



(19)
Bundesrepublik Deutschland
Deutsches Patent- und Markenamt

(10) **DE 699 20 461 T2 2005.12.01**

(12)

Übersetzung der europäischen Patentschrift

(97) **EP 1 096 471 B1**

(51) Int Cl.⁷: **G10L 15/20**

(21) Deutsches Aktenzeichen: **699 20 461.5**

(96) Europäisches Aktenzeichen: **99 203 613.7**

(96) Europäischer Anmeldetag: **29.10.1999**

(97) Erstveröffentlichung durch das EPA: **02.05.2001**

(97) Veröffentlichungstag

der Patenterteilung beim EPA: **22.09.2004**

(47) Veröffentlichungstag im Patentblatt: **01.12.2005**

(73) Patentinhaber:

**Telefonaktiebolaget LM Ericsson (publ),
Stockholm, SE**

(84) Benannte Vertragsstaaten:

**AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT,
LI, LU, MC, NL, PT, SE**

(74) Vertreter:

HOFFMANN & EITLE, 81925 München

(72) Erfinder:

**Brückner, Raymond, 90480 Nürnberg, DE; Hirsch,
Hans-Günther, 90562 Heroldsberg, DE; Klisch,
Rainer, 90419 Nürnberg, DE; Springer, Volker,
90491 Nürnberg, DE**

(54) Bezeichnung: **Verfahren und Vorrichtung zur robusten Merkmalsextraktion für die Spracherkennung**

Anmerkung: Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents kann jedermann beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist (Art. 99 (1) Europäisches Patentübereinkommen).

Die Übersetzung ist gemäß Artikel II § 3 Abs. 1 IntPatÜG 1991 vom Patentinhaber eingereicht worden. Sie wurde vom Deutschen Patent- und Markenamt inhaltlich nicht geprüft.

Beschreibung

[0001] Die Erfindung bezieht sich auf ein Verfahren und eine Einrichtung zum Ausführen einer robusten Merkmalsextrahierung für eine Spracherkennung in einem Umfeld mit Rauschen.

[0002] In dem Gebiet der Spracherkennung tritt ein Hauptproblem für eine akkurate Erkennung von Sprache in einem Fall eines Umfelds mit Rauschen auf. Alle möglichen verschiedenen Arten von Rauschen haben einen Einfluss auf die Spracherkennung, und sie können eine Erkennungsgenauigkeit drastisch verringern.

[0003] Vor allem in dem Gebiet des mobilen Fernsprechverkehrs oder von Zugangssystemen, die einen Zugang nach einer Erkennung eines gesprochenen Passworts erlauben, wird Spracherkennung immer wichtiger. Vor allem in diesen genannten Gebieten sind von den möglichen verschiedenen Arten von Rauschen die problematischsten ein zusätzliches stationäres oder ein nicht-stationäres Hintergrundrauschen. Eine andere Art von der Erkennungsgenauigkeit verringerndem Rauschen ist der Einfluss von Frequenzeigenschaften eines Übertragungskanal, falls die zu erkennende Sprache über solch einen Kanal übertragen wird. Zusätzliches Rauschen kann aus einem Hintergrundrauschen in Kombination mit einem auf einer Übertragungsleitung erzeugten Rauschen bestehen.

[0004] Deshalb ist es aus dem Stand der Technik bekannt, eine sogenannte lineare oder nicht-lineare spektrale Subtrahierung bereitzustellen. Die spektrale Subtrahierung ist eine Rauschunterdrückungstechnik, die die Effekte von zusätzlichem Rauschen zur Sprache reduziert. Sie schätzt die Größe oder das Leistungsspektrum von sauberer Sprache ab durch ein explizites Subtrahieren der Rauschgröße oder des Leistungsspektrums von der Größe mit Rauschen oder dem Leistungsspektrum. Solch eine Technik wurde zum Verbessern von Sprache in verschiedenen Kommunikationssituationen entwickelt.

[0005] Während die spektrale Subtrahierung eine Abschätzung des Rauschens während Pausen benötigt, wird auch angenommen, dass Rauscheigenschaften sich langsam ändern, um zu garantieren, dass die Rauschabschätzung noch gültig ist. Der Erfolg dieses Verfahrens benötigt die Verfügbarkeit eines robusten Endpunktes oder eines Sprachaktivierungsdetektors, um Sprache von Rauschen zu unterscheiden. Jedoch stellt eine gute Sprach- und Rauschtrennung eine notwendige Bedingung dar, aber sie ist bei einem geringen Signal-zu-Rauschverhältnis (SNR) schwer zu erreichen.

[0006] Zusätzlich ist selbst dann, wenn die spektrale Subtrahierung recheneffizient ist, da das Rauschen während Sprachpausen abgeschätzt wird, und selbst wenn diese Technik als eine Vorverarbeitungstechnik implementiert werden kann, die die anderen Verarbeitungsschritte unverändert lässt, die Leistungsfähigkeit des spektralen Subtrahierungsverfahrens stark von dem Rauschen abhängig und wie jedes Rauschen extrahiert wird. Das damit in Zusammenhang stehende Problem ist, dass, selbst wenn das Breitbandrauschen reduziert wird, ein Rauschrest verbleibt (Junqua et al; "Robustness in automatic speech recognition"; Kluwer Academic Publisher; 1996; Abschnitt 9.2 Speech Enhancement, Seiten 277 ff.)

[0007] Nichtsdestotrotz ist, selbst wenn mit den obengenannten Verfahren die Spracherkennung verbessert werden kann, für diese Herangehensweisen die Abschätzung der Rauscheigenschaften entscheidend. Wie oben bemerkt, wird eine Sprach-zu-Rausch-Unterscheidung zum Markieren solcher Segmente eines nur Rauschen enthaltenden Sprachsignals, gebraucht. Aber solch eine Unterscheidung kann nicht fehlerfrei sein, und sie ist schwer zu erreichen. Zusätzlich zu diesem können, wenn Segmente des Sprachsignals betrachtet werden, die das Überlagern von Sprache und stationärem Rauschen enthalten, diese Segmente durch die Überlagerung von entsprechenden Verteilungsfunktionen für eine spektrale Rauschkomponente und eine spektrale Sprachkomponente beschrieben werden. Diese Verteilungsfunktionen überlappen sich abhängig von dem SNR. Die Überlappung ist höher je geringer das SNR ist. Und deshalb kann es in diesem Fall nicht entschieden werden, ob kurzzeitige Spektren Sprache in spektralen Bereichen enthalten, wobei die spektrale Größenordnung der Sprache Werte der gleichen Größe oder geringeren Größe als das Rauschen annimmt.

[0008] Matsumoto H et al; "Smoothed Spectral Subtraction for a Frequency-Weighted HMM in Noisy Speech Recognition", Proceedings of the International Conference on Spoken Language Processing, XX, XX, 1. Januar 1996, Seiten 905-908, XP002065141 beschreibt ein Verfahren, in welchem ein sauberes Sprachleistungsspektrum durch eine spektrale Subtrahierung eines abgeschätzten Rauschleistungsspektrums von einem Sprachleistungsspektrum mit Rauschen abgeschätzt wird. Um eine Verzerrung in dem abgeschätzten log-Leistungsspektrum zu unterdrücken, das durch die spektrale Subtrahierung hervorgerufen wird, wird ein geglättetes log-Spektrum abgeleitet, und eine neue Abschätzung eines log-Leistungsspektrums wird dann berechnet als eine gewichtete Summe von dem nicht-geglätteten und dem geglätteten log-Leistungsspektrum.

Die gewichteten Koeffizienten werden als eine monoton zunehmende Funktion des abgeschätzten SNR ausgewählt, so dass hohe SNR-Komponenten im wesentlichen beibehalten werden, während niedrige SNR-Komponenten im wesentlichen durch solche des geglätteten log-Leistungsspektrums ersetzt werden. Jedoch ist vor allem die Berechnung des geglätteten log-Leistungsspektrums ziemlich komplex und verlangt eine große Menge an Rechenressourcen, die das Verfahren weniger geeignet für mobile Anwendungen machen.

[0009] Deshalb ist es eine Aufgabe der vorliegenden Erfindung, ein Verfahren und eine Vorrichtung bereitzustellen, die die Probleme überwinden, und die eine robustere Spracherkennung in einem Umfeld mit Rauschen erlauben.

[0010] Dies wird durch die Lehre der Ansprüche 1, 12, 17 und 19 erreicht.

[0011] Gemäß der Erfindung ist es vorteilhaft, dass ein kurzzeitiges nur Rauschen enthaltendes Spektrum geglättet wird und zusätzlich im Fall von Sprachsegmenten mit Rauschen unzuverlässige spektrale Komponenten durch sogenannte zuverlässige Komponenten interpoliert werden, was daher in einer verbesserten Spracherkennung resultiert oder noch deutlicher, in einer robusten Merkmalsextrahierung, die eine verbesserte Spracherkennung unterstützt.

[0012] Es ist vorteilhaft, die Interpolation basierend auf mindestens einem spektralen Komponenten eines angrenzenden kurzzeitigen Spektrums und/oder mindestens einem in der Zeit vorhergehenden spektralen Komponenten auszuführen, da es erwartet werden könnte, dass eine sogenannte unzuverlässige Sprachkomponente mit einer geringen Wahrscheinlichkeit Sprache zu enthalten, geglättet wird.

[0013] Weitere bevorzugte Ausführungsformen können in den Ansprüchen 2–11 und 13–16 und 20 gefunden werden.

[0014] Nach Anspruch 2 wird eine verbesserte Spracherkennung durch Aufnehmen von zwei angrenzenden spektralen Komponenten erreicht, und wobei eine in der Zeit vorangeht.

[0015] Ein weiterer Vorteil nach Anspruch 3 ist die berechnete Wahrscheinlichkeit mit einer Schwelle zu vergleichen, um eine Definition zu bekommen, welche spektrale Komponente zu interpolieren ist.

[0016] Nach Anspruch 5 ist es weiterhin vorteilhaft, die spektrale Komponente basierend auf rauschloser Sprache zu interpolieren.

[0017] Nach Anspruch 8 werden zwei Interpolationen ausgeführt, die in einer noch besseren Spracherkennung resultieren.

[0018] Es ist nach Anspruch 11 vorteilhaft, die Einteilung YYY der kurzzeitigen Spektren auf einem MEL-Frequenzbereich zu basieren, da der MEL-Frequenzbereich auf dem menschlichen Ohr basiert.

[0019] Weiterhin ist es vorteilhaft, das Verfahren für eine Spracherkennung zum Steuern von elektronischen Geräten anzuwenden, wie zum Beispiel für Mobiltelefone, Telefone oder ein Sprache verwendendes Zugangssystem, um Zugang oder Wählen etc. zu erlauben.

[0020] In dem folgenden wird die Erfindung weiterhin beschrieben und mittels Beispiele und durch die Figuren unterstützt.

[0021] Die folgenden Figuren zeigen:

[0022] [Fig. 1](#): Integration einer Interpolation in eine Merkmalsextrahierung nach Anspruch 1,

[0023] [Fig. 2](#): Integration einer Interpolation in eine Merkmalsextrahierung nach Anspruch 5,

[0024] [Fig. 3](#): Integration einer Interpolation in eine Merkmalsextrahierung nach Anspruch 8,

[0025] [Fig. 4](#): Diagramm einer Wahrscheinlichkeitsfunktion für die Existenz von Sprache.

[0026] Um eine Spracherkennung robuster gegenüber Rauschen zu machen, kann ein robustes Merkmalsextrahierungsschema angewandt werden. Dieses Schema versucht diese Merkmale aus dem Sprachsignal zu

extrahieren, die nicht für Rauschen empfindlich sind, oder die von dem Einfluss des Rauschens befreit sind. Zusätzlich zu diesem sind die Mehrheit der Merkmalsextrahierungsschemata auf einer kurzzeitigen spektralen Analyse basiert. Zusätzlich sind die meisten der Spracherkennungssysteme auf einer kurzzeitigen Analyse in dem MEL-Frequenzbereich basiert. Der MEL-Frequenzbereich basiert auf dem menschlichen Ohr und ist im Stand der Technik bekannt und wird deshalb nicht weiter in dieser Anmeldung beschrieben.

[0027] Der Ausdruck "robust" sollte eine Robustheit gegenüber stationärem und nicht-stationärem Hintergrundrauschen, wie in dem Stand der Technik genannt, abdecken. Zusätzlich zu diesem sollte die in der Anmeldung genannte Robustheit auch eine Robustheit gegenüber unbekannten Frequenzeigenschaften abdecken, die von irgendeiner Art eines elektronischen Gerätes resultieren, zum Beispiel einem Mikrofon und/oder einem digitalen oder analogen Filter durch Verwendung der Erfindung mit einem Mobiltelefon oder irgendeinem anderen Telefon.

[0028] In dem folgenden wird die Erfindung weiter mit der Hilfe von [Fig. 1](#) beschrieben. Ein Sprachsignal wird, nach einer Vorverarbeitung, zum Beispiel einem Framing bzw. einer Rahmenbildung und einer Transformation, die nicht auf dieser Erfindung basiert, und die im Stand der Technik gut bekannt ist, in eine Anzahl von kurzzeitigen spektralen Komponenten in L-Unterbänder aufgespaltet, wobei $L = 1, 2, \dots L$ bevorzugt in dem Bereich von 10 bis 30 MEL ist. Die Einteilung in kurzzeitige spektrale Komponenten kann auf einer kurzzeitigen spektralen Analyse in dem MEL-Frequenzbereich basieren. Die Ausgabe solch einer Filterbank in dem MEL-Frequenzbereich ist ein kurzzeitiges Größen- oder Leistungsspektrum in die Anzahl von L-Unterbändern mit $L = 1, 2, \dots$.

[0029] In einem nächsten Schritt wird eine Abschätzung eines Rauschspektrums von nur Rauschen enthaltenden Segmenten ausgeführt ES. Dies kann durch Anwenden eines Sprachaktivitätsdetektors durchgeführt werden. Das Durchschnittsspektrum wird aus allen Spektren berechnet, die zu einem Segment gehören, das als Rauschen durch einen Sprachaktivitätsdetektor markiert ist. Das Durchschnittsspektrum wird als die Abschätzung des Rauschspektrums genommen bzw. aufgefasst.

[0030] Dann wird eine spektrale Subtrahierung ausgeführt, durch Subtrahieren des abgeschätzten Rauschspektrums von dem kurzzeitigen Spektrum S_{sub} mit Rauschen, und was in einem Schätzergebnis resultiert. Eine Realisierung ist die Subtrahierung der spektralen Komponenten in dem Größenbereich, der wie folgt berechnet wird:

$$|\hat{S}_l(t)| = |X_l(t)| - |\hat{N}_l(t)| \quad (1)$$

mit $|\hat{S}_l|$ = das Abschätzungsergebnis einer spektralen Sprachkomponente;

mit $|X_l|$ = die tatsächliche Eingangskomponente mit Rauschen,

und $|\hat{N}_l|$ = die abgeschätzte Rauschkomponente zur Zeit t und für das Unterband $l = 1, \dots, L$.

[0031] Es sollte bemerkt werden, dass in allen Gleichungen die Werte $\hat{S}_l$, t etc. die gleiche Bedeutung haben sollten. In einem weiteren Schritt wird eine Wahrscheinlichkeit, Rauschen zu enthalten, für jede kurzzeitige Spektralkomponente berechnet. Die Berechnung der Wahrscheinlichkeit wird später detaillierter beschrieben.

[0032] Falls die berechnete Wahrscheinlichkeit für die kurzzeitige spektrale Komponente in einer geringen Wahrscheinlichkeit, Sprache zu enthalten, resultiert, wird in einem nächsten Schritt, diese spektrale Komponente interpoliert IP.

[0033] Die in den meisten der heute verwendeten Spracherkennungssystemen cepstrale Koeffizienten angewandt werden, können die wie vorher beschriebenen, geglätteten und interpolierten kurzzeitigen Spektren weiter wie folgt verarbeitet werden. Ein Logarithmus von jeder kurzzeitigen spektralen Komponente wird einer log-bildung unterzogen, und letztendlich wird eine Anzahl von cepstralen Koeffizienten aus ihr über eine diskrete Kosinustransformation DCT transferiert.

[0034] In dem folgenden wird eine Interpolation der spektralen Komponenten jedes kurzzeitigen Spektrums detaillierter beschrieben.

[0035] Wie zuvor beschrieben, wird die Wahrscheinlichkeit berechnet, dass eine spektrale Komponente Sprache enthält. Im Fall, dass die Wahrscheinlichkeitsberechnung in einer niedrigen Wahrscheinlichkeit resultiert, Sprache zu enthalten, wird die spektrale Komponente als unzuverlässig betrachtet. Diese unzuverlässigen Komponenten werden dann interpoliert. Diese Interpolation basiert zum Beispiel auf einem Betrachten der

spektralen Komponenten mindestens eines angrenzenden Unterbandes und/oder mindestens einer in der Zeit vorhergehenden oder nachfolgenden spektralen Komponente des kurzzeitigen Spektrums.

[0036] Als ein bevorzugtes Beispiel werden zwei angrenzende Unterbänder und eine in der Zeit vorhergehende spektrale Komponente für die Interpolation genommen.

[0037] Die Interpolation wird zum Beispiel durch Berechnen einer gewichteten Summe wie folgt durchgeführt:

$$|\tilde{S}_i(t_i)| = [\text{prob}_i(t_i) \times |\hat{S}_i(t_i)| + \text{prob}_{i-1}(t_i) \times |\hat{S}_{i-1}(t_i)| + \text{prob}_{i+1}(t_i) \times |\hat{S}_{i+1}(t_i)| + \text{prob}_i(t_{i-1}) \times |\hat{S}_i(t_{i-1})|] / (\sum \text{prob}) \quad (2)$$

wobei prob = Wahrscheinlichkeit eines Unterbandes I.

[0038] Trotzdem können weitere spektrale Komponenten von weiteren angrenzenden Unterbändern eine Basis für die Interpolation darstellen, und deshalb wird die Beschreibung nicht als den Umfang dieser einschränkend angesehen.

[0039] Wie zuvor bemerkt, wird die Wahrscheinlichkeit berechnet, dass jedes kurzzeitige Spektrum Rauschen enthält. Deshalb wird ein Verhältnis zwischen $|\hat{S}_i(t_i)|/|X_i(t_i)|$ ausgeführt. Dies enthält indirekt das Signal-zu-Rauschverhältnis in einem kurzzeitigen Spektrum zur Zeit t_i . Das Verhältnis kann Werte zwischen minus unendlich und 1 annehmen, wie in [Fig. 4](#) gezeigt, wo die X-Achse das Verhältnis zeigt und die Y-Achse die Wahrscheinlichkeit prob zeigt. Negative Werte treten in einem Fall auf, in dem die Rauschkomponente $|\hat{N}_i(t_i)|$ größer ist als die tatsächliche spektrale Komponente $X_i(t_i)$. Dies kennzeichnet Rauschen, und es kennzeichnet deshalb ferner eine unzuverlässige Komponente. Im Vergleich zu diesem wird ein Wert von 1 genommen, falls kein Rauschen abgeschätzt wird und wo reine Sprache existiert. [Fig. 4](#) zeigt ein Beispiel zum Definieren eines Maßes für die Wahrscheinlichkeit, dass eine spektrale Komponente Sprache enthält.

[0040] Wie vorher bemerkt, wird eine Interpolation in einem Fall ausgeführt, dass die berechnete Wahrscheinlichkeit für das kurzzeitige Spektrum dazu führt, dass das Spektrum eine geringe Wahrscheinlichkeit aufweist, Sprache zu enthalten. Da die berechnete Wahrscheinlichkeit nicht immer entweder 0 oder 1 ist, wird bevorzugt eine Schwelle definiert, um spektrale Komponenten mit ihrem entsprechenden Verhältnis zu definieren, das unter dieser Schwelle liegt und deshalb angenommen wird, dass es eine geringe Wahrscheinlichkeit aufweist, Sprache zu enthalten.

[0041] Bevorzugt ist die Schwelle 0,6, aber kann irgendeinen Wert zwischen 0 und 1 annehmen. Die Schwelle kann vordefiniert sein oder kann sich abhängig von der Umgebung ändern.

[0042] Im folgenden wird ein anderes Beispiel mit der Hilfe von [Fig. 2](#) beschrieben.

[0043] Es sollte hier bemerkt werden, dass Blöcke mit dem gleichen Index dazu gedacht sind, durchgehend für alle Figuren die gleiche Bedeutung zu haben.

[0044] Zusätzlich zu den genannten Schritten wird eine Anzahl M von spektralen Abständen für jedes kurzzeitige Spektrum bestimmt. Die spektralen Abstände werden zwischen dem kurzzeitigen Spektrum nach Subtrahierung und einer Anzahl M von Spektren bestimmt, die rauschlose Sprache enthalten, wobei $M = 1, 2, \dots$ ist. Weiterhin wird für jedes kurzzeitige Spektrum mindestens ein Spektrum bestimmt, welches rauschlose Sprache mit dem kleinsten spektralen Abstand zu dem kurzzeitigen Spektrum enthält MATCH.

[0045] Der spektrale Abstand kann wie folgt berechnet werden:

$$\text{Fehler}(|\hat{S}(t_i)|, |V(m)|) = \sum_{l=1}^L \text{prob}_l(t_i) \times (|\hat{S}_l(t_i)| - |V_l(m)|)^2 / \sum_{l=1}^L \text{prob}_l(t_i) \quad (3)$$

mit $V(m)$ = Spektrum mit rauschloser Sprache, und
wobei m einer der Codebucheinträge ist.

[0046] Der eine Wert, der berechnet wird und den kleinsten Fehler aufweist, wird als derjenige mit dem kleinsten spektralen Abstand betrachtet.

[0047] Diesem folgend wird die Interpolation durch Extrahieren von spektraler Information von den rauschlosen Sprachspektren mit dem kleinsten spektralen Abstand ausgeführt. Trotzdem können auch eine Anzahl von

rauschlosen Sprachspektren genommen werden, die alle einen kleinen spektralen Abstand aufweisen.

[0048] Eine Möglichkeit zum Ausführen der Interpolation ist eine gewichtete Summe von der tatsächliche abgeschätzten spektralen Komponente $|\hat{S}_i(t_i)|$ und der entsprechenden Komponente des nächsten Sektors zu berechnen. Dies könnte wie folgt ausgeführt werden:

$$|\tilde{S}_i(t_i)| = \text{prob}_i(t_i) |\hat{S}_i(t_i)| + (1 - \text{prob}_i(t_i)) \times V_i(m_{\text{closest}}) \quad (4)$$

[0049] Der Grund für diese Vorgehensweise basiert auf einem Wissen, dass Sprache nur in gewissen Gebieten eines L-dimensionalen spektralen Merkmalsraum vorliegt. Mit diesem Wissen basiert die Interpolation der spektralen Komponenten mit einer geringen Wahrscheinlichkeit, Sprache zu enthalten, auf der Extrahierung von spektraler Information von dem nächsten Eintrag oder mehr als einem nahen Eintrag in dem Codebuch CB.

[0050] In dem folgenden wird eine andere bevorzugte Ausführungsform der Erfindung mittels [Fig. 3](#) beschrieben.

[0051] Wie in [Fig. 3](#) gezeigt, wird zusätzlich zu einer ersten Interpolation IP1 eine andere Interpolation IP2 ausgeführt. In dem gezeigten Beispiel basiert die erste Interpolation auf spektralen Komponenten mindestens eines angrenzenden Unterbandes und/oder mindestens einer in der Zeit vorangehenden oder nachfolgenden spektralen Komponente des kurzzeitigen Spektrums, wie gemäß Anspruch 1 und [Fig. 1](#) bemerkt wurde.

[0052] Die zweite Interpolation IP2 basiert auf einem Extrahieren von spektraler Information von den rauschlosen Sprachspektren mit dem kleinsten spektralen Abstand gemäß Anspruch 6 und [Fig. 2](#).

[0053] Trotzdem wird dies nicht als die Erfindung einschränkend angesehen, da die Interpolation nach Anspruch 6 zuerst ausgeführt werden könnte, und dann die Interpolation nach Anspruch 1 ausgeführt werden kann. Trotzdem sollte wieder bemerkt werden, dass für beide Interpolationen der wesentliche Aspekt der ist, dass unzuverlässige kurzzeitige spektrale Komponenten interpoliert werden oder besser, die kurzzeitigen spektralen Komponenten mit einer geringen Wahrscheinlichkeit, Sprache zu enthalten. Im allgemeinen basiert die Interpolation der spektralen Komponenten mit Rauschen auf zuverlässigen spektralen Sprachkomponenten, die in der Nachbarschaft gefunden werden könnten, wie es für beide Interpolationen angewandt werden könnte.

[0054] In dem folgenden wird eine Vorrichtung für eine robuste Merkmalsextrahierung für Spracherkennung in einer Umgebung mit Rauschen gemäß [Fig. 1](#) beschrieben, wobei ein Sprachsignal segmentiert wird, und gekennzeichnet ist durch spektrale Komponenten. Die Vorrichtung umfasst eine Filterbank MEL zum Aufspalten des Sprachsignals in eine Anzahl von spektralen Komponenten in L-Unterbändern, mit $L = 1, 2, \dots$ Ferner werden Abschätzeinrichtungen ES vorgesehen zum Abschätzen eines Rauschspektrums aus Segmenten, die nur Rauschen enthalten.

[0055] Die Vorrichtung umfasst ferner einen Subtrahierer SSub zum Ausführen einer spektralen Subtrahierung des abgeschätzten Rauschspektrums von dem entsprechenden Kurzzeitspektrum, und einen Rechner zum Berechnen einer Wahrscheinlichkeit für jedes kurzzeitige Spektrum, Rauschen zu enthalten, wird vorgesehen.

[0056] Es werden in der Vorrichtung auch Interpolationseinrichtungen IP vorgesehen zum Interpolieren der spektralen Komponente jedes kurzzeitigen Spektrums, wenn die berechnete Wahrscheinlichkeit, Sprache zu enthalten, für die kurzzeitige Spektrumskomponente gering ist.

[0057] In einer anderen bevorzugten Ausführungsform der Anmeldung wird ein Vergleicher zum Vergleichen der berechneten Wahrscheinlichkeit mit einer Schwelle bereitgestellt, die die niedrige Wahrscheinlichkeit definiert, Sprache zu enthalten.

[0058] Um cepstrale Koeffizienten zu bekommen, werden letztendlich eine Einrichtung zum Logarithmieren log jeder spektralen Komponente, und ferner eine Einrichtung zum Ausführen einer diskreten Kosinustransformation DCT vorgesehen, und ihre Ausgabe ist eine Anzahl von cepstralen Koeffizienten, die für die Spracherkennung benutzt werden.

[0059] In dem folgenden wird eine andere bevorzugte Ausführungsform gemäß [Fig. 2](#) beschrieben. Es sollte

wieder bemerkt werden, dass die gleichen Abkürzungen für den gleichen Block beibehalten werden sollten.

[0060] Zusätzlich zu der mit Bezug auf [Fig. 1](#) genannten Einrichtung wird ein Codebuch CB mit rauschlosen Sprachspektren vorgesehen, und ferner Bestimmungseinrichtungen zum Bestimmen einer Anzahl M von spektralen Abständen für jedes kurzzeitige Spektrum durch Bestimmen der spektralen Abstände zwischen dem kurzzeitigen Spektrum nach Subtrahierung, und eine Anzahl M von Spektren, die rauschlose Sprachspektren enthalten, werden auch bereitgestellt.

[0061] Weiterhin werden zu dieser Bestimmungseinrichtung MATCH vorgesehen, um für jedes kurzzeitige Spektrum mindestens ein rauschlose Sprache enthaltendes Spektrum mit dem kleinsten spektralen Abstand zu dem kurzzeitigen Spektrum zu bestimmen, und ihr Ergebnis ist eine Eingabe an die Interpolationseinrichtung IP, um die Interpolation auszuführen.

[0062] Schließlich wird ein Computerprogrammprodukt, ladbar in einen Speicher eines elektronischen Geräts, beschrieben werden (ohne Figur). Ein Computerprogrammprodukt, ladbar in einen Speicher, zum Beispiel eines Mobiltelefons, sollte Software-Codeteile oder Module zum Ausführen der Schritte einer der Ansprüche 1 bis 11 umfassen.

[0063] Bevorzugt wird dieses Computerprogramm auf einem computerlesbaren Medium gespeichert.

[0064] Letztendlich können das Verfahren und die Vorrichtung und das Computerprogrammprodukt für eine Spracherkennung zum Steuern elektronischer Geräte benutzt werden. Solch ein elektronisches Gerät ist zum Beispiel ein Telefon oder ein Mobiltelefon oder eine Zugangseinrichtung.

Patentansprüche

1. Verfahren für eine robuste Merkmalsextrahierung für eine Spracherkennung in einem verrauschten Umfeld bzw. einem Umfeld mit Rauschen, wobei ein Sprachsignal segmentiert wird und gekennzeichnet ist durch spektrale Komponenten mit den Schritten:

– Aufspalten des Sprachsignals in eine Anzahl von kurzzeitigen spektralen Komponenten in L Unterbändern mit $L = 1, 2, \dots$

– Abschätzen eines Rauschspektrums aus Segmenten, die nur Rauschen enthalten,

– Ausführen einer spektralen Subtrahierung des abgeschätzten Rauschspektrums von dem entsprechenden kurzzeitigen Spektrum,

– Berechnen einer Wahrscheinlichkeit für jede kurzzeitige Spektralkomponente, Rauschen zu enthalten, und

– Interpolieren der spektralen Komponente jedes kurzzeitigen Spektrums, wenn die berechnete Wahrscheinlichkeit, Sprache zu enthalten, für die kurzzeitige Spektralkomponente gering ist,

dadurch gekennzeichnet, dass die Interpolation auf spektrale Komponenten mindestens eines angrenzenden Unterbandes und/oder mindestens auf eine in der Zeit vorhergehende oder nachfolgende spektrale Komponente des kurzzeitigen Spektrums basiert.

2. Verfahren nach Anspruch 1, wobei die Interpolation auf zwei angrenzende Unterbänder und eine in der Zeit vorhergehende spektrale Komponente basiert.

3. Verfahren nach Anspruch 1 oder 2, wobei die berechnete Wahrscheinlichkeit mit einer Schwelle verglichen wird, die durch die geringe Wahrscheinlichkeit, Sprache zu enthalten, definiert wird.

4. Verfahren nach einem der Ansprüche 1 bis 3, wobei jede spektrale Komponente logarithmiert wird und eine diskrete Kosinustransformation (DCT) ausgeführt wird.

5. Verfahren nach Anspruch 1, mit den zusätzlichen Schritten:

– Bestimmen einer Anzahl M spektraler Abstände für jedes kurzzeitige Spektrum durch Bestimmen der spektralen Abstände zwischen dem kurzzeitigen Spektrum nach Subtrahierung und einer Anzahl M von Spektren, die rauschlose Sprache enthalten,

– Bestimmen für jedes kurzzeitige Spektrum mindestens eines Spektrums, das rauschlose Sprache mit dem kleinsten spektralen Abstand zu dem kurzzeitigen Spektrum enthält.

6. Verfahren nach Anspruch 5, wobei

– eine Interpolation ausgeführt wird, durch Extrahieren von spektraler Information von den rauschlosen Sprachspektren mit dem kleinsten spektralen Abstand.

7. Verfahren nach Anspruch 5 oder 6, wobei der spektrale Abstand bestimmt wird durch Berechnen eines gewichteten mittleren quadratischen Fehlers (MSE).
8. Verfahren nach Anspruch 1, wobei eine erste Interpolation gemäß Anspruch 1 ausgeführt wird und eine zweite Interpolation gemäß Anspruch 6 ausgeführt wird.
9. Verfahren nach einem der Ansprüche 5 bis 7, wobei die rauschlose Sprache enthaltenden Spektren in einem mit rauschloser Sprache trainiertem Codebuch enthalten sind.
10. Verfahren nach Anspruch 9, wobei die in dem Codebuch gespeicherte spektrale Information aus cepstralen Parametern besteht.
11. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Einteilung der Sprache in eine Anzahl von Unterbändern auf einem MEL-Frequenzbereich basiert.
12. Vorrichtung für eine robuste Merkmalsextrahierung für Spracherkennung in einer Umgebung mit Rauschen, wobei ein Sprachsignal segmentiert wird und gekennzeichnet ist durch spektrale Komponenten, mit
 - einer Filterbank (MEL) zum Aufspalten des Sprachsignals in eine Anzahl von spektralen Komponenten in L Unterbändern, mit $L = 1, 2, \dots$,
 - einer Abschätzungseinrichtung (ES) zum Abschätzen eines Rauschspektrums aus Segmenten, die nur Rauschen enthalten,
 - einem Subtrahierer (SSub) zum Ausführen einer spektralen Subtrahierung des abgeschätzten Rauschspektrums von dem entsprechenden kurzzeitigen Spektrum,
 - einem Rechner zum Berechnen einer Wahrscheinlichkeit für jedes kurzzeitige Spektrum, Rauschen zu enthalten, und
 - einer Interpolationseinrichtung (IP) zum Interpolieren der spektralen Komponente jedes kurzzeitigen Spektrums, wenn die berechnete Wahrscheinlichkeit, Sprache zu enthalten, für die kurzzeitige Spektralkomponente gering ist, dadurch gekennzeichnet, dass die Interpolationseinrichtung (IP) dazu ausgebildet ist, die Interpolation basierend auf spektralen Komponenten von mindestens einem angrenzenden Unterband und/oder auf mindestens einer in der Zeit vorhergehenden oder nachfolgenden spektralen Komponente des kurzzeitigen Spektrums auszuführen.
13. Vorrichtung nach Anspruch 12, mit einem Vergleicher zum Vergleichen der berechneten Wahrscheinlichkeit mit einer Schwelle, die die niedrige Wahrscheinlichkeit, Sprache zu enthalten, definiert.
14. Vorrichtung nach Anspruch 12 oder 13, mit einer Einrichtung (log) zum Logarithmieren jeder spektralen Komponente und einer diskreten Kosinustransformationseinrichtung (DCT) zum Ausführen einer Transformation.
15. Vorrichtung nach einem der Ansprüche 12 bis 14 mit
 - einem Codebuch (CB), das rauschlose Sprachspektren enthält,
 - einer Bestimmungseinrichtung zum Bestimmen einer Anzahl M von spektralen Abständen für jedes kurzzeitige Spektrum durch Bestimmen der spektralen Abstände zwischen dem kurzzeitigen Spektrum nach Subtrahierung und einer Anzahl M von Spektren, die rauschlose Sprachspektren enthalten, und
 - einer Bestimmungseinrichtung (MATCH), um für jedes kurzzeitige Spektrum mindestens ein rauschloses Sprache enthaltendes Spektrum mit dem kleinsten spektralen Abstand zu dem kurzzeitigen Spektrum zu bestimmen.
16. Vorrichtung nach einem der Ansprüche 12 bis 15, wobei die Filterbank eine MEL-Filterbank ist.
17. Verwendung der Vorrichtung nach einem der Ansprüche 12 bis 16 für eine Spracherkennung, um elektronische Geräte zu steuern.
18. Verwendung nach Anspruch 17, wobei das elektronische Gerät ein Telefon oder ein Mobiltelefon oder eine Zugangseinrichtung ist.
19. Computerprogrammprodukt, ladbar in einen Speicher eines elektronischen Geräts, das Software-Codeteile zum Ausführen der Schritte einer der Ansprüche 1 bis 11 umfasst.

20. Computerprogrammprodukt nach Anspruch 19, wobei das Computerprogramm auf einem computerlesbaren Medium gespeichert ist.

Es folgen 4 Blatt Zeichnungen

FIG. 1

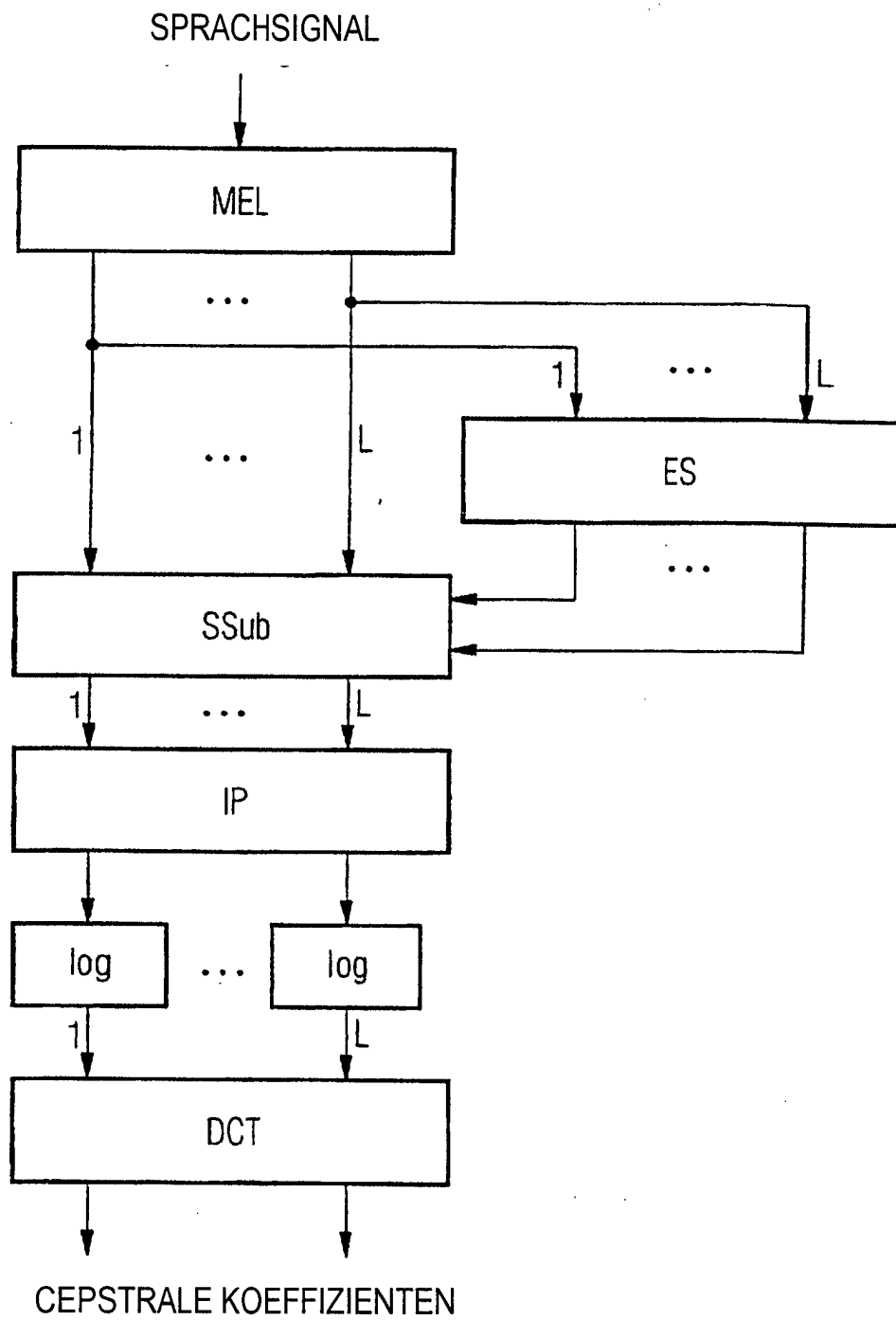


FIG. 2

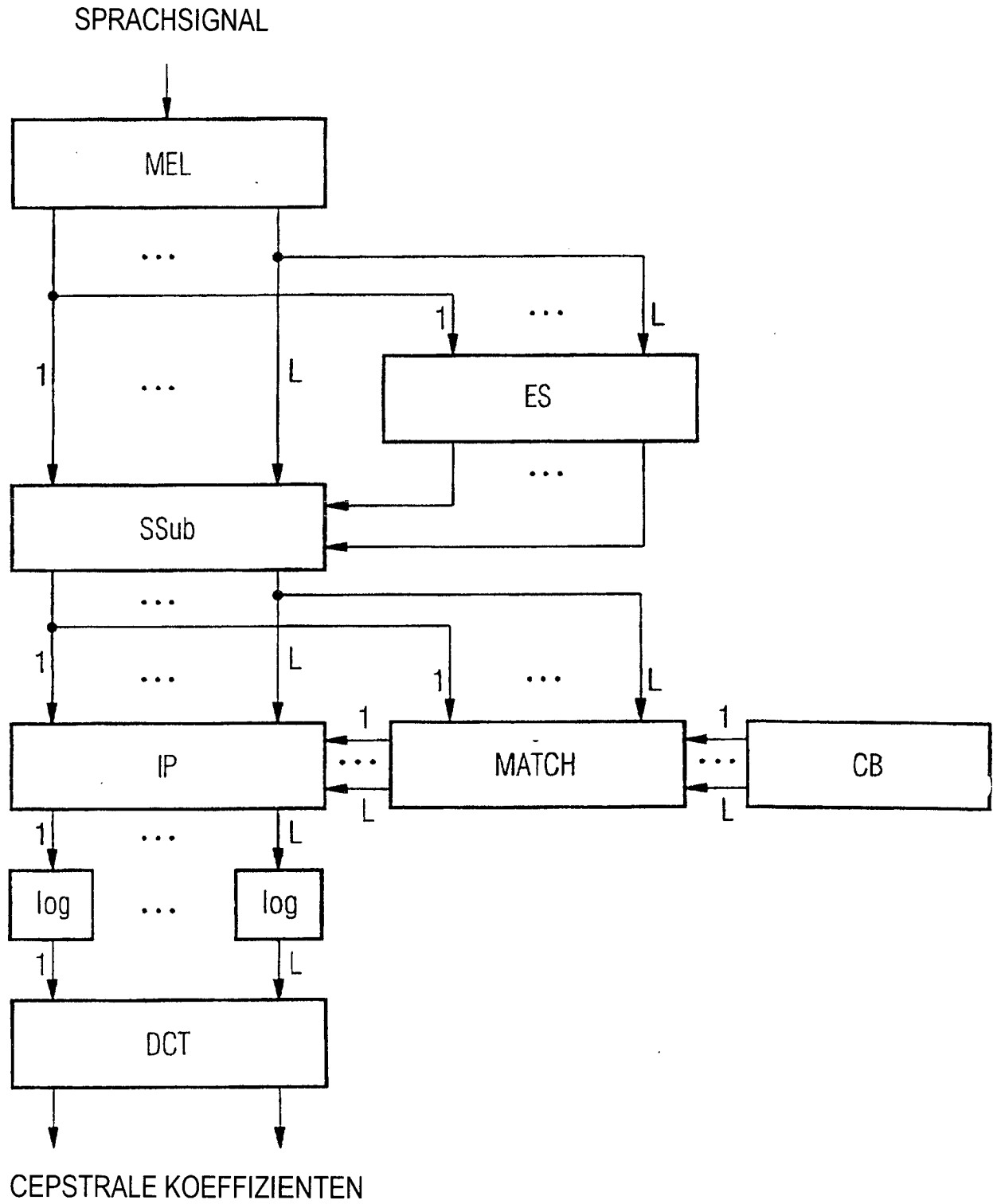


FIG. 3

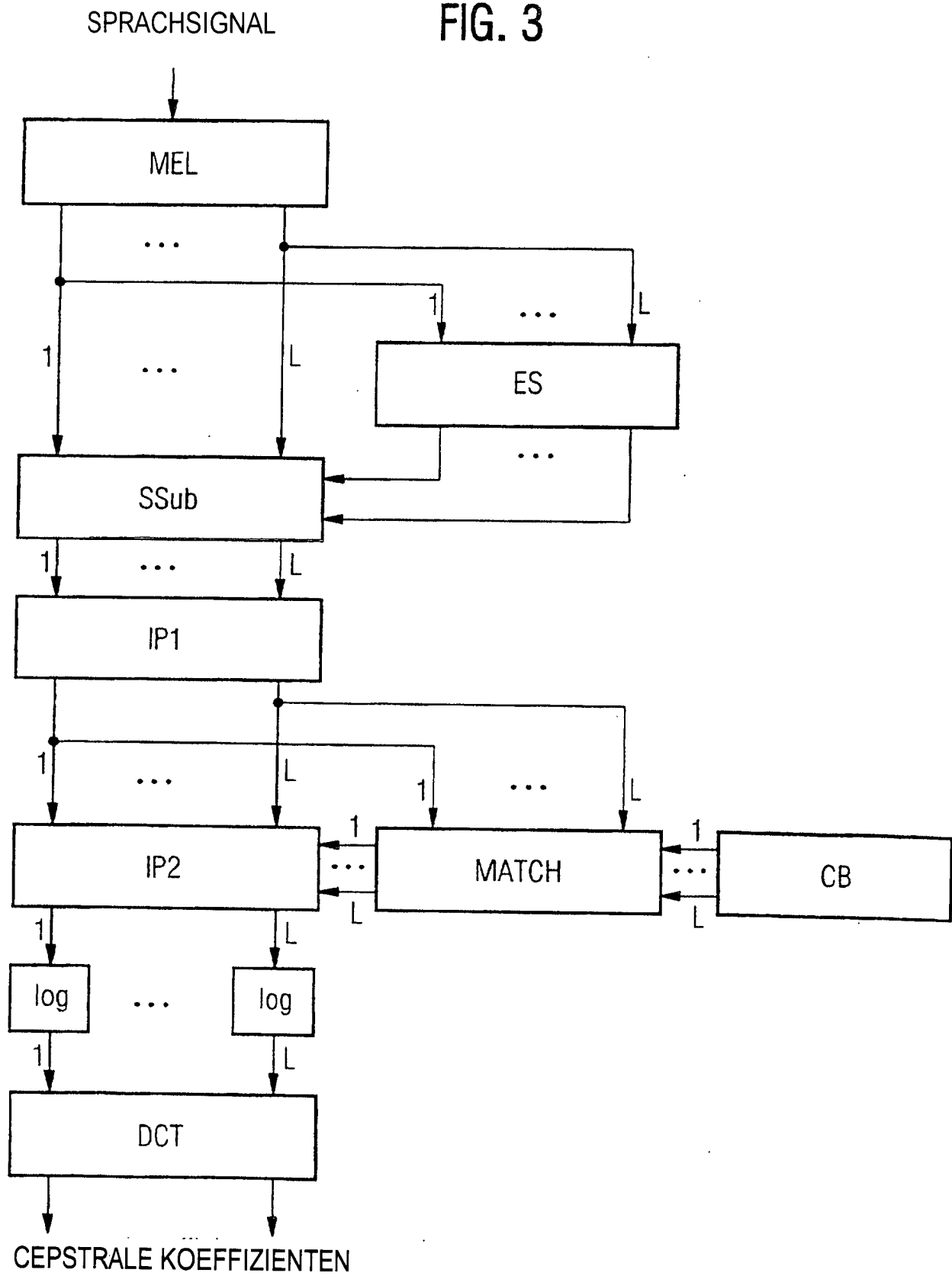


FIG. 4

