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

(10) **DE 699 20 047 T2** 2005.01.20

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 141 938 B1**
(21) Deutsches Aktenzeichen: **699 20 047.4**
(86) PCT-Aktenzeichen: **PCT/US99/28401**
(96) Europäisches Aktenzeichen: **99 968 458.2**
(87) PCT-Veröffentlichungs-Nr.: **WO 00/33294**
(86) PCT-Anmeldetag: **30.11.1999**
(87) Veröffentlichungstag der PCT-Anmeldung: **08.06.2000**
(97) Erstveröffentlichung durch das EPA: **10.10.2001**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **08.09.2004**
(47) Veröffentlichungstag im Patentblatt: **20.01.2005**

(51) Int Cl.[7]: **G10L 11/02**

(30) Unionspriorität:
201705 30.11.1998 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:
Microsoft Corp., Redmond, Wash., US

(72) Erfinder:
GU, Chuang, Bothell, US; LEE, Ming-Chieh, Bellevue, US; CHEN, Wei-ge, Issaquah, US

(74) Vertreter:
Grünecker, Kinkeldey, Stockmair & Schwanhäusser, 80538 München

(54) Bezeichnung: **DETEKTION VON REINER SPRACHE IN EINEM AUDIO SIGNAL, MIT HILFE EINER DETEKTIONS-GRÖSSE (VALLEY PERCENTAGE)**

Beschreibung

TECHNISCHES GEBIET

[0001] Die Erfindung betrifft die Detektion menschlicher Sprache durch einen Computer, und betrifft insbesondere die Detektion reiner Sprachsignale in einem Audiosignal, das Signale sowohl reiner Sprache als auch gemischter Sprache und Nicht-Sprache enthalten kann.

HINTERGRUND DER ERFINDUNG

[0002] Schall enthält typischerweise ein Gemisch aus Musik, Geräusch und/oder menschlicher Sprache. Die Fähigkeit, menschliche Sprache in Schall zu detektieren, hat wichtige Anwendungen auf vielen Gebieten, wie z. B. digitaler Audiosignal-Verarbeitung, -Analyse und -Codierung. Beispielsweise wurden spezialisierte Codec (Kompressions/Dekompressions-Algorithmen) für eine effizientere Kompression von reinem Schall, welcher entweder Musik oder Sprache, aber nicht beides enthält, entwickelt. Die meisten digitalen Audiosignalanwendungen verwenden daher eine gewisse Form der Sprachdetektion vor einer Anwendung eines spezialisierten Codec, um eine kompaktere Darstellung eines Audiosignals zur Speicherung, Rückgewinnung, Verarbeitung oder Übertragung zu erzielen.

[0003] Jedoch ist eine genaue Detektion von menschlicher Sprache durch einen Computer in einem Audiosignal, das von Schall erzeugt wird, der ein Gemisch von Musik, Geräusch und Sprache enthält, keine einfache Aufgabe. Die meisten existierenden Sprachdetektionsverfahren benutzen spektrale und statistische Analysen der von dem Audiosignal erzeugten Wellenformmuster. Die Herausforderung besteht in der Identifikation von Merkmalen der Wellenformmuster, welche zuverlässig die reinen Sprachsignale von den Nicht-Sprach- oder Gemischt-Sprachsignalen unterscheiden.

[0004] Beispielsweise ziehen einige existierende Verfahren der Sprachdetektion Vorteil aus einem als Null-Durchgangsrate (ZCR) bekannten spezifischen Merkmal. Siehe J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music", Proc. ICASSP'96, pp. 993–996, 1996. Das ZCR Merkmal liefert einen gewichteten Mittelwert der spektralen Energieverteilung in der Wellenform. Menschliche Sprache erzeugt typischerweise Audiosignale mit einer hohen ZCR, während anderer Schall wie Geräusch oder Musik dies nicht tut. Jedoch muß dieses Merkmal nicht immer zuverlässig sein, wie in dem Falle von Schall mit hoch rhythmischer Musik oder strukturiertem Geräusch, welche Audiosignale erzeugen können, welche ZCRs besitzen, die von denen der menschlichen Sprache nicht unterscheidbar sind.

[0005] Weitere existierende Verfahren verwenden mehrere Merkmale einschließlich des ZCR Merkmals in Verbindung mit einer ausgefeilten statistischen Merkmalsanalyse in dem Versuch, die Genauigkeit der Sprachdetektion zu verbessern. Siehe J. D. Hoyt and H. Wechsler, "Detection of Humane Speech Instructured Noise", Proc. ICASSP'94, Vol. II, 237–240, 1994; E. Scheirer and N. Slaney, "Construction and Evaluation of A Robust Multifeature Speech/Music Discriminator", Proc. ICASSP'97, 1997. Ein in dem Scheirer-Literaturhinweis beschriebenes Merkmal ist ein Prozentsatz eines 'Niederenergie'-Rahmenmerkmals – der Anteil von Rahmen mit einer RMS-Leistung kleiner als 50% der mittleren RMS-Leistung innerhalb eines Fensters.

[0006] Obwohl ein großer Forschungsaufwand auf die Detektion menschlicher Sprache fokussiert wurde, erfüllen all diese existierenden Verfahren eines oder mehreren der nachstehend erwünschten Eigenschaften eines Sprachdetektionssystems für moderne Multimediaanwendungen: Hohe Genauigkeit, Robustheit, kurze Verzögerungszeit und geringe Komplexität nicht.

[0007] Hohe Genauigkeit ist in digitalen Audiosignalanwendungen erwünscht, da es wichtig ist, die nahezu "genaue" Zeit zu erfassen, wann die Sprache beginnt und endet, oder die Begrenzungen, mit einer Genauigkeit innerhalb weniger als einer Sekunde. Robustheit ist dahingehend erwünscht, daß das Sprachdetektionssystem Audiosignale verarbeiten kann, welche ein Schallgemisch enthalten, das Geräusch, Musik, Gesang, Konversation, Werbungen, usw. enthält, wovon alle mit unterschiedlichen Raten ohne menschlichen Eingriff abgetastet werden können. Ferner sind die meisten digitalen Audiosignalanwendungen Echtzeitanwendungen. Somit ist es vorteilhaft, wenn das eingesetzte Sprachdetektionsverfahren Ergebnisse innerhalb weniger Sekunden und mit so geringer Komplexität wie möglich für eine Echtzeitimplementation zu vernünftigen Kosten liefert.

ZUSAMMENFASSUNG DER ERFINDUNG

[0008] Die Erfindung stellt ein verbessertes Verfahren zur Detektion menschlicher Sprache in einem Audiosignal bereit. Das Verfahren verwendet ein neues Merkmal des Audiosignals, welches als Tal-Prozentsatz bzw. Valley Percentage-(VP)-Merkmal bezeichnet wird, das reine Sprachsignale von den Nicht-Sprach- und Misch-Sprachsignalen genauer als existierende bekannte Merkmale unterscheidet. Obwohl das Verfahren in Software-Programmodulen implementiert wird, kann es auch in digitaler Hardwarelogik oder in einer Kombination von Hardware- und Softwarekomponenten implementiert werden.

[0009] Eine Implementation des Verfahrens arbeitet mit aufeinanderfolgenden Audioabtastwerten in einem Strom von Abtastwerten, indem eine vorbestimmte Anzahl von Abtastwerten durch ein sich bewegendes Zeitfenster hindurch betrachtet wird. Eine Merkmalberechnungs-Komponente berechnet den Wert des VP bei jedem Zeitpunkt durch Messen der Niederenergieanteile des Audiosignals (des Tals) im Vergleich zu den Hochenergieanteilen des Audiosignals (dem Berg) für einen speziellen Audioabtastwert in Bezug auf die umgebenden Audioabtastwerte in einem gegebenen Fenster. Intuitiv ist der VP wie der Talbereich zwischen Bergen. Der VP ist sehr nützlich bei der Detektion reiner Sprachsignale in Nicht-Sprach- und Gemischt-Sprach-Signalen, da menschliche Sprache tendenziell einen höheren VP als andere Typen von Schall, wie z. B. Musik oder Geräusch, aufweist.

[0010] Nachdem das Anfangsfenster der Abtastwerte verarbeitet ist, wird das Fenster bei dem nächsten folgenden Audioabtastwert in den Strom neu positioniert (auf diesen vorgeschoben). Die Merkmalberechnungs-Komponente wiederholt die Berechnung des VP zu diesem Zeitpunkt unter Verwendung des nächsten Fensters von Audioabtastwerten in dem Strom. Der Prozeß der Neupositionierung und Berechnung wird wiederholt, bis ein VP für jeden Abtastwert in dem Audiosignal berechnet worden ist. Eine Entscheidungsprozessor-Komponente klassifiziert die Audioabtastwerte in Klassifizierungen reiner Sprache oder Nicht-Sprache, indem er die berechneten VP-Werte gegenüber einem VP-Schwellenwert vergleicht.

[0011] In der tatsächlichen Praxis dauert menschliche Sprache üblicherweise wenigstens mehr als einige zusammenhängende Sekunden in realen digitalen Audiodaten an. Somit wird die Genauigkeit der Sprachdetektion verbessert, indem diejenigen isolierten Audioabtastwerte, die als reine Sprache klassifiziert werden, aber deren benachbarten Abtastwerte als Nicht-Sprache und umgekehrt klassifiziert wurden, entfernt werden. Jedoch ist es gleichzeitig erwünscht, die scharfe Begrenzung zwischen den Sprache- und Nicht-Sprache-Segmenten zu bewahren.

[0012] In der Implementation führt eine Nachentscheidungsprozessor-Komponente das Vorstehende durch Anwenden eines Filters auf die binäre Sprachentscheidungsmaske (welche eine Kette von "1"-en und "0"-en enthält), die von der Entscheidungsprozessor-Komponente erzeugt wird, aus. Insbesondere wendet die Nachentscheidungsprozessor-Komponente ein morphologisches Öffnungsfilter gefolgt von einem morphologischen Schließfilter auf die Werte der binären Entscheidungsmaske an. Das Ergebnis ist die Beseitigung aller isolierten reinen Sprache- oder Nicht-Sprache-Maskenwerte (Beseitigung der isolierten "1"-en und "0"-en). Was übrigbleibt ist die gewünschte Sprachdetektionsmaske, welche die Begrenzungen der reinen Sprach- und Nicht-Sprach-Abschnitte des Audiosignals kennzeichnet.

[0013] Implementationen des Verfahrens können weitere Merkmale zum Verbessern der Genauigkeit der Sprachdetektion enthalten. Beispielsweise enthält das Sprachdetektionsverfahren bevorzugt eine Vorprozessor-Komponente, um das Audiosignal zu reinigen, indem ungewolltes Geräusch vor der Berechnung des VP-Merkmals ausgefiltert wird. In einer Implementation reinigt die Vorprozessor-Komponente das Audiosignal, indem zuerst das Audiosignal in eine Energiekomponente umgewandelt wird, und dann ein morphologisches Schließfilter auf die Energiekomponente angewendet wird.

[0014] Das Verfahren implementiert die Detektion menschlicher Sprache effizient in Audiosignalen, welche ein Gemisch aus Musik, Sprache und Geräusch enthalten, unabhängig von der Abtastrate. Für bessere Ergebnisse kann jedoch eine Anzahl von Parametern, welche die Fenstergrößen und die Schwellenwerte steuern, durch das Verfahren implementiert werden. Obwohl es viele Alternativen für die Ermittlung dieser Parameter gibt, werden in einer Implementation, wie z. B. in überwachten digitalen Audiosignalanwendungen die Parameter durch ein Training der Anwendung a priori vorbestimmt. Ein Trainings-Audioabtastwert mit einer bekannten Abtastrate und bekannten Sprachbegrenzungen wird dazu verwendet, um die optimalen Werte der Parameter festzulegen. In anderen Implementationen wie z. B. einer Implementation in einer nicht überwachten Umgebung, ist eine adaptive Ermittlung dieser Parameter möglich.

[0015] Weitere Vorteile und Merkmale der Erfindung werden aus der nachstehenden detaillierten Beschreibung und den beigefügten Zeichnungen ersichtlich.

KURZBESCHREIBUNG DER ZEICHNUNGEN

[0016] **Fig.** 1 ist eine allgemeine Blockdarstellung, welche eine Übersicht einer Implementation des Detektionssystems für menschliche Sprache darstellt.

[0017] **Fig.** 2 ist eine Blockdarstellung, welche eine Implementation der Vorprozessor-Komponente des in **Fig.** 1 gezeigten Systems darstellt.

[0018] **Fig.** 3 ist eine Blockdarstellung, welche eine Implementation der Merkmalberechnungs-Komponente des in **Fig.** 1 gezeigten Systems darstellt.

[0019] **Fig.** 4 ist eine Blockdarstellung, welche eine Implementation der Entscheidungsprozessor-Komponente des in **Fig.** 1 gezeigten Systems darstellt.

[0020] **Fig.** 5 ist eine Blockdarstellung, welche eine Implementation der Nachentscheidungsprozessor-Komponente des in **Fig.** 1 gezeigten Systems darstellt.

[0021] **Fig.** 6 ist eine Blockdarstellung eines Computersystems, welche als eine Betriebsumgebung für eine Implementation der Erfindung dient.

DETAILLIERTE BESCHREIBUNG

Übersicht über ein Verfahren zur Detektion von menschlicher Sprache

[0022] Die nachfolgenden Abschnitte beschreiben ein verbessertes Verfahren zur Detektion von menschlicher Sprache in einem Audiosignal. Das Verfahren nimmt an, daß das Eingangsaudiosignal aus einem aufeinanderfolgenden Strom diskreter Audioabtastwerte mit einer festen Abtastrate besteht. Das Ziel des Verfahrens besteht in der Detektion des Vorhandenseins und der Dauer reiner Sprache in dem Eingangsaudiosignal.

[0023] Schall erzeugt Audiosignale, welche Wellenformmuster mit bestimmten charakteristischen Merkmalen abhängig von der Quelle des Schalls besitzen. Die meisten Sprachdetektionsverfahren ziehen einen Vorteil aus diesem Verhalten, indem sie versuchen zu identifizieren, welche Merkmale zuverlässig menschlichem Sprachschall zugeordnet sind. Im Gegensatz zu anderen Detektionsverfahren für menschliche Sprache, welche existierende bekannte Merkmale verwenden, verwendet dieses verbesserte Verfahren zur Detektion menschlicher Sprache ein als zuverlässig menschlichem Sprachschall zugeordnet erkanntes Merkmal, welches als das Tal-Prozentsatz-(VP)-Merkmal bezeichnet wird.

[0024] Bevor mit der Beschreibung einer Implementation des Sprachdetektionsverfahrens begonnen wird, ist es hilfreich, mit einer Reihe von Definitionen zu beginnen, welche durch den gesamten Rest der Beschreibung hindurch verwendet werden.

Definition 1: Fenster

[0025] Ein Fenster bezeichnet einen aufeinanderfolgenden Strom einer festen Anzahl von diskreten Audioabtastwerten (oder von diesen Audioabtastwerten abgeleiteten Werten). Das Verfahren arbeitet iterativ primär mit dem mittleren Abtastwert, welcher in der Nähe eines Mittelpunktes des Fensters angeordnet ist, jedoch immer in Relation zu den umgebenden Abtastwerten, welche durch das Fenster hindurch zu einem bestimmten Zeitpunkt betrachtet werden. Sobald das Fenster zu dem nächsten folgenden Audioabtastwert neu positioniert (verschoben) wird, wird der Audioabtastwert an dem Beginn des Fensters aus dem Blickfeld entfernt, und ein neuer Audioabtastwert dem Blickfeld an dem Ende des Fensters hinzugefügt. Fenster verschiedener Größen werden zur Lösung verschiedener Aufgaben verwendet. Beispielsweise wird das erste Fenster in der Vorprozessor-Komponente verwendet, um ein morphologisches Filter an die aus den Audioabtastwerten abgeleiteten Energiepegel anzulegen. Ein zweites Fenster wird in der Merkmalberechnungs-Komponente verwendet, um den maximalen Energiepegel innerhalb einer gegebenen Wiederholung des Fensters zu identifizieren. Ein drittes und ein viertes Fenster werden in der Nachentscheidungsprozessor-Komponente verwendet, um entsprechende morphologische Filter an die aus den Audioabtastwerten abgeleitete primäre Sprachentscheidungsmaske anzulegen.

Definition 2: Energiekomponente und Energiepegel

[0026] Die Energiekomponente ist der Absolutwert des Audiosignals. Der Energiepegel bezieht sich auf einen spezifischen Wert der Energiekomponente zum Zeitpunkt t_n, abgeleitet von einem entsprechenden Audioabtastzeitpunkt t_n. Somit gilt, wenn das Audiosignal durch S(t) dargestellt wird, die Abtastwerte zum Zeitpunkt t_n durch $S(t_n)$, dargestellt werden, die Energiekomponente durch I(t) dargestellt wird, die Pegel zum Zeitpunkt t_n durch $I(t_n)$ dargestellt wobei, und wenn $t = (t_1, t_2 \ldots t_n)$ ist:

$$I(t) = \begin{cases} S(t) & S(t) \geq 0 \\ -S(t) & S(t) \leq 0 \end{cases}$$

Definition 3: Binäre Entscheidungsmaske

[0027] Die binäre Entscheidungsmaske ist ein Klassifikationsschema, das zum Klassifizieren eines Wertes entweder in eine binäre 1 oder eine binäre 0 verwendet wird. Somit gilt beispielsweise, wenn die binäre Entscheidungsmaske durch B(t) dargestellt wird, und die binären Werte zum Zeitpunkt t_n als $B(t_n)$ dargestellt werden, und der Talprozentsatz durch VP(t) und die VP-Werte zum Zeitpunkt t_n durch $VP(t_n)$ dargestellt werden, und β einen VP-Schwellenwert darstellt, und wenn $t = (t_1, t_2 .. t_n)$ ist:

$$B(t) = \begin{cases} 1 \quad (Sprache) & VP(t) > \beta \\ 0 \quad (Nicht - Sprache) & VP(t) \leq \beta \end{cases}$$

Definition 4: Morphologisches Filter

[0028] Mathematische Morphologie ist ein leistungsfähiges nicht-lineares Signalverarbeitungswerkzeug, welches verwendet werden kann, um unerwünschte Eigenschaften aus den Eingangsdaten unter Beibehaltung ihrer Begrenzungsinformation zu entfernen. In dem Verfahren der Erfindung wird mathematische Morphologie effizient dazu verwendet, um die Genauigkeit der Sprachdetektion sowohl in der Vorprozessor-Komponente durch Ausfiltern von Geräusch aus dem Audiosignal und in der Nachentscheidungsprozessor-Komponente durch Ausfiltern isolierter binärer Entscheidungsmasken, die sich aus impulsartigen Audioabtastwerten ergeben, eingesetzt.

[0029] Insbesondere besteht das morphologische Schließfilter C(•) aus einem morphologischen Dilatationsoperator D(•), gefolgt von einem Erosionsoperator E(•) mit einem Fenster W. Wenn die Eingangsdaten durch I(t) dargestellt werden und die Datenwerte zum Zeitpunkt t_n durch $I(t_n)$ dargestellt werden und wenn $t = (t_1, t_2 .... t_n)$ ist, gilt:

C(I(t)) = E(D(I(t))) wobei

$$E(I(t)) = \min_i \{I(i(t) \mid t - W \leq i \leq t + W \}$$

$$D(I(t)) = \max_i \{I(i(t) \mid t - W \leq i \leq t + W \}$$

[0030] Das morphologische Öffnungsfilter O(•) besteht aus denselben Operatoren D(•)und E(•), welche aber in umgekehrter Reihenfolge angewendet werden. Somit gilt, wenn die Eingangsdaten durch I(t) dargestellt werden, und die Datenwerte zum Zeitpunkt t_n durch $I(t_n)$ dargestellt werden und wenn $t = (t_1, t_2 .... t_n)$ ist

O(I(t)) = D(E(I(t)))

Beispielimplementation

[0031] Der nachstehende Abschnitt beschreibt eine spezifische Implementation eines Detektionsverfahrens für menschliche Sprache detaillierter. **Fig.** 1 ist eine Blockdarstellung, welche die Hauptkomponenten der nachstehend beschriebenen Implementation darstellt. Jeder von den Blöcken in **Fig.** 1 repräsentiert Programmodule, welche Teile des vorstehend beschriebenen Detektionsverfahrens für menschliche Sprache implementieren. Abhängig von einer Vielzahl von Überlegungen, wie z. B. Kosten, Leistung und Konstruktionskomplexität kann jedes dieser Module auch in eine digitale Logik implementiert werden.

[0032] Unter Verwendung der vorstehend definierten Notation nimmt das in **Fig.** 1 dargestellte Sprachdetektionsverfahren als Eingangsgröße ein Audiosignal S(t) **110** auf. Die Vorprozessor-Komponente **114** reinigt das Audiosignal S(t) **110**, um Geräusche zu entfernen und wandelt es in eine Energiekomponente I(t) **112** um. Die Merkmalberechnungs-Komponente **116** berechnet einen Talprozentsatz VP(t) **118** aus der Energiekomponente I(t) **112** für das Audiosignal S(t) **110**. Die Entscheidungsprozessor-Komponente **120** klassifiziert den sich ergebenden Talprozentsatz VP(t) **118** in eine binäre Sprachentscheidungsmaske B(t) **122**, welche das Audiosignal S(t) **110** entweder als reine Sprache oder Nicht-Sprache identifiziert. Die Nachentscheidungsprozessor-Komponente **124** eliminiert isolierte Werte der binären Sprachentscheidungsmaske B(t) **122**. Das Ergebnis der Nachentscheidungsprozessor-Komponente ist die Sprachdetektionsmaske M(t) **126**.

Vorprozessor-Komponente

[0033] Die Vorprozessor-Komponente **114** des Verfahrens ist detaillierter in **Fig.** 2 dargestellt. In der aktuellen Implementation beginnt die Vorprozessor-Komponente **114** mit der Verarbeitung des Signals S(t) **110** durch Reinigen und Vorbereiten des Audiosignals S(t) **110** für die anschließende Verarbeitung. Insbesondere arbeitet die aktuelle Implementation iterativ an aufeinanderfolgenden Audioabtastwerten $S(t_n)$ **210** in einem Strom von Abtastwerten des Audiosignals S(t) **110** unter Verwendung der Fenstertechnik (wie vorstehend in der Definition 1 definiert). Die Vorprozessor-Komponente **114** beginnt mit der Durchführung des Energieumwandlungsschrittes **215**. In diesem Schritt wird jeder der Audioabtastwerte $S(t_n)$ **210** zum Zeitpunkt t_n in entsprechende Energiepegel $I(t_n)$ **220** zum Zeitpunkt t_n umgewandelt. Die Energiepegel $I(t_n)$ **220** zum Zeitpunkt t_n werden aus dem Absolutwert der Audioabtastwerte $S(t_n)$ **210** zum Zeitpunkt t_n, wobei $t = t_1, t_2, \ldots t_n$ ist, wie folgt aufgebaut:

$$I(t) = \begin{cases} S(t) & S(t) \geq 0 \\ -S(t) & S(t) \leq 0 \end{cases}$$

[0034] Die Vorprozessor-Komponente **114** führt anschließend einen Reinigungsschritt **225** zum Reinigen des Audiosignals S(t) **110** durch Filterung der Energiekomponente I(t) **102** zur Vorbereitung einer weiteren Bearbeitung durch. Bei der Auslegung der Vorprozessor-Komponente ist es zu bevorzugen, ein Reinigungsverfahren auszuwählen, das keine falschen Daten einführt. Die aktuelle Implementation verwendet ein morphologisches Schließfilter C(•) **230**, welches (wie vorstehend in Definition 4 definiert) die Kombination eines morphologischen Dilatationsoperators D(•) **235** gefolgt von einem Erosionsoperator E(•) **240** ist. Der Reinigungsschritt **225** wendet C(•) **230** auf das Eingangsaudiosignal S(t) **110**, indem es bei jedem der Energiepegel $I(t_n)$ **220** arbeitet, die jedem der Audioabtastwerte $S(t_n)$ **210** zum Zeitpunkt t_n entsprechen, indem ein erstes Fenster W_1 **245** mit einer vorbestimmten Größe verwendet wird, wobei $t = t_1, t_2, \ldots t$ ist, wie folgt an:

C(I(t)) = D(E(I(t))) wobei

$$E(I(t)) = \min_i \{I(i(t) \mid t - W_1 \leq i \leq t + W_1\}$$

$$D(I(t)) = \max_i \{I(i(t) \mid t - W_1 \leq i \leq t + W_1\}$$

[0035] Wie man sehen kann, berechnet das Schließfilter C(•) **230** den jeweiligen gefilterten Energiepegel $I'(t_n)$ **250**, indem zuerst die Energiepegel $I(t_n)$ **220** zu einem Zeitpunkt t_n auf die maximalen umgebenden Energiepegel in dem ersten Fenster W_1 **245** dilatiert, und dann die dilatierten Energiepegel auf die minimalen umgebenden Energiepegel in dem ersten Fenster W_1 **225** erodiert werden.

[0036] Das morphologische Schließfilter C(•) **230** entfernt unerwünschtes Geräusch aus dem Eingangsaudiosignal S(t) **110** ohne die Grenzen zwischen den unterschiedlichen Typen des Audioinhaltes zu verwischen. In einer Implementation kann die Anwendung des morphologischen Schließfilters C(•) **230** optimiert werden, indem die Größe des ersten Fensters W_1 **245** so bemessen wird, daß sie dem zu verarbeitenden spezifischen Audiosignal entspricht. In einer typischen Implementation wird die optimale Größe des ersten Fensters W_1 **245** durch ein Training der speziellen Anwendung vorbestimmt, in welchem das Verfahren mit Audiosignalen mit bekannten Sprachcharakteristiken verwendet wird. Demzufolge kann das Sprachdetektionsverfahren effizienter Grenzen zwischen einer Sprache und Nicht-Sprache in einem Audiosignal identifizieren.

Merkmalsberechnung

[0037] In der aktuellen Implementation berechnet, nachdem die Vorprozessor-Komponente das Eingangsaudiosignal S(t) **110** gereinigt hat, die Merkmalberechnungs-Komponente ein Unterscheidungsmerkmal.

[0038] Bei der Implementation einer Komponente zum Berechnen eines Merkmals eines Audiosignals, welches zuverlässig zwischen reiner Sprache von Nicht-Sprache unterscheidet, sind viele Probleme zu lösen. Erstens, welche Komponenten eines Audiosignals sind in der Lage zuverlässig Charakteristiken aufzudecken, welche das reine Sprachsignal von dem Nicht-Sprachsignal unterscheiden? Zweitens, wie kann diese Komponente manipuliert werden, um die Unterscheidungscharakteristik zu quantifizieren? Drittens, wie kann die Manipulation parametrisiert werden, um die Ergebnisse für eine Vielzahl von Audiosignalen zu optimieren?

[0039] Die Literatur bezüglich der Detektion menschlicher Sprache beschreibt eine Vielzahl von Merkmalen, welche verwendet werden können, um menschliche Sprache in einem Audiosignal zu unterscheiden. Beispielsweise verwenden die meisten existierenden Sprachdetektionsverfahren unter anderem Spektralanalyse, Cepstral-Analyse, die vorstehend erwähnte Null-Durchgangsrate, statistische Analyse, Formantenverfolgung, entweder alleine oder in Kombination, um nur einige zu nennen.

[0040] Diese existierenden Verfahren können zufriedenstellende Ergebnisse in einigen digitalen Audiosignalanwendungen liefern, garantieren jedoch kein genaues Ergebnis für eine große Vielzahl von Audiosignalen, die ein Gemisch von Schall, welcher Geräusch, Musik (strukturiertes Geräusch), Gesang, Gespräch, Werbung usw. enthält, welche alle mit unterschiedlichen Raten mit menschlichem Eingriff abgetastet werden können. Die Identifizierung eines zuverlässigen Merkmals ist schwierig, da die Genauigkeit, mit welcher das Audiosignal klassifiziert werden kann, von der Robustheit des Merkmals abhängig ist.

[0041] Bevorzugt hat nach der Durchführung der Merkmalberechnungs- und Entscheidungsprozessor-Komponenten das Sprachdetektionsverfahren alle Audioabtastwerte korrekt unabhängig von der Quelle des Audiosignals klassifiziert. Die Begrenzungen, welche den Start und das Ende von Sprachsignalen in einem Audiosignal angeben, sind von der korrekten Klassifizierung der benachbarten Abtastwerte abhängig, und die korrekte Klassifizierung ist nicht nur von der Zuverlässigkeit des Merkmals sondern auch von der Genauigkeit, mit welcher dieses berechnet wird, abhängig. Daher beeinflusst die Merkmalsberechnung direkt die Fähigkeit zur Detektion von Sprache. Wenn das Merkmal nicht korrekt ist, kann dann auch die Klassifizierung des Audioabtastwertes nicht korrekt sein. Demzufolge sollte die Merkmalberechnungs-Komponente des Verfahrens eine genaue Berechnung eines Unterscheidungsmerkmals liefern.

[0042] Unter Berücksichtigung des Vorstehenden ist es offensichtlich, daß die bestehenden Verfahren sehr schwierig in einer digitalen Echtzeitaudiosignalanwendung nicht nur wegen ihrer Komplexität, sondern auch, weil eine längere Zeitverzögerung zwischen der Eingabe des Audiosignals und der Detektion der Sprache vorliegt, die eine solche Komplexität unvermeidlich mit sich bringt, zu implementieren sein können. Ferner können die vorhandenen Verfahren zu einer Fein-Abstimmung der Sprachdetektionsfähigkeit aufgrund von Einschränkungen der verwendeten Unterscheidungsmerkmale und/oder der Unfähigkeit, die Implementation so zu parametrisieren, daß die Ergebnisse für eine spezielle Quelle des Audiosignals optimiert werden, nicht in der Lage sein. Die aktuelle Implementation einer Merkmalberechnungs-Komponente **116** löst diese Nachteile wie nachstehend detailliert beschrieben.

[0043] Die Merkmalberechnungs-Komponente **116** der aktuellen Implementation ist ferner in **Fig.** 3 dargestellt. Um den Wert des VP(t) **118** für das Eingangsaudiosignal S(t) **110** zu berechnen, berechnet die Merkmalberechnungs-Komponente **116** den Prozentsatz aller Audioabtastwerte $S(t_n)$ **210**, deren gefilterten Energiepegel $I'(t_n)$ **250** zum Zeitpunkt t_n unter einen Schwellenwert-Energiepegel **335** in dem zweiten Fenster W_2 **320** fallen.

[0044] Gemäß der Darstellung in **Fig.** 3 führt die Merkmalberechnungs-Komponente zuerst den Schritt **310** der Identifizierung des maximalen Energiepegels aus, um den maximalen Energiepegel Max **315** zu identifizieren, welcher in dem zweiten Fenster W_2 **320** unter allen gefilterten Energiepegeln $I'(t_n)$ **250** zum Zeitpunkt t_n auftritt. Der Schritt **330** der Energieschwellenwertberechnung berechnet den Schwellenwertenergiepegel **335** durch Multiplizieren des identifizierten maximalen Energiepegels Max **315** mit einem vordefinierten numerischen Faktor α **325**.

[0045] Schließlich berechnet der Schritt **340** der Talprozentsatz-Berechnung den Prozentsatz zum Zeitpunkt t_n, aller in dem zweiten Fenster W_2 **320** auftretenden gefilterten Energiepegel $I'(t_n)$ **250**, die unter den Schwellenwert-Energiepegel **335** fallen. Die sich ergebenden VP-Werte $VP(t_n)$ **345**, die dem jeweiligen Audioabtastwert $S(t_n)$ **210** zum Zeitpunkt t_n entsprechen, werden als das Talprozentsatzmerkmal VP(t) **118** des entsprechenden Audiosignals S(t) **110** bezeichnet.

[0046] Die Berechnung des Talprozentsatzmerkmals VP(t) **118** wird nachstehend unter Verwendung der

nachstehenden Notation ausgedrückt:
I'(t) für die gefilterte Energiekomponente **260**;
W_2 für das zweite Fenster **320**;
Max für den maximalen Energiepegel **315**;
α für den vorbestimmten numerischen Anteil **325**;
N(i) um eine Summierung der Anzahl von Energiepegeln unterhalb des Schwellenwertes anzuzeigen; und
VP(t) für den Talprozentsatz **118**.

$$VP(t) = \frac{\sum_{2W_2+1}^{t+W_2} N(i)}{2W_2+1}; \qquad N(i) = \begin{cases} 1 & I'(t) < \alpha * Max \\ 0 & I'8t) \geq \alpha * Max \end{cases}$$

$$Max = \max_i \{ I'(i) \mid t - W_2 \leq i \leq t + W_2 \}$$

[0047] Die Schritte **310**, **330** und **340** der Merkmalberechnungs-Komponente werden für jeden von den gefilterten Energiepegeln I'(t_n) **250** zum Zeitpunkt t_n wiederholt, indem das zweite Fenster W_2 **320** an jeden der nachfolgenden Audioabtastwerte S(t_{n+1}) **210** zum Zeitpunkt t_{n+1} in dem Eingangsaudiosignal S(t) **110** (und wie in Definition 1 definiert) vorgeschoben wird. Durch Modifizieren der Größe des zweiten Fensters W_2 **320** und des Wertes des numerischen Anteils α **325** kann die Berechnung des VP(t) **118** so optimiert werden, daß sie einer Vielfalt von Quellen von Audiosignalen entspricht.

Entscheidungsprozessor-Komponente

[0048] Die Entscheidungsprozessor-Komponente ist ein Klassifizierungsprozess, welcher direkt auf den von der Merkmalberechnungs-Komponente berechneten VP(t) **118** einwirkt. Die Entscheidungsprozessor-Komponente **120** klassifiziert den berechneten VP(t) **118** in reine Sprache- und Nicht-Sprache-Klassifizierungen durch Aufbauen einer binären Sprachentscheidungsmaske B(t) **122** für den VP(t) **118**, der dem Audiosignal S(t) **110** (siehe Definition der binären Entscheidungsmaske in Definition 3) entspricht.

[0049] **Fig.** 4 ist eine Blockdarstellung, welche den Aufbau der Sprachentscheidungsmaske B(t) **122** aus dem VP(t) **118** weiter veranschaulicht. Insbesondere führt die Entscheidungsprozessor-Komponente **120** einen binären Klassifizierungsschritt **420** aus, welcher jeden der VP-Werte VP(t_n) **345** zu einem Zeitpunkt t_n mit einem Schwellenwert-Talprozentsatz β **410** vergleicht. Wenn einer der VP-Werte VP(t_n) **345** zum Zeitpunkt t_n kleiner oder gleich dem Schwellenwert-Talprozentsatz β **410** ist, wird der entsprechende Wert der Sprachentscheidungsmaske B(t_n) **430** zum Zeitpunkt t_n gleich einem binären Wert "0" gesetzt. Wenn einer der VP-Werte VP(t_n) **345** zum Zeitpunkt t_n größer als der Schwellenwert-Talprozentsatz β **410** ist, wird der entsprechende Wert der Sprachentscheidungsmaske B(t_n) **430** zum Zeitpunkt t_n gleich dem binären Wert "1" gesetzt.

[0050] Die Klassifizierung des Talprozentsatzmerkmals VP(t) **118** in die binäre Sprachentscheidungsmaske B(t) **122** wird nachstehend unter Verwendung der folgenden Notation ausgedrückt:
VP(t) für den Talprozentsatz **118**;
B(t) für die binäre Sprachentscheidungsmaske **122**; und
β für den Schwellenwert-Prozentsatz **410**:

$$B(t) = \begin{cases} 1 & (Sprache) & VP(t) > \beta \\ 0 & (Nicht - Sprache) & VP(t) \leq \beta \end{cases}$$

[0051] Die Entscheidungsprozessor-Komponente **120** wiederholt den binären Klassifizierungsschritt **420**, bis alle VP-Werte VP(t_n) **345**, die dem jeweiligen Audioabtastwert S(t_n) **210** zum Zeitpunkt t_n entsprechen, entweder als reine Sprache oder Nicht-Sprache klassifiziert worden sind. Die sich ergebende Kette von binären Entscheidungsmasken B(t_n) **430** zum Zeitpunkt t_n wird als die Sprachentscheidungsmaske B(t) **122** des Audiosignals S(t) **110** bezeichnet. Der binäre Klassifizierungsschritt **420** kann durch Veränderung des Schwellenwert-Talprozentsatzes β **410** zur Anpassung an eine breite Vielfalt von Quellen des Audiosignals S(t) **110** optimiert werden.

[0052] Sobald die Entscheidungsprozessor-Komponente **120** die binäre Sprachentscheidungsmaske B(t) **122** für das Audiosignal S(t) **110** erzeugt hat, scheint es, als ob nur noch wenig zu tun wäre. Jedoch kann, wie vorstehend angemerkt, die Genauigkeit der Sprachdetektion weiter verbessert werden, indem der Nicht-Sprachklassifikation diejenigen isolierten Audioabtastwerte angepasst werden, die als reine Sprache klassifiziert werden, deren benachbarten Abtastwerte aber als Nicht-Sprache klassifiziert sind und umgekehrt. Dieses ergibt sich aus der vorstehend angegebenen Beobachtung, daß menschliche Sprache üblicherweise wenigstens für mehr als einige wenige zusammenhängende Sekunden in der Realität andauert.

[0053] Die Nachentscheidungsprozessor-Komponente **124** der aktuellen Implementation zieht einen Vorteil aus dieser Beobachtung, indem er ein auf die von der Entscheidungsprozessor-Komponente **120** erzeugte Sprachdetektionsmaske ein Filter anwendet. Anderenfalls wird die sich ergebende binäre Sprachentscheidungsmaske B(t) **122** wahrscheinlich mit anormalen kleinen isolierten "Aussetzern" oder "Spitzen" abhängig von der Qualität des Eingabeaudiosignals S(t) **110** übersät, und dadurch das Resultat für einige digitale Audiosignalanwendungen möglicherweise nutzlos gemacht.

[0054] Wie es in der aktuellen Implementation des in der Vorprozessor-Komponente **114** vorhandenen Reinigungsfilters beschrieben wird, verwendet die aktuelle Implementation des Nachentscheidungsprozessors auch eine morphologische Filtration, um bessere Ergebnisse zu erzielen. Insbesondere wendet die aktuelle Implementation zwei morphologische Filter in Aufeinanderfolge an, um den individuellen Sprachentscheidungsmaskenwert $B(t_n)$ **430** an seinen benachbarten Sprachentscheidungsmaskenwert $B(t_{n\pm1})$ zum Zeitpunkt t_n anzupassen (was die isolierten "1"-en und "0"-en beseitigt), während gleichzeitig die scharfe Begrenzung zwischen den reinen Sprach- und Nicht-Sprach-Abtastwerten erhalten bleibt. Ein Filter ist das morphologische Schließfilter C(•) **560**, ähnlich dem vorstehend beschriebenen Schließfilter **230** in der Vorprozessor-Komponente **114** (und wie es ferner in Definition 4 definiert ist). Das andere Filter ist das morphologische Öffnungsfilter O(•) **520**, welches dem Schließfilter **560** mit der Ausnahme ähnlich ist, daß die Erosions- und Dilatations-Operatoren in umgekehrter Reihenfolge – der Erosions-Operator zuerst, gefolgt von dem Dilatations-Operator als zweiter (und wie es ferner in Definition 4 definiert ist) angewendet werden. Gemäß **Fig.** 5 führt die Nachentscheidungsprozessor-Komponente den Filteranwendungsschritt **510** durch, welcher das morphologische Öffnungsfilter O(•) **520** auf jeden von den binären Sprachentscheidungsmaskenwerten $B(t_n)$ **430** zum Zeitpunkt t_n unter Verwendung eines dritten Fensters W_3 **540** mit einer vorbestimmten Größe durchführt:

O(B(t)) = D(E(B(t))) wobei

$$E(D(B(t))) = \min_i \{B(i(t) \mid t - W_3 \le i \le t + W_3 \}$$

$$D(B(t)) = \max_i \{B(i(t) \mid t - W_3 \le i \le t + W_3 \}$$

[0055] Wie man sehen kann, berechnet das morphologische Öffnungsfilter O(•) **520** den "geöffneten" Wert der binären Sprachentscheidungsmaske B(t) **122**, indem zuerst der Erosions-Operator E **525** und dann der Dilatations-Operator D **530** auf den binären Sprachentscheidungsmaskenwert $B(t_n)$ **430** zum Zeitpunkt t_n angewendet wird. Der Erosions-Operator E **535** erodiert den binären Entscheidungsmaskenwert $B(t_n)$ **430** zum Zeitpunkt t_n auf die minimalen umgebenden Maskenwerte in dem dritten Fenster W_3 **540**. Der Dilatations-Operator D **530** dilatiert den erodierten Entscheidungsmaskenwert $B(t_n)$ **430** zum Zeitpunkt t_n auf die maximalen umgebenden Maskenwerte in dem dritten Fenster W_3 **540**.

[0056] Die Nachentscheidungsprozessor-Komponente wendet dann das morphologische Schließfilter C(•) **560** auf jeden "geöffneten" binären Sprachentscheidungsmaskenwert $O(B(t_n))$ zum Zeitpunkt t_n unter Verwendung eines vierten Fensters W_4 **580** mit vorbestimmter Größe an:

C(O(B(t))) = E(D(O(B(t)))) wobei

$$D(O(B(t))) = \max_i \{O(B(i(t)) \mid t - W_4 \le i \le t + W_4 \}$$

$$E(D(O(B(t)))) = \min_i \{D(O(B(i(t))) \mid t - W_4 \le i \le t + W_4 \}$$

[0057] Wie man sehen kann, berechnet das morphologische Schließfilter C(•) **560** den "geschlossenen" Wert

der binären Sprachentscheidungsmaske B(t) **122**, indem zuerst der Dilatations-Operator D **530** und dann der Erosions-Operator E **525** auf den binären Sprachentscheidungsmaskenwert $B(t_n)$ **430** zum Zeitpunkt t_n angewendet wird. Der Dilatations-Operator D **565** dilatiert den "geöffneten" binären Entscheidungsmaskenwert $B(t_n)$ **430** zum Zeitpunkt t_n auf die maximalen umgebenden Maskenwerte in dem vierten Fenster W_4 **580**. Der Erosions-Operator E **570** erodiert den "geöffneten" binären Entscheidungsmaskenwert $B(t_n)$ **430** zum Zeitpunkt t_n auf die minimalen umgebenden Maskenwerte in dem vierten Fenster W_4 **580**.

[0058] Das Ergebnis der Durchführung der Nachentscheidungsprozessor-Komponente **124** ist die endgültige Abschätzung der binären Sprachdetektionsmaskenwerte $M(t_n)$ **590**, welche jedem Audioabtastwert $S(t_n)$ **210** zum Zeitpunkt t_n wie nachstehend ausgedrückt entsprechen:

$$M(t) = C(O(B(t)))$$

[0059] Durch die Anwendung morphologischer Filter, wie es in der Nachentscheidungsprozessor-Komponente beschrieben wird, können Abweichungen des Audiosignals S(t) **110** an benachbarte Abschnitte des Signals angeglichen werden, ohne die Begrenzungen der reinen Sprache und Nicht-Sprache zu verwischen. Das Ergebnis ist eine genaue Sprachdetektionsmaske M(t) **126**, welche die Start- und Stoppbegrenzungen menschlicher Sprache in dem Audiosignal S(t) **110** anzeigt. Ferner können die von der Nachentscheidungsprozessor-Komponente angewendeten morphologischen Filter optimiert werden, indem die Größe des dritten Fensters W_3 **540** und des vierten Fensters W_4 **580** angepasst werden, so daß sie dem zu verarbeitenden speziellen Audiosignal genügen. In einer typischen Implementation wird die optimale Größe des dritten Fensters W_3 **540** und des vierten Fensters W_4 **580** durch ein Training der speziellen Anwendung vorbestimmt, in welchem das Verfahren mit Audiosignalen mit bekannten Spracheigenschaften verwendet wird. Demzufolge kann das Sprachdetektionsverfahren effektiver die Begrenzungen von reinen Sprach und Nicht-Sprach-Signalen in einem Audiosignal S(t) **110** identifizieren.

Parametereinstellungen

[0060] Wie darauf im Hintergrundabschnitt angespielt, betrifft die Detektion menschlicher Sprache in einem Audiosignal eine digitale Audiokompression, da Audiosignale typischerweise reine Sprach- und Nicht-Sprach- oder Gemischt-Sprach-Signale enthalten. So wie die spezialisierten Sprach-Codec's ein reines Sprach-Signal genauer als ein Nicht-Sprach- oder ein Gemischt-Sprach-Signal komprimieren, detektiert die vorliegende Erfindung menschliche Sprache genauer in einem Audiosignal, welches vorverarbeitet oder gefiltert wurde, um ein Geräusch zu entfernen, als eines, welches nicht vorverarbeitet oder gefiltert wurde. Für die Zwecke dieser Erfindung ist das für die Vorverarbeitung oder Filterung von Geräusch aus dem Audiosignal angewendete genaue Verfahren nicht wichtig. Tatsächlich ist das für die Detektion von menschlicher Sprache in einem Audiosignal hierin beschriebene und nachstehend beanspruchte Verfahren unabhängig von der spezifischen Implementation der Geräuschreduzierung. In dem Zusammenhang der Erfindung kann es, obwohl es keine Rolle spielt, ob Geräusch vorhanden ist, dieses die Einstellung der in dem Verfahren implementierten Parameter verändern.

[0061] Wie im Hintergrundabschnitt angemerkt, sollte die Einstellung der Parameter für Fenstergrößen und Schwellenwerte so gewählt werden, daß die Genauigkeit der Detektion von reiner Sprache optimiert wird. In einer besseren Implementation ist die Genauigkeitsdetektion von reiner Sprache wenigstens 95%.

[0062] In einer Implementation können die Parameter durch Training ermittelt werden. Für das Trainings-Audiosignal sind die tatsächlichen Grenzen der reinen Sprach- und Nicht-Sprach-Abtastwerte bekannt, was hier als das ideale Ausgangssignal bezeichnet wird. Somit werden die Parameter für das ideale Ausgangssignal optimiert.

[0063] Beispielsweise werde angenommen, daß das ideale Ausgangssignal M(t) ist, und eine vollständige Suche in dem Parameterraum (W_1, W_2, W_3, W_4, α, β) zu der Einstellung dieser Werte führt:

$$\min_{W_1,W_2,W_3,W_4,\alpha,\beta} = \{\|M(t) - M(1t),(W_1,W_2,W_3,W_4,\alpha,\beta\ \}\|$$

[0064] Unter der weiteren Annahme, daß das von der speziellen Tonquelle erzeugte Trainingsaudiosignal eine Abtastrate von F kHz hat, ist die optimale Beziehung der Parameter für die und die Abtastrate nachstehend dargestellt.
$W_1 = 40*F/8$,
$W_2 = 2000*F/8$,

W_3 = 24*F/8,
W_4 = 32000*F/8, und
α = 10% und
β = 10%

Kurze Übersicht über ein Computersystem

[0065] **Fig.** 6 und die nachstehende Diskussion sind dafür gedacht, eine kurze, allgemeine Beschreibung einer geeigneten Berechnungsumgebung zu geben, in welcher die Erfindung implementiert werden kann. Obwohl die Erfindung oder ihre Aspekte in einer Hardwarevorrichtung implementiert werden können, ist das vorstehend beschriebene Nachführungssystem in Computer-ausführbaren Befehlen, die in Programmodulen organisiert sind, implementiert. Die Programmodule enthalten Routinen, Programme, Objekte, Komponenten und Datenstrukturen, welche die Aufgaben durchführen und die vorstehend beschriebenen Datentypen implementieren.

[0066] Obwohl **Fig.** 6 eine typische Konfiguration eines Desktop-Computers darstellt, kann die Erfindung in anderen Computersystemkonfigurationen einschließlich Handgeräten, Multiprozessorsystemen, in Mikroprozessor-basierender oder programmierbarer Konsumelektronik, Minicomputern, Großcomputern und dergleichen implementiert werden. Die Erfindung kann auch in verteilten Computerumgebungen angewendet werden, in welchen Aufgaben von entfernten Verarbeitungseinrichtungen ausgeführt werden, die über ein Kommunikationsnetz verknüpft sind. In einer verteilten Computerumgebung können Programmodule sowohl in lokalen als auch entfernten Speichervorrichtungen angeordnet sein.

[0067] **Fig.** 6 veranschaulicht ein Beispiel eines Computersystems, das als eine Betriebsumgebung für die Erfindung dient. Das Computersystem umfasst einen Personal Computer **620**, welcher eine Verarbeitungseinheit **621**, einen Systemspeicher **622** und einen Systembus **623**, der verschiedene Systemkomponenten einschließlich des Systemspeichers mit der Verarbeitungseinheit **621** verbindet, enthält. Der Systembus kann irgendeine von verschiedenen Typen von Busstrukturen einschließlich eines Speicherbusses oder einer Speichersteuerung eines Peripheriebusses und eines lokalen Busses unter Verwendung einer Busarchitektur, wie z. B. PCI, VESA, Microchannel (MCA), ISA und EISA enthalten, um nur einige wenige zu nennen. Der Systemspeicher umfasst einen Nur-Lese-Speicher (ROM) **624** und einen Speicher mit wahlfreiem Zugriff (RAM) **625**. Ein BasisEingabe/Ausgabe-System **626** (BIOS), das die Grundroutinen enthält, welche zur Übertragung von Information zwischen Elementen innerhalb des Personal Computers **620** beitragen, wie z. B. während des Startvorgangs, ist im ROM **624** gespeichert. Der Personal Computer **620** enthält ferner ein Festplattenlaufwerk **627**, ein Magnetplattenlaufwerk **628**, um beispielsweise von einer entfernbaren Platte **629** zu lesen oder darauf zu schreiben, und ein optisches Plattenlaufwerk **630**, z. B. zum Lesen einer CD-ROM-Disk **631** oder zum Lesen oder Schreiben an ein anderes Medium. Das Festplattenlaufwerk **627**, Magnetplattenlaufwerk **628** und das optische Plattenlaufwerk **630** sind mit dem Systembus **623** über eine Festplattenlaufwerk-Schnittstelle **632**, eine Magnetplattenlaufwerk-Schnittstelle **633** bzw. eine optische Plattenlaufwerk-Schnittstelle **634** verbunden. Die Laufwerke und deren zugeordneten Computer-lesbaren Medien stellen eine nicht flüchtige Speicherung von Daten, Datenstrukturen, Computer-ausführbaren Befehlen (Programmcode, wie z. B. dynamische Verknüpfungsbibliotheken und ausführbare Dateien) usw. für den Personal Computer **620** bereit. Obwohl sich die Beschreibung eines Computer-lesbares Mediums vorstehend auf eine Festplatte, eine entfernbare Magnetplatte und eine CD bezieht, kann dieses auch andere Medientypen umfassen, welche von einem Computer lesbar sind, wie z. B. Magnetkassetten, Flash-Speicherkarten, digitale Videoplatten, Bernoulli-Kassetten und dergleichen.

[0068] Eine Anzahl von Programmodulen kann in den Laufwerken und dem RAM **625** einschließlich eines Betriebssystems **635**, eines oder mehrerer Anwendungsprogramme **636**, weiterer Programmodule **637** und Programmdaten **638** gespeichert sein. Ein Benutzer kann Befehle und Information in dem Personal Computer **620** über eine Tastatur **640** und eine Zeigevorrichtung, wie z. B. eine Maus **642**, eingeben. Weitere (nicht dargestellte) Eingabevorrichtungen können ein Mikrophon, einen Joystick, ein Game-Pad, eine Satellitenschüssel, einen Scanner oder dergleichen umfassen. Diese und weitere Eingabevorrichtungen sind oft mit der Verarbeitungseinheit **621** über eine serielle Schnittstelle **646** verbunden, die mit dem Systembus gekoppelt ist, können jedoch auch über andere Schnittstellen wie z. B. einem Parallelanschluss, Spieleanschluss oder einen universellen seriellen Bus (USB) verbunden sein. Ein Monitor **647** oder anderer Typ von Anzeigevorrichtung ist ebenfalls mit dem Systembus **623** über eine Schnittstelle, wie z. B. eine Anzeigensteuerung oder einen Videoadapter **648**, verbunden. Zusätzlich zu dem Monitor enthalten Personal Computer typischerweise weitere (nicht dargestellte) Peripherieausgabevorrichtungen, wie z. B. Lautsprecher und Drucker.

[0069] Der Personal Computer **620** kann in einer Netzwerkumgebung unter Verwendung logischer Verbindungen zu einem oder mehreren entfernten Computern, wie z. B. einem entfernten Computer **659** betrieben, werden. Der entfernte Computer **659** kann ein Server, ein Router, ein gleichrangiges Gerät, oder ein weitere üblicher Netzwerkknoten sein, und enthält typischerweise viele oder alle von den bezüglich des Personal Computers **620** beschriebenen Elemente, obwohl nur eine Massenspeichervorrichtung **50** in **Fig.** 5 dargestellt worden ist. Die in **Fig.** 5 dargestellten logischen Verbindungen umfassen ein lokales Netzwerk (LAN) **651** und ein Weitbereichsnetz (WAN) **652**. Derartige Netzwerkumgebungen sind in Büros, unternehmensweiten Computernetzen, Intranets und im Internet üblich.

[0070] Wenn er in einer LAN-Netzwerkumgebung verwendet wird, ist der Personal Computer **620** mit dem lokalen Netzwerk **651** durch eine Netzwerkschnittstelle oder Adapter **653** verbunden. Wenn er in einer WAN-Netzwerkumgebung eingesetzt wird, enthält der Personal Computer **620** typischerweise ein Modem **654** oder andere Einrichtungen für den Aufbau von Kommunikationen über das Weitbereichsnetz **652**, wie z. B. das Internet. Das Modem **654**, welches intern oder extern vorliegen kann, ist mit dem Systembus **623** über die serielle Schnittstelle **646** verbunden. In einer Netzwerkumgebung können bezüglich des Personal Computers **620** dargestellte Programmodule oder Abschnitte davon in der entfernten Speichervorrichtung gespeichert sein. Die dargestellten Netzwerkverbindungen sind lediglich Beispiele und andere Einrichtungen für den Aufbau einer Kommunikationsverbindung zwischen den Computern können verwendet werden.

[0071] Im Hinblick auf die vielen möglichen Anwendungen, auf welche die Prinzipien unserer Erfindung angewendet werden können, betonen wird, daß die vorstehend beschriebenen Implementationen nur Beispiele der Erfindung sind, und nicht als eine Einschränkung des Schutzumfangs der Erfindung angesehen werden sollten. Statt dessen ist der Schutzumfang der Erfindung durch die nachstehenden Ansprüche definiert. Wir beanspruchen daher als unsere Erfindung alles, was in den Schutzumfang dieser Ansprüche fällt.

Patentansprüche

1. Verfahren zur Detektion von Sprache in einem Audiosignal mit einem Gemisch aus Sprache und Nicht-Sprache, wobei das Verfahren umfasst:
Berechnen eines Sprachdetektionsmerkmals aus dem Audiosignal, wobei das Sprachdetektionsmerkmal für einen Abtastwert in dem Audiosignal einen Anteil von mehreren umgebenden Abtastwerten, welche umgebende Abtastwerte mit niedriger Energie sind, repräsentiert, wobei ein umgebender Abtastwert mit niedriger Energie einen Energiepegel besitzt, welcher unter einen für die mehreren umgebenden Abtastwerte berechneten Schwellenenergiewert fällt;
Klassifizieren des Abtastwertes in dem Audiosignal, entweder in eine Sprach- oder Nicht-Sprach-Klassifizierung gemäß dem Sprachdetektionsmerkmal; und
Ermitteln einer Begrenzung zwischen einem Abschnitt des als Sprache klassifizierten Audiosignals und einem Abschnitt des als Nicht-Sprache klassifizierten Audiosignals basierend wenigstens zum Teil auf einer Vielzahl von Klassifizierungen.

2. Verfahren nach Anspruch 1, welches ferner umfasst:
vor dem Berechnen, Filtern des Audiosignals, um das Audiosignal zu reinigen, wobei Begrenzungsunterscheidungen in dem Audiosignal erhalten bleiben.

3. Verfahren nach Anspruch 2, wobei die Filterung ein Schließfilter verwendet, das einen Dilatations-Operator gefolgt von einem Erosions-Operator umfasst.

4. Verfahren nach Anspruch 1, welches ferner umfasst:
vor dem Berechnen, Umwandeln des Audiosignals in eine Energiekomponente mit einer Vielzahl von Energiepegeln, wobei jeder Energiepegel einem Audioabtastwert des Audiosignals entspricht.

5. Verfahren nach Anspruch 4, wobei die Energiekomponente des Audiosignals aufgebaut wird, indem jedem Energiepegel der Energiekomponente der Absolutwert des entsprechenden Audioabtastwertes des Audiosignals zugeordnet wird.

6. Verfahren nach Anspruch 4 oder 5, welches ferner umfasst:
vor dem Berechnen, Filtern des Audiosignals, um das Audiosignal unter Erhaltung von Begrenzungsunterscheidungen in dem Audiosignal zu reinigen, wobei die Filterung das Anlegen eines morphologischen Schließfilters an jeden Energiepegel der Energiekomponente umfasst, um eine gefilterte Energiekomponente des Audiosignals zu erzeugen.

7. Verfahren nach einem der Ansprüche 1–6, wobei die Berechnung des Sprachdetektionsmerkmals umfasst:
Ermitteln eines maximalen Energiepegels in den mehreren umgebenden Abtastwerten;
Berechnen des Schwellenwertenergiepegels als einen Anteil des maximalen Energiepegels; und
Festlegen des Sprachdetektionsmerkmals auf der Basis eines Prozentsatzes der mehreren umgebenden Abtastwerte, die einen Energiepegel besitzen, welcher unter den Schwellenwert-Energiepegel fällt.

8. Verfahren nach einem der Ansprüche 1–6, wobei die Klassifizierung auf dem Vergleich des berechneten Sprachdetektionsmerkmals mit einem Sprachdetektionsmerkmal-Schwellenwert beruht.

9. Verfahren nach einem der Ansprüche 1–8, wobei die Klassifizierung das Zuweisen eines binären Wertes zu einer Sprachentscheidungsmaske umfasst, um das Vorhandensein von Nicht-Sprache oder Sprache anzuzeigen.

10. Verfahren nach einem der Ansprüche 1 bis 9, welches ferner umfasst:
Filtern der Vielzahl von Klassifikationen, um isolierte Klassifikationen zu entfernen, wobei eine isolierte Klassifikation einen Wert besitzt, der sich von einem vorherrschenden Wert für umgebende Klassifikationen unterscheidet, und wobei die Filterung der Vielzahl von Klassifikationen eine oder mehrere morphologische Filter verwendet.

11. Verfahren nach Anspruch 10, wobei die Filterung der Vielzahl von Klassifikationen ein Öffnungsfilter gefolgt von einem Schließfilter verwendet.

12. Verfahren nach einem der Ansprüche 1–11, welches ferner das Wiederholen der Berechnung des Sprachdetektionsmerkmals für einen oder mehrere weitere Abtastwerte in dem Audiosignal umfasst.

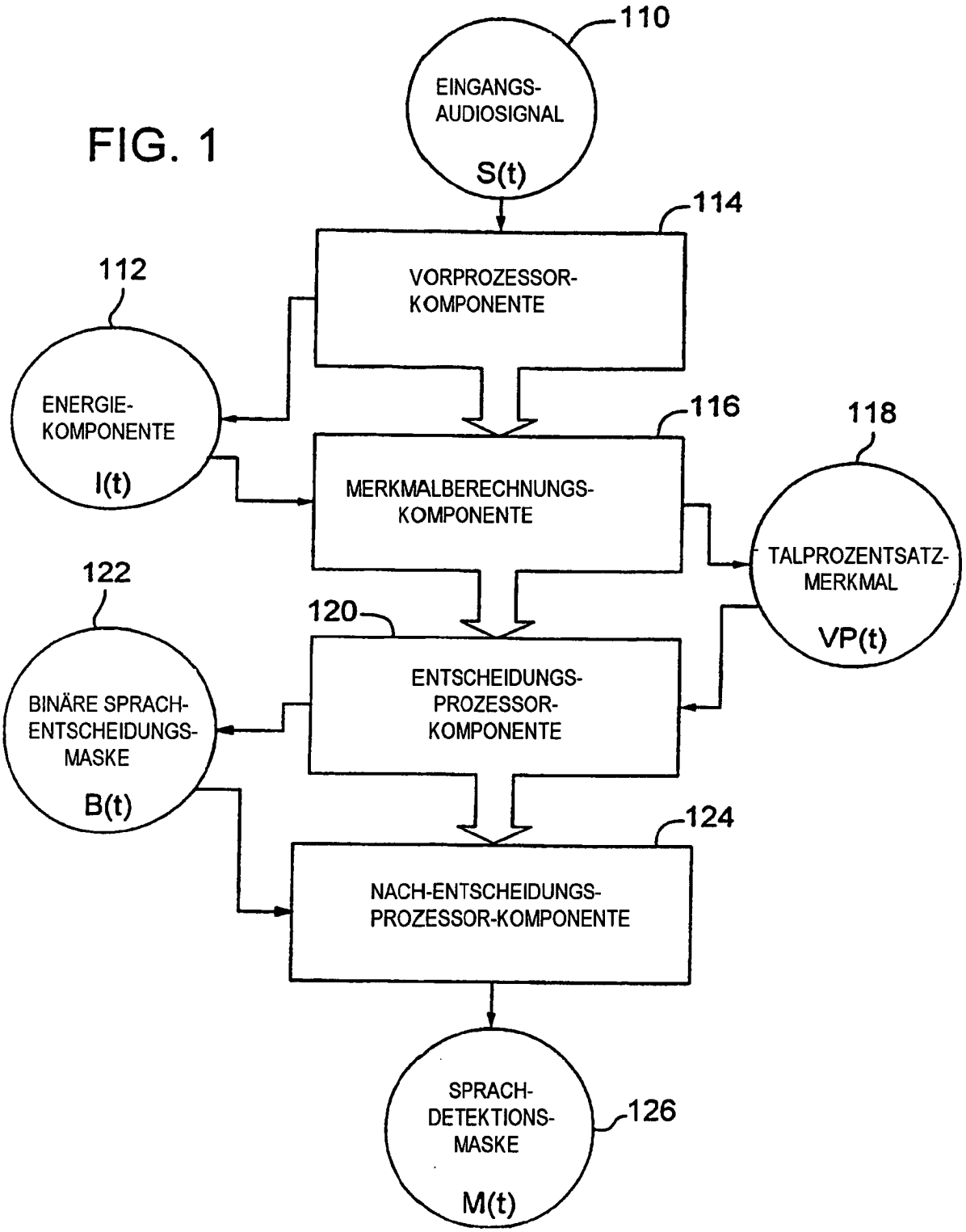
13. Computer-lesbares Medium mit darauf gespeicherten Computer-lesbaren Befehlen, um einen dadurch programmierten Computer zu veranlassen, das Verfahren nach einem der Ansprüche 1 bis 12 auszuführen.

14. Computersystem, das Einrichtungen umfasst, welche zum Durchführen des Verfahrens nach einem der Ansprüche 1 bis 12 angepasst sind.

Es folgen 6 Blatt Zeichnungen

Anhängende Zeichnungen

FIG. 1



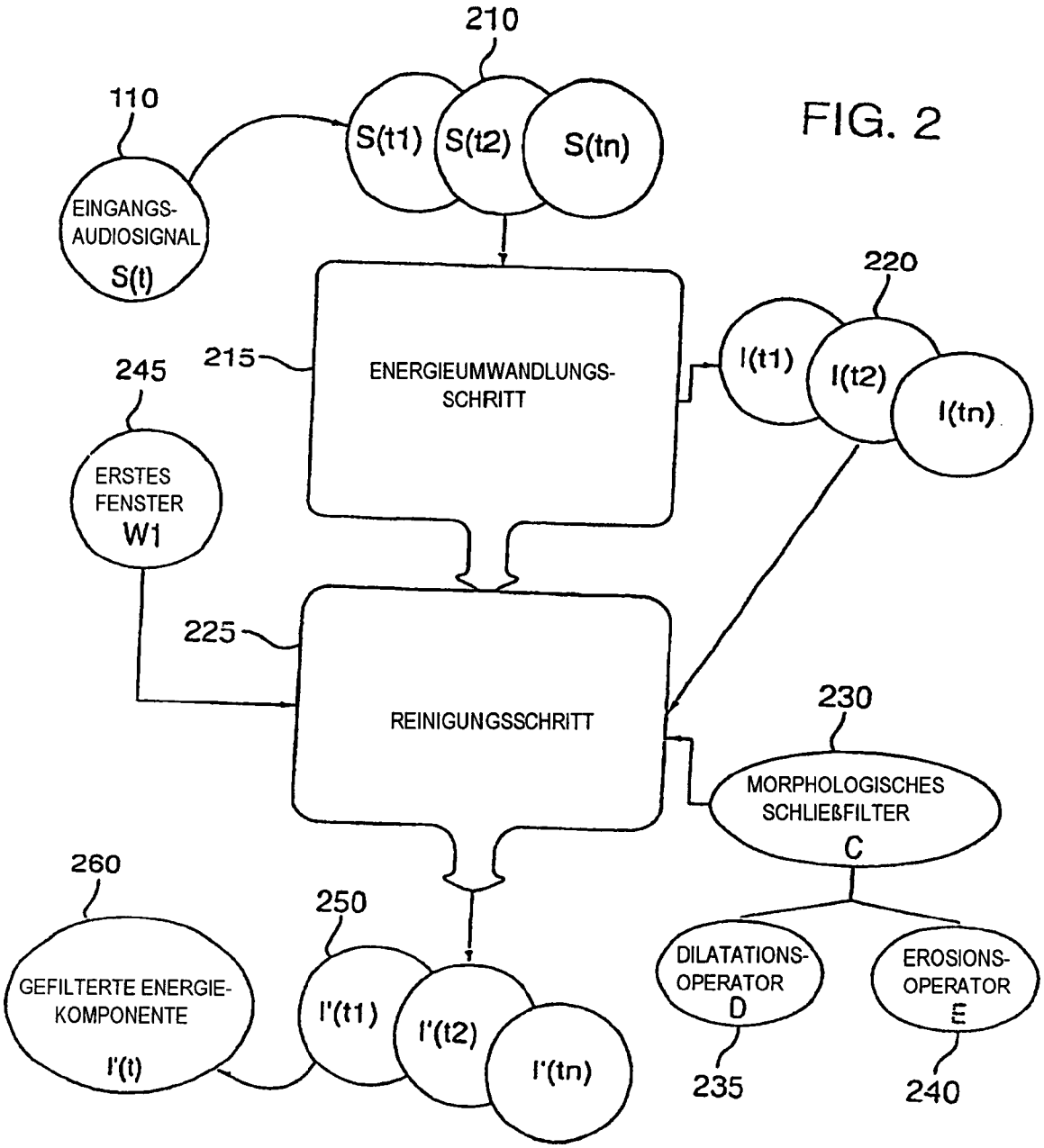

210
110
S(t1)
S(t2)
S(tn)
EINGANGS-
AUDIOSIGNAL
S(t)
220
215
ENERGIEUMWANDLUNGS-
SCHRITT
I(t1)
I(t2)
I(tn)
245
ERSTES
FENSTER
W1
225
REINIGUNGSSCHRITT
230
MORPHOLOGISCHES
SCHLIEßFILTER
C
260
250
GEFILTERTE ENERGIE-
KOMPONENTE
I'(t)
I'(t1)
I'(t2)
I'(tn)
DILATATIONS-
OPERATOR
D
EROSIONS-
OPERATOR
E
235
240

FIG. 2

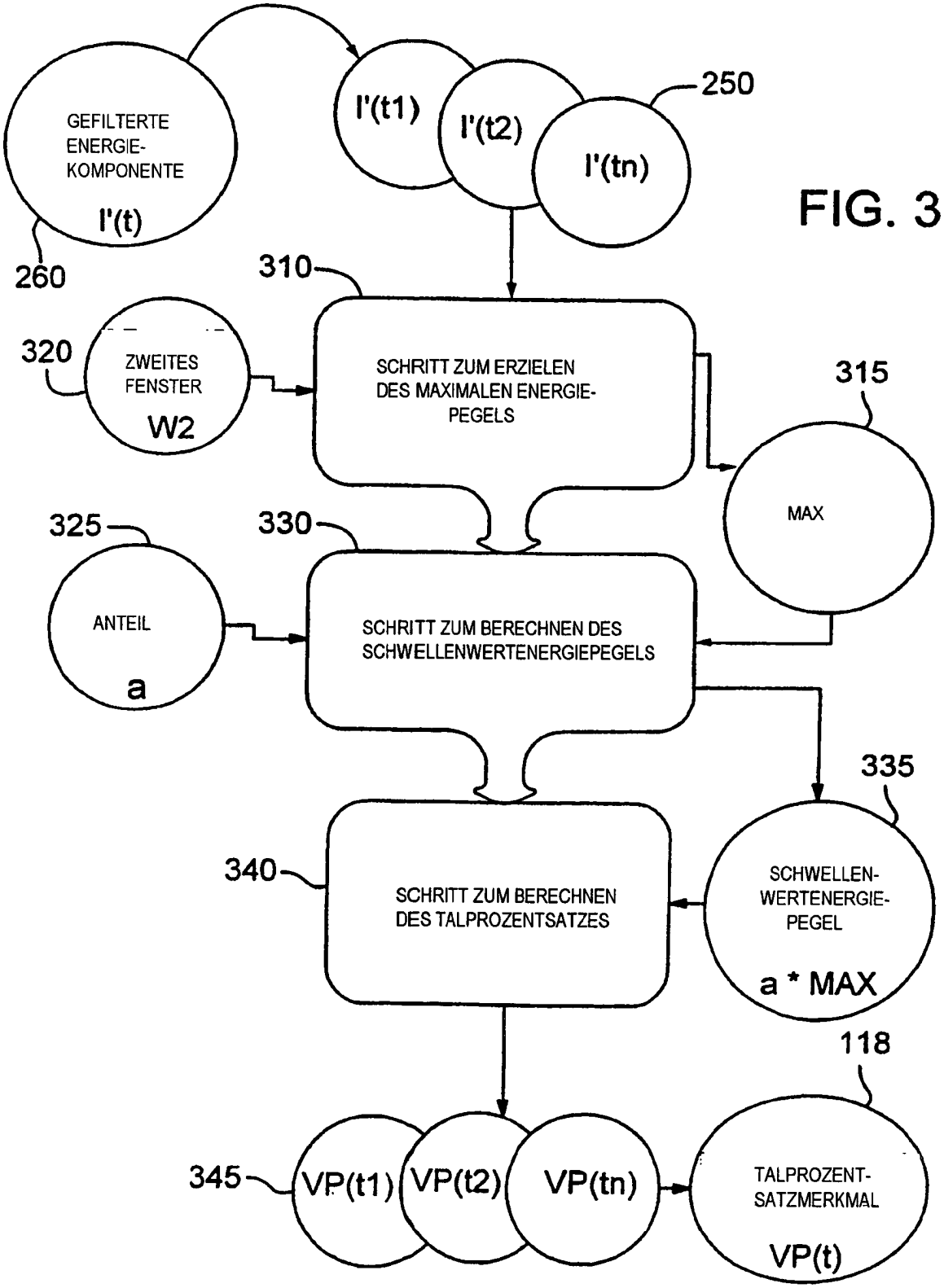

GEFILTERTE ENERGIE-KOMPONENTE
l'(t)
260
l'(t1)
l'(t2)
l'(tn)
250
310
320
ZWEITES FENSTER
W2
SCHRITT ZUM ERZIELEN DES MAXIMALEN ENERGIE-PEGELS
315
MAX
325
ANTEIL
a
330
SCHRITT ZUM BERECHNEN DES SCHWELLENWERTENERGIEPEGELS
335
340
SCHRITT ZUM BERECHNEN DES TALPROZENTSATZES
SCHWELLEN-WERTENERGIE-PEGEL
a * MAX
118
345
VP(t1)
VP(t2)
VP(tn)
TALPROZENT-SATZMERKMAL
VP(t)

FIG. 3

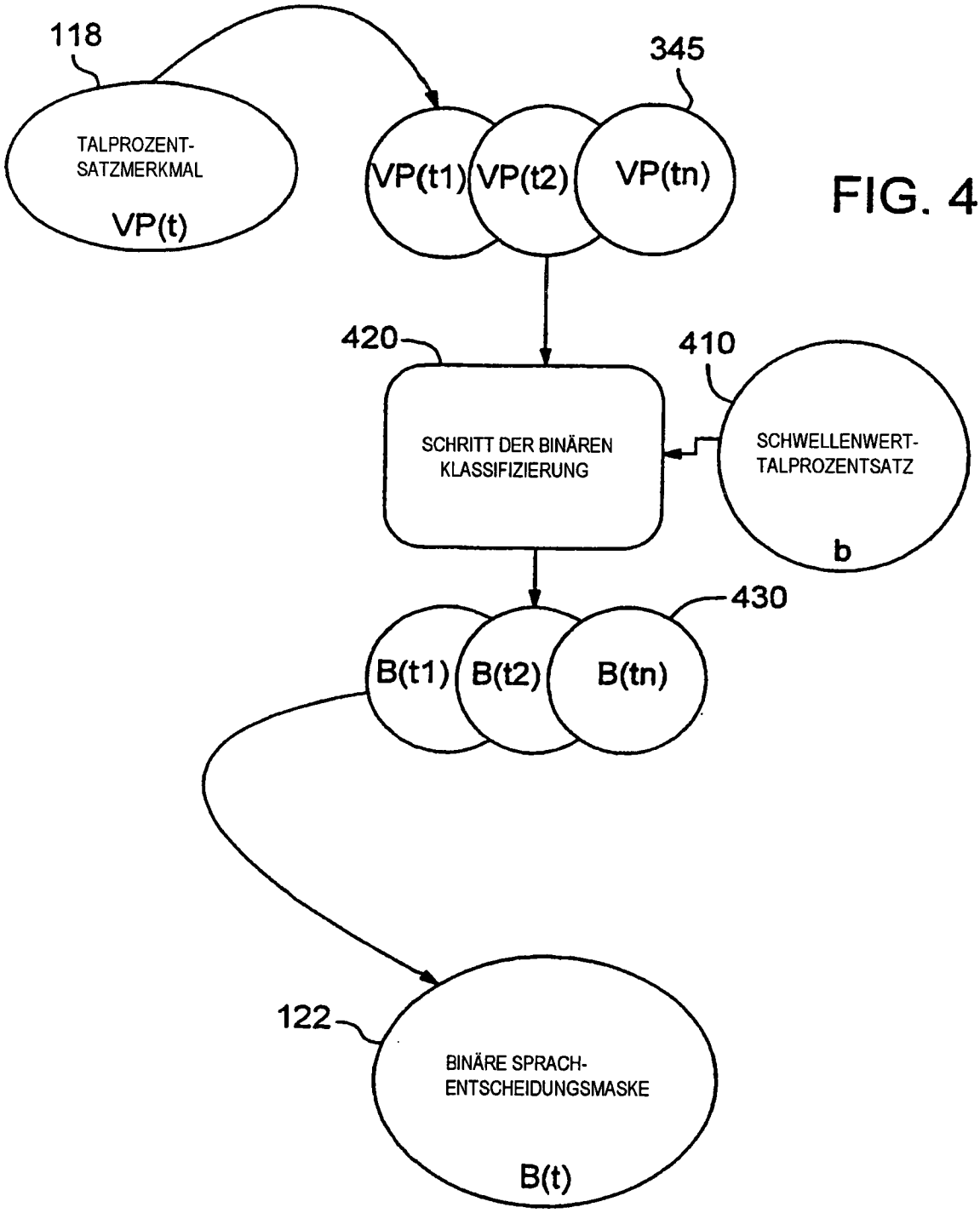

118
TALPROZENT-
SATZMERKMAL
VP(t)
345
VP(t1)
VP(t2)
VP(tn)
420
SCHRITT DER BINÄREN
KLASSIFIZIERUNG
410
SCHWELLENWERT-
TALPROZENTSATZ
b
430
B(t1)
B(t2)
B(tn)
122
BINÄRE SPRACH-
ENTSCHEIDUNGSMASKE
B(t)

FIG. 4

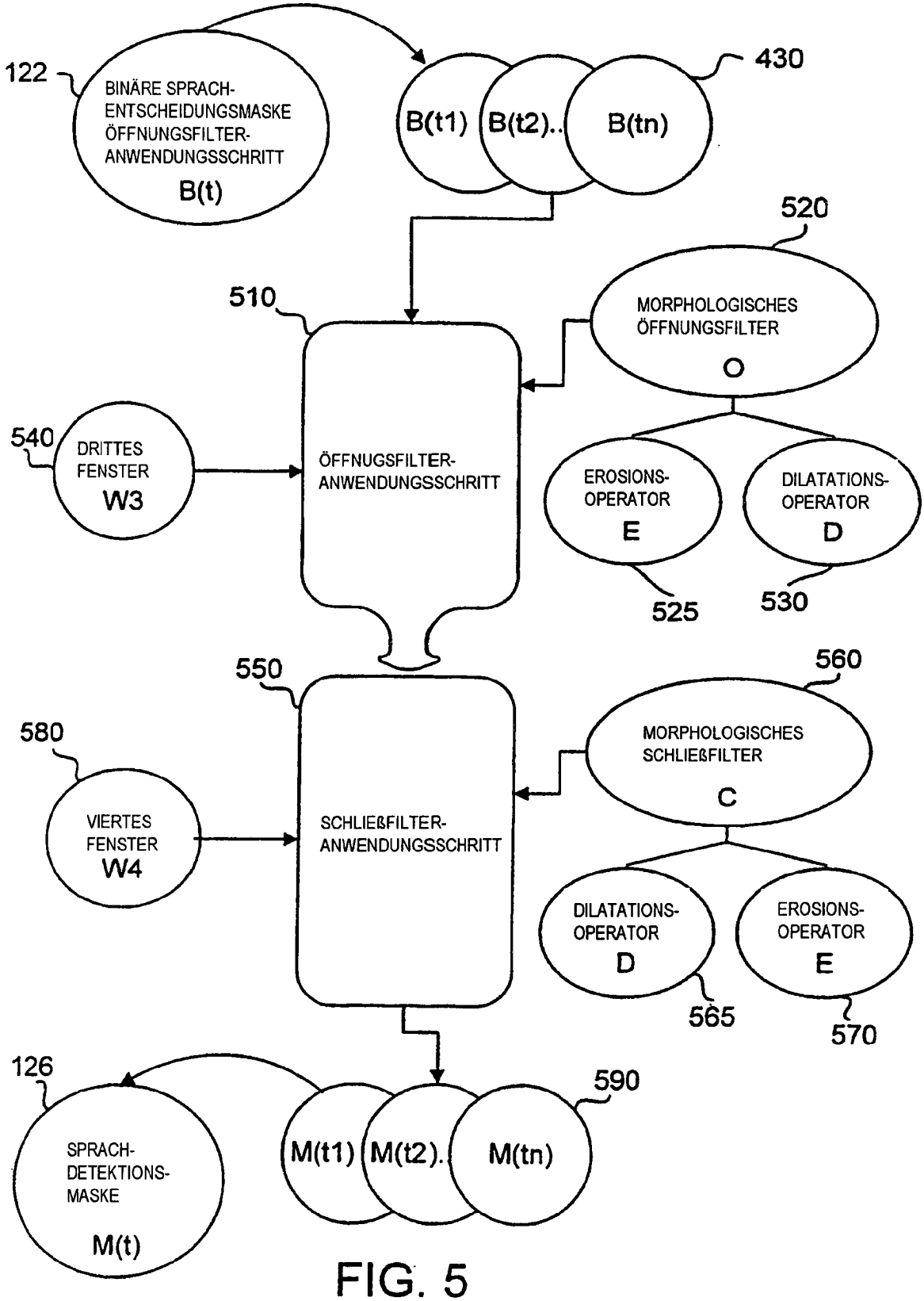
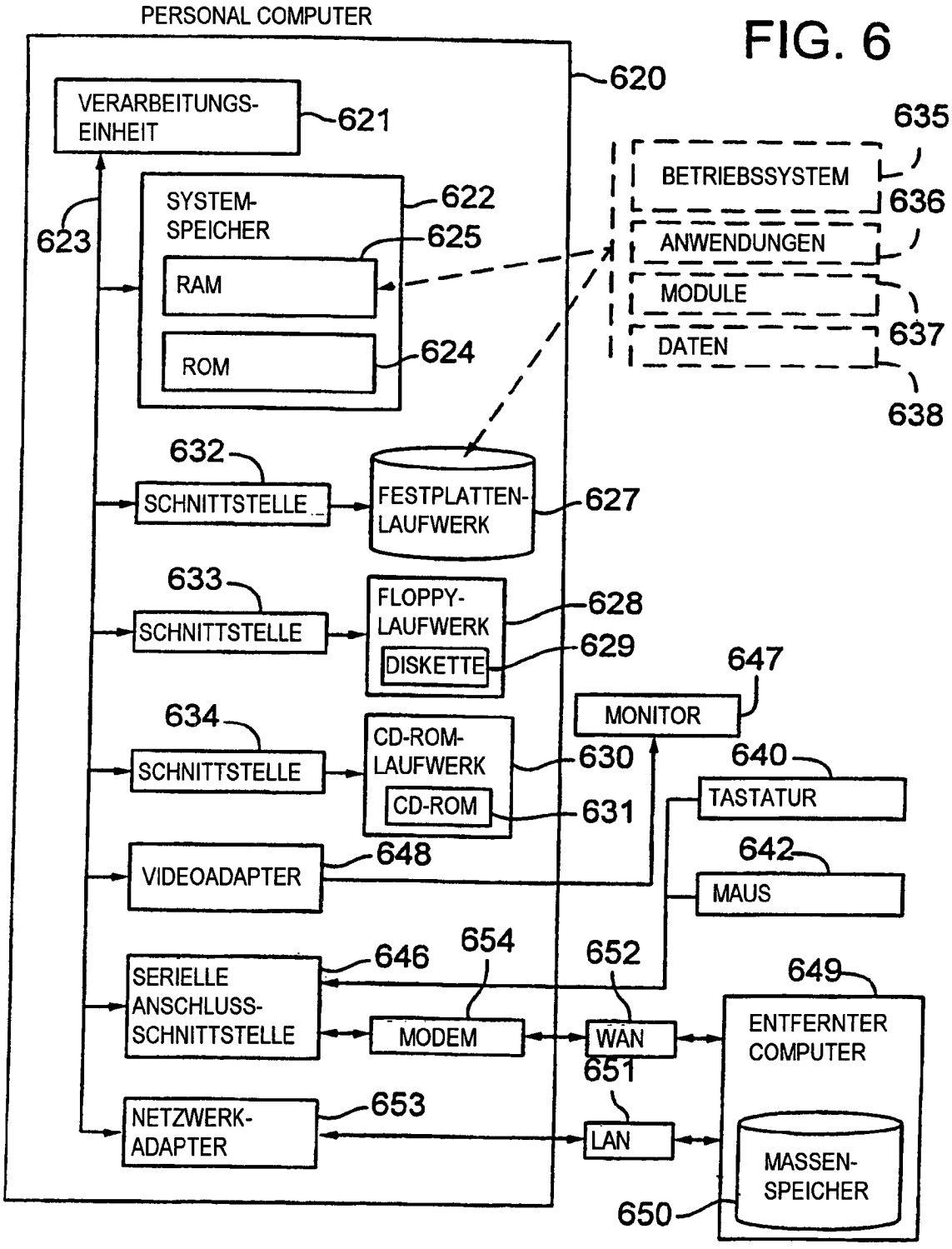

122
BINÄRE SPRACH-ENTSCHEIDUNGSMASKE ÖFFNUNGSFILTER-ANWENDUNGSSCHRITT
B(t)
B(t1) B(t2).. B(tn)
430
520
MORPHOLOGISCHES ÖFFNUNGSFILTER
O
510
ÖFFNUGSFILTER-ANWENDUNGSSCHRITT
540
DRITTES FENSTER
W3
EROSIONS-OPERATOR
E
525
DILATATIONS-OPERATOR
D
530
550
560
MORPHOLOGISCHES SCHLIEßFILTER
C
580
VIERTES FENSTER
W4
SCHLIEßFILTER-ANWENDUNGSSCHRITT
DILATATIONS-OPERATOR
D
565
EROSIONS-OPERATOR
E
570
126
SPRACH-DETEKTIONS-MASKE
M(t)
M(t1) M(t2).. M(tn)
590

FIG. 5

FIG. 6


PERSONAL COMPUTER
620
VERARBEITUNGS-EINHEIT
621
623
SYSTEM-SPEICHER
622
625
RAM
ROM
624
635
BETRIEBSSYSTEM
636
ANWENDUNGEN
MODULE
637
DATEN
638
632
SCHNITTSTELLE
FESTPLATTEN-LAUFWERK
627
633
SCHNITTSTELLE
FLOPPY-LAUFWERK
628
DISKETTE
629
647
MONITOR
634
SCHNITTSTELLE
CD-ROM-LAUFWERK
630
CD-ROM
631
640
TASTATUR
VIDEOADAPTER
648
642
MAUS
654
652
646
649
SERIELLE ANSCHLUSS-SCHNITTSTELLE
MODEM
WAN
ENTFERNTER COMPUTER
651
653
NETZWERK-ADAPTER
LAN
MASSEN-SPEICHER
650