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

(10) **DE 699 16 255 T2** 2005.04.14

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 058 925 B1**
(21) Deutsches Aktenzeichen: **699 16 255.6**
(86) PCT-Aktenzeichen: **PCT/US99/02280**
(96) Europäisches Aktenzeichen: **99 905 664.1**
(87) PCT-Veröffentlichungs-Nr.: **WO 99/040571**
(86) PCT-Anmeldetag: **03.02.1999**
(87) Veröffentlichungstag der PCT-Anmeldung: **12.08.1999**
(97) Erstveröffentlichung durch das EPA: **13.12.2000**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **07.04.2004**
(47) Veröffentlichungstag im Patentblatt: **14.04.2005**

(51) Int Cl.[7]: **G10L 15/20**
G10L 15/06

(30) Unionspriorität:
18257 **04.02.1998** **US**

(84) Benannte Vertragsstaaten:
DE, FR

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

(72) Erfinder:
SIH, C., Gilbert, San Diego, US; BI, Ning c/o QUALCOMM Incorporated, San Diego, US

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

(54) Bezeichnung: **SYSTEM UND VERFAHREN ZUR GERÄUSCHKOMPENSIERTEN SPRACHERKENNUNG**

Beschreibung

Hintergrund der Erfindung

1. Gebiet der Erfindung

[0001] Die vorliegende Erfindung bezieht sich auf Sprachverarbeitung. Spezieller bezieht sich die vorliegende Erfindung auf ein System und Verfahren für die automatische Erkennung gesprochener Wörter oder Sätze.

2. Beschreibung der verwandten Technik

[0002] Digitale Verarbeitung von Sprachsignalen hat eine weit verbreitete Anwendung gefunden, besonders bei Mobiltelefon- und PCS Anwendungen. Eine digitale Sprachverarbeitungstechnik ist die der Spracherkennung. Die Verwendung von Spracherkennung gewinnt Bedeutung aufgrund von Sicherheitsgründen. Beispielsweise kann Spracherkennung verwendet werden, um die manuelle Aufgabe des Drückens von Knöpfen auf der Tastatur eines Mobiltelefons zu ersetzen. Das ist besonders wichtig, wenn ein Nutzer einen Telefonanruf beginnen will, während er ein Auto fährt. Bei der Verwendung eines Telefons ohne Spracherkennung muss der Fahrer eine Hand vom Lenkrad nehmen und während dem Drücken der Knöpfe auf die Tastatur des Telefons schauen, um den Anruf zu wählen. Diese Handlungen erhöhen die Wahrscheinlichkeit eines Autounfalls. Spracherkennung erlaubt es dem Fahrer Telefonanrufe durchzuführen, während dem kontinuierlichen Beobachten der Straße und dem Halten beider Hände auf dem Lenkrad. Freisprecheinrichtungen für das Auto die Spracherkennung umfassen, werden wahrscheinlich eine gesetzliche Anforderung in zukünftigen Systemen aufgrund von Sicherheitsgründen sein.

[0003] Sprecherabhängige Spracherkennung, die heute am heutigsten verwendete Art, arbeitet in zwei Phasen: einer Trainingsphase und einer Erkennungsphase. In der Trainingsphase fordert das Spracherkennungssystem den Nutzer auf, jedes der Wörter in dem Vokabular einmal oder zweimal zu sprechen, so dass es die Charakteristiken der Sprache des Nutzers für diese speziellen Wörter oder Sätze bzw. Phrasen lernen kann. Die Erkennungsvokabulargrößen sind typischerweise klein (weniger als 50 Wörter) und das Spracherkennungssystem wird nur eine hohe Erkennungsgenauigkeit bei dem Nutzer erreichen, der es trainiert hat. Ein Beispiel eines Vokabulars für ein Freisprechsystem für das Auto würde folgendes umfassen: Die Zahlen auf der Tastatur, die Schlüsselwörter „Anruf", „Sende", „Wähle", „Abbrechen", „Freigeben" (clear), „Addiere", „Lösche", „Verlauf", „Programmiere", „Ja" und „Nein", sowie auch 20 Namen von häufig angerufenen Arbeitskollegen, Freunden oder Familienmitgliedern. Sobald das Training vollendet ist, kann der Nutzer Anrufe in der Erkennungsphase durch Sprechen der trainierten Schlüsselwörter veranlassen. Beispielsweise falls der Name „John" einer der trainierten Namen war, kann der Nutzer einen Anruf mit John beginnen durch sagen des Satzes „Anruf John". Das Spracherkennungssystem erkennt die Wörter „Anruf" und „John", und wählt die Nummer die der Nutzer vorher als Johns Telefonnummer eingegeben hat.

[0004] Ein Blockdiagramm einer Trainingseinheit **6** eines sprecherabhängigen Spracherkennungssystems ist in **Fig.** 1 gezeigt. Trainingseinheit **6** empfängt als Eingabe s(n), einen Satz von digitalisierten Sprachsamples oder -Abtastungen für das zu trainierende Wort oder den Satz bzw. das Satzglied. Das Sprachsignal s(n) wird durch Parameterbestimmungsblock **7** weitergeleitet, der ein Template bzw. eine Vorlage mit N Parametern {p(n) n = 1 ... N} erzeugt, das die Charakteristika bzw. Eigenschaften der Aussprache des Nutzers für das spezielle Wort oder Satzglied einfängt. Parameterbestimmungseinheit **7** kann irgendeine einer Zahl von Sprachparameterbestimmungstechniken implementieren, von denen viele in der Technik wohlbekannt sind. Ein beispielhaftes Ausführungsbeispiel einer Parameterbestimmungstechnik ist der Vocoder-Kodierer, der im U.S. Patent Nr. 5,414,796 beschrieben ist. Ein alternatives Ausführungsbeispiel einer Parameterbestimmungstechnik ist eine Fast Fourier Transformation (FFT), wobei die N Parameter, die N FFT Koeffizienten sind. Andere Ausführungsbeispiele leiten Parameter basierend auf den FFT Koeffizienten ab. Jedes gesprochene Wort oder Satzglied produziert eine Vorlage mit N Parametern, die in einer Vorlagendatenbank (template database) **8** gespeichert wird. Nach dem das Training mit M Vokabularwörtern vollendet ist, enthält die Vorlagendatenbank 8M Vorlagen, von denen jede N Parameter enthält. Vorlagendatenbank **8** wird in einer Art nichtflüchtigen Speicher gespeichert, so dass die Vorlagen resident bzw. gespeichert bleiben, wenn die Leistung bzw. Stromversorgung abgeschaltet wird.

[0005] **Fig.** 2 zeigt ein Blockdiagramm einer Spracherkennungseinheit **10**, die während der Erkennungsphase eines sprecherabhängigen Spracherkennungssystems arbeitet. Spracherkennungseinheit **10** umfasst eine Vorlagendatenbank **14**, die im Allgemeinen die Vorlagendatenbank **8** der Trainingseinheit **6** sein wird. Die Eingabe zur Spracherkennungseinheit **10** ist die digitalisierte Eingabesprache x(n), die die zu erkennende Spra-

che ist. Die Eingangssprache x(n) wird in den Parameterbestimmungsblock **12** weitergegeben, der die gleiche Parameterbestimmungstechnik wie der Parameterbestimmungsblock **7** der Trainingseinheit **6** durchführt. Der Parameterbestimmungsblock **12** erzeugt eine Erkennungsvorlage mit N Parametern {t(n) n = 1 ... N}, die die Charakteristika der Eingangssprache x(n) modeliert. Erkennungsvorlage t(n) wird dann zum Mustervergleichsblock **16** weitergeleitet, der einen Mustervergleich zwischen Vorlage t(n) und allen den in der Vorlagendatenbank **14** gespeicherten Vorlagen durchführt. Die Distanzen bzw. Abstände zwischen Vorlage t(n) und jeder der Vorlagen der Vorlagendatenbank **14** werden zum Entscheidungsblock **18** weitergegeben, der aus der Vorlagendatenbank **14** die Vorlage auswählt, die am nächsten bzw. am besten mit Erkennungsvorlage t(n) übereinstimmt. Die Ausgabe des Entscheidungsblocks **18** ist die Entscheidung darüber welches Wort aus dem Vokabular gesprochen wurde.

[0006] Erkennungsgenauigkeit ist ein Maß dafür wie gut ein Erkennungssystem gesprochene Wörter oder Sätze aus dem Vokabular korrekt erkennt. Beispielsweise gibt eine Erkennungsgenauigkeit von 95% an, dass die Erkennungseinheit Wörter aus dem Vokabular in 95 von 100 Fällen korrekt erkennt. In einem traditionellen Spracherkennungssystem wird die Erkennungsgenauigkeit in der Gegenwart von Rauschen bzw. Geräusch stark herabgesetzt. Der Hauptgrund für diesen Verlust an Genauigkeit ist, dass die Trainingsphase typischerweise in einer ruhigen Umgebung stattfindet, aber die Erkennung typischerweise in einer rausch- bzw. geräuschbehafteten Umgebung stattfindet. Beispielsweise wird ein Freisprechspracherkennungssystem fürs Auto gewöhnlich trainiert, während das Auto in einer Garage steht oder in der Einfahrt geparkt ist, so dass der Motor und die Klimaanlage nicht laufen und die Fenster für gewöhnlich hochgekurbelt bzw. geschlossen sind. Jedoch wird die Erkennung normalerweise verwendet während sich das Auto bewegt, so dass der Motor läuft, Straßen- und Windgeräusche bzw. Rauschen vorhanden sind, die Fenster können unten sein, etc. Als ein Ergebnis der Ungleichheit des Rauschpegels zwischen den Trainings- und Erkennungsphasen bildet die Enkennungsvorlage keine gute Übereinstimmung mit irgendeiner der während dem Training erhaltenen Vorlagen. Das erhöht die Wahrscheinlichkeit eines Erkennungsfehlers oder Versagens.

[0007] **Fig.** 3 erläutert eine Spracherkennungseinheit **20**, die Spracherkennung in der Gegenwart von Rauschen durchführen muß. Wie in **Fig.** 3 gezeigt, addiert ein Summierer **22** zum Sprachsignal x(n) ein Rauschsignal w(n) zum Produzieren eines rauschkorrumpierten bzw. rauschgeschädigten- bzw. geräuschgeschädigten Sprachsignals r(n). Es sollte verstanden werden, dass Summierer **22** nicht ein physikalisches Element des Systems ist, sondern ein Produkt einer lauten bzw. geräuschvollen Umgebung ist. Das rauschgeschädigte Sprachsignal r(n) wird dem Parameterbestimmungsblock **24** eingegeben, der eine rauschgeschädigte Vorlage t1(n) produziert. Mustervergleichsblock **28** vergleicht Vorlage t1(n) mit allen den Vorlagen der Vorlagedatenbank **26**, die in einer ruhigen Umgebung konstruiert wurde. Da die rauschgeschädigte Vorlage t1(n) nicht genau mit irgendeiner der Trainingsvorlagen übereinstimmt, gibt es eine hohe Wahrscheinlichkeit, dass die durch den Entscheidungsblock **30** produzierte Entscheidung ein Erkennungsfehler oder Versagen sein kann.

[0008] Gales M J F et al: "Robust speech recognition in additive and convolutional noise using parallel model combination", Computer Speech and Lanugage, Bd. 9, Nr. 4, 1. Oktober 1995, Seiten 289–307, XP000640904 offenbart ein Verfahren der parallelen Modellkombination (PMC) als eine Technik zum Kompensieren der Effekte von additivem Rauschen bei einem Spracherkenner. In diesem Schriftstück wird das PMC-Schema erweitert, um die Effekte von Faltungsrauschen einzubeziehen. Das wird gemacht durch Einführen einer modifizierten „Fehlanpassungs"-Funktion, die es erlaubt eine Schätzung der Differenz bzw. des Unterschieds der Kanalbedingungen oder Schieflage zwischen Training und Testumgebungen zu machen. Hat man diese Schieflage geschätzt, können Maximum Likelihood (ML) Schätzungen des geschädigten Sprachmodells in der üblichen Weise erhalten werden. Das Schema wird bewertet unter Verwendung der NOISEX-92 Datenbank, wobei die Leistungsfähigkeit bei der Anwesenheit sowohl von störendem additiven Rauschen und Faltungsrauschen nur leichte Degradation zeigt, verglichen mit dem was man erhält, wenn kein Faltungsrauschen vorhanden ist. Hat man die Form der Modelle entschieden, ist es nötig eine Methode zum Schätzen der neuen Modellparameter zu wählen. Die „optimale" Technik für additives Rauschen würde sein Samples des Hintergrundrauschens zu den reinen Trainingsdaten auf dem Wellenformniveau zu addieren. Eine neue, an die Testumgebung angepasste Trainingsdatenbank, könnte dann generiert und ein neuer Satz von Modellen, trainiert werden: Jedoch sollte bemerkt werden, dass um dieses Training durchzuführen das folgende nötig ist.

(1) Die gesamte Trainingsdatenbank ist online verfügbar.

(2) Ausreichende Rauschsamples sind verfügbar zum Addieren zu den reinen Daten.

(3) Rechenleistung ist verfügbar zum Durchführen der Rauschaddition, und Umschulen bzw. erneuten Training der Modellparameter, immer wenn sich das Hintergrundrauschen ändert.

[0009] Angesichts dieser Bedingungen wird es normalerweise unpraktisch sein, diese Art der Kompensation durchzuführen. Jedoch wenn man den reinen Sprachmodellen unterstellt, dass sie ausreichende Information

über die Statistiken der drei Trainingsdaten enthalten, können sie in dem Kompensationsschema anstelle der Daten selbst verwendet werden. Außerdem kann ein Modell des Hintergrundrauschens generiert werden und zwar unter Verwendung was auch immer an Rauschsamples verfügbar ist, um die Hintergrundrauschbedingungen zu repräsentieren. Das Problem ist dann ein Verfahren zum Kombinieren der zwei Modelle zu finden, um die geschädigten Sprachmodelle genau zu schätzen.

[0010] Ferner beschreibt Gong Y: „Speech recognition in noisy environments: A survey", Speech Communication, Bd. 16, Nr. 3, 1. April 1995, Seite 261–291, XP004013163, dass die Leistungsniveaus der meisten aktuellen Spracherkenner signifikant abnehmen, wenn Umgebungsrauschen während der Verwendung auftritt. Solche Leistungsdegradation wird hauptsächlich verursacht durch Fehlanpassungen zwischen Training und Betriebsumgebungen. Während der letzten Jahre wurde viel Aufwand auf das Reduzieren dieser Fehlanpassung gelenkt. Dieses Dokument untersucht Forschungsergebnisse auf dem Gebiet digitaler Techniken für rauschbehaftete Einzelmikrofon-Spracherkennung eingeordnet in drei Kategorien: Rauschwiderstehende Eigenschaften und Ähnlichkeitsmessung, Sprachverbesserung, und Sprachmodellkompensation für Rauschen. Der Überblick zeigt an, das die essentiellen Punkte bei rauschbehafteter Spracherkennung aus dem Verbinden von Zeit- und Frequenzkorrelationen bestehen, Vorsehen höherer Gewichtung für die hohen SNR Anteile der Sprache bei der Entscheidungsfindung, Ausnutzen aufgabenspezifischer a priori Kenntnis, sowohl der Sprache, als auch dem Rauschen, Verwenden klassenabhängiger Verarbeitung und Einbeziehen von Gehörmodellen bei der Sprachverarbeitung. Als ein Sonderfall der Modellkompensation ist eine andere Strategie, das Rauschen zu den Trainingszeichen zu addieren. Mit dieser Technik wird die Fehlanpassung zwischen Training und Betriebsumgebungen komplett verschwinden. Verglichen mit Rauschsubtraktion vom Beobachtungssignal ist das Addieren von Rauschen zu Trainingsdaten einfacher, weil es frei von dem negativen Leistungsspektrumsproblem ist. Verwenden von rauschverunreinigten Daten zum Trainieren eines Systems kann die Erkennungsgenauigkeit bei jener spezifischen Trainingsbedingung dramatisch verbessern. Für festes SNR sind die berichteten Ergebnisse besser als jene von anderen anspruchsvolleren bzw. technisch ausgefeilteren Verarbeitungstechniken wie beispielsweise spektraler Subtraktion, Kalman Filterung und spektraler Transformation. Offenbar jedoch können Techniken basierend auf Rauschaddition zur Sprache den Lombard Effekt nicht verkraften. Spracherkennung im Rauschen bedingt eine große Vielfalt an Fachwissen über jede Verarbeitungsstufe. Aufgrund der komplexen Natur der rauschbehafteten Spracherkennung sind Daten für genaue vergleichende Leistungsauswertung der Techniken nicht verfügbar.

Zusammenfassung der Erfindung

[0011] Die vorliegende Erfindung ist ein System und Verfahren für die automatische Erkennung von gesprochenen Wörter oder Sätzen in Gegenwart von Rauschen bzw. Geräusch, gemäß den unabhängigen Ansprüchen. Bevorzugte Ausführungsbeispiele der vorliegenden Erfindung können den abhängigen Ansprüchen entnommen werden.

[0012] Sprecherabhängige Spracherkennungssysteme arbeiten in zwei Phasen: einer Trainingsphase und einer Erkennungsphase. In der Trainingsphase eines herkömmlichen Spracherkennungssystems wird ein Nutzer aufgefordert, alle der Wörter oder Sätze bzw. Satzteile in einem angegebenen Vokabular zu sprechen. Die digitalisierten Sprachsamples bzw. Abtastungen für jedes Wort oder Satzglied werden zum Erzeugen einer Vorlage von Parametern verarbeitet, die die gesprochenen Wörter charakterisieren. Die Ausgabe der Trainingsphase ist eine Sammlung solcher Vorlagen. In der Erkennungsphase spricht der Nutzer ein spezielles Wort oder Satzglied zum Beginnen einer gewünschten Aktion. Das gesprochene Wort oder Satzglied wird digitalisiert und verarbeitet zum Erzeugen einer Vorlage, die mit allen den, während des Trainings erzeugten Vorlagen verglichen wird. Die nächste Übereinstimmung bestimmt die Aktion, die durchgeführt wird. Die Hauptbeeinträchtigung bzw. Schwäche, die die Genauigkeit von Spracherkennungssystemen begrenzt, ist die Gegenwart von Rauschen. Die Addition von Rauschen während der Erkennung reduziert die Erkennungsgenauigkeit stark, weil dieses Rauschen während dem Training, als die Vorlagedatenbank erzeugt wurde, nicht vorhanden war. Die Erfindung erkennt die Notwendigkeit spezielle Rauschbedingungen zu berücksichtigen, die während der Zeit der Erkennung vorhanden sind, um die Erkennungsgenauigkeit zu verbessern.

[0013] Statt Vorlagen von Parametern zu speichern, speichert das verbesserte Sprachverarbeitungssystem und Verfahren die digitalisierten Sprachsamples für jedes gesprochene Wort oder Satzglied in der Trainingsphase. Die Ausgabe der Trainingsphase ist deshalb eine digitalisierte Sprachdatenbank. In der Erkennungsphase werden die Rauschcharakteristika bzw. Rauscheigenschaften in der Audioumgebung kontinuierlich überwacht. Wenn der Nutzer ein Wort oder Satzglied spricht zum Beginnen der Erkennung wird eine rauschkompensierte Vorlagedatenbank konstruiert und zwar durch Addieren eines Rauschsignals zu jedem der Signale in der Sprachdatenbank und Durchführen von Parameterbestimmung mit jedem der Sprach- plus Rausch-

signale. Ein Ausführungsbeispiel dieses addierten Rauschsignals ist ein künstlich synthetisiertes Rauschsignal mit Eigenschaften ähnlich denen des aktuellen Rauschens. Ein alternatives Ausführungsbeispiel ist eine Aufzeichnung des Rauschzeitfensters, das gerade aufgetreten ist, bevor der Nutzer das Wort oder Satzglied zum Beginnen bzw. Initiieren der Erkennung gesprochen hat. Da die Vorlagendatenbank, unter Verwendung der gleichen Rauschart, die in dem gesprochenen zu erkennenden Wort oder Satzglied vorhanden ist, konstruiert wird, kann die Spracherkennungseinheit eine gute Übereinstimmung zwischen Vorlagen finden, die Erkennungsgenauigkeit verbessert.

Kurze Beschreibung der Zeichnungen

[0014] Die Merkmale, Ziele und Vorteile der vorliegenden Erfindung werden mit der unten angegebenen detaillierten Beschreibung offensichtlicher werden, wenn man diese zusammen mit den Zeichnungen betrachtet, in denen gleich Bezugszeichen durchweg das Gleiche bezeichnen und wobei:

[0015] **Fig.** 1 ist ein Blockdiagramm einer Trainingseinheit eines Spracherkennungssystems;

[0016] **Fig.** 2 ist ein Blockdiagramm einer Spracherkennungseinheit;

[0017] **Fig.** 3 ist ein Blockdiagramm einer Spracherkennungseinheit, die Spracherkennung mit einer durch Rauschen geschädigten Spracheingabe durchführt;

[0018] **Fig.** 4 ist ein Blockdiagramm einer verbesserten Trainingseinheit eines Spracherkennungssystems; und

[0019] **Fig.** 5 ist ein Blockdiagramm einer beispielhaften verbesserten Spracherkennungseinheit.

Detaillierte Beschreibung der bevorzugten Ausführungsbeispiele

[0020] Diese Erfindung liefert ein System und Verfahren zum Verbessern von Spracherkennungsgenauigkeit, wenn Rauschen vorhanden ist. Es macht sich die neuesten Fortschritte in Rechenleistung und Speicherintegration zu Nutze und modifiziert die Trainings- und Erkennungsphasen, um die Anwesenheit von Rauschen während der Erkennung zu berücksichtigen. Die Funktion einer Spracherkennungseinheit ist es die nächste Übereinstimmung mit einer Erkennungsvorlage zu finden, die mit rauschgeschädigter bzw. rauschkorrumpierter Sprache berechnet wird. Da die Charakteristika bzw. Eigenschaften des Rauschens mit Zeit und Ort variieren können, erkennt die Erfindung, dass die beste Zeit die Vorlagendatenbank zu konstruieren während der Erkennungsphase ist.

[0021] **Fig.** 4 zeigt ein Blockdiagramm einer verbesserten Trainingseinheit **40** eines Spracherkennungssystems. Im Gegensatz zu den herkömmlichen, in **Fig.** 1 gezeigten, Trainingsverfahren ist die Trainingseinheit **40** modifiziert, um den Parameterbestimmungsschritt zu eliminieren. Anstelle des Speicherns von Vorlagen von Parametern werden digitalisierte Sprachsamples der aktuellen Wörter und Sätze gespeichert. Somit empfängt die Trainingseinheit **40** als Eingabe Sprachsamples s(n) und speichert digitalisierte Sprachsamples s(n) in einer Sprachdatenbank **42**. Nach dem Training enthält die Sprachdatenbank 42M Sprachsignale, wobei M die Anzahl der Wörter in dem Vokabular ist. Während das bisherige System und Verfahren des Durchführens der Parameterbestimmung Information über die Sprachcharakteristika durch einzig Speichern von Sprachparametern verliert, kann dieses System und Verfahren alle Sprachinformation zur Verwendung in der Erkennungsphase aufbewahren bzw. konservieren.

[0022] **Fig.** 5 zeigt ein Blockdiagramm einer verbesserten Spracherkennungseinheit **50** zur Verwendung zusammen mit Trainingseinheit **40**. Die Eingabe zur Spracherkennungseinheit **50** ist ein rauschkorrumpiertes bzw. rauschgeschädigtes Sprachsignal r(n). Das rauschgeschädigte Sprachsignal r(n) wird vom Summierer **52** generiert durch Addieren des Sprachsignals x(n) mit dem Rauschsignal w(n). Wie zuvor, ist der Summierer **52** nicht ein physikalisches Element des Systems, sondern ein Produkt einer geräuschvollen Umgebung.

[0023] Spracherkennungseinheit **50** umfasst Sprachdatenbank **60**, die die digitalisierten Samples der Sprache bzw. Sprach-Samples enthält, die während der Trainingsphase aufgezeichnet wurden. Spracherkennungseinheit **50** umfasst auch Parameterbestimmungsblock **54**, durch den das rauschgeschädigte Sprachsignal r(n) gereicht bzw. gegeben wird zum Produzieren der rauschgeschädigten Vorlage t1(n). Wie in einem herkömmlichen Spracherkennungssystem kann der Parameterbestimmungsblock **54** irgendwelche einer Anzahl von Sprachparameterbestimmungstechniken implementieren.

[0024] Eine beispielhafte Parameterbestimmungstechnik verwendet linear prädiktive Codierung (linear predictive coding, LPC) Analysetechniken. LPC Analysetechniken modellieren den Vokaltrakt als ein Digitalfilter. Verwendet man LPC Analyse können LPC Cepstral-Koeffizienten c(m) berechnet werden als die Parameter zum Repräsentieren des Sprachsignals. Die Koeffizienten c(m) werden unter Verwendung der folgenden Schritte berechnet. Zuerst wird das rauschgeschädigte Sprachsignal r(n) über einen Rahmen von Sprach-Samples gefenstert (windowed) durch Anwenden einer Fensterfunktion v(n):

$$y(n) = r(n)v(n)\ 0 <= n <= N - 1 \qquad (1)$$

[0025] In dem beispielhaften Ausführungsbeispiel ist die Fensterfunktion v(n) ein Hamming-Fenster und die Rahmengröße N ist gleich 160. Als Nächstes weiden die Autokorrelationskoeffizienten über die gefensterten Samples berechnet unter Verwendung der Gleichung:

$$R(k) = \sum_{m=0}^{N-k} y(m)y(m+k) \qquad k=1,2, \ldots, P \qquad (2)$$

[0026] In dem beispielhaften Ausführungsbeispiel ist P, die Anzahl der zu berechnenden Autokorrelationskoeffizienten, gleich der Ordnung des LPC Prädiktors, die 10 ist. Die LPC Koeffizienten werden dann direkt von den Autokorrelationswerten unter Verwendung von Durbins Rekusionsalgorithmus berechnet. Der Algorithmus kann wie folgt angegeben werden:

1.

$$E^{(0)} = R(0),\ i = 1 \qquad (3)$$

2.

$$k_i = \left\{ R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right\} / E^{(i-1)} \qquad (4)$$

3.

$$\alpha_i^{(i)} = k_i \qquad (5)$$

4.

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}\ 1 <= j <= i - 1 \qquad (6)$$

5.

$$E^{(i)} = (1 - k_i^2)E^{(i-1)} \qquad (7)$$

6.

Falls i < P, dann gehe zu (2) mit i = i + 1. (8)

7. Die endgültige Lösung für die LPC Koeffizienten ist gegeben durch

$$a_j = \alpha_j^{(P)}\ 1 <= j <= P \qquad (9)$$

[0027] Die LPC Koeffizienten werden dann zu den LPC Cepstral-Koeffizienten konvertiert unter Verwendung der folgenden Gleichungen:

$$c(0) = \ln(R(0)) \qquad (10)$$

$$c(m) = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad 1 <= m <= P \qquad (11)$$

$$c(m) = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad m > P \qquad (12)$$

[0028] Es sollte verstanden werden, dass andere Techniken zur Parameterbestimmung anstelle der LPC Cepstral-Koeffizienten verwendet werden können.

[0029] Zusätzlich wird das Signal r(n) zum Sprachdetektionsblock **56** weitergegeben, der die Anwesenheit oder Abwesenheit von Sprache bestimmt. Sprachdetektionsblock **56** kann die Anwesenheit oder Abwesenheit von Sprache bestimmen unter Verwendung irgendeiner einer Anzahl von Techniken. Ein solches Verfahren ist offenbart in dem oben erwähnten U.S. Patent Nr. 5,414,796 mit dem Titel "VARIABLE RATE VOCODER". Diese Technik analysiert den Pegel der Sprachaktivität um die Bestimmung bezüglich der Anwesenheit oder Abwesenheit von Sprache zu machen. Der Pegel der Sprachaktivität basiert auf der Energie des Signals im Vergleich mit der Energieschätzung des Hintergrundrauschens. Zuerst wird die Energie E(n) für jeden Rahmen berechnet, der in einem bevorzugten Ausführungsbeispiel aus 160 Samples zusammengesetzt ist. Die Energieschätzung des Hintergrundrauschens B(n) kann dann berechnet werden unter Verwendung der Gleichungen:

$$B(n) = \min[E(n), 5059644, \max(1{,}00547*B(n-1), B(n-1)+1)] \qquad (13)$$

[0030] Falls B(n) < 160000 werden drei Schwellen unter Verwendung von B(n) wie folgt berechnet:

$$T1(B(n)) = -(5{,}544613 \times 10^{-6})*B^2(n) + 4{,}047152*B(n) + 362 \qquad (14)$$

$$T2(B(n)) = -(1{,}529733 \times 10^{-5})*B^2(n) + 8{,}750045*B(n) + 1136 \qquad (15)$$

$$T3(B(n)) = -(3{,}957050 \times 10^{-5})*B^2(n) + 18{,}89962*B(n) + 3347 \qquad (16)$$

Falls B(n) > 160000 werden die drei Schwellen berechnet als:

$$T1(B(n)) = -(9{,}043945 \times 10^{-8})*B^2(n) + 3{,}535748*B(n) - 62071 \qquad (17)$$

$$T2(B(n)) = -(1{,}986007 \times 10^{-7})*B^2(n) + 4{,}941658*B(n) + 223951 \qquad (18)$$

$$T3(B(n)) = -(4{,}838477 \times 10^{-7})*B^2(n) + 8{,}630020*B(n) + 645864 \qquad (19)$$

[0031] Dieses Sprachdetektionsverfahren zeigt die Anwesenheit von Sprache an, wenn die Energie E(n) größer als die Schwelle T2(B(n)) ist und zeigt die Abwesenheit von Sprache an, wenn die Energie E(n) kleiner als die Schwelle T2(B(n)) ist. In einem alternativen Ausführungsbeispiel kann dieses Verfahren erweitert werden zum Berechnen von Energieschätzungen des Hintergrundrauschens und Schwellen in zwei oder mehr Frequenzbändern. Zusätzlich sollte es verstanden werden, dass die in Gleichungen (13)–(19) gelieferten Werte experimentell bestimmt wurden, und in Abhängigkeit von den Umständen modifiziert werden können.

[0032] Wenn Sprachdetektionsblock **56** bestimmt, dass Sprache abwesend ist, sendet er ein Steuersignal, das Rauschanalyse, Modellierung und Syntheseblock **58** aktiviert. Es sollte bemerkt werden, dass in der Abwesenheit von Sprache das empfangene Signal r(n) das gleiche wie das Rauschsignal w(n) ist.

[0033] Wenn Rauschanalyse, Modellierung und Syntheseblock **58** aktiviert ist, analysiert er die Eigenschaften des Rauschsignals r(n), modelliert es und synthetisiert ein Rauschsignal w1(n) das gleiche Eigenschaften wie das aktuelle Rauschen w(n) hat. Ein beispielhaftes Ausführungsbeispiel zum Durchführen von Rauschanalyse, Modellierung und Synthese ist im U.S. Patent Nr. 5,646,991 offenbart. Dieses Verfahren führt Rauschanalyse durch und zwar durch Weitergeben des Rauschsignals r(n) durch ein Prädiktionsfehlerfilter, das gegeben ist durch:

$$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-i} \qquad (20)$$

wobei P, die Ordnung des Prädiktors, in dem beispielhaften Ausführungsbeispiel 5 ist. Die LPC Koeffizienten a_i, werden wie früher erklärt, unter Verwendung der Gleichungen (1) bis (9) berechnet. Sobald die LPC Koeffizienten erhalten wurden, können synthetisierte Rausch-Samples mit den gleichen spektralen Eigenschaften generiert werden und zwar durch weiterleiten von weißem Rauschen durch das Rauschsynthesefilter, dass gegeben ist durch:

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}} \qquad (21)$$

welches gerade die Umkehrfunktion des zur Rauschanalyse verwendeten Filters ist. Nach dem Anwenden eines Skalierungsfaktors auf jedes der synthetisierten Rausch-Samples, um die bis synthetisierte Rauschenergie gleich der aktuellen Rauschenergie zu machen, ist die Ausgabe das synthetisierte Rauschen w1(n).

[0034] Das synthetisierte Rauschen w1(n) wird zu jedem Satz der digitalisierten Sprach-Samples in der Sprachdatenbank **60** durch den Summierer **62** addiert, zum Erzeugen von Sätzen mit synthetisiertem Rauschen geschädigten Sprach-Samples. Dann wird jeder Satz mit synthetisiertem Rauschen geschädigten Sprach-Samples durch Parameterbestimmungsblock **64** durchgeleitet, der einen Satz von Parametern generiert für jeden Satz mit synthetisiertem Rauschen geschädigten Sprach-Samples unter Verwendung der gleichen Parameterbestimmungstechnik, die im Parameterbestimmungsblock **54** verwendet wird. Parameterbestimmungsblock **54** produziert eine Vorlage von Parametern für jeden Satz der Sprach-Samples und die Vorlagen werden in der rauschkompensierten Vorlagedatenbank **66** gespeichert. Rauschkompensierte Vorlagedatenbank **66** ist ein Satz von Vorlagen, der konstruiert wird als ob herkömmliches Training stattgefunden hätte und zwar bei der gleichen Art von Rauschen das während der Erkennung vorhanden ist. Man beachte, dass es viele mögliche Verfahren zum Erzeugen von geschätztem Rauschen w1(n) zusätzlich zu dem im US-Patent Nr. 5,646,991 offenbarten Verfahren gibt. Ein alternatives Ausführungsbeispiel ist einfach ein Zeitfenster des aktuell vorhandenen Rauschens aufzunehmen und zwar wenn der Nutzer still ist und verwenden dieses Rauschsignals als das geschätzte Rauschen w1(n). Das Zeitfenster des Rauschens, das aufgezeichnet wurde, genau bevor das zu erkennende Wort oder Satzglied gesprochen wurde, ist ein beispielhaftes Ausführungsbeispiel dieses Verfahrens. Noch ein anderes Verfahren ist es, über Verschiedene über eine spezifizierte Dauer erhaltene Rauschfenster zu mitteln.

[0035] Noch Bezug nehmend auf **Fig.** 5 vergleicht Mustervergleichsblock **68** die rauschgeschädigte Vorlage t1(n) mit allen den Vorlagen der rauschkompensierten Vorlagedatenbank **66**. Da die Rauscheffekte mit den Vorlagen der rauschkompensierten Vorlagedatenbank **66** eingeschlossen sind, ist Entscheidungsblock **70** fähig eine gute Übereinstimmung für t1(n) zu finden. Durch Berücksichtigen der Rauscheffekte in dieser Art und Weise wird die Genauigkeit des Spracherkennungssystems verbessert.

[0036] Die vorhergehende Beschreibung der bevorzugten Ausführungsbeispiele ist vorgesehen es einem Fachmann zu ermöglichen die vorliegende Erfindung nachzuvollziehen oder zu verwenden. Die verschiedenen Modifikationen an diesen Ausführungsbeispielen werden dem Fachmann leicht ersichtlich werden und die hierin definierten generischen Prinzipien können ohne die Verwendung erfinderischer Fähigkeit auf andere Ausführungsbeispiele 17932 angewendet werden.

Patentansprüche

1. Ein Spracherkennungssystem, das Folgendes aufweist:
eine Trainingseinheit (**40**) zum Empfangen von Signalen von zu trainierenden Wörtern oder Sätzen, zum Generieren von digitalisierten Samples für jedes der Worte oder Sätze, und zum Speichern der digitalisierten Samples in einer Sprachdatenbank (**42**) und
eine Spracherkennungseinheit (**50**) zum Empfangen eines Eingabesignals, das es zu erkennen gilt, wobei das Eingabesignal durch Rauschen korrumpiert bzw. geschädigt ist, Generieren einer rauschkompensierten Vorlagedatenbank (template data base) (**66**) durch Anwenden der Wirkungen bzw. Effekte des Rauschens auf die digitalisierten Samples der Sprachdatenbank, und Vorsehen eines Spracherkennungsergebnisses für das rauschbeschädigte Eingabesignal, basierend auf der rauschkompensierten Vorlagedatenbank,
wobei die Spracherkennungseinheit (**50**) weiterhin Folgendes aufweist:
eine Sprachdetektiereinheit (**56**) zum Empfangen des rauschgeschädigten Eingabesignals und zum Bestimmen ob Sprache in dem Eingabesignal vorliegt, wobei das Eingabesignal als ein Rauschsignal benannt wird, wenn bestimmt wird, dass keine Sprache in dem Eingabesignal vorliegt; und
eine Rauscheinheit (**58**), die nach Bestimmung, dass Sprache in dem Eingabesignal nicht vorliegt, aktiviert

wird, wobei die Rauscheinheit zum Analysieren des Rauschsignals und Synthetisieren eines synthetisierten Rauschsignals mit Charakteristika des Rauschsignals dient, wobei das synthetisierte Rauschsignal zum Anwenden der Rauscheffekte auf die digitalisierten Samples der Sprachdatenbank dient.

2. Das Spracherkennungssystem nach Anspruch 1, wobei die Sprachdetektiereinheit (**50**) das Vorliegen von Sprache durch Analysieren des Pegels der Sprachaktivität in dem Eingabesignal bestimmt.

3. Das Spracherkennungssystem nach Anspruch 1, wobei die Rauscheinheit (**58**) das synthetisierte Rauschsignal analysiert und synthetisiert mittels einer linearprädiktiven Kodierungstechnik (linear predictive coding (LPC) technique).

4. Das Spracherkennungssystem nach Anspruch 1, wobei das synthetisierte Rauschsignal einem Fenster des Rauschsignals entspricht, das gerade vor dem zu erkennenden Eingabesignal aufgenommen wurde.

5. Das Spracherkennungssystem nach Anspruch 1, wobei das synthetisierte Rauschsignal einem Durchschnitt verschiedener Fenster des Rauschsignals, aufgenommen über eine vorbestimmte Zeitperiode, entspricht.

6. Das Spracherkennungssystem nach einem der vorhergehenden Ansprüche, wobei die Spracherkennungseinheit (**50**) Folgendes aufweist:
eine erste Parameterbestimmungseinheit (**54**) zum Empfangen des rauschgeschädigten Eingabesignals und zum Generieren einer Vorlage bzw. Template von Parametern, die das Eingabesignal gemäß einer vorbestimmten Parameterbestimmungstechnik repräsentiert;
eine zweite Parameterbestimmungseinheit (**64**) zum Empfangen der Sprachdatenbank mit den Rauscheffekten, angewendet auf die digitalisierten Samples, und Generieren der rauschkompensierten Vorlagedatenbank (**66**), gemäß der vorbestimmten Parameterbestimmungstechnik; und
eine Mustervergleichseinheit (**68**) zum Vergleichen der Vorlage von Parametern, die das Eingabesignal repräsentieren, mit den Vorlagen der rauschkompensierten Vorlagedatenbank (**66**), um die beste Übereinstimmung zu bestimmen und hierdurch das Spracherkennungsergebnis zu identifizieren.

7. Das Spracherkennungssystem nach Anspruch 6, wobei die Parameterbestimmungstechnik eine linear prädiktive Kodierungsanalysetechnik (linear predictive coding (LPC) analysis technique) ist.

8. Eine Spracherkennungseinheit eines sprecherabhängigen Spracherkennungssystems zum Erkennen eines Eingabesignals, wobei die Spracherkennungseinheit (**50**) die Wirkungen bzw. Effekte einer verrauschten Umgebung berücksichtigt, wobei die Einheit Folgendes aufweist:
Mittel (**40**) zum Speichern digitalisierter Samples von Wörtern und Sätzen eines Vokabulars in einer Sprachdatenbank (**42**);
Mittel (**52**) zum Anwenden der Rauscheffekte auf die digitalisierten Samples des Vokabulars um rauschgeschädigte digitalisierte Samples des Vokabulars zu generieren;
Mittel (**50**) zum Generieren einer rauschkompensierten Vorlagedatenbank (**66**), basierend auf den rauschgeschädigten digitalisierten Samples; und
Mittel (**68**, **70**) zum Bestimmen eines Spracherkennungsergebnisses für das Eingabesignal, basierend auf der rauschkompensierten Vorlagedatenbank (**66**), wobei die Mittel zum Anwenden der Rauscheffekte Folgendes aufweisen:
Mittel (**56**) zum Bestimmen, ob Sprache im Eingabesignal vorliegt, wobei das Eingabesignal als ein Rauschsignal benannt wird, wenn bestimmt wird, dass Sprache nicht in dem Eingabesignal vorliegt; und
Mittel (**58**) zum Analysieren des Rauschsignals und zum Synthetisieren eines synthetisierten Rauschsignals, wobei das synthetisierte Rauschsignal zu den digitalisierten Samples des Vokabulars addiert wird.

9. Die Spracherkennungseinheit nach Anspruch 8, die weiterhin Folgendes aufweist:
erste Parameterbestimmungsmittel (**54**) zum Empfangen des Eingabesignals und zum Generieren einer Vorlage von Parametern, die das Eingabesignal repräsentieren, und zwar mittels einer vorbestimmten Parameterbestimmungstechnik; und
zweite Parameterbestimmungsmittel (**64**) zum Empfangen der rauschgeschädigten, digitalisierten Samples des Vokabulars und zum Generieren der Vorlagen der rauschkompensierten Vorlagedatenbank (**66**), und zwar gemäß der vorbestimmten Parameterbestimmungstechnik;
wobei die Mittel (**68**, **70**) zum Bestimmen des Spracherkennungsergebnisses die Vorlagen bzw. Templates der rauschkompensierten Vorlagedatenbank vergleicht um die beste Übereinstimmung zu bestimmen und hierdurch das Spracherkennungsergebnis zu identifizieren.

10. Ein Verfahren zur Spracherkennung, das die Effekte bzw. Wirkungen einer verrauschten Umgebung berücksichtigt, wobei das Verfahren die folgenden Schritte aufweist:
Generieren digitalisierter Samples eines jeden trainierten Wortes oder Satzes, wobei jedes der Worte oder Sätze zu einem Vokabular gehört;
Speichern der digitalisierten Samples in einer Sprachdatenbank (**42**); Empfangen eines zu erkennenden Eingabesignals;
Anwenden der Rauscheffekte auf die digitalisierten Samples des Vokabulars um rauschgeschädigte, digitalisierte Samples des Vokabulars zu generieren;
Generieren einer rauschkompensierten Vorlagedatenbank (**66**), basierend auf den rauschgeschädigten, digitalisierten Samples; und
Vorsehen eines Spracherkennungsergebnisses für das rauschgeschädigte Eingabesignal, basierend auf der rauschkompensierten Vorlagedatenbank,
wobei der Schritt des Anwendens der Rauscheffekte die folgenden Schritte aufweist:
Bestimmen, ob Sprache in dem Eingabesignal vorliegt, wobei das Eingabesignal als ein Rauschsignal benannt wird, wenn bestimmt wird, dass keine Sprache in dem Eingabesignal vorliegt; und
Analysieren des Rauschsignals und Synthetisieren eines synthetisierten Rauschsignals, wobei das synthetisierte Rauschsignal zu den digitalisierten Samples des Vokabulars addiert wird, um die rauschgeschädigten, digitalisieren Samples zu generieren.

11. Das Verfahren zur Spracherkennung nach Anspruch 10, das weiterhin die folgenden Schritte aufweist:
Generieren einer Vorlage von Parametern, die das Eingabesignal repräsentieren und zwar gemäß einer vorbestimmten Parameterbestimmungstechnik; und
Generieren von Vorlagen für die rauschkompensierte Vorlagedatenbank gemäß der vorbestimmten Parameterbestimmungstechnik;
wobei der Schritt des Vorsehens eines Spracherkennungsergebnisses die Vorlage der Parameter, was das Eingabesignal repräsentiert, mit den Vorlagen der rauschkompensierten Vorlagedatenbank vergleicht, um die beste Übereinstimmung zu bestimmen, und hierdurch das Spracherkennungsergebnis zu identifizieren.

Es folgen 2 Blatt Zeichnungen

Anhängende Zeichnungen

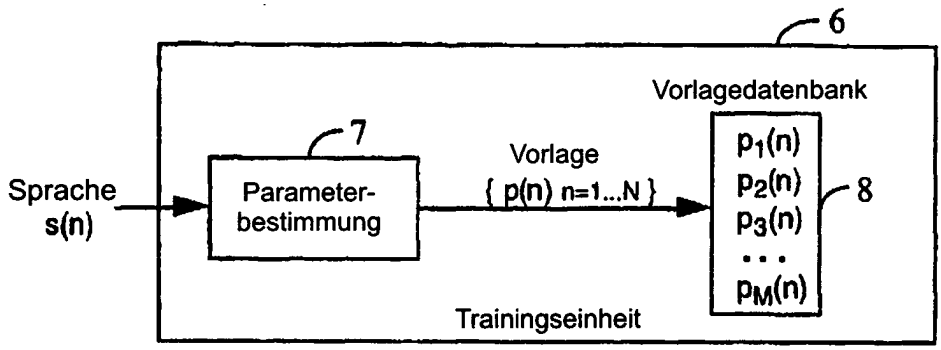

FIG. 1

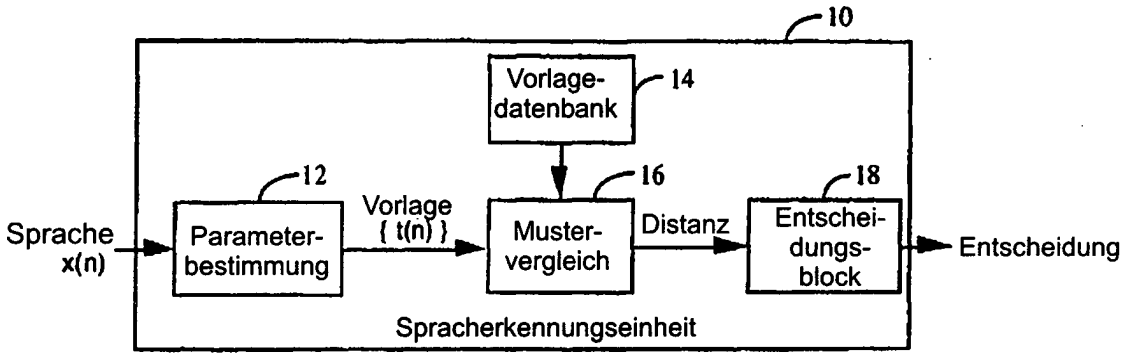

FIG. 2

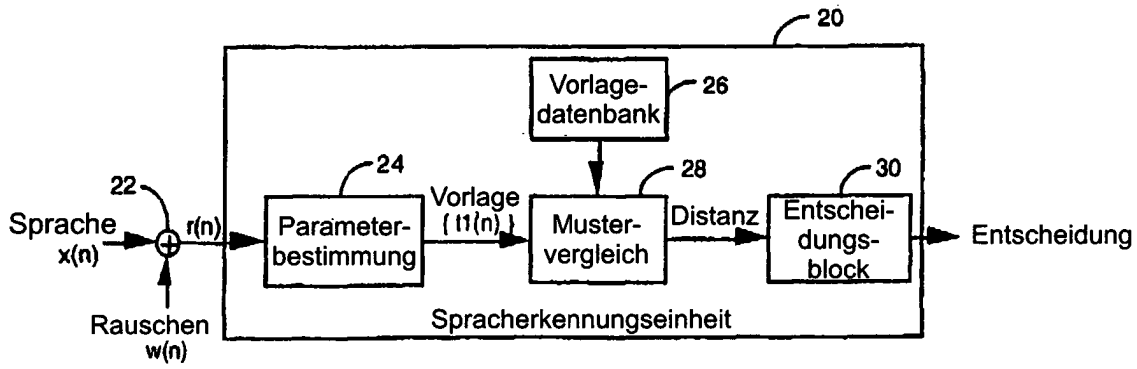

FIG. 3

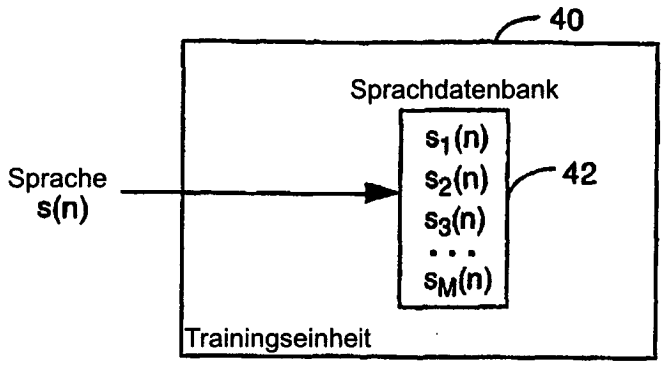
40
Sprachdatenbank
Sprache
s(n)
$s_1(n)$
$s_2(n)$
$s_3(n)$
$s_M(n)$
42
Trainingseinheit

FIG. 4

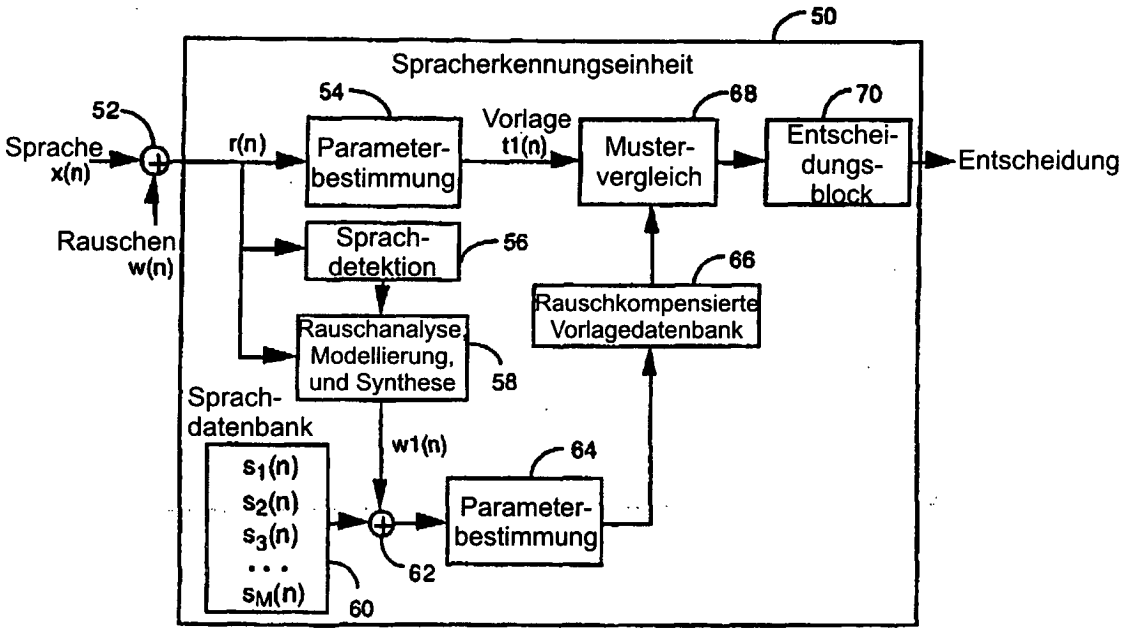
50
Spracherkennungseinheit
52
Sprache
x(n)
Rauschen
w(n)
r(n)
54
Parameter-
bestimmung
Vorlage
t1(n)
68
Muster-
vergleich
70
Entschei-
dungs-
block
Entscheidung
Sprach-
detektion
56
66
Rauschkompensierte
Vorlagedatenbank
Rauschanalyse,
Modellierung,
und Synthese
58
Sprach-
datenbank
w1(n)
64
$s_1(n)$
$s_2(n)$
$s_3(n)$
$s_M(n)$
60
Parameter-
bestimmung
62

FIG. 5