Sprachcodec in GSM- Mobilfunknetzen

Für die Mobilfunknetze nach dem GSM-Standard stehen drei verschiedene Sprachcodecs zur Verfügung, die auf einem gemeinsamen Modell der Spracherzeugung beruhen. Mit dem künftigen AMR-Codec ist eine weitere Steigerung der Sprachgüte zu erwarten.

Von Prof. Dr.-Ing. Peter Vary

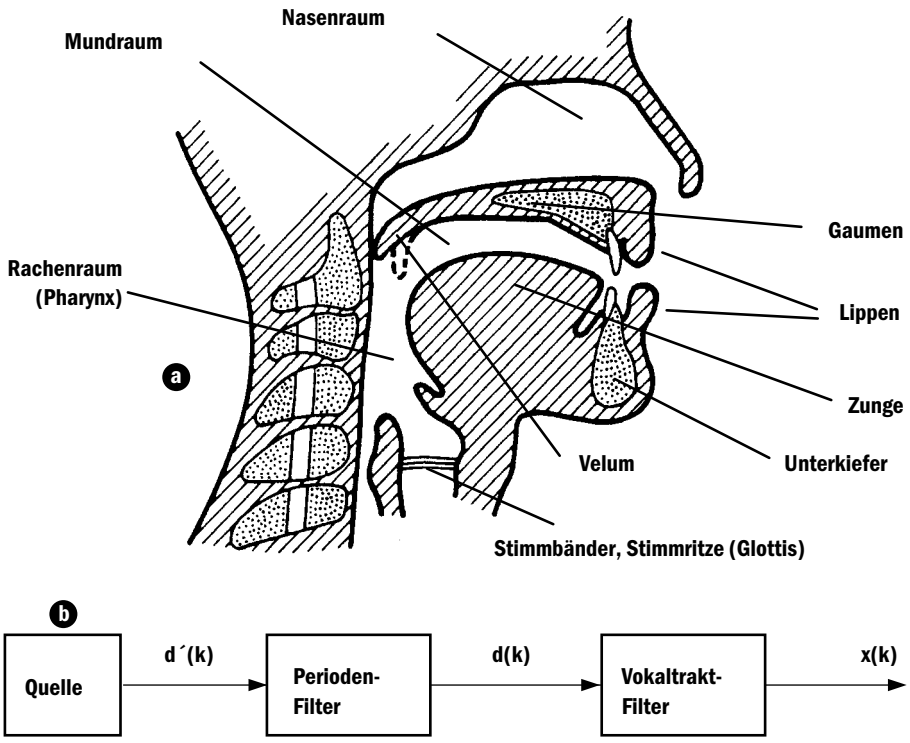
In digitalen leitungsgebundenen Telefonnetzen, wie dem ISDN, werden Sprachsignale mit einer Datenrate von 64 kbit/s übertragen. Dabei wird das Signal mit 8.000 Abtastwerten pro Sekunde und mit 8 bit pro Abtastwert beziehungsweise mit $2^8 = 256$ logarithmisch gestuften Quantisierungsniveaus dargestellt. Für die Übertragung in den digitalen D- und E-Mobilfunknetzen nach den GSM-900- und -1800-Standards steht dagegen aus Gründen der Frequenzknappheit nur eine Netto-Datenrate von maximal 13,0 kbit/s zur Verfügung. Wegen der Beeinträchtigung der Funkübertragung durch Störungen, die sich empfangsseitig als Bitfehler bemerkbar machen, werden sendeseitig durch Kanalcodierung redundante Bits hinzugefügt, die den Empfänger in die Lage versetzen, Bitfehler zu erkennen und zu korrigieren. Die Brutto-Datenrate für Sprach- und Kanalcodierung beträgt insgesamt 22,8 kbit/s für den Vollraten-Kanal und 11,4 kbit/s für den Halbraten-Kanal. Bisher stehen zwei Codecs für den Vollraten-Kanal und ein Codec für den Halbraten-Kanal zur Verfügung. An der Standardisierung eines weiteren Codecs, des sogenannten Adaptive Multirate Codecs (AMR-Codec), der dynamisch an die jeweils verfügbare Datenrate und die momentane Kanalqualität angepaßt werden kann, wird gearbeitet.

Modellgestützte prädiktive Codierung

Zur Sprachcodierung mit Datenraten von weniger als 16 kbit/s, das heißt weniger als effektiv 2 bit pro Abtastwert, haben sich in nahezu sämtlichen internationalen Telekommunikationsstandards Verfahren durchgesetzt, die auf einem stark vereinfachenden Quelle-Filter-Modell der Spracherzeugung beruhen. Dabei werden in unterschiedlicher Weise auch Eigenschaften des Gehörs, insbesondere der spektrale Verdeckungseffekt, ausgenutzt.

Bild 1a zeigt eine schematische Übersicht des menschlichen Sprechtraktes. In erster Näherung erfolgt die Bildung des Sprachsignals in den zwei Stufen Signalerzeugung (Anregung) und Signalformung (Filterung). Unter Signalerzeugung versteht man den Vorgang, bei dem im Luftstrom, der den Sprechtrakt durchströmt, Schwingungen oder Geräusche erzeugt werden, die nach weiterer Formung als Schall abgestrahlt werden. An der Signalerzeugung ist maßgeblich der Kehlkopf mit den Stimmbändern beteiligt. Die Signalformung erfolgt weitgehend im Rachenraum oberhalb des Kehlkopfes und im Mundraum, die zusammen den Vokaltrakt bilden. Bei Nasalen und nasalierten Vokalen trägt auch der Nasenraum zur Signalformung bei.

Mit der akustischen Theorie der Vokalartikulation, die den Vokaltrakt als schallhartes akustisches Rohr variablen Querschnitts beschreibt, läßt sich das zeitdiskrete Modell nach Bild 1b ableiten. Dabei wird der k-te Abtastwert x(k) des Sprachsignals aus den Abtastwerten d'(k) eines spektral flachen Anregungssignal durch zweistufige Filterung erzeugt. Das erste Filter, das in Bild 1b als Perioden-Filter bezeichnet wird, ist in stimmhaften Sprachabschnitten für die Ausformung der Grundperiode und der harmonischen spektralen Struktur verantwortlich, während das Vokaltrakt-Filter die spektrale Einhüllende bestimmt. Beide Filter sind zeitveränderlich, ihre Parameter lassen sich aus den Abtastwerten des Sprachsignals bestimmen. Bild 2 zeigt im Sinne einer Momentaufnahme das Betragsspektrum eines



alle Bilder: P.Vary

Bild 1: Modell der Spracherzeugung: a) menschlicher Sprechtrakt
b) zeitdiskretes Quelle-Filter-Modell

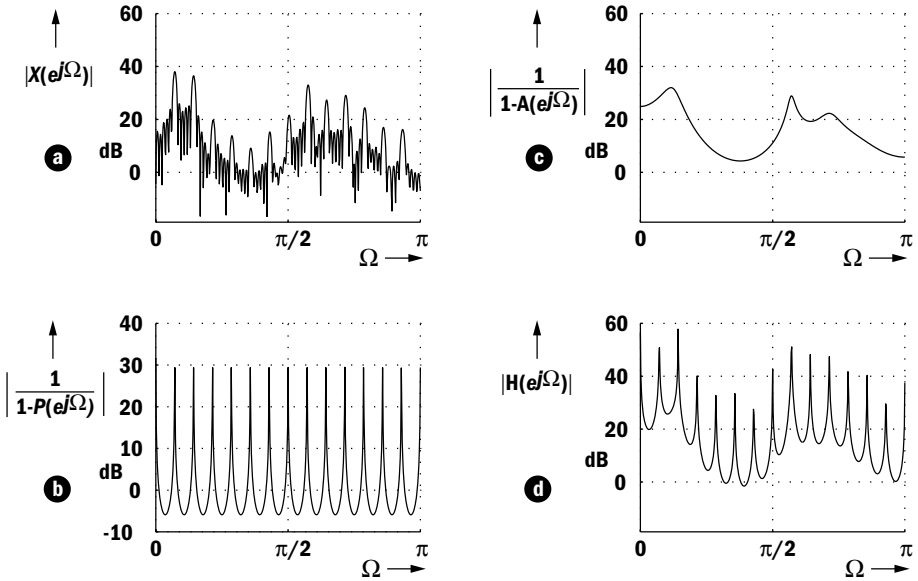


Bild 2: Zur Funktionsweise des Modells der Spracherzeugung
a) Betragsspektrum eines stimmhaften Signalabschnitts (20 ms) und Betragsfrequenzgänge für ($e^{j\Omega}$)
b) Perioden-Filter
c) Vokaltrakt-Filter
d) Gesamtfilter

stimmhaften Sprachabschnitts sowie die entsprechenden Betragsfrequenzgänge des Perioden-Filters, des Vokaltrakt-Filters und der Reihenschaltung der beiden Teilfilter. Die Frequenzachse wurde wie folgt auf die Abtastfrequenz f_A normiert:

$$\Omega = \frac{2\pi}{f_A} f$$

Die beiden Filter lassen sich durch die z-Transformierten P(z) und A(z) beschreiben. Für das Gesamtfilter gilt

$$H(z) = \frac{1}{1-P(z)} \cdot \frac{1}{1-A(z)}$$

Das Beispiel verdeutlicht, daß das momentane Signalspektrum aus Bild 2c relativ gut durch den Gesamtfrequenzgang nach Bild 2d nachgebildet werden kann.

Grundstruktur der Sprachdecoder

Die drei standardisierten Codecs beruhen gleichermaßen auf dem skizzierten Modell der Spracherzeugung. Die empfangsseitige Decodierung läßt sich deshalb einheitlich durch die Anordnung nach Bild 3 beschreiben. In dieser sich auf die prinzipiellen Eigenschaften konzentrierenden Sichtweise bestehen die wesentlichen Unterschiede, wie in Tabelle 1 angegeben, in den Datenraten und in den Aufteilungen der Datenraten auf die einzelnen Komponenten.

Die Decodierung läßt sich in allen drei Fällen auf das Modell aus Bild 1b zurückführen. Zusätzlich ist ein weiteres Filter mit der Übertragungsfunktion N(z) vorgesehen, das in unterschiedlicher Weise als Nachfilter (engl. Post-Filter) zur Verbesserung der auditiven Sprachqualität beiträgt. Trotz der gemeinsamen Grundstruktur bestehen zwischen den Decodieralgorithmen deutliche Unterschiede, so daß die Decoder zueinander nicht kompatibel sind.

Aus der Tabelle 1 ist in Verbindung mit Bild 3 ersichtlich, daß der Vollraten-Codec, der bisher in den D- und E-Netzen ausschließlich eingesetzt wird, eine Brutto-Datenrate von A = 22,8 kbit/s benötigt. Davon entfallen ein Anteil von B+C = 13,0 kbit/s auf die eigentliche Sprachcodierung und ein Anteil von 9,8 kbit/s (A - [B+C]) auf die Kanalcodierung. Die Netto-Bitrate von 13,0 kbit/s enthält 9,4 kbit/s für die Anregung (Anteil B) und 3,6 kbit/s für die Koeffizienten des zweistufigen Synthesefilters (Anteil C).

Der verbesserte Vollraten-Codec (EFR) wurde kürzlich für die neuen amerikanischen PCN-1900 Netze entwickelt. Bei entsprechender Erweiterung der Infrastruktur und der Endgeräte könnte er auch in den bereits etablierten Netzen eingesetzt werden. Im Vergleich zum Vollraten-Codec (FR) werden mit C = 4,2 kbit/s eine etwas höhere Datenrate zur Übertragung der Filterkoeffizienten und mit B = 8,0 kbit/s eine niedrigere Rate zur Codierung des Anregungssignals d'(k) aufgewendet. Die spektrale Information, repräsentiert durch P(z) und A(z), kann deshalb genauer dargestellt werden. Gleichzeitig wird auch die Anregungsinformation d'(k) besser quantisiert. Dies gelingt durch indirekte Vektorquantisierung des Restsignals, die allerdings einen deutlich erhöhten Rechen- und Speicheraufwand erfordert. Es wird die gleiche Kanalcodierung wie beim FR-Codec benutzt, was sich vorteilhaft in bezug auf die Netzinfrastruktur auswirkt. Weiterhin ist in den 13,0 kbit/s noch ein Anteil von D = 0,8 kbit/s für zusätzliche Paritätsbits enthalten.

Der Halbraten-Codec weist die halbe Brutto-Datenrate von A = 11,4 kbit/s auf. In stimmhaften Abschnitten werden die Filterparameter mit B = 3,15 kbit/s und in stimmlosen Abschnitten mit nur 1,65 kbit/s übertragen, da in den stimmlosen, das heißt nicht periodischen Abschnitten auf das Perioden-Filter verzichtet wird. Dementsprechend wird die Anregungsinformation mit B = 2,3 kbit/s beziehungsweise 3,8 kbit/s codiert. Die Quantisierung erfolgt, ähnlich wie beim EFR-Codec, vektoriell. Der Vergleich der Bitraten-Beiträge des Halbraten-Codecs mit denen der beiden Vollraten-Codecs zeigt, daß die Reduktion auf 5,6 kbit/s im wesentlichen durch eine Reduktion des Bitratenanteils B für das Anregungssignal erreicht wird.

Die sendeseitigen Komponenten dieser drei Codecs weisen größere strukturelle Unterschiede auf als die empfangsseitigen Decoder. Der FR-Codec arbeitet nach dem Prinzip der linear-prädiktiven Restsignal-Codierung (RELP = Residual Excited Linear Prediction), während der EFR- und der HR-Codec auf dem sogenannten CELP-Prinzip (Code Excited Linear Prediction) beruhen.

Grundlage der FR-Codierung: RELP

Beim RELP-Codec, dessen Grundprinzip Bild 4 a zeigt, werden zunächst für jeden Signalabschnitt, bestehend aus beispielsweise N = 160 Abtastwerten (20 ms), die Koeffizienten der beiden Filter A(z) und P(z) berechnet,

VERGLEICH

Bitraten

Codec	Bitraten in kbit/s				
	A	B	C	D	B+C+D
Vollraten-Codec: FR	22,8	9,4	3,6	—	13,0
Verbesserter Vollraten-Codec: EFR	22,8	8,0	4,2	0,8	13,0
Halbraten-Codec: HR	11,4	2,3*/3,8	3,15*/1,65	0,15	5,6

Tabelle 1: Bitraten der drei standardisierten Codecs: FR Fullrate Codec, EFR Enhanced Fullrate Codec und HR Halfrate Codec *bei stimmhaften Abschnitten

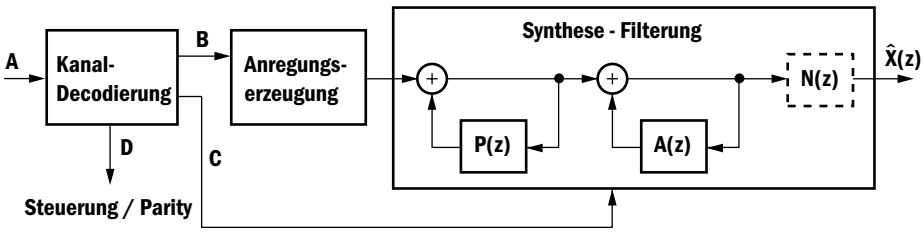


Bild 3: Struktur der empfangsseitigen Decodierung (siehe auch Tabelle 1)
(A: Brutto-Datenrate; B: Datenrate für die Anregung; C: Datenrate zur Einstellung des zweistufigen Synthesefilters)

wobei P(z) in der Regel jeweils nach 40 Werten (5 ms) aktualisiert wird. Die Koeffizienten werden als sogenannte Nebeninformation zum Empfänger übertragen. Anschließend wird das Anregungssignal in fünf Schritten gewonnen:

- Kurzzeitprädiktion bzw. adaptive Filterung von x(k) mit 1 - A(z)
- Langzeitprädiktion bzw. adaptive Filterung von d(k) mit 1 - P(z)
- Tiefpaßfilterung
- Taktreduktion um den Faktor r (r = 3 oder r = 4)
- skalare Quantisierung.

In der ersten Stufe wird durch Lineare Prädiktion der aktuelle Abtastwert x(k) aus n = 8...10 vorhergehenden Abtastwerten geschätzt entsprechend

$$\hat{x}(k) = \sum_{i=1}^{n} a_i \cdot x(k-1)$$

und das Prädiktions-Fehlersignal

$$d(k) = x(k) - \hat{x}(k)$$

gebildet. Diese Schätzung wird auch als Kurzzeitprädiktion bezeichnet. Zur Berechnung der Prädiktorkoeffizienten a_i (i=1,2..n) ist alle 20 ms ein Gleichungssystem der Dimension n · n zu lösen, das die n + 1 ersten Werte der Autokorrelationsfunktion des aktuellen Signalsegments enthält. Es läßt sich zeigen, daß durch diese erste Filterung mit der Übertragungsfunktion 1 - A(z) die Wirkung des akustischen Vokaltraktes im Rahmen der vereinfachten Modellierung nach Bild 1 aufgehoben wird und so in erster Näherung das Eingangssignal d(k) des Vokaltraktfilters nach Bild 1b zurückgewonnen wird. Dieses Signal d(k) zeichnet sich durch eine stark reduzierte Dynamik aus (Bild 4b) und weist in stimmhaften Sprachabschnitten impulsförmige Signalverläufe im Abstand der momentanen Grundperiode auf. Deshalb kann mit einer zweiten Filterstufe, die eine gesamte Grundperiode gemäß

$$d'(k) = d(k) - \hat{d}(k) = d(k) - b \cdot d(k - N_0)$$

berücksichtigt, eine weitere Dynamikreduktion erzielt werden. Diese zweite Filterstufe liefert im Rahmen des stark vereinfachenden Modells aus Bild 1b aufgrund ihrer Übertragungsfunktion 1 - P(z) das Anregungssignal d'(k) des Sprechtraktmodells. Die Größe N_0 entspricht der momentanen Sprachgrundperiode $T_0 = N_0 \cdot T$, wobei $T = 1/f_A$ die Dauer des Abtastintervalls bezeichnet. Da N_0 sehr viel größere Werte als n annimmt und dieser zweite Prädiktor somit ein größeres Zeitintervall „überblickt", wird er in der Literatur allgemein als Langzeitprädiktor (engl. Long Term Prediction, LTP) bezeichnet. Das Ergebnis der zweistufigen Analyse-Filterung ist ein spektral flaches Rest- bzw. Anregungssignal d'(k), das nach Tiefpaßfilterung und Taktreduktion um den Faktor r skalar quantisiert wird. Aufgrund der Taktreduktion läßt sich die reduzierte Anzahl der beibehaltenen Abtastwerte mit einer Rate von etwa 10 kbit/s genügend genau darstellen. Allerdings gehen durch die Tiefpaßfilterung die spektralen Anteile von d'(k) oberhalb von $f_g = f_A/2r$ verloren. Sie werden auf der Empfangsseite durch (r-1)-fache Spiegelung der Anteile aus dem Bereich $0 \le f \le f_g$ in das Intervall $f_g \le f \le f_A/2$ ersetzt. Die Spiegelung erfolgt, indem man die bei der Unterabtastung verlorengegangenen Abtastwerte durch Nullwerte ersetzt. Es entsteht wiederum ein breitbandiges Anregungssignal mit flachem Spektrum, das dem zweistufigen Synthesefilter zugeführt wird.

Wegen der Tiefpaßfilterung, die im FR-Codec mit r = 3 eine Grenzfrequenz von f_g = 1,33 kHz besitzt, ist dieser Codec-Typ nicht für Modem- und Musiksignale geeignet. Im FR-Codec wird sendeseitig die zweite Prädiktion in modifizierter Form als Rückwärtsprädiktion (engl. Closed Loop Prediction) ausgeführt und die Unterabtastung erfolgt auf einem vom momentanen Signalabschnitt abhängigen Raster (Regular Pulse Excitation Grid, RPE-Grid).

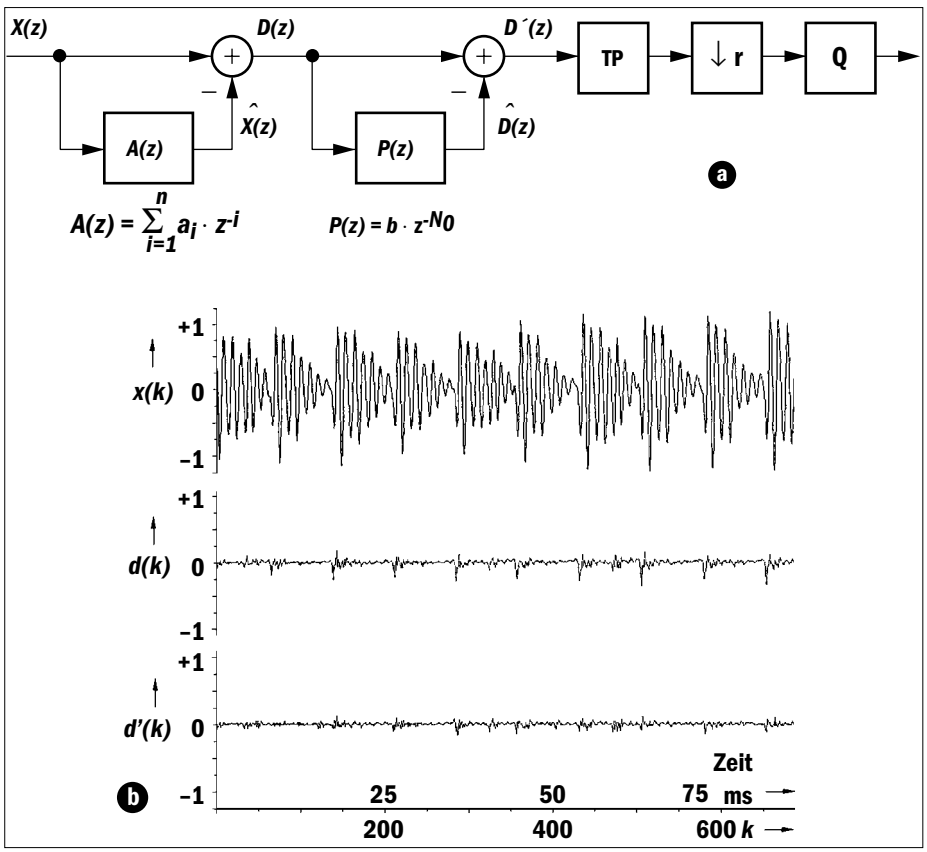


Bild 4: Prinzip der Restsignalcodierung mit Kurzzeit- und Langzeitprädiktion
a) Blockschaltbild b) Signalbeispiel für einen stimmhaften Laut

Grundlage der EFR- & HR-Codierung: CELP

Im Vergleich zum RELP-Algorithmus, der das Sprachsignal x(k) analysiert und filtert, um die Abtastwerte des Rest- beziehungsweise des Anregungssignals d'(k) zu gewinnen, geht man beim CELP-Ansatz (Code Excited Linear Prediction) umgekehrt vor. Es handelt sich um ein Analyse-durch-Synthese-Verfahren mit vektorieller Quantisierung des Anregungssignals.

Der sendeseitige Codierer enthält, wie Bild 5 zeigt, einen vollständigen Decoder mit zweistufigem Synthesefilter. Analyse-durch-Synthese bedeutet, daß ein kurzer Abschnitt des Sprachsignals x(k), bestehend aus zum Beispiel L = 40 Abtastwerten, mit unterschiedlichen Versionen synthetisierter Abschnitte $\hat{x}(k)$ verglichen wird. Der Vergleich erfolgt durch Differenzbildung x(k) - $\hat{x}(k)$, spektrale Gewichtung mit einem (meist festen) Filter W(z) und Berechnung der resultierenden Fehlerenergie gemäß

$$E = \sum_{k=1}^{L} e^2(k)$$

Die Abtastwerte $\hat{x}(k)$ werden durch Filterung von gespeicherten Anregungsfolgen erzeugt, die in einem Vektorcodebuch abgelegt sind. Dieses Codebuch enthält beispielsweise K = 1.024 typische Folgen (Vektoren $\vec{c}_i$, i = 0, 1,..., 1.023) der Länge L = 40, die aus einer Rauschfolge mit Gaußverteilung gewonnen wurden. Bei der angegebenen Dimensionierung sind versuchsweise alle 1.024 Vektoren mit einem jeweils zu optimierenden Verstärkungsfaktor g zu multiplizieren und anschließend zu filtern, um 1.024 Varianten $\hat{x}$ (k) zu erzeugen. Der Vektor mit der geringsten Fehlerenergie E wir schließlich ausgewählt. Zum Empfänger wird nicht der Vektor $\vec{c}_i$, sondern nur dessen Adresse übermittelt, die in diesem Fall mit 10 bit darzustellen ist.

Der Empfänger verfügt über das gleiche Vektorcodebuch, das Bestandteil des Codec-Standards ist. Er wählt den entsprechenden Vektor aus, um einen kurzen Signalabschnitt durch Filterung zu synthetisieren. Durch diese indirekte vektorielle Quantisierung gelingt es hier, L = 40 Anregungswerte d'(k) mit 10 bit, das heißt mit effektiv nur 1/4 bit pro Wert zu codieren. Hinzukommen vier bis sechs bits für den Verstärkungsfaktor g.

Neben der Anregungsinformation werden natürlich auch die Koeffizienten des zweistufigen Synthesefilters (Anteil C in Bild 3) zum Empfänger übertragen. Das Fehlergewichtungsfilter mit der Übertragungsfunktion W(z) wird aus A(z) abgeleitet und bewirkt, daß bei der Auswahl des besten Vektors der psychoakustische Effekt der Fehlermaskierung bis zu einem gewissen Grade berücksichtigt wird.

Dieses Verfahren ist extrem rechenintensiv, da zur Codierung eines Signalabschnitts x(k) der Länge L sendeseitig K Signalabschnitte $\hat{x}$(k) zu synthetisieren sind.

In der Literatur findet man zahlreiche Vorschläge, den Rechenaufwand unter anderem durch spezielle Wahl des Codebuchs stark zu reduzieren. In dieser Hinsicht existieren große Unterschiede zwischen den verschiedenen Standards.

Im HR- und EFR-Codec wird das „Analyse-durch-Synthese-Prinzip" noch dahingehend erweitert, daß auch die beiden Parameter b und N_0 der ersten Filterstufe in der Syntheseschleife variiert werden, um die beste Einstellung zu finden. Diese Vorgehensweise wird als Closed-Loop Long-Term Prediction bezeichnet. Eine ausführlichere Darstellung würde den Rahmen dieses Artikels sprengen.

Qualitätsvergleich

Zur objektiven beziehungsweise instrumentellen Beurteilung der Sprachgüte von Sprachcodecs existiert bisher noch kein allgemeingültiges Verfahren, obwohl in jüngster Zeit auf diesem Gebiet große Fortschritte gemacht wurden (vgl. Funkschau 3/98). Im Rahmen der Standardisierung werden daher nach wie vor aufwendige auditive Tests in unterschiedlichen Sprachen durchgeführt.

Die Sprachqualitäten und die Klangcharakteristika der drei Mobilfunk-Codecs sind in gewissen Grenzen unterschiedlich. Diese Unterschiede sind bei manchen Stimmen deutlich, bei anderen weniger deutlich bemerkbar. In allen drei Fällen handelt es sich um modellgestützte Codierverfahren, die für die Codierung von Sprachsignalen optimiert wurden. Zur Übertragung von Musiksignalen sind sie alle wegen stark hörbaren Verzerrungen nur eingeschränkt geeignet.

Der FR-Codec erreicht nicht die Qualität der ISDN-Übertragung. Wegen des zugrundeliegenden RELP-Prinzips mit variabler Unterabtastung und spektraler Spiegelung des tiefpaßgefilterten Restsignals ergibt sich insbesondere bei Frauen- und Kinderstimmen eine Rauhigkeit. Auch bei der Übertragung von Hintergrundgeräuschen besteht keine volle Transparenz.

Demgegenüber bietet der EFR-Codec für Sprachsignale eine deutliche Qualitätssteigerung. Er klingt natürlicher und „glatter". ISDN-Qualität erreicht aber auch er nicht. Hintergrundgeräusche klingen natürlicher als beim FR-Codec.

Der Half-Rate-Codec liefert im Vergleich zum Full-Rate-Codec bei ungestörter Übertragung eine etwas geringere Sprachqualität. Unter dem Einfluß funkspezifischer Störungen verhält er sich jedoch wegen des relativ stärkeren Fehlerschutzes etwas robuster. Die Transparenz bezüglich der Hintergrundgeräusche ist nochmals etwas reduziert.

Ausblick: AMR-Codec

Die Sprachqualität der drei Codec wird maßgeblich durch die jeweils zur Verfügung stehende Netto-Datenrate bestimmt. Beim Codec-Entwurf wurde diese in Verbindung mit der Kanalcodierung im Sinne eines Kompromisses festgelegt, der die ungünstigsten Störverhältnisse auf dem Funkweg berücksichtigt. Die Sprachqualität ist dadurch auch bei guten Übertragungsbedingungen zwangsläufig schlechter als im Festnetz. Dieser Nachteil soll mit einem neuen Sprachcodec, dem sogenannten Adaptive Multirate Codec umgangen werden, der zur Zeit unter der Federführung der ETSI-Arbeitsgruppe SMG11 entwickelt und voraussichtlich bis Anfang 1999 standardisiert wird. Dieser Codec wird über mehrere Datenraten verfügen und die insgesamt verfügbare Datenrate von 11,4 kbit/s (HR-Kanal) oder 22,8 kbit/s (FR-Kanal) in Abhängigkeit von der momentanen funktechnischen Versorgungsqualität variabel auf die Quellen- und die Kanalcodierung aufteilen. Dadurch soll bei guter Funkversorgung eine deutlich höhere Sprachqualität und bei schlechter Funkverbindung eine robustere Übertragung als in den heutigen Netzen erreicht werden. Zusätzlich wird ein sogenannter Breitband-Modus mit einer analogen Bandbreite des Sprachsignals von 7 kHz vorgesehen. Der AMR-Codec soll auch als Basis-Codec in den künftigen UMTS-Mobilfunkstandard (Universal Mobile Telecommunication System) übernommen werden. *(GG)*

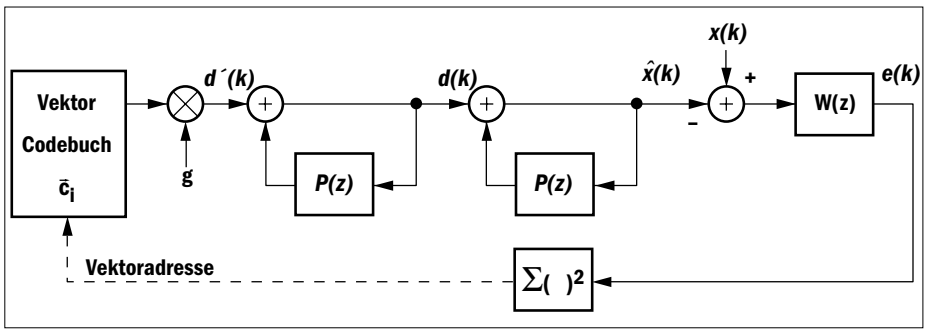


Bild 5: Prinzip des CELP-Codierers

■ Literatur: P. Vary, U.Heute, W. Hess: „Digitale Sprachsignalverarbeitung", Teubner Verlag, 1998