(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 836 176 A2

(12) EUROPÄISCHE PATENTANMELDUNG

(43) Veröffentlichungstag:
15.04.1998 Patentblatt 1998/16

(51) Int. Cl.[6]: **G10L 5/06**, G10L 7/08, G10L 9/06, G10L 9/18

(21) Anmeldenummer: **97116746.5**

(22) Anmeldetag: **25.09.1997**

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

(30) Priorität: **09.10.1996 DE 19641619**

(71) Anmelder:
NOKIA MOBILE PHONES LTD.
02150 Espoo (FI)

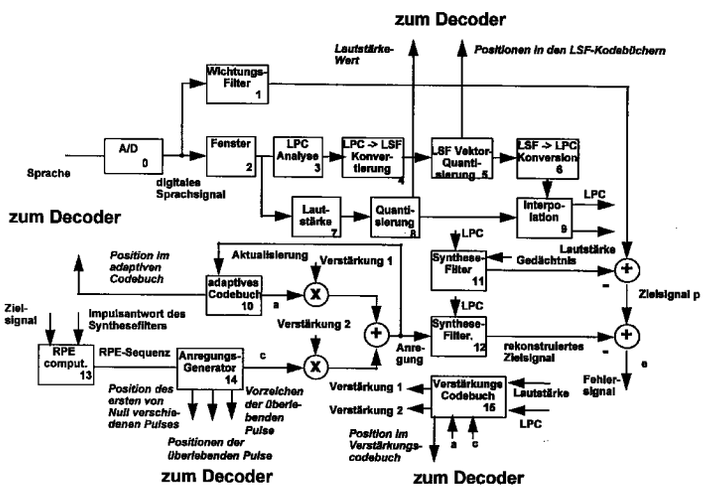
(72) Erfinder: **Görtz, Udo**
44797 Bochum (DE)

(74) Vertreter:
TER MEER STEINMEISTER & PARTNER GbR
Mauerkircherstrasse 45
81679 München (DE)

(54) Verfahren zur Synthese eines Rahmens eines Sprachsignals

(57) Die Erfindung beschreibt eine Möglichkeit, ein Sprachsignal ohne das aufwendige Absuchen eines stochastischen Codebuchs zu kodieren. Eine "ideale" RPE dient als Ausgangspunkt des Verfahrens. Die fünf größten RPE-Pulse werden betragsgleich quantisiert und durch ihre Vorzeichen unterschieden. Die restlichen RPE-Pulse werden Null gesetzt. Mit diesem einfachen und schnell durchführbaren Verfahren erzielt man die gleiche Sprachqualität wie mit deutlich aufwendigeren "closed-loop" Methoden.

Fig. 3



EP 0 836 176 A2

Beschreibung

Zeitbereichssprachkodierer, von denen im folgenden die Rede sein soll, arbeiten im wesentlichen alle nach dem gleichen Prinzip: Ein lineares Synthesefilter wird mit einem Anregungssignal derart beaufschlagt, daß dessen Ausgangssignal das zu übertragende Sprachsignal im Sinne eines festzulegenden Fehlermaßes möglichst gut approximiert. Oft besteht das Anregungssignal aus zwei Komponenten. Erstere soll die harmonischen, meist stimmhaften Sprachanteile nachbilden helfen, letztere die rauschförmigen Sprachanteile. Die eigentliche Lautformung, die beim realen Sprachtrakt durch den Rachen-Mund-Nasenraum geschieht, erfolgt durch das Synthesefilter. Die erzielbare Sprachqualität hängt dabei wesentlich von der Anregung des Synthesefilters ab.

Sogenannte Restsignalkodierer, wie zum Beispiel der im digitalen Mobilfunk derzeit eingesetzte RPE-LTP-Sprachkodierer, erreichen bei vergleichsweise geringer Komplexität und Bitraten deutlich oberhalb 10 kB/s nicht die heutzutage erforderliche Sprachqualität. Nach dem CELP-Prinzip (CELP = Code Excited Linear Prediction) arbeitende Analyse-durch-Synthese Sprachkodierer hingegen, die nicht das Sprachsignal selbst, sondern es beschreibende Parameter übertragen, erreichen im gleichen Bitratenbereich zwar eine deutlich bessere Sprachqualität als Restsignalkodierer, dies aber auf Kosten einer erheblich höheren Komplexität, wobei dieser Aufwand wesentlich durch das Absuchen von Codebüchern zur Bestimmung der stochastischen Anregung mitverursacht wird.

Wünschenswert wäre demnach, die Anregungsbestimmung zu vereinfachen, ohne die Sprachqualität zu reduzieren. Merkliche Vereinfachungen sind dann zu erwarten, wenn man das Absuchen von Codebüchern mittels eines guten, einfach bestimmbaren Vorauswahlkriteriums auf eine geringe Anzahl von Codevektoren beschränken oder sogar völlig auf das Absuchen des stochastischen Codebuchs verzichten kann, wenn es möglich wäre, die stochastische Anregung unmittelbar aus dem Sprachsignal abzuleiten, ohne die Übertragungsrate dadurch zu erhöhen. Bislang scheitert dieses Verfahren z. B. bei Bitraten um 13kB/s, weil es nicht gelingt, das Restsignal mit der verfügbaren Datenrate ausreichend gut zu quantisieren, weswegen man auch bei Zeitbereichs-Ansätzen im Bitratenbereich um 13 kB/s die Bestimmung der stochastischen Anregung nach dem CELP-Prinzip findet.

Aus der DE 90 067 17 U1 ist bereits eine Sprachsynthese bekannt, die ein RPE-Codewort verwendet.

Den Ausgangspunkt der Erfindung bildet eine "ideale RPE-Sequenz. Diese wird, wie seinerzeit von P. Kroon in seiner Dissertation "Time-domain coding of (near) toll quality speech at rates below 16 kb/s". Delft University of Technology, March 1985, angegeben, ermittelt. Zunächst wird daher auf die Bestimmung der RPE und die im RPE-LTP-Kodierer eingesetzte Variante dieses Anregungstyps eingegangen.

Berechnung der "idealen RPE"

Der zu bestimmende Anregungsvektor sei N Abtastwerte lang. Allgemein hat jeder dieser Abtastwerte eine eigene Amplitude und ein eigenes Vorzeichen. In der Praxis ist es aus Aufwandsgründen jedoch nötig, die Anzahl von Null verschiedener Pulse zu beschränken. Eine Möglichkeit, diese Aufwandsreduktion vorzunehmen, ist die sogenannte Regulärpuls-Anregung (Regular Pulse Excitation). Wenn beispielsweise jeder zweite Puls von Null verschieden ist, gibt es zwei Möglichkeiten, N/2 Pulse in einem Vektor der Länge N so zu plazieren, daß sich zwischen zwei von Null verschiedenen Pulsen jeweils eine Null befindet. Der erste-, dritte-, ..., Puls ist ungleich Null oder der zweite-, vierte-, ... Puls ist ungleich Null. Gibt es L von Null verschiedene Pulse, wobei L <= N, dann ist jeder (N/L)-te Puls ungleich Null und es gibt (N-(N/L)*(L-1)) Möglichkeiten, eine RPE-Sequenz zu erzeugen (beide Divisionen sind Integer-Divisionen). Der erste von Null verschiedene Puls kann sich auf (N-(N/L)*(L-1)) verschiedenen Positionen befinden. Der im Sinne eines zu approximierenden Zielvektors beste Satz von Amplituden berechnet sich wie folgt. Zunächst sei definiert:

p	Zielvektor, (1*N)-Matrix
h	Impulsantwort des Synthesefilters, (1*N)-Matrix
H	Impulsantwort-Matrix, (N*N)-Matrix
M	Verteilung der von Null verschiedenen Pulse im Anregungsvektor, (N*L)-Matrix
b	von Null verschiedene Pulsamplituden, (1*L)-Matrix
c	Anregungsvektor, (1*N)-Matrix
c'	gefilterte Anregung, (1*N)-Matrix
e	Differenz zwischen gefilterter Anregung und Zielsignal (Fehlervektor), (1*N)-Matrix
E	Fehlermaß, Skalar

Der Anregungsvektor ist gegeben durch

$$c = b \cdot M,$$

der gefilterte Anregungsvektor lautet

$$c' = b \cdot M \cdot H.$$

Der zu minimierende Fehler ist

$$E = p - c'.$$

Als Abstandsmaß dient die Summe der quadratischen Fehler.

$$E = e \cdot e^T.$$

Ersetzen von e in der Gleichung durch die vorgenannten Beziehungen liefert

$$E = p \cdot p^T - 2 \cdot H^T \cdot M^T \cdot b^T + b \cdot M \cdot H \cdot H^T \cdot M^T \cdot b^T.$$

Die partielle Ableitung nach den Komponenten des Pulsamplitudenvektors b

$$\frac{\partial E}{\partial b^T} = 0$$

führt auf den Satz bester Amplituden für die jeweilige Verteilung der von Null verschiedenen Pulse (Matrix M).

$$b^T = p \cdot H^T \cdot M^T \cdot (M \cdot H \cdot H^T \cdot M^T)^{-1}.$$

Die Impulsantwortmatrix hat nachfolgende Gestalt

$$H = \begin{matrix} h(0) & h(1) & h(2) & h(3) & .. & h(N-1) \\ 0 & h(0) & h(1) & h(2) & .. & h(N-2) \\ 0 & 0 & h(0) & h(1) & .. & h(N-3) \\ 0 & 0 & 0 & h(0) & .. & h(n-4) \\ .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & h(0) \end{matrix}.$$

Für den Fall L = N/2 ist M durch die beiden nachfolgenden Matrizen gegeben

$$M^{(1)} = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & .. & .. & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & .. & .. & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & .. & .. & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .. & 1 & 0 \end{matrix}$$

$$M^{(2)} = \begin{matrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & .. & .. & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & .. & .. & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & .. & .. & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .. & .. & 1 \end{matrix}$$

Allgemein befindet sich bei einer RPE in jeder Zeile von M nur ein einziges von Null verschiedenes Element, wobei die n-te Zeile die Position des n-ten Pulses der RPE angibt. Wenn es m Möglichkeiten gibt, mittels L von Null verschiedener Pulse eine RPE zu bilden, nimmt auch die Matrix M m verschiedene Gestalten an. Die "ideale RPE-Sequenz" ist diejenige, die gemäß obiger Rechnung das Fehlermaß E minimiert.

Bestimmung der RPE beim RPE-LTP-Kodierer

Die zuvor beschriebene Bestimmung der RPE erfordert das Lösen eines verkoppelten, linearen Gleichungssystems. Als der RPE-LTP-Kodierer definiert wurde, stand nicht genug Rechenleistung zur Verfügung, um den Algorithmus in einem für den Massenmarkt vorgesehenen Mobiltelefon zu implementieren. Deshalb kommt eine vereinfachte RPE-Variante zum Einsatz. Nach Dekorrelationsfilterung des zu übertragenden Sprachsignals verbleibt ein Restsignal, das im interessierenden Frequenzbereich ein theoretisch weißes Spektrum aufweist. Wenn alle Spektralkomponenten von gleicher Intensität sind, kann man auf die Übertragung des gesamten Bandes verzichten, es genügt, das Basisband zu übertragen, welches man durch Unterabtastung des Restsignals nach vorheriger Tiefpaß-Filterung gewinnt. Dadurch reduziert sich die Anzahl der zu übertragenden Pulse und damit die Übertragungsrate. Decodierseitig kann durch Interpolationsfilterung das nicht übertragene hohe Band zurückgewonnen werden.

Das Restsignal war bei der Berechnung der"idealen RPE" im vorigen Abschnitt nicht explizit erforderlich, so daß beide Verfahren zunächst recht unterschiedlich aussehen. Tatsächlich jedoch ist das beim RPE-LTP-Kodierer eingesetzte Verfahren als Näherung des zuvor beschriebenen Verfahrens interpretierbar. Die oben beschriebene RPE-Berechnung läßt sich gleichwertig durchführen, wenn man sie unter Einbeziehung des Restsignals in folgende Schritte untergliedert:

- Filterung des Restsignals r(n) mit einem FIR-Filter F(z) der Länge N→y(n),
- Abtastung (Dezimierung) des gefilterten Restsignals → z(n),
- Erhöhung der Abtastrate von z(n) auf die ursprüngliche → c(n),
- Synthesefilterung dieses Signals → v(n),
- Berechnung des Synthesefehlers → E,
- Minimierung des Synthesefehlers durch geeignete Wahl der Koeffizienten von F(z) → $\{f_0, f_1, ..., f_{N-1}\}$.

Gesucht sind also diejenigen N Filterkoeffizienten, die bei Filterung und Abtastung des gegebenen Restsignals den minimalen Fehler verursachen. In Matrixschreibweise ergibt sich:

$$y = f \cdot R$$

$$z = y \cdot M^t$$

$$c = z \cdot M$$

$$v = c \cdot H = f \cdot R \cdot M^t \cdot M \cdot H$$

$$E = p \cdot p^t - 2p \cdot v^t + v \cdot v^t$$
$$= p \cdot p^t - 2p \cdot H^t \cdot M^t \cdot M \cdot R^t \cdot f^t + f \cdot R \cdot M^t \cdot M \cdot H \cdot H^t \cdot M^t \cdot M \cdot R^t \cdot f^t$$

$$\frac{\partial E}{\partial f^t} = 0 \Rightarrow f \cdot R \cdot M^t \cdot M \cdot H \cdot H^t \cdot M^t \cdot M \cdot R^t = p \cdot H^t \cdot M^t \cdot M \cdot R^t$$

$$\text{sei } A = R \cdot M^t \cdot M \cdot H$$

$$f \cdot A \cdot A^t = p \cdot A^t$$

wobei

f (1xN)-Matrix.
R (NxN)-Matrix,
M (NpxN)-Matrix,
p (1xN)-Matrix

und

$$R = \begin{pmatrix} r(0) & r(1) & .. & r(N\text{-}1) \\ r(\text{-}1) & r(0) & .. & r(N\text{-}2) \\ .. & .. & .. & .. \\ r(\text{-}(N\text{-}1)) & r(\text{-}(N\text{-}2)) & .. & r(0) \end{pmatrix}.$$

Die Werte r(0), r(1), ..., r(N-1) stellen das aktuelle Restsignal dar, r(-(N- 1)), r(-(N-2)), ..., r(-1) sind Werte aus der Signalvergangenheit.

M ist beispielhaft für den Fall angegeben, daß der erste von Null verschiedene Puls auf der ersten Position im RPE-Vektor sitzt und jeder zweite Puls von Null verschieden ist:
$\{a_0, 0, a_1, 0, a_3, 0, .. a_{N-2}, 0\}$. Allgemein wird M konstruiert wie oben angegeben.

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & .. & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & .. & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & .. & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & .. & 1 & 0 \end{pmatrix}$$

Aus $f \cdot A \cdot A' = p \cdot A'$ kann nun nicht unmittelbar durch Nachmultiplizieren beider Gleichungsseiten mit $(A \cdot A')^{-1}$ der Koeffizientenvektor f bestimmt werden, weil die Inverse aufgrund der Konstruktion von A unabhängig vom Restsignal und von der Impulsantwort des Synthesefilters nicht existiert, da die Determinante von A stets Null ist: Ist A symmetrisch, gilt $\det(A) = \det(A^t)$. Weiterhin ist $\det(A \cdot B) = \det(A) \cdot \det(B)$ und $\det(A \cdot B) \neq 0$ und $\det(B) \neq 0$. R, M^tM und H sind quadratische Matrizen gleicher Dimension. Die Restsignalmatrix R kann bei ausreichender Sprachaktivität als invertierbar angenommen werden. Die Impulsantwortmatrix H ist ebenfalls invertierbar, weil sie eine Dreiecksmatrix ist, auf deren Hauptdiagonale stets von Null verschiedene Elemente stehen. M^tM hingegen ist nie invertierbar, sie enthält Nullspalten und Nullzeilen. Wenn z. B. der zweite, vierte, sechste, ... Puls in der RPE Null ist, dann befinden sich in der zweiten, vierten, sechsten, ... Spalte und Zeile in M^tM nur Nullen. Fortgesetzte Anwendung von $\det(A \cdot B) = \det(A) \cdot$

det(B) ergibt det(A • A') = 0 ∀ R, H.

Ein FIR-Filter F(z) der Länge N, mit dem man das Restsignal vor seiner Abtastung filtern müßte, um den kleinstmöglichen Synthesefehler zu erhalten, ist durch Vorgabe der Positionierung der von Null verschiedenen Pulse, das Synthesefilter, das Zielsignal und das Restsignal nicht eindeutig bestimmt. Werden nach Filterung des Restsignals m Pulse willkürlich nullgesetzt, fehlen zur Bestimmung der N Filterkoeffizienten m linear unabhängige Gleichungen. Der Rang von A ist nur so groß wie die Anzahl von null verschiedenere Pulse.

Bei der Berechnung der "idealen RPE" (siehe oben) wird ebenfalls das hier eingesetzte Fehlermaß benutzt. Die Minimierung des Fehlers muß bei beiden Verfahren auf den gleichen resultierenden Synthesefehler führen, denn durch das gewählte Fehlerkriterium ist sichergestellt, daß es, von Randextrema abgesehen, nur ein Minimum gibt. Die Anregungssignale der beiden exakt gleichen Synthesefilter müssen also in beiden Fällen exakt übereinstimmen: der Vektor z aus diesem Abschnitt und der Vektor b aus dem vorigen Abschnitt sind demnach gleich. Setzt man in

$$f \cdot R \cdot M^t \cdot M \cdot H \cdot H^t \cdot M^t \cdot M \cdot R^t = p \cdot H^t \cdot M^t \cdot M \cdot R^t$$

$$b = f \cdot R \cdot M^t$$

und nachmultipliziert mit $R \cdot M^t$, so erhält man

$$b \cdot (M \cdot H \cdot H^T \cdot M^T) = p \cdot H^T \cdot M^T,$$

wenn man die Invertierbarkeit von M • R' • R • M' voraussetzt, also die Gleichungen zur Berechnung der "idealen RPE". Das Gleichungssystem in f läßt sich formal in das System in b überführen. Umgekehrt läßt sich das System in b in das System in f überführen, wenn man statt b fRM^t einsetzt und die Gleichung mit MR^t nachmultipliziert.

Beispielhaft sei der Fall für N/2 von null verschiedene Pulse angegeben, wobei sich der erste von null verschiedene Puls auf der ersten Position im RPE-Vektor befindet.

$$b = z = f \cdot R \cdot M'$$

$$= f \cdot \begin{pmatrix} r(0) & r(1) & r(2) & r(3) & .. & r(N-1) \\ r(-1) & r(0) & r(1) & r(2) & .. & r(N-2) \\ r(-2) & r(-1) & r(0) & r(1) & .. & r(N-3) \\ r(-3) & r(-2) & r(-1) & r(0) & .. & r(N-4) \\ .. & ,, & .. & .. & .. & .. \\ r(-(N-1)) & r(.(N-2)) & r(-(N-3)) & r(-(N-4)) & .. & r(0) \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 0 & 0 & 0 & .. & 0 \\ 0 & 1 & 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 1 & 0 & 0 & .. & 0 \\ .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & 0 & .. & 1 \end{pmatrix}$$

$$= \begin{pmatrix} f_0 & f_1 & f_2 & .. & f_{N-1} \end{pmatrix} \cdot \begin{pmatrix} r(0) & r(2) & r(4) & .. & .. \\ r(-1) & r(1) & r(3) & .. & .. \\ r(-2) & r(0) & r(2) & .. & .. \\ .. & .. & .. & .. & .. \\ r(-(N-1)) & r(-(N-3)) & r(-(N-5)) & .. & .. \end{pmatrix}$$

Als Gleichungssystem in *f* geschrieben folgt

$$\begin{pmatrix} r(0) & r(-1) & r(-2) & .. & r(-(N-1)) \\ r(2) & r(1) & r(0) & .. & r(-(N-3)) \\ r(4) & r(3) & r(2) & .. & r(-(N-5) \\ .. & .. & .. & .. & .. \\ r(N-2) & r(N-3) & r(N-4) & .. & r(-1) \end{pmatrix} \cdot \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ .. \\ f_{N-1} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ .. \\ b_{N/2-1} \end{pmatrix}.$$

Zur Berechnung der N Filterkoeffizienten stehen nur N/2 Gleichungen zur Verfügung. Das System ist mit beliebig vielen verschiedenen Koeffizientenvektoren f erfüllbar. Weil es zur Minimierung des Synthesefehlers aber genügt, das Gleichungssystem irgendwie zu erfüllen, wählt man sinnvollerweise einen "bequemen" Koeffizientensatz für die (N-m) wählbaren Koeffizienten, m = Rang (A), multipliziert mit der vorstehenden Matrix und schafft die entstehenden Konstanten auf die rechte Gleichungsseite. Damit ist das verbleibende System reduzierter Ordnung eindeutig lösbar.

Beim RPE-LTP Kodierer wird das Filter F(z) nicht neu berechnet, wenn sich Zielsignal und Impulsantwort des Synthesefilters verändert haben. Die Filterkoeffizienten sind konstant. Der Betragsfrequenzgang dieses Filters hat den Verlauf eines als "typisch" angesehenen Sprachspektrums. Hierbei handelt es sich um einen Tiefpaß mit weichem Übergang vom Durchlaß- in den Sperrbereich. Die Grenzfrequenz liegt im Bereich um 1300 Hz. Das Filter F(z) kann als Tiefpaßfilter vor dem Abtaster angesehen werden. Allerdings entstehen durch den weichen Übergang vom Durchlaß- in den Sperrbereich Alias-Komponenten. Insgesamt stellt diese Vorgehensweise eine recht grobe Näherung dar. Der Betragsfrequenzgang von F(z) variiert nämlich nicht unerheblich.

In der Praxis läßt sich das Sprachsignal durch lineare Dekorrelationsfilterung nicht völlig dekorrelieren. Das Spektrum ist deshalb nicht weiß, sondern gegenüber dem Ursprungsspektrum lediglich flacher und insgesamt von geringerer Intensität. Die Annahme, bereits durch Kenntnis des Basisbandes das gesamte Band zu kennen, ist eine grobe Näherung und verursacht besonders bei Sprechern mit hohen Stimmen einen nicht unerheblichen Fehler, der beim RPE-LTP-Kodierer deutlich in Erscheinung tritt, weil nur das unterste Drittel des Gesamtbandes übertragen wird, was der Unterabtastung mit dem Faktor 3 entspricht.

Dennoch werden zur Übertragung der stochastischen Anregung 45 Bit/5ms benötigt, entsprechend 9 kB/s. Eine weniger genaue Quantisierung der einzelnen Pulse führt auf eine klar schlechtere Sprachqualität, eine Verbesserung derselben ist durch Reduktion des Unterabtastungsfaktors möglich, erhöht aber die Übertragungsrate. Dieser Weg scheidet zur Verbesserung des RPE-LTP-Kodierers deshalb aus. Neben den durch die Art und Weise, wie die RPE ermittelt wird, verursachten Qualitätseinbußen reduzieren weitere beim RPE-LTP-Kodierer aus Aufwandsgründen seinerzeit erforderlichen Beschränkungen die Qualität. So findet ein Synthesefilter von nur achter Ordnung Anwendung. Die Langzeitprädiktion erfolgt mittels eines einstufigen Prädiktors. Die zugehörige Verstärkung wird grobstufig skalar quantisiert.

Versuche, den RPE-LTP-Kodierer zu verbessern, schienen deshalb bei der Suche nach einem Algorithmus für einen deutlich verbesserten Sprachkodierer für das digitale Mobilfunknetz nicht sinnvoll. Diese weit verbreitete Annahme hat dazu geführt, daß auch der Anregungstyp RPE für moderne Zeitbereichskodierer de facto nicht mehr in Betracht gezogen wurde und die nach dem RPE-LTP-Kodierer entwickelten Zeitbereichssprachkodierer im wesentlichen nach dem CELP-Prinzip arbeiten und ihre stochastische Anregung durch aufwendiges Absuchen von trainierten oder algebraisch konstruierten Codebüchern ermittelten.

CELP-Prinzip

Abbildung 1 zeigt das CELP-Prinzip, wie es typischerweise eingesetzt wird. Ein zu approximierendes Zielsignal wird durch Absuchen (mindestens) zweier Codebücher nachgebildet. Dabei unterscheidet man zwischen einem adaptiven Codebuch (a2), dessen Aufgabe die Nachbildung der harmonischen Sprachanteile ist und einem/oder mehreren stochastischen Codebüchern (a4), die zur Synthese der nicht aus Prädiktion gewinnbaren Sprachanteile dienen. Das adaptive Codebuch (a2) ändert sich abhängig vom Sprachsignal, während das stochastische Codebuch (a4) zeitinvariant ist. Die Suche nach den besten Codevektoren läuft derart ab, daß nicht eine gemeinsame, d. h. gleichzeitige Suche in den Codebüchern stattfindet, wie es für eine optimale Auswahl der Codevektoren erforderlich wäre, sondern aus Aufwandsgründen zunächst das adaptive Codebuch (a2) durchsucht wird. Ist der gemäß des Fehlerkriteriums beste Codevektor gefunden, subtrahiert man dessen Beitrag zum rekonstruierten Zielsignal vom Zielvektor (Zielsignal) und erhält den durch einen Vektor aus dem stochastischen Codebuch (a4) noch zu rekonstruierenden Teil des Zielsignals. Die Suche in den einzelnen Codebüchern erfolgt nach dem gleichen Prinzip. In beiden Fällen wird der Quotient aus dem Quadrat der Korrelation des gefilterten Codevektors mit dem Zielvektor und der Energie des gefilterten Zielvektors für alle Codevektoren berechnet. Derjenige Codevektor, der diesen Quotienten maximiert, wird als bester

Codevektor angesehen, der das Fehlerkriterium (a5) minimiert. Die vorgeschaltete Fehlergewichtung (a6) gewichtet den Fehler entsprechend der Charakteristik des menschlichen Gehörs. Seine Position wird an den Decoder übertragen. Durch die Berechnung des genannten Quotienten wird für jeden Codevektor implizit der richtige Verstärkungsfaktor (Verstärkung 1, Verstärkung 2) ermittelt. Nachdem der beste Kandidat aus beiden Codebüchern bestimmt ist, kann man durch eine gemeinsame Optimierung der Verstärkung den qualitätsmindernden Einfluß des sequentiell durchgeführten Codebuchsuchens reduzieren. Dabei gibt man den ursprünglichen Zielvektor erneut vor und berechnet zu den nun ausgewählten Codevektoren passend die besten Verstärkungen, die sich meist geringfügig von denen unterscheiden, die während des Codebuchsuchens ermittelt wurden.

Das CELP-Prinzip ist dadurch gekennzeichnet, daß zum Auffinden des besten Codevektors jeder Kandidatenvektor einzeln gefiltert (a3) und mit dem Zielsignal verglichen werden muß. Dieser Vorgang verursacht trotz des sequentiellen Absuchens beider Codebücher einen erheblichen Aufwand, der bei der in der ersten CELP-Veröffentlichung vorgeschlagenen Codebuchgröße von 1024 Vektoren nicht einmal auf leistungsfähigen Fließkomma-Signalprozessoren in Echtzeit zu bewältigen ist. Der Schwerpunkt der Arbeiten an CELP-Kodierern beschäftigte (und beschäftigt) sich deshalb mit der Frage, wie man die Vorteile des CELP-Prinzips nutzen kann, ohne den Nachteil des hohen Rechenaufwands inkaufnehmen zu müssen.

Der Erfindung liegt die Aufgabe zugrunde, ein Verfahren zur Sprachsynthese zu schaffen, bei dem man im angegebenen Bitratenbereich auf das Absuchen stochastischer Codebücher völlig verzichten kann, ohne die Sprachqualität zu beeinträchtigen und ohne die Übertragungsrate verglichen mit dem Einsatz stochastischer Codebücher zu erhöhen.

Die Lösung der gestellten Aufgabe ist im Anspruch 1 angegeben. Vorteilhafte Weiterbildungen der Erfindung sind den Unteransprüchen zu entnehmen.

Gemäß der Erfindung wird ein Verfahren zur Synthese eines Rahmens eines Sprachsignals in einem Sprachcoder/-decoder, z. B. vom CELP-Typ, zur Verfügung gestellt, bei dem einem Synthesefilter des Sprachcoders ein Anregungsvektor zugeführt wird, der aus einer adaptiven Anregungskomponente a und einer stochastischen Anregungskomponente c besteht, wobei die stochastische Anregungskomponente c durch folgende Parameter gebildet wird, die einer zuvor errechneten idealen RPE-Sequenz entnommen werden:

a) Die Position des ersten von Null verschiedenen Pulses in der idealen RPE-Sequenz,

b) die Positionen einer vorgewählten Anzahl von betragsgrößten Pulsen der idealen RPE-Sequenz,

c) die Amplituden dieser betragsgrößten Pulse, und

d) die Vorzeichen dieser betragsgrößten Pulse,
und wobei diese Parameter ferner zum Sprachdecoder übertragen werden, um auch dort die stochastische Anregungskomponente c zu erzeugen.

Fast alle Zeitbereichskodierer besitzen heutzutage eine ähnliche Struktur. Die Synthesefilter-Koeffizienten eines Filters zehnter Ordnung werden oft in Reflexionsfaktoren oder in "Line Spectrum Frequencies" (LSFs) umgerechnet und (vektor-) quantisiert. Die Anregung des Synthesefilters setzt sich aus der gewichteten Überlagerung der adaptiven Anregung und der stochastischen Anregung zusammen. Beide Anregungsbestandteile werden sequentiell durch eine mehr oder weniger suboptimal durchgeführte Codebuchsuche bestimmt, wobei die adaptive Anregung, also der durch Wiederholung alter Anregungswerte gewinnbare Anregungsbestandteil, zuerst ermittelt wird. Der Grad der Suboptimalität beim Absuchen der Codebücher entscheidet über Rechenaufwand und Sprachqualität. Das Ziel ist, möglichst wenige Codevektoren innerhalb der Analyse-durch-Syntheseschleife zu untersuchen, damit der Rechenaufwand begrenzt wird. Dazu ist eine einfache aber gute Vorauswahl der innerhalb der Schleife zu untersuchender Codevektoren erforderlich. Die Vektorquantisierung der Anregung erlaubt einerseits die Reduktion der Übertragungsrate und verursacht andererseits bei gleicher Übertragungsrate wie eine Skalarquantisierung einen geringeren Quantisierungsfehler.

Das hier beschriebene neue Verfahren nach der Erfindung zur Ermittlung der stochastischen Anregung unterscheidet sich von diesem Ansatz erheblich. Es wird kein Vorauswahlkriterium benutzt und die stochastische Anregung wird auch nicht vektorquantisiert. Es handelt sich auch nicht um eine Skalarquantisierung im herkömmlichen Sinne, bei der man bestrebt ist, die übertragenen Pulse möglichst genau zu quantisieren. Das wesentliche Qualitätsproblem beim RPE-LTP-Kodierer ist, daß die RPE eine um den Faktor Drei unterabgetastete Version des dekorrelierten Sprachsignals ist. Selbst eine exakte Quantisierung der RPE-Pulse erhöht die Qualität nur unwesentlich. Eine Reduktion des Unterabtastungsfaktors auf zwei erhöht die Qualität zwar merklich, bedingt aber auch eine erheblich höhere Übertragungsrate. Weil die Übertragungsrate des Kodierers aber nicht steigen darf, scheidet dieser Weg aus.

Beim RPE-LTP-Kodierer kommt eine recht grobe Langzeitprädiktion zum Einsatz, so daß auch die RPE noch harmonische Sprachanteile beitragen muß. In heutigen Analyse-durch Synthese-Kodierern hingegen erfolgt die Langzeit-

prädiktion erheblich genauer als im RPE-LTP-Kodierer, so daß die verbleibende stochastische Anregung tatsächlich im wesentlichen rauschförmigen Charakter besitzt und die richtige Phasenlage der stochastischen Anregung wesentlich wichtiger ist als eine genaue Quantisierung der Amplituden. Diesem Umstand ist es auch zu verdanken, daß ACELPs (Algebraic Code Exicited Linear Prediction) mit Codeworten, die nur eine oder zwei Amplitudenstufen zulassen, gute Ergebnisse liefern. Bei einem ACELP beantwortet eine Codebuchsuche die Frage, auf welchen Pulspositionen Pulse plaziert werden müssen. Die Beantwortung dieser Frage verursacht allgemein einen erheblichen Aufwand, auch wenn die Codeworte nur aus Nullen und Einsen bestehen und die Vorzeichen schon durch suboptimale Methoden vorab ermittelt wurden.

Dieser Aufwand ist zumindest z. B. im Bitratenbereich um 13 kB/s entbehrlich. Die Positionen, auf denen von Null verschiedene Pulse sitzen müssen, lassen sich auch ohne hörbare Qualitätseinbuße einer mit erheblich geringerem Aufwand berechneten "idealen RPE'" entnehmen.

Um den Rechenaufwand bei der Lösung des Gleichungssystems zur Bestimmung der "idealen" RPE zu reduzieren, kann man erfindungsgemäß die stochastische Anregung z. B. alle 2,5 ms neu bestimmen. Das entspricht einer Unterrahmenlänge von N = 20 Abtastwerten. In diesem Fall ist ein Gleichungssystem zehnter Ordnung zu lösen. Die sich ergebenden Amplituden der "idealen RPE" werden nun betrachtet, um die "uberlebenden Pulse" zu finden. Zumindest die Hälfte der RPE-Amplituden sind relativ klein. Nur einige wenige Amplituden besitzen eine große Amplitude. Es ist ausreichend, die großen Amplituden überleben zu lassen, sie z. B. betragsmäßig gleich zu machen und dann nur noch deren Position und Vorzeichen an den Decoder zu übertragen. Drei bis fünf der betragsgrößten Pulse genügen für gute/sehr gute Sprachqualität. Die auf diesem Weg erhaltene Anregung hat die Gestalt einer Pseudo-MPE (Multi Pulse Excitation).

Die Erfindung wird nachfolgend unter Bezugnahme auf die Zeichnung näher beschrieben. Es zeigen:

Figur 1 die Darstellung des CELP-Prinzips, wie es herkömmlicherweise eingesetzt wird;
Figur 2 eine Darstellung zur erfindungsgemäßen Erzeugung einer stochastischen Anregung (Figur 2b) in Abhängigkeit einer idealen RPE Sequenz (Figur 2a);
Figur 3 ein beim erfindungsgemäßen Verfahren verwendeter Sprachcoder; und
Figur 4 ein beim erfindungsgemäßen Verfahren verwendeter Sprachdecoder.

Die Abbildung 2 zeigt, wie bei einem Ausführungsbeispiel der Erfindung aus einer idealen RPE nach Figur 2a eine stochastische Anregung nach Figur 2b erzeugt wird. Aus der idealen RPE werden dazu folgende Parameter bzw. Größen entnommen:

- Die Position des ersten von Null verschiedenen Pulses in der idealen RPE;
- die Positionen der überlebenden Pulse, also derjenigen Pulse, deren Amplitude größer als eine vorgegebene Schwelle ist; und
- die Vorzeichnen dieser überlebenden Pulse.

Dabei sind vorzugsweise die Amplituden der überlebenden Pulse alle gleich bzw. normiert, z. B. auf Eins, so daß die Vorzeichenangabe auch gleich der Amplitudenangabe ist, was dem Coder mitgeteilt werden muß.

Zur Bestimmung der Anregung ist eine exakte Bestimmung der Amplituden durch Lösung eines verkoppelten Gleichungssystems nicht unbedingt erforderlich. Die entsprechenden Pulspositionen und Vorzeichen lassen sich auch einem suboptimal gelösten System entnehmen. Hier kommen alle Möglichkeiten in Betracht, bei denen Amplituden, Positionen und Vorzeichen der großen Pulse weitgehend erhalten bleiben. Eine dieser Möglichkeiten ist, die Pulse sequentiell zu ermitteln, indem man zunächst den ersten Puls bestimmt, dessen Beitrag am rekonstruierten Zielsignal vom Zielsignal p subtrahiert, dann den zweiten Puls berechnet, usw.

Das beschriebene Verfahren zur Gewinnung einer Pseudo-MPE aus einer "idealen" RPE ist eine kombinierte "closed-loop" / "open-loop" Methode. Die "ideale" RPE ist bezüglich des zu approximierenden Zielsignals optimal (closed-loop"), wahrend die Quantisierung der "idealen" RPE nicht mit Blick auf dieses Zielsignal erfolgt, sondern von den Positionen der maximalen Pulse im RPE-Vektor abhängt ("open-loop"). Dadurch wird der Rechenaufwand zur Quantisierung vernachlässigbar klein. Das in Sprachkodierern in diesem Bitratenbereich ansonsten übliche sehr aufwendige Absuchen stochastischer Codebücher entfällt.

Die Anwendung dieses Verfahrens wird anhand eines beispielhaften Sprachkodierers im folgenden demonstriert, ist aber nicht auf diesen beschränkt.

Abbildung 3 zeigt den Sprachkodierer. Nach Abtastung des analogen Sprachsignals im Block 0 wird das digitale Sprachsignal einer Fensterung 2 unterzogen, bevor die LPC-Analyse 3 zur Ermittlung der Koeffizienten des Synthesefilters 11, 12 durchgeführt wird. Zweck der Fensterung ist, die Abschneideeffekte durch die endliche Länge des LPC-Analyseintervalls zu verringern. Das Synthesefilter ist in zwei Blöcke aufgeteilt, wobei Block 11 den Ausschwinganteil des Filters aufgrund der Werte im Filtergedächtnis darstellt und Block 12 das Synthesefilter mit nullgesetztem Gedächt-

nis zu Beginn einer jeden Filterung. Die Überlagerung beider Ausgangssignale ist das Ausgangssignal des Synthesefilters. Vor ihrer Quantisierung 5 erfolgt die Umrechnung 4 der direkten Filterkoeffizienten in "Line-Spectrum-Frequencies" (LSF), die günstigere Eigenschaften bezüglich der Quantisierung aufweisen als direkte Filterkoeffizienten. Die LSFs werden dann quantisiert 5 und die Positionen in den entsprechenden LSF-Code-büchern werden an den Decoder übertragen. Das gefensterte digitale Sprachsignal wird durch einen Lautstärkewert charakterisiert 7, der proportional der im Signal enthaltenen Energie ist. Dieser Wert wird logarithmisch quantisiert 8 und ebenfalls an den Decoder übertragen. Im Kodierer werden die quantisierten Werte der LSFs und der Lautstärke benutzt, ebenso wie im Decoder. Vor ihrer Verwendung werden die quantisierten LSFs wieder in direkte Filterkoeffizienten umgerechnet 6 und wie die Lautstärke mit den entsprechenden Werten des letzten Analyseintervalls linear interpoliert 9. Die vorgenannten Rechnungen erfolgen einmal pro Analyserahmen, der hier 160 Abtastwerte entsprechend 20 ms lang ist.

Die nachfolgenden Berechnungen erfolgen acht mal je Analyserahmen, also alle 2,5 ms. Der erste Schritt ist die Berechnung des aktuellen Zielsignals, welches nachgebildet werden soll. Dazu subtrahiert man zunächst den Ausschwinganteil des Synthesefilters 11 aufgrund vorangegangener Anregungen vom wichtungsgefilterten digitalen Sprachsignal aus Block 1. Die Wichtungsfilterung betont für das Gehör wichtige Bereiche im Sprachsignal. Nun erfolgt die Bestimmung der adaptiven Anregung a. Sie wird dem adaptiven Codebuch 10 entnommen, das eine bestimmte Anzahl vergangener Anregungswerte des Synthesefilters enthält. Dieses Codebuch 10 verändert nach jedem Unterrahmen seinen Inhalt. Es wird derjenige Anregungsvektor a aus dem adaptiven Codebuch ausgewählt, dessen gefilterte und mit einem Verstärkungsfaktor (Verstärkung 1) skalierte Version den im Sinne eines willkürlich gewählten Fehlerkriteriums, hier quadratischer Fehler, kleinsten Abstand zum Zielvektor p aufweist. Nach Bestimmung der gefilterten und skalierten adaptiven Anregung a wird diese vom Zielvektor p subtrahiert. Es verbleibt der durch den stochastischen Anregungsvektor c zu minimierende Restfehler. Dieser Anregungsvektor c wird nun keinem Codebuch entnommen, wie es bei solchen Kodierern normalerweise üblich ist, sondern aus dem Zielsignal p und der Impulsantwort h des Synthesefilters unmittelbar berechnet: Aus den genannten Signalen wird, wie oben erläutert, die "ideale" RPE in Block 13 bestimmt. Der Anregungsgenerator 14 bestimmt die Positionen z. B. der fünf betragsgrößten Pulse und dessen Vorzeichen und setzt die restlichen BPE-Pulse zu Null. Die überlebenden Pulse erhalten den gleichen Betrag und werden nur noch durch ihre Vorzeichen unterschieden. Nachdem beide Anregungsteilvektoren (adaptiver a und stochastischer c Anregungsvektor) bekannt sind, werden die Verstärkungsfaktoren gemeinsam optimiert und vektorquantisiert 15.

Beim Sprachdecoder nach Figur 4 befindet sich anstelle des sonst vorhandenen stochastischen Codebuchs ein Anregungsgenerator 24, der die o. e. Parameter vom Sprachcoder empfängt, also die Position des ersten von Null verschiedenen Pulses der idealen RPE-Sequenz, die Positionen der überlebenden Pulse sowie die Vorzeichen der überlebenden Pulse. Aus diesen Parametern wird der stochastische Anregungsvektor c gebildet, der nach Verstärkung dem Synthesefilter 21 zugeführt wird.

Die vom Decoder sonst durchzuführenden Verarbeitungsschritte entsprechen im wesentlichen denen, die auch schon im Coder ausgeführt wurden, allerdings mit der Ausnahme, daß die zur Konstruktion der Filterkoeffizienten und der Anregung erforderlichen Codevektoren aus den verschiedenen Codebüchern aufgrund der vom Coder übermittelten Positionsangaben unmittelbar entnommen werden. Darüber hinaus erfolgt noch eine Nachverarbeitung des synthetischen Sprachsignals, das sich am Ausgang des LPC-Synthesefilters 21 ergibt. Das Nachverarbeitungsfilter 22 betont die für den Höreindruck wichtigen Regionen im Sprachsignal und hilft, Störungen, die sich durch die Kodierung selbst und durch mögliche Übertragungsfehler ergeben haben, zumindest teilweise zu verdecken. Nach abschließender D/A-Wandlung 23 steht wieder ein analoges Sprachsignal zur Verfügung.

Patentansprüche

1. Verfahren zur Synthese eines Rahmens eines Sprachsignals in einem Sprachcoder/-decoder, bei dem einem Synthesefilter (12) des Sprachcoders ein Anregungsvektor zugeführt wird, der aus einer adaptiven Anregungskomponente (a) und einer stochastischen Anregungskomponente (c) besteht, wobei die stochastische Anregungskomponente (c) durch folgende Parameter gebildet wird, die einer zuvor errechneten idealen RPE-Sequenz (Regular Pulse Excitation Sequence) entnommen werden:

   a) Die Position des ersten von Null verschiedenen Pulses in der idealen RPE-Sequenz,

   b) die Positionen einer vorgewählten Anzahl von betragsgrößten Pulsen der idealen RPE-Sequenz.

   c) die Amplituden dieser betragsgrößten Pulse, und

   d) die Vorzeichen dieser betragsgrößten Pulse,
   und wobei diese Parameter ferner zum Sprachdecoder übertragen werden, um auch dort die stochastische

Anregungskomponente (c) zu erzeugen.

2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet**, daß die Amplituden der entnommenen betragsgrößten Pulse den gleichen, beliebig wählbaren Betrag erhalten.

3. Verfahren nach Anspruch 1 oder 2, **dadurch gekennzeichnet**, daß die vorgewählte Anzahl von betragsgrößten Pulsen im Bereich von N/6 ... N/4 liegt, wobei N die Anzahl von Abtastwerten in einem Unterrahmen eines Analyserahmens ist.

4. Verfahren nach Anspruch 3, **dadurch gekennzeichnet**, daß die stochastische Anregungskomponente (c) für jeden Unterrahmen neu berechnet wird.

Fig. 1

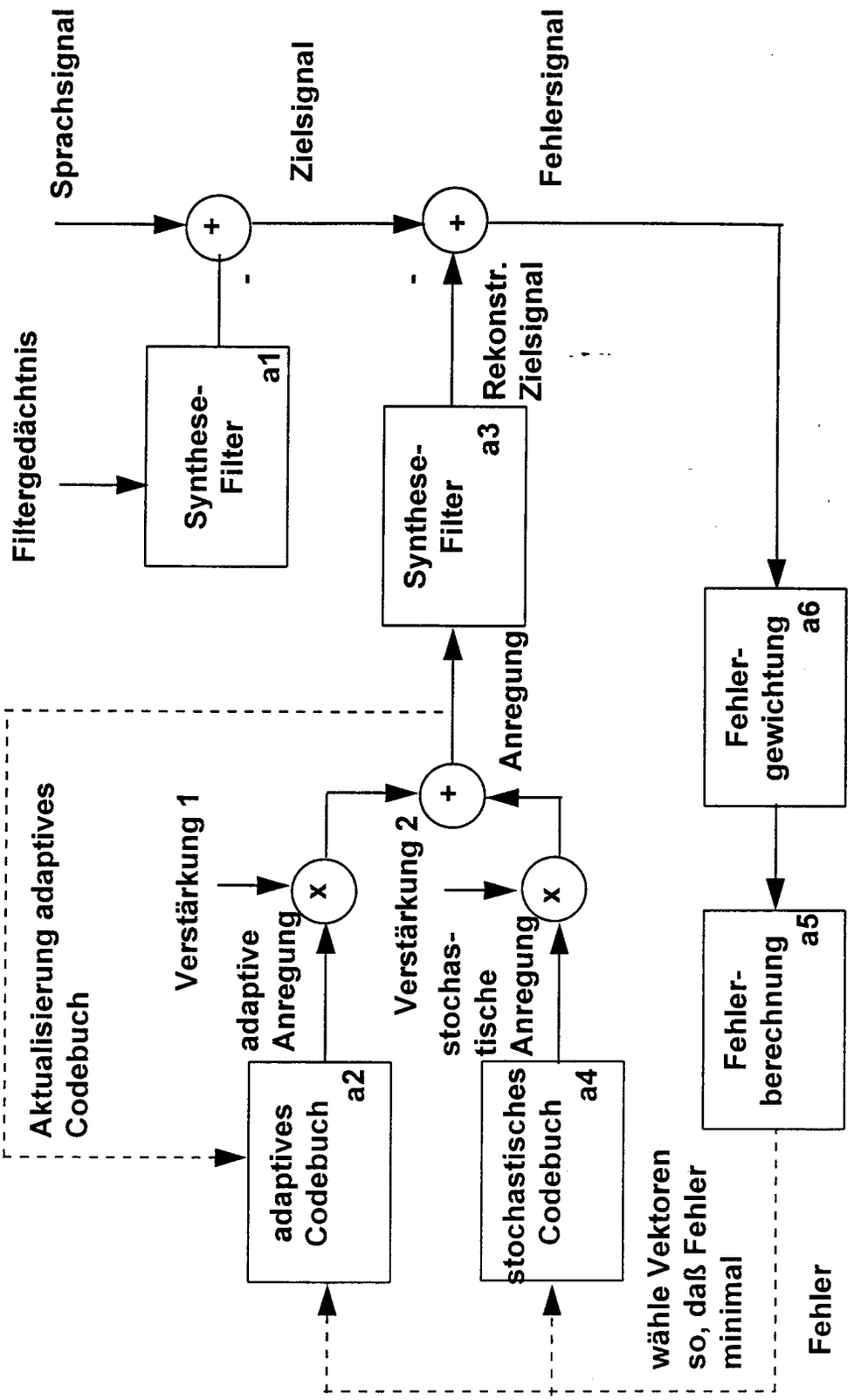

Sprachsignal
Zielsignal
Fehlersignal
Filtergedächtnis
Synthese-Filter
a1
Synthese-Filter
a3
Rekonstr. Zielsignal
Anregung
Aktualisierung adaptives Codebuch
Verstärkung 1
adaptive Anregung
Verstärkung 2
stochas-tische Anregung
adaptives Codebuch
a2
stochastisches Codebuch
a4
Fehler-gewichtung
a6
Fehler-berechnung
a5
wähle Vektoren so, daß Fehler minimal
Fehler

Fig. 2

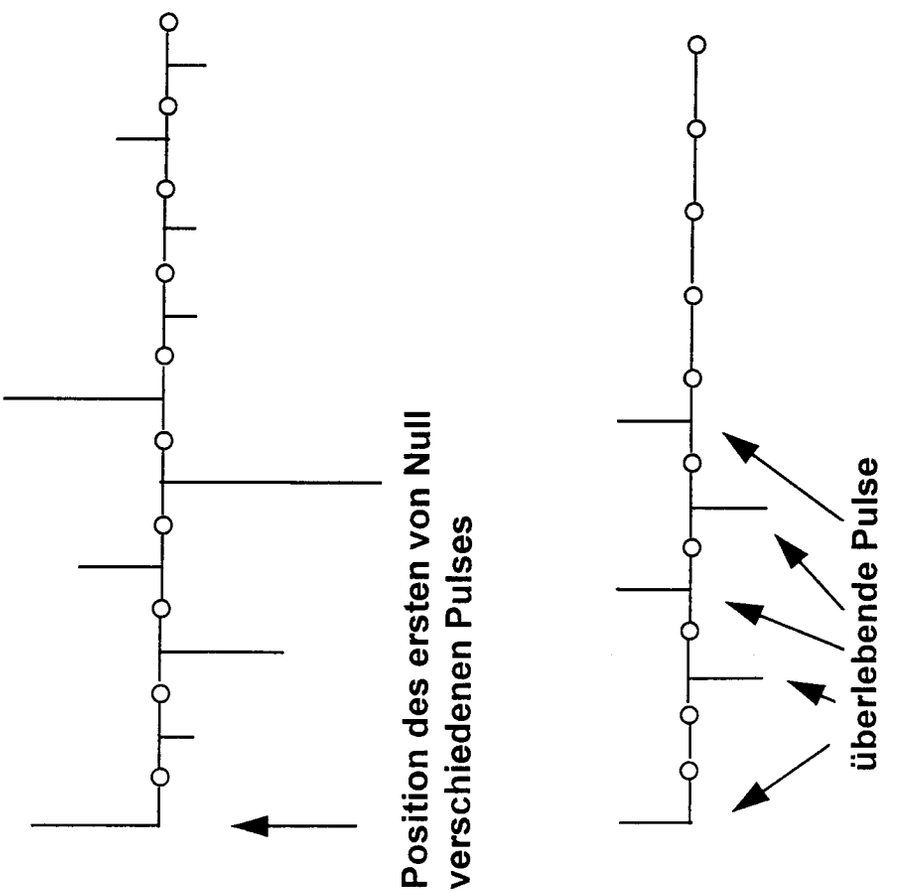

ideale RPE-Sequenz
(2a)
Position des ersten von Null verschiedenen Pulses
Anregung (=zum Decoder übertragene Sequenz)
(2b)
überlebende Pulse

Fig. 3

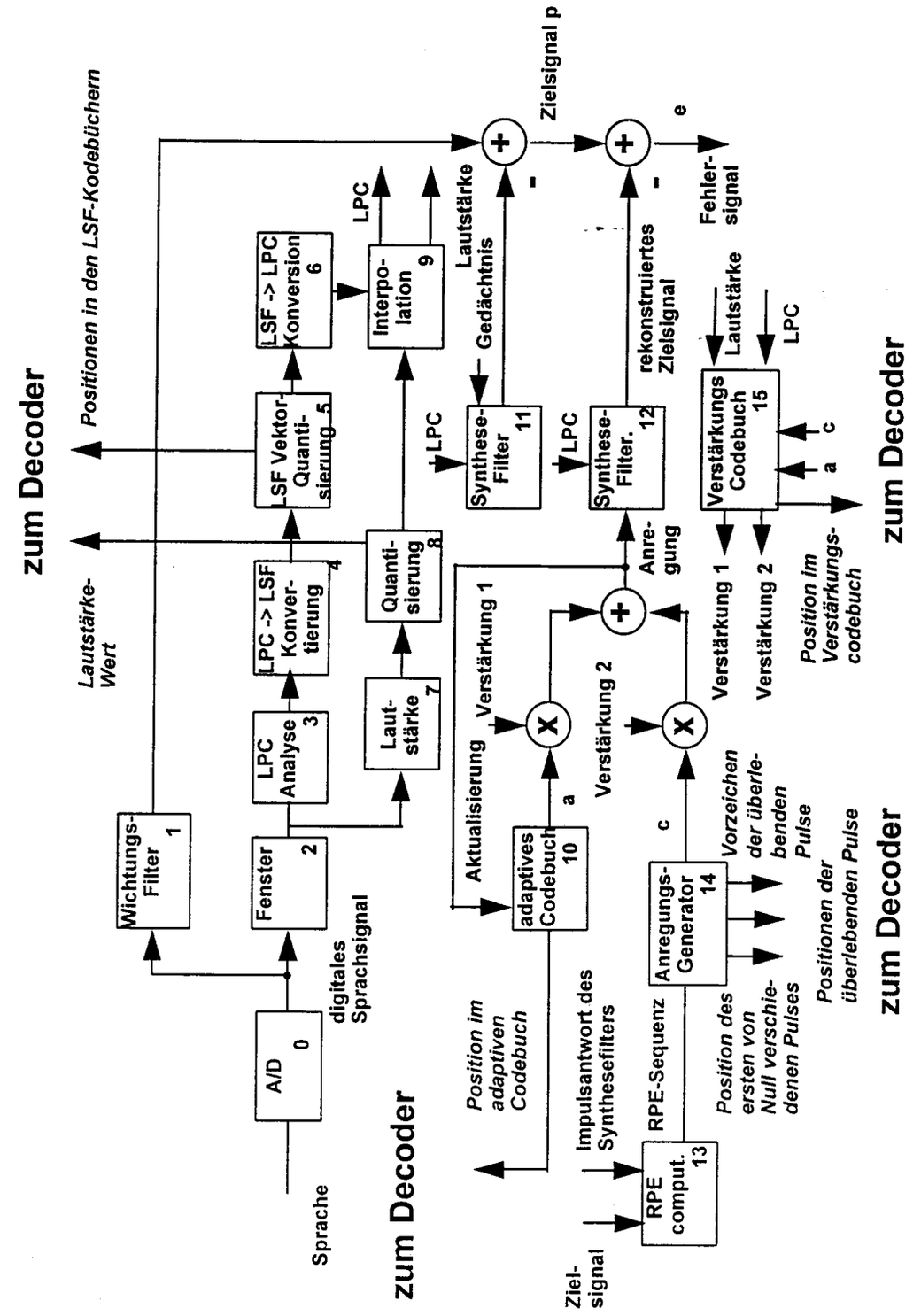
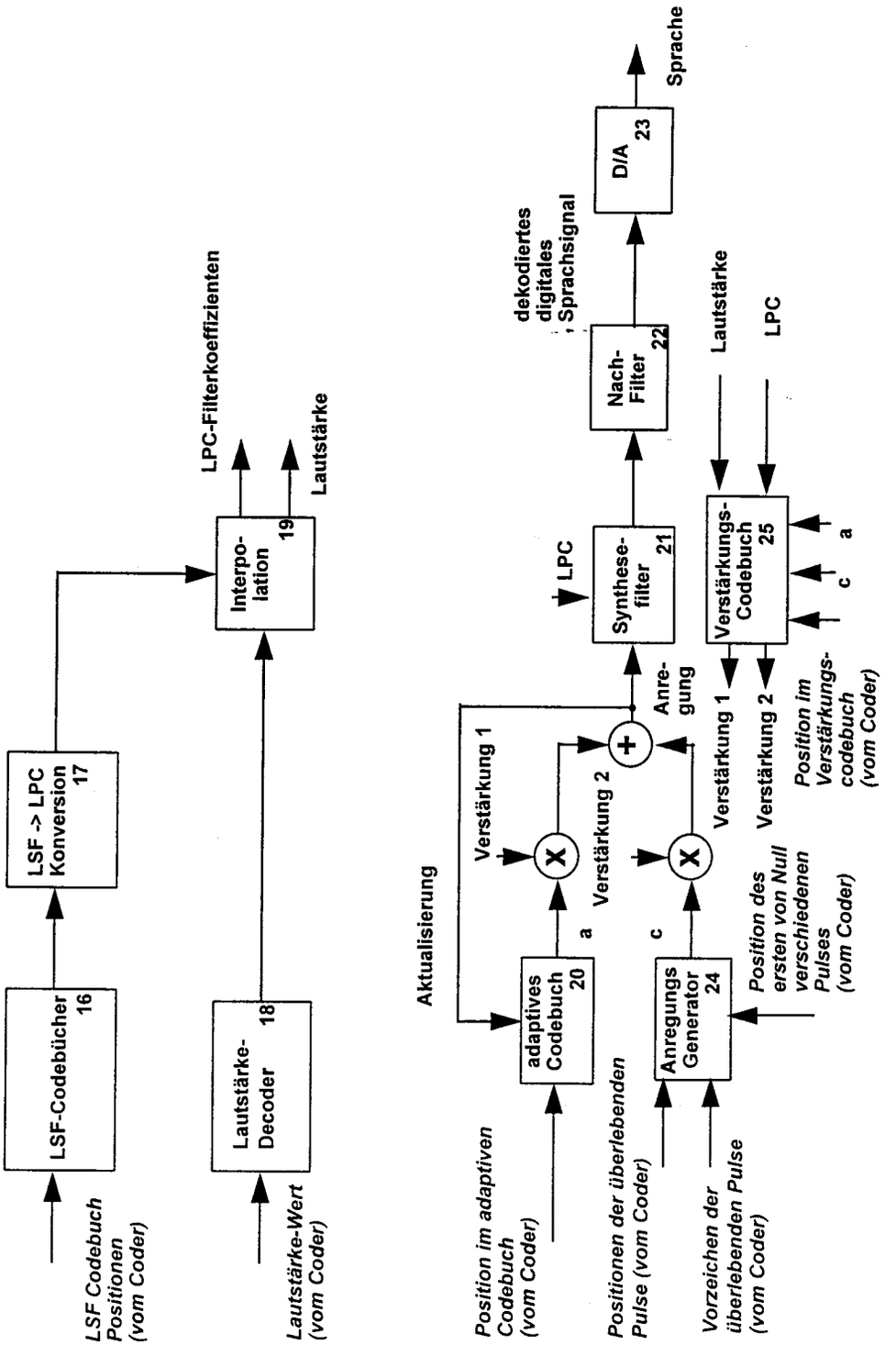

zum Decoder
Positionen in den LSF-Kodebüchern
Lautstärke-Wert
Wichtungs-Filter 1
Fenster 2
LPC Analyse 3
LPC -> LSF Konvertierung 4
LSF Vektor-Quantisierung 5
LSF -> LPC Konversion 6
Laut-stärke 7
Quantisierung 8
Interpolation 9
LPC
Lautstärke
Gedächtnis
Synthese-Filter 11
Synthese-Filter. 12
rekonstruiertes Zielsignal
Zielsignal p
Fehler-signal
e
Verstärkungs Codebuch 15
Lautstärke
LPC
a c
Position im Verstärkungs-codebuch
Verstärkung 1
Verstärkung 2
Anre-gung
Aktualisierung
adaptives Codebuch 10
a
Position im adaptiven Codebuch
Impulsantwort des Synthesefilters
RPE comput. 13
RPE-Sequenz
c
Anregungs-Generator 14
Vorzeichen der überle-benden Pulse
Position des ersten von Null verschie-denen Pulses
Positionen der überlebenden Pulse
A/D 0
Sprache
digitales Sprachsignal
Ziel-signal
zum Decoder
zum Decoder
zum Decoder

Fig. 4


LSF Codebuch Positionen (vom Coder)
LSF-Codebücher 16
LSF -> LPC Konversion 17
Lautstärke-Wert (vom Coder)
Lautstärke-Decoder 18
Interpo-lation 19
LPC-Filterkoeffizienten
Lautstärke
Position im adaptiven Codebuch (vom Coder)
adaptives Codebuch 20
Aktualisierung
Verstärkung 1
Verstärkung 2
a
Positionen der überlebenden Pulse (vom Coder)
Anregungs Generator 24
c
Vorzeichen der überlebenden Pulse (vom Coder)
Position des ersten von Null verschiedenen Pulses (vom Coder)
Anre-gung
LPC
Synthese-filter 21
Nach-Filter 22
dekodiertes digitales Sprachsignal
D/A 23
Sprache
Verstärkung 1
Verstärkung 2
Verstärkungs-Codebuch 25
Lautstärke
LPC
Position im Verstärkungs-codebuch (vom Coder)
c
a