung versehene Schwellenwert und die Differenz geringer ist als die Differenzschwelle.

3. Verfahren nach Anspruch 1 oder 2, wobei die Steigung des ersten mit Steigung versehenen Schwellenwerts und die Steigung des zweiten mit Steigung versehenen Schwellenwerts dieselbe ist.

4. Verfahren nach Anspruch 1, wobei die Differenz einer Veränderung der Punktzahl zwischen der ersten Punktzahl und der zweiten Punktzahl entspricht.

5. Verfahren nach Anspruch 1, wobei das erste gespeicherte Wort einen besten Kandidaten in einem Vokabular eines Spracherkennungssystems (**10**) aufweist, und das zweite gespeicherte Wort einen nächstbesten Kandidaten in einem Vokabular eines Spracherkennungssystems (**10**) aufweist.

6. Verfahren nach Anspruch 1, wobei die erste Punktzahl ein nächstliegendes Vergleichsergebnis aufweist, und die zweite Punktzahl ein nächstes nächstliegendes Vergleichsergebnis aufweist.

7. Verfahren nach Anspruch 1, wobei die erste Punktzahl und die zweite Punktzahl linear prädiktive Codierungskoeffizienten aufweisen.

8. Verfahren nach Anspruch 1, wobei die erste Punktzahl und die zweite Punktzahl Cepstral-Koeffizienten aufweisen.

9. Verfahren nach Anspruch 1, wobei die erste

Punktzahl und die zweite Punktzahl Bandpassfilterausgaben aufweisen.

10. Verfahren nach Anspruch 1, wobei die Differenz eine Differenz zwischen einem nächstliegenden Vergleichsergebnis und dem nächsten nächstliegenden Vergleichsergebnis aufweist.

11. Ein Spracherkennungssystem (**10**), das Folgendes aufweist:
Mittel zum Vergleichen (**18**) der Äußerung mit einem ersten gespeicherten Wort, um eine erste Punktzahl zu generieren;
Mittel zum Vergleichen (**18**) der Äußerung mit einem zweiten gespeicherten Wort, um eine zweite Punktzahl zu generieren; und
Mittel zum Bestimmen (**18**) einer Differenz zwischen der ersten Punktzahl und der zweiten Punktzahl;
Mittel zum Verarbeiten (**20**) der Äußerung, basierend auf der ersten Punktzahl und der bestimmten Differenz, und zwar betriebsmäßig zum:
Vergleichen der ersten Punktzahl mit einem ersten, mit Steigung versehenen Schwellenwert bzw. eines Steigungsschwellenwert und Zurückweisen der Äußerung, wenn die erste Punktzahl größer ist als der mit Steigung versehene erste Schwellenwert;
andernfalls Vergleichen der ersten Punktzahl mit einem zweiten, mit Steigung versehenen Schwellenwert und Anwenden eines N-Beste-Algorithmus um die Äußerung zu verifizieren, wenn die zweite Punktzahl größer ist als der zweite, mit Steigung versehene Schwellenwert;
andernfalls Akzeptieren der Äußerung;
wobei die ersten und zweiten mit Steigung versehenen Schwellenwerte gemäß der bestimmten Differenz variieren.

12. Spracherkennungssystem nach Anspruch 11, wobei die Mittel zum Verarbeiten (**20**) weiterhin betriebsmäßig dienen zum Vergleich der ersten Punktzahl mit einem zweiten, mit Steigung versehenen Wert und zum Vergleichen der bestimmten Differenz mit einer Differenzschwelle, und wobei der N-Beste-Algorithmus angewendet wird, um die Äußerung zu verifizieren, wenn die erste Punktzahl größer ist als der zweite mit Steigung versehene Schwellenwert, und die Differenz geringer ist als die Differenzschwelle.

13. Spracherkennungssystem nach Anspruch 11 oder 12, wobei die Steigung der ersten mit Steigung versehenen Schwellenwerte und die Steigung der zweiten mit Steigung versehenen Schwellenwerte die gleiche ist.

14. Spracherkennungssystem (**10**) nach Anspruch 11, das Folgendes aufweist:
Mittel (**14**) zum Extrahieren von Sprachparametern von digitalisierten Sprachabtastungen der Äußerung,
wobei die Mittel (**18**) zum Vergleichen der Äußerung mit einem ersten gespeicherten Wort, die Mittel zum Vergleichen (**18**) der Äußerung mit einem zweiten gespeicherten Wort, die Mittel (**18**) zum Bestimmen einer Differenz, die Mittel (**20**) zum Bestimmen eines Verhältnisses und die Mittel (**20**) zum Verarbeiten aller Teile eines einzigen Mittels sind.

15. Spracherkennungssystem (**10**) nach Anspruch 14, wobei:
die Mittel zum Extrahieren (**14**) einen Akustikprozessor (**14**) aufweisen; und
das einzelne Mittel einen Prozessor aufweist, gekoppelt an den Akustikprozessor (**14**).

16. Spracherkennungssystem (**10**) nach Anspruch 11, wobei das erste gespeicherte Wort einen besten Kandidaten in einem Vokabular des Spracherkennungssystems (**10**) aufweist, und das zweite gespeicherte Wort einen nächstbesten Kandidaten in einem Vokabular des Spracherkennungssystems (**10**) aufweist.

17. Spracherkennungssystem (**10**) nach Anspruch 11, wobei die erste Punktzahl ein nächstliegendes Vergleichsergebnis aufweist, und die zweite Punktzahl ein nächstes nächstliegendes Vergleichsergebnis aufweist.

18. Spracherkennungssystem (**10**) nach Anspruch 11, wobei die erste Punktzahl und die zweite Punktzahl linear prädiktive Codierungskoeffizienten aufweisen.

19. Spracherkennungssystem (**10**) nach Anspruch 11, wobei die erste Punktzahl und die zweite Punktzahl Cepstral-Koeffizienten aufweisen.

20. Spracherkennungssystem (**10**) nach Anspruch 11, wobei die erste Punktzahl und die zweite Punktzahl Bandpassfilterausgaben aufweisen.

21. Spracherkennungssystem (**10**) nach Anspruch 11, wobei die Differenz eine Differenz zwischen einem nächstliegenden Vergleichsergebnis und einem nächsten nächstliegenden Vergleichsergebnis aufweist.

Es folgen 2 Blatt Zeichnungen

Anhängende Zeichnungen

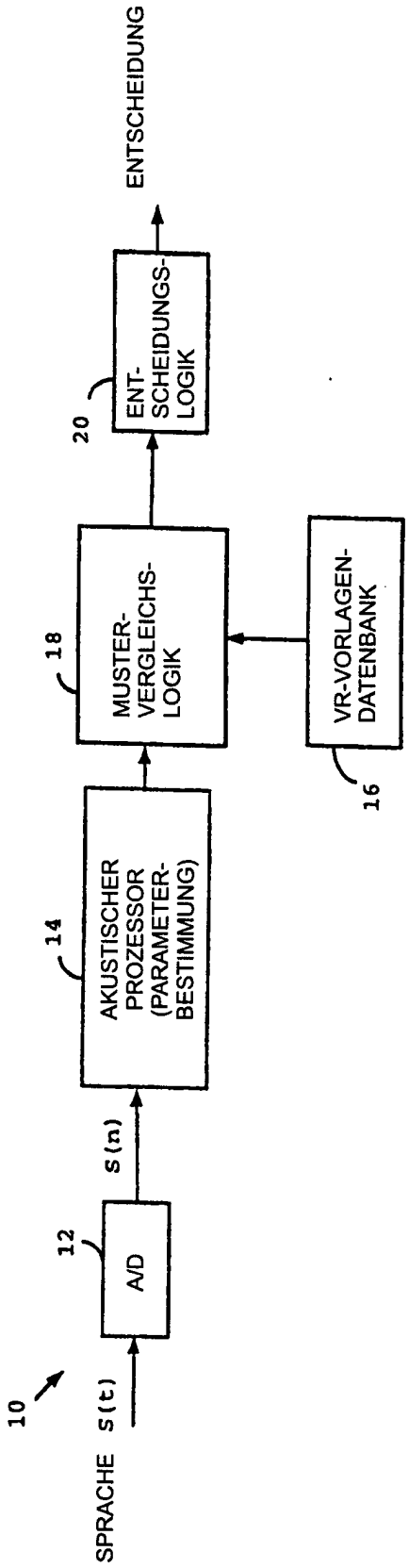
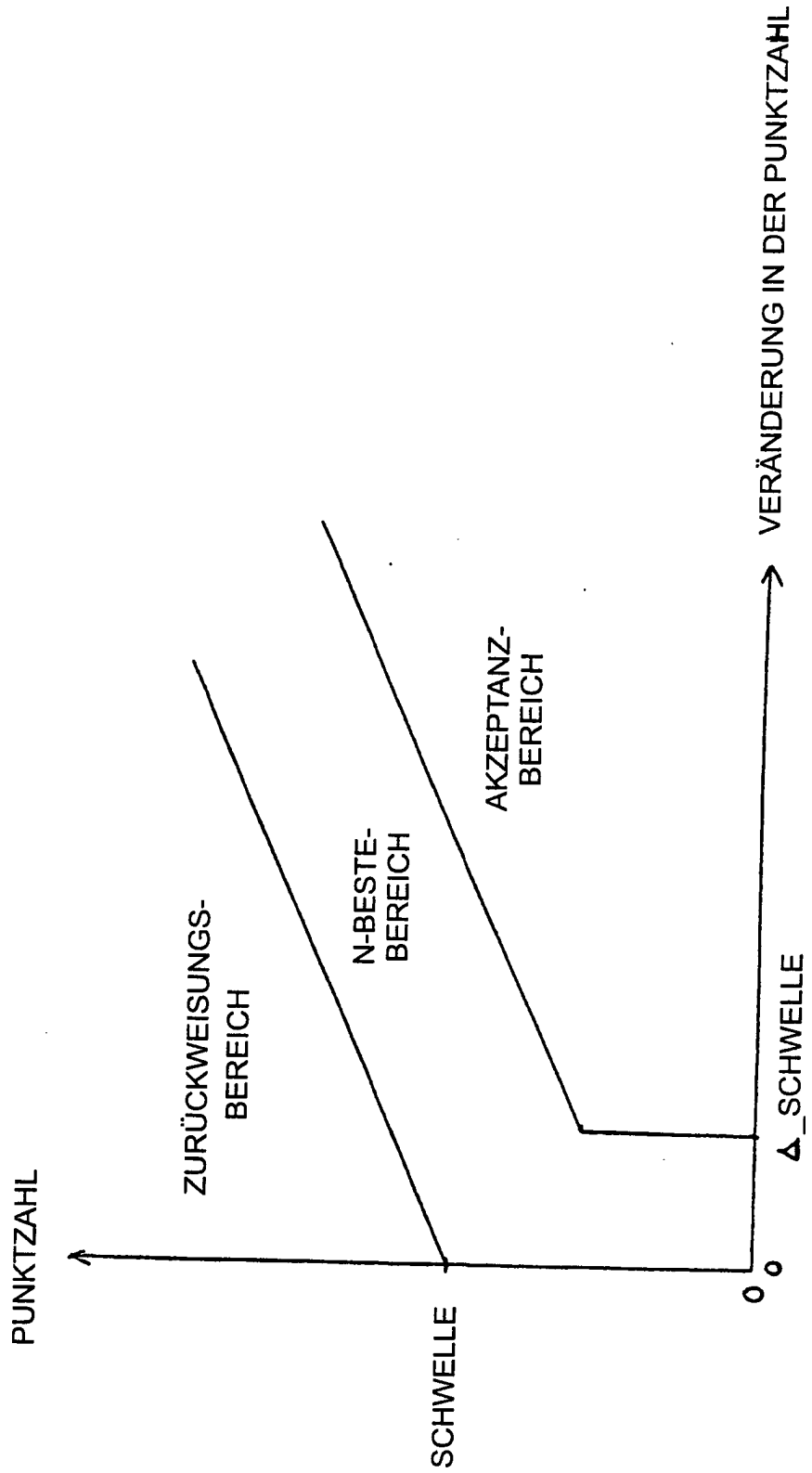


FIG. 1


PUNKTZAHL
ZURÜCKWEISUNGS-
BEREICH
N-BESTE-
BEREICH
AKZEPTANZ-
BEREICH
VERÄNDERUNG IN DER PUNKTZAHL
Δ_SCHWELLE
SCHWELLE
0
0